C

Caching

▸ OLAP Results, Distributed Caching

CAD and GIS Platforms

▸ Computer Environments for GIS and CAD

Cadaster

▸ Cadastre

Cadastre

Erik Stubkjær
Department of Development and Planning,
Aalborg University, Aalborg, Denmark

Synonyms

Cadaster; Land administration system; Land information system; Land policy; Land registry; Property register; Spatial reference frames

Definition

A cadastre may be defined as an official geographic information system (GIS) which identifies geographical objects within a country, or more precisely, within a jurisdiction. Just like a land registry, it records attributes concerning pieces of land, but while the recordings of a land registry is based on deeds of conveyance and other rights in land, the cadastre is based on measurements and other renderings of the location, size, and value of units of property. The cadastre and the land registry in some countries, e.g., the Netherlands and New Zealand, are managed within the same governmental organization. From the 1990s, the term "land administration system" came into use, referring to a vision of a complete and consistent national information system, comprising the cadastre and the land registry.

The above definition of cadastre accommodates to the various practices within continental Europe, the British Commonwealth, and elsewhere. Scientific scrutiny emerged from the 1970s, where the notions of a land information system or property register provided a frame for comparison of cadastres across countries. However, GIS and sciences emerged as the main approach of research in the more technical aspects of the cadastre. In fact, the

S. Shekhar et al. (eds.), *Encyclopedia of GIS*,
DOI 10.1007/978-3-319-17885-1

above definitional exercise largely disregards the organizational, legal, and other social science aspects of the cadastre. These are more adequately addressed when the cadastre is conceived of as a sociotechnical system, comprising technical, intentional, and social elements.

Historical Background

The notion of the cadastre has been related to Byzantine ledgers, called *katastichon* in Greek, literally "line by line". A Roman law of 111 BC required that land surveyors (*agrimensores*) should complete maps and registers of certain tracts of Italy. Also, an archive, a *tabularium*, was established in Rome for the deposit of the documents. Unfortunately, no remains of the *tabularium* seem to have survived the end of the Roman Empire.

The Cadastre reemerged in the fifteenth century in some Italian principalities as a means of recording tax liabilities. This seems part of a more general trend of systematic recording of assets and liabilities, e.g., through double-entry bookkeeping, and spread to other parts of Europe. In order to compensate for the lack of mapping skills, landed assets and their boundaries were described through a kind of written maps or *cartes parlantes*. During the sixteenth century, landscape paintings or so called "picture maps" were prepared e.g., for the court in Speyer, Germany, for clarifying argumentation on disputed land. During the same period in the Netherlands, the need for dike protection from the sea called for measurements and the organization of work and society; the practice of commissioning surveys for tax collection became increasingly common there.

A new phase was marked by new technology, the plane table with the related methods for distance measurement, mapping, and area calculation. The technology was introduced in 1720 in Austrian Lombardy through a formal trial against alternative mapping methods. The resulting Milanese census, the *Censimento*, with its integrated recordings in ledgers and maps became a model for other European principalities and kingdoms.

The uneven diffusion of cadastral technology reveals a power struggle between the ruling elite and the landed gentry and clerics, who insisted on their tax exemption privileges. In the early modern states, the cadastre was motivated by reference to a "God-given principle of equality" (German: *gottgefällige Gerechtigkeit'* or *gottgefällige Gleichheit* (Kain and Baigent 1993)). Generally, absolutist monarchs were eager to establish accounts of the assets of their realm, as the basis for decisions concerning their use in wars and for the general benefit of the realm. A continental European version of mercantilism, "cameralism", was lectured at universities, seeking a quasirational exploitation of assets and fair taxation, for which the cadastre was needed, as well as regulations and educational programs, for example in agriculture, forestry, and mining. Cadastral technology, the related professions, and the centralized administration together became an instrument of unification of the country, providing the technical rationale for greater equality in taxation. Taxation thus gradually became controlled by the central authority, rather than mediated through local magnates. This change was recognized by Adam Smith, by physiocrats, and by political writers in France. The administrative technology was complemented by codification, that is, a systematic rewriting of laws and by-laws that paved the way for the modern state where individual citizens are facing the state, basically on equal terms.

The reorganization of institutions after the French revolution of 1789 also changed the role of the cadastre as introduced by Enlightenment monarchs. In part, this was due to university reforms, e.g., by Wilhelm von Humboldt in Berlin. Cameralism was split into economics, which increasingly became a mathematically based discipline and a variety of disciplines lectured at agricultural and technical universities. Cadastral expertise was largely considered a subfield of geodesy, the new and rational way of measuring and recording the surface of the Earth. From the end of eighteenth century, cadastral maps were increasingly related to or based on geodetic

triangulations, as was the case for the Napoleonic cadastre of France, Belgium and the Netherlands.

The same cadastral reform included the intention of using the measured boundaries and areas and the parcel identification for legal purposes, primarily by letting the cadastral documentation with its "fixed boundaries" prove title to land and become the final arbiter in case of boundary disputes. However, the courts that were in charge of land registries, generally for a more than a century, and the legal profession were reluctant to adopt what might be considered an encroachment on their professional territory. During the nineteenth century, most countries improved their deeds recording system, and German-speaking countries managed to develop them into title systems backed by legal guaranties. Similarly, from South Australia the so-called Torrens system, adopted in 1858, influenced the English-speaking world. However, with few exceptions, the integration of cadastral and legal affairs into one information system had to await the introduction of computer technology and the adoption of business approaches in government.

The above historical account is Eurocentric and bypasses possible scientific exchange with South-Eastern neighbors during the twelfth to fifteenth centuries. Also, it leaves out the development in England and its colonies worldwide. The account describes how the notion of the cadastre emerged and varied across time and place. This calls for special care in scientific communications, since no standardized and theory-based terminology has been established.

Scientific Fundamentals

Spatial Reference Frames

The center and rotational axis of the Earth, together with the Greenwich meridian, provide a reference frame for the location of objects on the surface of the Earth. Furthermore, a map projection relates the three-dimensional positions to coordinates on a two-dimensional map plane. The skills of the geodesist and land surveyor are applied in the cadastral field to record agreed legal boundaries between property units, the assumption being that such recordings substantially reduce the number of boundary disputes.

Application of the above-mentioned "fixed-boundary" solution raises serious problems, even if the assumption may be largely confirmed. In the cases of land slides and land drifting due to streams, the solution is insensitive to the owner's cost of getting access to all parts of the property. Furthermore, the solution does not accommodate for later and better measurements of the boundary. Moreover, the boundary may have been shifted years ago for reasons that have become too costly to elicit, relative to the value of the disputed area and even acknowledging the fact that justice may not be served. Some jurisdictions hence allow for "adverse possession", that is: an official recognition of neighbors' agreement on the present occupational boundary, even if it differs from previous cadastral recordings. Likewise, legal emphasis on merestones and other boundary marks, as well as the recording of terrain features which determine permanent and rather well defined boundary points may supplement the pure fixed-boundary approach.

The geodetic surveyors' reference frames locate points in terms of coordinates, but the naming and identification of parcels relates to names, which is a subfield of linguistics. The cadastral identifier is a technical place name, related to the place names of towns, roads, parishes, and topographic features. Hierarchically structured administrative units or jurisdictions and their names provide a means of location of property units. Even if such ordinal structuring of a jurisdiction through place names is coarse, relative to the metric precision of measured boundary points, it provides in many cases for a sufficient localization of economic activities and it reduces the dependency of specialized and costly competence.

The linguistic approach to localization refers to another spatial reference frame than the Earth, namely the human body. The everyday expressions of "left" and "right", "up" and "down" all refer to the body of the speaker or the listener, as practiced when giving directions to tourists or explaining where to buy the best offers of the day. Important research areas include the

balancing of nominal, ordinal and metric means of localization and the consideration of relations amongst various spatial reference frames.

Communication and Standardization

The national information systems (cadastre, land registry, or whatever may be the name) and organizational structure of real property information systems depend on databases and related archives and need updating if they are to render trustworthy minutes of the spatial and legal situation of the property units. Computer and communication sciences provide the theoretical background for these structures and tasks. However, until recently the methods provided in terms of systems analysis and design, data modeling, etc. addressed the information system within a single decision body, while the situation pertaining to real property information depends on an interplay between ministries, local government, and the private sector. The notion of a geospatial data infrastructure is used to describe this scope. The modeling of this infrastructure compares to the modeling of an industrial sector for e-business, an emergent research issue that includes the development of vocabulary and ontology resources of the domain.

The specification of the property unit is a fundamental issue. Often, the unit is supposed to be in individual ownership, but state or other public ownership is common enough to deserve consideration, as are various forms of collective ownership. Collective ownership is realized by attributing rights and obligations to a so-called legal person, which may be an association, a limited company, or another social construct endorsed by legislation or custom. Comparative studies reveal that the property unit itself can be specified in a host of variations: Is the unit a single continuous piece of land, or is it defined as made up of one or more of such entities? Relations among pieces of land can be specified in other ways: In Northern and Central Europe, a construct exists where a certain share of a property unit is owned by the current owners of a number of other property units. In cadastral parlance, such unit is said to be owned by the share-holding property units. Furthermore, land and the building erected on it may establish one unit, yet alternatively the building always or under certain conditions constitutes a unit in its own right. Variations also occur as to whether parts of buildings can become units, for example in terms of condominiums, which may depend on conditions related to use for housing or business purposes. Research efforts under the heading of "standardization of the core cadastral unit" have contributed substantially to the understanding of the complexity of the property unit.

Updating the information in property registers is as essential as the specification of units. From an informatics point of view, a survey of the information flows in what may be called the "geodata network" may reveal uncoordinated, and perhaps duplicated, efforts to acquire information and other suboptimal practices. However, from the end users' point of view, what takes place is a transaction of property rights and related processes, for example subdivision of property units. The updating of property registers is from this point of view a by-product of the transaction.

The end-user point of view is taken also by economists, who offer a theoretical basis for investigations of the mentioned processes, a field known as "institutional economics". New institutional economics (NIE) introduces "transaction costs" as an expense in addition to the cost of producing a commodity to the market. In the present context, the transaction costs are the fees and honoraries, etc. to be paid by the buyer and seller of a property unit, besides the cost of the property itself. Buyers' efforts to make sure that the seller is in fact entitled to dispose of the property unit concerned can be drastically reduced, that is: transaction costs are lowered, where reliable title information from land registries is available.

The NIE approach, as advocated by Nobel laureate Douglass C North, was applied in recent, comparative research. His notion of "institution": the norms which restrict and enable human behavior, suggested research focused on practices rather than legal texts. Methods were developed and applied for the systematic description and comparison of property transactions, including a formal, ontology-based approach. The methods developed were feasible and

initial national accounts of transaction costs were drafted. However, advice for optimal solutions are not to be expected, partly because many of the agents involved, both in the private and the public sector, have other preferences than the minimizing of transaction costs. Moreover, property transactions are, for various political reasons, often regulated, for example through municipal preemption rights or spatial planning measures. The NIE approach does however offer an adequate theoretical basis for analyzes of the infrastructure of real property rights, analyzes which assist in the identification and remedy of the most inappropriate practices.

Cadastral Development Through Institutional Transactions

Institutional economics divides into two strands: NIE, and institutional political economy, respectively. The former may be applied to the cadastre and its related processes, conceived as a quasi-rational, smooth-running machine that dispatches information packets between agents to achieve a certain outcome, e.g., the exchange of money against title to a specific property unit. However, this approach does not account for the fact that the various governmental units, professional associations, etc. involved in the cadastral processes have diverse mandates and objectives. Development projects pay attention to these conflicting interests through stakeholder analyzes. Research in organizational sociology suggests identification of "policy issue networks" and investigation of the exchange of resources among the actors involved, for example during the preparation and passing of a new act. The resources are generally supposed not to be money, although bribery occurs, but more abstract entities such as legal-technical knowledge, access to decision centers, organizational skills, reputation, and mobilizing power.

Institutional economics provides a frame for relating this power game to the routine transactions of conveyance of real estate. It is done by introducing two layers of social analysis: the layer of routine transactions, and the layer of change of the rules which determine the routine transactions (Williamson 2000). In economic terms, conveyance is an example of a transaction in commodities or other assets. These transactions are performed according to a set of rules: acts, by-laws, professional codes of conduct, etc. which are, in the terminology of Douglass North, a set of institutions. The process of change of these institutions is the object of analysis on the second layer and, following Daniel Bromley, called "institutional transactions". Institutional transactions may reduce transaction costs within the jurisdiction concerned, but Bromley shows at length that this need not be so; generally, the initiator of an institutional transaction cannot be sure whether the intended outcome is realized. Among other things, this is because the transaction is open to unplanned interference from actors on the periphery and also because the various resources of the actors are only partly known at the outset.

The strand of institutional political economy is researching such institutional transactions in order to explain why some countries grow rich, while others fail to develop their economy. Here we have the theoretical basis for explaining the emergence and diffusion of the cadastre in early modern Europe, cf. "Historical Background" above. The pious and enlightened absolutist monarchs and their advisors established a set of norms that framed institutional transactions in a way that encouraged a growing number of the population to strive for "the common weal".

Key Applications

The definition of cadastre specifies the key application: the official identification and recording of information on geographical objects: pieces of land, buildings, pipes, etc. as well as documents and extracts from these on rights and restrictions pertaining to the geographical objects. Cadastral knowledge is applied not only in the public sector, but also in private companies, as we shall see form the following overview of services:

- Facilitation of financial inflow to central and local government by specification of property

units and provision of information on identification, size, and boundaries and by providing standardized assessment of property units (mass-appraisal)
- Supporting the market in real estate by providing trustworthy recordings of title, mortgages, and restrictions
- Supporting sustainable land use (urban, rural, and environmental aspects) by providing data for the decision process and applying the decision outcome in property development
- Supporting the construction industry and utilities with data and related services, and cooperating on the provision of statistics
- Assisting in continuous improvement of administration, often called "good governance", through:
  - Considering data definitions and data flows with other bodies in government and industry, by mutual relating of post addresses, units of property, units for small-area statistics and spatial planning, as well as road segments and other topographic elements
  - Applying cost-effective technology
  - Contributing towards good governance through participating in national high-level commissions that aim at change in the institutional structure concerning real property rights, housing, and land use.

The taxation issue is mentioned first among public sector issues, partly for historical reasons, but more essentially because the operation of the cadastre cost money. In a specific country, the list of tasks or functions may include more or less operations, and the grouping of tasks may differ. However, the list may give an idea of career opportunities, as does the corresponding list of services within the private sector:

- Assisting bodies in need of recording and management of a property portfolio, for example utilities, road, railroad and harbor agencies, defense, property owners' associations and hunting societies, charitable trusts, ecclesiastical property; supporting the selection of antenna sites for wireless networks
- Supporting property development by assisting owners in reorganizing the shape of and rights in the property unit and its surroundings, in compliance with the existing private and public restrictions, for example easements and zoning. Delivering impartial expertise in boundary disputes.
- Assisting bodies in need of measuring and recording of location-specific themes with legal implications, including cultural and natural heritage and opencast mining
- Facilitate the diffusion of open source and proprietary GIS software by adapting it to local needs

Again, in a specific country the structure of the construction industry and the utilities, as well as the status and scope of professional associations like those of geodetic surveyors and property developers may vary.

Future Directions

Formalization of Practice

Much of cadastral knowledge is still tacit. Thus a large task ahead is to elicit this knowledge and integrate it with relevant existing conceptual structures and ontologies. Cadastral work is embedded in legal prescripts and technical standards. The hypothesis supported by institutional economics is that the needed formalization will be achieved better through observation and analysis of human routine behavior than through legal analysis of prescripts or model building of information systems. Furthermore, the need for formalization is not only in regard to the production of cadastral services, but also the articulation of the intricate structure of process objectives or functions. The description of realized functions and function clusters depends to a certain extent on the local expertise that has an obvious interest in staying in business. The description effort thus has to face a value-laden context. Good places to begin research are with Kuhn (2001), Frank (2001), Stubkjær (2003), van Oosterom et al. (2005), and Zevenbergen et al. (2007).

Prioritized Issues

The International Federation of Surveyors (FIG) provides a framework for exchanges among professionals and academics. Of the ten commissions, commission 7 on Cadastre and Land Management at the International FIG Congress in Munich, October 2006, adopted a Work Plan 2007–2010. Major points of the plan include:

- Pro-poor land management and land administration, supporting good governance
- Land administration in post conflict areas
- Land administration and land management in the marine environment
- Innovative technology and ICT support in land administration and land management
- Institutional aspects for land administration and land management

Current working relations with UN bodies will continue and possibly include the UN Habitat Global Land Tool Network (GLTN). A good starting point would be http://www.fig.net/commission7/ and Kaufmann and Steudler (1998).

Development Research

From a global perspective, the past decades have been marked by an attempt, on the part of donors and experts, to install the Western model of individual real property rights and cadastre in developing or partner countries. The efforts have not had the intended effects for a variety of reasons. One may be that cadastral processes interfere with the preferences and habits of end users, including the large number of owners, lease-holders, tenants, mortgagees, etc. This is the case especially where the general economy and the structure of professions and university departments hamper the provision of skilled and impartial staff (in public and private sector) who could mediate between the wishes of the right holders and the technicalities of property rights, transactions, and recordings.

The provision of resources for such cadastral development likely depends on one or more global networks which comprises university departments, central and local governments, nongovernmental organizations (NGOs) and international organizations, for example as demonstrated in the Cities Alliance. Furthermore, the United Nations University (UNU) includes UNU-IIST, the International Institute for Software Technology, enabling software technologies for development, as well as UNU-WIDER which analyses property rights and registration from the point of view of development economics.

The Netherlands-based International Institute for Geo-Information Science and Earth Observation (ITC) has become an associated institution of UNU. With the Netherlands Cadastre, Land Registry and Mapping Agency, the ITC is establishing a School for Land Administration Studies at ITC. The school will among others execute a joint land administration programme with UNU, consisting of a series of seminars, short courses and networking. Bruce (1993), de Janvry and Sadoulet (2001), Palacio, and North (1990) are good places to start further reading on this.

Spatial Learning: Naming Objects of the Environment

In autumn 2006, the US National Science Foundation awarded a $3.5 million grant to establish a new research center to investigate spatial learning and use this knowledge to enhance the related skills students will need. Research goals include how to measure spatial learning.

Cadastral studies combine the verbal and the graphical-visual-spatial modes of communication and need reflection of the teaching and learning methods, also with web-supported distance learning in mind. More specifically, the need of balancing of nominal, ordinal and metric means of localization, etc. which was mentioned above in “Spatial Reference Frames”, would benefit from such reflection.

References

Bruce J (1993) Review of tenure terminology. Land tenure center report. http://agecon.lib.umn.edu/ltc/ltctb01.pdf. Accessed 14 Aug 2007

de Janvry A, Sadoulet E (2001) Access to land and land policy reforms. Policy brief no. 3, Apr

2001. http://www.wider.unu.edu/publications/pb3.pdf. Accessed 14 Aug 2007
Frank AU (2001) Tiers of ontology and consistency constraints in geographic information systems. Int J Geogr Inf Sci 15:667–678
Kain RJP, Baigent E (1993) The cadastral map in the service of the state: a history of property mapping. The University of Chicago Press, Chicago
Kaufmann J, Steudler D (1998) Cadastre 2014. http://www.fig.net/cadastre2014/index.htm. Accessed 14 Aug 2007
Kuhn W (2001) Ontologies in support of activities in geographical space. Int J Geogr Inf Sci 15:613–631
North DC (1990) Institutions, institutional change and economic performance. Cambridge University Press, Cambridge
Palacio A, Legal empowerment of the poor: an action Agenda for the world bank. Available via http://rru.worldbank.org/Documents/PSDForum/2006/background/legal_empowerment_of_poor.pdf. Accessed 14 Aug 2007
Stubkjær E (2003) Modelling units of real property rights. In: Virrantaus K, Tveite H (eds) ScanGIS'03 proceedings, 9th Scandinavian research conference on geographical information sciences, Espoo, June 2003
van Oosterom P, Schlieder C, Zevenbergen J, Hess C, Lemmen C, Fendel E (eds) (2005) Standardization in the cadastral domain. In: Proceedings, standardization in the cadastral domain, Bamberg, Dec 2004. The International Federation of Surveyors, Frederiksberg
Williamson OE (2000) The new institutional economics: taking stock, looking ahead. J Econ Literat 38:595–613
Zevenbergen J, Frank A, Stubkjær E (eds) (2007) Real property transactions: procedures, transaction costs and models. IOS Press, Amsterdam

Recommended Reading

Chang H-J Understanding the relationship between institutions and economic development: some key theoretical issues. UNU/WIDER Discussion Paper 2006/05. July 2006. http://www.wider.unu.edu/publications/dps/dps2006/dp2006-05.pdf. Accessed 14 Aug 2007
de Janvry A, Gordillo G, Platteau J-P, Sadoulet E (eds) (2001) Access to land, rural poverty and public action. UNU/WIDER studies in development economics. Oxford University Press, Oxford/New York

Camera Model

▸ Photogrammetric Methods

Carbon Emissions

▸ Climate Risk Analysis for Financial Institutions

Carbon Finance

▸ Climate Risk Analysis for Financial Institutions

Carbon Trading

▸ Climate Risk Analysis for Financial Institutions

Cardinal Direction Relations

▸ Directional Relations

Cartographic Data

▸ Photogrammetric Products

Cartographic Generalization

▸ Abstraction of Geodatabases

Cartographic Information System

▸ Atlas Information Systems

Catalog Entry

▸ Metadata and Interoperability, Geospatial

Catalogue Information Model

Liping Di[1] and Yuqi Bai[2]
[1]Center for Spatial Information Science and Systems (CSISS), George Mason University, Fairfax, VA, USA
[2]Center for Spatial Information Science and Systems (CSISS), George Mason University, Greenbelt, MD, USA

Synonyms

Catalogue information schema; Catalogue metadata schema; Registry information model

Definition

The catalogue information model is a conceptual model that specifies how metadata is organized within the catalogue. It defines a formal structure representing catalogued resources and their interrelationships, thereby providing a logical schema for browsing and searching the contents in a catalogue.

There are multiple and slightly different definitions of the catalogue information model used by various communities. The Open Geospatial Consortium (OGC) defines the catalogue information model in the OGC Catalogue Services Specification (OGC) as an abstract information model that specifies a BNF grammar for a minimal query language, a set of core queryable attributes (names, definitions, conceptual data types), and a common record format that defines the minimal set of elements that should be returned in the brief and summary element sets. The Organization for the Advancement of Structured Information Standards (OASIS) defines the registry information model in the ebXML Registry Information Model (ebRIM) specification (OASIS) as the information model which provides a blueprint or high-level schema for the ebXML registry. It provides the implementers of the ebXML registry with information on the type of metadata that is stored in the registry as well as the relationships among metadata classes. The registry information model defines what types of objects are stored in the registry and how stored objects are organized in the registry. The common part from these two definitions is the schema for describing the objects catalogued/registered in and for organizing the descriptions in a catalogue/registry – the catalogue metadata schema.

Historical Background

The first catalogues were introduced by publishers serving their own business of selling the books they printed. At the end of the fifteenth century, they made lists of the available titles and distributed them to those who frequented the book markets. Later on, with the increasing volume of books and other inventories, the library became one of the earliest domains providing a detailed catalogue to serve their users. These library catalogues hold much of the reference information (e. g., author, title, subject, publication date, etc.) of bibliographic items found in a particular library or a group of libraries.

People began to use the term "metadata" in the late 1960s and early 1970s to identify this kind of reference information. The term "meta" comes from a Greek word that denotes "alongside, with, after, next." More recent Latin and English usage would employ "meta" to denote something transcendental or beyond nature (Using Dublin Core).

The card catalogue was a familiar sight to users for generations, but it has been effectively replaced by the computerized online catalogue which provides more advanced information tools helping to collect, register, browse, and search digitized metadata information.

Scientific Fundamentals

Metadata can be thought of as data about other data. It is generally used to describe the characteristics of information-bearing entities to aid in the identification, discovery, assessment, management, and utilization of the described

entities. Metadata standards have been developed to standardize the description of information-bearing entities for specific disciplines or communities. For interoperability and sharing purposes, a catalogue system usually adopts a metadata standard used in the community the system intends to serve as its catalogue information model.

A metadata record in a catalogue system consists of a set of attributes or elements necessary to describe the resource in question. It is an example of the catalogue information model being used by the catalogue system. A library catalogue, for example, usually consists of the author, title, date of creation or publication, subject coverage, and call number specifying the location of the item on the shelf. The structures, relationships, and definitions for these queryable attributes – known as conceptual schemas – exist for multiple information communities. For the purposes of interchange of information within an information community, a metadata schema may be created that provides a common vocabulary which supports search, retrieval, display, and association between the description and the object being described.

A catalogue system needs to reference an information model for collecting and manipulating the metadata of the referenced entities catalogued in the system. The information model provides specific ways for users to browse and search them. Besides the metadata information that directly describes those referenced entities themselves, a catalogue might hold another type of metadata information that describes the relationship between these entities.

Some catalogue services may only support one catalogue information model, each with the conceptual schema clearly defined, while others can support more than one catalogue information model. For example, in the US Geospatial Data Clearinghouse, the affiliated Z39.50 catalogue servers only support US Content Standard for Digital Geospatial Metadata (CSDGM) standard in their initial developing stage. While in OGC Catalogue Service base specification, what catalogue information model can be used is undefined. Developers are encouraged to propose their own catalogue information model as profiles. However, to facilitate the interoperability between diverse OGC-compliant catalogue service instances, a set of core queryable parameters originated from Dublin Core is proposed in the base specification and is desirable to be supported in each catalogue service instance. OGC further endorsed the OASIS ebRIM (e-Business Registry Information Model) as the preferred basis for future profiles of OGC Catalogue (OGC).

How a catalogue information model can be formally discovered and described in a catalogue service is another issue. Some catalogue services do not provide specific operations for automatic discovery of the underlying catalogue information model, while others support particular operations to fulfill this task. In the OGC Catalogue Services Specification, the names of supported information model elements can be listed in the capabilities files, and a mandatory *DescribeRecord* operation allows the client to discover elements of the information model supported by the target catalogue service. This operation allows some of or the entire information model to be described.

Key Applications

The concept of the catalogue information model has been widely applied in many disciplines for information management and retrieval. Common metadata standards are widely adopted as the catalogue information model. Among them, the Dublin Core is one of the most referenced and commonly used metadata information models for scientific catalogues. In the area of geographic information science, ISO 19115 is being widely adopted as the catalogue information model for facilitating the sharing of a large volume of geospatial datasets.

Dublin Core

The Dublin Core metadata standard is a simple yet effective element set for describing a wide range of networked resources (Dublin Core). The "Dublin" in the name refers to Dublin, Ohio, USA, where the work originated from

a workshop hosted by the Online Computer Library Center (OCLC), a library consortium which is based there. The "Core" refers to the fact that the metadata element set is a basic but expandable "core" list (Using Dublin Core).

The Simple Dublin Core Metadata Element Set (DCMES) consists of 15 metadata elements: title, creator, subject, description, publisher, contributor, date, type, format, identifier, source, language, relation, coverage, and rights. Each element is optional and may be repeated.

The Dublin Core Metadata Initiative (DCMI) continues the development of exemplary terms or "qualifiers" that extend or refine these original 15 elements. Currently, the DCMI recognizes two broad classes of qualifiers: element refinement and encoding scheme. Element refinement makes the meaning of an element narrower or more specific. Encoding scheme identifies schemes that aid in the interpretation of an element value.

There are many syntax choices for Dublin Core metadata, such as SGML, HTML, RDF/XML, and key-value pair TXT file. In fact, the concepts and semantics of Dublin Core metadata are designed to be syntax independent and are equally applicable in a variety of contexts.

Earth Science

With the advances in sensor and platform technologies, the Earth science community has collected a huge volume of geospatial data in the past 30 years via remote sensing methods. To facilitate the archival, management, and sharing of these massive geospatial data, the Earth science community has been one of the pioneers in defining metadata standards and using them as information models in building catalogue systems.

FGDC Content Standard for Digital Geospatial Metadata

The Federal Geographic Data Committee (FGDC) of the USA is a pioneer in setting geospatial metadata standards for the US federal government. To provide a common set of terminology and definitions for documenting digital geospatial data, FGDC initiated work on setting the Content Standard for Digital Geospatial Metadata (CSDGM) in June of 1992 through a forum on geospatial metadata. The first version of the standard was approved on June 8, 1994, by the FGDC.

Since the issue of Executive Order 12906, "Coordinating Geographic Data Acquisition and Access: The National Spatial Data Infrastructure," by President William J. Clinton on April 11, 1994, this metadata standard has been adopted as the catalogue information model in numerous geospatial catalogue systems operated by US federal, state, and local agencies as well as companies and groups. It has also been used by other nations as they develop their own national metadata standards.

In June of 1998, the FGDC approved the CSDGM version 2, which is fully backward compatible with and supersedes the June 8, 1994, version. This version provides for the definition of profiles (Appendix E) and extensibility through user-defined metadata extensions (Appendix D). The June 1998 version also modifies some production rules to ease implementation.

The Content Standard for Digital Geospatial Metadata (CSDGM) (FGDC, CSDGM) identifies and defines the metadata elements used to document digital geospatial data sets for many purposes, which includes (1) preservation of the meaning and value of a dataset, (2) contribution to a catalogue or clearinghouse, and (3) aid in data transfer. CSDGM groups the metadata information into the following seven types:

- Identification_Information
- Data_Quality_Information
- Spatial_Data_Organization_Information
- Spatial_Reference_Information
- Entity_and_Attribute_Information
- Distribution_Information
- Metadata_Reference_Information

For each type, it further defines composed elements and their type, short name, and/or domain information.

To provide a common terminology and set of definitions for documenting geospatial data obtained by remote sensing, the FGDC defined the Extensions for Remote Sensing Metadata within the framework of the June 1998 version of the

CSDGM (FGDC, Content Standard for Digital Geospatial Metadata). These *remote sensing extensions* provide additional information particularly relevant to remote sensing: the geometry of the measurement process, the properties of the measuring instrument, the processing of raw readings into geospatial information, and the distinction between metadata applicable to an entire collection of data and those applicable only to component parts. For that purpose, these *remote sensing extensions* establish the names, definitions, and permissible values for new data elements and the compound elements of which they are the components. These new elements are placed within the structure of the base standard, allowing the combination of the original standard and the new extensions to be treated as a single entity (FGDC, Content Standard for Digital Geospatial Metadata).

ISO 19115

In May 2003, ISO published ISO 19115: Geographic Information Metadata (ISO/TC 211). The international standard was developed by ISO Technical Committee (TC) 211 as a result of consensus among TC national members as well as its liaison organizations on geospatial metadata. ISO 19115, rooted at FGDC CSDGM, provides a structure for describing digital geographic data. Actual clauses of 19115 cover properties of the metadata: identification, constraints, quality, maintenance, spatial representation (grid and vector), reference systems, content (feature catalogue and coverage), portrayal, distribution, extensions, and application schemas. Complex data types used to describe these properties include extent and citations. ISO 19115 has been adopted by OGC as a catalogue information model in its Catalogue Service for Web-ISO 19115 Profile (OGC). Figure 1 depicts the top-level UML model of the metadata standard.

ISO 19115 defines more than 300 metadata elements (86 classes, 282 attributes, 56 relations). The complex, hierarchical nested structure and relationships between the components are shown using 16 UML diagrams.

To address the issue whether a metadata entity or metadata element shall always be documented in the metadata or sometimes be documented, ISO 19115 defines a descriptor for each package and each element. This descriptor may have the following values:

- M (mandatory)
- C (conditional)
- O (optional)

Mandatory (M) means that the metadata entity or metadata element shall be documented.

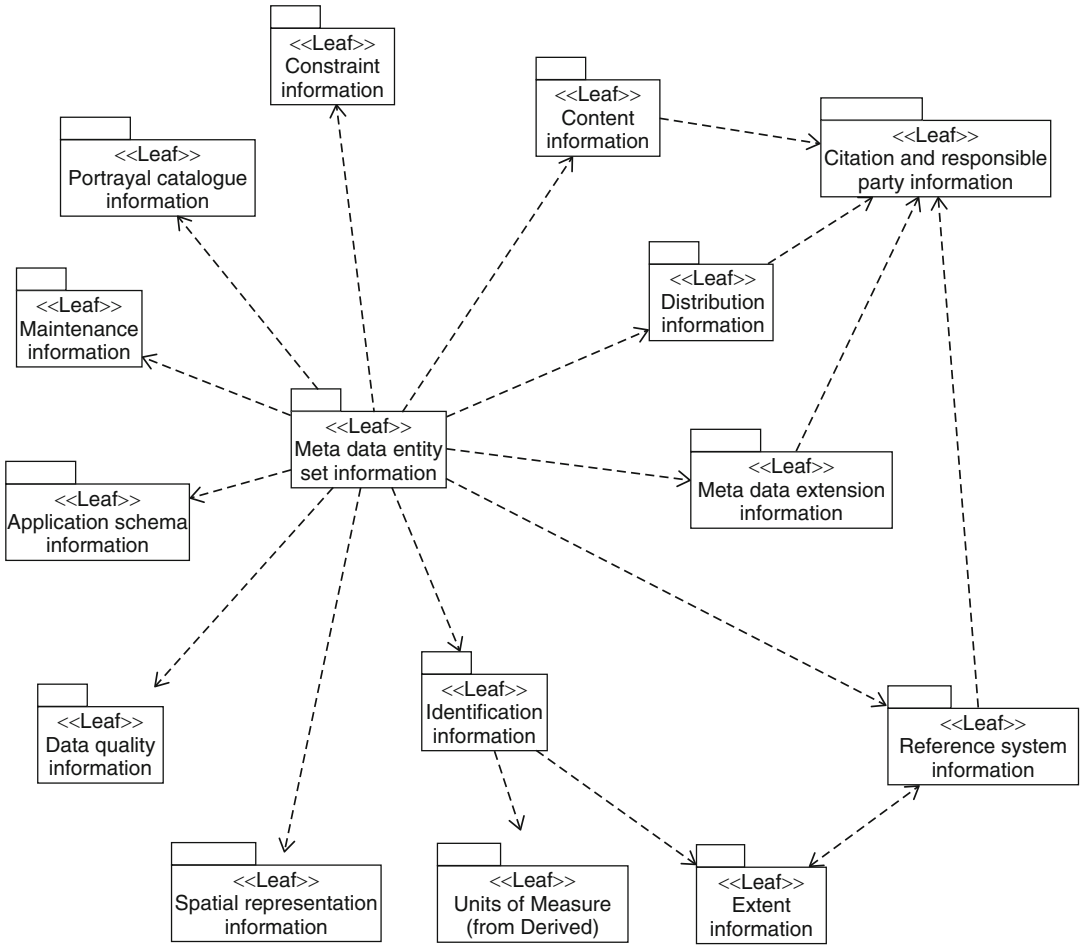
Conditional (C) specifies an electronically manageable condition under which at least one metadata entity or a metadata element is mandatory. Conditions are defined in the following three possibilities:

- Expressing a choice between two or more options. At least one option is mandatory and must be documented.
- Documenting a metadata entity or a metadata element if another element has been documented.
- Documenting a metadata element if a specific value for another metadata element has been documented. To facilitate reading by humans, plain text is used for the specific value. However, the code shall be used to verify the condition in an electronic user interface.

In short, if the answer to the condition is positive, then the metadata entity or the metadata element shall be mandatory.

Optional (O) means that the metadata entity or the metadata element may be documented or may not be documented. Optional metadata entities and optional metadata elements provide a guide to those looking to fully document their data. If an optional entity is not used, the elements contained within that entity (including mandatory elements) will also not be used. Optional entities may have mandatory elements; those elements only become mandatory if the optional entity is used.

ISO 19115 defines the core metadata that consists of a minimum set of metadata required to serve the full range of metadata applications. All the core elements must be available in



Catalogue Information Model, Fig. 1 ISO 19115 metadata UML package

a given metadata system. The optional ones need not be instantiated in a particular dataset. These 22 metadata elements are shown in Table 1.

Currently, ISO is developing ISO 19115-2, which extends ISO 19115 for imagery and gridded data. Similar to the FGDC efforts and using FGDC CSDGM Extensions for Remote Sensing Metadata as its basis, ISO 19115-2 will define metadata elements particularly for imagery and gridded data within the framework of ISO 19115. According to the ISO TC 211 program of work, the final CD was posted in March 2007; barring major objection, it will be published as DIS in June 2007.

US NASA ECS Core Metadata Standard

To enable an improved understanding of the Earth as an integrated system, in 1992, the National Aeronautics and Space Administration (NASA) of the USA started the Earth Observing System (EOS) program, which coordinates efforts to study the Earth as an integrated system. This program, using spacecraft, aircraft, and ground instruments, allows humans to better understand climate and environmental changes and to distinguish between natural and human-induced changes. The EOS program includes a series of satellites, a science component, and a data system for long-term global observations of the land surface, biosphere, solid Earth, atmosphere,

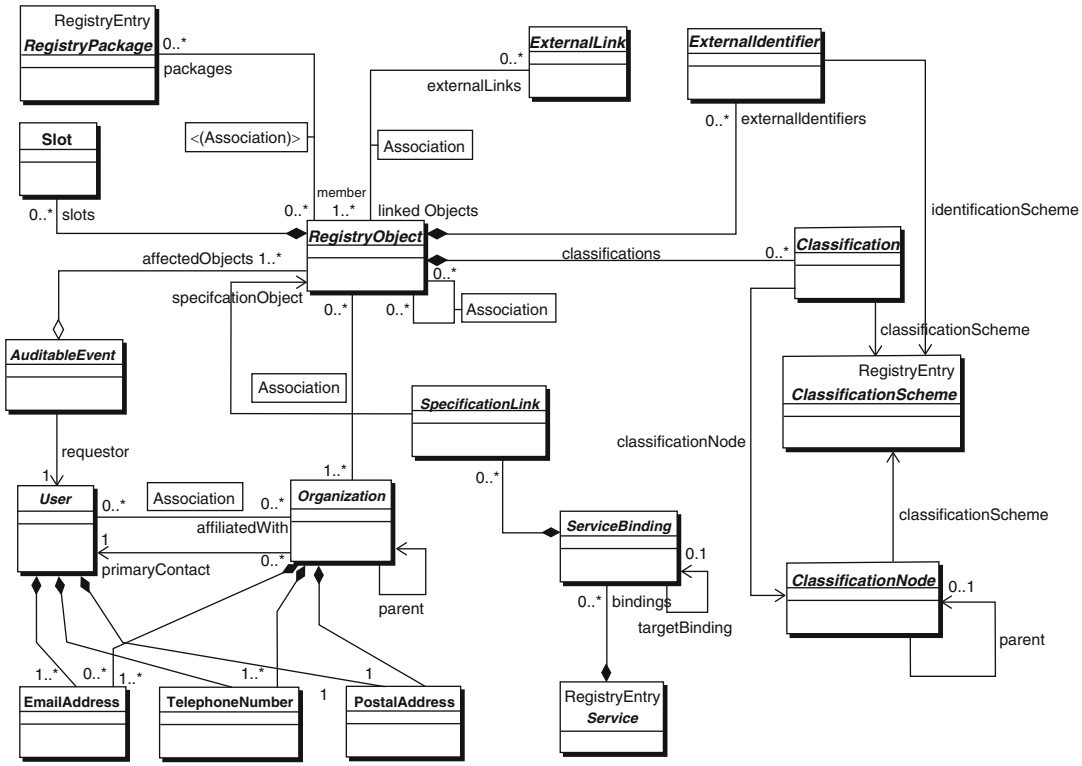
Catalogue Information Model, Table 1 ISO 19115 core metadata

Dataset title (M)	**Spatial representation type (O)**
(MD_Metadata > MD_DataIdentification.citation > CI_Citation.title)	(MD_Metadata > MD_DataIdentification.spatialRepresentationType)
Dataset reference date (M)	**Reference system (O)**
(MD_Metadata > MD_DataIdentification.citation > CI_Citation.date)	(MD_Metadata > MD_ReferenceSystem)
Dataset responsible party (O)	**Lineage (O)**
(MD_Metadata > MD_DataIdentification.pointOfContact > CI_ ResponsibleParty)	(MD_Metadata > DQ_DataQuality.lineage > LI_Lineage)
Geographic location of the dataset (by four coordinates or by geographic identifier) (C)	**Online resource (O)**
(MD_Metadata > MD_DataIdentification.extent > EX_Extent > EX_GeographicExtent > EX_GeographicBoundingBox or EX_GeographicDescription)	(MD_Metadata > MD_Distribution > MD_DigitalTransferOption.onLine > CI_OnlineResource)
Dataset language (M)	**Metadata file identifier (O)**
(MD_Metadata > MD_DataIdentification.language)	(MD_Metadata.fileIdentifier)
Dataset character set (C)	**Metadata standard name (O)**
(MD_Metadata > MD_DataIdentification.characterSet)	(MD_Metadata.metadataStandardName)
Dataset topic category (M)	**Metadata standard version (O)**
(MD_Metadata > MD_DataIdentification.topicCategory)	(MD_Metadata.metadata.StandardVersion)
Spatial resolution of the dataset (O)	**Metadata language (C)**
(MD_Metadata > MD_DataIdentification.spatial Resolution > MD_ Resolution.equivalentScale or MD_Resolution.distance)	(MD_Metadata.language)
Abstract describing the dataset (M)	**Metadata character set (C)**
(MD_Metadata > MD_DataIdentification.abstact)	(MD_Metadata.characterSet)
Distribution format (O)	**Metadata point of contact (M)**
(MD_Metadata > MD_Distribution > MD_Format.name and MD_Format.version)	(MD_Metadata.contact > CI_ResponsibleParty)
Additional extent information for the dataset (vertical and temporal) (O)	**Metadata date stamp (M)**
(MD_Metadata > MD_DataIdentification.extent > EX_Extent > EX_TemporalExtent or EX_VerticalExtent)	(MD_Metadata.dateStamp)

and oceans. The program aims at accumulating 15 years of Earth observation data at a rate of over 2 terabytes per day. To support data archival, distribution, and management, NASA has developed an EOS Data and Information System (EOSDIS) and its core system (ECS), the largest data and information system for Earth observation in the world.

In order to standardize the descriptions of data collected by the EOS program, NASA has developed the ECS Core Metadata Standard. The standard defines metadata in several areas: algorithm and processing packages, data sources, references, data collections, spatial and temporal extent, and content. The ECS Core Metadata Standard has been used as the catalogue information model for EOSDIS Data Gateway (EDG) and the EOS ClearingHOuse (ECHO). The ECS Core Metadata Standard was the basis for the development of FGDC CSDGM Extensions for Remote Sensing Metadata.

With new satellites being launched and instruments being operational, this standard will incorporate new keywords from them into the new version. The current version is 6B, released on October 2002. The 6B version is logically segmented into eight modules for the purpose of readability, including data originator, ECS collection, ECS



Catalogue Information Model, Fig. 2 The high-level UML model of ebRIM (OASIS)

data granule, locality spatial, locality temporal, contact, delivered algorithm package, and document (ECS).

ebRIM

The ebXML Registry Information Model (ebRIM) was developed by OASIS to specify the information model for the ebXML *registry* (OASIS). The goal of the ebRIM specification is to communicate what information is in the *registry* and how that information is organized. The high-level view of the information model is shown in Fig. 2.

ebRIM has been widely used in the world of e-business web services as the standardized information model for service registries. In the geospatial community, OGC has adopted the ebRIM specification as one of the application profiles of its Catalogue Service for Web Specification (OGC). And this model has been further approved as the preferred meta-model for future OGC CS-W Catalogue Application Profiles in the OGC technical meeting in December 2007 (OGC).

Future Directions

Catalogue information models define the organization of metadata in the catalogue. Each catalogue system normally adopts a metadata standard as its catalogue information model. With the wide acceptance of the concept of metadata, there are usually multiple related metadata standards used by different catalogue systems in the same application domain. Metadata crosswalks are needed to build maps between two metadata elements and/or attributes so that the interoperability among the legacy catalogue systems becomes possible. Besides these crosswalks, another direction is to organize a formal representation of every metadata concept into a geospatial ontology for enabling semantic interoperability between catalogues.

In addition to the catalogue information models, there are other necessary parts to compose an operational catalogue service, such as query language, query and response protocol, etc. Related research has been directed in the OGC, FGDC, and other agencies.

References

Dublin Core, http://en.wikipedia.org/wiki/Dublin_Core. Accessed 13 Sept 2007

ECS, Release 6B Implementation Earth Science Data Model for the ECS Project. http://spg.gsfc.nasa.gov/standards/heritage/eosdis-core-system-data-model. Accessed 13 Sept 2007

FGDC, CSDGM. http://www.fgdc.gov/standards/standards_publications/index_html. Accessed 13 Sept 2007

FGDC, Content Standard for Digital Geospatial Metadata, Extensions for Remote Sensing Metadata. http://www.fgdc.gov/standards/standards_publications/index_html. Accessed 13 Sept 2007

ISO/TC 211, ISO 19115 geographic information: metadata

OASIS, ebXML Registry Information Model (ebRIM). http://www.oasis-open.org/committees/regrep/documents/2.0/specs/ebrim.pdf. Accessed 13 Sept 2007

OGC, OGC™ Catalogue Services Specification, OGC 04-021r3. http://portal.opengeospatial.org/files/?artifact_id=5929&Version=2 . Accessed 13 Sept 2007

OGC, OpenGIS® Catalogue Services Specification 2.0 – ISO19115/ISO19119 Application Profile for CSW 2.0. OGC 04-038r2. https://portal.opengeospatial.org/files/?artifact_id=8305. Accessed 13 Sept 2007

OGC, EO Products Extension Package for ebRIM (ISO/TS 15000-3) Profile of CSW 2.0. OGC 06-131. http://portal.opengeospatial.org/files/?artifact_id=17689. Accessed 13 Sept 2007

OGC, OGC Adopts ebRIM for Catalogues. http://www.opengeospatial.org/pressroom/pressreleases/655. Accessed 13 Sept 2007

Using Dublin Core, http://www.dublincore.org/documents/2001/04/12/usageguide/. Accessed 13 Sept 2007

Catalogue Information Schema

▸ Catalogue Information Model

Catalogue Metadata Schema

▸ Catalogue Information Model

Category, Geographic

▸ Geospatial Semantic Integration

Central Perspective

▸ Photogrammetric Methods

Central Projection

▸ Photogrammetric Methods

Centrographic Measures

▸ CrimeStat: A Spatial Statistical Program for the Analysis of Crime Incidents

CGI

▸ Web Mapping and Web Cartography

CGIS

▸ Geocollaboration

Chain

▸ Spatial Data Transfer Standard (SDTS)

Chain of Space-Time Prisms

▸ Space-Time Prism Model

Change Detection

Jérôme Théau
GIS Training and Research Center, Idaho State University, Pocatello Idaho, ID, USA

Synonyms

Detection of changes; Digital change detection methods; Land cover change detection

Definition

Change Detection can be defined as the process of identifying differences in the state of an object or phenomenon by observing it at different times (Singh 1989). This process is usually applied to earth surface changes at two or more times. The primary source of data is geographic and is usually in digital format (e.g., satellite imagery), analog format (e.g., aerial photos), or vector format (e.g., feature maps). Ancillary data (e.g., historical, economic, etc.) can also be used.

Historical Background

Change detection history starts with the history of remote sensing and especially the first aerial photography taken in 1859 by Gaspard Felix Tournachon, also known as Nadar. Thereafter, the development of change detection is closely associated with military technology during world wars I and II and the strategic advantage provided by temporal information acquired by remote sensing. Civilian applications of change detection were developed following these events in the twentieth century using mostly interpretation and analog means. However, civilian availability of data was limited until the 1970s and 1980s due to military classification of imagery.

The development of digital change detection era really started with the launch of Landsat-1 (called first: Earth Resources Technology Satellite) in July 1972. The regular acquisition of digital data of the earth surface in multispectral bands allowed scientists to get relatively consistent data over time and to characterize changes over relatively large area for the first time. The continuity of this mission as well as the launch of numerous other ones ensured the development of change detection techniques from that time.

However, the development of digital change detection techniques was limited by data processing technology capacities and followed closely the development of computer technologies. The situation evolves from the 1960s when a few places in the world were equipped with expensive computers to the present when personal computers are fast and cheap enough to apply even complex algorithms and change detection techniques to satellite imagery. The computer technology also evolved from dedicated hardware to relatively user-friendly software specialized for image processing and change detection.

Based on published literature, the algebra techniques such as image differencing or image ratioing were the first techniques used to characterize changes in digital imagery during the 1970s (Lunetta and Elvidge 1998). These techniques are simple and fast to perform and are still widely used today. More complex techniques were developed since then with the improvement of processing capacities but also with the development of new theoretical approaches. Change detection analysis of the earth surface is a very active topic due to the concerns about consequences of global and local changes. This field of expertise is constantly progressing.

Scientific Fundamentals

Changes on Earth Surface

The earth surface is changing constantly in many ways. First, the time scales, at which changes can occur, are very heterogeneous. They may vary from catastrophic events (e.g., flood) to geological events (e.g., continental drift) which correspond to a gradient between punctual and continuous changes respectively. Secondly, the spatial scales, at which changes can occur, are

also very heterogeneous and may vary from local events (e.g., road construction) to global changes (e.g., ocean water temperature). Due to this very large spatio-temporal range, the nature and extent of changes are complex to determine because they are interrelated and interdependent at different scales (spatial, temporal). Change detection is, therefore, a challenging task.

Imagery Characteristics Regarding Changes

Since the development of civilian remote sensing, the earth benefits from a continuous and increasing coverage by imagery such as: aerial photography or satellite imagery. This coverage is ensured by various sensors with various properties. First, in terms of the time scale, various *temporal resolutions* (i.e., revisit time) and mission continuities allow coverage of every point of the earth from days to decades. Secondly, in terms of the spatial scale, various spatial resolutions (i.e., pixel size, scene size) allow coverage of every point of the earth at a sub-meter to a kilometer resolution. Thirdly, sensors are designed to observe the earth surface using various parts of the electromagnetic spectrum (i.e., spectral domain) at different resolutions (i.e., spectral resolution). This diversity allows the characterization of a large spectrum of earth surface elements and change processes. However, change detection is still limited by data availability and data consistency (i.e., multisource data).

Changes in Imagery

Changes in imagery between two dates translate into changes in radiance. Various factors can induce changes in radiance between two dates such as changes in: sensor calibration, solar angle, atmospheric conditions, seasons, or earth surface. The first premise of using imagery for change detection of the earth surface is that change in the earth surface must result in a change in radiance values. Secondly, the change in radiance due to earth surface changes must be large compared to the change in radiance due to other factors. A major challenge in change detection of the earth surface using imagery is to minimize these other factors. This is usually performed by carefully selecting relevant multidate imagery and by applying pre-processing treatments.

Data Selection and Pre-processing

Data selection is a critical step in change detection studies. The acquisition period (i.e., season, month) of multidate imagery is an important parameter to consider in image selection because it is directly related to phenology, climatic conditions, and solar angle. A careful selection of multidate images is therefore needed in order to minimize the effects of these factors. In vegetation change studies (i.e., over different years), for example, summer is usually used as the target period because of the relative stability of phenology, solar angle, and climatic conditions. The acquisition interval between multidate imagery is also important to consider. As mentioned before, earth surface changes must cause enough radiance changes to be detectable. However, the data selection is often limited by data availability and the choice is usually a compromise between the targeted period, interval of acquisition, and availability. The cost of imagery is also a limiting factor in data selection.

However, a careful data selection is usually not enough to minimize radiometric heterogeneity between multidate images. First, atmospheric conditions and solar angle differences usually need additional corrections and secondly other factors such as sensor calibration or geometric distortions need to be considered. In change detection analysis, multidate images are usually compared on a pixel basis. Then, very accurate *registrations* need to be performed between images in order to compare pixels at the same locations. Mis*registration* between multidate images can cause significant errors in change interpretation. The sensitivity of change detection approaches to mis*registration* is variable though. The minimization of radiometric heterogeneity (due to sources other than earth surface change) can be performed using different approaches depending on the level of correction required and the availability of atmospheric data. The techniques such as dark object subtraction, relative

radiometric normalization or radiative transfer code can be used.

Change Detection Methods

Summarized here are the most common methods used in change detection studies (Singh 1989; Lunetta and Elvidge 1998; Coppin et al. 2004; Lu et al. 2004; Mas 1999). Most of these methods use image processing approaches applied to multidate satellite imagery.

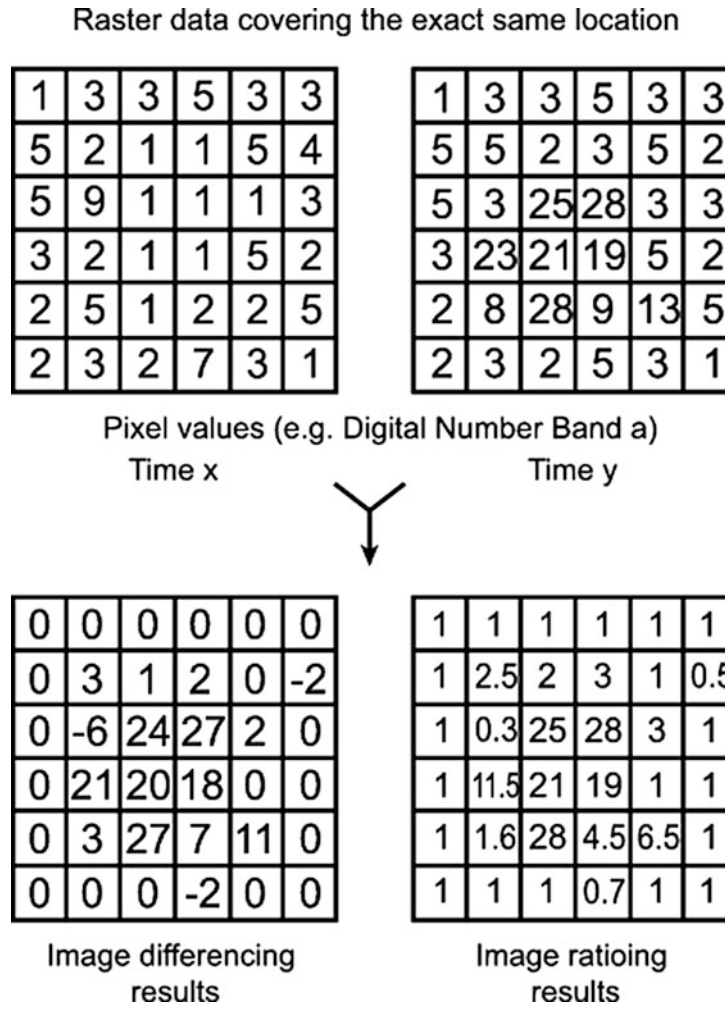
Image differencing: This simple method is widely used and consists of subtracting registered images acquired at different times, pixel by pixel and band by band. No changes between times result in pixel values of 0, but if changes occurred these values should be positive or negative (Fig. 1). However, in practice, exact image *registration* and perfect radiometric corrections are never obtained for multidate images. Residual differences in radiance not caused by land cover changes are still present in images. Then the challenge of this technique is to identify threshold values of change and no-change in the resulting images. Standard deviation is often used as a reference values to select these thresholds. Different normalization, histogram matching, and standardization approaches are used on multidate images to reduce scale and scene dependent effects on differencing results. The image differencing method is usually applied to single bands but can be also applied to processed data such as multidate vegetation indices or principal components.

Image ratioing: This method is comparable to the image differencing method in terms of its simplicity and challenges. However, it is not as widely used. It is a ratio of registered images acquired at different times, pixel by pixel and band by band. Changes are represented by pixel values higher or lower than 1 (Fig. 1). Pixels with no change will have a value of one. In practice, for the same reasons as in image differencing, the challenge of this technique is in selecting threshold values between change and no change. This technique is often criticized because the non-normal distribution of results limits the validity of threshold selection using the standard deviation of resulting pixels.

Post-classification: This method is also commonly referred to as "Delta classification". It is widely used and easy to understand. Two images acquired at different times are independently classified and then compared. Ideally, similar thematic classes are produced for each classification. Changes between the two dates can be visualized using a change matrix indicating, for both dates, the number of pixels in each class (Fig. 2). This matrix allows one to interpret what changes occurred for a specific class. The main advantage of this method is the minimal impacts of radiometric and geometric differences between multidate images. However, the accuracy of the final result is the product of accuracies of the two independent classifications (e.g., 64 % final accuracy for two 80 % independent classification accuracies).

Direct multidate classification: This method is also referred to as "Composite analysis", "Spectral-temporal combined analysis", "Spectral-temporal change classification", "Multidate clustering", or "Spectral change pattern analysis". Multidate images are combined into a single dataset on which a classification is performed (Fig. 3). The areas of changes are expected to present different statistics (i.e., distinct classes) compared to the areas with no changes. The approach can be unsupervised or supervised and necessitates only one classification procedure. However, this method usually produces numerous classes corresponding to spectral changes within each single image but also to temporal changes between images. The interpretation of results is often complex and requires a good knowledge of the study area. Combined approaches using principal component analysis or Bayesian classifier can be performed to reduce data dimensionality or the coupling between spectral and temporal change respectively.

Linear transformations: This approach includes different techniques using the same

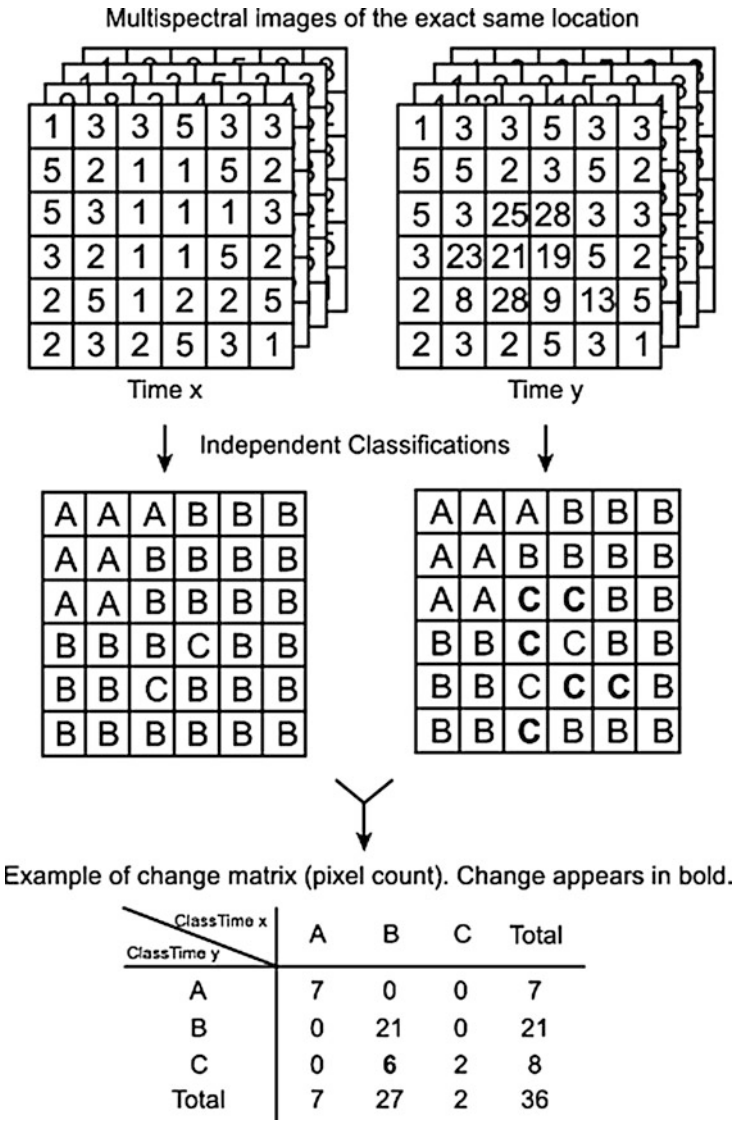


Change Detection, Fig. 1 Example of image differencing and image ratioing procedures

theoretical basis. The Principal Component Analysis (PCA) and the Tasseled-Cap transformations are the most common ones. Linear transformations are often used to reduce spectral data dimensionality by creating fewer new components. The first components contain most of the variance in the data and are uncorrelated. When used for change detection purposes, linear transformations are performed on multidate images that are combined as a single dataset (Fig. 3).

After performing a PCA, unchanged areas are mapped in the first component (i.e., information common to multidate images) whereas areas of changes are mapped in the last components (i.e., information unique to either one of the different dates). Usually the PCA is calculated from a variance/co-variance matrix. However, standardized matrix (i.e., correlation matrix) is also used. The PCA is scene dependent and results can be hard to interpret. The challenging steps are to label changes from principal components and to select thresholds between change and no-change areas. A good knowledge of the study area is required.

The Tasseled-Cap is also a linear transformation. However, unlike PCA, it is independent of the scene. The new component directions are selected according to pre-defined spectral properties of vegetation. Four new components are computed and oriented to enhance brightness, greenness, wetness, and yellowness. Results are

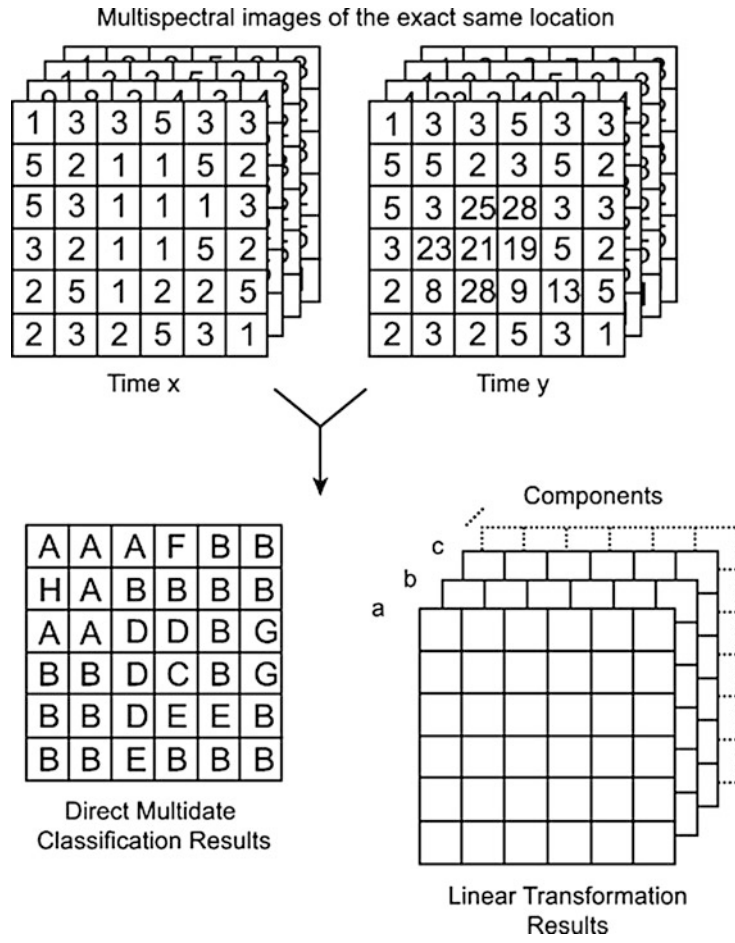


Change Detection, Fig. 2 Example of a post-classification procedure

also difficult to interpret and change labeling is challenging. Unlike PCA, Tasseled-Cap transformation for change detection requires accurate atmospheric calibration of multidate imagery.

Other transformations such as multivariate alteration detection or Gramm-Schmidt transformation were also developed but used to a lesser extent.

Change vector analysis: This approach is based on the spatial representation of change in a spectral space. When a pixel undergoes a change between two dates, its position in n-dimensional spectral space is expected to change. This change is represented by a vector (Fig. 4) which is defined by two factors, the direction which provides information about the nature of change and the magnitude which provides information about the level of change. This approach has the advantage to process concurrently any number of spectral bands. It also provides detailed information about change. The challenging steps are to define thresholds of magnitude, discriminating between change and no change, and to interpret vector direction in

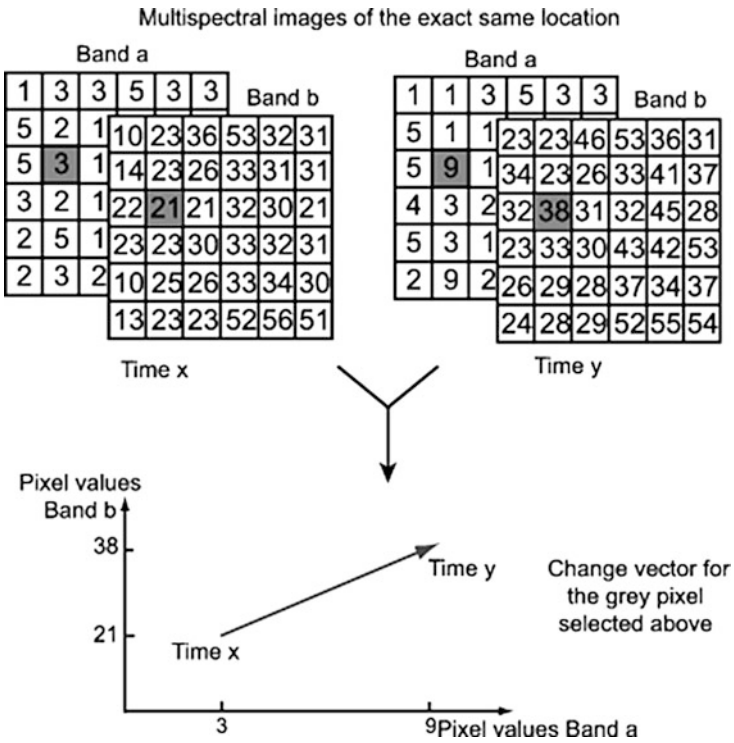


Change Detection, Fig. 3 Example of direct multidate classification and linear transformation procedures

relation with the nature of change. This approach is often performed on transformed data using methods such as Tasseled-Cap.

Image regression: This approach assumes that there is a linear relationship between pixel values of the same area at two different times. This implies that a majority of the pixels did not encounter changes between the two dates (Fig. 5). A regression function that best describes the relationship between pixel values of each spectral band at two dates is developed. The residuals of the regression are considered to represent the areas of changes. This method has the advantage of reducing the impact of radiometric heterogeneity (i.e., atmosphere, sun angle, sensor calibration) between multidate images. However, the challenging steps are to select an appropriate regression function and to define thresholds between change and no change areas.

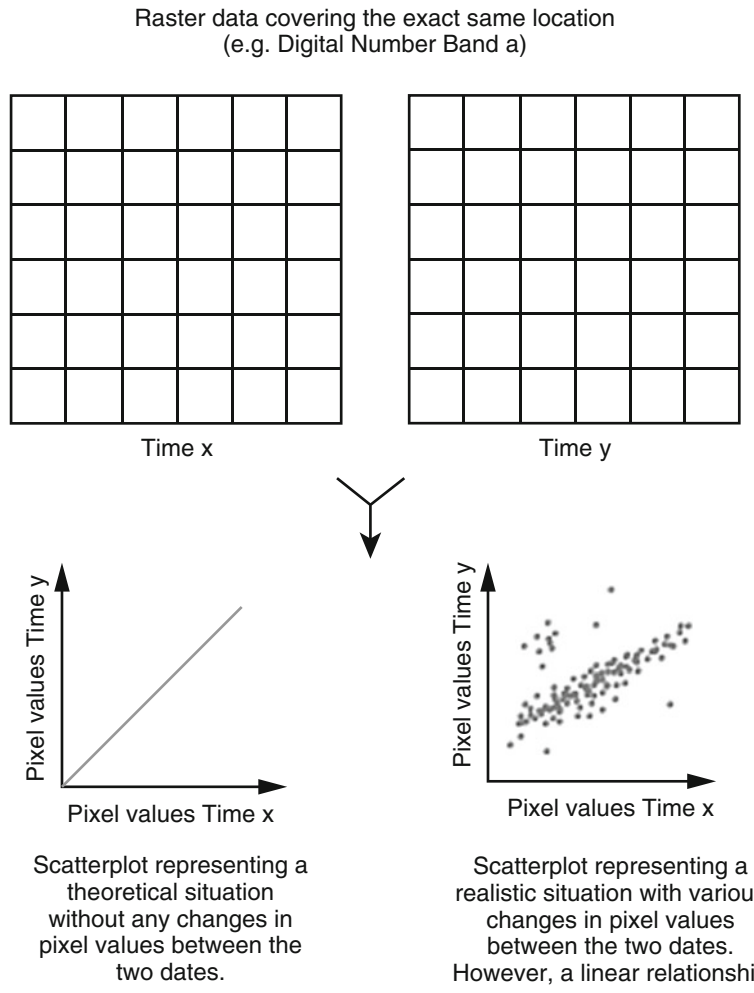
Multitemporal spectral mixture analysis: The spectral mixture analysis is based on the premise that a pixel reflectance value can be computed from individual values of its composing elements (i.e., end-members) weighted by their respective proportions. This case assumes a linear mixing of these components. This method allows retrieving sub-pixel information (i.e., surface proportions of end-members) and can be used for change detection purposes by performing separate analysis and comparing results at different dates (Fig. 6). The advantage of this method is to provide precise and repeatable results. The challenging step of this approach is to select suitable end-members.



Change Detection, Fig. 4 Example and principle of the change vector procedure

Combined approaches: The previous techniques represent the most common approaches used for change detection purposes. They can be used individually, but are often combined together or with other image processing techniques to provide more accurate results. Numerous combinations can be used and they will not be described here. Some of them include the combination of vegetation indices and image differencing, change vector analysis and principal component analysis, direct multidate classification and principal component analysis, multitemporal spectral analysis and image differencing, or image enhancement and post-classification.

Example of change detection analysis: Mapping changes in caribou habitat using multitemporal spectral mixture analysis: The George River Caribou Herd (GRCH), located in northeastern Canada, increased from about 5,000 in the 1950s to about 700,000 head in the 1990s. This has led to an over-utilization of summer habitat, resulting in degradation of the vegetation cover. This degradation has had a direct impact on health problems observed in the caribou (*Rangifer tarandus*) population over the last few years and may also have contributed to the recent decline of the GRCH (404,000 head in 2000–2001). Lichen habitats are good indicators of caribou herd activity because of their sensitivity to overgrazing and overtrampling, their widespread distribution over northern territories, and their influence on herd nutrition. The herd range covers a very large territory which is not easily accessible. As a result, field studies over the whole territory are limited and aerial surveys cannot be conducted frequently. Satellite imagery offers the synoptic view and temporal resolution necessary for mapping and monitoring caribou habitat. In this example, a change detection approach using Landsat imagery was used. The procedure was based on spectral mixture analysis and produced maps showing the lichen proportion inside each pixel. The procedure was applied to multidate imagery to monitor the spatio-temporal evolution of the lichen resource over the past three decades



Change Detection, Fig. 5 Example and principle of the image regression procedure

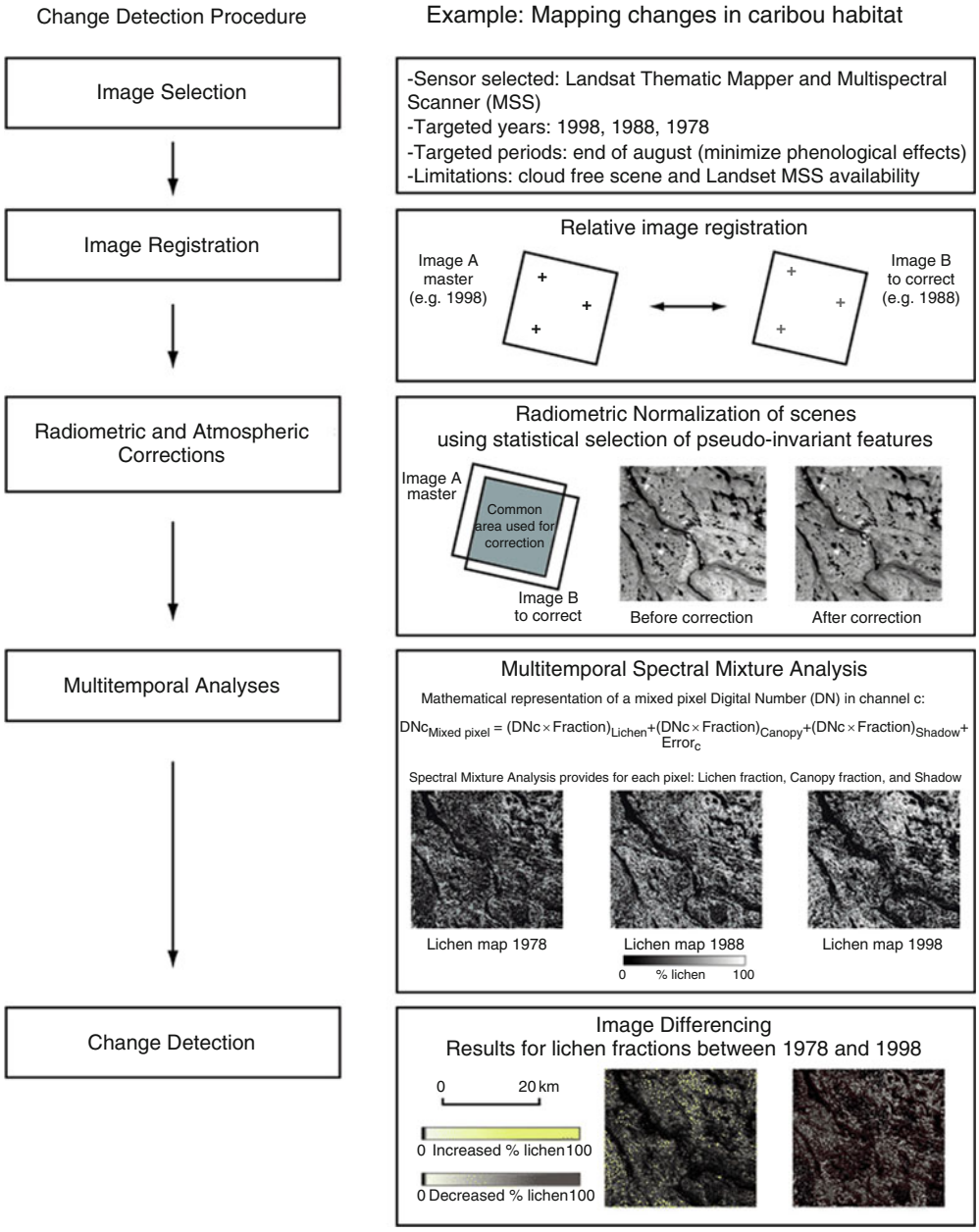
and gave new information about the habitat used by the herd in the past, which was very useful to better understand population dynamics. Figure 6 summarizes the approach used in this study and illustrates the steps typical of a change detection procedure.

Key Applications

The earth surface is changing constantly in many ways. Changes occur at various spatial and temporal scales in numerous environments. Change detection techniques are employed for different purposes such as research, management, or business (Lunetta and Elvidge 1998; Canada Centre for Remote Sensing; Diversitas; ESA; Global Change Master Directory; IGBP; IHDP; WCRP). Monitoring changes using GIS and remote sensing is therefore used in a wide field of applications. A non-exhaustive list of key applications is presented here.

Forestry

- Deforestation (e.g., clear cut mapping, regeneration assessment)
- Fire monitoring (e.g., delineation, severity, detection, regeneration)
- Logging planning (e.g., infrastructures, inventory, biomass)
- Herbivory (e.g., insect defoliation, grazing)
- Habitat fragmentation (e.g., landcover changes, heterogeneity)



For more details see: Théau and Duguay (2004) Mapping Lichen Habitat Changes inside the Summer Range of the George River Caribou Herd (Québec-Labrador, Canada) using Landsat Imagery (1976-1998). Rangifer. 24: 31-50.

Change Detection, Fig. 6 Example of a change detection procedure. Case study of mapping changes in caribou habitat using multitemporal spectral mixture analysis

Agriculture and Rangelands

- Crop monitoring (e.g., growing, biomass)
- Invasive species (e.g., detection, distribution)
- Soil moisture condition (e.g., drought, flood, landslides)
- Desertification assessment (e.g., bare ground exposure, wind erosion)

Urban

- Urban sprawl (e.g., urban mapping)

- Transportation and infrastructure planning (e.g., landcover use)

Ice and Snow

- Navigation route (e.g., sea ice motion)
- Infrastructure protection (e.g., flooding monitoring)
- Glacier and ice sheet monitoring (e.g., motion, melting)
- Permafrost monitoring (e.g., surface temperature, tree line)

Ocean and Coastal

- Water quality (e.g., temperature, productivity)
- Aquaculture (e.g., productivity)
- Intertidal zone monitoring (e.g., erosion, vegetation mapping)
- Oil spill (e.g., detection, oil movement)

Future Directions

In the past decades, a constant increase of remotely sensed data availability was observed. The launch of numerous satellite sensors as well as the reduction of product costs can explain this trend. The same evolution is expected in the future. The access to constantly growing archive contents also represents a potential for the development of more change detection studies in the future. Long-term missions such as Landsat, SPOT (Satellite pour l'Observation de la Terre), AVHRR (Advanced Very High Resolution Radiometer) provide continuous data for more than 20–30 years now. Although radiometric heterogeneity between sensors represents serious limitation in time series analysis, these data are still very useful for long term change studies. These data are particularly suitable in the development of temporal trajectory analysis which usually involves the temporal study of indicators (e.g., vegetation indices, surface temperature) on a global scale.

Moreover, as mentioned before in the Historical Background section, the development of change detection techniques are closely linked with the development of computer technologies and data processing capacities. In the future, these fields will still evolve in parallel and new developments in change detection are expected with the development of computer technologies.

Developments and applications of new image processing methods and geospatial analysis are also expected in the next decades. Artificial intelligence systems as well as knowledge-based expert systems and machine learning algorithms represent new alternatives in change detection studies (Coppin et al. 2004). These techniques have gained considerable attention in the past few years and are expected to increase in change detection approaches in the future. One of the main advantages of these techniques is that they allow the integration of existing knowledge and non-spectral information of the scene content (e.g., socio-economic data, shape, and size data). With the increasing interest in using integrated approaches such as coupled human-environment systems, these developments look promising.

The recent integration of change detection and spatial analysis modules in most GIS software also represents a big step towards integrated tools in the study of changes on the earth surface. This integration also includes an improvement of compatibility between image processing software and GIS software. More developments are expected in the future which will provide new tools for integrating multisource data more easily (e.g., digital imagery, hard maps, historical information, vector data).

Cross-References

- ▸ Co-location Pattern Discovery
- ▸ Correlation Queries in Spatial Time Series Data
- ▸ Spatiotemporal Change Footprint Pattern Discovery

References

Canada Centre for Remote Sensing, http://ccrs.nrcan.gc.ca/index_e.php. Accessed Nov 2006

Coppin P, Jonckheere I, Nackaerts K, Muys B, Lambin E (2004) Digital change detection methods in ecosystem monitoring: a review. Int J Remote Sens 25(9):1565–1596
Diversitas – Integrating biodiversity science for human well-being, http://www.diversitas-international.org/. Accessed Nov 2006
ESA – Observing the Earth, http://www.esa.int/esaEO/index.html. Accessed Nov 2006
Global Change Master Directory, http://gcmd.nasa.gov/index.html. Accessed Nov 2006
IGBP – International Geosphere-Biosphere Programme, http://www.igbp.net/. Accessed Nov 2006
IHDP – International Human Dimensions Programme on Global Environmental Change, http://www.ihdp.org/. Accessed Nov 2006
Lu D, Mausel P, Brondízios E, Moran E (2004) Change detection techniques. Int J Remote Sens 25(12):2365–2407
Lunetta RS, Elvidge CD (1998) Remote sensing change detection: environmental monitoring methods and applications. Ann Arbor Press, Chelsea, p 318
Mas J-F (1999) Monitoring land-cover changes: a comparison of change detection techniques. Int J Remote Sens 20(1):139–152
Singh A (1989) Digital change detection techniques using remotely-sensed data. Int J Remote Sens 10(6):989–1003
WCRP – World Climate Research Programme, http://wcrp.wmo.int/. Accessed Nov 2006

Change of Support Problem

▸ Error Propagation in Spatial Prediction

Channel Modeling and Algorithms for Indoor Positioning

Muzaffer Kanaan, Bardia Alavi, Ahmad Hatami, and Kaveh Pahlavan
Center for Wireless Information Network Studies, Worcester Polytechnic Institute, Worcester, MA, USA

Synonyms

Indoor geolocation; Indoor location estimation; Indoor position estimation

Definition

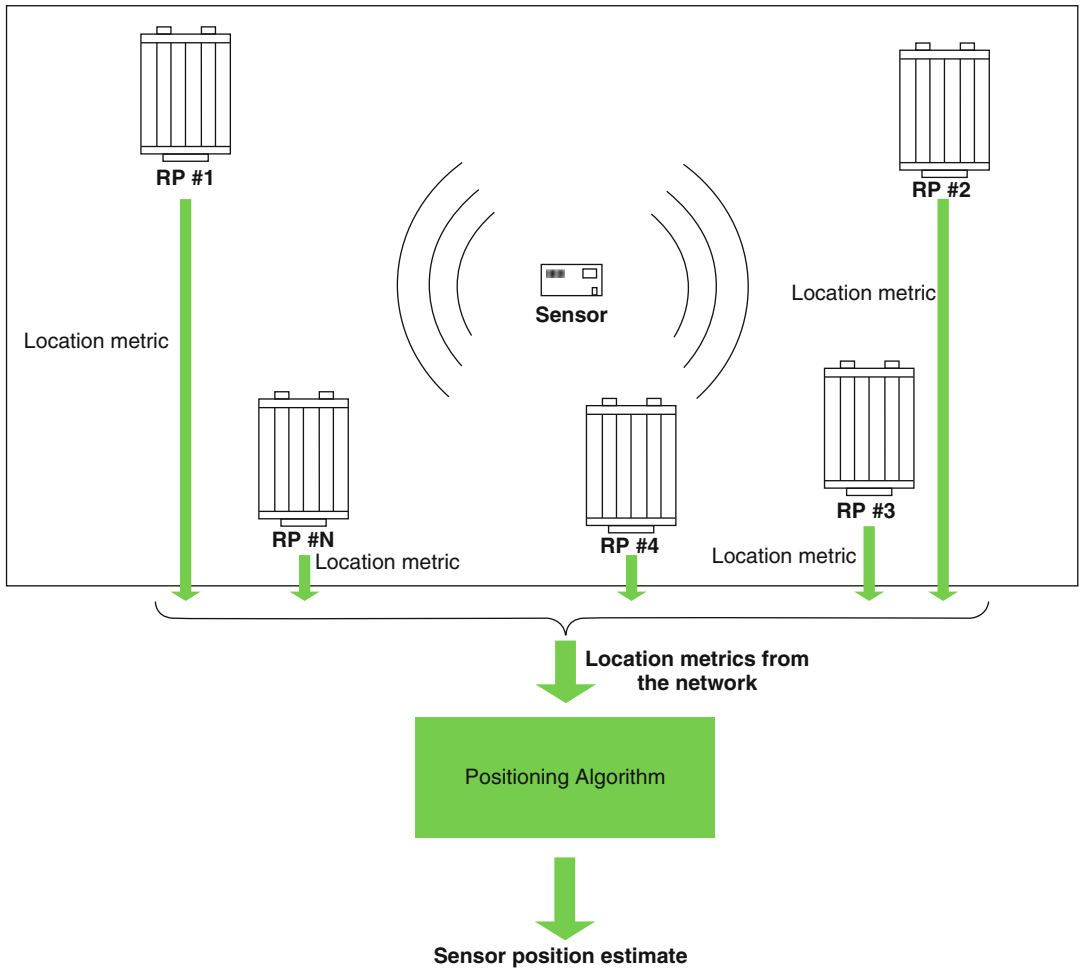
One of the new frontiers in wireless networking research is location awareness. Knowledge of a user's location enables a number of location-based services (LBS) to be delivered to that user. However, while this problem has been largely addressed for the outdoor environment, indoor positioning is an open area of research. Here, the problem of accurate indoor positioning is discussed, and the current state of the art in accurate position estimation techniques is reviewed.

Historical Background

Serious research in the field of positioning first began in the 1960s, when several US government agencies, including the Department of Defense (DoD), National Aeronautics and Space Administration (NASA), and the Department of Transportation (DOT) expressed interest in developing systems for position determination (Kaplan 1996). The result, known as the Global Positioning System (GPS), is the most popular positioning system in use today. Activity in this area continued after cellular networks flourished in the 1990s, driven largely by regulatory requirements for position estimation, such as E-911.

While these developments were taking place, similar research and development activity started in the field of indoor positioning, due to emerging applications in the commercial as well as public safety/military areas. In the commercial space, indoor positioning is needed for applications such as tracking people with special needs (such as people who are sight impaired), as well as locating equipment in warehouses and hospitals. In the public safety and military space, very accurate indoor positioning is required to help emergency workers as well as military personnel effectively complete their missions inside buildings. Some of these applications also require simple, low-power user terminals such as those that might be found in ad hoc sensor networks.

Positioning techniques developed for GPS and cellular networks generally do not work well



Channel Modeling and Algorithms for Indoor Positioning, Fig. 1 General structure of an indoor geolocation system. *RP* reference point

in indoor areas, owing to the large amount of signal attenuation caused by building walls. In addition, the behavior of the indoor radio channel is very different from the outdoor case, in that it exhibits much stronger multipath characteristics. Therefore, new methods of position estimation need to be developed for the indoor setting. In addition, the accuracy requirements of indoor positioning systems are typically a lot higher. For an application such as E-911, an accuracy of 125 m for 67 % of the time is considered acceptable (FCC 1996), while a similar indoor application typically requires an accuracy level on the order of only a few meters (Sayed et al.). In the next few sections, an overview of positioning techniques is provided for the indoor environment.

Scientific Fundamentals

Structure of a Positioning System

The basic structure of a positioning system is illustrated in Fig. 1, where a sensor (whose location is to be determined) is shown. The system consists of two parts: reference points (RPs) and the positioning algorithm. The RPs are radio transceivers, whose locations are assumed to be known with respect to some coordinate system.

Each RP measures various characteristics of the signal received from the sensor, which is referred to in this entry as *location metrics*. These location metrics are then fed into the positioning algorithm, which then produces an estimate of the location of the sensor.

The location metrics are of three main types:

- Angle of arrival (AOA)
- Time of arrival (TOA)
- Received signal strength (RSS)

This section is organized in four subsections; in the first three, each of these location metrics is discussed in greater detail, while the last is devoted to a nonexhaustive survey of position estimation techniques using these metrics.

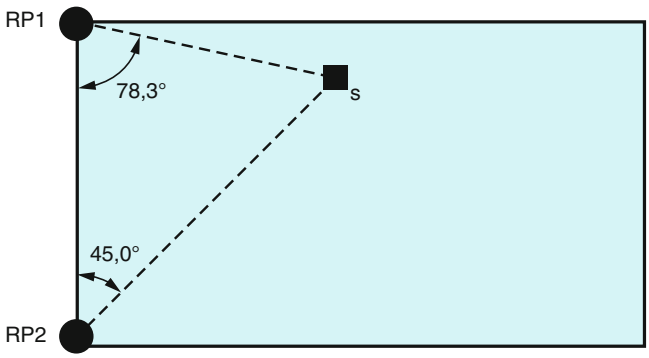
Angle of Arrival

As its name implies, AOA gives an indication of the direction the received signal is coming from. In order to estimate the AOA, the RPs need to be equipped with special antennae arrays. Figure 2 shows an example of the AOA estimation in an ideal nonmultipath environment. The two RPs measure the AOAs from the sensor as 78.3° and 45°, respectively. These measurements are then used to form lines of position, the intersection of which is the position estimate.

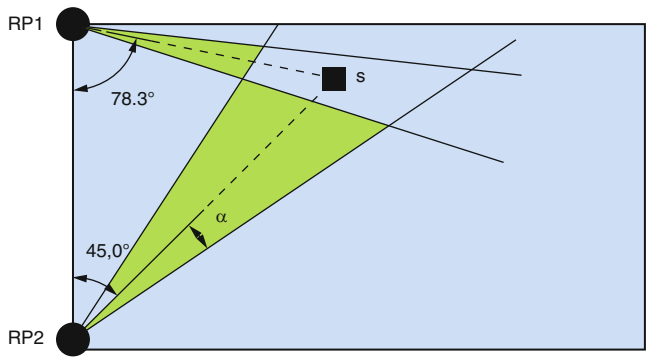
In real-world indoor environments, however, multipath effects will generally result in AOA estimation error. This error can be expressed as

$$\hat{\theta} = \theta_{\text{true}} \mp \alpha \tag{1}$$

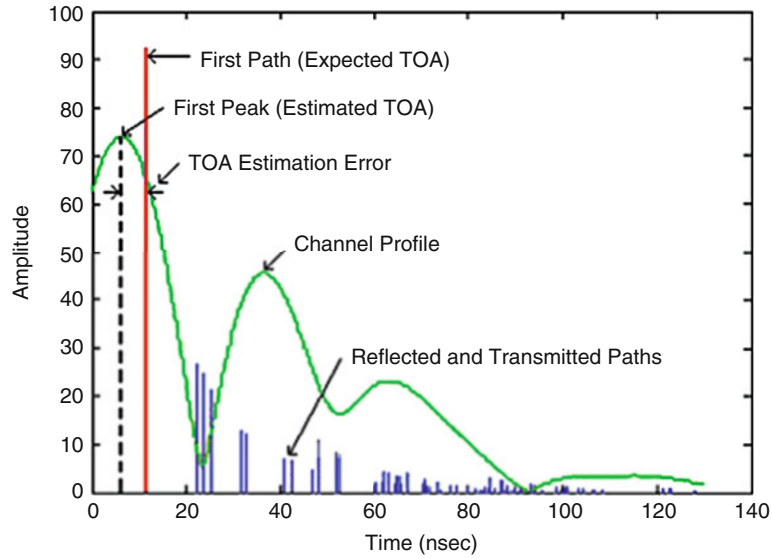
where θ_{true} is the true AOA value, generally obtained when the sensor is in the line-of-sight (LOS) path from the RP. In addition, $\hat{\theta}$ represents the estimated AOA, and α is the AOA estimation error. As a result of this error, the sensor position is restricted over an area defined with an angular spread of 2α, as illustrated in Fig. 3 below for the two-RP scenario. This clearly illustrates that in order to use AOA for indoor positioning, the sensor has to be in the LOS path to the RP, which is generally not possible.



Channel Modeling and Algorithms for Indoor Positioning, Fig. 2 Illustration of angle of arrival (AOA). *S* sensor



Channel Modeling and Algorithms for Indoor Positioning, Fig. 3 Illustration of AOA in the presence of multipath



Channel Modeling and Algorithms for Indoor Positioning, Fig. 4
Illustrating basic time of arrival (TOA) principles for positioning

Time of Arrival (TOA)

TOA gives an indication of the range (i.e., distance between a transmitter and a receiver). The basic concept can be illustrated with reference to the channel profile of Fig. 4 below. Since the speed of light in free space, c, is constant, the TOA of the direct path (DP) between the transmitter and the receiver, τ, will give the true range between the transmitter and receiver as defined by the equation:

$$d = c \times \tau \,. \tag{2}$$

In practice, the TOA of the DP cannot be estimated perfectly, as illustrated in Fig. 4. The result is *ranging error* [also referred to as the *distance measurement error* (DME) in the literature], given as

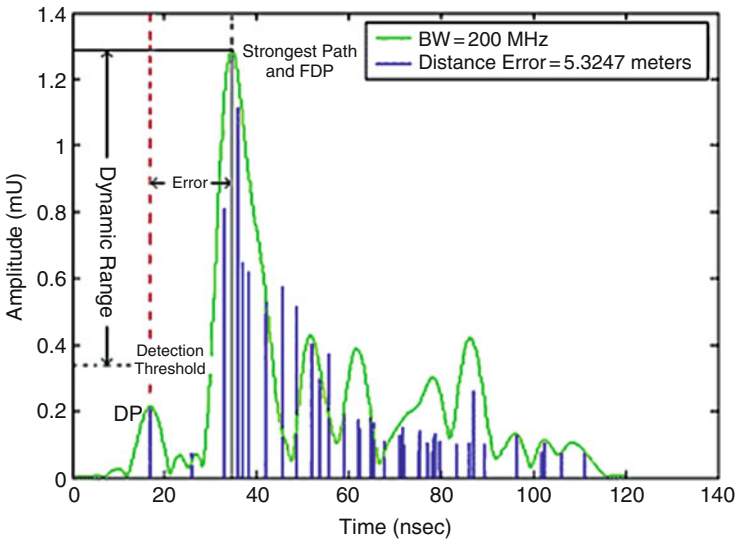
$$\varepsilon = \hat{d} - d \tag{3}$$

where $\hat{d}$ is the estimated distance and d is the true distance.

There are two main sources of ranging error: multipath effects and undetected direct path (UDP) conditions. Multipath effects will result in the DP, as well as reflected and transmitted paths to be received. It has been shown empirically that multipath ranging error can be reduced by increasing the bandwidth of the system used for the TOA estimation (Alavi and Pahlavan 2005). UDP conditions, on the other hand, refer to cases where the DP cannot be detected at all, as shown in Fig. 5 below. UDP conditions generally occur at the edge of coverage areas, or in cases where there are large metallic objects in the path between the transmitter and the receiver. As a result, the difference between the first detected path (FDP) and the DP is beyond the dynamic range of the receiver, and the DP cannot be detected, as shown in Fig. 5. Unlike multipath-based ranging error, UDP-based ranging error typically cannot be reduced by increasing the bandwidth. In addition, the occurrence of UDP-based ranging error is itself random in nature (Alavi and Pahlavan 2005).

Through UWB measurements in typical indoor areas, it has been shown that both multipath ranging error and UDP-based ranging error follow a Gaussian distribution, with mean and variance that depend on the bandwidth of operation (Alavi and Pahlavan 2005). The overall model can be expressed as follows:

$$\hat{d} = d + G(m_{\mathrm{w}}, \sigma_{\mathrm{w}}) \log(1 + d) + \zeta \cdot G(m_{\mathrm{UDP,w}}, \sigma_{\mathrm{UDP,w}}) \tag{4}$$



Channel Modeling and Algorithms for Indoor Positioning, Fig. 5 Illustration of undetected direct path (UDP)-based distance measurement error (DME) at a bandwith (BW) of 200 MHz. *FDP* stands for first detected path

where $G(m_w, \sigma_w)$ and $G(m_{UDP,w}, \sigma_{UDP,w})$ are the Gaussian random variable (RV) that refer to multipath and UDP-based ranging error, respectively. The subscript w in both cases denotes the bandwidth dependence. The parameter ζ is a binary RV that denotes the presence or absence of UDP conditions, with a probability density function (PDF) given as

$$f(\zeta) = (1 - P_{UDP,w})\,\delta(\zeta - 1) + P_{UDP,w}\delta(\zeta) \tag{5}$$

where $P_{UDP,w}$ denotes the probability of occurrence of UDP-based ranging error.

Received Signal Strength

RSS is a simple metric that can be measured and reported by most wireless devices. For example, the MAC layer of IEEE 802.11 WLAN standard provides RSS information from all active access points (APs) in a quasiperiodic beacon signal that can be used as a metric for positioning (Bahl and Padmanabhan 2000). RSS can be used in two ways for positioning purposes.

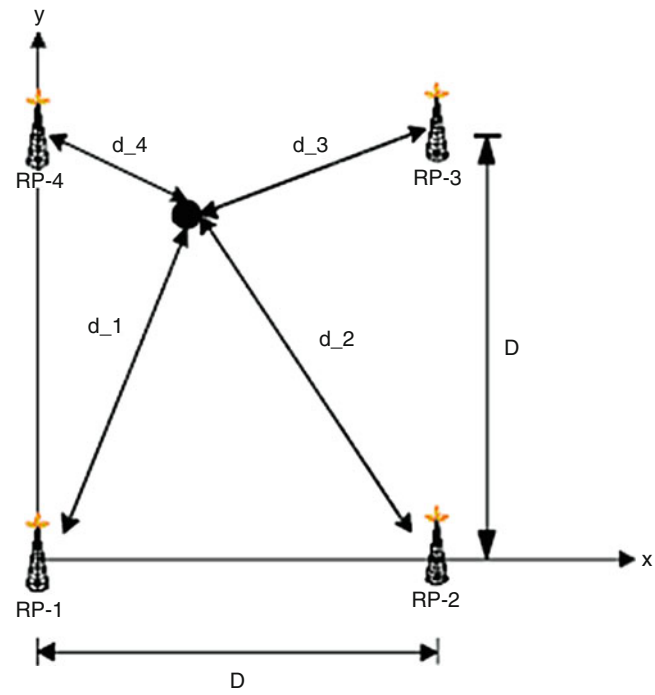
If the RSS decays linearly with the log-distance between the transmitter and receiver, it is possible to map an observed RSS value to a distance from a transmitter and consequently determine the user's location by using distances from three or more APs. In other words:

$$RSS_d = 10\log_{10} P_r = 10\log_{10} P_t - 10\alpha \log_{10} d + X \tag{6}$$

where α is the distance-power gradient, X is the shadow fading (a lognormal distributed random variable), P_r is the received power, and P_t is the transmitted power. While simple, this method yields a highly inaccurate estimate of distance in indoor areas, since instantaneous RSS inside a building varies over time, even at a fixed location; this is largely due to shadow fading and multipath fading. If, on the other hand, the RSS value to expect at a given point in an indoor area is known, then the location can be estimated as the point where the expected RSS values approximate the observed RSS values most closely. This is the essence of the *pattern recognition* approach to position estimation, which will be discussed in greater detail in the following section.

Position Estimation Techniques

Position estimation techniques can be categorized in a number of different ways. They can be grouped in terms of whether the sensing infrastructure used for measuring location metrics is deployed in a fixed or an ad hoc manner. They can also be grouped according to how the position computations are performed. In the category of *centralized algorithms*, all the location metrics



Channel Modeling and Algorithms for Indoor Positioning, Fig. 6
System scenario for performance evaluation

are sent to one central node, which then carries out the computations. In contrast, the term *distributed algorithms* refers to a class of algorithms where the computational load for the position calculations are spread out over all the nodes in the network. In the next few sections, some examples of centralized and distributed positioning algorithms for both fixed positioning and ad hoc scenarios will be discussed. Owing to space limitations, the treatment is by no means exhaustive; the interested reader is referred to Hightower and Borriello (2001) and Niculescu as well as any associated references contained therein.

Centralized Algorithms

In this section, two algorithms for fixed position estimation and one algorithm from ad hoc positioning are discussed. For fixed location estimation, the closest neighbor with TOA grid (CN-TOAG) (Kanaan and Pahlavan 2004) and ray-tracing assisted closest neighbor (RT-CN) algorithms (Hatami and Pahlavan 2006) are discussed. For ad hoc positioning, a distributed version of the least-squares (LS) algorithm is presented (Di Stefano et al. 2003).

CN-TOAG Algorithm

The CN-TOAG algorithm leverages the fact that at any given point in an indoor covered by a number of RPs, the exact value of the TOA is known Kanaan and Pahlavan (2004). Consider the grid arrangement of RPs in an indoor setting, as shown in Fig. 6. Each of these RPs would perform a range measurement, $d_i (1 < qi < qN$, where N is the number of RPs in the grid), to the user to be located.

Let $\mathbf{D}$ represent the vector of range measurements that are reported by the RPs, and let $\mathbf{Z}$ represent the vector of expected TOA-based range measurements at a certain point, $\mathbf{r} = (x, y)$. For the purposes of this algorithm, $\mathbf{Z}$ is known as the *range signature* associated with the point $\mathbf{r}$. An estimate of the user's location, $\hat{\mathbf{r}}$, can be obtained by finding that point $\mathbf{r}$, where $\mathbf{Z}$ most closely approximates $\mathbf{D}$. The error function, e($\mathbf{r}$)=e(x, y), is defined as

$$e(\mathbf{r}) = e(x, y) = ||\mathbf{D} - \mathbf{Z}(\mathbf{r})|| = ||\mathbf{D} - \mathbf{Z}(x, y)|| \tag{7}$$

where ||.|| represents the vector norm. Equation (7) can also be written as

$$e(x,y) = \sqrt{\sum_{k=1}^{N}\left(d_k - \sqrt{(x-X_k)^2+(y-Y_k)^2}\right)^2} \tag{8}$$

where N is the number of RPs, d_k is the range measurement performed by kth RP ($1 < qk < qN$), and (X_k, Y_k) represents the location of the kth RP in Cartesian coordinates (assumed to be known precisely). The estimated location of the mobile, $\hat{\mathbf{r}}$, can then be obtained by finding the point (x, y) that minimizes (8). This point can be found by using the gradient relation:

$$\nabla e(x,y) = \mathbf{0} \tag{9}$$

Owing to the complexity of the function in (2), it is not possible to find an analytical solution to this problem. CN-TOAG provides a numerical method of solving (9), detailed in Kanaan and Pahlavan (2004).

RT-CN Algorithm

The RT-CN algorithm is based on the RSS metric. The general idea is that the RSS characteristics of the area covered by the RPs are characterized in a data structure known as a *radio map*. Generally, the radio map is generated using on-site measurement in a process called training or fingerprinting. On-site measurement is a time- and labor-consuming process in a large and dynamic indoor environment. In Hatami and Pahlavan (2006) two alternative methods to generate a radio map without on-site measurements are introduced. The RT-CN algorithm uses two-dimensional ray-tracing (RT) computations to generate the reference radio map. During localization, mobile station (MS) applies the nearest neighbor (NN) algorithm to the simulated radio map and the point that is the closest in signal space to the observed RSS values. In this way, a very high-resolution radio map can be generated and higher localization accuracy results. In order to generate an accurate radio map in this technique, the localization system requires knowledge of the location of access points within the coverage area. In addition, a powerful central entity is required, both to perform the RT computations for the radio map and to execute the NN algorithm on the radio map to come up with the final position estimate. As such, it is an example of a centralized pattern recognition algorithm.

Distributed LS Algorithm

The algorithm that is featured in Di Stefano et al. (2003) is a distributed implementation of the steepest descent LS algorithm. The system scenario assumes ultrawide band (UWB) communications between sensor nodes. The sensor nodes perform range measurements between themselves and all the neighbors that they are able to contact. Then the following objective function is minimized using the distributed LS algorithm:

$$E = \frac{1}{2}\sum_{i}\sum_{j\in N(i)}\left(d_{ij} - \hat{d}_{ij}\right)^2 \tag{10}$$

where d_{ij} is the actual distance between two nodes i and j and $\hat{d}_{ij}$ is the estimated distance between the same two nodes. Assuming some transmission range R for every sensor node, $N(i)$ represents the set of neighbors for node i, i.e., $N(i) = \{j : d_{ij} < qR, i \neq j\}$.

Effects of the Channel Behavior on TOA-Based Positioning Algorithm Performance

Channel behavior is intimately linked with the performance of the positioning algorithms. As already noted above, the main effect of the channel is to introduce errors into the measurement of the metrics used for the positioning process. The precise manner in which these errors are introduced is determined by the quality of link (QoL) between the sensor and all the RPs that it is in contact with. In a TOA-based system, the exact amount of error from a given RP depends on whether UDP conditions exist or not. In this case, the channel is said to exhibit *bipolar* behavior, i.e., it suddenly switches from the detected direct path (DDP) state to the UDP state from time to

time and this results in large DME values. These will then translate to large values of estimation error; in other words, the quality of estimation (QoE) will be degraded (Kanaan et al. 2006).

Owing to the site-specific nature of indoor radio propagation, the very occurrence of UDP conditions is random and is best described statistically (Alavi and Pahlavan 2005). That being the case, the QoE (i.e., location estimation accuracy) will also need to be characterized in the same manner. Different location-based applications will have different requirements for QoE. In a military or public safety application (such as keeping track of the locations of firefighters or soldiers inside a building), high QoE is desired. In contrast, lower QoE might be acceptable for a commercial application (such as inventory control in a warehouse). In such cases, it is essential to be able to answer questions like: "What is the probability of being able to obtain a mean square error (MSE) of $1\,\mathrm{m}^2$ from an algorithm x over different building environments that give rise to different amounts of UDP?" or "What algorithm should be used to obtain an MSE of $0.1\,\mathrm{cm}^2$ over different building environments?" Answers to such questions will heavily influence the design, operation, and performance of indoor geolocation systems.

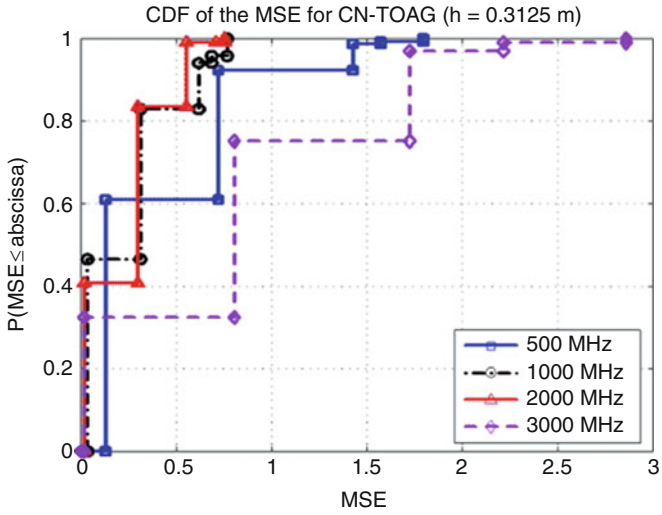
Given the variability of the indoor propagation conditions, it is possible that the distance measurements performed by some of the RPs will be subject to DDP errors, while some will be subject to UDP-based errors. Various combinations of DDP and UDP errors can be observed. To illustrate, consider the example system scenario shown in Fig. 6. For example, the distance measurements performed by RP-1 may be subject to UDP-based DME, while the measurements performed by the other RPs may be subject to DDP-based DME; this combination can be denoted as *UDDD*. Other combinations can be considered in a similar manner.

Since the occurrence of UDP conditions is random, the performance metric used for the location estimate (such as the MSE) will also vary stochastically and depends on the particular combination observed. For the four-RP case shown in Fig. 6, it is clear that the following distinct combinations will have to be used: *UUUU*, *UUUD*, *UUDD*, *UDDD*, and *DDDD*. Each of these combinations can be used to characterize a different *QoL class*. The occurrence of each of these combinations will give rise to a certain MSE value in the location estimate. This MSE value will also depend on the specific algorithm used. There may be more than one way to obtain each DDP/UDP combination. If UDP conditions occur with probability P_{udp}, then the overall probability of occurrence of the ith combination P_i can be generally expressed as

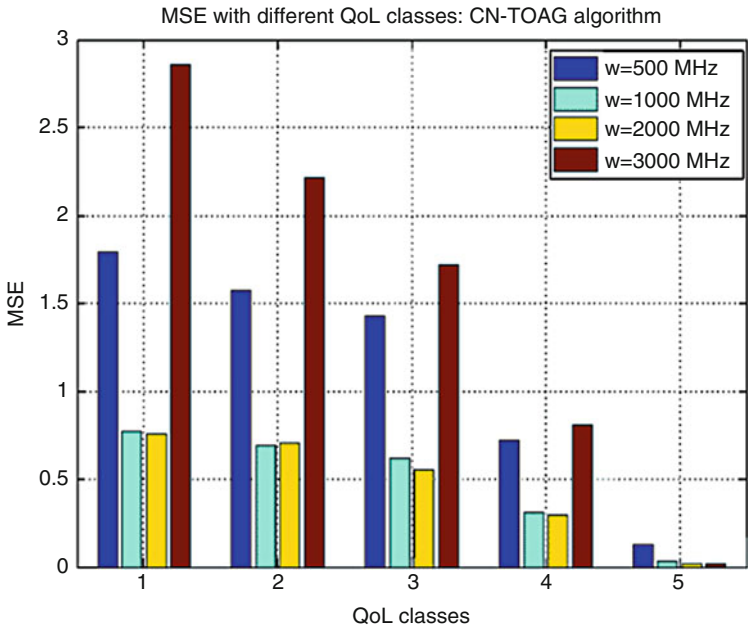
$$P_i = \binom{N}{N_{\mathrm{udp,i}}} P_{\mathrm{udp}}^{N_{\mathrm{udp,i}}} \left(1 - P_{\mathrm{udp}}\right)^{N - N_{\mathrm{udp,i}}} \quad (11)$$

where N is the total number of RPs (in this case four), and $N_{\mathrm{udp,i}}$ is the number of RPs where UDP-based DME is observed. Combining the probabilities, P_i, with the associated MSE values for each QoL class, a discrete cumulative distribution function (CDF) of the MSE can be obtained. This discrete CDF is known as the *MSE profile* (Kanaan et al. 2006). The use of the MSE profile will now be illustrated with examples, focusing on the CN-TOAG algorithm.

The system scenario in Fig. 6 is considered with $D = 20\,\mathrm{m}$. A total of 1,000 uniformly distributed random sensor locations are simulated for different bandwidth values. In line with the FCC's formal definition of UWB signal bandwidth as being equal to or more than 500 MHz (US Federal Communications Commission 2004), the results are presented for bandwidths of 500, 1,000, 2,000, and 3,000 MHz. For each bandwidth value, different QoL classes are simulated, specifically *UUUU*, *UUUD*, *UUDD*, *UDDD*, and *DDDD*. Once a sensor is randomly placed in the simulation area, each RP calculates TOA-based distances to it. The calculated distances are then corrupted with UDP- and DDP-based DMEs in accordance with the DME model based on UWB measurements as given in Alavi and Pahlavan (2005). The positioning algorithm is then applied to estimate the sensor location. Based on 1,000 random trials, the MSE is calculated for each bandwidth value and the corresponding combinations of UDP- and DDP-based



Channel Modeling and Algorithms for Indoor Positioning, Fig. 7 (MSE) Profile for the closest neighbor with TOA grid (CN-TOAG) algorithm



Channel Modeling and Algorithms for Indoor Positioning, Fig. 8 Quality of link (QoE) variation across the various QoL classes

DMEs. The probability of each combination is also calculated in accordance with (11).

The results are shown in Figs. 7 and 8. Figure 7 shows the MSE profiles for the CN-TOAG algorithm. From this plot, it is observed that as the bandwidth increases from 500 MHz to 2,000 MHz, the range of MSE profile values gets smaller. This correlates with the findings of Alavi and Pahlavan (2005), where it was observed that the overall DME goes down over this specific range of bandwidths. Above 2,000 MHz, however, the MSE profile becomes wider as a result of increased probability of UDP conditions (Alavi and Pahlavan 2005), which increases the overall DME. This, in turn, translates into an increase in the position estimation error. In order to gain further insight into the variation of the QoE across the different QoL classes, again considering bandwidth as a parameter, just the MSE is plotted, as seen in Fig. 8.

Key Applications

The applications of indoor localization technology are vast and can be broadly classified into two categories: commercial and public safety/military. Commercial applications range from inventory tracking in a warehouse to tracking children, elderly, and people with special needs (McKelvin et al. 2005). Location-sensitive web browsing and interactive tour guides for museums are other examples (Koo et al. 2003). In the public safety/military space, the most prevalent application is to help emergency workers (police, firefighters, etc.).

Accurate indoor localization is also an important part of various personal robotics applications (Jensfelt 2001) as well as in the more general context of context-aware computing (Ward et al.). More recently, location sensing has found applications in location-based handoffs in wireless networks (Pahlavan et al. 2000), location-based ad hoc network routing (Ko and Vaidya 1998), and location-based authentication and security. Many of these applications require low-cost, low-power terminals that can be easily deployed with little or no advanced planning; this is the basis for developments in ad hoc sensor networks. Recent developments in integrated circuit (IC) technology as well as microelectromechanical systems (MEMS) have made it possible to realize such low-cost, low-power terminals. In the next few years, there will undoubtedly be scores of new applications for indoor localization.

Future Directions

Indoor positioning is a relatively new area of research, and, as such, there are a number of different problems to be solved. Among these are questions such as: "What algorithms and techniques should be used to obtain a certain level of positioning error performance?" and "What is the best performance that can be obtained from a given positioning algorithm under UDP conditions?". Issues such as these will need to be looked at in order for the indoor positioning field to mature.

Cross-References

▸ Indoor Positioning

References

Alavi B, Pahlavan K (2005) Indoor geolocation distance error modeling with UWB channel measurements. In: Proceedings of the IEEE personal indoor mobile radio communications conference (PIMRC), Berlin, 11–14 Sept 2005

Bahl P, Padmanabhan VN (2000) RADAR: an in-building RF-based user location and tracking system. In: Proceedings of the IEEE INFOCOM 2000, Tel Aviv, 26–30 March 2000

Di Stefano G, Graziosi F, Santucci F (2003) Distributed positioning algorithm for ad-hoc networks. In: Proceedings of the IEEE international workshop on UWB systems, Oulu, June 2003

FCC Docket No. 94-102. Revision of the commissions rules to insure compatibility with enhanced 911 emergency calling systems. Federal Communications Commission Technical report RM-8143, July 1996

Hatami A, Pahlavan K (2006) Comparative statistical analysis of indoor positioning using empirical data and indoor radio channel models. In: Proceedings of the consumer communications and networking conference, Las Vegas, 8–10 Jan 2006

Hightower J, Borriello G (2001) Location systems for ubiquitous computing. IEEE Comput Mag 34(8): 57–66

Jensfelt P (2001) Approaches to mobile robot localization in indoor environments. Ph.D. thesis, Royal Institute of Technology, Stockholm

Kanaan M, Pahlavan K (2004) CN-TOAG: a new algorithm for indoor geolocation. In: Proceedings of the IEEE international symposium on personal, indoor and mobile radio communications, Barcelona, 5–8 Sept 2004

Kanaan M, Akgül FO, Alavi B, Pahlavan K (2006) Performance benchmarking of TOA-based UWB indoor geolocation systems using MSE profiling. In: Proceedings of the IEEE vehicular technology conference, Montreal, 25–28 Sept 2006

Kaplan ED (1996) Understanding GPS. Artech, Boston

Ko Y, Vaidya NH (1998) Location-aided routing (LAR) in mobile ad hoc networks. In: Proceedings of the ACM/IEEE international conference on mobile computing and networking, 1998 (MOBICOM'98), Dallas, 25–30 Oct 1998

Koo SGM, Rosenberg C, Chan H-H, Lee YC (2003) Location-based e-campus web services: from design to deployment. In: Proceedings of the first IEEE international conference on pervasive computing and communications, Fort Worth, 23–26 Mar 2003

McKelvin ML, Williams ML, Berry NM (2005) Integrated radio frequency identification and wireless sensor network architecture for automated inventory manage-

ment and tracking applications. In: Proceedings of the Richard Tapia celebration of diversity in computing conference (TAPIA'05), Albuquerque, 19–22 Oct 2005
Niculescu D. Positioning in ad-hoc sensor networks. IEEE Netw 18(4):24–29
Pahlavan K, Krishnamurthy P, Hatami A, Ylianttila M, Mäkelä J, Pichna R, Vallström J (2000) Handoff in hybrid mobile data networks. IEEE Personal Commun Mag 7:34–47
Sayed AH, Tarighat A, Khajehnouri N. Network-based wireless location. IEEE Signal Process Mag 22(4):24–40
US Federal Communications Commission (2004) Revision of Part 15 of the commission's rules regarding ultra-wideband transmission systems, FCC 02–48, First Report & Order, April 2004
Ward A, Jones A, Hopper A. A new location technique for the active office. IEEE Personal Commun Mag 4(5):42–47

Characteristic Travel Time

▸ Dynamic Travel Time Maps

Check-Out

▸ Smallworld Software Suite

Chronological

▸ Time-Aware Personalized Location Recommendation

Classification Integration

▸ Integration of Spatial Constraint Databases

Clementini Operators

▸ Dimensionally Extended Nine-Intersection Model (DE-9IM)

Climate Adaptation

▸ Climate Extremes and Informing Adaptation

C

Climate Adaptation, Introduction

Shahed Najjar, Udit Bhatia, and Auroop R. Ganguly
Sustainability and Data Sciences Laboratory (SDS Lab), Department of Civil and Environmental Engineering, Northeastern University, Boston, MA, USA

Synonyms

Adaptation; Learning; RFC; Scenario planning; Transformations

Definitions

Climate Change

Climate change is defined as changes in the state of the climate variables that can be identified (by using statistical tests) by changes in the mean and/or the variability of its properties and that persists for an extended period. An extended period in climate context implies decades or an even longer time scale (IPCC 2014). Climate change may be due to natural internal processes or external forcings and persistent anthropogenic changes in the composition of the atmosphere or in land use (Stocker et al. 2013).

Adaptation

In the context of climate and climate-related extremes, IPCC's Special Report on Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation (SREX) defines adaptation as, "the process of adjustment to actual or expected climate and its effects, in order to moderate harm or exploit benefit opportunities" (Field 2012).

Reasons for Concerns (RFCs)

Five integrative reasons for concerns (RFCs) provide a framework for summarizing key risks across sectors and regions. RFCs provide one starting point for evaluating dangerous anthropogenic interference with the climate system. IPCC's fifth assessment report highlights the following five RFCs (IPCC 2014):

1. Unique and threatened systems: Ecosystems and cultures that are already at risk from changing climate.
2. Extreme weather events: Amplified risks of extreme weather events such as precipitation extremes, heat waves, and coastal flooding in a changing climate.
3. Distribution of impacts: Uneven distribution of risk and pronounced impacts of these risks for disadvantages population and communities.
4. Global aggregate impacts: Mass extinctions in biodiversity, perturbed global transportation, communication, and trade networks as a consequence of increased risks from climate- and weather-related events in a changing climate.
5. Large-scale singular events: For example, there is medium confidence that sustained warming greater than a certain threshold could result in near-complete loss of the Greenland ice, which in turn will contribute up to 7 m of global mean sea level rise.

Nonstationarity

As climate change became an increasingly prominent topic, technical terms like "stationarity" and "nonstationarity" also became more noticeable. Nonstationarity can be defined as processes that have statistical properties that are deterministic functions of time. Demonstrating nonstationarity is more complex than stationarity because it is necessary to do so through analysis of the process physics. In the context of risk management and informing adaptation in the context of climate, flood frequency analysis of variables such as precipitation events, floods, temperature extremes, etc. is marked by an assumption that these variables conform to stationary, independent, and identically distributed (i.i.d.) random process. These assumptions are at odds with the recognition that there is high statistical confidence in the fact that climate varies naturally at all scales, both spatially and temporally, as well as in response to the anthropogenic activities.

Transformational Adaptation

IPCC SREX defines transformations as "The altering of fundamental attributes of a system (including value systems; regulatory, legislative, or bureaucratic regimes; financial institutions; and technological or biological systems)." While incremental adaptation aims to improve efficiency within existing systems, transformational adaptation may involve changing the fundamental attributes of these systems. Where vulnerability is high and adaptive capacity low, changes in climate extremes can make it difficult for systems to adapt sustainably without transformational changes.

Historical Background

Across the globe, specifically in urban coastal areas, some extreme weather events such as heat waves have become more frequent, cold extremes have become less frequent, and patterns of rainfall are likely changing (Mishra et al. 2015). Even if future greenhouse gases emissions were to be committed to lower levels, there is moderate to high confidence that the climate would continue to change for decades to come. It has been estimated that a heat wave of the same magnitude as the 2003 European heat wave could prove five times more lethal in a large American city in terms of mortality. In ecosystems, changing climate could reduce the productivity and abundance of species and induce mass extinctions and habitat changes (Urban 2015).

Society's need to manage changing environmental conditions is not new; people have been adjusting to their environment since the emergence of civilizations. Modern efforts to stabilize and protect our homes, livelihoods, and resources

in the face of a variable climate include the development of floodplain regulations, insurance, wildlife reserves, drinking water reservoirs, and building codes. However, these actions have been taken in response to a climate that has been relatively stable for many centuries.

While climate mitigation deals with energy and economic policies to "avoid the unmanageable," climate adaptation is about engineering the coupled natural-built human system to "manage the unavoidable." A survey at the World Economic Forum ranked failure of climate adaptation and mitigation as well as greater incidences of extremes weather as two of the top ten global risks of highest concern. Two other concerns among the top ten were water and food crises, which in turn are strongly influenced by climate. The United States Department of Defense, an agency with a broad mission space encompassing the political, military, economic, social, informational, and infrastructural sectors, has called climate change a "threat multiplier" in their Quadrennial Defense Review Report.

Climate mitigation relies on the perceived urgency of the climate challenge by nations of the world. Developing nations may find the trade-offs particularly difficult to justify, given the immediate impacts on the poor and on the aspirations of their middle class, as well as the disparities in per capita and in historical emissions when compared to developed nations. However, the poorer sections of the developing nations are expected to have to bear the brunt of climate-related hazards and resource scarcity. Conversely, developed nations do not consistently rank climate change as among the highest of policy concerns but expect the developing nations to bear the burden of emission reductions by appealing not to per capita but to the total emissions per country. Belief systems, ranging from political ideologies and the ability of technological innovations to solve society's problems to humanity's manifest destiny, in addition to a host of cultural and historical experiences on individuals and nations, color the perceptions around climate mitigation. What adds to the complexity is that the costs of mitigation have to be borne by the current generation but the perceived benefits are to future generations and to the planet at large.

However, if mitigation is not an easy problem to address, climate adaptation is perhaps relatively better poised. In addition, adaptation becomes an absolute necessity if mitigation pathways fail to fully materialize and/or if historical emissions are already causing significant damage. The impacts of natural hazards such as hurricanes, floods, and heat or cold waves are felt immediately, and the scarcity of water, food, and energy resources affects lives and well-being. Thus, the need to adapt is often acceptable to even those who may not perceive climate change as a major threat or even as a driver of weather extremes or resource constraints. This is especially true if adaptation measures are "low regret" (e.g., reinforces what needed to be done anyway irrespective of the nature of climate change in any given region), even though there may be occasions when transformational adaptations may be the only strategy. The importance and near centrality of water to adaptation have been well recognized, both directly through water security and through their impacts on energy and food security. The resilience of interdependent critical lifelines and infrastructures, in sectors such as water (including waste water), transportation, energy (or power and fuel), and communications, has been recognized as an urgent societal need. Natural ecosystems, which may act as soft infrastructures (e.g., in coastal and/or urban regions where marine ecosystems could help slow down the impacts of sea level rise or reduce the strength of storm surges), may have interdependencies with built, or the hard, infrastructures. One step to adaptation is what has been sometimes called "translational climate science" or the ability to develop actionable yet credible insights from climate science through computational modeling and data sciences.

Incremental and Transformative Adaptation

Incremental adaptations to climate change can be thought as extensions of actions and behaviors that already reduce the loss and enhance the benefits of variations in changing climate and weather extremes. Incremental adaptations are

doing slightly more of what is already being done to deal with natural variation in climate and with extreme events. However, transformative adaptation measures seek to change the basic attributes of the systems that are affected or likely to be affected by the variations. Kates et al. classify transformation adaptation into three broad categories which are discussed in further detail in this section (Kates et al. 2012):

1. **Adaptation new to resources/location**: Examples include introduction of technologies into places where they have been not used before. This can either be done through technology transfer or by technological inventions relevant for the location. Environmental, human-induced changes and mass migrations have posed serious challenges or a serious challenge to urban coastal megacities. As a consequence, these cities are facing an ever-growing challenge of food and water security. For instance, the population of Chennai, an urban coastal metropolitan city on the Eastern coast of India, has increased by fourfolds in the last five decades (Chennai City Population Census 2011). When the city was facing an unprecedented crisis of drinking water, introduction of seawater desalination plants has proved to be transformative adaptation to drinking water problem.
2. **Enlarged scale or intensity**: Incremental adaptations can become transformative when they are used at a greater scale with much larger effects. This kind of adaptation measures generally requires a system-level view with an underlying philosophy that the whole is greater than the sum of its parts.
3. **Different places and locations**: Some adaptations collectively transform place-based hu-

Climate Adaptation, Introduction, Table 1 Summary of sectorial risks in changing climate and potential adaptation strategies

Sector	Key risks	Potential adaptation strategies
Water resources	• Drought frequency likely to increase by the end of the twentieth century • Raw water quality likely to reduce • Increased concentration of pollutants during droughts	• Adaptive water management strategies • Scenario planning • Low regret solutions (IPCC 2014)
Ecosystems • Terrestrial • Marine • Inland	• Increasing ocean acidification in medium to high emission scenarios to impact population dynamics, physiology, and behavior of marine species • Carbon stored in terrestrial biosphere susceptible to loss to atmosphere as a consequence of deforestation and climate change • Increased risk of species extinction and habitat migrations	• Promote genetic diversity • Assisted migration and dispersal from severely impacted ecosystems • Manipulation of disturbing regimes (such as forest fires, coastal flooding)
Urban areas	• Heat stress • Inland and coastal flooding • Drought and water scarcity • Risks amplified by lack of resilient infrastructure systems	• Multilevel urban risk governance • Including voice of low-income groups in informing policy • Building resilient infrastructure systems
Rural areas	Moderate to severe impact on: • Food security • Agriculture income • Shifts in production area of crops • Freshwater availability	• Trade reforms and investments in rural areas • Adaptations for agriculture and water through policies taking account of rural decision-making contexts

man environment systems or shift such systems to other locations. Resettlement associated with climate variability, and, by some accounts, climate change per se, is already under way in a few locations. This category of transformations becomes imperative when risks have increased beyond the threshold, where incremental transformations and even technology transformations may not result in a significant positive effect.

Future Directions: Risks and Opportunities for Adaptation

In the context of climate change, key risks refer to dangerous human-induced interferences with changing climate. Identifications of key risks are based on the following criteria:

(a) Large magnitude
(b) High probability or irreversibility of impacts
(c) Timing of impact
(d) Persistent vulnerability
(e) Limited potential to reduce risks through mitigation or adaptation

Some examples of key risks in changing climate include:

(a) Risk of disruption of livelihoods in low-lying coastal zones and small islands due to sea level rise, storms, and coastal flooding (Aerts et al. 2014).
(b) Risk of deaths and mass migrations of large urban populations as a consequence of inland flooding.
(c) Systematic risks due to extremes leading to breakdown of infrastructure networks (Bhatia et al. 2015).
(d) Increased risk of mortality and illness during extreme period of heats, particularly for vulnerable urban population (Meehl and Tebaldi 2004).
(e) Risk of loss of biodiversity in terrestrial, marine, and/or inland water ecosystems.
(f) Risk of food insecurity linked to warming, drought, flooding, and extreme precipitation events, particularly for poorer populations in urban and rural settings.

Sectors that are likely to be impacted by the key risks include freshwater resources, ecosystems (include terrestrial, marine, and inland),

Climate Adaptation, Introduction, Table 2 Summary of regional risks in changing climate, climate drivers, and potential adaptation strategies. Climate drivers and adaptation strategies for given key risk are identified by the same number (Adapted from IPCC 2014)

Region	Key risk	Climate driver	Potential adaptation strategies
Australasia	I. Increased risk in riverine, coastal, and urban flood II. Heat waves: increased risk of heat-wave-related mortality, forest fires, decreasing crop output/hectare III. Significant change in community composition and structure of coral reef systems in Australia IV. Increased risk of drought-related water and food shortage in South Asia and the Indian subcontinent	I. Extreme precipitation, cyclones, and sea level rise II. Warming trends and extreme temperature events III. Warming trends and cyclones IV. Extreme temperature, drying trends, and warming trends	I. Exposure reduction and protecting natural barriers (e.g., mangroves) II. Heat health warning systems, urban planning to reduce heat islands, and new work practices to avoid heat stress among outdoor population III. Direct interventions such as assisted colonization and shading and reducing stresses such as fishing, tourism IV. Integrated water resource management, water infrastructure and reservoir development, water reuse, and desalinated sea water usage in coastal areas

(continued)

Climate Adaptation, Introduction, Table 2 (continued)

Region	Key risk	Climate driver	Potential adaptation strategies
Central and South America	I. Decreased food production and quality II. Par Waterborne diseases III. Water availability in Central America and extreme precipitation events resulting in floods and landslides	I. Precipitation, extreme precipitation, temperature, and warming trends II. Warming trends and extreme precipitation III. Drying trends, warming trends, and extreme precipitation	I. Strengthening traditional indigenous systems and developing a new variety of crops more adaptable to temperature and droughts II. Developing early warning systems for disease control III. Programs to extend public health services
North America	I. Wildfire induces loss of ecosystem integrity and human mortality II. Heat-related human mortality III. Urban floods in riverine and coastal area resulting in ecosystem damage, human mortality, mass migrations, and infrastructure damage	I. Warming trends and drying trend II. Extreme temperature and warming trends III. Extreme precipitation and cyclones	I. Introducing resilient vegetation and prescribed burning II. Early heat warning systems, cooling centers, residential air conditioning, and community and household scale adaptations through family support III. Low impervious surface pavement designs, updating old rainfall-based infrastructure design to reflect current and changing climate conditions, and protecting natural flood barriers (e.g., mangroves)
Europe	I. Significant reduction in water availability from river abstraction and from groundwater resources, combined with increased water demand (e.g., for irrigation, energy and industry, domestic use) and with reduced water drainage and runoff as a result of increased evaporative demand II. Increased economic losses and people affected by extreme heat events: impacts on health and well-being, labor productivity, crop production, air quality, and increasing risk of wildfires in southern Europe and in Russian boreal region	I. Drying trend, warming trend, and extreme temperature II. Extreme temperature	I. Implementation of best practices and governance instruments in river basin management plans and integrated water management II. Implementation of warning systems III. Adaptation of dwellings and workplaces and of transport and energy infrastructure IV. Reductions in emissions to improve air quality V. Improved wildfire management VI. Development of insurance products against weather-related yield variations

food sector, infrastructure sector, urban and rural areas, and human health. Table 1 summarizes the selected key risks and possible adaptation scenarios for these risks in changing climate.

Risks will vary through time across regions and populations, dependent on innumerable factors including the extent of adaptation and mitigation. Moreover, adaptation is region and context specific, and with no universal strategy to reduce the risk, characterizing the risks and understanding context and place-based adaptation strategies are critical to inform adaptation strategies. Table 2 summarizes the selected regional risks and feasible adaptation scenarios for Australasia

(Australia Asia), Europe, North America, and Central America (IPCC 2014).

Cross-References

- ▸ Climate Change and Developmental Economies
- ▸ Climate Extremes and Informing Adaptation

References

Aerts JCJH, Botzen WJW, Emanuel K, Lin N, de Moel H, Michel-Kerjan EO (2014) Evaluating flood resilience strategies for coastal megacities. Science 344:473–475. doi:10.1126/science.1248222

Bhatia U, Kumar D, Kodra E, Ganguly AR (2015) Network science based quantification of resilience demonstrated on the Indian railways network. PLoS ONE 10:e0141890. doi:10.1371/journal.pone.0141890

Chennai City Population Census 2011 | Tamil Nadu n.d. http://www.census2011.co.in/census/city/463-chennai.html. Accessed 3 Sept 2016

Field CB (2012) Managing the risks of extreme events and disasters to advance climate change adaptation: special report of the intergovernmental panel on climate change. Cambridge University Press, New York

IPCC (2014) Climate change 2014: impacts, adaptation, and vulnerability. Part B: regional aspects. Contribution of working group II to the fifth assessment report of the intergovernmental panel on climate change [Barros VR, Field CB, Dokken DJ, Mastrandrea MD, Mach KJ, Bilir TE, Chatterjee M, Ebi KL, Estrada YO, Genova RC, Girma B, Kissel ES, Levy AN, MacCracken S, Mastrandrea PR, White LL (eds)]. Cambridge University Press, Cambridge/New York

Kates RW, Travis WR, Wilbanks TJ (2012) Transformational adaptation when incremental adaptations to climate change are insufficient. Proc Natl Acad Sci 109:7156–7161. doi:10.1073/pnas.1115521109

Meehl GA, Tebaldi C (2004) More intense, more frequent, and longer lasting heat waves in the 21st century. Science 305:994–997. doi:10.1126/science.1098704

Mishra V, Ganguly AR, Nijssen B, Lettenmaier DP (2015) Changes in observed climate extremes in global urban areas. Environ Res Lett 10:24005. doi:10.1088/1748-9326/10/2/024005

Stocker TF, Qin D, Plattner GK, Tignor M, Allen SK, Boschung J et al (2013) Climate change 2013: the physical science basis. Intergovernmental panel on climate change, working group I contribution to the IPCC fifth assessment report (AR5), New York

Urban MC (2015) Accelerating extinction risk from climate change. Science 348:571–573. doi:10.1126/science.aaa4984

Climate and Human Stresses on the Water-Energy-Food Nexus

Laura Blumenfeld[1], Tyler Hall[1], Hayden Henderson[1,2], Lindsey Bressler[1], Catherine Moskos[1], Udit Bhatia[1], Poulomi Ganguli[1], Devashish Kumar[1], and Auroop R. Ganguly[1]

[1]Sustainability and Data Sciences Laboratory (SDS Lab), Department of Civil and Environmental Engineering, Northeastern University, Boston, MA, USA

[2]Department of Mechanical and Industrial Engineering, Northeastern University, Boston, MA, USA

Synonyms

Food security; Integrated assessment models; Mass migration; Resource scarcity; Resources management; WEF

Definition

Water, energy, and food are indistinguishably linked. Water is an input for producing agricultural goods in the fields and along the entire agro-food supply chain. Energy is required for food production and water distribution: to power agricultural and irrigation machinery and for processing and transportation of agricultural goods. Agriculture accounts for nearly 70 % of total water withdrawal across the globe, while food production and processing accounts for nearly 30 % energy consumption worldwide (Water 2014).

The synergies and trade-offs between water consumption, food production, and energy consumption are manifold:

- Using water to irrigate crops might promote food production, but it can also reduce river flows and hydropower potential.
- Growing bioenergy crops under irrigated agriculture can increase overall water withdrawals and endanger food security.

- Converting surface irrigation into high-efficiency pressurized irrigation may save water but may also result in higher energy use.

Water, energy, and food (WEF) represent the greatest global risks because they are expected to be highly impacted by climate change, demographic shifts including mass migrations, aging infrastructure, global trade networks, and other challenges of the twenty-first century (Andrews-Speed et al. 2012). The nexus approach considers the different dimensions of water, energy, and food equally and recognizes the interdependencies of different resource uses to develop sustainably (Bazilian et al. 2011).

Historical Background

In 2011, Texas experienced its most extreme drought on record, stressing the ability of the power grid to meet demand. A 2013 study discussed the water-energy nexus in the context of Texas droughts. The electricity supply grid in Texas is unique in that it is almost entirely separated from the rest of the power infrastructure in the USA. As such, there is limited capacity to purchase power from other geographic regions in case of a generation deficit. In the event of a statewide drought, this isolation presents vulnerability. Texas also encompasses a range of climates. In the subhumid eastern half of the state, most power plants (70 %) use once-through cooling and draw from surface water, most often reservoirs. In the semiarid west, power plants use wet cooling systems to minimize water demand, which is met mostly with groundwater. During 2011 Texas experienced 100 days of above 100°F temperatures and a record low level of precipitation. The combination of high demand, low rainfall, and higher temperatures increased evaporation, lowering the state's reservoirs by 30 % compared to the previous year. At one point 88 % of the state was experiencing "exceptional" drought. The drought was accompanied by greater electricity demand for air conditioning, leading to a 6 % rise in electricity use. For three days in August, peak demand was so high that utilities shut off 1.5 GW of nonessential industrial loads to avoid instating rolling blackouts (Texas).

Given the severity of the drought, the Texas power system demonstrated exceptional resiliency. As a state prone to such dry weather, most power plants have either been built or retrofitted with equipment to ensure operability with restricted water use. Natural gas, for example, has become a major source of electricity in Texas and requires no cooling water if used in a combustion turbine. The construction of efficient combined-cycle power plants reduces water use per unit generation. Many plants that rely primarily on once-through cooling have supplementary cooling towers for use during drought conditions. One plant even has an 8.5-mile pipeline to bring cooling water from a secondary source. Lastly, wind power has seen enormous growth in Texas during the past decade, with 10 GW capacity now installed. Wind power requires no cooling water. As evidence of the effectiveness of these alternatives, not a single power plant was cited for water discharges above the allowable level during the drought.

The population of Texas is projected to increase dramatically in the coming decades, and infrastructure planners are working on new ways to ensure that electricity demand is met even under extreme drought. Some have suggested adding supplemental cooling towers to all plants, but critics argue that this option is too costly. Rather, those critics recommend wisely choosing what type of new generation and cooling systems to build. These include dry cooling systems that, while expensive, use air rather than water for cooling. To meet the rising demand for cooling water, the Texas State Water Plan calls for the construction of 26 new reservoirs. Some are advocating the increased utilization of groundwater resources, specifically using aquifers to store water and eliminate evaporative losses. This is a common practice in California, Arizona, and Florida but has yet to be implemented in Texas. Another option is drawing water from non-freshwater sources, including treated wastewater, brackish water, and seawater.

Seawater in particular is an untapped resource, accounting for 30 % of cooling withdrawals in the whole USA, but only 2 % in Texas.

A 2014 Pew survey at the World Economic Forum ranked the top ten global risks; water crises ranked third after fiscal crises and unemployment, closely followed by failure of climate mitigation and adaptation, as well as greater incidence of extreme events such as floods, storms, and fires, in fifth and sixth positions, respectively, and food crises in the eighth. Water and climate are not only interrelated to each other but as noted in the recent US Department of Defense Quadrennial Defense Review report act as threat multipliers for energy and food crises, in addition to influencing governance failures and global conflicts.

Certain areas of the world are at a greater risk of experiencing water stress than others, with larger potential for stresses on the water-food-energy nexus. Unfortunately, many of the water-stressed regions happen to be in geopolitically sensitive regions of the world. The region encompassing the Middle East and North Africa is particularly at risk; it is home to 6.3 % of the world's population but only 1.4 % of the world's renewable fresh water. Researchers quantify water availability by measuring the amount of annual renewable freshwater per person. Less than one thousand cubic meters per person per year is considered "water scarce" while having between one thousand and one thousand eight hundred cubic meters per person classifies a country as "water stressed." Currently, there are 15 water-scarce countries in the world, and 12 of them are located in the Middle East and North Africa. On the other hand, Israel happens to be among the most technologically advanced nations in terms of water technologies related to water use, including for agriculture, as well as desalination. The relatively unique perspective of and about Israel, with advanced water technologies, in the water-scarce Middle East region, leads to the possibility of intriguing treaties occasionally negotiated in outside of full public view. Water sharing considerations lead to intriguing geopolitical considerations in the Indian subcontinent as well, with the Himalayan water towers supplying a majority of freshwater to India, Pakistan, and Bangladesh, including the volatile flashpoints of the India-Pakistan borders. However, a 1960 Indus Water Treaty signed by both nations has been in place for over 50 years and is considered to be one of the most successful water sharing agreements. The issues are made more complicated by the fact that the major rivers with sources in the Himalayas originate from Tibet in China. Other flashpoints include the Nile river basin (Fig. 1).

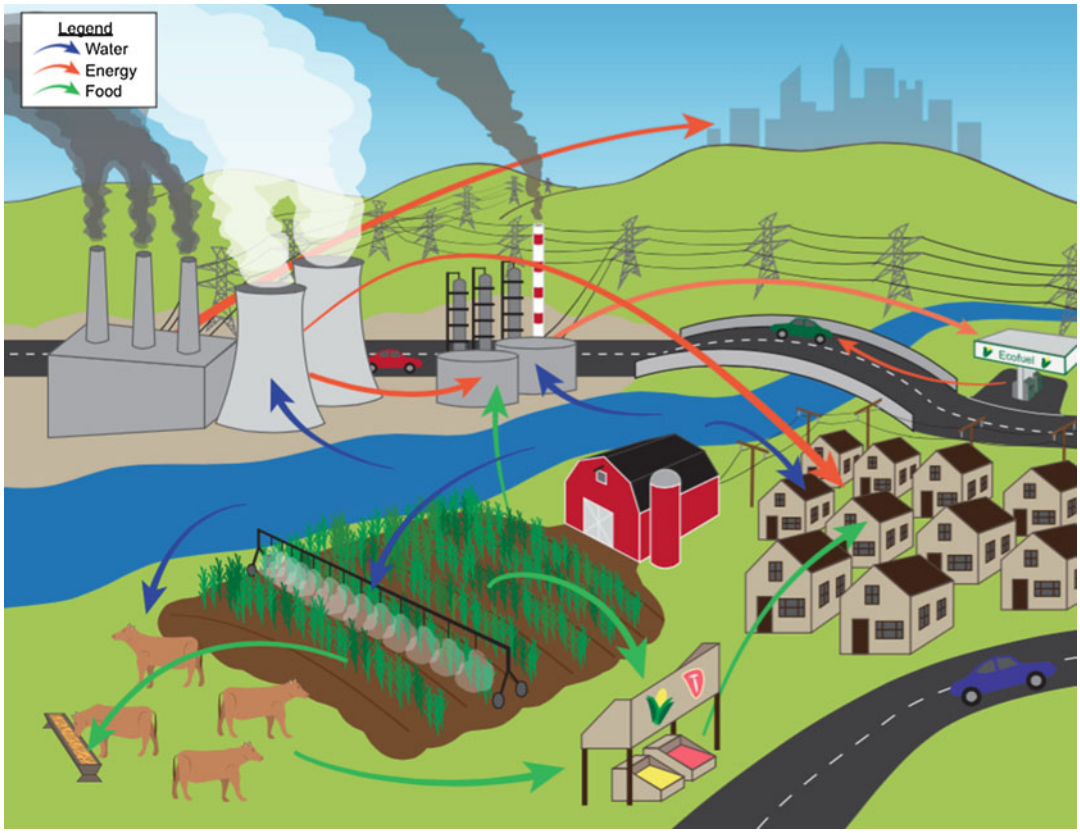
Scientific Fundamentals

Water-Energy Nexus

Nearly 90 % of the electricity produced in the USA requires water for cooling, amounting to about 170 billion gallons of water withdrawn and six billion gallons consumed daily. Thermoelectric power plants boil water into steam to drive a turbine and then discharge the remaining heat energy into a flow of cooling water (Fig. 2).

Thermoelectric power plants get energy from a variety of fuels and utilize several types of cooling systems. The most basic distinction among water use in these systems is withdrawal versus consumption. Withdrawal refers to water that is pumped through the plant and discharged back to the source. Withdrawn water remains available for other uses downstream, such as irrigation. Consumption refers to water that is used in the cooling process and not returned to the source. For example, cooling systems that evaporate water have high consumption rates (Rutberg and Michael 2012).

The amount of water consumed by a power plant depends on the type of the cooling system installed. "Once-through" systems pull water from the source, use it for cooling, and then return it to the source. As this method withdraws very large quantities, it requires an abundant water supply. The benefit of such systems is that little water is consumed. One other typical cooling system is "wet cooling," which runs water through a cooling tower. Cooling towers make use of evaporation to reduce the temperature of the water. While this method withdraws much less water

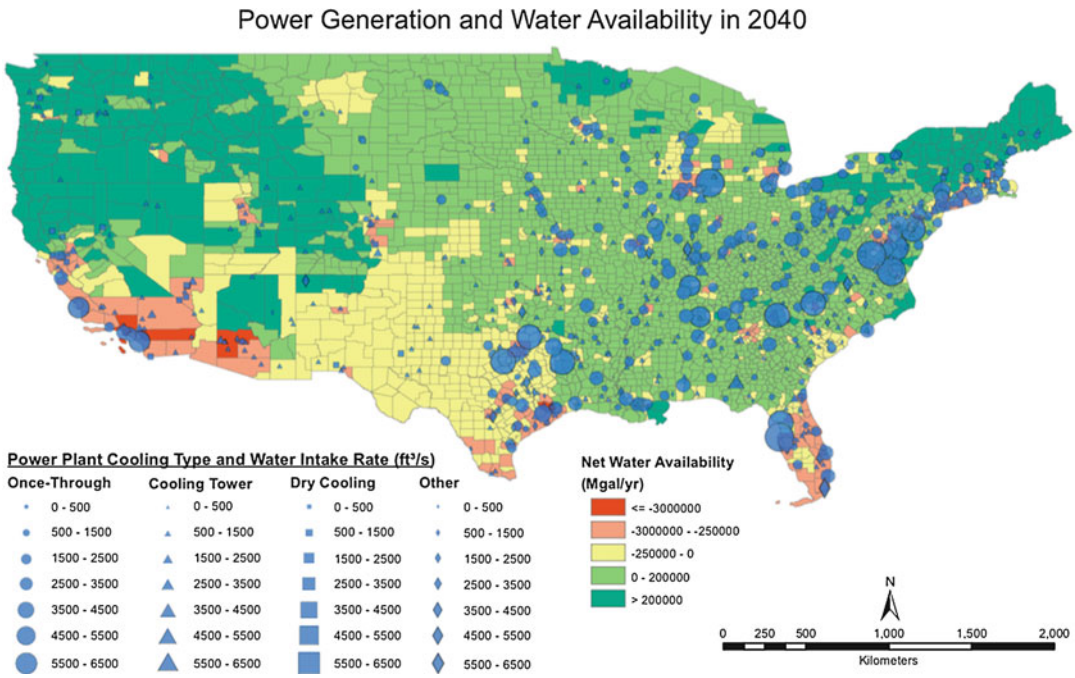


Climate and Human Stresses on the Water-Energy-Food Nexus, Fig. 1 A simplified schematic view of the relationship between water, food, and energy. Water is used for cooling thermoelectric power plants, which produce electricity for other activities. Water is also essential for agriculture and food. Some agricultural products are refined into fuels in a process that connects water, energy, and food

than once-through systems, it has a higher rate of consumption. A third type of cooling method is known as "dry cooling." These systems circulate water in a closed loop and dissipate the heat to the air through a large heat exchanger, similar to the radiator in a car. "Dry cooling" technology has the benefit of consuming virtually no water but is expensive, large, and dependent on the local climate. Some power plants use a combination of wet and dry cooling. These are called "hybrid" systems. Dammed hydroelectric power production relies on an adequate supply of water; lower water levels in reservoirs correlates to reduced generating capacity. Extracting a single gallon of oil requires between 2 and 350 gallons of water. Growing biofuel crops has a higher water intensity than any method of fossil fuel extraction: corn and soy irrigation each consume upward of 10,000 gallons of water for each MMBTU of fuel produced. The refining of petroleum and plant-based fuels also requires large amounts of water, in the range of 1–2 billion gallons every day. While this chapter focuses on water for cooling, water plays a vital role throughout the extraction, refining, and transport of fuels as well (DoE US 2006).

Climate Impacts on Water and Consequences on Power Generation

As mentioned at the outset of this chapter, climate change will have wide reaching effects on the water cycle, altering the temperature and quantity of water available in all regions of the country. Also as discussed, electricity generation is highly



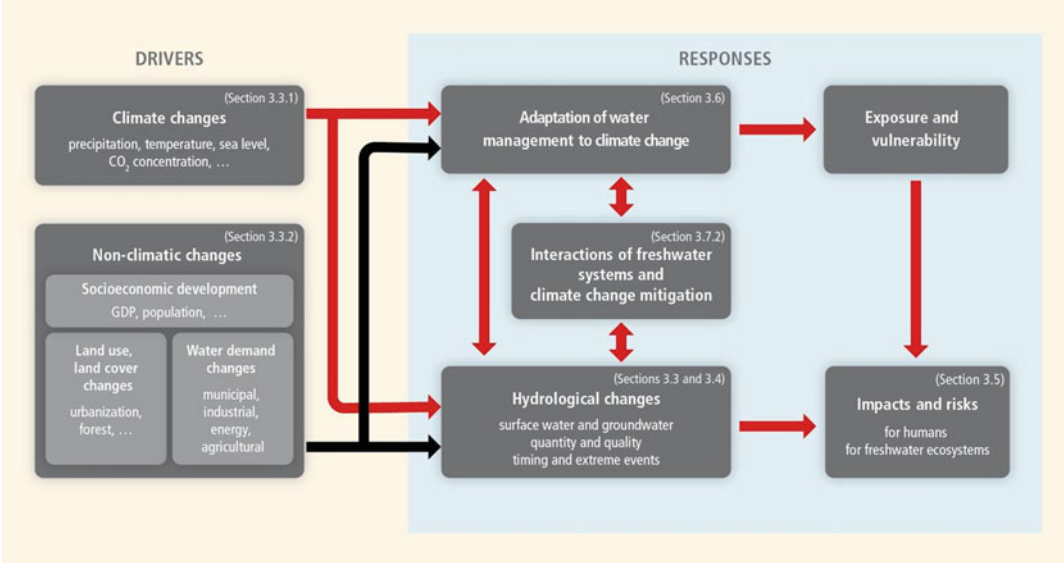
Climate and Human Stresses on the Water-Energy-Food Nexus, Fig. 2 Power plant information from the Energy Information Agency (EIA) overlaid with county-level water availability projections for 2040. Power plants are broken down by the type of cooling system employed (Adapted from Ganguly et al. 2015) and the rate of water intake. Water availabilities are median values taken from an ensemble of precipitation-minus-evaporation models. Values are less domestic demand, taken as per capita demand times the projected 2040 population

dependent on abundant sources of water for cooling. The confluence of these two relationships forms a water-energy nexus that will stress water supplies and potentially limit power production capacity in the future (Förster and Lilliestam 2009).

The relationship between energy production and water for once-through cooling systems can be broken down into following sub-relations:

- The temperature of the water discharged by a plant cannot exceed a specified level
- The mixed river and discharged water temperature must not exceed a specified level
- The temperature difference of discharged versus river water must not exceed a specified level
- The plant can only withdraw up to a certain fraction of the available stream flow
- The cooling system must not exceed its pumping capacity

With the exception of the last criterion, which is a mechanical limit, these regulations are put in place at the local, state, and federal levels to protect the aquatic environment. The first three criteria are temperature dependent. As water temperatures increase due to climate change, it becomes increasingly difficult for power plants to meet these discharge requirements. If the intake rate of cooling water is kept constant, higher intake temperatures equate to higher discharge temperatures. In some cases, the higher discharge temperature will exceed criteria 1 or 2. Alternatively, a cooling system may compensate for higher intake temperatures by increasing the intake rate. In this case criterion 4 limits plant operation, especially if the availability of water is reduced, as in drought conditions. Additionally, criterion 5 prevents the plant from drawing in more cooling water than the capability of the plant's equipment (Kimmell 2009). Water availability can also become an issue in light of criteria



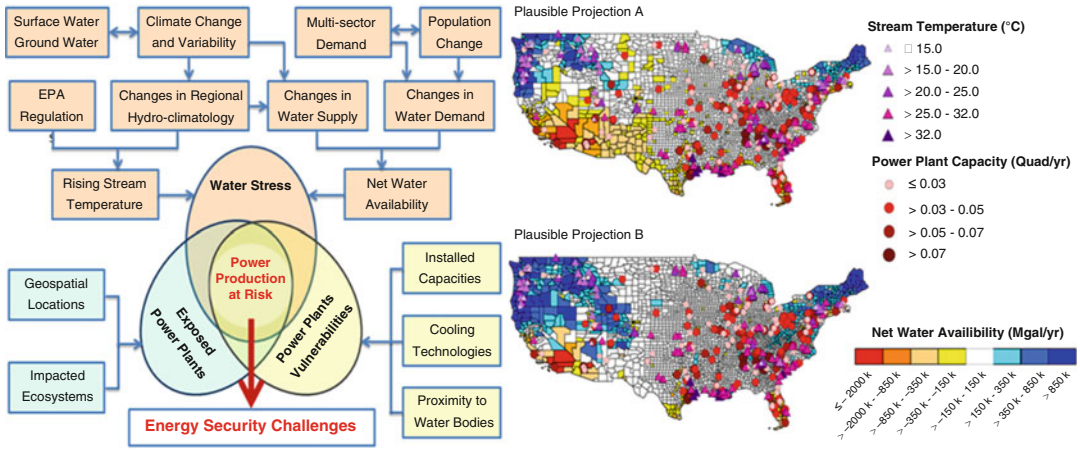
Climate and Human Stresses on the Water-Energy-Food Nexus, Fig. 3 A flowchart of water availability with both its drivers and the responses (Source: IPCC WG-II 2014 report Chapter 3 IPCC 2014)

3 and 4. In the event of low water levels, a plant may be forced to withdraw less water so as to satisfy criterion 4. As a result of the lower intake rate, the temperature of the discharged water will be higher than usual. In this case, criterion 1 limits the plant's generating capacity. In addition to the above 5 criteria, in certain cases reduced water availability may lower the water level in bodies of water from which power plants withdraw cooling water. If the water level falls below the intake level, the plant will be unable to intake sufficient quantities of cooling water. Many systems have intakes at depths shallower than 10 feet and may run dry under drought conditions. In any of these situations, power plants are forced to either reduce generation or shut down entirely. As a result, the availability of electricity is reduced. In regions at high risk of increased water temperatures and/or reduced water availability induced by climate change, power production is especially vulnerable (Fig. 3).

Climate Change Consequences on the Marine Food Web

Agriculture and water systems have an obvious connection: plants need water to grow and yield food. Aside from this main connection, climate influences the agriculture sector with changes in temperature, carbon dioxide levels, and water levels. Climate disruptions will cause a significant decrease in the yield of most crops and livestock because of changes in the atmosphere and the water availability. Changes in water availability will affect what crops will grow in certain areas and the amount of yield and hence influence food production around the world.

In a world with a growing population, the demand for food is growing, and the agriculture sector needs to increase production to match this projected growth. Without the necessary increase in agricultural lands, there would either need to be fundamental improvements in yield and management, consumption patterns, or both. However, consumption patterns are expected to increase, with large middle class populations in developing nations aspiring to the standards of living available in the First World countries. Water requirements vary by crop type, and scarcity or variability in water availability may reduce yield substantially. In a study on wheat, rice, maize, and soybean conducted by Parry et al., carbon dioxide levels, temperature, and water availability were projected to future levels to determine how the yield would be affected (Parry et al. 2004). The results illustrate slight to moderate negative impact worldwide with the most



Climate and Human Stresses on the Water-Energy-Food Nexus, Fig. 4 A proof of concept on aspects of the water-energy nexus. The process flow (*left*) considered water stress resulting from scarcer and warmer freshwater (Adapted from Ganguly et al. 2015) on thermoelectric power production based on climate model and population projections, which combined with stream temperature sensor and GIS data on energy infrastructures resulted in new insights (*right*)

potential changes in yield in Asia and Africa. The yields, according to this study, will fall by 10 % combined overall throughout all of the regions, which has the potential to affect the food security of the growing population, especially in developing economies.

Climate change is severely impacting marine food webs. Marine life is comprised within an intricate network; currents make nutrient-rich waters available to marine life, sustaining single-celled plants called phytoplankton, the zooplankton that feed on them, and the larger fish that eat the zooplankton and may be consumed as food by humans.

Climate change and associated changes in ocean circulations, sea level rise, and coastal winds are altering the patterns of nutrient upwelling and thereby changing the timing of fish spawning and the yield from fisheries (Wang et al. 2015). Rising sea temperatures are threatening the habitats of sea life by straining the ability of these creatures to cope with temperature changes. Marine mammals are being forced to travel longer distances to find food or to rely on less nutritious, energy-expending substrate for survival. Changing marine patterns are also affecting human food production and the fishing industry.

Climate, Humans, and WEF

Changes in population and lifestyles, changes in climate, resilience of infrastructures and societies, and regional land use and urbanization are the main stressors of the water-food-energy nexus. They impact water resources (both quantity and quality), water-related hazards (floods and droughts), built and natural infrastructures, and coupled natural-engineered human systems across many scales. The influence and impact of these stressors are critically dependent on time horizons, spatiotemporal resolutions, and the resilience of the coupled systems. As climate change influences water availability, additional stress will be placed on the balance between food and biofuel crops. As discussed earlier, biofuel crops are much more water intensive than traditional fuels. As such, the energy and food needs of people must be carefully considered in light of changes in the geographic distribution of water availability (Fig. 4).

Planning Horizons and Spatial Resolutions

The spectrum of water challenges operates at multiple space and time scales, and across disparate planning horizons, over which scientific insights, projections, and policy insights need

to be generated. They include near-real time to weeks, seasonal to interannual, and decadal to mid-century.

High-resolution predictions are possible over near-real time to weekly time scales. Beyond this range, chaotic characteristics and/or extreme sensitivity to initial conditions limits the predictability of hydrological and meteorological systems. Within these time frames, short-term monitoring and predictions may lead to urgent and immediate events and emergency management. Specific examples include flood and flash flood warnings, monitoring water quality of source lakes to generate advance warning of possible pathogens in drinking water, urgent warnings about crop destruction, or stoppage of power plant operations owing to lack of water during a drought or excess water because of floods.

Seasonal to interannual time scales encompass changes in climate patterns such as El Niño, seasonal changes in monsoons in the Southwest USA, seasonal floods in Iowa, the severity of winter in the Northeast USA, the number of tropical cyclones striking the Gulf Coast during the hurricane season, seasonal droughts in California, and regional energy production. Specific weather events are not predictable at these time scales, but average seasonal and annual patterns can be characterized.

Decadal to mid-century time scales, ranging from about 5–30 years in the future, are not predictable on a seasonal or even annual basis. However, physics-based climate models can project climate trends and variability based on assumed emission scenarios. Trends in global warming and changes in weather patterns start becoming prominent at these horizons. Also at this scale, variability in mean and extreme climate is expected to be large. While climate change and the associated uncertainty are expected to be a major factor in water challenges at these time horizons, demographic changes may dominate the impact on water resources. The intensity and severity of floods and droughts may not be predictable on a single-event, seasonal, or annual basis at these time horizons, but duration and frequency changes may be predictable. The nexus of water with food and energy can be examined and characterized together with their uncertainties, and appropriate changes can be made in emergency preparedness. Reinforcements and remediation measures to build the resilience of communities, engineered systems, and natural ecosystems may be revived or made more resilient based on this knowledge.

Beyond decadal to mid-century time scales, climate trends are expected to dominate over uncertainties. Projections for population and human systems in this range are not available and difficult to create. Similarly, infrastructure and technology changes are nearly impossible to foresee. Most stakeholder decision horizons do not extend to these scales. The climate change adaptation community, as well as the related integrated assessments community, has been working at these ranges with a variety of simplified models. Data and geographic information science may be useful at these scales primarily to develop predictability and uncertainty studies in climate, evaluate the performance of models of water, energy, and food systems, develop enhanced projections with uncertainty, and examine aggregate impacts in terms relevant for adaptation and mitigation. Water in the atmosphere, oceans, land, and biosphere significantly impact climate variability yet remain among the largest sources of uncertainty. Climate impacts regional hydrometeorology and the water balance, including availability and quality of surface and groundwater, and influences natural hazards such as floods and droughts. Water availability and temperature affect terrestrial and marine ecosystems, as well as energy and food security. Climate influences the probability of hazards such as floods and droughts at multiple space-time resolutions. Adapting to changing hazards requires the resilient design of critical infrastructures and effective management of key resources.

Key Applications

The challenge of understanding energy, water, and food policy interactions, and addressing them in an integrated manner, appears daunting. Comprehensive understanding of WEF nexus and subsequent impacts of the human and climate is required to:

- Assess the current state and pressures on natural and human resources systems
- Forecast expected demands, trends, and drivers on resources systems and interactions between water, energy, and food systems
- Delineate different sectorial goals, policies, and strategies in regard to water, energy, and food. This includes an analysis of the degree of coordination and coherence of policies, as well as the extent of regulation of uses.
- Need for planned investments, acquisitions, reforms, and large-scale Infrastructure;
- Inform key stakeholders, decision-makers, and user groups.

Future Directions

The key challenges/questions for future directions are fivefold:

1. What are the relationships among the interlinked stressors and the stressed systems across the components of the water-climate-energy-food-ecosystem nexus at different time and spatial scales?
2. Can remotely sensed and other information about lakes or rivers including water levels and quality, capacity, location and water use of power production, resilience of energy infrastructures, food crops and biofuels, as well as freshwater or marine ecosystems be related to each other through graphical dependencies to form interconnected network structures across the disparate systems of the nexus, with a view to understanding their systemic dependencies, feedback, and resilience?
3. What are the characteristics of the stressors, especially the attributes of their extremes, how do they impact the nexus as well as the individual components of the nexus, and how do failures or loss of functionality propagate along the tightly interconnected system of systems?
4. Can the future flows, feedback, and vulnerability along the nexus network, as well as the perturbations of the nexus owing to possible non-stationary behavior of the extreme stressors, be predicted across multiple time scales to enable short-term recovery and long-term preparedness?
5. How would uncertainties along the interconnected networks of the nexus be quantified, including in how the impacts of changes in the stressors and their extremes propagate along the nexus?

Cross-References

- ▸ Climate Extremes and Informing Adaptation
- ▸ Climate Hazards and Critical Infrastructures Resilience

References

Andrews-Speed P, Bleischwitz R, Boersma T, Johnson C, Kemp G, VanDeveer SD (2012) The global resource nexus: the struggles for land, energy, food, water, and minerals. Transatlantic Academy, Washington, DC

Bazilian M, Rogner H, Howells M, Hermann S, Arent D, Gielen D et al (2011) Considering the energy, water and food nexus: towards an integrated modelling approach. Energy Policy 39:7896–906. doi:10.1016/j.enpol.2011.09.039

DoE US (2006) Energy demands on water resources: report to Congress on the interdependency of energy and water, vol 1. U.S. Department of Energy, Washington, DC

Förster H, Lilliestam J (2009) Modeling thermoelectric power generation in view of climate change. Reg Environ Change 10:327–338. doi:10.1007/s10113-009-0104-x

IPCC (2014) Climate change 2014: impacts, adaptation, and vulnerability. Part B: regional aspects. Contribution of working group II to the fifth assessment report of the intergovernmental panel on climate change [Barros VR, Field CB, Dokken DJ, Mastrandrea MD, Mach KJ, Bilir TE, Chatterjee M, Ebi KL, Estrada YO, Genova RC, Girma B, Kissel ES, Levy AN, MacCracken S, Mastrandrea PR, White LL (eds)]. Cambridge University Press, Cambridge/New York

Ganguly AR, Kumar D, Ganguli P, Short G, Klausner J (2015) Climate adaptation informatics: water stress on power production. Comput Sci Eng 17:53–60. doi:10.1109/MCSE.2015.106

Kimmell TA, Veil JA, Division ES (2009) Impact of drought on U.S. steam electric power plant cooling water intakes and related water resource management issues. Argonne National Laboratory (ANL), Washington, DC

Parry ML, Rosenzweig C, Iglesias A, Livermore M, Fischer G (2004) Effects of climate change on global food production under SRES emissions and socio-economic scenarios. Glob Environ Change 14:53–67. doi:10.1016/j.gloenvcha.2003.10.008

Rutberg MJ, Michael J (2012) Modeling water use at thermoelectric power plants. Thesis, Massachusetts Institute of Technology
Texas NS. Dried out: confronting the Texas drought n.d. https://stateimpact.npr.org/texas/drought/. Accessed 22 Aug 2016
Wang D, Gouhier TC, Menge BA, Ganguly AR (2015) Intensification and spatial homogenization of coastal upwelling under climate change. Nature 518:390–394. doi:10.1038/nature14235
Water UN (2014) The United Nations world water development report 2014: water and energy. UNESCO, Paris

Climate Change

▸ Climate Extremes and Informing Adaptation
▸ Climate Risk Analysis for Financial Institutions

Climate Change and Developmental Economies

Lindsey Bressler[1], Kara Morgan[1], Allison Traylor[1], Hayden Henderson[1], Udit Bhatia[1], Babak Fard[1], Devashish Kumar[1], Rajarshi Majumder[2], Sourav Mukherji[3], Joyashree Roy[4], Matthias Ruth[5,6], and Auroop R. Ganguly[1]
[1]Sustainability and Data Sciences Laboratory (SDS Lab), Department of Civil and Environmental Engineering, Northeastern University, Boston, MA, USA
[2]Department of Economics, The University of Burdwan, Burdwan, West Bengal, India
[3]Organisational Behaviour & Human Resources Management, Indian Institute of Management Bangalore, Bengaluru, India
[4]Global Change Programme, Department of Economics, Jadavpur University, Kolkata, India
[5]Resilient Cities Lab, School of Public Policy and Urban Affairs, Boston, MA, USA
[6]Department of Civil and Environmental Engineering, Northeastern University, Boston, MA, USA

Synonyms

Adaptation; Kuznets curve; Mitigation; Sustainable development

Definitions

Climate Change: Climate change is defined as changes in the state of the climate variables that can be identified (by using statistical tests) by changes in the mean and/or the variability of its properties and that persists for an extended period. Extended period in climate context implies three decades or even longer time scale (IPCC 2014). Climate change may be due to natural internal processes or external forcings and persistent anthropogenic changes in the composition of the atmosphere and/or in land use (Stocker et al. 2013).

Adaptation: In the context of climate and climate-related extremes, Intergovernmental Panel on Climate Change Special Report on Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation (SREX) defines adaptation as "the process of adjustment to actual or expected climate and its effects, in order to moderate harm or exploit benefit opportunities" (Field 2012).

Mitigation: IPCC defines mitigation in its third assessment report as "an anthropogenic intervention to reduce the sources or enhance the sinks of greenhouse gases" (IPCC-TAR 2001). Although it is hard to completely distinguish mitigation from adaptation measures, the key difference is that mitigation reduces all impacts (positive and negative) of climate change and thus reduces the adaptation challenge, whereas adaptation is selective; it can take advantage of positive impacts and reduce negative ones.

Gini coefficient: Gini coefficient is a measure of statistical dispersion intended to represent the income distribution of residents in a nation. Developed in 1912, this is the most commonly used measure of inequality (van Ginneken 2003).

Sustainable development: Fundamental to our understanding of climate change and developmental economics is the term "sustainable development." The most commonly used definition for sustainable development comes from *Our*

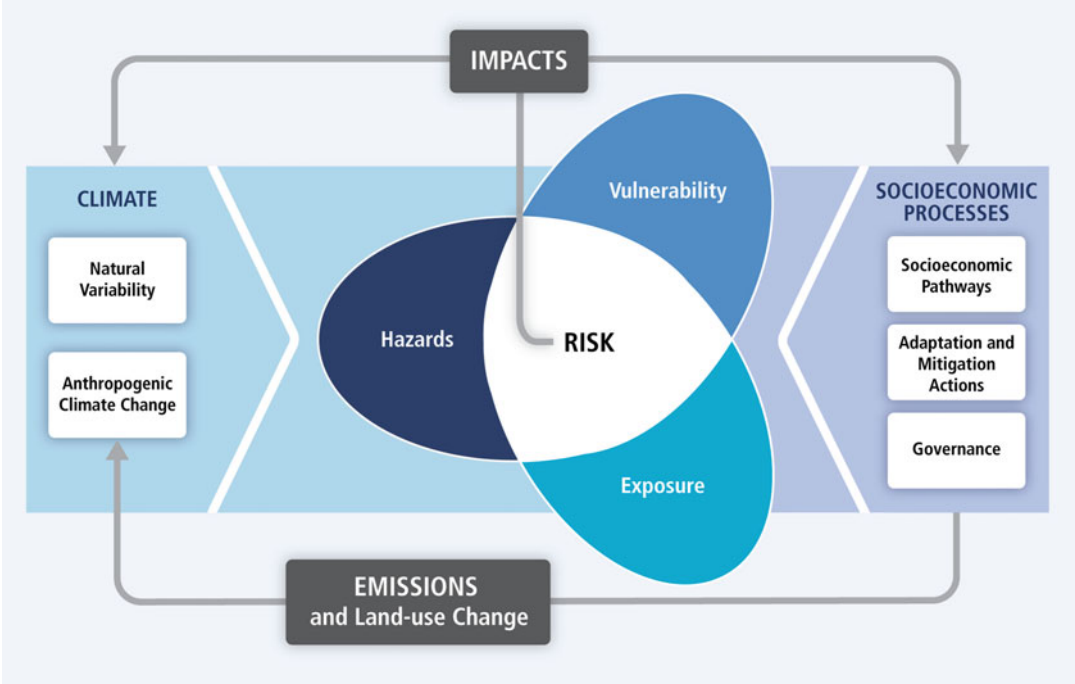
Common Future, a paper released by Brundtland Commission. The commission was established in the 1980s as a response to the inadequate response of the 1970s environmental movement. Sustainable development, according to the commission, is "development that meets the needs of the present without compromising the ability of future generations to meet their own needs" (Brundtland 1985). The commission further specified this definition by adding three pillars: economic growth, environmental protection, and social equality. Although many focus on the second pillar, the commission emphasized the interconnectivity of all three. Only when each pillar is achieved can sustainable development truly be realized.

Historical Background

Climate change is expected to worsen the scarcity of water, food, and energy resources and exacerbate hazards such as heat waves and heavy precipitation. Developing economies are characterized by growing and vulnerable populations, ever increasing income and wealth inequality, deteriorating or inadequate infrastructures, as well as the inability to mobilize resources for relief, rescue, or recovery efforts. Thus, while no region is immune to climate change, developing economies have been hit the hardest and will continue to be disproportionately impacted in the coming decades. However, climate adaptation and mitigation remain hotly debated, especially in situations where economic inequalities are severe and the disparity between future societal aspirations and the current status is large. Reduced reliance on fossil fuels may be viewed as less urgent compared to industrialization and economic growth, while land use and urban planning may be viewed as hindrances to improving quality of life. Adapting to projected disasters or resource scarcity may seem a case of misplaced priorities amidst a lack of adequate investments in education, health, or food security. However, it is within these developing economies that low regrets or transformational adaptation may significantly reduce loss of lives and economic devastation of the underprivileged. In the longer term, renewable energy or other low-carbon investments may make good business sense while lessening the more severe effects of climate change. The Intergovernmental Panel on Climate Change (IPCC) findings on mitigation are outlined in its Working Group III report (Summary for Policymakers 2014). According to the IPCC, mitigation efforts, by limiting the impacts of climate change, can enhance sustainable development, allow for more equitable distribution of resources, and assist with poverty. IPCC Working Group III describes the challenge of distributive justice, or the division of mitigation efforts equitably, to account for the greater impact of climate change on developing nations with historically lower emissions. For the 1.3 billion people in the world without access to electricity (Summary for Policymakers 2014), sustainable development will be required for this population's economic growth to not simultaneously expanding fossil fuel usage. The report presents developing, urbanizing cities as a significant mitigation target to reduce emissions, as many developed nations remain locked into excessive emissions by existing infrastructure. Mitigation policy will require a systemic approach, innovation, and difficult decisions: effective greenhouse gas emission reductions cannot be done if all nations do not act. The systemic change required will necessitate significant public, private, and institutional spending at an global level that takes into consideration local and historical practices and the potential for injustice between both developing and developed nations.

The IPCC's WGII report (IPCC 2014) identifies different areas of society that are at risk and the impact of climate change on different populations in the face of remaining uncertainty on the exact timing and extent of its impacts.

The IPCC report finds the most at risk vulnerable populations to be those that are already the most disadvantaged in society: socially, economically, politically, culturally, institutionally, or otherwise. The WGII report recommends first evaluating levels of vulnerability and risk. Once risks are defined, policy makers must evaluate resilience; the ability of a society to recover from disasters and hazardous events, efficiently and effectively. Climate resilience, specifically, is the ability to manage the impacts of climate change,



Climate Change and Developmental Economies, Fig. 1 Risk of climate-related impacts results from the interaction of climate-related hazards with the vulnerability and exposure of human and natural systems. Changes in both the climate system (left) and socioeconomic processes including adaptation and mitigation (right) are drivers of hazards, exposure, and vulnerability (Source: IPCC 2014, WGII AR5 SPM, pp 26)

reducing disruptions and expanding opportunities. In Fig. 1, the IPCC's iterative risk management process is illustrated. Risk, according to the IPCC is the intersection of hazards, vulnerability, and exposure. While climate causes hazards, risk arises out of vulnerability and exposure due socioeconomic processes.

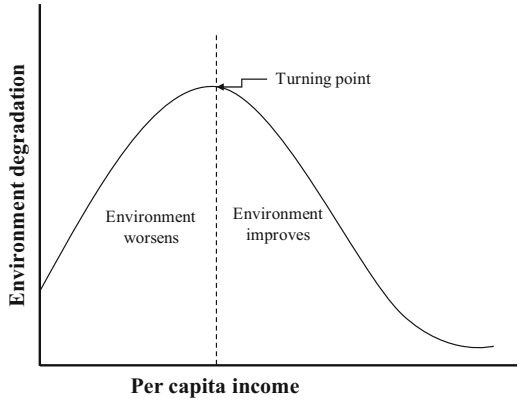
Among the uncertainty of climate change and its impacts, it is widely accepted that effective adaptation requires localized and targeted approaches. The IPCC recommends investing in infrastructures, development assistance, and existing disaster risk management institutions. However, the report emphasizes that a one-size-fits-all solution will be ineffective globally due to varying social values, interests, expectations, and circumstances (Smith et al. 2014).

Scientific Fundamentals

The relationship between economic growth and environmental degradation is a fundamental one for sustainable development. Many hope that this relationship can be decoupled, that is, that economic growth should not be conditional on environmental degradation. Two ideas, the environmental Kuznets curve (EKC) and the more recent Ecomodernist Manifesto, explore the topic of decoupling further.

The Environmental Kuznets Curve

A country's level of economic development is a key component in the debate on climate change mitigation. Do developed countries need to take on the majority of emissions reductions? Is there such a thing as a far share carbon space? Does a country's GDP growth inherently lead to a worse off environment? Questions like these are at the heart of the climate change-development conundrum. The environmental Kuznets curve, developed in 1991 by Grossman and Krueger, shows a graphical relationship between development and environmental degradation. Based on the economic Kuznets curve, which described the relationship between a country's gross domestic



Climate Change and Developmental Economies, Fig. 2 Hypothetical environmental Kuznets curve. Actual EKC may not be smooth or symmetrical

product (GDP) per capita and Gini coefficient, the EKC, too, is shaped like an inverted parabola, which is shown in Fig. 2. A key difference between the two curves is that the Kuznets curve went from observation to a theory, whereas the EKC is the reverse.

The EKC is important because it challenges the idea that environmental degradation necessarily continues with economic growth and instead provides a paradigm for how growth and sustainability can work simultaneously. It is noted that EKC is true for many local pollutants but not for CO_2. EKC has not have been interpreted as a justification for maintaining "business as usual" or encouraging greater GDP growth. Under the logic of the EKC, developing countries will, with economic development, ultimately reach an optimal level of pollution. This theory supports the idea that in order to lower emissions overall, a country's economy must be stronger. However, the concept that some countries are "too poor to be green" misinterprets development's relationship to the environment in two key ways. First, it leaves out movements that are directed against the environmental degradation that are caused, rather than mitigated, by increasing wealth. Second, it disregards the environmentalism of the poor (Martinez-Alier 1995). This type of environmentalism is centered on maintaining community resources that are threatened by the state or market system. In addition, Stern et al. point out that an EKC-type relationship may be the result of the effects of free trade (Stern 2004). As countries' economies begin to grow, they specialize in human capital-based activities and trade with countries that specialize in resource-intensive ones. Because of specialization, the results of the EKC may not accurately reflect a true improvement in environmental outcomes.

Climate Resilient Pathways

Effective policy to address climate change must extend beyond adaptation. IPCC's Working Group II recommends the implementation of climate-resilient pathways, an approach to sustainable development that combines both adaptation and mitigation strategies. The pathways require transformations on social, economic, political, and technological levels (IPCC 2014). Mitigation strives to reduce the rate and magnitude of climate change, allowing more time, even decades, to implement effective adaptation.

The IPCC implores policy makers to implement deliberative policy, strive for innovation, and make changes as mistakes are made and lessons are learned. Policy makers face challenges of limited resources, poor planning, misinformation, miscalculation, and the prioritization of short-term consequences over the long term. Though the task of climate mitigation is daunting, the impact is predicted to do more than reduce climate change risks, but to improve lives, well-being, and global management of the environment.

New Methods of Measuring Development

Although the empirical validity of EKC specifically for CO_2 has been disputed, there is not one official alternative to measuring the relationship between degradation and the environment economy. Barrett and Graddy proposed that there exists an inverted U-shape relationship between environmental degradation and civil and

political freedoms, rather than economic development (Barrett et al. 2000).

Some ecological economists argue that a better measure of human well-being is reflected in human development rather than in the measures of GDP growth alone. Steinberger and Roberts looked at the relationships between four different measures of human development (life expectancy, literacy, GDP per capita, and Human Development Index) and two measures of resource use (primary energy use and carbon emissions). They concluded that human well-being over time is becoming steadily more efficient. This challenges the perception that increased energy usage and increased emissions are necessary for better living conditions.

Ultimately, Steinberger and Roberts' research reflects new possibilities of dissociation between raising the standard of living while degrading the environment. In their words, "high human development can be generated at lower and lower energy and carbon emissions costs, and the quality of life is steadily decoupling from its material underpinnings." They found that different measures of development can be achieved at different rates of energy usage. Literacy, for instance, for instance, requires far less energy output than GDP (Steinberger and Roberts 2010). This new paradigm of development is useful. The justification for decoupling is reflected in other literature as well, including the Ecomodernist Manifesto.

Ecomodernist Manifesto

The Ecomodernist Manifesto rejects the idea that humanity will run out of resources and that increased human development is a problem. According to the Manifesto, the real issues are the misuse of energy sources, inefficient technology, and excessive carbon emissions. Predicted outcomes of the current path of resource usage, including ocean acidification, the loss of ozone in the earth's stratosphere, and climate shifts, could result in economic, population, and ecological loss. Along with long-term effects, the Manifesto points out well-known immediate impacts on populations, including water and air pollution. The way for humanity to develop further while preserving the surface of the earth is for society to "decouple" from nature and decrease resource dependency. Simply put, "nature unused is nature spared" (Asafu-Adjaye et al. 2015).

The Manifesto defines "decoupling" as the decrease in rate of environmental impact of a process, as economic output rates increase. The authors set a goal of "absolute decoupling," when the rate of human consumption of resources and energy peaks and declines. The Manifesto suggests humanity is on this path to peak environmental impact by the end of this century, due to trends of increased urbanization, agricultural technology expansion, and the introduction of efficient technology. The Manifesto's recommended strategy to reduce the human dependency on natural resources and strive for rapid decarbonization include urbanization, aquaculture, agricultural intensification, nuclear power, and desalination.

A strength of the manifesto is that it recommends innovation, not a return to earlier practices. It rejects nostalgia in environmentalism, and the idea that humanity has previously lived "lighter on the land," since three-fourths of global deforestation occurred before the industrial revolution. It calls for the expansion of electricity in the developing world, in contrast to environmentalists who theorize that resources are not available for such development. It presents the idea that global poverty is an environmental problem, one that cannot be ignored.

However, the Manifesto lacks a concrete strategy for more efficient use of resources. It presents scalable, power dense technologies as an alternative to carbon energy sources, even though present technologies are not yet capable of achieving that transition. It relies on the discovery of such technologies, but even admits that such progress is not inevitable.

Case Study: Paradox of Development and Mitigation

The relationship between development and its conflict with sustainable development is well

documented both in theoretical and empirical literature. Though the start was made in developed countries and most of the empirical studies are still in the context of high-income industrialized countries, developing countries are slowly rising up to analytical research that explores the link between economic activities, environmental impacts, climate change, and sustainability. Several such studies as well as the reports brought out by the UNFCC have put developing countries, especially India and China, in a dilemma. On one hand lie the aspirations of its people to achieve a decent standard of living and come out of poverty, which requires these economies to achieve a high macroeconomic growth rate. On the other is the global responsibility of these countries to check GHG emissions and adopt mitigation measures to delay climate change. Matters are made more complex by the domestic heterogeneity of these countries – one part having a lifestyle and values akin to the first world and another whose troubles, struggles, and aspirations resemble the least developed countries. While the terms of and solutions to challenges at the macro level remain gray, there is some evidence that at the microlevel policy makers prefer development at any cost. We shall refer to some case studies to understand how some of these activities are degrading the environment in developing countries.

The Asansol-Durgapur region in the eastern part of India has been an economic downturn and industries shuttered. The region came to be known as the Ruhr of India because of the large number of large industrial units set up in the region postindependence. These units were set up with adequate attention to environmental standards and impact on local pollution. However, since the early 1990s the region faced an economic downturn and industries shuttered. By the early 2000, the government started a new industrialization drive in the region by providing several concessions to entrepreneurs including fiscal benefits, a map of deregulated land use, and fast licensing. A large number of industrial units came up over the next five years, most of them in the earlier green belts and close to the residential areas. Most of them were also polluting in nature (sponge iron, carbon products, and smelting units, among others). The industrial units were belching fumes containing substantial amount of harmful gases, toxic chemicals, and Suspended Particulate Matter (SPM). This has resulted in pollution of air and water in the surrounding areas, health- and other-related damages to both the workers and residents, and another impetus to global climate change. SPM10 level at Durgapur stood at 350 mg/m^3 in August 2014, more than five times of the permissible standard level of 60 mg/m^3. The nitrogen dioxide level stood at 51.5 mg/m^3 in 2014 and continues to increase. It has been observed by researchers that the neo-industrialization drive has raised pollution levels and there are substantial costs involved as evident by using Willingness To Pay and Willingness to Accept methods. More than one-third of the respondents report the pollution level as unbearable while half of them say that it has increased over the last decade. Estimated value of environmental damage is about 2 % of the gross output of the new industries – clearly pointing at the substantial cost of development without looking at environmental impacts.

It is not that only industrialization in developing countries is to be blamed. Agriculture has its own way of affecting the environment and bringing about climate change. Biswas (2010) looked at the impact of extensive agricultural expansion on water availability and LULC changes in the rice bowl of Eastern India (Biswas 2010). It was observed that over a period of three decades [1971–2001 – roughly coinciding with the period of agricultural revolution in the region that turned mono-crop land to three-crop land through intensive irrigation, mechanization, improved seeds, pesticides, and chemical fertilizer], land use shifted strongly in favor of agriculture. However, the strategy was water intensive and resulted in lowering of the groundwater table from 8 m to 15 m in just 10 years. The number of surface water bodies decreased drastically, as did their total surface area. Markov chain modeling predicts a 50 % decline in water bodies over the next 25 years along with a corresponding rise in cultivated area and settlements. However, this situation is not

sustainable as reducing water availability leads to increased cost of cultivation, making agriculture a nonprofitable and vulnerable livelihood option, defeating the initial objective of agricultural expansion. Another associated threat is that of arsenic contamination of groundwater – a serious health risk in the lower Gangetic plain of Bangladesh and Eastern India – mainly caused by groundwater depletion for irrigation (Chakraborti et al. 2010).

Key Application

Social Entrepreneurship

As outlined in previous sections, climate change's consequences could lead to even greater vulnerabilities and worse outcomes for all of us, even more so for the world's most vulnerable populations. For instance, in developing economies such as India, the government's priorities seem to be at odds with those of climate mitigation. Economic development, more often than not, is likely to result in adverse impact on climate in this context. It is also not practical to expect business organizations to do much for climate mitigation apart from what is mandated by the law or regulatory authority. After all, their objective function is maximization of shareholders' wealth and climate related priorities can at best be looked as one of their several constraints. Does that mean all is lost for countries such as India, as far as climate mitigation is concerned? Are climate mitigation priorities going to be necessarily sacrificed at the altar of developmental priorities, or is there any hope? *Social enterprises*, an emerging class of organizations that work at the intersection of economic and environmental sustainability might offer such hope. While a consistent definition of a social enterprise is yet to emerge, what is being generally accepted is the fact that these organizations pursue a social objective such as poverty alleviation or environmental sustainability as their primary goal (objective function) and leverage market principles to become financially sustainable. Their objective of financial sustainability distinguishes them from the non-governmental/not-for-profit organizations and the fact that they do not want to *maximize* profits makes them different from the commercial enterprises.

Key application of the mitigation and enabling resilience in the resource constrained setting is exemplified through a story of social enterprise – Selco that provides solar lighting systems to the economically underprivileged. A significant number of India's 620,000 villages that accommodates nearly 60% of her population do not enjoy benefits of electricity. The electric grid has not yet reached many of them. Even those villages where there is grid connectivity, a lack of adequate power generation results in long periods without a supply. As a consequence, life and livelihood activities in these villages are governed by availability of sunlight. Availability of sources of energy at an affordable price is a critical problem for the rural poor in India, the negative impact of which is disproportionately borne by the womenfolk. The compromises that they make in order to overcome this challenge had deleterious effects both on their health as well as on the environment. In 1994, Harish founded Selco as a for-profit enterprise that would sell solar lights to the rural poor in India. While organizations and institutions in those days that wanted to address the needs of the poor, typically operated as not-for-profits, depending on grants and philanthropies, Harish was confident that he could create a financially sustainable enterprise that had a social objective. He has often been quoted saying that he set up Selco to prove three things – the poor can afford technology, the poor can maintain technology and it is possible to run a commercial venture that fulfills a social objective. In 2014, after nearly two decades of operation, Selco has sold 200,000 solar lighting systems across five different Indian states and has remained a financially sustainable enterprise. Selco's success clearly demonstrates that it is possible to tackle the twin problems of sustainable energy and poverty alleviation simultaneously, even while maintaining the bottom line of an enterprise.

However, the journey in the beginning was arduous, to say the least. Even though Harish

and his team were able to put together a solar lighting system that was suitable for the harsh environment of rural Karnataka, at about INR 5000 per light, it was not affordable by those at the base of the economic pyramid. They could only buy the product if they were provided credit. However, the banks were not ready to lend to the poor, especially because lighting systems were viewed as a consumer durable product and banks were instructed to provide loans to the poor only for income generating activities. Thus, Harish realized that it was necessary to link the purchase of solar lights to a stream of income. That was not too difficult to do because solar lights could increase the number of business hours for those who had to close shop after sundown because of lack of electricity. Moreover, there were others who were purchasing kerosene to do business after sunset. Selco could structure a financing plan for them such that the money they saved from not having to buy kerosene was more than the money they would have to pay for loan repayment. Finally, after a lot of convincing, banks started to provide credit for purchasing solar lights and Harish's dream of selling solar lighting systems to the poor took concrete shape. Harish also realized that apart from financing, Selco also needed to provide prompt service to its customers. Since customers would depend on its lights to run their business, any downtime would imply loss of livelihood opportunity and thereby loss of credibility. Selco therefore established a wide network of service centers all across Karnataka so that service engineers could reach even the most remotely located customer within a reasonable amount of time.

Selco's journey, apart from being inspirational, holds a lot of lessons for social entrepreneurs and others who engage with the problem of seeking market based solutions for poverty alleviation. However, changes need to be systemic in order to have any perceptible impact on resilience or mitigation. Therefore, such efforts need to be scaled in multiple domains such as healthcare, education and livelihood generation. This, on one hand, will reduce the economic vulnerability of a large section of the population, giving them opportunity for self-determination. On the other hand, it will make them resilient to deal with the adverse impact of climate change.

Future Directions

First and foremost, Ecomodernist Manifesto are calls to action and discussion. Though the Manifesto identifies the challenges of global climate change to be technological ones, it recognizes the need to adopt certain values in society to fully address them, including democracy, tolerance, and pluralism. However, in order to reach the "great Anthropocene" era that the Manifesto strives for, private businesses and state institutions must invest in technological research and embrace regulations to mitigate emissions. Technologies including nuclear power, wind power, solar power, and desalination remain either unsustainable and carbon intensive or economically inefficient. Scalable, power dense alternatives to carbon energy must be developed in order to both urbanize and intensify agriculture and simultaneously reduce human impacts on the environment. The environmental Kuznets curve, too, reflects a certain sense of optimism in its own modeling. One could argue that either the Ecomodernist Manifesto or the EKC simplifies the idea of decoupling too much. However, the concept – that economic growth does not necessarily need to be the cause of environmental degradation – may be a positive framework to encourage action on sustainable development.

Cross-References

- Climate Adaptation, Introduction
- Climate and Human Stresses on the Water-Energy-Food Nexus

References

Asafu-Adjaye J, Blomquist L, Brand S, Brook BW, Defries R, Ellis E, Foreman C, Keith D, Lewis M, Lynas M, Nordhaus T, Pielke R, Pritzker R, Roy J, Sagoff M, Shellenberger M, Stone R, Teague P (2015) An ecomodernist manifesto, pp. 32

Barrett S, Graddy K (2000) Freedom, growth, and the environment. Env Dev Econ. core/journals/environment-and-development-economics/article/freedom-growth-and-the-environment/393DCC0CAB23F8A9837DCC892B3CB90A. Accessed 6 Sept 2016
Biswas B (2010) Changing water resources study using GIS and spatial model – a case study of Bhatar Block, district Burdwan, West Bengal, India. J Indian Soc Remote Sens 37:705–717. doi:10.1007/s12524-009-0049-z
Brundtland GH (1985) World commission on environment and development. Env Policy Law 14:26–30
Chakraborti D, Rahman MM, Das B, Murrill M, Dey S, Chandra Mukherjee S et al (2010) Status of groundwater arsenic contamination in Bangladesh: a 14-year study report. Water Res 44:5789–5802. doi:10.1016/j.watres.2010.06.051
Field CB (2012) Managing the risks of extreme events and disasters to advance climate change adaptation: special report of the intergovernmental panel on climate change. Cambridge University Press, Cambridge
IPCC. Climate Change (2014) Impacts, adaptation, and vulnerability. Part B: regional aspects. Contribution of working group II to the fifth assessment report of the intergovernmental panel on climate change [Barros VR, Field CB, Dokken DJ, Mastrandrea MD, Mach KJ, Bilir TE, Chatterjee M, Ebi KL, Estrada YO, Genova RC, Girma B, Kissel ES, Levy AN, MacCracken S, Mastrandrea PR, White LL (eds)]. Cambridge University Press, Cambridge/New York
IPCC-TAR M (2001) Third assessment report of the intergovemmental panel on climate change. Cambridge University Press, New York
Martinez-Alier J (1995) The environment as a luxury good or "too poor to be green"? Ecol Econ 13:1–10
Smith KR, Woodward A, Campbell-Lendrum D, Chadee DD, Honda Y, Liu Q et al (2014) Human health: impacts, adaptation, and co-benefits. In: Field CB, Barros VR, Dokken DJ, Mach KJ, Mastrandrea MD, Bilir TE et al (eds) Climate change 2014 impacts adapt. Vulnerability Part Glob. Sect. Asp. Contrib. Work. Group II Fifth Assess. Rep. Intergov. Panel Clim. Change. Cambridge University Press, Cambridge/New York, pp 709–54
Steinberger JK, Roberts JT (2010) From constraint to sufficiency: the decoupling of energy and carbon from human needs, 1975–2005. Ecol Econ 70:425–33. doi:10.1016/j.ecolecon.2010.09.014
Stern DI (2004) The rise and fall of the environmental Kuznets curve. World Dev 32:1419–1439. doi:10.1016/j.worlddev.2004.03.004
Stocker TF, Qin D, Plattner GK, Tignor M, Allen SK, Boschung J et al (2013) Climate change 2013: the physical science basis. Intergovernmental panel on climate change, working group I contribution to the IPCC fifth assessment report (AR5), New York
Summary for Policymakers. Clim. Change (2014) Mitig. Clim. Change Contrib. Work. Group III Fifth Assess. Rep. Intergov. Panel Clim. Change Edenhofer O R Pichs-Madruga Sokona E Farahani Kadner K Seyboth Adler Baum Brunner P Eickemeier B Kriemann J Savol. Schlömer C, Von Stechow T, Zwickel JC x (eds) Cambridge University Press, Cambridge/New York
van Ginneken W (2003) Extending social security: policies for developing countries. Int Labour Rev 142:277–294. doi:10.1111/j.1564-913X.2003.tb00263.x

Climate Extremes

▸ Climate Extremes and Informing Adaptation

Climate Extremes and Informing Adaptation

Hayden Henderson[1,2], Laura Blumenfeld[1], Allison Traylor[1,3], Udit Bhatia[1], Devashish Kumar[1], Evan Kodra[1,4], and Auroop R. Ganguly[1]
[1]Sustainability and Data Sciences Laboratory (SDS Lab), Department of Civil and Environmental Engineering, Northeastern University, Boston, MA, USA
[2]Department of Mechanical and Industrial Engineering, Northeastern University, Boston, MA, USA
[3]Department of Political Science, Northeastern University, Boston, MA, USA
[4]risQ Incorporated, Cambridge, MA, USA

Synonyms

Climate adaptation; Climate change; Climate impacts; Climate resilience; Climate risks; Climate variability; Disaster risks; Floods and droughts; Weather extremes

Definitions

Climate Extremes

Climate extremes may be defined inclusively as severe hydrological or weather events, as well as significant regional changes in hydrometeorology, which are caused or exacerbated by climate change and which may in turn cause

severe stresses on regional resources, economy, and the environment. While regional warming and heat waves, and perhaps heavy precipitation, can be attributed to climate change with a degree of credibility and projected relatively reliably, significant uncertainties continue to exist for regional hydrology, including floods and soil moisture, as well as tropical cyclones or hurricanes and droughts. Intergovernmental Panel of Climate Change (IPCC) Special Report on Extremes 2012 (SREX) has adopted an event-based definition of climate extremes. It defines extremes as "occurrence of a value of a weather or climate variable above (or below) a threshold value near the upper (or lower) ends of the range of observed values of the variable" (Field 2012). It is noted that our definition of the extremes includes stresses induced by severe events as well as regional changes in hydrometeorology.

Adaptation

IPCC SREX 2012 defines adaptation as "the process of adjustment to actual or expected climate and its effects, in order to moderate harm or exploit beneficial opportunities." Adaptation measures can range from local actions including regulated water usage in households or farmers planting eco-friendly crops to large-scale infrastructure changes – such as building defenses (e.g., levees, natural barriers such as mangroves in coastal area) – to protect against rising sea levels or improving the quality of infrastructure to stand against high intensity hurricane events.

The two primary responses of policy makers and stakeholders to climate change are mitigation and adaptation. Both mitigation and adaptation are essential, because even if emissions are drastically reduced in future decades, adaptation measures will still be required to cope up with the changes that have already been induced. Mitigation addresses the root causes of factors inducing climate change. For example, measures to reduce the emissions of greenhouse gases are considered mitigation efforts. Adaptation seeks to reduce the risks posed by consequences of environmental changes, weather extremes exacerbated by climate change, or natural variability.

Historical Background

The United States experienced 32 weather extreme events including floods, hurricanes, and droughts between 2011 and 2013. Each of these events caused at least one billion dollars in damages. 2012 ranks as the second costliest year on record, with more than $110 billion in damages (Karl et al. 2009). While heat waves are expected to grow more intense, frequent, and longer duration despite considerable uncertainties and geographic variability, cold snaps are expected to reduce in frequency but expected to persist with current intensity and duration. Overall, the tails (or extremes) of temperature distributions at regional and seasonal scales have been shown to change asymmetrically under warming scenarios. Precipitation extremes, specifically high rainfall events, are expected to intensify under warming scenarios, although design (intensity-duration-frequency, or IDF) curves exhibit considerable uncertainty especially from local to regional scales (Kao and Ganguly 2011). Snowfall averages may decrease over certain regions but the extremes of snowfall may not reduce in intensity (Kodra et al. 2011). Wind speeds appear to have shown a decline globally although projections of wind extremes under climate scenarios are difficult and certain regional wind-derived marine circulation has intensified and is expected to continue to do so under warming scenarios (Kulkarni et al. 2016). Analysis of climate extremes averaged over global urban regions suggests more intense heat waves, lack of consistent patterns in precipitation extremes, and reduction in wind extremes in cities.

In addition to intensification of weather extremes in changing environments damage and losses from weather-related events have markedly increased over the past 30 years, mostly due to increase in exposure owing to mass population migrations and increased value of properties in urban coastal areas (Aerts et al. 2014). According to one reinsurance company (Munich Re), approximately $150 billion in economic losses were caused by weather-related events in the year 2012 alone. The situation is further exacerbated by lack of adaptation, i.e., reactive responses and

anticipatory planning. Weather and hydrologic hazards may be caused or exacerbated by natural climate variability and climate change. However, the hazards turn into disasters and indeed catastrophic events when infrastructures and lifelines are vulnerable and when exposure to hazards is high.

For example, in 2005 during Hurricane Katrina, the eye of the hurricane passed east of the city of New Orleans without causing catastrophic damage to buildings and structures. However, flood walls and levees designed to protect the city from floods were breached at more than 50 locations leaving approximately 8 % of New Orleans flooded. Hence, how much human population are affected by changes in extreme weather also depends on level of adaptability and preparedness in addition to exposure and vulnerability.

The major constraints in translating climate extreme science to adaptation-relevant insights are the uncertainties in our understanding and in projections at (local to regional) scales and (decadal) planning horizons relevant to stakeholders. At regional and decadal scales process understanding and model projections are less accurate, while at decadal scales the uncertainties are dominated by natural variability and hence difficult to translate to risk-based design principles. While there is strong evidence of human influence in the warming of the atmosphere and the ocean and in changes in the global water cycle and changes in climatic extremes (Qin et al. 2013), the low confidence in the presence of trends in certain extreme events such as intensification of hurricanes, droughts, and the subsequent attribution to human activities makes adaptation and planning for these extreme events a daunting task (Table 1).

Scientific Methods

IPCC's fifth assessment calls for more attention to how adaptation is implemented in response to climate risks with special focus on the role of extremes in the adaptation process (Change IP on C 2014). However, future climate simulations display large uncertainty in mean changes. As a result, the uncertainty in future changes of extreme events, especially at the local and larger scale, is great. The uncertainty created by a changing climate and dynamic development trajectories poses challenges for decision-making. This section outlines methods that can be used to quantify, characterize, and attribute extremes to inform adaptation and policies. In the context of climate, while there are different types of extremes, heat waves and cold snaps are the most difficult to quantify, and hence we focus on the methods related to these. Methods to quantify extremes are classified into three broad categories: (a) impact relevant metrics, (b) methods to quantify trends in time and space, and (c) extreme attribution.

(a) Impact relevant metrics

Impact relevant metrics include heat waves (defined as prolonged period of excessively hot weather. While definitions vary, a heat wave is measured relative to the usual weather in the area and relative to normal temperatures for the season) and cold spells (defined as rapid fall in temperature within a 24-h period requiring substantially increased protection to agriculture, industry, commerce, and social activities. The precise criterion for a cold wave is determined by the rate at which the temperature falls and the minimum to which it falls. This minimum temperature is dependent on the geographical region and time of year).

(b) Methods to quantify trends in time and space

A few examples of these methods include (but not limited to) generalized extreme value theory (GEV), trend analysis, and covariates in extremes.

Generalized Extreme Value

Generalized extreme value (GEV) theory is a family of continuous distribution that combine type I (Gumbel), type II (Fréchet), and type III (Weibull) extreme value distributions. The GEV is the only possible limit distribution of sequence

Climate Extremes and Informing Adaptation, Table 1 Summary of global-scale assessment of recent observed changes and human contribution to the extremes, both in terms of detection of change and attribution to humans for the changes. Note the increase in uncertainty as we move down the table (Source: IPCC AR5 (Field 2012) Working Group I (WGI) Summary for Policy Makers, Table SP)

Phenomenon and direction of trend	Assessment that changes occurred (typically since 1950 unless otherwise indicated)	Assessment of a human contribution to observed changes
Warmer and/or fewer cold days and nights over most land areas	*Very likely*	*Very likely*
	Very likely	*Likely*
	Very likely	*Likely*
Warmer and/or more frequent hot days and nights over most land areas	*Very likely*	*Very likely*
	Very likely	*Likely*
	Very likely	*Likely* (nights only)
Warm spells/heat waves. Frequency and/or duration increases over most land areas	*Medium confidence* on a global scale *Likely* in large parts of Europe, Asia and Australia	*Likely*[a]
	Medium confidence in many (but not all) regions	Not formally assessed
	Likely	*More likely than not*
Heavy precipitation events, Increase in the frequency, intensity, and/or amount of heavy precipitation	*Likely* more land areas with increases than decreases[c]	*Medium confidence*
	Likely more land areas with increases than decreases	*Medium confidence*
	Likely over most land areas	*More likely than not*
Increases in intensity and/or duration of drought	*Low confidence* on a global scale *Likely* changes in some regions[d]	*Low confidence*
	Medium confidence in some regions	*Medium confidence*[f]
	Likely in many regions, since 1970[e]	*More likely than not*
Increases in intense tropical cyclone activity	*Low confidence* in long term (centennial) changes *Virtually certain* in North Atlantic since 1970	*Low confidence*[i]

(*continued*)

Climate Extremes and Informing Adaptation, Table 1 (*continued*)

Phenomenon and direction of trend	Assessment that changes occurred (typically since 1950 unless otherwise indicated)	Assessment of a human contribution to observed changes
	Low confidence	*Low confidence*
	Likely in some regions, since 1970	*More likely than not*
Increased incidence and/or magnitude of extreme high sea level	*Likely* (since 1970)	*Likely*[k]
	Likely (late twentieth century)	*Likely*[k]
	Likely	*More likely than not*[k]

[a]Attribution is based on available case studies. It is *likely* that human influence has more than doubled the probability of occurrence of some observed heat waves in some locations.
[b]Models project near-term increases in the duration, intensity and spatial extent of heat waves and warm spells.
[c]In most continents, *confidence* in trends is not higher than *medium* except in North America and Europe where there have been *likely* increases in either the frequency or intensity of heavy precipitation with some seasonal and/or regional variation. It is *very likely* that there have been increases in central North America.
[d]The frequency and intensity of drought has *likely* increased in the Mediterranean and West Africa, and *likely* decreased in central North America and north-west Australia.
[e]AR4 assessed the area affected by drought.
[f]SREX assessed *medium confidence* that anthropogenic influence had contributed to some changes in the drought patterns observed in the second half of the 20th century, based on its attributed impact on precipitation and temperature changes. SREX assessed *low confidence* in the attribution of changes in droughts at the level of single regions.
[g]There is *low confidence* in projected changes in soil moisture.
[h]Regional to global-scale projected decreases in soil moisture and increased agricultural drought are *likely (medium confidence)* in presently dry regions by the end of this century under the RCP8.5 scenario. Soil moisture drying in the Mediterranean, Southwest US and southern African regions is consistent with projected changes in Hadley circulation and increased surface temperatures, so there is *high confidence* in *likely* surface drying in these regions by the end of this century under the RCP8.5 scenario.
[i]There is *medium confidence* that a reduction in aerosol forcing over the North Atlantic has contributed at least in part to the observed increase in tropical cyclone activity since the 1970s in this region.
[j]Based on expert judgment and assessment of projections which use an SRES A1B (or similar) scenario.
[k]Attribution is based on the close relationship between observed changes in extreme and mean sea level.
[l]There is *high confidence* that this increase in extreme high sea level will primarily be the result of an increase in mean sea level. There is *low confidence* in region-specific projections of storminess and associated storm surges.
[m]SREX assessed it to be *very likely* that mean sea level rise will contribute to future upward trends in extreme coastal high water levels

independent and identically distributed random variables' maxima that are properly normalized. The GEV has cumulative distribution function:

$$F(x;\mu,\sigma,\xi) = \exp\left\{-[1+\xi\left(\frac{x-\mu}{\sigma}\right)]^{-1\backslash\xi}\right\} \quad (1)$$

It is the three parameter distribution where μ, σ, and ξ *represent* location parameter, scale parameter, and the shape parameter, respectively. In statistics, location parameter determines the shift of distribution, scale parameter quantifies spread (or variability) of the distribution, and shape parameter controls symmetry of the distribution (Coles 2001). In the context of climate, Fig. 1 (top row) shows how changes in the location parameter would impact the distribution of extremes, and similarly middle and bottom rows show the corresponding changes in extremes (or tails) when scale and shape factors are changed (Kodra and Ganguly 2014).

To model series of extremes, a series of independent observations $X_1, X_2, \ldots X_n$ is considered for some large value of n. Data is blocked into such sequences and a series of block maxima $M_{n1}, M_{n2}, \ldots N_{nm}$ to which GEV is generated. For example, if n corresponds to the number of observations in each year and m number of years are considered, block maxima corresponds to annual maxima.

Estimates of extreme quantiles of the annual maximum distribution are then obtained by inverting (1):

$$x_p = \mu - \frac{\sigma}{\xi}\left[1 - y_p^{-\xi}\right] \quad (2)$$

where x_p is the return level associated with return period $1/p$. In other words, x_p is exceeded by the annual maxima in a given year by probability p.

Analysis of Trends

The detection, estimation, and prediction of trends and associated statistical significance are important aspects of climate extremes to analyze extremes. For example, given a time series of temperature, the trend is the rate at which temperature changes over a given period of time, which may be linear or non-linear. Trend may be linear or nonlinear. To test the presence of trends, simple linear regression is most commonly used to estimate the slope in combination with significance tests such as parametric Student's t-test or nonparametric Mann-Kendall test (to test both linear and nonlinear significance) with the underlying null hypothesis that no trend is present.

(c) Extreme attribution

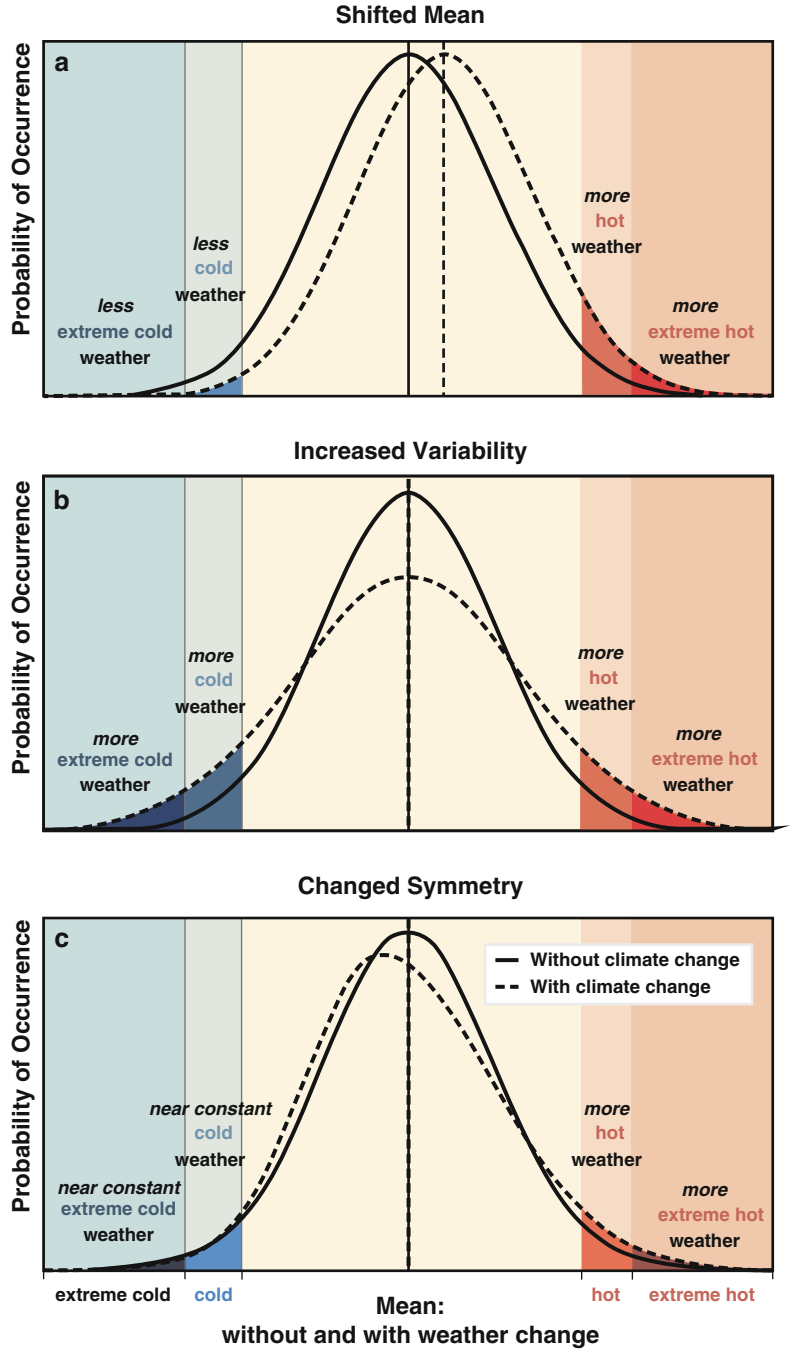
Weather and climate extremes occur all the time, with or without climate change. However, as shown in Table 1, there is a justifiable and strong sense that some of these extremes are evolving and becoming more frequent, and the primary reason can be attributed to human-induced changes in climate. However, given the small signal-to-noise ratio and uncertain nature of forced changes, attributing changes solely to human-induced changes or natural variability can be misleading (Trenberth et al. 2015). Extreme attribution studies aim to determine to what extent human-induced climate change has altered the probability or magnitude of particular events with significant confidence levels (Stott et al. 2016). This section discusses some of the methods used for extreme attribution.

Fractional Attributable Risk

If A is the probability of a climatic event occurring in the presence of human-induced forcing, and $B$is the probability of it occurring if the same forcing had not been present, then the fraction of the current risk that is attributable to past greenhouse gas emissions (fraction of attributable risk; FAR) is given by FAR $= 1 - A/B$.

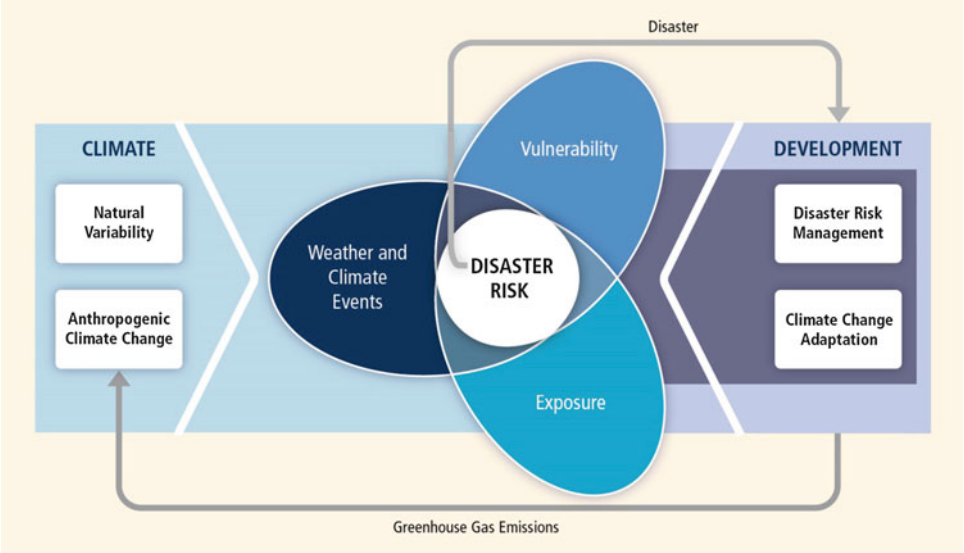
Model Approaches

General circulation models (GCMs), which often include biological, chemical, geological, atmospheric, and oceanic processes, provide the most comprehensive simulations of the climate system. Data from model experiments with different climate forcing combinations are available from Climate Research Program's Coupled Model



Climate Extremes and Informing Adaptation, Fig. 1 The IPCC SREX discussed the potential global warming consequences on (temperature, in this case) extremes through three representational images and assuming Gaussian (normal bell shaped, symmetrical distribution). The first image (*top*) depicts a shift in the mean without any other change in the temperature distribution, leading to more hot extremes but less cold extremes. The *middle figure* shows no change in mean but changes in variability, leading to more extremes on either tails, i.e., hotter and colder extremes. The *last figure* shows a change in the distributional symmetry with or without climate change (Source: *IPCC SREX Field 2012*)

Intercomparison Project Phase 5 (CMIP5) (Taylor et al. 2012). This data typically involves pooling data from multimodel ensembles of simulations with and without anthropogenic influences to generate large samples of the relevant variables such as temperature, precipitation, humidity, etc. The distribution of variables in the world with human influences and the world without these influences thus can be constructed from which estimates of FAR can be obtained (Stott et al. 2016).



Climate Extremes and Informing Adaptation, Fig. 2 The 2012 IPCC SREX depicts the connection between climate-related hazards and vulnerability or exposure. While hazards have traditionally been considered acts of God, disasters are caused by the very human failure to be prepared. In the current era of greenhouse gas-driven climate change, even hazards are not immune to human influences (Figure source: IPCC SREX (1.1.2, figure 1–1) Lavell et al. 2012)

Informing Adaptation: Risk Management

Adaptation to climate extremes and preparedness to disaster seek to reduce factors and modify environmental and human contexts that contribute to climate-related risk, to promote sustainability in social and economic development (Lavell et al. 2012). The promotion of adequate preparedness for disaster is also a function of disaster risk management and adaptation to climate change.

One of the many ways in which climate change is likely to affect societies and ecosystems around the world is through extremes and changes in extreme events (Fig. 2). As a result, regularly updated appraisals of evolving climate conditions and extreme weather would be immensely beneficial for adaptation planning. In fact, in a conventional risk framework, one of the components is probability of occurrence of hazard. Risk analysis methods identify the vulnerabilities of specific components of a system to an adverse event and quantify the loss of functionality of the system as a consequence of that event. In the context of climate and weather extremes, hazard (H) can be visualized as an outcome of an extreme event. For example, in the context of planning for transportation systems, H may represent the severe snowstorm or hurricane that can potentially deviate the system from its normal functionality. In addition to hazard (H) and its subsequent probability of occurrence p(H), risk to the system also depends on likelihood of vulnerability p(V) and chances of the system getting exposed to these risks. Mathematically, risk can be quantified as:

$$\text{Risk} = p(H) \times p(E) \times p(V) \tag{3}$$

Risk in a system is interpreted as total reduction in functionality and is related to the temporal effect of an extreme event on the system (Linkov et al. 2014).

Resilience Framework

As discussed in the previous section, adaptation to climate extremes seeks to reduce factors

contribute to climate-related risk. While a risk management system framework focuses on strengthening the specific components within a system, increasing interconnectivity and complexity of systems makes risk analysis of individual components constituting these complex systems unrealistic (Linkov et al. 2014; Bhatia et al. 2015). Moreover, uncertainties associated with the three components of risk challenge our ability to fully comprehend the risk related to the system. To address these challenges, resilience must be built into systems to help them quickly recover and adapt when adverse events do occur. The National Academy of Sciences defines resilience as "the ability to prepare and plan for, absorb, recover from, and more successfully adapt to adverse events" (Disaster Resilience: A National Imperative n.d.). Resilient systems are able to minimize the negative impacts of adverse events on impacted societies and even improve their functionality by adapting to and learning from fundamental changes caused by the extreme events.

Key Applications

Concepts to quantify extremes play a fundamental role in understanding the evolving nature of weather and climatic extremes in global environmental change. This understanding is translated to planning and managing diverse critical infrastructure sectors including 4 lifelines, namely, transportation (Bhatia et al. 2015), water resources (Kao and Ganguly 2011), healthcare (Semenza et al. 1995), and energy (Pryor and Barthelmie 2010). Given the increase in interdependencies, complexities and interconnectivity of infrastructure systems, concepts to quantify the resilience of complex systems ranging from ecosystems to power grid management (Gao et al. 2016), and communication and transportation (Bhatia et al. 2015) networks are gaining attention in scientific community.

Future Directions

Leveraging Physics and Data Science-Based Methods

Geographic information sciences (GI science or GIS) are necessary to develop data science methods that can translate the science of climate extremes to insights and metrics that are actionable by stakeholders and policy makers. Note that here GIS is used broadly to denote two different connotations. This section describes how GIS in both forms can address stakeholder-relevant crucial gaps in climate (extremes) science.

The first major gap is to help generate deeper understanding of the processes that may relate to climate extremes. A combination of data and process understanding, "physics-guided data mining" may help us gain a better understanding of relevant processes such as convection and aerosols, El Nino and other climate oscillators, and the monsoons (Ganguly et al. 2014). A second gap is the ability to characterize extremes and develop downscaling strategies. A comprehensive assessment of uncertainty, including those arising from multiple emissions scenarios, climate models and initial conditions, may be developed by bringing together physical understanding and data sciences, characterizing predictability and natural variability, and balancing historical skills with multimodel consensus. The third gap is the development of metrics for geospatial and temporal climate extremes and their spatial heterogeneity and temporal change, as well as decision support tools and policy aids for enabling effective decisions. The massive volume and complexity of data take the data and decision science challenges to the realm of what has been called "Big Data." However, the need to examine extremes and large changes as well as their early indicators makes this a "small data" concern. This "Big Data–small data" problem is perhaps the enduring GIS challenge of climate extremes. Newer big data-driven methods in rare or extreme events and in the analysis of complex data are likely to lead to innovative solutions.

Arguably the most significant knowledge gap in climate science relevant for informing

stakeholders and policy makers is the inability to produce credible assessments of local to regional climate extremes. Results from the latest generation of global climate model runs do not suggest the possibility of significant improvements in the near future, while regional climate models remain promising. However, ultrahigh-resolution models and physical understanding continue to improve process models. On the other hand, climate-related data, from archived model simulations, and remote or in situ sensors, have already moved into the petabyte scale and are projected to reach 350 PB by 2030. Thus, data-driven hypothesis examination and hypothesis generation need to leverage methods for handling massive and complex data. Geographical information science, comprising both geospatial process models and data science developments, can help address these challenges.

Role of Big Data in Extreme Event Mining

Generalized extreme value distribution is the only possible limit distribution of properly normalized maxima of a sequence of independent and identically distributed random variables. However, climate extreme events that can be correlated with space and time may deviate from the assumption of proper normalization. Hence, statistical approaches have not been well developed for a majority of climate extremes. Nonlinear dynamical approaches are better at characterizing the climate system rather than generating projections and, even so, are not well developed for predictability assessment in climate. Traditional spatial and spatiotemporal data mining in computer science, while well suited to certain kinds of geographic data, cannot handle the complex dependence structures, low-frequency variability, and nonlinear data generation processes relevant for predicting climate extremes. The barriers are particularly challenging given the so-called deep uncertainties in climate arising from both natural variability in the climate system, such as from oceanic oscillators combined with our lack of understanding of the relevant processes. The so-called Big Data methods can succeed in the context of climate extremes if in addition to handling massive data volumes, nonlinear data generation processes, complex proximity based as well as long-memory and long-range dependence in time and space, and extreme events or change can be directly addressed.

Cross-References

- ▸ Climate Adaptation, Introduction
- ▸ Informing Climate Adaptation with Earth System Models and Big Data

References

Aerts JCJH, Botzen WJW, Emanuel K, Lin N, Moel H de, Michel-Kerjan EO (2014) Evaluating flood resilience strategies for coastal megacities. Science 344:473–475. doi:10.1126/science.1248222

Bhatia U, Kumar D, Kodra E, Ganguly AR (2015) Network science based quantification of resilience demonstrated on the Indian Railways network. PLoS ONE 10:e0141890. doi:10.1371/journal.pone.0141890

Change IP on C (2014) Climate change 2014–impacts, adaptation and vulnerability: regional aspects. Cambridge University Press, New York

Coles S (2001) An introduction to statistical modeling of extreme values. Springer, London

Disaster Resilience: A National Imperative n.d. http://www.nap.edu/openbook.php?record_id=13457. Accessed 1 July 2015

Field CB (2012) Managing the risks of extreme events and disasters to advance climate change adaptation: special report of the intergovernmental panel on climate change. Cambridge University Press, New York

Ganguly AR, Kodra EA, Agrawal A, Banerjee A, Boriah S, Chatterjee S et al (2014) Toward enhanced understanding and projections of climate extremes using physics-guided data mining techniques. nonlinear Process Geophys 21:777–795. doi:10.5194/npg-21-777-2014

Gao J, Barzel B, Barabási A-L (2016) Universal resilience patterns in complex networks. Nature 530:307–312. doi:10.1038/nature16948

Kao S-C, Ganguly AR (2011) Intensity, duration, and frequency of precipitation extremes under 21st-century warming scenarios. J Geophys Res Atmos 116:D16119. doi:10.1029/2010JD015529

Karl TR, Melillo JT, Peterson TC (2009) Global climate change impacts in the United States. Cambridge University Press, New York

Kodra E, Ganguly AR (2014) Asymmetry of projected increases in extreme temperature distributions. Sci Rep 4. doi:10.1038/srep05884

Kodra E, Steinhaeuser K, Ganguly AR (2011) Persisting cold extremes under 21st-century warming scenarios. Geophys Res Lett 38:L08705. doi:10.1029/2011GL047103

Kulkarni S, Deo MC, Ghosh S (2016) Evaluation of wind extremes and wind potential under changing climate for Indian offshore using ensemble of 10 GCMs. Ocean Coast Manag 121:141–152. doi:10.1016/j.ocecoaman.2015.12.008

Lavell A, Oppenheimer M, Diop C, Hess J, Lempert R, Li J et al (2012) Climate change: new dimensions in disaster risk, exposure, vulnerability, and resilience. In: Field CB (ed) Managing the risks of extreme events and disasters to advance climate change adaption. Cambridge University Press, New York, pp 25–64

Linkov I, Bridges T, Creutzig F, Decker J, Fox-Lent C, Kröger W et al (2014) Changing the resilience paradigm. Nat Clim Change 4:407–409. doi:10.1038/nclimate2227

Pryor SC, Barthelmie RJ (2010) Climate change impacts on wind energy: a review. Renew Sustain Energy Rev 14:430–437. doi:10.1016/j.rser.2009.07.028

Qin D, Plattner GK, Tignor M, Allen SK, Boschung J, Nauels A et al (2013) Summary for policymakers. Climate change 2013: the physical science basis. Contribution of working group I to the fifth assessment report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge/New York

Semenza JC, Rubin CH, Falter KH, Selanikio JD, Flanders WD, Howe HL et al (1996) Heat-related deaths during the July 1995 heat wave in Chicago. N Engl J Med 335:84–90. doi:10.1056/NEJM199607113350203

Stott PA, Christidis N, Otto FEL, Sun Y, Vanderlinden J-P, van Oldenborgh GJ et al (2016) Attribution of extreme weather and climate-related events. Wiley Interdiscip Rev Clim Change 7:23–41. doi:10.1002/wcc.380

Taylor KE, Stouffer RJ, Meehl GA (2012) An overview of CMIP5 and the experiment design. Httpdxdoiorg101175BAMS–11-000941. http://journals.ametsoc.org/doi/abs/10.1175/BAMS-D-11-00094.1. Accessed 10 Dec 2015

Trenberth KE, Fasullo JT, Shepherd TG (2015) Attribution of climate extreme events. Nat Clim Change 5:725–730. doi:10.1038/nclimate2657

Climate Finance

▸ Climate Risk Analysis for Financial Institutions

Climate Hazards and Critical Infrastructures Resilience

Udit Bhatia[1], Allison Traylor[1], Catherine Moskos[1], Laura Blumenfeld[1], Lindsey Bressler[1], Tyler Hall[1], Rachael Heiss[1], Kevin D. Clark[1,4], Nan Deng[1], Devashish Kumar[1], Evan Kodra[1], Stephen E. Flynn[2], Haris N. Koutsopoulos[5], Jerome F. Hajjar[5], and Auroop R. Ganguly[1]

[1]Sustainability and Data Sciences Laboratory (SDS Lab), Department of Civil and Environmental Engineering, Northeastern University, Boston, MA, USA

[2]College of Social Sciences and Humanities, Northeastern University, Boston, MA, USA

[3]Northeastern University, Boston, MA, USA

[4]risQ Corporation, Cambridge, MA, USA

[5]Department of Civil and Environmental Engineering, Northeastern University, Boston, MA, USA

Introduction

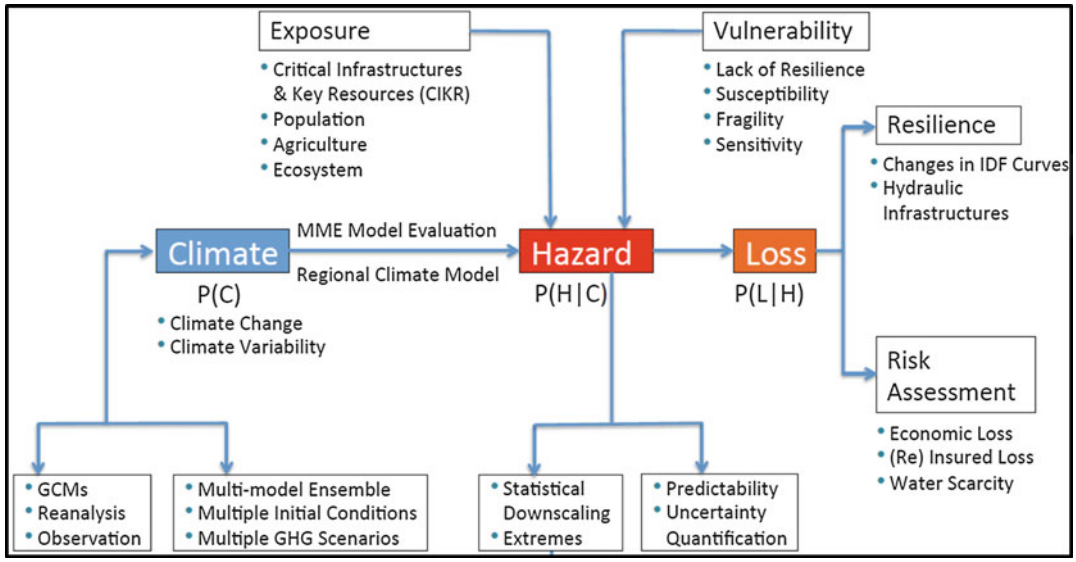
Climate Hazards and Critical Infrastructures Resilience

Civil and environmental engineers design structures such as buildings, bridges, and levees based on implicit or explicit assessments of risks. Thus, dead and live loads are considered along with acute stressors including natural hazards, in addition to the strength of materials and the fragility of components. Safety factors in engineering design essentially attempt to account for unknown or perhaps even unknowable uncertainties and change. The consequences of failure, measured in terms of economic damage and danger to human lives, result in safety factor assignments. Probabilistic risk assessments and reliability analyses attempt to explicitly consider risks in engineering design. Performance-based engineering and fuse-based systems attempt to develop design paradigms that rely on anticipated stresses to structural components. In certain implementations, system level functionality may

be maintained while allowing for a systematic failure of components. The methods have been successfully applied to earthquake engineering. However, the challenges and the opportunities become significantly different when concepts from engineering design need to be generalized to embedding resilience in critical infrastructures, especially in the context of adapting to threats resulting from climate change. The current state of practice and research consider three related issues, specifically, the nature of the climate and related stressors, the definition of the stressed systems under consideration, as well as the evolving concept of resilience. Resilience in this context goes beyond robustness to the immediate effects of a hazard as well as the ability to gracefully recover from the aftermath in a timely, cost-effective, and efficient manner. In other words, resilience is defined as the ability of the entire system to maintain essential functionality despite acute or chronic stressors and, in the event of failure or loss of functions, gets back to normalcy quickly and easily. The stressed systems of primary concern are what have been called critical infrastructures and lifeline infrastructure networks. The United States Department of Homeland Security defines 16 critical infrastructure sectors, specifically, chemical, commercial facilities, communications, critical manufacturing, dams, defense industrial base, emergency services, energy, financial services, food and agriculture, government facilities, healthcare and public health, information technology, nuclear reactors and materials and waste, transportation, and water and wastewater. The National Infrastructure Advisory Council lists four critical lifeline infrastructure networks: transportation, electricity and power, communications, and water and wastewater. Developing resilience across these lifelines and sectors requires an understanding of the cascading interdependencies across infrastructure elements and networks, the ability to design systems for effective response and recovery, the ability to design for greater resilience, the availability of appropriate metrics and financial instruments or economic incentives, as well as the ability to effectively govern across organizational and jurisdictional barriers. Adaptation to climate change and climate-related weather or hydrologic extremes, especially over the lifetime of infrastructure sectors and lifelines, requires an understanding on both the nonstationary nature of climate stressors and the deep uncertainties. The earth's climate system is fundamentally changing in ways such that the past is no longer an effective guide to the future in terms of design parameters. Uncertainties resulting from both our lack of understanding and the intrinsic variability of the climate system cannot be assigned likelihoods. The situation calls for flexible design principles, which remain risk informed and resilience centric. Case studies discuss urban heat islands, sea level rise and land subsidence, hurricanes and storm surge in coastal megacities, and severe droughts with consequences for the nexus of food-energy-water.

Probabilistic Risk Assessments and Climate Hazards

The Special Report on Extremes (IPCC 2012) as well as the Intergovernmental Panel on Climate Change's Fifth Assessment Report (AR5) (IPCC 2014a, b) published in 2013–2014 depicts how climate extremes may turn into disasters depending on vulnerability and exposure. The framework relies on risk computations, where three aspects are considered: hazards, or the probability of threats; vulnerability, or the probability of damage conditional on hazards; and consequences or economic damages and/or losses of human lives. Climate hazards may be broadly defined to include either extreme weather or hydrological events or changes in regional hydrometeorology, which may be caused or exacerbated by climate variability or change and which could stress all or parts of the coupled natural-engineered-built systems (Fig. 1). Recent climate hazards in the United States include hurricanes Katrina in New Orleans in 2005 and Sandy in New York/New Jersey in 2012, floods in Iowa in 2013, the 2010 (ongoing) droughts in California, the 2014 cold snaps in the Northeast,



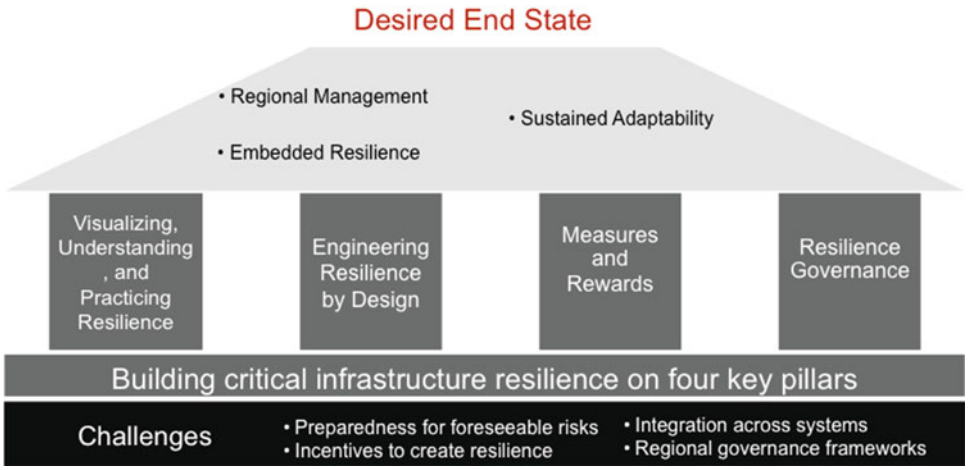
Climate Hazards and Critical Infrastructures Resilience, Fig. 1 Schematic representation of probabilistic risk assessment (PRA) methods in context of climate-related hazards. PRA methods such as this can be used to identify the vulnerabilities of both natural and built environments to an expected climate-related hazard and quantify the losses as a result of consequences of these events

and 2012 summer heat waves across the United States. Hurricane Katrina was a Category 5 over the Gulf of Mexico but reduced to a Category 3 by the time it made landfall on the Gulf Coast. However, the natural phenomena, the hurricane hazard itself in this case, was not the sole reason why Hurricane Katrina was the costliest natural disaster in US history. In fact, post-landfall news for a while appeared to suggest that the storm was moving northward over land, but no major destruction was reported. However, it was then that the levee, which was known to be highly vulnerable to start with, broke from the weight of the water. The resulting floodwater devastated New Orleans, where to start with the human settlement patterns were susceptible. This is where the hazard (Hurricane Katrina) interacted with the vulnerability of a critical infrastructure (levee) as with exposure (e.g., human settlements in this case) to result in levels of losses that were historically unprecedented and thus far unsurpassed within the United States. Probabilistic risk assessments (PRA) thus remain important to extract a comprehensive characterization of climate hazards, understand how and when the hazards may turn into catastrophic disasters, and perhaps even used to examine the impacts of strategic policy and tactical interventions. Figure 1 shows a comprehensive depiction of PRA and PRA-inspired methods, which have been or could be used in the context of climate hazards. In the context of climate-change impacts, risk is often represented as probability of occurrence of hazards, including but not limited to extremes such as heat waves, droughts, floods, cold snaps multiplied by the impacts these events may cause on natural and human systems. Climate observations from in situ and remote sensors such as satellites, reanalysis data (Kalnay et al. 1996), and data from general circulation model (GCMs) are assimilated together with Greenhouse emission scenarios, multi-model ensembles, and multiple initial conditions of GCMs (see next sections) to project the changes and variability in climate and climate-related extremes. However, global circulation models are run at a coarse spatial resolution, typically of the order of 100 km and are unable to delve information at the local to regional scales relevant to policymakers and stakeholders. As a result, GCM output cannot be directly used for impact assessment at regional or local scales. To overcome this problem, downscaling is often

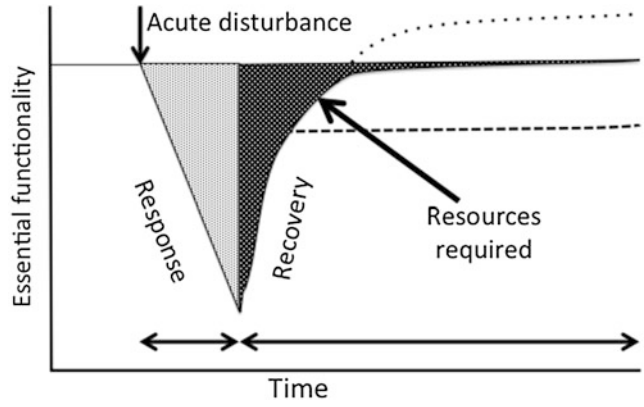
used to obtain local-scale climate projections at finer resolution from atmospheric variables provided by GCMs (Ghosh and Mujumdar 2008). Interaction of the climate-related hazards with the exposure and vulnerabilities of critical infrastructures and population put these systems at risk, resulting in the loss of economy or/and human lives. However, quantification of climate-related hazards and related risk is associated with uncertainties arising out of natural variability, anthropogenic climate change, or a combination of both. Hence, uncertainty quantification and characterization forms a crucial part of PRA-inspired methods before they can be deployed to motivate strategic policy changes and resilient design practices.

Resilience Paradigm: Beyond Probabilistic Risks

While critical infrastructure systems and lifelines were built as isolated entities, in actuality they are functionally interdependent. Disasters ranging from hurricane to large-scale power outages have shown how failure in one system may trigger a cascade of failures in interdependent infrastructure systems. Although investigation of resilience in infrastructure systems has triggered enormous interest, most of the research endeavors have focused on the isolated systems. However, critical infrastructures including lifelines exhibit a large number of interdependencies. These interdependencies could be cyber or cyber-physical (Buldyrev et al. 2010), geographical (Solé et al. 2008) or political, and so on. Traditional risk analysis methods focus on identification of vulnerabilities of specific system components. Subsequent risk management frameworks, hence, focus on strengthening these specific components to prevent overall system failure (Linkov et al. 2014). However, the factors which make traditional risk assessment tools unviable are as follows: (1) complexity and interconnectedness of infrastructure networks including lifelines and (2) nonstationarity and deep uncertainty associated with climate hazards. However, the development of resilience at system level faces the following challenges: the lack of consensus over defining and quantifying resilience, lack of preparedness for foreseeable and unforeseeable risks under changing climate, absence of incentive structure for public and private infrastructure owners to create resilience, and organizational barriers to creating resilience. Figure 2 sums up the barriers and plausible solutions to overcome these in order to translate resilience from



Climate Hazards and Critical Infrastructures Resilience, Fig. 2 Deficiencies in critical infrastructure resilience arise from four broad challenges. The four pillars outline the elements of solution to overcome these challenges to embed resilience in functioning of critical infrastructure and lifeline systems. Visualization and understanding resilience is an obligatory part of the framework to enforce resilient engineering and policy practices, which in turn, requires exhaustive understanding of interdependencies of various infrastructure systems



Climate Hazards and Critical Infrastructures Resilience, Fig. 3 Resilience management framework adapted from the commentary in *Nature* by Linkov et al. While probabilistic risk assessment-enabled methods give the probability of system hitting the lowest point of its essential functionality and thus help the system prepare and plan for adverse events, resilience management goes beyond and integrates the capacity of a system to absorb and recover from adverse events, and then adapt. The *dashed line* suggests that state of the system after recovery may be better or worse with respect to the initial performance, depending upon the system resilience

a mere "buzzword to operational paradigm for system management" (Linkov et al. 2014). As highlighted in the correspondence piece (Fisher 2015), resilience has been defined in more than 70 ways in the literature. While the National Academy of Sciences (Disaster Resilience 2015) defines resilience as "the ability to prepare and plan for, absorb, recover from, and more successfully adapt to adverse events", many scientists have just focused on the recovery part (Fig. 3a) to define resilience as the system's ability to bounce back after stress. Long-term policies based on the two extreme ends of definitions are likely to be very different and would be associated with different costs, depending on the definitions and metrics we adopt to measure resilience. At the regional scale, the structure and function of infrastructure systems—particularly in the lifeline sectors—are appropriately represented using network models and network science tools (Solé et al. 2008; Albert et al. 2000; Sen et al. 2003; Guimerà et al. 2005). A key issue for assessing and improving the resilience of infrastructure systems is to understand the behavior of the lifeline sectors during normal operating conditions, as well as in the presence of both nondeliberate hazards (e.g., natural hazards, human accidents, technology failures) and deliberate threats (e.g., terrorism, sabotage). Over the last decade, there have been considerable advances in the understanding of cascading interdependencies of the lifeline networks (Buldyrev et al. 2010; Koç et al. 2013; Hernandez-Fajardo and Dueñas-Osorio 2013). However, they had relatively little impact on the design of resilient interconnected infrastructures to mitigate the risk of cascading failures because the applicability of these frameworks on real-life networked infrastructures is not a trivial task, because the oversimplified assumptions on which these models are based may not be valid for the inextricable interdependent systems (Vespignani 2010).

Climate Hazards: Variability and Deep Uncertainty

As discussed in previous sections, both PRAs and resilience management framework include risk analysis as a central component. However, climate change might produce extreme events that cannot be predicted precisely, particularly at the spatial resolutions and time horizons relevant to the infrastructure owners and managers. Time horizons to be considered for emergency(Aerts et al. 2014) management

and infrastructure planning are near real time, seasonal to interannual, decadal to mid-century and multi-decadal to centennial, and so on. Infrastructures and lifelines are expected to remain the same on near real-time horizons and seasonal to interannual time horizons. However, weather predictions at these time scales may inform emergency response and management. On decadal to mid-century time scale, significant changes in reinforcements and remediation measures to build resilience of communities are expected. However, predicting climate-related hazards would be a major challenge as nonstationarity, including trends in global warming and changes in the statistics of weather pattern and relations (Salvi et al. 2015), and variability in mean and extreme climate (Ghosh et al. 2012) are expected to be predominant at these time scales (Hawkins and Sutton 2009). For example, Ganguly et al. (2009) demonstrated that increased trends in temperature and heat waves suggest for urgent mitigation and adaptation strategies, but these projections are concurrent with large uncertainty and variability making the decision making process complicated.

Since projections of weather and climate come from the numerical models that resolve the relevant processes, uncertainties in applying these models may result from:

1. **Internal variability:** Arises out of initial condition uncertainty and is more relevant for the short time scales (Palmer et al. 2005)
2. **Multimodal uncertainty:** As we know, climate system is highly complex, and it is practically impossible to model all the processes and parameters that govern the climate. Depending upon the choice of the parameters to model these processes, different GCMs differ substantially in their projections in the future.
3. **Boundary condition uncertainty:** The source of this uncertainty resides in the assumption over the future world economic and social development, leading to alternative sources of greenhouse emissions whose relative likelihood cannot be easily accessed (Tebaldi and Knutti 2007).

Case Studies

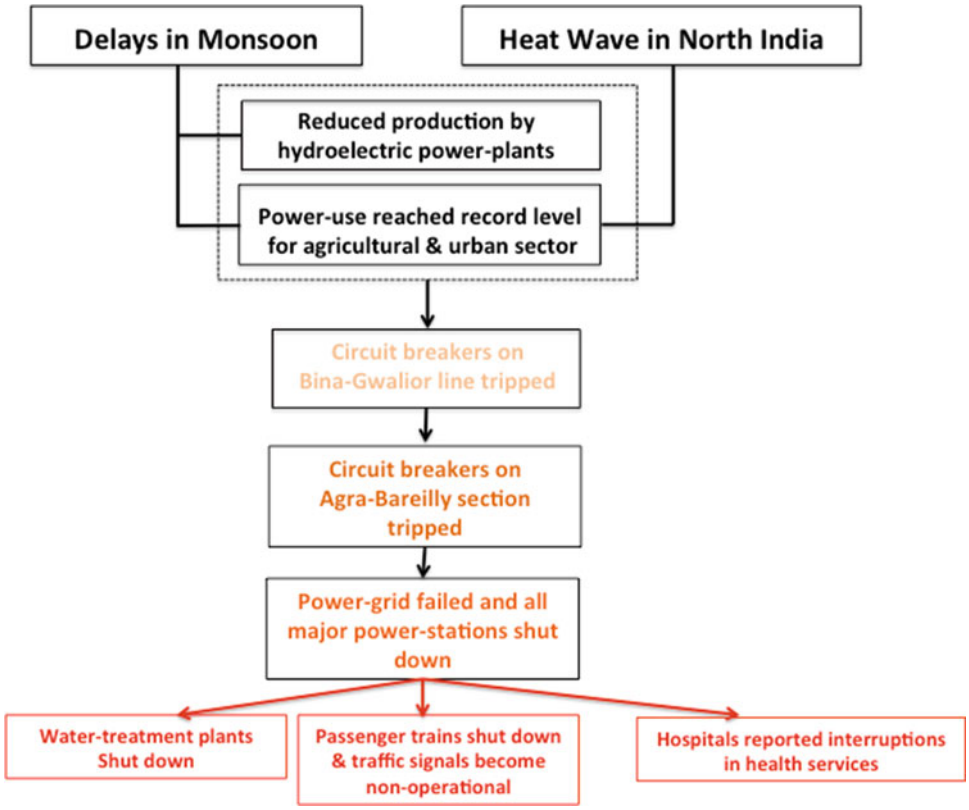
This section elucidates three case studies in the context of lifeline infrastructures, where climate extremes act as a stressor. In case study I, the impact of blizzards on the national airspace system of the United States is discussed. Particularly, this case study highlights that a regional climate event such as blizzard impacts the entire air traffic across the nation.

The second case study demonstrates the cascading interdependencies that exist between the various lifeline networks. In 2012, the power-grid failure in the northern and eastern parts of India resulted in catastrophic impacts on various lifelines dependent upon the power grid, directly or indirectly. Triggered by the combination of climatic events, delayed monsoon and heat waves, in this case, and manmade error, this has been recorded as the worst power blackout in the global history in terms of number of people affected.

Finally, the vulnerability of a Massachusetts Bay Transit Authority (MBTA) system of Boston is discussed, which lost its essential functionality on its rail system when hit by severe blizzards in 2015.

Case Study 1: Blizzard 2015 and Its Impact on National Airspace System of the United States of America

FlightAware (2015) reported that 1,200 flights were expected to be cancelled on January 26, 2015, to reduce air traffic volume in the northeast prior to a forecasted heavy winter storm. Delta airlines preemptively cancelled 600 flights; furthermore, a dozen flights from London Heathrow to New York, Philadelphia, and Boston were cancelled on the same date. These are just a few of the steps commercial airlines, the Federal Aviation Administration (FAA), State Officials, and Airport Authorities took in preparation for the January 27, 2015, blizzard, designated "Juno" by the National Weather Service (NWS). Across New England, contingency plans were implemented to sustain critical functions and return air transportation to normal operations as soon as possible. The impact of the storm was felt



Climate Hazards and Critical Infrastructures Resilience, Fig. 4 Flowchart showing events resulting in 2012 blackouts and resulting consequences on other interdependent lifeline services, including water distribution and wastewater distribution networks, transportation networks, and healthcare services

locally and nationally. This study addresses air traffic delays, diversions, and flight cancellations caused by extreme winter weather events and system recovery. An airport that is better prepared to respond to weather hazards operates more efficiently for passengers and airlines and can avoid significant negative impact to the NAS as a whole.

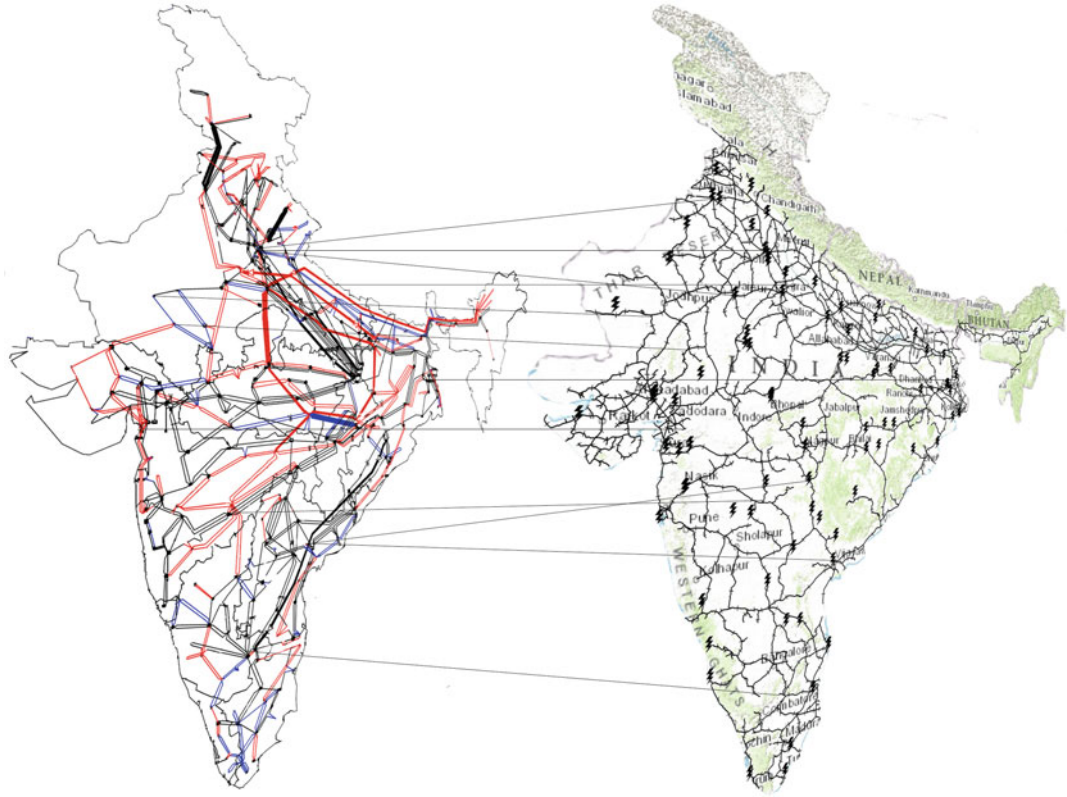
As of August 21, 2014, there were 19,453 airports in the United States (IPCC 2014a). Five of the busiest are located in the Eastern Service Area (ESA) of the National Airspace System (NAS): Atlanta, New York's JFK, Boston, Philadelphia, and Washington DC. Numerous studies (Jarrah et al. 1993; Abdelghany et al. 2004) have shown that convective weather in/around airports are a major cause of flight delays and a significant causal factor in aircraft accidents. In 2012, FlightStats.com (FlightStats 2015) issued a report stating that from October 27 to November 1 in North America alone, 20,254 flights were canceled due to Hurricane Sandy. Roughly 9,978 flights were canceled at New York area airports alone. United stands as the airline with the most cancellations by Sandy (2,149) followed by JetBlue (1,469), US Airways (1,454), Southwest (1,436), Delta (1,293), and American (759). In an examination of weather events over the past 7 years, Sandy comes in second in terms of total number of cancelled flights, behind the North American Blizzard of February 2010 (22,441 flights), for which the Blizzard of January 2015, designated "Juno," is compared in this report. Airport system capacity directly relates to NAS capacity, and Juno adversely affected airports and air traffic in the system.

Climate Hazards and Critical Infrastructures Resilience, Fig. 5 Massachusetts Bay Transit System: light rail routes (*Green*, *orange*, *blue*, *red lines*) and bus route to international airport (Adapted from Massacheusetts Bay Transportation Authority, Boston)

Case Study 2: 2012 India Blackouts

On July 30–31, 2012, two severe blackouts hit northern and eastern India, which impacted over 620 million people, across 22 out of 29 states of the nation. Given the population size affected, this has been recorded as the largest power outage in the history. Figure 1 shows how both non-intentional manmade and natural events resulted in the collapse of the power grid. In the summer of 2012, extreme heat caused record power consumption in northern India. The situation was further exacerbated by delayed monsoons, which resulted in drawing of increased power from the grid for running water pumps to irrigate the paddy fields in Kharif season.

On July 30, circuit breakers on a 400 kV line between cities of Bina and Gwalior got tripped. As this line fed into another transmission section (Agra-Bareilly), circuit breakers at that section also tripped. As a result of this sequential tripping, power failure cascaded through the grid. The system failed again on the afternoon of July 31 due to relay problem. As a result, power stations across the affected parts went offline, resulting in the shortage of 32 GW of power. The failure cascaded through other dependent



Climate Hazards and Critical Infrastructures Resilience, Fig. 6 Illustrative representation of interdependencies between power grid and Indian Railways network. 2012 power blackout brought more than 300 trains in northern and eastern India to a standstill, leaving people confined in the trains

infrastructures, hence severely affecting the functioning of lifeline systems including transportation, water distribution and wastewater treatment units, and health care services. Several hospitals faced interruptions in providing health services. Water treatment plants in affected regions were shut down for several hours. More than 300 trains, which include both long distance trains and local trains, were stalled, leaving passengers stuck midway. An illustration of cascading independencies between the power grid and Indian Railways Network is shown in Fig. 3. This case study highlights the imperative need to address the model the complexities of integrated systems to embed resilient design practices into large-scale lifeline infrastructure networks (Linkov et al. 2014). Also, the role of geographic information systems is implicit and ubiquitous to model and visualize complex systems such as these, operating at spatial scales ranging from local to regional to global (Fig. 4).

Case Study 3: 2012 Blizzard 2015 and Massachusetts Bay Transit System

In 2015, Boston confronted the snowiest winter ever in the history of recording climate events. In February alone, four storms had brought Boston record-breaking snowfall of over 100 in. Thousands of citizens' lives were affected. Boston's transportation system undertook an unprecedented test: highways blocked, flights canceled, and train service shut down. After a thorough analysis on dwell time and boarding data of the northern stations of the Orange Line (shown in Fig. 5) provided by MBTA Overhead Contact System center, the ridership decreased dramatically by nearly 30 % on the first day after the blizzard and recovered rapidly on the next

one. Meanwhile, the travel time and dwell time among, between, and within stations increased almost 50 %, which means the Orange Line train system lost one-third of its capacity. According to the boarding record and peak hour statistics, the remaining capacity can just meet the highest demand of current ridership. However, with growing population and transit use, the capacity limit might become a bottleneck in front of extreme weather or an emergency event, let alone worse weather conditions. The subsequent snowstorms during February 2015 have proved this hypothesis and resulted in system shutdown at times (Fig. 6).

Given that the capacity one train can carry is equivalent to almost 15 buses, it is almost impossible to completely replace the train service by using the shuttle bus. As a result, passengers have to turn away transit and resort to driving cars as their commuter mode, which brought even more congestion on the highways. The transition from SOV (single-occupancy vehicle) to HOV (high-occupancy vehicle) usage cannot be widely accepted if robust and reliable transit service is not being provided. Given the fragility and unreliability of current rail service that the northern part of MBTA Orange Line presented, a more comprehensive evaluation of the whole MBTA transit system, including its capacity, resilience, and future evolution, is recommended.

Cross-References

▸ Internet-Based Spatial Information Retrieval

References

Abdelghany KF, Shah SS, Raina S, Abdelghany AF (2004) A model for projecting flight delays during irregular operation conditions. J Air Transp Manag 10:385–394. doi:10.1016/j.jairtraman.2004.06.008

Aerts JCJH, Botzen WJW, Emanuel K, Lin N, Moel H de, Michel-Kerjan EO (2014) Evaluating flood resilience strategies for coastal megacities. Science 344: 473–475. doi:10.1126/science.1248222

Albert R, Jeong H, Barabási A-L (2000) The Internet's Achilles' Heel: error and attack tolerance of complex networks. Nature 406:200–0

Buldyrev SV, Parshani R, Paul G, Stanley HE, Havlin S (2010) Catastrophic cascade of failures in interdependent networks. Nature 464:1025–1028. doi:10.1038/nature08932

Disaster Resilience: A National Imperative (2015) [Internet]. [cited 1 Jul 2015]. Available: http://www.nap.edu/openbook.php?record_id=13457

Fisher L (2015) Disaster responses: More than 70 ways to show resilience. Nature 518:35–35. doi:10.1038/518035a

FlightAware – Flight Tracker/Flight Status/Flight Tracking. In: FlightAware [Internet]. [cited 1 Jul 2015]. Available: http://flightaware.com/

FlightStats – Global Flight Tracker, Status Tracking and Airport Information [Internet]. [cited 1 Jul 2015]. Available: http://www.flightstats.com/go/Home/home.do

Ganguly AR, Steinhaeuser K, Erickson DJ, Branstetter M, Parish ES, Singh N et al (2009) Higher trends but larger uncertainty and geographic variability in 21st century temperature and heat waves. Proc Natl Acad Sci 106:15555–15559. doi:10.1073/pnas.0904495106

Ghosh S, Mujumdar PP (2008) Statistical downscaling of GCM simulations to streamflow using relevance vector machine. Adv Water Resour 31:132–146. doi:10.1016/j.advwatres.2007.07.005

Ghosh S, Das D, Kao S-C, Ganguly AR (2012) Lack of uniform trends but increasing spatial variability in observed Indian rainfall extremes. Nat Clim Change 2:86–91. doi:10.1038/nclimate1327

Guimerà R, Mossa S, Turtschi A, Amaral LAN (2005) The worldwide air transportation network: anomalous centrality, community structure, and cities' global roles. Proc Natl Acad Sci USA 102:7794–7799 doi:10.1073/pnas.0407994102

Hawkins E, Sutton R (2009) The potential to narrow uncertainty in regional climate predictions. Bull Am Meteorol Soc 90:1095–1107. doi:10.1175/2009BAMS2607.1

Hernandez-Fajardo I, Dueñas-Osorio L (2013) Probabilistic study of cascading failures in complex interdependent lifeline systems. Reliab Eng Syst Saf 111:260–272. doi:10.1016/j.ress.2012.10.012

IPCC (2012) Managing the risks of extreme events and disasters to advance climate change adaptation: special report of the intergovernmental panel on climate change [Internet]. Available: https://www.ipcc.ch/pdf/special-reports/srex/SREX_Full_Report.pdf

IPCC (2014a) Climate change 2014 – impacts, adaptation and vulnerability: part A: global and sectoral aspects [Internet]. Cambridge University Press. Available: http://www.cambridge.org/us/academic/subjects/earth-and-environmental-science/climatology-and-climate-change/climate-change-2014-impacts-adaptation-and-vulnerability-part-global-and-sectoral-aspects-working-group-ii-contribution-ipcc-fifth-assessment-report-volume-1?format=PB

IPCC (2014b) Climate change 2014 – impacts, adaptation and vulnerability: part B: regional aspects [Internet]. Cambridge University Press. Available:

http://www.cambridge.org/us/academic/subjects/earth-and-environmental-science/climatology-and-climate-change/climate-change-2014-impacts-adaptation-and-vulnerability-part-b-regional-aspects-working-group-ii-contribution-ipcc-fifth-assessment-report-volume-2?format=PB#contentsTabAnchor

Jarrah AIZ, Yu G, Krishnamurthy N, Rakshit A (1993) A decision support framework for airline flight cancellations and delays. Transp Sci 27:266–280. doi:10.1287/trsc.27.3.266

Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L et al (1996) The NCEP/NCAR 40-year reanalysis project. Bull Am Meteorol Soc 77:437–471. doi:10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2

Koç Y, Warnier M, Kooij RE, Brazier FMT (2013) An entropy-based metric to quantify the robustness of power grids against cascading failures. Saf Sci 59:126–134. doi:10.1016/j.ssci.2013.05.006

Linkov I, Bridges T, Creutzig F, Decker J, Fox-Lent C, Kröger W et al (2014) Changing the resilience paradigm. Nat Clim Change 4:407–409. doi:10.1038/nclimate2227

Palmer TN, Shutts GJ, Hagedorn R, Doblas-Reyes FJ, Jung T, Leutbecher M (2005) Representing model uncertainty in weather and climate prediction. Annu Rev Earth Planet Sci 33:163–193. doi:10.1146/annurev.earth.33.092203.122552

Salvi K, Ghosh S, Ganguly AR (2015) Credibility of statistical downscaling under nonstationary climate. Clim Dyn 1–33 doi:10.1007/s00382-015-2688-9

Sen P, Dasgupta S, Chatterjee A, Sreeram PA, Mukherjee G, Manna SS (2003) Small-world properties of the Indian railway network. Phys Rev E 67:036106. doi:10.1103/PhysRevE.67.036106

Solé RV, Rosas-Casals M, Corominas-Murtra B, Valverde S (2008) Robustness of the European power grids under intentional attack. Phys Rev E 77:026102. doi:10.1103/PhysRevE.77.026102

Tebaldi C, Knutti R (2007) The use of the multi-model ensemble in probabilistic climate projections. Philos Trans R Soc Lond Math Phys Eng Sci 365:2053–2075. doi:10.1098/rsta.2007.2076

Vespignani A (2010) Complex networks: the fragility of interdependency. Nature 464:984–985. doi:10.1038/464984a

Climate Impacts

▸ Climate Extremes and Informing Adaptation

Climate Resilience

▸ Climate Extremes and Informing Adaptation

Climate Risk Analysis for Financial Institutions

Farid Razzak
Rutgers Business School, Rutgers University, New Brunswick, NJ, USA

Synonyms

Carbon Emissions; Carbon Finance; Carbon Trading; Climate Change; Climate Finance; Climate Trend Analysis; Emissions Trading; GHG; GIS Mobile Remote Sensors; MRV; REDD+; Sequestration; Sustainability Risk

Definition

The climate change phenomenon is widely understood to be magnified by harmful greenhouse gases (GHGs) that are by-products of emissions yielded from advances in human engineering in the energy, technology, transportation, and land development industries. Effectively, the pollution that is being generated from human activities is actively contributing to the imbalance in the planet's climate, therefore creating the scenario where human prosperity may be severely hindered in the near future. Global industrial incentives, regulations, and policies have been formed to mitigate the climate change phenomenon in the form of monetized financial instruments that can help manage the amount of global pollution permitted, financial climate risk disclosures that keep investors informed about climate-related impacts to investments, and environmental sustainability analysis that validates the business continuity of an investment impacted by environmental risks.

The management of future pollution that may contribute to furthering climate change by financially incentivizing more prudent business practices and climate-friendly organizational strategies has created opportunities for climate change investment research. Geographical Information Systems that can provide insight into different aspects of global climate change facilitates data-driven financial investment

decisions which can create a dynamic and robust relationship between the financial and scientific aspects of leveraging climate change mitigation.

This chapter will explore the historical background of global climate change polices and legislation over recent decades; the financial instruments, markets, and risk disclosures that resulted from these policies; the relevant scientific and investment approaches regarding climate change mitigation; how Geographical Information Systems can serve as a crucial tool in the financial applications of climate change mitigation and the future prospects of Geographical Information Systems in this domain.

Historical Background

United Nations Climate Mitigation Polices

The environmental impacts of climate change were not clearly understood by the nations of the world in the early 1980s. The United States was the first government to lead an exploratory study of international environmental risks which included a thorough analysis on climate change effects. This study brought significant awareness to the potential impacts of climate change warranting a more specified scientific study of climate change to illuminate the future risks that nations of the world may have to encounter (Moore 2012).

In 1988, the United Nations World Meteorological Organization (WMO) and the United Nations Environment Program (UNEP) established the Intergovernmental Panel on Climate Change (IPCC) to provide research on the science of climate change, analyze the societal and economical risks due to climate change, and produce strategies to mitigate the impacts that climate change presents for further discussion on the international topic (Moore 2012).

The first assessment report from the IPCC was delivered on 1990 and provided ample evidence to suggest that climate change would be of crucial importance for the near future of environmental risks and policy planning. With subsequent reports from the IPCC echoing similar sentiments and additional evidence supporting the analysis, it was decided at the 1992 United Nations Conference on Environment and Development (UNCED) to formally begin action to create policies for climate change mitigation by commissioning the United Nations Framework Convention on Climate Change (UNFCCC) (Moore 2012; Raufer and Iyer 2012).

The purpose of the UNFCCC was to establish a voluntary commitment from the United States and 153 other nations to reduce harmful greenhouse gas (GHG) emissions to environmentally acceptable levels within the next few decades, to find strategies to reduce the global warming epidemic, and to assess viable options to address inevitable climate change effects on the environment (Moore 2012; Raufer and Iyer 2012). Annual meetings of the parties involved with the UNFCCC have been conducted since the inception of the convention onward, formally referred to as the UNFCCC conference of parties (COP), yielding progressive legislation and policies toward the mitigation of climate change (Moore 2012).

The meetings of most significant and considered progressive milestones for climate change mitigation policies have been that of the COP of 1997 in Kyoto, Japan, and the COP of 2009 in Copenhagen, Denmark (Moore 2012; Raufer and Iyer 2012; Alexander 2013).

Kyoto Protocol

On December 11, 1997, during an annual UNFCC COP in Kyoto, Japan, the Kyoto Protocol was adopted and given an effective date of February 16, 2005. The Kyoto Protocol is widely seen as the first significant step toward an internationally standardized GHG emissions reduction plan that seeks to manage harmful emissions and provide a formalized scalability platform to continuously improve on climate change mitigation strategies (Moore 2012; Raufer and Iyer 2012; Baranzini and Carattini 2014).

The Kyoto Protocol facilitated the reduction of emissions by the establishment of binding agreements among 37 industrialized nations and European nations which committed the nations to reduce their GHG emissions output by an average

of about 5–8 % from the year 1990 emissions output by a 5-year span of 2008–2009 (Moore 2012; Raufer and Iyer 2012; Baranzini and Carattini 2014). More importantly, the Kyoto Protocol placed a larger responsibility and burden on the developed countries due to the accepted notion that they were the primary contributors to the current amount of GHG emissions in the atmosphere (Moore 2012).

Enforcement of the Kyoto Protocol was generally conducted through industrial policies and regulations at the federal and local government levels of each respective participatory nation (Moore 2012). However, the Kyoto Protocol also offered market-based financial and economical incentives to achieve promotion of environment-friendly investments, business practices, and technologies as well as to meet GHG reduction targets via economic and efficient options. The market-based options that the Kyoto Protocol introduced were GHG Emissions Trading, the Clean Development Mechanism (CDM), and the Joint Implementation (JI). Each of the options followed a cap and trade framework in which there was a "cap" or quota on the allowed amount of commodities (emissions allowed to be produced) that were in the market. The "trade" aspect refers to the ability and platform to trade the commodities as an instrument with other market participants (Moore 2012; Raufer and Iyer 2012; Baranzini and Carattini 2014; Henríquez 2013; Kossoy and Guigon 2012).

Greenhouse Gas Emissions Trading

As previously mentioned, the Kyoto Protocol receives commitments from participatory nations to reduce GHG emission levels by a 5-year span in 2008–2012. Some countries may be able to facilitate these targets while being well under target emissions levels, but some may require additional allowance of emissions to meet practical industrial and economic demands. To address this potential issue, Article 17 of the Kyoto Protocol allows for different market-based financial instruments that can allow for trade of excess emissions allowances to countries that may exceed emissions targets (Raufer and Iyer 2012; Baranzini and Carattini 2014; Henríquez 2013; Kossoy and Guigon 2012).

The provision in the Kyoto Protocol also allows the trade of other equally important environmental reduction targets such as the removal units (RMU) based on land use, land use change, and forestry (LULUCF) activities to help mitigate deforestation activities which directly contribute to the natural mitigation of climate change (Moore 2012). Additionally the Kyoto Protocol also offers the global Clean Development Mechanisms that acts as the authority of GHG emissions offset programs which allows industrialized or developing countries that engage in qualified local projects that are designed to help reduce GHG emissions or to provide environmental sustainability to earn certified emissions credits (CER).

A CER is the equivalent of 1 ton of carbon dioxide (CO_2) allowed to be emitted into the atmosphere. CO_2 is one of the harmful GHG emissions that contributes to climate change. With these earned CER credits, they can be traded, sold, or purchased on international markets for the benefit of nations to meet or exceed their GHG emissions reduction targets (Moore 2012; Raufer and Iyer 2012). It is also important to note that 2 % of the income proceeds from CDM projects goes toward the Kyoto Protocol Adaptation Fund which financially backs projects and programs for countries that are impacted most adversely from climate change effects without the ability to mitigate them (Moore 2012). Lastly, the Joint Implementation provision in the Kyoto Protocol under Article 6 allows participating nations to engage in qualified projects that reduce GHG emissions in other countries to earn emissions reduction credits which can be used toward the participatory nations GHG emissions reduction targets. Joint Implementations allows for mutually beneficial partnerships that help to foster prosperity in developing nations while also keeping a focus on the mitigation of climate change (Moore 2012). It is important to note that all of these market-based mechanisms are heavily dependent on accurate analysis, measurement, and forecasting of GHG emissions to be considered a viable climate change mitigation strategy (Raufer and Iyer 2012; Rosenqvist et al. 2003).

GHG emissions trading relies on the overall calculated emissions quotas for each respective nation to determine the appropriate amount of commoditized GHG emissions to be allowed into the market. For this market-based platform to be successful, accurate monitoring and measurements of actual GHG emissions from each nation is required though regulated carbon registries and authorities (Moore 2012; Rosenqvist et al. 2003).

Once regulated appropriately and accurately, national and regional marketplaces are allowed to be established so long as they follow the Kyoto Protocol's fundamental stipulations. This has allowed for emissions marketplaces such as the then Chicago Climate Exchange (CCX) and European Climate Exchange (ECX), both of which operated as a trading platform similar to that of other financial commodities exchanges, and now the Intercontinental Exchange Futures Europe which is now the leading market in emissions trading. All of which followed the European Union's emissions trading scheme (EU ETS) (Raufer and Iyer 2012; Baranzini and Carattini 2014; Rosenqvist et al. 2003; Kossoy and Guigon 2012).

Reducing Emissions from Deforestation and Forest Degradation

At the 11th Conference of Parties (COP11) of the UNFCCC in the year 2005, the reducing emissions from deforestation and forest degradation (REDD) program was established to assist with the reduction of carbon emissions and preservation of the forests (Alexander 2013; Tänzler and Ries 2012). The program was initially developed to support the Clean Development Mechanism (CDM) policies under the Kyoto Protocol to allow developing countries to gain funds for projects focused around the conservation, afforestation, and reforestation leading to the reduction of GHG emissions. The IPCC had earlier concluded that the continual degradation of terrestrial and wetland forests has direct impacts on the mitigation of climate change (Alexander 2013; Tänzler and Ries 2012; Plugge et al. 2011; Nzunda and Mahuve 2011; Wertz-Kanounnikoff et al. 2008). At the 12th Conference of Parties (COP15) of the UNFCC in the year 2007, the Bali Action Plan was ratified, yielding the REDD+ program (REDDplus). REDD+ included all of the original REDD stipulations but also incorporated a focus on funding projects that created sustainable management of forests and further enhancement of forest carbon stocks in developing countries (Alexander 2013). REDD+ programs are based on the science that terrestrial forests, wetland forests, and biodiversity are capable of natural carbon sequestration, where GHG emissions such as carbon dioxide (CO_2) is captured by plant life and where carbon is stored in the soil beneath the plant life (Alexander 2013; Tänzler and Ries 2012; Plugge et al. 2011; Nzunda and Mahuve 2011; Wertz-Kanounnikoff et al. 2008).

Copenhagen Accord

As the Kyoto Protocol's framework approached its expiration date of 2012, a mounting need to develop a new framework that may extend and/or enhance the Kyoto Protocol's principles for climate change mitigation was direly needed. The December 2009 United Nations Climate Change Conference in Copenhagen, Denmark, addressed the concerns of the expiration of the Kyoto Protocol by developing, negotiating, and ratifying the Copenhagen Accord. The Copenhagen Accord committed 186 nations (including the United States) to reduce GHG emissions levels, engage in clean energy projects, and put focus on adaptation projects due to the impacts of climate change. The Copenhagen Accord also requests a technical analysis due in 2015 to determine the need of a new potential CO_2 atmospheric concentration level to maintain to achieve the underlying goals behind climate change mitigation (Moore 2012). The main highlights of the Copenhagen Accord included continued action by countries to manage global temperature increases to under 2 °C, submission of GHG emissions reduction goals by January 2010 from each participatory country, reports from developing countries about climate mitigation actions, and financial funding for environmental conservation projects in developing countries (Moore 2012). The Copenhagen Accord also stipulates that the UNFCCC will continue its

role for financial governance, GHG emissions reporting and monitoring, and scientific climate analysis for the years beyond the expiration of the Kyoto Protocol and will conduct meetings as necessary to achieve appropriate mitigation of climate change (Moore 2012).

Scientific Fundamentals

Given the importance of the climate change phenomenon evidenced by the international climate change mitigation policies mentioned in the previous section, a substantial focus on accurately assessing, measuring, forecasting, and validating the variables of climate change emerges. All the policies and strategies to mitigate climate change fundamentally require measurement and validation methodologies in order to succeed. The fundamental scientific approaches to analyzing climate change and its mitigation provide a key perspective of the future and how to maneuver accordingly to adapt to the potential impacts from climate change.

Proper scientific analysis can benefit all stakeholders within the climate change mitigation framework by providing relative perspective and data interpretation that can potentially drive strategic decisions. This section will briefly review the popular scientific methods that examine aspects of climate change and its mitigation.

Climate Trend Analysis

Climate can be defined as the weather conditions that reveal over an arbitrary period of time, which is usually supported through conventional statistical analysis or statistical diagnostics. Trend, relative to climate, can be defined as the gradual differences of certain climate-related variables over some period of time (Shea 2014).

Traditional statistical time series analysis can be conducted on temperature changes, rainfall measurement, snow patterns, and flooding, and other climate change indicators to detect, estimate, and predict possible emerging climate trends are significant scientific tools to better understand climate change (Shea 2014).

More advanced statistical techniques can be applied to derive more specific data analysis, such as Taylor diagrams, to graphically compare statistical correlation summaries between individual climate patterns(observed or modeled), empirical orthogonal function(EOF), and rotated EOF analysis to interpret potential spatial modes or patterns of variability changes over time (Shea 2014).

All the fundamental Climate Trend Analysis techniques are important for statistical analysis and modeling that can help produce climate change projections for the near future. These projections directly impact climate change mitigation policies, adaptation projects, and business decisions of respective stakeholders.

Surface and Air Temperature Analysis for Land and Sea

Measurements of land air temperature and sea surface temperature (SST) are of significant importance to understand climate conditions in respective regions. This is evidenced by the many decades of available data of the measurements that exist previously to the climate change mitigation conversation. These measurements can provide the data necessary to corroborate findings from climate data models by serving as the ground truth validation source (Hansen et al. 2006). More importantly, the temperature measurements over land and sea can be coordinated in a spatiotemporal plane for pattern analysis, data modeling, and statistical analysis.

Land surface air temperature weather stations are usually stationed in strategic locations throughout a specified region to collect appropriate data and summarize as the highest and lowest temperature recorded for a particular day, which is then reported to a central station which may collect the raw data and combine it with other regional surface temperature weather stations for further analysis. Appropriate standards are followed in the placement of the temperature sensors which ensure they are impartial to influences that may be in close proximity (Hansen et al. 2006).

Similarly, sea surface temperatures (SST) can be collected by remote stations on ships or buoys

equipped with sensors that take measurements of the water surface and summarize the highs and lows of daily water temperature and levels which can be later polled and combined at a central station (Reynolds et al. 2007). Climate scientists rely on statistical anomaly analysis of the water temperature and levels to assess potential inclement weather in the form of cyclones, hurricanes, and tropical storms. With the mentioned techniques, climatologists can develop statistical models that can help estimate, detect, and project future weather patterns (Reynolds et al. 2007).

Satellite resolution imaging may give a broader, less granular depiction of the overall temperature ranges worldwide to help focus on particular patterns or regions of interest, but they are unable to produce the amount of detail that surface level temperature sensors can provide (Kungvalchokechai and Sawada 2013).

The monitoring and analysis of land surface temperature is scientifically linked to the planet's weather and climate patterns, which can be a direct result of increasing atmospheric GHGs. The temperature increases in certain regions can have effects on global glaciers, arctic ice sheets, and vegetation on the planet. Accurately understanding the aspects of the surface temperatures can give scientists a clearer picture about adaptation needs and climate impact projections.

Emissions Analysis

The term greenhouse gases emissions refer directly to the emissions produced from industrial processes, transportation by-products, agricultural by-products, and societal waste products. The gases in questions are the following: carbon dioxide (CO_2), methane (CH_4), nitrous oxide (N_2O), perfluorocarbons ($PFCs$), hydrofluorocarbons ($HFCs$), sulfur hexafluoride (SF_6), as well as the indirect gases that will not be mentioned here (Raufer and Iyer 2012). As mentioned in the previous section, the success of climate change mitigation policies rely directly on the accurate measurements of past, present, and future GHG emissions that could reach the atmosphere, thereby increasing the global temperatures.

GHG emissions control stipulated from climate change mitigation policies require monitoring sensors that can accurately audit the amount of GHG emissions produced. These policies are driven by the science that each GHG has a direct impact on the climate change to the planet. These greenhouse gases that are emitted to the atmosphere create a barrier which does not allow solar heat received from the sun to escape the planet's atmosphere once it has reached surface level, thereby increasing the climate (Myhre et al. 2013).

Scientific methods to derive the atmospheric lifetime, which is the amount of time a gas may stay in the atmosphere; GHG concentrations, which are the estimated values measured in respective until current GHG emissions in the atmosphere; radiative forcing, which is the amount of heat energy the gases absorb and keep in the earth's atmosphere rather than allow it to leave back to space; and global warming potential (GWP), which is the a derived ratio from the atmospheric lifetime and radiative forcing over a specified timescale to determine the impact of the gas on global warming relative to carbon dioxide (CO_2); give climate scientists quantifiable metrics to weigh and assess the impact of each GHG emission to appropriate mathematical models and climate data models (Myhre et al. 2013). These methodology and analysis give climate scientists quantifiable terms to weight and assess the different intensities and impacts of each GHG emission to appropriate mathematical models and climate data models. An important note about emissions that impact climate change include both natural (water vapor) and anthropogenic (pollution or pollutants from human activity) sources which both need to be accurately quantified and analyzed (Myhre et al. 2013).

Carbon Capture and Sequestration Analysis

The term carbon sequestration refers to the natural or synthetic process of capturing and/or storing carbon dioxide (CO_2) emissions, thereby mitigating climate change by reducing the amount of the GHG emission to reach or remain in the atmosphere. The natural process of achieving a balance of CO_2 emissions and climate change comes in the form of forested wetlands, terrestrial

forests, and plant life, all of which have the capability to capture CO_2 emissions for consumption and store carbon into the soil which their roots are deeply entrenched (Freedman 2014; Alexander 2013). The synthetic process captures carbon-based emissions at the point of production from industrial facilities that produce the emissions and transport it deep underneath land or sea where it may dissolve or be stored indefinitely (Katzer et al. 2007).

Both the natural and synthetic carbon sequestration processes require accurate calculations and depictions of the amount of CO_2 being captured and/or stored to determine the effectiveness of the mitigation (Freedman 2014; Alexander 2013; Katzer et al. 2007). To achieve this feat synthetically, scientists need to mathematically calculate the amount of CO_2 in units of metric tons that can be properly captured and stored under the planet's land and sea without causing adverse effects to the environment. The terrestrial or natural approach would require scientists to determine the amount of CO_2 that plant life from forested areas can capture and store the emissions to achieve a substantial mitigation to climate change (Freedman 2014; Katzer et al. 2007). This is evidenced by the ratification of the REDD+ policy mentioned in the previous section.

Geographical Information Systems

The scientific analysis techniques, data sources, and respective stakeholder interests in climate change mitigation have created a demand for platforms that can dynamically bring together the different aspects that are required to perform effective climate change mitigation analysis. Advances in information technology, accessibility to data sources, and economic costs of data storage have allowed for the availability of Geographical Information Systems (GIS) platforms to be developed for robust analysis requirements of climate change mitigation research. GIS serves as a tool for scientific-based climate research by practically combining the many different scientific analysis techniques with appropriate data streams and visualizations to provide data-driven insights for climate change stakeholders.

Geographical Information Systems can be developed and customized to successfully achieve the feature requirements for different climate analysis purposes, but some of the conceptual fundamentals that a GIS system developed to analyze climate change usually revolve around the following abilities.

Mapping

A GIS system for climate change analysis should have the ability to render a data canvas of the geographical region of interest or global map where data overlays can be produced based on appropriate data streams to represent appropriate depictions of the said data.

Gridding and Regridding

Gridded data can be high-resolution images of a certain geographical region that does not give the total perspective of surrounding regions due to computational or storage limitations. Segments or fragments of a larger overall high-resolution image are provided, which is a part of a sequenced grid of neighboring images that can be examined individually. Due to the nature of the high-resolution image, data overlays, points of interests, and data streams can still be integrated using GIS technologies but only specific to the gridded image provided (Shea 2014; Reynolds and Smith 1994).

Regridding refers to the interpolation of one grid resolution image to a different grid resolution image, usually that of a sequence that depicts the immediate neighboring resolutions of a specific geographical region. Different methods such as temporal, vertical, or horizontal interpolation is used to combine the resolutions, but most commonly spatial (horizontal) interpolation is utilized (Shea 2014; Reynolds and Smith 1994). Depending on the type of analysis and data, appropriate interpolation techniques are required. To perform quantitative analysis on data points across many gridded resolutions, regretting across a common grid is required to avoid misleading numerical calculations among the data from different grid images. GIS applications and platforms provide many different interpolation techniques for regridding

which allows for more accurate data analysis (Shea 2014; Reynolds and Smith 1994). This is a crucial tool that can ensure accuracy of very computationally large amounts of geographical data.

Monitoring and Measurement

GIS applications and systems can be configured to dynamically operate with real-time data streams from third-party data vendors or remote sensors that may provide climate-based or emissions-based information. A platform that can actively receive the data streams from the sensors and spatially visualize and overlay the data on a geographical plane relative to the sensor's logistical location can provide an automated monitoring system to detect potentially interesting climate or emissions patterns which can be practically interpreted depending on stakeholder interests (Rosenqvist et al. 2003; Reynolds and Smith 1994; Gibbs et al. 2007; Palmer Fry 2011). The monitoring aspect of GIS indicates the ability to process large amounts of data, store the data, and visualize the data in minimal amounts of time to provide insight to the stakeholder. Without this aspect or ability of a GIS platform or system, climate change analysis techniques would not benefit greatly from GIS technologies.

Reporting

The ability to retrieve information and analysis dynamically in an easy to interpret format is a key fundamental for a GIS system that may be developed for the purposes of climate analysis. The reporting mechanism allows the user of the system to gather important data and intelligence that could lead insight-driven decision. Both monitoring and reporting are crucial aspects of a GIS system designed for climate analysis due to the fact that reporting is based on the data derived from monitoring, and the insights from reporting are the primary output that analysis will be conducted on. Inaccuracies or inconsistencies in reporting may deem the GIS system obsolete, but accurate reporting could mean a substantial increase in productivity, efficiency, and progress in conducting relevant analysis and research on climate change (Rosenqvist et al. 2003; Reynolds and Smith 1994; Gibbs et al. 2007; Palmer Fry 2011).

Verification

The ability to monitor and report on different aspects of climate change based on statistical models or projections derived from historical data may not always accurately portray the actual observational data. Scientific analysis requires corroborated ground truth data to validate if the data models developed from historical data or data from a different region is statistically significant enough to be accurate. Verification is a critical factor in climate change mitigation policies due to the reliance on the ability to correctly determine climate change and emissions levels to properly incentivize global participants to achieve the common goal to degrade atmospheric temperature increases (Moore 2012). To achieve ground truth validations, climate change mitigation policies emphasize the requirement of approved sensors that can accurately verify the integrity of measurements taken at the point of production. This can be interpreted as remote sensors that are capable of measuring the "ground truth" data that is required in climate-based analysis scenarios. (Rosenqvist et al. 2003; Reynolds and Smith 1994; Gibbs et al. 2007; Palmer Fry 2011). GIS systems need to be scalable and adaptable to incorporate regulatory ground truth data or provide the appropriate information technology that meet the standards of climate mitigation policies.

Key Applications

Some of the aspects of climate change mitigation policies discussed offer financial instruments, incentives, and platforms for interested investors, impacted industrial stakeholders and participating nations to explore opportunities and strategies that can directly, indirectly, or residually impede the global temperature increase. Stakeholders who may decide to participate in the incentives offered by climate mitigation polices are aware that proper knowledge and analysis of climate

change aspects that may be related to respective interests may provide a competitive edge for potential investment decisions (Kossoy and Guigon 2012). This section will explore some practical examples of Geographical Information Systems that perform climate science-related analysis and their application to different financial investment research.

Climate Finance

The term climate finance represents the financial mechanisms set in place by climate change mitigation policies, such as Kyoto Protocol and Copenhagen Accord, which allow for national, regional, and international parties to have access to financing channels specifically for climate change mitigation and adaptation projects and programs (Kossoy and Guigon 2012; Buchner et al. 2011). These projects and programs are developed based on achieving minimal carbon-based emissions footprints and resiliency to climate change through appropriate research and economic development. The term had been originally coined to refer to the obligations that developed countries committed to developing countries under the ratified UNFCCC policies; however, the term is now more synonymous with the all financial procedures and flows relating to climate change mitigation and adaptation projects and programs (Kossoy and Guigon 2012; Buchner et al. 2011). Financial funding can be provided from government budgets, domestic budgets, capital markets, and public and/or private sectors mediated through bilateral financial institutions, multilateral financial institutions, and development cooperation agencies or directly from the UNFCC itself via the Green Climate Fund, NGOs(nongovernmental organizations), and/or private sector. Investments decisions and strategies in renewable energy can potentially be considered climate finance if the renewable energy projects and programs qualify under the UNFCC guidelines (Kossoy and Guigon 2012; Buchner et al. 2011).

The financing projects and programs designed to mitigate or adapt the effects of climate change, such as the previously discussed Carbon Offset Programs, Clean Development Mechanisms, and Joint Implementation programs, have reached billions of dollars a year on average which is forecasted to grow into the trillions in the near future (Buchner et al. 2011; Moore 2012). To effectively monitor the funding needs, progress, success, and completion of projects and programs, the intermediaries of the climate finance framework rely on GIS-based tools and analysis to make informed data-driven decisions.

Carbon Finance

The UNFCCC stipulations of pollution and emissions control creates a realm in which carbon footprints and greenhouse gases are constrained to limit the potential increase in global climate change (Moore 2012). This constraint creates a commodity out of the amount of carbon-based or GHG emissions permitted for industrial and national interests (Raufer and Iyer 2012). Climate mitigation policy frameworks have promoted the investments in projects and programs that reduce the previously mentioned GHG emissions as well as provided a platform where the commodity of allowed emissions amounts are monetized into financial instruments that are tradable in a market-based cap and trade framework. The platform where the commoditized emissions allowances are exchanged is typically referred to the carbon market, while the overall concept of investing and trading these commodities can be represented by the term carbon finance. (Moore 2012; Raufer and Iyer 2012; Kossoy and Guigon 2012).

Carbon finance leverages the Kyoto Protocol's Clean Development Mechanisms and Joint Implementation framework to help facilitate the investments into emissions reductions projects to earn or trade emissions allowances or credits (Kossoy and Guigon 2012; Henríquez 2013). The World Bank facilitates carbon finance through its own carbon finance unit which purchases carbon credits or GHG emissions reductions generated from projects or programs in developing countries or transitioning economies to their fund contributors that employ their services, usually in the form of governments or companies with an interest in attaining or trading the carbon credits (Lewis 2010). The World Bank can achieve this by providing carbon funds

and facilities which contribute to projects and programs that can yield carbon credits or GHG emissions reductions according to the Kyoto Protocol's Clean Development Mechanism and Joint Implementation frameworks (Lewis 2010; Henríquez 2013; Kossoy and Guigon 2012; Moore 2012). Essentially, the World Bank invests and supports projects and programs that qualify to earn carbon credits, which the World Bank can acquire and sell to interested parties through their carbon finance business (Lewis 2010; Henríquez 2013). Carbon credits are the official allowance of 1 metric ton of CO_2 or equivalent gases earned through approved projects or programs that progress the climate change mitigation agenda (Moore 2012).

The pricing of allowed carbon-based emissions is based on the limited supply and high demand for carbon credits (Litterman 2013; Henríquez 2013; Kossoy and Guigon 2012). To make informed investment decisions from the buyers and sellers' positions, proper financial analysis and careful investment research need to occur on the projects and programs that yield the carbon credits. Climate finance shares elements with carbon finance with respect to the dependence on GIS tools and analysis to determine if projects and programs properly qualify and succeed to earn carbon credits. Carbon finance, however, depends on both financial analysis and scientific research to make proper investment decisions (Litterman 2013; Henríquez 2013; Kossoy and Guigon 2012). GIS tools that combine both give stakeholders in carbon finance more insight when making critical decisions. Investment research can potentially incorporate GIS-based tools to understand estimates or forecasts of potential deficit or surplus in carbon emissions. GIS serves as an important investment research tool in the carbon finance domain because of the similarities to the commodities markets. Understanding the fundamentals of the commodity may help provide beneficial insight when investing in such a commodity. In the case of carbon finance, the commodity are the carbon credits or GHG emissions allowances yielded from climate mitigation projects or purchased from a carbon market at a market-competitive price.

Sustainability Risk Management

The awareness of the effects climate change may have on society and economy creates a legitimate business concern for investor and stakeholder confidence. Organizations which employ business strategies that do not take environmental risk factors, such as climate change, into consideration when planning, operating, or expanding may be adversely impacted by evolving climate change mitigation policies or effects of climate change. Organizations must solidify confidence with investors and stakeholders by engaging in business strategies that align operations and revenue goals with environmentally friendly policies. Global and national compliance regulations stemming from global climate change mitigation efforts call for renewable energy, environmentally friendly business practices, and sustainability initiatives. Organizations may not have a strategic initiative or outlook to align their business strategies with considerations for regulatory and environmental risks associated with climate change which can lead to lack of investment confidence and appeal (Zu 2013; Baumast 2013; Schmiedeknecht 2013). An example of this can be evidenced by organizational climate risk disclosures that are disseminated as public information for investors and stakeholders to review potential liabilities and assets of the respective business that can be impacted by environmental risk factors.

Sustainability Risk Management considers the optimal business strategy for an organization to achieve an effective and efficient balance between the prosperity of a business and its adherence to environmentally friendly policies. Traditional risk management and climate science techniques may be performed on business assets, interests, and liabilities to assess relative impacts to the organizational profit goals (Zu 2013; Baumast 2013). Sustainability and vulnerability assessments conducted in depth consider environmental risk factors such as floods, natural disasters, and climate change and how they may be detrimental to the business (Zu 2013; Baumast 2013; Schmiedeknecht 2013). They also consider adaptation strategies to achieve residency in the wake of such environmental risks for business

continuity and prosperity which can be translated into long-term confidence for investors. To achieve such assessments, GIS tools and analysis can be employed to analyze environmental risk factors to business operations, supply chains, and other applicable business assets. Forecasting and simulation models of risk factors are considered as well as financial burdens that may be experienced by the impacted business (Zu 2013). After examination of each business process and potential environmental risks that may impact them are analyzed, strategies are developed to minimize the risks (Zu 2013). Integrated technologies utilizing GIS-based analysis and data management tools can be employed to conduct automated monitoring, auditing, and reporting on sustainability models and goals to achieve compliance. The identification of potential risks and issues impacting business interests early on can help the business maneuver its directional strategy to avoid costly regulatory failures or repetitional damage (Zu 2013; Schmiedeknecht 2013).

Sustainability Risk Management extends into the financial markets by allowing organizations that satisfy corporate sustainability assessments to be held in Sustainability Indices (Schmiedeknecht 2013). Sustainability indices represent an index of organizations considered to be socially responsible, environmentally friendly, and sustainable in the event of environmental risks. Investment firms that offer Sustainability Indices may market them as safer and resilient to climate change to potential investors who seek investment confidence relative to environmental risks (Schmiedeknecht 2013). Organizations who are able to reach Sustainability Indices may be considered a safer investment option compared to organizations that cannot achieve the same qualifications.

Future Directions

Climate change analysis and Geographical Information Systems share a relationship that will only evolve as the awareness and applications of climate change mitigation become more prevalent. GIS systems and tools are used to provide meaningful insight and intelligence for monitoring, reporting, and verification applications of climate change analysis. GIS technologies and systems may combine climate-related research and analysis data sources on geographical planes that can help perform traditional analytical techniques to yield data models that can be used to make strategic decisions. The continued emergence and demand for GIS and geospatial analysis skills in the climate science and investment research markets can be expected to grow as the relevance of climate-related applications, such as climate finance, carbon finance, and sustainability management, increases.

Important trends in climate change research, GIS, and financial applications are briefly discussed in the following sections.

Mobile GIS Remote Sensor Networks

Mobile GIS remote sensor networks are considered an important topic in both climate research and GIS. To optimally and efficiently design, monitoring, reporting, and verification GIS systems that can potentially be incorporated into REDD+ projects and programs or other climate finance-funded projects are crucial to attain accurate data at the highest integrity standards (Samek et al. 2013; Rosenqvist et al. 2003; Patenaude et al. 2004). Much research is being conducted to optimize and propose equipment and techniques to achieve an economically and practically feasible approach to achieving GIS remote sensor networks that can potentially become a standardized method to collect and validate data such as emissions, carbon storage, carbon sequestration rates, air temperatures, and other climate change-related datapoint.

Data-Driven GIS Decision-Making Tools

GIS systems that can properly collect data from multiple data sources and conduct application-specific analysis on the said data with potential business logic are an area that climate change stakeholders are seeking to expand (Sizo et al. 2014; Benz et al. 2004; Ganguly et al. 2005). Climate-based financial and regulatory applications to automate business intelligent GIS

systems that can perform dynamic analytical observations to yield insights to assist in decision-making scenarios can directly provide value-added service for climate change stakeholders. Automated sustainability assessments for organizations or GIS-based systems that can signal important investment research analysis are some of the many applications that data-driven geospatial analysis and technology is making available to the climate research and finance-based industries (Tomlinson 2007; Zu 2013).

Cross-References

- ArcGIS: General-Purpose GIS Software
- Climate Adaptation, Introduction
- Climate Change and Developmental Economies
- Climate Extremes and Informing Adaptation
- Climate Hazards and Critical Infrastructures Resilience
- Data Models in Commercial GIS Systems
- Financial Asset Analysis with Mobile GIS
- Geosensor Networks, Formal Foundations
- GPS Data Processing for Scientific Studies of the Earth's Atmosphere and Near-Space Environment

References

Alexander S (2013) Reducing emissions from deforestation and forest degradation. In: Finlayson M, McInnes R, Everard M (eds) Encyclopedia of wetlands: wetland management, vol 2. Springer, Berlin/Heidelberg

Baranzini A, Carattini S (2014) Taxation of emissions of greenhouse gases. In: Freedman B (ed) Global environmental change. Handbook of global environmental pollution, vol 1. Springer, Berlin/Heidelberg, pp 543–560

Baumast A (2013) Carbon disclosure project. In: Idowu SO, Capaldi N, Zu L, Gupta AD (eds) Encyclopedia of corporate social responsibility. Springer, Berlin/Heidelberg, pp 302–309

Benz UC, Hofmann P, Willhauck G, Lingenfelder I, Heynen M (2004) Multi-resolution, object-oriented fuzzy analysis of remote sensing data for gis-ready information. ISPRS J Photogramm Remote Sens 58(3):239–258

Buchner B, Falconer A, Hervé-Mignucci M, Trabacchi C, Brinkman M (2011) The landscape of climate finance. Climate Policy Initiative, Venice, p 27

Freedman B (2014) Maintaining and enhancing ecological carbon sequestration. In: Freedman B (ed) Global environmental change. Handbook of global environmental pollution, vol 1. Springer, Berlin/Heidelberg, pp 783–801

Ganguly AR, Gupta A, Khan S (2005) Data mining technologies and decision support systems for business and scientific applications. In: Encyclopedia of data warehousing and mining. Idea Group Publishing

Gibbs HK, Brown S, Niles JO, Foley JA (2007) Monitoring and estimating tropical forest carbon stocks: making REDD a reality. Environ Res Lett 2(4):045023

Hansen JE, Ruedy R, Sato M, Lo K (2006) NASA GISS surface temperature (GISTEMP) analysis. Trends: a compendium of data on global change

Henríquez BLP (2013) Environmental commodities markets and emissions trading, towards a low carbon future. Routledge

Katzer J, Ansolabehere S, Beer J, Deutch J, Ellerman AD, Friedmann SJ, Herzog H, Jacoby HD, Joskow PL, McRae G et al (2007) The future of coal: options for a carbon-constrained world. Massachusetts Institute of Technology, Boston

Kossoy A, Guigon P (2012) State and trends of the carbon market. World Bank, Washington DC

Kungvalchokechai S, Sawada H (2013) The filtering of satellite imagery application using meteorological data aiming to the measuring, reporting and verification (MRV) for REDD. Asian J Geoinf 13(3)

Lewis JI (2010) The evolving role of carbon finance in promoting renewable energy development in China. Energy Policy 38(6):2875–2886

Litterman B (2013) What is the right price for carbon emissions. Regulation 36:38

Moore C (2012) Climate change legislation: current developments and emerging trends. In: Chen W-Y, Seiner J, Suzuki T, Lackner M (eds) Handbook of climate change mitigation. Springer, Berlin/Heidelberg, pp 43–87

Myhre G, Shindell D, Bréon F-M, Collins W, Fuglestvedt J, Huang J, Koch D, Lamarque J-F, Lee D, Mendoza B, Nakajima T, Robock A, Stephens G, Takemura T, Zhang H (2013) Climate change 2013: the physical science basis. Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change, book section 8. Cambridge University Press, Cambridge/New York, pp 659–740

Nzunda EF, Mahuve TG (2011) A swot analysis of mitigation of climate change through REDD. In: Filho WL (ed) Experiences of climate change adaptation in Africa. Climate change management. Springer, Berlin/Heidelberg, pp 201–216

Palmer Fry BP (2011) Community forest monitoring in REDD+: the "M" in MRV? Environ Sci Policy 14(2):181–187

Patenaude G, Hill RA, Milne R, Gaveau DLA, Briggs BBJ, Dawson TP (2004) Quantifying forest above ground carbon content using LiDAR remote sensing. Remote Sens Environ 93(3):368–380

Plugge D, Baldauf T, Köhl M (2011) Reduced emissions from deforestation and forest degradation (REDD): why a robust and transparent monitoring, reporting and verification (MRV) system is mandatory. In: Climate change–research and technology for adaptation and mitigation. InTech, Rijeka, pp 155–170
Raufer R, Iyer S (2012) Emissions trading. In: Chen W-Y, Seiner J, Suzuki T, Lackner M (eds) Handbook of climate change mitigation. Springer, New York, pp 235–275
Reynolds RW, Smith TM (1994) Improved global sea surface temperature analyses using optimum interpolation. J Clim 7(6):929–948
Reynolds RW, Smith TM, Liu C, Chelton DB, Casey KS, Schlax MG (2007) Daily high-resolution-blended analyses for sea surface temperature. J Clim 20(22):5473–5496
Rosenqvist Å, Milne A, Lucas R, Imhoff M, Dobson C (2003) A review of remote sensing technology in support of the kyoto protocol. Environ Sci Policy 6(5):441–455
Samek JH, Skole DL, Thongmanivong S, Lan DX, Van Khoa P (2013) Deploying internet-based MRV tools and linking ground-based measurements with remote sensing for reporting forest carbon. APN Sci Bull Issue 3, 4
Schmiedeknecht MH (2013) Dow jones sustainability indices. In: Idowu SO, Capaldi N, Zu L, Gupta AD (eds) Encyclopedia of corporate social responsibility. Springer, Berlin/Heidelberg, pp 832–838
Shea D (2014) Climate data guide retrieved from https://climatedataguide.ucar.edu/climate-data-tools-and-analysis/
Sizo A, Bell S, Noble B (2014) Automated gis routine for strategic environmental assessment: a spatiotemporal analysis of urban and wetland change
Tänzler D, Ries F (2012) International climate change policies: the potential relevance of REDD+ for peace and stability. In: Scheffran J, Brzoska M, Brauch HG, Link PM, Schilling J (eds) Climate change, human security and violent conflict. Hexagon Series on Human and Environmental Security and Peace, vol 8. Springer, Berlin/Heidelberg, pp 695–705
Tomlinson RF (2007) Thinking about GIS: geographic information system planning for managers. ESRI, Inc., Redlands
Wertz-Kanounnikoff S, Verchot LV, Kanninen M, Murdiyarso D (2008) How can we monitor, report and verify carbon emissions from forests. Moving ahead with REDD: issues, options, and implications. Center for International Forestry Research (CIFOR), Bogor, pp 87–98
Zu L (2013) Sustainability risk management. In: Idowu SO, Capaldi N, Zu L, Gupta AD (eds) Encyclopedia of corporate social responsibility. Springer, Berlin/Heidelberg, pp 2395–2407

Climate Risks

▸ Climate Extremes and Informing Adaptation

Climate Trend Analysis

▸ Climate Risk Analysis for Financial Institutions

Climate Variability

▸ Climate Extremes and Informing Adaptation

Cloaking Algorithms

Chi-Yin Chow
Department of Computer Science, City University of Hong Kong, Hong Kong, China

Definition

Spatial cloaking is a technique used to blur a user's exact location into a spatial region in order to preserve her location privacy. The blurred spatial region must satisfy the user's specified privacy requirement. The most widely used privacy requirements are k-anonymity and minimum spatial area. The k-anonymity requirement guarantees that a user location is indistinguishable among k users. On the other hand, the minimum spatial area requirement guarantees that a user's exact location must be blurred into a spatial region with an area of at least $\mathcal{A}$, such that the probability of the user being located in any point within the spatial region is $1/\mathcal{A}$. A user location must be blurred by a spatial cloaking algorithm either on the client side or a trusted third-party before it is submitted to a location-based database server.

Main Text

This article surveys existing spatial cloaking techniques for preserving users' location privacy in location-based services (LBS) where users have to continuously report their locations to the database server in order to obtain the service. For example, a user asking about the nearest gas station has to report her exact location. With untrustworthy servers, reporting the location information may lead to several privacy threats. For example, an adversary may check a user's habit and interest by knowing the places she visits and the time of each visit. The key idea of a spatial cloaking algorithm is to perturb an exact user location into a spatial region that satisfies user specified privacy requirements, e.g., a k-anonymity requirement guarantees that a user is indistinguishable among k users.

Cross-References

- ▸ Location-Based Services: Practices and Products
- ▸ Privacy Preservation of GPS Traces

Cloaking Algorithms for Location Privacy

Chi-Yin Chow
Department of Computer Science, City University of Hong Kong, Hong Kong, China

Synonyms

Anonymity; Location anonymization; Location blurring; Location perturbation; Location-based services; Location-privacy; Nearest neighbor; Peer to peer; Privacy

Definition

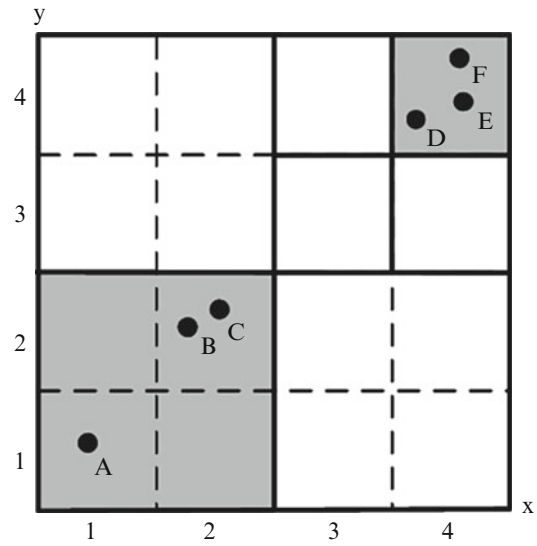
Spatial cloaking is a technique to blur a user's exact location into a spatial region in order to preserve her location privacy. The blurred spatial region must satisfy the user's specified privacy requirement. The most widely used privacy requirements are k-anonymity and minimum spatial area. The k-anonymity requirement guarantees that a user location is indistinguishable among k users. On the other hand, the minimum spatial area requirement guarantees that a user's exact location must be blurred into a spatial region with an area of at least $\mathcal{A}$, such that the probability of the user being located in any point within the spatial region is $\frac{1}{\mathcal{A}}$. A user location must be blurred by a spatial cloaking algorithm either on the client side or a trusted third party before it is submitted to a location-based database server.

Historical Background

The emergence of the state-of-the-art location-detection devices, e.g., cellular phones, global positioning system (GPS) devices, and radio-frequency identification (RFID) chips, has resulted in a location-dependent information access paradigm, known as location-based services (LBS). In LBS, mobile users have the ability to issue snapshot or continuous queries to the location-based database server. Examples of snapshot queries include *where is the nearest gas station* and *what are the restaurants within one mile of my location*, while examples of continuous queries include *where is the nearest police car for the next one hour* and *continuously report the taxis within one mile of my car location*. To obtain the precise answer of these queries, the user has to continuously provide her exact location information to a database server. With untrustworthy database servers, an adversary may access sensitive information about

individuals based on their location information and queries. For example, an adversary may identify a user's habits and interests by knowing the places she visits and the time of each visit.

The k-anonymity model (Sweeney 2002a, b) has been widely used in maintaining privacy in databases (Bayardo and Agrawal 2005; LeFevre et al. 2006, 2005; Meyerson and Williams 2004). The main idea is to have each tuple in the table as k-anonymous, i.e., indistinguishable among other $k-1$ tuples. However, none of these techniques can be applied to preserve user privacy for LBS, mainly for the reason that these approaches guarantee the k-anonymity for a snapshot of the database. In LBS, the user location is continuously changing. Such dynamic behavior requires continuous maintenance of the k-anonymity model. In LBS, k-anonymity is a user-specified privacy requirement which may have a different value for each user.



Cloaking Algorithms for Location Privacy, Fig. 1 Adaptive interval cloaking ($k = 3$)

Scientific Fundamentals

Spatial cloaking algorithms can be divided into two major types: k-anonymity spatial cloaking (Chow et al. 2006; Gedik and Liu 2005; Gruteser and Grunwald 2003; Gruteser and Liu 2004; Kalnis et al. 2006; Mokbel et al. 2006) and uncertainty spatial cloaking (Cheng et al. 2006). k-anonymity spatial cloaking aims to blur user locations into spatial regions which satisfy the user's specified k-anonymity requirement, while uncertainty spatial cloaking aims to blur user locations into spatial regions which stratify the user's specified minimum spatial area requirement.

Adaptive Interval Cloaking

This approach assumes that all users have the same k-anonymity requirements (Gedik and Liu 2005). For each user location update, the spatial space is recursively divided in a KD-tree-like format until a minimum k-anonymous subspace is found. Such a technique lacks scalability as it deals with each single movement of each user individually. Figure 1 depicts an example of the adaptive interval cloaking algorithm in which the k-anonymity requirement is three. If the algorithm wants to cloak user A's location, the system space is first divided into four equal subspaces, $\langle(1,1),(2,2)\rangle$, $\langle(3,1),(4,2)\rangle$, $\langle(1,3),(2,4)\rangle$, and $\langle(3,3),(4,4)\rangle$. Since user A is located in the subspaces $\langle(1,1),(2,2)\rangle$, which contains at least k users, these subspaces are further divided into four equal subspaces, $\langle(1,1),(1,1)\rangle$, $\langle(2,1),(2,1)\rangle$, $\langle(1,2),(1,2)\rangle$, and $\langle(2,2),(2,2)\rangle$. However, the subspace containing user A does not have at least k users, so the minimum suitable subspace is $\langle(1,1),(2,2)\rangle$. Since there are three users, D, E, and F, located in the cell (4,4), this cell is the cloaked spatial region of their locations.

CliqueCloak

This algorithm assumes a different k-anonymity requirement for each user (Gedik and Liu 2005). CliqueCloak constructs a graph and cloaks user locations when a set of users forms a clique in the graph. All users share the same cloaked spatial region which is a minimum bounding box covering them. Then, the cloaked spatial region is reported to a location-based database server as their locations. Users can also specify the maximum area of the cloaked region which is

considered as a constraint on the clique graph, i.e., the cloaked spatial region cannot be larger than the user's specified maximum acceptable area.

k-Area Cloaking

This scheme keeps suppressing a user location into a region which covers at least $k-1$ other sensitive areas, e.g., restaurants, hospitals, and cinemas around the user's current sensitive area (Gruteser and Liu 2004). Thus, the user resident area is indistinguishable among k sensitive areas. This spatial cloaking algorithm is based on a map which is partitioned into zones, and each zone contains at least k sensitive areas. Thus, the continuous movement of users is just abstracted as moving between zones. Users can specify their own privacy requirements by generalizing personalized sensitivity maps.

Hilbert *k*-Anonymizing Spatial Region (hilbASR)

Here, users are grouped together into variant buckets based on the Hilbert ordering of user locations and their own k-anonymity requirements (Kalnis et al. 2006). Using the dynamic hilbASR, the cloaked spatial regions of users A to F can be determined by using two equations, $start(u)$ and $end(u)$, which are depicted in Fig. 2, where $start(u)$ and $end(u)$ indicate the start and end rankings of a cloaked spatial region, respectively, u is a user identity, and the dotted line represents the Hilbert ordering.

Nearest-Neighbor *k*-Anonymizing Spatial Region (nnASR)

This is the randomized version of a k-nearest neighbor scheme (Kalnis et al. 2006). For a user location u, the algorithm first determines a set S of k-nearest neighbors of u, including u. From S, the algorithm selects a random user u' and forms a new set S' that includes u' and the $k-1$ nearest neighbors of u'. Then, another new set S'' is formed by taking a union between S and S''. Finally, the required cloaked spatial region is the bounding rectangle or circle which covers all the users of S''.

Uncertainty

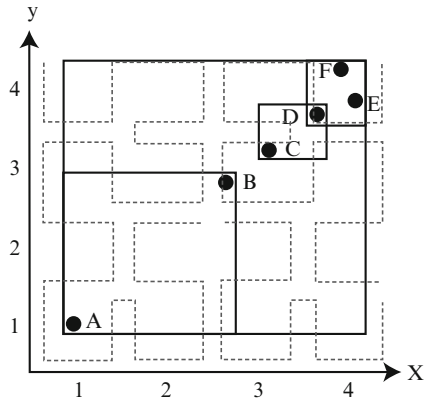
This approach proposes two uncertainty spatial cloaking schemes, *uncertainty region* and *coverage of sensitive area* (Cheng et al. 2006). The uncertainty region scheme simply blurs a user location into an uncertainty region at a particular time t, denoted as $U(t)$. The larger region size means a more strict privacy requirement. The coverage of sensitive area scheme is proposed for preserving the location privacy of users who are located in a sensitive area, e.g., hospital or home. The coverage of sensitive area for a user is defined as $\text{Coverage} = \frac{\text{Area(sensitive/area)}}{\text{Area(uncertainty/region)}}$. The lower value of the coverage indicates a more strict privacy requirement.

Casper

Casper supports both the k-anonymity and minimum spatial area requirements (Mokbel et al. 2006). System users can dynamically change their own privacy requirements at any instant. It proposes two grid-based pyramid structures to improve system scalability, *complete pyramid* and *incomplete pyramid*.

Complete Pyramid

Figure 3a depicts the complete pyramid data structure which hierarchically decomposes the spatial space into H levels where a level of height h has 4^h grid cells. The root of the pyramid is of height zero and has only one grid cell that covers the whole space. Each pyramid cell is represented as (cid, N), where cid is the cell identifier and N is the number of mobile users within the cell boundaries. The pyramid structure is dynamically maintained to keep track of the current number of mobile users within each cell. In addition, the algorithm keeps track of a hash table that has one entry for each registered mobile user with the form $(uid, profile, cid)$, where oid is the mobile user identifier, $profile$ contains the user-specified privacy requirement, and cid is the cell identifier in which the mobile user is located. The cid is always in the lowest level of the pyramid (the shaded level in Fig. 3a).



Users	A	B	C	D	E	F
k_u	6	2	2	3	3	3
$Rank(u)$	0	1	2	3	4	5
$Start(u)$	0	0	2	3	3	3
$End(u)$	5	1	3	5	5	5

$$Start(u) = Rank(u) - (Rank(u) \bmod k_u)$$
$$End(u) = Start(u) + k_u - 1$$

Cloaking Algorithms for Location Privacy, Fig. 2 hilbASR

Incomplete Pyramid

The main idea of the incomplete pyramid structure is that not all grid cells are appropriately maintained. The shaded cells in Fig. 3b indicate the lowest level cells that are maintained.

Cloaking Algorithm

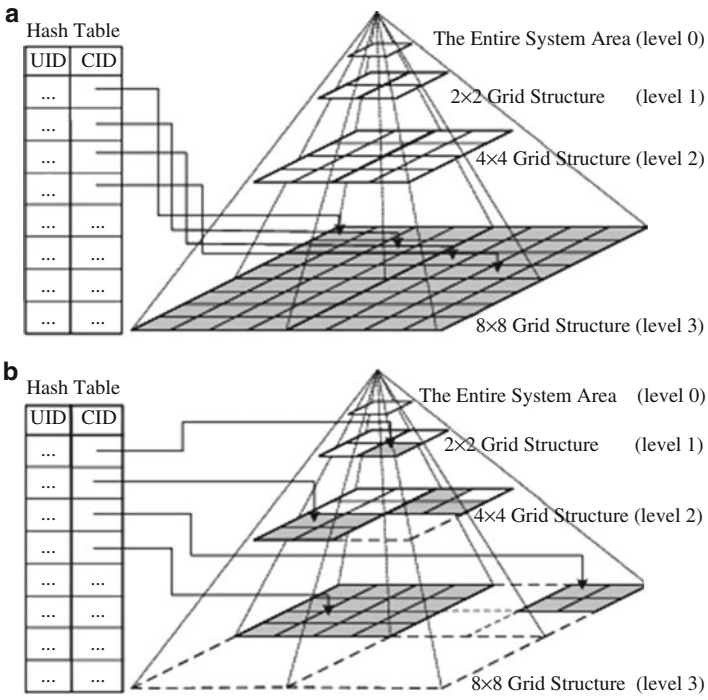
Casper adopts a bottom-up cloaking algorithm which starts at a cell where the user is located at from the lowest maintained level and then traverses up the pyramid structure until a cell satisfying the user-specified privacy requirement is found. The resulting cell is used as the cloaked spatial region of the user location. In addition to the regular maintenance procedures as that of the basic location anonymizer, the adaptive location anonymizer is also responsible for maintaining the shape of the incomplete pyramid. Due to the highly dynamic environment, the shape of the incomplete pyramid may have frequent changes. Two main operations are identified in order to maintain the efficiency of the incomplete pyramid structure, namely, *cell splitting* and *cell merging*.

In the cell splitting operation, a cell *cid* at level i needs to be split into four cells at level $i + 1$ if there is at least one user u in *cid* with a privacy profile that can be satisfied by some cell at level $i + 1$. To maintain such criterion, Casper keeps track of the most relaxed user u_r for each cell. If a newly coming object u_{new} to the cell *cid* has a more relaxed privacy requirement than u_r, the algorithm checks if splitting cell *cid* into four cells at level $i + 1$ would result in having a new cell that satisfies the privacy requirements of u_{new}. If this is the case, the algorithm will split cell *cid* and distribute all its contents to the four new cells. However, if this is not the case, the algorithm just updates the information of u_r. In case one of the users leaves cell *cid*, the algorithm will just update u_r if necessary.

In the cell merging operation, four cells at level i are merged into one cell at a higher level $i - 1$ only if all the users in the level i cells have strict privacy requirements that cannot be satisfied within level i. To maintain this criterion, the algorithm keeps track of the most relaxed user u'_r for the four cells of level i together. If such a user leaves these cells, the algorithm has to check upon all existing users and make sure that they still need cells at level i. If this is the case, the algorithm just updates the new information of u'_r. However, if there is no need for any cell at level i, the algorithm merges the four cells together into their parent cell. In the case of a new user entering cells at level i, the algorithm just updates the information of u'_r if necessary.

Peer-to-Peer Spatial Cloaking

This algorithm also supports both the k-anonymity and minimum spatial area requirements (Chow et al. 2006). The main idea is that before requesting any location-based service, the mobile user will form a group from her peers via single-hop and/or multi-hop communication. Then, the



Cloaking Algorithms for Location Privacy, Fig. 3 Grid-based pyramid data structures. (**a**) Complete pyramid. (**b**) Incomplete pyramid

spatial cloaked area is computed as the region that covers the entire group of peers. Figure 4 gives an illustrative example of peer-to-peer spatial cloaking. The mobile user A wants to find her nearest gas station while being five anonymous, i.e., the user is indistinguishable among five users. Thus, the mobile user A has to look around and find four other peers to collaborate as a group. In this example, the four peers are B, C, D, and E. Then, the mobile user A cloaks her exact location into a spatial region that covers the entire group of mobile users A, B, C, D, and E. The mobile user A randomly selects one of the mobile users within the group as an *agent*. In the example given in Fig. 4, the mobile user D is selected as an agent. Then, the mobile user A sends her query (i.e., what is the nearest gas station) along with her cloaked spatial region to the agent. The agent forwards the query to the location-based database server through a base station. Since the location-based database server processes the query based on the cloaked spatial region, it can only give a list of candidate answers that includes the actual answers and some false positives. After the agent receives the candidate answers, it forwards the candidate answers to the mobile user A. Finally, the mobile user A gets the actual answer by filtering out all the false positives.

Key Applications

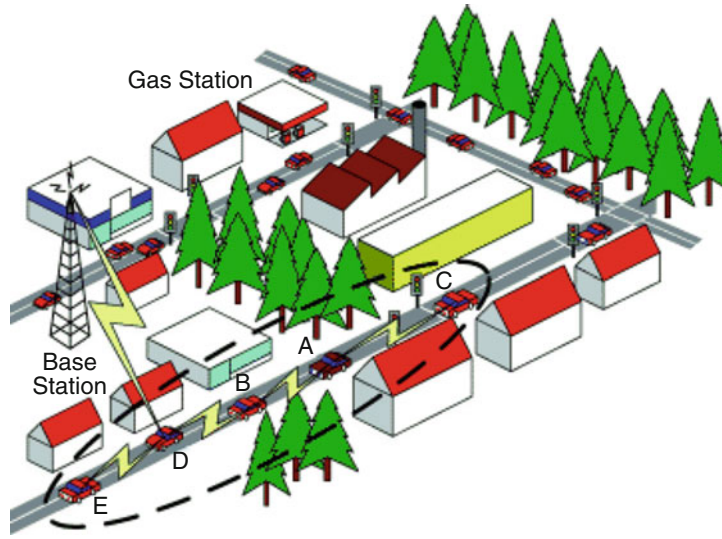
Spatial cloaking techniques are mainly used to preserve location privacy, but they can be used in a variety of applications.

Location-Based Services

Spatial cloaking techniques have been widely adopted to blur user location information before it is submitted to the location-based database server, in order to preserve user location privacy in LBS.

Spatial Database

Spatial cloaking techniques can be used to deal with some specific spatial queries. For example, given an object location, find the minimum area which covers the object and other $k - 1$ objects.



Cloaking Algorithms for Location Privacy, Fig. 4
An example of peer-to-peer spatial cloaking

Data Mining

To perform data mining on spatial data, spatial cloaking techniques can be used to perturb individual location information into lower resolution to preserve their privacy.

Sensor-Based Monitoring System

Wireless sensor networks (WSNs) promise to have a vast significant academic and commercial impact by providing real-time and automatic data collection, monitoring applications, and object positioning. Although sensor-based monitoring or positioning systems clearly offer convenience, the majority of people are not convinced to use such systems because of privacy issues. To overcome this problem, an in-network spatial cloaking algorithm can be used to blur user locations into spatial regions which satisfy user-specified privacy requirements before location information is sent to a sink or base station.

Future Directions

Existing spatial cloaking algorithms have limited applicability as they are:

(a) *Applicable only for snapshot locations and queries.* As location-based environments are characterized by the *continuous* movements of mobile users, spatial cloaking techniques should allow continuous privacy preservation for both user locations and queries. Currently, existing spatial cloaking algorithms only support snapshot location and queries.

(b) *Not distinguishing between location and query privacy*. In many applications, mobile users do not mind that their exact location information is revealed; however, they would like to hide the fact that they issue some location-based queries as these queries may reveal their personal interests. Thus far, none of the existing spatial cloaking algorithms support such a relaxed privacy notion where it is always assumed that users have to hide both their locations and the queries they issue.

Examples of applications that call for such a new relaxed notion of privacy include:

(1) *Business operation*. A courier business company has to know the location of its employees in order to decide which employee is the nearest one to collect a certain package. However, the company is not allowed to keep track of the employees' behavior in terms of their location-based queries. Thus, company employees reveal their location information, but not their query information.

(2) *Monitoring system.* Monitoring systems (e.g., transportation monitoring) rely on

the accuracy of user locations to provide their valuable services. In order to convince users to participate in these systems, certain privacy guarantees should be imposed on their behavior through guaranteeing the privacy of their location-based queries even though their locations will be revealed.

Cross-References

- ▸ Location-Based Services: Practices and Products
- ▸ Privacy and Security Challenges in GIS
- ▸ Privacy Preservation of GPS Traces

References

Bayardo RJ Jr, Agrawal R (2005) Data privacy through optimal k-anonymization. In: Proceedings of the international conference on data engineering (ICDE), Tokyo, pp 217–228

Cheng R, Zhang Y, Bertino E, Prabhakar S (2006) Preserving user location privacy in mobile data management infrastructures. In: Proceedings of privacy enhancing technology workshop, Cambridge, pp 393–412

Chow CY, Mokbel MF, Liu X (2006) A peer-to-peer spatial cloaking algorithm for anonymous location-based services. In: Proceedings of the ACM symposium on advances in geographic information systems (ACM GIS), Arlington, pp 171–178

Gedik B, Liu L (2005) A customizable k-anonymity model for protecting location privacy. In: Proceedings of the international conference on distributed computing systems (ICDCS), Columbus, pp 620–629

Gruteser M, Grunwald D (2003) Anonymous usage of location-based services through spatial and temporal cloaking. In: Proceedings of the international conference on mobile systems, applications, and services (MobiSys), San Francisco, pp 31–42

Gruteser M, Liu X (2004) Protecting privacy in continuous location-tracking applications. IEEE Secur Priv 2(2):28–34

Kalnis P, Ghinita G, Mouratidis K, Papadias D (2006) Preserving anonymity in location based services. Technical report TRB6/06, Department of Computer Science, National University of Singapore

LeFevre K, DeWitt DJ, Ramakrishnan R (2005) Incognito: efficient full-domain k-anonymity. In: Proceedings of the ACM international conference on management of data (SIGMOD), Baltimore, pp 29–60

LeFevre K, DeWitt D, Ramakrishnan R (2006) Mondrian multidimensional k-anonymity. In: Proceedings of the international conference on data engineering (ICDE), Atlanta

Meyerson A, Williams R (2004) On the complexity of optimal K-anonymity. In: Proceedings of the ACM symposium on principles of database systems (PODS), Paris, pp 223–228

Mokbel MF, Chow CY, Aref WG (2006) The new casper: query processing for location services without compromising privacy. In: Proceedings of the international conference on very large data bases (VLDB), Seoul, pp 763–774

Sweeney L (2002a) Achieving k-anonymity privacy protection using generalization and suppression. Int J Uncertain Fuzziness Knowl Sys 10(5):57–88

Sweeney L (2002b) k-anonymity: a model for protecting privacy. Intern J Uncertain Fuzziness Knowl-based Syst 10(5):557–570

Close Range

- ▸ Photogrammetric Applications

Closest Point Query

- ▸ Nearest Neighbor Query

Closest Topological Distance

- ▸ Conceptual Neighborhood

Cloud

- ▸ Medical Image Dataset Processing over Cloud/MapReduce with Heterogeneous Architectures

Cluster Analysis

- ▸ Geodemographic Segmentation

Clustering of Geospatial Big Data in a Distributed Environment

Thomas Triplet and Samuel Foucher
Computer Research Institute of Montreal, Montreal, QC, Canada

Synonyms

Distributed computing; Machine learning; Spatiotemporal clustering; Unsupervised learning

Historical Background

Clustering, sometimes called unsupervised learning/classification or exploratory data analysis, is one of the most fundamental steps in understanding a dataset, aiming to discover the unknown nature of data through the separation of a finite dataset, with little or no ground truth, into a finite and discrete set of "natural," hidden data structures. Given a set of n points in a two-dimensional space, the purpose of clustering is to group them into a number of sets based on similarity measures and distance vectors. Clustering is also useful for compression purpose in large databases (Daschiel and Datcu 2005). The term *Unsupervised Learning* is sometimes used in some fields (i.e., in Machine Learning and Data Mining). Clustering will usually aim at creating homogeneous groups that are maximally separable. It is a fundamental tool in Knowledge Discovery and Data (KDD) mining when looking for meaningful patterns (Alam et al. 2014). Geographical Knowledge Discovery (GKD) is seen as an extension of KDD to the case of spatial data (Miller 2010).

Geospatial data are data coupled with some information about the location where the data were collected or measured. For example, a photography may be associated with the location where it was shot or a temperature reading may be associated with the location of the sensor.

Many geospatial applications now rely on massive amounts of data that may require processing in real time. It is possible to scale a system *vertically* to some extent, that is, add extra storage, more memory, or a faster CPU to a single machine. This approach is however usually costly, remains sensible to failures, and does not scale well with the number of users.

As an alternative, it is also possible to scale a system *horizontally*, that is, combine several commodity servers to solve a problem too complex for a single machine. Besides the sheer amount of data, distributed systems have many benefits compared to a monolithic system, including greater reliability, higher availability, and better performance depending on the use cases. However, distributed systems also present a number of challenges, network latency, and hardware failure, to name a few.

To be effective, this horizontal approach also requires specially designed algorithms that can be efficiently distributed between the different machines. After reviewing distributed computing systems, this article presents clustering algorithms suitable for geospatial big data in a distributed environment. The last section describes several key applications that can benefit from large-scale geospatial clustering.

Scientific Fundamentals

We can distinguish two main types of geospatial data: rasters and vectors. While the same data can usually be represented in both models, there are key differences between the two models, resulting in specific clustering algorithms for the two cases. In this section, we first describe those two geospatial data types and their specificities, and then we detail key architectural differences between major distributed databases and computing systems. Last, we present geospatial clustering algorithms designed to perform efficiently on those distributed computing systems.

Geospatial Data Types

The **raster data model** relies a discrete regular grid of individual and usually square cells, where each cell represents a spatial position and each piece of data is associated with one or more cells.

Raster models are best suited to represent data that vary continuously, for example, aerial and satellite imagery or elevation surfaces. The spatial resolution of raster data depends on the resolution of the grid and is determined at the data acquisition phase.

Data in raster format is basically a matrix of data points on a regular grid. It comes in various forms depending on the source, satellite imagery, Digital Elevation Model, or grid data in meteorology. Data volumes are generally significant as the spatial coverage can be extensive combined sometimes with a high resolution (Miller 2010). In addition, the number of raster dimensions can be fairly high such as hyperspectral imagery.

The **vectorial data model** relies on geometric shapes such as points, lines, or polygons that can be defined by mathematical functions: points are defined by their coordinates, latitude, and longitude typically in the 2D space. Altitude or depth may also be used to define coordinates in the 3D space. Points can be joined together in a specific order to define a line. A closed line, where the last point corresponds to the first point of the line, defines a polygon. The vector model is most useful to represent data with discrete and well-defined boundaries such as country borders, parcels, or streets. Various data structures can be used to store vector data, in particular the *spaghetti* model, which simply describes the objects independently of the others, and more sophisticated *topological* models where each object includes information about the elements it is related to. For example, using the spaghetti model, a polygon is defined by the coordinates of its boundary points; using a topological approach, a polygon can be described as a series of connected lines; each line of the polygon was previously defined in the model as a series of points.

The vector data model was standardized together by the Open Geospatial Consortium (OGC) and the International Organization for Standardization (ISO 19125). This standard called *Simple Features* defines the specifications of vector data (coordinates, points, lines, and polygons) (ISO 2004), as well as a number of spatial operators including an extension of the SQL for traditional relational databases (ISO 2008).

Recently, with the rapid increase of geospatial archives and the availability of large temporal stacks, research efforts have focused on *spatiotemporal clustering* in order to extract meaningful spatiotemporal patterns (Kisilevich et al. 2010b). The direct approach is to simply add the time dimension in the distance metric between points. With the addition of a time dimension on a single spatial entity, the notion of trajectory appears. The time measurements can be regular or irregular. Geo-referenced time series are common in meteorology such as sea surface temperature. Moving spatial objects and trajectories such as sea surface temperature, or to represent moving spatial clouds. Kisilevich et al. (2010a) distinguish three kinds of spatialtemporal data according to the way they are collected: *movement, cellular networks, and environmental*. Movement datasets are typically associated with location-based services or sometimes video surveillance applications (Kuijpers et al. 2008); patterns will be formed by grouping similar trajectories (Andrienko 2008). Environmental data collected either from a network of sensors or satellite imagery are used in many applications (seismology, meteorology, remote sensing, etc.).

Distributed Systems for Geospatial Big Data

Many geospatial applications (see next section) now rely on massive amounts of data that may require processing in real time. It is possible to scale a single machine vertically to some extent by adding extra storage, more memory, or a faster CPU. This approach is however usually costly, is sensible to failures, and does not scale well with the number of users.

CAP Theorem

To overcome those limitations, it is possible to scale a system horizontally, that is, combine several machines to form a cluster. The cluster is leveraged by distributing data and processing algorithms across the different machines. Distributed systems are typically used to process amounts of data that are typically too large for a

single machine. Besides the sheer amount of data, distributed systems have many benefits compared to a monolithic system, including greater reliability, higher availability and better performance depending on the use-cases. However, distributed systems also present a number of challenges, network latency and hardware failure to name a few.

In general, distributed systems should feature the following characteristics:

Consistency (C): all nodes in the cluster see the same data;

- Availability (A): all requests get a success or error notification, even if one or several nodes are unavailable (failure or planned maintenance);
- Partition tolerance (P): the system remains fully functional, even if one or several nodes are unavailable.

However, Brewer's CAP theorem (Brewer 2012; Gilbert and Lynch 2002) stipulates that a distributed system can present at most two of the three traits above. It is therefore possible to design CA, CP, or AP systems, and the choice of an architecture over another largely depends on the intended usage of the system.

Distributed Geospatial Databases

The fundamental principle to allow the processing of spatial big data in a reasonable time is parallelism, which is often not trivial and requires new algorithms that can be distributed across the different machines of the cluster. For example, while many geospatial systems rely on PostgreSQL (http://www.postgresql.org/) and PostGIS (http://postgis.net/) to store geospatial data, the traditional relational model of database is difficult to distribute and parallelizing complex SQL queries is challenging. New paradigms such as Not-Only-SQL (Cattell 2011) (NoSQL) have thus emerged. Most NoSQL systems relax the transactional properties of traditional databases, which guarantee consistency, to favor high availability and partition tolerance (type AP). They usually implement a *Shared-Nothing* architecture (Stonebraker et al. 1986) where all the nodes are independent, which facilitates horizontal scalability.

NoSQL systems can be broadly classified into 5 families:

Key Value (KV): the simplest form of NoSQL system, where data are represented as a list of pairs $<key, value>$, similar to a hash table. In many systems, this list is stored in memory for better performance. This family includes in particular MemcacheDB, (http://memcachedb.org/) Redis (http://redis.io/), and Amazon DynamoDB (DeCandia et al. 2007).

- **Column**: those systems are used to logically organize $<key, value>$ pairs into tables, conceptually similar to tables in the relational model. Notable column-based systems include Google's proprietary Bigtable (Chang et al. 2006), used in particular to index massive amounts of geospatial data from Google Earth (https://www.google.com/earth/), and its open-source derivatives from the Apache foundation: HBase (http://hbase.apache.org/), Cassandra (http://cassandra.apache.org/), and Accumulo (https://accumulo.apache.org/).
- **Document**: this family is the most common among NoSQL systems and is used to store semi-structured data, typically in XML or JSON format. The main systems of this type are MongoDB (http://www.mongodb.org/), Apache CouchDB (http://couchdb.apache.org/), and ElasticSearch (http://www.elasticsearch.org/).
- **Graph**: Graph systems can efficiently represent strongly connected data. The most popular graph database is Neo4J (Webber 2012).
- **Constraint**: Constraint databases (Kanellakis et al. 1995) rely on constraint programming to represent geospatial data and reason about them. While they can represent raster data, their capabilities are best leveraged with vectorial data sets. Unlike other NoSQL families, the constraint programming paradigm inherently difficult to parallelize and distribute efficiently. As a result, current systems do not adopt a Shared-Nothing architecture and do

not scale well. NoSQL constraint databases will thus not be considered in the remaining of this article.

Most NoSQL databases do not natively support geospatial operations or raster data types. Notable exceptions are ElasticSearch, MongoDB, and Amazon DynamoDB, which now feature a limited set of geospatial data types (points, lines, and polygons), indexes (geohashes, quadtrees, or R-trees), and queries (bounding-box, radius, or arbitrary shapes). However, a number of NoSQL databases now support geospatial features using lightweight extensions similar to PostGIS for PostgreSQL, for example, GeoCouch (https://github.com/couchbase/geocouch/) for Apache CouchDB.

Distributed Computing and Clustering

While they have been studied for many years, horizontally distributed computing systems became widely popular in 2004 when Google described the architecture of its distributed file system GFS (Ghemawat et al. 2003) and the MapReduce (Dean and Ghemawat 2004) paradigm they modernized to process and index billions of web pages. Apache Hadoop (http://hadoop.apache.org/) is the most popular open-source framework that implements and reproduces the distributed architecture developed by Google.

The MapReduce paradigm relies on a form of divide-and-conquer technique to simplify some of the challenges of distributed computing: a complex task is first decomposed into simpler tasks that are executed on the machines of the cluster (*map*), and the individual results are then aggregated (*reduce*) to produce the desired output. This approach is very effective for batch processing or when subsets of the data can be processed individually. For example, to process geospatial data, geohashes or quadtrees can be leveraged to effectively divide a large dataset into smaller pieces that can be processed by the machines in the cluster.

A number of geospatial frameworks for Hadoop have been developed recently, in particular SpatialHadoop (Eldawy and Mokbel 2015), GeoCloud2 (http://www.mapcentia.com/en/geocloud/), and HadoopGIS (Chen et al. 2014). PostGIS users can also leverage Hadoop and Cassandra by using BigSQL (http://www.bigsql.org/se/) While those frameworks feature geospatial indexes and operators, they do not provide tools for more advanced analysis using machine learning algorithms. It is however possible to integrate a third-party machine learning library for that purpose and benefit from the MapReduce paradigm to apply distributed classification and clustering algorithms to geospatial big data. One of the most popular libraries is Apache Mahout (http://mahout.apache.org/). It includes a wide variety of algorithms covering supervised learning (random forest, naive Bayes, hidden Markov models, etc.), collaborative filtering (user and item-based collaborative filtering, matrix factorization, etc.), dimensionality reduction (Lanczos, stochastic, principle component analysis), natural language processing (latent Dirichlet allocation, TF-IDF vectors), and clustering (k-means algorithm and its fuzzy and streaming variants, spectral clustering).

More recently, the MapReduce paradigm was generalized to handle streaming data, which are very frequent in GIS applications (see next section), and several alternative frameworks supporting this use case are now available, including Storm (https://storm.apache.org/) and Apache Spark (Zaharia et al. 2010). The key feature of Spark is the introduction of Resilient Distributed Datasets (Zaharia et al. 2012), a distributed memory abstraction that lets programmers perform in-memory computations on large clusters in a fault-tolerant manner. As a result, Spark can improve the performance of MapReduce by two orders of magnitude, and its popularity is now surpassing Hadoop (According to Google Trends, Spark became more popular than Hadoop in September 2014. Source: http://www.google.ca/trends/explore#q=ApacheSpark,Apache_Hadoop). While Spark does not offer any geospatial features, several geospatial frameworks are making use of its distributed computing capabilities to process geospatial big data in real time. Notable systems include

GeoMesa (Fox et al. 2013) to process vectorial data and GeoTrellis (http://geotrellis.io/), more adequate for raster data. Both libraries provide a geospatial extension of the standard Spark RDDs. This key feature allows programmers to leverage other Spark features and apply them to geospatial data. For example, an important component of Spark is MLLib, a library for data analysis similar to Mahout. It implements several key machine learning algorithms, including the k-means clustering algorithm and its streaming variant to build data clusters in real time as new data feed the system.

Distributed Clustering Algorithms

This section details properties of the algorithms that are important for their parallelization in a distributed environment. It is however beyond the scope of this article to extensively review all characteristics and use cases of each clustering algorithm. This article also does not aims at comparing the accuracy of those algorithms: all clustering algorithms rely on some assumption on the distribution of the data (clusters of similar shapes, similar densities, etc.), and the best method to cluster some data depends on the actual distribution of the data.

Clustering techniques can be categorized into two broad categories: density-based and distance-based algorithms.

Distance-based algorithms rely on the distance between data points in the feature space to establish the clusters. Distance-based algorithms assume that clusters to find are of similar shapes and will perform well if this hypothesis is verified by the actual data. Those algorithms are also well suited to cluster raster data types: While we can look at a raster dataset simply as a collection of points, clustering techniques specific to raster will try to take into account or enforce a certain notion of spatial homogeneity between neighborhood points. Spatial homogeneity that states that nearby points are more similar than far-apart points is often captured via the estimation of the local autocorrelation function (Hagenauer and Helbich 2013). High spatial correlation values between spatially close points often limit the number of independent points in a local neighborhood which can clash with the independence assumptions of some clustering methods.

Density-based algorithms on the other hand make use of the density of data points within a region to discover the clusters. Unlike distance-based techniques, density-based algorithms can uncover clusters of various shapes but assume that they are of similar density. The choice of a clustering algorithm therefore depends on the distribution of the data. Density-based algorithms are easier to parallelize and more scalable as they usually rely on local search techniques to identify dense regions in the feature space. One of the most popular density-based algorithm is DBSCAN (Ester et al. 1996), and many distributed variants (He et al. 2014; Noticewala and Vaghela 2014; Kisilevich et al. 2010a; Patwary et al. 2012) have been implemented using MapReduce and show significant running-time improvements, even when handling billions of data points. Another popular example of density-based algorithm is DenClue (Hinneburg and Keim 1998) and its recent improvements (Hinneburg and Gabriel 2007), which is also suitable for segmenting and clustering raster data. More recently, Cludoop (Yu et al. 2015) was implemented using MapReduce, and experiments on geospatial data showed significant improvements in terms of performance and scalability over MR-DBSCAN (He et al. 2014).

In addition, multiple dense regions can be explored simultaneously by discretizing the input feature space into a finite number of grid cells and applying the clustering method within each cell. Existing algorithms include STING (Wang et al. 1997), WaveCluster (Sheikholeslami et al. 1998; Jestes et al. 2011), and Clique (Agrawal et al. 1998). Parallel grid-based clustering further divides cells into sub-cells, processes each sub-cell, and combines the individual results to build the final clusters (Xiaoyun et al. 2009; Zhang et al. 2010). More recently, PatchWork (Gouineau et al. 2016) was implemented using Apache Spark to distribute local density computations and sowed significant performance

improvements over MapReduce implementations of DBScan. This approach is particularly useful to mine geospatial data as the cell grid can be defined using Hilbert or Z-order space filling curves (Dai and Su 2003; Hong-bo et al. 2009) which are implemented in the distributed GeoMesa framework (see the section above).

It is also possible to further categorize clustering techniques depending on the output of the algorithm (Hruschka et al. 2009): hierarchical or partitioning.

Partitioning algorithms may define mutually exclusive hard clusters or soft clusters that allow a certain degree of overlap measured by a membership function. Fuzzy clustering techniques are typical of this kind of soft partitioning approach (Ehrlich et al. 1984). The most popular partitioning algorithm is k-means clustering (MacQueen 1967; Lloyd et al. 1982): given k clusters to find, the technique determines the centers of the clusters and updates the membership of each cluster iteratively using the distance to the center of the cluster. The approach can be easily parallelized and distributed given that computing the distance to the cluster centers has no dependencies. Several distributed implementations are available, in particular for MapReduce and Spark as part of the Mahout and MLLib libraries, respectively. For the same reason, distributed implementations of the streaming variant of k-means are available to cluster spatiotemporal data in real time. Distributed implementations are also available for related algorithms such as CLARANS (Ng et al. 2005).

Hierarchical algorithms represent input data as a nested set of partitions, that is, as a tree also called a dendrogram. Hierarchical techniques implementing a divisive top-down strategy, where a larger cluster is split into several subclusters, have been proposed. However, hierarchical agglomerative clustering (HAC) is the most popular strategy. It is a bottom-up strategy that iteratively groups together the two most similar clusters to form a new cluster. The computation of the similarity measurement between two clusters depends on the linkage method and is one of the main factors that differentiate HAC methods. In single-linkage clustering, the link between two clusters is made by a single element pair of the two elements (one in each cluster) that are closest to each other. In complete-linkage clustering, the link between two clusters considers all element pairs, and the distance between clusters equals the distance between those two elements (one in each cluster) that are farthest away from each other. Other linkage methods such as UPGMA have been proposed and often used in bioinformatics for phylogenetic studies. However, HAC algorithms rely on a global distance matrix and are notoriously difficult to parallelize. Furthermore, most of those algorithms have a computational complexity in $O(N^2 \log N)$ or $O(N^2)$ with N the number of data points and do not scale well. Clustering of geospatial big data using naive HAC algorithms will therefore quickly become problematic, and more sophisticated methods have been proposed, including DISC (Jin et al. 2013), MR-VPSOM (Gao et al. 2010), etc. Those HAC methods can be efficiently distributed and were implemented using the MapReduce framework for batch computations and NoSQL databases for the storage of large-distance matrices.

Key Applications

This section presents various key use cases of clustering algorithms that facilitate the analysis of spatial and spatiotemporal big data.

Internet of Things

The Internet of Things (IoT) is one of the key applications of spatial big data and machine learning. It is a recent domain that emerged from the proliferation of devices that are connected to the Internet or to other devices. Examples of such connected devices include smartphones, drones, wearable electronics (e.g., clothes and watches), light bulbs, home appliances, etc. The IoT has a tremendous potential in numerous industries such as domotics, transportation, retail, health, or resource consumption.

Those devices are usually equipped with a variety of sensors that can collect data in real time or several times per minute. They also often include a geolocation tracker or can be paired with a device that has a geolocation tracker. Those connected devices thus generate very large amounts of data with a strong spatiotemporal component. Those devices rely on the integration of many spatiotemporal data sources, including the geolocation of the user (from their connected smartphone) and meteorological data.

One of the key benefits of the IoT is to enable machine-to-machine communication thereby facilitating the automation of various tasks. Automation relies on the spatiotemporal data collected by the devices, as well as rules to define triggers and actions. Web services to facilitate the implementation of those rules have emerged and are gaining popularity as new devices become connected. Some devices are now relying on machine learning algorithms to learn how they are being used. Connected home thermostats, for instance, are now capable of learning from the habits of the home owners to automatically define rules and triggers to adjust the temperature.

Smart Cities

Another key application of large-scale geospatial information systems is the modeling of public infrastructures for the development of smarter cities. Smart cities such as Barcelona, Stockholm, or Montreal heavily rely on digital technologies to reduce resource consumption and to engage more effectively with their citizens.

An example of smart city projects, which also relies on the Internet of Things, is the public bicycle sharing systems such as Bixi (http://www.publicbikesystem.com/) that are implanted in many large cities. In such a system, bikes are equipped with GPS trackers, allowing the operator to monitor the usage and adjust the service accordingly (Wood et al. 2011). Studies of the usage of public bicycle sharing systems using spatial clustering algorithms (Austwick et al. 2013) were also conducted to reveal structures of social communities in major cities.

As another example, distributed computing systems could also be leveraged to optimize the engineering and planning of new infrastructures. For example, the city of Riyadh, Saudi Arabia, modeled the entire transportation network, including constraints for transit time between major activity centers. The goal was to analyze the network to highlight infrastructure deficiencies where usage exceeds capacity and predict future travel demand and potential congestion areas under different scenarios and network topologies. Similar studies were conducted in Jaipur, India (Gahlot et al. 2012), and Vancouver, Canada (Foth 2010), for the design of their public transit network.

Remote Sensing

Remote sensing is the science of obtaining information about objects or areas from a distance, typically from aircraft, boats, or satellites, without making physical contact with the object and thus in contrast to on-site observation. It refers to the use of aerial sensor technologies to detect and classify objects on Earth (both on the surface and in the atmosphere and oceans) by means of propagated signals, including electromagnetic (RADAR, LiDAR, etc.), acoustic (SONAR, seismograms, etc.), and geodetic (gravitational field measurement). The remote sensor can collect the signal passively emitted from a surface of interest (e.g., a photometer measuring sunlight) or actively transmit a signal and collect its reflection (e.g., RADARs in airplanes).

Remote sensing has an immense range of applications: agriculture (e.g., crop monitoring), geology (terrain analysis, topography, etc.), hydrology (flood monitoring), environment (sea ice coverage, biomass mapping, forestry, land usage, etc.), and oceanography (oil spill detection, tsunami detection, phytoplankton concentration, etc.), to name a few.

For example, many oceanographic characteristics (such as currents) vary over both time and space. At a fixed location, an important spatial coordinate is the vertical axis through the water column – or profile – from the surface to the ocean bottom. An individual profile can be viewed as a vertical-line plot. A time series of profiles is best viewed by stacking sequential profiles next to each other to form an image.

Remote sensors thus usually produce massive amounts data of type raster, which may also be combined with data from other on-site sensors. Distributed clustering algorithms (Lv et al. 2010) and color coding then reveal both the vertical and temporal structure of the measured quantity, which depends on the sensor.

Ocean Networks Canada has developed several types of sensors, including an active zooplankton acoustic profiler (ZAP). This sensor emits an acoustic pulse through water; when it encounters fishes, suspended particulate, or zooplankton floating in the water, a part of the sound is reflected back. By gating the reflected signals in time, the vertical distribution of scatterers is recorded and provides useful information about marine life.

Medical Area and Disaster Monitoring

As transportation means have allowed to travel faster around the globe, more effective infection monitoring tools are needed to help in the control of disease outbreaks as illustrated by the recent H1N1 flu and Ebola pandemics. The ability to quickly analyze the evolution of the disease, and to discover patterns in the data, is critical to understand the root cause of the pandemics and take appropriate measures to control an emerging disease situation and prevent their further spread.

Epidemiological data consist of spatiotemporal data describing the evolution of the disease in both space and time. Key challenges of epidemiological data are the ability of analyzing new trends and patterns in pseudo-real time as the disease spreads, as well as the recursive nature of those patterns, that is, patterns from previous pandemics are likely to give important clues for the prediction of the evolution of the current pandemics. Note that those challenges are not unique to pandemics, and other disasters that have strong geospatial and temporal dimensions, such as tornadoes, water flooding, or oil spills, share the same characteristics.

Crisis detection and management can also be facilitated by integrating traditional geospatial data sources with alternative sources, in particular from social media. The underlying idea is that social sensing, which is the set of information obtained from a group of people, can provide information similar to those obtainable from a sensor network. Twitter (https://twitter.com/), one of the most popular social networks, allows people to share short messages in real time, many of them are now associated with a geolocation. Using geospatial data mining and natural language processing techniques, it is thus possible to leverage Twitter as an effective data source for social sensing. This approach has been successfully applied to the identification of outbreaking seismic events (Avvenuti et al. 2014): the system is able to detect earthquake within seconds of the event and to notify people far earlier than official channels.

Future Directions

While data clustering has been extensively studied and proved to be tremendously useful in data mining and knowledge discovery, clustering of geospatial big data presents several challenges to be addressed in the future.

First, an increasing number of geospatial applications are now generating very large volumes of data. Those applications include remote sensing, drones, and the Internet of Things. The number of sensors that collect geocoded data is increasing exponentially, from 500 million devices in 2003 to an anticipated 50 billion sensors in the next 5 years, resulting in volumes of data far exceeding the computing power of a single machine. However, many clustering algorithms which were developed in the past two decades rely on an iterative approach, which is inherently difficult to parallelize and distribute efficiently. While a few parallelized variants of popular algorithms, such as k-means, have been proposed and implemented using the MapReduce paradigm, the variety of geospatial clustering algorithms that can be efficiently distributed is limited. As a result, scaling a system horizontally to accommodate larger amounts of data or to reduce the running time of the algorithms remains challenging.

In addition to the rapidly increasing number of sensors, a large portion of those sensors

can collect data several times per second or per minute: geospatial datasets thus often present a strong temporal dimension. A growing number of applications require real-time or near real-time processing of those spatiotemporal data, for example, traffic optimization or crime prevention in smart cities. For those applications, a batch-oriented approach to distributed computing such as the popular MapReduce paradigm is not suitable because of latency issues. Alternative distributed computing frameworks such as Apache Storm or Spark significantly can handle continuous streams of data and reduce the latency of the system. However, very few clustering algorithms suitable for geospatial applications have been implemented for those frameworks so far.

Last, most popular distributed computing frameworks, Hadoop and Spark in particular, were developed only recently and are still under active development. The sets of features of those systems are often not stable or mature yet. In addition, with the notable exception of Accumulo, most distributed systems have not yet emphasized development on data access control and privacy concerns, which can be critical for geospatial applications.

Cross-References

- Big Data and Spatial Constraint Databases
- Distributed Geospatial Computing (DGC)
- Irregular Shaped Spatial Clusters: Detection and Inference
- k-NN Search in Time-dependent Road Networks
- Movement Patterns in Spatio-Temporal Data
- Outlier Detection
- Outlier Detection, Spatial
- Patterns, Complex

References

Agrawal R, Gehrke J, Gunopulos D, Raghavan P (1998) Automatic subspace clustering of high dimensional data for data mining applications. In: Proceedings of the 1998 ACM SIGMOD international conference on management of data (SIGMOD'98), New York. ACM, pp 94–105

Alam S, Dobbie G, Koh YS, Riddle P, Rehman SU (2014) Research on particle swarm optimization based clustering: a systematic review of literature and techniques. Swarm Evol Comput 17(0):1–13

Andrienko G (2008) Spatio-temporal aggregation for visual analysis of movements. In: Proceedings of IEEE symposium on visual analytics science and technology (VAST 2008), Columbus, pp 51–58

Austwick MZ, O'Brien O, Strano E, Viana M (2013) The structure of spatial networks and communities in bicycle sharing systems. PLoS ONE 8(9):e74685, 09

Avvenuti M, Cresci S, Marchetti A, Meletti C, Tesconi M (2014) Ears (earthquake alert and report system): a real time decision support system for earthquake crisis management. In: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining (KDD'14), New York. ACM, pp 1749–1758

Brewer E (2012) Cap twelve years later: how the "rules" have changed. Computer 45(2):23–29

Cattell R (2011) Scalable SQL and NoSQL data stores. SIGMOD Rec 39(4):12–27

Chang F, Dean J, Ghemawat S, Hsieh WC, Wallach DA, Burrows M, Chandra T, Fikes A, Gruber RE (2006) Bigtable: a distributed storage system for structured data. In: Proceedings of the 7th symposium on operating systems design and implementation (OSDI'06), Berkeley. USENIX Association, pp 205–218

Chen X, Vo H, Aji A, Wang F (2014) High performance integrated spatial big data analytics. In: Proceedings of the 3rd ACM SIGSPATIAL international workshop on analytics for big geospatial data (BigSpatial'14), New York. ACM, pp 11–14

Dai H-K, Su H-C (2003) Approximation and analytical studies of inter-clustering performances of space-filling curves. In: Banderier C, Krattenthaler C (eds) Discrete random walks (DRW'03), Paris, Sept 1–5 2003. Discrete mathematics and theoretical computer science proceedings, vol AC. DMTCS, pp 53–68

Daschiel H, Datcu M (2005) Information mining in remote sensing image archives: system evaluation. IEEE Trans Geosci Remote Sens 43(1):188–199

Dean J, Ghemawat S Mapreduce: simplified data processing on large clusters. In: Proceedings of the 6th conference on symposium on opearting systems design & implementation (OSDI'04), vol 6, Berkeley. USENIX Association, pp 10–10

DeCandia G, Hastorun D, Jampani M, Kakulapati G, Lakshman A, Pilchin A, Sivasubramanian S, Vosshall P, Vogels W (2007) Dynamo: Amazon's highly available key-value store. In: Proceedings of twenty-first ACM SIGOPS symposium on operating systems principles (SOSP'07), New York. ACM, pp 205–220

Ehrlich R, Bezdek JC, Fullh W (1984) Fcm: the fuzzy c-means clustering algorithm. Comput Geosci 10(2–3):191–203

Eldawy A, Mokbel MF (2015) Spatialhadoop: a mapreduce framework for spatial data. In: Proceedings of the

31st IEEE international conference on data engineering (ICDE), Seoul
Ester M, Kriegel H-P, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: Simoudis E, Han J, Fayyad UM (eds) Second international conference on knowledge discovery and data mining. AAAI Press, Palo Alto, pp 226–231
Foth N (2010) Long-term change around skytrain stations in Vancouver, Canada: a demographic shift-share analysis. Geograph Bull 51:37–52
Fox A, Eichelberger C, Hughes J, Lyon S (2013) Spatio-temporal indexing in non-relational distributed databases. In: 2013 IEEE international conference on big data, Santa Clara, pp 291–299
Gahlot V, Swami BL, Parida M, Kalla P (2012) User oriented planning of bus rapid transit corridor in GIS environment. Int J Sustain Built Environ 1:102–109
Gao H, Jiang J, She L, Fu Y (2010) A new agglomerative hierarchical clustering algorithm implementation based on the map reduce framework. J Digit Content Technol Appl 4(3):95–100
Ghemawat S, Gobioff H, Leung S-T (2003) The google file system. In: Proceedings of the 19th ACM symposium on operating systems principles (SOSP '03), New York. ACM, pp 29–43
Gilbert S, Lynch N (2002) Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services. SIGACT News 33(2):51–59
Gouineau F, Landry T, Triplet T (2016) PatchWork: a scalable density-grid clustering algorithm. In: Proceedings of the 31st ACM symposium on applied computing, data mining track, Pisa
Hagenauer J, Helbich M (2013) Contextual neural gas for spatial clustering and analysis. Int J Geograph Inf Sci 27:251–266
He Y, Tan H, Luo W, Feng S, Fan J (2014) MR-DBSCAN: a scalable mapreduce-based DBSCAN algorithm for heavily skewed data. Front Comput Sci 8(1): 83–99
Hinneburg A, Gabriel H-H (2007) Denclue 2.0: fast clustering based on kernel density estimation. In: Proceedings of the 7th international conference on intelligent data analysis (IDA'07). Springer, Berlin/Heidelberg, pp 70–80
Hinneburg A, Keim DA (1998) An efficient approach to clustering in large multimedia databases with noise. In: Agrawal R, Stolorz PE, Piatetsky-Shapiro G (eds) Proceedings of the fourth international conference on knowledge discovery and data mining (KDD-98), New York, 27–31 Aug 1998. AAAI Press, pp 58–65
Hong-bo X, Zhong-xiao H, Qi-Long H (2009) A clustering algorithm based on grid partition of space-filling curve. In: 2009 fourth international conference on internet computing for science and engineering (ICICSE), Harbin, pp 260–265
Hruschka ER, Campello RJGB, Freitas AA, de Carvalho ACPLF (2009) A survey of evolutionary algorithms for clustering. IEEE Trans Syst Man Cybern Part C Appl Rev 39(2):133–155
ISO (2004) Geographic information—simple feature access—Part 1: common architecture. ISO 19125–1:2004, International Organization for Standardization, Geneva
ISO (2008) Geographic information—simple feature access—Part 2: SQL option. ISO 19125–2:2004, International Organization for Standardization, Geneva
Jestes J, Yi K, Li F (2011) Building wavelet histograms on large data in mapreduce. Proc VLDB Endow 5(2):109–120
Jin C, Patwary MMA, Agrawal A, Hendrix W, Liao W-k, Choudhary A (2013) Disc: a distributed single-linkage hierarchical clustering algorithm using mapreduce. In: Proceedings of the 4th international SC workshop on data intensive computing in the clouds, Denver. (http://datasys.cs.iit.edu/events/DataCloud2013/)
Jin C, Liu R, Chen Z, Hendrix W, Agrawal A, Choudhary A (2015) A scalable hierarchical clustering algorithm using spark. In: IEEE first international conference on big data computing service and applications, Redwood City, pp 418–426
Kanellakis PC, Kuper GM, Revesz P (1995) Constraint query languages. J Comput Syst Sci 51(1):26–52
Kisilevich S, Mansmann F, Keim D (2010a) P-dbscan: a density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos. In: Proceedings of the 1st international conference and exhibition on computing for geospatial research & application (COM.Geo '10), Washington, DC. ACM, Springer, pp 1–4. (http://www.springer.com/us/book/9780387098227)
Kisilevich S, Mansmann F, Nanni M, Rinzivillo S (2010b) Spatio-temporal clustering. In: Maimon O, Rokach L (eds) Data mining and knowledge discovery handbook. Springer, pp 855–874. http://www.springer.com/us/book/9780387098227
Kuijpers B, Alvares LO, Palma AT, Bogorny V (2008) A clustering-based approach for discovering interesting places in trajectories. In: Proceedings of the 2008 ACM symposium on applied computing, Fortaleza, pp 863–868
Lloyd S (1982) Least squares quantization in PCM. IEEE Trans Inf Theory 28(2):129–137
Lv Z, Hu Y, Zhong H, Wu J, Li B, Zhao H (2010) Parallel k-means clustering of remote sensing images based on MapReduce. In: Proceedings of the 2010 international conference on web information systems and mining (WISM'10). Springer, Berlin/Heidelberg, pp 162–170
MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the 5th Berkeley symposium on mathematical statistics and probability, Berkeley/Los Angeles
Miller HJ (2010) The data avalanche is here. Shouldn't we be digging? J Reg Sci 50:181–201
Ng RT, Han J, Ieee Computer Society (2005) Clarans: a method for clustering objects for spatial data mining. IEEE Trans Knowl Data Eng 1003–1017
Noticewala M, Vaghela D (2014) Article: Mr-idbscan: efficient parallel incremental dbscan algorithm using mapreduce. Int J Comput Appl 93(4):13–18

C

Patwary MA, Palsetia D, Agrawal A, Liao W-k, Manne F, Choudhary A (2012) A new scalable parallel dbscan algorithm using the disjoint-set data structure. In: Proceedings of the international conference on high performance computing, networking, storage and analysis (SC'12), Los Alamitos. IEEE Computer Society Press, pp 62:1–62:11

Sheikholeslami G, Chatterjee S, Zhang A (1998) Wavecluster: a multi-resolution clustering approach for very large spatial databases. Proc Int Confer Very Large Data Bases 24:428–439

Stonebraker M (1986) The case for shared nothing. IEEE Database Eng Bull 9(1):4–9

Wang W, Yang J, Muntz RR (1997) Sting: a statistical information grid approach to spatial data mining. In: Proceedings of the 23rd international conference on very large data bases (VLDB'97), San Francisco. Morgan Kaufmann Publishers Inc, pp 186–195

Webber J (2012) A programmatic introduction to neo4j. In: Proceedings of the 3rd annual conference on systems, programming, and applications: software for humanity (SPLASH'12), New York. ACM, pp 217–218

Wood J, O'Brien O, Slingsby A, Dykes J (2011) Visualizing the dynamics of London's bicycle-hire scheme. Cartogr Int J Geograph Inf Geovis 46(4):239–251

Xiaoyun C, Yi C, Xiaoli Q, Min Y, Yanshan H (2009) PGMCLU: a novel parallel grid-based clustering algorithm for multi-density datasets. In: 1st IEEE symposium on web society, 2009 (SWS'09), Lanzhou, pp 166–171

Yu Y, Zhao J, Wang X, Wang Q, Zhang Y (2015) Cludoop: an efficient distributed density-based clustering for big data using Hadoop. Int J Distrib Sensor Netw 2015(2):1–13

Zaharia M, Chowdhury M, Franklin MJ, Shenker S, Stoica I (2010) Spark: cluster computing with working sets. In: Proceedings of the 2Nd USENIX conference on hot topics in cloud computing (HotCloud'10), Berkeley. USENIX Association, pp 10–10

Zaharia M, Chowdhury M, Das T, Dave A, Ma J, McCauley M, Franklin MJ, Shenker S, Stoica I (2012) Resilient distributed datasets: a fault-tolerant abstraction for in-memory cluster computing. In: Proceedings of the 9th USENIX conference on networked systems design and implementation (NSDI'12), Berkeley. USENIX Association, pp 2–2

Zhang H, Zhou Y, Li J, Wang X, Yan B (2010) Analyze the wild birds' migration tracks by MPI-based parallel clustering algorithm. In: Proceedings of the 6th international conference on advanced data mining and applications: Part I (ADMA'10). Springer, Berlin/Heidelberg, pp 383–393

Cognition

▸ Hierarchies and Level of Detail

Cognitive Engineering

▸ Geospatial Semantic Web: Personalization

Cognitive Mapping

▸ Wayfinding, Landmarks

Cognitive Psychology

▸ Wayfinding: Affordances and Agent Simulation

Collaborative Geographic Information Systems

▸ Geocollaboration

Collaborative Tracking

▸ Feature Detection and Tracking in Support of GIS

Collocation Pattern

▸ Co-location Pattern

Collocation, Spatiotemporal

▸ Movement Patterns in Spatio-Temporal Data

Co-location

▸ Patterns, Complex

Co-location Mining

▸ Co-location Pattern Discovery

Co-location Pattern

Nikos Mamoulis
Department of Computer Science, University of Hong Kong, Hong Kong, China

Synonyms

Collocation pattern; Spatial association pattern

Definition

A (spatial) *co-location pattern* P can be modeled by an undirected connected graph where each node corresponds to a nonspatial feature and each edge corresponds to a neighborhood relationship between the corresponding features. For example, consider a pattern with three nodes labeled "timetabling," "weather," and "ticketing" and two edges connecting "timetabling" with "weather" and "timetabling" with "ticketing." An *instance* of a pattern P is a set of objects that satisfy the unary (feature) and binary (neighborhood) constraints specified by the pattern's graph. An instance of an example pattern is a set $\{o_1, o_2, o_3\}$ of three spatial locations where $label(o_1)$ = "timetabling," $label(o_2)$ = "weather," $label(o_3)$ = "ticketing" (unary constraints), and $dist(o_1, o_2) \leq \varepsilon$, $dist(o_1, o_3) \leq \varepsilon$ (spatial binary constraints). In general, there may be an arbitrary spatial (or spatiotemporal) constraint specified at each edge of a pattern graph (e.g., topological, distance, direction, and time-difference constraints).

Main Text

Co-location patterns are used to derive co-location rules that associate the existence of nonspatial features in the same spatial neighborhood. An example of such a rule is "if a water reservoir is contaminated, then people who live in nearby houses have high probability of having a stomach disease." The interestingness of a co-location pattern is quantized by two measures: the *prevalence* and the *confidence*. Co-location patterns can be mined from large spatial databases with the use of algorithms that combine (multi-way) spatial join algorithms with spatial association rule mining techniques.

Cross-References

▸ Patterns, Complex
▸ Retrieval Algorithms, Spatial

Co-location Pattern Discovery

Wei Hu
International Business Machines Corp., Rochester, MN, USA

Synonyms

Co-location mining; Co-location rule discovery; Co-location rule finding; Co-location rule mining; Co-occurrence; Spatial association; Spatial association analysis

Definition

Spatial co-location rule discovery or *spatial co-location pattern discovery* is the process that identifies spatial co-location patterns from large spatial datasets with a large number of Boolean spatial features.

Historical Background

The co-location pattern and rule discovery are part of the spatial data mining process. The

differences between spatial data mining and classical data mining are mainly related to data input, statistical foundation, output patterns, and computational process. The research accomplishments in this field are primarily focused on the output pattern category, specifically the predictive models, spatial outliers, spatial co-location rules, and clusters (Shekhar et al. 2003).

The spatial pattern recognition research presented here, which is focused on co-location, is also most commonly referred to as the spatial co-location pattern discovery and co-location rule discovery. To understand the concepts of spatial co-location pattern discovery and rule discovery, we will have to first examine a few basic concepts in spatial data mining.

The first word to be defined is Boolean spatial features. *Boolean spatial features* are geographic object types. They either are absent or present regarding different locations within the domain of a two-dimensional or higher (three)-dimensional metric space such as the surface of the earth (Shekhar et al. 2003). Some examples of Boolean spatial features are categorizations such as plant species, animal species, and types of roads, cancers, crimes, and businesses.

The next concept relates to co-location patterns and rules. *Spatial co-location patterns* represent the subsets of Boolean spatial features whose instances are often located in close geographic proximity (Shekhar et al. 2003). It resembles frequent patterns in many aspects. Good examples are symbiotic species. The Nile crocodile and Egyptian plover in ecology prediction (Fig. 1) are one good illustration of a point spatial co-location pattern representation. Frontage roads and highways (Fig. 2) in specified metropolitan road maps could be used to demonstrate line-string co-location patterns.

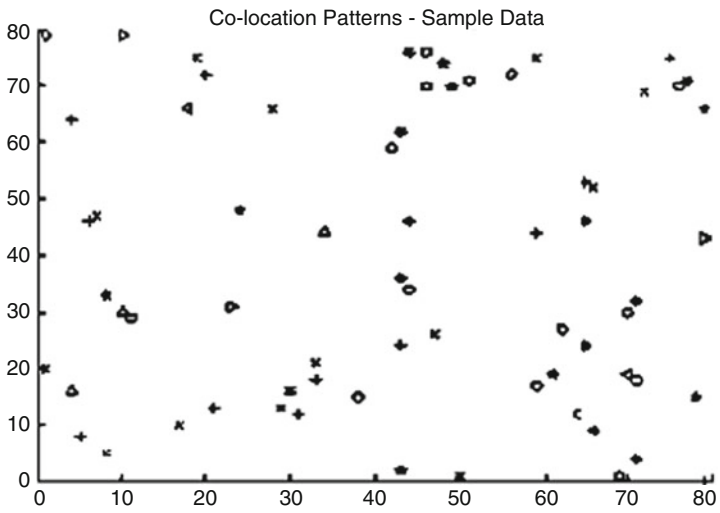
Examples of various categories of spatial co-location patterns are given in Table 1. We can see that the domains of co-location patterns are distributed in many interesting fields of science research and daily services, which proves their great usefulness and importance.

Spatial co-location rules are models to associate the presence of known Boolean spatial features referencing the existence of instances of other Boolean spatial features in the neighborhood. Figure 1 also provides good examples of spatial co-location rules. As can be seen, rule "Nile crocodiles → Egyptian plover" can predict the presence of Egyptian plover birds in the same areas where Nile crocodiles live. A dataset consisting of several different Boolean spatial feature instances is marked on the space. Each type of Boolean spatial features is distinguished by a distinct representation shape. A careful examination reveals two co-location patterns: ('+', 'x') and ('o', '*') (Shekhar et al. 2003). Spatial co-location rules can be further classified into popular rules and confident rules, according to the frequency of cases showing in the dataset. The major concern here is the difference of dealing with rare events and popular events. Usually, rare events are ignored, and only the popular co-location rules are mined. So if there is a need to identify the confident co-location rules, then special handling and a different approach must be taken to reach them (Huang et al. 2003).

Spatial co-location rule discovery is the process that identifies spatial co-location patterns from large spatial datasets with a large number of Boolean spatial features (Shekhar et al. 2003). The problems of spatial co-location rule discovery are similar to the *spatial association rule mining* problem, which identifies the interrelationships or associations among a number of spatial datasets. The difference between the two has to do with the concept of transactions.

An example of association rule discovery can be seen with market basket datasets, in which transactions represent sets of merchandise item categories purchased altogether by customers (Shekhar et al. 2003). The association rules are derived from all the associations in the data with support values that exceed a user-defined threshold. In this example, we can define in detail the process of mining association rules as to identify frequent item sets in order to plan store layouts or marketing campaigns as a part of related business intelligence analysis.

On the other hand, in a spatial co-location rule discovery problem, we usually see that the transactions are not explicit (Shekhar et al. 2003). There are no dependencies among the



Co-location Pattern Discovery, Fig. 1 Illustration of point spatial co-location patterns. *Shapes* represent different spatial feature types. Spatial features in sets {'+', 'x'} and {'o', '*'} tend to be located together (Shekhar et al. 2003)

Co-location Pattern Discovery, Fig. 2 Illustration of line-string co-location patterns. Highways, e.g., Hwy100, and frontage roads, e.g., Normandale Road, are co-located (Shekhar et al. 2003)

transactions analyzed in market basket data, because the transaction data do not share instances of merchandise item categories but rather instances of Boolean spatial features instead. These Boolean spatial features are distributed into a continuous space domain and thus share varied spatial types of relationships, such as overlap, neighbor, etc., with each other.

Although spatial co-location patterns and co-location rules differ slightly, according to the

Co-location Pattern Discovery, Table 1 Examples of co-location patterns (Xiong et al. 2004)

Domains	Example features	Example co-location patterns
Ecology	Species	Nile crocodile, Egyptian plover
Earth science	Climate and disturbance events	Wildfire, hot, dry, lightning
Economics	Industry types	Suppliers, producers, consultants
Epidemiology	Disease types and environmental events	West Nile disease, stagnant water sources, dead birds, mosquitoes
Location-based service	Service type requests	Tow, police, ambulance
Weather	Fronts, precipitation	Cold front, warm front, snow fall
Transportation	Delivery service tracks	US Postal Service, UPS, newspaper delivery

previous definitions, it can be said that *spatial co-location pattern discovery* is merely another phrasing for spatial co-location rule finding. Basically, the two processes are the same and can be used in place of each other. Both are used to find the frequent co-occurrences among Boolean spatial features from given datasets.

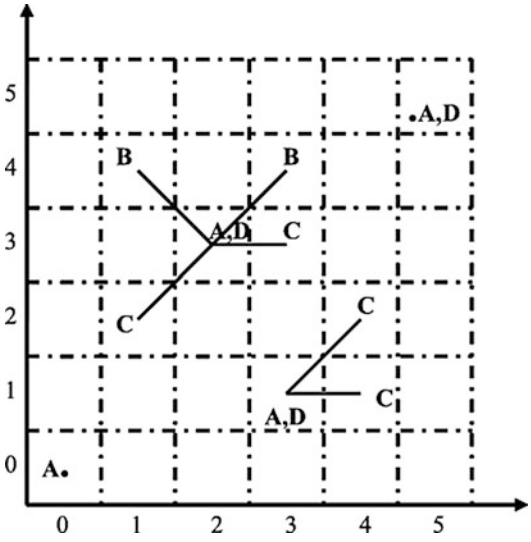
Co-location Pattern Discovery, Table 2 Boolean feature A and the defined transactions related to B and C

Instance of A	Transaction
(0,0)	Ø
(2,3)	{B,C}
(3,1)	{C}
(5,5)	Ø

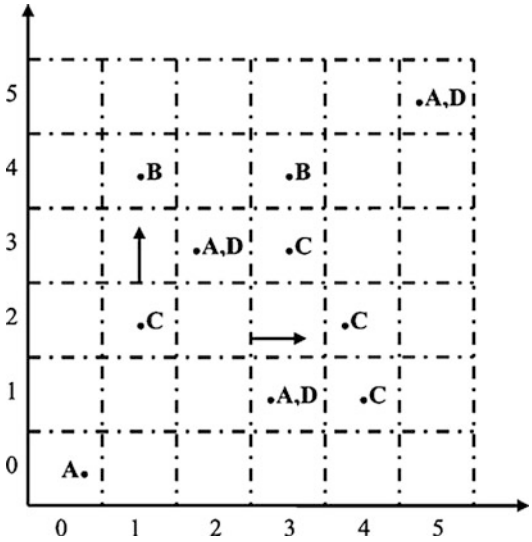
Scientific Fundamentals

According to one categorization, there are three methods of finding co-location patterns in spatial datasets, depending on the focus of the search. These three categories are the *reference feature-centric model*, the *window-centric model*, and the *event-centric model* (Shekhar and Huang 2001). The *reference feature-centric model* is relevant to application domains that focus on a specific Boolean spatial feature such as cancer. The goal of the scientists is to find the co-location patterns between this Boolean spatial feature and other task-related features such as asbestos or other substances. This model uses the concept of neighborhood relationship to materialize the transactions from datasets. Measurements of support and confidence can be used to show the degree of interestingness (Shekhar and Huang 2001). For example, if there are two features, A and B, and if A is the relevant feature, then B is said to be close to A if B is a neighbor of A. But how can we tell that B is a neighbor of A? Here we can use either the Euclidean distance or the Manhattan distance, depending on the type of application domain we are investigating. Then with the corresponding definition of the distance between the features, we could declare them to be neighbors. Thus by considering A, all the other Boolean spatial features surrounding A are used as transactions. Once the data is materialized as above, the support and confidence are computed and used to measure the degree of interestingness (Shekhar and Huang 2001). Table 2 shows an instance of data, and Fig. 3 illustrates the layout and process we described with detailed data from Table 2.

The second method of finding co-location patterns is a *window-centric model* or *data partitioning model*. The process defines proper sized windows and then enumerates all possible windows as transactions. Each window is actually a partition of the whole space, and the focus is on the local co-location patterns, which are bounded by the window boundaries. Patterns across multiple windows are of no concern. Each window



Co-location Pattern Discovery, Fig. 3 Transactions are defined around instances of feature A, relevant to B and C (Shekhar and Huang 2001)



Co-location Pattern Discovery, Fig. 4 Example of window-centric model (Shekhar and Huang 2001)

is a transaction, and the process tries to find which features appear together the most number of times in these transactions, alias, and windows, i.e., using support and confidence measurements (Shekhar and Huang 2001).

Figure 4 shows the processing with window partitions on data similar to that shown in Fig. 3. As this is a local model, even though here the A and C could have been a pattern, these features are completely ignored since they are not within a single window.

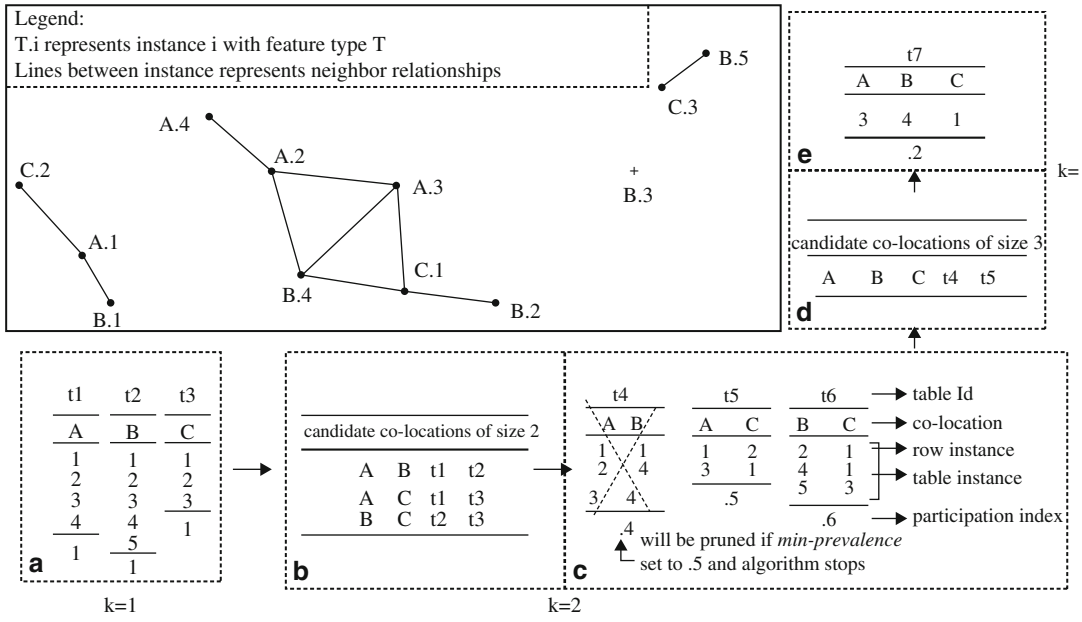
The third modeling method is the *event-centric model*. This model is mostly related to ecology- specific domains where scientists want to investigate specific events such as drought, El Nino, etc. The goal of this model is to find the subsets of spatial features likely to occur in the neighborhood of a given event type. One of the assumptions of this algorithm is that the neighbors are reflexive, that is, interchangeable. For example, if A is a neighbor of B, then B is also a neighbor of A.

The event centric defines key concepts as follows: "A *neighborhood* of l is a set of locations L = {l1, l2, l3,. . . ,lk} such that li is a neighbor of l" (Shekhar and Huang 2001). "I = {I1,. . . ,Ik} is a *row instance* of a co-location C ={f1,. . . ,fk} if Ij is an instance of feature fj" (Shekhar and Huang 2001). The *participation ratio* and *participation index* are two measures which replace support and confidence here. The participation ratio is the number of row instances of co-location C divided by number of instances of Fi. Figure 5 shows an example of this model.

Table 3 shows a summary of the interest measures for the three different models.

With different models to investigate different problems of various application domains, there are also multiple algorithms used in the discovery process. Approaches to discover co-location rules can be categorized into two classes, *spatial statistics* and *data mining approaches*.

Spatial statistics-based approaches use measures of spatial correlation to characterize the relationship between different types of spatial features. Measures of spatial correlation include the cross-K function with Monte Carlo simulation, mean nearest-neighbor distance, and spatial regression models. Computing spatial correlation measures for all possible co-location patterns can be computationally expensive due to the exponential number of candidate subsets extracted from a large collection of spatial Boolean features that we are interested in Huang et al. (2004).



Co-location Pattern Discovery, Fig. 5 Event-centric model example (Huang et al. 2004)

Co-location Pattern Discovery, Table 3 Interest measures for different models (Shekhar et al. 2003)

Model	Items	Transactions defined by	Interest measures for $C_1 \rightarrow C_2$	
			Prevalence	Conditional probability
Reference feature centric	Predicates on reference and relevant features	Instances of reference feature C_1 and C_2 involved with	Fraction of instance of reference feature with $C_1 \cup C_2$	Pr(C_2 is true for an instance of reference features given C_1 is true for that instance of reference feature)
Data partitioning	Boolean feature types	A partitioning of spatial dataset	Fraction of partitions with $C_1 \cup C_2$	Pr(C_2 in a partition given C_1 in that partition)
Event centric	Boolean feature types	Neighborhoods of instances of feature types	Participation index of $C_1 \cup C_2$	Pr(C_2 in a neighborhood of C_1)

Data mining approaches can be further divided into two categories: the *clustering-based map overlay approach* and the *association rule-based approaches*.

Clustering-based map overlay approach regards every spatial attribute as a map layer and considers spatial clusters (regions) of point data in each layer as candidates for mining the associations among them. *Association rule-based approaches* again can be further divided into two categories: the *transaction-based approaches* and the *distance-based approaches*.

Transaction-based approaches aim to define transactions over space such that an a priori-like algorithm can be used just as in the association rule discovery process. Transactions over space can be defined by a reference-centric model as discussed previously, which enables the derivation of association rules using the a priori algorithm. There are few major shortcomings of this approach: generalization of this paradigm is nontrivial in the case where no reference feature is specified, and duplicate counts for many candidate associations may result when defining transactions around locations of instances of all features.

Distance-based approaches are relatively novel. A couple of different approaches have

been presented by different research groups. One proposes the participation index as the prevalence measure, which possesses a desirable anti-monotone property (Huang et al. 2003). Thus, a unique subset of co-location patterns can be specified with a threshold on the participation index without consideration of detailed algorithm applied such as the order of examination of instances of a co-location. Another advantage of using the participation index is that it can define the correctness and completeness of co-location mining algorithms.

Key Applications

The problem of mining spatial co-location patterns can be applied to many useful science research or public interest domains.

As shown in Table 1, one of the top application domains is location-based services. With advances such as GPS and mobile communication devices, many location-based services have been introduced to fulfill users' increasing desires for convenience. Many of the services requested by service subscribers from their mobile devices see benefit from the support of spatial co-location pattern mining. The location-based service provider needs to know which requests are submitted frequently together and which are located in spatial proximity (Xiong et al. 2004).

Ecology is another good field to apply this technology because ecologists are very interested in finding frequent co-occurrences among spatial features, such as drought, EI Nino, substantial increase/drop in vegetation, and extremely high precipitation (Xiong et al. 2004).

A third important domain whose future cannot be imagined without spatial data mining is weather services. The identification of correct and valuable co-location patterns or rules from huge amounts of collected historical data can be expected to lead to better predictions about incoming weather, deeper insights into environmental impacts on weather patterns, and suggestions of possible effective steps to prevent the future deterioration of the environment.

A final example in our list of applications is traffic control or transportation management. With the knowledge of co-location rules discovered from existing datasets, better supervising and management could be carried out to make transportation systems run in the most efficient way, as well as to gain clearer foresights of future road network development and expansion.

There are many more interesting fields related to the spatial co-location application domain, such as disease research, economics, earth science, etc. (Shekhar et al. 2002). With the availability of more spatial data from different areas, we can expect more research and studies to benefit from this technology.

Future Directions

Spatial co-location pattern discovery and co-location rule mining are very important, even essential tasks of a *spatial data mining systems* (SDMS), which extract previously unknown but interesting spatial patterns and relationships from large spatial datasets. These methods have the potential to serve multiple application domains and have a wide impact on many scientific research fields and services. Current approaches to mine useful *co-location patterns* are still evolving with new studies carried out in the field. We can expect extended development of such techniques to improve the algorithms and efficiency in future studies.

A first potential direction of research is to find more efficient algorithms against extended spatial data types other than points, such as line segments and polygons (Huang et al. 2004).

Second direction is that as only *Boolean spatial features* are mined here, the future studies can extend the *co-location mining* framework to handle categorical and continuous features that also exist in the real world (Huang et al. 2004).

Third potential extension can be on the notion of co-location pattern to be de-colocation pattern or co-incidence pattern (Xiong et al. 2004).

References

Huang Y, Xiong H, Shekhar S, Pei J (2003) Mining confident colocation rules without a support threshold. In: Proceedings of the 18th ACM symposium on applied computing (ACM SAC), Melbourne

Huang Y, Shekhar S, Xiong H (2004) Discovering colocation patterns from spatial datasets: a general approach. IEEE Trans Knowl Data Eng (TKDE) 16(12) December

Shekhar S, Huang Y (2001) Discovering spatial colocation patterns: a summary of results. In: Proceedings of 7th international symposium on spatial and temporal databases (SSTD), Redondo Beach

Shekhar S, Schrater P, Raju W, Wu W (2002) Spatial contextual classification and prediction models for mining geospatial data. IEEE Trans Multimed

Shekhar S, Zhang P, Huang Y, Vatsavai RR (2003) Trends in spatial data mining. In: Kargupta H, Joshi A, Sivakumar K, Yesha Y (eds) Data mining: next generation challenges and future directions. AAAI/MIT Press, Cambridge, MA

Xiong H, Shekhar S, Huang Y, Kumar V, Ma X, Yoo J (2004) A framework for discovering co-location patterns in data sets with extended spatial objects. In: Proceedings of SIAM international conference on data mining (SDM)

Co-location Patterns

▸ Co-location Patterns, Interestingness Measures

Co-location Patterns, Algorithms

Nikos Mamoulis
Department of Computer Science, University of Hong Kong, Hong Kong, China

Synonyms

Association; Co-occurrence; Mining collocation patterns; Mining spatial association patterns; Participation index; Participation ratio; Reference-feature centric

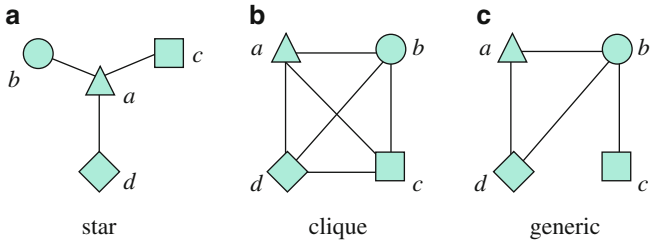
Definition

A spatial co-location pattern associates the coexistence of a set of non-spatial features in a spatial neighborhood. For example, a co-location pattern can associate contaminated water reservoirs with a certain disease within 5 km distance from them. For a concrete definition of the problem, consider number n of spatial datasets $R_1, R_2, \ldots, R_n$, such that each R_i contains objects that have a common non-spatial feature f_i. For instance, R_1 may store locations of water sources, R_2 may store locations of appearing disease symptoms, etc. Given a distance threshold ε, two objects on the map (independent of their feature labels) are *neighbors* if their distance is at most ε. We can define a *co-location pattern* P by an undirected connected graph where each node corresponds to a feature and each edge corresponds to a neighborhood relationship between the corresponding features. Figure 1 shows examples of a *star* pattern, a *clique* pattern and a generic one. A variable labeled with feature f_i is only allowed to take instances of that feature as values. Variable pairs that should satisfy a spatial relationship (i.e., constraint) in a valid pattern instance are linked by an edge. In the representations of Fig. 1, we assume that there is a single constraint type (e.g., close to), however in the general case, any spatial relationship could label each edge. Moreover, in the general case, a feature can label more than two variables. Patterns with more than one variable of the same label can be used to describe *spatial autocorrelations* on a map.

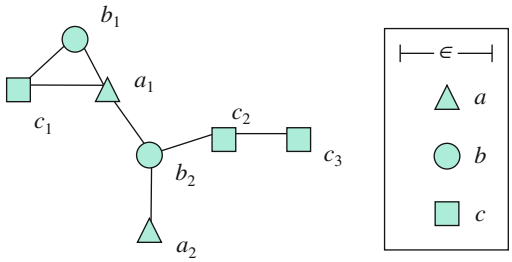
Interestingness measures (Huang et al. 2003; Shekhar and Huang 2001) for co-location patterns express the statistical significance of their instances. They can assist the derivation of useful rules that associate the instances of the features.

Historical Background

The problem of mining association rules based on spatial relationships (e.g., adjacency, proximity, etc.) of events or objects was first discussed in Koperski and Han (1995). The spatial data are



Co-location Patterns, Algorithms, Fig. 1 Three pattern representations. (**a**) Star. (**b**) Clique. (**c**) Generic



Co-location Patterns, Algorithms, Fig. 2 Mining example

converted to transactional data according to a *reference feature* model. Later, the research interest shifted toward mining *co-location patterns*, which are feature centric sets with instances that are located in the same neighborhood (Huang et al. 2003; Morimoto 2001; Munro et al. 2003; Shekhar and Huang 2001; Zhang et al. 2004). Huang et al. (2003), Morimoto (2001), Munro et al. (2003), and Shekhar and Huang (2001) focused on patterns where the closeness relationships between features form a complete graph (i.e., every pair of features should be close to each other in a pattern), whereas Zhang et al. (2004) extended this model to feature-sets with closeness relationships between arbitrary pairs and proposed an efficient algorithm for mining such patterns (which is herein reviewed). Yang (2005) extended the concept of co-locations for objects with extend and shape, whereas Wang et al. (2005) studied the mining of co-location patterns that involve spatio-temporal topological constraints.

Scientific Fundamentals

Consider a number n of spatial datasets R_1, R_2, ..., R_n, such that each R_i contains all objects that have a particular non-spatial feature f_i. Given a feature f_i, we can define a transactional database as follows. For each object o_i in R_i a spatial query is issued to derive a set of features $I = \{f_j : f_j \neq f_i \wedge \exists\, o_j \in R_j(dist(o_i, o_j) \leq \varepsilon)\}$. The collection of all feature sets I for each object in R_i defines a transactional table T_i. T_i is then mined using some itemsets mining method (e.g., Agrawal and Skrikant 1994; Zaki and Gouda 2003). The *frequent* feature sets I in this table, according to a minimum support value, and can be used to define rules of the form:

$$(label(o) = f_i) \Rightarrow (o \text{ close to some } o_j \in R_j, \quad \forall f_j \in I).$$

The support of a feature set I defines the confidence of the corresponding rule. For example, consider the three object-sets shown in Fig. 2. The lines indicate object pairs within a distance ε from each other. The shapes indicate different features. Assume that one must extract rules having feature a on their left-hand side. In other words, find features that occur frequently close to feature a. For each instance of a, generate an itemset; a_1 generates $\{b, c\}$ because there is at least one instance of b (e.g., b_1 and b_2) and one instance of c (e.g., c_1) close to a_1. Similarly, a_2 generates itemset $\{b\}$ (due to b_2). Let 75% be the minimum confidence. One first discovers frequent itemsets (with minimum support 75%) in $T_a = \langle\{b, c\}, \{b\}\rangle$, which gives us a sole itemset $\{b\}$. In turn, one can generate the rule

$$(label(o) = a) \Rightarrow (o \text{ close to } o_j \text{ with } label(o_j) = b),$$

with confidence 100%. For simplicity, in the rest of the discussion, $f_i \Rightarrow I$ will be used to denote rules that associate instances of feature f_i with instances of feature sets I, $f_i \notin I$, within its proximity. For example, the rule above can be expressed by $a \Rightarrow \{b\}$. The mining process for feature a can be repeated for the other features (e.g., b and c) to discover rules having them on their left side (e.g., one can discover rule $b \Rightarrow \{a, c\}$ with conf. 100%). Note that the features on the right hand side of the rules are not required to be close to each other. For example, rule $b \Rightarrow \{a, c\}$ does not imply that for each b the nearby instances of a and c are close to each other. In Fig. 2, observe that although b_2 is close to instances a_1 and a_2 of a and instance c_2 of c, c_2 is neither close to a_1 nor to a_2.

A co-location *clique* pattern P of length k is described by a set of features $\{f_1, f_2, \ldots, f_k\}$. A valid instance of P is a set of objects $\{o_1, o_2, \ldots, o_k\} : (\forall 1 \leq i \leq k, o_i \in R_i) \wedge (\forall 1 \leq i < j \leq k, dist(o_i, o_j) \leq \varepsilon)$. In other words, all pairs of objects in a valid pattern instance should be close to each other, or else the closeness relationships between the objects should form a *clique graph*. Consider again Fig. 2 and the pattern $P = \{a, b, c\}$. $\{a_1, b_1, c_1\}$ is an instance of P, but $\{a_1, b_2, c_2\}$ is not.

Huang et al. (2003) and Shekhar and Huang (2001) define some useful measures that characterize the interestingness of co-location patterns. The first is the *participation ratio* $pr(f_i, P)$ of a feature f_i in pattern P, which is defined by the following equation:

$$pr(f_i, P) = \frac{\#\text{ instances of } f_i \text{ in any instance of } P}{\#\text{ instances of } f_i}. \quad (1)$$

Using this measure, one can define *co-location rules* that associate features with the existences of other features in their neighborhood. In other words, one can define rules of the form $(label(o) = f_i) \Rightarrow (o$ participates in an instance of P with confidence $pr(f_i, P))$. These rules are similar to the ones defined in Koperski and Han (1995); the difference here is that there should be neighborhood relationships between all pairs of features on the right hand side of the rule. For example, $pr(b, \{a, b, c\}) = 0.5$ implies that 50% of the instances of b (i.e., only b_1) participate in some instance of pattern a, b, c (i.e., $\{a_1, b_1, c_1\}$).

The *prevalence* $prev(P)$ of a pattern P is defined by the following equation:

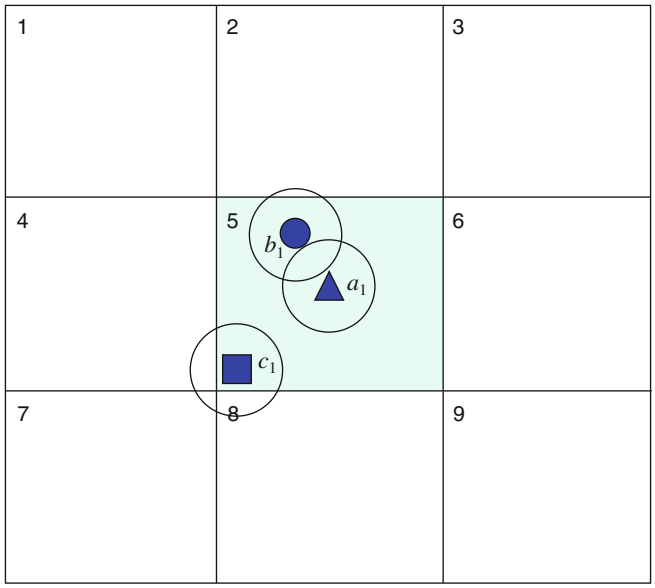
$$prev(P) = \min\{pr(f_i, P), f_i \in P\}. \quad (2)$$

For example, $prev(\{b, c\}) = 2/3$ since $pr(b, \{b, c\}) = 1$ and $pr(c, \{b, c\}) = 2/3$. The prevalence captures the minimum probability that whenever an instance of some $f_i \in P$ appears on the map, it will then participate in an instance of P. Thus, it can be used to characterize the strength of the pattern in implying co-locations of features. In addition, prevalence is monotonic; if $P \subseteq P'$, then $prev(P) \geq prev(P')$. For example, since $prev(\{b, c\}) = 2/3$, we know that $prev(\{a, b, c\}) \leq 2/3$. This implies that the a priori property holds for the prevalence of patterns and algorithms like generalized (Agrawal and Skrikant 1994) can be used to mine them in a level-wise manner (Shekhar and Huang 2001).

Finally, the *confidence* $conf(P)$ of a pattern P is defined by the following equation:

$$conf(P) = \max\{pr(f_i, P), f_i \in P\}. \quad (3)$$

For example, $conf(b, c) = 1$ since $pr(b, \{b, c\}) = 1$ and $pr(c, \{b, c\}) = 2/3$. The confidence captures the ability of the pattern to derive co-location rules using the participation ratio. If P is confident with respect to a minimum confidence threshold, then it can derive at least one co-location rule (for the attribute f_i with $pr(f_i, P) = conf(P)$). In Fig. 2, $conf(\{b, c\})=1$ implies that we can find one feature in $\{b, c\}$ (i.e., b), every instance of which participates in an instance of $\{b, c\}$. Given a collection of spatial



Co-location Patterns, Algorithms, Fig. 3 A regular grid and some objects

objects characterized by different features, a minimum prevalence threshold *min_prev*, and a minimum confidence threshold *min_conf*, a data analyst could be interested in discovering prevalent and/or confident patterns and the co-location rules derived by them. The confidence of a co-location rule between two patterns, $P_1 \rightarrow P_2$, $P_1 \cap P_2 = \varnothing$, can be defined by the conditional probability that an instance of P_1 participates in some instance of $P_1 \cup P_2$ (given that $P_1 \cup P_2$ is prevalent with respect to *min_prev*) (Shekhar and Huang 2001).

It is now discussed how co-location patterns are mined from a spatial database. Star-like patterns are the first are of focus (as seen in Fig. 1a). As an example, consider the rule: "given a pub, there is a restaurant and a snack bar within 100 meters from it with confidence 60%". Assume that the input is n datasets $R_1, R_2, \ldots, R_n$, such that for each i, R_i stores instances of feature f_i.

The mining algorithm, a high-level description of which is shown in Fig. 3, operates in two phases; the hashing phase and the mining phase. During the hashing phase, each dataset R_i is read and the instances of the corresponding feature are spatially partitioned with the help of a regular grid. Each object is extended by the distance threshold ε to form a disk and hashed into the partitions intersected by this disk. Figure 4 shows an example. The space is partitioned into 3×3 cells. Object a_1 (which belongs to dataset R_a, corresponding to feature a) is hashed to exactly one partition (corresponding to the central cell C_5). Object b_1 is hashed to two partitions (C_2 and C_5). Finally, object c_1 is hashed into four partitions (C_4, C_5, C_7, and C_8).

The mining phase employs a main memory algorithm to efficiently find the association rules in each cell. This method is in fact a multi-way main memory spatial join algorithm based on the plane sweep technique (Brinkhoff et al. 1993; Mamoulis and Papadias 2001; Preparata and Shamos 1985). The **synch_sweep** procedure extends the plane sweep technique used for pair-wise joins to (i) apply for multiple inputs and (ii) for each instance of one input, find if there is at least one instance from other inputs close to it.

synch_sweep takes a feature f_i as input and a set of partitions of all feature instances hashed into the same cell C, and finds the maximal patterns each feature instance is included directly (without computing their sub-patterns first). The objects in the partition R_i^C (corresponding to feature f_i) in cell C are scanned in sorted order of their x-value. For each object o_i, we initialize the *maximal* star pattern L where o_i can participate as L's center. Then for each other feature, we sweep a vertical line along the x-axis to find

Co-location Patterns, Algorithms, Fig. 4 An algorithm for reference feature co-locations

```
/* R_i stores the coordinates of all objects with feature f_i */
Algorithm find_centric_co-locations(R_1, R_2, ..., R_n)
1.   /* 1. Spatial-hashing phase */
2.   super-impose a regular grid G over the map;
3.   for each feature f_i
4.      hash the objects from R_i to a number of buckets:
5.         each bucket corresponds to a grid cell;
6.         each object o is hashed to the cell(s) intersected by
7.            the disk centered at o with radius ε;
8.   /* 2. Mining phase */
9.   for each cell C of the grid;
10.     for each feature f_i
11.        load bucket R_i^C in memory;
12.        /* R_i^C containing objects in R_i hashed into C */
13.        sort points of R_i^C according to their x co-ordinate;
14.     for each feature f_i
15.        synch_sweep(f_i, R_1^C, R_2^C, ..., R_n^C);
```

if there is any instance (i.e., object) within ε distance from o_i; if there is, we add the corresponding feature to L. Finally, L will contain the maximal pattern that includes f_i; for each subset of it we increase the support of the corresponding co-location rule. For more details about this process, the reader can refer to Zhang et al. (2004).

Overall, the mining algorithm requires two database scans; one for hashing and one for reading the partitions, performing the spatial joins and counting the pattern supports, provided that the powerset of all features but f_i can fit in memory. This is a realistic assumption for typical applications (with 10 or less feature types). Furthermore, it can be easily extended for arbitrary pattern graphs like those of Fig. 1b and c.

Key Applications

Sciences

Scientific data analysis can benefit from mining spatial co-location patterns (Salmenkivi 2004; Yang 2005). Co-location patterns in census data may indicate features that appear frequently in spatial neighborhoods. For example, residents of high income status may live close to areas of low pollution. As another example from geographical data analysis, a co-location pattern can associate contaminated water reservoirs with a certain decease in their spatial neighborhood. Astronomers may use spatial analysis to identify features that commonly appear in the same constellation (e.g., low brightness, similar colors). Biologists may identify interesting feature combinations appearing frequently in close components of protein or chemical structures.

Decision Support

Co-location pattern analysis can also be used for decision support in marketing applications. For example, consider an E-commerce company that provides different types of services such as weather, timetabling and ticketing queries (Morimoto 2001). The requests for those services may be sent from different locations by (mobile or fix line) users. The company may be interested in discovering types of services that are requested by geographically neighboring users in order to provide location-sensitive recommendations to them for alternative products. For example, having known that ticketing requests are frequently asked close to timetabling requests, the company may choose to advertise the ticketing service to all customers that ask for a timetabling service.

Future Directions

Co-location patterns can be extended to include the temporal dimension. Consider for instance, a database of moving objects, such that each object

is characterized by a feature class (e.g., private cars, taxis, buses, police cars, etc.). The movements of the objects (trajectories) are stored in the database as sequences of timestamped spatial locations. The objective of spatio-temporal co-location mining is to derive patterns composed by combinations of features like the ones seen in Fig. 1. In this case, each edge in the graph of a pattern corresponds to features that are close to each other (i.e., within distance ε) for a large percentage (i.e., large enough support) of their locations during their movement. An exemplary pattern is "ambulances are found close to police cars with a high probability". Such extended spatial co-location patterns including the temporal aspect can be discovered by a direct application of the existing algorithms. Each temporal snapshot of the moving objects database can be viewed as a segment of a huge map (that includes all frames) such that no two segments are closer to each other than ε. Then, the spatio-temporal co-location patterns mining problem is converted to the spatial co-locations mining problem we have seen thus far.

A more interesting (and more challenging) type of spatio-temporal collocation requires that the closeness relationship has a duration of at least τ time units, where τ is another mining parameter. For example, we may consider, as a co-location instance, a combination of feature instances (i.e., moving objects), which move closely to each other for τ *continuous* time units. To count the support of such *durable* spatio-temporal patterns, we need to slide a window of length τ along the time dimension and for each position of the window, find combinations of moving objects that qualify the pattern. Formally, given a durable pattern P, specified by a feature-relationship graph (like the ones of Fig. 1) which has a node f_i and distance/duration constraints ε and τ, the participation ratio of feature f_i in P is defined by:

$$pr(f_i, P) = \frac{\#\ \tau\text{-length windows with an instance of } P}{\#\ \tau\text{-length windows with a moving object of type } f_i}. \quad (4)$$

Thus, the participation ratio of a feature f_i in P is the ratio of window positions that define a sub-trajectory of at least one object of type f_i which also defines an instance of the pattern. Prevalence and confidence in this context are defined by (2) and (3), as for spatial co-location patterns. The efficient detection of such patterns from historical data as well as their on-line identification from streaming spatio-temporal data are interesting problems for future research.

Cross-References

- ▸ Co-location Pattern
- ▸ Patterns, Complex
- ▸ Retrieval Algorithms, Spatial

References

Agrawal R, Skrikant R (1994) Fast algorithms for mining association rules. In: Proceedings of the 20th international conference on very large data bases, pp 487–499

Brinkhoff T, Kriegel HP, Seeger B (1993) Efficient processing of spatial joins using r-trees. In: Proceedings of the ACM SIGMOD international conference

Huang Y, Xiong H, Shekhar S, Pei J (2003) Mining confident co-location rules without a support threshold. In: Proceedings of the 18th ACM symposium on applied computing (ACM SAC) (2003)

Koperski K, Han J (1995) Discovery of spatial association rules in geographic information databases. In: Proceedings of the 4th international symposium on advances in spatial databases (SSD), vol. 951, pp. 47–66

Mamoulis N, Papadias D (2001) Multiway spatial joins. ACM Trans Database Syst 26(4):424–475

Morimoto Y (2001) Mining frequent neighboring class sets in spatial databases. In: Proceedings of the ACM SIGKDD international conference knowledge discovery and data mining

Munro R, Chawla S, Sun P (2003) Complex spatial relationships. In: Proceedings of the 3rd IEEE international conference on data mining (ICDM)

Preparata FP, Shamos MI (1985) Computational geometry: an introduction. Springer, New York

Salmenkivi M (2004) Evaluating attraction in spatial point patterns with an application in the field of cultural history. In: Proceedings of the 4th IEEE international conference on data mining

Shekhar S, Huang Y (2001) Discovering spatial co-location patterns: a summary of results. In: Proceedings of the 7th international symposium on advances in spatial and temporal databases (SSTD)

Wang J, Hsu W, Lee ML (2005) A framework for mining topological patterns in spatio-temporal databases. In: Proceedings of the 14th ACM international conference on Information and knowledge management. Full paper in IEEE Trans. KDE 16(12), 2004

Yang H, Parthasarathy S, Mehta S (2005) Mining spatial object associations for scientific data. In: Proceedings of the 19th International Joint Conference on Artificial Intelligence

Zaki MJ, Gouda K (2003) Fast vertical mining using diffsets. In: Proceedings of the ACM SIGKDD Conference

Zhang X, Mamoulis N, Cheung, DWL, Shou Y (2004) Fast mining of spatial collocations. In: Proceedings of the ACM SIGKDD Conference

Co-location Patterns, Interestingness Measures

Marko Salmenkivi
HIIT Basic Research Unit, Department of Computer Science, University of Helsinki, Helsinki, Finland

Synonyms

Association Measures; Co-location Patterns; Interestingness Measures; Selection Criteria; Significance Measures

Definition

Interestingness measures for spatial *co-location patterns* are needed to select from the set of all possible patterns those that are in some (quantitatively measurable) way, characteristic for the data under investigation, and, thus, possibly, provide useful information.

Ultimately, interestingness is a subjective matter, and it depends on the user's interests, the application area, and the final goal of the spatial data analysis. However, there are properties that can be objectively defined, such that they can often be assumed as desirable. Typically, these properties are based on the frequencies of pattern instances in the data.

Spatial association rules, co-location patterns and co-location rules were introduced to address the problem of finding associations in spatial data, and in a more general level, they are applications of the problem of finding *frequent patterns* on spatial domain. Interestingness of a pattern in data is often related to its frequency, and that is the reason for the name of the problem.

In practice, a pattern is considered as interesting, if the values of the interestingness measures (possibly only one) of the pattern exceed the thresholds given by the user.

Historical Background

Finding patterns in data and evaluating their interestingness has traditionally been an essential task in statistics. Statistical data analysis methods cannot always be applied to large data masses. For more detailed discussion of the problems, see Scientific Fundamentals. Data mining, or knowledge discovery from databases, is a branch of computer science that arose in the late 1980s, when classical statistical methods could no longer meet the requirements of analysis of the enormously increasing amount of digital data. Data mining develops methods for finding trends, regularities, or patterns in very large datasets. One of the first significant contributions of data mining research was the notion of association rule, and algorithms, e.g., Apriori (Agrawal and Ramakrishnan 1994), for finding all interesting association rules from transaction databases. Those algorithms were based on first solving the subproblem of the *frequent itemset discovery*. The interesting association rules could easily be deduced from the frequent itemsets.

When applying association rules in spatial domain, the key problem is that there is no natural notion of transactions, due to the continuous two-dimensional space. Spatial association rules were first introduced in Koperski and Han (1995). They were analogous to association rules with the exception that at least one of the predicates

in a spatial association rule expresses spatial relationship (e.g., *adjacent_to, within, close_to*). The rules always continue a reference feature. Support and confidence were used as interestingness measures similarly to the transaction-based association rule mining. Another transaction-based approach was proposed in Morimoto (2001): spatial objects were grouped into disjoint partitions. One of the drawbacks of the method is that different partitions may result in different sets of transactions, and, thus, different values for the interestingness measures of the patterns. As a solution to the problem, co-location patterns in the context of the event-centric model were introduced in Shekhar and Huang (2001).

Scientific Fundamentals

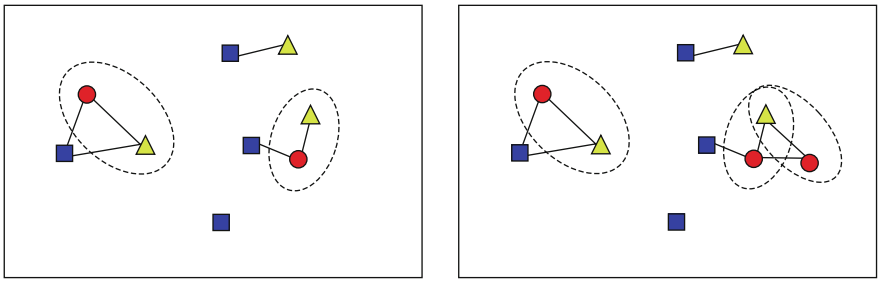
Different models can be employed to model the spatial dimension, and the interpretation of co-location patterns as well as the interestingness measures are related to the selected model. The set of proposed models include at least the window-centric model, reference feature-centric model, event-centric model, and buffer-based model (Xiong et al. 2004).

Co-location patterns and co-location rules can be considered in the general framework of *frequent pattern* mining as pattern classes. Other examples of pattern classes are itemsets and association rules (in relational databases), episodes (in event sequences), strings, trees and graphs (Mannila et al. 1995; Zaki 2002).

In the window-centric model the space is discretized by a uniform grid, and the set of all the possible windows of size $k \times k$ form the set of transactions. The items of the transaction are the features present in the corresponding window. Thus, support can be used as the interestingness measure. The interpretation of the confidence of the rule $A \rightarrow B$ is the conditional probability of observing an instance of B in an arbitrary $k \times k$-window, given that an instance of feature A occurs in the window.

The reference feature-centric model focuses on a specific *Boolean spatial feature*, and all the discovered patterns express relationships of the reference feature and other features. The spatial association rules introduced in Koperski and Han (1995) are based on selecting a reference feature, and then creating transactions over space. Transactions make it possible to employ the interestingness measures introduced for transaction databases in the context of *frequent itemset discovery*: *support* of a feature set (analogously to the support of an itemset in transaction databases), and *confidence* (or *conditional probability*) of an association rule.

In the event-centric model introduced in Shekhar and Huang (2001), the spatial proximity of objects is modeled by using the notion of *neighborhood*. The neighborhood relation $R(x, y)$, $x, y \in O$, where O is the set of spatial objects, is assumed to be given as input. The objects and the neighborhood relation can be represented as an undirected graph, where nodes correspond to objects, and an edge between nodes indicates that the objects are neighbors (see Fig. 1). A limitation of the event-centric model is that it can be used only when the objects are points. An advantage is that the



Co-location Patterns, Interestingness Measures, Fig. 1 Examples of (row) instances of co-location patterns in the event-centric model

pattern discovery is not restricted to patterns with a reference feature. Furthermore, no explicit transactions need to be formed. This fact also has consequences as to the choice of relevant interestingness measures. In a transaction-based model a single object can only take part in one transaction, whereas in the event-centric model it is often the case that a single object participates in several instances of a particular pattern.

Figure 1 shows an example. There are nine spatial point objects. The set of features consists of three features indicated by a triangle (denote it by A), circle (B), and rectangle (C). In this example only one feature is assigned to each object, in general there may be several of them. There are three instances of feature A, two instances of B, and four instances of C. The solid lines connect the objects that are neighbors. Cliques of the graph indicate the instances of co-location patterns. Hence, there is only one instance of pattern $\{ABC\}$ containing all the features.

The participation ratio of a feature f in a co-location pattern $\mathcal{P}$ is the number of instances of the feature that participate in an instance of $\mathcal{P}$ divided by the number of all instances of f. For instance, in the example data on the left panel of Fig. 1 the participation ratio of feature A in pattern $\{AB\}$, $pr(A, \{AB\}) = 2/3$, since two out of three instances of feature A also participate in instances of $\{AB\}$. Correspondingly $pr(B, \{AB\}) = 2/2 = 1$, since there is no instance of B that is not participating in $\{AB\}$. The objects on the right panel of Fig. 1 are equal to those of the left panel, except for an additional point with feature B. Now, there are two different points with feature B such that they both are neighbors of the same instance of A. The instances of pattern $\{A, B\}$ have been indicated by the dashed lines. Thus, one instance of A participates in two instances of $\{A, B\}$. The participation ratios are equal to the left-side case: $pr(A, \{AB\}) = 2/3$ and $pr(B, \{AB\}) = 3/3 = 1$.

Prevalence of a co-location pattern is defined as $prev(P) = \min\{pr(f, \mathcal{P}), f \in \mathcal{P}\}$. A co-location pattern is *prevalent*, if its prevalence exceeds the user-specified threshold value. Prevalence is a monotonous interestingness measure with respect to the pattern class of co-location patterns, since adding features to P can clearly only decrease $prev(P)$.

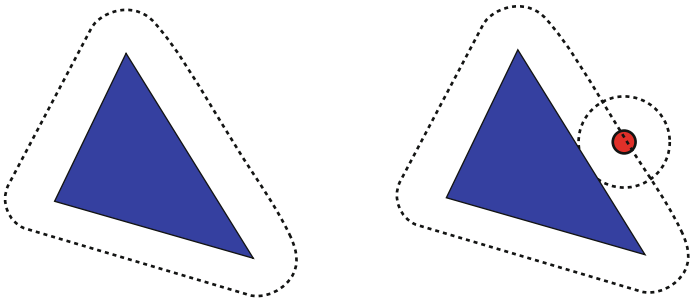
Let $\mathcal{P}$ and $\mathcal{Q}$ be co-location patterns, and $\mathcal{P} \cap \mathcal{Q} = \emptyset$. Then $\mathcal{P} \rightarrow \mathcal{Q}$ is a *co-location rule*. The *confidence* (or *conditional probability*) of $\mathcal{P} \rightarrow \mathcal{Q}$ (in a given dataset) is the fraction of instances of $\mathcal{P}$ such that they are also instances of $\mathcal{P} \cup \mathcal{Q}$. A co-location rule is *confident* if the confidence of the rule exceeds the user-specified threshold value. A "sufficiently" high prevalence of a co-location pattern indicates that the pattern can be used to generate confident co-location rules. Namely, assume that the user-specified confidence threshold for interesting co-location rules is min_conf. Then, if $prev(P) \geq min\_conf$, rule $f \rightarrow f_1, \ldots, f_n$ is confident for all $f \in \mathcal{P}$.

In the example of Fig. 1 the prevalence $prev(AB) = \min(2/3, 1) = 2/3$. Thus, one can generate rules $A \rightarrow B$, the confidence of rule being 2/3, and $B \rightarrow A$ (confidence 1).

Another interestingness measure proposed for co-location patterns is *maximum participation ratio* (MPR). Prevalence of a pattern is the minimum of the participation rations of its features, whereas MPR is defined as the *maximum* of them. Correspondingly, a "sufficiently" high MPR implies that at least one of the features, denote it by T, rarely occurs outside $\mathcal{P}$. Hence, the co-location rule $\{T\} \rightarrow \mathcal{P} \setminus \{T\}$ is confident (Huang et al. 2003). The motivation of using the MPR is that rare features can more easily be included in the set of interesting patterns.

A drawback of MPR is that it is not monotonous. However, a weaker property ("weak monotonicity") can be proved for MPR. This property is utilized in Huang et al. (2003) to develop a level-wise search algorithm for mining confident co-location rules.

The buffer-based model extends the co-location patterns to polygons and line strings (Xiong et al. 2004). The basic idea is to introduce a buffer, which is a zone of a specified distance, around each spatial object. The boundary of the buffer is the isoline of equal distance to the edge of the objects (see Fig. 2). The (Euclidean) neighborhood N(o) of



Co-location Patterns, Interestingness Measures, Fig. 2 Examples of neighborhoods in the buffer-based model

an object o is the area covered by its buffer. The (Euclidean) neighborhood of a feature f is the union of $\mathcal{N}(o_i)$, where $o_i \in O_f$, and O_f is the set of instances of f. Further, the (Euclidean) neighborhood $\mathcal{N}(C)$ for a feature set $C = \{f_1, f_2, \ldots, f_n\}$ is defined as the intersection of $\mathcal{N}(f_i)$, $f_i \in C$.

The coverage ratio $Pr(C)$, where $C = \{f_1, f_2, \ldots, f_n\}$ is a feature set is defined as $\frac{\mathcal{N}(C)}{Z}$, where Z is the total size of the investigation area. Intuitively, the coverage ratio of a set of features measures the fraction of the investigation area that is influenced by the instances of the features.

The coverage ratio is a monotonous interestingness measure in the pattern class of co-location patterns in the buffer-based model, with respect to the size of the co-location pattern (Xiong et al. 2004). Now in the buffer-based model the conditional probability (confidence) of a co-location rule $P \rightarrow Q$ expresses the probability of finding the neighborhood of Q in the neighborhood of P. Due to the monotonicity of coverage ratio, it can be computed as $\frac{\mathcal{N}(P \cup Q)}{\mathcal{N}(P)}$. Xiong et al. also demonstrate that the definition of conditional probability (confidence) of a co-location rule in the event-centric model does not satisfy the law of compound probability: it is possible that $Prob(BC|A) \neq Prob(C|AB)Prob(B|A)$, where $Prob(BC|A)$ is equal to the confidence of the rule $A \rightarrow BC$. They show, however, that in the buffer-based model this law holds.

Statistical Approaches

An essential difference in the viewpoints of *spatial statistics* and co-location pattern mining is that in statistics the dataset is considered as a sample. The aim in statistics is typically to infer, based on the sample, knowledge of properties of the "reality", that is, the phenomenon, that generated the data. The goal of co-location pattern mining is to find descriptions of the data, that is, only the content of the available database is the object of investigation. In a sense, statistical analysis is more ambitious. However, sophisticated statistical data analysis methods cannot always be applied to large data masses. This may be due to the lack of computational resources, expert knowledge, or other human resources needed to preprocess the data before statistical analysis is possible.

Furthermore, depending on the application, treating the content of a spatial database as a sample may be relevant, or not. Consider, for instance, roads represented in a spatial database. Clearly, it is usually the case that (practically) all of them are included in the database, not only a sample. On the other hand, in an ecological database that includes the known locations of nests of different bird species, it is obvious that not all the nests have been observed, and thus a part of the information is missing from the database. Another example is a linguistic database that contains dialect variants of words in different regions. Such variants cannot in practice be exhaustively recorded everywhere, and, thus, the data in the database is a sample.

Statistical analysis of spatial point patterns is closely related to the problem of finding interesting co-location patterns (see, e.g., Bailey and Gatrell 1995; Diggle 1983). In statistics, features are called *event types*, and their instances are *events*. The set of events in the investigation area form a *spatial point pattern*. Point patterns

of several event types (called *marked point patterns*) may be studied, for instance, to evaluate spatial correlation (either positive, i.e., clustering of events, or negative, i.e., repulsion of events). Analogously, the point pattern of a single event type can be studied for evaluating possible *spatial autocorrelation*, that is, clustering or repulsion of the events of the event type.

In order to evaluate spatial (auto)correlation, point patterns, that is the data, are modeled as realizations (samples) generated by spatial point processes. A spatial point process defines a joint probability distribution over all point patterns. The most common measures of spatial correlation in point patterns are the $G(h)$, and $K(h)$-functions. For a single event type the value of $G(h)$-function in data is the number of events such that the closest other event is within a distance less than h divided by the number of all events. For two event types, instead of the closest event of the same type, the closest event of the *other* event type is considered. Thus, the confidence of the co-location rule $A \rightarrow B$, where A and B are single features in the event-centric model, is equal to the value of $G_{A,B}$ (h)-function in the data, when the neighborhood relation is defined as the maximum distance of h between objects.

The statistical framework implies that the relationship of the phenomenon and the data, which is a sample, has to be modeled in some way. In spatial statistics, the interestingness measures can be viewed from several perspectives, depending on the statistical framework, and the methods used in the data analysis. One of the most common frameworks is the hypothesis testing.

Hypothesis testing sets up a null hypothesis, typically assuming no correlation between features, and an alternative hypothesis that assumes spatial correlation. A test statistic, e.g., $G(h)$ or $K(h)$-function, for measuring spatial correlation is selected, denote it by T. The value of the test statistic in data, denote it by t, is compared against the theoretical distribution of the test statistic, assuming that the null hypothesis holds. Then, a natural interestingness measure of the observed spatial correlation is based on the so-called p-value, which is defined as $Pr(T > t|H_0)$. The smaller the p-value, the smaller the probability that the observed degree of spatial correlation could have been occurred by chance. Thus, the correlation can be interpreted as interesting if the p-value is small. If the p-value is not greater than a predefined α, the deviation is defined to be *statistically significant* with the *significance level* α.

The correlation patterns introduced in Salmenkivi (2006) represent an intermediate approach between spatial point pattern analysis and co-location pattern mining. Correlation patterns are defined as interesting co-location patterns (in the event-centric model) of the form $A \rightarrow B$, where A and B are single features. The interestingness is determined by the statistical significance of the deviation of the observed $G(h)$-value from a null hypothesis assuming no spatial correlation between features A and B.

Key Applications

Large spatial databases and spatial datasets. Examples: digital road map (Shekhar and Ma), census data (Malerba et al. 2001), place name data (Leino et al. 2003; Salmenkivi 2006).

Future Directions

A collection of interesting patterns can be regarded as a summary of the data. However, the pattern collections may be very large. Thus, condensation of the pattern collections and pattern ordering are important challenges for research on spatial co-location patterns.

Co-location patterns and rules are local in the sense that, given a pattern, only the instances of the features that appear in the pattern are taken into account when evaluating the interestingness of the pattern. However, the overall distribution and density of spatial objects and features may, in practice, provide significant information as to the interestingness of a pattern. This challenge is to some extent related to the challenge of integrating statistical and data mining approaches.

Cross-References

- ▸ Co-location Pattern
- ▸ Co-location Pattern Discovery
- ▸ Data Analysis, Spatial
- ▸ Frequent Itemset Discovery
- ▸ Frequent Pattern
- ▸ Statistical Descriptions of Spatial Patterns

References

Agrawal R, Ramakrishnan S (1994) Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th international conference on very large data bases, Santiago, 12–15 Sept, pp 487–499

Bailey TC, Gatrell AC (1995) Interactive spatial data analysis. Longman, Harlow

Diggle PJ (1983) Statistical analysis of spatial point patterns. Mathematics in biology. Academic, London

Huang Y, Xiong H, Shekhar S, Pei J (2003) Mining confident co-location rules without a support threshold In: Proceedings of the 2003 ACM symposium on applied computing (ACM SAC March, 2003), Melbourne, pp 497–501

Koperski K, Han J (1995) Discovery of spatial association rules in geographic information databases. In: Proceedings of 4th international symposium on large spatial databases (SSD95), Portlane, pp 47–66

Leino A, Mannila H, Pitkänen R (2003) Rule discovery and probabilistic modeling for onomastic data. In: Lavrac N, Gamberger D, Todorovski L, Blockeel H (eds) Knowledge discovery in databases: PKDD 2003. Lecture notes in artificial intelligence, vol 2838. Springer, Heidelberg, pp 291–302

Malerba D, Esposito F, Lisi FA (2001) Mining spatial association rules in census data. In: Proceedings of 4th international seminar on new techniques and technologies for statistics (NTTS 2001), Crete

Mannila H, Toivonen H, Verkamo AI (1995) Discovering frequent episodes in sequences. In: First international conference on knowledge discovery and data mining (KDD'95, August), pp. 210–215, Montreal. AAAI Press

Morimoto Y (2001) Mining frequent neighboring class sets in spatial databases. In: International proceedings of the 7th ACM SIGKDD conference on knowledge and discovery and data mining, San Francisco, pp 353–358

Salmenkivi M (2006) Efficient mining of correlation patterns in spatial point data. In: Fürnkranz J, Scheffer T, Spiliopoulou M (eds) Knowledge discovery in databases: PKDD-06, Berlin, Proceedings. Lecture notes in computer science, vol 4213. Springer, Berlin, pp 359–370

Shekhar S, Huang Y (2001) Discovering spatial co-location patterns: a summary of results. In: Proceedings of 7th international symposium on advances in spatial and temporal databases (SSTD 2001), Redondo Beach

Shekhar S, Ma X. GIS subsystem for a new approach to accessing road user charges

Xiong H, Shekhar S, Huang Y, Kumar V, Ma X, Yoo JS (2004) A framework for discovering co-location patterns in data sets with extended spatial objects. In: Proceedings of the fourth SIAM international conference on data mining (SDM04), Lake Buena Vista

Zaki MJ (2002) Efficiently mining frequent trees in a forest. In: Proceedings of 8th ACM SIGKDD international conference on knowledge discovery and data mining, Edmonton

Recommended Reading

Mannila H, Toivonen H (1997) Levelwise search and borders of theories in knowledge discovery. Data Min Knowl Disc 1(3):241–258

Co-location Rule Discovery

- ▸ Co-location Pattern Discovery

Co-location Rule Finding

- ▸ Co-location Pattern Discovery

Co-location Rule Mining

- ▸ Co-location Pattern Discovery

COM/OLE

- ▸ Smallworld Software Suite

Combinatorial Map

- ▸ Geosensor Networks, Qualitative Monitoring of Dynamic Fields

Complex Event Processing

▸ Data Stream Systems, Empowering with Spatiotemporal Capabilities

Components; Reuse

▸ Smallworld Software Suite

Composite Geographic Information Systems Web Application

▸ GIS Mashups

Computational Grid

▸ Grid

Computational Infrastructure

▸ Grid

Computer Cartography

▸ Conflation of Geospatial Data

Computer Environments for GIS and CAD

Joe Astroth
Autodesk Location Services, San Rafael, CA, USA

Synonyms

CAD and GIS platforms; Convergence of GIS and CAD; Evolution of GIS and LBS; Geo–mashups; Technological inflection points in GIS and CAD development

Definition

In the last few decades, computing environments have evolved to accommodate the need for integrating the separate, and often incompatible, processes of Geographic Information Systems (GIS) and Computer Assisted Design (CAD). This chapter will explore the evolution of GIS and CAD computing environments-from desktop to Web, and finally to wireless-along with the industry requirements that prompted these changes.

Historical Background

Before the 1980s, Computer Assisted Design (CAD) and Geographic Information Systems (GIS) functions were performed primarily on minicomputers running 32-bit operating systems such as VAX, VMS, or UNIX. Since minicomputers were expensive (approximately $200,000), many CAD and GIS solutions were bundled with hardware and offered as turnkey solutions. Although the combination of software and hardware was a popular option, it was still prohibitively expensive for small organizations, making CAD and GIS affordable only for government, academic institutions, and major corporations.

With the advent of the personal computer, particularly the IBM PC in 1981, which sold for approximately $1600, GIS and CAD became affordable for small- and medium-sized organizations.

Soon after the introduction of affordable personal computers, Autodesk developed the first PC-based CAD software, AutoCAD®, which sold for approximately $1000. Desktop GIS products appeared on the market shortly thereafter. Rather than being retrofitted from minicomputer programs, the most successful of these applications were engineered specifically for the PC. With the availability of these powerful desktop programs, small- to medium-sized organizations that had previously relied on analog mapping and drafting had access to the wealth of information and time-saving tools formerly available only to large organizations.

Scientific Fundamentals

CAD and GIS During the Workstation Phase

Although both GIS and CAD were introduced on the PC at around the same time, they were considered completely separate, and often incompatible, applications, making data sharing difficult. For example, precision and accuracy in GIS, unlike that of CAD, is variable, depending on scale. In CAD, precision was represented as 64-bit units (double precision), and in GIS as 32-bit units (single precision). Positional accuracy, which indicates the proximity of a feature on a map to its real location on earth, is quite high for CAD relative to GIS. For example, a 1:10,000 scale map might have a positional accuracy of 2.5 m (8.2 ft).

Another barrier to data sharing between CAD and GIS on the PC was the process of data collection. Many GIS survey instruments, such as Total Stations and GPS, collect data in ground units, rather than grid units. Ground units, which represent features on the earth's surface exactly, are both longer and bigger than grid units. In addition, elevations and scale are not factored into ground units.

Since a great deal of map data was collected in the field, maps drawn in CAD were stored in ground units, and scale and coordinate systems were added afterwards. CAD engineers found that using ground units, rather than grid units, was advantageous. For example, assuming that dimensions on the map were accurate, if the line on the map measured 100 m (328 ft), it corresponded to 100 m on the ground. With GIS grid units, 100 m on the map might actually correspond to 100.1 m (328 ft 3 in) on the ground.

Another significant difference between CAD and GIS applications is that in CAD, unlike in GIS, points and polylines represent objects in the real world, but contain no attached information. In GIS, points, polylines, and polygons can represent wells, roads, and parcels, and include attached tables of information. In many cases, when spatial information was transferred from CAD to GIS applications, features were "unintelligent" and had to be assigned meaningful information, such as topology and attributes, manually. Because of this manual step, translation from CAD to GIS was extremely difficult, even with automated import tools.

GIS and CAD applications also provided different types of tools, making it difficult for users to switch systems. In GIS applications, tools were designed for data cleanup, spatial analysis, and map production, whereas tools in CAD were intended for data entry, drafting, and design. Since CAD drafting tools were much easier to use, CAD technicians were wary of using GIS software for creation of design drawings.

CAD drawings themselves also made it difficult to transfer data to GIS. A typical CAD drawing contains objects made up of arcs and arbitrarily placed label text. However, in GIS, text can be generated based on attributes or database values, often producing a result that is not aesthetically pleasing to a cartographer.

The representation of land parcels, common in GIS applications for municipalities, presented another challenge for integrating CAD drawings and GIS. Polylines portraying lot and block lines in a survey plan need to be translated into meaningful polygons in GIS that represent the parcel. Cleanup tools are used to ensure the accuracy of the lot and block lines. Each parcel must also be associated with its appropriate attributes, and a polygon topology must be created so that Parcel Identification Numbers (PINs or PIDs) inside each polygon are linked to the parcel database.

These barriers to integrating GIS and CAD led to the development of software solutions in each phase of the technological advancement in computing environments.

Bridging the Gap Between CAD in GIS in the Workstation Phase

In the initial workstation phase, the only way to integrate GIS data with AutoCAD data was to use DXF™ (drawing exchange format). This process was extremely time-consuming and error-prone. Many CAD drawings were drawn for a particular project or plan and never used again. Often these drawings were not in the same coordinate system as the GIS and had to be transformed on import. Even today, a GIS enterprise is built

and maintained by importing data from CAD drawings. Graphic representations of layers of a formation, such as water, sewer, roads and parcels, are imported into the GIS using the file-based method.

To better merge the CAD world with the GIS world, a partnership was formed between Autodesk, Inc. and ESRI, leading to the creation of ArcCAD®. ArcCAD was built on AutoCAD and enabled users to create GIS layers and to convert GIS layers into CAD objects. This tool also facilitated data cleanup and the attachment of attributes. Because ArcCAD enabled GIS data to be shared with a greater number of people, the data itself became more valuable.

Although ArcCAD solved some of the integration problems between CAD and GIS, it still did not provide full GIS or CAD functionality. For example, overlay analysis still had to be performed in ArcInfo® and arcs and splines were not available in the themes created by ArcCAD.

In order to provide a fully functional GIS built on the AutoCAD platform, Autodesk developed AutoCAD Map® (now called Autodesk Map®), which made it simple for a CAD designer to integrate with external databases, build topology, perform spatial analysis, and utilize data cleaning, without file translation or lost data. In AutoCAD Map, lines and polygons were topologically intelligent with regard to abstract properties such as contiguity and adjacency. Since DWG™ files were already file-based packets of information, they became GIS-friendly when assigned topology and connected to databases. Precision was enforced instantly, since the DWG files could now store coordinate systems and perform projections and transformations. AutoCAD Map represented the first time a holistic CAD and GIS product was available for the PC Workstation environment.

Although AutoCAD Map could import and export the standard GIS file types (circa 1995: ESRI SHP, ESRI Coverage, ESRI E00, Microstation DGN, MapInfo MID/MIF, Atlas BNA) users began to request real-time editing of layers from third-party GIS files. To meet this demand, Autodesk created a new desktop GIS/CAD product called Autodesk World®. World was designed for users who were not GIS professionals or AutoCAD engineers, and offered the basic tools of both systems: precision drafting and the capability to query large geospatial data and perform rudimentary analysis and reports.

World used a Microsoft Office interface to access and integrate different data types, including geographic, database, raster, spreadsheet, and images, and supported Autodesk DWG as a native file format, increasing the value of maps created in AutoCAD and AutoCAD Map. World enabled users to open disparate GIS data files simultaneously and perform analysis regardless of file type. Autodesk World could access, analyze, edit and save data in all the standard formats without import or export.

Although Autodesk World represented a real breakthrough in integrating GIS and CAD files, it lacked an extensive CAD design environment. AutoCAD was still the CAD environment of choice, and AutoCAD Map continued to offer better integration of GIS within a full CAD environment. Autodesk World filled a need, much like other desktop GIS solutions at the time, but there was still a gap between the CAD design process and analysis and mapping within the GIS environment.

In the same time period, AutoCAD Map continued to evolve its GIS capabilities for directly connecting, analyzing, displaying, and theming existing GIS data (in SDE, SHP, DGN, DEM, and Raster formats, for example) without import or export. In support of the Open GIS data standard, AutoCAD Map could read OpenGIS information natively. GIS and CAD integration continues to be one of key features of AutoCAD Map.

CAD and GIS During the Web Phase

The next significant inflection point in technology was the World Wide Web, which increased the number of users of spatial data by an order of magnitude. With the advent of this new technology and communication environment, more people had access to information than ever before.

Initially, CAD and GIS software vendors responded to the development of the Web by Web-enabling existing PC applications. These Web-enabled applications offered the ability to assign

Universal Resource Locators (URLs) to graphic objects or geographic features, such as points, lines and polygons, and enabled users to publish their content for viewing in a browser as an HTML (Hypertext Markup Language) page and a series of images representing maps or design.

Software developers also Web-enabled CAD and GIS software by providing a thin client or browser plug-in, which offered rich functionality similar to the original application.

CAD for the Web

In the early Web era, slow data transfer rates required thin clients and plug-ins to be small (less than one megabyte) and powerful enough to provide tools such as pan and zoom. In light of this, Autodesk's developed a CAD plug-in called Whip! which was based on AutoCAD's ADI video driver.

Although the Whip! viewer today has evolved into the Autodesk DWF™ Viewer, the file format, DWF (Design Web Format) remains the same. DWF files can be created with any AutoCAD based product, including AutoCAD Map, and the DWF format displays the map or design on the Web as it appears on paper. DWF files are usually much smaller than the original DWGs, speeding their transfer across the Web. With the development of DWF, Internet users had access to terabytes of information previously available only in DWG format. This was a milestone in information access.

From a GIS perspective, 2D DWF files were useful strictly for design and did not represent true coordinate systems or offer GIS functionality. Although Whip!-based DWF was extremely effective for publishing digital versions of maps and designs, GIS required a more comprehensive solution.

Note: Today, DWF is a 3D format that supports coordinate systems and object attributes.

GIS for the Web

As the Web era progressed, it became clear that a simple retrofit of existing applications would not be sufficient for Web-enabled GIS. In 1996, Autodesk purchased MapGuide® from Argus Technologies. MapGuide viewer was a browser plug-in that could display full vector-format GIS data streamed from an enormous repository using very little bandwidth. Each layer in MapGuide viewer could render streamed data from different MapGuide Servers around the Internet. For example, road layers could be streamed directly from a server in Washington, DC, while the real-time location of cars could be streamed directly from a server in Dallas, Texas. MapGuide managed its performance primarily with scale-dependent authoring techniques that limited the amount of data based on the current scale of the client map.

MapGuide could perform basic GIS functions such as buffer and selection analysis, as well as address-matching navigation with zoom-goto. One of the more powerful aspects of MapGuide was the generic reporting functionality, in which MapGuide could send a series of unique IDs of selected objects to any generic Web page for reporting. Parcels, for example, could be selected in the viewer and the Parcel IDs could be sent to a server at City Hall that had the assessment values. A report was returned, as a Web page, containing all the information about the selected parcels. Again, the report could reside on any server, anywhere. The maps in MapGuide were just stylized pointers to all the potential servers around the Internet, containing spatial and attribute data. MapGuide was revolutionary at the time, and represented, in the true sense, applications taking advantage of the distributed network called the Web.

MapGuide continued to evolve, using ActiveX controls for Microsoft Internet Explorer, a plug-in for Netscape and a Java applet that could run on any Java-enabled browser. Initially, MapGuide used only its own file format, SDF, for geographic features. Later, MapGuide could natively support DWG, DWF, SHP, Oracle Spatial, and ArcSDE.

Although MapGuide was an extremely effective solution, it could run only on Microsoft Windows servers. The development of MapGuide OpenSource and Autodesk MapGuide Enterprise was inspired by the need to move toward a neutral server architecture and plug-in-free client experience. MapGuide could be now be used either without a plug-in or with the newest DWF Viewer as a thin client.

Within AutoCAD Map, users could now publish directly to the MapGuide Server and maintain the data dynamically, further closing the GIS-CAD gap.

CAD and GIS During the Wireless Phase

Wireless CAD and GIS marked the beginning of the next inflection point on the information technology curve, presenting a new challenge for GIS and CAD integration. Since early wireless Internet connection speeds were quite slow-approximately one quarter of wired LAN speed-Autodesk initially decided that the best method for delivering data to handheld device was "sync and go," which required physically connecting a handheld to a PC and using synchronization software to transfer map and attribute data to the device. GIS consumers could view this data on their mobile devices in the field without being connected to a server or desktop computer. Since handheld devices were much less expensive than PCs, mobile CAD and GIS further increased the number of people who had access to geospatial information.

Wireless CAD

Autodesk OnSite View (circa 2000) allowed users to transfer a DWG file to Palm-OS handheld and view it on the device. When synchronized, the DWG file was converted to an OnSite Design file (OSD), and when viewed, allowed users to pan, zoom and select features on the screen.

With the advent of Windows CE support, OnSite View allowed redlining, enabling users to mark up a design without modifying the original. Redlines were saved as XML (Extensible Markup Language) files on the handheld and were transferred to the PC on the next synchronization or docking. These redline files could be imported into AutoCAD, where modifications to the design could be made.

Autodesk OnSite View could be considered more mobile than wireless, since no direct access to the data was available without connecting the mobile device to the PC. OnSite View filled a temporary niche before broadband wireless connections became available.

Wireless GIS and Location-Based Services

Initially, the mobile GIS solution at Autodesk was OnSite Enterprise, which leveraged the mobility of OnSite and the dynamism of MapGuide. OnSite Enterprise created handheld MapGuide maps in the form of OSD files that users could simply copy off the network and view on their mobile devices with OnSite.

In 2001, when true broadband wireless came on the horizon, Autodesk created a new corporate division focused solely on Location-Based Services (LBS). The burgeoning Wireless Web required a new type of software, designed specifically to meet the high transaction volume, performance (+ 40 transactions per second), and privacy requirements of wireless network operators (WNOs). The next technological inflection point had arrived, where maps and location-based services were developed for mass-market mobile phones and handheld devices.

Autodesk Location Services created LocationLogic™, a middleware platform that provides infrastructure, application services, content provisioning, and integration services for deploying and maintaining location-based services. The LocationLogic platform was built by the same strong technical leadership and experienced subject matter experts that worked on the first Autodesk GIS products. The initial version of LocationLogic was a core Geoserver specifically targeted for wireless and telecom operators that required scalability and high-volume transaction throughput without performance degradation.

The LocationLogic Geoserver was able to provide:

- Point of Interest (POI) queries
- Geocoding and reverse geocoding
- Route planning
- Maps
- Integrated user profile and location triggers

Points of Interest (POIs) usually comprise a set of businesses that are arranged in different categories. POI directories, which can include hundreds of categories, are similar to Telecom Yellow Pages, but with added location intelligence.

Common POI categories include Gas Stations, Hotels, Restaurants, and ATMs, and can be customized for each customer. Each listing in the POI tables is spatially indexed so users can search for relevant information based on a given area or the mobile user's current location.

Geocoding refers to the representation of a feature's location or address in coordinates (x,y) so that it can be indexed spatially, enabling proximity and POI searches within a given area. Reverse geocoding converts x, y coordinates to a valid street address. This capability allows the address of a mobile user to be displayed once their phone has been located via GPS or cell tower triangulation. Applications such as "Where am I?" and friend or family finders utilize reverse geocoding.

Route planning finds the best route between two or more geographical locations. Users can specify route preferences, such as shortest path based on distance, fastest path based on speed limits, and routes that avoid highways, bridges, tollways, and so on. Other attributes of route planning include modes of transportation (such as walking, subway, car), which are useful for European and Asian countries.

The maps produced by the LocationLogic's Geoserver are actually authored in Autodesk MapGuide. Although the Geoserver was built "from the ground up," LocationLogic was able to take advantage of MapGuide's effective mapping software.

LocationLogic also supports user profiles for storing favorite routes or POIs. Early versions of LocationLogic also allowed applications to trigger notifications if the mobile user came close to a restaurant or any other point of interest. This capability is now used for location-based advertising, child zone notifications, and so on.

Key Applications

Early LBS applications built on LocationLogic included traffic alerts and friend finder utilities. For example, Verizon Wireless subscribers could receive TXT alerts about traffic conditions at certain times of day and on their preferred routes. Friend finder utilities alerted the phone user that people on their list of friends were within a certain distance of the phone.

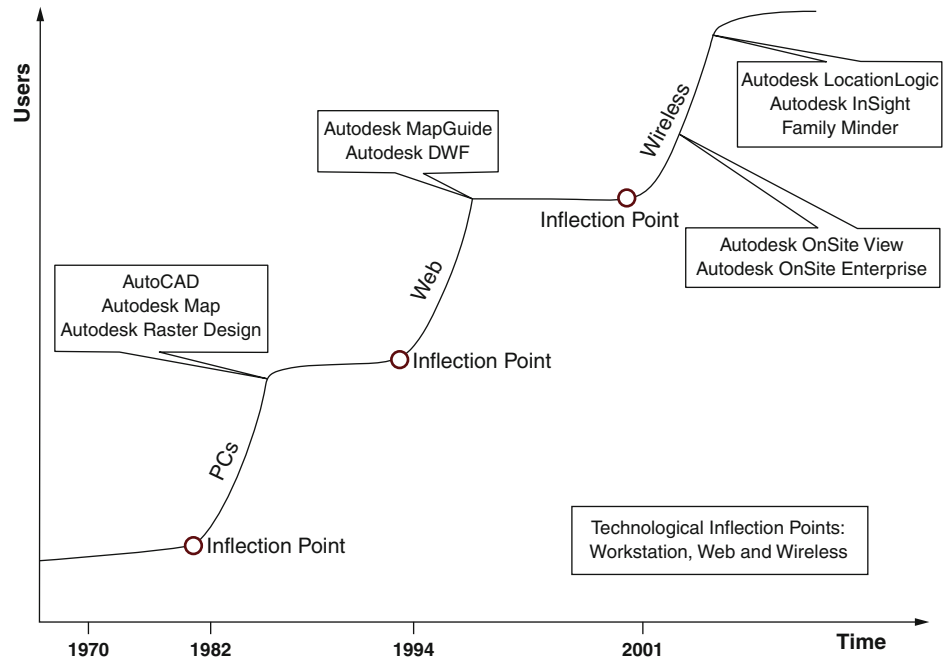
More recently, Autodesk Location Services has offered two applications built on LocationLogic that can be accessed on the cell phone and via a Web browser: Autodesk Insight™ and Autodesk Family Minder.

Autodesk Insight is a service that enables any business with a PC and Web browser to track and manage field workers who carry mobile phones. Unlike traditional fleet tracking services, Insight requires no special investment in GPS hardware. Managers and dispatchers can view the locations of their staff, determine the resource closest to a customer site or job request, and generate turn-by-turn travel instructions from the Web interface. Managers can also receive alerts when a worker arrives at a given location or enters or leaves a particular zone. Reports on travel, route histories, and communications for groups or individuals, over the last 12 or more months, can be generated from the Web interface.

Family Minder allows parents and guardians to view the real-time location of family members from a Web interface or their handset. Parents and guardians can also receive notifications indicating that a family member has arrived at or left a location. The recent advances in mobile phone technology, such as sharper displays, increased battery life and strong processing power, make it possible for users to view attractive map displays on regular handsets.

Enterprise GIS: Workstation, Web and Wireless Synergy

In 1999, Autodesk acquired VISION*®, along with its expertise in Oracle and enterprise GIS integration. This was a turning point for Autodesk GIS. File-based storage of information (such as DWG) was replaced with enterprise database storage of spatial data. Currently, Autodesk has integrated VISION* into its development, as seen in Autodesk GIS Design server. Autodesk Topobase™, which also stores its data in Oracle, connects to AutoCAD Map and MapGuide to provide enterprise GIS Public Works and Municipal solutions.



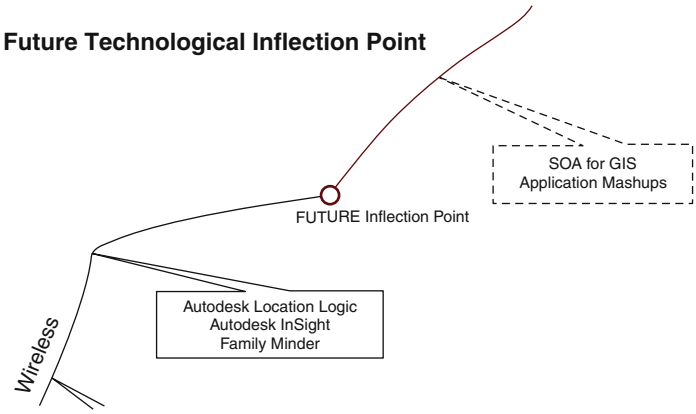
Computer Environments for GIS and CAD, Fig. 1 Technological inflection points along the information technology curve – exponential jumps in access to geospatial information

MapGuide and AutoCAD Map support Oracle Spatial and Locator, which allow all spatial data to be stored in a central repository. All applications can view the data without duplication and reliance on file conversion. AutoCAD Map users can query as-built information from the central repository for help in designs, and any modifications are saved and passed to GIS users. The central GIS database can also be published and modified from Web-based interfaces, such as MapGuide. Real-time wireless applications, such as Autodesk Insight, can use the repository for routing and mobile resource management.

Summary

At each technological inflection point-workstation, Web and wireless-Autodesk has leveraged infrastructural changes to exponentially increase the universe of potential consumers of geospatial information. The shift from minicomputer to PC saw Autodesk create AutoCAD and AutoCAD Map to enable sharing of geographic and design information. The next inflection point, workstation to Web, spurred another jump in the number of spatial data consumers, and the CAD and GIS gap continued to close. The most recent inflection point, Web to wireless, saw the number of spatial data users reach a new high, as GIS applications were embedded in the users' daily tools, such as cell phones (see Fig. 1). At this point in the technology curve, the need for synergy between CAD and GIS is apparent more than ever. Since the value of spatial data increases exponentially with the number of users who have access to it, Autodesk's enterprise GIS solution, with its centralized spatial database, provides significant value to a wide variety of spatial data consumers.

Autodesk has a history of leveraging inflection points along the computing and communication technology curve to create exciting and innovative solutions. For over two decades, Autodesk's mission has been to spearhead the "democratization of technology" by dramatically increasing the accessibility of heretofore complex and expensive software. This philosophy has been pervasive in the GIS and LBS solutions that it has brought to a rapidly growing geospatial user community.



Computer Environments for GIS and CAD, Fig. 2 Future technological inflection point (Continued from Fig. 1)

Future Directions

The next potential inflection point will emerge with the development of Service Oriented Architecture (SOA), built upon a Web 2.0 and Telco 2.0 framework. Not only will the distributed data and application architecture continue to increase the number of geospatial data consumers, but it will increase the use and accessibility of powerful analytical and visual tools as well.

Historically, the Web was leveraged to distribute data with wireless and Web technology. Now, the "geo-mashups" between tools such as Google Earth and AutoCAD Civil 3D, make use of the interaction of Web-based applications and data. A simple example of an SOA application is LocationLogic's geocoder, which performs geocoding and reverse-geocoding via Asynchronous JavaScript and XML (AJAX) calls to a URL that return sets of coordinates and addresses, respectively.

As GIS applications become integrated into current technologies (such as cars and courier boxes), demand for rapid data and application processing will apply pressure to all aspects of the distribution model. One challenge will be to provide rapid information updates, such as current aerial photographs and the latest traffic conditions. These "just-in-time" applications will require a massive increase in scale to accommodate the large number of business and personal users. At each technological inflection point, the accessibility to this vital information will increase exponentially (see Fig. 2).

CAD and GIS will soon be so integrated that the location on the timeline from design to physical feature or survey to map will be the only way to determine which technology is currently being used. Seamless and transparent services and data distribution will bring subsets of CAD and GIS utilities together to produce dynamic applications on demand. Servers will no longer host only databases, but will run self-supported applications, functions, and methods that are CAD, GIS, database, and business oriented. These services will be offered through the new Web 2.0 to provide powerful solutions.

Transparent GIS Services and integrated geospatial data will affect a larger segment of the population. No longer will the technology just be "cool," but will be completely integral to daily life. Autodesk's role will be to continue to provide tools that will leverage this new reality and meet the coming new demands in information and technology.

Cross-References

- Data Models in Commercial GIS Systems
- Internet GIS
- Internet-Based Spatial Information Retrieval
- Location Intelligence
- Location-Based Services: Practices and Products
- Oracle Spatial, Raster Data
- Privacy Threats in Location-Based Services
- Vector Data

- ▸ Web Mapping and Web Cartography
- ▸ Web Services, Geospatial

Recommended Reading

Autodesk Geospatial. http://images.autodesk.com/adsk/files/autodesk_geospatial_white_paper.pdf
Autodesk Inc. (2007) Map 3D 2007 essentials. Autodesk Press, San Rafael
Barry D (2003) Web services and service-orientated architectures, the Savvy Manager's guide, your road map to emerging IT. Morgan Kaufmann Publishers, San Francisco
Best Practices for Managing Geospatial Data. http://images.autodesk.com/adsk/files/%7B574931BD-8C29-4A18 -B77C-A60691A06A11%7D_Best_Practices. pdf
CAD and GIS – Critical Tools, Critical Links: Removing Obstacles Between CAD and GIS Professionals. http://images.autodesk.com/adsk/files/3582317_Critical Tools0.pdf
Hjelm J (2002) Creating location services for the wireless guide. Professional Developer's guide series. Wiley Computer Publishing, New York
Jagoe A (2003) Mobile location services, the definitive guide. Prentice Hall, Upper Saddle River
Kolodziej K, Hjelm J (2006) Local positioning systems, LBS applications and services. Taylor and Francis group. CRC Press, Boca Raton.
Laurini R, Thompson D (1992) Fundamentals of spatial information systems. The APIC series. Academic, London/San Diego
Longley P, Goodchild M, Maguire D, Rhind D (1999) Geographical information systems, 2nd edn. Principles and technical issues, vol 1; Management issues and applications, vol 2. Wiley, New York
Sharma C (2001) Wireless internet enterprise applications. Wiley tech brief series. Wiley Computer Publishing, New York
Schiller J, Voisard A (2004) Location-based services. Morgan Kaufmann, San Francisco
Plewe B (1997) GIS ONLINE; information retrieval, mapping and the internet. OnWord Press, Santa Fe
Vance D, Smith C, Appell J (1998) Inside autodesk world. OnWord Press, Santa Fe
Vance D, Walsh D, Eisenberg R (2000) Inside AutoCAD map 2000. Autodesk Press, San Rafael
Vance D, Eisenberg R, Walsh D (2000) Inside AutoCAD map 2000, the ultimate how-to resource and desktop reference for AutoCAD map. OnWord Press, Florence

Computer Supported Cooperative Work

- ▸ Geocollaboration

Computer Vision Augmented Geospatial Localization

Ashish Gupta
Department of Civil, Environmental, and Geodetic Engineering, Ohio State University, Columbus, OH, USA

Synonyms

Autonomous navigation; Global navigation satellite systems; GPS-denied geo-localization; Simultaneous localization and mapping; Unmanned aerial vehicles; Visual odometry

Definition

Geospatial localization is the estimation of global geographic location using, in part, geospatial analysis. Geospatial analysis uses statistical and other analytic techniques for data that has a geographic or spatial context to it, typically available in geographic information systems (GIS). Geographic location is typically ascertained using Global Navigation Satellite Systems (GNSS) like GPS and GLONASS, which requires simultaneous line-of-sight connection with multiple satellites to estimate location within an error margin of a few meters. These constraints limit the use of GNSS-based localization to outdoors with few obstructing structures in close proximity and a tolerance to uncertainty in exact location. In addition to these constraints, in many environments such as indoors, urban canyons, under dense foliage, underwater, and underground, there is limited or no GPS access. Besides these naturally occurring constraints, GPS access can be easily blocked by jamming, spoofing, and other GPS-denial threats in adversarial environments. Consequently, for positioning, navigation, and timing (PNT) applications, GPS must be augmented or supplanted by other sensors and systems. In such cases GPS is used for an approximate localization within a geographic region, which can range from tens to thousands of square

meters based on the environment. Alternate techniques are used to ascertain exact location within this geographic region. Simultaneous localization and mapping (SLAM) techniques are popularly used in robotics to estimate the location of a robot in real time based on information acquired from the environment using sensors on board the robot (Lategahn et al. 2011). It is typical to use multiple sensors like inertial measurement units (IMU), single or double video cameras for monocular or stereo vision, light detection and ranging (LIDAR), and sound navigation and ranging (SONAR). The choice of sensor suite is based on the environment (aerial, ground, underwater, indoor), the type of mobile platform, and the performance, processing, and cost budget. Vision-based sensors are among the most popular in SLAM techniques since they are informative and cost-effective. In addition to acquiring sensor information of its vicinity and estimating its position, SLAM also builds a map of the geographic region in real time. There are different types of maps. However, with navigation being the principal objective, topological maps are most relevant. A topological map focuses on the connectivity between important entities in the environment with disregard to their exact location (Paul and Newman 2010). Metric mapping can be used in conjunction with topological maps to compute a topometric map which is used to compute exact localization within that geographic region (Badino et al. 2011). This technique records sensor information with its registered location in a map database. Subsequently, while moving through that geographic region, sensor information can be used as a query to the recorded map database to retrieve geospatial location in real time.

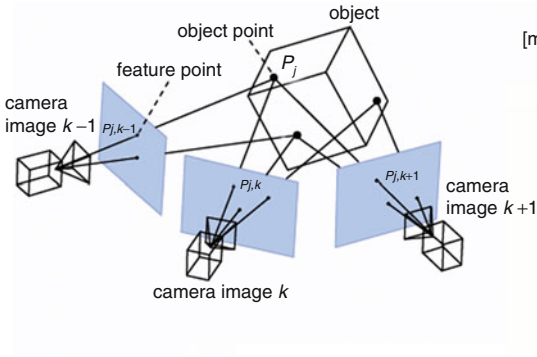
Historical Background

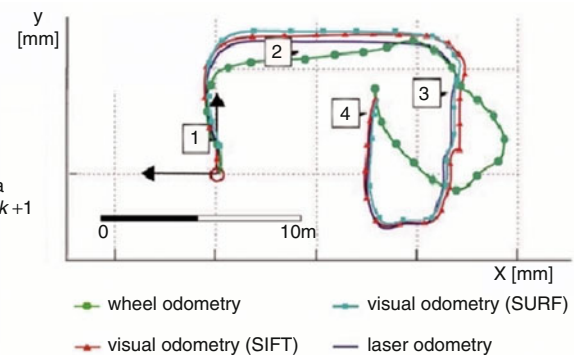
Long-range navigation (LORAN) was a hyperbolic radio navigation system developed prior to the advent of GNSS-based PNT. It used low-frequency radio waves and covered several thousand miles, but had a poor accuracy of hundreds of meters. It was used for localization of naval vessels while cruising oceans. Such systems are too inaccurate for precision demands of present-day PNT-dependent systems. To overcome the accuracy issues of GNSS for precision aviation, a Local Area Augmentation System (LAAS) is used for precision aircraft landing in all weather conditions (Enge 1999). A VHF signal link from airport transmitters is used by aircraft to correct GPS signal for precise localization. Cellular capable devices can use Assisted GPS (A-GPS) for improved localization using information provided by the cellular network in conjunction with satellite signal for a quicker estimation of location. Localization using cellular tower triangulation is another alternative with a reasonable error of several tens of meters, but it is only feasible outdoors in urban areas. For indoor navigation, the IEEE 802.11 wireless LAN (WLAN) location tracking system is an option (Emery and Denko 2007). It uses received signal strength indication (RSSI) on mobile devices and estimates location by comparison with a precomputed database of RSSI measurements in that indoor environment. It accounts for signal propagation loss and can provide accuracy of a few meters.

These PNT systems are currently operational but are incumbent with high infrastructure cost and have other limitations like availability exclusively in urban environments. Moreover, they have a natural limitation of localization accuracy. In comparison, computer vision-based pose estimation techniques used in robotics for SLAM have a comparatively high localization accuracy, but have historically been used for mapping small-sized environments. However, success in the DARPA Grand Challenges for autonomous navigation of driverless vehicles over large distances using SLAM-based techniques established the viability of computer vision augmented geospatial localization as a viable PNT alternative in GPS-denied or GPS-degraded environments (Thrun et al. 2006).

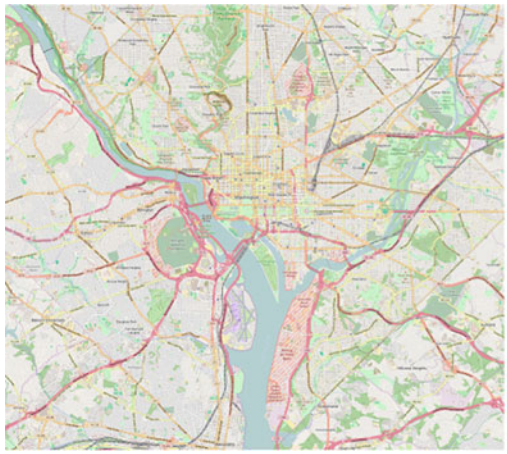
Scientific Fundamentals

Visual sensor-based localization using computer vision comprises two main parts in its approach: metric and topological localization. Metric





Computer Vision Augmented Geospatial Localization, Fig. 1 Estimation of pose of sensors onboard moving vehicle. Trajectory of moving vehicle is estimated based on sensor locations in current and previous frames. Localization accuracy depends on the type of sensor and type of features computed from the data stream. The graph is illustrative of the difference between trajectories recovered using different types of odometry techniques



Computer Vision Augmented Geospatial Localization, Fig. 2 Map data is abstracted as graph data structure. The transport network layer from OpenStreetMap of Washington, DC, is abstracted as a graph, where *edges* typically represent roads and *nodes* represent intersections and other relevant points in the map

location is estimated by computing the coordinates of the location of the sensor on board the vehicle. These could be geographic coordinates of latitude and longitude. The coordinates of the vehicle pose are typically computed by triangulation, using methods like structure from motion (SfM) (Koenderink and van Doorn 1991) or Visual Odometry (VO) (Alonso et al. 2012). An illustrative example is shown in Fig. 1. Pose of the sensor is estimated based on matching features across different frames in the data stream of a moving vehicle. The sequence of poses is used to estimate a 3D trajectory of this moving vehicle. In topological localization, the position of the sensor on board the vehicle is retrieved from a finite set of possible locations. Topological localization provides a coarse location estimate. Topological maps are typically stored as graph structures, where nodes indicate possible locations and edges are connections between locations. An example is shown in Fig. 2 for the city of Washington, DC, where the transport network layer acquired from OpenStreetMap for the county area has been abstracted as a graph. The weight of an edge can indicate the similarity or proximity between locations. The size of the finite set of locations is typically kept small so that efficient retrieval in real-time applications is a tractable

problem. While metric approaches provide accurate localization results, they tend to fail and drift over time as the vehicle traverses big distances in its geographic region. On the other hand, due to its finite state space, topological approaches provide a robust localization but only rough position estimates. A fusion of the metric and topological approaches achieves accurate metric results while maintaining the robustness of topological matching, which is a technique typically referred to as topometric localization. It uses a fine-grained topological map, where each node has an associated coordinate of its real metric location. Such topological maps can be acquired from sources like GIS databases for outdoor navigation. Finding the node of the current location translates to finding the metric coordinate of the vehicle.

A generic topometric localization algorithm involves the two stages of map creation and then localization. A vehicle equipped with cameras, IMU, and GNSS-capable device first traverses the routes to be mapped. GPS and inertial sensors are used to create a graph of this environment. The graph is metric in the sense that the nodes contain the exact location of the vehicle. From the acquired images using an onboard camera in monocular or stereo configuration, visual local features are extracted. These features are processed and stored in a database with a reference to the node corresponding to its real location. At runtime, the vehicle drives over the routes included in the a priori map. Video imagery is processed online to obtain features. As the vehicle moves, these visual features are matched with those in the database. Since there are potentially multiple feature matches from different parts of the mapped region, a method like Bayesian filtering is utilized to estimate the probability density function of the position of the vehicle. This facilitates pruning of false-positive matches and provides accurate localization and a smooth estimated trajectory of the moving vehicle.

Key Applications

Geospatial localization using a sensor suite that includes visual sensors in a GNSS-denied environment has several applications in numerous scenarios. Since it provides very accurate location information in real time, it can be used for autonomous navigation for self-driving cars, unmanned aerial vehicles (UAV) and unmanned ground vehicles (UGV). A vision-based system provides rich real-time information that allows an autonomous navigation system to tackle a dynamic environment, such as appearance of unexpected objects in the vicinity of the vehicle that were absent during the mapping phase, which is not available in other PNT systems.

A robust GPS alternative is particularly important for military applications since GPS signals can easily be denied to mission critical navigation systems on several assets, especially in contested territory. Relatively cost-effective vision-based geo-localization can be alternatively used by guidance systems on weapons platforms like missiles, drones, and UGVs.

Community-driven map generation projects like OpenStreetMap are extremely popular (Floros et al. 2013). Accurate and information-rich vision-based localization can simultaneously correct registration errors in these maps and also annotate the maps with geo-referenced objects like buildings, road signs, vegetation, and other geographic entities.

Vision-based localization is typically unhampered by its environment, unlike radio signals which suffer issues of multipath and propagation path losses by absorption. Since it can be used in most environments, it can also be used ubiquitously with disregard to change in environments like transitioning from outdoors to indoors, driving through tunnels, etc. which otherwise typically require a hand-off between different PNT systems operational in their respective environments.

Future Directions

Computer vision augmented geospatial localization is a rapidly emerging technology. Future developments include improved sensor fusion where vision, inertial, LIDAR, SONAR, magnetometer, and gravimeter sensors will be

efficiently combined for ubiquitous navigation while traveling across different environments without degradation in localization accuracy. The quality of information in GIS databases and accuracy of geospatial localization are synergistic where one improves the other and vice versa. Cross-referencing and registration of visual information from different mobile platforms including UAV and UGV can improve GIS databases and provide a ground and aerial map of a geographic region for accurate 3D geospatial localization.

Cross-References

- Bayesian Network Integration with GIS
- Feature Detection and Tracking in Support of GIS
- Indoor Localization
- OpenStreetMap
- Optimal Location Queries on Road Networks
- Road Network Data Model
- Spatial Analysis along Networks

References

Alonso IP, Llorca DF, Gavilan M, Pardo SA, Garcia-Garrido MA, Vlacic L, Sotelo MA (2012) Accurate global localization using visual odometry and digital maps on urban environments. IEEE Trans Intell Transp Syst 13(4):1535–1545

Badino H, Huber D, Kanade T (2011) Visual topometric localization. In: IEEE intelligent vehicles symposium, proceedings (2011), Baden-Baden, pp 794–799

Emery M, Denko M (2007) Ieee 802.11 wlan based real-time location tracking in indoor and outdoor environments. In: Canadian conference on electrical and computer engineering, CCECE 2007, Apr 2007, Vancouver, pp 1062–1065

Enge P (1999) Local area augmentation of gps for the precision approach of aircraft. Proc IEEE 87(1): 111–132

Floros G, Van Der Zander B, Leibe B (2013) OpenStreetSLAM: global vehicle localization using OpenStreetMaps. In: Proceedings – IEEE international conference on robotics and automation, Karlsruhe, pp 1054–1059

Koenderink JJ, van Doorn AJ (1991) Affine structure from motion. J Opt Soc Am A 8(2):377–385

Lategahn H, Geiger A, Kitt B (2011) Visual SLAM for autonomous ground vehicles. In: Proceedings – IEEE international conference on robotics and automation, Shanghai, pp 1732–1737

Paul R, Newman P (2010) FAB-MAP 3D: topological mapping with spatial and visual appearance. In: 2010 IEEE international conference on robotics and automation, Anchorage, pp 2649–2656

Thrun S, Montemerlo M, Dahlkamp H, Stavens D, Aron A, Diebel J, Fong P, Gale J, Halpenny M, Hoffmann G, Lau K, Oakley C, Palatucci M, Pratt V, Stang P, Strohband S, Dupont C, Jendrossek LE, Koelen C, Markey C, Rummel C, van Niekerk J, Jensen E, Alessandrini P, Bradski G, Davies B, Ettinger S, Kaehler A, Nefian A, Mahoney P (2006) Stanley: the robot that won the darpa grand challenge: research articles. J Robot Syst 23(9):661–692

Computing Fitness of Use of Geospatial Datasets

Leen-Kiat Soh and Ashok Samal
Department of Computer Science and Engineering, The University of Nebraska at Lincoln, Lincoln, NE, USA

Synonyms

Conflict Resolution; Dempster Shafer Belief Theory; Evidence; Frame of Discernment; Information Fusion; Plausibility; Quality of Information; Timeseries Data

Definition

Geospatial datasets are widely used in many applications including critical decision support systems. The goodness of the dataset, called the Fitness of Use (FoU), is used in the analysis and has direct bearing on the quality of derived information from the dataset that ultimately plays a role in decision making for a specific application. When a decision is made based on different sources of datasets, it is important to be able to fuse information from datasets of different degrees of FoU. Dempster-Shafer belief theory is used as the underlying conflict resolution mechanism during information fusion. Furthermore, the Dempster-Shafer belief theory is demonstrated as a viable approach to fuse information derived

from different approaches in order to compute the FoU of a dataset.

Historical Background

In most applications, sometimes it is assumed that the datasets are perfect and without any blemish. This assumption is, of course, not true. The data is merely a representation of a continuous reality both in space and time. It is difficult to measure the values of a continuous space and time variable with infinite precision. Limitations are also the result of inadequate human capacity, sensor capabilities and budgetary constraints. Therefore, the discrepancy exists between the reality and the datasets that are derived to represent it. It is especially critical to capture the degree of this discrepancy when decisions are made based on the information derived from the data. Thus, this measure of quality of a dataset is a function of the purpose for which it is used, hence it is called its *fitness of use* (FoU). For a given application, this value varies among the datasets. Information derived from high-FoU datasets is more useful and accurate for the users of the application than that from low-FoU datasets. The challenge is to develop appropriate methods to fuse derived information of varying degrees of FoU as well as of derived information from datasets of varying degrees of FoU. This will give insights as to how the dataset can be used or how appropriate the dataset is for a particular application (Yao 2003).

An information theoretic approach is used to compute the FoU of a dataset. The Dempster-Shafer belief theory (Shafer 1976) is used as the basis for this approach in which the FoU is represented as a range of possibilities and integrated into one value based on the information from multiple sources. There are several advantages of the Dempster-Shafer belief theory. First, it does not require that the individual elements follow a certain probability. In other words, Bayes theory considers an event to be either true or untrue, whereas the Dempster-Shafer allows for unknown states (Konks and Challa 2005). This characteristic makes the Dempster-Shafer belief theory a powerful tool for the evaluation of risk and reliability in many real applications when it is impossible to obtain precise measurements and results from real experiments. In addition, the Dempster-Shafer belief Theory provides a framework to combine the evidence from multiple sources and does not assume disjoint outcomes (Sentz and Ferson 2002). Additionally, the Dempster-Shafer's measures are not less accurate than Bayesian methods, and in fact reports have shown that it can sometimes outperform Bayes' theory (Cremer et al. 1998; Braun 2000).

Scientific Fundamentals

Assume that there is a set of geospatial datasets, $S = \{S_1, S_2, \cdots, S_n\}$. A dataset S_i may consist of many types of information including (and not limited to) spatial coordinates, metadata about the dataset, denoted by aux_i, and the actual time series data, denoted by ts_i.

The metadata for a dataset may include the type of information being recorded (e.g., precipitation or volume of water in a stream), the period of record, and the frequency of measurement. Thus,

$$aux_i = \langle type_i, tb_i, te_i, int_i \rangle,$$

where tb_i and te_i denote the beginning and the ending time stamps for the measurements, and int_i is the interval at which the measurements are made. Other metadata such as the type and age of the recording device can also be added.

The time series data in a dataset may consist of a sequence of measurements,

$$ts_i = \langle m_{i,1}, m_{i,2}, \ldots, m_{i,p} \rangle.$$

Each measurement stores both the time the measurement was taken and the actual value recorded by the sensor. Thus, each measurement is given by

$$m_{i,j} = \langle t_{i,j}, v_{i,j} \rangle.$$

It is assumed that the measurements in the dataset are kept in chronological order. Therefore,

$$t_{i,j} < t_{i,k}, \quad \text{for} \quad j < k.$$

Furthermore, the first and last measurement times should match the period of record stored in the metadata,

$$tb_i = t_{i,1} \quad \text{and} \quad te_i = t_{i,p}.$$

The problem of finding the suitability of a dataset for a given application is to define a function for the FoU that computes the fitness of use of a dataset described above. The function *FoU* maps S_i to a normalized value between 0 and 1:

$$FoU\,(S_i, A) = [0, 1]\,,$$

where S_i is a single dataset and A is the intended application of the data. The application A is represented in the form of domain knowledge that describes how the goodness of a dataset is viewed. A set of rules may be used to specify this information. Thus,

$$A = \{R_1, R_2, \cdots, R_d\}\,,$$

where R_i is a domain rule that describes the goodness of a dataset and d is the number of rules. Therefore, the *FoU* function is defined with respect to an application domain. Different applications can use different rules for goodness and derive different *FoU* values for the same dataset.

Dempster-Shafer Belief Theory

The two central ideas of the Dempster-Shafer belief theory are: (a) obtaining degrees of belief from subjective probabilities for a related question, and (b) Dempster's rule for combining such degrees of belief when they are based on independent items of evidence. For a given proposition P, and given some evidence, a confidence interval is derived from an interval of probabilities within which the true probability lies within a certain confidence. This interval is defined by the *belief* and *plausibility* supported by the evidence for the given proposition. The lower bound of the interval is called the *belief* and measures the strength of the evidence in favor of a proposition. The upper bound of the interval is called the *plausibility*. It brings together the evidence that is compatible with the proposition and is *not inconsistent* with it. The values of both belief and plausibility range from 0 to 1. The belief function (*bel*) and the plausibility function (*pl*) are related by:

$$pl(P) = 1 - bel(\overline{P}),$$

where $\overline{P}$ is the negation of the proposition P. Thus, $bel(\overline{P})$ is the extent to which evidence is in favor of $\overline{P}$.

The term *Frame of Discernment* (FOD) consists of all hypotheses for which the information sources can provide evidence. This set is finite and consists of mutually exclusive propositions that span the hypotheses space. For a finite set of mutually exclusive propositions (θ), the set of possible hypotheses is its power set (2^θ), i.e., the set of all possible subsets including itself and a null set. Each of these subsets is called a focal element and is assigned a confidence interval (belief, plausibility).

Based on the evidence, a probability mass is first assigned to each focal element. The masses are *probability-like* in that they are in the range [0, 1] and sum to 1 over all hypotheses. However, they represent the belief assigned to a focal element. In most cases, this basic probability assignment is derived from the experience and the rules provided by some experts in the application domain.

Given a hypothesis H, its belief is computed as the sum of all the probability masses of the subsets of H as follows:

$$bel(H) = \sum_{e \subset H} m(e),$$

where $m(e)$ is the probability mass assigned to the subset e. The probability mass function distributes the values on subsets of the frame of discernment. Only to those hypotheses, for which it has direct evidence, are the none-zero values assigned. Therefore, the Dempster-Shafer belief theory allows for having a single piece of evidence supporting a set of multiple propositions being true. If there are multiple sources of information, probability mass functions can be derived for each data source. These mass

values are then combined using Dempster's Combination Rule to derive joint evidence in order to support a hypothesis from multiple sources. Given two basic probability assignments, m_A and m_B for two independent sources (A and B) of evidence in the same frame of discernment, the joint probability mass, m_{AB}, can be computed according to Dempster's Combination Rule:

$$m_{AB}(C) = \frac{\sum_{A \cap B = C} m(A) * m(B)}{1 - \sum_{A \cap B = \emptyset} m(A) * m(B)}.$$

Furthermore, the rule can be repeatedly applied for more than two sources sequentially, and the results are order-independent. That is, combining different pieces of evidence in different sequences yields the same results.

Finally, to determine the confidence in a hypothesis H being true, belief and plausibility are multiplied together:

$$confidence(H) = bel(H) \cdot pl(H).$$

Thus, the system is highly confident regarding a hypothesis being true if it has high belief and plausibility for that hypothesis being true.

Suppose that there are three discrete FoU outcomes of the datasets *suitable* (s), *marginal* (m), and *unsuitable* (u), and $\theta = \{s, m, u\}$. Then, the frame of discernment is

$$\begin{aligned} FOD &= 2^{\theta} \\ &= \{\emptyset, \{s\}, \{m\}, \{u\}, \{s, m\}, \{s, u\}, \\ &\quad \{m, u\}, \{s, m, u\}\}. \end{aligned}$$

To illustrate how the Dempster-Shafer belief theory can be used to fuse derived information from geospatial databases, two novel approaches to derive information from geospatial databases are herein presented – i.e., (1) heuristics, and (2) statistics – before fusing them. For each approach, the computation of the FoU of the dataset based on the derived information is demonstrated.

In the first approach, a set of domain heuristics for this purpose and then the combination rule to compute the FoU of the datasets are used. The heuristics can be based on common sense knowledge or can be based on expert feedback. The following criteria are used:

- *Consistency* – A dataset is consistent if it does not have any gaps. A consistent dataset has a higher fitness value
- *Length* – The period of record for the dataset is also an important factor in the quality. Longer periods of record generally imply a higher fitness value
- *Recency* – Datasets that record more recent observations are considered to be of a higher fitness value
- *Temporal Resolution* – Data are recorded at different time scales (sampling periods). For example, the datasets can be recorded daily, weekly or monthly. Depending on the application, higher or lower resolution may be better. This is also called the granularity (Mihaila et al. 1999)
- *Completeness* – A data record may have many attributes, e.g., time, location, and one or more measurements. A dataset is complete if all the relevant attributes are recorded. Incomplete datasets are considered to be inferior (Mihaila et al. 1999)
- *Noise* – All datasets have some noise due to many different factors. All these factors may lead to data not being as good for use in applications.

For each of the above criteria, one or more heuristics can be defined to determine the probability mass for different data quality values. The heuristics in the form of rules are specified as follows:

$$\begin{aligned} &C_1(S_i) \wedge C_2(S_i) \wedge \cdots \wedge C_n(S_i) \rightarrow mass \\ &\quad (S_i, \{qtype\}) = m, \end{aligned}$$

where C_i specifies a condition of the dataset, C_j (S_i) evaluates to true if the condition C_j holds for the dataset S_i, and $mass(S_i, \{qtype\})$ denotes the mass of evidence that the dataset S_i contributes to the FoU outcome types in $\{qtype\}$. Then, a rule is triggered or fires if all the conditions are met. When the rule fires, the right-hand side of

the rule is evaluated which assigns a value m to the probability mass for a given set of outcome types which in the example, $\{qtype\} \subseteq \{suitable, marginal, unsuitable\}$.

Applying a set of rules as defined above to dataset S_i thus yields a set of masses for different combinations of outcome types. These masses are then combined using the Dempster's Combination Rule to yield a coherent set of masses for each element of FOD. The result may be further reduced by considering only the singletons: $\{suitable\}$,$\{marginal\}$, and $\{unsuitable\}$, which allows one to compute the belief, plausibility, and confidence values on only these three outcome types.

Now, how one may compute the FoU of a dataset using information derived statistically is shown. In the following, the statistical method used is called temporal variability analysis. Suppose that S_i has the following measurements:

$$ts_i = \langle m_{i,1}, m_{i,2}, \ldots, m_{i,p} \rangle, \text{ and}$$
$$m_{i,j} = \langle t_{i,j}, v_{i,j} \rangle .$$

Suppose that the measurements are collected periodically at some regular intervals. Suppose that the *period* of a data series is defined as the time between two measurements collected at the same spatial location at the same time mark. Given this notion of periodicity, the average value of all measurements at each particular time mark over the interval of measurements can be computed. Formally,

$ts_i = \langle m_{i,1}, m_{i,2}, \cdots, m_{i,p} \rangle$ can be re-written as:

$$ts_i = \langle m_{i,1}, m_{i,2}, \ldots, m_{i,period}, m_{i,period+1}, m_{i,period+2}, \cdots, m_{i,2*period}, m_{i,2*period+1}, \cdots, m_{i,k*period} \rangle,$$

such that $t_{i,k*period} - t_{i,1} = \text{int}_i$. Given the above representation, the periodic mean can be derived at each time mark j as

$$mean_{i,j} = \frac{\sum_{p=0}^{k} m_{i,p*period+j}}{k} .$$

Likewise, the periodic variance can be derived for the time marks j as

$$var_{i,j} = \frac{k \sum_{p=0}^{k} m^2_{i,p*period+j} - \left(\sum_{p=0}^{k} m_{i,p*period+j} \right)^2}{k\,(k-1)} .$$

Given the set of means and variances for all time marks in a period, the coefficient of variation at each time mark j can be further computed as

$$cov_{i,j} = \frac{\sqrt{var_{i,j}}}{mean_{i,j}} .$$

The temporal variability of the dataset S_i can then be defined as the average value of coefficient of variation for all time marks:

$$\bar{c}\,(S_i) = \frac{\sum_{j=1}^{period}}{cov_{i,j}} period .$$

Heuristics can then be used to assign probability masses to the different outcomes based on the value of $\bar{c}$. For example, to assign probability masses to the outcomes, the temporal variability can be divided into three ranges: the upper (largest) one-third, the middle one-third and the lower (smallest) one-third. For each range, one or more heuristics are defined to determine the probability mass for different FoU values. The heuristics are specified in the form of rules as

$$(\bar{c}\,(S_i) \;\; within\ range\ k) \rightarrow mass\,(S_i, \{qtype\}) = m,$$

where $\bar{c}(S_i)$ is the average coefficient of variation of the dataset S_i, and the range k is one the three ranges mentioned above. For a given dataset S_i, the right hand side of the above rule is evaluated and a value m to the probability mass is assigned for a given type (*suitable*, *marginal*, or *unsuitable*). These probability masses can also be combined using Dempster's Combination Rule.

Thus, at this point, there are two pieces of derived FoU values for a dataset. One is through the heuristic approach and the other through the statistical approach. Both FoU values represent a confidence in the dataset belonging to a particular type (suitable, marginal, or unsuitable). To obtain one single composite FoU out of the two values, yet another fusion can be performed. That is, to fuse the two derived information of varying FoU, one may simply treat each FoU as a mass for the dataset to belong to a particular *qtype*. Thus, by employing Dempster's Combination Rule, one can repeat the same process in order to obtain different mass values that support the notion that the dataset has a FoU of a certain type. This allows the FoUs to be fused at subsequent levels as well.

The idea of FoU has also been applied in several different contexts. De Bruin et al. (2001) have proposed an approach based on decision analysis where value of information and value of control were used. Value of information is the expected desirability of reducing or eliminating uncertainty in a chance node of a decision tree while value of control is the expected amount of control that one could affect the outcome of an uncertain event. Both of these values can be computed from the probability density of the spatial data. Vasseur et al. (2003) have proposed an ontology-driven approach to determine fitness of use of datasets. The approach includes conceptualization of the question and hypothesis of work to create the ontology of the problem, browsing and selecting existing sources informed by the metadata of the datasets available, appraisal of the extent of these databases matching or missing the expected data with the quality expected by the user, translation of the corresponding (matched) part into a query on the database, reformulation of the initial concepts by expanding the ontology of the problem, querying the actual databases with the query formulae, and final evaluation by the user to accept or reject the retrieved results. Further, Ahonen-Rainio and Kraak (2005) have investigated the use of sample maps to supplement the fitness for use of geospatial datasets.

Key Applications

The key category of applications for the proposed technique is to categorize or cluster spatio-temporal entities (objects) into similar groups. Through clustering, the techniques can be used to group a dataset into different clusters for knowledge discovery and data mining, classification, filtering, pattern recognition, decision support, knowledge engineering and visualization.

- *Drought Mitigation:* Many critical decisions are made based on examining the relationship between the current value of a decision variable (e.g., precipitation) and its historic norms. Incorporating the fitness of use in the computation will make the decision making process more accurate.
- *Natural Resource Management:* Many natural resources are currently being monitored using a distributed sensor network. The number of these networks continues to grow as the sensors and networking technologies become more affordable. The datasets are stored as typical time series data. When these datasets are used in various applications, it would be useful to incorporate the fitness of use values in the analysis.

Future Directions

The current and future work focuses on extending the above approach to compute the fitness of use for derived information and knowledge. FoU, for example, is applied directly to raw data. However, as data is manipulated, fused, integrated, filtered, cleaned, and so on, the derived metadata or information appears, and with it, an associated measure of fitness. This fitness can be based on the intrinsic FoU of the data that the information is based on and also on the *technique* that derives the information. For example, one may say that the statistical approach is more rigorous than the heuristic approach and thus should be given more mass or confidence. Likewise, this notion can

be extended to knowledge that is the result of using information. A piece of knowledge is, for example, a decision. Thus, by propagating FoU from data to information, and from information to knowledge, one can tag a decision with a confidence value.

Cross-References

- ▸ Crime Mapping and Analysis
- ▸ Data Collection, Reliable Real-Time
- ▸ Error Propagation in Spatial Prediction
- ▸ Indexing and Mining Time Series Data

References

Ahonen-Rainio P, Kraak MJ (2005) Deciding on fitness for use: evaluating the utility of sample maps as an element of geospatial metadata. Cartogr Geogr Inf Sci 32(2):101–112

Braun J (2000) Dempster-Shafer theory and Bayesian reasoning in multisensor data fuSion. In: Sensor fusion: architectures, algorithms and applications IV. Proceedings of SPIE, vol 4051, Orlando, pp 255–266

Cremer F, den Breejen E, Schutte K (1998) Sensor data fusion for antipersonnel land mine detection. In: Proceedings of EuroFusion98, Great Malvern, pp 55–60

De Bruin S, Bregt A, Van De Ven M (2001) Assessing fitness for use: the expected value of spatial data sets. Int J Geogr Inf Sci 15(5):457–471

Konks D, Challa S (2005) An introduction to Bayesian and Dempster-Shafer data fusion available via DSTO-TR-1436, Edinburgh, Nov 2005. http://www.dsto.defence.gov.au/publications/2563/DSTO-TR-1436.pdf

Mihaila G, Raschid L, Vidal ME (1999) Querying, "Quality of Data" metadata. In: Proceedings of the third IEEE meta-data conference, Bethesda, Apr 1999

Sentz K, Ferson S (2002) Combination of evidence in Dempster-Shafer belief theory. Available via SANDIA technical report SAND2002-0835. http://www.sandia.gov/epistemic/Reports/SAND2002-0835.pdf

Shafer G (1976) A mathematical theory of evidence. Princeton University Press, Princeton

Vasseur B, Devillers R, Jeansoulin R (2003) Ontological approach of the fitness of use of geospatial datasets. In: Proceedings of 6th AGILE conference on geographic information science, Lyon, pp 497–504

Yao X (2003) Research issues in spatiotemporal data mining. A white paper submitted to the University Consortium for Geographic Information Science (UCGIS) workshop on geospatial visualization and knowledge discovery, Lansdowne, 18–20 Nov 2003

Computing Performance

▸ Network GIS Performance

Conceptual Generalization of Databases

▸ Abstraction of Geodatabases

Conceptual Model

▸ Application Schema

Conceptual Modeling

▸ Spatiotemporal Database Modeling with an Extended Entity-Relationship Model

Conceptual Modeling of Geospatial Databases

▸ Modeling with ISO 191xx Standards

Conceptual Neighborhood

Anthony G. Cohn
School of Computing, University of Leeds, Leeds, UK

Synonyms

Closest topological distance; Continuity network; Qualitative similarity

Definition

A standard assumption concerning reasoning about spatial entities over time is that change is continuous. In qualitative spatial calculi, such

as the mereotopological RCC or 9-intersection calculi in which a small finite set of *jointly exhaustive and pairwise disjoint* sets of relations are defined, this can be represented as a *conceptual neighborhood* diagram (also known as a *continuity network*). A pair of relations R_1 and R_2 are conceptual neighbors if it is possible for R_1 to hold at a certain time, and R_2 to hold later, with no third relation holding in between. The diagram to be found in the definitional entry for *mereotopology* illustrates the conceptual neighborhood for RCC-8.

Cross-References

- ▸ Knowledge Representation, Spatial
- ▸ Mereotopology
- ▸ Representing Regions with Indeterminate Boundaries

Recommended Reading

Cohn AG, Hazarika SM (2001) Qualitative spatial representation and reasoning: an overview. Fundam Inf 46(1–2):1–29

Cohn AG, Renz J (2007) Qualitative spatial representation and reasoning. In: Lifschitz V, van Harmelen F, Porter F (eds) Handbook of knowledge representation, Ch. 13. Elsevier, München

Conceptual Schema

- ▸ Application Schema

Concurrency Control for Spatial Access

Jing (David) Dai[1] and Chang-Tien Lu[2]
[1]Google, New York City, NY, USA
[2]Department of Computer Science, Virginia Tech, Falls Church, VA, USA

Synonyms

Concurrency control protocols; Concurrent spatial operations; Simultaneous spatial operations

Definition

Concurrency control for spatial access method refers to the techniques providing the serializable operations in multi-user spatial databases. Specifically, the concurrent operations on spatial data should be safely executed and follow the ACID rules (i.e., Atomicity, Consistency, Isolation, and Durability). With concurrency control, multi-user spatial databases can process the search and update operations correctly without interfering with each other.

The concurrency control techniques for spatial databases have to be integrated with particular spatial access methods to process simultaneous operations. There are two major concerns in concurrency control for the spatial access method. One is how to elevate the throughput of concurrent spatial operations and the other is concerned with preventing phantom access.

Main Text

In the last several decades, spatial data access methods have been proposed and developed to manage multi-dimensional databases as required in GIS, computer-aided design, and scientific modeling and analysis applications. In order to apply the widely studied spatial access methods in real applications, particular concurrency control protocols are required for multi-user environments. The simultaneous operations on spatial databases need to be treated as exclusive operations without interfering with one another.

The existing concurrency control protocols mainly focus on the R-tree family. Most of them were developed based on the concurrency protocols on the B-tree family. Based on the locking strategy, these protocols can be classified into two categories, namely, link-based approaches and lock-coupling methods.

Concurrency control for spatial access methods is generally required in commercial database management systems. In addition, concurrency control methods are required in many specific spatial applications, such as the taxi management

systems that need to continuously query the locations of taxies. The study on spatial concurrency control is far behind the research on spatial query processing approaches. There are two interesting and emergent directions in this field. One is to apply concurrency control methods on complex spatial operations, such as nearest neighbor search and spatial join; the other to design concurrency control protocols for moving object applications.

Cross-References

▸ Indexing, Hilbert R-Tree, Spatial Indexing, Multimedia Indexing

Recommended Reading

Chakrabarti K, Mehrotra S (1999) Efficient concurrency control in multi-dimensional access methods. In: Proceedings of ACM SIGMOD international conference on management of data, Philadelphia

Kornacker M, Mohan C, Hellerstein J (1997) Concurrency and recovery in generalized search trees. In: Proceedings of ACM SIGMOD international conference on management of data, Tucson

Song SI, Kim YH, Yoo JS (2004) An enhanced concurrency control scheme for multidimensional index structure. IEEE Trans Knowl Data Eng 16(1):97–111

Concurrency Control for Spatial Access Method

Jing (David) Dai[1] and Chang-Tien Lu[2]
[1]Google, New York City, NY, USA
[2]Department of Computer Science, Virginia Tech, Falls Church, VA, USA

Synonyms

Concurrency control protocols; Concurrent spatial operations; Phantom update protection; Simultaneous spatial operations

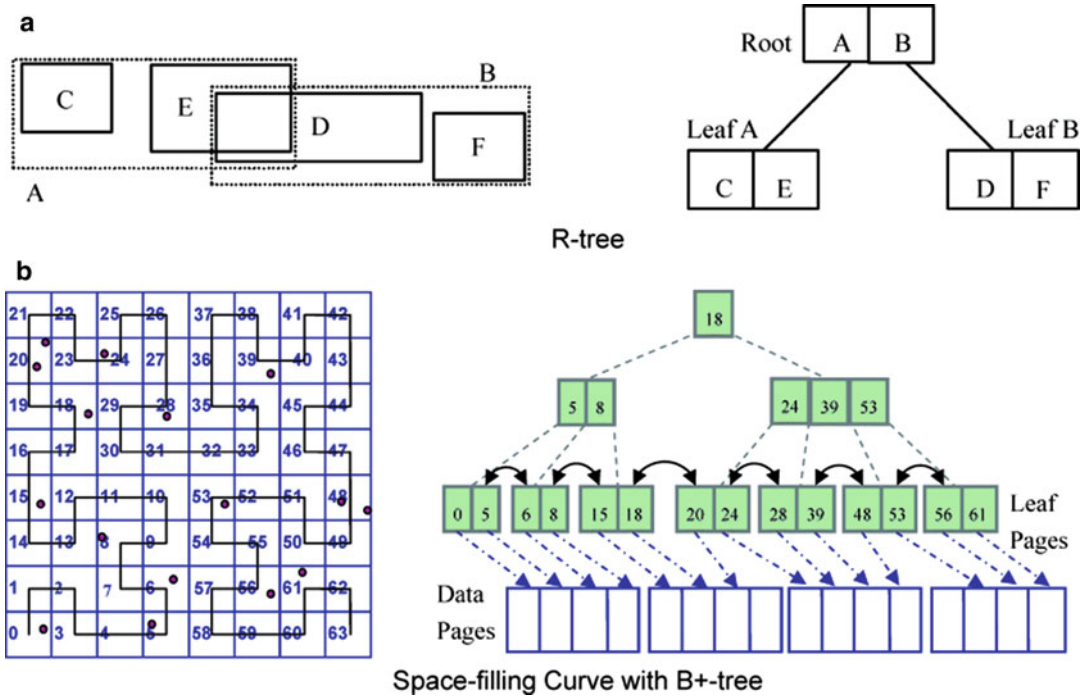
Definition

The concurrency control for spatial access method refers to the techniques providing the serializable operations in multi-user spatial databases. Specifically, the concurrent operations on spatial data should be safely executed and follow the ACID rules (i.e., Atomicity, Consistency, Isolation, and Durability). With concurrency control, multi-user spatial databases can perform the search and update operations correctly without interfering with each other.

There are two major concerns in the concurrency control for spatial data access. One is the throughput of concurrent spatial operations. The throughput refers to the number of operations (i.e., search, insertion, and deletion) that are committed within each time unit. It is used to measure the efficiency of the concurrency control protocols. The other concern is to prevent phantom access. Phantom access refers to the update operation that occurs before the commitment and in the ranges of a search/deletion operation, while not reflected in the results of that search/deletion operation. The ability to prevent phantom access can be regarded as a certain level of consistency and isolation in ACID rules.

Historical Background

In the last several decades, spatial data access methods have been proposed and developed to manage multi-dimensional databases as required in GIS, computer-aided design, and scientific modeling and analysis applications. Representative spatial data access methods are R-trees (Guttman 1984), and space-filling curve with B-trees (Gaede and Gunther 1998). As shown in Fig. 1a, the R-tree groups spatial objects into Minimum Bounding Rectangles (MBR), and constructs a hierarchical tree structure to organize these MBRs. Differently, Fig. 1b shows the space-filling curve which splits the data space into equal-sized rectangles and uses their particular curve (e.g., Hilbert curve) identifications to index the objects in the cells



Concurrency Control for Spatial Access Method, Fig. 1 Representative spatial access methods

into one-dimensional access methods, e.g., the B-tree family.

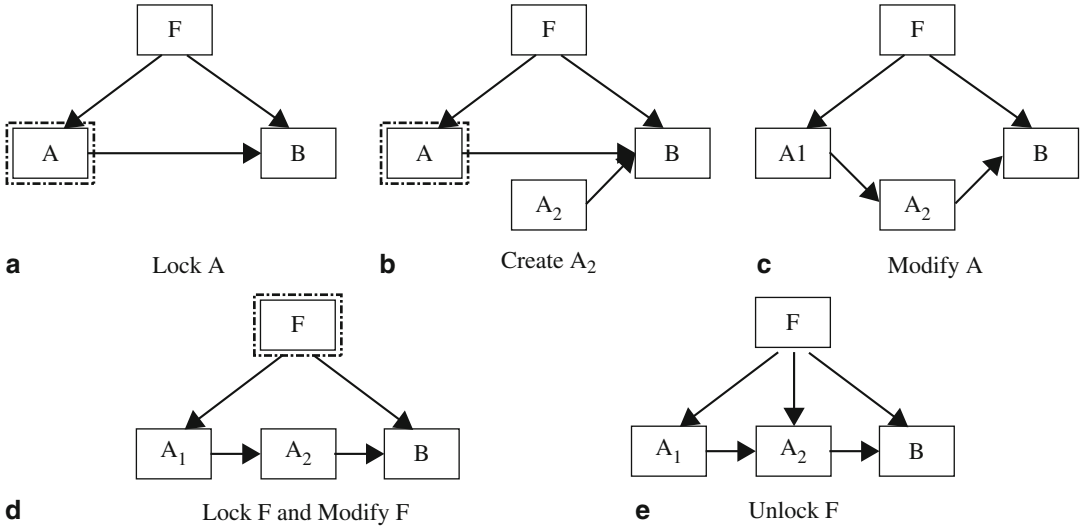
In order to apply the widely studied spatial access methods to real applications, particular concurrency control protocols are required for the multi-user environment. The simultaneous operations on spatial databases need to be treated as exclusive operations without interfering with each other. In other words, the results of any operation have to reflect the current stable snapshot of the spatial database at the commit time.

The concurrency control techniques for spatial databases have to be integrated with spatial access methods to process simultaneous operations. Most of the concurrency control techniques were developed for one-dimensional databases. However, the existing spatial data access methods, such as R-tree family and grid files, are quite different from the one-dimensional data access methods (e.g., overlaps among data objects and among index nodes are allowed). Therefore, the existing concurrency control methods are not suitable for these spatial databases. Furthermore, because the spatial data set usually is not globally ordered, some traditional concurrency control techniques such as link-based protocols are difficult to adapt to spatial databases.

Spatial Concurrency Control Techniques

Since the last decade of the twentieth century, concurrency control protocols on spatial access methods have been proposed to meet the requirements of multi-user applications. The existing concurrency control protocols mainly focus on the R-tree family, and most of them were developed based on the concurrency protocols on the B-tree family. Based on the locking strategy, these protocols can be classified into two categories, namely, link-based methods and lock-coupling methods.

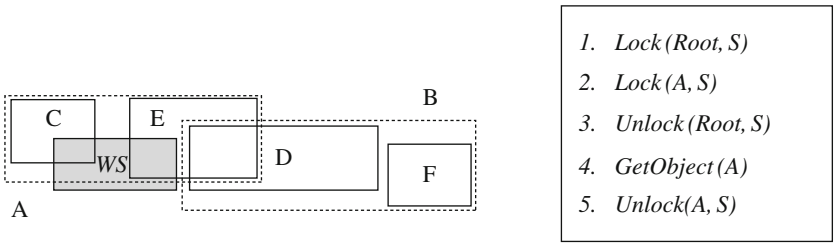
The link-based methods rely on a pseudo global order of the spatial objects to isolate each concurrent operation. These approaches process update operations by temporally disabling the links to the indexing node being updated so that the corresponding search operations will not retrieve any inconsistent data. For instance, to split node A into A_1 and A_2 in Fig. 2, a lock



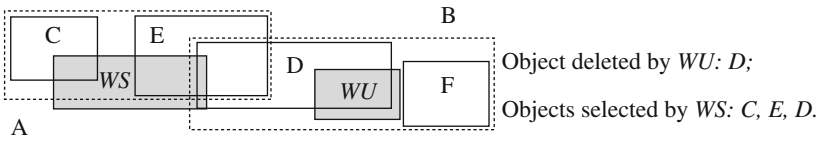
Concurrency Control for Spatial Access Method, Fig. 2 Example of node split in link-based protocol

will be requested to disable the link from A to its right sibling node B (step a) before the actual split is performed. Then, a new node A_2 will be created in step b by using the second half of A, and linked to node B. In step c, A will be modified to be A_1 (by removing the second half), and then unlocked. Node F will be locked before adding a link from F to A_2 in step d. Finally, F will be unlocked in step e, and thus the split is completed. Following this split process, no search operations can access A_2, and no update operations can access $A(A_1)$ before step c. Therefore, the potential confliction caused by concurrent update operations on node A can be prevented. As one example of the link-based approach, R-link tree, a right-link style algorithm (Kornacker and Banks 1995), has been proposed to protect concurrent operations by assigning logical sequence numbers (LSNs) on the nodes of R-trees. This approach assures each operation has at most one lock at a time. However, when a propagating node splits and the MBR updates, this algorithm uses lock coupling. Also, in this approach, additional storage is required to maintain additional information, e.g., LSNs of associated child nodes. Concurrency on the Generalized Search Tree (CGiST) (Kornacker et al. 1997) protects concurrent operations by applying a global sequence number, the Node Sequence Number (NSN). The counter of the NSN is incremented in a node split and a new value is assigned to the original node with the new sibling node receiving the original node's prior NSN and right-link pointer. In order for an insert operation to execute correctly in this algorithm, multiple locks on two or more levels must be held. Partial lock coupling (PLC) (Song et al. 2004) has been proposed to apply a link-based technique to reduce query delays due to MBR updates for multi-dimensional index structures. The PLC technique provides high concurrency by using lock coupling only in MBR shrinking operations, which are less frequent than expansion operations.

The lock-coupling-based algorithms (Chen et al. 1997; Ng and Kamada 1993) release the lock on the current node only when the lock on the next node to be visited has been granted while processing search operations. As shown in Fig. 3, using the R-tree in Fig. 1a, suppose objects C, E, D, and F are indexed by an R-tree with two leaf nodes A and B. A search window WS can be processed using the lock-coupling approach. The locking sequence in Fig. 3 can protect this search operation from reading the intermediate results of update operations as well as the results of update operations submitted after WS. During node splitting and MBR updating,



Concurrency Control for Spatial Access Method, Fig. 3 Example of locking sequence using lock-coupling for WS



Concurrency Control for Spatial Access Method, Fig. 4 Example of phantom update

this scheme holds multiple locks on several nodes simultaneously. The dynamic granular locking approach (DGL) has been proposed to provide phantom update protection (discussed later) in R-trees (Chakrabarti and Mehrotra 1998) and GiST (Chakrabarti and Mehrotra 1999). The DGL method dynamically partitions the embedded space into lockable granules that can adapt to the distribution of the objects. The lockable granules are defined as the leaf nodes and external granules. External granules are additional structures that partition the non-covered space in each internal node to provide protection. Following the design principles of DGL, each operation requests locks only on sufficient granules to guarantee that any two conflicting operations will request locks on at least one common granule.

Scientific Fundamentals

ACID Rules

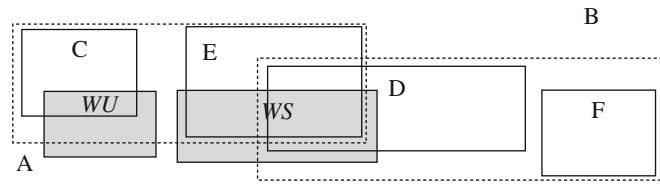
Concurrency control for spatial access methods should assure the spatial operations are processed following the ACID rules (Ramakrishnan and Gehrke 2001). These rules are defined as follows.

- Atomicity – Either all or no operations are completed.
- Consistency – All operations must leave the database in a consistent state.
- Isolation – Operations cannot interfere with each other.
- Durability – Successful operations must survive system crashes.

The approaches to guarantee Atomicity and Durability in traditional databases can be applied in spatial databases. Current research on spatial concurrency control approaches mainly focus on the Consistency and Isolation rules. For example, in order to retrieve the valid records, spatial queries should not be allowed to access the intermediate results of location updates. Similarly, the concurrent location updates with common coverage have to be isolated as sequential execution; otherwise, they may not be processed correctly.

Phantom Update Protection

In addition to the ACID rules, phantom update protection is used to measure the effectiveness of a concurrency control. An example of phantom update is illustrated in Fig. 4, where C, E, D, and F are objects indexed in an R-tree, and leaf nodes A, B are their parents, respectively. A deletion with the window WU is completed before the commitment of the range query WS. The range query returns the set $\{C, E, D\}$, even object D



Concurrency Control for Spatial Access Method, Fig. 5 Example of efficient concurrency control

should have been deleted by *WU*. A solution to prevent phantom update in this example is to lock the area affected by *WU* (which is $D \cup WU$) in order to prevent the execution of *WS*.

Measurement

The efficiency of concurrency control for spatial access methods is measured by the throughput of concurrent spatial operations. The issue to provide high throughput is to reduce the number of unnecessary conflicts among locks. For the example shown in Fig. 5, even if the update operation with window *WU* and the range query with window *WS* intersect with the same leaf node *A*, they will not affect each other's results. Therefore, they should be allowed to access *A* simultaneously. Obviously, the smaller the lockable granules, the more concurrency operations will be allowed. However, this may significantly increase the number of locks in the database, and therefore generate additional overhead on lock maintenance. This is a tradeoff that should be considered when designing concurrency control protocols.

Key Applications

Concurrency control for spatial access methods are generally required in commercial multidimensional database systems. These systems are designed to provide efficient and reliable data access. Usually, they are required to reliably handle a large amount of simultaneous queries and updates. Therefore, sound concurrency control protocols are required in these systems.

In addition, concurrency control methods are required in many specific spatial applications which have frequent updates or need fresh query results. For instance, a mobile advertise/alarm system needs to periodically broadcast time-sensitive messages to cell phone users within a certain range. Concurrency control methods should be employed to protect the search process from frequent location updates, because the updates are not supposed to reveal their intermediate or expired results to the search process. Another example is a taxi management system that needs to assign a nearest available taxi based on a client's request. Concurrency control methods need to be applied to isolate the taxi location updating and queries so that the query results are consistent to the up-to-date snapshot of the taxi locations.

Future Directions

The study on spatial concurrency control is far behind the research on spatial query processing approaches. There are two interesting and emergent directions in this field. One is to apply concurrency control methods on complex spatial operations; the other is to design concurrency control protocols for moving object applications.

Complex spatial operations, such as spatial join, k-nearest neighbor search, range nearest neighbor search, and reverse nearest neighbor search, require special concern on concurrency control to be applied in multi-user applications. For example, how to protect the changing search range, and how to protect the large overall search range have to be carefully designed. Furthermore, the processing methods of those complex operations may need to be redesigned based on the concurrency control protocol in order to improve the throughput.

Spatial applications with moving objects have attracted significant research efforts. Even though many of these applications assume that the query processing is based on main memory, their frequent data updates require

sophisticated concurrency control protocols to assure the correctness of the continuous queries. In this case, concurrency access framework will be required to support the frequent location updates of the moving objects. Frequent update operations usually result in a large number of exclusive locks which may significantly degrade the throughput. Solutions to improve the update speed and reduce the coverage of operations have to be designed to handle this scenario.

Cross-References

▸ Indexing, Hilbert R-tree, Spatial Indexing, Multimedia Indexing

References

Chakrabarti K, Mehrotra S (1998) Dynamic granular locking approach to phantom protection in R-trees. In: Proceedings of IEEE international conference on data engineering, Orlando

Chakrabarti K, Mehrotra S (1999) Efficient concurrency control in multi-dimensional access methods. In: Proceedings of ACM SIGMOD international conference on management of data, Philadelphia

Chen JK, Huang YF, Chin YH (1997) A study of concurrent operations on R-trees. Inf Sci Int J 98(1–4):263–300

Gaede V, Gunther O (1998) Multidimensional access methods. ACM Comput Surv 30(2):170–231

Guttman A (1984) R-trees: a dynamic index structure for spatial searching. In: Proceedings of ACM SIGMOD international conference on management of data, Boston

Kornacker M, Banks D (1995) High-concurrency locking in R-trees. In: Proceedings of international conference on very large data bases, Zurich

Kornacker M, Mohan C, Hellerstein J (1997) Concurrency and recovery in generalized search trees. In: Proceedings of ACM SIGMOD international conference on management of data, Tucson

Ng V, Kamada T (1993) Concurrent accesses to R-trees. In: Proceedings of symposium on advances in spatial databases, Singapore

Ramakrishnan R, Gehrke J (2001) Database management systems, 2nd edn. McGraw-Hill, New York

Song SI, Kim YH, Yoo JS (2004) An enhanced concurrency control scheme for multidimensional index structure. IEEE Trans Knowl Data Eng 16(1): 97–111

Concurrency Control Protocols

▸ Concurrency Control for Spatial Access
▸ Concurrency Control for Spatial Access Method

Concurrent Processing

▸ Spatial Data Analytics on Homogeneous Multi-Core Parallel Architectures

Concurrent Spatial Operations

▸ Concurrency Control for Spatial Access
▸ Concurrency Control for Spatial Access Method

Conditional Spatial Regression

▸ Spatial and Geographically Weighted Regression

Conflation

▸ Ontology-Based Geospatial Data Integration
▸ Positional Accuracy Improvement (PAI)

Conflation of Features

Sharad Seth and Ashok Samal
Department of Computer Science and Engineering, The University of Nebraska at Lincoln, Lincoln, NE, USA

Synonyms

Automated Map Compilation; Data Integration; Entity Integration; Feature Matching; Realignment; Rubber-Sheeting; Vertical Conflation

Definition

In GIS, conflation is defined as the process of combining geographic information from overlapping sources so as to retain accurate data, minimize redundancy, and reconcile data conflicts (Longley et al. 2001). The need for conflation typically arises in updating legacy data for accuracy or missing features/attributes by reference to newer data sources with overlapping coverage. For example, the street-name and address-range data from the US Census Bureau can be conflated with the spatially accurate USGS digital-line-graph (DLG) to produce a more accurate and useful source than either dataset. Conflating vector GIS data with raster data is also a common problem.

Conflation can take many different forms. Horizontal conflation refers to the matching of features and attributes in adjacent GIS sources for the purpose of eliminating positional and attribute discrepancies in the common area of the two sources. Vertical conflation solves a similar problem for GIS sources with overlapping coverage. As features are the basic entities in a GIS, the special case of feature conflation has received much attention in the published research. The data used for conflation are point, line, and area features and their attributes. Figure 1 illustrates a problem solved by feature conflation. The first two GIS data layers show a digital ortho-photo and a topographic map of the Mall area in Washington D.C. In the third layer on the right, showing an overlay of the two sources, the corresponding features do not exactly line up. With conflation, these discrepancies can be minimized, thus improving the overall accuracy of the data sources.

Historical Background

Until the 1980s, the collection of geographical information in digital form was expensive enough that having multiple sources of data for the same region, as required for conflation, was possible only for large governmental organizations. It is not surprising, therefore, that early use of conflation was initiated by governmental agencies. The applications related to the automation of *map compilation* for transferring positional information from a *base map* to a non-georeferenced target map. An iterative process, called alignment or rubber-sheeting, was used to bring the coordinates of the two maps into mutual consistency. The latter term alludes to stretching the target map that is printed on a rubber sheet so as to align it with the base map at all points. Although contemplated many years earlier (White 1981), the first semi-automated systems for alignment came into existence only in the mid-1980s. These interactive systems were screen-based and image-driven (Lynch and Saalfield 1985). The operator was allowed, and even assisted, to select a pair of intersections to be matched. With each additional selected pair, the two maps were brought into closer agreement.

Fully automated systems, later developed in a joint project between the US Geological Society (USGS) and the Bureau of Census, aimed at consolidating the agencies' 5,700 pairs of metropolitan map sheet files (Saalfield 1988). This development was facilitated by parallel advances in computer graphics devices, computational-geometry algorithms, and statistical pattern recognition. Automation of the process required replacing the operator's skills at discerning like features with an analogous feature-matching algorithm on the computer. Alignment may be thought of as a mathematical transformation of one image that preserves topology. A single global transformation may be insufficient to correct errors occurring due to local distortions, thus necessitating local alignments for different regions of the image. Delauney triangulation defined by selected points is preferred for rubber-sheeting because it minimizes the formation of undesirable thin, long triangles.

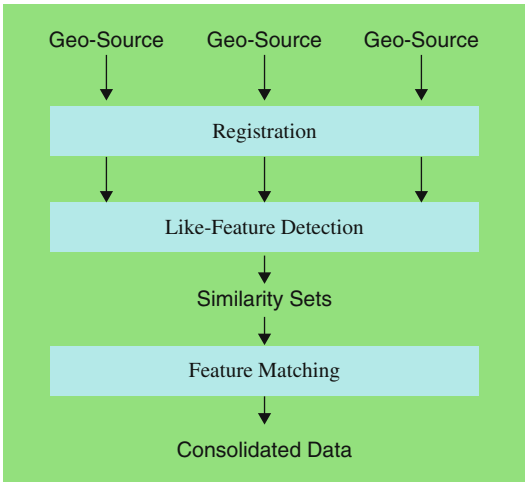
Early work in feature conflation was based on proximity of features in non-hierarchical ("flattened") data sources. Because GIS data are typically organized into a hierarchy of classes and carry much information that is not position-related, such as, names, scalar quantities, and

Conflation of Features, Fig. 1 Two GIS data layers for Washington DC and their overlay

geometrical shapes, the methods used to discover identical objects can go beyond proximity matches and include rule-based approaches (Cobb et al. 1998), string matching, and shape similarity (Samal et al. 2004).

Scientific Fundamentals

A prototype of a feature conflation system is shown in Fig. 2. Such a system would typically form the back end of a geographic information and decision-support system used to respond to user queries for matched features. Some systems may not implement all three steps while others may further refine some of the steps, e.g., like-feature-detection may be split into two steps that either use or ignore the geographical context of the features during comparison. Further details of the basic steps appear in the following sections.



Conflation of Features, Fig. 2 Feature conflation steps

Registration and Rectification

Registration refers to a basic problem in remote sensing and cartography of realigning a recorded digital image with known ground truth or another image. An early survey in geographic data processing (Nagy and Wagle 1979) formulates the registration problem in remote sensing as follows:

The scene under observation is considered to be a 2D intensity distribution $f(x, y)$. The recorded digital, another 2-D distribution $g(u, v)$, is related to the "true" scene $f(x, y)$ through an unknown transformation T:

$$g(u, v) = T(f(x, y)).$$

Thus, in order to recover the original information from the recorded observations, we must first determine the nature of the transformation T, and then execute the inverse operation T^{-1} on this image.

Often, because only indirect information is available about T, in the form of another image or map of the scene in question, the goal of registration becomes finding a mathematical transformation on one image that would bring it into concurrence with the other image. Geometric

distortions in the recorded image, which affect only the position and not the magnitude, can be corrected by a rectification step that only transforms the coordinates.

Like-Feature Detection

The notion of similarity is fundamental to matching features, as it is to many other fields, including, pattern recognition, artificial intelligence, information retrieval, and psychology. While the human view of similarity may be subjective, automation requires objective (quantitative) measures.

Similarity and distance are complementary concepts. It is often intuitively appealing to define a distance function $d(A, B)$ between objects A and B in order to capture their dissimilarity and convert it to a normalized similarity measure by its complement:

$$s(A, B) = 1 - \frac{d(A, B)}{U}, \qquad (1)$$

where the normalization factor U may be chosen as the maximum distance between any two objects that can occur in the data set. The normalization makes the value of similarity a real number that lies between zero and one.

Mathematically, any distance function must satisfy the properties if minimality ($d(a,b) \geq d(a,a) \geq 0$), symmetry ($d(a,b) = d(b,a)$), and triangular inequality ($d(a,b) + d(b,c) \geq d(a,c)$). However, in human perception studies, the distance function must be replaced by the "judged distance" for which all of these mathematical axioms have been questioned (Santini and Jain 1999). Tversky (1977) follows a set-theoretic approach in defining similarities between two objects as a function of the attributes that are shared by the two or by one but not the other. His definition is not required to follow any of the metric axioms. It is particularly well suited to fuzzy attributes with discrete overlapping ranges of values.

In GIS, the two objects being compared often have multiple attributes, such as name, location, shape, and area. The name attribute is often treated as a character string for comparison using the well-known Hamming or Levenshtein metric aimed at transcription errors. An alternative measure of string comparison, based on their phonetic representation (Hall and Dowling 1980), may be better suited to transcription errors. However, the names are often word phrases that may look very different as character strings, but connote the same object, e.g., "National Gallery of Art" and "National Art Gallery". Table 1 (Samal et al. 2004) shows the type of string errors that string matching should accommodate:

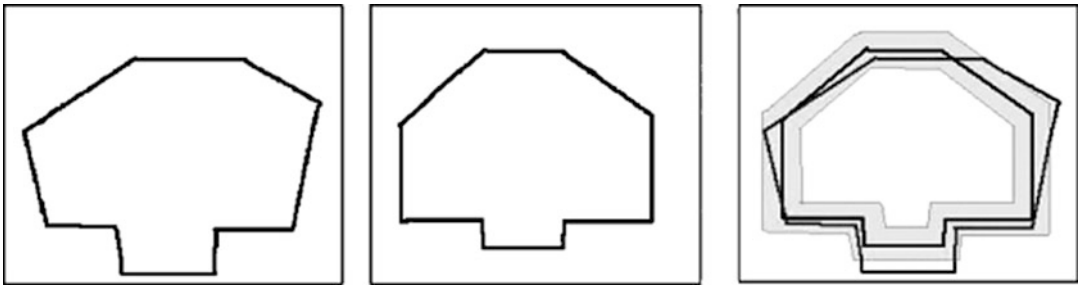
For locations or points, the Euclidean distance is commonly used for proximity comparison. A generalization to linear features, such as streets or streams, is the Hausdorff distance, which denotes the largest minimum distance between the two linear objects. Goodchild and Hunter (1997) describe a less computer-intensive and robust method that relies on comparing two representations with varying accuracy. It estimates the percentage of the total length of the low-accuracy representation that is within a specified distance of the high-accuracy representation.

The shape is an important attribute of polygonal features in GIS, such as building outlines and region boundaries. As polygons can be regarded as linear features, the Goodchild and Hunter approach may be adapted to define shape comparison. A two-step process is described for this purpose by Samal et al. (2004). First, a veto is imposed if the aspect ratios are significantly different. Otherwise, the shapes are scaled to match the lengths of their major axes and overlaid by aligning their center points. The similarity of a less accurate shape A to a more accurate shape B is the percentage A within the buffer zone of B (see Fig. 3). When the accuracy of the two sources is comparable, the measure could be taken as the average value of the measure computed both ways.

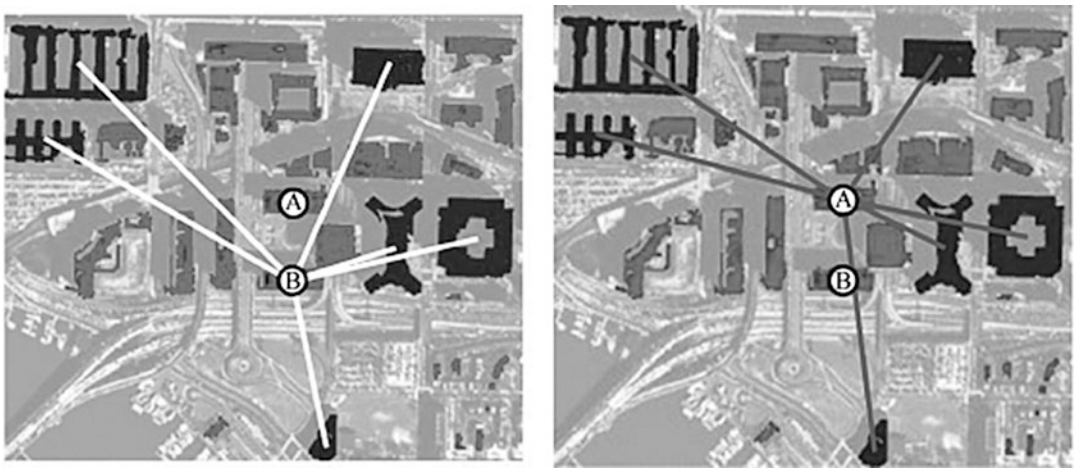
Comparing scalars (reals or integers) seems to be straightforward: take the difference as their distance and convert to similarity by using Eq. (1). The normalization factor U, however, must be chosen carefully to match intuition. For example, one would say that the pair of numbers 10 and 20 is less similar to the pair 123,010 and 123020, even though the difference is the

Conflation of Features, Table 1 Typical string errors and differences that matching should accommodate

Error type	Examples	
	Sample 1	Sample 2
Word omission	Abraham Lincoln Memorial	Lincoln Memorial
Word substitution	Reagan National Airport	Washington National Airport
Word transposition	National Art Gallery	National Gallery of Art
Word abbreviation	National Archives	Nat'l Archives
Character omission	Washington Monument	Washington Monument
Character substitution	Frear Gallery	Freer Gallery



Conflation of Features, Fig. 3 Two polygons and their buffered intersection



Conflation of Features, Fig. 4 Two similar features and their geographic contexts

same in both cases. Hence, the normalization factor should be equated with the magnitude of the range of values defined for scalars a and b.

The hierarchical nature of GIS data makes it possible to also assess the similarities of two objects along their categorical structure. A knowledge base of spatial categories can be built using Wordnet (Fellbaum 1998) and Spatial Data Transfer Standard (SDTS) (Rodriguez and Egenhofer 2003).

Context

Clearly, context plays an important role in the human perception of similarity. The similarity measures described above, however, are *context-independent*: two features are compared in isolation without reference to other features on their respective sources. Context-independent similarity measures alone are not always sufficient enough to determine feature matches unambiguously, necessitating the use of some form of context to resolve such cases.

The *geographic context* is defined as the spatial relationships between objects in an area (Samal et al. 2004). Examples of such relationships include topologies, distances, and directions. Topological relationships, such as disjoint, meet, overlap, and covering are used by the researchers at the University of Maine to model the context of areal features in applications involving query by sketch and similarity of spatial scenes (Bruns and Eggenhofer 1996). Distances and angles of a feature to other features have also been used to represent the geographic context (Samal et al. 2004). Figure 4 shows the geographic contexts of two nearby features with similar shapes. The contexts can be seen to be different enough to disambiguate these two features when they are compared with a candidate feature in another source. Further, to keep the cost of context-dependent matching under control, it may be enough to define the geographic context with respect to only a small number of well chosen landmark features.

Feature Matching

The similarity measures discussed above for individual attributes of a feature must be combined in some fashion to provide overall criteria for feature matching. According to Cobb et al. (1998), "The assessment of feature match criteria is a process in which evidence must be evaluated and weighed and a conclusion drawn – not one in which equivalence can be unambiguously determined … after all, if all feature pairs matched exactly, or deviated uniformly according to precise processes, there would be no need to conflate the maps!"

The problem can be approached as a restricted form of the classification problem in pattern recognition: Given the evidence provided by the similarity scores of different attributes of two features, determine the likelihood of one feature belonging to the same class as the other feature. Because of this connection, it is not surprising that researchers have used well-known techniques from pattern recognition to solve the feature matching problem. These techniques include clustering (Baraldi and Blonda 1999) and fuzzy logic (Zadeh 1965).

For example, similarity of two buildings appearing in different GIS data layers (as in Fig. 1a, b) could be established by comparing their individual attributes, such as shape and coordinates. These context-independent measures, however, may not be sufficient and it may become necessary to use the geographical context to resolve ambiguities or correct errors.

Key Applications

Coverage Consolidation:	Data gathering is the most expensive part of building a geographical information system (GIS). In traditional data gathering, this expense is directly related to the standards of rigor used in data collection and data entry. Feature conflation can reduce the cost of GIS data acquisition by combining inexpensive sources into a superior source. With the widespread use of the Web and GPS, the challenge in consolidation is shifting from improving accuracy to integrating an abundance of widely distributed sources by automated means
Spatial data update:	By identifying common and missing features between two sources through feature conflation, new features can be added to an old source or their attributed updated from a newer map
Coverage registration:	Non-georeferenced spatial data must be registered before it can be stored in a GIS. Good registration requires choosing a number of features for which accurate geo-positional information is available and which are also spatially accurate on the source. Spatial data update can help in identifying good candidate features for registration
Error detection:	Feature conflation can not only tell which features in two sources are alike, but also provide a degree of confidence for these assertions. The pairs with low confidence can be checked manually for possible errors

Future Directions

Conflation in GIS can be thought of as part of the broader problems in the information age of searching, updating, and integration of data.

Because the sense of place plays such an important role in our lives, all kinds of non-geographical data related to history and culture can be tied to a place and thus become a candidate for conflation. In this view, geographical reference becomes a primary key used by search engines and database applications to consolidate, filter, and access the vast amount of relevant data distributed among many data sources. The beginnings of this development can already be seen in the many applications already in place or envisaged for Google Earth and other similar resources. If the consolidated data remains relevant over a period of time and finds widespread use, it might be stored and used as a new data source, much in the same fashion as the results of conflation are used today.

The traditional concern in conflation for positional accuracy will diminish in time with the increasing penetration of GPS in consumer devices and the ready availability of the accurate position of all points on the earth. The need for updating old data sources and integrating them with new information, however, will remain an invariant.

Cross-References

- ▸ Conflation of Geospatial Data
- ▸ Geospatial Semantic Integration
- ▸ Ontology-Based Geospatial Data Integration

References

Baraldi A, Blonda P (1999) A survey of fuzzy clustering algorithms for pattern recognition. IEEE Trans Syst Man Cybern I II 29(6):778–785

Bruns H, Eggenhofer M (1996) Similarity of spatial scenes. In: Molenaar M, Kraak MJ (eds) Proceedings of the 7th international symposium on spatial data handling. Taylor and Francis, London, pp 31–42

Cobb M, Chung MJ, Foley H III, Petry FE, Shaw KB (1998) A rule-based approach for the conflation of attributed vector data. Geoinformatica 2(1):7–35

Fellbaum C (ed) (1998) Wordnet: an electronic lexical database. MIT, Cambridge

Goodchild MF, Hunter GJ (1997) A simple positional accuracy measure for linear features. Int J Geogr Inf Sci 11(3):299–306

Hall P, Dowling G (1980) Approximate string matching. ACM Comput Surv 12(4):381–402

Longley PA, Goodchild MF, Maguire DJ, Rhind DW (2001) Geographic information systems and science. Wiley, Chichester

Lynch MP, Saalfield A (1985) Conflation: automated map compilation – a video game approach. In: Proceedings of the auto-carto 7 ACSM/ASP, Falls Church, 11 Mar 1985

Nagy G, Wagle S (1979) Geographic data processing. ACM Comput Surv 11(2):139–181

Rodriguez A, Egenhofer M (2003) Determining semantic similarity among entity classes from different ontologies. IEEE Trans Knowl Data Eng 15:442–456

Saalfield A (1988) Conflation: automated map compilation. Int J GIS 2(3):217–228

Samal A, Seth S, Cueto K (2004) A feature-based approach to conflation of geospatial sources. Int J GIS 18(5):459–589

Santini S, Jain R (1999) Similarity measures. IEEE Trans Pattern Anal Mach Intell 21(9):87–883

Tversky A (1977) Features of similarity. Psychol Rev 84:327–352

White M (1981) The theory of geographical data conflation. Internal Census Bureau draft document

Zadeh LA (1965) Fuzzy sets. Inf Control 8:338–353

Recommended Reading

Rodriguez A, Egenhofer M (1999) Assessing similarity among geospatial feature class definitions. In: Vckorski A, Brasel KE, Schek HJ (eds) Lecture notes in computer science, vol 1580. Springer, Berlin, pp 189–202

Conflation of Geospatial Data

Ching-Chien Chen[1] and Craig A. Knoblock[2]
[1]Geosemble Technologies, El Segundo, CA, USA
[2]Department of Computer Science, University of Southern California, Marina del Rey, CA, USA

Synonyms

Computer cartography; Geospatial data alignment; Geospatial data reconciliation; Imagery conflation

Definition

Geospatial data conflation is the compilation or reconciliation of two different geospatial datasets covering overlapping regions (Saalfeld 1988). In general, the goal of conflation is to combine the best quality elements of both datasets to create a composite dataset that is better than either of them. The consolidated dataset can then provide additional information that cannot be gathered from any single dataset.

Based on the types of geospatial datasets dealt with, the conflation technologies can be categorized into the following three groups:

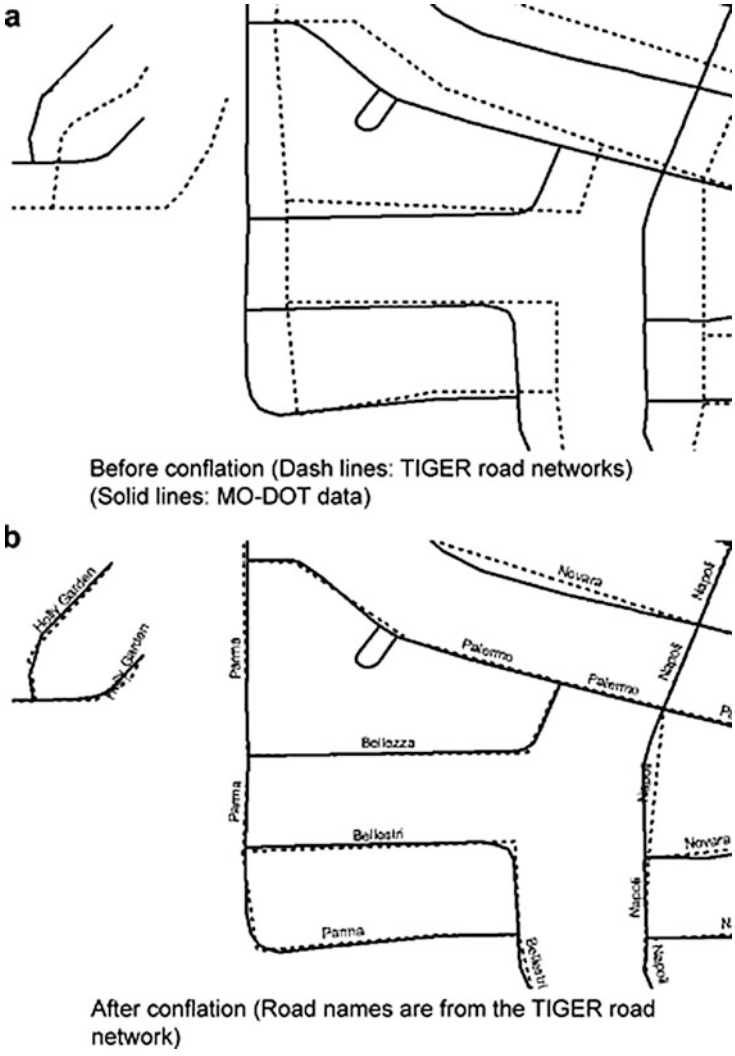
- Vector to vector data conflation: A typical example is the conflation of two road networks of different accuracy levels. Figure 1 shows a concrete example to produce a superior dataset by integrating two road vector datasets: road network from US Census TIGER/Line files, and road network from the department of transportation, St. Louis, MO (MO-DOT data).
- Vector to raster data conflation: Fig. 2 is an example of conflating a road vector dataset with a USGS 0.3 m per pixel color image. Using the imagery as the base dataset for position, the conflation technique can correct the vector locations and also annotate the image with appropriate vector attributes (as Fig. 2b).
- Raster to raster data conflation: Fig. 3 is an example of conflating a raster street map (from MapQuest) with a USGS image. Using the imagery as the base dataset for position, the conflation technique can create intelligent images that combine the visual appeal and accuracy of imagery with the detailed attributes often contained in maps (as Fig. 3b).

Also note that although the examples shown in Figs. 1, 2, and 3 are the conflation of datasets covering the same region (called vertical conflation), the conflation technologies can also be applied to merge adjacent datasets (called horizontal conflation).

Historical Background

For a number of years, significant manual effort has been required to conflate two geospatial datasets by identifying features in two datasets that represent the same real-world features, then aligning spatial attributes and non-spatial attributes of both datasets. Automated vector and vector conflation was first proposed by Saalfeld (1988), and the initial focus of conflation was using geometrical similarities between spatial attributes (e.g., location, shape, etc.) to eliminate the spatial inconsistency between two overlapping vector maps. In particular, in Saalfeld (1988), Saalfeld discussed mathematical theories to support the automatic process. From then, various vector to vector conflation techniques have been proposed (Walter and Fritsch 1999; Ware and Jones 1998) and many GIS systems (such as Conflex (http://www.digitalcorp.com/conflex.htm)) have been implemented to achieve the alignments of geospatial datasets. More recently, with the proliferation of attributed vector data, attribute information (i.e., non-spatial information) has become another prominent feature used in the conflation systems, such as ESEA MapMerger (http://www.esea.com/products/) and the system developed by Cobb et al. (1998).

Most of the approaches mentioned above focus on vector to vector conflation by adapting different techniques to perform the matching. However, due to the rapid advances in remote sensing technology from the 1990s to capture high resolution imagery and the ready accessibility of imagery over the Internet, such as Google Maps (http://maps.google.com/) and Microsoft TerraService (http://terraservice.net/), the conflation with imagery (such as *vector to imagery* conflation, *imagery to imagery* conflation and *raster map to imagery* conflation) has become one of the central issues in GIS. The objectives of these imagery-related conflation are, of course, to take full advantages of updated high resolution imagery to improve out-of-date GIS data and to display the ground truth in depth with attributes inferred from other data sources (as the examples shown in Figs. 2b and 3b). Due to



Conflation of Geospatial Data, Fig. 1 An example of vector to vector conflation

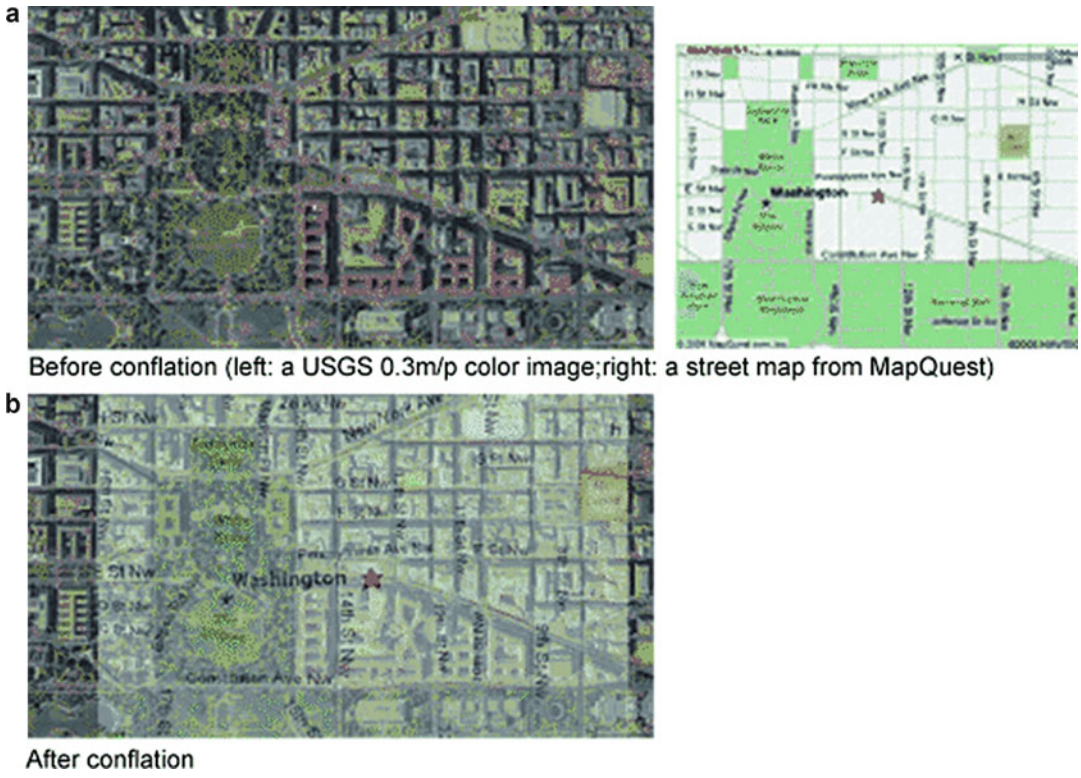
the natural characteristics of imagery (or, more generally, geospatial raster data), the matching strategies used in conflation involve more image-processing or pattern recognition technologies. Some proposed approaches (Cobb et al. 1998; Flavie et al. 2000) rely on edge detections or interest-point detections to extract and convert features from imagery to vector formats, and then apply *vector to vector* conflation to align them. Other approaches (Agouris et al. 2001; Chen et al. 2006a; Eidenbenz et al. 2000), however; utilize the existing vector data as prior knowledge to perform a vector-guided image processing. Conceptually, the spatial information on the vector data represents the existing knowledge about the approximate location and shape of the counterpart elements in the image, thus improving the accuracy and running time to detect matched features from the image. Meanwhile, there are also numerous research activities (Chen et al. 2004a; Seedahmed and Martucci 2002; Dare and Dowman 2000) focusing on conflating different geospatial raster datasets. Again, these approaches perform diverse image-processing techniques to detect and match counterpart elements, and then geometrically align these raster datasets so that the respective pixels or their derivatives (edges, corner point, etc.) representing the same underlying spatial structure are fused.

Conflation of Geospatial Data, Fig. 2 An example of vector to raster data conflation (Modified figure from Chen et al. 2006a)

Today, with the popularity of various geospatial data, automatic geospatial data conflation is rather an area of active research. Consequently, there are various commercial products, such as MapMerger and Conflex, supporting automatic *vector to vector* data conflation with limited human intervention. However, there are no commercial products to provide automatic *vector to raster* or *raster to raster* conflation.

Scientific Fundamentals

A geospatial data conflation system requires efficient and robust geometric and statistical algorithms, and image processing and pattern recognition techniques to implement a rather broad spectrum of mathematical theories. The framework of conflation process can be generalized into the following steps: (1) Feature matching: Find a set of conjugate point pairs, termed control



Conflation of Geospatial Data, Fig. 3 An example of raster map to imagery conflation (Modified figure from Chen et al. 2004a)

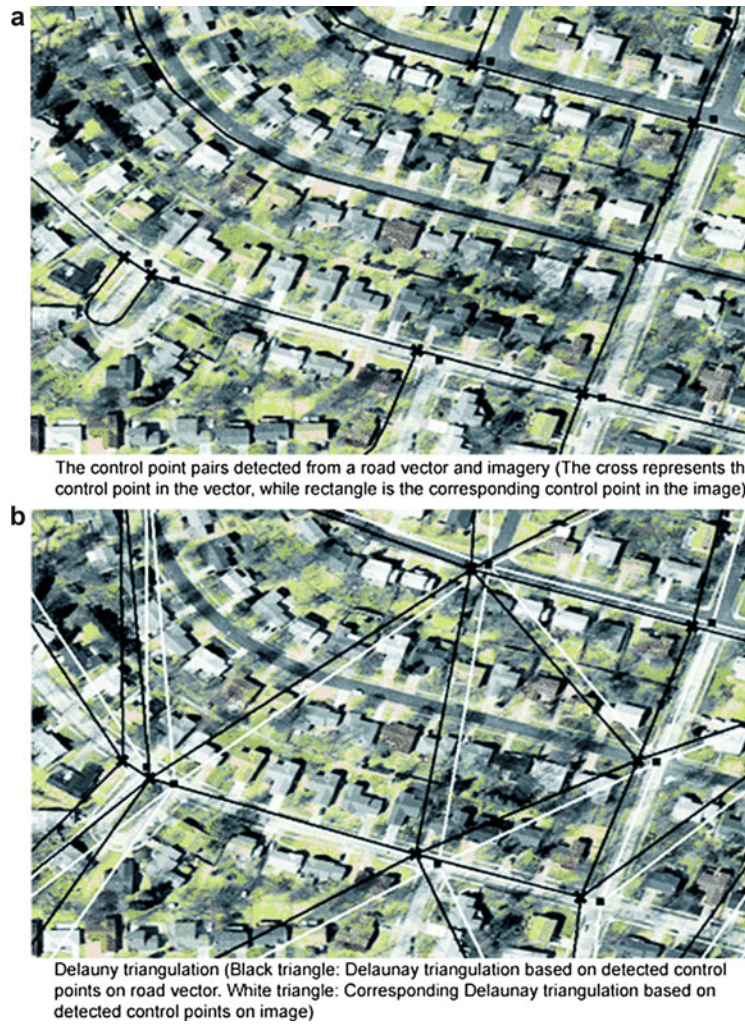
point pairs, in two datasets, (2) Match checking: Filter inaccurate control point pairs from the set of control point pairs for quality control, and (3) Spatial attribute alignment: Use the accurate control points to align the rest of the geospatial objects (e.g., points or lines) in both datasets by using space partitioning techniques (e.g., triangulation) and geometric interpolation techniques.

During the late 1980s, Saalfeld (1988) initialized the study to automate the conflation process. He provided a broad mathematical context for conflation theory. In addition, he proposed an iterative conflation paradigm based on the above-mentioned conflation framework by repeating the matching and alignment, until no further new matches are identified. In particular, he investigated the techniques to automatically construct the influence regions around the control points to reposition other features into alignment by appropriate local interpolation (i.e., to automate the third step in the above-mentioned conflation framework). The conclusion of Saalfeld's work is that Delaunay triangulation is an effective strategy to partition the domain space into triangles (influence regions) to define local adjustments (see the example in Fig. 4). A Delaunay triangulation is a triangulation of the point set with the property that no point falls in the interior of the circumcircle of any triangle (the circle passing through the three triangle vertices). The Delaunay triangulation maximizes the minimum angle of all the angles in the triangulation, thus avoiding elongated, acute-angled triangles. The triangle vertices (i.e., control points) of each triangle define the local transformation within each triangle to reposition other features. The local transformation used for positional interpolation is often the affine transformation, which consists of a linear transformation (e.g., rotation and scaling) followed by a translation. An

affine transformation can preserve collinearity and topology. The well-known technique, rubber-sheeting (imagine stretching a dataset as if it were made of rubber), typically refers to the process comprising triangle-based space partition and the transformation of features within each triangle.

What Saalfeld discovered had a profound impact upon conflation techniques. From then on, the rubber-sheeting technique (with some variants) is widely used in conflation algorithms, because of the sound mathematical theories and because of its success in many practical examples. In fact, these days, most of commercial conflation products support the piecewise rubber-sheeting. Due to the fact that rubber-sheeting has become commonly known strategy to geometrically merge datasets based on the control points, many algorithms have been invented around this conflation paradigm with a major focus on solving the matching (correspondence) problem to find accurate control point pairs (i.e., to automate the first two steps in the above-mentioned conflation framework). However, feature matching algorithms differ with the types of datasets undergoing the match operation. In the following, we discuss existing conflation (matching) technologies based on the types of geospatial datasets dealt with.

- *Vector to vector* **conflation:** There have been a number of efforts to automatically or semi-automatically accomplish vector to vector



Conflation of Geospatial Data, Fig. 4 An example of Delaunay triangulation based on control points (Modified figure from Chen et al. 2006a)

conflation. Most of the existing vectorvector conflation algorithms are with a focus on road vector data. These approaches are different, because of the different methods utilized for locating the counterpart elements from both vector datasets. The major approaches include:

- Matching vector data based on the similarities of geometric information (such as nodes and lines) (Saalfeld 1988; Walter and Fritsch 1999; Ware and Jones 1998).
- Matching attributeannotated vector data based on the similarities of vector shapes as well as the semantic similarities of vector attributes (Cobb et al. 1998).
- Matching vector data with unknown coordinates based on the feature point (e.g., the road intersection) distributions (Chen et al. 2006b).

- *Vector to imagery* **conflation:** *Vector to imagery* (and *Vector to raster*) conflation, on the other hand, mainly focus on developing effective and efficient image processing techniques to resolve the correspondence problem. The major approaches include:
  - Detecting all salient edges from imagery and then comparing with vector data (Filin and Doytsher 2000).
  - Utilizing vector data to identify corresponding image edges based on (modified) Snakes algorithm (Agouris et al. 2001; Kass et al. 1987).
  - Utilizing stereo images, elevation data and knowledge about the roads (e.g., parallellines and road marks) to compare vector and imagery (Eidenbenz et al. 2000).
  - Exploiting auxiliary spatial information (e.g., the coordinates of imagery and vector, the shape of roads around intersections, etc.) and non-spatial information (e.g., the image color/resolution and road widths) to perform a localized image processing to compute the correspondence (Chen et al. 2004b, 2006a). Figure 2b is the example result based on this technology.
- *Raster to raster* **conflation:** In general, *raster to raster* conflation (e.g., *imagery to imagery* conflation *and map to imagery* conflation) requires more data-specific image processing techniques to identify the corresponding features from raster data. Some exiting approaches, for example, include:
  - Conflating two images by extracting and matching various features (e.g., edges and feature points) across images (Seedahmed and Martucci 2002; Dare and Dowman 2000).
  - Conflating a raster map and imagery by computing the relationship between two feature point sets detected from the datasets (Chen et al. 2004a). In this approach, especially, these feature points are generated by exploiting auxiliary spatial information (e.g., the coordinates of imagery, the orientations of road segments around intersections from the raster map, etc.) and non-spatial information (e.g., the image resolution and the scale of raster maps). Figure 3b is the example result based on this technology.

Key Applications

Conflation technologies are used in many application domains, most notably the sciences and domains using high quality spatial data such as GIS.

Cartography

It is well known that computers and mathematical methods have had a profound impact upon cartography. There has been a massive proliferation of geospatial data, and no longer is the traditional paper map the final product. In fact, the focus of cartography has shifted from map production to the presentation, management and combination of geospatial data. Maps can be produced on demand for specialized purposes. Unfortunately, the data used to produce maps may not always be consistent. Geospatial data conflation can be used to address this issue. For example, we can conflate to out-of-date maps with up-to-date imagery to identify inconsistencies.

GIS

Geographic information provides the basis for many types of decisions ranging from economic and community planning, land and natural resource management, health, safety and military services. Improved geographic data should lead to better conclusions and better decisions. In general, superior data would include greater positional accuracy, topological consistency and abundant attribution information. Conflation technology, of course, plays a major role in producing high quality data for various GIS applications requiring high-quality spatial data.

Computational Geometry

Although originally the conflation technology is intended for consolidating geospatial datasets that are known to contain the same features, the methods employed in conflation can be adapted for other applications. For example, the variants of rubber-sheeting techniques are widely used to support general spatial interpolation. The point or line matching algorithms, in turn, can be used in a broad spectrum of geometric object comparisons.

Aerial Photogrammetry

With the wide availability of high resolution aerial photos, there is a pressing need to analyze aerial photos to detect changes or extract up-to-date features. In general, the problem of extracting features from imagery has been an area of active research for the last 25 years and given the current state-of-the-art will be unlikely to provide near-term fully-automated solutions to the feature extraction problem. For many regions, there are detailed feature datasets that have already been constructed, but these may need to be conflated with the current imagery. Conflating and correlating vector data with aerial photos is more likely to succeed over a pure feature extraction approach since we are able to exploit significant prior knowledge about the properties of the features to be extracted from the photos. Furthermore, after conflation, the attribution information contained in vector dataset can be used to annotate spatial objects to better understand the context of the photos.

Homeland Security

The conflation of geospatial data can provide insights and capabilities not possible with individual data. It is important to the national interest that this automatic conflation problem be addressed since significant statements concerning the natural resources, environment, urban settlements, and particularly internal or Homeland Security, are dependent on the results of accurate conflation of geospatial datasets such as satellite images and geospatial vector data including transportation, hydrographic and cadastral data.

Military Training and Intelligence

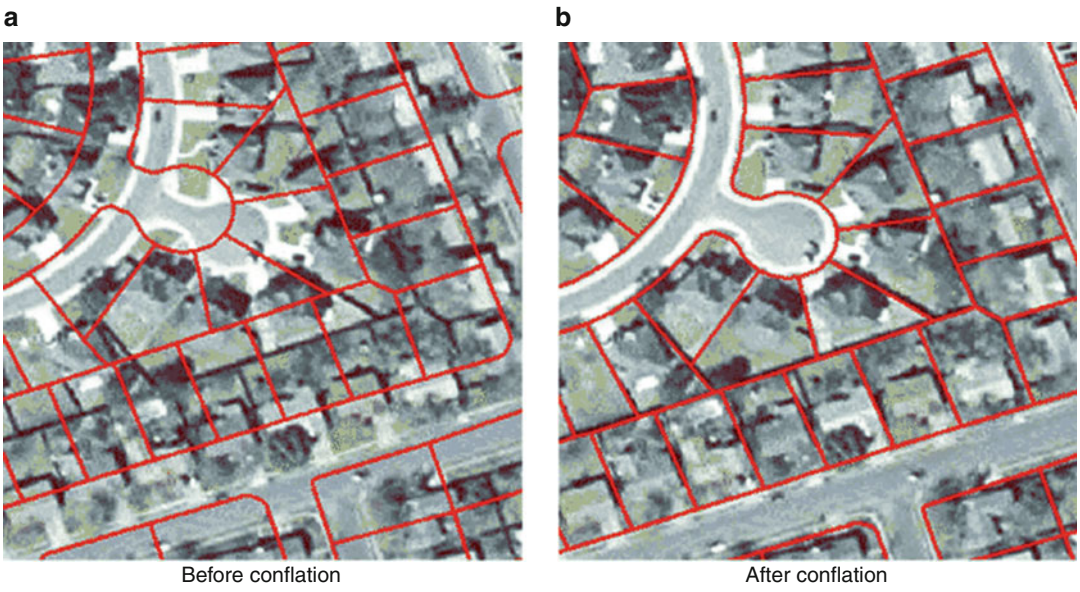
Many military training and preparation systems require high quality geospatial data for correctly building realistic training environments across diverse systems/applications. An integrated view of geographic datasets (especially satellite imagery and maps) can also help military intelligence analysts to more fully exploit the information contained in maps (e.g., road/railroad networks and textual information from the map, such as road names and gazetteer data) for analyzing imagery (i.e., identify particular targets, features, and other important geographic characteristics) and use the information in imagery to confirm information in maps. The geospatial data conflation technique is the key technology to accomplish this.

Crisis Management

In a crisis, such as a large fire, a category 5 hurricane, a dirty bomb explosion, emergency personnel must have access to relevant geographic information quickly. Typically, geographic data, such as maps and imagery are important data sources for personnel who are not already familiar with a local area. The conflation technology enables emergency personnel to rapidly integrate the maps, vector data, and imagery for a local area to provide an integrated geographic view of an area of interest.

Transportation Data Update

Many GIS applications require the road vector data for navigation systems. These days, up-to-date high resolution imagery is often utilized to



Conflation of Geospatial Data, Fig. 5 An example of parcel vector data to imagery conflation

verify and update road vector data. The ability to automatically conflate the original road vector data with images supports more efficient and accurate updates of road vector.

Real Estate

With the growth of the real estate market, there are many online services providing real estate records by superimposing the parcel boundaries on top of high-resolution imagery to show the location of parcels on imagery. However, as is typically the case in integrating different geospatial datasets, a general problem in combining parcel vector data with imagery from different sources is that they rarely align (as shown in Fig. 5a). These displacements can mislead the interpretation of parcel and land use data. As the example shown in Fig. 5, parcel data are often represented as polygons and include various attributes such as ownership information, mailing address, acreage, market value and tax information. The cities and counties use this information for watershed and flood plain modelling, neighborhood and transportation planning. Furthermore, various GIS applications rely on parcel data for more accurate geocoding. By conflating parcel vector data and imagery, the detailed attribution information provided by the parcel data (as an example shown Fig. 5b) can be combined with the visible information provided by the imagery. Therefore, the conflation of these datasets can provide cost savings for many applications, such as county, city, and state planning, or integration of diverse datasets for more accurate address geocoding or emergency response.

Future Directions

With the rapid improvement of geospatial data collection techniques, the growth of Internet and the implementation of Open GIS standards, a large amount of geospatial data is now readily available. There is a pressing need to combine these datasets together using conflation technology. Although there has been significant progress on automatic conflation technology in the last few years, there is still much work to be done. Important research problems include, but are not limited to the following: (1) resolving discrepancies between datasets with very different levels of resolution and thematic focus, (2) extending existing technologies to handle a broad range of datasets (in addition to road networks), such as

elevation data and hydrographic data, (3) allowing for uncertainty in the feature matching stage, and (4) improving the processing time (especially for raster data) to achieve conflation on the fly.

Cross-References

- ▸ Change Detection
- ▸ Intergraph: Real-Time Operational Geospatial Applications
- ▸ Photogrammetric Applications
- ▸ Uncertain Environmental Variables in GIS
- ▸ Voronoi Diagram

References

Agouris P, Stefanidis A, Gyftakis S (2001) Differential snakes for change detection in road segments. Photogramm Eng Remote Sens 67(12):1391–1399

Chen C-C, Knoblock CA, Shahabi C, Chiang Y-Y, Thakkar S (2004a) Automatically and accurately conflating orthoimagery and street maps. In: Proceedings of the 12th ACM international symposium on advances in geographic information systems, Washington, DC

Chen C-C, Shahabi C, Knoblock CA (2004b) Utilizing road network data for automatic identification of road intersections from high resolution color orthoimagery. In: Proceedings of the second workshop on spatiotemporal database management (co-located with VLDB2004), Toronto

Chen C-C, Knoblock CA, Shahabi C (2006a) Automatically conflating road vector data with orthoimagery. Geoinformatica 10(4):495–530

Chen C-C, Shahabi C, Knoblock CA, Kolahdouzan M (2006b) Automatically and efficiently matching road networks with spatial attributes in unknown geometry systems. In: Proceedings of the third workshop on spatiotemporal database management (co-located with VLDB2006), Seoul

Cobb M, Chung MJ, Miller V, Foley H III, Petry FE, Shaw KB (1998) A rule-based approach for the conflation of attributed vector data. GeoInformatica 2(1):7–35

Dare P, Dowman I (2000) A new approach to automatic feature based registration of SAR and SPOT images. Int Arch Photogramm Remote Sens 33(B2):125–130

Eidenbenz C, Kaser C, Baltsavias E (2000) ATOMI – automated reconstruction of topographic objects from aerial images using vectorized map information. Int Arch Photogramm Remote Sens 33(Part 3/1):462–471

Filin S, Doytsher Y (2000) A linear conflation approach for the integration of photogrammetric information and GIS data. Int Arch Photogramm Remote Sens 33:282–288

Flavie M, Fortier A, Ziou D, Armenakis C, Wang S (2000) Automated updating of road information from aerial images. In: Proceedings of American society photogrammetry and remote sensing conference, Amsterdam

Kass M, Witkin A, Terzopoulos D (1987) Snakes: active contour models. Int J Comput Vis 1(4): 321–331

Saalfeld A (1988) Conflation: automated map compilation. Int J Geogr Inf Sci 2(3): 217–228

Seedahmed G, Martucci L (2002) Automated image registration using geometrical invariant parameter space clustering (GIPSC). In: Proceedings of the photogrammetric computer vision, Graz

Walter V, Fritsch D (1999) Matching spatial data sets: a statistical approach. Int J Geogr Inf Sci 5(1): 445–473

Ware JM, Jones CB (1998) Matching and aligning features in overlayed coverages. In: Proceedings of the 6th ACM symposium on geographic information systems, Washington, DC

Conflict Resolution

- ▸ Computing Fitness of Use of Geospatial Datasets
- ▸ Smallworld Software Suite

Consequence Management

- ▸ Emergency Evacuations, Transportation Networks

Conservation Medicine

- ▸ Exploratory Spatial Analysis in Disease Ecology

Constrained Nearest Neighbor Queries

- ▸ Variations of Nearest Neighbor Queries in Euclidean Space

Constraint Data, Visualizing

Shasha Wu
Department of Mathematics, Computer Science and Physics, Spring Arbor University, Spring Arbor, MI, USA

Synonyms

Constraint database visualization; Isometric color bands displays

Definition

In general, visualization is any technique for creating images, diagrams, or animations in order to present any message. Scientific visualization is an application of computer graphics which is concerned with the presentation of potentially huge quantities of laboratory, simulation, or abstract data to aid cognition, hypotheses building, and reasoning.

In a spatial database system, spatial information is usually stored in the format of *raster data* or *vector data*. To visualize *raster data*, the visualization application has to convert the geographical information associated with each pixel into a specific color and present each pixel individually. To visualize *vector data*, the visualization application has to identify the geometrical primitives such as points, lines, curves, and polygons, convert the original geospatial coordinate system to screen coordinate system, associate a particular color to each shape, and then output those shapes through the drawing functions provided by the operating system.

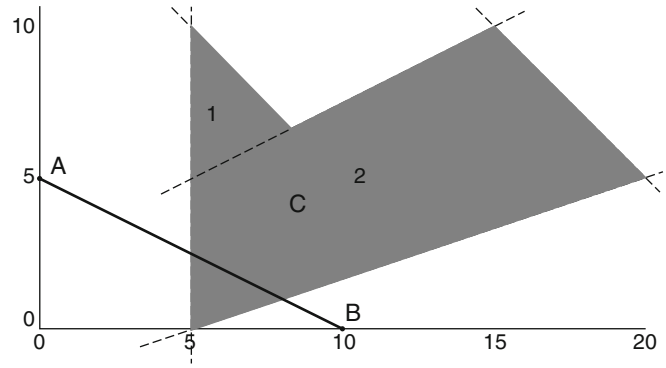
The geometrical primitives used by vector data are all based upon mathematical equations to represent images in computer graphics. The constraint databases use linear equality and inequality constraints as its primitive data type to represent spatial data. That makes constraint databases a natural solution for storing, retrieving, and displaying vector-based spatial data.

The visualization of spatial data in a constraint database system is the process of transforming the vector-based spatial data, which is represented by linear equality and linear inequality constraints in a disjunctive normal form (DNF) (Revesz 2002; Rigaux et al. 2003), into a set of points, segments, and convex polygons, associating a color with the individual shapes, and then projecting those shapes onto the output devices.

For example, Fig. 1 is the visualization result of the following three linear constraint relations:

$$
\begin{aligned}
\text{PointA}(x, y) :- \quad & x = 0,\ y = 5. \\
\text{LineAB}(x, y) :- \quad & x \geq 0,\ y \geq 0,\ x + 2y = 10. \\
\text{PolygonC}(i, x, y) :- \quad & i = 1,\ x - 2y \leq -5,\ x + y \leq 15, \\
& -x \leq -5. \\
\text{PolygonC}(i, x, y) :- \quad & i = 2,\ -x + 2y \leq 5\, x + y \leq 25, \\
& x - 3y \leq 5, -x \leq -5.
\end{aligned}
$$

Constraint databases are well suited for animation because they allow any granularity for the animation without requiring much data storage (Revesz 2002). Beyond that, the ability of representing spatiotemporal and non-spatiotemporal data in an identical format and the support of recursive queries make constraint databases a good approach for many difficult visualization problems, such as the visualization of recursively defined



Constraint Data, Visualizing, Fig. 1 Visualization of point, polyline, and polygon in constraint databases

spatiotemporal concepts discussed in Revesz and Wu (2004, 2006).

Although most existing constraint database systems can only visualize 2-D spatiotemporal objects, they can be extended to visualize three or even higher-dimensional spatiotemporal objects. By introducing new variables into the linear constraints, constraint databases can represent higher-dimensional objects similar to 2-D objects. The visualization of those objects is reduced to a process of visualizing the union of basic higher-dimensional blocks.

Historical Background

Constraint databases, including spatial constraint databases, were proposed by Kanellakis, Kuper, and Revesz in (1990). They showed in Kanellakis et al. (1995) that "efficient, declarative database programming can be combined with efficient constraint solving" and suggested that the constraint database framework can be applied to manage spatial data. A few years later, several spatial constraint databases systems, such as the MLPQ system (Revesz and Li 1997), the CCUBE system (Brodsky et al. 1997), the DEDALE system (Grumbach et al. 1998), and the CQA/CDB system (Goldin et al. 2003), were developed. During the development of those systems, convex polygons were the major visualization blocks presenting the outputs of the spatial constraint databases systems. Extreme point data models like the rectangles data model (Revesz 2002) and Worboys' data model are also implemented in some constraint database systems. For example, the MLPQ system implements both the regular polygon visualization and the parametric rectangle visualization. The last one, named the PReSTO system, implements several special animation features like *Collide* and *Block*. With the increased number of applications developed from the spatial constraint database systems, the aim of efficiently and naturally visualizing sophisticated spatial or spatiotemporal constraint data attracts more and more attention.

Scientific Fundamentals

Static Displays

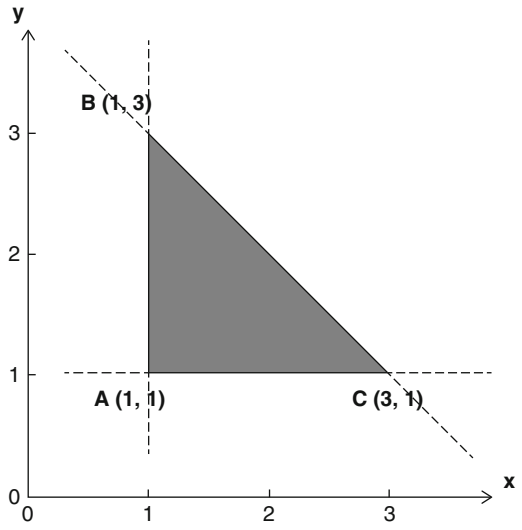
Any 2-D static display can be reduced to the visualization of points, polylines, and polygons. In constraint databases, a point can be directly represented by linear equations over two variables (x, y). For example, point $A(1,1)$ in Fig. 2 can be represented as

$$\mathrm{A(x, y) : -} \quad \mathrm{x} = 1, \mathrm{y} = 1.$$

It is a trivial problem to visualize a point with x and y coordinations. Things are a little bit more complex for polylines and polygons.

The line segment between points $B(1,3)$ and $C(3,1)$ can be represented as

$$\mathrm{BC(x, y) : -} \quad \mathrm{x + y} = 4, \mathrm{x} \geq 1, \mathrm{x} \leq 3,$$
$$\mathrm{y} \geq 1, \mathrm{y} \leq 3,$$



Constraint Data, Visualizing, Fig. 2 Representing spatial object by convex polygon(s)

A polygon can be either a convex polygon or non-convex polygon. A convex polygon can be directly represented by a set of conjunctive linear inequality constraints over the two variables (x, y). A non-convex polygon must be first partitioned into convex components. Then, it can be represented by constraint databases and visualized through the union of the convex components.

This way, any vector data can be represented by disjunctive normal form formulas of linear equations and linear inequality constraints over x and y (Revesz 2002). For example, the triangle formed by vertices $A(1,1)$, $B(1,3)$, and $C(3,1)$ can be represented by the following conjunction of linear inequality constraints:

$$\mathrm{ABC(x,\ y)} : - \quad \mathrm{x} \geq 1,\ \mathrm{y} \geq 1,\ \mathrm{x} + \mathrm{y} \leq 4. \quad (1)$$

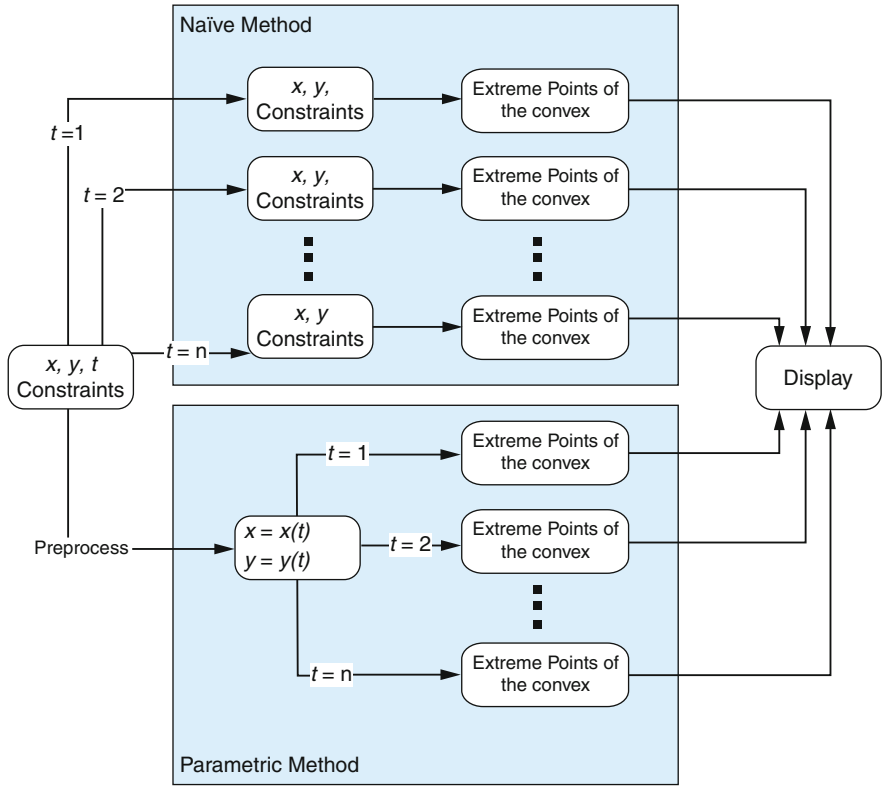
Most of the original spatial objects are described by polygons. To represent those objects, constraint databases have to first decompose polygons into convex components. For some kind of polygons, it is a NP-hard problem (O'Rourke and Supowit 1983). But for most of the commonly used polygons, it is possible to find polynomial polygon partition algorithms (Chazelle and Dobkin 1979; Keil 1985; Schachter 1978). Among all of them, the simplest algorithm is polygon triangulation, which represents a polygon through a set of triangles. However, triangulation results in a large, sometimes prohibitive, number of convex components in the partition. Given a polygon with n vertices, the number of triangles in the partition is $n - 2$. To solve the problem, Keil (1985) proposed an algorithm that can generate an optimal number of convex components in the partition for most types of polygons. However, the time complexity of his algorithm is $O(N^2 n \log n)$. To reduce the time complexity, a less optimal algorithm is proposed and implemented in Rigaux et al. (2003), in which the polygon is first triangulated and then the adjacent triangles are merged to reduce the number of convex components in the result.

Animation of Spatiotemporal Objects

Each *spatiotemporal object* has a *spatial extent* and a *temporal extent*. The spatial extent represents the set of points in space that belong to the object. The temporal extent represents the set of time instances when the object exists. The shape and the location of a 2-D spatiotemporal object may change over time. In constraint databases, each 2-D spatiotemporal object is represented by linear constraints over the three variables (x, y, t) in disjunctive normal form formula (Revesz 2002).

There are two different animation methods, as shown in Fig. 3, to visualize the spatiotemporal constraints:

The *naive animation method* works directly on constraint databases. It finds the linear constraint relations that have only two spatial variables named as x and y to represent the extreme points of the polygon for each time instance t_i, by instantiating the variable t to t_i in the original linear constraint tuple. Then it calls the graphic library functions provided by the operating system to output the polygon. The whole computation will be executed every time the user requests an animation display. It is a time-consuming process and often times causes many delays and jumps in the animation.



Constraint Data, Visualizing, Fig. 3 Naive and parametric animation methods (See Fig. 16.11 in Revesz 2002)

The *parametric animation method* has a preprocessing step and a display step to speed up the animation. The preprocessing step is executed at the time the constraint relation is loaded or constructed. It first computes the *extreme points* of each polygon based on its constraint tuple. Then, each polygon is describable by a sequence of extreme points. Finally, each extreme point is represented by parametric functions $x = x(t)$ and $y = y(t)$, which will be kept in memory until the close of the constraint relation. The display step is executed every time the user requests an animation display. After the user specifies the range and the granularity of time for the animation and sends the request to the system, the extreme point parametric functions are loaded and the time variable t is instantiated several times based on the required granularity. It generates a sequence of polygon outputs and sequentially and smoothly outputs them onto the monitor. This way, the spatiotemporal data are visualized as an animation display.

Key Applications

The visualization of spatial constraint databases is similar to the visualization of other GIS systems, such as the ARC/GIS system. However, the power of efficiently describing infinite spatial and spatiotemporal data and the support of recursive queries make the visualization of spatial constraint databases more attractive for complex problems like visualization of the recursively defined spatiotemporal concepts (Revesz and Wu 2004). These applications typically include problems where various kinds of spatial and spatiotemporal information such as maps, population, meteorology phenomena, and moving objects are represented and visualized. The following are some examples of such applications.

Visualization Functions of the MLPQ Constraint Database System

The MLPQ constraint database system implemented many visualization operators. For example, the *Complement* operator returns the complement of the given spatial object. The *Difference* operator generates the difference between two spatial objects. Three commonly used visualization operators are described as follows.

2D Animation

The MLPQ constraint database system can display the spatiotemporal relations in animations. It provides a set of buttons for the user to control the displaying of the animation. The animation button allows the user to set the start and end time, the time interval of two frames, and the speed of the animation. The play and playback buttons play the animation forward and backward, respectively. The first, forward, next, and last buttons allow the user to navigate between frames.

Block

Some spatial objects like light and fire are formed by a set of independent points. If some of the points are blocked by the presence of another object, the rest of the points just continue moving along the trajectory determined by the transformation function. The block operator is designed to visualize such situations in the constraint databases. It takes two relations and a time instance t_k as the inputs and returns a new relation that represents the points of the first relation at time instance t_k that are not blocked by the second relation at any time before t_k. Based on the block operator, people can easily visualize the spatial objects such as the shadow of a ball or show how a lake can block the movement of a fire in the forest.

Collide

A common scenario between moving objects is the collision. The *Collide* operator is designed to visualize the collision situation in animation. It assumes an extra attribute for spatiotemporal objects called mass. Suppose there are two objects that do not change their shape and that are traveling at uniform speed defined by the transition function in the constraint databases. The *Collide* operator will generate a new relation that expresses the motion of the objects before and after collision. This operator can be used to visualize applications like the crash of two cars or the contact of two billiard balls.

Applications Based on Recursively Defined Concepts

Visualization of recursively defined concepts is a general problem that appears in many areas. For example, drought areas based on the Standardized Precipitation Index (SPI) and long-term air pollution areas based on safe and critical level standards are recursively defined concepts. In Revesz and Wu (2004), a general and efficient representation and visualization method was proposed to display recursively defined spatiotemporal concepts. Sample applications such as visualization of drought and pollution areas were implemented to illustrate the method.

Applications for Epidemiology

Efficient computerized reasoning about epidemics is important to public health and national security, but it is a difficult task because epidemiological data are usually spatiotemporal, recursive, and fast changing, hence hard to handle in traditional relational databases and geographic information systems. In Revesz and Wu (2006), a particular epidemiological system called WeNiVIS was implemented based on the visualization of spatiotemporal constraint databases. It enables the visual tracking of the West Nile virus epidemic in Pennsylvania and helps people to predicate the high-risk areas.

Future Directions

Spatial constraint databases provide an efficient way to store and query spatial or spatiotemporal data over the Internet. There are a growing number of web-based spatial constraint database applications. Most of the applications ask for high-level visualization methods of constraint data to improve their user interfaces. Methods to enhance

the performance of visualizing spatial constraint databases over the Internet are being developed.

By adding one more parameter in the database, spatial constraint databases can easily represent 3-D data. However, by the time of writing this entry, only primitive visualization methods like isometric color bands are implemented in some existing constraint database systems. For example, the map can be visualized by discrete color zones according to the values of variable z, which can be used to represent the value of elevation, precipitation, or temperature. Using 2-D images to visualize 3-D objects is a temporary solution for this problem. Compared to the 3-D visualization of other commercial GIS systems, this method has many restrictions on the 3-D objects to be visualized, and the result is not impressive. The implementations of real 3-D visualization of constraint databases are being developed.

Cross-References

- Constraint Database Queries
- Constraint Databases and Data Interpolation
- Constraint Databases and Moving Objects
- Constraint Databases, Spatial
- MLPQ Spatial Constraint Database System
- Raster Data
- Vector Data

References

Brodsky A, Segal V, Chen J, Exarkhopoulo P (1997) The CCUBE constraint object-oriented database system. Constraints 2(3–4):245–277

Chazelle B, Dobkin D (1979) Decomposing a polygon into its convex parts. In: Proceedings of 11th annual ACM symposium on theory of computing, Atlanta, pp 38–48

Goldin D, Kutlu A, Song M, Yang F (2003) The constraint database framework: lessons learned from CQA/CDB. In: Proceedings of international conference on data engineering, Bangalore, pp 735–737

Grumbach S, Rigaux P, Segoufin L (1998) The {DEDALE} system for complex spatial queries. In: Proceedings of ACM SIGMOD international conference on management of data, Seattle, pp 213–224

Kanellakis PC, Kuper GM, Revesz P (1990) Constraint query languages. In: Proceedings of ACM symposium on principles of database systems, Nashville, pp 299–313

Kanellakis PC, Kuper GM, Revesz P (1995) Constraint query languages. J Comput Syst Sci 51(1):26–52

Keil JM (1985) Decomposing a polygon into simpler components. SIAM J Comput 14(4):799–817

O'Rourke J, Supowit KJ (1983) Some NP-hard polygon decomposition problems. IEEE Trans Inf Theory 29(2):181–190

Revesz P (2002) Introduction to constraint databases. Springer, New York

Revesz P, Li Y (1997) MLPQ: a linear constraint database system with aggregate operators. In: Proceedings of 1st international database engineering and applications symposium. IEEE Press, Washington DC, pp 132–137

Revesz P, Wu S (2004) Visualization of recursively defined concepts. In: Proceedings of the 8th international conference on information visualization. IEEE Press, Washington DC, pp 613–621

Revesz P, Wu S (2006) Spatiotemporal reasoning about epidemiological data. Artif Intell Med 38(2):157–170

Rigaux P, Scholl M, Segoufin L, Grumbach S (2003) Building a constraint-based spatial database system: model, languages, and implementation. Inf Syst 28(6):563–595

Schachter I (1978) Decomposition of polygons into convex sets. IEEE Trans Comput 27(11):1078–1082

Constraint Database Queries

Lixin Li
Department of Computer Sciences, Georgia Southern University, Statesboro, GA, USA

Synonyms

Constraint query languages; Datalog, SQL; Logic programming language

Definition

A database query language is a special-purpose programming language designed for retrieving information stored in a database. Structured query language (SQL) is a very widely used commercially marketed query language for relational databases. Different from conventional

programming languages such as C, C++, or Java, a SQL programmer only needs to specify the properties of the information to be retrieved, but not the detailed algorithm required for retrieval. Because of this property, SQL is said to be *declarative*. In contrast, conventional programming languages are said to be *procedural*.

To query spatial constraint databases, any query language can be used, including SQL. However, Datalog is probably the most popularly used rule-based query language for spatial constraint databases because of its power of recursion. Datalog is also declarative.

Historical Background

The Datalog query language is based on logic programming language Prolog. The history of Datalog queries and logic programming is discussed in several textbooks such as Ramakrishnan (1998), Silberschatz et al. (2006), and Ullman (1989). Early work on constraint logic programming has been done by Jaffar and Lassez (1987). The concepts of constraint data model and query language have been explored by Kanellakis et al. (1990, 1995). Recent books on constraint databases are Kuper et al. (2000) and Revesz (2002).

Scientific Fundamentals

A Datalog query program consists of a set of rules of the following form (Revesz 2002):

$$R_0(x_1, \ldots, x_k) :- R_1(x_{1,1}, \ldots, x_{1,k}), \ldots R_n(x_{n,1}, \ldots, x_{n,k}).$$

where each R_i is either an input relation name or a defined relation name and the xs are either variables or constants.

Query 1 *For the ultraviolet radiation example in Scientific Fundamentals in Entry "▸ Constraint Databases and Data Interpolation", find the amount of ultraviolet radiation for each ground location (x,y) at time t.*

Since the input relations in Tables 1 and 2 in Entry "Constraint Databases and Data Interpolation" only record the incoming ultraviolet radiation u and filter ratio r on a few sample points, these cannot be used directly to answer the query. Therefore, to answer this query, the interpolation results of *INCOMING(y, t, u)* and *FILTER(x, y, r)* are needed. To write queries, it is not necessary to know precisely what kind of interpolation method is used and what are the constraints used in the representation interpolation. The above query can be expressed in Datalog as follows (Li 2003):

$$GROUND(x, y, t, i) : - INCOMING(y, t, u), FILTER(x, y, r), i = u(1 - r).$$

The above query could be also expressed in SQL. Whatever language is used, it is clear that the evaluation of the above query requires a join of the *INCOMING* and *FILTER* relations. Unfortunately, join operations are difficult to express in simple GIS, including the ArcGIS system. However, join processing is very natural in constraint database systems.

If the IDW interpolation is used (see section "Scientific Fundamentals," "Key Applications," Entry "▸ Constraint Databases and Data Interpolation"), the final result of the Datalog query, *GROUND(x, y, t, i)*, can be represented by Table 1. Since there are five second-order Voronoi regions for *Incoming* and four regions for *Filter*, as shown in Figs. 8 and 9 in Entry "▸ Constraint Databases and Data Interpolation," there should be 20 tuples in *GROUND(x, y, t, i)* in Table 1. Note that the constraint relations can be easily joined by taking the conjunction of the constraints from each pair of tuples of the two input relations. Finally, in a constraint database system, the constraint in each tuple is automatically simplified by eliminating the unnecessary variables u and r.

Constraint Database Queries, Table 1 *GROUND (x, y, t, i)* using IDW

X	Y	T	I	
x	y	t	i	$2x-y-20 < q0, 12x+7y-216 < q0, 13y+7t-286 < q0, 2y-3t-12 < q0, y < q15,$
				$((x-2)^2+(y-14)^2)0.9+((x-2)^2+(y-1)^2)0.5$ $=(2(x-2)^2+(y-14)^2+(y-1)^2)r,$
				$((y-13)^2+(t-22)^2)60+(y^2+(t-1)^2)20$ $=((y-13)^2+(t-22)^2+y^2+(t-1)^2)u,$
				$i=u(1-r)$
x	by	bt	bi	b$2x-y-20 \geq 0, 12x+7y-216 < q0, 13y+7t-286 < q0, 2y-3t-12 < q0, y < q15,$
				$((x-25)^2+(y-1)^2)0.9+((x-2)^2+(y-1)^2)0.8$ $=(2(y-1)^2+(x-25)^2+(x-2)^2)r,$
				$((y-13)^2+(t-22)^2)60+(y^2+(t-1)^2)20$ $=((y-13)^2+(t-22)^2+y^2+(t-1)^2)u,$
				$i=u(1-r)$
x	y	t	y	$\cdots$
				$\cdots$
x	y	t	i	$2x-y-20 < q0, 12x+7y-216 \geq 0, y \geq 15, y+3t-54 < q0, 7y-t-136 < q0, 2y+5t-60 \geq 0,$
				$((x-25)^2+(y-14)^2)0.5+((x-2)^2+(y-14)^2)0.3$ $=(2(y-14)^2+(x-25)^2+(x-2)^2)r$
				$((y-29)^2+t^2)20+((y-13)^2+(t-22)^2)40$ $=((y-29)^2+t^2+(y-13)^2+(t-22)^2)u,$
				$i=u(1-r)$

Key Applications

There are many possible queries for a particular set of GIS data. For example, a very basic query for a set of spatiotemporal data would be, "What is the value of interest at a specific location and time instance?" With good interpolation results and efficient representation of the interpolation results in constraint databases, many spatiotemporal queries can be easily answered by query languages. In the following, some examples of Datalog queries are shown.

Ozone Data Example

Based on the ozone data example in section "Historical Background", "Key Applications," Entry "▸ Constraint Databases and Data Interpolation," some sample spatiotemporal queries are given below. Assume that the input constraint relations are (Li et al. 2006):

- *Ozone_orig(x, y, t, w)*, which records the original measured ozone value w at monitoring site location (x, y) and time t;

Constraint Database Queries, Table 2 Sample *(x, y, t, p)*

X	Y	T	P (price/square foot)
888	115	4	56.14
888	115	76	76.02
1630	115	118	86.02
1630	115	123	83.87
...	...	...	...
2240	2380	51	91.87
2650	1190	43	63.27

- *Ozone_interp(x, y, t, w)*, which stores the interpolation results of the ozone data by any spatiotemporal interpolation method, such as 3-D shape function or IDW;
- *Ozone_loocv(x, y, t, w)*, which stores the interpolated ozone concentration level at each monitoring site (x, y) and time t after applying the leave-one-out cross-validation.

Note

The leave-one-out cross-validation is a process that removes one of the n observation points and uses the remaining $n-1$ points to estimate

its value, and this process is repeated at each observation point (Hjorth 1994). The observation points are the points with measured original values. For the experimental ozone data, the observation points are the spatiotemporal points (x, y, t), where (x, y) is the location of a monitoring site and t is the year when the ozone measurement was taken. After the leave-one-out cross-validation, each of the observation points will not only have its original value but also will have an interpolated value. The original and interpolated values at each observation point can be compared for the purpose of an error analysis. The interpolation error at each data point by calculating the difference between its original and interpolated values is as follows:

$$E_i = \frac{| I_i - O_i |}{O_i} \quad (1)$$

where E_i is the interpolation error at observation point i, I_i is the interpolated value at point i, and O_i is the original value at point i.

Query 2 *For a given location with longitude x and latitude y, find the ozone concentration level in year t.*

This can be expressed in Datalog as follows:

$$Ozone\_value(w) - Ozone\_interp(x, y, t, w).$$

Query 3 *Suppose that in the future years, there will be a budget increase so that new ozone monitoring sites can be added. Find the best areas where new monitoring sites should be installed.*

In order to decide the best locations to add new monitoring sites, it is necessary to first find those monitoring sites that have average large interpolation errors according to equation (1), for example, over 20 %. Then, do a *buffer* operation on the set of monitoring sites with big errors to find out the areas within certain distance to each site, for example, 50 miles. Since the buffered areas are the areas with poor interpolation result, these areas can be considered the possible areas where new monitoring sites should be built. To find the monitoring sites with more than 20 % interpolation errors, perform the following Datalog queries:

$$Error(x, y, t, r) : -Ozone\_orig(x, y, t, w1),$$
$$Ozone\_loocv(x, y, t, w2),$$
$$r = | w1 - w2 | / w1.$$
$$Avg\_error(x, y, avg(r)) : -Error(x, y, t, r).$$
$$Sites\_Chosen(x, y) : -Avg\_error(x, y, ae),$$
$$ae >= 0.2.$$

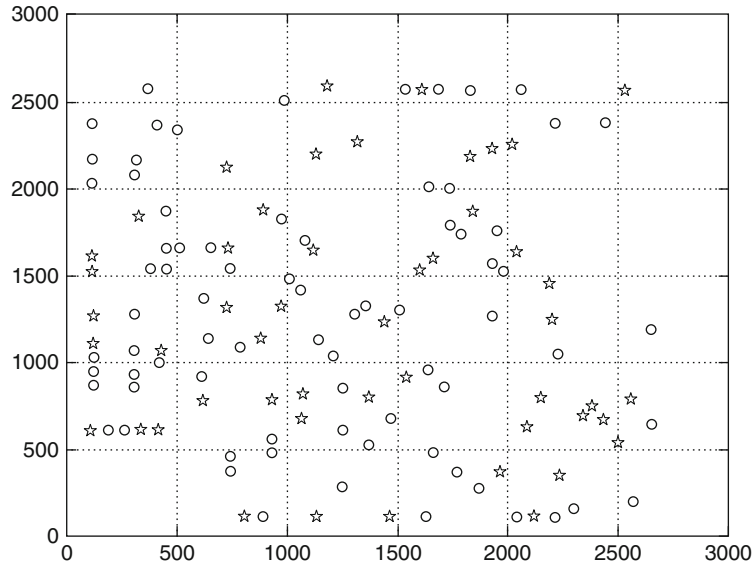
To find the areas within 50 miles to the sites with more than 20 % interpolation errors, a GIS *buffer* operation on the relation *Sites_Chosen* should be performed. The buffer operation is provided by many GIS software packages and the MLPQ constraint database system. After performing the buffer operation, an output relation will be created which contains a 50-mile buffer around the locations stored in the *Sites_Chosen* relation.

Similarly, if there will be a budget cut, similar queries to find out and shut down the monitoring sites with small interpolation errors can be designed.

House Price Data Example

The house price data consist of a set of real estate data obtained from the Lancaster County assessor's office in Lincoln, Nebraska. House sale histories since 1990 are recorded in the real estate data set and include sale prices and times. In the experiment, 126 residential houses are randomly selected from a quarter of a section of a township, which covers an area of 160 acres. Furthermore, from these 126 houses, 76 houses are randomly selected as sample data, and the remaining 50 houses are used as test data. Figure 1 shows the 76 houses with circles and the 50 remaining houses with stars.

Tables 2 and 3 show instances of these two data sets. Based on the fact that the earliest sale of the houses in this neighborhood is in 1990, the time is encoded in such a way that 1 represents January 1990, 2 represents February 1990, ..., and 148 represents April 2002. Note that some houses are sold more than once in the past, so they have more than one tuple in Table 2. For example, the house at the location (888, 115) was sold three



Constraint Database Queries, Fig. 1
Seventy-six sample houses (∘) and 50 test houses (★)

Constraint Database Queries, Table 3 Test *(x, y, t)*

X	Y	T
115	1525	16
115	1525	58
115	1525	81
115	1610	63
...	...	...
120	1110	30
615	780	59

times in the past at time 4 and 76 (which represent 4/1990 and 4/1996) (Li and Revesz 2002).

Assume that the input constraint relations are $House(x, y, t, p)$ and $Built(x, y, t)$. $House(x, y, t, p)$ represents the interpolation result of house price data, and $Built(x, y, t)$ records the time t (in month) when the house at location (x, y) was built. The *Built* relation can be usually easily obtained from real estate or city planning agencies.

Query 4 *For each house, find the starting sale price when the house was built.*

This can be expressed as follows:

$$Start(x, y, p) : -Built(x, y, t), \quad House(x, y, t, p).$$

Query 5 *Suppose it is known that house prices in general decline for some time after the first sale. For each house, find the first month when it becomes profitable, that is, the first month when its price exceeded its initial sale price.*

This can be expressed as follows:

$$not\_Profitable(x, y, t) : -Built(x, y, t).$$
$$not\_Profitable(x, y, t_2) : -not\_Profitable(x, y, t_1), House(x, y, t_2, p_2), Start(x, y, p), t_2 = t_1 + 1, p_2 > qp.$$
$$Profitable(x, y, t_2) : -not\_Profitable(x, y, t_1), House(x, y, t_2, p_2), Start(x, y, p), t_2 = t_1 + 1, \ p_2 > p.$$

Query 6 *How many months did it take for each house to become profitable?*

This translates as

$$Time\_to\_Profit(x, y, t_3) : -Built(x, y, t_1), Profitable(x, y, t_2), t_3 = t_2 - t_1.$$

All of the above queries could be a part of a more complex data mining or decision support task. For example, a buyer may want to find out which builders tend to build houses that become profitable in a short time or keep their values best.

Future Directions

Interesting directions for the future work could be to continue to design more interesting queries in spatial constraint databases which can be a valuable part of decision support systems.

Cross-References

- ▸ Constraint Databases and Data Interpolation
- ▸ Constraint Databases and Moving Objects
- ▸ Constraint Databases, Spatial
- ▸ MLPQ Spatial Constraint Database System

Recommended Reading

Hjorth U (1994) Computer intensive statistical methods, validation, model selection, and bootstrap. Chapman and Hall/CRC, London/New York

Jaffar J, Lassez JL (1987) Constraint logic programming. In: Proceedings of the 14th ACM symposium on principles of programming languages, Munich, pp 111–119

Kanellakis PC, Kuper GM, Revesz P (1990) Constraint query languages. In: ACM symposium on principles of database systems, Nashville, pp 299–313

Kanellakis PC, Kuper GM, Revesz P (1995) Constraint query languages. J Comput Syst Sci 1:26–52

Kuper GM, Libkin L, Paredaens J (eds) (2000) Constraint databases. Springer, Berlin/Heidelberg

Li L (2003) Spatiotemporal interpolation methods in GIS. Ph.D thesis, University of Nebraska-Lincoln, Lincoln

Li L, Revesz P (2002) A comparison of spatio-temporal interpolation methods. In: Proceedings of the second international conference on GIScience 2002. Lecture notes in computer science, vol 2478. Springer, Berlin/Heidelberg/New York, pp 145–160

Li L, Zhang X, Piltner R (2006) A spatiotemporal database for ozone in the conterminous US. In: Proceedings of the thirteenth international symposium on temporal representation and reasoning, Washington, DC. IEEE, pp 168–176

Ramakrishnan R (1998) Database management systems. McGraw-Hill, New York

Revesz P (2002) Introduction to constraint databases. Springer, New York

Silberschatz A, Korth H, Sudarshan S (2006) Database system concepts, 5th edn. McGraw-Hill, New York

Ullman JD (1989) Principles of database and knowledge-base systems. Computer Science Press, New York

C

Constraint Database Systems

- ▸ Linear Versus Polynomial Constraint Databases
- ▸ Polynomial Spatial Constraint Databases

Constraint Database Visualization

- ▸ Constraint Data, Visualizing

Constraint Databases and Data Interpolation

Lixin Li
Department of Computer Sciences, Georgia Southern University, Statesboro, GA, USA

Synonyms

Contraint relations; Data approximation; Delaunay triangulation; Fourier series; Inverse distance weighting; Nearest neighbors; Shape function; Spatial interpolation; Spatiotemporal interpolation; Splines; Trend surfaces

Definition

Constraint databases generalize relational databases by finitely representing infinite relations. In the constraint data model, each attribute is associated with an attribute variable, and the value of an attribute in a relation is specified implicitly using constraints. Compared with the traditional relational databases, constraint databases offer an extra layer of data abstraction, which is called the *constraint level* (Revesz

2002). It is the constraint level that makes it possible for computers to use finite number of tuples to represent infinite number of tuples at the logical level.

It is very common in GIS that sample measurements are taken only at a set of points. Interpolation is based on the assumption that things that are close to one another are more alike than those that are farther apart. Interpolation is needed in order to estimate the values at unsampled points.

Constraint databases are very suitable for representing spatial/spatiotemporal interpolation results. In this entry, several spatial and spatiotemporal interpolation methods are discussed, and the representation of their spatiotemporal interpolation that results in constraint databases is illustrated by some examples. The performance analysis and comparison of different interpolation methods in GIS applications can be found in Li and Revesz (2002), Li and Revesz (2004), and Li et al. (2006).

Historical Background

There exist a number of spatial interpolation algorithms, such as *inverse distance weighting (IDW)*, *Kriging*, *splines*, *trend surfaces*, and *Fourier series*. Spatiotemporal interpolation is a growing research area. With the additional *time* attribute, the above traditional spatial interpolation algorithms are insufficient for spatiotemporal data, and new spatiotemporal interpolation methods must be developed. There have been some papers addressing the issue of spatiotemporal interpolation in GIS. Gao (2006), Li et al. (2003), and Revesz and Wu (2006) deal with the use of spatiotemporal interpolations for different applications. Li et al. (2004) and Li and Revesz (2004) discuss several newly developed *shape function* based spatial/spatiotemporal interpolation methods. There have been some applications on the shape function-based methods. For example, Li et al. (2006) applies a shape function interpolation method to a set of ozone data in the conterminous USA, and Li and Revesz (2004) compares shape function-, IDW-, and Kriging-based spatiotemporal interpolation methods by using an actual real estate data set with house prices. Revesz and Wu (2006) also uses a shape function-based interpolation method to represent the West Nile virus data in constraint databases and implements a particular epidemiological system called WeNiVIS that enables the visual tracking of and reasoning about the spread of the West Nile virus epidemic in Pennsylvania.

Scientific Fundamentals

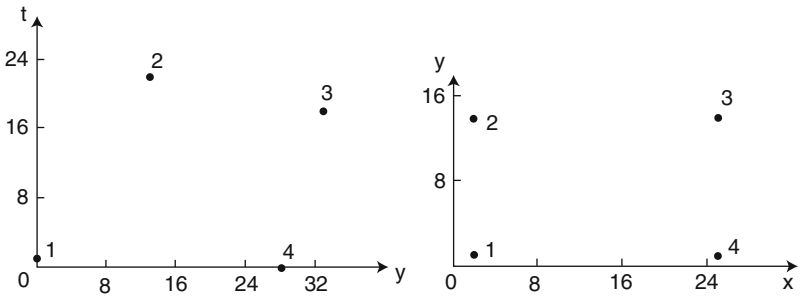
Suppose that the following two sets of sensory data are available in the database (Revesz and Li 2002):

- *Incoming* (y, t, u) records the amount of incoming ultraviolet radiation u for each pair of latitude degree y and time t, where time is measured in days.
- *Filter* (x, y, r) records the ratio r of ultraviolet radiation that is usually filtered out by the atmosphere above location (x, y) before reaching the earth.

Suppose that Fig. 1 shows the locations of the (y, t) and (x, y) pairs where the measurements for u and r, respectively, are recorded. Then Tables 1 and 2 could be instances of these two relations in a relational database.

The above relational database can be translated into a constraint database with the two constraint relations shown in Tables 3 and 4.

Although any relational relation can be translated into a constraint relation as above, not all the constraint relations can be converted back to relational databases. This is because a constraint relation can store infinite number of solutions. For example, the infinite number of interpolation results of u and r for all the points in the domains for *Incoming* (y, t, u) and *Filter* (x, y, r) can be represented in a constraint database by a finite number of tuples. The representation of interpolation results in constraint databases by different methods for *Incoming* and *Filter* will be given in Key Applications.



Constraint Databases and Data Interpolation, Fig. 1 The spatial sample points for *Incoming* (*left*) and *Filter* (*right*)

Constraint Databases and Data Interpolation, Table 1 Relational incoming (y, t, u)

ID	Y	T	U
1	0	1	60
2	13	22	20
3	33	18	70
4	29	0	40

Constraint Databases and Data Interpolation, Table 2 Relational filter (x, y, r)

ID	X	Y	R
1	2	1	0.9
2	2	14	0.5
3	25	14	0.3
4	25	1	0.8

Constraint Databases and Data Interpolation, Table 3 Constrain incoming (y, t, u)

ID	Y	T	U	
id	y	t	u	id=1, $y = 0, t = 1, u = 60$
id	y	t	u	id=2, $y = 13, t = 22, u = 20$
id	y	t	u	id=3, $y = 33, t = 18, u = 70$
id	y	t	u	id=4, $y = 29, t = 0, u = 40$

Constraint Databases and Data Interpolation, Table 4 Constraint filter (x, y, r)

ID	X	Y	R	
id	x	y	r	id=1, $x = 2, y = 1, r = 0.9$
id	x	y	r	id=2, $x = 2, y = 14, r = 0.5$
id	x	y	r	id=3, $x = 25, y = 14, r = 0.3$
id	x	y	r	id=4, $x = 25, y = 1, r = 0.8$

Key Applications

Applications Based on Shape Function Spatial Interpolation

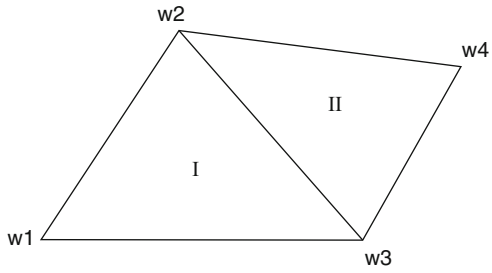
Shape functions, which can be viewed as a spatial interpolation method, are popular in engineering applications, for example, in finite element algorithms (Zienkiewics and Taylor 2000). There are various types of 2-D and 3-D shape functions. 2-D shape functions for triangles and 3-D shape functions for tetrahedra are of special interest, both of which are linear approximation methods. Shape functions are recently found to be a good interpolation method for GIS applications, and the interpolation results are very suitable to be represented in linear constraint databases (Li et al. 2004; Li and Revesz 2002, 2004; Li et al. 2006; Revesz and Li 2002).

2-D Shape Function for Triangles

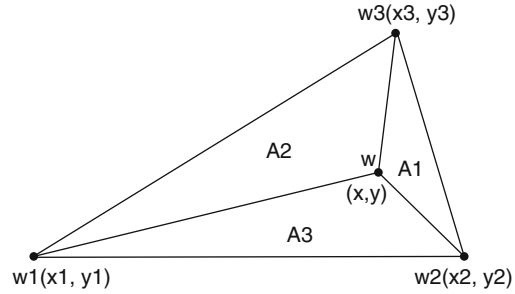
When dealing with complex two-dimensional geometric domains, it is convenient to divide the total domain into a finite number of simple sub-domains which can have triangular or quadrilateral shapes. Mesh generation using triangular or quadrilateral domains is important in finite element discretization of engineering problems. For the generation of triangular meshes, quite successful algorithms have been developed. A popular method for the generation of triangular meshes is the "Delaunay triangulation" (Preparata and Shamos 1985).

A linear interpolation function for a triangular area can be written in terms of three shape functions N_1, N_2, N_3, and the corner values w_1, w_2, w_3. In Fig. 2, two triangular finite elements, I and II, are combined to cover the whole domain considered (Li and Revesz 2004).

In this example, the function in the whole domain is interpolated using four discrete values w_1, w_2, w_3, and w_4 at four locations. A particular feature of the chosen interpolation method is that the function values inside the sub-domain I can



Constraint Databases and Data Interpolation, Fig. 2
Linear interpolation in space for triangular elements



Constraint Databases and Data Interpolation, Fig. 3
Computing shape functions by area divisions

be obtained by using only the three corner values w_1, w_2 and w_3, whereas all function values for the sub-domain II can be constructed using the corner values w_2, w_3, and w_4. Suppose $\mathcal{A}$ is the area of the triangular element I. The linear interpolation function for element I can be written as

$$w(x,y) = N_1(x,y)w_1 + N_2(x,y)w_2 + N_3(x,y)w_3 = [N_1\ N_2\ N_3]\begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} \tag{1}$$

where N_1, N_2 and N_3 are the following shape functions:

$$N_1(x,y) = \frac{[(x_2y_3 - x_3y_2) + x(y_2 - y_3) + y(x_3 - x_2)]}{2\mathcal{A}}$$
$$N_2(x,y) = \frac{[(x_3y_1 - x_1y_3) + x(y_3 - y_1) + y(x_1 - x_3)]}{2\mathcal{A}}$$
$$N_3(x,y) = \frac{[(x_1y_2 - x_2y_1) + x(y_1 - y_2) + y(x_2 - x_1)]}{2\mathcal{A}} \tag{2}$$

It should be noted that for every sub-domain, a local interpolation function similar to expression (1) is used. Each local interpolation function is constrained to the local triangular sub-domain. For example, the function w of expression (1) is valid only for sub-domain I. For sub-domain II, the local approximation takes a similar form as the expression (1) with replacing the corner values w_1, w_2 and w_3 with the new values w_2, w_3 and w_4.

Alternatively, considering only sub-domain I, the 2-D shape function (2) can also be expressed as follows (Revesz and Li 2002):

$$N_1(x,y) = \frac{\mathcal{A}_1}{\mathcal{A}},$$
$$N_2(x,y) = \frac{\mathcal{A}_2}{\mathcal{A}}, \tag{3}$$
$$N_3(x,y) = \frac{\mathcal{A}_3}{\mathcal{A}}$$

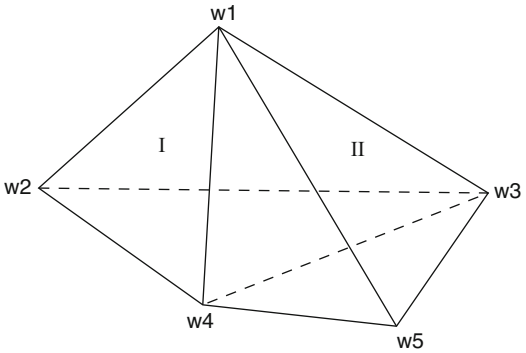
where $\mathcal{A}_1$, $\mathcal{A}_2$ and $\mathcal{A}_3$ are the three sub-triangle areas of sub-domain I as shown in Fig. 3, and $\mathcal{A}$ is the area of the outside triangle $w_1w_2w_3$.

3-D Shape Function for Tetrahedra

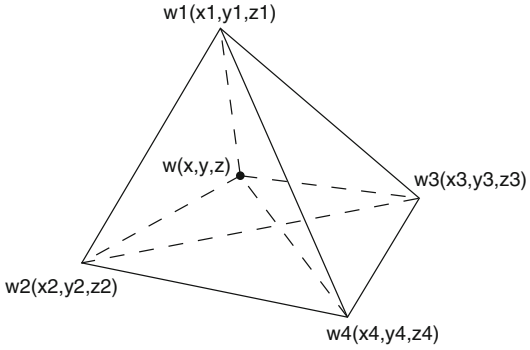
Three-dimensional domains can also be divided into a finite number of simple sub-domains, such as tetrahedral or hexahedral sub-domains. Tetrahedral meshing is of particular interest. With a large number of tetrahedral elements, complicated 3-D objects can be approximated. There exist several methods to generate automatic tetrahedral meshes, such as the 3-D Delaunay tetrahedralization and some tetrahedral mesh improvement methods to avoid poorly shaped tetrahedra.

A linear interpolation function for a 3-D tetrahedral element can be written in terms of four shape functions N_1, N_2, N_3, N_4 and the corner values w_1, w_2, w_3, w_4. In Fig. 4, two tetrahedral elements, I and II, cover the whole domain considered (Li and Revesz 2004).

In this example, the function in the whole domain is interpolated using five discrete values w_1, w_2, w_3, w_4, and w_5 at five locations in space. To obtain the function values inside the tetrahedral element I, the four corner values w_1, w_2, w_3, and w_4 can be used. Similarly, all function



Constraint Databases and Data Interpolation, Fig. 4
Linear interpolation in space for tetrahedral elements



Constraint Databases and Data Interpolation, Fig. 5
Computing shape functions by volume divisions

values for element II can be constructed using the corner values w_1, w_3, w_4, and w_5. Suppose $\mathcal{V}$ is the volume of the tetrahedral element I. The linear interpolation function for element I can be written as:

$$\begin{aligned} w(x,y,z) &= N_1(x,y,z)w_1 + N_2(x,y,z)w_2 \\ &\quad + N_3(x,y,z)w_3 + N_4(x,y,z)w_4 \\ &= [N_1\ N_2\ N_3\ N_4]\begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix} \end{aligned} \tag{4}$$

where N_1, N_2, N_3 and N_4 are the following shape functions:

$$\begin{aligned} N_1(x,y,z) &= \frac{a_1 + b_1x + c_1y + d_1z}{6\mathcal{V}}, \\ N_2(x,y,z) &= \frac{a_2 + b_2x + c_2y + d_2z}{6\mathcal{V}}, \\ N_3(x,y,z) &= \frac{a_3 + b_3x + c_3y + d_3z}{6\mathcal{V}}, \\ N_4(x,y,z) &= \frac{a_4 + b_4x + c_4y + d_4z}{6\mathcal{V}}. \end{aligned} \tag{5}$$

By expanding the other relevant determinants into their cofactors, there exists

$$a_1 = \det\begin{bmatrix} x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \\ x_4 & y_4 & z_4 \end{bmatrix} \quad b_1 = -\det\begin{bmatrix} 1 & y_2 & z_2 \\ 1 & y_3 & z_3 \\ 1 & y_4 & z_4 \end{bmatrix}$$

$$c_1 = -\det\begin{bmatrix} x_2 & 1 & z_2 \\ x_3 & 1 & z_3 \\ x_4 & 1 & z_4 \end{bmatrix} \quad d_1 = -\det\begin{bmatrix} x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \\ x_4 & y_4 & 1 \end{bmatrix}$$

with the other constants defined by cyclic interchange of the subscripts in the order 4, 1, 2, 3 (Zienkiewics and Taylor 2000).

Alternatively, considering only the tetrahedral element I, the 3-D shape function (5) can also be expressed as follows (Li and Revesz 2004):

$$\begin{aligned} N_1(x,y,z) &= \frac{\mathcal{V}_1}{\mathcal{V}}, N_2(x,y,z) = \frac{\mathcal{V}_2}{\mathcal{V}}, \\ N_3(x,y,z) &= \frac{\mathcal{V}_3}{\mathcal{V}}, N_4(x,y,z) = \frac{\mathcal{V}_4}{\mathcal{V}}. \end{aligned} \tag{6}$$

$\mathcal{V}_1$, $\mathcal{V}_2$, $\mathcal{V}_3$ and $\mathcal{V}_4$ are the volumes of the four sub-tetrahedra $ww_2w_3w_4$, $w_1ww_3w_4$, $w_1w_2ww_4$, and $w_1w_2w_3w$, respectively, as shown in Fig. 5; and $\mathcal{V}$ is the volume of the outside tetrahedron $w_1w_2w_3w_4$.

Representing Interpolation Results in Constraint Databases

In traditional GIS, spatial data are represented in the relational data model, which is the most popular data model. Many database systems are based on the relational model, such as Oracle and MySQL. However, the relational model has disadvantages for some applications, which may lead to infinite relational databases (Revesz 2002). An infinite relational database means the database has relations with infinite number of tuples. In reality, only a finite

set of the tuples can be stored in a relation. Therefore, a finite set of tuples has to be extracted, which leads to data incompleteness. Using constraint databases can solve this infinity problem.

The sensory data of the ultraviolet radiation example in scientific fundamentals will be used to illustrate how to represent 2-D shape function spatial interpolation results in constraint databases. In this example, *Incoming*(y, t, u) is treated as if it contains a set of 2-D spatial data. Let *INCOMING*(y, t, u) be the constraint relation that represents the shape function interpolation result of the *Incoming* relation. Similarly, let *FILTER*(x,y,r) be the constraint relation that represents the shape function interpolation result of the *Filter* relation.

Triangulation of the set of sampled points is the first step to use 2-D shape functions. Figure 6 shows the Delaunay triangulations for the sample points in *Incoming*(y, t, u) and *Filter*(x, y, r) illustrated in Fig. 1.

The domain of a triangle can be represented by a conjunction C of three linear inequalities corresponding to the three sides of the triangle. Then, by the shape function (2), the value w of any point x, y inside a triangle can be represented by the following linear constraint tuple:

$$
\begin{aligned}
R(x, y, w) &: - C, \\
w = [((y_2 - y_3)w_1 + (y_3 - y_1)w_2 + (y_1 - y_2)w_3)/(2\mathcal{A})]x \\
+[((x_3 - x_2)w_1 + (x_1 - x_3)w_2 + (x_2 - x_1)w_3)/(2\mathcal{A})]y \\
+[((x_2 y_3 - x_3 y_2)w_1 + (x_3 y_1 - x_1 y_3)w_2 \\
+(x_1 y_2 - x_2 y_1)w_3)/(2\mathcal{A})].
\end{aligned}
$$

where $\mathcal{A}$ is a constant for the area value of the triangle. By representing the interpolation in each triangle by a constraint tuple, a constraint relation to represent the interpolation in the whole domain can be found in linear time.

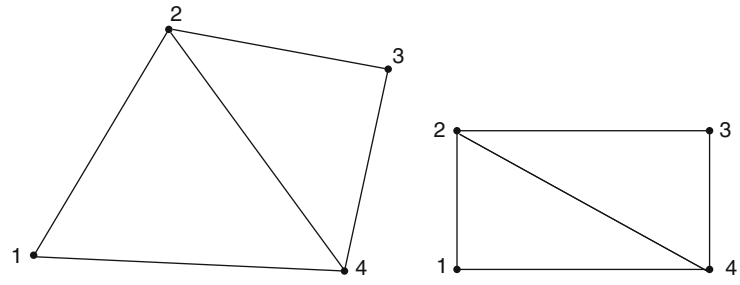
Table 5 illustrates the constraint representation for the interpolation result of *FILTER* using 2-D shape functions. The result of *INCOMING* is similar, and the details can be found in reference Revesz and Li (2002).

Applications Based on Shape Function Spatiotemporal Interpolation

There are two fundamentally different ways for spatiotemporal interpolation: reduction and extension (Li and Revesz 2002). These methods can be described briefly as follows:

Reduction This approach reduces the spatiotemporal interpolation problem to a regular spatial interpolation case. First, interpolate (using any 1-D interpolation in time) the measured value over time at each sample point. Then get spatiotemporal interpolation results by substituting the desired time instant into some regular spatial interpolation functions

Extension This approach deals with time as another dimension in space and extends the



Constraint Databases and Data Interpolation, Fig. 6 Delaunay triangulations for *Incoming* (*left*) and *Filter* (*right*)

Constraint Databases and Data Interpolation, Table 5 FILTER (x, y, r) using 2-D shape functions

X	Y	R	
x	y	r	$13x - 23y + 296 \geq 0, x \geq 2, y \geq 1,$
			$r = 0.0004x - 0.0031y + 0.1168$
x	y	r	$13x - 23y + 296 < q0, x < q25, y < q14,$
			$r = 0.0013x - 0.0038y + 0.1056$

spatiotemporal interpolation problem into a one-higher dimensional spatial interpolation problem

Reduction Approach

This approach for 2-D space and 1-D time problems can be described by two steps: 2-D spatial interpolation by shape functions for triangles and approximation in space and time. The second step, interpolation in space and time, can be implemented by combining a time shape function with the space approximation function (1).

Assume the value at node i at time t_1 is w_{i1}, and at time t_2 the value is w_{i2}. The value at the node i at any time between t_1 and t_2 can be interpolated using a 1-D time shape function in the following way:

$$w_i(t) = \frac{t_2 - t}{t_2 - t_1} w_{i1} + \frac{t - t_1}{t_2 - t_1} w_{i2}. \qquad (7)$$

Using the example shown in Fig. 2 and utilizing formulas (1) and (7), the interpolation function for any point constraint to element I at any time between t_1 and t_2 can be expressed as follows (Li and Revesz 2004):

$$\begin{aligned} w(x, y, t) &= N_1(x, y)\left[\frac{t_2 - t}{t_2 - t_1} w_{11} + \frac{t - t_1}{t_2 - t_1} w_{12}\right] + N_2(x, y)\left[\frac{t_2 - t}{t_2 - t_1} w_{21} + \frac{t - t_1}{t_2 - t_1} w_{22}\right] \\ &\quad + N_3(x, y)\left[\frac{t_2 - t}{t_2 - t_1} w_{31} + \frac{t - t_1}{t_2 - t_1} w_{32}\right] \\ &= \frac{t_2 - t}{t_2 - t_1} \cdot [N_1(x, y) w_{11} + N_2(x, y) w_{21} + N_3(x, y) w_{31}] \\ &\quad + \frac{t - t_1}{t_2 - t_1} \cdot [N_1(x, y) w_{12} + N_2(x, y) w_{22} + N_3(x, y) w_{32}]. \end{aligned}$$

The reduction approach for 3-D space and 1-D time problems can be developed in a similar way by combining the 3-D interpolation formula (4) and the 1-D shape function (7). Using the example shown in Fig. 4, the interpolation function for any point constraint to the sub-domain I at any time between t_1 and t_2 can be expressed as follows (Li and Revesz 2004):

$$\begin{aligned} w(x, y, z, t) = \quad & N_1(x, y, z)\left[\tfrac{t_2 - t}{t_2 - t_1} w_{11} + \tfrac{t - t_1}{t_2 - t_1} w_{12}\right] + N_2(x, y, z)\left[\tfrac{t_2 - t}{t_2 - t_1} w_{21} + \tfrac{t - t_1}{t_2 - t_1} w_{22}\right] \\ & + N_3(x, y, z)\left[\tfrac{t_2 - t}{t_2 - t_1} w_{31} + \tfrac{t - t_1}{t_2 - t_1} w_{32}\right] + N_4(x, y, z)\left[\tfrac{t_2 - t}{t_2 - t_1} w_{41} + \tfrac{t - t_1}{t_2 - t_1} w_{42}\right] \\ = \quad & \tfrac{t_2 - t}{t_2 - t_1}[N_1(x, y, z) w_{11} + N_2(x, y, z) w_{21} + N_3(x, y, z) w_{31} + N_4(x, y, z) w_{41}] \\ & + \tfrac{t - t_1}{t_2 - t_1}[N_1(x, y, z) w_{12} + N_2(x, y, z) w_{22} + N_3(x, y, z) w_{32} + N_4(x, y, z) w_{42}]. \end{aligned} \qquad (8)$$

Since the 2-D/3-D space shape functions and the 1-D time shape function are linear, the spatiotemporal interpolation function (110) is not linear but quadratic.

Extension Approach

For 2-D space and 1-D time problems, this method treats *time* as a regular third dimension. Since it extends 2-D problems to 3-D problems, this method is very similar to the linear approximation by 3-D shape functions for tetrahedra. The only modification is to substitute the variable z in Eqs. (4), (5), and (6) by the time variable t.

For 3-D space and 1-D time problems, this method treats *time* as a regular fourth dimension. New linear 4-D shape functions based on 4-D Delaunay tessellation can be developed to solve this problem. See reference Li (2003) for details on the 4-D shape functions.

Representing Interpolation Results in Constraint Databases

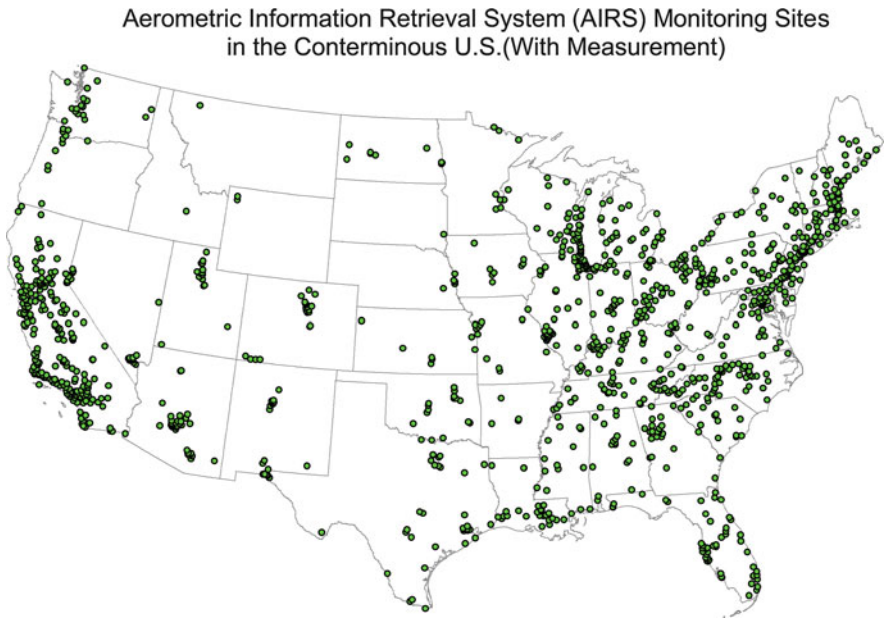
The previous section pointed out the infinity problem for relational databases to represent spatial data. The relational data model shows more disadvantages when handling spatiotemporal data. For example, using the relational model, the current contents of a database (database instance) is a snapshot of the data at a given instant in time. When representing spatiotemporal data, frequent updates have to be performed in order to keep the database instance up to date, which erases the previous database instance. Therefore, the information in the past will be lost. This irrecoverable problem makes the relational data model impractical for handling spatiotemporal data. Using constraint data model can solve this problem. A set of Aerometric Information Retrieval System (AIRS) data will be used to illustrate how spatiotemporal interpolation data can be represented accurately and efficiently in constraint databases.

The experimental AIRS data is a set of data with annual ozone concentration measurements in the conterminous USA (website www.epa.gov/airmarkets/cmap/data/category1.html). AIRS is a computer-based repository of information about airborne pollution in the US and various World Health Organization (WHO) member countries. The system is administered by the US Environmental Protection Agency (EPA). The data coverage contains point locations of the monitoring sites for which AIRS data are collected, the annual concentration level measurements of ozone (O3), and the years of the measurement. Several datasets from the US EPA (website http://cfpub.epa.gov/gdm) were obtained and reorganized into a dataset with schema (x, y, t, w), where x and y attributes are the longitude and latitude coordinates of monitoring site locations, t is the year of the ozone measurement, and w is the O34MAX (4th Max of 1-h Values for O3) value of the ozone measurement. The original dataset has many zero entries for ozone values, which means no measurements available at a particular site. After filtering out all the zero entries from the original dataset, there are 1209 sites left with measurements. Figure 7 shows the locations of the 1209 monitoring sites Li et al. (2006).

Among the 1209 monitoring sites with measurements, some sites have complete measurements of yearly ozone values from 1994 to 1999, while the other sites have only partial records. For example, some sites only have measurements of ozone values in 1998 and 1999. In total, there are 6135 ozone value measurements recorded. Each measurement corresponds to the ozone value at a spatiotemporal point (x, y, t), where (x, y) is the location of one of the 1209 monitoring sites, and t is a year between 1994 and 1999.

The spatiotemporal interpolation extension method based on 3-D shape functions is implemented into a Matlab program and applied to the AIRS ozone data. The Matlab function *delaunayn* is used to compute the tetrahedral mesh with the 6135 spatiotemporal points as corner vertices. There are 30,897 tetrahedra in the resulting mesh. Using the mesh and the original 6135 original ozone values measured at its corner vertices, the annual ozone value at any location and year can be interpolated, as long as the spatiotemporal point is located inside the domain of the tetrahedral mesh.

Since the 3-D shape function based spatiotemporal interpolation Eq. (4) is linear, the interpolation results can be stored in a linear constraint database. Suppose the constraint relation *Ozone_interp* is used to store the interpolation results. Table 6 shows one sample tuple of *Ozone_interp*. The other omitted tuples



Constraint Databases and Data Interpolation, Fig. 7 1209 AIRS monitoring sites with measurements in the conterminous US

Constraint Databases and Data Interpolation, Table 6 The constraint relation *Ozone_interp*(x, y, t, w), which stores the 3-D shape function interpolation results of the ozone data

X	Y	R	W	
				$0.002532x + 0.003385y + 0.000511t \geq 1,$
				$0.002709x + 0.003430y + 0.000517t \geq 1,$
				$0.002659x + 0.003593y + 0.000511t < q1,$
				$0.002507x + 0.003175y + 0.000515t < q1,$
x	y	t	w	$v = 0.0127,$
				$v_1 = 1/6 \mid 1.71x + 2.17y + 0.35t - 682.87 \mid,$
				$v_2 = 1/6 \mid 2.10x + 2.84y + 0.40t - 790.39 \mid,$
				$v_3 = 1/6 \mid 1.28x + 1.63y + 0.24t - 474.05 \mid,$
				$v_4 = 1/6 \mid 2.53x + 3.38y + 0.51t - 999.13 \mid,$
				$wv = 0.063v_1 + 0.087v_2 + 0.096v_3 + 0.074v_4$
x	y	t	w	$\vdots$
				$\vdots$

are of similar format. Since there are 30,897 tetrahedra generated in the tetrahedral mesh, there should be 30,897 tuples in *Ozone_interp*.

The tuple shown in Table 6 corresponds to the interpolation results of all the points located in the tetrahedron with corner vertices (−68.709, 45.217, 1996), (−68.672, 44.736, 1999), (−67.594, 44.534, 1995), and (−69.214, 45.164, 1999). The ozone values measured at these four points are 0.063, 0.087, 0.096, and 0.074, respectively. In this constraint tuple, there are 10 constraints. The relationship among these constraints is AND. The first four constraints define the four facets of the tetrahedron, the next five constraints give the volume values, and the last constraint is the interpolation function.

Applications Based on IDW Spatial Interpolation

Inverse distance weighting (IDW) interpolation (Shepard 1968) assumes that each measured point has a local influence that diminishes with distance. Thus, points in the near neighborhood are given high weights, whereas points at a far distance are given small weights. Reference Revesz and Li (2003) uses IDW to visualize spatial interpolation data.

The general formula of IDW interpolation for 2-D problems is the following:

$$w(x, y) = \sum_{i=1}^{N} \lambda_i w_i$$
$$\lambda_i = \frac{(\frac{1}{d_i})^p}{\sum_{k=1}^{N} (\frac{1}{d_k})^p} \tag{9}$$

where $w(x, y)$ is the predicted value at location (x, y), N is the number of nearest known points surrounding (x, y), λ_i are the weights assigned to each known point value w_i at location (x_i, y_i), d_i are the 2-D Euclidean distances between each (x_i, y_i) and (x, y), and p is the exponent, which influences the weighting of w_i on w.

For 3-D problems, the IDW interpolation function is similar as formula (9), by measuring 3-D Euclidean distances for d_i.

Representing Interpolation Results in Constraint Databases

To represent the IDW interpolation, the nearest neighbors for a given point should be found. The idea of higher-order Voronoi diagrams (or kth order Voronoi diagrams) can be borrowed from computational geometry to help find the nearest neighbors. Higher-order Voronoi diagrams generalize ordinary Voronoi diagrams by dealing with k closest points. The ordinary Voronoi diagram of a finite set S of points in the plane is a partition of the plane so that each region of the partition is the locus of points which are closer to one member of S than to any other member (Preparata and Shamos 1985). The higher-order Voronoi diagram of a finite set S of points in the plane is a partition of the plane into regions such that points in each region have the same closest members of S. As in an ordinary Voronoi diagram, each Voronoi region is still convex in a higher-order Voronoi diagram. From the definition of higher-order Voronoi diagrams, it is obvious to see that the problem of finding the k closest neighbors for a given point in the whole domain, which is closely related to the IDW interpolation method with $N = k$, is equivalent to constructing kth order Voronoi diagrams.

Although higher-order Voronoi diagrams are very difficult to create by imperative languages, such as C, C++, and Java, they can be easily constructed by declarative languages, such as Datalog. For example, a second-order Voronoi region for points (x_1, y_1), (x_2, y_2) can be expressed in Datalog as follows.

At first, let $P(x, y)$ be a relation that stores all the points in the whole domain. Also let $Dist(x, y, x_1, y_1, d_1)$ be a Euclidean distance relation where d_1 is the distance between (x, y) and (x_1, y_1). It can be expressed in Datalog as:

$$Dist(x, y, x_1, y_1, d_1) :$$
$$-d_1 = \sqrt{(x - x_1)^2 + (y - y_1)^2}.$$

Note that any point (x, y) in the plane does *not* belong to the second-order Voronoi region of the sample points (x_1, y_1) and (x_2, y_2) if there exists another sample point (x_3, y_3) such that (x, y) is closer to (x_3, y_3) than to either (x_1, y_1) or (x_2, y_2). Using this idea, the complement can be expressed as follows:

$$\begin{aligned} Not\_2Vor(x, y, x_1, y_1, x_2, y_2) : -P(x_3, y_3), \\ Dist(x, y, x_1, y_1, d_1), \\ Dist(x, y, x_3, y_3, d_3), \\ d_1 > d_3. \end{aligned}$$

$$\begin{aligned} Not\_2Vor(x, y, x_1, y_1, x_2, y_2) : -P(x_3, y_3), \\ Dist(x, y, x_2, y_2, d_2), \\ Dist(x, y, x_3, y_3, d_3), \\ d_2 > d_3. \end{aligned}$$

Finally, the negation of the above can be taken to get the second-order Voronoi region as follows:

$$2Vor(x, y, x_1, y_1, x_2, y_2) \\ :- not\ Not\_2Vor(x, y, x_1, y_1, x_2, y_2)\ . \tag{10}$$

The second-order Voronoi diagram will be the union of all the nonempty second-order Voronoi regions. Similarly to the second order, any kth-order Voronoi diagram can be constructed.

After finding the closest neighbors for each point by constructing higher-order Voronoi diagrams, IDW interpolation in constraint databases can be represented. The representation can be obtained by constructing the appropriate Nth-order Voronoi diagram and using formula (9).

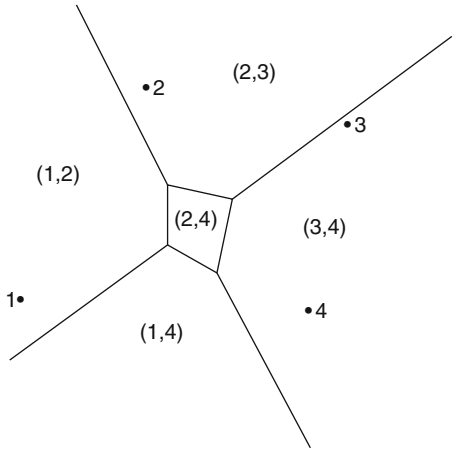
Based on formula (10), assume that the second-order Voronoi region for points (x_1, y_1), (x_2, y_2) is stored by the relation $Vor_2nd(x, y, x_1, y_1, x_2, y_2)$, which is a conjunction C of some linear inequalities corresponding to the edges of the Voronoi region. Then, using IDW interpolation with $N = 2$ and $p = 2$, the value w of any point (x, y) inside the Voronoi region can be expressed by the constraint tuple as follows:

$$\begin{aligned} R(x, y, w) :- &\big((x - x_2)^2 + (y - y_2)^2 \\ &+ (x - x_1)^2 + (y - y_1)^2\big)w \\ &= \big((x - x_2)^2 + (y - y_2)^2\big)w_1 \\ &+ \big((x - x_1)^2 + (y - y_1)^2\big)w_2, \\ &Vor_2nd(x, y, x_1, y_1, x_2, y_2). \end{aligned} \tag{11}$$

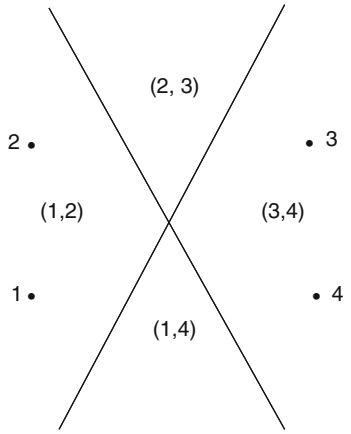
or equivalently as,

$$\begin{aligned} R(x, y, w) :- &\big((x - x_2)^2 + (y - y_2)^2 \\ &+ (x - x_1)^2 + (y - y_1)^2\big)w \\ &= \big((x - x_2)^2 + (y - y_2)^2\big)w_1 \\ &+ \big((x - x_1)^2 + (y - y_1)^2\big)w_2, \\ &C. \end{aligned} \tag{12}$$

In the above polynomial constraint relation, there are three variables: x, y, and w. The highest-order terms in the relation are $2x^2w$ and $2y^2w$, which are both cubic. Therefore, this is a cubic constraint tuple. The sensory data of the ultraviolet radiation example in scientific fundamentals will be used to illustrate how to represent IDW spatial interpolation results in constraint databases. Figures 8 and 9 show the second-order Voronoi diagrams for the sample points in *Incoming(y, t, u)* and *Filter(x, y, r)*, respectively. Please note that some second-order Voronoi regions are empty. For example, there is no (1,3) region in Fig. 8, and there are no (1,3) and (2,4) regions in Fig. 9 (Revesz and Li 2002).



Constraint Databases and Data Interpolation, Fig. 8
The 2nd order Voronoi diagram for *Incoming*



Constraint Databases and Data Interpolation, Fig. 9
The 2nd order Voronoi diagram for *Filter*

Let *INCOMING(y, t, u)* and *FILTER(x, y, r)* be the constraint relations that store the IDW interpolation results of *Incoming(y, t, u)* and *Filter(x, y, r)*. Based on formula (12), Table 7 shows the result of *FILTER*. Note that the four tuples in

Constraint Databases and Data Interpolation, Table 7 FILTER (x, y, r) using IDW

X	Y	R	
x	y	r	$2x - y - 20 < q0,\ 12x + 7y - 216 < q0,$
			$((x-2)^2 + (y-14)^2)0.9 + ((x-2)^2 + (y-1)^2)0.5$
			$= (2(x-2)^2 + (y-14)^2 + (y-1)^2)r$
x	y	r	$2x - y - 20 < q0,\ 12x + 7y - 216 < q0,$
			$((x-25)^2 + (y-1)^2)0.9 + ((x-2)^2 + (y-1)^2)0.8$
			$= (2(y-1)^2 + (x-25)^2 + (x-2)^2)r$
x	y	r	$2x - y - 20 \geq 0,\ 12x + 7y - 216 \geq 0,$
			$((x-25)^2 + (y-14)^2)0.8 + ((x-25)^2 + (y-1)^2)0.3$
			$= (2(x-25)^2 + (y-14)^2 + (y-1)^2)r$
x	y	r	$2x - y - 20 < q0,\ 12x + 7y - 216 \geq 0,$
			$((x-25)^2 + (y-14)^2)0.5 + ((x-2)^2 + (y-14)^2)0.3$
			$= (2(y-14)^2 + (x-25)^2 + (x-2)^2)r$

Table 7 represent the four second-order Voronoi regions in Fig. 9. The result of *INCOMING* is similar and the details can be found in reference Li (2003).

Applications Based on IDW Spatiotemporal Interpolation

Similar as shape functions, IDW is originally a spatial interpolation method, and it can be extended by reduction and extension approaches to solve spatiotemporal interpolation problems (Li 2003).

Reduction Approach

This approach first finds the nearest neighbors of for each unsampled point and calculates the corresponding weights λ_i. Then, it calculates for each neighbor the value at time t by some time interpolation method. If 1-D shape function interpolation in time is used, the time interpolation will be similar to (7). The formula for this approach can be expressed as:

$$w(x, y, t) = \sum_{i=1}^{N} \lambda_i w_i(t), \quad \lambda_i = \frac{(\frac{1}{d_i})^p}{\sum_{k=1}^{N} (\frac{1}{d_k})^p} \tag{13}$$

where $d_i = \sqrt{(x_i - x)^2 + (y_i - y)^2}$ and $w_i(t) = \frac{t_{i2} - t}{t_{i2} - t_{i1}} w_{i1} + \frac{t - t_{i1}}{t_{i2} - t_{i1}} w_{i2}$.

Extension Approach

Since this method treats time as a third dimension, the IDW-based spatiotemporal formula is in the form of (9) with

$$d_i = \sqrt{(x_i - x)^2 + (y_i - y)^2 + (t_i - t)^2}.$$

Future Directions

Interesting directions for the future work could be to represent more interpolation methods in spatial constraint databases, apply more interesting data sets to the interpolation methods, compare the performances of the methods, and animation/visualize the interpolation results.

Cross-References

- Constraint Database Queries
- Constraint Databases and Moving Objects
- Constraint Databases, Spatial
- Voronoi Diagram

References

Gao J, Revesz P (2006) Voting prediction using new spatiotemporal interpolation methods. In: Proceedings of the seventh annual international conference on digital government research, San Diego

Li J, Narayanan R, Revesz P (2003) A shape-based approach to change detection and information mining in remote sensing. In: Chen CH (ed) Frontiers of remote sensing information processing. WSP, Singapore/River Edge, pp 63–86

Li L (2003) Spatiotemporal interpolation methods in GIS. Ph.D thesis, University of Nebraska-Lincoln, Lincoln

Li L, Li Y, Piltner R (2004) A new shape function based spatiotemporal interpolation method. In: Proceedings of the first international symposium on constraint databases 2004. Lecture notes in computer science, vol 3074. Springer, Berlin/Heidelberg/New York, pp 25–39

Li L, Revesz P (2002) A comparison of spatio-temporal interpolation methods. In: Proceedings of the second international conference on GIScience 2002. Lecture notes in computer science, vol 2478. Springer, Berlin, Heidelberg, New York, pp 145–160

Li L, Revesz P (2004) Interpolation methods for spatio-temporal geographic data. J Comput Environ Urban Syst 28(3):201–227

Li L, Zhang X, Piltner R (2006) A spatiotemporal database for ozone in the conterminous U.S. In: Proceedings of the thirteenth international symposium on temporal representation and reasoning, Budapest. IEEE, pp 168–176

Preparata FP, Shamos MI (1985) Computational geometry: an introduction. Springer, Berlin/Heidelberg/New York

Revesz P (2002) Introduction to constraint databases. Springer, New York

Revesz P, Li L (2002) Constraint-based visualization of spatial interpolation data. In: Proceedings of the sixth international conference on information visualization. IEEE Press, London/England, pp 563–569

Revesz P, Li L (2002) Representation and querying of interpolation data in constraint databases. In: Proceedings of the second national conference on digital government research. Los Angeles, California, pp 225–228

Revesz P, Li L (2003) Constraint-based visualization of spatiotemporal databases. In: Advances in geometric modeling, chapter 20. Wiley, England, pp 263–276

Revesz P, Wu S (2006) Spatiotemporal reasoning about epidemiological data, Artificial Intelligence in Medicine, 38(2):157–170

Shepard D (1968) A two-dimensional interpolation function for irregularly spaced data. In: 23nd national conference ACM, Las Vegas. ACM, pp 517–524

Zienkiewics OC, Taylor RL (2000) Finite element method. The basis, vol 1. Butterworth Heinemann, London

Constraint Databases and Moving Objects

Lixin Li
Department of Computer Sciences, Georgia Southern University, Statesboro, GA, USA

Synonyms

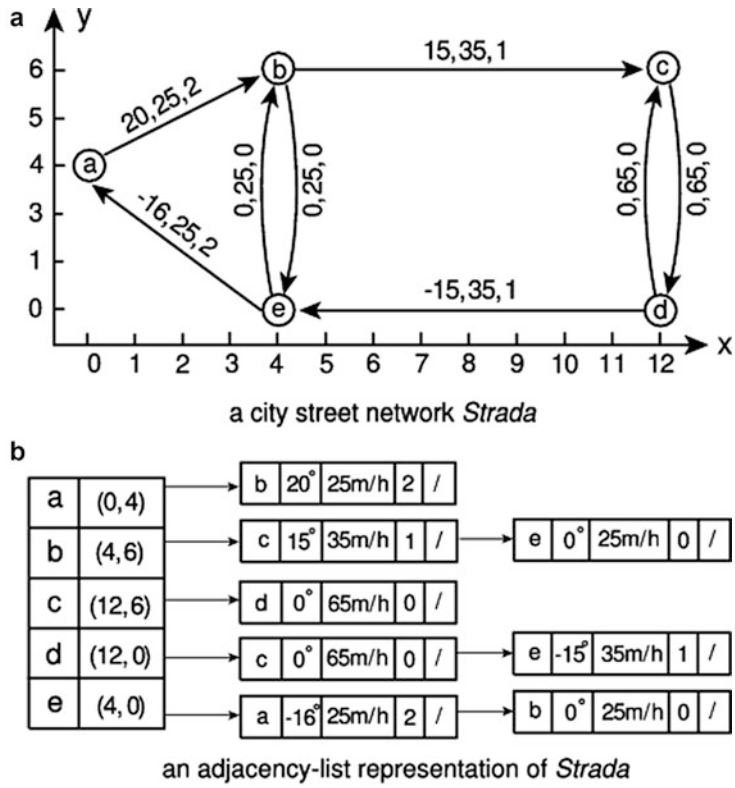
Continuously changing maps; Moving object constraint databases; Moving points; Moving regions; Spatio-temporal objects

Definition

Moving objects can be represented in spatial constraint databases, given the trajectory of the objects.

Historical Background

In general, there are two fundamental ways to abstract moving objects: *moving points* and *moving regions* (Güting 1994). Moving points can describe objects for which only the time-dependent position is of interest, while moving regions are able to describe those for which both time-dependent position and spatial extent are of interest. Parametric Rectangles (PReSTO) (Revesz and Cai 2002) belong to moving region abstraction. They use growing or shrinking parametric rectangles to model spatiotemporal objects in constraint databases. One advantage in using moving regions is the ability to represent spatial dimensions of objects. Moving points can also be a proper abstraction for moving objects,



Constraint Databases and Moving Objects, Fig. 1 A city street network *Strada* and its adjacency-list representation

such as people, animals, stars, cars, planes, ships, and missiles (Erwig et al. 1997). In reference Saglio and Moreira (1999), Saglio and Moreira argued that moving points are possibly the simplest class of continuously changing spatial objects and there are many systems, including those dealing with the position of cars, ships, or planes, which only need to keep the position of the objects.

Continuously changing maps are special cases of moving regions. There are many applications that need to be visualized by continuously changing maps which will be illustrated in Key Applications.

Scientific Fundamentals

The following example describes how to illustrate the movements of snow removal vehicles in a city street network by moving points in spatial constraint databases.

Suppose that snow removal vehicles are going to clear the snow on the streets of a city. Adjacency-list representation (Cormen et al. 1999) of directed weighted graphs can be applied to model such networks. The city street network *Strada* is shown in Fig. 1a and its adjacency-list representation list is shown in Fig. 1b. Each street has the following attributes: slope, speed limit, and snow clearance priority (the less the value, the higher the priority). These three attributes are shown as labels of each edge in Fig. 1a. They are also displayed in the property fields of each node in Fig. 1b. For example, for the street segment $\overrightarrow{s_{bc}}$, the slope is 15°, the speed limit is 35 mph, and the clearance priority value is 1. The movements of snow removal vehicles in *Strada* can be represented in Fig. 2 by eight Datalog rules in constraint databases.

Key Applications

There are many applications that need to be modeled as moving objects, such as continuously changing maps. Constraint databases are capable of handling such applications. The MLPQ

```
Snow_removal(from,to,x,y,t,priority) :- from = "a", to = "b", x =16.8t, y = 4+8.4t,
                                        t >= 0, t <= 14, priority = 2.
Snow_removal(from,to,x,y,t,priority) :- from = "b", to = "c", x = 4+19.3t, y = 6,
                                        t >= 0, t <= 25, priority = 1.
Snow_removal(from,to,x,y,t,priority) :- from = "b", to = "e", x = 4, y = 6-20t,
                                        t >= 0, t <= 18, priority = 0.
Snow_removal(from,to,x,y,t,priority) :- from = "c", to = "d", x = 12, y = 6-20t,
                                        t >= 0, t <= 18, priority = 0.
Snow_removal(from,to,x,y,t,priority) :- from = "d", to = "c", x = 12, y = 20t,
                                        t >= 0, t <= 18, priority = 0.
Snow_removal(from,to,x,y,t,priority) :- from = "d", to = "e", x = 12-20.7t, y = 0,
                                        t >= 0, t <= 23, priority = 1.
Snow_removal(from,to,x,y,t,priority) :- from = "e", to = "a", x = 4-14.7t, y = 14.7t,
                                        t >= 0, t <= 16, priority = 2.
Snow_removal(from,to,x,y,t,priority) :- from = "e", to = "b", x = 4, y = 20t,
                                        t >= 0, t <= 18, priority = 0.
```

Constraint Databases and Moving Objects, Fig. 2 Constraint databases representation of the snow removal vehicles for *Strada*

(Management of Linear Programming Queries) system is a good example. The MLPQ system is a constraint database system for linear constraint databases (Kanjamala et al. 1998; Revesz 2002; Revesz and Li 1997). This system has a graphic user interface (GUI) which supports Datalog-based and icon-based queries as well as visualization and animations. The MLPQ system can outdo the popular ArcGIS system by powerful queries (such as recursive queries) and the ability to display continuously changing maps. A few examples are given below.

SPI Spatiotemporal Data

The point-based spatiotemporal relation *Drought_Point (x, y, year, SPI)* stores the average yearly SPI (Standardized Precipitation Index) values sampled by 48 major weather stations in Nebraska from 1992 to 2002. SPI is a common and simple measure of drought which is based solely on the probability of precipitation for a given time period. Values of SPI range from 2.00 and above (extremely wet) to −2.00 and less (extremely dry), with near normal conditions ranging from 0.99 to −0.99. A drought event is defined when the SPI is continuously negative and reaches a value of −1.0 or less, and continues until the SPI becomes positive. The

Constraint Databases and Moving Objects, Table 1 A point-based spatiotemporal relation

Drought_Point			
x (easting)	y (northing)	Year	SPI
−315515.56	2178768.67	1992	0.27
−315515.56	2178768.67	1993	−0.17
⋮	⋮	⋮	⋮

Drought_Point relation, as shown in Table 1, was obtained from the Unified Climate Access Network (UCAN) (Li 2003).

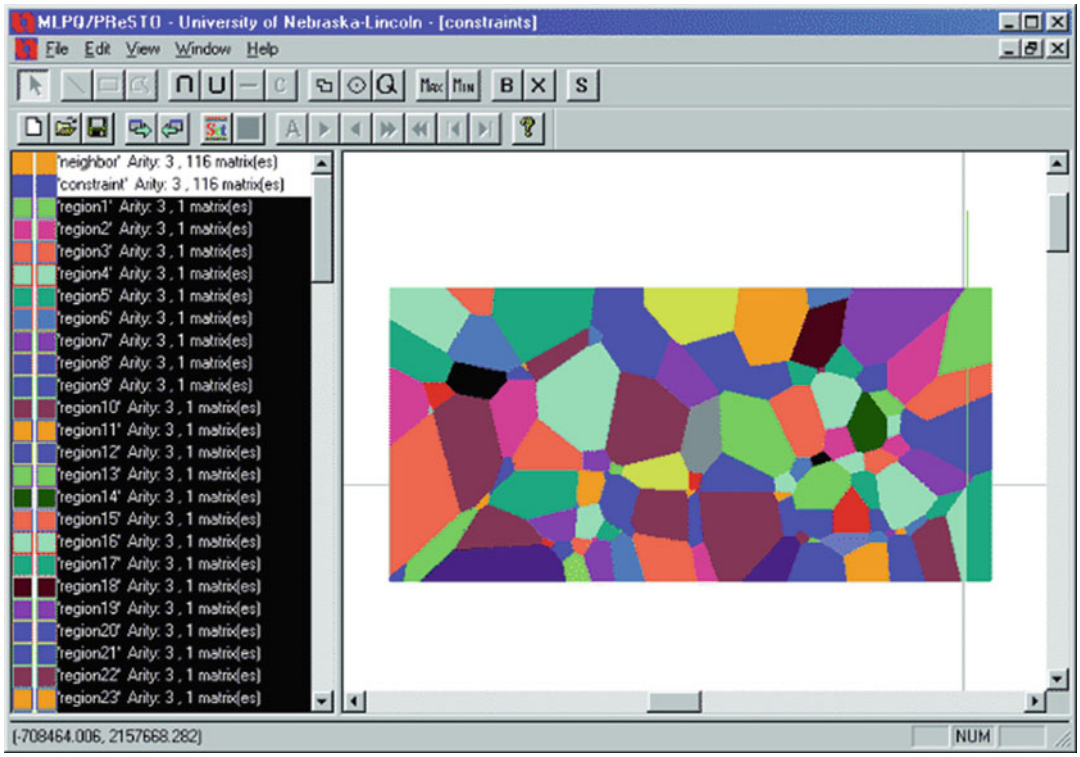
Assume that in the point-based spatiotemporal relation *Drought_Point*, the 48 weather stations have not changed their locations for the last 10 years and measured SPI values every year. The spatial and temporal parts of the 2nd-order Voronoi region-based relation of *Drought_Point* are shown in Table 2.

Continuously changing maps in MLPQ can be used to visualize the 2nd-order Voronoi diagrams. Users need to push the color animation button in the MLPQ GUI and input the following three parameters: the beginning time instance, ending time instance and step size. Then, the color of each region of the map will be animated according to its value at a specific time instance. Figure 3 shows the 2nd-order Voronoi diagram for the 48 weather stations in Nebraska at the snapshot when $t = 1992$ (Li 2003).

Constraint Databases and Moving Objects, Table 2 A 2nd-order Voronoi region-based database

Drought_Vo2_Space	
$(x_1, y_1), (x_2, y_2)$	Boundary
$(-9820.18, 1929867.40), (-42164.88, 1915035.54)$	$(-17122.48, 2203344.58)$, (3014.51, 2227674.50) (330-51.50, 2227674.50), (33051.5, 2140801.51)
⋮	⋮

Drought_Vo2_Time		
$(x_1, y_1), (x_2, y_2)$	Year	avgSPI
$(-9820.18, 1929867.4), (-42164.88, 1915035.54)$	1992	-0.47
$(-9820.18, 1929867.4), (-42164.88, 1915035.54)$	1993	0.71
⋮	⋮	⋮
$(-507929.66, 2216998.17), (-247864.81, 1946777.44)$	2002	-0.03



Constraint Databases and Moving Objects, Fig. 3 The 2rd order Voronoi diagram for 48 weather stations in Nebraska which consists of 116 regions

NASS Spatiotemporal Data

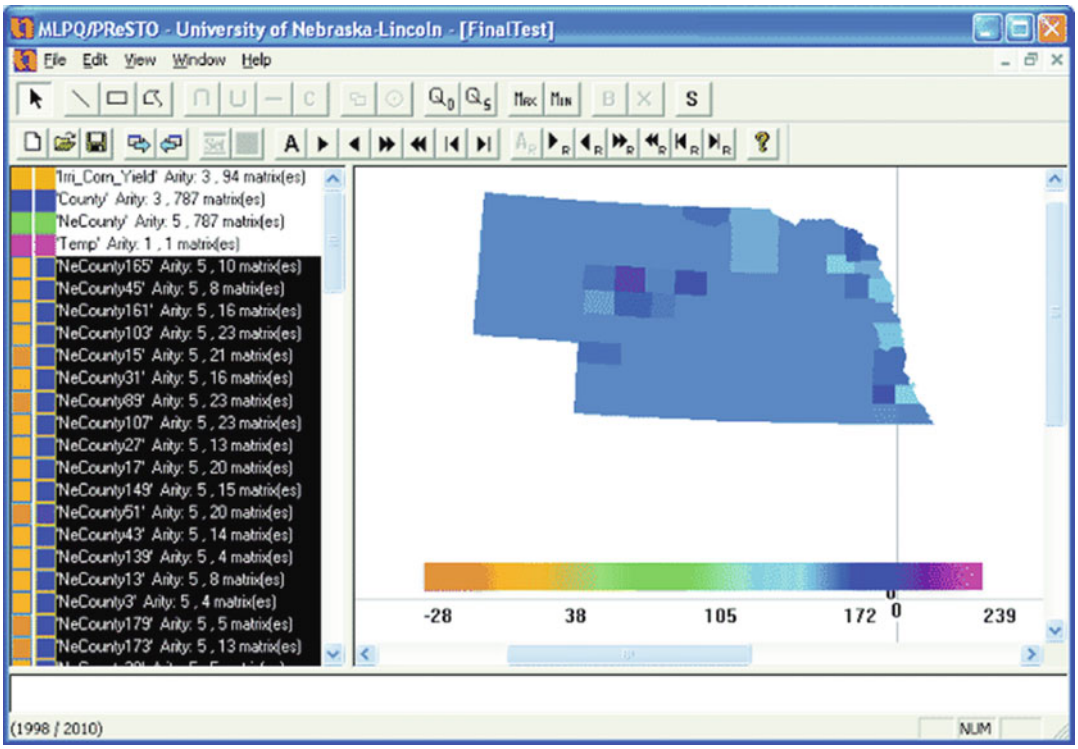
A NASS (National Agricultural Statistics Service) region-based spatiotemporal database shows the yearly corn yield and production in each county of the state of Nebraska. The spatial part of the database is shown in the upper half of Table 3 which uses the vector representation of counties in Nebraska, while the temporal part is shown in the lower half of Table 3 (Li and Revesz 2003).

Continuously changing maps can be used to animate the total corn yield in each county in Nebraska during a given period. First, each county

Constraint Databases and Moving Objects, Table 3 A region-based spatiotemporal database with separate spatial and temporal relations

Nebraska_Corn_Space_Region	
County	Boundary
1	(−656160.3, 600676.8), (−652484.0, 643920.3), (−607691.1, 639747.6), (−608934.8, 615649.0), (−607875.6, 615485.8), (−610542.0, 576509.1), (−607662.7, 576138.5), (−611226.9, 537468.5), (−607807.7, 536762.1), (−608521.1, 527084.0), (−660885.4, 531441.2), (−661759.8,532153.1)
⋮	⋮

Nebraska_Corn_Time_Region					
County	Year	Practice	Acres	Yield	Production
1	1947	Irrigated	2700	49	132300
1	1947	Non-irrigated	81670	18	1470060
1	1947	Total	84370	19	1602360
⋮	⋮	⋮	⋮	⋮	⋮



Constraint Databases and Moving Objects, Fig. 4 A snapshot of continuously changing maps for county-based corn yield in Nebraska when $t = 1998$

polygon needs to be represented in MLPQ. Although such county vector data in the US are usually available in ArcView shape file format, a program can be implemented to convert ArcView shape files to MLPQ input text files. The conversion from MLPQ files to shape files can also be implemented. Figure 4 shows the snapshot during the color map animation when $t = 1998$ (Li 2003).

Future Directions

Interesting directions for future work could be to continue the discovery of moving objects that are difficult to model in relational databases but can be conveniently modeled in spatial constraint databases, and to extend the MLPQ system so as to improve the visualization/animation power of the system.

Cross-References

- Constraint Database Queries
- Constraint Databases and Data Interpolation
- Constraint Databases, Spatial
- MLPQ Spatial Constraint Database System
- Visualization of Spatial Constraint Databases

References

Cormen TH, Leiserson CE, Rivest R (1999) Introduction to algorithms. McGraw-Hill, New York

Erwig M, Güting R, Schneider M, Vazirgiannis M (1997) Spatio-temporal data types: an approach to modeling and querying moving objects in databases. ChoroChronos Research Project, Technical report CH-97-08

Güting R (1994) An introduction to spatial database systems. VLDB J 3(4):357–399

Kanjamala P, Revesz P, Wang P (1998) MLPQ/GIS: a GIS using linear constraint databases. In: Prabhu CSR (ed) Proceedings of the 9th COMAD international conference on management of data, Hyderabad, pp 389–393

Li L (2003) Spatiotemporal interpolation methods in GIS. Ph.D thesis, University of Nebraska-Lincoln, Lincoln

Li L, Revesz P (2003) The relationship among GIS-oriented spatiotemporal databases. In: Proceedings of the third national conference on digital government research, Boston

Revesz P (2002) Introduction to constraint databases. Springer, New York

Revesz P, Cai M (2002) Efficient querying of periodic spatiotemporal objects. Ann Math Artif Intell 36(4):437–457

Revesz P, Li Y (1997) MLPQ: a linear constraint database system with aggregate operators. In: Proceedings of the 1st international database engineering and applications symposium. IEEE Press, Washington, DC, pp 132–137

Saglio J-M, Moreira J (1999) Oporto: a realistic scenario generator for moving objects. In: Proceedings of the DEXA'99 workshop on spatio-temporal data models and languages (STDML), Florence, pp 426–432

Constraint Databases, Spatial

Peter Z. Revesz
Department of Computer Science and Engineering, University of Nebraska-Lincoln, Lincoln, NE, USA

Synonyms

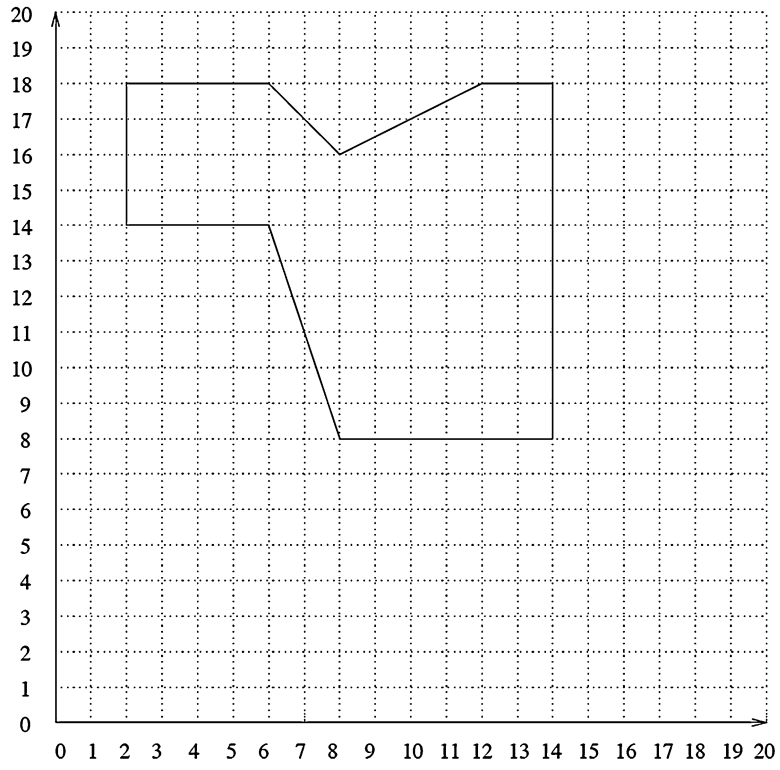
Databases, Relational; Query, Datalog

Definition

Spatial constraint databases form a generalization of relational databases for the purpose of representing spatial and spatiotemporal data. Whereas in a relational database each table is a finite set of tuples, in a spatial constraint database, each table is a finite set of quantifier-free conjunctions of atomic constraints. In spatial constraint databases, the most frequent type of atomic constraints used are linear equations and linear inequalities. The variables of the atomic constraints correspond to the attributes in the relation; hence, they are called attribute variables.

As an example from Revesz (2010), consider the highly simplified map of the town of Lincoln, Nebraska, shown in Fig. 1.

This map can be represented in a spatial constraint database with linear inequality constraints as follows (Table 1).



Constraint Databases, Spatial, Fig. 1 A map of Lincoln, Nebraska

Constraint Databases, Spatial, Table 1 Lincoln

Name	X	Y	
Lincoln	x	y	$y \geq x + 8,\ y \geq 14,\ x \geq 2,\ y \leq 18,\ y \leq -x + 24$
Lincoln	x	y	$y \leq x + 8,\ y \leq 0.5x + 12,\ y \leq 18,\ x \leq 14,\ y \geq 8,\ y \geq -3x + 32$

In the above the attribute variables x and y represent the longitude and latitude, respectively, as measured in units from the $(0, 0)$ point in the above map. In general, any polygonal shape can be represented by first dividing it into a set of convex polygons and then representing each convex polygon with n sides by a conjunction of n linear inequality constraints. Note that the above town map is a concave polygon, but it can be divided along the line $y = x + 8$ into two convex polygons. The convex pentagon above line $y = x + 8$ is represented by the first row of the constraint table, while the convex hexagon below line $y = x + 8$ is represented by the second row of the constraint table. Within any row, the atomic constraints are connected by commas, which simply mean conjunction. While the atomic constraints can be given in any order in each row, it is customary to present them in an order that corresponds to a clockwise ordering of the sides of the convex polygon that they together represent.

Historical Background

Constraint databases, including spatial constraint databases, were proposed by Kanellakis et al. in 1990. A much-delayed journal version of their original conference paper appeared in Kanellakis et al. (1995). These papers considered a number of constraint database query languages and challenged researchers to investigate further their properties. Benedikt et al. (1998) showed that relational calculus queries of constraint databases when the constraint database contains polynomial

constraints over the reals cannot express even simple Datalog expressible queries. On the other hand, Datalog queries with linear constraints can already express some computationally hard or even undecidable problems. Only in special cases, such as with gap-order constraints of the form $xy \geq c$ where x and y are integer or rational variables and c is a nonnegative constant, can an algorithm be given for evaluating Datalog queries (Revesz 1993).

The above results influenced researchers to implement several spatial constraint database systems with non-recursive query languages, usually some variation of non-recursive SQL and linear equality and linear inequality constraints. These systems include, in historical order, the MLPQ system (Brodsky et al. 1997), the CCUBE system (Brodsky et al. 1997), the DEDALE system (Grumbach et al. 1998), and the CQA/CDB system (Goldin et al. 2003). The MLPQ system implements both SQL and Datalog queries.

Constraint databases are reviewed in a number of books. Chapter 5.6 of Abiteboul et al. (1995), a standard reference in database theory, is a compact description of the main ideas of constraint databases. Kuper et al. (2000) is a collection of research articles devoted to constraint databases. It is a good introduction to already advanced researchers. Revesz (2002) is the standard textbook for the subject. It is used at many universities. Chapter 4 of Rigaux et al. (2002), which is an excellent source on all aspects of spatial databases, is devoted exclusively to constraint databases. Chapter 6 of Gting and Schneider (2005), which is a sourcebook on moving object databases, is also devoted exclusively to constraint databases.

Scientific Fundamentals

The semantics of logical models of a spatial constraint database table is a relational database that contains all the rows that can be obtained by substituting values into the attribute variables of any constraint row such that the conjunction of constraints in that row is true. For example, it is easy to see that the semantics of the spatial constraint database table Lincoln is a relation that contains all (x, y) points that belong to the town map. Since there are an infinite number of such (x, y) points when x and y are real numbers, spatial constraint databases are also called finitely representable infinite relational databases.

Spatial constraint databases can represent not only areas but also boundaries by using linear equality constraints. Representing the boundary of an n-ary (concave or convex) polygonal area requires n rows in a constraint table. For example, the boundary of the town of Lincoln, Nebraska, can be represented as shown in Table 2.

In the above each range constraint of the form $a \leq x \leq b$ is an abbreviation of $a \leq x, x \leq b$ where x is any variable and a and b are constants.

Spatial constraint databases can be extended to higher dimensions. For example, a Z attribute for height or a T attribute for time can be added. As an example, suppose that Fig. 1 shows the map of Lincoln, Nebraska, in year 2000, and since then the town has expanded to the east continuously at the rate of one unit per year. Then the growing town area between years 2000 and 2007 can be represented as shown in Table 3.

Spatial constraint databases with polynomial constraints are also possible. With the increased complexity of constraints, more complex spatial and spatiotemporal objects can be represented. For example, suppose that an airplane flies over Lincoln, Nebraska. Its shadow can be represented as a spatial constraint database

Constraint Databases, Spatial, Table 2 Lincoln_Boundary

X	Y	
x	y	$x \geq 2,\ x \leq 6,\ y = 18$
x	y	$x \geq 6,\ x \leq 8,\ y = -x + 24$
x	y	$x \geq 8,\ x \leq 12,\ y = 0.5x + 12$
x	y	$x \geq 12,\ x \leq 14,\ y = 18$
x	y	$y \geq 8,\ y \leq 18,\ x = 14$
x	y	$x \geq 8,\ x \leq 14,\ y = 8$
x	y	$x \geq 6,\ x \leq 8,\ y = -3x + 32$
x	y	$x \geq 2,\ x \leq 6,\ y = 14$
x	y	$y \geq 14,\ y \leq 18,\ x = 2$

relation Airplane_Shadow using polynomial constraints over the variables x, y and t. (Here the time unit t will be measured in seconds and not years as in the Lincoln_Growing example.)

Spatial constraint databases can be queried by the same query languages that relational databases can be queried. For example, the popular Structured Query Language (SQL) for relational databases is also applicable to spatial constraint databases. For example, the following query finds when the towns of Lincoln, Nebraska, and Omaha, Nebraska, will grow into each other.

```
SELECT Min(Lincoln_Growing.T)
FROM   Lincoln_Growing, Omaha_Growing
WHERE  Lincoln_Growing.X=Omaha_Growing.X AND
               Lincoln_Growing.Y=Omaha_Growing.Y AND
               Lincoln_Growing.T=Omaha_Growing.T
```

Suppose that the airplane flies over Lincoln, Nebraska. The next query finds when the shadow of the airplane will leave the town.

```
SELECT Max(Airplane_Shadow.T)
FROM   Lincoln, Airplane_Shawdow
WHERE  Lincoln.X = Airplane_Shadow.X AND
               Lincoln.Y = Airplane_Shadow.Y
```

Besides SQL queries, spatial constraint databases can be queried by relational calculus queries (i.e., first-order logic queries) and by Datalog queries.

Key Applications

Many applications of spatial constraint database systems are similar to the applications of other GIS systems, such as the ARC/GIS system. These applications typically include problems where various kinds of maps, road networks, and land utilization information are represented and overlaying of different maps plays a key role in information processing and querying. However, efficiently describing and querying spatial or geographic data is just one application area of spatial constraint databases.

Spatial constraint databases are also useful in applications that go beyond traditional GIS systems. These applications typically require the representation of moving objects or spatiotemporal data and high-level, often recursive, SQL or Datalog queries. The following are some examples of such applications.

Applications Based on Interpolated Spatiotemporal Data

Spatiotemporal data, just like spatial data, often contain missing pieces of information that need to be estimated based on the known data. For example, given the prices of the houses when they were sold during the past 30 years in a town, one may need to estimate the prices of those houses which were not sold at any time within the past 30 years. In such applications the results of the interpolation can be represented as a spatial constraint database with x, y fields for the location of houses, t field for time, and a p field for price (Li and Revesz 2004). The value of p will be estimated as some function, often a linear equation, of the other variables. If the spatiotemporal interpolation result is represented

Constraint Databases, Spatial, Table 3 Lincoln_Growing

X	Y	T	
x	y	t	$y \geq x + 8,\ y \geq 14,\ x \geq 2,\ y \leq 18,\ y \leq -x + 24,\ 2000 \leq t \leq 2007$
x	y	t	$y \leq x + 8, y \leq 0.5x + 12, y \leq 18, x \leq 14 + (t - 2000), y \geq 8, y \geq -3x + 32, 2000 \leq t \leq 2007$

in a spatial constraint database like MLPQ, then it becomes easy to use the data for applications like estimating price and tax payments for particular houses, estimating total taxes received by the town from sale of houses in any subdivision of the town, etc.

Applications Based on Continuously Changing Maps

In many applications maps are not static but continuously changing. For example, the regions where drought occurs or where an epidemic occurs change with time within every country or state. Such changing maps are conveniently represented in spatial constraint databases similarly to the representation of the growing town area (Revesz and Wu 2006). In these cases too, when the changing map representations are available in a spatial constraint database, many types of applications and specific queries can be developed. For example, a drought monitoring application can estimate the damage done to insured agricultural areas. Here the changing drought map needs to be overlaid with static maps of insured areas and particular crop areas. Another example is tracking the spread of an epidemic and estimating when it may reach certain areas of the state or country and how many people may be affected.

Applications Based on Moving Objects

Moving objects can be animate living beings, that is, people and animals, natural phenomena such as hurricanes and ocean currents, or man-made moving objects such as airplanes, cars, missiles, robots, and trains. Moving objects can also be represented in spatial constraint databases. The representation can then be used in a wide range of applications from weather prediction to airplane and train scheduling or some combination application. For example, one may need to find which airplanes are in danger of being influenced by a hurricane. Endangered airplanes can be given a warning and rerouted if need be. Another application is checking whether the airspace surrounding an airport ever gets too crowded by the arriving and departing airplanes.

Future Directions

Spatial constraint database systems that implement polynomial constraints are being developed. There are a growing number of spatial constraint database applications. Improved algorithms for indexing and querying spatial constraint data are also being developed. Finally, improved high-level visualization methods of constraint data are sought after to enhance the user interface of spatial constraint database systems.

Cross-References

- Constraint Database Queries
- Constraint Databases and Data Interpolation
- Constraint Databases and Moving Objects
- Linear Versus Polynomial Constraint Databases
- MLPQ Spatial Constraint Database System
- Spatial Constraint Databases, Indexing
- Visualization of Spatial Constraint Databases

References

Abiteboul S, Hull R, Vianu V (1995) Foundations of databases. Addison-Wesley, Reading, Massachusetts

Benedikt M, Dong G, Libkin L, Wong L (1998) Relational expressive power of constraint query languages. J ACM 45(1):1–34

Brodsky A, Segal V, Chen J, Exarkhopoulo P (1997) The CCUBE constraint object-oriented database system. Constraints 2(3-4):245–277

Goldin D, Kutlu A, Song M, Yang F (2003) The constraint database framework: lessons learned from CQA/CDB.

In: Proceedings of international conference on data engineering, pp 735–737
Grumbach S, Rigaux P, Segoufin L (1998) The DEDALE system for complex spatial queries. In: Proceedings of ACM SIGMOD international conference on management of data, pp 213–24
Gting RH, Schneider M (2005) Moving objects databases. Morgan Kaufmann, Amsterdam
Kanellakis PC, Kuper GM, Revesz PZ (1990) Constraint query languages. In: Proceedings of ACM symposium on principles of database systems, pp 299–313
Kanellakis PC, Kuper GM, Revesz PZ (1995) Constraint query languages. J Comput Syst Sci 51(1):26–52
Kuper GM, Libkin L, Paredaens J (eds) (2000) Constraint databases. Springer, Berlin
Li L, Revesz PZ (2004) Interpolation methods for spatiotemporal geographic data. Comput Environ Urban Syst 28:201–227
Revesz PZ (1993) A closed-form evaluation for Datalog queries with integer (gap)-order constraints. Theor Comput Sci 116(1):117–49
Revesz PZ (2002) Introduction to constraint databases. Springer, New York
Revesz PZ (2010) Introduction to databases: from biological to spatio-temporal. Springer, New York
Revesz P, Wu S (2006) Spatiotemporal reasoning about epidemiological data. Artif Intell Med 38(2):157–170
Rigaux P, Scholl M, Voisard A (2002) Spatial databases with application to GIS. Morgan Kaufmann, San Francisco

Recommended Reading

Revesz P, Li Y (1997) MLPQ: a linear constraint database system with aggregate operators. In: Proceedings of 1st international database engineering and applications symposium. IEEE Press, Washington, pp 132–137

Constraint Programming

▸ Integration of Spatial Constraint Databases

Constraint Query Languages

▸ Constraint Database Queries
▸ Linear Versus Polynomial Constraint Databases
▸ Polynomial Spatial Constraint Databases

Constraints, Authority

▸ Time Geography

Constraints, Capability

▸ Time Geography

Constraints, Coupling

▸ Time Geography

Content Metadata

▸ Feature Catalogue

Context-Aware

▸ User Interfaces and Adaptive Maps

Context-Aware Dynamic Access Control

▸ Security Models, Geospatial

Context-Aware Presentation

▸ Information Presentation, Dynamic

Context-Aware Role-Based Access Control

▸ Security Models, Geospatial

Context-Sensitive Visualization

▸ Information Presentation, Dynamic

Contextualization

▸ Geospatial Semantic Web: Personalization

Contingency Management System

▸ Emergency Evacuation Plan Maintenance

Continuity Matrix

▸ Spatial Weights Matrix

Continuity Network

▸ Conceptual Neighborhood

Continuous Location-Based Queries

▸ Continuous Queries in Spatio-Temporal Databases

Continuous Queries

▸ Indexing, Query and Velocity-Constrained
▸ Queries in Spatiotemporal Databases, Time Parameterized

Continuous Queries in Spatio-Temporal Databases

Xiaopeng Xiong[1], Mohamed F. Mokbel[2], and Walid G. Aref[1]
[1]Department of Computer Science, Purdue University, West Lafayette, IN, USA
[2]Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN, USA

Synonyms

Continuous location-based queries; Continuous query processing; Long-running spatiotemporal queries; Moving queries

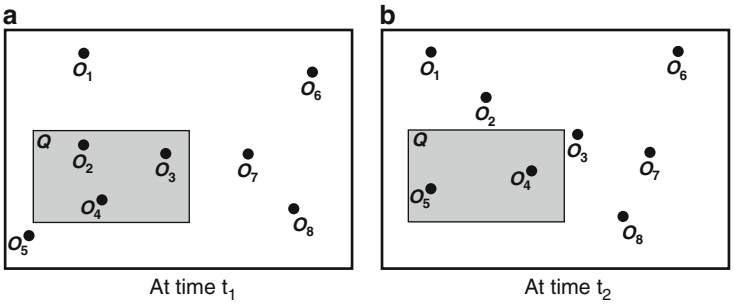
Definition

A continuous query is a new query type that is issued once and is evaluated continuously in a database server until the query is explicitly terminated. The most important characteristic of continuous queries is that their query result does not only depend on the present data in the databases but also on continuously arriving data. During the execution of a continuous query, the query result is updated continuously when new data arrives. Continuous queries are essential to applications that are interested in transient and frequently updated objects and require monitoring query results continuously. Potential applications of continuous queries include but are not limited to real-time location-aware services, network flow monitoring, online data analysis and sensor networks.

Continuous queries are particularly important in *Spatiotemporal Databases*. Continuous spatiotemporal queries are evaluated continuously against spatiotemporal objects and their results are updated when interested objects change spatial locations or spatial extents over time. Figure 1 gives an example of a continuous query in a spatiotemporal database. In Fig. 1, o_1 to o_8 are objects moving in the data space and Q is a continuous spatiotemporal query that tracks moving objects within the shaded query region. As plotted in Fig. 1a, the query answer of Q with respect to time t_1 consists of three objects: $\{o_2, o_3, o_4\}$. Assume that at a later time t_2, the objects change their locations as shown in Fig. 1b. Particularly, o_2 and o_3 move out of the query region while o_5 moves inside the query region. o_4 also moves, however it remains inside the query region. Due to the continuous evaluation of Q, the query answer of Q will be updated to $\{o_4, o_5\}$ at time t_2.

Historical Background

The study of continuous spatiotemporal queries started in the 1990s as an important part of the study of *Spatiotemporal Databases*. Since then, continuous spatiotemporal queries have received



Continuous Queries in Spatio-Temporal Databases, Fig. 1 An example of continuous query

increasing attention due to the advances and combination of portable devices and locating technologies. Recently, the study of continuous queries in spatiotemporal databases has become one of the most active fields in the database domain.

Scientific Fundamentals

There are various types of continuous queries that can be supported in spatiotemporal databases. In general, continuous spatiotemporal queries can be classified into various categories based on different classifying criteria. The most common classifying criteria are based on the type of query interest, the mobility of query interest and the time of query interest (Mokbel et al. 2003).

According to the type of query interest, there are a wide variety of continuous spatiotemporal queries. The following describes some interesting query types that are widely studied.

- *Continuous range queries* (Kalashnikov et al. 2002; Mokbel et al. 2004). This type of continuous query is interested in spatiotemporal objects inside or overlapping with a given spatial region. Continuous range queries have many important applications and are sometimes used as a filter step for other types of continuous queries.

Examples:

> Continuously report the number of trucks on Road US-52.
>
> Continuously show me all the hotels within 3 miles during my drive.

- *Continuous k-nearestneighbor (CkNN) queries* (Tao et al. 2002; Xiong et al. 2005; Li and Han 2004). A CkNN query is a query tracking continuously the k objects that are the nearest ones to a given query point. The objects of interest and/or the query point may move during query evaluation.

Examples:

> Continuously track the nearest maintenance truck to my vehicle (in the battlefield).
>
> Continuously show me the 10 nearest hotels during my drive.

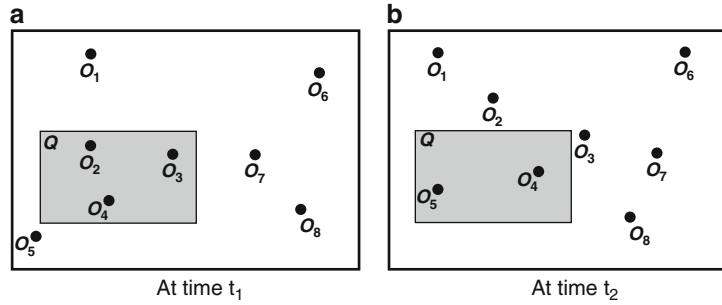
- *Continuous Reverse Nearestneighbor (CRNN) queries* (Kang et al. 2007; Xia and Zhang 2006). A CRNN query continuously identifies a set of objects that have the querying object as their nearest neighbor object.

Examples:

> Continuously find soldiers who need my help (I am the nearest doctor to him/them).
>
> Continuously send electronic advertisement of our hotel to vehicles that have our hotel as their nearest hotel.

Based on the mobility of interest, continuous spatiotemporal queries can be classified as *moving queries over static objects*, *static queries over moving objects* and *moving queries over moving objects* (Mokbel et al. 2003).

- *Moving queries over static objects*. In this query category, the objects of interest are



Continuous Queries in Spatio-Temporal Databases, Fig. 2 Continuous spatiotemporal query types based on mobility

static while the query region or query point of the continuous spatiotemporal query may change over time. This query type is abstracted in Fig. 2a. In Fig. 2a, the query o_2 moves along with time (e.g., at t_1, t_2 and t_3) and the objects (represented by black dots) are stationary.

Examples:

> Continuously return all the motels within 3 miles to John during John's road-trip.
>
> Continuously find the nearest gas station to Alice while she is driving along Highway IN-26.

In the examples above, the objects of interest (i.e., the gas stations and the motels) are static objects and the query regions/points (i.e., the 3-mile region surrounding John's location and the location of Alice) are continuously changing.

- *Static queries over moving objects*. In this query category, the query region/point of continuous spatiotemporal query remains static while the objects of interest are continuously moving. This query type is abstracted in Fig. 2b. As plotted in Fig. 2b, the objects keep moving along with time (e.g., at t_1, t_2 and t_3) while the query Q is stationary.
- Examples:
- "Continuously monitor the number of buses in the campus of Purdue University."
- "Continuously find the nearest 100 taxis to a certain hotel."
- In the above examples, the query region (i.e., the university campus) or the query point (i.e., the hotel) does not move while the objects of interest (i.e., the buses and the taxis) continuously move.
- *Moving queries over moving objects*. In this query category, both the query region/point of continuous spatiotemporal query and the objects of interest are capable of moving. This query type is abstracted in Fig. 2c. As shown in Fig. 2c, the query Q and the objects are both moving over time (e.g., at t_1, t_2 and t_3).

Example:

> Continuously report all the cars within 1 mile when the sheriff drives along the State street.

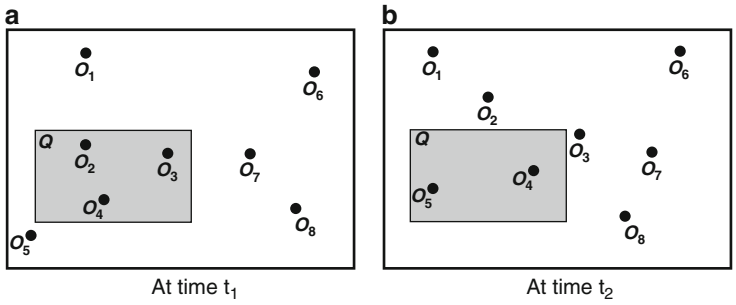
In this example, the query region (i.e., 1-mile region surrounding the location of the sheriff) and the objects of interest (i.e., the cars) are both moving.

Based on the time of query interest, continuous spatiotemporal queries can be classified as *historical queries*, *present queries* and *future queries* (Mokbel et al. 2003). Figure 3 plots the three types of queries. In Fig. 3, the gray dots represent historical object locations and the black dots represent current object locations. The dotted lines with arrows represent anticipated object movements in the future based on the objects' current velocities.

- *Historical queries*. These types of continuous queries are interested in events of the past. Historical queries are especially interesting to applications in data warehouse analysis and business intelligence.

Example:

> Continuously calculate the average number of trucks on Highway I-65 for the past 2 hours.



Continuous Queries in Spatio-Temporal Databases, Fig. 3 Continuous spatiotemporal query types based on the time of interest

In the above example, the query result depends on the historical data of the past 2 h and is continuously updated when time evolves.

- *Present queries*. In this query category, continuous queries are evaluated against only the current status of spatiotemporal objects. Present queries are important to real-time monitoring and location-based services.

Examples:

> Continuously return the total number of vehicles in the monitored area.
>
> Send an alarm once the child steps out of the neighbor's home.

The query results of the examples above depend only on the current locations of the objects of interest (i.e., the locations of vehicles and the location of the child).

- *Future queries*. In this query type, the query results are based on the predication of future events. The evaluation of future queries usually relies on the knowledge of expected object movement, such as the velocity information of the objects. Future queries are particularly useful for alarm systems and risk prevention.

Example:

> Send an alarm if two aircrafts are becoming less than 5 miles apart from each other in 3 minutes.

The above query can be evaluated continuously based on the velocity information of the aircrafts. When the velocities of the aircrafts are changed, the query is re-evaluated and the query result may change accordingly.

Key Applications

Continuous spatiotemporal queries have numerous applications in a wide range of fields.

Traffic Monitoring

Continuous spatiotemporal queries can be applied to monitor real-time traffic conditions in interested areas. Abnormal traffic conditions such as vehicle collision and traffic congestion can be detected and monitored by continuously analyzing incoming traffic data. Besides traffic detection and monitoring, recommendations of alternative routes can be sent to drivers, allowing them to bypass the slow-traffic roads.

Traffic Pattern Detection

Traffic usually demonstrates a repeated pattern with respect to location and time. Detection of such a pattern is important to predict the traffic conditions in the future at the interested area. Continuous spatiotemporal queries can be used to detect such patterns by continuously analyzing traffic data and maintaining spatio-temporal histograms.

Danger Alarming

Depending on the present or predicted locations of interested objects, continuous queries can trigger alarms to prevent potential dangers. Danger alarming has applications in both daily life and national security. For example, alarms can be sent

when kids step out of the home backyard or when an unidentified flight is predicted to fly over a military base in 5 min.

Digital Battlefield

In the digital battlefield, continuous queries can help commanders make decisions by continuously monitoring the context of friendly units such as soldiers, tanks and flights.

Road-Trip Assistance

Travel by driving will become more convenient with the aid of continuous spatiotemporal queries. A driver can be continuously informed about information such as nearby hotels, gas stations and grocery stores. More dynamic information such as nearby traffic conditions and weather alarms based on the current location can also be supported by integrating corresponding information in spatiotemporal databases.

Location-Based E-Commerce

Continuous spatiotemporal queries can be utilized to shift E-commerce from web-based E-commerce to location-based E-commerce. Location-based E-commerce is E-commerce associated with the locations of potential customers. One example of location-based E-commerce is "sending coupons to all vehicles that are within twenty miles of my hotel".

Climate Analysis and Predicting

Climatology study can benefit from employing continuous spatiotemporal queries over climate data. Climate phenomena such as hurricanes, storms and cumulonimbus clouds can be modeled as spatiotemporal objects with changing spatial extents and moving locations. Climate phenomena can be continuously monitored and it is plausible to analyze historical climate phenomena and to predict future climate phenomena.

Environmental Monitoring

Continuous spatiotemporal queries can work with sensor networks to monitor environmental phenomena such as forest fires or polluted water domains. Sensors in a sensor network continuously detect environmental events and feed the data into spatiotemporal databases. Then, the properties of the environmental phenomena (e.g., the shape and the movement of the fire or the polluted water area) can be continuously monitored.

Future Directions

The study of continuous spatiotemporal queries is steadily progressing. Current efforts focus on a variety of challenging issues (Roddick et al. 2004; Sellis 1999a, b). The following provides some directions among a long list of research topics.

- *Query Language.* This direction is to define an expressive query language so that any continuous spatiotemporal queries can be properly expressed.
- *Novel Query Types.* More query types are proposed based on new properties of spatiotemporal data. Examples of new continuous spatiotemporal query types include continuous probabilistic queries, continuous group nearest queries etc.
- *Query Evaluation.* Due to the continuous evaluation of queries, efficient query evaluation algorithms are needed to process queries incrementally and continuously whenever data is updated.
- *Data Indexing.* Traditional data indexing structures usually do not perform well under frequent object updates. There is a challenging issue on designing updatetolerant data indexing to cope with continuous query evaluation.
- *Scalability.* When the number of moving objects and the number of continuous queries become large, the performance of spatiotemporal databases will degrade and cannot provide a timely response. Increasing the scalability of spatiotemporal databases is an important topic to address.

Cross-References

- ▸ Queries in Spatiotemporal Databases, Time Parameterized
- ▸ Spatiotemporal Database Modeling with an Extended Entity-Relationship Model

References

Kalashnikov DV, Prabhakar S, Hambrusch SE, Aref WA (2002) Efficient evaluation of continuous range queries on moving objects. In: DEXA '02: proceedings of the 13th international conference on database and expert systems applications. Springer, Heidelberg, pp 731–740

Kang JM, Mokbel MF, Shekhar S, Xia T, Zhang D (2007) Continuous evaluation of monochromatic and bichromatic reverse nearest neighbors. In: ICDE, Istanbul

Li JYY, Han J (2004) Continuous k-nearest neighbor search for moving objects. In: SSDBM '04: proceedings of the 16th international conference on scientific and statistical database management (SSDBM'04). IEEE Computer Society, Washington, DC, p 123

Mokbel MF, Aref WA, Hambrusch SE, Prabhakar S (2003) Towards scalable location-aware services: requirements and research issues. In: GIS, New Orleans

Mokbel MF, Xiong X, Aref WA (2004) SINA: scalable incremental processing of continuous queries in spatiotemporal databases. In: SIGMOD, Paris

Roddick JF, Hoel E, Egenhofer ME, Papadias D, Salzberg B (2004) Spatial, temporal and spatiotemporal databases – hot issues and directions for PHD research. SIGMOD Rec 33(2):126–131

Sellis T (1999) Chorochronos – research on spatiotemporal database systems. In: DEXA '99: proceedings of the 10th international workshop on database & expert systems applications. IEEE Computer Society, Washington, DC, p 452

Sellis TK (1999) Research issues in spatiotemporal database systems. In: SSD'99: proceedings of the 6th international symposium on advances in spatial databases. Springer, London, pp 5–11

Tao Y, Papadias D, Shen Q (2002) Continuous nearest neighbor search. In: VLDB, Hong Kong

Xia T, Zhang D (2006) Continuous reverse nearest neighbor monitoring. In: ICDE '06: proceedings of the 22nd international conference on data engineering (ICDE '06), Washington, DC, p 77

Xiong X, Mokbel MF, Aref WG (2005) SEA-CNN: scalable processing of continuous K-nearest neighbor queries in spatiotemporal databases. In: ICDE, Tokyo

Continuous Query Processing

▸ Continuous Queries in Spatio-Temporal Databases

Continuously Changing Maps

▸ Constraint Databases and Moving Objects

Contraflow for Evacuation Traffic Management

Brian Wolshon
Department of Civil and Environmental Engineering, Louisiana State University, Baton Rouge, LA, USA

Synonyms

All-lanes-out; Emergency preparedness; Evacuation planning; Merge designs; One-way-out evacuation; Reversible and convertible lanes; Split designs

Definition

Contraflow is a form of reversible traffic operation in which one or more travel lanes of a divided highway are used for the movement of traffic in the opposing direction (The common definition of contraflow for evacuations has been broadened over the past several years by emergency management officials, the news media, and the public to include the reversal of flow on any roadway during an evacuation.) (American Association of State Highway and Transportation Officials 2001). It is a highly effective strategy because it can both immediately and significantly increase the directional capacity of a roadway without the time or cost required to plan, design, and construct additional lanes. Since 1999, contraflow has been widely applied to evacuate regions of the southeastern United States (US) when under threat from hurricanes. As a result of its recent demonstrated effectiveness during Hurricane Katrina (Wolshon 2006), it also now looked upon as a potential preparedness measure for other mass-scale hazards.

Contraflow segments are most common and logical on freeways because they are the highest capacity roadways and are designed to facilitate high speed operation. Contraflow is also more practical on freeways because these routes do not incorporate at-grade intersections that interrupt flow or permit unrestricted access into the reversed segment. Freeway contraflow can also be implemented and controlled with fewer manpower resources than unrestricted highways.

Nearly all of the contraflow strategies currently planned on US freeways have been designed for the reversal of all inbound lanes. This configuration, shown schematically in Inset 1d of Fig. 1, is commonly referred to as a "One-Way-Out" or "All-Lanes-Out" evacuation. Though not as popular, some contraflow plans also include options for the reversal of only one of the inbound lanes (Inset 1b) with another option to use one or more of the outbound shoulders (Inset 1c) (Wolshon 2001). Inbound lanes in these plans are maintained for entry into the threat area by emergency and service vehicles to provide assistance to evacuees in need along the contraflow segment.

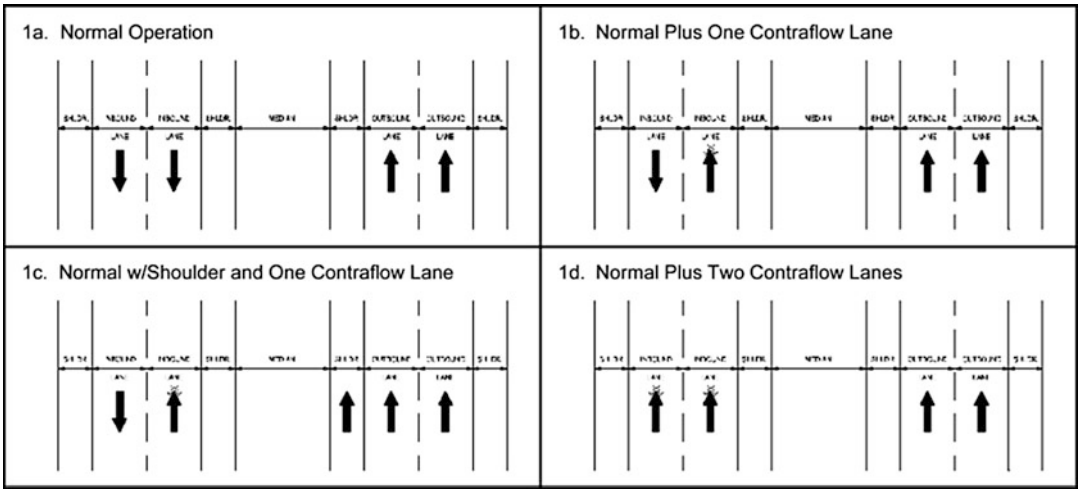
Historical Background

Although evacuation-specific contraflow is a relatively recent development, its application for other types of traffic problems is not new (Wolshon and Lambert 2004). In fact, various forms of reversible traffic operation have been used throughout the world for decades to address many types of directionally unbalanced traffic conditions. They have been most common around major urban centers where commuter traffic is heavy in one direction while traffic is light in the other. Reverse and contraflow operations have also been popular for managing the infrequent, but periodic and predictable, directionally imbalanced traffic patterns associated with major events like concerts, sporting events, and other public gatherings. Reversible lanes have also been cost effective on bridges and in tunnels where additional directional capacity is needed, but where additional lanes can not be easily added.

While the date of the first use of contraflow for an evacuation is not known with certainty, interest in its potential began to be explored after Hurricane Andrew struck Florida in 1992. By 1998, transportation and emergency management officials in both Florida and Georgia had plans in place to use contraflow on segments of Interstate freeways. Ultimately, the watershed event for evacuation contraflow in the United States was Hurricane Floyd in 1999. Since then, every coastal state threatened by hurricanes has developed and maintains plans for the use of evacuation contraflow.

Hurricane Floyd triggered the first two major implementations of contraflow, one on a segment of Interstate (I) 16 from Savannah to Dublin, Georgia and the other on I-26 from Charleston to Columbia, South Carolina. The results of both of these applications were generally positive, although numerous areas for improvement were also identified. The contraflow application in South Carolina was particularly interesting because it was not pre-planned. Rather, it was implemented on an improvisational basis after a strong public outcry came from evacuees trapped for hours in congested lanes of westbound I-26 seeking ways to use the near-empty eastbound lanes.

The first post-Floyd contraflow implementations occurred in Alabama for the evacuation of Mobile and Louisiana for the evacuation New Orleans. Once again, many lessons were learned and numerous improvements in both physical and operational aspects of the plans were suggested. The timing of these events was quite fortuitous for New Orleans. Within 3 months of the major changes that were implemented to the Louisiana contraflow plan after Hurricane Ivan, they were put into operation for Hurricane Katrina. The changes, so far the most aggressive and far-ranging of any developed until that time (Wolshon et al. 2006), involved the closure of lengthy segments of interstate freeway, forced traffic onto alternative routes, established contraflow segments across the state boundary into Mississippi, coordinated parallel non-freeway routes, and reconfigured several interchanges to more effectively load traffic from



Contraflow for Evacuation Traffic Management, Fig. 1 Freeway contraflow lane use configurations for evacuations (Wolshon 2001)

surface streets. The results of these changes were reflected in a clearance time for the city that was about half of the previous prediction (Wolshon and McArdle).

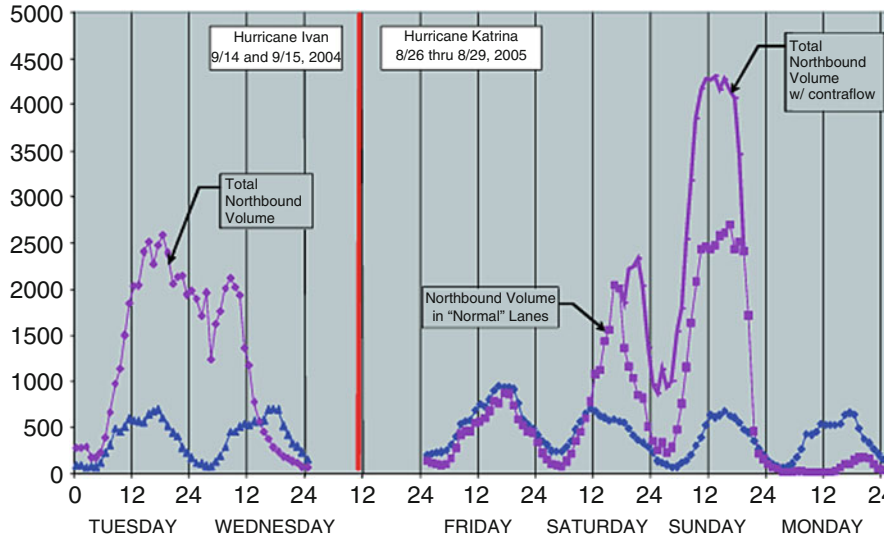
Scientific Fundamentals

Although the basic concept of contraflow is simple, it can be complex to implement and operate in actual practice. If not carefully designed and managed, contraflow segments also have the potential to be confusing to drivers. To insure safe operation, improper access and egress movements must be prohibited at all times during its operation. Segments must also be fully cleared of opposing traffic prior to initiating contraflow operations. These are not necessarily easy to accomplish, particularly in locations where segments are in excess of 100 miles and where interchanges are frequent. For these reasons some transportation officials regard them to be risky and only for use during daylight hours and under the most dire situations. They are also the reason why contraflow for evacuation has been planned nearly exclusively for freeways, where access and egress can be tightly controlled.

To now, contraflow evacuations have also been used only for hurricane hazards and wildfires and no other type of natural or manmade hazard. The first reason for this is that these two hazards affect much greater geographic areas and tend to be slower moving relative to other hazards. Because of their scope they also create the need move larger numbers of people over greater distances than other types of hazards. The second reason is that contraflow requires considerable manpower and materiel resources as well as time to mobilize and implement. Experiences in Alabama and Louisiana showed that the positioning of traffic control devices and enforcement personnel takes at least 6 h not including the time to plan and preposition equipment for the event. In Florida, where needs are great and manpower resources are stretched thin, evacuation contraflow requires involvement from the Florida National Guard. For this reason (among others), Florida officials require a minimum of 49 h of advanced mobilization time for contraflow to be implemented (Wolshon et al. 2005).

Operational Effects of Contraflow

As the goal of an evacuation is to move as many people as quickly out of the hazard threat zone as possible, the primary goal of contraflow is to increase the rate of flow and decrease the travel time from evacuation origins and destinations. Prior to field measurement, it was hypothesized



Contraflow for Evacuation Traffic Management, Fig. 2 Northbound traffic volume – I-55 at Fluker Louisiana (Data source: LA DOTD)

that the flow benefits of contraflow would be substantial, but less than that of an equivalent normally flowing lane (Wolshon 2001). These opinions were based on measurements of flow on I-26 during the Hurricane Floyd evacuation and the theory that drivers would drive at slower speeds and with larger spacing in contraflow lanes.

The highest flow rates measured by the South Carolina Department of Transportation (DOT) during the Floyd evacuation were between 1500 to 1600 vehicles per hour per lane (vphpl) (United States Army Corps of Engineers 2000). Traffic flows measured during the evacuations for Hurricanes Ivan and Katrina on I-55 in Louisiana were somewhat less than the South Carolina rates. Flows in the normal-flow lanes of I-55 averaged about 1230 vphpl during the peak 10 h of the evacuation. Flow rates in the contraflow lanes during the same period averaged about 820 vphpl. These volumes compare to daily peaks of about 400 vphpl during routine periods and a theoretical capacity of 1800–2000 vphpl for this segment.

The graph of Fig. 2 illustrates the hourly traffic flow on I-55 during the evacuations for Hurricanes Ivan (when contraflow was not used) and Katrina (when contraflow was used). During the 48 h period of the Ivan evacuation (shown on the left side of the graph) a total of 60,721 vehicles traveled northbound through this location. During the Katrina evacuation, the total volume was 84,660 vehicles during a corresponding 48 h period. It is also worthy to note that the duration of the peak portion of the evacuation (i.e., when the volumes were noticeably above the prior 3 week average) was about the same for both storms.

The data in Fig. 2 are also of interest because they are consistent with prior analytical models of evacuation that have estimated maximum evacuation flow on freeways with contraflow to be about 5000 vph. One of the difficulties in making full analyses of evacuation volume in general, and of contraflow volume in specific, has been a lack of speed data. Although the flow rates recorded during the two recent Louisiana hurricane evacuations are considerably below than the theoretical capacity of this section of freeway, it can not be determined with certainty if the conditions were congested with low operating speeds and small headways or relatively free flowing at more moderate levels of demand. It is also interesting to note that empirical observation of speed at a point toward the end of the segment did not appear to support the popular theory of elevated driver caution during contraflow. In fact, traffic enforcement personnel in Mississippi measured

speeds well in excess of posted speed limits as the initial group of drivers moved through the newly opened lanes.

Elements of Contraflow Segments

Reversible roadways have a number of physical and operational attributes that common among all applications. The principle physical attributes are related to spatial characteristics of the design, including its overall length, number of lanes, as well as the configuration and length of the inbound and outbound transition areas. The primary operational attributes are associated with the way in which the segment will be used and include the temporal control of traffic movements. The temporal components of all reversible lane segments include the frequency and duration of a particular configuration and the time required to transition traffic from one direction to another. The duration of peak-period commuter reversible applications, for example, typically last about 2 h (not including set-up, removal, and transition time) with a twice daily frequency. Evacuation contraflow, however, may only be implemented once in several years, its duration of operation may last several days.

Like all reversible flow roadways, contraflow lanes need to achieve and maintain full utilization to be effective. Although this sounds like an obvious fact, it can be challenging to achieve in practice. The most common reason for underutilization has been inadequate transitions into and out of the contraflow segment. Contraflow requires a transition section at the inflow and outflow ends to allow drivers to maneuver into and out of the reversible lanes from the unidirectional lanes on the approach roadways leading into it. Since these termini regulate the ingress and egress of traffic entering and exiting the segment and they are locations of concentrated lane changing as drivers weave and merge into the desired lane of travel, they effectively dictate the capacity of the entire segment.

Through field observation and simulation studies (Theodoulou 2003; Williams et al. 2007) it has been shown that contraflow entry points with inadequate inflow transitions result in traffic congestion and delay prior to the contraflow segment and prohibit the segment from carrying capacity-level demand. This was illustrated by I-10 contraflow segment in New Orleans during the Hurricane Ivan evacuation. At that time, evacuating traffic vehicles in the left and center outbound lanes of I-10 were transitioned across the median and into the contraflow lanes using a paved crossover. However, the combination of the crossover design, temporary traffic control devices, presence of enforcement personnel, and weaving vehicles created a flow bottleneck that restricted inflow into the contraflow lanes. This caused two problems. First, it limited the number of vehicles that could enter the contraflow lanes limiting flow beyond the entry point significantly below its vehicle carrying capability. The other was that it caused traffic queues upstream of the crossover that extended back for distances in excess of 14 miles. This plan was significantly improved prior to the Katrina evacuation 1 year later by permitting vehicles to enter the contraflow lanes at multiple points, spatially spreading the demand over a longer distance and reducing the length and duration amount of the congested conditions (Wolshon et al. 2006).

Inadequate designs at the downstream end of contraflow segments can also greatly limit its effectiveness. Prior experience and simulation modeling (Lim 2003) have shown that an inability to move traffic from contraflow lanes back into normally flowing lanes will result in congestion backing up from the termination transition point in the contraflow lanes. Under demand conditions associated with evacuations, queue formation can occur quite rapidly and extend upstream for many miles within hours. To limit the potential for such scenarios, configurations that require merging of the normal and contraflowing lanes are discouraged; particularly if they also incorporate lane drops. Two popular methods that are used to terminate contraflow include routing the two traffic streams at the termination on to separate routes and reducing the level of outflow demand at the termination by including egress point along the intermediate segment. Several of the more common configurations are discussed in the following section.

Contraflow Plans and Designs

The primary physical characteristics of contraflow segments are the number of lanes and the length. A 2003 study (Urbina and Wolshon 2003) of hurricane evacuation plans revealed that 18 controlled access evacuation contraflow flow segments and three additional arterial reversible roadway segments have been planned for use in the US. Currently, all of the contraflow segments are planned for a full "One-Way-Out" operation. The shortest of the contraflow freeway segments was the I-10 segment out of New Orleans at about 25 miles long. The longest were two 180 segments of I-10 in Florida; one eastbound from Pensacola to Tallahassee and the other westbound from Jacksonville to Tallahassee. Most of the others were between 85 and 120 miles.

In the earliest versions of contraflow, nearly all of the planned segments that were identified in the study were initiated via median crossovers. Now that single point loading strategies have been shown to be less effective, many locations are changing to multi-point loading. Most popular of these are median crossovers, with supplemental loading via nearby reversed interchange ramps.

The termination configurations for the reviewed contraflow segments were broadly classified into one of two groups. The first were *split designs*, in which traffic in the normal and contraflowing lanes were routed onto separate roadways at the terminus. The second group were the *merge designs* in which the separate lane groups are reunited into the normal-flow lanes using various geometric and control schemes. The selection of one or the other of these termination configurations at a particular location by an agency has been a function of several factors, most importantly the level of traffic volume and the configuration and availability of routing options at the end of the segment.

In general, split designs offer higher levels of operational efficiency of the two designs. The obvious benefit of a split is that it reduces the potential for bottleneck congestion resulting from merging four lanes into two. Its most significant drawback is that it requires one of the two lane groups to exit to a different route, thereby eliminating route options at the end of the segment. In some older designs, the contraflow traffic stream was planned to be routed onto an intersecting arterial roadway. One of the needs for this type of split design is adequate capacity on the receiving roadway.

Merge termination designs also have pros and cons. Not surprisingly, however, these costs and benefits are nearly the exact opposite of split designs in their end effect. For example, most merge designs preserve routing options for evacuees because they do not force vehicles on to adjacent roadways and exits. Unfortunately, the negative side to this is that they also have a greater potential to cause congestion since they merge traffic into a lesser number of lanes. At first glance it would appear illogical to merge two high volume roadways into one. However, in most locations where they are planned exit opportunities along the intermediate segment will be maintained to decrease the volumes at the end of the segment.

Key Applications

The list of applications for contraflow continues to grow as transportation and emergency preparedness agencies recognize its benefits. As a result, the number of locations that are contemplating contraflow for evacuations is not known. However, a comprehensive study of contraflow plans (Urbina and Wolshon 2003) in 2003 included 21 reverse flow and contraflow sections. The locations and distances of these locations are detailed in Table 1.

Future Directions

As experiences with contraflow increase and its effectiveness becomes more widely recognized, it is likely that contraflow will be accepted as a standard component emergency preparedness planning and its usage will grow. Several recent high profile negative evacuation experiences have prompted more states to add contraflow options to their response plans. The most notable of

Contraflow for Evacuation Traffic Management, Table 1 Planned contraflow/reverse flow evacuation routes (Urbina and Wolshon 2003)

State	Route(s)	Approx. distance (miles)	Origin location	Termination location
New Jersey	NJ-47/ NJ-347[a] Atlantic City Expressway NJ-72/ NJ-70[a] NJ-35[a] NJ-138/I-195	19 44 29.5 3.5 26	Dennis Twp Atlantic City Ship Bottom Boro Mantoloking Boro Wall Twp	Maurice River Twp Washington Twp Southampton Pt. Pleasant Beach Upper Freehold
Maryland	MD-90	11	Ocean City	US 50
Virginia[a]	I-64	80	Hampton Roads Bridge	Richmond
North Carolina	I-40	90	Wilmington	Benson (I-95)
South Carolina	I-26	95	Charleston	Columbia
Georgia	I-16	120	Savannah	Dublin
Florida	I-10 Westbound I-10 Eastbound SR 528 (Beeline) I-4 Eastbound I-75 Northbound FL Turnpike I-75 (Alligator Alley)	180 180 20 110 85 75 100	Jacksonville Pensacola SR 520 Tampa Charlotte County Ft. Pierce Coast	Tallahassee Tallahassee SR 417 Orange County I-275 Orlando Coast
Alabama	I-65	135	Mobile	Montgomery
Louisiana	I-10 Westbound I-10/I-59 (east/north)	25 115[a]	New Orleans New Orleans	I-55 Hattiesburg[a]
Texas	I-37	90	Corpus Christi	San Antonio

[a]Notes: Delaware and Virginia contraflow plans are still under development. The actual length of the New Orleans, LA to Hattiesburg, MS contraflow segment will vary based on storm conditions and traffic demand. Since they are undivided highways, operations on NJ-47/NJ-347, NJ-72/NJ-70, and NJ-35 are "reverse flow" rather than contraflow

these was in Houston, where scores of evacuees (including 23 in a single tragic incident) reportedly perished during the highly criticized evacuation for Hurricane Rita in 2005 (Senior Citizens From Houston Die When Bus Catches Fire 2005). Plans for contraflow are currently under development and should be ready for implementation by the 2007 storm season. Contraflow is also being evaluated for use in some of the larger coastal cities of northeast Australia.

In other locations where hurricanes are not a likely threat, contraflow is also being studied. Some of these examples include wildfires in the western United States (Wolshon and Marchive 2007) and tsunamis and volcanoes in New Zealand. Greater emphasis on terrorism response have also resulted in cities with few natural hazards to begin examining contraflow for various accidental and purposeful manmade hazards (Sorensen and Vogt 2006).

It is also expected that as contraflow gains in popularity, the application of other developing technologies will be integrated into this strategy. Such has already been the case in South Carolina, Florida, and Louisiana where various intelligent transportation systems (ITS) and other remote sensing technologies have been applied to monitor the state and progression of traffic on contraflow sections during an evacuation. In Washington DC, where reversible flow has been evaluated for use on primary arterial roadways during emergencies, advanced control systems for modifying traffic signal timings have also been studied (Chen et al.).

Cross-References

▸ Contraflow in Transportation Network
▸ Dynamic Travel Time Maps

References

American Association of State Highway and Transportation Officials (2001) A policy on geometric design of highways and streets, 5th edn. American Association of State Highway and Transportation Officials, Washington, DC

Chen M, Chen L, Miller-Hooks E, Traffic signal timing for urban evacuation. ASCE J Urban Plan Dev Spec Emerg Transp Issue 133(1):30–42

Lim YY (2003) Modeling and evaluating evacuation contraflow termination point designs. Master's thesis, Department of Civil and Environmental Engineering, Louisiana State University

Senior Citizens from Houston Die When Bus Catches Fire (2005) Washington post staff writer, Saturday, Sept 24, p A09. Also available online at: http://www.washingtonpost.com/wp-dyn/content/article/2005/09/23/AR200509230 0505.html

Sorensen J, Vogt B (2006) Interactive emergency evacuation planning guidebook. Chemical Stockpile Emergency Preparedness Program, Department of Homeland Security. Available online at: http://emc.ornl.gov/CSEPPweb/evac_files/index.htm

Theodoulou G (2003) Contraflow evacuation on the westbound I-10 out of the city of New Orleans. Master's thesis, Department of Civil and Environmental Engineering, Louisiana State University

United States Army Corps of Engineers (2000) Southeast United States hurricane evacuation traffic study. Buckeley, Schuh, and Jernigan, Inc., Tallahassee (Performed by post)

Urbina E, Wolshon B (2003) National review of hurricane evacuation plans and policies: a comparison and contrast of state practices. Transp Res Part A Policy Pract 37(3):257–275

Williams B, Tagliaferri PA, Meinhold SS, Hummer JE, Rouphail NM (2007) Simulation and analysis of freeway lane reversal for coastal hurricane evacuation. ASCE J Urban Plan Dev Spec Emerg Transp Issue 133(1):61–72

Wolshon B (2006) Planning and engineering for the Katrina evacuation. Bridge Natl Acad Sci Eng 36(1):27–34

Wolshon B (2001) One-way-out: contraflow freeway operation for hurricane evacuation. Natl Hazard Rev Am Soc Civil Eng 2(3):105–112

Wolshon B, Lambert L (2004) Convertible lanes and roadways. National Cooperative Highway Research Program, Synthesis 340, Transportation Research Board, National Research Council, Washington, DC, 92pp

Wolshon B, Marchive E (2007) Evacuation planning in the urban-wildland interface: moving residential subdivision traffic during wildfires. ASCE J Urban Plan Dev Spec Emerg Transp Issue 133(1):73–81

Wolshon B, McArdle B (in press) Temporospatial analysis of hurricane Katrina regional evacuation traffic patterns. ASCE J Infrastruct Syst Spec Infrastruct Plan Design Manag Big Events Issue

Wolshon B, Catarella-Michel A, Lambert L (2006) Louisiana highway evacuation plan for hurricane Katrina: proactive management of regional evacuations. ASCE J Transp Eng 132(1):1–10

Wolshon B, Urbina E, Levitan M, Wilmot C (2005) National review of hurricane evacuation plans and policies, part II: transportation management and operations. ASCE Natl Hazard Rev 6(3):142–161

Contraflow in Transportation Network

Sangho Kim
Rancho Cucamonga, CA, USA

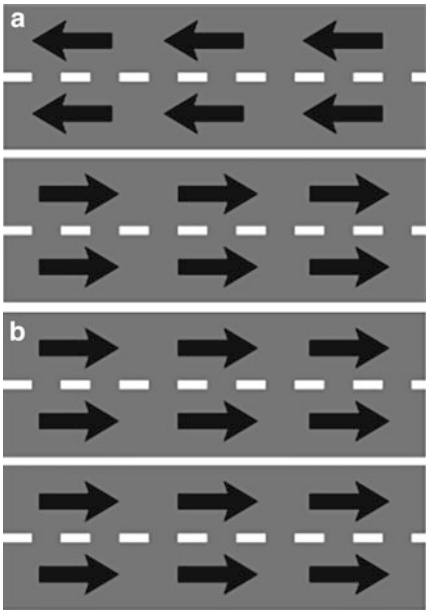
Synonyms

Counterflow; Emergency response; Evacuation routes; Lane reversal; Networks, spatial; Road networks

Definition

Contraflow is a method designed to increase the capacity of transportation roads toward a certain direction by reversing the opposite direction of road segments. Figure 1 shows a change of road direction under contraflow operation. Contraflow has been primarily used as a part of evacuation schemes. Incoming lanes are reversed to an outbound direction from an affected area during an evacuation to increase the efficiency of traffic movement. When contraflow is implemented on a road network, a significant amount of resources is required in terms of manpower and safety facilities. Automated contraflow execution requires controlled access at both starting and end points of a road. Manual execution requires police officers and barricade trucks. Contraflow planners also need to take into account other factors from the perspectives of planning, design, operation, and financial cost.

Today, there are many attempts to generate automated contraflow plans with the advanced computing power. However, computerized contraflow



Contraflow in Transportation Network, Fig. 1 Contraflow road direction. (**a**) Normal operation. (**b**) Contraflow operation

involves a combinatorial optimization problem because the number of possible contraflow network configurations is exponentially increasing with the number of road segments (i.e., edges in a graph). In addition, a direction change of a road segment affects the flow (or scheduling) of overall traffic movement at the system level. Thus, it is very hard to find an optimal contraflow network configuration among a huge number of possibilities.

Historical Background

The use of contraflow methods on road networks is not a new concept. Since the beginning of the modern roadway system, there has been a search for solutions to resolve unbalanced flow of traffic due to limited capacity. As the need for massive evacuations of population began to increase around southeastern coastal states threatened by hurricanes, more efficient methods of moving surface transportation have been discussed over the past 20 years (Wolshon et al. 2002). However,

Contraflow in Transportation Network, Table 1 Evacuation traffic flow rates with varying number of contraflow lanes (Source: Wolshon 2001; FEMA 2000)

Use configuration	Estimated average outbound flow rate (vehicle/h)
Normal (two-lanes outbound)	3,000
Normal plus one contraflow lane	3,900
Normal and shoulder plus one contraflow lane	4,200
Normal plus two contraflow lanes	5,000

the utilization of contraflow has been considered and executed in recent years. During the evacuation of hurricane Floyd in 1999, the South Carolina Department of Transportation measured the traffic flow of Interstate Highway 26 with varying numbers of contraflow lanes (Wolshon 2001). Table 1 summarizes their results. The first important finding is that flow rate increases as contraflow lanes (either a lane of opposing traffic or a shoulder) are added. Second, the amount of increased flow per lane is less than the average flow rate under normal condition. It is known that the flow rate per lane under normal operation is about 1,500 vehicles/h. However, the increased flow rate per lane in the table is under 1,000 vehicles/h. The limited increases are caused by the unfamiliarity of drivers and their uneasiness driving through an opposite or shoulder lane. Finally, it is observed that the use of shoulder lanes is not as effective as that of normal lanes for contraflow.

During the Rita evacuation in 2005, many evidences showed how ill-planned contraflow negatively affected traffic flow. The following are quoted observations (Litman 2006) of the traffic problems during the Rita evacuation: "High-occupancy-vehicle lanes went unused, as did many inbound lanes of highways, because authorities inexplicably waited until late Thursday to open some up. … As congestion worsened state officials announced that contraflow lanes would be established on Interstate Highway 45 (Fig. 2), US Highway 290 and Interstate Highway 10. But by mid-afternoon,

Contraflow in Transportation Network, Fig. 2 Hurricane Rita evacuation required contraflow on Interstate Highway 45. Notice that traffic on both sides of I-45 is going north (Source: dallasnews.com)

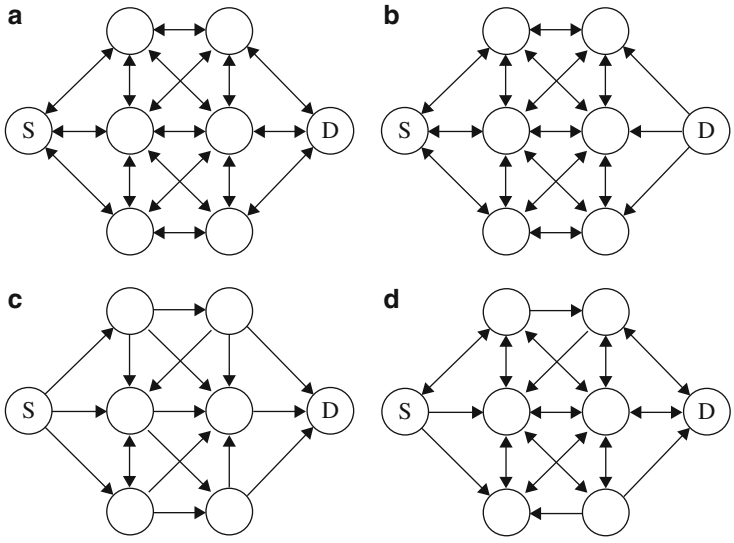
with traffic immobile on 290, the plan was dropped, stranding many and prompting others to reverse course. 'We need that route so resources can still get into the city,' explained an agency spokeswoman."

Scientific Fundamentals

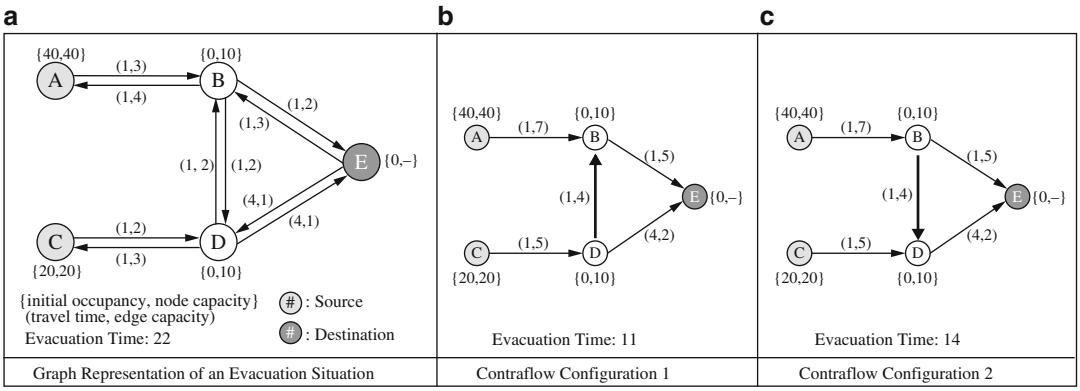
Why is Planning Contraflow Difficult?: Figuring out an optimal contraflow network configuration is very challenging due to the combinatorial nature of the problem. Figure 3 shows examples of contraflow network configurations. Suppose that people (e.g., evacuees) in a source node S want to escape to destination node D on the network. Figure 3a is a road network with all edges in two way directions. In other words, no edge is reversed in the network. Figure 3b is an example of a so called 'Infeasible' contraflow configuration because no evacuee can reach destination node D due to the ill-flipped road segments. The network in Fig. 3c allows only two types of flippings (i.e., $\uparrow, \downarrow$). A network in Fig. 3d allows three types of flippings (i.e., $\uparrow, \downarrow, \uparrow\downarrow$).

Each network used in these examples has 17 edges. If two types of flippings are allowed as shown in Fig. 3c, the number of possible network configurations is 2^{17}, that is, 131,072. Among them, 89,032 configurations are feasible. An experiment was conducted by assigning some number of evacuees on node S and travel time/capacity attributes on edges. If evacuation time is measured for all feasible configurations (89,023), only 346 configurations have minimum (i.e., optimal) evacuation time, which corresponds to 0.26 % out of total possible configurations. For the same network with three types of flippings as shown in Fig. 3d, the number of possible networks is 3^{17}, which is more than 100 million. It is impossible to handle such exponentially large number of configurations even with the most advanced computing system. These examples with such a small size network show why it is difficult to find an optimal contraflow network. The problem is classified as an NP-hard problem in computer science domain.

Modeling Contraflow using Graph: It is often necessary to model a contraflow problem using a mathematical graph. S. Kim et al. (Kim and Shekhar 2005) presented a modeling approach for the contraflow problem based on graph and flow network. Figure 4 shows a simple evacuation situation on a transportation network. Suppose that each node represents a city with initial occupancy and its capacity, as shown in Fig. 4a. City A has 40 people and also capacity 40. Nodes A and C are modeled as source nodes, while node E is modeled as a destination node (e.g., shelter). Each edge represents a road between two cities with travel time and its capacity. For example, a highway segment between cities A and B has travel time 1 and capacity 3. If a time unit is 5 min, it takes 5 min for evacuees to travel from A to B and a maximum of 3 evacuees can simultaneously travel through the edge. Nodes B and D have no initial occupancy and only serve as



Contraflow in Transportation Network, Fig. 3 Examples of infeasible contraflow network and 2 or 3 types of flippings. (**a**) All two way (**b**) Infeasible configuration (**c**) Two types flippings (**d**) Three types flippings

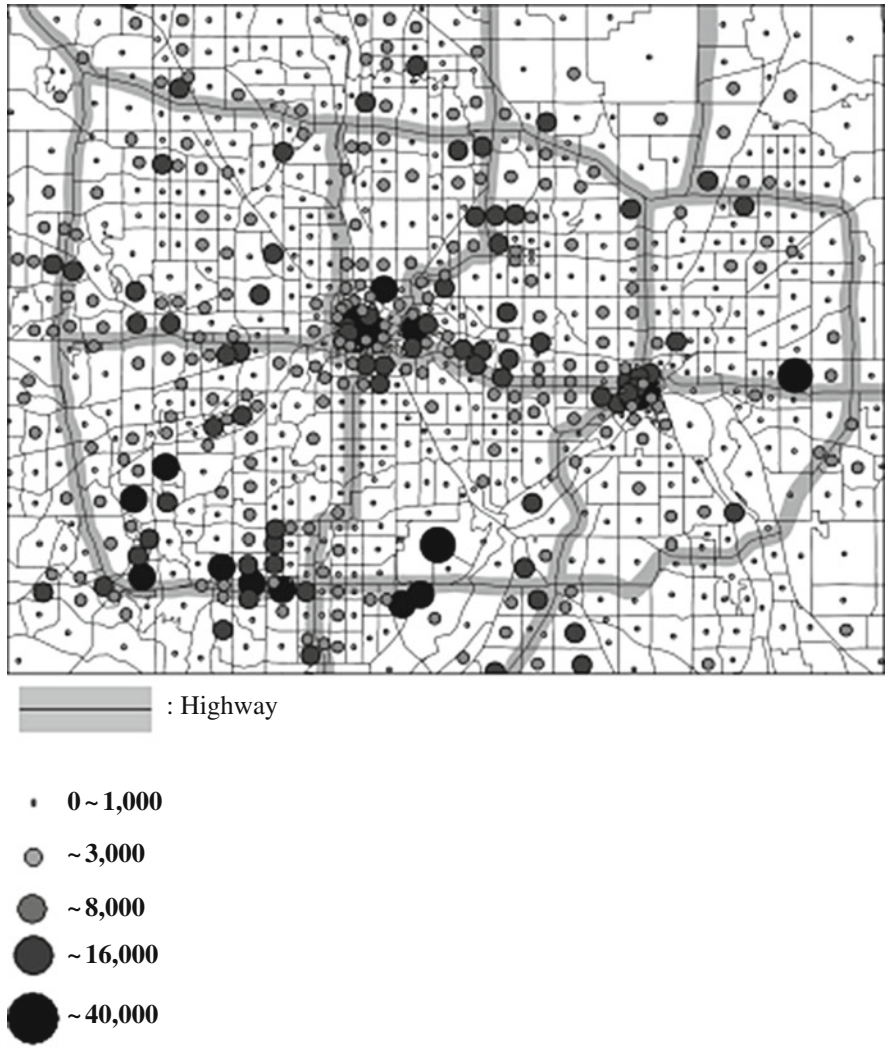


Contraflow in Transportation Network, Fig. 4 Graph representation of a simple evacuation situation and two following contraflow configuration candidates

transshipment nodes. The evacuation time of the original network in Fig. 4a is 22, which can be measured using minimum cost flow algorithm.

Figure 4b and c illustrate two possible contraflow configurations based on the original graph. All the two-way edges used in the original configuration are merged by capacity and directed in favor of increasing outbound evacuation capacity. There are two candidate configurations that differ in the direction of edges between nodes B and D. If the evacuation times of both configurations are measured, the configuration in Fig. 4b has evacuation time 11, while the configuration in Fig. 4c has evacuation time 14. Both configurations not only reduce but also differ in evacuation time. Even though the time difference is just 3 in this example, the difference may be significantly different in the case of a complicated real network. This example illustrates the importance of choice among possible network configurations. In addition, there are critical edges affecting the evacuation time, such as edge (B, D) in Fig. 4.

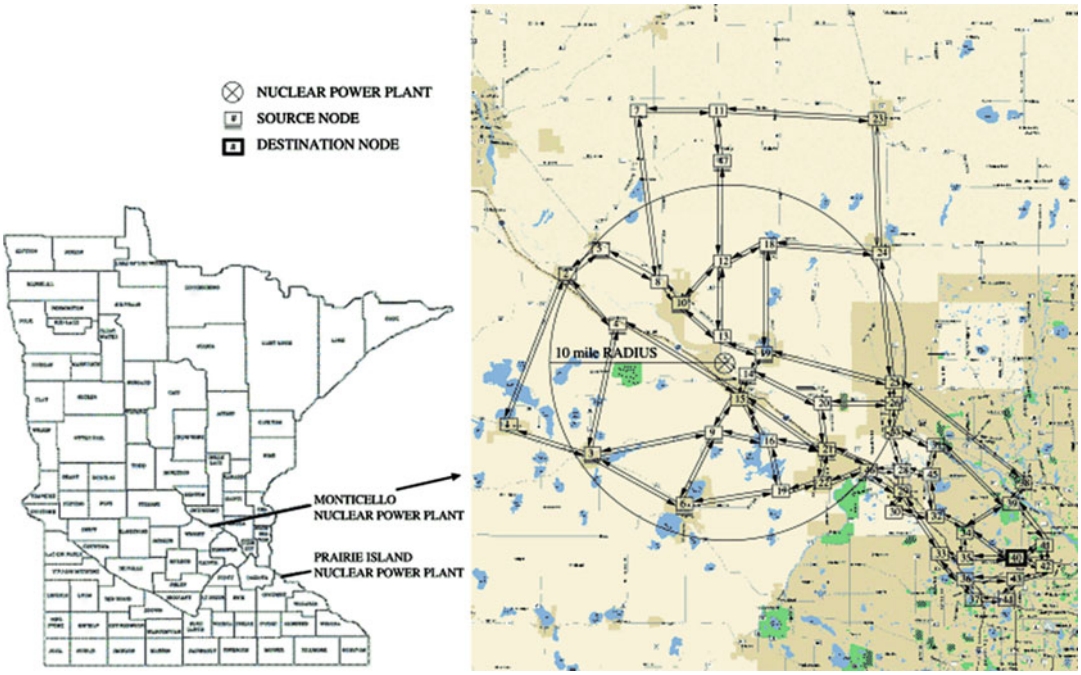
Solutions for Contraflow Planning: S. Kim et al. (Kim and Shekhar 2005) presented heuristic approaches to find a sub-optimal contraflow network configuration from a given network. Their approaches used the congestion status of a road network to select the most effective target



Contraflow in Transportation Network, Fig. 5 Day-time population distribution in the Twin Cities, Minnesota

road segments. The experimental results showed that reversing less than 20 % of all road segments was enough to reduce evacuation time by more than 40 %. Tuydes and Ziliaskopoulos (2004) proposed a mesoscopic contraflow network model based on a dynamic traffic assignment method. They formulated capacity reversibility using a mathematical programming method. Theodoulou and Wolshon (2004) used CORSIM microscopic traffic simulation to model the freeway contraflow evacuation around New Orleans. With the help of a micro scale traffic simulator, they were able to suggest alternative contraflow configurations at the level of entry and termination points.

Datasets for Contraflow Planning: When emergency managers plan contraflow schemes, the following datasets may be considered. First, population distribution is important to predict congested road segments and to prepare resources accordingly. Figure 5 shows a day-time population distribution in the Twin Cities, Minnesota. The dataset is based on Census 2000



Contraflow in Transportation Network, Fig. 6 Monticello nuclear power plant located around Twin Cities, Minnesota

and employment Origin-Destination estimate from the Minnesota Department of Employment and Economic Development, 2002.

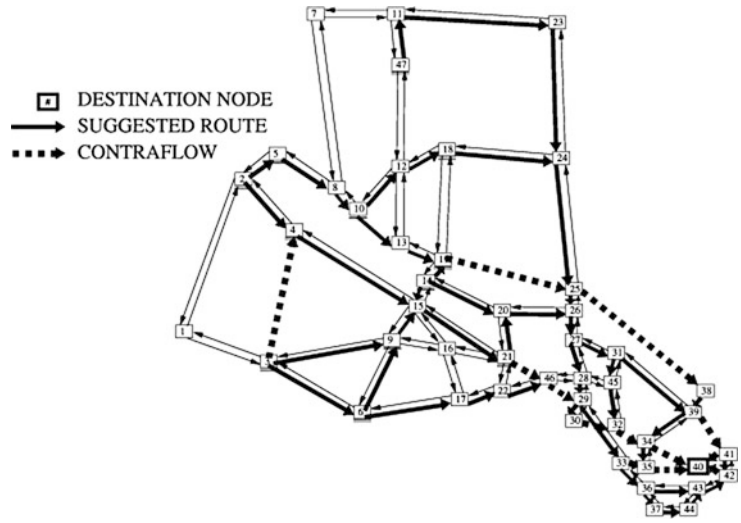
Second, a scenario dataset needs to be prepared. The scenario may include road network, accident location, affected area, destination (e.g., evacuation shelter). Figure 6 shows a virtual scenario of a nuclear power plant failure in Monticello, Minnesota. There are twelve cities directly affected by the failure within 10 miles of the facility and one destination shelter. The affected area is highly dynamic in this case because wind direction can change the shape of the affected area. The road network in the scenario is based on Interstate highway (I-94) and major arterial roads.

Figure 7 shows a possible contraflow scheme based on the Twin Cities population distribution and Monticello nuclear power plant scenario. In this scheme, the dotted road segments represent suggested contraflow. If the suggested road segments are reversed as contraflow, the evacuation time can be reduced by a third from the results of computer simulation.

Key Applications

Evacuation under Emergency: When a contraflow program is executed under an emergency situation, several factors should be taken into account: traffic control, accessibility, merging of lanes, use of roadside facilities, safety, labor requirements, and cost (Wolshon 2001). Among these factors, there is a tradeoff between contraflow and safety because most freeways and arterial roads are originally designed for one way direction. An easy example would be a driver who cannot see a traffic light if he drives in the opposite direction. Thus, considerable resources (e.g., police officers, barricade trucks, etc) and cost are required when a contraflow program is implemented. Most coastal states threatened by hurricanes every year prepare contraflow schemes. According to Wolshon (2001), 11 out of 18 coastal states have contraflow plans in place.

The application of contraflow for various desaster types is somewhat limited due to the following reason. Table 2 presents various types



Contraflow in Transportation Network, Fig. 7 A possible contraflow scheme for Monticello nuclear power plant scenario

Contraflow in Transportation Network, Table 2 Different types of disasters present different types of evacuation properties (Source: Litman 2006)

Type of disaster	Geographic scale	Warning	Contraflow before	Contraflow after
Hurricane	Very large	Days	√	√
Flooding	Large	Days	√	√
Earthquake	Large	None		√
Tsunami	Very large	Short		√
Radiation/toxic release	Small to large	Sometimes		√

of disasters and their properties. According to Litman (2006), evacuation route plans should take into account the geographic scale and length of warning. Contraflow preparedness is most appropriate for disasters with large geographic scale and long warning time, which gives responders time to dispatch resources and establish reversed lanes. Thus, hurricane and flooding are the most appropriate candidates to apply contraflow plans before disaster. Other types of disasters with relatively short warning time may consider contraflow only after disaster to resolve traffic congestion for back home traffic.

Automated Reversible Lane System: Washington D. C. has an automated reversible lane system to address the daily traffic jams during morning and evening peak time (Metropolitan Washington Council of Governments 2004). For example, Interstate Highway 95 operates 28 miles of reversible lanes during 6:00–9:00 AM and 3:30–6:00 PM. The reversible lane system has proven to provide substantial savings in travel time. Reversible lanes are also commonly found in tunnels and on bridges. The Golden Gate Bridge in San Francisco has 2 reversible lanes. Figure 8 shows a controlled access of reversible lane system.

Others: Contraflow programs are used for events with high density population such as football games, concerts, and fireworks on the Fourth of July. Highway construction sometimes requires contraflow. Figure 9 shows an example of contraflow use for highway construction.

Future Directions

For emergency professionals, there are many issues to be solved with regard to contraflow. Currently many planners rely on educated guesses with handcrafted maps to plan contraflow. They need computerized contraflow tools to acquire precise quantification (e.g., evacuation time) of

Contraflow in Transportation Network, Fig. 8 Controlled access of automated reversible lane system in Adelaide (Source: wikipedia.org)

Contraflow in Transportation Network, Fig. 9 Use of contraflow for highway construction on I-10 in Arizona (Source: map.google.com)

contraflow networks with geographic and demographic data. Other assessments are also required to plan efficient resource use, appropriate termination of contraflow, road markings, post disaster re-entry, etc. For researchers, the development of efficient and scalable contraflow simulation models and tools are urgent tasks. As shown in the Rita evacuation, a large scale evacuation (i.e., three million evacuees) is no longer an unusual event. Efficient tools available in the near future should handle such large scale scenarios with high fidelity traffic models.

Cross-References

- ▸ Emergency Evacuation, Dynamic Transportation Models
- ▸ Emergency Evacuations, Transportation Networks

References

Kim S, Shekhar S (2005) Contraflow network reconfiguration for evacuation planning: a summary of results. In: Proceedings of the 13th ACM symposium on

advances in geographic information systems, Bremen, pp 250–259
Litman T (2006) Lessons from Katrina and Rita: what major disasters can teach transportation planners. J Trans Eng 132(1):11–18
Metropolitan Washington Council of Governments (2005) 2004 performance of regional high-occupancy vehicle facilities on freeways in the Washington region. Analysis of person and vehicle volumes
Theodoulou G, Wolshon B (2004) Alternative methods to increase the effectiveness of freeway contraflow evacuation. J Trans Res Board Transp Res Rec 1865:48–56
Tuydes H, Ziliaskopoulos A (2004) Network re-design to optimize evacuation contraflow. Technical report 04-4715, Presented at 83rd Annual Meeting of the Transportation Research Board
Wolshon B (2001) One-way-out: contraflow freeway operation for hurricane evacuation. Natl Hazard Rev 2(3):105–112
Wolshon B, Urbina E, Levitan M (2002) National review of hurricane evacuation plans and policies. Technical report, Hurricane Center, Louisiana State University, Baton Rouge

Contraint Relations

▸ Constraint Databases and Data Interpolation

Convergence of GIS and CAD

▸ Computer Environments for GIS and CAD

Converging

▸ Movement Patterns in Spatio-Temporal Data

Co-occurrence

▸ Co-location Pattern Discovery
▸ Co-location Patterns, Algorithms
▸ Statistically Significant Co-location Pattern Mining

Coordinate Systems

▸ Reference Frames

Coregistration

▸ Registration

Co-registration

▸ Registration

Correlated

▸ Patterns, Complex

Correlated Frailty Models

▸ Spatial Survival Analysis

Correlated Walk

▸ CrimeStat: A Spatial Statistical Program for the Analysis of Crime Incidents

Correlation and Spatial Autocorrelation

Sang-Il Lee
Department of Geography Education, Seoul National University, Seoul, South Korea

Definition

Spatial autocorrelation or spatial dependence can be defined as a particular relationship between the spatial proximity among observational units and the numeric similarity among their values; positive spatial autocorrelation refers to situations in which the nearer the observational units, the more similar their values (and vice versa for

its negative counterpart). The presence of spatial autocorrelation or dependence means that a certain amount of information is shared and duplicated among neighboring locations, and thus, an entire data set possesses a certain amount of redundant information. This feature violates the assumption of independent observations upon which many standard statistical treatments are predicated. This entry revolves around what happens to the nature and statistical significance of correlation coefficients (e.g., Pearson's r) when spatial autocorrelation is present in both or either of the two variables under investigation.

Historical Background

A lack of independence results in reduced degrees of freedom or effective sample size; the greater the level of spatial autocorrelation, the smaller the number of degrees of freedom or effective sample size. This means that any type of statistical test based on an original sample size could be flawed in the presence of spatial autocorrelation, thus heightening the probability of committing a Type I error. Suppose that $n!$ different map patterns are generated from n observations. Because the $n!$ different map patterns are identical in terms of sample mean and variance, any statistical inferences based on these values are identical. However, all of the map patterns possess different degrees of freedom or effective sample size, and thus $n!$ different statistical estimations should be obtained.

This type of problem occurs in situations dealing with the correlation between two variables, which has long been known (Bivand 1980; Griffith 1980; Haining 1980; Richardson and Hémon 1981). The presence of spatial autocorrelation in both or either of two variables under investigation (i.e., bivariate spatial dependence) means that when the nature of a bivariate association at a location is known, one can guess the nature of bivariate associations at nearby locations. For example, if a location has a pair of higher-than-average values for two variables, there is a more-than-random chance to observe similar pairs in nearby locations. This feature again violates the assumption of independent observations and reduces the number of degrees of freedom or effective sample size. In this context, standard inferential tests tend to underestimate the true sampling variance of the Pearson's correlation coefficient when positive spatial autocorrelation is present in two variables under investigation, resulting in a heightened chance of committing a Type I error. One can generate $n!$ different pairs of spatial patterns from the original variables; all of the pairs are identical in terms of Pearson's correlation coefficient, but they are different in terms of the number of degrees of freedom or effective sample size (Clifford and Richardson 1985; Clifford et al. 1989; Haining 1991; Dutilleul 1993). These notions can extend to situations dealing with a pair of regression residuals (Tiefelsdorf 2001).

Two different approaches exist, addressing the problem of spatial autocorrelation in bivariate correlation. One is to seek to remedy the problem by providing modified hypothesis testing procedures taking the degree of spatial autocorrelation into account (for a comprehensive review and discussion, see Griffith and Paelinck 2011). The other is to develop bivariate spatial autocorrelation statistics to capture the degree of spatial *co-patterning* between two map patterns and, further, to propose some techniques for exploratory spatial data analysis (ESDA) that allow the detecting of bivariate spatial clusters (among others, Lee 2001; Anselin et al. 2002; Lee 2012).

Scientific Fundamentals

For this section, I seek to conceptualize and illustrate the concept of *bivariate spatial dependence* with which the problems of correlation in the presence of spatial autocorrelation are better captured and tackled. For simplicity, subsequent discussions about spatial autocorrelation tend to refer to its positive component.

Nearly all studies about spatial autocorrelation focus on univariate cases, i.e., on the similarity/dissimilarity in nearby locations in a single map pattern in terms of their values. However, *correlation* could be a legitimate statistical

concept endemic to bivariate situations. A correlation coefficient should gauge the nature (direction and magnitude) of the relationship between two variables under investigation. Interestingly, spatial autocorrelation can be viewed as a particular case of correlation, although only a single variable is involved, which is why it is known as *auto*correlation. Because any type of correlation should entail two vectors, another vector should be *spatially* derived for spatial autocorrelation to be a type of correlation. One of the most commonly used concepts for this case is a *spatial lag* vector, each element of which represents a weighted mean of a location's neighbors. In this sense, spatial autocorrelation could be rephrased as the correlation between one variable and its spatial lag vector (Lee 2001).

But, what kinds of issues can arise when we combine the two concepts, correlation and spatial autocorrelation? This question might be better captured by a rather new concept known as *bivariate spatial dependence*, which is a simple extension of the general concept of spatial dependence, and can be defined as "a particular relationship between the spatial proximity among observational units and the numeric similarity of their bivariate associations" (Lee 2001, 2012). In a bivariate situation, each observational unit contains a pair of values, and the nature of the bivariate association is assumed to be conceptually defined and numerically evaluated. If the distribution of bivariate associations is not spatially random, then we might legitimately state that bivariate spatial dependence exists.

Before attempting to illustrate the concept of bivariate spatial dependence, we begin with univariate spatial dependence. Any local set composed of a reference observational unit and its neighbors takes on one of the following four types of *univariate spatial association*:

$$\mathbf{H}-\tilde{\mathbf{H}} \qquad \mathbf{H}-\tilde{\mathbf{L}} \qquad \mathbf{L}-\tilde{\mathbf{H}} \qquad \mathbf{L}-\tilde{\mathbf{L}} \quad (1)$$

Here, **H** denotes a value at a reference unit that is greater than or equal to a threshold value (usually the average) or a positive z-score (original values having the mean subtracted and then divided by the standard deviation), and **L** denotes the opposite. $\tilde{\mathbf{H}}$ denotes a *spatial lag* that is greater than or equal to the global average, and $\tilde{\mathbf{L}}$ denotes the opposite. The symbol '$-$' denotes a univariate horizontal relationship. The symbol "$\sim$" is introduced here to make a clear distinction between an original value at a location and a derived value from a set of locations. This conceptualization is the basis for the Moran's I statistic.

If another concept (i.e., the *spatial moving average*) is introduced, the situation changes substantively. Unlike the spatial lag, this concept treats the reference unit itself as one of its neighbors. Consider

$$\tilde{\mathbf{H}}^* \quad \tilde{\mathbf{L}}^* \quad (2)$$

Here, $\tilde{\mathbf{H}}^*$ and $\tilde{\mathbf{L}}^*$ denote the spatial moving averages at each location. This conceptualization forms the foundation for the Getis-Ord's G_i^* statistic. The four types of univariate spatial association listed in (1) reduce to the two values in (2); $\mathbf{H}-\tilde{\mathbf{H}}$ and $\mathbf{L}-\tilde{\mathbf{L}}$ respectively are linked to $\tilde{\mathbf{H}}^*$ and $\tilde{\mathbf{L}}^*$, but $\mathbf{H}-\tilde{\mathbf{L}}$ and $\mathbf{L}-\tilde{\mathbf{H}}$ can point either way, depending on the differences in values and/or spatial weights. These two values can be conceptualized as two different types of *univariate spatial clusters* (Lee and Cho 2013). This distinction between spatial association types and spatial cluster types is critical because it can represent the two contrasting perspectives of *spatial modeling* and *spatial exploration*. This distinction plays a pivotal role in addressing various issues about multivariate spatial dependence, a particular case of which is bivariate spatial dependence.

We now move to bivariate situations in which two variables, denoted by X and Y, are under investigation. Each observational unit should take on one of the following four types of *bivariate association* (Lee 2012):

$$\begin{array}{cccc} \mathbf{H} & \mathbf{H} & \mathbf{L} & \mathbf{L} \\ | & | & | & | \\ \mathbf{H} & \mathbf{L} & \mathbf{H} & \mathbf{L} \end{array} \quad (3)$$

In this work, the symbol '|' denotes a bivariate vertical relationship at a location. Pearson's correlation coefficient is predicated upon this conceptualization and is *aspatial* in nature in the

sense that it does not consider the spatial distribution of the pair-wise local bivariate associations.

Suppose that a location has only one neighbor at which the four different types of bivariate association are possible, resulting in the following 16 different types of bivariate spatial association:

$$\begin{array}{cccc}
\begin{array}{c}\mathbf{H-H}\\ |\ \ |\\ \mathbf{H-H}\end{array} & \begin{array}{c}\mathbf{H-H}\\ |\ \ |\\ \mathbf{H-L}\end{array} & \begin{array}{c}\mathbf{H-L}\\ |\ \ |\\ \mathbf{H-H}\end{array} & \begin{array}{c}\mathbf{H-L}\\ |\ \ |\\ \mathbf{H-L}\end{array} \\
\begin{array}{c}\mathbf{H-H}\\ |\ \ |\\ \mathbf{L-H}\end{array} & \begin{array}{c}\mathbf{H-H}\\ |\ \ |\\ \mathbf{L-L}\end{array} & \begin{array}{c}\mathbf{H-L}\\ |\ \ |\\ \mathbf{L-H}\end{array} & \begin{array}{c}\mathbf{H-L}\\ |\ \ |\\ \mathbf{L-L}\end{array} \\
\begin{array}{c}\mathbf{L-H}\\ |\ \ |\\ \mathbf{H-H}\end{array} & \begin{array}{c}\mathbf{L-H}\\ |\ \ |\\ \mathbf{H-L}\end{array} & \begin{array}{c}\mathbf{L-L}\\ |\ \ |\\ \mathbf{H-H}\end{array} & \begin{array}{c}\mathbf{L-L}\\ |\ \ |\\ \mathbf{H-L}\end{array} \\
\begin{array}{c}\mathbf{L-H}\\ |\ \ |\\ \mathbf{L-H}\end{array} & \begin{array}{c}\mathbf{L-H}\\ |\ \ |\\ \mathbf{L-L}\end{array} & \begin{array}{c}\mathbf{L-L}\\ |\ \ |\\ \mathbf{L-H}\end{array} & \begin{array}{c}\mathbf{L-L}\\ |\ \ |\\ \mathbf{L-L}\end{array}
\end{array} \tag{4}$$

This association is both bivariate *and* spatial because two pairs (bivariate) in adjacent locations (spatial) are compared. The four main diagonal elements clearly show positive bivariate spatial dependence because exactly the same types of pairs of bivariate association are connected. In contrast, the four anti-diagonal elements can be viewed as examples of negative bivariate spatial dependence because rather different types of bivariate association are placed next to each other.

We next consider additional neighbors. If one more neighbor is added, then we have $4^3 = 64$ different types of bivariate spatial association for each local set. The situation becomes more complicated, although a decent chance still exists for observational units to show perfect and positive bivariate spatial dependence. Because areal units in the real world (i.e., administrative units, school districts, and other types of functional regions) have been reported to have approximately six contiguous neighbors on average, we consider $4^7 = 16{,}384$ different types of bivariate spatial associations at each location, and little chance exists of identifying those showing typical positive bivariate spatial dependence.

We might be able to simplify the situation by applying the notion of spatial lag as seen in (1). Because each variable has four different types of univariate spatial association at a location, we always have only 16 different types of *bivariate spatial association* (Lee 2012; Lee and Cho 2013), no matter how many neighbors are involved. Consider

$$\begin{array}{cccc}
\begin{array}{c}\mathbf{H-\tilde{H}}\\ |\ \ |\\ \mathbf{H-\tilde{H}}\end{array} & \begin{array}{c}\mathbf{H-\tilde{H}}\\ |\ \ |\\ \mathbf{H-\tilde{L}}\end{array} & \begin{array}{c}\mathbf{H-\tilde{H}}\\ |\ \ |\\ \mathbf{L-\tilde{H}}\end{array} & \begin{array}{c}\mathbf{H-\tilde{H}}\\ |\ \ |\\ \mathbf{L-\tilde{L}}\end{array} \\
\begin{array}{c}\mathbf{H-\tilde{L}}\\ |\ \ |\\ \mathbf{H-\tilde{H}}\end{array} & \begin{array}{c}\mathbf{H-\tilde{L}}\\ |\ \ |\\ \mathbf{H-\tilde{L}}\end{array} & \begin{array}{c}\mathbf{H-\tilde{L}}\\ |\ \ |\\ \mathbf{L-\tilde{H}}\end{array} & \begin{array}{c}\mathbf{H-\tilde{L}}\\ |\ \ |\\ \mathbf{L-\tilde{L}}\end{array} \\
\begin{array}{c}\mathbf{L-\tilde{H}}\\ |\ \ |\\ \mathbf{H-\tilde{H}}\end{array} & \begin{array}{c}\mathbf{L-\tilde{H}}\\ |\ \ |\\ \mathbf{H-\tilde{L}}\end{array} & \begin{array}{c}\mathbf{L-\tilde{H}}\\ |\ \ |\\ \mathbf{L-\tilde{H}}\end{array} & \begin{array}{c}\mathbf{L-\tilde{H}}\\ |\ \ |\\ \mathbf{L-\tilde{L}}\end{array} \\
\begin{array}{c}\mathbf{L-\tilde{L}}\\ |\ \ |\\ \mathbf{H-\tilde{H}}\end{array} & \begin{array}{c}\mathbf{L-\tilde{L}}\\ |\ \ |\\ \mathbf{H-\tilde{L}}\end{array} & \begin{array}{c}\mathbf{L-\tilde{L}}\\ |\ \ |\\ \mathbf{L-\tilde{H}}\end{array} & \begin{array}{c}\mathbf{L-\tilde{L}}\\ |\ \ |\\ \mathbf{L-\tilde{L}}\end{array}
\end{array} \tag{5}$$

Each observational unit is assigned to one of these 16 types in terms of *local* bivariate spatial dependence. Certain interesting findings are drawn from this illustration. First, the four cases of perfect positive spatial dependence are observed in the four corners, and their four negative counterparts are observed in the middle. Second, the four cases in the main diagonal are more closely associated with a positive aspatial correlation, measured by Pearson's correlation coefficient, and the four cases in the anti-diagonal are more strongly associated with a negative aspatial correlation. This notion is not confined to Pearson's r, but can extend to other linear correlation coefficients (see Griffith and Amrhein 1991).

By combining these two aspects, we can make certain general statements. First, with a decent level of positive Pearson's r, the main diagonal

cases are expected to be more observable than the anti-diagonal cases. If the first and last cases prevail for a local set, a positive bivariate spatial dependence can be said to exist; if the second and third cases prevail, a negative bivariate spatial dependence can be said to exist. Second, with a decent level of negative Pearson's r, the anti-diagonal cases are expected to be more observable than the main diagonal counterparts. If the first and last cases prevail for a local set, a positive bivariate spatial dependence can be said to exist; if the second and third cases prevail, a negative bivariate spatial dependence can be said to exist. In an overall sense, if no bivariate spatial autocorrelation exists, the 16 different types of bivariate spatial association (occurrences of which are subordinate to the nature of the global aspatial correlation) must be randomly distributed; otherwise, they should show a certain degree of spatial clustering.

These situations are further simplified by incorporating the notion of spatial moving average. Because the four different types of univariate spatial association defined in (1) reduce to the two different values seen in (2), the 16 different types of bivariate spatial association defined in (5) can reduce to the following four:

$$\begin{array}{cccc} \tilde{\mathbf{H}}^* & \tilde{\mathbf{H}}^* & \tilde{\mathbf{L}}^* & \tilde{\mathbf{L}}^* \\ | & | & | & | \\ \tilde{\mathbf{H}}^* & \tilde{\mathbf{L}}^* & \tilde{\mathbf{H}}^* & \tilde{\mathbf{L}}^* \end{array} \tag{6}$$

These classifications can be referred to as four different types of *bivariate spatial clusters* (Lee and Cho 2013). The cases in the four corners in (5) represent typical examples of the four types; the others are classified into one of the four cases, depending on differences in values and/or spatial weights.

Key Applications

For this section, I focus on two strands of endeavors that have been undertaken in this particular field: one is to develop a means to remedy the problem of correlation in the presence of bivariate spatial dependence; the other is to devise bivariate spatial autocorrelation statistics for the bivariate counterparts of Moran's I and Getis-Ord's G_i^* statistics.

The test statistic for Pearson's r is given by

$$t = r\sqrt{n-2}\Big/\sqrt{1-r^2} \tag{7}$$

with $n-2$ degrees of freedom when the following two assumptions are satisfied: pairs of observations are drawn from the same, approximately bivariate normal, distribution with constant expectation and finite variance (Haining 1991) and observations of each variable are mutually independent. This standard hypothesis testing procedure for the correlation coefficient might not hold for spatial data. The first assumption of a constant mean structure cannot be assumed because of the potential presence of a global trend. More importantly, the second assumption cannot be sustained because of the usual presence of univariate spatial autocorrelation for both or either of the variables under investigation, which alludes to bivariate spatial dependence.

The standard error of Pearson's r, which is also a part of (7), is given by

$$\hat{\sigma}_r = \sqrt{\frac{1-r^2}{n-2}}, \tag{8}$$

where the denominator is associated with the number of degrees of freedom. This standard error should be adjusted according to the degree of spatial autocorrelation in the variables; it should be larger when positive spatial autocorrelation prevails (and vice versa for negative spatial autocorrelation) (Haining 1991). This outcome can be shown in (8); the lack of independence among pairs of observations due to positive bivariate spatial dependence reduces the number of degree of freedom or effective sample size, thus making the standard error larger.

Several approaches have been proposed in order to remedy or at least alleviate the problem of underestimation of the true sampling variance that the standard inferential test commits (Clifford and Richardson 1985; Dutilleul 1993). In this entry, we focus solely

on the Clifford-Richardson's solution (for a more comprehensive treatment, see Griffith and Paelinck 2011). They redefine the equation for the standard error by replacing n in (8) with n', their "effective sample size," which arguably refers to the number of equivalent, independent samples:

$$\hat{\sigma}_r = \sqrt{\frac{1-r^2}{n'-2}}. \tag{9}$$

They also provide the equation for computing the effective sample size as

$$n' = 1 + n^2 \left[trace \left(\hat{\mathbf{R}}_X \hat{\mathbf{R}}_Y \right) \right]^{-1}, \tag{10}$$

where $\hat{\mathbf{R}}_X$ and $\hat{\mathbf{R}}_Y$ are the estimated $n \times n$ spatial autocorrelation matrices for the two variables and the *trace* is a matrix operation which is the sum of the diagonal elements. Because each diagonal element of matrix $\hat{\mathbf{R}}_X \hat{\mathbf{R}}_Y$ can be seen as the relative degree of spatial autocorrelation at each location (1 for no spatial autocorrelation, more than 1 for positive spatial autocorrelation), $trace \left(\hat{\mathbf{R}}_X \hat{\mathbf{R}}_Y \right)$ captures the overall degree of bivariate spatial dependence. If no spatial autocorrelation is present for either of the two variables across all locations, each diagonal element of matrix $\hat{\mathbf{R}}_X \hat{\mathbf{R}}_Y$ is 1, $trace \left(\hat{\mathbf{R}}_X \hat{\mathbf{R}}_Y \right) = n$, and thus $n' \cong n$ (Haining 1991). If a positive bivariate spatial dependence prevails, n' is less than n, resulting in a reduced effective sample size or a lesser number of degrees of freedom.

Suppose, for example, that we have 50 pairs of observations and a Pearson's r of 0.3. The test statistic and the number of degrees of freedom according to the standard hypothesis testing method as shown in (7) are, respectively, 2.179 and 48, which implies that r is statistically significant ($p = 0.0343$). If we have a positive bivariate spatial autocorrelation of 2.0 on average across locations, then we have $t = 1.541$ with the effective sample size of 26 ($1 + 50^2/100$) according to (10), which is not statistically significant ($p = 0.1365$). This Clifford-Richardson's solution is implemented in an *R* package named SpatialPack (Vallejos et al. 2013).

Any bivariate spatial autocorrelation statistic should capture the degree of *spatial co-patterning* by measuring both pair-wise covariance and spatial clustering (Lee 2001). One of the most important considerations in determining how to measure bivariate spatial dependence might be the fact that both Pearson's r and Moran's I are cross-product statistics (Getis 1991), which take the form of an average of the sum of products of two vectors. Pearson's r is defined as an average of the cross-product of two standardized vectors, $\mathbf{z}_X$ and $\mathbf{z}_Y$; similarly, Moran's I can be defined as an average of the cross-product of two standardized vectors, $\mathbf{z}_X$ and $\tilde{\mathbf{z}}_X$ (a standardized spatial lag vector), when a spatial weights matrix is row standardized (Lee 2001):

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}} \tag{11}$$

$$= \frac{1}{n} \sum_i z_{X_i} z_{Y_i}, \text{ and}$$

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij} (x_i - \bar{x}) (x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

$$= \frac{\sum_i (x_i - \bar{x}) \sum_j w_{ij} (x_j - \bar{x})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (x_i - \bar{x})^2}}$$

$$= \frac{1}{n} \sum_i z_{X_i} \tilde{z}_{X_i}. \tag{12}$$

Predicated upon all of the discussions about univariate statistics for spatial autocorrelation, Lee (2012) identifies six vectors that may play roles in defining bivariate spatial autocorrelation statistics conforming to the general form of cross-product statistic: $\mathbf{z}_X$, $\tilde{\mathbf{z}}_X$, and $\tilde{\mathbf{z}}_X^*$ (a standardized spatial moving average vector) for the X variable and $\mathbf{z}_Y$, $\tilde{\mathbf{z}}_Y$, and $\tilde{\mathbf{z}}_Y^*$ for the Y variable. Using these two sets of vectors, one can obtain various types of bivariate spatial autocorrelation statistics. In this entry, only the following two are discussed (i.e., the cross-Moran or bivariate Moran statistic denoted by CM and Lee's L^* statistic):

$$CM = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij} (x_i - \bar{x}) (y_j - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}} = \frac{1}{n} \sum_i z_{X_i} \tilde{z}_{Y_i}, \text{ and}$$

$$L^* = \frac{n}{\sum_i \left(\sum_j w_{ij}^* \right)^2} \frac{\sum_i \left[\left(\sum_j w_{ij}^* (x_j - \bar{x}) \right) \left(\sum_j w_{ij}^* (y_j - \bar{y}) \right) \right]}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}} = \frac{1}{n} \sum_i \tilde{z}_{X_i}^* \tilde{z}_{Y_i}^*. \quad (13)$$

Here, w_{ij} and w_{ij}^* are elements from a zero diagonal and nonzero diagonal spatial weights matrix, respectively. The former statistic is one derived from a multivariate spatial correlation matrix proposed by Wartenberg (1985) and is a simple extension of univariate Moran's I, thus gauging the correlation between one variable at original locations and the other variable at the neighboring locations (a spatial lag vector). In contrast, the latter, which was proposed by Lee (2001, 2004, 2009), is defined as the correlation between one variable and the other variable's spatial moving average vectors. In comparison, cross-Moran is more congruent with the concept of *cross-correlation*, whereas Lee's L^* deals more directly with the concept of *co-patterning* by considering not only bivariate association at the original locations but also their spatial association with neighboring locations.

In examining the different advantages and weaknesses, one can conclude that the bivariate Moran's statistic is more congruent with the spatial modeling perspective, whereas Lee's statistic is more strongly associated with the spatial exploration perspective. For example, many situations might exist in which one should postulate that a dependent variable at a given set of locations is influenced by independent variables in the neighboring locations. However, if the main interest lies in measuring the spatial similarity between the two map patterns, and exploring and detecting possible bivariate spatial clusters, L^* might be the better option. In addition, L^* is much more congruent with what is documented in (6). The higher the Pearson's aspatial correlation coefficient, and at the same time the higher the level of spatial clustering of bivariate association, the higher the L^* statistic. Certain exploratory spatial data analysis (ESDA) techniques using Lee's local L_i^* (see Eq. 14) can be developed like ones using cross-Moran (Anselin et al. 2002), which is beyond the scope of this entry (see Lee 2012; Lee and Cho 2013):

$$L_i^* = \frac{n^2}{\sum_i \left(\sum_j w_{ij}^* \right)^2} \frac{\left(\sum_j w_{ij}^* (x_j - \bar{x}) \right) \left(\sum_j w_{ij}^* (y_j - \bar{y}) \right)}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}} = \tilde{z}_{X_i}^* \tilde{z}_{Y_i}^* \quad (14)$$

The distributional properties for all bivariate spatial autocorrelation statistics have been established with the randomization assumption (Lee 2004, 2009), which might be crucial to develop certain kinds of ESDA techniques, such as bivariate cluster maps.

Future Directions

Bivariate spatial dependence points to situations in which nearby observational units carry shared information in terms of bivariate association, thus violating the assumption of independent sampling, and the shared information spuriously strengthens (or weakens) the nature of correlation between two variables under investigation, making any conventional statistical inferences or judgments considerably questionable.

The notion and procedure of correlation coefficient decomposition based on the eigenvector spatial filtering (ESF) technique (Griffith and Paelinck 2011; Chun and Griffith 2013) provides

an invaluable insight into our understanding of correlation with spatial autocorrelation. It allows an aspatial correlation coefficient to be decomposed into five *sub*-correlations between spatially filtered variables, common spatial autocorrelation components, unique spatial autocorrelation components, one's spatially filtered variable and the other's unique spatial autocorrelation component, and one's unique spatial autocorrelation component and the other's spatially filtered variable.

Bivariate spatial dependence or autocorrelation is a special case of multivariate spatial dependence or autocorrelation (Wartenberg 1985). For example, "trivariate" spatial dependence is simply defined as "a particular relationship between the spatial proximity among observational units and the numeric similarity of their trivariate associations." Thus, we have $4^3 = 64$ different types of *trivariate spatial association*, similar to (5), and $2^3 = 8$ different types of *trivariate spatial clusters*, similar to (6).

Because each pair of variables in a multivariate data set can be viewed as a building block for statistical treatments, the notion of bivariate spatial dependence should have certain implications in spatializing any form of multivariate statistical techniques, e.g., spatial principal components analysis (e.g., Griffith 1988; Dray et al. 2008; Lee and Cho 2014; Lee 2015) and spatial canonical correlation analysis.

Cross-References

- Spatial Autocorrelation and Spatial Interaction
- Spatial Autocorrelation Measures
- Spatial Filtering
- Spatial Statistics and Geostatistics: Basic Concepts

References

Anselin L, Syabri I, Smirnov O (2002) Visualizing multivariate spatial correlation with dynamically linked windows. In: Anselin L, Rey S (eds) New tools for spatial data analysis: proceedings of the specialist meeting, Center for Spatially Integrated Social Science (CSISS), University of California, Santa Barbara

Bivand R (1980) A Monte Carlo study of correlation coefficient estimation with spatially autocorrelated observations. Quaest Geogr 6:5–10

Clifford P, Richardson S (1985) Testing the association between two spatial processes. Stat Decis Suppl Issue 2:155–160

Clifford P, Richardson S, Hémon D (1989) Assessing the significance of the correlation between two spatial processes. Biometrics 45:123–134

Chun Y, Griffith DA (2013) Spatial statistics & geostatistics: theory and applications for geographic information science & technology. Sage, Los Angeles

Dray S, Sonia S, François D (2008) Spatial ordination of vegetation data using a generalization of Wartenberg's multivariate spatial correlation. J Veg Sci 19:45–56

Dutilleul P (1993) Modifying the t test for assessing the correlation between two spatial processes. Biometrics 49:305–314

Getis A (1991) Spatial interaction and spatial autocorrelation: a cross-product approach. Environ Plan A 23:1269–1277

Griffith DA (1980) Towards a theory of spatial statistics. Geogr Anal 12:325–339

Griffith DA (1988) Advanced spatial statistics: special topics in the exploration of quantitative spatial data series. Kluwer, Dordrecht

Griffith DA, Amrhein CG (1991) Statistical analysis for geographers. Prentice-Hall, Englewood Cliffs

Griffith DA, Paelinck, JH (2011) Non-standard spatial statistics and spatial econometrics. Springer, New York

Haining RP (1980) Spatial autocorrelation problems. In: Herbert DT, Johnston RJ (eds) Geography and the urban environment, vol 3. Wiley, New York, pp 1–44

Haining RP (1991) Bivariate correlation with spatial data. Geogr Anal 23:210–227

Lee S-I (2001) Developing a bivariate spatial association measure: an integration of Pearson's r and Moran's I. J Geogr Syst 3:369–385

Lee S-I (2004) A generalized significance testing method for global measures of spatial association: an extension of the Mantel test. Environ Plan A 36:1687–1703

Lee S-I (2009) A generalized randomization approach to local measures of spatial association. Geogr Anal 41:221–248

Lee S-I (2012) Exploring bivariate spatial dependence and heterogeneity: a comparison of bivariate measures of spatial association. Paper presented at the annual meeting of the association of American geographers, New York, 24–28 Feb

Lee S-I (2015) Some elaborations on spatial principal components analysis. Paper presented at the annual meeting of the association of American geographers, Chicago, 21–25 Apr

Lee S-I, Cho D (2013) Delineating the bivariate spatial clusters: a bivariate AMOEBA technique. Paper presented at the annual meeting of the association of American geographers, Los Angeles, 9–13 Apr

Lee S-I, Cho D (2014) Developing a spatial principal components analysis. Paper presented at the annual meeting of the association of American geographers, Tampa, 8–12 Apr

Richardson S, Hémon D (1981) On the variance of the sample correlation between two independent lattice processes. J Appl Probab 18:943–948

Tiefelsdorf M (2001) Specification and distributional properties of the spatial cross-correlation coefficient $C_{\varepsilon_1,\varepsilon_2}$. Paper presented at the Western Regional Science Conference, Palm Springs, 26 Feb

Vallejos R, Osorio F, Cuevas F (2013) SpatialPack – an R package for computing spatial association between two stochastic processes defined on the plane. Available via DIALOG. http://rvallejos.mat.utfsm.cl/Time%20Series%20I%202013/paper3.pdf. Accessed 12 Feb 2016

Wartenberg D (1985) Multivariate spatial correlation: a method for exploratory geographical analysis. Geogr Anal 17:263–283

Correlation Queries

▸ Correlation Queries in Spatial Time Series Data

Correlation Queries in Spatial Time Series Data

Pusheng Zhang
Microsoft Corporation, Redmond, WA, USA

Synonyms

Correlation Queries; Spatial Cone Tree; Spatial Time Series

Definition

A **spatial framework** consists of a collection of locations and a neighbor relationship. A **time series** is a sequence of observations taken sequentially in time. A **spatial time series dataset** is a collection of time series, each referencing a location in a common spatial framework. For example, the collection of global daily temperature measurements for the last 10 years is a spatial time series dataset over a degree-by-degree latitude-longitude grid spatial framework on the surface of the Earth.

Correlation queries are the queries used for finding collections, e.g. pairs, of highly correlated time series in spatial time series data, which might lead to find potential interactions and patterns. A strongly correlated pair of time series indicates potential movement in one series when the other time series moves.

Historical Background

The massive amounts of data generated by advanced data collecting tools, such as satellites, sensors, mobile devices, and medical instruments, offer an unprecedented opportunity for researchers to discover these potential nuggets of valuable information. However, correlation queries are computationally expensive due to large spatio-temporal frameworks containing many locations and long time sequences. Therefore, the development of efficient query processing techniques is crucial for exploring these datasets.

Previous work on query processing for time series data has focused on dimensionality reduction followed by the use of low dimensional indexing techniques in the transformed space. Unfortunately, the efficiency of these approaches deteriorates substantially when a small set of dimensions cannot represent enough information in the time series data. Many spatial time series datasets fall in this category. For example, finding anomalies is more desirable than finding well-known seasonal patterns in many applications. Therefore, the data used in anomaly detection is usually data whose seasonality has been removed. However, after transformations (e.g., Fourier transformation) are applied to deseasonalize the data, the power spectrum spreads out over almost all dimensions. Furthermore, in most spatial time series datasets, the number of spatial locations is much greater than the length of the time series. This makes it possible to improve the performance of query processing of spatial time series data by exploiting spatial proximity in the design of access methods.

In this chapter, the spatial cone tree, an spatial data structure for spatial time series data, is dis-

cussed to illustrate how correlation queries are efficiently supported. The spatial cone tree groups similar time series together based on spatial proximity, and correlation queries are facilitated using spatial cone trees. This approach is orthogonal to dimensionality reduction solutions. The spatial cone tree preserves the full length of time series, and therefore it is insensitive to the distribution of the power spectrum after data transformations.

Scientific Fundamentals

Let $x = \langle x_1, x_2, \ldots, x_m \rangle$ and $y = \langle y_1, y_2, \ldots, y_m \rangle$ be two time series of length m. The correlation coefficient of the two time series is defined as: $\text{corr}(x, y) = \frac{1}{m-1} \sum_{i=1}^{m} (\frac{x_i - \overline{x}}{\sigma_x}) \cdot (\frac{y_i - \overline{y}}{\sigma_y}) = \hat{x} \cdot \hat{y}$, where $\overline{x} = \frac{\sum_{i=1}^{m} x_i}{m}$, $\sigma_x = \sqrt{\frac{\sum_{i=1}^{m} (x_i - \overline{x})^2}{m-1}}$, $\overline{y} = \frac{\sum_{i=1}^{m} y_i}{m}$, $\sigma_y = \sqrt{\frac{\sum_{i=1}^{m} (y_i - \overline{y})^2}{m-1}}$, $\widehat{x_i} = \frac{1}{\sqrt{m-1}} \frac{x_i - \overline{x}}{\sigma_x}$, $\widehat{y_i} = \frac{1}{\sqrt{m-1}} \frac{y_i - \overline{y}}{\sigma_y}$, $\hat{x} = \langle \hat{x}_1, \hat{x}_2, \ldots, \widehat{x_m} \rangle$, and $\hat{y} = \langle \hat{y}_1, \widehat{y_2}, \ldots, \hat{y}_m \rangle$.

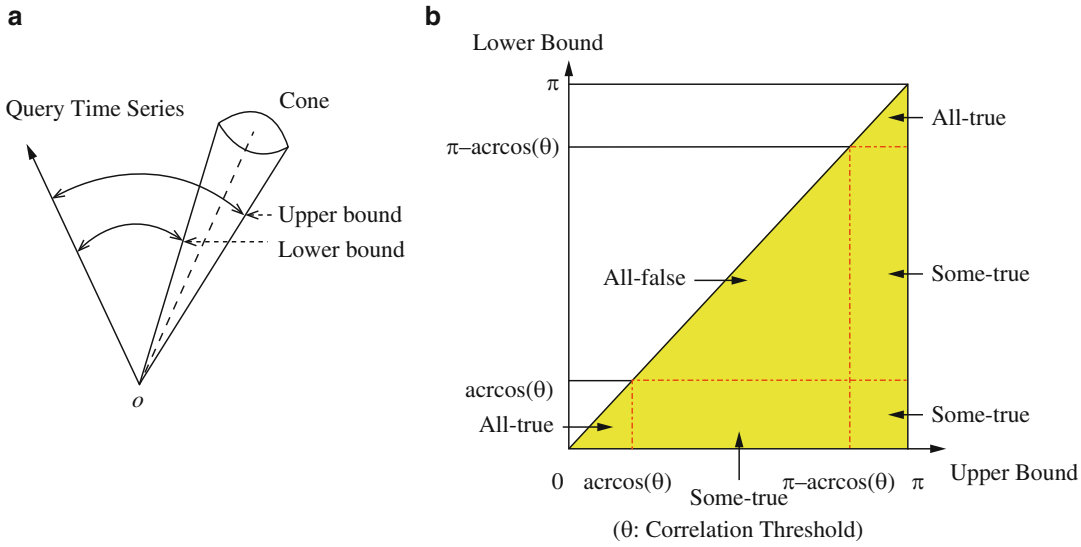
Because the sum of the $\widehat{x_i}^2$ is equal to 1: $\sum_{i=1}^{m} \widehat{x_i}^2 = \sum_{i=1}^{m} \left(\frac{1}{\sqrt{m-1}} \frac{x_i - \overline{x}}{\sqrt{\frac{\sum_{i=1}^{m} (x_i - \overline{x})^2}{m-1}}} \right)^2 = 1$, $\hat{x}$ is located in a multi-dimensional unit sphere. Similarly, $\hat{y}$ is also located in a multi-dimensional unit sphere. Based on the definition of $corr(x, y)$, $corr(x, y) = \hat{x} \cdot \hat{y} = \cos(\angle(\hat{x}, \hat{y}))$. The correlation of two time series is directly related to the angle between the two time series in the multi-dimensional unit sphere. *Finding pairs of time series with an absolute value of correlation above the user given minimal correlation threshold θ is equivalent to finding pairs of time series $\hat{x}$ and $\hat{y}$ on the unit multi-dimensional sphere with an angle in the range of* $(0, \arccos(\theta))$ *or* $(180° - \arccos(\theta), 180°)$.

A **cone** is a set of time series in a multi-dimensional unit sphere and is characterized by two parameters, the center and the span of the cone. The center of the cone is the mean of all the time series in the cone. The span τ of the cone is the maximal angle between any time series in the cone and the cone center.

A **spatial cone tree** is a spatial data structure for correlation queries on spatial time series data. The spatial cone tree uses a tree data structure, and it is formed of nodes. Each node in the spatial cone tree, except for the root, has one parent node and several-zero or more-child nodes. The root node has no parent. A node that does not have any child node is called a leaf node and a non-leaf node is called an internal node.

A leaf node contains a cone and a data pointer p_d to a disk page containing data entries, and is of the form $\langle$*(cone.span, cone.center)*, $p_d \rangle$. The cone contains one or multiple normalized time series, which are contained in the disk page referred by the pointer p_d. The *cone.span* and *cone.center* are made up of the characteristic parameters for the cone. The data pointer is a block address. An internal node contains a cone and a pointer p_i to an index page containing the pointers to children nodes, and is of the form $\langle$ *(cone.span, cone.center)*, $p_i \rangle$. The *cone.span* and *cone.center* are the characteristic parameters for the cone, which contains all normalized times series in the subtree rooted at this internal node. Multiple nodes are organized in a disk page, and the number of nodes per disk page is defined as the blocking factor for a spatial cone tree. Notice that the blocking factor, the number of nodes per disk page, depends on the sizes of cone span, cone center, and data pointer.

Given a minimal correlation threshold θ $(0 < \theta < 1)$, the possible relationships between a cone C and the query time series, T_q, consist of all-true, all-false, or some-true. All-true means that all times series with a correlation over the correlation threshold; all-false means all time series with a correlation less than the correlation threshold; some-true means only part of time series with a correlation over the correlation threshold. The upper bound and lower bound of angles between the query time series and a cone is illustrated in Fig. 1a. Let T is any normalized time series in the cone C and $\angle(\overrightarrow{T_q}, \overrightarrow{T})$ is denoted for the angle between the query time series vector $\overrightarrow{T_q}$ and the time series vector $\overrightarrow{T}$ in the multi-dimensional sphere. The following properties are satisfied:



Correlation Queries in Spatial Time Series Data, Fig. 1 (**a**) Upper bound and lower bound, (**b**) properties of spatial cone tree

1. If $\gamma_{\max} \in (0, \arccos(\theta))$, then $\angle(\overrightarrow{T_q}, \overrightarrow{T}) \in (0, \arccos(\theta))$;
2. If $\gamma_{\min} \in (180° - \arccos(\theta), 180°)$, then $\angle(\overrightarrow{T_q}, \overrightarrow{T}) \in (180° - \arccos(\theta), 180°)$;
3. If $\gamma_{\min} \in (\arccos(\theta), 180°)$ and $\gamma_{\max} \in (\gamma_{\min}, 180° - \arccos(\theta))$, then $\angle(\overrightarrow{T_q}, \overrightarrow{T}) \in (\arccos(\theta), 180° - \arccos(\theta))$.

If either of the first two conditions is satisfied, the cone C is called an all-true cone (all-true lemma). If the third condition is satisfied, the cone C is called an all-false cone (all-false lemma). If none of the conditions is satisfied, the cone C is called a some-true cone (some-true lemma). These lemma are developed to eliminate cones with all times series satisfying/dissatisfying the correlation threshold in query processing.

The key idea of query processing is to process a correlation query in a filter-and-refine style on the cone level, instead on the individual time series level. The *filtering* step traverses the spatial cone tree, applying the all-true and all-false lemmas on the cones. Therefore, the cones satisfying all-true or all-false conditions are filtered out. The cones satisfying some-true are traversed recursively until all-true or all-false is satisfied or a leaf cone is reached. The *refinement* step exhaustively checks the some-true leaf cones.

Key Applications

The explosive growth of spatial data and widespread use of spatial databases emphasize the need for the automated discovery of spatial knowledge. The complexity of spatial data and intrinsic spatial relationships limits the usefulness of conventional data mining techniques for extracting spatial patterns. Efficient tools for extracting information from geo-spatial data are crucial to organizations which make decisions based on large spatial datasets, including the National Aeronautics and Space Administration (NASA), the National Geospatial-Intelligence Agency (NGA), the National Cancer Institute (NCI), and the United States Department of Transportation (USDOT). These organizations are spread across many application domains including Earth science, ecology and environmental management, public safety, transportation, epidemiology, and climatology. The application of correlation queries used in Earth science is introduced in details as follows.

NASA Earth observation systems currently generate a large sequence of global snapshots of the Earth, including various atmospheric, land, and ocean measurements such as sea surface temperature (SST), pressure, and precipitation. These data are spatial time series data in nature. The climate of the Earth's land surface is strongly influenced by the behavior of the oceans. Simultaneous variations in climate and related processes over widely separated points on the Earth are called teleconnections. For instance, every three to seven years, an El Nino event, i.e., the anomalous warming of the eastern tropical region of the Pacific Ocean, may last for months, having significant economic and atmospheric consequences worldwide. El Nino has been linked to climate phenomena such as droughts in Australia and heavy rainfall along the eastern coast of South America. To investigate such land-sea teleconnections, time series correlation queries across the land and ocean is often used to reveal the relationship of measurements of observations.

Future Directions

In this chapter, the spatial cone tree on spatial time series data was discussed, and how correlation queries can be efficiently supported using the spatial cone tree was illustrated. In future work, more design issues on the spatial cone tree should be further investigated, e.g., the blocking factor and balancing of the tree. The spatial cone tree should be investigated to support complex correlation relationships, such as time lagged correlation. The generalization of spatial cone trees to non-spatial index structures using spherical k-means to construct cone trees is also an interesting research topic.

Recommended Reading

Agrawal R, Faloutsos C, Swami A (1993) Efficient similarity search in sequence databases. In: Lecture notes in computer science, vol 730. Berlin/Heidelberg

Box G, Jenkins G, Reinsel G (1994) Time series analysis: forecasting and control. Prentice Hall, Upper Saddle River

Dhillon I, Fan J, Guan Y (2001) Efficient clustering of very large document collections. In: Grossman R, Kamath C, Kegelmeyer P, Kumar V, Namburu R (eds) Data mining for scientific and engineering applications. Kluwer Academic, Dordrecht

Chan FK, Fu AW (2003) Haar wavelets for efficient similarity search of time-series: with and without time warping. IEEE Trans Knowl Data Eng 15(3):678–705

Guttman A (1984) R-trees: a dynamic index structure for spatial searching. ACM, pp 47–57

Kahveci T, Singh A, Gurel A (2002) Similarity searching for multi-attribute sequences. IEEE, p 175

Keogh E, Pazzani M (1999) An indexing scheme for fast similarity search in large time series databases. IEEE, pp 56–67

National Oceanic and Atmospheric Administration. El Nino Web Page. www.elnino.noaa.gov/

Rafiei D, Mendelzon A (2000) Querying time series data based on similarity. IEEE Trans Knowl Data Eng 12(5):675–693

Rigaux P, Scholl M, Voisard A (2001) Spatial databases: with application to GIS. Morgan Kaufmann Publishers, Reading

Samet H (1990) The design and analysis of spatial data structures. Addison-Wesley Publishing Company, San Francisco

Shekhar S, Chawla S (2003) Spatial databases: a tour. Prentice Hall, Upper Saddle River. ISBN:0130174807

Zhang P, Huang Y, Shekhar S, Kumar V (2003a) Correlation analysis of spatial time series datasets: a filter-and-refine approach. In: Lecture notes in computer science, vol 2637. Springer, Berlin/Heidelberg

Zhang P, Huang Y, Shekhar S, Kumar V (2003b) Exploiting spatial autocorrelation to efficiently process correlation-based similarity queries. In: Lecture notes in computer science, vol 2750. Springer, Berlin/Heidelberg

COSP

▸ Error Propagation in Spatial Prediction

Counterflow

▸ Contraflow in Transportation Network

Coverage Standards and Services, Geographic

Wenli Yang[1] and Liping Di[2]
[1]Center for Spatial Information Science and Systems, College of Science, George Mason University, Fairfax, VA, USA
[2]Center for Spatial Information Science and Systems (CSISS), George Mason University, Fairfax, VA, USA

Definition

A geographic coverage is a representation of a phenomenon or phenomena within a bounded spatiotemporal region by assigning a value or a set of values to each position within the spatiotemporal domain. Geographic coverage standards specify schema and frameworks for geographic coverage or coverage components. Geographic coverage services are those having standard interfaces defined by widely recognized standardization bodies.

Main Text

Geographic phenomena can be observed in two forms, one is discrete and the other is continuous. Discrete phenomena are usually objects that can be directly recognized due to the existence of their geometrical boundaries with other objects. Continuous phenomena usually do not have observable boundaries and vary continuously over space.

The information on the discrete phenomena and that on continuous phenomenon are often used differently and operations performed on the data recording of these two categories of information are usually also different. Thus, there are often differences in data structure designs, data encoding approaches, data accessing and processing methods for these two types of geographic phenomena. Geographic coverage is a concept for continuous phenomena. Geographic coverage standards defines conceptual schema for coverage and analyze coverage types and components, e.g., ISO: ISO/TC211 (2005). These include characteristics of spatiotemporal domain coverage and attribute range, major coverage types, and operations on coverages. Geographic coverage standards provide a common technology language and guide the development of interoperable services on coverage data. Geographic coverage services perform various functionalities for coverage including collecting, archiving, cataloging, publishing, distributing, and processing of coverage data. Geographic coverage services compliant with standard schema and interfaces are interoperable. They can be described, published and found in standard service catalogues, be accessed by all compliant clients, and be connected in order to construct service chains to accomplish complex geospatial modeling tasks.

Cross-References

▸ Geographic Coverage Standards and Services

References

ISO: ISO/TC211 (2005) ISO 19123 geographic information – schema for coverage geometry and functions

CPU-GPU

▸ Medical Image Dataset Processing over Cloud/MapReduce with Heterogeneous Architectures

Crime Mapping

▸ CrimeStat: A Spatial Statistical Program for the Analysis of Crime Incidents
▸ Hotspot Detection, Prioritization, and Security

Crime Mapping and Analysis

Ronald E. Wilson and Katie M. Filbert
Mapping and Analysis for Public Safety Program & Data Resources, National Institute of Justice, Washington, D.C, MD, USA

Synonyms

Environmental criminology; First law of geography; Geographical analysis; Rational choice; Route activity; Social disorganization; Spatial analysis of crime; Spatial aspects of crime; Statistical techniques

Definition

The term "crime mapping" is inaccurate as it is overly simplistic. Crime mapping is often associated with the simple display and querying of crime data using a Geographic Information System (GIS). Instead, it is a general term that encompasses the technical aspects of visualization and statistical techniques, as well as practical aspects of geographic principles and criminological theories.

From a technical standpoint, the term is a combination of visualization and statistical techniques manifested as software. This combination of techniques is shared between mapping, spatial analysis and spatial data analysis. Mapping is simply a visualization tool that is used to display raw geographic data and output from analysis, which is done through a GIS. Spatial analysis is the statistical testing of geographic features in relation to other geographic features for patterns, or lack there of. Spatial data analysis is the combination of spatial analysis with associated attribute data of the features to uncover spatial interactions between features.

From a practical standpoint, crime mapping is a hybrid of several social sciences, which are geography, sociology and criminology. It combines the basic principles of geographic analysis, sociological and criminological theory and makes it possible to test conjectures from these disciplines in order to confirm or refute them as actual. In essence it has developed into an applied science, with its own tools, that examines a range of issues about society and its relationship with the elements that contribute to crime. Thus, crime mapping is interdisciplinary, involving other disciplines that incorporate the spatial perspectives of social phenomena related to crime, such as inequality, residential stability, unemployment, resource depravation, economic opportunities, housing availability, migration, segregation, and the effects of policy. Using a geographic framework often leads to a more comprehensive understanding of the factors that contribute to or suppress crime.

Even though the term "crime mapping" is a misnomer it will continue to be widely used as a general term with regard to the study of the spatial aspects of crime. Users of the term need to let the context of their work dictate which standpoint is being referred to.

Historical Background

Starting in the 1930s, crime mapping was used with limited success in the United States due to lack of data and the means to analyze that data, computational capacity. Thus, its value was simple depictions on paper of where crimes were occurring. For social science these depictions were not important until researchers from the Chicago School of Sociology combined criminological theory with geographic theory on a map. The result was the theory of social disorganization. Using a map, Shaw and McKay (1942) overlaid residences of juvenile offenders with the (Park et al. 1925) concentric zone model of urban land uses, including demographic characteristics. They discovered a geographic correlation between impoverished and blighted places with those that had most of the juvenile offender residences. This fostered a new line of research that examined the spatial aspects of crime that spanned from 1950 to the late 1970s. Despite the

impact this had on furthering spatial theories of crime there was not much more that could be done because the basic principles of geographic analysis had not yet been operationalized into what geographer Jerome (Dobson 1983) would call, "Automated Geography." Personal computers soon came afterwards, but software permitting empirical testing of these theories did not come until much later. It was at this point that crime mapping became useful to law enforcement, primarily to depict where crimes were occurring in order to focus resources (Weisburd and McEwen 1997). However, there was not yet a relationship between academic institutions and law enforcement agencies to couple theories with actual observations from the street.

Crime mapping with computers made an entrance in the mid 1960s allowing the production of maps of crime by city blocks shaded by volume of incidents. This was still of little interest to researchers studying crime. Even though criminologists were becoming interested in the spatial analysis of crime they were not looking to other disciplines, including geography, for help in analyzing data using a spatial framework. A manifold of software programs from geography were available that could have been used, but there is little evidence in any of the social science literature that demonstrate that these programs were being used. Also neglected were principles of geographic analysis, to analyze the spatial aspects of data. With practitioners, their struggle was different. To produce maps of crime required serious computing infrastructure that, at the time, was only available within larger city government agencies, which did not hold making crime maps in high priority (Weisburd and McEwen 1997).

The growth of environmental criminology in the 1980s, spearheaded by Paul and Patricia Brantingham, allowed the discipline of geography to make inroads into criminological theory (La Vigne and Groff 2001). Environmental criminology fused geographical principles and criminological theory together with GIS and provided opportunities to empirically test the theories it was purporting. Significant contributions by George Rengert (1989) (Rengert and Simon 1981), Jim LeBeau (1987, 1992) and Keith Harries (1974, 1980), to environmental criminology using GIS and spatial statistics software continued, thereafter, to strengthen the role of geography in the study of crime. As a result, criminology now has several geographic theories of crime, including rational choice (Cornish and Clarke 1986), routine activity (Cohen and Felson 1979), and crime pattern theory (Brantingham and Brantingham 1981). Social disorganization theory was also extended with geographical principles through the incorporation of simultaneous social interactions between adjacent neighborhoods. For a brief and succinct listing of these theories see Paulsen and Robinson (2004). At this point, crime mapping branched out to become useful in a new practitioner-based area beyond law enforcement, the criminal justice agency. The confluence of geographic principles, criminological theory and advancing technology led to the development of crime prevention programs based on empirical evidence, such as 'Hot Spot' Policing.

In the late 1980s the Federal government played a role in advancing the use of the crime mapping. The National Institute of Justice (NIJ) funded several efforts under the Drug Market Analysis Program (DMAP) that brought together academic institutions with law enforcement agencies in five cities in the United States (La Vigne and Groff 2001). The purpose was to identify drug markets and activities associated with them by tracking movement of dealers and users in and out of them. These grants were the first to promote working relationships between practitioners and researchers in the area of crime mapping to move them beyond the limitations each was facing not having the other as a partner.

Continuing improvements in GIS throughout the 1990s, and into the 2000s, made it possible to better assemble, integrate, and create new data. This is probably the greatest impact that GIS has had on crime mapping. Not only could a GIS assemble multiple and disparate sets of demographic, economic and social data with crime data, it could also create new units of analysis that better modeled human behavior. This capability

afforded a more accurate understanding of the spatial interactions among offenders, victims and their environments that could be captured and analyzed in ways that more accurately represented human settlement and activity. This freed criminologists from being confined to the standard units of analysis, such as administrative boundaries from the US Census Bureau or other local governmental agencies. GIS provided the unique opportunity to represent boundaries of human activity more accurately through the creation of more distinct partitions, such as police beats or land use, as well as asymmetrical boundaries of human interaction created with buffers or density surfaces. In this regard, there is nothing else like GIS in the study of crime. For practitioners this freed them from having to depend on other government agencies to produce crime maps for law enforcement purposes, as well as provide opportunities to produce custom "on demand" maps for specific purposes, including search warrants or patrol deployment.

The late 1990s saw the advancement of crime mapping in not only both academic departments that study crime and law enforcement agencies, but also in the Federal government. NIJ established the Crime Mapping Research Center in 1997, now the Mapping and Analysis for Public Safety (MAPS) Program, for the purpose of conducting research and evaluation of the spatial aspects of crime. One year later NIJ provided money to the National Law Enforcement and Corrections Technology Center (NLECTC) Rocky Mountain Division to establish the Crime Mapping and Analysis Program (CMAP). This program was to provide assistance to law enforcement and criminal justice agencies specifically in the use of crime mapping. Into the 2000s all large agencies and most medium-sized agencies are using GIS as part of their analysis and operations efforts and are using crime mapping far beyond just the simple mapping of where crime is occurring. Research has continued to refine spatial theories of crime based on better coordination with practitioners, funding from Federal agencies and the development of software for the further understanding of crime through geography.

Scientific Fundamentals

Crime mapping, as an applied science, is ultimately about *where*. As a result, there are contributions from primarily two social science disciplines that make up the foundations of crime mapping. The first provides a set of principles that sets the stage for the study of crime within a spatial framework, geography. The second provides a set of specific spatial theories about criminal activity and environmental conditions that form the foundation of the spatial aspects of crime, criminology.

Geographic Principles

A complete understanding of crime is facilitated by two sets of factors: individual and contextual. Crime mapping deals with the contextual. Therefore, geographic principles are necessary to understand that context. These principles provide a framework for measuring the interactions between places. Analysis in that framework is possible combining long standing geographic principles that have been implemented through GIS and spatial data analysis software. GIS facilitates the visualization of raw data and the results from statistical analysis. Spatial statistical techniques extend traditional statistics to form a more complete approach toward understanding social problems, including crime. The following are the three basic geographic principles that are the foundation for the contextual analysis of crime.

Place

Criminology has a long history of looking at the geographical influences on crime. Some of the most significant pieces of work were in regards to the study of crime in neighborhoods, communities, cites, regions and even across the United States (Brantingham and Brantingham 1981; Reiss et al. 1986; Bursik and Grasmick 1993; Weisburd et al. 1995). These studies identify "places" in which criminology seeks to understand criminal activity. The focus of

studying crime in place demonstrates the use of geography as a framework for contextual analysis that no other discipline can offer. Place becomes the cornerstone because it allows for the categorizing of space by defining a geographic unit of analysis for the systematic measurement of human and environmental characteristics in relation to neighboring places.

Tobler's First Law of Geography

Places are not isolated islands of activity. Interactions, such as social, demographic, or economic occur within and between places. These interactions form spatial relationships based on the concept that those things closer together in space are more related. That is, changes in human activity and physical environments change slowly across space, with abrupt changes being out of the ordinary. Named after Waldo (Tobler 1970), this law forms the theoretical foundation for the concept of distance decay that is used for analysis of these spatial interactions and relationships which then allows for measurement in the strength of interactions between places.

Spatial Processes

Human interactions that occur within, and between, geographic places form two concepts: spatial heterogeneity and spatial dependence. Spatial heterogeneity is the variability of human and environmental conditions across space. At the local level this is change across a defined space where conditions, such as racial composition, economic stability, housing conditions, land use, or migration vary. These things are not evenly distributed across space and form various patterns, at different scales, and in multiple directions, all of which are asymmetric. Spatial dependence represents the strength of a relationship of some phenomenon between places that have influence on each other, a concept known as spatial autocorrelation. These patterns range from clusters to randomly distribution to dispersed to uniform. These are indications that human activity and the environments which they develop have a wide range of variability, one that usually follows systemic patterns.

Criminological Theories

Criminology has developed a set of spatial theories of crime that have utilized all three of the geographic principles listed.

Rational Choice

Rational choice theory is based on classical ideas that originated in the 1700s, with the work of Cesare Beccaria and others who took a utilitarian view of crime (Beccaria 1764). This perspective suggests that criminals think rationally and make calculated decisions, weighing costs and risks of committing a crime against potential benefits while being constrained by time, cognitive ability and information available resulting in a 'limited' rather than 'normal' rationality (Cornish and Clarke 1986). In this sense, rational choice theory also brings in economic ideas and theories into criminology.

Routine Activities

Routine activities theory helps explain why crime occurs at particular places and times. The theory suggests that crime opportunities are a function of three factors that converge in time and place, including a motivated offender, suitable target or victim, and lack of a capable guardian (Cohen and Felson 1979). A fourth aspect of routine activities theory, suggested by John Eck, is place management. Rental property managers are one example of place managers (Eck and Wartell 1997). They have the ability to take nuisance abatement and other measures to influence behavior at particular places. Criminals choose or find their targets within context of their routine activities, such as traveling to and from work, or other activities such as shopping, and tend not to go that far out of their way to commit crimes (Felson 1994).

Crime Pattern

Crime pattern theory looks at the opportunities for crime within context of geographic space, and makes a distinction between crime events and criminality, that is, the propensity to commit crime (Brantingham and Brantingham 1981). Crime pattern theory integrates rational choice

and routine activities theories, with a geographic framework, place. The theory works at various geographic scales, from the macro-level with spatial aggregation at the census tract or other level, to the micro-scale with focus on specific crime events and places. Crime pattern theory focuses on situations or places where there is lack of social control or guardianship over either the suspect or victim, combined with a concentration of targets. For example, a suburban neighborhood can become a hot spot for burglaries because some homes have inadequate protection and nobody home to guard the property.

Social Disorganization

Social disorganization theory emphasizes the importance of social controls in neighborhoods on controlling behavior, particularly for individuals with low self-control or a propensity to commit crime. Social controls can include family, as well as neighborhood institutions such as schools and religious places. When identifying places with social disorganization, the focus is on ability of local residents to control social deviancy (Bursik and Grasmick 1993). Important factors include poverty, as well as turnover of residents and outmigration, which hinder the development of social networks and neighborhood institutions that lead to collective efficacy (Sampson et al. 1997).

Key Applications

There are five key applications in crime mapping. These applications are thematic mapping, non-graphical indicators, hot spots, spatial regression and geographic profiling. They make up a full compliment of techniques from elementary to advanced.

Thematic Mapping

Thematic maps are color coded maps that depict the geographic distribution of numeric or descriptive values of some variable. They reveal the geographic patterns of the underlying data. A variable can be quantitative or qualitative. Quantitative maps provide multiple techniques for categorizing the distribution of a variable. Qualitative maps provide a mechanism for classification of some description, or label, of a value. They are often shaded administrative or statistical boundaries, such as census blocks, police beats or neighborhoods. For example, robbery rates based on population can be derived for neighborhood boundaries giving an indication of the neighborhoods that pose the highest risk. However, locations can be symbolized to show quantities based on size or color of the symbol. For example, multiple crime events at a particular location give an indication of repeat victimization, such as common in burglary. However, simple visualization of values and rates can be misleading, especially since the method of classification can change the meaning of a map. Spatial statistics are then used to provide more rigorous and objective analysis of spatial patterns in the data.

Non-graphical Indicators

Non-graphical statistical tests produce a single number that represents the presence of the clustering of crime incidents or not. These are global level statistics indicating the strength of spatial autocorrelation, but not its location. They compare actual distributions of crime incidents with random distributions. Positive spatial autocorrelation indicates that incidents are clustered, while negative indicates that incidents are uniform. Tests for global spatial autocorrelation within a set of points include Moran's I, (Chakravorty 1995), Geary's C statistic, and Nearest Neighbor Index (Levine 2005). After visualizing data in thematic maps these are the first statistical tests conducive to determining whether there are any local level relationships between crime and place exist.

Hot Spots

Hot spots are places with concentrations of high crime or a greater than average level of crime. The converse of a hot spot is a *cold spot*, which are places that are completely, or almost, devoid of crime. Identification and analysis of hot spots is often done by police agencies, to provide guidance as to where to place resources and target crime reduction efforts. Hot spot analysis

can work at different geographic levels, from the macro-scale, looking at high crime neighborhoods, or at the micro-scale to find specific places such as particular bars or street segments that are experiencing high levels of crime (Eck et al. 2005). Depending on the level of analysis, police can respond with specific actions such as issuing a warrant or focusing at a neighborhood level to address neighborhood characteristics that make the place more criminogenic. A variety of spatial statistical techniques are used for creating hot spots, such as density surfaces (Levine 2005), location quotients (Isserman 1977; Brantingham and Brantingham 1995; Ratcliffe 2004), local indicators of spatial autocorrelation (LISA) (Anselin 1995; Getis and Ord 1996; Ratcliffe and McCullagh 1998), and nearest neighborhood hierarchical clustering (Levine 2005).

Spatial Regression

Regression techniques, such as Ordinary Least Squares (OLS), have been used for quite some time in criminology as explanatory models. This technique has a major limitation, in that it does not account for spatial dependence inherent in almost all data. Holding to geographic principles, a place with high crime is most likely surrounded by neighbors that also experience high crime, thereby displaying spatial autocorrelation, i.e. a spatial effect. Spatial regression techniques, developed by Luc (Anselin 2002), take into account spatial dependence in data. Not factoring these spatial effects into models makes them biased and less efficient. Tests have been created for identifying spatial effects in the dependent variable (spatial lag) and among the independent variables (spatial error). If tests detect the presence of spatial lag or error, this form of regression adjusts the model so that spatial effects do not unduly affect the explanatory power of the model.

Geographic Profiling

Geographic profiling is a technique for identifying the likely area where a serial offender resides or other place such as their place of work, that serves as an anchor point. Geographic profiling techniques draw upon crime place theory and routine activities theory, with the assumption that criminals do not go far out of their daily routines to commit crimes. Geographic profiling takes into account a series of crime locations that have been linked to a particular serial criminal and creates a probability surface that identifies the area where the offender's anchor point may be (Rossmo 2000; Canter 2003). Geographic profiling was originally developed for use in serial murder, rapes, and other rare but serious crimes. However, geographic profiling is being expanded to high-volume crimes such as serial burglary (Chainey and Ratcliffe 2005).

Future Directions

The advancement of research and practice in crime mapping rests on continuing efforts in three areas: efforts by research and technology centers, software development, and expansion into law enforcement and criminal justice.

Crime mapping research and technology centers, such as the MAPS Program, the CMAP and the Crime Mapping Center at the Jill Dando Institute (JDI), are primary resources for research, development and application of GIS, spatial data analysis methodologies and geographic technologies. These three centers serves as conduits for much of the work conducted in both the academic and practitioner communities. The MAPS Program is a grant funding and applied research center that serves as a resource in the use of GIS and spatial statistics used in crime studies. The program awards numerous grants for research and development in the technical, applied and theoretical aspects of using GIS and spatial data analysis to study crime, as well as conduct research themselves. As a counterpart to the MAPS Program, CMAP's mission is to serve practitioners in law enforcement and criminal justice agencies by developing tools and training materials for the next generation and crime analysts and applied researchers in the use of GIS and spatial analysis. In the UK the Jill Dando Institute of Crime Science has a Crime Mapping Center that contributes to the advancement in understanding the spatial aspects of crime with an approach

called "crime science." This approach utilizes theories and principles from many scientific disciplines to examine every place as a unique environment for an explanation of the presence or absence of crime. They conduct applied research and provide training with their unique approach on a regular basis. The MAPS Program and the Crime Mapping Center at the JDI hold conferences on a regular basis. These events form the nexus for practitioners and researchers to work together in the exchange of ideas, data, experiences and results from analysis that create a more robust applied science.

Software programs are vital to the progression of the spatial analysis of crime. These programs become the scientific instruments that researchers and practitioners need in understanding human behavior and environmental conditions as they relate to crime. Software, such as CrimeStat, GeoDa and spatial routines for 'R' are being written to include greater visualization capabilities, more sophisticated modeling and mechanisms for seamless operation with other software. For example, in version three of CrimeStat the theory of travel demand was operationalized as a set of routines that apply to criminals as mobile agents in everyday life. GeoDa continues to generate robust tools for visualization based on the principles of Exploratory Data Analysis (EDA). New and cutting edge tools for geographic visualization, spatial statistics and spatial data analysis are being added to the open statistical development environment 'R' on a regular basis. All of these programs provide a rich set of tools for testing theories and discovering new patterns that reciprocally help refine what is known about patterns of crime. The emergence of spatial statistics has proven important enough that even the major statistical software packages, such as SAS, SPSS, and Stata are all incorporating full sets of spatial statistics routines.

The application of crime mapping is expanding into broader areas of law enforcement and criminal justice. In law enforcement mapping is taking agencies in new directions toward crime prevention. For example, the Computer Mapping, Planning and Analysis of Safety Strategies (COMPASS) Program, funded by the NIJ, combines crime data with community data where crime is a characteristic of populations rather than a product. That is to say crime is, at times, a cause of conditions rather than the result of conditions. It is an indicator of the "well being" of neighborhoods, communities or cities. Shared with local level policy makers, COMPASS provides a view into this "well being" of their communities. Resources can be directed to those places that are not "well" and helps to understand what makes other places "well." Combined with problem-oriented policing, a strategy that addresses specific crime problems, this approach can be effective in reducing crime incidents and a general reduction in social disorder (Braga et al. 1999). Coupled with applications in criminal justice, mapping can be utilized to understand the results of policy and the outcomes. This includes topics important to community corrections in monitoring or helping returning offenders, including registered sex offenders. Or, mapping can be of use in allocating probation and parole officers to particular geographic areas, directing probationers and parolees to community services, and selecting sites for new community services and facilities (Karuppannan 2005). Finally, mapping can even help to understand the geographic patterns of responses to jury summons to determine if there are racial biases are occurring in some systematic way across a jurisdiction (Ratcliffe 2004).

These three elements will persist and intertwine to evermore incorporate the geographic aspects of basic and applied research of crime through technology. The advancement of knowledge that crime mapping can provide will require continued reciprocation of results between research and practice through technology (Stokes 1997). The hope is that researchers will continue to create new techniques and methods that fuse classical and spatial statistics together to further operationalize geographic principles and criminological theory to aid in the understanding of crime. Practitioners will implement new tools that are developed for analyzing crime with geographic perspectives. They will also continue to take these tools in new directions as "improvers of technology" (Stokes 1997) and discover new

patterns as those tools become more complete in modeling places.

Acknowledgements We would like to thank Keith Harries, Dan Helms, Chris Maxwell and Susan Wernicke-Smith for providing comments on this entry in a very short time. They provided valuable comments that were used toward crafting this entry.

The views expressed in this paper are those of the authors, and do not represent the official positions or policies of the National Institute of Justice or the US Department of Justice.

Cross-References

- Autocorrelation, Spatial
- Constraint Data, Visualizing
- CrimeStat: A Spatial Statistical Program for the Analysis of Crime Incidents
- Data Analysis, Spatial
- Exploratory Visualization
- Hotspot Detection, Prioritization, and Security
- Patterns, Complex
- Spatial Econometric Models, Prediction
- Spatial Regression Models
- Statistical Descriptions of Spatial Patterns
- Time Geography

References

Anselin L (1995) Local indicators of spatial association – LISA. Geogr Anal 27:93–115

Anselin L (2002) Under the hood: issues in the specification and interpretation of spatial regression models. http://sal.uiuc.edu/users/anselin/papers.html

Braga AA, Weisburd DL, et al (1999) Problemoriented policing in violent crime places: a randomized controlled experiment. Criminology 7:541–580

Brantingham P, Brantingham P (1981) Environmental criminology. Waverland Press, Prospect Heights

Brantingham P, Brantingham P (1995) Location quotients and crime hotspots in the city. In: Block C, Dabdoub M, Fregly S (eds) Crime analysis through computer mapping. Police Executive Research Forum, Washington, DC

Beccaria C (1764) Richard Davies, translator: on crimes and punishments, and other writings. Cambridge University Press

Bursik RJ, Grasmick HG (1993) Neighborhoods and crime: the dimensions of effective community control. Lexington Books, New York

Canter D (2003) Mapping murder: the secrets of geographic profiling. Virgin Publishing, London

Chainey S, Ratcliffe J (2005) GIS and crime mapping. Wiley, Hoboken

Chakravorty S (1995) Identifying crime clusters: the spatial principles. Middle States Geogr 28:53–58

Cohen L, Felson M (1979) Social change and crime rate trends. Am Soc Rev 44(4):588–608

Cornish D, Clarke RV (1986) The reasoning criminal. Springer

Dobson JE (1983) Automated geography. Prof Geogr 35(2):135–143

Eck J, Chainey S, Cameron J, Leitner M, Wilson RE (2005) Mapping crime: understanding hot spots. National Institute of Justice, Washington, DC

Eck J, Wartell J (1997) Reducing crime and drug dealing by improving place management: a randomized experiment. National Institute of Justice

Felson M (1994) Crime and everyday life. Pine Forge

Getis A, Ord JK (1996) Local spatial statistics: an overview. In: Longley P, Batty M (eds) Spatial analysis: modelling in a GIS environment. Geoinformation International, Cambridge, pp 261–277

Harries KD (1974) The geography of crime and justice. McGraw-Hill, New York

Harries KD (1980) Crime and the environment. Charles C Thomas Press, Spingfield

Isserman AM (1977) The location quotient approach for estimating regional economic impacts. J Am Inst Plan 43:33–41

Karuppannan J (2005) Mapping and corrections: management of offenders with geographic information systems. Corrections Compendium. http://www.iaca.net/Articles/drjaishankarmaparticle.pdf

La Vigne NG, Groff ER (2001) The evolution of crime mapping in the United States: from the descriptive to the analytic. In: Hirschfield A, Bowers K (eds) Mapping and analyzing crime data. University of Liverpool Press, Liverpool, pp 203–221

LeBeau JL (1987) Patterns of stranger and serial rape offending: factors distinguishing apprehended and at large offenders. J Crim Law Criminol 78(2):309–326

LeBeau JL (1992) Four case studies illustrating the spatialtemporal analysis of serial rapists. Police Stud 15:124–145

Levine N (2005) CrimeStat III version 3.0, a spatial statistics program for the analysis of crime incident locations

Park RE, Burgess EW, McKenzie RD (1925) The city: suggestions for investigation of human behavior in the urban environment. University of Chicago Press, Chicago

Paulsen DJ, Robinson MB (2004) Spatial aspects of crime: theory and practice. Allyn and Bacon, Boston

Ratcliffe JH, McCullagh MJ (1998) The perception of crime hotspots: a spatial study in Nottingham, UK. In: Crime mapping case studies: successes in the field. National Institute of Justice, Washington, DC

Ratcliffe JH (2004) Location quotients and force-field analysis. In: 7th annual international crime mapping research conference, Boston
Reiss AJ, Tonry M (eds) (1986) Communities and crime, vol 8. University of Chicago Press, Chicago
Rengert GF, Simon H (1981) Crime spillover. Sage Publications, Beverley Hills
Rengert GF (1989) Behavioral geography and criminal behavior. In: Evans DJ, Herbert DT (eds) The geography of crime. Routledge, London
Rossmo DK (2000) Geographic profiling. CRC Press, Boca Raton
Sampson RJ, Raudenbush SR, Earls F (1997) Neighborhoods and violent crime: a multilevel study of collective efficacy. Science 227:918–924
Shaw CR, McKay HD (1942) Juvenile delinquency and urban areas. University of Chicago Press, Chicago
Stokes DE (1997) Pasteur's quadrant: basic science and technological innovation. Brookings Institution Press, Washington, DC
Tobler WR (1970) A computer movie simulating urban growth in the Detroit region. Econ Geogr 46: 234–240
Weisburd D, Eck JE (eds) (1995) Crime and place: crime prevention studies, vol 4. Police Executive Research Forum/Willow Tree Press, Washington, DC
Weisburd D, McEwen T (eds) (1997) Introduction: crime mapping and crime prevention. In: Crime mapping and crime prevention. Criminal Justice Press, Monsey, pp 1–23

Recommended Reading

Clarke RV (1992) Situational crime prevention: successful case studies. Harrow and Heston, New York
Cresswell T (2004) Place: a short introduction. Blackwell Publishing Ltd, Malden
Haining R (2003) Spatial data analysis: theory and practice. Cambridge University Press, Cambridge/New York
Ray JC (1977) Crime prevention through environmental design. Sage Publications, Beverly Hills
Ronald CV (1992) Situational crime prevention. Harrow and Heston, New York
Weisburd D, Green L (1995) Policing drug hot spots: the Jersey city drug market analysis experiment. Justice Q 12(4):711–736

Crime Travel Demand

▸ CrimeStat: A Spatial Statistical Program for the Analysis of Crime Incidents

CrimeStat

▸ CrimeStat: A Spatial Statistical Program for the Analysis of Crime Incidents

CrimeStat: A Spatial Statistical Program for the Analysis of Crime Incidents

Ned Levine
Ned Levine & Associates, Houston, TX, USA

Synonyms

Centrographic measures; Correlated walk; Crime mapping; CrimeStat; Crime travel demand; Geographic profiling; Hotspot; Interpolation; Journey to crime analysis; Knox test; Mantel test; Space–time interaction; Spatial statistics program

Definition

CrimeStat is a spatial statistics and visualization program that interfaces with desktop GIS packages. It is a stand-alone Windows program for the analysis of crime incident locations and can interface with most desktop GIS programs. Its aim is to provide statistical tools to help law enforcement agencies and criminal justice researchers in their crime mapping efforts. The program has many statistical tools, including centrographic, distance analysis, hot spot analysis, space-time analysis, interpolation, Journey-to-Crime estimation, and crime travel demand modeling routines. The program writes calculated objects to GIS files that can be imported into a GIS program, including shape, MIF/MID, BNA, and ASCII. The National Institute of Justice is the distributor of CrimeStat and makes it available for free to analysts, researchers, educators, and students (The program is available at http://www.icpsr.umich.edu/crimestat). The program is distributed along

with a manual that describes each of the statistics and gives examples of their use (Levine 2007a).

Historical Background

CrimeStat has been developed by Ned Levine and Associates since the late 1990s under grants from the National Institute of Justice. It is an outgrowth of the Hawaii Pointstat program that was UNIX-based (Levine 1996). CrimeStat, on the other hand, is a Windows-based program. It is written in C++ and is multi-threading. To date, there have been three major versions with two updates. The first was in 1999 (version 1.0) with an update in 2000 (version 1.1). The second was in 2002 (CrimeStat II) and the third was in 2004 (CrimeStat III). The current version is 3.1 and was released in March 2007.

Scientific Fundamentals

The current version of CrimeStat covers seven main areas of spatial analysis: centrographic; spatial autocorrelation, hot spot analysis, interpolation, space-time analysis, Journey-to-Crime modeling, and crime travel demand modeling.

Centrographic Measures

There are a number of statistics for describing the general properties of a distribution. These include central tendency of the overall spatial pattern, dispersion and directionality. Among the statistics are the mean center, the center of minimum distance, the standard distance deviation, the standard deviational ellipse, the harmonic mean, the geometric mean, and the directional mean (Ebdon 1988).

Spatial Autocorrelation

There are several statistics for describing spatial autocorrelation, including Moran's I, Geary's C, and a Moran Correlogram (Moran 1948; Geary 1954; Ebdon 1988). There are also several statistics that describe spatial autocorrelation through the properties of distances between incidents including the nearest neighbor statistic (Clark and Evans 1954), the linear nearest neighbor statistic, the K-order nearest neighbor distribution (Cressie 1991), and Ripley's K statistic (Ripley 1981). The testing of significance for Ripley's K is done through a Monte Carlo simulation that estimates approximate confidence intervals.

Hot Spot Analysis

An extreme form of spatial autocorrelation is a *hot spot*. While there is no absolute definition of a 'hot spot', police are aware that many crime incidents tend to be concentrated in a limited number of locations. The Mapping and Analysis for Public Safety Program at the National Institute of Justice has sponsored several major studies on crime hot spot analysis (Harries 1999; LaVigne and Wartell 1998; Eck et al. 2005).

CrimeStat includes seven distinct 'hot spot' analysis routines: the mode, the fuzzy mode, nearest neighbor hierarchical clustering (Everitt et al. 2001), risk-adjusted nearest neighbor hierarchical clustering (Levine 2004), the Spatial and Temporal Analysis of Crime routine (STAC) (Block 1994), K-means clustering, and Anselin's Moran statistic (Anselin 1995).

The *mode* counts the number of incidents at each location. The *fuzzy mode* counts the number of incidents at each location within a specified search circle; it is useful for detecting concentrations of incidents within a short distance of each other (e.g., at multiple parking lots around a stadium; at the shared parking lot of multiple apartment buildings).

The *nearest neighbor hierarchical clustering* routine defines a search circle that is tied to the random nearest neighbor distance. First, the algorithm groups incidents that are closer than the search circle and then searches for a concentration of multiple incidents within those selected. The center of each concentration is identified and all incidents within the search circle of the center of each concentration are assigned to the cluster. Thus, incidents can belong to one-and-only-one cluster, but not all incidents belong to a cluster. The process is repeated until the distribution is stable (first-order clusters). The user can specify a minimum size for the cluster to eliminate very small clusters (e.g., 2 or 3 incidents at the

same location). Once clustered, the routine then clusters the first-order clusters to produce second-order clusters. The process is continued until the grouping algorithm fails. The *risk-adjusted nearest neighbor hierarchical clustering* routine follows the same logic but compares the distribution of incidents to a baseline variable. The clustering is done with respect to a baseline variable by calculating a cell-specific grouping distance that would be expected on the basis of the baseline variable, rather than a single grouping distance for all parts of the study area.

The *Spatial and Temporal Analysis of Crime* hot spot routine (STAC) is linked to a grid and groups on the basis of a minimum size. It is useful for identifying medium-sized clusters. The *K-means* clustering algorithm divides the points into K distinct groupings where K is defined by the user. Since the routine will frequently create clusters of vastly unequal size due to the concentration of incidents in the central part of most metropolitan areas, the user can adjust them through a separation factor. Also, the user can define specific starting points (seeds) for the clusters as opposed to allowing the routine to find its own.

Statistical significance of these latter routines is tested with a Monte Carlo simulation. The nearest neighbor hierarchical clustering, the risk-adjusted nearest neighbor hierarchical clustering, and the STAC routines each have a Monte Carlo simulation that allows the estimation of approximate confidence intervals or test thresholds for these statistics.

Finally, unlike the other hot spot routines, *Anselin's Local Moran* statistic is applied to aggregates of incidents in zones. It calculates the similarity and dissimilarity of zones relative to nearby zones by applying the Moran's I statistic to each zone. An approximate significance test can be calculated using an estimated variance.

Interpolation

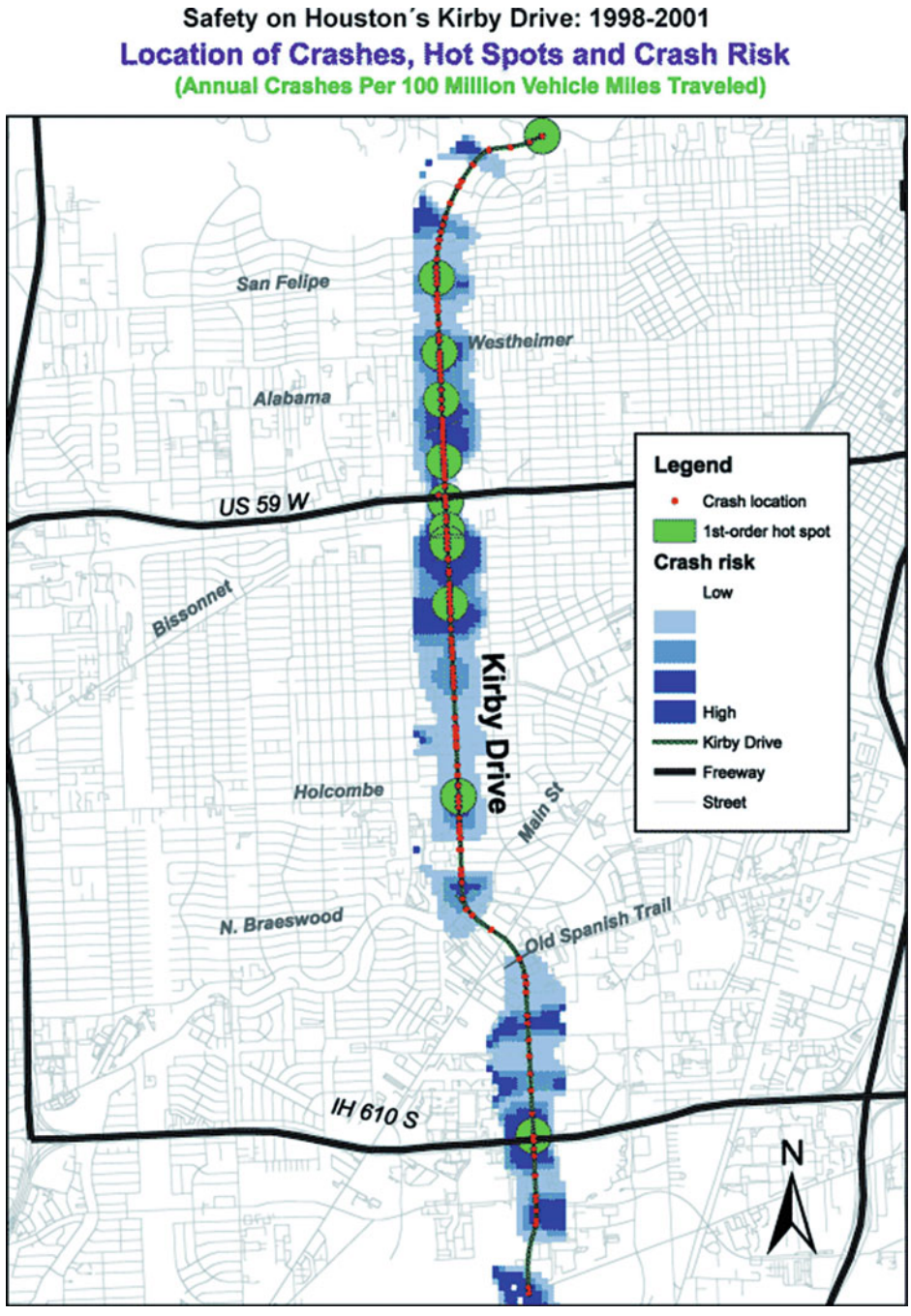
Interpolation involves extrapolating a density estimate from individual data points. A fine-mesh grid is placed over the study area. For each grid cell, the distance from the center of the cell to each data point is calculated and is converted into a density using a mathematical function (a kernel). The densities are summed over all incidents to produce an estimate for the cell. This process is then repeated for each grid cell (Bailey and Gatrell 1995). *CrimeStat* allows five different mathematical functions to be used to estimate the density. The particular dispersion of the function is controlled through a bandwidth parameter and the user can select a fixed or an adaptive bandwidth. It is a type of hot spot analysis in that it can illustrate where there are concentrations of incidents. However it lacks the precision of the hot spot routines since it is smoothed. The hot spot routines will show exactly which points are included in a cluster.

CrimeStat has two different kernel function, a single-variable kernel density estimation routine for producing a surface or contour estimate of the density of incidents (e.g., the density of burglaries) and a dual-variable kernel density estimation routine for comparing the density of incidents to the density of an underlying baseline (e.g., the density of burglaries relative to the density of households).

As an example, Fig. 1 shows motor vehicle crash risk along Kirby Drive in Houston for 1999–2001. *Crash risk* is defined as the annual number of motor vehicle crashes per 100 million vehicle miles traveled (VMT) and is a standard measure of motor vehicle safety. The duel-variable kernel density routine was used to estimate the densities with the number of crashes being the incident variable and VMT being the baseline variable. In the map, higher crash risk is shown as darker. As a comparison, hot spots with 15 or more incidents were identified with the nearest neighbor hierarchical clustering routine and are overlaid on the map as are the crash locations.

Space-Time Analysis

There are several routines for analyzing clustering in time and in space. Two are global measures – the *Knox* and *Mantel* indices, which specify whether there is a relationship between time and space. Each has a Monte Carlo simulation to estimate confidence intervals around the calculated statistic.



CrimeStat: A Spatial Statistical Program for the Analysis of Crime Incidents, Fig. 1 Safety on Houston's Kirby Drive: 1998–2001

The third space-time routine is a specific tool for predicting the behavior of a serial offender called the *Correlated Walk Analysis* module. This module analyzes periodicity in the sequence of events committed by the serial offender by distance, direction, and time interval. It does this by analyzing the sequence of lagged incidents. A diagnostic correlogram allows the user to analyze

periodicity by different lags. The user can then specify one of several methods for predicting the next incident that the serial offender will commit, by location and by time interval. Error is, of course, quite sizeable with this methodology because serial offenders don't follow strict mathematical rules. But the method can be useful for police because it can indicate whether there are any repeating patterns that the offender is following.

Journey-to-Crime Analysis

A useful tool for police departments seeking to apprehend a serial offender is *Journey-to-crime analysis* (sometimes known as *Geographic Profiling*). This is a method for estimating the likely residence location of a serial offender given the distribution of incidents and a model for travel distance (Brantingham and Brantingham 1981; Canter and Gregory 1994; Rossmo 1995; Levine 2007b). The method depends on building a typical travel distance function, either based on empirical distances traveled by known offenders or on an a priori mathematical function that approximates travel behavior (e.g., a negative exponential function, a negative exponential function with a low use 'buffer zone' around the offender's residence).

CrimeStat has a Journey-to-Crime routine that uses the travel distance function and a Bayesian Journey-to-Crime routine that utilizes additional information about the likely origins of offenders who committed crimes in the same locations. With both types – the traditional distance-based and the Bayesian, there are both calibration and estimation routines. In the calibration routine for the Journey-to-Crime routine, the user can create an empirical travel distance function based on the records of known offenders where both the crime location and the residence location were known (typically from arrest records). This function can then be applied in estimating the likely location of a single serial offender for whom his or her residence location is not known.

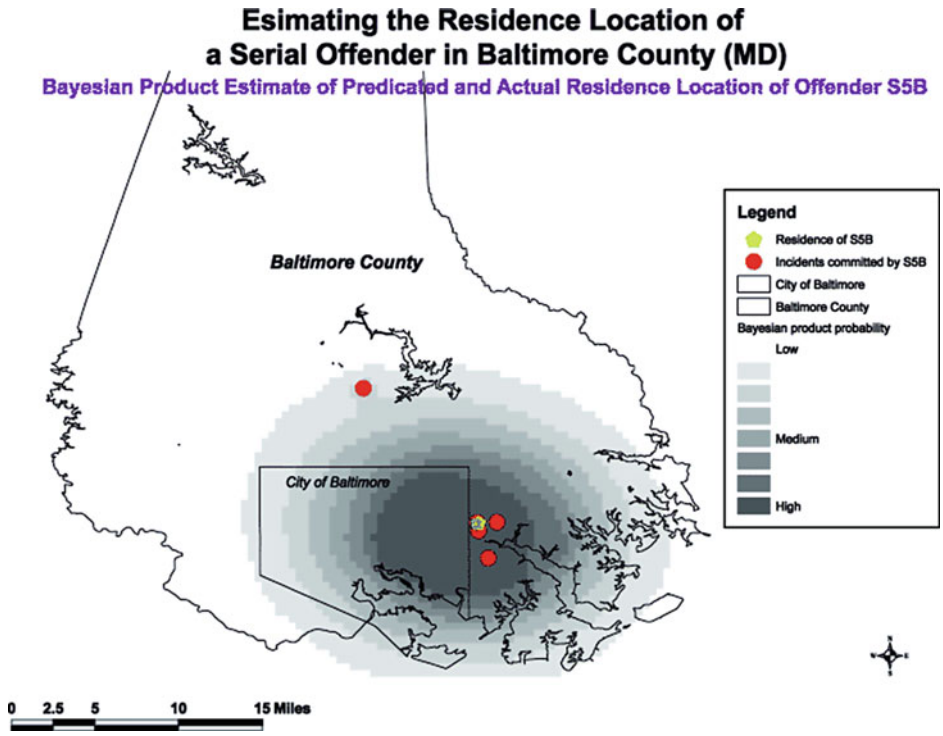
The Bayesian Journey-to-Crime routine utilizes information about the origins of other offenders who committed crimes in the same locations as a single serial offender. Again, based on a large set of records of known offenders, the routine estimates the distribution of origins of these offenders. This information can then be combined with the travel distance function to make estimates of the likely location of a serial offender where the residence location is not known. Early tests of this method suggest that it is 10–15% more accurate than the traditional travel distance only method in terms of estimating the distance between the highest probability location and the location where the offender lived.

As an example, Fig. 2 shows a Bayesian probability model of the likely residence location of a serial offender who committed five incidents between 1993 and 1997 in Baltimore County, Maryland (two burglaries and three larceny thefts). The grid cell with the highest probability is outlined. The location of the incidents is indicated as is the actual residence location of the offender when arrested. As seen, the predicted highest probability location is very close to the actual location (0.14 of a mile error).

Crime Travel Demand Modeling

CrimeStat has several routines that examine travel patterns by offenders. There is a module for modeling crime travel behavior over a metropolitan area called *Crime Travel Demand* modeling. It is an application of travel demand modeling that is widely used in transportation planning (Ortuzar and Willumsen 2001). There are four separate stages to the model. First, predictive models of crimes occurring in a series of zones (crime destinations) and originating in a series of zones (crime origins) are estimated using a non-linear (Poisson) regression model with a correction for over-dispersion (Cameron and Trivedi 1998). Second, the predicted origins and destinations are linked to yield a model of crime *trips* from each origin zone to each destination zone using a gravity-type spatial interaction model. To estimate the coefficients, the calibrated model is compared with an actual distribution of crime trips.

In the third stage, the predicted crime trips are separated into different travel modes using an approximate multinomial utility function (Domencich and McFadden 1975). The aim is



CrimeStat: A Spatial Statistical Program for the Analysis of Crime Incidents, Fig. 2 Estimating the residence location of a serial offender in Baltimore County (MD)

to examine possible strategies used by offenders in targeting their victims. Finally, the predicted crime trips by travel mode are assigned to particular routes, either on a street network or a transit network. The cost of travel along the network can be estimated using distance, travel time, or a generalized cost using the A* shortest path algorithm (Sedgewick 2002).

Once calibrated, the model can be used to examine possible interventions or policy scenarios. For example, one study examined the travel behavior of individuals who were involved in Driving-while-Intoxicated (DWI) motor vehicle crashes in Baltimore County. Neighborhoods where a higher proportion of DWI drivers involved in crashes were identified as were locations where many DWI crashes had occurred. Interventions in both high DWI driver neighborhoods and the high DWI crash locations were simulated using the model to estimate the likely reduction in DWI crashes that would be expected to occur if the interventions were actually implemented.

Key Applications

CrimeStat is oriented mostly toward the law enforcement and criminal justice fields, but it has been used widely by researchers in other fields including geography, traffic safety, urban planning, sociology, and even fields like botany and forestry. The tools reflect a range of applications that criminal justice researchers and crime analysts might find useful, some describing the spatial distribution and others being targeted to particular offenders.

For example, hot spot analysis is particularly useful for police departments. Police officers, crime analysts and researchers are very familiar with the concentration of crime or other incidents that occur in small areas. Further they are aware that many offenders live in certain neighborhoods that are particularly poor and lacking in social amenities. There is a large literature on high crime areas so that the phenomenon is very well known (e.g., see Cohen and Felson 1979; Wilson and Kelling 1982). The hot spot tools can be

useful to help police systematically identify the high crime areas as well as the areas where there are concentrations of offenders (which are not necessarily the same as the high crime locations). For example, the hot spot tools were used to identify locations with many red light running crashes in Houston as a prelude for introducing photo-enforcement. The Massachusetts State Police used the neighbor nearest hierarchical clustering algorithm to compare heroin and marijuana arrest locations with drug seizures in one small city (Bibel 2004).

Another criminal justice application is the desire to catch serial offenders, particularly high visibility ones. The Journey-to-Crime and Bayesian Journey-to-Crime routines can be useful for police departments in that it can narrow the search that police have to make to identify likely suspects. Police will routinely search through their database of known offenders; the spatial narrowing can reduce that search substantially. The CrimeStat manual has several examples of the Journey-to-Crime tool being used to identify a serial offender. As an example, the Glendale (Arizona) Police Department used the Journey-to-Crime routine to catch a felon who had committed many auto thefts (Hill 2004).

Many of the other tools are more relevant for applied researchers such as the tools for describing the overall spatial distribution or for calculating risk in incidents (police typically are interested in the volume of incidents) or for modeling the travel behavior of offenders. Two examples from the CrimeStat manual are given. First, the spatial distribution of *"Man With A Gun"* calls for service during Hurricane Hugo in Charlotte, North Carolina was compared with a typical weekend (LeBeau 2004). Second, the single-variable kernel density routine was used to model urbanization changes in the Amazon between 1996 and 2000 (Amaral et al. 2004).

Future Directions

Version 4 of CrimeStat is currently being developed (CrimeStat IV). The new version will have a complete restructuring to modernize it consistent with trends in computer science. First, there will be a new GUI interface that will be more Windows Vista-oriented. Second, the code is being revised to be consistent with the .NET framework and selected routines will be compiled as objects in a library that will be available for programmers and third-party applications. Third, additional statistics relevant for crime prediction are being developed. These include a spatial regression module using Markov Chain Monte Carlo methods and an incident detection module for identifying emerging crime hot spot spots early in their sequence. Version 4 is expected to be released early in 2009.

Cross-References

- Autocorrelation, Spatial
- Crime Mapping and Analysis
- Data Analysis, Spatial
- Emergency Evacuation, Dynamic Transportation Models
- Hotspot Detection, Prioritization, and Security
- Movement Patterns in Spatio-temporal Data
- Nearest Neighbor Problem
- Movement Patterns in Spatio-Temporal Data
- Public Health and Spatial Modeling
- Routing Vehicles, Algorithms
- Statistical Descriptions of Spatial Patterns

References

Amaral S, Monteiro AMV, Câmara G, Quintanilha JA (2004) Evolution of the urbanization process in the Brazilian Amazonia. In: Levine N (ed) CrimeStat III: a spatial statistics program for the analysis of crime incident locations (version 3.0), Chapter 8. Ned Levine & Associates, Houston; National Institute of Justice, Washington, DC

Anselin L (1995) Local indicators of spatial association – LISA. Geogr Anal 27(2):93–115

Bailey TC, Gatrell AC (1995) Interactive spatial data analysis. Longman Scientific & Technical/Burnt Mill, Essex

Bibel B (2004) Arrest locations as a means for directing resources. In: Levine N (ed) CrimeStat III: a spatial statistics program for the analysis of crime incident locations (version 3.0), Chapter 6. Ned Levine & Associates, Houston; National Institute of Justice, Washington, DC

Block CR (1994) STAC hot spot areas: a statistical tool for law enforcement decisions. In: Proceedings of the workshop on crime analysis through computer mapping. Criminal Justice Information Authority, Chicago
Brantingham PL, Brantingham PJ (1981) Notes on the geometry of crime. In: Brantingham PJ, Brantingham PL (eds) Environmental criminology. Waveland Press, Inc., Prospect Heights, pp 27–54
Cameron AC, Trivedi PK (1998) Regression analysis of count data. Cambridge University Press, Cambridge
Canter D, Gregory A (1994) Identifying the residential location of rapists. J Forens Sci Soc 34(3):169–175
Clark PJ, Evans FC (1954) Distance to nearest neighbor as a measure of spatial relationships in populations. Ecology 35:445–453
Cohen LE, Felson M (1979) Social change and crime rate trends: a routine activity approach. Am Soc Rev 44:588–608
Cressie N (1991) Statistics for spatial data. Wiley, New York
Domencich T, McFadden DL (1975) Urban travel demand: a behavioral analysis. North-Holland Publishing Co. Reprinted 1996. Available at: http://emlab.berkeley.edu/users/mcfadden/travel.html
Ebdon D (1988) Statistics in geography, 2nd edn. (with corrections) Blackwell, Oxford
Eck J, Chainey S, Cameron J, Leitner M, Wilson RE (2005) Mapping crime: understanding hot spots. Mapping and Analysis for Public Safety/National Institute of Justice, Washington, DC
Everitt BS, Landau S, Leese M (2001) Cluster analysis, 4th edn. Oxford University Press, New York
Geary R (1954) The contiguity ratio and statistical mapping. Inc Stat 5:115–145
Harries K (1999) Mapping crime: principle and practice. NCJ 178919, National Institute of Justice/US Department of Justice, Washington, DC. Available at http://www.ncjrs.org/html/nij/mapping/pdf.html
Hill B (2004) Catching the bad guy. In: Levine N (ed) CrimeStat III: a spatial statistics program for the analysis of crime incident locations (version 3.0), Chapter 10. Ned Levine & Associates, Houston; National Institute of Justice, Washington, DC
LaVigne N, Wartell J (1998) Crime mapping case studies: success in the field, vol 1. Police Executive Research Forum and National Institute of Justice/US Department of Justice, Washington, DC
LeBeau JL (2004) Distance analysis: man with a gun calls for Servicein Charlotte, N.C., 1989. In: Levine N (ed) CrimeStat III: a spatial statistics program for the analysis of crime incident locations (version 3.0), Chapter 4. Ned Levine & Associates, Houston; National Institute of Justice, Washington, DC
Levine N (1996) Spatial statistics and GIS: software tools to quantify spatial patterns. J Am Plan Assoc 62(3):381–392
Levine N (2004) Risk-adjusted nearest neighbor hierarchical clustering. In: Levine N (ed) CrimeStat III: a spatial statistics program for the analysis of crime incident locations (version 3.0), Chapter 6. Ned Levine & Associates, Houston; National Institute of Justice, Washington, DC
Levine N (2007a) CrimeStat III: a spatial statistics program for the analysis of crime incident locations (version 3.1). Ned Levine & Associates, Houston; National Institute of Justice, Washington, DC
Levine N (2007b) Bayesian journey to crime estimation (update chapter). In: Levine N (ed) CrimeStat III: a spatial statistics program for the analysis of crime incident locations (version 3.1). Ned Levine & Associates, Houston; National Institute of Justice, Washington, DC. Available at http://www.icpsr.umich.edu/crimestat
Moran PAP (1948) The interpretation of statistical maps. J R Stat Soc B 10:243–251
Ortuzar JD, Willumsen LG (2001) Modeling transport, 3rd edn. Wiley, New York
Ripley BD (1981) Spatial statistics. Wiley, New York
Rossmo DK (1995) Overview: multivariate spatial profiles as a tool in crime investigation. In: Block CR, Dabdoub M, Fregly S (eds) Crime analysis through computer mapping. Police Executive Research Forum, Washington, DC, pp 65–97
Sedgewick R (2002) Algorithms in C++: part 5 graph algorithms, 3rd edn. Addison-Wesley, Boston
Wilson JQ, Kelling G (1982) Broken windows: the police and neighborhood safety. Atl Mon 29(3):29–38

Cross-Covariance Models

▸ Hurricane Wind Fields, Multivariate Modeling

CSCW

▸ Geocollaboration

Cuda/GPU

Cheng-Zhi Qin
State Key Laboratory of Resources & Environmental Information System, Institute of Geographic Sciences & Natural Resources Research, Chinese Academy of Sciences, Beijing, P.R. China

Synonyms

General-purpose computing on graphics processing units (GPGPUs)

Definition

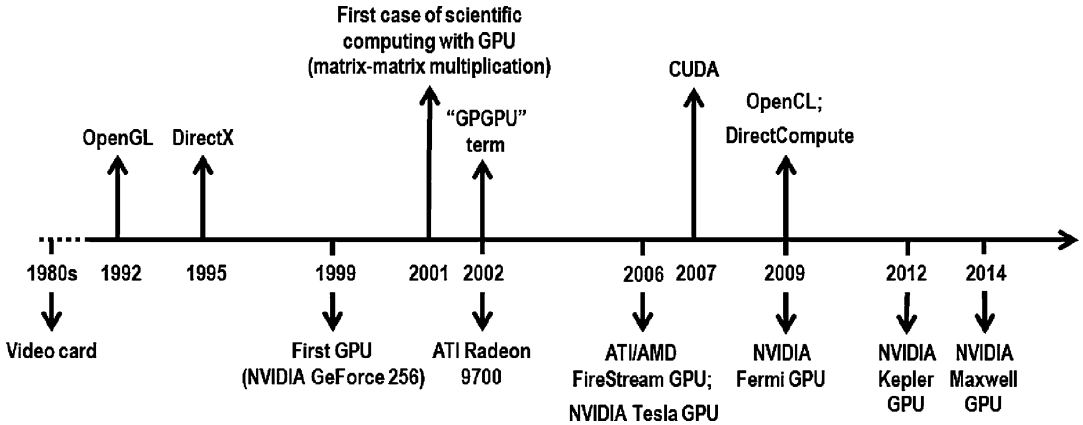
A graphics processing unit (GPU) is an electronic circuit originally designed to accelerate real-time computation for computer graphics. As one component of the basic hardware inside a modern personal computer, the GPU is connected to the central processing unit (CPU) through a system bus. For the purpose of fast image rendering, which requires that the whole process of image rendering should be completed within one frame (typically 1/30 s), the GPU has been inherently designed as a highly parallelized processor containing many cores, high memory bandwidth, and single-instruction multiple-data (SIMD) execution (Lindholm et al. 2008; Garland and Kirk 2010).

In recent years, the high performance of modern GPUs has motivated researchers to explore general-purpose computing on GPUs (GPGPUs). This has resulted in GPUs taking over the computational tasks traditionally performed by CPUs, especially compute-intensive, data-parallel tasks (Owens et al. 2007). The successful and wide use of GPGPUs is due to several attractive features of GPUs that have evolved rapidly in recent years. First, modern GPUs permit data to be bidirectionally transferred between GPU and CPU, where previously data could only be transferred from the CPU to the GPU. Second, state-of-the-art GPUs can achieve one to several orders of magnitude higher floating-point operations per second (FLOP(s)) than CPUs. Third, the programmability of modern GPUs allows programmers to develop GPU-available algorithms with the aid of GPU programming models, such as the compute unified device architecture (CUDA) for NVIDIA Corporation's GPUs, platform-independent OpenCL, and so on. Furthermore, their comparatively low cost and good cost-performance ratios make GPGPUs popular in current parallel computation.

Scientific computing benefited from the use of GPGPUs, in which matrix operation was one of the first successful cases. During the first 10 years of the twenty-first century, developers explored the significant acceleration performance of GPUs in a wide range of application domains, such as physical simulation, computational chemistry, medical image processing, as well as geocomputation.

Historical Background

The GPU was first proposed by NVIDIA through its release of GeForce 256 in 1999 (Fig. 1). A GPU is also referred to as a visual processing unit (VPU) by ATI Technologies, e.g., with the release of its Radeon 9700 in 2002. Although there are many GPU producers, most GPUs are produced by NVIDIA and ATI (ATI was acquired by AMD Inc. in 2006). Driven by demand for 3D graphics from the game industry, GPU producers



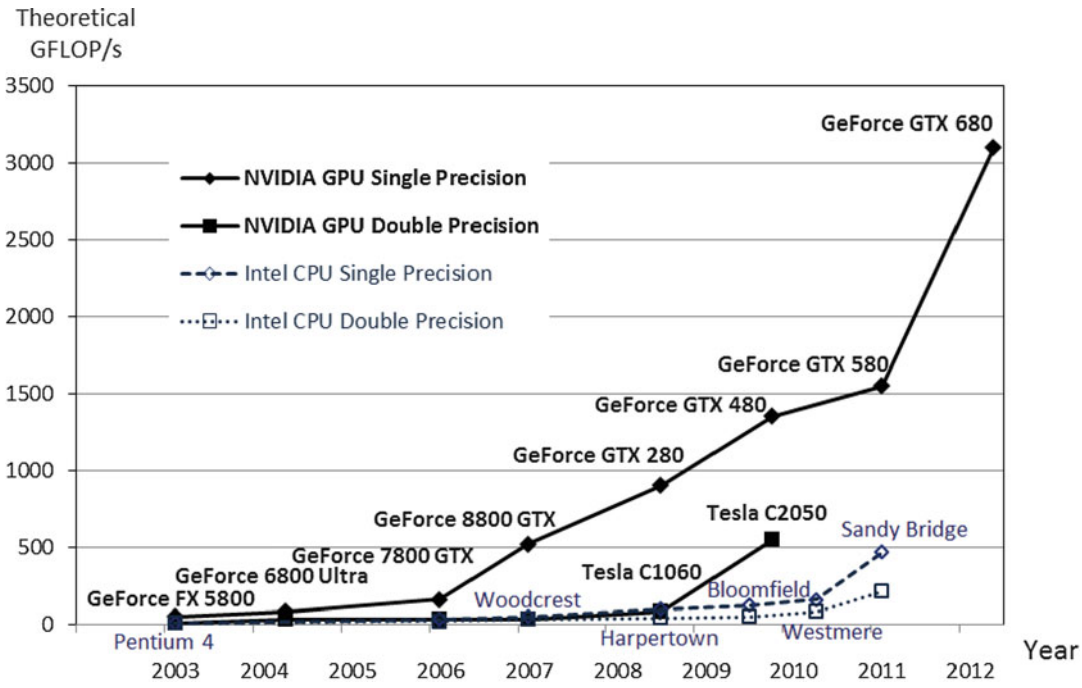
Cuda/GPU, Fig. 1 Evolution of GPUs

have continually improved their performance and capacities.

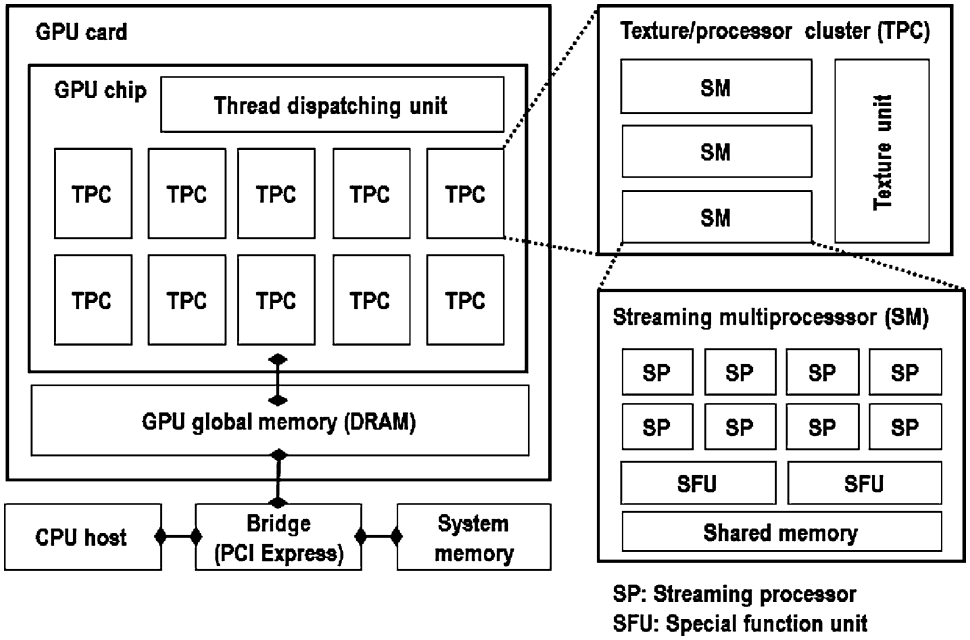
In the early 2000s, both NVIDIA and ATI added programmable capacity and floating-point support to GPUs. These improvements make it possible to off-load the non-graphical calculations from CPUs to GPUs. One of the first attempts to use GPUs in scientific computing was the matrix multiplication function developed in 2001 (Larsen and McAllister 2001). This new trend was represented by the term GPGPU, which was proposed by Dr. Mark Harris in 2002 (See http://GPGPU.org/about. Accessed on 8 Jan 2015.).

Without easy-to-use GPGPU programming tools, the use of GPGPUs would not be practical, nor would it have received such widespread acceptance. Although graphics application programming interfaces (APIs) such as OpenGL and DirectX were released in the 1990s to aid in the development of graphics applications, these graphics APIs were inconvenient for developing non-graphical applications. Instead CUDA, which was first released by NVIDIA in 2007, brought GPGPU widespread popularity. CUDA is a C-language extension for general-purpose programming used exclusively for recent NVIDIA GPUs (NVIDIA Corp. 2012). Other main programming models for GPGPU include DirectCompute from Microsoft, which is specific to newer Windows operating system, and OpenCL, which is designed by Apple Inc. and maintained by Khronos Group. Since first being released in 2009, OpenCL has been an industry standard for GPGPUs because it not only provides capabilities similar to CUDA but also has programming portability across GPUs, multicore processors, and operating systems (Stone et al. 2010; Munshi 2012).

With the many applications of GPGPU, GPU producers continually enhance their computational performance. The computing capacity of GPUs has been doubled every 12–18 months and is several times higher than that of contemporary CPUs (Lindholm et al. 2008) (Fig. 2). NVIDIA's GPUs with Fermi



Cuda/GPU, Fig. 2 Comparison of computational capacity (unit: 10 billion FLOP(s) or GFLOP(s)) between NVIDIA GPU and Intel CPU (Adapted from NVIDIA Corp. 2012)



Cuda/GPU, Fig. 3 Architecture of a typical modern GPU, Tesla unified graphics, and computing architecture of the NVIDIA GeForce GTX 280

architecture, which was first released in 2009, improved the precision of GPGPUs by supporting floating-point arithmetic being in compliance with IEEE 754 standard and providing error correction code protection for GPU memory (Du et al. 2012). Subsequent generations of NVIDIA GPUs were mainly improved with the addition of extra cores in GPUs, as well as having lower energy consumption. Nowadays, GPUs play an important role in high-performance computation. Currently, three of the ten fastest supercomputers in the world use GPUs for parallel computing (See http://www.top500.org/, Nov 2014. Accessed on 8 Jan 2015.).

Scientific Fundamentals

The high-performance computing power of GPGPU is mainly due to the GPU's highly parallel and throughput-oriented microprocessor architecture (Garland and Kirk 2010). Throughput means the average number of operations executed per unit time, which is often measured in op(s), giga-op(s) (GOP(s)), FLOP(s), or gigaFLOP(s) (GFLOP(s)). Modern GPUs are designed to have hundreds to thousands of simple, and thus small, scalar streaming processor cores organized in many computing units. Each streaming processor core can execute many computational tasks simultaneously. Their synchronization and communication is controlled by hardware. For example, GeForce GTX 280, one of NIVIDIA GPUs that was Tesla unified graphics and computing architecture, has 240 scalar "streaming processors" that are organized as 30 "streaming multiprocessors" (Fig. 3; Garland et al. 2008). Each streaming processor can execute 128 concurrent "threads." The threads in GPUs are very lightweight, and thus the overhead of thread creation, scheduling, and destroying is negligible (Garland and Kirk 2010). Modern GPUs also possess very high memory bandwidth. Bandwidth means the rate of data transfer between the memory and processor, which is often measured in gigabytes(s) (GB(s)). The high memory bandwidth of GPUs is able to support SIMD execution on streaming processor cores. Current GPUs from different producers are based on similar architecture (Cruz et al.

2011), which supports pipelined processing very efficiently and maximizes the throughput for GPGPU.

The architectures of GPUs and CPUs are very different, which are the result of different ways of processing computational tasks. Traditionally, CPUs were designed using a latency-oriented architecture, which minimized the execution time of sequential program (i.e., its latency) (Garland and Kirk 2010). The number of cores in modern CPUs is also very few (typically 2–8 cores), which is far fewer than the number of streaming processor cores contained within a typical modern GPU. Compared with the core(s) in a CPU, the streaming processor cores in GPUs are simpler and physically smaller. Although the simple core in a GPU means that a single thread executed in the GPU might be slower than if executed in a CPU, the smaller core makes it possible to contain many more parallel processing units in GPU. Thus, GPGPUs are able to achieve high total throughput via SIMD execution on GPUs.

GPGPU programming tools have been developed to aid programmers to easy exploit the strengths of GPU architecture in general-purpose applications. Most of current GPGPU programming tools, in which CUDA and OpenCL are representative, have similar programming models. Within the CUDA programming model (Fig. 4), a programmer writes a serial C function ("kernel") for a single thread that describes the computational task for one thread using private local memory. CUDA then maps the thread to the physical thread in a GPU. The programmer organizes the threads into blocks (which are mapped to streaming multiprocessors in the NVIDIA GPU) and then onto a 1D, 2D, or 3D grid. The threads in a block are grouped into warps that have shared memory. Each warp has 32 threads which are executed in a single-instruction, multiple-thread (SIMT) way (Garland et al. 2008). Blocks cannot communicate each other. This hierarchical grouping provides a good way to scale the program's parallelism with the increasing number of processor cores (Nickolls et al. 2008).

The basic features of current GPU architecture and GPGPU programming models indicate that GPGPUs are suitable especially for applications that are compute intensive and required a high degree of homogeneous computation and data parallelism (Cruz et al. 2011). When programming a GPGPU, programmers need to create a design that obtain as high a throughput as possible by mapping the parallelizability of their applications to the logic framework of the GPGPU, when the sequential computing part of their applications is to be executed on a CPU.

Key Applications

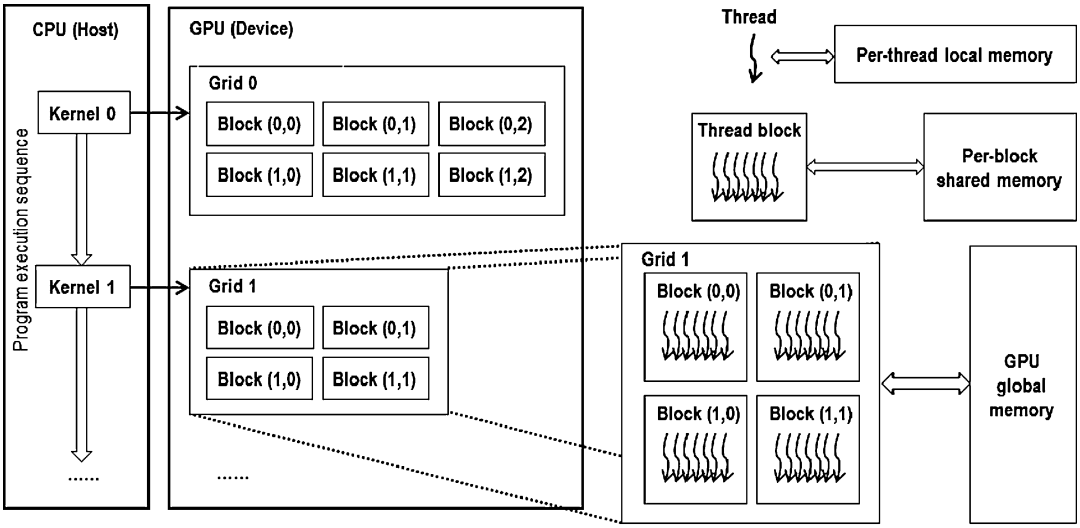
GPGPUs have been used in the parallelization of many geocomputation tasks, from single algorithms to complex spatial process simulations.

Raster-Based Geocomputation Algorithm

Raster structure is widely used in geocomputation applications and is inherently suitable for GPU architectures. Raster-based geocomputation often shows better data parallelism than vector-based geocomputation, and the domain decomposition during GPGPU for the raster-based geocomputation is often easier to reach load balancing than for vector-based geocomputation. Therefore, GPGPUs have been effectively used to accelerate many raster-based geocomputation algorithms, such as Kriging interpolation (Cheng 2013), digital terrain analysis based on the gridded digital elevation model (Qin and Zhan 2012), and remote-sensing image processing (Bernabé et al. 2012). Note that it is still challenging to use GPGPUs for some raster-based algorithms with regional operations that have complex data dependencies (e.g. Qin and Zhan 2012).

Vector-Based Geocomputation Algorithm

Vector-based geocomputation algorithms often have more complex data dependencies than raster-based algorithms. This feature makes them comparatively difficult to design GPU-available parallel algorithms for vector-based geocomputation. Some successful attempts include GPGPU algorithms used for reducing LiDAR data (Oryspayev et al. 2012), estimating



Cuda/GPU, Fig. 4 Architecture of CUDA

the roofs' solar potential based on the LiDAR point cloud (Lukač and Žalik 2013), constructing circular cartogram (Tang 2013), and finding flock patterns from spatiotemporal trajectory dataset (Fort et al. 2014).

Spatial Process Simulation and Decision Support

GPGPUs are also used in more computationally intensive spatial process simulations (often with grid structure), such as shallow-water wave simulation (Brodtkorb et al. 2010), flood modeling (Kalyanapu et al. 2011), hydrological modeling (Tristram et al. 2014), and marine ecosystem modeling (Siewertsen et al. 2013). Furthermore, the spatial decision support benefits from the powerful computational capability of GPGPUs. Such examples include fast response for releases of chemical or biological agents in urban areas (Singh et al. 2011) and parallel agent-based diffusion of spatial opinions from individuals (Tang and Bennett 2011).

Future Directions

In the computer science community, GPGPU programming tools will be continually improved. For example, OpenCL is designed for programming portability; however, its performance portability is not guaranteed (Du et al. 2012). This issue is worth devoting more research efforts in order to help OpenCL-based programs obtain better performance across different hardware configurations.

In the geocomputation domain, geocomputation developers would gladly welcome a domain-specific parallel programming library that hides the details of parallel programming as much as possible to developers and can be provided to help simplify coding of the parallel geocomputation program. One recent attempt on this issue is a parallel strategy proposed for raster-based geocomputation using GPU and other popular types of parallel computing platforms (Qin et al. 2014).

Meanwhile, more work devoted to designing GPGPUs for specific geocomputation algorithms is needed, especially for vector-based algorithms and raster-based algorithms with regional operation.

Because GPGPUs are a powerful means for accelerating computational tasks with specific features, GPU clusters and heterogeneous GPU-CPU clusters have been developed to deal with large-scale complex computations. However, it is not so easy nor trivial to decide the correct parallel strategies to fully exploit these platforms

(Feichtinger et al. 2011). In particular, for large-scale geosimulation coupling multiple processes, more practices are needed.

Cross-References

- CUDA
- Distributed Geocomputation
- MPI
- Parallel Computing

References

Bernabé S, Plaza A, Marpu PR et al (2012) A new parallel tool for classification of remotely sensed imagery. Comput Geosci 46:208–218

Brodtkorb AR, Hagen TR, Lie K-A et al (2010) Simulation and visualization of the Saint-Venant system using GPUs. Comput Vis Sci 13:341–353

Cheng T (2013) Accelerating universal Kriging interpolation algorithm using CUDA-enabled GPU. Comput Geosci 54:178–183

Cruz FA, Layton SK, Barba LA (2011) How to obtain efficient GPU kernels: an illustration using FMM & FGT algorithms. Comput Phys Commun 182:2084–2098

Du P, Weber R, Luszczek P et al (2012) From CUDA to OpenCL: towards a performance-portable solution for multi-platform GPU programming. Parallel Comput 38:391–407

Feichtinger C, Habich J, Kostler H et al (2011) A flexible patch-based lattice Boltzmann parallelization approach for heterogeneous GPU-CPU clusters. Parallel Comput 37:536–549

Fort M, Sellares A, Valladares N (2014) A parallel GPU-based approach for reporting flock patterns. Int J Geogr Inf Sci 28(9):1877–1903

Garland M, Kirk DB (2010) Understanding throughput-oriented architectures. Commun ACM 53:58–66

Garland M, LeGrand S, Nickolls J et al (2008) Parallel computing experiences with CUDA. IEEE Micro 28(4):13–27

Kalyanapu AJ, Shankar S, Pardyjak ER et al (2011) Assessment of GPU computational enhancement to a 2D flood model. Environ Model Softw 26:1009–1016

Larsen ES, McAllister D (2001) Fast matrix multiplies using graphics hardware. Paper presented at Supercomputing, Denver, 10–16 Nov 2001

Lindholm E, Nickolls J, Oberman S et al (2008) NVIDIA tesla: a unified graphics and computing architecture. IEEE Micro 28(2):39–55

Lukač N, Žalik B (2013) GPU-based roofs' solar potential estimation using LiDAR data. Comput Geosci 52:34–41

Munshi A (2012) The OpenCL specification (Version 1.2). Khronos OpenCL Working Group

Nickolls J, Buck I, Garland M et al (2008) Scalable parallel programming with CUDA. ACM Queue 6(2):40–53

NVIDIA Corp. (2012) NVIDIA CUDA C programming guide (Version 4.2)

Oryspayev D, Sugumaran R, DeGroote J et al (2012) LiDAR data reduction using vertex decimation and processing with GPGPU and multicore CPU technology. Comput Geosci 43:118–125

Owens JD, Luebke D, Govindaraju N et al (2007) A survey of general-purpose computation on graphics hardware. Comput Graph Forum 26(1):80–113

Qin C-Z, Zhan L (2012) Parallelizing flow-accumulation calculations on Graphics Processing Units—from iterative DEM preprocessing algorithm to recursive multiple-flow-direction algorithm. Comput Geosci 43:7–16

Qin C-Z, Zhan L-J, Zhu A-X et al (2014) A strategy for raster-based geocomputation under different parallel computing platforms. Int J Geogr Inf Sci 28(11):2127–2144

Siewertsen E, Piwonski J, Slawig T (2013) Porting marine ecosystem model spin-up using transport matrics to GPUs. Geosci Model Dev 6:17–28

Singh B, Pardyjak ER, Norgren A et al (2011) Accelerating urban fast response Lagrangian dispersion simulations using inexpensive graphics processor parallelism. Environ Model Softw 26:739–750

Stone JE, Gohara D, Shi G (2010) OpenCL: a parallel programming standard for heterogeneous computing systems. Comput Sci Eng 12(3):66–73

Tang W (2013) Parallel construction of large circular cartograms using graphics processing units. Int J Geograph Inf Sci 27(11):2182–2206

Tang W, Bennett DA (2011) Parallel agent-based modeling of spatial opinion diffusion accelerated using graphics processing units. Ecol Model 222:3605–3615

Tristram D, Hughes D, Bradshaw K (2014) Accelerating a hydrological uncertainty ensemble model using graphics processing units (GPUs). Comput Geosci 62:178–186

Customization

- Mobile Usage and Adaptive Visualization

Cyberinfrastructure for Spatial Data Integration

Ilya Zaslavsky
San Diego Supercomputer Center, University of California San Diego, San Diego, CA, USA

Synonyms

E-Science; Heterogeneity; Standards; Virtualization, resource

Definition

The term cyberinfrastructure (CI) refers to a new research environment that supports integration of geographically distributed computing and information processing services to enable a new level of data-intensive collaborative science enterprise. It includes high-performance data management and storage hardware and software, combined with secure data access and advanced information- and knowledge-management technologies and a variety of search, analysis, visualization, modeling and collaboration tools linked over high-speed networks, to create an enabling end-to-end framework for scientific discovery. CI applications in earth sciences span such disciplines as earthquake modeling and prediction, ecology, atmospheric sciences, hydrology and oceanography.

Historical Background

The term was first articulated at a press briefing on the Presidential Decision Directive (PDD-63) in 1998, in reference to information systems as the major component of the nation's critical infrastructures in need of protection (http://www.fas.org/irp/offdocs/pdd/pdd-63.htm). In 2003, the National Science Foundation's (NSF's) blue ribbon panel used the term in outlining the need to efficiently connect high-performance computing resources, information resources, and researchers, to support scientific discovery. Several large information technology projects were funded by the NSF, focused on CI development in earth science and other domains. In June 2005, an Office of Cyberinfrastructure was created at the NSF (http://www.nsf.gov/dir/index.jsp?org=OCI). At the same time, the National Institutes of Health supported advanced CI projects in biomedical sciences, and a range of infrastructure projects were developed in industry. These developments were accompanied by the emergence of relevant information exchange standards, more importantly *web services*, and *service-oriented architecture* (SOA), which now form the backbone of the large CI projects.

Several of these projects have been using spatial information technologies for monitoring distributed resources, searching for resources and extracting data fragments based on their spatial properties, integrating spatial data of different types, creating composite maps, and serving spatial data in standard formats, etc.

Scientific Fundamentals

The challenges of integrating spatial information from physically distributed spatial data sources derive from:

- Extreme heterogeneity in how web-accessible information is collected, represented, described, interpreted and queried
- Exponentially increasing volumes of available data, numbers of users and applications
- Volatility of the web, with autonomously managed data sources and services
- The need to transform data into standard agreed-upon forms suitable for spatial representation and analysis

The mechanisms proposed within CI projects to address these challenges follow the idea

of *resource virtualization*, that is, decoupling information sources and services from their specific implementation, geographic location, or physical configuration. Such resources can be pooled together to address computation- or data-intensive problems. For spatial data, this typically involves:

- Standardization of spatial metadata, specifically following FGDC Content Standard for Digital Geospatial Metadata, ISO 19115 and 19139, and standardscompliant registration of different types of spatial data and services to catalogs that can be accessed and queries in a standard fashion
- Standardization of spatial data transmission on the web (GML (Geographic Markup Language), in particular), and XML-based (eXtensible Markup Language) standards for graphic rendering on the web (e.g., SVG (Scalable Vector Graphics))
- Wrapping of spatial data into standardscompliant GIS services, most notably following specification developed within the Open Geospatial Consortium
- Publication of geographic information systems (GIS) *web services* representing common GIS functionality which can be invoked from different client environments; and development of convenient platforms for generation of such services (e.g., ESRI's ArcGIS Server)
- Creation of advanced query-processing clients, in particular implementing *spatial information mediation* techniques for assembling spatial data fragments into composite query responses
- automatic map assembly services that support 'on-the-fly' generation of thematic maps from multiple grid-enabled data sources
- Single sign-on authentication, whereupon login users are issued a certificate that specifies their access rights with respect to different registered resources, which may have different security models
- Replication and caching of commonly used spatial data collections and metadata, to support quality of service and alleviate 'single point of failure' issues
- Development of spatial query processing architectures for distributing geospatial processing to multiple processing nodes (including the emerging P2P-type architectures for spatial data processing)
- Development of techniques for capturing data and service semantics, to enable semantic mediation across disparate sources using different vocabularies, and orchestration of services into processing pipelines (workflows)
- Development of principles for archiving large volumes of spatial data, and instantiating the state of GIS databases for a given time
- Development of map-based *grid* monitoring services.

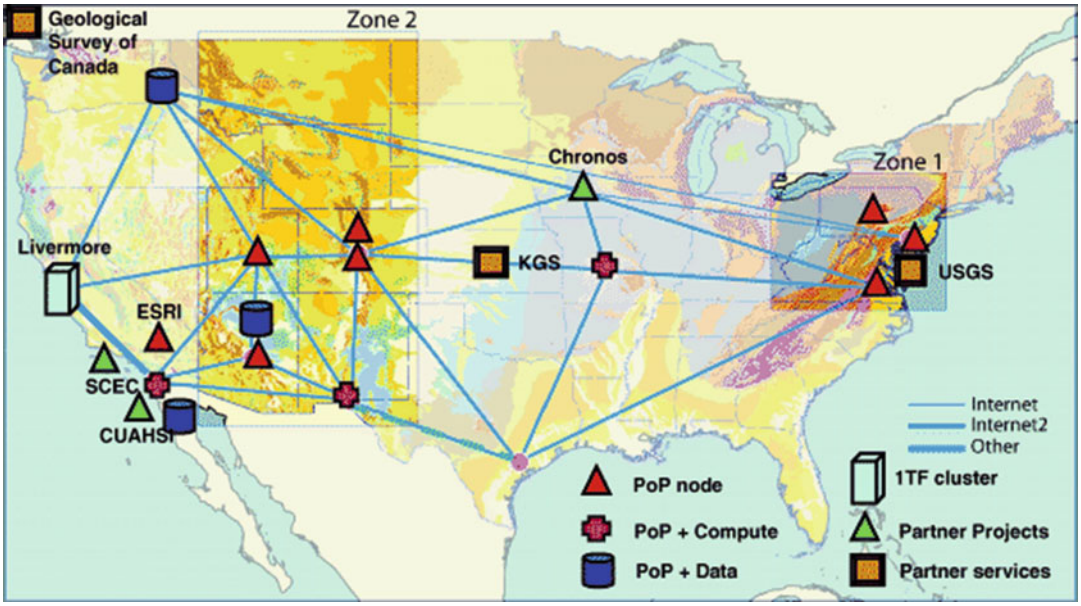
Given the traditional GIS focus on spatial data infrastructure, standard description and reuse of secondary data sources, and support of spatial data integration, individual components of the CI agenda have been long explored in GIS literature. However, the experience of several large CI projects reviewed below and a series of interoperability initiatives led by the Open Geospatial Consortium, have demonstrated that these components can be integrated into functioning computing systems supporting discovery, integration and analysis of distributed spatial data resources and services.

Key Applications

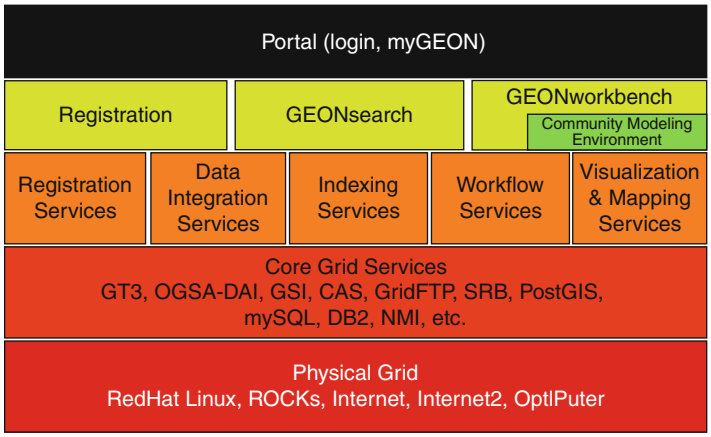
Geology

The Geosciences Network (GEON) (www.geongrid.org) is a large multiuniversity project supported by the NSF (Fig. 1). Its mission is development of infrastructure for data integration in the geosciences. GEON's SOA includes tools for publication, search, analysis and integration of several types of spatial data (Fig. 2). GIS-related innovations in GEON include:

- Map-based *grid* node monitoring service
- Web-based environment for publishing and registering shapefiles, Web Map Service/Web Feature Service (WMS/WFS), grids, and other types of spatial data



Cyberinfrastructure for Spatial Data Integration, Fig. 1 Organization of the Geosciences Network (GEON) *grid*, showing locations of GEON point-of-presence (*PoP*) nodes, compute nodes, and data nodes, as well as partner institutions and services



Cyberinfrastructure for Spatial Data Integration, Fig. 2 GEON *grid* software layers. Software components spelled out: *ROCKS* SDSC clustering toolkit, www.rocksclusters.org; *OptIPuter* novel infrastructure integrating optical networking, internet protocol, computer storage, analysis and visualization, http://www.optiputer.net/; *GT3* GLOBUS Toolkit, v.3, www.globus.org; *OGSA-DAI* middleware supporting web service access to data resources, www.ogsadai.org.uk; *CAS* Community Authorization Service, http://www.globus.org/grid_software/security/cas.php; *GridFTP* (high-performance data transfer protocol, http://www.globus.org/grid_software/data/gridftp.php; *SRB* SDSC storage resource broker, http://www.sdsc.edu/srb/; *PostGIS* GIS extension to the PostgreSQL database management system, http://postgis.refractions.net/; *mySQL* open source database management system, www.mysql.com/; *DB2* IBM's DBMS, www.ibm.com/db2; *NMI* NSF middleware initiative tools, http://www.nsf-middleware.org/

- Ability to semantically annotate spatial data and services, and ontology-based search
- Map assembly services, for automatic generation of online thematic maps from fragments extracted from several registered data sources
- Ability to securely engage high-performance computing and networking resources (clusters, supercomputers) in computation- and data-intensive processing (e.g., with LIDAR (Light Detection And Ranging) datasets).

Ecology

The National Ecological Observatory Network (NEON) (www.neoninc.org) is designed as a nationwide measurement and observation system enabling regional- to continental-scale multidisciplinary ecological modeling and forecasting. While in its design phase at the time of writing, NEON is envisioned as an environment integrating sensor networks and advanced computing resources based on SOA and *grid* computing approaches.

Hydrology and Environmental Engineering

The Water And Environmental Research Systems (WATERS) network initiative includes two interrelated CI efforts supported by NSF: the Consortium of Universities for the Advancement of Hydrologic Sciences (CUAHSI) hydrologic information system (HIS) (www.cuahsi.org/his/), and the Collaborative Large-Scale Engineering Analysis Network for Environmental Research (CLEANER) (cleaner.ncsa.uiuc.edu) projects. The core of CUAHSI HIS, for example, is development of uniform data models and web service wrappers for heterogeneous repositories of observation data available from federal (e.g., United States Geological Survey (USGS), Environment Protection Agency (EPA), National Oceanic and Atmospheric Administration (NOAA), state, and local sources. Both projects have been actively using GIS for online browsing, search and visualization, and developed GIS-based client interfaces, both online and desktop, for the observation *web services* available from CUAHSI HIS.

Biomedical Sciences

The Biomedical Informatics Research Network (BIRN) (www.nbirn.net) project, supported by the National Institutes of Health, has pioneered many *grid* developments, by creating a production-level infrastructure for integrating neuroscience data across multiple participating universities. BIRN has a significant spatial information integration component, focused on combining information from multiple distributed atlases of animal and human brains. In particular, it supports spatial and semantic annotation of neuroscience images and segmentations, querying across distributed image collections registered to a common stereotaxic coordinate system and available through the BIRN data *grid*, and retrieving neuroscience images from GIS map services into GIS-based client interfaces. An extension of BIRN is the National Institute of Environmental Health Sciences (NIEHS) Hurricane Response Portal (http://www-apps.niehs.nih.gov/katrina/), where online GIS is integrated with portal technologies so that user access to various spatial data layers is controlled by credentials evaluated via the project's portal.

Other projects that focus on CI development within various earth science domains, and use GIS approaches, include the Ocean Research Interactive Observatory Networks (ORION) (http://www.orionprogram.org/), Real-Time Observatories, Applications, and Data Management Network (ROADnet) (roadnet.ucsd.edu), Linked Environments for Atmospheric Discovery (LEAD) (lead.ou.edu). A range of similar efforts are supported by federal agencies such as USGS, EPA and NOAA, where relevant geographic data can be accessed via the Geospatial One Stop portal (GOS) (www.geodata.gov). Similar developments exist in Europe, for example, the UK e-Science program, with several GIS-focused efforts within the EDINA project, an on-line abstract and indexing database at Edinburgh University Library (http://edina.ac.uk/).

Future Directions

CI and services for spatial information integration is a rapidly developing area, expanding to new research domains that span from integration of archaeological data to plant sciences CI (judging from recently awarded or announced grant competitions). While each project focuses on a unique combination of domain and computer science research, several topics have emerged that will likely define CI development for several years to come. They include:

- Support for environmental observatories, and handling of heterogeneous streams of observation data
- Creation and management of large earth science ontologies, and ontology tagging of important data sets
- Efficient integration of multimodal and geographically distributed data sources
- Information integration across spatial scales
- Mining of web-accessible sources for place-related information and knowledge
- Efficient organization of distributed computational infrastructure for massive data-intensive computations, fault-tolerance and rapid response
- Spatiotemporal indexing and scheduling of grid resources
- Secure access to heterogeneous grid resources
- Social aspects of organizing geographically distributed research groups; and many more.

Cross-References

▸ Grid
▸ Service-Oriented Architecture
▸ Spatial Information Mediation
▸ Web Services

Recommended Reading

Berman F, Hey A, Fox G (2003) Grid computing: making the global infrastructure a reality. Wiley, Indianapolis

Catlett C, Smarr L (1992) Metacomputing. Commun ACM 35:44–52

Chervenak A, Foster I, Kesselman C, Salisbury C, Tuecke S (2001) The data grid: towards an architecture for the distributed management and analysis of large scientific datasets. J Netw Comput Appl 23:187–200

DeVogele T, Parent C, Spaccapietra S (1998) On spatial database integration. Int J Geogr Inf Sci 12:335–352

Erl T (2005) Service oriented architecture: concepts, technology, and design. Prentice Hall, Upper Saddle River

Foster I, Kesselman C, Tuecke S (2001) The anatomy of the grid: enabling scalable virtual organizations. Int J Supercomput Appl 15:200–222

Foster I, Kesselman C, Nick J, Tuecke S (2002) The physiology of the grid: an open grid services architecture for distributed systems integration. Open Grid Service Infrastructure WG, Global Grid Forum, 22 June 2002

Lacroix Z, Boucelma O, Essid M (2002) A WFS-based mediation system for GIS interoperability. In: Workshop on advances in geographic information systems. ACM, New York, pp 23–28

National Science Foundation (1993) NSF blue ribbon panel on high performance computing, 19 Oct 1993 (NSB 93-205). Available via http://www.nsf.gov/pubs/stis1993/nsb93205/nsb93205.txt. Accessed 30 Sept 2005

O'Reilly T (2005) What is Web 2.0. http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html. Accessed 2 Nov 2004

Reed C (2004) Data integration and interoperability: ISO/OGC standards for geo-information. www.directionsmag.com/article.php?article_id=687. Accessed

Zaslavsky I, Baru C, Bhatia K, Memon A, Velikhov P, Veytser V (2005) Grid-enabled mediation services for geospatial information. In: Agouris P, Croitoru A (eds) Next generation geospatial information. From digital image analysis to spatiotemporal databases. Taylor and Francis, London, pp 15–24