



Semi-automatic Description of Named Rivers and Bays for their Representation in a Terminological Knowledge Base

Juan Rojas-Garcia¹ 

Department of Translation and Interpreting, University of Granada, Spain
juanrojas@ugr.es

Pamela Faber 

Department of Translation and Interpreting, University of Granada, Spain
pfaber@ugr.es

Abstract

EcoLexicon (<http://ecolexicon.ugr.es>) is a terminological knowledge base on environmental science, whose design permits the geographic contextualization of data. For the geographic contextualization of landform concepts, this paper presents a semi-automatic method for extracting terms associated with named rivers (e.g., *Pearl River*) and named bays (e.g., *San Francisco Bay*). Terms were extracted from an English specialized corpus on Coastal Engineering, where named rivers and bays were automatically identified. Statistical procedures were applied for selecting terms, rivers, and bays in distributional semantic models to construct the conceptual structures underlying the usage of named rivers and bays in Coastal Engineering texts. The rivers sharing associated terms were also automatically clustered and represented in the same conceptual network. The same was done for named bays sharing associated terms. The results showed that the method successfully described the semantic frames for named rivers and bays with explanatory adequacy, according to the premises of Frame-based Terminology. Furthermore, the semantic networks unveiled that the named rivers and bays mentioned in the Coastal Engineering corpus are both thematically related to sediment concentration and sediment transport in rivers, sediment discharge into bays and seas, and the negative effects of sediment supply decrease on coastal erosion because of human activities.

2012 ACM Subject Classification Computing methodologies → Information extraction; Computing methodologies → Lexical semantics; Computing methodologies → Semantic networks

Keywords and phrases Named river, Named bay, Conceptual information extraction, Geographic contextualization, Text mining, Frame-based Terminology

Acknowledgements This research was carried out as part of project FFI2017-89127-P, Translation-Oriented Terminology Tools for Environmental Texts (TOTEM), funded by the Spanish Ministry of Economy and Competitiveness. Funding was also provided by an FPU grant given by the Spanish Ministry of Education to the first author.

1 Introduction

EcoLexicon is a multilingual, terminological knowledge base on environmental science (<http://ecolexicon.ugr.es>) that is the practical application of Frame-based Terminology ([12]). Since most concepts designated by environmental terms are multidimensional ([11]), the flexible design of EcoLexicon permits the contextualization of data so that they are more relevant to specific subdomains, communicative situations, and geographic areas ([19]). However, the geographic contextualization of landform concepts, namely, named landforms, is barely tackled in terminological resources because of two reasons in our opinion: (1) they

¹ Corresponding author. E-mail: juanrojas@ugr.es

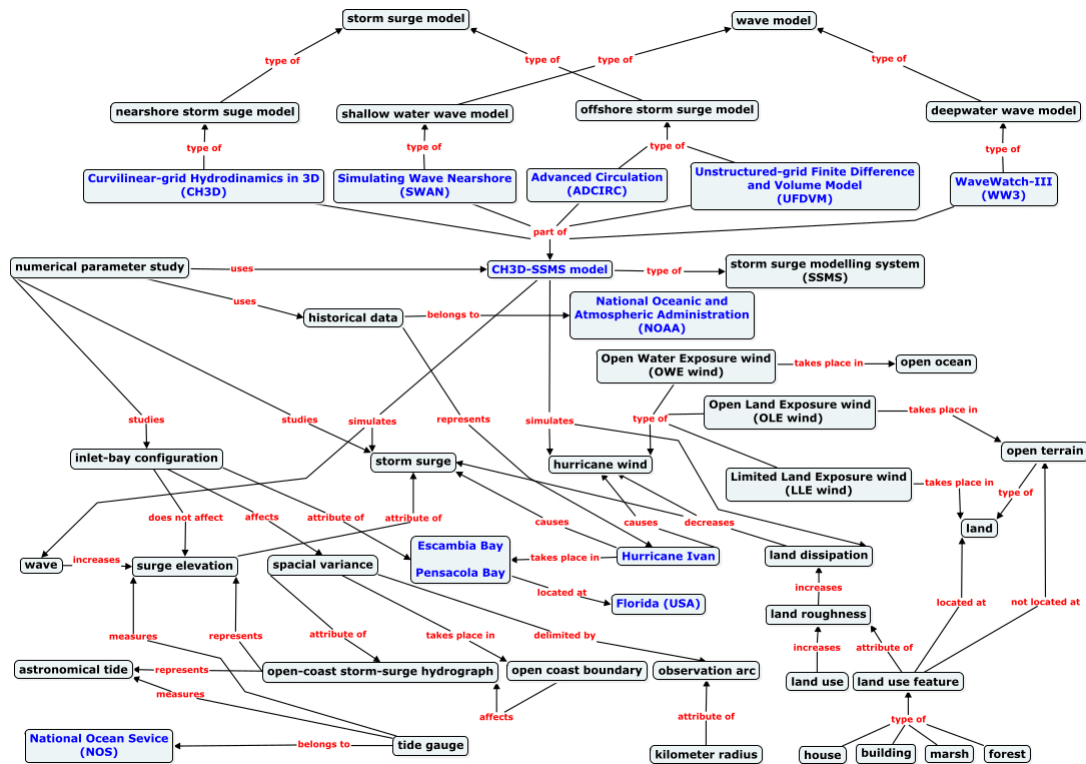


Figure 1 Semantic network of the terms associated with *Escambia* and *Pensacola* bays in an English Coastal Engineering corpus.

are considered mere instances of concepts such as RIVER, BAY, or BEACH, and their specific relational behaviour with other concepts in a specialized knowledge domain is thus neglected and not semantically described; (2) their semantic representation depends on knowing which terms are related to each named landform, and how these terms are related to each other, a time-consuming task taking into account that terminologists do not often resort to natural language processing systems beyond corpus tools such as Sketch Engine ([17]).

Consequently, this paper presents a semi-automatic method of extracting terms associated with named rivers (e.g., *Omaru River*) and named bays (e.g., *Suisun Bay*) as types of landform from a corpus of English Coastal Engineering texts. The aim is to represent that knowledge in semantic networks in EcoLexicon according to the theoretical premises of Frame-based Terminology. Hence, on the hypothesis that named rivers and bays should be considered concepts rather than instances in the Coastal Engineering domain, each named river and bay should appear in the context of a specialized semantic frame that highlights both its relation to other terms and the relations between those terms.

These semantic frames, such as that shown in Figure 1 underlying the usage of *Escambia* and *Pensacola* bays in Coastal Engineering texts, provide the background knowledge about named rivers and bays necessary in communicative situations, such as specialized translation to appropriately render terms into another language ([12]). Moreover, they make the semantic and syntactic behavior of terms explicit by means of the description of conceptual relations and term combinations ([10]).

The rest of this paper is organized as follows. Section 2 provides motivations for the research, and background on distributional semantic models and clustering techniques.

Section 3 explains the materials and methods applied in this study, namely, the automatic identification of named rivers and bays, the selection procedures for terms, bays, and rivers in distributional semantic models, and the clustering technique for bays and rivers sharing associated terms. Section 4 shows the results obtained. Finally, Section 5 discusses the results and presents the conclusions derived from this work as well as plans for future research.

2 Background and literature review

2.1 Motivations for the research

Despite the fact that named landforms, among other named entities, are frequently found in specialized texts on environment, their representation and inclusion in knowledge resources has received little research attention, as evidenced by the lack of named landforms in terminological resources for the environment such as DiCoEnviro², GEMET³ or FAO Term Portal⁴. In contrast, AGROVOC⁵ basically contains a list of named landforms with hyponymic information, whereas ENVO⁶ provides descriptions of the named landforms with only geographic details, and minimal semantic information consisting of the relation *located_in* (and *tributary_of* in the case of named rivers and bays).

Up to the present, knowledge resources have limited themselves to representing concepts such as BAY, RIVER or BEACH, on the assumption that the concepts linked to each of them are also appropriate, respectively, to all instances of named bays, rivers and beaches in the real world. This issue is evident in the following description of forcing mechanisms acting on suspended sediment concentrations (SSC) in bays and rivers.

According to [26], temporal variations in the SSC of bays and rivers are the result of a variety of forcing mechanisms. River discharge is a primary controlling factor, as well as tides, meteorological forcing (i.e., wind-wave resuspension, offshore winds, storm and precipitation), and human activities. Several of these mechanisms tend to act simultaneously. Nonetheless, the specific mix of active mechanisms is different in each bay and river. For example, SSC in San Francisco Bay is controlled by spring-neap tidal variability, winds, freshwater runoff, and longitudinal salinity differences, whereas precipitation and river discharge are the mechanisms in Suisun Bay. In Yangtze River, SSC is controlled by tides and wind forcing, whereas river discharge, tides, circulation, and stratification are the active forcing mechanisms in York River.

Consequently, in a knowledge resource, a list of forcing mechanism concepts semantically linked to BAY and RIVER concepts would not represent the knowledge really transmitted in specialized texts. To cope with this type of situation, terminological knowledge bases should include the semantic representation of named landforms.

To achieve that aim in EcoLexicon regarding named rivers and bays, the knowledge should be represented in a semantic network according to the theoretical premises of Frame-based Terminology ([12]), which propose knowledge representations with explanatory adequacy for enhanced knowledge acquisition in communicative situations such as specialized translation ([10]). Hence, on the hypothesis that named rivers and bays should be considered concepts rather than instances, each named river and bay should appear in the context of a specialized

² http://olst.ling.umontreal.ca/cgi-bin/dicoenviro/search_enviro.cgi

³ <https://www.eionet.europa.eu/gemet/en/themes/>

⁴ <http://www.fao.org/faoterm/en/>

⁵ <http://aims.fao.org/en/agrovoc>

⁶ <http://www.environmentontology.org/Browse-Env0>

semantic frame that highlights both its relation to other terms and the relations between those terms. The construction of these semantic networks and the semi-automatic extraction of terms from a specialized corpus are described in this paper. As far as we know, this framework has not been studied in the context of specialized lexicography, which is an innovative aspect of this work. Needless to say that the extraction and description of named landforms from text corpora have been applied in the field of Geographic Information Retrieval ([8]; [33]), but not with the purposes of the Frame-based Terminology.

2.2 Distributional semantic models

Distributional semantic models (DSMs) represent the meaning of a term as a vector, based on its statistical co-occurrence with other terms in the corpus. According to the distributional hypothesis, semantically similar terms tend to have similar contextual distributions ([24]). The semantic relatedness of two terms is estimated by calculating a similarity measure of their vectors, such as Euclidean distance, or cosine similarity, *inter alia*.

Depending on the language model ([3]), DSMs are either count-based or prediction-based. Count-based DSMs calculate the frequency of terms within a term's context (i.e., a sentence, paragraph, document, or a sliding context window spanning a given number of terms on either side of the target term). Correlated Occurrence Analogue to Lexical Semantics (COALS) ([29]) is an example of this type of model.

Prediction-based models exploit neural probabilistic language models, which represent terms by predicting the next term on the basis of previous terms. Examples of predictive models include continuous bag-of-words (CBOW) and skip-gram (SG) models ([23]).

DSMs have been used in combination with clustering. Work on lexical semantics applying DSMs and clustering techniques includes identification of semantic relations ([5]), word sense discrimination and disambiguation ([28]), automatic metaphor identification ([31]), and classification of verbs into semantic groups ([14]).

3 Materials and methods

3.1 Materials

3.1.1 Corpus data

The terms related to named rivers and bays were extracted from a subcorpus of English texts on Coastal Engineering, comprising roughly 7 million tokens and composed of specialized texts (scientific articles and PhD dissertations) and semi-specialized texts (textbooks on Coastal Engineering). This subcorpus is part of the English EcoLexicon corpus (23.1 million tokens) (see [21] for a detailed description).

3.1.2 GeoNames geographic database

The automatic detection of the named rivers and bays in the corpus was performed with a GeoNames database dump. GeoNames (<http://www.geonames.org>) has over 10 million proper names for 645 different geographic entities, such as bays, beaches, rivers, mountains, etc. For each entity, information about their normalized designations, alternate designations, latitude, longitude, and location name is stored. A daily GeoNames database dump is publicly available as a worldwide text file.

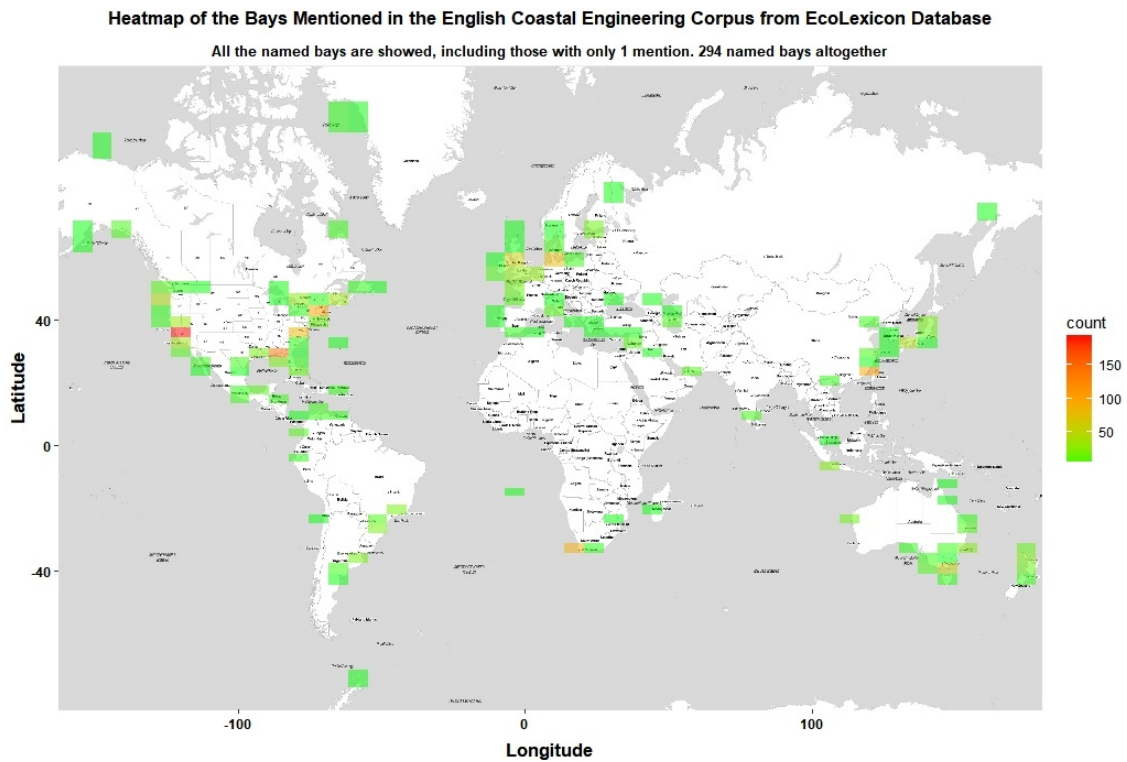


Figure 2 Map with the location and color-coded frequency of the 294 named bays.

3.2 Methodology

3.2.1 Pre-processing

After their compilation and cleaning, the corpus texts were tokenized, tagged with parts of speech, lemmatized, and lowercased in R programming language. The multi-word terms in EcoLexicon were then automatically matched in the lemmatized corpus and joined with underscores.

3.2.2 Named landform recognition

Both normalized and alternate names of the rivers and bays in GeoNames were searched in the lemmatized corpus. A total of 681 designations for rivers and 306 for bays were recognized and listed. Nevertheless, since various designations can refer to the same river or bay because of syntactic variation (e.g., *Mersey River* and *River Mersey*; *Bay of Ingleses* and *Ingleses Bay*), and orthographic variation (e.g., *Yangtze* and *Yangtse River*), a procedure was created to identify variants and give them a single designation in the corpus. Because of space constraints, the procedure is not described.

Once the variants were normalized in the lemmatized corpus and joined with underscores, the number of named rivers was 662, and 294 for bays. The bays are shown on the map in Figure 2, with color-coded rectangles that depict their frequency in the corpus. Their latitudes and longitudes were retrieved from the GeoNames database dump.

The occurrence frequency for the named rivers ranged from 118 to one mention, and from 127 to one mention for the bays. In our study, only those rivers and bays with a frequency greater than 9 were considered. Figure 3 shows the 55 named rivers that fulfilled

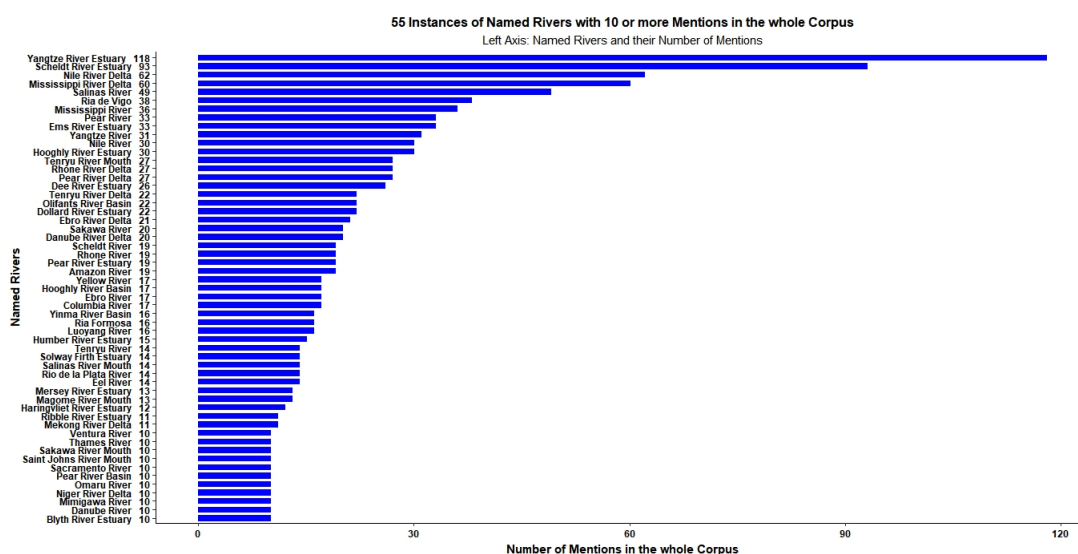


Figure 3 Designations and mentions of the 55 named rivers with frequency higher than 9.

this condition, along with their number of mentions. In the case of the bays, 29 designations fulfilled the condition.

3.2.3 Construction of two term-term matrices for named rivers and bays

Two count-based DSMs for named rivers and bays, respectively, were selected to obtain term vectors since this type of DSM outperforms prediction-based ones on small-sized corpora ([2]; [30]).

For the construction of both DSMs, terms with fewer than 3 characters, numbers and punctuation marks were removed. Additionally, the minimal occurrence frequency was set to 5 ([9]). The sliding context window spanned 30 terms on either side of the target term because large windows improve the DSM performance for small corpora ([29]; [7]), and capture more semantic relations ([15]). We followed standard practice and did not use stopwords (i.e., determiners, conjunctions, relative adverbs, and prepositions) as context words ([16]). Since only nouns are represented in the semantic networks, adjectives, adverbs, and verbs were also disregarded as context words.

For the rivers, the resulting DSM was a $4,705 \times 4,705$ matrix, whose row vectors represented the 55 named rivers plus the 4,650 terms inside the context windows of 30 terms on either side of those rivers. For the bays, a $3,867 \times 3,867$ matrix was obtained, which represented the 29 named bays plus the 3,838 terms inside their context windows.

3.2.4 Term selection procedure and weighting schemes

Subsequently, for the rivers, a $55 \times 4,650$ submatrix was extracted, where the rows represented the 55 named rivers, and the columns represented the 4,650 terms co-occurring with them. For the bays, a $29 \times 3,838$ submatrix was extracted. To cluster rivers sharing associated terms, the terms that best discriminated different groups of rivers were selected. This was done by applying Moisl's statistical criteria ([25]), whereby only the column vectors with the highest values in *raw frequency*, *variance*, *variance-to-mean ratio* (vmr), and *term frequency-inverse*

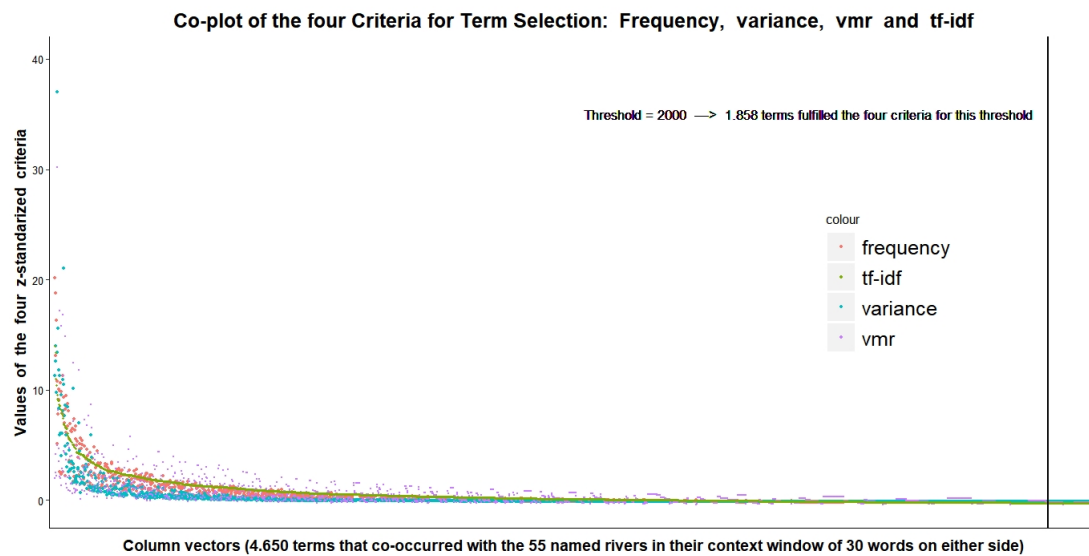


Figure 4 Co-plot of the 4 statistical criteria for term selection in the case of the named rivers.

document frequency (tf-idf) were retained. Figure 4 shows the co-plot of the four criteria for the rivers in descending order of magnitude. A threshold of 2000 was set. This meant that only 1,858 column terms fulfilled all criteria for the rivers. In the case of the bays, a threshold of 1000 was set, and only 847 column terms fulfilled thus all criteria.

Accordingly, in the case of the rivers, a reduced matrix of $1,913 \times 1,913$ dimensions was obtained (1,858 terms plus 55 named rivers). For the bays, the reduced matrix consisted of 876×876 dimensions (847 terms plus 29 named bays). Both matrices were then subjected to two weighting schemes. First, the statistical log-likelihood measure calculated the association score between all term pairs, since it captures syntagmatic and paradigmatic relations ([4]; [18]) and achieves better performance for small-sized corpora ([1]). Secondly, the scores were transformed by applying natural logarithm to reduce skewness ([18]).

3.2.5 Clustering of named rivers and bays

A hierarchical clustering technique was applied to both weighted, reduced matrices, using cosine distance as the intervector distance measure, and Ward's Method as the clustering algorithm.

Since it is not clear how strong a cluster is supported by data, a means for assessing the certainty of the existence of a cluster in corpus data was devised. For this, probability values (*p*-values) were computed for each hierarchical cluster using multiscale bootstrap resampling, implemented in the R package *pvclust* ([32]). For the rivers, thirteen groups with *p*-values higher than 95% were strongly supported by corpus data, as marked by the red rectangles in the dendrogram in Figure 5. For the bays, Figure 6 shows five clusters.

3.2.6 Terms characterizing each cluster

To ascertain the terms strongly associated with each of the clusters, the following procedure was used:

1. For each of the named rivers and bays in their corresponding clusters, a set of the top-30 terms, most semantically related to each, was extracted from the corresponding DSM

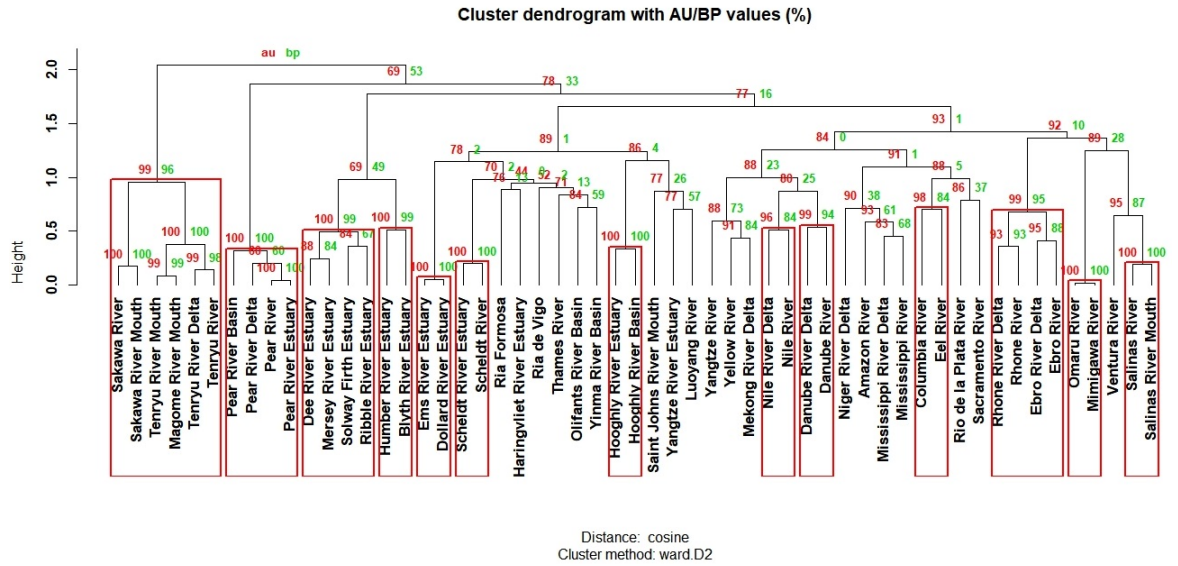


Figure 5 Dendrogram of the hierarchical clustering of the 55 named rivers with 13 clusters.

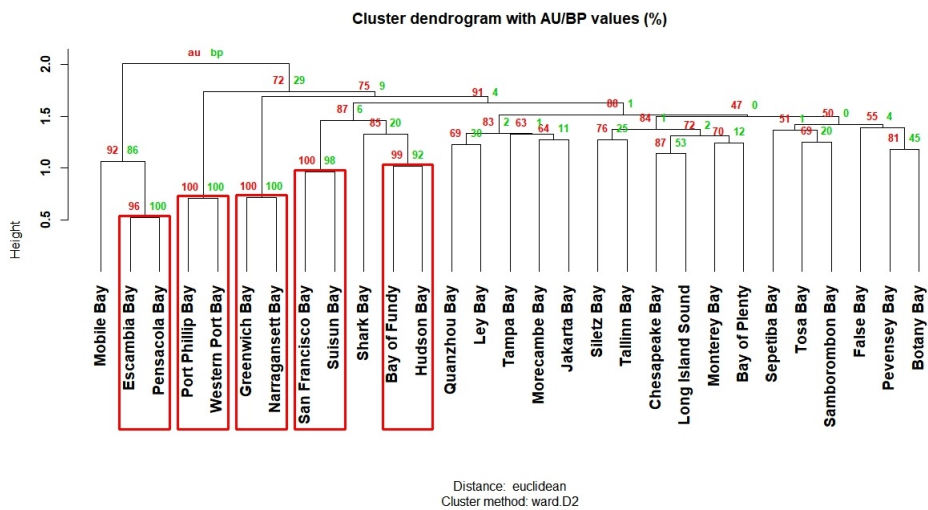


Figure 6 Dendrogram of the hierarchical clustering of the 29 named bays with 5 clusters.

using cosine similarity.

2. For each cluster, the mathematical operation *set intersection* was applied to the sets of the top-30 terms, most semantically related to the rivers and bays in the same cluster. Only the shared terms with a cosine similarity higher than 0.55 were selected.

A reduced set of terms was thus obtained for each cluster to describe the named rivers and bays.

4 Results

Because of space constraints, only the results for some clusters are provided. Numbering the clusters in Figures 5 and 6 from left to right, the second and twelfth cluster for rivers (Figure 5), and the fourth cluster for bays (Figure 6) are described. As shown in Figure 5, the second cluster is formed by the basin, delta, and estuary of the *Pearl River*, and the river itself, located in China. The *Omaru* and the *Mimigawa* rivers, placed in Japan, comprise the twelfth cluster. The fourth cluster in Figure 6 consists of the *San Francisco* and the *Suisun* bays, in California (the USA). These clusters were selected because the named rivers and bays, despite being different landforms and located in different world areas, are related to the same topic as explained in the following subsections, namely, the sediment concentration and sediment transport in rivers, sediment discharge into bays and seas, and the negative effects of sediment supply decrease on coastal erosion because of human activities.

For the description of the semantic networks, the semantic relations were manually extracted by querying the corpus in Sketch Engine ([17]), and analysing knowledge-rich contexts, namely, a context indicating at least one item of domain knowledge that could be useful for conceptual analysis ([22]). The query results were concordances of any elements between the river/bay in a cluster and related terms in a ± 40 span. The semantic relations were those in EcoLexicon ([13]), with the addition of *supplies*, *prevents*, *accumulates_in*, *simulates*, *tributary_of*, *increases*, *decreases*, *belongs_to*, *uses*, and *instance_of*, necessary for the explanatory adequacy of the frames ([10]). Furthermore, the semantic frames shown in the following were validated by Coastal Engineering experts from the University of Granada (Spain).

4.1 Second cluster in Figure 5: *Pearl River*

Predicting *sediment load* in a river system have long been a goal of earth scientists for numerous reasons, including alternation of fish habitats, changes in the load from anthropogenic effects, and the evolution of deltas, estuaries, and coastal environments. Hence, hydrologists have made efforts in applying *sediment rating curves* that can empirically describe the relationship between *suspended sediment concentration* (g/km^3) and *water discharge* (m^3/s) for a certain location. In *sediment rating curves*, *sediment rating parameters* also intervene, which are often associated with *river bed morphology* and *soil erodibility*. Engineers use *sediment rating curves* for predicting the life span of a *dam* on a river, and earth scientists use them to study the erosional and depositional environments.

Dam and *reservoir construction* are regarded as the main cause of the decline in *sediment load*. For that reason, the issue of *sediment load* in the *Pearl River Delta* was studied. Attention was paid to the *sediment rating parameters* of the *sediment rating curves*. The parameters reflected a temporal relationship between *water discharge* and *suspended sediment concentration* due to *human activities*, such as *land use* and *reservoir construction*. These

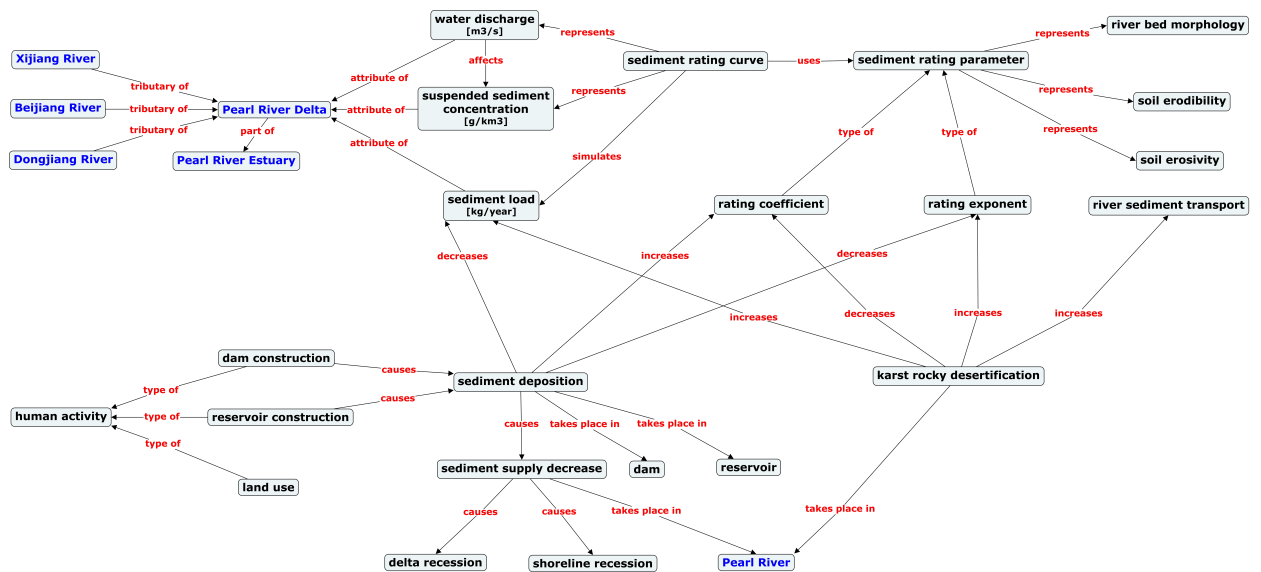


Figure 7 Semantic network of the terms associated with the *Pearl River*.

activities are causing a decrease in sediment supply from the *Pearl River*, with grave consequences on the coast (see Figure 7).

4.2 Twelfth cluster in Figure 5: *Omaru River* and *Mimigawa River*

Owing to the interruption of sediment flow at dams, the degradation of riverbed was observed on the downstream of the *Omaru*, *Mimigawa*, *Hitotsuse* and *Ooyodo* rivers. Sediment discharge through these four rivers is thus considered to decrease considerably, causing *coastal erosion* on the *Miyazaki Coast*. The *Sumiyoshi Beach*, located on this coast, is thus a severe eroded beach because of the decrease in sediment supply from the four rivers, and the blocking of *longshore sand transport* by the *breakwater* of the *Miyazaki Port* (see Figure 8).

4.3 Fourth cluster in Figure 6: *San Francisco Bay* and *Suisun Bay*

San Francisco and *Suisun* bays are involved in research studies to determine whether the timescale dependence of *forcing mechanisms* on *suspended sediment concentration* (SSC) is typical in bays and estuaries, based on *SSC data*. Of the *forcing mechanisms*, several tend to be concurrently active in bays and estuaries, rather than only one. Multiple active *forcing mechanisms* have been observed in bays and estuaries, but the specific mix of active mechanisms is different in each. It poses the question whether named estuaries and bays should be considered either instances of the ESTUARINE WATER concept or concepts for themselves (see Figure 9).

5 Conclusions

To extract knowledge for the semantic frames or conceptual structures ([12]) that underlie the usage of named rivers and bays in Coastal Engineering texts, a semi-automated method for the extraction of terms and semantic relations was devised. The semantic relations linking concepts in the semantic frames were manually extracted by querying the corpus in Sketch Engine, and analysing knowledge-rich contexts. It was a time-consuming task, although

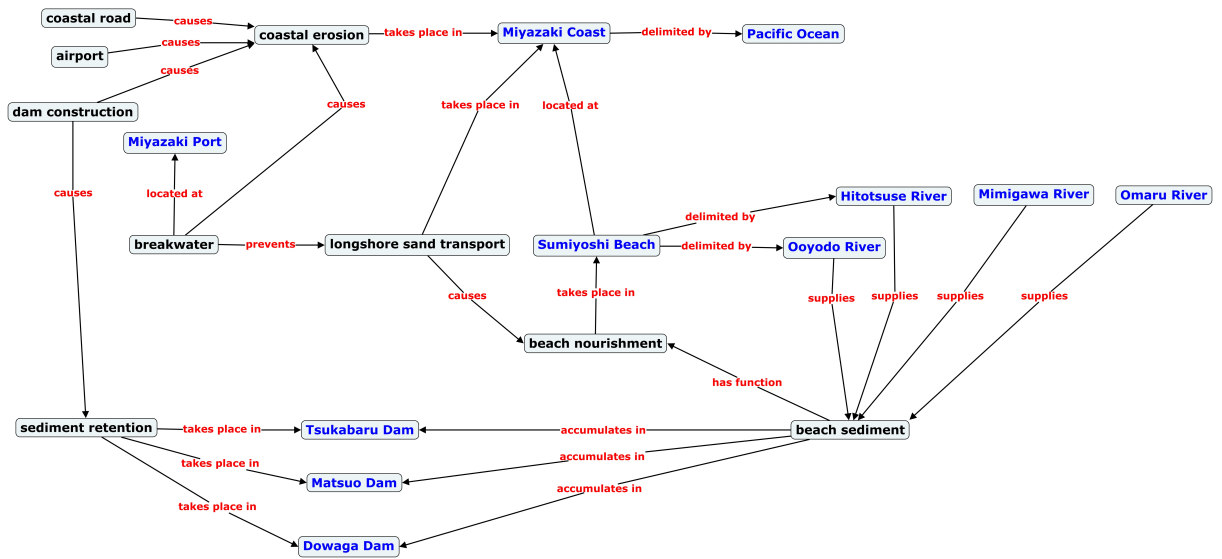


Figure 8 Semantic network of the terms associated with *Omaru* and *Mimigawa* rivers.

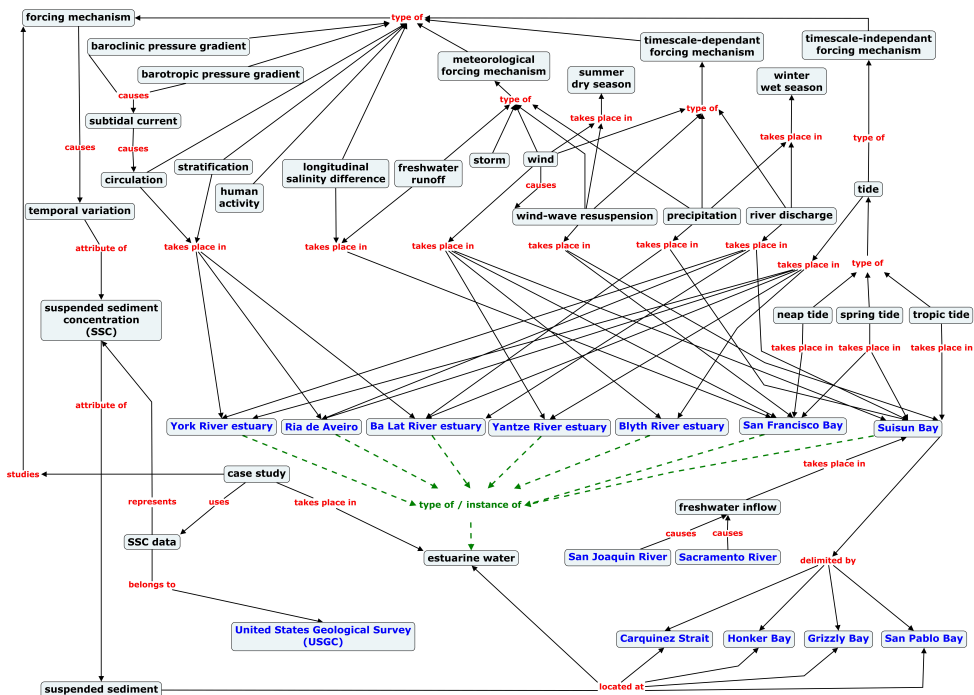


Figure 9 Semantic network of the terms associated with *San Francisco* and *Suisun* bays.

essential for the explanatory adequacy of frames ([10]). In future research, the knowledge patterns by [20] for the automatic extraction of semantic relations will be tested.

The method for the extraction of terms closely associated with named rivers and bays combined, on the one hand, the use of a count-based DSM, weighted by the log-likelihood association measure, and on the other hand, a selection procedure for terms based on four statistical criteria. Although this term selection procedure offered successful results to construct the semantic frames, Topic Modelling ([6]), a domain-specific dimension reduction technique for texts, will be also applied, and a comparison of both methods will be carried out.

The semantic frames in the previous section reflect that most terms related to named rivers and bays are multi-word terms (MWT) since specialized language units are mostly represented by such compound forms ([27]). The MWT extraction was possible because they were previously matched and joined by means of underscoring in the lemmatized corpus, thanks to the list of MWTs stored in EcoLexicon. This confirms that EcoLexicon is a valuable resource for any natural language processing tasks related to specialized corpora on environmental science.

Finally, the conceptual structures also highlighted that Coastal Engineering texts attach great importance to the study of the processes that each named river triggers, the processes that affect a certain named river, the crucial role that a named river plays to prevent coastal erosion, and the close relation between rivers and bays in sediment concentration and transport. On the evidence of these findings supporting our working hypothesis, it would be more appropriate for named rivers and bays in the Coastal Engineering domain to be considered concepts for themselves rather than mere instances of the RIVER and BAY concepts to be semantically represented in terminological resources.

References

- 1 Maha Arabia, Nawal Alhelewh, Abdul Malik Al-Salman, and Eric Atwell. An empirical study on the holy quran based on a large classical arabic corpus. *International Journal of Computational Linguistics*, 5(1):1–13, 2014.
- 2 Fatemeh Ars, Jon Willits, and Michael Jones. Comparing predictive and co-occurrence based models of lexical semantics trained on child directed speech. In *38th Annual Conference of the Cognitive Science Society, Austin, Texas, USA, August 10-13, 2016*, pages 1092–1097, 2016.
- 3 Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, MD, USA, June 22-27, 2014, vol. 1*, pages 238–247, 2014. doi:10.3115/v1/P14-1023.
- 4 Gabriel Bernier-Colborne and Patrick Drouin. Evaluation of distributional semantic models: a holistic approach. In *5th International Workshop on Computational Terminology (Computerm2016), Osaka, Japan, December 12, 2016*, pages 52–61, 2016.
- 5 Ann Bertels and Dirk Speelman. Clustering for semantic purposes: Exploration of semantic similarity in a technical corpus. *Terminology*, 20(2):279–303, 2014. doi:10.1075/term.20.2.07ber.
- 6 David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
- 7 John A. Bullinaria and Joseph P. Levy. Extracting semantic representations from word co occurrence statistics: A computational study. *Behavior Research Methods*, 39(3):510–526, 2007. doi:10.3758/BF03193020.
- 8 Curdin Derungs and Ross S. Purves. From text to landscape: locating, identifying and mapping the use of landscape features in a swiss alpine corpus. *International Journal of Geographical Information Science*, 28(6):1272–1293, 2014. doi:10.1080/13658816.2013.772184.

- 9 Stefan Evert. Corpora and collocations. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An International Handbook, vol. 2*, pages 1212–1248, Berlin, 2009. Mouton de Gruyter.
- 10 Pamela Faber. The cognitive shift in terminology and specialized translation. *MonTI. Monografías de Traducción e Interpretación*, 1:107–134, 2009.
- 11 Pamela Faber. The dynamics of specialized knowledge representation: Simulational reconstruction or the perception action interface. *Terminology*, 17(1):9–29, 2011.
- 12 Pamela Faber. *A Cognitive Linguistics View of Terminology and Specialized Language*. De Gruyter Mouton, Berlin/Boston, 2012.
- 13 Pamela Faber, Pilar León-Araúz, and Juan Antonio Prieto. Semantic relations, dynamicity, and terminological knowledge bases. *Current Issues in Language Studies*, 1:1–23, 2009.
- 14 Stefan Gries and Anatol Stefanowitsch. Cluster analysis and the identification of collexeme classes. In Sally Rice and John Newman, editors, *Empirical and experimental methods in cognitive/functional research*, pages 73–90, Stanford (California), 2010. CSLI.
- 15 Daniel Jurafsky and James H. Martin. Vector semantics. In *Speech and Language Processing. Draft of September 23, 2018*, pages 1–32, 2018.
- 16 Douwe Kiela and Stephen Clark. A systematic study of semantic vector space model parameters. In *2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, Gothenburg, Sweden, April 26-30, 2014, pages 21–30, 2014. doi:10.3115/v1/W14-1503.
- 17 Adam Kilgarriff, Pavel Rychlý, Pavel Smrz, and David Tugwell. The sketch engine. In *11th EURALEX International Congress, Lorient, France, July 6-10, 2004*, pages 105–115, 2004.
- 18 Gabriella Lapesa, Stefan Evert, and Sabine Schulte im Walde. Contrasting syntagmatic and paradigmatic relations: Insights from distributional semantic models. In *3rd Joint Conference on Lexical and Computational Semantics (SEM'2014)*, Dublin, Ireland, August 23-24, 2014, pages 160–170, 2014. doi:10.3115/v1/S14-1020.
- 19 Pilar León-Araúz, Arianne Reimerink, and Pamela Faber. Multidimensional and multimodal information in ecollexicon. In Adam Przepiórkowski, Maciej Piasecki, Krzysztof Jassem, and Piotr Fuglewicz, editors, *Computational Linguistics. Studies in Computational Intelligence, vol. 458*, pages 143–161, Berlin/Heidelberg, 2013. Springer. doi:10.1007/978-3-642-34399-5_8.
- 20 Pilar León-Araúz, Antonio San Martín, and Pamela Faber. Pattern-based word sketches for the extraction of semantic relations. In *5th International Workshop on Computational Terminology (Computerm2016)*, Osaka, Japan, December 12, 2016, pages 73–82, 2016.
- 21 Pilar León-Araúz, Antonio San Martín, and Arianne Reimerink. The ecollexicon english corpus as an open corpus in sketch engine. In *18th EURALEX International Congress, Ljubljana, Slovenia, July 17-21, 2018*, pages 893–901, 2018.
- 22 Ingrid Meyer. Extracting knowledge-rich contexts for terminography: A conceptual and methodological framework. In Didier Bourigault, Christian Jacquemin, and Marie Claude L’Homme, editors, *Recent Advances in Computational Terminology*, pages 279–302, Amsterdam/Philadelphia, 2001. John Benjamins.
- 23 Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations (ICLR)*, Scottsdale, Arizona, USA, May 2-4, 2013, 2013. doi:arXiv:1301.3781v3.
- 24 George Miller and Walter Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28, 1991. doi:10.1080/01690969108406936.
- 25 Hermann Moisl. *Cluster Analysis for Corpus Linguistics*. De Gruyter Mouton, Berlin, 2015.
- 26 Susanne Moskalski and Raymond Torres. Influences of tides, weather, and discharge on suspended sediment concentration. *Continental Shelf Research*, 37:36–45, 2012. doi:10.1016/j.csr.2012.01.015.
- 27 Preslav Nakov. On the interpretation of noun compounds: Syntax, semantics, and entailment. *Natural Language Engineering*, 19(3):291–330, 2013. doi:10.1017/S1351324913000065.

- 28 Patrick Pantel and Dekang Lin. Discovering word senses from text. In *ACM Conference on Knowledge Discovery and Data Mining (KDD-02)*, Edmonton, Canada, July 23-26, 2002, pages 613–619, 2002. doi:10.1145/775047.775138.
- 29 Douglas Rohde, Laura Gonnerman, and David Plaut. An improved model of semantic similarity based on lexical co occurrence. . *Communications of the ACM*, 8:627–633, 2006.
- 30 Magnus Sahlgren and Alessandro Lenci. The effects of data size and frequency range on distributional semantic models. In *2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas, USA, November 1-5, 2016*, pages 975–980, 2016. doi:10.18653/v1/D16-1099.
- 31 Ekaterina Shutova, Lin Sun, and Anna Korhonen. Metaphor identification using verb and noun clustering. In *23rd International Conference on Computational Linguistics, vol. 2, Beijing, China, August 23-27, 2010*, pages 1002–1010, 2010.
- 32 Ryota Suzuki and Hidetoshi Shimodaira. Pvcust: An r package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, 22(12):1540–1542, 2006. doi:10.1093/bioinformatics/btl117.
- 33 Flurina M. Wartmann, Elise Acheson, and Ross S. Purves. Describing and comparing landscapes using tags, texts, and free lists: an interdisciplinary approach. *International Journal of Geographical Information Science*, 32(8):1572–1592, 2018. doi:10.1080/13658816.2018.1445257.