Title
Development of city buildings dataset for urban building energy modeling

Permalink
https://escholarship.org/uc/item/21x4121t

Authors
Chen, Yixing
Hong, Tianzhen
Luo, Xuan
et al.

Publication Date
2019

DOI
10.1016/j.enbuild.2018.11.008

Development of City Buildings Dataset for Urban Building Energy Modeling

Yixing Chen[1], Tianzhen Hong[1, *], Xuan Luo[1], Barry Hooper[2]

[1] Building Technology and Urban Systems Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[2] Department of Environment, City of San Francisco, California, USA

* Corresponding author (T. Hong). thong@lbl.gov; 1(510) 486-7082

Abstract

Urban building energy modeling (UBEM) is becoming a proven tool to support energy efficiency programs for buildings in cities. Development of a city-scale dataset of the existing building stock is a critical step of UBEM to automatically generate energy models of urban buildings and simulate their performance. This study introduces data needs, data standards, and data sources to develop city building datasets for UBEM. First, a literature review of data needs for UBEM was conducted. Then, the capabilities of the current data standards for city building datasets were reviewed. Moreover, the existing public data sources from several pioneer cites were studied to evaluate whether they are adequate to support UBEM. The results show that most cities have adequate public data to support UBEM; however, the data are represented in different formats without standardization, and there is a lack of common keys to make the data mapping easier. Finally, a case study is presented to integrate the diverse data sources from multiple city departments of San Francisco. The data mapping process is introduced and discussed. It is recommended to use the unique building identifiers as the common keys in the data sources to simplify the data mapping process. The integration methods and workflow are applied to other U.S. cities for developing the city-scale datasets of their existing building stock, including San Jose, Los Angeles, and Boston.

Keywords

City building dataset, CityGML, Urban building energy modeling, data standards, data mapping

1. Introduction

Buildings in cities of the United States consume up to 70% of primary energy. Reducing energy use of building stock in cities becomes a critical strategy to achieving cities' energy and environmental goals. The City of San Francisco has established some of the most competitive climate and sustainability targets in the world, covering a broad range of sectors, including energy efficiency, renewable energy, transportation, water, green infrastructure, and waste. With robust goals to measure progress, San Francisco aims to reduce greenhouse gas (GHG) emissions by 25% below 1990 levels by 2017, 40% by 2025, and 80% by 2050 [1]. San Francisco has been making great progress towards its ambitious GHG emission reduction goal. By 2015, San Francisco's GHG emission was 28.4% below 1990 levels, equivalent to 1.8 million metric tons of carbon dioxide equivalent (CO_2e) emission ($mtCO_2e$) reduction. San Francisco has approximately 180,000 buildings, which contribute to 52% of the city's total GHG emissions [2]. The building sector holds great potential to reduce energy use and GHG emissions through the proliferation of new, energy efficient buildings and by retrofitting existing buildings. The building sector's 2015 GHG emissions were reduced by 38%, or 1.3 million $mtCO_2e$ compared to the 1990 level, which contributed to 73% of San Francisco's total GHG emission reduction.

San Francisco provides various incentive and financing programs to help residents and building owners save investment and operating costs, minimize energy waste, and lower their property's environmental impact [3]. San Francisco's Energy Watch program [4], supported by local utility company Pacific Gas and Electric, offers incentives to commercial and multifamily buildings for energy efficiency upgrades to lighting, refrigeration equipment and controls, network-level computer power management software and so on. San Francisco’s Property Assessed Clean Energy (PACE) financing program [5] helps homeowners finance energy-saving, renewable energy and water-saving home upgrades. GoSolarSF [6], managed by the San Francisco Public Utilities Commission, provides cash incentives for installing eligible solar electric systems. The Energy Upgrade California Multifamily Program [7] in San Francisco offers $750 per unit in rebates to help multifamily property owners (5+ units) lower the cost of energy efficiency upgrades. Those incentive and financing programs contribute significantly to GHG reductions in San Francisco’s buildings sector; however, they are mainly implemented at the individual building level, which limits their broad adoption and requires a significant amount of staff effort to manage the programs. The incentive and financing programs should be analyzed and implemented on a larger scale to boost the energy renovation rate of the building stock. Future programs should consider not only the technologies for individual buildings but also the opportunities of district-scale technologies, such as district heating and cooling systems, combined heat and power systems, and community-scale photovoltaic (PV) systems.

Urban building energy modeling (UBEM) refers to the application of bottom-up physics-based building energy models to predict operational energy use, as well as indoor and outdoor environmental conditions, for groups of buildings in the urban context [8]. UBEM is an excellent tool to explore opportunities for energy conservation measures (ECMs) when applying to a large group of buildings in the urban context. Delmastro et al. [9] leveraged UBEM to aid decision-makers in the planning process by simulating and analyzing the evolution of the building stock from an energetic, economic, and social perspective over long-term horizons. In particular, their approach: (1) identified the cost-optimal mix of successful renovation packages; (2) identified buildings that need to be prioritized; and (3) considered the impact of socioeconomic factors on policies implementation. Chen et al. [10] presented a case study using UBEM to analyze the potential energy and cost savings of five individual ECMs and two measure packages for 940 office and retail buildings in San Francisco. UBEM can also be used to evaluate the district-scale technologies. Yamaguchi et al. [11] presented a simulation model based on the bottom-up UBEM approach to evaluating different technology implementation scenarios, including distributed electricity generators and district heating and cooling systems.

UBEM is becoming a proven tool to support energy efficiency programs for buildings in cities. Development of a city-scale dataset of the existing building stock is a critical step of UBEM to automatically generate energy models of urban buildings and simulate their performance. Monteiro et al. [12] presented the process of collecting, mapping, cleaning, and integrating data to create an urban building dataset for 3,259 buildings with 18,484 residential dwellings and 33,659 inhabitants to support an information system for smart cities. Davila et al. [13] collaborated with the Boston Redevelopment Authority to develop a citywide UBEM based on official GIS datasets and a custom building archetype library for 83,541 buildings.

More and more cities in the world are moving to provide open data via web portals to empower their use to support cities' energy and environmental goals. For example, San Francisco's open data portal [12] provides geographic information system (GIS) building geometry information, including the footprint and height of each building in San

Francisco. It also includes building characteristics, such as year built, number of stories, and building type. Similar building data can be found in other cities, such as Chicago [13] and New York City [14].

Cities are the main sources to provide the input data for UBEM and the major adopters of UBEM tools in the future. Cities spend lots of effort to collect the data and make them publicly available. However, those data are not collected specifically for UBEM and some important information for UBEM may be ignored. For example, San Francisco provides the permit database to record the changing history of buildings; however, that information is presented in "text" format without standardized description, which makes them less useful to support UBEM. It is very important to make sure that cities are collecting enough data in a standardized format to support UBEM in the future.

This study first conducts a literature review to understand the data needs for current UBEM studies and the current data standards to represent those city building datasets. It then studies the status of the public building data sources from several pioneer cities in the United States to answer three questions: (1) Are the existing public data from cities adequate to support UBEM? (2) Are there easy ways to integrate those diverse data sources? (3) How to standardize the data for interoperability? Finally, a case study is presented to develop a standardized city building dataset for San Francisco by integrating publically available buildings datasets from multiple city departments.

2. Data Needs for UBEM

Reinhart and Davila [8] reviewed emerging simulation methods and implementation workflows for UBEM. The data inputs for UBEM were also discussed, which included the climate data and the building data. The climate datasets in the typical meteorological year (TMY) format for building performance simulation are widely available for more than 2100 cities worldwide [15]. This study focuses on the building data for UBEM, including the geometry data and the non-geometric properties. A literature review was conducted to understand the building data used to model the energy performance of building stocks. Table 1 provides a summary of the building data organized into three categories: geometry, segmentation parameters, and energy use data. For the geometry data, cases 1 to 8 used the GIS-based building footprint, building height and the number of stories to create the building geometry for each building. Case 8 derived the number of stories based on the building height. Case 9 used the total floor area to scale the rectangular box geometry. Cases 10 to 17 used the total floor area to scale the energy performance results.

None of the studies has the detailed information about the building systems and their efficiencies. Instead, the information is assumed based on the archetype. Several segmentation parameters are used to identify the archetypes, including the age (year built), use type, and heating type. The shape/size of the building derived from the geometry is also used in several studies as segmentation parameters.

Energy data was available for several studies, typically at the annual resolution. In additional, several studies require more information of the segmentation parameters. Cases 2, 3, and 9 require the number of stories above ground as well as the number of stories below ground (basement). Cases 2, 9, and 16 use the heated floor area while the other cases use the total floor area. Cases 1 and 2 need both the year of construction and the year of refurbishment.

In summary, the building data needs for UBEM typically include the GIS footprint, building height, number of stories above ground, number of stories below ground, total floor area, heated floor area, number of dwellings, year of construction, year of refurbishment, use type (building type), heating system type, annual electricity use, and annual natural gas use.

Table 1. Summary of data needs for UBEM

Case ID	Each Building?*	Building sector**	Geometry					Segmentation			Energy Use			Other data	Reference
			Footprint	Building Height	No. of Stories	Floor area	No. of dwellings	Age (year built)	Use type	Heating type	Annual	Monthly	Time series		
1	Y	R	√	√	√	√		√	√	√		√			[16]
2	Y	R	√	√	√	√		√	√	√	√			***	[17]
3	Y	A	√	√	√		√	√	√	√				****	[18]
4	Y	C	√	√	√			√	√		√				[19]
5	Y	C	√	√	√			√	√						[10]
6	Y	A	√	√	√			√	√						[20]
7	Y	A	√	√	√			√	√						[21]
8	Y	A	√	√				√	√	√	√			*****	[22]
9	Y	R			√	√		√	√	√			√		[23]
10	N	R			√	√	√	√		√					[24]
11	N	R		√	√	√		√	√						[25]
12	N	R				√		√	√						[26]
13	N	A				√		√	√	√	√				[27]
14	N	R				√	√	√	√						[28]
15	N	A				√		√	√						[29]
16	N	A				√		√	√	√					[30]
17	N	A	√	√		√		√	√						[31]

Note:
* Model each building or not: Y – Yes, N – No
** R – Residential, C – Commercial, A – All (Residential & Commercial)
*** Number of staircases, attachment to other buildings
**** Number of persons per building, volume, type of hot water supply
***** Measured heat demand at the substations

3. Data standards for city building datasets

More and more cities in the world are moving to provide open data via web portals to empower their use to support cities' energy and environmental goals. However, there is a lack of consistency, semantics, and standards among the shared data to enable interoperability for various types of urban applications. For San Francisco, the building GIS-based footprint data are provided in the Shapefile format, while the building characteristics are stored in multiple files with Shapefile, fixed-width text, or comma-separated values (CSV) format. Moreover, different terms are used to represent the same data elements among different datasets. Table 2 lists some of the terms used for the same data elements in the building datasets from San Francisco, Chicago, and Portland. In addition, the same data element in different datasets may represent slightly different things. For example, in Table 2, the building height in San Francisco dataset is the median value of the building height; while the building height in Portland dataset is the average value of the building height.

Table 2. Different terms used for the same data elements among different buildings datasets in three U.S. cities: San Francisco, Chicago, and Portland

Terms	**San Francisco**	**Chicago**	**Portland**
Building Type	LANDUSE	Property classification	BLDG_USE
Year Built	YRBUILT	Year_built	YEAR_BUILT
Number of Floors	STOREYNO	Stories	NUM_STORY
Building Height	gnd1st_delta_m	N/A	AVG_HEIGHT

It is essential to gather building asset data at the city scale from a wide range of sources (e.g., surveys, city projects, city datasets, and public records) and assemble them into a single open database with standardized formats and terms. The primary data formats to support UBEM include Shapefile/FileGDB, GeoJSON, and CityGML. The ESRI Shapefile [32] and FileGDB [33] formats are popular geospatial vector data format used by GIS software tools. They typically include two-dimensional (2D) GIS-based building footprint information and a table of building properties or attributes. GeoJSON [34] is a data format based on JSON (JavaScript Object Notation) for encoding a variety of 2D GIS data structures, which is friendly to web applications built upon JavaScript. However, the Shapefile/FileGDB and GeoJSON formats do not provide a schema to define the building properties, leading to inconsistency among different datasets.

Building Energy Data Exchange Specification (BEDES) [35], developed by the U.S. Department of Energy (DOE) and Lawrence Berkeley National Laboratory (LBNL), is a dictionary of terms and definitions commonly used in tools and activities that help stakeholders make energy investment decisions, track building performance, and implement energy efficiency policies and programs. BEDES provides common terms and definitions for building energy data, which different tools, databases, and data formats can share. More than 50 projects, programs, and applications are involved in the development of BEDES. Table 3 shows the BEDES terms for the terms used in the literature for UBEM. For city building data in FileGDB or GeoJSON format, BEDES can be used to provide more standardized terms.

Table 3. BEDES terms for the terms used in the literature

Terms used in the literature	**BEDES terms**
Building height	Building Height
Number of stories above ground	Above Grade Floor Quantity
Number of stories below ground	Below Grade Floor Quantity
Total floor area	Gross Floor Area
Heated floor area	Heated Gross Floor Area
Number of dwellings	Apartment Unit Quantity

Year of construction	Completed Construction Status Date
Year of refurbishment	Completed Major Remodel Date
Use type (building type)	Occupancy Classification
Heating system type	Heating Type
Annual electricity use	Annual Electricity Resource Value
Annual natural gas use	Annual Natural Gas Resource Value
Annual site energy use	Annual Site Energy Resource Value
Annual source energy use	Annual Source Energy Source Value

CityGML is an international Open Geospatial Consortium (OGC) standard that provides an open data model to represent and exchange digital three-dimensional (3D) models of cities and landscapes [36,37]. Many UBEM projects selected CityGML as the data model to represent and exchange 3D city models, especially for European research projects. CityGML was used to represent the semantic 3D city for predicting the photovoltaic potential and heating energy demand of urban districts [38] and analyzing strategies for improving building standards [39]. TEASER, an open framework for urban energy modeling of building stocks, includes a ready-to-use interface for CityGML [40]. The Open Source City Database (CityDB) is a flexible framework to create and run city-scale building energy simulations with the building datasets in CityGML or GeoJSON formats [41]. City Building Energy Saver (CityBES) [42,43], developed by LBNL, is a web-based data and computing platform, focusing on energy modeling and analysis of the building stock of a city to support district or city-scale building energy efficiency programs. CityBES accepts building stock data in both CityGML and GeoJSON formats.

CityGML defines the 3-D geometry, topology, semantics, and appearance of urban objects, including buildings and their components, bodies of water, city furniture (street lighting, traffic lights), transportation infrastructure (streets, roads, bridges, tunnels), and vegetation. Figure 1 shows some examples of CityGML objects. For many of these attributes describing 3-D city models, CityGML provides its standard external code list enumerating the values for each attribute type, such as standard lists of land use type (LandUseClassType) and building usage type (BuildingUsageType).

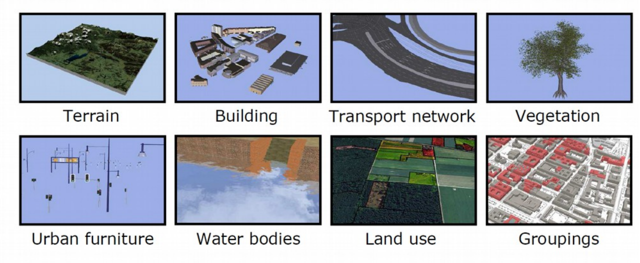


Figure 1. Examples of CityGML objects [44]

CityGML enables flexible representation of objects at various levels of detail, which is critical as data availability varies widely for a large number of buildings and other urban infrastructure. Figure 2 shows a building can be represented at five levels of details: a simple 2-D footprint, a box shape, adding slope roofs, adding exterior shades and windows and doors, and full details of interior layout and zoning. CityGML version 1.0 was released in 2008, and an extended version 2.0 was adopted in March 2012.

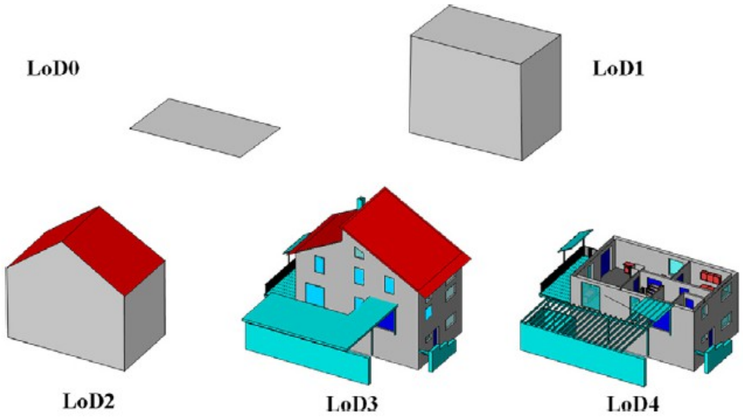


Figure 2. Five levels of details (LODs) to represent a building in CityGML [36]

CityGML has the concept of Application Domain Extension (ADE) to model user-defined objects and attributes. The CityGML Energy ADE extends the CityGML Standard by features and properties, which are necessary to perform an energy simulation and to store the corresponding results [45]. Table 4 listed the mapping of the terms to the standardized CityGML and Energy ADE elements. Several terms are straightforward, including building height, number of stories above ground, number of stories below ground, total floor area, heated floor area, year of construction, and use type (building type). Some terms are not available in CityGML or Energy ADE, as it requires the detailed systems information, including number of dwellings, year of refurbishment, and heating system type. The EnergyDemand element in the Energy ADE is designed for time series data. Although the EnergyDemand element can be used to represent the annual electricity and natural gas use, it is too tedious. Moreover, the EnergyDemand element cannot cover the annual site and source energy use.

Table 4. CityGML elements for the terms used in the literature

Terms used in the literature	**CityGML and Energy ADE examples**
Building height	<bldg:measuredHeight uom="m">6.52</… >
Number of stories above ground	<bldg:storeysAboveGround>2</…>
Number of stories below ground	<bldg:storeysBelowGround>0</…>

Total floor area	<energy:FloorArea>
	  <energy:type>grossFloorArea</…>
	  <energy:value uom="m2">240</…>
	</energy:FloorArea>

Heated floor area	<energy:FloorArea> <energy:type>energyReferenceArea</…> <energy:value uom="m2">240<…> </energy:FloorArea> *Note: energyReferenceArea is referred as heated or cooled area in some European reports.*
Number of dwellings	Not available, need to specify each unit/dwelling
Year of construction	<bldg:yearOfConstruction>2010</…>
Year of refurbishment	Not available, need to specify the energy conservation measures
Use type (building type)	<bldg:usage>1000</…> *Note: code 1000 is for "residential building". The codes*

	are defined in the BuildingUsageType.xml, according to the dictionary concept of GML3.
Heating system type	Not available, need to specify the heating system

Annual electricity use	<energy:EnergyDemand gml:id="…">
	  <energy:energyAmount>
	    <energy:RegularTimeSeries>
	      <energy:variableProperties>
	        <energy:TimeValuesProperties>
	          <energy:acquisitionMethod>measurement</…>
	          <energy:interpolationType>succeedingTotal</…>
	        </energy:TimeValuesProperties>
	      </energy:variableProperties>
	      <energy:temporalExtent>
	        <gml:TimePeriod>
	          <gml:beginPosition>2017-01-01T00:00:00</… >
	          <gml:endPosition>2017-12-31T23:00:00</…>
	        </gml:TimePeriod>
	      </energy:temporalExtent>
	      <energy:timeInterval unit="year">1</…>
	      <energy:values uom="kWh">24000</…>
	    </energy:RegularTimeSeries>
	  </energy:energyAmount>
	  <energy:endUse>otherOrCombination</…>
	  <energy:energyCarrierType>electricity</…>
	</energy:EnergyDemand>

Annual natural gas use	Similar to Annual electricity use. Change the “electricity” to “naturalGas” in the energy:energyCarrierType element.
Annual site energy use	Not available
Annual source energy use	Not available

4. City Building Data Sources

Many cities in the United States provide public building data to support building energy efficiency programs and research. This section reviews the public data sources provided by six cities to check whether those data are adequate to support UBEM. Table 5 shows several public building data sources for the six cities, including San Francisco (SF), Chicago (CHI), Los Angeles (LA), Boston (BOS), San Jose (SJ), and Portland in Oregon (PDX). The public building data are typically provided in Shapefile or GeoJSON format when the building or parcel footprint data are available. The building characteristic data are typically stored in CSV format. The detailed data mapping among different data sources is introduced in Section 5.

Table 5. Public building data sources for six U.S. cities

City	Data source name	File format	Records	Primary key for mapping
San Francisco, CA (SF)	Building Footprints (BF)	Shapefile	177023	Building footprint
	Land Use (LU)	Shapefile	155468	Parcel ID, parcel footprint
	Assessor Record (AR)	Fix-width text	207850	Parcel ID
	Energy Benchmarking (EB)	CSV	1630	Parcel ID
Chicago, IL (CHI)	Building Footprints	GeoJSON	820606	Building ID, building footprint
	Energy Benchmarking	CSV	2718	Building ID
	Assessor Record	Website	165752	Parcel footprint
Los Angeles, CA (LA)	Building Footprints	Shapefile	1122422	Building ID, Assessor ID
	Assessor Record	CSV	2397615	Assessor ID
	Energy Benchmarking	CSV	6489	Building ID
Boston, MA (BOS)	Building Footprints	Shapefile	129370	Building footprint, building ID
	Property Assessment (PA)	CSV	172841	Parcel ID
	Energy Benchmarking	CSV	1800	Building ID
San Jose, CA (SJ)	Building Footprints	Shapefile	324217	Building footprint, parcel ID
	Zoning (ZO)	Shapefile	12295	Zoning district footprint
	Annexations (AN)	Shapefile	2370	Annexation footprint
	Assessor Record	CSV	106452	Parcel ID
Portland, OR (PDX)	Building Footprints	Shapefile	712334	Building ID
	Energy Benchmarking	CSV	410	Building ID

Table 6 shows the data availability to support UBEM of the six cities. All the cities have the data of building footprint, gross floor area, number of dwellings, year of construction, and building type. The Chicago datasets do not include the building height, while the number of stories information is missing in San Jose datasets. For UBEM, users can assume the floor-to-floor height to derive the building height or the number of stories from each other. The number of stories above ground, the number of stories below ground, and the heated floor area are missing in all the datasets.

Most of the cities have energy benchmarking data for a small portion of the buildings. The results show that most cities have adequate public data to support UBEM; however, the data are represented in different formats without standardization and there is a lack of common keys to map the data between datasets.

Table 6. Public building data sources to support UBEM

City	Building Footprint	Building Height	No. of Stories (total)	Gross floor area	No. of dwellings	Year of Construction	Year of Refurbishment	Use type (building type)	Heating system type	Annual electricity use	Annual natural gas use	Annual site energy use	Annual source energy use
SF	BF	BF	AR	LU, AR, EB	LU, AR	LU, AR	AR	LU, AR, EB				EB	EB
CHI	BF		BF, AR	BF, EB, AR	BF, AR	BF, EB, AR		EB, AR		EB	EB	EB	EB
LA	BF	BF		BF, AR, EB	AR	AR	AR	AR, EB		EB	EB	EB	EB
BOS	BF	BF	PA	BF, PA, EB	PA	PA, EB	PA	PA, EB	PA, EB	EB	EB	EB	
SJ	BF	BF		BF	ZO, AR	AN, AR	AR	ZO, AR					
PDX	BF	BF	BF	BF, EB	BF	BF, EB		BF, EB				EB	EB

Note: There are no data for the three fields: number of stories above ground, number of stories below ground, and heated floor area.

5. Case Study: Development of City Buildings Dataset for San Francisco

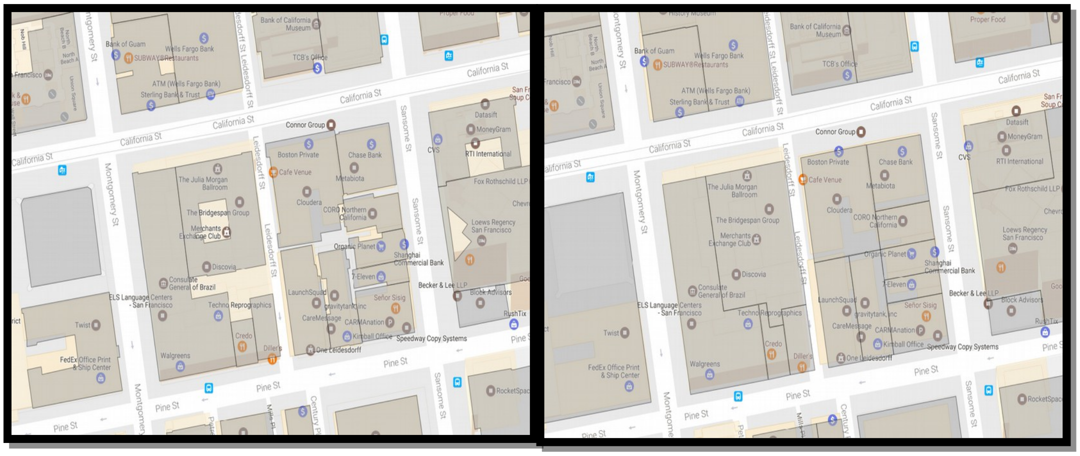
This section presents a case study to integrate the city building datasets from multiple city departments of San Francisco. A master dataset was created to include all the original data, while a simplification and standardization process was performed to produce the building dataset in various formats, including CityGML, GeoJSON, and FileGDB/Shapefile.

5.1. Data Sources

The city of San Francisco provides many public building datasets from multiple city departments, including Building Footprint data from the Department of Technology, Land Use data from the Department of Planning, Assessor Records from San Francisco County, and Energy Disclosure data from the Department of Environment.

5.1.1. Building footprint dataset

The Building Footprint dataset is available at the San Francisco's open data portal [12]. It includes the footprints of 177,023 buildings in San Francisco. Figure 3a shows a sample of the footprint data in gray. There are 43 attributes associated with each footprint polygon. The dataset includes multiple statistical attributes (the minimum, maximum, range, standard deviation, variety, minority, majority, and median) related to the altitude of ground and roof and the distances between the ground and the roof. The median value of the distance between the roof and the ground can be used as the building height.



(a) Building footprint (b) Parcel polygon

Figure 3. A sample of building footprint and parcel polygon data in San Francisco

5.1.2. Land use dataset

The Land Use dataset is also available at the San Francisco's open data portal [12]. There are 15 land use attributes associated with each parcel. The land use data records the address, the land use category, the building gross floor area, and the year built. However, those attributes are associated with the parcel information (Figure 3b) rather than the building footprint (Figure 3a).

5.1.3. Assessor recorder dataset

The Assessor Records dataset is maintained by the Office of the Assessor-Recorder [46]. The data can be viewed at the San Francisco's property information map portal [47]. There are 57 attributes associated with each assessor record, including the land value, personal property value, prior sales price, property usage type, number of stories, number of rooms (for residential), year built, and so on. As with the land use dataset, those attributes are associated with the parcel information rather than the building footprint (Figure 3).

5.1.4. Energy disclosure dataset

Passed in 2011, the San Francisco's Existing Commercial Buildings Energy Performance Ordinance, referred to as the energy disclosure dataset, requires annual energy benchmarking, periodic energy efficiency assessment, and public disclosure of benchmarking information for commercial buildings with 10,000 square feet (929 m^2) or more of heated and cooled space [48]. The energy disclosure data for 2010 to 2016 are available at San Francisco's open data portal [12]. It currently includes 1652 buildings. The address and parcel number of the energy ordinance results are available. The energy ordinance results for each building include the data from 2011 to 2016. Each ordinance result includes benchmark status, the reason for exemption, ENERGY STAR score, site and source energy use intensities (EUIs), percentage better than the national median site and source EUI, total GHG emissions, total GHG emission intensity, and weather-normalized site and source EUIs.

5.2. Data Mapping

The land use, assessor records, and energy disclosure databases use the Assessor Parcel Number (APN) as parcel identifiers to store the building data. We first consolidated the parcel-related data and mapped them with the building footprint data to create a master building dataset with all the fields/attributes from each dataset. Next, the master dataset was simplified and standardized to create 3-D city models for all the San Francisco buildings. BEDES was then used to standardize the terms in the building dataset. The final dataset products were produced in CityGML, GeoJSON, and FileGDB formats that can be used by various urban modeling and analysis tools.

5.2.1. Consolidating the parcel-related datasets

The three parcel-related datasets were stored in three different formats with separated metadata files in text or Microsoft Word documents. The parcel identifier appeared in each row of the three datasets as the key to mapping them. Figure 4 shows the workflow of the parcel-related dataset consolidation. The land use dataset was provided in the

Shapefile format, which includes both the parcel geometry and the related attributes. The land use dataset was first split using QGIS to create parcel geometry only and the land use-related attributes. QGIS [49] is a free and open source GIS tool. A script written in Ruby [50] was developed to merge the land use attributes in the CSV format, the energy disclosure in the CSV format, and the assessor records in a fixed-width text format. Finally, the merged attributes and the parcel geometry were joined together using QGIS to create the parcel-related dataset in the Shapefile format.

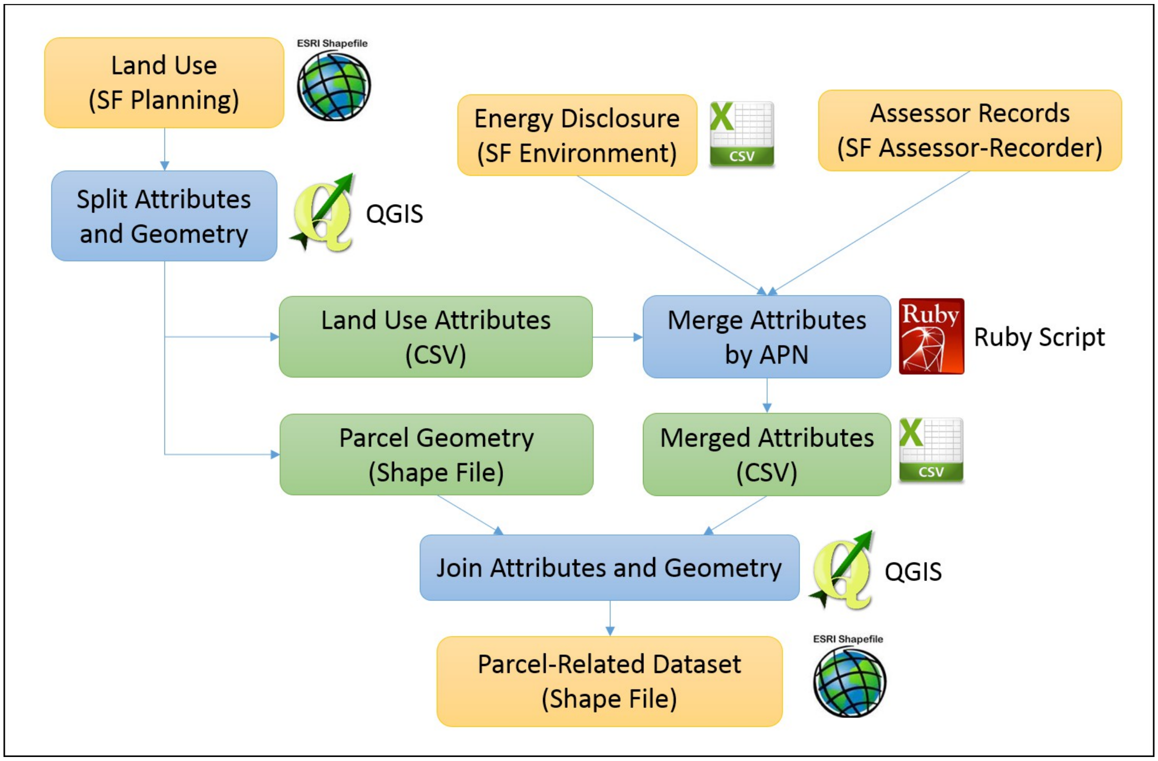
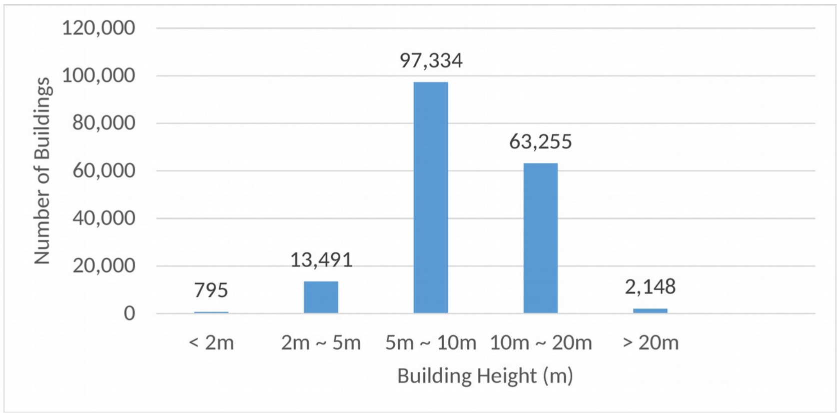


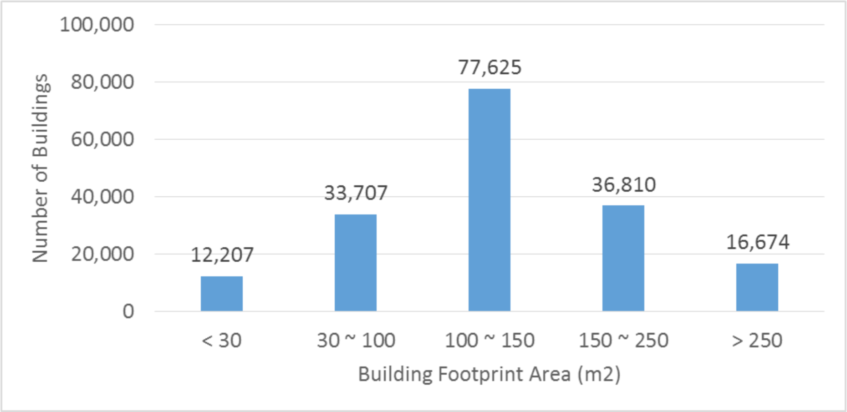
Figure 4. Workflow of parcel-related dataset consolidation

5.2.2. Mapping the building footprint with parcel polygon

There is no existing unique building identifier for different city departments to use to link their data directly with the buildings. The Pacific Northwest National Laboratory is currently working on a project to create unique building identifiers for all the buildings in the United States. Among the available data sources, most of the building-related information is associated with the parcel number. Therefore, it is necessary to map the building footprint with the parcel polygon to link the building datasets. One building footprint may overlap with multiple parcel polygons, while one parcel polygon may also overlap with multiple building footprints. It makes the mapping procedure complicated. There are 177,023 buildings in the San Francisco building footprint dataset. Figure 5 shows the distribution of their height and footprint area. We eliminated buildings with a lower than 2.5 m height and a floor area of less than 30 m^2, which resulted in 171,474 remaining buildings.



(a) Building Height



(b) Building Footprint Area

Figure 5. Building height and footprint area distributions of all San Francisco buildings

Two methods were used to map the building footprint with the parcel polygon. The first method is straightforward and uses the central point of a building to find the corresponding parcel polygon, which contains the building's central point. Using this method, we successfully found one parcel for each building. However, it may not be accurate when the building is overlapped with multiple parcels.

The second method is to do polygon clipping and find the overlap areas of the building with each parcel. We set the minimum overlap percentage to 10% of the building footprint area to eliminate those overlaps with small area due to the slight shifting in the data layer. Figure 6 shows the number of parcels per building using the polygon clipping method. It shows that 87.4% of the buildings belong to only one parcel, while 12.4% of the buildings are mapped with two parcels. Only 0.2% of the buildings are overlapped with more than two parcels. For the buildings overlapped with multiple parcels, we chose the parcel with the most significant overlap area.

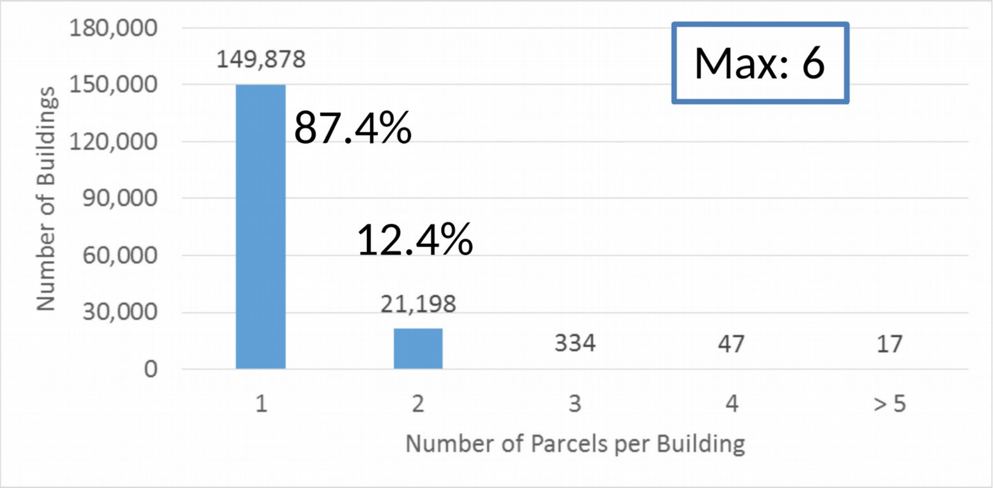


Figure 6. Number of parcels per building using the polygon clipping method

The results generated by both methods are very close. The same 154,813 buildings (94.4%) were found using either method. The second method provides more detailed information than the first one; however, it is much more challenging to implement.

As a starting point, the first method was adopted by the San Francisco Department of Technology to assign the parcel for each building. For the following steps, we used the first method to generate the mapping between building footprints and the parcel polygons and created the master dataset with all the properties of each building.

5.3. Data Standardization

5.3.1. Simplifying and standardizing the dataset

There are 183 attributes for each building in the master dataset. To make the dataset more concise, we exclude 77 attributes in the final product (Table 7). There are six reasons for the exclusion of those attributes:

(1) There are too many geometry statistics in the building footprint dataset. For the final products, the building height, building perimeter, and footprint floor area are included, and the rest of 36 geometry statistics are excluded;
(2) There are several fields from different data sources for the same data. The data fields with more detailed information are kept, while the others are excluded;
(3) There are 12 fields without data. Those empty fields are excluded;
(4) We excluded 12 fields related to the assessor’s closed roll (property tax);
(5) We excluded nine fields related to the property values as they change every year and do not directly relate to energy modeling; and
(6) One field is used to link the energy disclosure data with the San Francisco property information map but could not be used for other applications. We excluded this field.

Table 7. Reasons and examples of fields to be excluded

Reason for exclusion	No. of fields	Example fields	Description	Data source
Exclude geometry statistics	36	gnd_MINcm	Minimum ground elevation	Building Footprint
		STDcm_1st	Standard deviation of first return (roof altitude)	Building Footprint
		hgt_MAXcm	Maximum height	Building Footprint
More detailed data available from other sources	7	Building Address		Energy Disclosure
		YRBLT	Year Built	Assessor Recorder
No data and/or no field description	12	REPRIPRVAL	Prior Sales Price	Assessor Recorder
		LEASEHOLD	Leasehold Notation Flag	Assessor Recorder
		WORKFVLAND		Assessor Recorder
Exclude assessor's closed roll (property tax)	12	ROLLYEAR	Closed Roll Year	Assessor Recorder
		RP1LNDVAL	Closed Roll Assessed Land Value	Assessor Recorder
Exclude property sale information	9	RECURRPRIC	Current Sales Price	Assessor Recorder
		RECURRSALD	Current Sales Date (YYMMDD)	Assessor Recorder
Specific for certain application	1	PIM Link	Link to San Francisco Property Information Map	Energy Disclosure

After the simplification, there are 106 attributes left in the final dataset, including seven from the building footprint dataset, 17 from the land use dataset, 21 from the assessor recorder dataset, and 61 from the energy disclosure dataset. One BEDES term is used for each attribute. Table 8 shows a list of example attributes in the final master dataset. The results are stored in FileGDB and GeoJSON formats.

Table 8. Example attributes in the final master dataset

Original filed	BEDES term
sf_MBLR	Assessor parcel number
gnd1st_delta_m	Building Height

STREET	Street Name
RESUNITS	Residential Units
BLDGSQFT	Gross Floor Area
YRBUILT	Completed Construction Status Date
RP1CLACDE	Property Class Code
CONSTTYPE	Construction Type
ZONE	Zoning Code
FBA	Basement Floor Area
STOREYNO	Number of Floors
UNITS	Number of Units
ROOMS	Number of Rooms
BEDS	Number of Bedrooms
BATHS	Number of Bathrooms
RP1LSTMOD	Last Modified Date
Benchmark 2015 Status	2015 Benchmark Compliance Status
2015 Reason for Exemption	2015 Benchmark Reason for Exemption
2015 ENERGY STAR Score	2015 ENERGY STAR Assessment Value
2015 Site EUI (kBtu/ft2)	2015 Annual Site Energy Resource Intensity
2015 Source EUI (kBtu/ft2)	2015 Annual Source Energy Resource Intensity
2015 Total GHG Emissions Intensity (kgCO2e/ft2)	2015 Direct Annual CO$_2$e Emissions Intensity

5.3.2. Creating the CityGML with Energy ADE datasets

The Shapefile/FileGDB and GeoJSON formats can standardize the 2D building footprint data; however, there are not schemas for the building attributes. Although the BEDES terms can make the terms more readable, a standardized and machine-readable dataset is still necessary. Table 4 shows the CityGML and Energy ADE elements of the data needs for the UBEM. As not every attribute can be mapped to a standard CityGML or Energy ADE element, many attributes were named as CityGML generic types (gen::_GenericsAttribute) to keep the records of the collected information. For example, the annual site energy use intensity (EUI) of buildings in the year of 2015, available from the disclosure dataset named “SiteEUI_15”, is represented using a generic attribute defined in the generic schema with an element as *<gen::doubleAttribute name = "SiteEUI_15">*.

As a single CityGML file for San Francisco is too large (2.75 GB) to view or edit in general GIS or city building data visualization and analysis tools, the master buildings dataset was transformed into 16 CityGML files (at various sizes from 20 MB to 368 MB) according to the partition of the 16 planning districts of San Francisco, considering the efficient management of the CityGML files. When compressed, the total size of these 16 files was 116 MB. These planning districts are groups of census tracts and are used in various areas of the planning process, including analysis, management, and some parts of the general plan. Figure 7 shows the geographical locations and names of these districts and provides an example of the 2-D visualization of three CityGML files partitioned by planning districts: namely, Downtown, South of Market, and Mission.

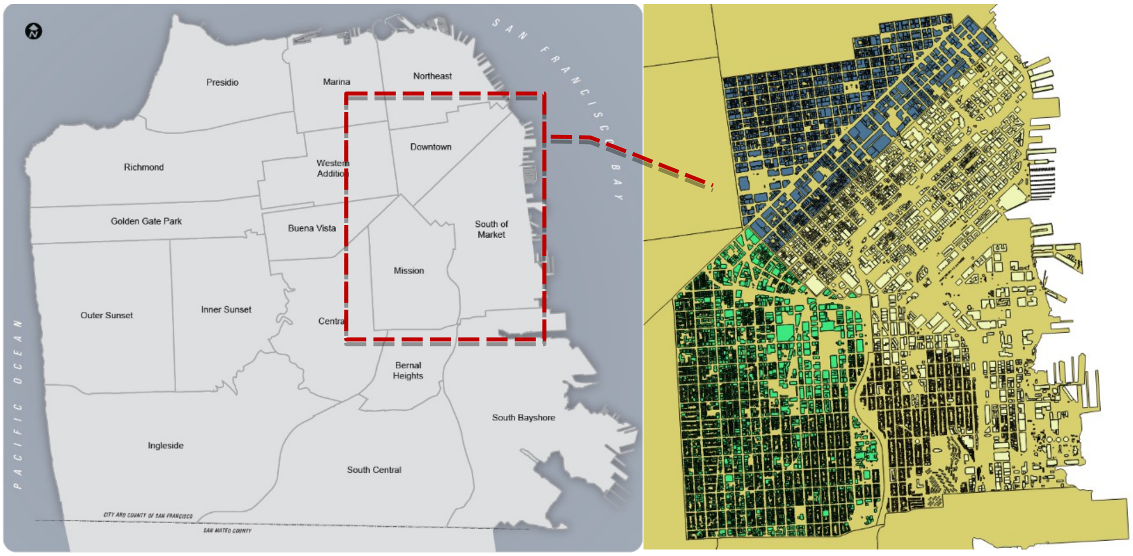


Figure 7. Partitioning of the CityGML files according to the 16 planning districts in San Francisco

Since the CityGML files were generated and validated by the standard CityGML 2.0 and Energy ADE schemas, the transformed 16 files for San Francisco can generally be used by urban visualization, analysis, modeling, and data management software.

5.4. Final products

The final products are the San Francisco buildings dataset covering the entire existing building stock, represented in multiple formats, including CityGML with Energy ADE, GeoJSON, and Shapefile/FileGDB. The final products are freely available to the public. In the future, the datasets could be enriched to include data from other building-related sources (e.g., changes/retrofits of buildings based on the building permits) and from other sectors (e.g., transportation, city water body, and city furniture such as light poles and plant pots). The methods and process used to develop the buildings dataset for San Francisco are generic and can be adopted by other cities.

6. Discussion

6.1. Applications of the city building dataset

The developed city buildings dataset can be used by multiple applications in multiple ways. Two examples are illustrated as follows.

6.1.1. Urban scale energy modeling

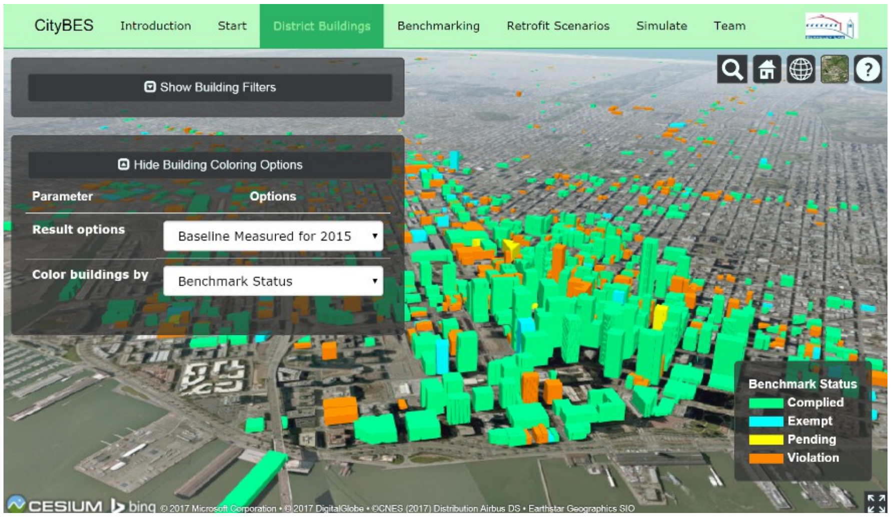
Chen et al. [10] presented a case study using LBNL's CityBES[1] to analyze the potential retrofit energy use and energy cost savings of five individual ECMs and two measure packages for 940 office and retail buildings in six city districts in northeast San

[1] https://citybes.lbl.gov

Francisco, California. A subset of the final products (the San Francisco building dataset) was used in CityBES to perform the UBEM to evaluate building retrofits.

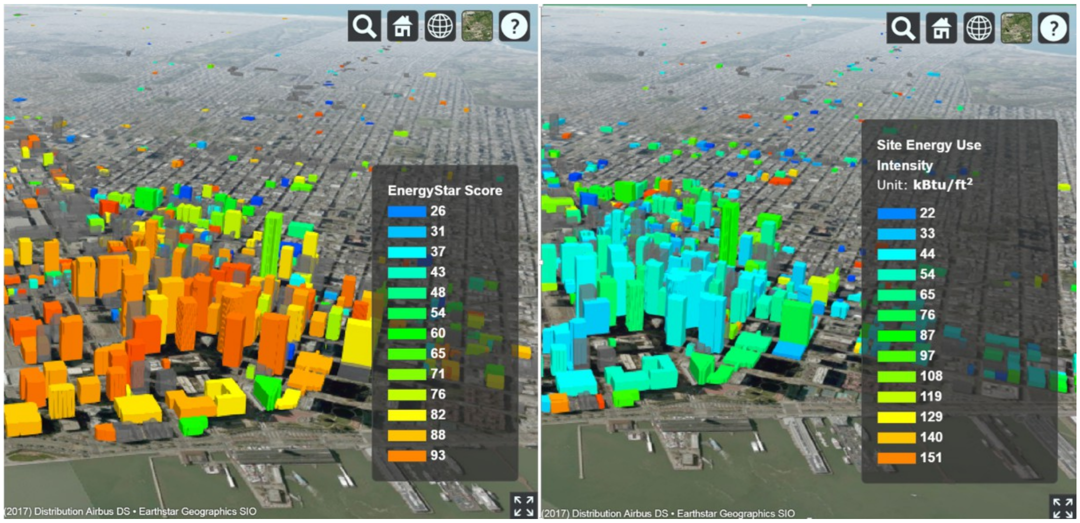
6.1.2. Visualization of energy disclosure dataset

Figure 8 shows the visualization feature using the San Francisco's energy disclosure dataset. The original energy disclosure dataset is presented in CSV/Excel format. Through the data consolidation procedure, each record of the energy disclosure dataset was linked to the associated building. The energy disclosure dataset thus can be visualized in a better way with the color-coded 3-D building geometry and map. Figure 8 (a) shows the benchmark status of each building in 2015, including Complied, Exempt, Pending, and Violation; while Figure 8 (b) and (c) present the ENERGY STAR score and site energy use intensity of each building in 2015.



Parameter	Options
Result options	Baseline Measured for 2015
Color buildings by	Benchmark Status

(a) Benchmark Status


EnergyStar Score
26
31
37
43
48
54
60
65
71
76
82
88
93
(2017) Distribution Airbus DS • Earthstar Geographics SIO
Site Energy Use Intensity
Unit: kBtu/ft²
22
33
44
54
65
76
87
97
108
119
129
140
151
(2017) Distribution Airbus DS • Earthstar Geographics SIO

(b) ENERGY STAR Score (c) Site EUI

Figure 8. Visualization of San Francisco's energy disclosure dataset

6.2. Data quality

The quality of the building dataset needs to be improved over time. For example, some of the building footprints include the yard and garden area, which makes the median building height smaller than the real median building height. The source datasets have common data issues, such as missing or invalid data.

For urban building energy modeling, some critical data are not available in the dataset, e.g., window-to-wall ratio, construction type, and energy system type (e.g., HVAC, lighting). Advanced urban sensing technologies need to be developed and applied to obtain such information at the city scale. For example, we can use drones (unmanned aerial vehicles) and cars to take photos and videos, use infrared images, and apply machine learning to extract those detailed building data.

6.3. CityGML and Energy ADE data model

CityGML is an effective way to represent 3-D geometry information. It covers several high-level building characteristics, but it does not have the detailed information necessary for building energy modeling. The Energy ADE for CityGML is currently under development, to integrate the building spatial and physics properties for urban energy simulation [51,52]. When representing the same amount of information for a 3-D model, the size of a CityGML file is typically larger than the GeoJSON or FileGDB format. Therefore, powerful computing resources are necessary to process CityGML files. Splitting a city into multiple CityGML files can be more feasible.

6.4. Data sources and ownership

The current building data are static characteristics or historical data. With the increasing adoption of the Internet of Things, more and more real-time dynamic sensing data are becoming available, which are a rich data source for urban applications.

The case study integrates the data from public sources. However, lots of private building data, e.g., Google Map, OpenStreetMap, CoStar, are available with a different licensing policy. Developing a system to handle the public and private data is necessary for long-term data management.

6.5. Limitations

Although datasets of multiple U.S. cities have been developed using the presented data sources, methods and workflow, their application to cities in other countries still needs to be investigated. Part of the authors' on-going research is looking at other data sources, such as building permits which can provide good information on changes to buildings. Integrating these additional existing and new sources can create new data challenges.

7. Conclusions

The building data needs for UBEM typically include the GIS building footprint, building height, total number of stories, number of stories above ground, number of stories below ground, total floor area, heated floor area, number of dwellings, year of construction, year of refurbishment, use type (building type), heating system type, annual electricity use, annual natural gas use, annual site energy sue, and annual source energy use.

The data standards/formats used in UBEM mainly include the Shapefile/FileGDB, GeoJSON, and CityGML. The current data standards can provide a standardized representation of the 2D or 3D building geometry information. However, the Shapefile/FileGDB and GeoJSON files do not provide schemas for the building attributes. The CityGML and Energy ADE provide the standardized presentation for several necessary data fields and future enhancements are necessary to cover more high-level building information.

The existing public data sources from several pioneer cites are adequate to support UBEM. However, the data are represented in different formats without standardization and there lack common keys to map the data from diverse sources. The mapping of building footprint and parcel polygons to link multiple datasets is the most complicated and challenging step for the data integration. In future, city's buildings datasets can use the standardized unique building identifiers for indexing which makes the mapping and linking of diverse building datasets straightforward.

A city-scale building dataset is a key to urban building energy modeling. Today, cities put an enormous amount of effort into collecting and sharing building data via open web-based data portals. When this is done, it is essential to provide the data in a standardized way, to enable interoperability and adoption by various types of urban applications. CityGML, an international standard for 3-D city models, is an excellent tool for representing and exchanging city data among different users and different tools. This paper presented methods and tools that can be used to integrate city-scale building data from multiple city departments. The data are represented in the CityGML format, as well as in the GeoJSON and Shapefile/FileGDB formats, to support existing urban modeling and analysis tools, as well as future developments.

The buildings dataset is open access and can be used by a variety of urban/city applications, including retrofit analysis of existing buildings, urban planning, and visualizing the energy performance and code compliance status of building stock. The developed scripts, tools, and tutorials, although based on the city of San Francisco, have been applied to datasets in other U.S. cities including San Jose, Los Angeles, Chicago, New York City, and Boston, enabling researchers and city consultants to create standardized buildings datasets for their urban applications.

Acknowledgments

This work was supported by the United States Department of Energy, Office of Energy Efficiency and Renewable Energy (EERE) under Contract No. DE-AC02-05CH11231. Various departments of the City of San Francisco provided the source datasets.

References

[1] San Francisco Environment. San Francisco Citywide Greenhouse Gas Reduction Actions and Goals 2018. https://sfenvironment.org/article/citywide-actions-and-goals (accessed March 10, 2018).

[2] San Francisco Environment. San Francisco Climate Action Strategy 2013 Update. San Francisco, CA: 2013.

[3] San Francisco Environment. Energy for San Francisco Residents 2017. https://sfenvironment.org/energy-for-san-francisco-residents (accessed November 2, 2017).

[4] San Francisco Environment. San Francisco Energy Watch. SF Dep Environ 2017. https://sfenvironment.org/energy/energy-efficiency/commercial-and-multifamily-properties/sf-energy-watch (accessed November 26, 2017).

[5] San Francisco Environment. San Francisco's Property Assessed Clean Energy financing program 2017. https://sfenvironment.org/residentialpace (accessed March 3, 2017).

[6] San Francisco Environment. Solar Incentives for Homeowners 2017. https://sfenvironment.org/solar-incentives-homeowners (accessed November 2, 2017).

[7] San Francisco Environment. Energy Upgrade California Multifamily Program 2017. https://sfenvironment.org/energy-upgrade-california-multifamily (accessed November 2, 2017).

[8] Reinhart CF, Davila CC. Urban building energy modeling - A review of a nascent field. Build Environ 2016;97:196–202.

[9] Delmastro C, Mutani G, Corgnati SP. A supporting method for selecting cost-optimal energy retrofit policies for residential buildings at the urban scale. Energy Policy 2016;99:42–56. doi:10.1016/j.enpol.2016.09.051.

[10] Chen Y, Hong T, Piette MA. Automatic generation and simulation of urban building energy models based on city datasets for city-scale building retrofit analysis. Appl Energy 2017;205:323–35. doi:10.1016/j.apenergy.2017.07.128.

[11] Yamaguchi Y, Shimoda Y, Mizuno M. Transition to a sustainable urban energy system from a long-term perspective: Case study in a Japanese business district. Energy Build 2007;39:1–12. doi:10.1016/j.enbuild.2006.03.031.

[12] City of San Francisco. DataSF 2018. https://datasf.org/ (accessed July 27, 2018).

[13] City of Chicago. Chicago Data Portal 2018. https://data.cityofchicago.org/ (accessed July 27, 2018).

[14] City of New York. NYC OpenData 2018. https://opendata.cityofnewyork.us/ (accessed July 27, 2018).

[15] U.S. Department of Energy. EnergyPlus Weather Data 2018. https://energyplus.net/weather (accessed August 3, 2018).

[16] Sokol J, Cerezo Davila C, Reinhart CF. Validation of a Bayesian-based method for defining residential archetypes in urban building energy models. Energy

Build 2017;134:11–24. doi:10.1016/j.enbuild.2016.10.050.

[17] Österbring M, Mata É, Thuvander L, Mangold M, Johnsson F, Wallbaum H. A differentiated description of building-stocks for a georeferenced urban bottom-up building-stock model. Energy Build 2016;120:78–84. doi:10.1016/j.enbuild.2016.03.060.

[18] Aksoezen M, Daniel M, Hassler U, Kohler N. Building age as an indicator for energy consumption. Energy Build 2015;87:74–86. doi:10.1016/j.enbuild.2014.10.074.

[19] Chen Y, Hong T. Impacts of building geometry modeling methods on the simulation results of urban building energy models. Appl Energy 2018;215:717–35. doi:10.1016/j.apenergy.2018.02.073.

[20] Dogan T, Reinhart C. Shoeboxer: An algorithm for abstracted rapid multi-zone urban building energy model generation and simulation. Energy Build 2017;140:140–53. doi:10.1016/j.enbuild.2017.01.030.

[21] Cerezo Davila C, Reinhart CF, Bemis JL. Modeling Boston: A workflow for the efficient generation and maintenance of urban building energy models from existing geospatial datasets. Energy 2016;117:237–50. doi:10.1016/j.energy.2016.10.057.

[22] Nageler P, Zahrer G, Heimrath R, Mach T, Mauthner F, Leusbrock I, et al. Novel validated method for GIS based automated dynamic urban building energy simulations. Energy 2017;139:142–54. doi:10.1016/j.energy.2017.07.151.

[23] Heine Kristensen M, Elbaek Hedegaard R, Petersen S. Hierarchical calibration of archetypes for urban building energy modeling. Energy Build 2018;175:219–34. doi:10.1016/j.enbuild.2018.07.030.

[24] Fracastoro GV, Serraino M. A methodology for assessing the energy performance of large scale building stocks and possible applications. Energy Build 2011;43:844–52. doi:10.1016/j.enbuild.2010.12.004.

[25] Theodoridou I, Papadopoulos AM, Hegger M. A typological classification of the Greek residential building stock. Energy Build 2011;43:2779–87. doi:10.1016/j.enbuild.2011.06.036.

[26] Firth SK, Lomas KJ. Investigating Co2 Emission Reductions in Existing Urban Housing Using a Community Domestic Energy Model. Build Simul 2009:2098–105.

[27] Heiple S, Sailor DJ. Using building energy simulation and geospatial modeling techniques to determine high resolution building sector energy consumption profiles. Energy Build 2008;40:1426–36. doi:10.1016/j.enbuild.2008.01.005.

[28] Ballarini I, Corgnati SP, Corrado V. Use of reference buildings to assess the energy saving potentials of the residential building stock: The experience of TABULA project. Energy Policy 2014;68:273–84. doi:10.1016/j.enpol.2014.01.027.

[29] Tuominen P, Holopainen R, Eskola L, Jokisalo J, Airaksinen M. Calculation method and tool for assessing energy consumption in the building stock. Build Environ 2014;75:153–60. doi:10.1016/j.buildenv.2014.02.001.

[30] Mata É, Sasic Kalagasidis A, Johnsson F. Building-stock aggregation through archetype buildings: France, Germany, Spain and the UK. Build Environ 2014;81:270–82. doi:10.1016/j.buildenv.2014.06.013.

[31] Caputo P, Costa G, Ferrari S. A supporting method for defining energy strategies in the building sector at urban scale. Energy Policy 2013;55:261–70. doi:10.1016/j.enpol.2012.12.006.

[32] Wikipedia. Shapefile 2017. https://en.wikipedia.org/wiki/Shapefile (accessed November 4, 2017).

[33] GDAL. ESRI File Geodatabase 2017. http://www.gdal.org/drv_filegdb.html.

[34] GeoJSON WG. GeoJSON 2017. http://geojson.org/ (accessed November 26, 2017).

[35] U.S. DOE, Lawrence Berkeley National Laboratory. Building Energy Data Exchange Specification (BEDES) 2017. https://bedes.lbl.gov/ (accessed January 7, 2018).

[36] OGC. CityGML 2017. https://www.citygml.org/ (accessed February 22, 2017).

[37] Gröger G, Plümer L. CityGML - Interoperable semantic 3D city models. ISPRS J Photogramm Remote Sens 2012;71:12–33. doi:10.1016/j.isprsjprs.2012.04.004.

[38] Eicker U, Nouvel R, Duminil E, Coors V. Assessing passive and active solar energy resources in cities using 3D city models. Energy Procedia 2014;57:896–905. doi:10.1016/j.egypro.2014.10.299.

[39] Strzalka A, Bogdahn J, Coors V, Eicker U. 3D City modeling for urban scale heating energy demand forecasting. HVAC&R Res 2011;17:37–41. doi:10.1080/10789669.2011.582920.

[40] Remmen P, Lauster M, Mans M, Fuchs M, Müller D. TEASER : an open tool for urban energy modelling of building stocks. J Build Perform Simul 2017;1493. doi:10.1080/19401493.2016.1283539.

[41] Macumber D, Gruchalla K, Brunhart-lupo N, Gleason M, Abbot-Whitley J, Robertson J, et al. City Scale Modeling With OpenStudio. SimBuild 2016, Salt Lake City, UT, U.S.: 2016.

[42] Hong T, Chen Y, Lee SH, Piette MA. CityBES : A Web-based Platform to Support City-Scale Building Energy Efficiency. Urban Comput. 2016, San Francisco, San Francisco, California USA: 2016.

[43] Chen Y, Hong T, Piette MA. City-Scale Building Retrofit Analysis: A Case Study using CityBES. Build. Simul. 2017, San Francisco, CA USA: 2017.

[44] Laurini R. Visual Information Systems Chapter V: Virtual 3D Cities. 2015.

[45] Agugiaro G, Benner J, Cipriano P, Nouvel R. The Energy Application Domain Extension for CityGML: enhancing interoperability for urban energy simulations. Open Geospatial Data, Softw Stand 2018;3:2. doi:10.1186/s40965-018-0042-y.

[46] San Francisco Office of the Assessor-Recorder. San Francisco Assessor Records 2017. http://www.sfassessor.org/ (accessed November 4, 2017).

[47] San Francisco Planning Department. San Francisco Property Information Map 2017. http://propertymap.sfplanning.org/ (accessed November 26, 2017).

[48] San Francisco Environment. San Francisco Existing Commercial Buildings Energy Performance, 2010-2014. San Francisco, CA USA: 2015.

[49] QGIS Community. QGIS 2017. http://www.qgis.org/en/site/ (accessed May 24, 2017).

[50] Ruby Community. Ruby 2017. https://www.ruby-lang.org/en/ (accessed May

24, 2017).

[51] Nouvel R, Kaden R, Bahu J, Kaempf J, Cipriano P, Lauster M, et al. Genesis of the CityGML Energy ADE. CISBAT 2015, Lausanne, Switzerland: 2015.

[52] Benner J, Geiger A, Häfele K. Virtual 3D City Model Support for Energy Demand Simulations on City Level – The CityGML Energy Extension. REAL CORP 2016, Hamburg, Germany: 2016.