https://dx.doi.org/10.1016/j.enbuild.2023.112911

A Review and Reflection on Open Datasets of City-level Building Energy Use and their Applications

Xiaoyu Jin[1], Chong Zhang[1], Fu Xiao[1, 2, *] Ao Li[1] and Clayton Miller[3]

[1.] Department of Building Environment and Energy Engineering, The Hong Kong Polytechnic University

[2.] Research Institute for Smart Energy, The Hong Kong Polytechnic University

[3.] College of Design and Engineering, National University of Singapore

[*] Corresponding author: linda.xiao@polyu.edu.hk

Abstract

Data related to building energy use fuel the research and applications on building energy efficiency, which is an essential measure to address global energy and environmental challenges. However, in most cities, there is a lack of comprehensive and publicly accessible building energy use datasets with necessary temporal and spatial granularities to support urban building energy modeling, regional energy planning, city-level building performance benchmarking, and policymaking on building energy efficiency. Data owners and governments are facing challenges in determining which and how building energy use data should be disclosed at the city level and how to protect data privacy. This review paper provides insights to answer these questions based on a comprehensive and critical review of worldwide open datasets and their applications in the built environment context. Detailed information about the collected 33 building energy datasets is summarized and categorized. Studies identified into 11 subdomains using these open datasets are critically analyzed and compared. Potential policy implications based on the studies are also proposed. Moreover, non-energy datasets that are frequently used in research relevant to urban building energy use are also introduced. Solutions to privacy issues are discussed to address concerns from data publishers. Finally, significant conclusions are made to support the proper disclosure of city-level building energy data. This review study is valuable to urban building energy data disclosure, urban building energy modeling, and data-driven energy policymaking.

Highlights

• Over 193 worldwide open datasets of city-level building energy consumption are summarized.

• Representative applications of the collected 33 datasets are analyzed and compared.

• Policy implications are presented for using the open datasets to support policymaking.

• Useful non-energy datasets concerning weather information, building characteristics, urban forms, and occupant-related factors are recommended.

• Solutions to privacy issues are explored, including anonymizing methods and proper data aggregation.

Keywords: open data / urban building energy / energy modelling / energy policy / data disclosure

1. Introduction

Currently, more than half of the world's population lives in urban areas. It is estimated that with population growth and rapid urbanization, the global urban population will increase by 1.5 times by the middle of the 21st century. Contributing to more than 70% of final energy use and greenhouse gas emissions, the urban area has also been recognized as crucial for mitigating energy risks and climate change [1]. Therefore, urban energy planning has become an important research topic for the sustainable development of cities and society. Moreover, as the building sector is responsible for high energy consumption and environmental impact [2], it is a vital part of urban energy analysis [1].

Compared with simulated datasets, measured datasets can explain unpredictable and complicated real-life situations, which can help to provide a more comprehensive understanding and improve the efficiency of building energy use [1]. Meanwhile, city-level building datasets can be used in many specific fields, including urban building energy modeling (UBEM) [3], renewable energy analysis, regional energy planning and building performance analysis [4].

However, at present, the research fields of energy, environment, and sustainable development are all facing the urgency of developing publicly accessible datasets [1]. The difficulty of data collection and limited data disclosure makes it hard for researchers to obtain city-level building data. In a survey carried out by Kjærgaard et al. (2020) [5], 78% of respondents reported that the limited availability of datasets is the most significant barrier to open data-driven research, and 53% of respondents reported the lack of available data as the second largest barrier. Though open data shows significant research value, challenges such as data collection and enrichment still need to be addressed.

The establishment of energy-related policies requires realistic energy data and corresponding analysis. However, government departments, as the most important collectors and holders of city-level building energy data, still have not disclosed their datasets in most countries. In fact, data openness can bring many benefits and opportunities for energy saving at the city or country level. Firstly, peer pressure caused by the disclosure of energy data can incentivize more energy-efficient actions. Meanwhile, disclosure can promote more effective collaboration across scientific and policy fields and therefore increase productivity through collaboration and burden sharing [6]. A relevant study indicates that a 3-8% reduction in energy consumption can be achieved over 2 to 4 years in 6 American cities after data disclosure [7]. Flores [8] pointed out that energy use in large buildings in Australia has decreased by up to 35% in just nine years due to the introduction of mandatory disclosure. Meng et al. [9] evaluated the impact of a novel difference-in-difference strategy on building energy use in a New York City benchmarking program. The results show that the policy implementation achieved energy savings of 14% in 4 years.

In the past few years, the climate sciences, community geosciences, public health, and biomedical research communities have all made significant advancements in software and data, including overcoming the difficulty of anonymizing private information [6]. Despite the considerable benefits of open data, the openness of data in the urban energy domain still seems to need to catch up to other fields. For the building energy field, the existing literature reviews about open building energy datasets mainly focus on high-resolution data (time interval from 1 second to 1 hour), of limited buildings [2] [9], or city-level data only from the United States [3, 10-13]. So far, there has yet to be an aggregation of worldwide building energy datasets at the city-level for providing a comprehensive overview of the characteristics of these datasets and the relevant studies using these datasets.

Therefore, this review's primary purpose is to collect and summarize open urban building energy datasets worldwide systematically and the relevant studies and potential applications based on these open datasets to support researchers in relative fields, offer policy implications for governments and promote data disclosure.

The rest of the paper is organized as follows. The dataset-searching methods and relevant research are introduced in Section 2. An overview of the collected datasets is presented in Section 3. Section 4 categorizes these datasets and corresponding research in detail according to the application purposes and introduces the potential policy implications. Discussion on non-energy datasets supporting energy research, as well as privacy issues, are discussed in Section 5. Section 6 presents the main conclusion and prospects for future work.

2. Methodology of Review

The methodology of this review mainly consists of the following three steps: searching review papers about building energy-related open datasets, collecting open datasets from the review papers, and searching for research papers using these datasets. In this review, a total of 33 datasets and 198 research papers using these datasets are collected. A flowchart of the review methodology is shown in Figure 1.

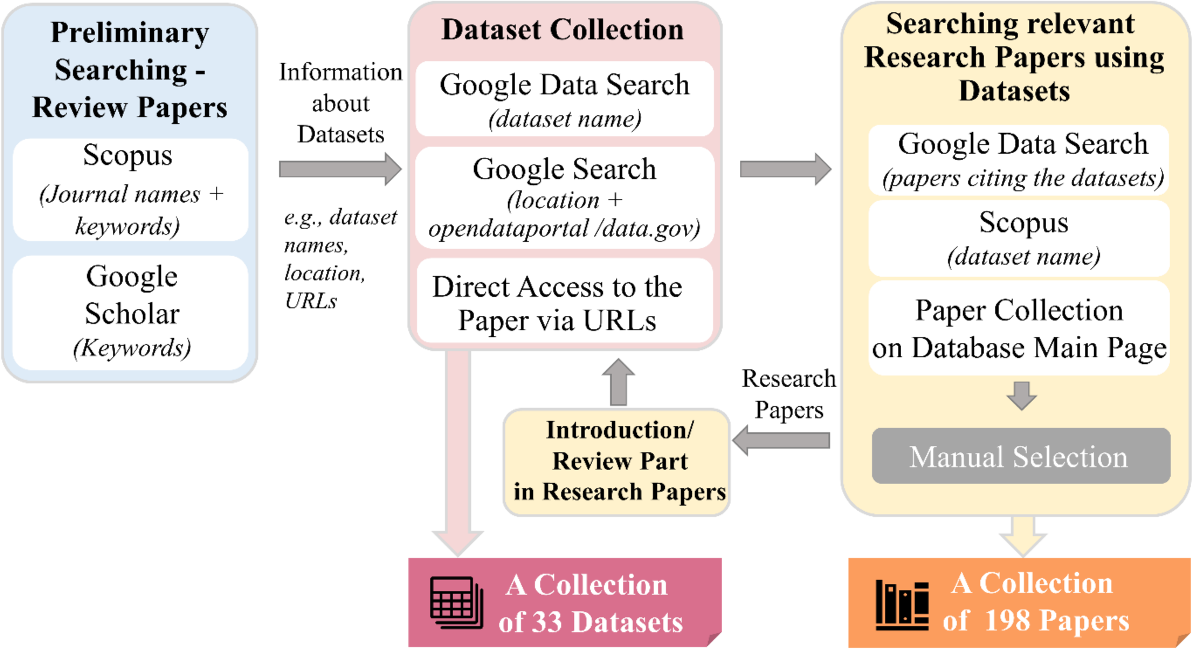
Preliminary Searching - Review Papers
Scopus
(Journal names + keywords)
Google Scholar
(Keywords)
Information about Datasets
e.g., dataset names, location, URLs
Dataset Collection
Google Data Search
(dataset name)
Google Search
(location + opendataportal /data.gov)
Direct Access to the Paper via URLs
Introduction/ Review Part in Research Papers
Research Papers
Searching relevant Research Papers using Datasets
Google Data Search
(papers citing the datasets)
Scopus
(dataset name)
Paper Collection on Database Main Page
Manual Selection
A Collection of 33 Datasets
A Collection of 198 Papers

Figure 1. Flowchart of Review Methodology

To streamline the search process, the review starts with a preliminary search for review papers related to building energy which used open datasets, as shown in Figure 1. The reason for starting with review papers is because “database/dataset” is seldom used as a keyword or in the titles of research papers, even if these papers use open datasets. Therefore, the keywords “database/dataset” cannot be directly used to do a literature search, which is different from how most other review papers searched for relevant literature. Articles published before June 2021 were collected. Scopus database was used for searching within journals. As shown in Table 1, the following search code was used for screening the review papers in Scopus: TS = (“Open Data” OR “Data Disclosure”). In this step, building-energy-related journals were examined for review paper collection. The initial screening finds 11 journals with lots of potentials to contain the target review papers. After the first round of examination, five of them were eliminated because there were too few available papers in these journals. The six remaining journals include *Applied Energy*, *Energy and Buildings*, *Building and Environment*, *Energy Policy*, *Renewable & Sustainable Energy Reviews*, and *Energy Procedia*. The papers were further screened for two rounds: the selection was based on titles and contents. Papers were collected corresponding to the keywords “Open Data” (31) and “Data Disclosure” (3). Also, Google Scholar provided supplement searching in case of any omissions. The search code was: TS = (building energy open data) and TS = (building energy data disclosure). Twelve papers were collected using Google Scholar. Finally, as shown in Table 1, a total of 46 useful review papers were collected.

Next, according to information provided by these review papers, datasets were collected and aggregated. A few papers provide links directed to the data sources, while this is an ideal but uncommon case. Many of the links in the articles are expired, or only names/locations of the datasets were provided. For invalid URLs, some of their home pages could still generally be traced, allowing the original datasets to be further searched. The Google Data Search engine was used to identify datasets mentioned only by name. Where only the location (i.e., city/state/country) was provided, keywords following the format "location name + open data portal/data.gov" were used.

Table 1. Number of Review Papers Found in the Preliminary Searching using Scopus and Google Scholar

Scopus						**Google Scholar**
Journal Title	**Key Words (searching with Quotation Marks)**					**Key Words**
	open data			*data disclosure*		*building energy open data/ data disclosure*
	Primary Result	Selection based on research titles	Selection based on contents	Primary Result	Selection based on contents	
Applied Energy	60	17	13	7	2	12
Energy and Buildings	42	10	6	2	1	
Energy Procedia	34	10	4	0	0	
Building and Environment	36	9	3	1	0	
Energy Policy	32	10	3	7	0	
Renewable & Sustainable Energy Reviews	27	7	2	3	0	
Sub-total	231	63	31	20	3	12
Total Number of Useful Review Papers						**46**

Figure 1 shows that the third step is to search for research papers using these datasets. The Google Data Search not only shows data sources but also presents how many articles have used the datasets. This function is helpful in the collection of relevant research papers. However, there are some limitations to this method. Some documents in Google Data Search didn't use the datasets in their studies but only mentioned them as potential sources or cited aggregated statistics such as the total recorded building number or energy consumption. Some datasets had names that were too generic, leading to misidentification when using Google Data Search. Meanwhile, the names of some datasets were too long or abbreviated which made it challenging to find the corresponding datasets. As a result, few relevant papers were collected for those datasets. To carry out a more comprehensive investigation of studies using these datasets, the authors further adjusted the keywords to conduct a search on Scopus for each dataset. It is worth mentioning that there is one dataset, i.e., Dataport database (DATAPORT), which provides access to university faculty, staff and students with large amount of data for academic research purposes and fuels great amount of building energy research, listing over 300 research papers published on its official website [14]. To refine and summarize the documents from this data directory, the latest (published after 2018) and most representative articles (with citations > 10 or the paper has a unique research direction) for DATAPORT have been selected for further analysis.

More datasets were discovered based on the review section of the collected relevant research articles, as shown in Figure 1. Ultimately, 33 datasets were found, and the detailed information of each dataset is listed in Appendix A, including publisher, covered cities/areas, data amount, time interval, time range, and website. Since the research direction varies with time intervals of data, the collected research papers were also categorized according to sampling interval frequency. There were 198 papers collected in total that were all published before April 2022. The research article number for each dataset is found in Appendix A.

These collected papers were also classified by application purposes for further analysis. This process required fully comprehending the articles to categorize these studies by their research purposes further. Since most of the research outcomes are helpful for energy-related policies, the policy implications of these studies were summed up separately. While examining the publications, it was found that many studies also used datasets in other fields besides building energy consumption, such as weather and geographical information datasets. Therefore, those valuable and publicly available datasets were also gathered for future analysis. Finally, an exploratory investigation was conducted in terms of privacy issues of open datasets; the location accuracy and time intervals were compared and discussed.

3. Overview of City-level Building Energy Datasets

This section summarizes the data disclosure status of different countries worldwide, presents the spatial and temporal distributions of the collected city-level building energy datasets, and discusses the variables included in each dataset. Differing building energy data disclosure policies result in varying amounts of available data in different cities. It was found that the open datasets mainly come from the United States, Europe, Australia, and Singapore.

Table 2. (a) Dataset Distributions of Countries

District/Country		Data Opening Status	Relevant Policies	Dataset Number
The United States		Over 15 cities provide publicly accessible building data.	Benchmarking and transparency policies [7]	22
Europe	The United Kingdom	Large number of energy performance record, but relatively less public data; UK and Ireland have higher openness levels [15, 16].	European union energy performance of buildings directive [17]	2
	Ireland			2
	Italy			1
	Denmark			1
Australia		Energy consumption information of commercial office space with 1000 square meters or more should disclose their energy data.	Commercial building information disclosure (CBD) program [18]	1
Singapore		A phased approach in disclosure of building energy performance was rolled out since 2014.	Building and construction authority data disclosure plan [19]	4

Table 2. (b) Dataset Distributions of Time Intervals

Time Interval	Dataset Number
Year	25
Month	6
Sub-hour	2

As shown in Table 2 (a), 33 open datasets from eight countries were collected, with the United States having the most significant number, which may attribute to the stricter regulations on building energy data disclosure in the United States. The imbalance issue of data availability in different countries results that more related research focused on cities in the United States.

As shown in Table 2 (b), most of these datasets disclose building energy data by year, and open datasets with higher time frequencies account for the lowest proportion. This is reasonable as higher frequency can lead to more difficulties with collection and privacy issues for data holders. It should be noted that even if yearly data has the lowest time frequency, it always discloses the highest number of building characteristic variables, which means it can provide more information than data with a higher frequency.

In general, the variables in the datasets could be categorized into13 types: 1) property identification, including address/ unique coding of the building; 2) energy consumption, including electricity, gas, and other kinds of energy consumed in the building; 3) building category, also expressed as property use type in some datasets; 4) floor area, some datasets also have a more detailed record of floor area values of different property use type in one building; 5) year of construction; 6) energy ratings, like EnergyStar or LEED scores; 7) energy device installation conditions; 8) greenhouse gas emissions; 9) occupancy rate; 10) water consumption values; 11) building envelope characteristics; 12) household condition; 13) other variables. A summary of the variables in each dataset is provided in Appendix B.

As shown in Figure 2, the number of variables in the yearly datasets mainly ranges from 20 to 80; even though the mean number of the variables is lower than the sub-hourly datasets, there are several datasets with many variables, the maximum which reaches one thousand, indicating the feasibility of including numerous variables into the yearly datasets. It can also be found that the average variable number of high-frequency data is higher than that of the monthly datasets.

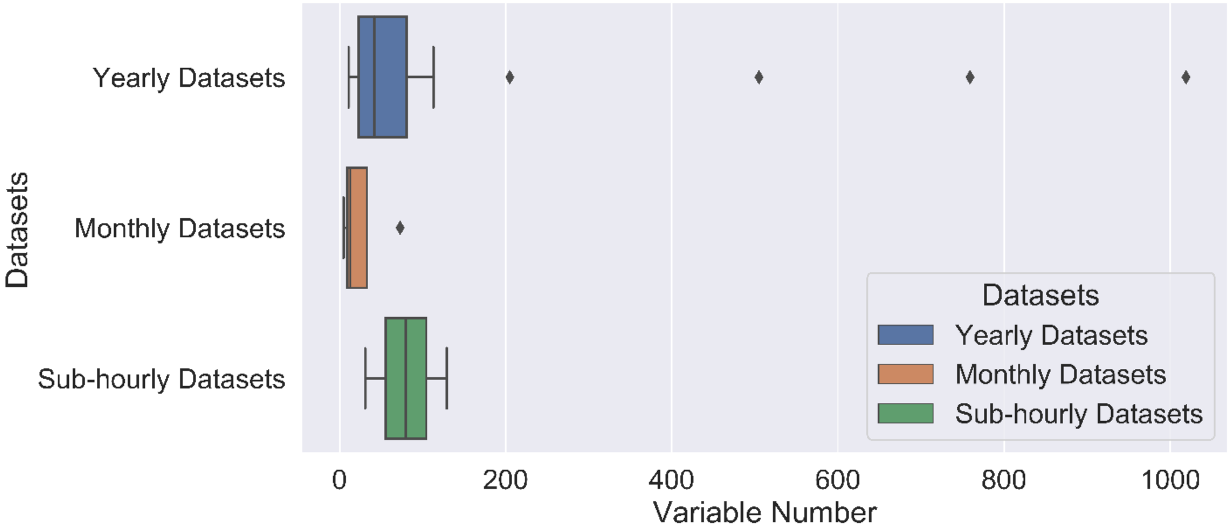

Figure 2. Distribution of Variable Numbers in the Collected Datasets

Figure 3 presents the variable distribution of the collected datasets. The statistics show that the essential variables, such as building identification and energy consumption are recorded in all the building energy use datasets. Most datasets record building category, year built, and energy rating data. Other variables are included in different datasets for different purposes. A small number of the datasets cover household condition, building envelope characteristics, and occupancy rates, among others. This might be because these kinds of information are difficult to obtain due to privacy issues.

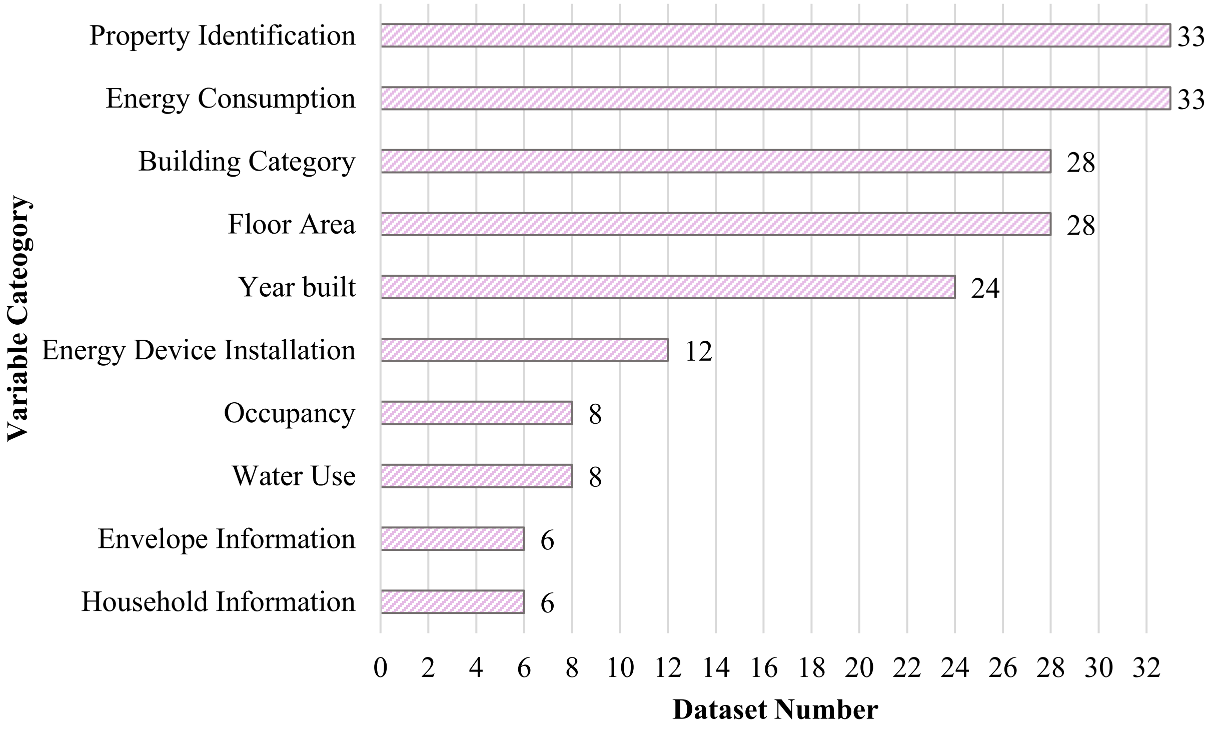

Figure 3. Variable Category Distribution of All the Collected Datasets

4. Applications of City-level Building Energy Datasets

This section categorizes the research and applications based on the collected city-level building energy datasets and further summarizes the potential applications of these datasets. A total of 198 relevant papers have been collected based on the search methodology mentioned in Section 2. As shown in Figure 4, these papers can be grouped into three research directions by their application types: building energy management, grid management and social economics analysis, which are analyzed in detail in Section 4.1 ~ 4.3. Research papers are further subdivided into each section. Moreover, since providing policy support is an essential goal of this study, the policy implications will be introduced and summarized in the Section 4.4.

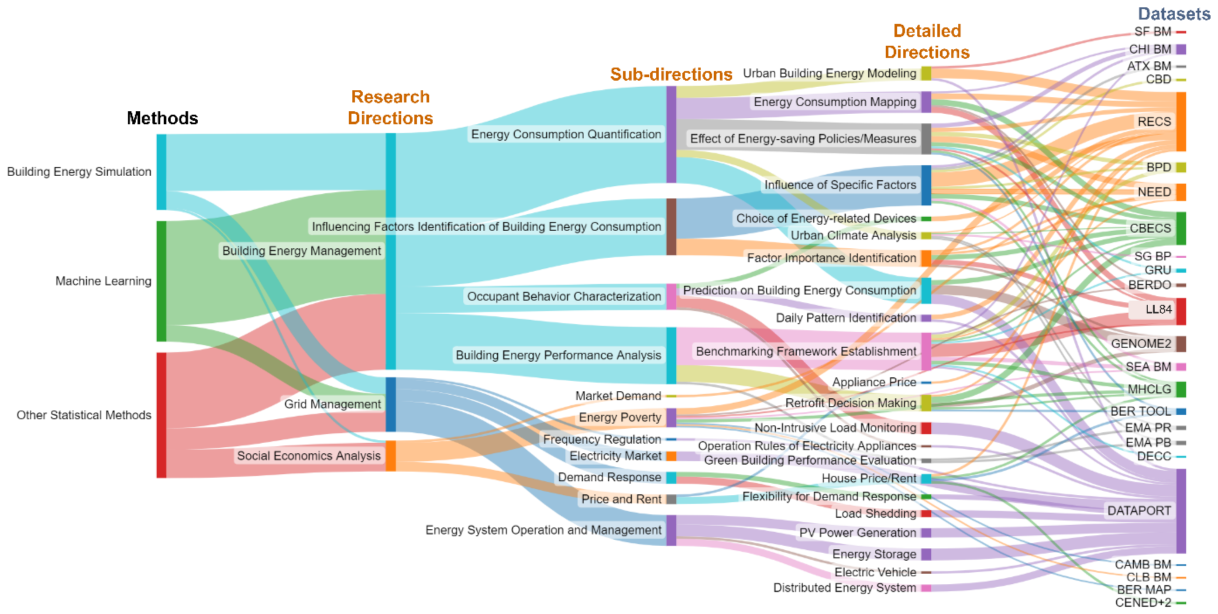
Methods
Research Directions
Sub-directions
Detailed Directions
Datasets
Building Energy Simulation
Machine Learning
Other Statistical Methods
Building Energy Management
Grid Management
Social Economics Analysis
Energy Consumption Quantification
Influencing Factors Identification of Building Energy Consumption
Occupant Behavior Characterization
Building Energy Performance Analysis
Market Demand
Energy Poverty
Frequency Regulation
Electricity Market
Demand Response
Price and Rent
Energy System Operation and Management
Urban Building Energy Modeling
Energy Consumption Mapping
Effect of Energy-saving Policies/Measures
Influence of Specific Factors
Choice of Energy-related Devices
Urban Climate Analysis
Factor Importance Identification
Prediction on Building Energy Consumption
Daily Pattern Identification
Benchmarking Framework Establishment
Appliance Price
Retrofit Decision Making
Non-Intrusive Load Monitoring
Operation Rules of Electricity Appliances
Green Building Performance Evaluation
House Price/Rent
Flexibility for Demand Response
Load Shedding
PV Power Generation
Energy Storage
Electric Vehicle
Distributed Energy System
SF BM
CHI BM
ATX BM
CBD
RECS
BPD
NEED
CBECS
SG BP
GRU
BERDO
LL84
GENOME2
SEA BM
MHCLG
BER TOOL
EMA PR
EMA PB
DECC
DATAPORT
CAMB BM
CLB BM
BER MAP
CENED+2

Figure 4. Sankey Diagram of City-level Building Energy Datasets and Their Applications

4.1 Building Energy Management

Building energy management is the most important application of the collected datasets, which includes wide usage purposes on energy consumption prediction, influencing factor analysis of energy consumption, occupant behavior research, and building energy performance analysis. And this section will be presented according to the subsections under building energy management shown in Figure 4. Generally, these studies are conducted with predictive modelling of energy consumption, but the ultimate application purposes of these research can be various. Except for solely targeting on energy consumption prediction some papers utilize the prediction model to investigate other topics such as the influencing factors on building energy consumption, occupant behavior, and building energy performance.

4.1.1 Energy Consumption Quantification

The summary of papers regarding energy consumption prediction is shown in Table 3. There are various application purposes, including energy consumption mapping, effect of energy-saving policies/measures, urban building energy modeling, urban climate, and time series prediction on building energy consumption.

Table 3. Application Summary: Research of Energy Consumption Prediction using Open Datasets

Research Category	**Output**	**Data Granularity**	**Typical Literature (dataset)**
Energy Consumption Mapping	Energy Use Intensity (Total/End Use)/Electricity Profile/Energy Consumption (by Community/Capita)	year	[20] (RECS,CBECS), [21] (CHI BM), [22] (LL84), [23] (BERDO), [24] (CBECS), [25] (LL84), [88] [26] (CBECS, LL84), [27] (RECS), [28] (RECS)
Effect of Energy-saving Policies/Measures	Energy Consumption Change/GHG Emission/Probability Density Functions of Energy Use	year/month	[29] (GRU), [30] (RECS,CBECS), [31] (BER TOOL), [32] (NEED), [33] (BPD), [34] (NEED), [35] (NEED), [36] (MHCLG), [37] (LL84), [38] (CHI BM), [39] (CHI BM), [40] (RECS)
Urban Building Energy Modeling	Generated Hourly Consumption Profile of a City/District/Site EUI by month/Appliances in a Household	year/15 mins	[41] (RECS, CBECS), [42] (CBECS,BPD), [43] (RECS), [44] (CBECS), [45] (CBECS), [46] (RECS), [47] (SF BM), [48] (RECS), [49]DATAPORT, [50] (RECS)
Urban Climate Analysis	Building Anthropogenic Heat	year/month	[51] (SG BP, EMA PR, EMA PB), [52] (SG BP, EMA PR, EMA PB)
Prediction on Building Energy Consumption	Electricity Consumption Profile of Buildings (day/week/month/year)/	hour/15 minutes	[53] GENOME2, [96] [54]DATAPORT, [55]DATAPORT, [56]GENOME2, [57]GENOME2, [58]GENOME2, [59]DATAPORT, [60]DATAPORT

The studies of energy consumption mapping generally use existing building energy data to estimate the energy consumption of other buildings in cities or districts. The outputs of these studies reflect the spatial distribution of energy use in the cities, which can be energy use intensity (EUI)[20 , 21] [25] [33], electricity profile [41] [22] [23] of buildings, as well as energy consumption of communities [28]. Commonly used datasets include RECS, CBECS, LL84, BER MAP, CHI BM, and BERDO， which are all coarse-grained datasets with yearly or monthly intervals. The results of these studies can help policymakers better understand the energy use of buildings to take more targeted measures. For example, Constantine et al. [25] developed a data-driven predictive model of electricity and gas use at the building, district, and city scales using training data, including energy consumption values from LL84 and predictors from the widely-available property and zoning information. The energy consumption condition of many buildings without disclosing energy consumption data can be estimated. The study found that building use, size, and form were reliable predictors of energy use at the building and zip code levels. In addition, Roth et al. [22] generated an hourly electricity profile for each building in New York City by using the LL84 dataset. Yearly electricity consumption was mapped to each building using data provided in LL84. The aggregated hourly electricity consumption in the city and the prototype simulation data were used for convex optimization to get the hourly electricity profile for each building. According to the mapping results, the locations of distributed energy resources and district energy systems in New York City can be determined appropriately.

The city-level building energy consumption prediction can also be used for evaluating the effects of energy-saving policies or measures. Before implementing a policy, relevant studies are conducted to quantify the energy-saving potential. For example, Lannon et al. [32] simulated the necessary retrofitting measures to achieve the UK government's carbon emission reduction targets in 2050, using the building prototypes model and the retrofit data on existing building stock provided by the NEED dataset. These datasets can also be used for examining the outcome of a policy to adjust the policy promptly. For example, Boampong et al. [29] estimated the energy saving effect of a demand-side management program of air conditioner usage in the Gainesville region using the monthly electricity consumption dataset GRU. The outcome demonstrates notable seasonal variations in the effects of this demand-side management program.

The open datasets can not only be directly analyzed as input or outputs of the predictive models of energy consumption, but also can be used for providing essential parameters on urban building energy modeling (UBEM), and developing the city-scale simulation tools. Shen et al. [43] used the RECS dataset to provide features such as house type, floor area, energy consumption by fuel types, HVAC equipment types, and hot water fuel types for the building modeling of the five developed archetypical building types, to explore the energy saving potential for residential buildings in the United States. In addition, the empirical energy consumption data can be used for calibrating the simulation models. Leibowicz et al. [49] established building models in CitySim simulation software to quantify the energy-saving strategies in the aspects of fuel type transition, efficient appliance and thermal property improvement in residential buildings, and the demands of electrical appliance usage were calibrated using the DATAPORT dataset. The CityBes web app was also developed using multi-source open data [61]. By applying CityBes, Yixing Chen et al. used the building performance dataset from San Francisco (SF BM) to develop an automatic and rapid urban building energy model calibration method [62].

As the EUI can be converted into building anthropogenic heat, the energy consumption prediction can also contribute to urban climate research. For example, Satos et al. [52] evaluated and compared the spatial variability of buildings' anthropogenic heat between the different land use methods in Singapore, using the metered data from more than 13000 buildings in the datasets of EMA PR and EMA PB.

Datasets of high granularity (i.e., short time intervals) promote time series consumption prediction, which uses energy consumption data of a few time steps to predict future energy consumption. As fine-grained city-level datasets with high quality, DATAPORT and GENOME2 are widely used to carry out time series predictions on building energy consumption. The outputs of relevant research papers can be the electricity consumption profile of buildings at the time intervals from a day to a year. Generally, these papers use data-driven methods, such as Random Forests [53], Convolution Neural Network [54-56], and Long Short-term Memory Neural Network [56, 63]. The city-level datasets enable more comprehensive studies than those that only include energy consumption data for one building or a few buildings. For example, some scholars have also established a benchmarking scheme for these data-driven models with data from GENOME2 [64]. In addition, they attempted to solve deficiencies of machine learning, including poor interpretability and high dependence on data quantity. For model interpretability, Miller [65] tried to recognize and cluster the energy consumption behavior of groups of buildings, providing a basis for developing explainable models. For the data shortage problem, a Recurrent Generative Adversarial Network was applied to generate more data for the training [66]. Some scholars also used transfer learning techniques to make predictions for buildings lacking sufficient data [67]. It can be said that these datasets essentially promote the development of machine learning models to predict energy demand in buildings.

4.1.2 Influencing Factors of Building Energy Consumption

Table 4 summarizes the existing research on the influencing factors of building energy consumption. These studies established predictive models for building energy consumption and further analyzed the influencing factors. As shown in Table 4, some studies target examining influence from certain elements, and some try identifying essential factors. The categorization of literature may have a little overlap since some studies focus on more than one influencing factor.

Table 4. Application Summary: Research on Influencing Factors of Building Energy Consumption using Open Datasets

Research Category		**Typical Variables**	**Data Granularity**	**Typical Literature (Datasets)**
Influence of Specific Factors	Urban Form	Horizontal Compactness, Vertical Density, Variation of Building Heights, Urban Porosity, Roughness Length, Vegetation	year/month	[68] (SEA BM), [69] (CHI BM)
	Climate Change	Weather Condition, Climate Zone	year	[70] (BPD), [71] (RECS)
	Housing Type	Housing Type, Size, Building Age	year	[72] (NEED), [73] (CBECS),
	Occupants	Occupancy, Education, Income, Age Groups	year	[74] (RECS), [75] (CBD), [76] (RECS), [77] (RECS), [78] (RECS), [79] (RECS)
	Building Envelope and Geometry	Envelope Type (Cavity, Solid Wall, Attic Insulation), Window-to-wall Ratio	year	[80] (NEED), [81] (CBECS)
	Energy Appliance	Equipment Installation, Air-conditioner Use Frequency	year	[82] (ATX BM), [76] (RECS), [77] (RECS)
	Economic Factors	House Rent, Incentives,	year	[75] (CBD), [79] (RECS)
Factor Importance Identification		-	year/hour	[65] (GENOME2), [83] (CBECS), [84] (MHCLG), [85] (LL84), [86] (LL84), [87] (CBECS), [88] (RECS), [89] (NEED)

Influence from several factors on energy consumption can be investigated through open data. Here are some typical examples. For urban form, Ahn et al. [68] established a regression model to determine EUI using the building benchmarking dataset SEA BM in Seattle, considering factors related to urban form, including horizontal compactness, vertical density, and variation of building heights. For climate change, Fonseca et al. [70] combined building energy data from the BPD dataset, extensive weather data with Bayesian statistics, and first-principles building energy models to predict the potential impacts of climate change on buildings in 96 U.S. cities in the 21st century. The results showed that commercial buildings in hot/warm and humid climates should be at the top of the U.S. building sector climate action agenda. For house type, Summerfield et al. [72] used a sample of over 2.5 million gas-heated dwellings in England recorded in the dataset of NEED, gas consumption as the target variable, quantified differences in gas consumption by dwelling type, size, and age. For occupants, many factors can be investigated using the RECS dataset, which includes significant contextual information of the occupants, such as age, income and education information. For example, using RECS, Estiri Field [78] assessed the presence and shape of an age-energy consumption profile in the US residential sector. For building envelope, Gillich et al. [80] calculated expected energy saving in the UK through cavities, solid walls, and attic insulation and discussed the impact on future carbon budgets using the NEED dataset. For energy appliances, Rhodes et al. [82] estimated the effects of resolving typical air-conditioner design and installation issues such as low efficiency, oversizing, duct leakage, and low measured capacity on peak power demand and cooling energy consumption by using the building benchmarking dataset ATX BM in Austin. The influence of economic factors such as house rent and incentives can also be evaluated; Gui et al. [75] analyzed the data from the CBD dataset in Australia to study the correlation between EUI and commercial real estate factor and found that the "green building" brand effect makes office buildings more attractive for leasing and their energy performance is more reflective of changes in the commercial real estate market.

Studies for factor importance identification generally use data-driven methods such as Ridge regression, LASSO regression [85], and Random Forests [86] [87] for energy consumption prediction, then identify essential features. The yearly data research generally focuses on the contextual information of the occupants, buildings, and urban environment. For example, the feature importance analysis based on Random forests using the LL84 dataset [86] shows that residentials with lower household income and more residential complaints per capita are highly correlated with higher site EUI. And fine-grained datasets such as GENOME2 can enable researchers to study the impact of the time-related features [65], such as seasonal energy consumption features or different fluctuation characteristics in electricity consumption profiles.

4.1.3 Occupant Behavior

Occupant behavior is an important influencing factor of energy consumption in buildings, which is somewhat subjective and has lots of uncertainty. Therefore, many studies target analyzing occupant behaviors to predict energy consumption more accurately. Different from the studies of quantifying the impact of occupant-related factors mentioned in the above subsection, the occupant behavior here focuses more on the residents' subjective choices or behaviors, including choosing energy-related devices and daily consumption patterns, as shown in Table 5, rather than their contextual information such as income and age groups. For example, using the RECS dataset, Adua et al. Field [90] investigated the relationship between householders' energy-related behavior and residential energy consumption. The findings indicate that the impacts of efficiency technologies on energy consumption and the associated behaviors differ depending on the technology type. And the DATAPORT dataset has been widely adopted for Non-Intrusive Load Monitoring (NILM), which means load disaggregation. Zarabie et al. [90] adopted a machine learning approach to disaggregate the electricity load into fixed and shiftable components. Since DATAPORT also provides occupants' usage information of water-related devices, the researchers can also identify water-related events such as the usage of showers, clothes washers, and dishwashers [91]. To determine the occupants' daily pattern, yearly datasets such as RECS can be used to get higher-level family and appliance statistics, and the lower-level, individual activity data from other sources can be connected to each occupant category to identify typical consumption profiles by occupant types [92]. And with the fine-grained data from DATAPORT, the occupants' usage features, including frequency, consistency, and peak time, can be determined [93] to quantify the flexibility potential of the building, and the correlations in usage patterns of appliances can also be identified [94], to help distribution network operators with power network modeling and management.

Table 5. Application Summary: Research of Occupant Behavior using Open Datasets

Application Tasks	**Outputs**	**Data Granularity**	**Typical Literature (dataset)**
Choice of Energy-related Devices	Energy Consumption of Households, Augmented Data for Appliance Survey	year	[95] (RECS), [96] (RECS)
Non-Intrusive Load Monitoring	Disaggregated Electricity loads, Disaggregation of Water Event	15 mins	[97] (DATAPORT), [90] (DATAPORT), [98] (DATAPORT), [99] (DATAPORT), [91] (DATAPORT)
Daily Pattern Identification	Appliance Use Frequency, Appliance Consumption Profile by Occupant Types, Clusters by Appliance Use Schedules, Appliance Usage at Peak Time	year/15 mins	[100] (RECS), [92] (RECS), [93] (DARAPORT)
Operation Rules of Electricity Appliances	Correlations in Usage Patterns of Appliances	15 mins	[94] (DATAPORT)

4.1.4 Building Energy Performance Analysis

Most collected datasets aim to build energy performance benchmarking and improve supervision or management. For example, the Building Performance Database (BPD) is mainly used for: 1) expanding public access to building energy information; 2) peer group benchmarking for screening buildings that require high energy saving measures; 3) impact estimation of energy technologies and retrofit methods; 4) supporting risk-analysis based investment [4].

Table 6. Application Summary: Research on Building Energy Performance Analysis using Open Datasets

Category		**Outputs**	**Data Granularity**	**Typical Literature (dataset)**
Benchmarking Framework Establishment		Building Groups by Electricity Profiles/ Building Groups by Rating Bands/ Energy Consumption Baseline	year/month/15 mins	[101] (DATAPORT), [4] (BPD), [102](LL84), [103] (CBECS, LL84), [104] (MHCLG), [105] (SEA BM), [106] (LL84), [107] (LL84), [108] (GRU), [109] (LL84), [110] (LL84), [111] (BEBR), [112] (LL84)
Green Building Performance Evaluation		Energy Performance of Buildings with Green Marks	month	[113] (EMA PR, EMA PB)
Retrofit Decision Making	Multi-Matrix Retrofit Potential Evaluation	Equipment with Low Performance/ Predicted Building Performance	year	[84] (MHCLG), [114] (BER TOOL)
	Building Characteristic Determination	Energy System Types/ Building Types/ General operation strategies	hour/15 mins	[115] (DATAPORT), [116] (GENOME2), [117] (GENOME2)
	HVAC System Selection	Optimal Choice of HVAC System	year	[118] (CBECS), [119] (CBECS), [120] (CBECS, RECS)

With these purposes, the studies related to building energy performance analysis take a significant proportion. Table 6 summarizes building energy performance analysis research, which includes three general directions: benchmarking framework establishment, retrofit decision-making, and green building evaluation.

The city-level building samples with energy information enable horizontal comparison of the building's energy performance and promote the establishment of a building benchmarking framework. The outputs of the models in benchmarking frameworks can be building groups categorized by electricity profiles [101], rating bands [4], or energy consumption baseline for comparing with the actual energy consumption, which can be the average energy consumption for the peer group [110] or the expecting energy consumption of the building [112]. It is noteworthy that the open data provide essential resources for developing a novel data-driven benchmarking framework, which enables the policymakers to accurately and quickly determine the energy performance of many buildings at the city scale. Since the year 2018, data-driven benchmarking frameworks based on machine learning have been developed, such as DUE-B [110], which uses a Classification Regression Tree to perform the classification of building groups, then identify the abnormal building samples; EnergyStar++ [103], which applies Random Forests and XGB to calculate the expecting EUI, and then presents variable importance and model interpretations for policymakers.

Besides, the disclosed energy data can help evaluate green buildings' performance. For example, Agarawal et al. [121] examined the energy performance of the buildings with “Green Mark” (GM) certification in Singapore with the difference-indifference method, using the dataset of EMA PR and EMA PB. The results show no significant differences in energy consumption between GM-certified and non-GM-certified buildings.

Decision-making in building retrofit is an important application of building energy performance analysis. Multiple matrixes developed from the open datasets were used to evaluate the potential of building retrofit. Usman et al. [114] utilized the dataset BER MAP to get the building energy rating labels for a substantial number of buildings in Ireland. A machine learning model was trained using the data of the established building prototype and the building energy rating labels provided by BER MAP. The predicted building energy rating is combined with information from other resources (social, economic, and environmental data) to build up a multi-dimensional matrix to identify buildings with high retrofit value to determine the target buildings and areas for policymakers. Besides, before making retrofit decisions, the open datasets can be used to identify the principle building use, performance class, HVAC types, or general operation strategies of buildings, to correct the label of primary space usage type for a building [117], and reduce the expert efforts of inferring information for raw data [115, 116]. In addition, the abundant knowledge of buildings in the dataset can be used for training machine learning models to select optimal HVAC systems for building retrofit. For example, by using hundreds of high energy-efficient building samples with over 100 valuable features provided in the CBECS dataset, Tian et al. [118] developed a Bayesian Network to select the optimal HVAC systems for case buildings.

4.2 Grid Management

Buildings are the primary consumers in the power grids. Building energy datasets usually include power consumptions, which are essential for grid management. The city-level electrical load profiles are valuable for making reasonable grid operation decisions. The DATAPORT dataset is a high-quality building energy consumption dataset of fine-grained data. It provides energy consumption data by electric circuits in each household and includes indoor temperature, PV power generation, and energy storage data. Therefore, DATAPORT facilitates comprehensive studies on grid management, as shown in Table 7.

The flexibility potential can be quantified for demand response in grid management using DATAPORT. Yue et al. [99] identified the load patterns of intelligent home appliances and quantified their demand response flexibility. Afzalan et al. [93] proposed a data-driven method to quantify the potential of a single flexible load user to participate in demand response and used the data of DATAPORT to test the model empirically. The models for load shedding can also be established based on DATAPORT. Cole et al. [122] built up a dynamic model of a 900-home community using the empirical data of DATAPORT, conducted coordinated control based on a reduced-order modeling strategy and economic model for many residential air-conditioning systems, and achieved substantial reductions in the peak electricity demand.

Table 7. Application Summary: Research of Grid Management using Open Datasets

Research Category		**Outputs**	**Dataset**	**Data Granularity**	**Typical Literature**
Demand Response	Flexibility for Demand Response	Load Patterns, Flexibility Potential	DATAPORT	15 mins	[93], [99]
	Load Shedding	Peak Demand, Total Demand, Scheduling of Appliances, Cost			[50], [122], [123]
Frequency Regulation		Regulation Strategies on HVAC, Regulation Qualification Signals			[124]
Electricity Market		Electricity Consumption/Price, Thermal Comfort			[125], [126], [127], [128], [129]
Energy System Operation and Management	PV Power Generation	Size/Tilt/Azimuth of Solar Panel, Anomalies of Solar Arrays			[130], [131], [132], [133]
	Energy Storage	Demand Peak, Discharging Strategy			[134], [135], [136], [137], [138]
	Electric Vehicle	Peak demand, Charging Schedule			[139]
	Distributed Energy System	Operation Strategies, Optimization Scheme, Cost, Profit			[140], [141]

The HVAC devices on the demand side can also be aggregated to support the frequency regulation of the grid. Since DATAPORT also provides information about the power of HVAC devices, it can also be used for frequency regulation. Abbas et al. [124] established a model of 400 households based on DATAPORT. It used the samples from DATAPORT to train a data-driven model for determining the relationship between setpoint offset, total HVAC power consumption, and HVAC power variation for devices within a specific cluster. This relationship was used for frequency regulation.

The research on the electricity market includes the studies about electricity pricing strategy [125] [127] and tariff [126], the impact of pricing strategy on thermal comfort [128], as well as characteristic analysis of a simulated blockchain electricity market [129]. Some studies used DATAPORT as field experiment data on pricing the electricity consumption because there was a summer program during the data collection period that compared information provision and conservation appeals to crucial peak pricing throughout two summers when the wholesale price of power is significantly higher than the retail tariff [125]. Like the studies of demand response, some of the studies also used the building information in DATAPORT to build up dynamic models to support investigations about the electricity market. For example, Devine et al. [129] proposed a demurrage mechanism for the blockchain electricity market by creating a simulated market from solar energy generation data in DATAPORT.

The R&D on energy system operation and management typically include PV power generation, energy storage, electric vehicle, and distributed energy system. For example, Mason et al. [130] used the features from customer electricity profiles as the inputs to train the deep neural networks and get the parameters of PV panels. For energy storage, Odonkor et al. [134] designed an efficient charging dispatch policy for a shared battery system using the reinforcement learning method. For an electric vehicle, Dang et al. [139] presented a mechanism for managing electric vehicle charging behavior to mitigate grid impacts by using the continuous electric vehicle load data from DATAPORT. For distributed energy system, Jones et al. [141] maximized the benefit of a distributed energy and water and energy system on the community scale by developing a mixed-integer linear program that optimizes distributed technology capacities and hourly dispatch and using the data from DATAPORT for testing, which includes empirical data on rooftop solar power generation, as well as water and electricity demand profiles.

4.3 Socio-economic Analysis

Energy usage is an important economic issue for the residents and an influencing factor for house prices and rent. For example, the acquisition of a heating or cooling service can sometimes be a burden to a family; the rent of the house can be influenced by the energy performance. Therefore, the energy data of the building is required for socio-economic analysis. The research directions can be generally divided into three aspects: energy poverty, market demand, price, and rent, as summarized in Table 8.

A few datasets are rich In social, and economic information about the occupants, such as Residential Energy Consumption Survey (RECS) and Commercial Building Energy Consumption Survey (CBECS) datasets; they are convenient to be used for social and economic research. For example, the RECS dataset includes the residents' income and fuel expenditure levels, the occupants’ identities as owners or tenants, and whether or not the occupants pay the utility bill themselves; therefore, it can promote the social economics analysis in most aspects. Other datasets with energy benchmarking data, such as LL84, or energy performance certificate information, such as MHCLG, can also be aggregated with datasets of social surveys for analysis.

Table 8. Application Summary: Research of Social Economics Analysis using Open Datasets

Category		Outputs	Data Granularity	Typical Literature (dataset)
Energy Poverty		Marginal Residential Energy Prices, Home-heating Energy-poverty Risk, Geographical Distribution of Energy Poverty/ Burden, Relationship of Energy Demand and Price/Racial, Socioeconomic Factors, Relationship of Fuel Poverty and Neighborhood Characteristics	year	[142] (RECS), [143] (MHCLG), [144] (RECS), [145] (RECS), [146] (RECS), [147] (BERDO, CAMB BM, LL84, SEA BM, CLB BM), [148] (MHCLG), [149] (BER MAP)
Price and Rent	House Price/Rent	Influence of Energy Performance Certification/Energy Efficiency on Property Sell/Rent Price, Influence of Landlords' Underinvestment on Energy Consumption	year	[150] (MHCLG), [151] (BER TOOL), [152] (RECS), [153] (CENED+2)
	Appliance Price	Rebound Effect on EnergyStar-labeled Devices	year	[154] (RECS)
Market Demand		Market Demand Potential for Insulation Materials	year	[155] (NEED)

As shown in Table 8, energy poverty takes a significant proportion in socio-economics studies. Energy poverty can be defined as a family's in-affordability of warmth/cooling because they live in an energy-inefficient home [146]. The geographical distribution of energy poverty in a country is an essential output of the most relevant studies. Mohr et al. [142] used the RECS dataset to quantify energy poverty in each household with the variables about energy demand and income. Then, using a logit prediction model, they identified significant factors that can lead to energy poverty in different areas of the United States. Ahmed et al. [144] also used RECS to investigate the energy poverty status of public housing residents of the United States. The findings suggest that the housing authorities can help to reduce energy consumption effectively and then ease the energy poverty status of the residents.

There are studies investigating the relationship between energy-related factors and the price and rent of houses and appliances. Sejas et al. [150] studied the influence of energy performance certification on the market values of properties by using the MHCLG dataset, which includes a large number of energy performance certificates in the UK. It made several recommendations for how to best use these threshold effects to set the energy performance certificate standard better. For electricity appliances, Sun [154] studied the rebound effects of EnergyStar-labeled devices using the RECS dataset. The rebound effect means that the increased efficiency of the appliances often leads to lower costs, making it possible to purchase more improved products or other products or services. The data from RECS can reflect the energy efficiency and frequency of use of appliances (e.g., dishwasher, air conditioner). They can aid in the development of an empirical model to test the predicted relationships between EnergyStar appliances and their frequency of use. Negative and positive rebound effects were found for different appliances.

Besides, the building energy dataset can be applied to estimating the market demand potentials of the insulation materials of the building envelope. Varriale et al. [155] used the statistics in the NEED dataset to model the distribution of dwelling type, age, and size and to estimate the demand for certain building insulation measures from the domestic sector in Wales from 2016 to 2050.

4.4 Policy Implications

Since urban-level building energy data can provide building energy characteristics for a district or city, the outcome of relevant research can be utilized to support policymakers in many aspects. Although policy impact hasn't been referred to directly in some papers, policy implications can always be derived based on their results. In other words, most of the papers offer policy implications in a direct or indirect way. Therefore, this section will summarize policymakers' insight based on the collected literature results. Relevant policies can be categorized as the target group and target factor identification for new policy-making, impact analysis of policy, and the pricing of energy and electrical appliances. The corresponding research directions are listed in Table 9.

Research in many directions contributes to determining the targeted groups for new policies, such as city or regional building energy-saving renovation plans, energy regulations for electrical appliances and household energy-saving renovation subsidies. The target groups can be specific clusters of buildings, appliances, or residents. The research directions of energy consumption mapping, benchmarking framework, and retrofit decision-making have a similar way of showing their policy implications for target group identification. Many studies of these directions conduct analysis based on peer group comparison to identify the performance and outliers of each cluster. Policymakers can target underperformers to establish energy efficiency requirements [38, 156] or determine the building groups with the most considerable retrofit potential integrating the energy performance with social economics factors [114]. For urban climate, as mentioned in Section 4.1, the output of urban heat flux of high spatial resolution can show hotspots and the biggest energy consumer in the city so that the climate controllers can incorporate additional information about urban typology and conduct corresponding reactions [52]. The studies of occupant behavior in [95, 96],appliance choices can help policymakers regulate target occupant clusters' preference for appliances or consumption habits by incentivizing field [95, 96] and adjusting the energy regulations for certain energy-efficient electrical appliances. For the research of energy poverty, the quantified energy cost burden of high spatial resolution enables researchers to develop targeted, and proactive policies, including energy efficiency incentives and regulations based on measurable energy performance, subsidies for specific energy retrofits connected to building attributes, and affordable housing schemes that include rental subsidy amounts [147].

Besides target groups, policymakers must also be aware of the influence of different factors and determine the priority for policy making. As mentioned in Section 4.1, the study of influencing factors can quantify the impact of other factors on energy consumption. The target factors include urban form, housing type, building envelope, and economic factors.

Table 9. Policy Categories and Directions of Corresponding Studies

Policy Category	**Research Direction**
Target Group Identification for New Policy Making	Energy Consumption Mapping
	Urban Climate
	Occupant Behavior
	Benchmarking Framework
	Retrofit Decision Making
	Energy Poverty
Target Factor Identification for New Policy Making	Influencing Factors
Impact Analysis of Policy	Effect of Energy-saving Policies
Energy Pricing	Energy Poverty
	Demand Response
	Electricity Market

Policymakers also need to evaluate the effects of established policies or policies not yet published to understand their impacts in practice and estimate feasibility, respectively. There is, therefore, a need for studies on the policy response. As discussed in Section 4.1, the effect of energy-saving policies or measures is an important research direction. The large-scale energy-saving program is the most common measure to be evaluated, including the retrofit strategies on the envelope and various energy devices. The effect of benchmarking policies and data disclosure schemes can also be evaluated [157].

According to the above illustration, research based on urban-level open datasets can significantly support policymakers in many aspects. This review also proposes a novel perspective, using open datasets to guide policy formulation. The summary and categorization of policy implications can provide an informative and convenient guideline for policymakers and researchers.

5. Discussion

Most urban energy studies based on open datasets use building energy use datasets and take advantage of other open data sources, such as geographical information, occupant status, social and economic factors. This section discusses non-energy-related datasets and privacy issues related to data disclosure to highlight the opportunities and challenges of open-source datasets for urban building energy research. Section 5.1 will briefly introduce other valuable public datasets and use cases. Considering privacy is always a critical issue in data disclosure, Section 5.2 summarize and proposes approaches to protecting privacy in data disclosure.

5.1 Values of Heterologous Data Sources

As mentioned in Section 3, the open energy consumption datasets may contain several non-energy variables such as building category, floor area, occupancy, and household information. These variables play essential roles in research of most directions, whether it is energy consumption forecasting, building energy benchmarking, grid management, or social or economic analysis. Sometimes these non-energy variables are used to establish physical models for simulation or train the data-driven models to get more reliable results and investigate their underlying relationships with energy consumption. However, it can be concluded from Figure 3 of Section 3 that the proportion of energy consumption datasets that can additionally provide detailed building information is tiny. Therefore, the energy datasets themselves often need to be more capable of providing sufficient information to support relevant research. For this case, external non-energy datasets are required.

Non-energy datasets are commonly used to provide important information not included in the open energy datasets, such as weather data, building characteristics, urban forms, and occupant-related factors. This section will introduce datasets from heterologous sources, widely used in previous studies as supplement data for energy consumption datasets.

Weather is an essential factor in building energy use. Many cities provide publicly available weather data on local observatories or weather station websites. For global weather, some platforms offer data to the public. The Meteonorm dataset [158] involves data from 8,325 weather stations worldwide, which offers monthly weather information for every location on the earth; variables of temperature, humidity, wind data, the duration of the sun, and the number of rainy days are recorded. These weather features are essential in cross-city analysis, especially when the locations are in different climate zones. The dataset MERRA2 [159] also provides worldwide weather information. This dataset is based on satellite imaging and records data in spaced grids, making it independent of ground observatories. And for studies aiming at assessing the impact of climate change, the Climate Change World Weather File Generator (CCWorldWeatherGen) [160] can be a helpful tool, as it generates future climate change weather files for any place in the world.

There are some widely used datasets for building information, such as Better Building Neighborhood Program [161] and Primary Land Use Tax Lot Output (PLUTO) [10] datasets. The Better Building Neighborhood Program dataset contains records on over 75,000 upgraded single-family buildings by region and zip code, covering information about facilities and their energy-related characteristics. The PLUTO dataset presents comprehensive and detailed building information on New York City, including the location of the census, property type, area by usage, floor number and land value. With many useful features and samples, PLUTO is widely used to establish data-driven models of building energy consumption prediction. When merged with the LL84 dataset, it has been used for predicting EUI for 1.1 million buildings in NYC with a very high accuracy [20].

In addition, building permits, which can be found on the open data portal of many cities' governments, can sometimes provide helpful information, such as altered building type, system categories of upgrading, and construction year. However, these records have problems, such as ununified format, lack of quantitative description, and neglect of the energy system renovation [162].

Data from some geographical information systems (GIS) is also quite helpful. For example, OpenStreetMap, as a world map, offers data that can be exported for selected areas. A list of tags such as building name, height, type, and 2D footprint is involved. Ali et al. (2020) aggregated building energy data and GIS data from the Irish Energy Performance Certificates dataset and OpenStreetMap, respectively, producing maps of four dimensions of building performance, environment, and social and economic condition [109]. After coupling, a multi-criteria decision analysis map is made for energy planning and analysis and can be an intuitive tool to support decision-making. With information from GIS, policymakers can have a broader perspective to make optimal decisions.

Moreover, researchers can investigate more factors related to the occupants using survey data. For example, the Living in Wales Household Survey [111] contains a population of 2741 Welsh dwellings and provides statistical information about households and the financial condition of homes in Wales. The Census American Community Survey Data [112] provides detailed demographic, social, economic, commuting, and housing statistics in the United States. These datasets can be used for building prototype establishment and quantifying energy burden. Besides, The American Time Use Survey [36] provides nationally representative estimates of how, where, and with whom Americans aged 15 years old and up spend their time; the occupant behavior models can also be built with these data.

5.2 Privacy Issues

Privacy is an inherent concern of data owners to determine whether the data can be published or to what extent detailed building information can be collected and disclosed. Assigning identifiers to buildings that help to conceal real building information is a widely adopted approach to protecting privacy. This section will summarize the identifiers of individual buildings in the collected datasets and discuss the factors that may contain privacy-sensitive information, including the positioning accuracy of buildings and the time interval of sample data. A few possible privacy-protecting approaches during data disclosure are proposed below.

Building identifiers, such as their name, address, zip code, longitude and latitude, In some datasets, the buildings could be precisely located. Still, in others, they could only be found roughly at the district or city level. Appendix C summarizes the building identifiers used by the reviewed datasets. The parameters for identifying the properties are listed, and the locating accuracy is classified into four levels: city/region, district, small area (community), and individual property.

According to the identifier summary, most datasets' buildings can be located precisely to individual properties. The authors found that the positioning accuracy is also related to other sensitive information such as household data (e.g., income, gender, and age groups) and the sampling interval of the dataset since occupant behaviors can be conducted from high-frequency data. According to Appendix B, datasets with household information include BPD, RECS, DOE SP, MHCLG, and DK HEAT. Among these datasets, only MHCLG can locate each sample in individual buildings, while the rest can only be traced at the district or city level. And for the time interval, as summarized in Figure 5, almost all datasets that can be used to locate buildings to the individual property have a time interval of a year. High-frequency datasets have very low positioning accuracy, as buildings from these datasets are all strictly anonymized and can only be located at city or region levels. A straightforward implication can be derived from this: for data disclosure, there is a trade-off between spatial accuracy and sensitive information. Data cannot be disclosed with accurate location and sensitive information at the same time while respecting privacy, especially for a massive number of buildings in cities. For buildings where energy consumption supervision is required, it is reasonable to disclose yearly or monthly data to make the public aware of energy performance.

In cases where buildings can be located precisely, geographical features, such as urban form and vegetation coverage, can also be used to establish a more convincing model. Also, accurate location enables spatial visualization, providing more intuitive insight to policymakers. So, a trade-off should be made. Providing some information at the district or census area level may be a feasible choice: this is another data aggregation method. For instance, the BER map published by SEAI includes 18,580 small areas consisting of buildings with similar characteristics in Ireland. Also, Ma and Cheng [109] established a GIS-based urban energy use model by leveraging household information aggregated by blocks and energy data of individual buildings in NYC.

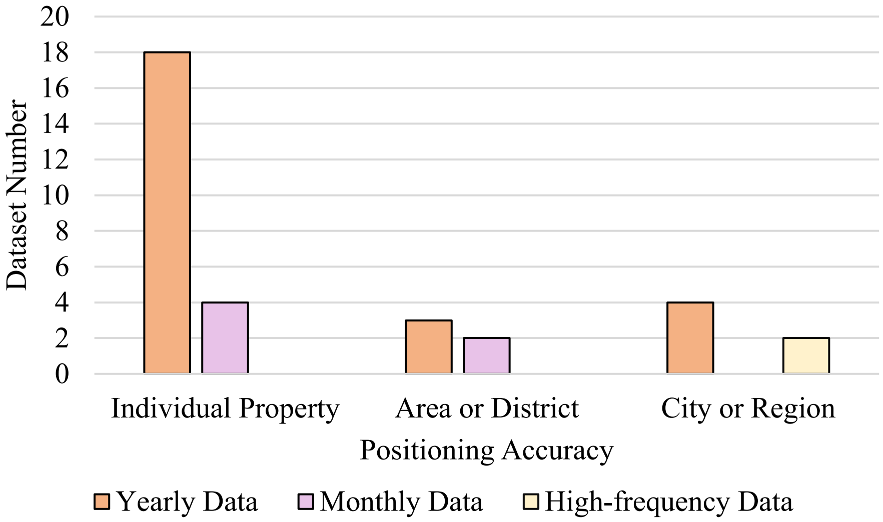

Figure 5. Positioning Accuracy for Datasets with Different Time Intervals

Proper anonymizing is also required for privacy protection. Replacing building names or addresses with codes can be one possible way. Furthermore the cross-dataset analysis would be much more convenient if each public dataset contained unique building identification codes [163]. Based on encoding, some models exist for the protection of time-series data. For example, the PAD framework applies a model of k-anonymity to protect privacy [164]. Publishers can determine the data use purpose and privacy level, and this framework will then choose the proper encoding method to minimize sensitive information.

Moreover, for data collection, privacy issues can be settled with data aggregation. The data from tenants' individual-metered accounts for commercial and multifamily buildings are generally protected. But according to researchers, it is possible to approximate average building consumption based on only a proportion of meters. Previous studies have made significant attempts on aggregating utility meter data at certain thresholds according to the quantified privacy risk metrics [165]. And in several states in the United States, the aggregation thresholds of meters have been adopted by a few utility companies [7].

The privacy issue is a crucial problem to be solved in promoting building energy data disclosure. Generally, data should not be published with sensitive information and the accurate location simultaneously to protect individuals' privacy. For example, suppose the data contains household information such as income, gender, and age groups or very high-frequency data. In that case, it’s better to publish it without accurate building locations and vice versa. And by using the methods of anonymous encoding and data aggregation, privacy can be protected more effectively.

6. Conclusion and Reflection

This paper presents a comprehensive, systematic literature review on worldwide open city-level building energy use datasets. Since most cities only disclose aggregated data of the whole town, this review targeted datasets containing individual buildings' information. The used approaches for collecting city-level open data are introduced, which could provide an effective guideline for researchers. There are 33 datasets collected in total, and their details and URLs (where available) have been summarized in this paper, providing great convenience for scholars in relevant areas. Overall, the U.S. has a much greater number of datasets than other countries. Most data have a yearly frequency, and the higher the time frequency, the fewer available datasets.

Relevant research using these datasets is also collected and summarized to clarify the potential applications based on these open datasets and bring more insight for both researchers and policymakers. Over 198 papers are categorized by application scenarios of open datasets. It can be found that most of the studies using city-level building energy data are related to energy policy in the aspects of determining target groups and factors for new regulations, impact analysis for policies and pricing for energy or electricity devices. The summary of the corresponding research category for each policy direction can provide guidelines for both policymakers and researchers, informing them of what kind of analysis they should perform for a specific scenario.

In addition, this paper gives a brief introduction to non-energy datasets used in the research above, including weather datasets and building/household information data. By merging these datasets with building energy data, more dimensions will be covered for the relevant research, and more accurate and comprehensive models will be established.

The authors also conducted an exploratory analysis of building identifiers in the collected datasets to address privacy issues and promote disclosure. The statistics imply that for privacy protection, the frequency and sensitive information in the data should not be high at the same time. A few methods of protecting sensitive information have also been introduced.

The main conclusions and reflections of this paper are as follows:

(1) Even though there is a significant amount of open data available in developed countries such as the U.S., the U.K., Ireland and Singapore, city-level building energy data disclosure is still at a preliminary stage in most other countries. It is undeniable that more research papers on various topics related to urban building energy use are published for cities with more open data. As a result, these cities benefit significantly in terms of urban building planning, design, and operation, and energy policymaking. This further verifies the need for more data disclosure.

(2) Currently, the focus of data disclosure is yearly data. In contrast, the publishing amount of monthly data still has great potential to enable more research and potential applications that yearly data cannot cover.

(3) Data publishers can consider more utilization cases and establish multiple usage purposes, like building energy benchmarking, decarbonization in buildings, and district energy system planning,

before collecting and publishing data so that variables required for corresponding analysis will be available. Moreover, the level of privacy can also be defined based on the usage purpose.

(4) Assigning an identifier to an individual building and using the identifier rather than the name of the building in the open datasets is a promising way to protect building privacy. Moreover, a unified identifier for each building is beneficial for conducting studies using multiple datasets in one city or district. The categorizing methods across datasets should also be unified. For example, the building and household categories should share the same classification standard in all datasets from the same city to avoid controversy when merging datasets.

It is expected that urban building energy data disclosure will be implemented in more cities in the near future, and more efficient models for policymaking will be established based on open data. Reliable data-driven decisions should result concerning building energy-efficient retrofit, energy policy and energy planning, and more effective building benchmarking methods. Energy conservation and decarburization in cities will also be attained in more reliable and cost-effective ways.

Acknowledgments

The authors gratefully acknowledge the support of this research by the National Key Research and Development Program of China (2021YFE0107400) and the Research Grant Council of the Hong Kong SAR (152133/19E).

Appendices

Appendix A. Detailed Information of Collected Open Datasets

Title	Organization	City/Area	Sample Number	Time Interval	Time Range	URL	Research Paper Number *
Building Performance Dataset (BPD)	U.S Department of Energy	The United States: 11 Cities/Districts	800-40000 residential/commercial units	Year	1 year – 5 years	https://bpd.lbl.gov/	10
Commercial Building Energy Consumption Survey (CBECS)	U.S. Energy Information Administration	The United States	6720 commercial buildings	Year	31 years (Since 1989)	https://www.eia.gov/consumption/commercial/data/2012/index.php?view=microdata	16
Residential Energy Consumption Survey (RECS)	U.S. Energy Information Administration	The United States	5600 residential units	Year	42 years (since 1978)	https://www.eia.gov/consumption/residential/index.php	32
DOE Buildings Performance Database, [sample Residential data] (DOE SP)	Office of Energy Efficiency and Renewable Energy	U.S.: Dayton, Gainesville	29393 residential units	Month	1 year	https://data.openei.org/submissions/182	2
Dataport Database (DATAPORT)	The Pecan Street Research Institute, University of Texas	U.S.: Texas, Colorado, California, New York City	1115 residential units	1 sec - 15 mins	6 years	https://dataport.pecanstreet.org/ Free access for university faculty, staff and students for academic research purposes.	42
Building Data Genome Project 2 （GENOME2）	BUDS-LAB	U.S.: Eastern, Central, Mountain, Pacific Districts; London, Dublin	3053 energy meters from 1636 buildings	1 hour	2 years	https://github.com/buds-lab/building-data-genome-project-2	11
Energy and Water Data Disclosure for Local Law 84 （LL84）	Mayor's Office of Sustainability	U.S.: New York City	28067 buildings	Year & Month	9 years (2012 - 2020)	https://data.cityofnewyork.us/Environment/Energy-and-Water-Data-Disclosure-for-Local-Law-84-/usc3-8zwd	16
Local Law 87 Energy Audit Data (LL87)	Mayor's Office of Sustainability	U.S.: New York City	6176 buildings	Year	7 years (2012 - 2018)	https://data.cityofnewyork.us/Environment/LL87-Energy-Audit-Data/au6c-jqvf	3
GRU Customer Electric Consumption (GRU)	Gainesville Regional Utilities	U.S.: Gainesville	100405 consumers	Month	8 years (2012-2019)	https://data.cityofgainesville.org/Utilities/GRU-Customer-Electric-Consumption-2012-2021/ba7j-nifw	1
Chicago Energy Benchmarking (CHI BM)	City of Chicago Sustainability Program	U.S.: Chicago	3485 buildings	Year	2014-2018	https://data.cityofchicago.org/Environment-Sustainable-Development/Chicago-Energy-Benchmarking/xq83-jr8c	2
National Energy Efficiency Data-Framework (NEED)	Department for Business, Energy and Industrial Strategy, UK	The United Kingdom	55154 records	Year	13 years (2005-2017)	https://www.gov.uk/government/statistics/national-energy-efficiency-data-framework-need-anonymised-data-2019	17
Energy Performance of Buildings Data of England and Wales (MHCLG)	Ministry of Housing, Communities & Local Government	UK: England, Wales	20125562 domestic units, 923599 non-domestic units, 390812 Display Energy Certificates	Year	15 years (2005-2019)	https://epc.opendatacommunities.org/domestic/search	6
Ireland Building Energy Rating Map (BER MAP)	Sustainable Energy Authority of Ireland	Ireland	18580 small areas	Year	1 year	https://www.seai.ie/technologies/seai-maps/ber-map/	6
Database CENED+2 (CENED+2)	Lombardy Region, Lombard Infrastructures	Italy: Lombardy Region	947814 records	Year	1 year	https://www.dati.lombardia.it/Energia/Database-CENED-2-Certificazione-ENergetica-degli-E/bbky-sde5	2

Title	Organization	City/Area	Sample Number	Time Interval	Time Range	URL	Research Paper Number

Austin Energy Conservation Audit and Disclosure Data (ATX BM)	Austin Energy	U.S.: Austin	9288 residential units; 1493 multifamily units; 2696 commercial units	Year	11 years (since 2009)	https://catalog.data.gov/dataset/2015-2017-ecad-residential-audit-data;
						https://catalog.data.gov/dataset/2019-multifamily-ecad;
						https://catalog.data.gov/dataset/2016-ecad-commercial-reported-data
							16

Berdo disclosure: REPORTED ENERGY AND WATER METRICS (BERDO)	Environment Department of Boston	U.S.: Boston	2427 non-residential and residential units	Year	2015-2019	https://data.boston.gov/dataset/building-energy-reporting-and-disclosure-ordinance/resource/033c30b4-8d28-40ad-9572-43d8455aaab6	
Chicago Energy Usage 2010 (CHI Usage)	Aggregated from ComEd and Peoples Natural Gas by Accenture	U.S.: Chicago	67051 community units	Month	1 year	https://data.cityofchicago.org/Environment-Sustainable-Development/Energy-Usage-2010/8yq3-m6wp	
Existing Buildings Energy & Water Efficiency Program (LA EBEWE)	LA Department of Building and Safety (LADBS)	U.S.: Los Angeles	40798 records	Year	-	https://data.lacity.org/City-Infrastructure-Service-Requests/Existing-Buildings-Energy-Water-Efficiency-EBEWE-P/9yda-i4ya	
Existing Buildings Energy Performance Ordinance Report (SF BM)	San Francisco Department of Environment	U.S.: San Francisco	2630 buildings	Year	Since 2010	https://data.sfgov.org/Energy-and-Environment/Existing-Buildings-Energy-Performance-Ordinance-Re/j2j3-acqj	
Building Energy Benchmarks (CLB BM)	Department of Energy and Environment (DOEE)	U.S.: District of Columbia	15335 buildings	Year	-	https://opendata.dc.gov/datasets/building-energy-benchmarking/explore?location=38.890801%2C-77.021832%2C12.37	
BESO Large Building Energy Data and Compliance Status (BESO)	Office of Energy & Sustainable Development (OESD)	U.S.: City of Berkley	307 large buildings	Year	3 years (2017, 2018, 2019)	https://data.cityofberkeley.info/Energy-and-Environment/BESO-Large-Building-Energy-Data-and-Compliance-Sta/5vy5-rwja	
Building Performance Program (BLDR BM)	City of Boulder Government	U.S.: City of Boulder	420 City-owned buildings	Year	1 year	https://open-data.bouldercolorado.gov/datasets/af7d4e7ad9cc40debbe0251b21236907_0?_ga=2.131433807.1346296851.1609396523-1634115722.1609396523	
Cambridge Building Energy and Water Use Data Disclosure 2016-2019 (CAMB BM)	Community Development Department	U.S.: City of Cambridge	1049 buildings	Year	4 years (2016-2019)	https://data.cambridgema.gov/Energy-and-the-Environment/Cambridge-Building-Energy-and-Water-Use-Data-Discl/72g6-j7aq	
Building Energy Benchmarking Results (MCPL BM)	Department of Environmental Protection	U.S.: Montgomery County	156 county-owned buildings	Year	1 year	https://data.montgomerycountymd.gov/Environment/Building-Energy-Benchmarking-Results/izzs-2bn4	
Large Building Energy Benchmarking Data (PHI BM)	Office of Sustainability	U.S.: Philadelphia	4204 properties	Year	6 years (2013-2018)	https://www.opendataphilly.org/dataset/large-commercial-building-energy-benchmarking	
Building Energy Benchmarking (SEA BM)	Office of Sustainability and Environment	U.S.: Seattle	3581 buildings	Year	5 years (2015-2019)	https://data.seattle.gov/dataset/2019-Building-Energy-Benchmarking/3th6-ticf	
National BER Research Tool (BER TOOL)	Sustainable Energy Authority of Ireland	Ireland	947200 residential units	Year	1 year	https://ndber.seai.ie/BERResearchTool/ber/search.aspx	
District heating energy efficiency of Danish building typologies: Datasets and supplementary materials (DK HEAT)	Aarhus University, Aarhus Municipality	Denmark	42969 buildings located in Aarhus, Denmark	Year	1 year	https://data.mendeley.com/datasets/v8mwvy7p6r/1	1

Title	Organization	City/Area	Sample Number	Time Interval	Time Range	URL	Research Paper Number
Building Efficiency Register - Commercial Building Disclosure (CBD)	Department of the Environment and Energy	Australia	more than 10000 properties	Year	10 years (2011-2020)	https://www.data.gov.au/dataset/ds-dga-5b667f9c-48c8-4b14-af1b-670c92f63def/details	7
Voluntary Disclosed Building Energy Performance Data (SG BP)	Building and Construction Authority (BCA) Singapore	Singapore	1244 buildings	Year	2 years (2017,2018)	https://data.gov.sg/dataset/building-energy-performance-data?resource_id=17c5cf53-9b5c-428b-a5b7-d88b8071d4af	2
Average Monthly Household Electricity Consumption by Postal Code (private apartment) (EMA PR)	Energy Market Authority Singapore	Singapore	9912 buildings of private apartment	Month	2 years (2015,2016)	https://www.ema.gov.sg/cmsmedia/Publications_and_Statistics/Statistics/2RSU.pdf	5
Average Monthly Household Electricity Consumption by Postal Code (public housing) (EMA PB)	Energy Market Authority Singapore	Singapore	9540 buildings of public housing	Month	6 months (2016 month 1-6)	https://data.gov.sg/dataset/hdb-property-information?resource_id=482bfa14-2977-4035-9c61-c85f871daf4e	
BCA Building Energy Benchmarking Report (BEBR)	Building and Construction Authority (BCA) Singapore	Singapore	500 – 1000 buildings per year	Year	8 years (2013 – 2020)	https://www.bca.gov.sg/BESS/BenchmarkingReport/BenchmarkingReport.aspx	1

Research paper number [*]: the number of research papers using the corresponding dataset.

Appendix B. Variable Information of each Dataset

Abbreviation of Dataset	Variable Number	Categories of Variables												
		Property Identification	**Building Type**	**Floor Area**	**Energy Usage**	**Water Usage**	**Energy Rating**	**Year Built**	**Occupancy**	**Envelope Condition**	**Household Information**	**Energy Device Installation**	**GHG Emission**	**Other**
BPD	31	√	√	√	√		√	√			√	√		
CBECS	1019	√	√	√	√			√	√	√		√		
RECS	759	√	√	√	√			√	√	√	√	√		√ *1
DOE SP	33	√	√	√	√			√			√	√		
DATAOPORT	129	√	√	√	√			√				√		√ *2
GENOME2	31	√		√	√	√	√	√	√					
LL84	67	√	√	√	√	√	√	√	√					
LL87	693	√	√	√	√	√	√	√		√		√		
ATX BM	78	√	√	√	√		√	√		√		√		√ *3
BERDO	28	√	√	√	√	√	√	√					√	
CHI BM	29	√	√	√	√	√	√	√					√	
CHI Usage	73	√	√		√			√	√					
LA EBEWE	73	√	√	√	√	√	√	√	√				√	
SF BM	113	√	√	√	√		√	√					√	
CLB BM	71	√	√	√	√	√	√	√					√	
BESO	23	√	√	√	√		√							
BLDR BM	38	√	√	√	√		√							
CAMB BM	42	√	√	√	√		√	√						
MCPL	19	√	√	√	√		√	√						
PHI BM	23	√	√	√	√		√	√	√					
SEA BM	42	√	√	√	√		√	√					√	
GRU	9	√			√									
NEED	18	√	√	√	√			√				√		
MHCLG	90	√	√	√	√	√			√	√	√	√	√	√ *4
BER MAP	505	√	√	√	√		√	√		√		√		
BER TOOL	20	√	√	√	√		√			√			√	
CENED+2	205	√	√	√	√		√	√				√	√	√ *5
DK HEAT	13	√	√	√	√			√			√			√ *6
CBD	62	√			√		√						√	
SG BP	11	√	√	√	√		√							
EMA PR	5	√			√									
EMA PB	13	√			√									
BEBR	23	√	√	√	√		√							

*1 Facilities (hot tub, swimming pool, etc.)
*2 Indoor temperature
*3 Energy bill, toilet type, water system information (duct, plumb)
*4 Energy bills, improvement recommendation
*5 Appliance efficiency
*6 Annual relative daily change, annual relative cooling temperature

Appendix C. Building Identifiers of each Dataset

Dataset Title	Identification Variables (number of the variables)	Building Type		Positioning Accuracy			
		Residential	Commercial	Individual Property	Small Area	Urban District	City/ Region
LL84, LL87	Property Id, Property Name, Parent Property Id, Parent Property Name, City Building, BBL - 10 digits, NYC Borough, Block and Lot (BBL) self-reported, NYC Building Identification Number (BIN), Address 1 (self-reported), Address 2, Postal Code, Street Number, Street Name, Borough (14)	√	√	√			
MHCLG	LMK_KEY, ADDRESS1, ADDRESS2, ADDRESS3, POSTCODE, BUILDING_REFERENCE_NUMBER (6)	√		√			
ATX BM	Property ID，Property Name，Property Management Company Name， Full Street Address (4)	√	√	√			
BERDO	Property Name, Address, ZIP (3)	√	√	√			
CHI BM	ID, Property Name, Address, ZIP code (4)	√	√	√			
GRU	Service Address, Service City (2)	√	√	√			
LA EBEWE	BUILDING ADDRESS, BUILDING ID (2)	√	√	√			
SF BM	Parcel(s), Building Name, Building Address, Postal Code, Full Address (5)	√	√	√			
CLB BM	Longitude, latitude, OBJECTID，PID，DC REAL PROPERTY ID，PM PROPERTY ID， PROPERTY NAME,PMPARENT PROPERTY ID，PARENT PROPERTY NAME (9)	√	√	√			
BESO	BESO ID， Building Name, Building Address (3)	√	√	√			
BLDR BM	ObjectId, Building ID, Street (3)		√	√			
MCPL BM	Building ID, MapLot, Buildings Included, Parent Building ID, Assessor Address, Address Point GIS (6)	√	√	√			
CAMB BM	Property Name, Address, City, Zip code, Property ID (6)		√	√			
PHI BM	objected, portfolio_manager_id, street_address, property_name, postal_code (5)		√	√			
SEA BM	OSEBuildingID, BuildingName, TaxParcelIdentificationNumber, Address, City, State, ZipCode, Latitude, Longitude,Neighborhood,CouncilDistrictCode (11)	√	√	√			
SG BP	buildingname，buildingaddress (2)		√	√			
EMA PR	Postcode (1)	√		√			
EMA PB	Postcode (1)	√		√			
BEBR	Building Name, Building Address (2)		√	√			
CBD	Building Hashed Key, Building Full Name (Address), Building Short Name, Building Street Address, Suburb, Building Post Code, State, Geocode, Longitude, Latitude (10)		√	√			
CENED+2	Climate zone, Local zone, Coordinates, Address (6)	√	√	√			
BER MAP	Small Area ID (1)	√	√		√		
CHI Usage	COMMUNITY AREA NAME, CENSUS BLOCK (2)	√	√		√		
DOE SP	DataJamID, recorded ID, zip code (3)	√				√	
BPD	ID, zip code, city, state (4)	√	√			√	
BER TOOL	County Name (1)	√				√	
CBECS	Building identifier (PUBID), Census region, Census division (3)		√				√
RECS	Unique identifier for each respondent, Census Region, Census Division, Housing unit in Census Metropolitan Statistical Area or Micropolitan Statistical Area, Census 2010 Urban Type (5)	√					√
DK HEAT	Application Code, Application Codemean (2)	√	√				√
NEED	Code of region, Code of tax band (2)	√	√				√
DATAPORT	unique identifier, city, state (3)	√					√
GENOME2	Building id, site id, building id kaggle, site_id kaggle (4)	√	√				√

References

[1] F Johari, G Peronato, P Sadeghian, X Zhao, and J Widén, Urban building energy modeling: State of the art and future prospects. Renewable and Sustainable Energy Reviews, 2020. **128**: p. 109902. https://doi.org/10.1016/j.rser.2020.109902.

[2] Y Himeur, A Alsalemi, F Bensaali, and A Amira, Building power consumption datasets: Survey, taxonomy and future directions. Energy and Buildings, 2020. **227**: p. 110404. https://doi.org/10.1016/j.enbuild.2020.110404.

[3] Y Chen, T Hong, X Luo, and B Hooper, Development of city buildings dataset for urban building energy modeling. Energy and Buildings, 2019. **183**: p. 252-265. https://doi.org/10.1016/j.enbuild.2018.11.008.

[4] PA Mathew, LN Dunn, MD Sohn, A Mercado, C Custudio, and T Walter, Big-data for building energy performance: Lessons from assembling a very large national database of building energy use. Applied Energy, 2015. **140**: p. 85-93.

[5] MB Kjærgaard, O Ardakanian, S Carlucci, B Dong, SK Firth, N Gao, GM Huebner, A Mahdavi, MS Rahaman, and FD Salim, Current practices and infrastructure for open data based research on occupant-centric design and operation of buildings. Building and Environment, 2020. **177**: p. 106848.

[6] S Pfenninger, J DeCarolis, L Hirth, S Quoilin, and I Staffell, The importance of open data and software: Is energy research lagging behind? Energy Policy, 2017. **101**: p. 211-215.

[7] N Mims, SR Schiller, E Stuart, L Schwartz, C Kramer, and R Faesy, Evaluation of US building energy benchmarking and transparency programs: Attributes, impacts, and best practices. 2017, Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States).

[8] C FLORES, APEC Workshop on Energy Intensity Reduction in the APEC Regions, in Asia-Pacific Economic Cooperation (APEC) Workshop, Y.P. Kamsing WONG, Editor. 2021, Hong Kong Environment Bureau,Hong Kong Electrical and Mechanical Services Department: Hong Kong

[9] B Najafi, S Moaveninejad, and F Rinaldi, Data analytics for energy disaggregation: methods and applications, in Big data application in power systems. 2018, Elsevier. p. 377-408.

[10] D Ribeiro, Developments in local energy efficiency policy: a review of recent progress and research. Current Sustainable/Renewable Energy Reports, 2018. **5**(1): p. 109-115.

[11] J Lin, Implementations of Energy Benchmarking Disclosures. 2017, Stanford University.

[12] K Palmer and M Walls, Using information to close the energy efficiency gap: a review of benchmarking and disclosure ordinances. Energy Efficiency, 2017. **10**(3): p. 673-691.

[13] MAG Zotano and H Bersini, A data-driven approach to assess the potential of Smart Cities: the case of open data for Brussels Capital Region. Energy Procedia, 2017. **111**: p. 750-758.

[14] Academic Papers Using Pecan Street Data, https://www.pecanstreet.org/dataport/papers/; [Accessed 2022.12.05]

[15] data.gov.uk, https://data.gov.uk/; [Accessed 2022.12.05]

[16] Sustainable Energy Authority of Ireland, https://www.seai.ie/; [Accessed 2022.12.05]

[17] E Commission. Energy performance of buildings directive, https://energy.ec.europa.eu/topics/energy-efficiency/energy-efficient-buildings/energy-performance-buildings-directive_en; [Accessed 2022.12.05]

[18] The Commercial Building Disclosure (CBD) Program - Overview, https://www.cbd.gov.au/program/overview/overview [Accessed 2022.12.05]

[19] BaC Authority. Building Energy Benchmarking, https://www1.bca.gov.sg/buildsg/sustainability/bca-building-energy-benchmarking-and-disclosure [Accessed 2022.12.05]

[20] B Howard, L Parshall, J Thompson, S Hammer, J Dickinson, and V Modi, Spatial distribution of urban building energy consumption by end use. Energy and Buildings, 2012. **45**: p. 141-151.

[21] N Abbasabadi, M Ashayeri, R Azari, B Stephens, and M Heidarinejad, An integrated data-driven framework for urban energy use modeling (UEUM). Applied energy, 2019. **253**: p. 113550.

[22] J Roth, A Martin, C Miller, and RK Jain, SynCity: Using open data to create a synthetic city of hourly building energy estimates by integrating data-driven and physics-based methods. Applied Energy, 2020. **280**: p. 115981.

[23] M Han, Z Wang, and X Zhang, An Approach to Data Acquisition for Urban Building Energy Modeling Using a Gaussian Mixture Model and Expectation-Maximization Algorithm. Buildings 2021. **11**(1): p. 30.

[24] M Lokhandwala and RJSP Nateghi, Leveraging advanced predictive analytics to assess commercial cooling load in the US. Sustainable Production and Consumption, 2018. **14**: p. 66-81.

[25] CE Kontokosta and C Tull, A data-driven predictive model of city-scale energy use in buildings. Applied Energy, 2017. **197**: p. 303-317. https://doi.org/10.1016/j.apenergy.2017.04.005.

[26] C Robinson, B Dilkina, J Hubbs, W Zhang, S Guhathakurta, MA Brown, and RM Pendyala, Machine learning approaches for estimating commercial building energy consumption. Applied Energy, 2017. **208**: p. 889-904. https://doi.org/10.1016/j.apenergy.2017.09.060.

[27] W Zhang, S Guhathakurta, R Pendyala, V Garikapati, and C Ross, A Generalizable Method for Estimating Household Energy by Neighborhoods in US Urban Regions. Energy Procedia, 2017. **143**: p. 859-864. https://doi.org/10.1016/j.egypro.2017.12.774.

[28] W Zhang, C Robinson, S Guhathakurta, VM Garikapati, B Dilkina, MA Brown, and RM Pendyala, Estimating residential energy consumption in metropolitan areas: A microsimulation approach. Energy, 2018. **155**: p. 162-173. https://doi.org/10.1016/j.energy.2018.04.161.

[29] R Boampong, Evaluating the Energy Savings Effect of a Utility Demand-Side Management Program:A Difference-in-Difference Coarsened Exact Matching Approach. The Energy Journal, 2020. **41**(4).

[30] P T Agami Reddy PhD, An actuarial approach to retrofit savings in buildings. ASHRAE Transactions, 2014. **120**: p. 308.

[31] D Dineen, F Rogan, and BÓ Gallachóir, Improved modelling of thermal energy savings potential in the existing residential stock using a newly available data source. Energy 2015. **90**: p. 759-767.

[32] S Lannon, H Iorwerth, M Eames, and M Hunt, Regional modelling of domestic energy consumption using stakeholder generated visions as scenarios, in Urban Energy Simulation. 2018: Glasgow, UK.

[33] T Walter and MD Sohn, A regression-based approach to estimating retrofit savings using the Building Performance Database. Applied Energy, 2016. **179**: p. 996-1005. https://doi.org/10.1016/j.apenergy.2016.07.087.

[34] H Adan and F Fuerst, Do energy efficiency measures really reduce household energy consumption? A difference-in-difference analysis. Energy Efficiency, 2016. **9**(5): p. 1207-1219. 10.1007/s12053-015-9418-3.

[35] I Hamilton, A Summerfield, T Oreszczyn, and P Ruyssevelt, Using epidemiological methods in energy and buildings research to achieve carbon emission targets. Energy and Buildings, 2017. **154**: p. 188-197.

[36] E Green, S Lannon, J Patterson, F Varriale, and H Iorwerth, Decarbonising the Welsh housing stock: from practice to policy. Buildings and Cities, 2020. **1**(1): p. 277-292.

[37] VJ Reina and C Kontokosta, Low hanging fruit? Regulations and energy efficiency in subsidized multifamily housing. Energy Policy, 2017. **106**: p. 505-513. https://doi.org/10.1016/j.enpol.2017.04.002.

[38] JH Scofield and J Doane, Energy performance of LEED-certified buildings from 2015 Chicago benchmarking data. Energy and Buildings, 2018. **174**: p. 402-413. https://doi.org/10.1016/j.enbuild.2018.06.019.

[39] L Shang, HW Lee, S Dermisi, and Y Choe, Impact of energy benchmarking and disclosure policy on office buildings. Journal of Cleaner Production, 2020. **250**: p. 119500. https://doi.org/10.1016/j.jclepro.2019.119500.

[40] N Mostafavi, M Farzinmoghadam, and S Hoque, Urban residential energy consumption modeling in the Integrated Urban Metabolism Analysis Tool (IUMAT). Building and Environment, 2017. **114**: p. 429-444.

[41] S Heiple and DJ Sailor, Using building energy simulation and geospatial modeling techniques to determine high resolution building sector energy consumption profiles. Energy and Buildings, 2008. **40**(8): p. 1426-1436.

[42] Y Ye, K Hinkelman, J Zhang, W Zuo, and G Wang, A methodology to create prototypical building energy models for existing buildings: A case study on US religious worship buildings. Energy and Buildings, 2019. **194**: p. 351-365.

[43] P Shen, Z Wang, and Y Ji, Exploring potential for residential energy saving in New York using developed lightweight prototypical building models based on survey data in the past decades. Sustainable Cities and Society, 2021. **66**: p. 102659. https://doi.org/10.1016/j.scs.2020.102659.

[44] H Deng, D Fannon, and MJ Eckelman, Predictive modeling for US commercial building energy use: A comparison of existing statistical and machine learning algorithms using CBECS microdata. Energy and Buildings, 2018. **163**: p. 34-43. https://doi.org/10.1016/j.enbuild.2017.12.031.

[45] M Heidarinejad, N Mattise, M Dahlhausen, K Sharma, K Benne, D Macumber, L Brackney, and J Srebric, Demonstration of reduced-order urban scale building energy models. Energy and Buildings, 2017. **156**: p. 17-28. https://doi.org/10.1016/j.enbuild.2017.08.086.

[46] P Torres, M Blackhurst, and N Bouhou, Cross comparison of empirical and simulated models for calculating residential electricity consumption. Energy and Buildings, 2015. **102**: p. 163-171. https://doi.org/10.1016/j.enbuild.2015.05.015.

[47] Y Chen and T Hong, Impacts of building geometry modeling methods on the simulation results of urban building energy models. Applied Energy, 2018. **215**: p. 717-735.

[48] DR Carlson, HS Matthews, and M Bergés, One size does not fit all: Averaged data on household electricity is inadequate for residential energy policy and decisions. Energy and Buildings, 2013. **64**: p. 132-144.

[49] BD Leibowicz, CM Lanham, MT Brozynski, JR Vázquez-Canteli, NC Castejón, and Z Nagy, Optimal decarbonization pathways for urban residential building energy services. Applied Energy, 2018. **230**: p. 1311-1325. https://doi.org/10.1016/j.apenergy.2018.09.046.

[50] A Mammoli, M Robinson, V Ayon, M Martínez-Ramón, C-f Chen, and JM Abreu, A behavior-centered framework for real-time control and load-shedding using aggregated residential energy resources in distribution microgrids. Energy and Buildings, 2019. **198**: p. 275-290.

[51] LGR Santos, VK Singh, MO Mughal, E Riegelbauer, JA Fonseca, LK Norford, and I Nevat, Building Anthropogenic heat flux in Singapore. 2020.

[52] LGR Santos, VK Singh, MO Mughal, I Nevat, LK Norford, and JA Fonseca. Estimating building's anthropogenic heat: a joint local climate zone and land use classification method. in eSIM Conference 2021. 2020.

[53] A-D Pham, N-T Ngo, TTH Truong, N-T Huynh, and N-S Truong, Predicting energy consumption in multiple buildings using machine learning for improving energy efficiency and sustainability. Journal of Cleaner Production, 2020. **260**: p. 121082.

[54] K Muralitharan, R Sakthivel, and R Vishnuvarthan, Neural network based optimization approach for energy demand prediction in smart grid. Neurocomputing, 2018. **273**: p. 199-208.

[55] M Voß, C Bender-Saebelkampf, and S Albayrak. Residential short-term load forecasting using convolutional neural networks. in 2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm). 2018. IEEE.

[56] C Nichiforov, G Stamatescu, I Stamatescu, I Făgărăşan, and SS Iliescu. Intelligent load forecasting for building energy management systems. in 2018 IEEE 14th International Conference on Control and Automation (ICCA). 2018. IEEE.

[57] MN Fekri, AM Ghosh, and K Grolinger, Generating energy data for machine learning with recurrent generative adversarial networks. Energies, 2019. **13**(1): p. 130.

[58] A Li, F Xiao, C Fan, and M Hu. Development of an ANN-based building energy model for information-poor buildings using transfer learning. Springer.

[59] KR Mestav, J Luengo-Rozas, and L Tong, Bayesian state estimation for unobservable distribution systems via deep learning. IEEE T Power Syst, 2019. **34**(6): p. 4910-4920.

[60] A Thomas, Y Guo, Y Kim, B Aksanli, A Kumar, and TS Rosing. Hierarchical and distributed machine learning inference beyond the edge. IEEE.

[61] CityBes, https://citybes.lbl.gov/; [Accessed 2023.02.06]

[62] Y Chen, Z Deng, and T Hong, Automatic and rapid calibration of urban building energy models by learning from energy performance database. Applied Energy, 2020. **277**: p. 115584.

[63] N Somu, GR MR, and K Ramamritham, A hybrid model for building energy consumption forecasting using long short term memory networks. Applied Energy, 2020. **261**: p. 114131.

[64] C Miller, More buildings make more generalizable Models—Benchmarking prediction methods on open electrical meter data. Machine Learning and Knowledge Extraction, 2019. **1**(3): p. 974-993.

[65] C Miller, What's in the box?! towards explainable machine learning applied to non-residential building smart meter classification. Energy and Buildings, 2019. **199**: p. 523-536.

[66] MN Fekri, AM Ghosh, and K Grolinger, Generating energy data for machine learning with recurrent generative adversarial networks. Energies, 2020. **13**(1): p. 130.

[67] A Li, F Xiao, C Fan, and M Hu. Development of an ANN-based building energy model for information-poor buildings using transfer learning. in Building Simulation. 2021. Springer.

[68] Y Ahn and D-W Sohn, The effect of neighbourhood-level urban form on residential building energy use: A GIS-based model using building energy benchmarking data in Seattle. Energy and Buildings, 2019. **196**: p. 124-133.

[69] C Li, Y Song, N Kaza, and RJJoPE Burghardt, Explaining Spatial Variations in Residential Energy Usage Intensity in Chicago: The Role of Urban Form and Geomorphometry. Journal of Planning Education and Research, 2019: p. 0739456X19873382.

[70] JA Fonseca, I Nevat, and GW Peters, Quantifying the uncertain effects of climate change on building energy consumption across the United States. Applied Energy, 2020. **277**: p. 115556.

[71] P Shen and B Yang, Projecting Texas energy use for residential sector under future climate and urbanization scenarios: A bottom-up method based on twenty-year regional energy use data. Energy 2020. **193**: p. 116694.

[72] A Summerfield, T Oreszczyn, J Palmer, I Hamilton, F Li, J Crawley, and R Lowe, What do empirical findings reveal about modelled energy demand and energy ratings? Comparisons of gas consumption across the English residential sector. Energy Policy, 2019. **129**: p. 997-1007.

[73] MD Sohn and LN Dunn, Exploratory Analysis of Energy Use Across Building Types and Geographic Regions in the United States. Frontiers in Built Environment, 2019. **5**: p. 105.

[74] K Hellman Miller, F Colantuoni, and C Lasco Crago, An empirical analysis of residential energy efficiency adoption by housing types and occupancy. 2014.

[75] X Gui and Z Gou, Understanding green building energy performance in the context of commercial estates: A multi-year and cross-region analysis using the Australian commercial building disclosure database. Energy 2021. **222**: p. 119988.

[76] K Steemers and GY Yun, Household energy consumption: a study of the role of occupants. Building Research and Information, 2009. **37**(5-6): p. 625-637.

[77] S Karatasou and M Santamouris, Socio-economic status and residential energy consumption: A latent variable approach. Energy and Buildings, 2019. **198**: p. 100-105.

[78] H Estiri and E Zagheni, Age matters: Ageing and household energy demand in the United States. Energy Research and Social Science, 2019. **55**: p. 62-70.

[79] E Hewitt and Y Wang, Understanding the Drivers of National-Level Energy Audit Behavior: Demographics and Socioeconomic Characteristics. Sustainability, 2020. **12**(5): p. 2059.

[80] A Gillich, EM Saber, and E Mohareb, Limits and uncertainty for energy efficiency in the UK housing stock. Energy Policy, 2019. **133**: p. 110889.

[81] L Troup, R Phillips, MJ Eckelman, and D Fannon, Effect of window-to-wall ratio on measured energy consumption in US office buildings. Energy and Buildings, 2019. **203**: p. 109434.

[82] JD Rhodes, B Stephens, and ME Webber, Using energy audits to investigate the impacts of common air-conditioning design and installation issues on peak power demand and energy consumption in Austin, Texas. Energy and Buildings, 2011. **43**(11): p. 3271-3278.

[83] D Choi and C Kim. Diagnosis of building energy consumption in the 2012 CBECS data using heterogeneous effect of energy variables: A recursive partitioning approach. in Building Simulation. 2021. Springer.

[84] M Nazeriye, A Haeri, and F Martinez-Alvarez, Analysis of the impact of residential property and equipment on building energy efficiency and consumption—a data mining approach. Applied Sciences, 2020. **10**(10): p. 3589.

[85] D Hsu, Identifying key variables and interactions in statistical models of building energy consumption using regularization. Energy, 2015. **83**: p. 144-155.

[86] J Ma and JC Cheng, Identifying the influential features on the regional energy use intensity of residential buildings based on Random Forests. Applied Energy, 2016. **183**: p. 193-201.

[87] E Wang, Decomposing core energy factor structure of US commercial buildings through clustering around latent variables with Random Forest on large-scale mixed data. Energy Conversion and Management, 2017. **153**: p. 346-361.

[88] L Adua and B Clark, Even for the environment, context matters! States, households, and residential energy consumption. Environmental Research Letters, 2019. **14**(6): p. 064008.

[89] J Urquizo, C Calderón, and P James. Engineering modelling of building energy consumption in cities: Identifying key variables and their interactions with the built environment. in International Conference on Computational Science and Its Applications. 2019. Springer.

[90] AK Zarabie, S Das, and H Wu, A Data-Driven Machine Learning Approach for Consumer Modeling with Load Disaggregation. arXiv e-prints, 2020: p. arXiv-2011.

[91] JS Vitter and M Webber, A non-intrusive approach for classifying residential water events using coincident electricity data. Environmental modelling & software, 2018. **100**: p. 302-313.

[92] B Aksanli, AS Akyurek, and TS Rosing, User behavior modeling for estimating residential energy consumption, in Smart City 360°. 2016, Springer. p. 348-361.

[93] M Afzalan and F Jazizadeh, Residential loads flexibility potential for demand response using energy consumption patterns and user segments. Applied Energy, 2019. **254**: p. 113693.

[94] NA Funde, MM Dhabu, A Paramasivam, and PS Deshpande, Motif-based association rule mining and clustering technique for determining energy usage patterns for smart meter data. Sustainable Cities and Society, 2019. **46**: p. 101415.

[95] LJER Adua, Reviewing the complexity of energy behavior: Technologies, analytical traditions, and household energy consumption data in the United States. Energy Research and Social Science, 2020. **59**: p. 101289.

[96] H-C Yang, SM Donovan, SJ Young, JB Greenblatt, and L-B Desroches, Assessment of household appliance surveys collected with Amazon Mechanical Turk. Energy Efficiency, 2015. **8**(6): p. 1063-1075.

[97] AK Zarabie and S Das, An l0-norm constrained non-negative matrix factorization algorithm for the simultaneous disaggregation of fixed and shiftable loads. arXiv e-prints, 2019. **1980**: p. 00142.

[98] O Parson, G Fisher, A Hersey, N Batra, J Kelly, A Singh, W Knottenbelt, and A Rogers. Dataport and NILMTK: A building data set designed for non-intrusive load monitoring. in 2015 ieee global conference on signal and information processing (globalsip). 2015. IEEE.

[99] H Yue, K Yan, J Zhao, Y Ren, X Yan, and H Zhao. Estimating Demand Response Flexibility of Smart Home Appliances via NILM Algorithm. in 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). 2020. IEEE.

[100] L Diao, Y Sun, Z Chen, and J Chen, Modeling energy consumption in residential buildings: A bottom-up analysis based on occupant behavior pattern clustering and stochastic simulation. Energy and Buildings, 2017. **147**: p. 47-66.

[101] JY Park, X Yang, C Miller, P Arjunan, and Z Nagy, Apples or oranges? Identification of fundamental load shape profiles for benchmarking buildings using a large and diverse dataset. Applied Energy, 2019. **236**: p. 1280-1295.

[102] S Papadopoulos and CE Kontokosta, Grading buildings on energy performance using city benchmarking data. Applied Energy, 2019. **233-234**: p. 244-253. https://doi.org/10.1016/j.apenergy.2018.10.053.

[103] P Arjunan, K Poolla, and C Miller, EnergyStar++: Towards more accurate and explanatory building energy benchmarking. Applied Energy, 2020. **276**: p. 115413. https://doi.org/10.1016/j.apenergy.2020.115413.

[104] Y Wu, L Blunden, and A Bahaj, City-wide building energy efficiency assessment using EPC data. Future Cities and Environment, 2017. **1**(7): p. 1-7.

[105] Y Pan and L Zhang, Data-driven estimation of building energy consumption with multi-source heterogeneous data. Applied Energy, 2020. **268**: p. 114965.

[106] S Papadopoulos, B Bonczak, and CE Kontokosta, Pattern recognition in building energy performance over time using energy benchmarking data. Applied Energy, 2018. **221**: p. 576-586.

[107] CEJAE Kontokosta, A market-specific methodology for a commercial building energy performance index. The Journal of Real Estate Finance and Economics, 2015. **51**(2): p. 288-316.

[108] PH Jones, NW Taylor, MJ Kipp, and HS Knowles, Quantifying household energy performance using annual community baselines. International Journal of Energy Sector Management, 2010.

[109] J Ma and JC Cheng, Estimation of the building energy use intensity in the urban scale by integrating GIS and big data technology. Applied Energy, 2016. **183**: p. 182-192.

[110] Z Yang, J Roth, and RK Jain, DUE-B: Data-driven urban energy benchmarking of buildings using recursive partitioning and stochastic frontier analysis. Energy and Buildings, 2018. **163**: p. 58-69.

[111] P Arjunan, K Poolla, and C Miller, BEEM: Data-driven building energy benchmarking for Singapore. Energ Buildings, 2022. **260**: p. 111869.

[112] X Jin, F Xiao, C Zhang, and A Li, GEIN: An interpretable benchmarking framework towards all building types based on machine learning. Energ Buildings, 2022. **260**: p. 111909. https://doi.org/10.1016/j.enbuild.2022.111909.

[113] S Agarwal, TF Sing, and Z Yang, Are Green Buildings Really'Greener'? Energy Efficiency of Green Mark Certified Buildings in Singapore. Energy Efficiency of Green Mark Certified Buidlings in Singapore, 2017.

[114] U Ali, MH Shamsi, M Bohacek, K Purcell, C Hoare, E Mangina, and J O'Donnell, A data-driven approach for multi-scale GIS-based building energy modeling for analysis, planning and support decision making. Applied Energy, 2020. **279**: p. 115834. https://doi.org/10.1016/j.apenergy.2020.115834.

[115] P Westermann, C Deb, A Schlueter, and R Evins, Unsupervised learning of energy signatures to identify the heating system and building type using smart meter data. Applied Energy, 2020. **264**: p. 114715.

[116] C Miller and F Meggers, Mining electrical meter data to predict principal building use, performance class, and operations strategy for hundreds of non-residential buildings. Energy and Buildings, 2017. **156**: p. 360-373.

[117] M Quintana, P Arjunan, and C Miller. Islands of misfit buildings: Detecting uncharacteristic electricity use behavior using load shape clustering. in Building Simulation. 2021. Springer.

[118] Z Tian, B Si, X Shi, and Z Fang, An application of Bayesian Network approach for selecting energy efficient HVAC systems. Journal of Building Engineering, 2019. **25**: p. 100796.

[119] Z Tian, S Wei, and X Shi, Developing data-driven models for energy-efficient heating design in office buildings. Journal of Building Engineering, 2020. **32**: p. 101778.

[120] S Cho, S Ray, P Im, H Honari, and J Ahn, Methodology for energy strategy to prescreen the feasibility of Ground Source Heat Pump systems in residential and commercial buildings in the United States. Energy strategy reviews, 2017. **18**: p. 53-62.

[121] S Agarwal, TF Sing, and Z Yang, Are Green Buildings Really'Greener'? Energy Efficiency of Green Mark Certified Buildings in Singapore. Energy Efficiency of Green Mark Certified Buildings in Singapore (October 2017), 2017.

[122] WJ Cole, JD Rhodes, W Gorman, KX Perez, ME Webber, and TF Edgar, Community-scale residential air conditioning control for effective grid management. Applied Energy, 2014. **130**: p. 428-436. https://doi.org/10.1016/j.apenergy.2014.05.067.

[123] J Leitao, P Gil, B Ribeiro, and A Cardoso. Application of Bees Algorithm to Reduce Household's Energy Costs via Load Deferment. in 2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE). 2018. IEEE.

[124] A Abbas and B Chowdhury. A Data-Driven Approach for Providing Frequency Regulation with Aggregated Residential HVAC Units. in 2019 North American Power Symposium (NAPS). 2019.

[125] J Burkhardt, K Gillingham, and PK Kopalle, Experimental evidence on the effect of information and pricing on residential electricity consumption. 2019, National Bureau of Economic Research.

[126] G Fridgen, M Kahlen, W Ketter, A Rieger, and M Thimmel, One rate does not fit all: An empirical analysis of electricity tariffs for residential microgrids. Applied Energy, 2018. **210**: p. 800-814. https://doi.org/10.1016/j.apenergy.2017.08.138.

[127] B Xia, H Ming, K-Y Lee, Y Li, Y Zhou, S Bansal, S Shakkottai, and L Xie, EnergyCoupon: A Case Study on Incentive-based Demand Response in Smart Grid, in Proceedings of the Eighth International Conference on Future Energy Systems. 2017, Association for Computing Machinery: Shatin, Hong Kong. p. 80–90.

[128] KS Cetin, L Manuel, and A Novoselac, Effect of technology-enabled time-of-use energy pricing on thermal comfort and energy use in mechanically-conditioned residential buildings in cooling dominated climates. Building and Environment, 2016. **96**: p. 118-130. https://doi.org/10.1016/j.buildenv.2015.11.012.

[129] MT Devine and P Cuffe, Blockchain Electricity Trading Under Demurrage. IEEE Transactions on Smart Grid, 2019. **10**(2): p. 2323-2325. 10.1109/TSG.2019.2892554.

[130] K Mason, MJ Reno, L Blakely, S Vejdan, and S Grijalva, A deep neural network approach for behind-the-meter residential PV size, tilt and azimuth estimation. Solar Energy, 2020. **196**: p. 260-269. https://doi.org/10.1016/j.solener.2019.11.100.

[131] DL Donaldson and D Jayaweera, Effective solar prosumer identification using net smart meter data. International Journal of Electrical Power & Energy Systems, 2020. **118**: p. 105823. https://doi.org/10.1016/j.ijepes.2020.105823.

[132] S Iyengar, S Lee, D Sheldon, and P Shenoy, SolarClique: Detecting Anomalies in Residential Solar Arrays, in Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies. 2018, Association for Computing Machinery: Menlo Park and San Jose, CA, USA. p. Article 38.

[133] TA Deetjen, JS Vitter, AS Reimers, and ME Webber, Optimal dispatch and equipment sizing of a residential central utility plant for improving rooftop solar integration. Energy, 2018. **147**: p. 1044-1059. https://doi.org/10.1016/j.energy.2018.01.110.

[134] P Odonkor and K Lewis, Automated design of energy efficient control strategies for building clusters using reinforcement learning. Journal of Mechanical Design, 2019. **141**(2).

[135] A Bartolini, F Carducci, CB Muñoz, and G Comodi, Energy storage and multi energy systems in local energy communities with high renewable energy penetration. Renewable Energy, 2020. **159**: p. 595-609. https://doi.org/10.1016/j.renene.2020.05.131.

[136] E Barbour, D Parra, Z Awwad, and MC González, Community energy storage: A smart choice for the smart grid? Applied Energy, 2018. **212**: p. 489-497. https://doi.org/10.1016/j.apenergy.2017.12.056.

[137] DF Quintero Pulido, G Hoogsteen, MV Ten Kortenaar, JL Hurink, RE Hebner, and GJM Smit, Characterization of Storage Sizing for an Off-Grid House in the US and the Netherlands. Energies, 2018. **11**(2): p. 265.

[138] G Henri and N Lu, A Supervised Machine Learning Approach to Control Energy Storage Devices. IEEE Transactions on Smart Grid, 2019. **10**(6): p. 5910-5919. 10.1109/TSG.2019.2892586.

[139] Q Dang. Electric Vehicle (EV) charging management and relieve impacts in grids. in 2018 9th IEEE International Symposium on Power Electronics for Distributed Generation Systems (PEDG). 2018. IEEE.

[140] P Odonkor and K Lewis, Data-Driven Design of Control Strategies for Distributed Energy Systems. Journal of Mechanical Design, 2019. **141**(11).

[141] EC Jones and BDJSC Leibowicz, Co-optimization and community: Maximizing the benefits of distributed electricity and water technologies. Sustainable Cities and Society, 2021. **64**: p. 102515.

[142] TM Mohr, Fuel poverty in the US: evidence using the 2009 Residential Energy Consumption Survey. Energy Economics, 2018. **74**: p. 360-369.

[143] K Mahoney, JP Gouveia, and PJER Palma, (Dis) United Kingdom? Potential for a common approach to energy poverty assessment. Energy Research and Social Science, 2020. **70**: p. 101671.

[144] R Ahmed, KJ Stater, and M Stater, The Effect of Poverty Status and Public Housing Residency on Residential Energy Consumption in the US. Energy Studies Review, 2013. **20**(1).

[145] S Chaitkin and R Van Buskirk. Estimating marginal residential energy prices in the analysis of proposed appliance energy efficiency standards. in Proceedings ACEEE Summer Study on Energy Efficiency in Buildings. 2000.

[146] DJ Bednar, TG Reames, and GA Keoleian, The intersection of energy and justice: Modeling the spatial, racial/ethnic and socioeconomic patterns of urban residential heating consumption and efficiency in Detroit, Michigan. Energy and Buildings, 2017. **143**: p. 25-34.

[147] CE Kontokosta, VJ Reina, and B Bonczak, Energy cost burdens for low-income and minority households: Evidence from energy benchmarking and audit data in five US cities. Journal of the American Planning Association, 2020. **86**(1): p. 89-105.

[148] R Vonnak and Q Zhao, Fuel Poverty and Income Deprivation in Bristol, UK. eprints.gla.ac.uk, 2020.

[149] JA Kelly, JP Clinch, L Kelleher, and S Shahab, Enabling a just transition: A composite indicator for assessing home-heating energy-poverty risk and the impact of environmental policy measures. Energy Policy, 2020. **146**: p. 111791.

[150] R Sejas-Portillo, D Comerford, M Moro, and T Stowasser, Limited attention in the housing market: Threshold effects of energy-performance certificates on property prices and energy-efficiency investments. 2020, CESifo.

[151] M Hyland, RC Lyons, and S Lyons, The value of domestic building energy efficiency—evidence from Ireland. Energy economics, 2013. **40**: p. 943-952.

[152] J Melvin, The split incentives energy efficiency problem: Evidence of underinvestment by landlords. Energy Policy, 2018. **115**: p. 342-352.

[153] L Tagliabue, FR Cecconi, N Moretti, and M Dejaco. The Influence of Energy Performance Certification the Market Value of Residential Buildings. in IOP Conference Series: Earth and Environmental Science. 2019. IOP Publishing.

[154] B Sun, Heterogeneous direct rebound effect: Theory and evidence from the Energy Star program. Energy Economics, 2018. **69**: p. 335-349.

[155] F Varriale, Forecasting future demand for domestic thermal insulation in Wales. Indoor and Built Environment, 2016. **25**(7): p. 1096-1113.

[156] D Hsu, How much information disclosure of building energy performance is necessary? Energy Policy, 2014. **64**: p. 263-272.

[157] OI Asensio and MA Delmas, The effectiveness of US energy efficiency building labels. Nature Energy, 2017. **2**(4): p. 1-9.

[158] Meteonorm data and program, https://www.pvsyst.com/help/meteo_source_meteonorm.htm [Accessed 2022.12.05]

[159] G Ramadevi, S Parvathi, A Kumaresan, and K Vijayakumar, Modern-Era Retrospective analysis for Research and Applications, Version 2. Advances in Natural and Applied Sciences, 2017. **11**(6 SI): p. 109-113.

[160] Climate Change World Weather File Generator, https://energy.soton.ac.uk/ccweathergen-climate-change-weather-file-generator-for-the-uk/; [Accessed 2022.12.05]

[161] Better Buildings Neighborhood Program Single-family home upgrade project dataset. Office of Energy Efficiency & Renewable Energy (EERE).

[162] T Hong, Y Chen, X Luo, N Luo, and SH Lee, Ten questions on urban building energy modeling. Building and Environment, 2020. **168**: p. 106508. https://doi.org/10.1016/j.buildenv.2019.106508.

[163] N Wang, A Vlachokostas, M Borkum, H Bergmann, and S Zaleski, Unique Building Identifier: A natural key for building data matching and its energy applications. Energy and Buildings, 2019. **184**: p. 230-241. https://doi.org/10.1016/j.enbuild.2018.11.052.

[164] MB Kjærgaard, O Ardakanian, S Carlucci, B Dong, SK Firth, N Gao, GM Huebner, A Mahdavi, MS Rahaman, FD Salim, FC Sangogboye, JH Schwee, D Wolosiuk, and Y Zhu, Current practices and infrastructure for open data based research on occupant-centric design and operation of buildings. Building and Environment, 2020. **177**: p. 106848. https://doi.org/10.1016/j.buildenv.2020.106848.

[165] OV Livingston, TC Pulsipher, DM Anderson, A Vlachokostas, and N Wang, An analysis of utility meter data aggregation and tenant privacy to support energy use disclosure in commercial buildings. Energy, 2018. **159**: p. 302-309.