Article

Quality Control Methods for Advanced Metering Infrastructure Data

Eric Garrison [1] and Joshua New [2]

[1] Department of Electrical Engineering and Computer Science, University of Tennessee Knoxville 37996, USA; egarris4@vols.utk.edu
[2] Energy and Transportation Science Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA; newjr@ornl.gov
* Correspondence: egarris4@vols.utk.edu

Version January 23, 2021 submitted to Smart Cities

Abstract: While urban-scale building energy modeling is becoming increasingly common, it currently lacks standards, guidelines, or empirical validation against measured data. Empirical validation necessary to enable best practices is becoming increasingly tractable. The growing prevalence of advanced metering infrastructure has led to significant data regarding the energy consumption within individual buildings, but is something utilities and countries are still struggling to analyze and use wisely. In partnership with the Electric Power Board of Chattanooga, Tennessee, a crude OpenStudio/EnergyPlus model of over 178,000 buildings has been created and used to compare simulated energy against actual, 15-minute, whole-building electrical consumption of each building. In this study, classifying building type is treated as a use case for quantifying performance associated with smart meter data. This article attempts to provide guidance for working with advanced metering infrastructure for buildings related to: quality control, pathological data classifications, statistical metrics on performance, a methodology for classifying building types, and assess accuracy. Advanced metering infrastructure was used to collect whole-building electricity consumption for 178,333 buildings, define equations for common data issues (missing values, zeros, and spiking), propose a new method for assigning building type, and empirically validate gaps between real buildings and existing prototypes using industry-standard accuracy metrics.

Keywords: urban-scale energy modeling; multi-scale building energy modeling; empirical validation; virtual utility; building energy modeling; EnergyPlus; OpenStudio

1. Introduction

In the United States, there are approximately 125 million residential and commercial buildings. Collectively, these buildings consumed approximately 40% of nation's primary energy use, 73% of the electricity, 80% of demand during critical generation hours, and totalled approximately $419 billion in energy bills during 2019. Buildings constitute more than any other energy-consuming sector, and is often referred to as the "built environment." Many locations are attempting to stimulate intelligent and efficient use of energy by assessing smart city developement [1] in service to climate action plans [2]. In order to facilitate private sector application of energy efficiency in these buildings, enhanced decision making tools and financing instruments are becoming available. Tools are becoming available from the subfield of urban-scale building energy modeling [3], where a digital twin of a city-sized area is created and leveraged for many emerging use cases [4]. Such a tool can be used by a city's sustainability officers to evaluate and prioritize attractive energy-saving technologies prior to incentivization or updating building codes. Likewise, a utility could use it to facilitate deployment of energy- and demand-saving technologies to customers through its energy efficiency program. There are many specific instances of urban-scale modeling demonstrations that have been developed recently by universities and U.S. national laboratories.

Massachusetts Institute of Technology (MIT) created a model of 83,541 buildings in Boston, Massachusetts by leveraging publicly available GIS data from tax assessor records [5]. Many urban-scale energy modeling techniques use such locally-rich data sources, but a more general method is needed for assigning building types and properties that doesn't rely on small regions. Stanford University was able to assess 22 modeled buildings in California including comparison to measured data using Mean Absolute Percent Error (MAPE) [6]. This work builds on the same DOE prototypes and comparisons while also leveraging more industry-standard metrics and delving into the challenges of pre-processing energy use data. University of College London in the United Kingdom has set about the ambitious goal of modeling London with 98,000 building energy models built on datasets of building descriptors and energy use not typically found in the United States [7]. Among other notable urban-scale modeling efforts is University of Applied Sciences Stuttgart in Germany that has defined a flexible workflow for ingesting, simulating, and analyzing city-scale data [8]. For the study of London as well as Stuttgart's SimStadt, the authors are not aware of any statistical analysis or summary of advanced metering infrastructure data used or issues encountered.

Lawrence Berkeley National Laboratory's City Building Energy Saver (CityBES) team have created a modern visualization tool for analyzing city-scale building energy models online using building energy models from traditional data sources (e.g. tax assessor data) but also includes thermal and radiant coupling between neighboring buildings [9]. Their analysis has been applied to 940 office and retail buildings in northeast San Francisco with estimates on potential energy savings. Like many such efforts, scalable methods for assigning building type are not needed and details for empirical validation against measured data are lacking. National Renewable Energy Laboratory's URBANopt team have created an open-source repository to facilitate urban-scale energy analysis for buildings [10]. This software repository is flexible and scalable, but relies on the user to provide necessary data, does not provide tools for analyzing energy data, and has not yet been involved in any case studies comparing to measured data. Many of these studies leverage geographically-limited datasets to create building-specific energy models, but some have begun to grapple with the challenges of scalability and empirical validation. At a larger scale, modeling of buildings allows benchmarking of the existing building stock, cost-optimization of energy technologies, and renewables that could offset remaining energy use. This simulation-informed benchmark, reduce, offset approach could help actualize a sustainable built environment.

In a 5-year vision to create a model of every U.S. building, a larger team has set about the task of identify, comparing, and extracting building-specific descriptors from nation-scale data sources. In order to quantify the value of specific data layers or algorithms, the team has partnered with the Electric Power Board of Chattanooga, TN (EPB) which has provided 15-minute electricity use for each building. EPB's service area covers 8 counties and approximately 1400 km^2 in East Tennessee and Georgia. The data sources and algorithms, which we collectively refer to as "Automatic Building Energy Modeling (AutoBEM)," has been used to create 178,368 distinct OpenStudio and EnergyPlus models for every building in EPB's service territory. The models have since quantified energy, demand, emissions, and cost-reductions under nine monetization scenarios for the utility and is being used to inform programmatic rollout of energy efficiency, demand management, product/service lines, and new business models. Previous work has focused on peer review[11], scalable data sources [12], assessment of value propositions [13], virtual utility with buildings as thermal batteries [14], and microclimate interaction [15] detailing the development and application of the building energy models. These research areas are provided for context, but are explicitly outside the scope of the current article.

There exists a software vs. reality technical gap that can lead to distrust in models when applied as digital twins to inform city-scale decisions. Individuals that create software models of real-world objects are often attacked for failing to empirically validate the model with measured data from the real-world. While this is more difficult and costly to maintain, real-world data can expose gaps both in software inputs or underlying algorithms. While there is a tendency for modelers to trust "ground truth" data, those that collect data often prefer to rely on models. This can be due to sensor drift/failure, placement, measurement uncertainty, data acquisition challenges, or formatting/conversion issues.

There is also a research gap for accurately defining the building type of a structure. Traditional urban-scale building energy modeling approaches use tax assessor's data and attempt to map land use or other codes to a canonical set of prototype models. This meta-parameter of a building, combined with the assumption that the building was built to code at the time of construction, is subsequently used to fill out building details (e.g. HVAC type/efficiency, insulation levels, linear feet of refrigeration cases) necessary to perform physics-based energy calculations. This paper discloses results of a methodology based on Energy Use Intensity (EUI) for assigning building type.

This article presents a few simple methods for performing quality control assessment on advanced metering infrastructure data, comparison to prototypical building types, and quantification of error between models and whole-building electricity use. To the authors' knowledge, comparison of building energy models to measured data from over 100,000 buildings has never been published. As such, we hope the crude models, quality control, and industry-standard error metrics will stimulate comparison and improvement of empirical validation techniques for urban-scale modeling. The rest of this article will provide details of the sub-hourly, whole-building electric use information and mathematical methods for comparing this data to building energy models in Materials and Methods. Results follow summarizing statistical analysis of unusual data patterns, methods for correction, industry-standard error metrics for comparison between measured and modeled data, and error rates for building type assignments.

2. Materials and Methods

EPB provided measured data taken from revenue-grade electrical meters for 178, 377 premise IDs. This data was subject to many of the metering issues described above. Technical challenges arise when working with such large data sets, including organization, filtering, and transcription. This paper attempts to address some of these issues and represents an expansion of analysis from our previous paper [16]. An overview of the preprocessing methods is given in this paper, but for more specific details, refer to [16]. Three patterns of outlier data are investigated: missing, 0-vectors, and spiking.

2.1. Data

The first goal of this research was to perform a quality-control analysis on meter data for the nearly 180, 000 customers in EPB's service area. The metered data was collected in calendar year 2015 and initially presented as 50 gigabytes of unsorted tuples in the format <*time*, *premise ID*, *energy use*>. In each tuple, *time* indicates a 15-minute interval during the calendar year, *premise ID* an un-linked property ID, and *energy use* the reported amount of kW hours consumed by the property during the indicated 15-minute interval. This data was sorted by premise ID and chronologically for easier analysis.

An initial look at the data revealed a number of issues:

- Many premise ids have missing data. Almost all premise IDs had at least one 15-minute interval missing from the year, other premise IDs had significantly more data missing.
- Some data is not formatted properly. Date/time formats may have been invalid, or non-numeric values may have been given for energy use. Anything not formatted properly was ignored and treated as missing data.
- There is duplication in the data. Certain premise id and time combinations were entered several times. In these cases, the first properly formatted energy value encountered during sorting was used.
- Some premise IDs may have changed sometime during the year. This is likely due to customers changing rate structures or buildings having new owners. The result is that some premise IDs

had no energy values beyond a certain time of the year or have their first energy values late in the year.

As a consequence of these issues, nearly all premise IDs are missing some data. For the particular year of data, there are exactly $35,040$ 15-minute intervals corresponding to the start of the year (January 1, 00:00-00:15) and every 15 minutes until the end of the year (December 31, 23:45-24:00). Ideally, each premise ID would have the exact number of data points as there are 15-minute intervals in the year. Instead, most premise IDs have at least some missing data, resulting in fewer energy use values. Analysis showed that over 93% of the premise IDs were missing less than 2% of their data (i.e. missing fewer than 701 of 35,040 data points for the year). For the purpose of this research, this was sufficient data to continue the comparison. Any premise IDs missing an excessive amount of data could be individually filtered during later analysis. We refer to *missing vectors* as premise IDs missing in excess of 90% of their data.

In addition to missing data, a quick scan through the premise IDs revealed two unusual trends. The first, which we are calling *0-vectors*, are premise IDs in which all given energy use values for the entire year are zero. The second, which we are calling *spiking vectors*, are similar to the 0-vectors but with one or more 15-minute energy use values exceeding $10,000$ kWh (and in some cases, exceeding $10,000,000$ kWh). Clearly, neither of these patterns represents normal operation of a standard building type. Premise IDs displaying either of these trends could also be filtered during later analysis.

2.2. Comparison

The second goal of this research was to compare crude building simulations to the metered data to determine the value of crude simulations. A total of 97 different prototype building and vintage combinations were simulated using climate zone ASHRAE-169-2006-4A building codes and Actual Meteorological Year (AMY) weather data matching the year of the metered data. The simulations produced the energy use of each building/vintage combination in 15-minute intervals for the same calendar year, resulting in $35,040$ 15-minute intervals for each building/vintage combination. Each premise ID was compared to the 97 prototype vectors to determine a level of similarity. Comparing buildings requires finding the energy use intensity (EUI) of each building, given by the kWh use normalized by area. We were able to obtain square footage of $178,333$ of the initial $178,377$ buildings, allowing us to perform comparisons on nearly all premise IDs in the service area.

For this analysis, Euclidean distance (Eq. 1) was used to determine similarity between each premise ID and the prototype vectors. A smaller Euclidean distance indicates a higher similarity between two values. Every premise ID was individually compared to each of the 97 prototype vectors using Euclidean distance. From there, each premise ID was assigned the building type and vintage corresponding to the prototype vector for which it had the smallest distance:

$$d(p,v) = \sqrt{\sum_{i=1}^{n}(p_i - v_i)^2} \tag{1}$$

where: d is the distance, n is the number of values in a vector, p is the chronological energy use of a premise ID, and v is the chronological energy use of a prototype building. Both p and v have values in 15-minute intervals, given by Eq. (2). The time intervals begin at January 1, 00:00 - 00:15 of the calendar year and continue in 15-minute intervals. Thus, p_1 represents January 1, 00:00 - 00:15, p_2 represents January 1, 00:15-00:30, etc., until $p_{35,040}$ at December 31, 23:45 - 24:00. The value of n in these calculations is $35,040$.

$$p = \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ \dots \\ p_{35,040} \end{bmatrix} \qquad v = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ \dots \\ v_{35,040} \end{bmatrix} \tag{2}$$

For the prototype buildings, each v_i is a positive numeric value. In the case of the premise IDs, missing values at time i are given the value $p_i = NaN$ to differentiate them from actual values of zero and to help with coding. With this representation, the 0-vectors discussed above will have all p_i as either 0 or NaN. As NaN values represent missing data, these values were "skipped" in the distance calculation. For Eq. (1), $p_i = NaN$ "skips" the corresponding v_i value in the prototype vector.

Once each premise ID had been assigned to a building type and vintage, error rates were calculated between the premise ID and prototype. CV(RMSE) (coefficient of variation of the root mean square error) and NMBE (normalized mean bias error) are industry standards for comparing simulated and measured data and were measured based on ASHRAE Guideline 14 [17].

$$CVRMSE = 100 \times \frac{\sqrt{\sum_{i=1}^{n} \frac{(p_i - v_i)^2}{n-1}}}{\bar{p}} \tag{3}$$

$$NMBE = 100 \times \frac{\sum_{i=1}^{n} (p_i - v_i)^2}{n \times \bar{p}} \tag{4}$$

As with previous equations, n is the number of values in a vector, p is the chronological energy use of a premise ID, and v is the chronological energy use of a prototype building. The value $\bar{p}$ is the mean of non-NaN values in p.

3. Results

The results are broken down based on the two different analysis performed. First, several statistical analysis were performed on the metered data to determine the effects of removing the premise IDs matching previously identified patterns. Second, CV(RMSE) and NMBE measurements are given using the same filtering criteria.

3.1. Statistical Analysis

One issue in dealing with a real-world data set is determining how trustworthy the data is. In situations where no ground truth is available, statistical information can be analyzed to determine the consistency of the data set. For this research, several statistics were analyzed with and without filtering. These statistics are RMSE (root-mean square error), RE (relative error), AE (absolute error), average, standard deviation, and the minimum/maximum values. Threshold analysis can be a more accurate way to determine an average of a series of data with missing values than several other methods [18]. In this study, we compute the average, but then use a sliding window of 1.5 hours and standard deviation of 3 ($c = 3$, $n = 6$) to discard any electricity use outside of that range. See [16] for the full implementation of threshold averaging.

$$AE = \sum_{i=1}^{n} |y_i - \bar{y}_i| \tag{5}$$

$$RE = \frac{\sum_{i=1}^{n} \frac{|\bar{y}_i - y_i|}{y_i}}{n} \tag{6}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2}{n}} \tag{7}$$

$$threshold = \mu \pm c\sigma \tag{8}$$

where y_i is the utility data, $\bar{y}_i$ the simulated data, μ the threshold window average, and σ the threshold window standard deviation.

For the statistical analysis, three different filters were applied to the initial data set containing 178, 377 premise IDs. These filters removed premise IDs with the criteria listed below. It is possible that a premise ID could belong to more than on filter. In those cases, filters were applied in the order of *Missing*, *Zeros*, *Spiking*.

1. *Missing*: 90% or more data points of the premise ID were missing (indicating that 90% or more of the data consisted of NaN values).
2. *Zeros*: the maximum value of any 15-minute energy interval did not exceed 0.001 kWh.
3. *Spiking*: contained a maximum value that was over 10, 000 and 50 times larger than the threshold average value. The threshold average value is calculated from Eq. 8.

Effects of individual filters applied to the data are shown in Table 1. For the zeros filter, a value slightly higher than 0 was used to account for conversion issues between data types, such as string to float. Spiking values were selected by the authors based on inspection of this specific dataset and applied equally to all buildings. The values used are with acknowledgment of data-specific behaviors and reported here for completeness, without implying a best practice for flagging buildings with volatile vacillations of energy use.

Table 1. Averages of General Statistical information of metered data based on filters.

Filter	RMSE	RE	AE	Threshold Avg.	Min.	Max.	Raw Avg.	Raw Std. Dev.	Total Values
No Filter	67.55	1.15	105,774.20	1.64	0.08	3,439.39	15.04	147.28	178,333
Remove Missing	67.79	1.16	106,147.48	1.64	0.08	3,450.12	15.08	147.71	177,703
Remove Zeros	69.00	1.18	108,030.75	1.68	0.09	3,512.76	15.36	150.42	174,607
Remove Spikes	0.32	0.92	4,436.47	0.82	0.08	9.54	0.86	0.69	178,195
All Filters	0.33	0.94	4,546.49	0.84	0.08	8.31	0.87	0.61	173,839

The most effective filter by far was removing the spiking data. All metrics, except for the average minimum value, were reduced. This is especially interesting because the spiking filter removed only 138 premise IDs, far fewer than either the missing or zeros filters. Removing spiking premise IDs also reduced the error measurements significantly, while missing and zero premise IDs had almost no effect. It is also worth noting that the threshold average was reduced by roughly $1/2$ while the raw average was reduced by $1/17$, indicating that threshold averaging can be useful for data with unusually high outliers.

3.2. Industry-standard Error Metrics

With the available square footage for premise IDs to perform the Euclidean distance calculations, 178, 333 premise IDs were compared with their matched prototype vectors. The raw data, with no filters applied, is given in Table 2 and shows the average values for each building type for distance, valid data points (the number of non-*NaN* values in the premise ID), CV(RMSE), NMBE, and total number of premise IDs matched to that building type. To clarify, the *Valid Data Points* is the average number of datapoints used to classify each premise; this would be 35,040 if every building had data for every 15-minute period during the year. Also, *Total Matches* refers to the number of buildings assigned to that building type based on Euclidean distance between the building's actual *EUI* compared to the prototype building.

The table reveals several concerning outliers: the IECC and Warehouse building types have error rates exceeding one million percent. A low distance value represents a closer, or better, match between a premise ID and the prototype. The distance value is incredibly high for the QuickServiceRestaurant (QSR), and relatively high for Outpatient and PrimarySchool building types. This indicates that the premise IDs matching to QSR and PrimarySchool do not match *EUI* as well as other premise IDs match their building types.

Table 2. Average Values for Each Building Prototype. No Premise IDs removed.

Building Type	Distance	Valid Data Points	CV(RMSE) (%)	NMBE (%)	Total Matches
FullServiceRestaurant	3.19	33,454.58	780.84	0.77	52
HighriseApartment	0.05	34,313.91	94.16	-8.42	2,068
Hospital	0.21	33,769.13	91.83	6.35	319
IECC	0.02	34,354.81	1,301,192.86	-1,170,353.50	171,821
LargeHotel	0.33	34,160.53	215.37	7.27	408
LargeOffice	0.24	32,162.15	193.09	6.83	41
MediumOffice	6.16	34,422.75	5,678.22	16.61	4
MidriseApartment	0.12	33,928.00	205.77	-21.63	851
Outpatient	18.60	32,643.27	880.82	15.94	59
PrimarySchool	13.17	30,649.00	12,215.67	10.19	2
QuickServiceRestaurant	922.84	33,324.95	1,341.74	56.38	318
RetailStandalone	0.06	23,356.33	68.10	4.77	3
RetailStripmall	2.30	34,962.38	1,579.86	3.92	26
SecondarySchool	0.63	10,318.00	952.27	5.37	2
SmallHotel	0.13	34,380.43	161.19	2.15	1,557
SmallOffice	0.04	12,622.33	508.85	2.03	3
Warehouse	0.06	12,373.44	2,581,773.49	-2,212,138.46	799

Once the initial data was analyzed, filters were applied. After filtering out premise IDs with the *missing*, *zeros*, or *spiking* patterns discussed above, a total of 173, 839 premise IDs remain. The same values are reported in Table 3.

Table 3. Average Values for Each Building Prototype (Missing, Zeros, and Spiking Data removed).

Building Type	Distance	Valid Data Points	CV(RMSE) (%)	NMBE (%)	Total Matches
FullServiceRestaurant	0.22	34,717	78.24	0.00	48
HighriseApartment	0.05	34,406	89.82	-8.42	2,060
Hospital	0.14	33,974	71.85	6.36	316
IECC	0.02	34,454	286,475.32	-257,943.01	167,893
LargeHotel	0.10	34,291	75.82	7.13	400
LargeOffice	0.25	33,692	197.68	7.34	39
MediumOffice	3.59	34,861	4,959.92	1.42	3
MidriseApartment	0.04	34,126	89.16	-21.66	837
Outpatient	0.19	33,992	52.69	10.89	53
QuickServiceRestaurant	2.98	34,902.15	75.72	50.45	256
RetailStandalone	0.08	35,033.50	90.59	9.36	2
RetailStripmall	0.08	34,994.91	91.07	1.98	23
SecondarySchool	1.26	20,618.00	1,880.51	13.71	1
SmallHotel	0.08	34,546.61	95.82	2.15	1,540
SmallOffice	0.07	35,027.00	107.24	8.55	1
Warehouse	0.04	19,997.73	1,758,724.16	-1,499,840.30	367

When the original data is filtered, many of the results are improved. The filtering generally reduces the CV(RMSE) for each building type, while NMBE remains largely unchanged. Comparing the filtered and unfiltered FullServiceRestaurant (FSR), the unfiltered CV(RMSE) of 780.84% is reduced to the filtered value of 78.24%, an improvement by a factor of 10. This is especially notable because the total matches changed from 52 to 48, indicating that the data for the 4 filtered premise IDs was enough to warp the CV(RMSE) of the FSR significantly. Such a decrease is likely the result of removing one or more *Spiking* premise IDs. By definition of the *Spiking* category, at least one value needs to be exceptionally high, which would yield a large error value for that time interval and increase CV(RMSE) compared to other premise IDs with the same building type. When *Spiking* premise IDs are removed, the average CV(RMSE) for that building type will decrease.

Three building types, LargeOffice, RetailStandalone, and SecondarySchool, have their CV(RMSE) and NMBE values increased rather than decreased. When the CV(RMSE) increases, it indicates that the remaining premise IDs have larger outliers than the removed premise IDs, while an NMBE increase indicates a higher average error in the remaining premise IDs. These changes indicate that the removed premise IDs were likely from the *Missing* category. Removing a premise ID in the *Zeros* category would likely reduce the NMBE values, although this might not happen if the building type's *EUI* is very close to 0. This effect is especially noticable on the RetailStandalone and SecondarySchool building types as both had only a single premise ID removed from their *Total Matches*.

Both matches to PrimarySchool in the unfiltered data are removed with filtering, indicating that the unfiltered measurements matched one of the unusual data patterns and were not likely good matches for the building type.

The distance value for the QSR decreased from 922.84 to 2.98 after filtering, indicating that the premise IDs that remain after filtering are significantly more likely to be represented by the QSR building type. This can also be seen in the CV(RMSE), which drops from $1,341.74$ to 75.72 after filtering. This is the result of removing *Spiking* premise IDs, which have an extremely high CV(RMSE) due to the nature of their outliers. The largest outliers within the filtered data are still the IECC and Warehouse building types, whose error measurements exceed $1,000,000$. Although filtering reduced these error measurements significantly, they still far exceed a desirable value and warrant additional investigation in the future.

Generally, the quality control methods appear to be successful in reducing the error rates. The most significant effect comes from removing *Spiking* premise IDs, which lowers CV(RMSE) significantly due to the removal of extreme outliers. The effect on NMBE exists but is less intense. Removing the other two outlier patterns, *Zeros* and *Missing*, have a varied effect on the error measurements that will depend on the building type's energy profile. However, these energy patterns do not accurately represent a building type, and premise IDs matching the patterns should still be filtered prior to analysis.

4. Conclusions

This article attempts to provide guidance for working with advanced metering infrastructure for buildings related to: quality control, pathological data classifications (and their equations), statistical metrics on performance, a methodology for classifying building types, and industry-standard accuracy metrics. Common problems (missing, 0-vectors, and spiking) observed with advanced metering infrastructure data, and the mathematical definitions of these issues, has been shared along with methods for handling these, or similar, data quality problems. Actual 15-minute electricity use from over 178,000 customers has been used to assign building type. The provided statistics can inform the time-of-use energy match between building energy models and real buildings. While advanced metering infrastructure data may become more prevalent, the approaches in this study generally are not feasible since organizations, other than utilities, typically do not have such energy use for buildings at city-scale. CV(RMSE) and NMBE error metrics are used to quantify improvement of the match between modeled and measured building energy use when applying the quality control methods.

Future work will share distributions of error by building type, vintage, and other characteristics to show improvement and remaining challenges in driving down the error in urban-scale energy modeling for both electricity and demand. involves ongoing work to generalize features and Artificial Intelligence-based prediction of building type. Authors

References

1. Patrão, C.; Moura, P.; Almeida, A.T.d. Review of Smart City Assessment Tools. *Smart Cities* **2020**, *3*, 1117–1132. doi:10.3390/smartcities3040055.
2. Bassett, E.; Shandas, V. Innovation and Climate Action Planning. *Journal of the American Planning Association* **2010**, *76*, 435–450. doi:10.1080/01944363.2010.509703.

3. Reinhart, C.F.; Cerezo Davila, C. Urban building energy modeling – A review of a nascent field. *Building and Environment* **2016**, *97*, 196 – 202. doi:https://doi.org/10.1016/j.buildenv.2015.12.001.
4. Ang, Y.Q.; Berzolla, Z.M.; Reinhart, C.F. From concept to application: A review of use cases in urban building energy modeling. *Applied Energy* **2020**, *279*, 115738. doi:https://doi.org/10.1016/j.apenergy.2020.115738.
5. Cerezo Davila, C.; Reinhart, C.F.; Bemis, J.L. Modeling Boston: A workflow for the efficient generation and maintenance of urban building energy models from existing geospatial datasets. *Energy* **2016**, *117*, 237 – 250. doi:https://doi.org/10.1016/j.energy.2016.10.057.
6. Nutkiewicz, A.; Yang, Z.; Jain, R.K. Data-driven Urban Energy Simulation (DUE-S): A framework for integrating engineering simulation and machine learning methods in a multi-scale urban energy modeling workflow. *Applied Energy* **2018**, *225*, 1176 – 1189. doi:https://doi.org/10.1016/j.apenergy.2018.05.023.
7. Steadman, P., E.S.L.R.G.S.D.R.P.; Humphrey, D. Building stock energy modelling in the UK: the 3DStock method and the London Building Stock Model. *Buildings and Cities* **2020**, *1*, 100 – 119. doi:http://doi.org/10.5334/bc.52.
8. Nouvel, R.; Brassel, K.H.; Bruse, M.; Duminil, E.; Coors, V.; Eicker, U.; ROBINSON, D. SimStadt, a new workflow-driven urban energy simulation platform for CityGML city models. Proceedings of International Conference CISBAT 2015 Future Buildings and Districts Sustainability from Nano to Urban Scale. LESO-PB, EPFL, 2015, number CONF, pp. 889–894.
9. Chen, Y.; Hong, T.; Piette, M.A. Automatic generation and simulation of urban building energy models based on city datasets for city-scale building retrofit analysis. *Applied Energy* **2017**, *205*, 323 – 335. doi:https://doi.org/10.1016/j.apenergy.2017.07.128.
10. Polly, B.; El Kontar, R.; Charan, T.; Fleming, K.; Moore, N.; Goldwasser, D.; Long, N. URBANopt: An Open-Source Software Development Kit for Community and Urban District Energy Modeling: Preprint **2020**.
11. Ingraham, J.; New, J. Virtual EPB. Building Technologies Office Peer Review, 2018.
12. J.R. New, M.B. Adams, E.G.B.B.; Guo, T. Scaling Beyond Tax Assessor Data. ASHRAE/IBPSA-USA 2020 Building Performance Analysis Conference & SimBuild (BPACS), 2020.
13. New, J.; Adams, M.; Im, P.; Yang, H.; Hambrick, J.; Copeland, W.; Bruce, L.; Ingraham, J. Automatic Building Energy Model Creation (AutoBEM) for Urban-Scale Energy Modeling and Assessment of Value Propositions for Electric Utilities. Proceedings of the International Conference on Energy Engineering and Smart Grids (ESG), 2018.
14. New, J. Creating a Virtual Utility: Energy and Demand Opportunities via Automatic Building Energy Modeling (AutoBEM). DistribuTech International, 2020.
15. Allen-Dumas, M.; Rose, A.; New, J.; Omitaomu, O.; Yuan, J.; Branstetter, M.; Sylvester, L.; Seals, M.; Carvalhaes, T.; Adams, M.; Bhandari, M.; Shrestha, S.; Sanyal, J.; Berres, A.; Kolosna, C.; Fu, K.; Kahl, A. Impacts of the Morphology of New Neighborhoods on Microclimate and Building Energy Use **2020**.
16. Garrison, E.; New, J.; Adams, M. Accuracy of a Crude Approach to Urban Multi-Scale Building Energy Models Compared to 15-min Electricity Use. *ASHRAE Transactions* **2019**, *125*, 11–19.
17. Guideline, A. Guideline 14-2014. *Measurement of energy, demand, and water savings* **2014**.
18. Castello, C.C.; New, J.R.; Smith, M.K. Autonomous correction of sensor data applied to building technologies using filtering methods. 2013 IEEE Global Conference on Signal and Information Processing, 2013, pp. 121–124.

Author Contributions: Formal analysis, Eric Garrison; Funding acquisition, Joshua New; Methodology, Joshua New; Project administration, Joshua New; Resources, Joshua New; Software, Eric Garrison; Supervision, Joshua New; Validation, Eric Garrison; Writing – original draft, Eric Garrison and Joshua New.

Funding: This work was funded by field work proposal CEBT105 under US Department of Energy Building Technology Office Activity Number BT0201000.

Acknowledgments: The authors would like to thank Amir Roth and Madeline Salzman for their support and review of this project.

Conflicts of Interest: Disclaimer: This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan)