Energy Proceedings
Vol 28, 2022
ISSN 2004-2965

A Computationally Efficient Algorithm to Enable Privacy Preserving Urban Energy Data Sharing Under the “15/15” Rule[#]

Marco Miotti[1], Rishee Jain[1*]

1 Department of Civil and Environmental Engineering, Stanford University, Stanford CA, UA

ABSTRACT

Detailed energy consumption data is key to data-driven urban energy modeling efforts. However, privacy considerations often prevent such data to be shared directly with researchers. Here, we present an approach based on the “15/15 rule,” used in several states in the United States, to enable energy data to be shared while protecting the information of individual customers. We do so based on a case study typical for urban energy data, where public information is combined with privacy-sensitive data. We compare two implementations, showing that our custom algorithm achieves a 1,000 times higher computational speed at only a 10% increase in information loss compared to a previously published clustering method. Our work aims to provide a mechanism to accelerate broader energy data sharing and serve as a baseline for similar efforts in different regulatory contexts, including potential future policy frameworks based on differential privacy.

Keywords: energy data, urban building energy models, privacy, data sharing, clustering, algorithms

NONMENCLATURE

Abbreviations	
SVCE	Silicon Valley Clean Energy
r-km	Restricted k-means clustering
k-unique-nn	k-unique-nearest neighbors clustering (our custom clustering algorithm)
Symbols	
Y	Energy use data for each of N customers (privacy-sensitive)
X	$N \times D$ matrix, indexed $x_{i,j}$, containing building features (public knowledge)
K	Minimum group/cluster size
Ξ	The set of rows still available for clustering
Θ_q	The set of rows belonging to cluster q
σ_i^2	Contribution of row i to variance in data
$\Delta_{i,\xi}^2$	Difference between rows i and ξ

1. INTRODUCTION

The analysis of large datasets can contribute to our understanding of urban energy use patterns and decarbonization pathways. In the context of urban building energy modeling, for example, such datasets usually contain information on energy consumption for a set of customers or buildings along with features for each of those buildings. Relating those features to the energy consumption profiles using a quantitative model can then inform efforts to reduce urban energy use and decarbonize the energy sector through urban planning, building design, and retrofit measures (e.g. [1]–[3]).

However, sharing detailed energy consumption data along with contextual features for each profile is subject to privacy concerns and regulations (e.g. [3]–[7]). This is particularly the case for data containing energy consumption profiles with high temporal resolution and data containing detailed contextual features—the kind of data often most useful to researchers. Recent work has privacy-preserving approaches to sharing energy data (e.g. [5], [8]–[11]), but short-term applications must also fit into the existing policy context and be agreed on by the data sharing partners.

Currently applied in California and certain other states in the United Sates, the “15/15 rule” states that shared energy data must be aggregated such that each data point include a minimum of 15 customers with no one individual customer’s load exceeding 15 percent of the group’s energy consumption [4], [5]. The goal of this rule is to protect the energy consumption profiles of individual customers. An alternative iteration of the rule, applied in some cases to residential customers, imposes a larger minimum group size but with no limitation to individual contributions to that group. Similar protections exist or are being developed in other parts of the world [10]. For example, Europe’s General Data Protection Regulation (GDPR) covers building energy consumption, such as smart meter data [7], [11].

Through privacy-preserving data aggregation, some amount of information in the data is inherently lost [12]. The goal is to aggregate or modify data to comply with privacy standards while minimizing the information loss incurred for subsequent modeling efforts.

Here, we present and compare two approaches to aggregate data into groups and then calculate summary statistics for each group to comply with the 15/15 rule. Our results are based on a real-world data collaboration case with a public-private community choice aggregator utility, Silicon Valley Clean Energy in California, USA. We focus on a dataset that consists of protected information (each customer's energy consumption) as well as public information (properties of the building that each customer is located in).

We describe two approaches to determine the groups, compare their performance in terms of computational time and incurred information loss, and discuss other privacy-preserving measures such as differential privacy. The methodological considerations, findings, and discussions resulting from this case study may also be helpful in different scenarios and contexts.

2. ALGORITHM

2.1 Data structure

The raw dataset contains energy consumption data Y (specifically, 15-minute interval electricity and natural gas consumption for one year) for each of N residential and commercial customers. This data has been supplemented with D features of the buildings and their urban context that each of these customers are located in. We name this set of feature data X. The features include building square footage, floor-to-area ratio, the year of construction, the population density in the building's surroundings, and similar properties. Features with a heavily tailed distribution, such as building square footage, have been log-transformed. Nominal data, such as building type (residential single-unit, residential multi-unit, commercial, etc.) have been converted to dummies. Analysis using this data will aim to infer the relationship between building features X, along with weather data, and energy consumption Y, using regression-type models.

The privacy-sensitive part of the data is the energy consumption vector Y. The building features X are assumed to be public knowledge, as most of them are publicly available through tax assessor and other data. Even if they were not, an adversary possessing auxiliary information on X gathered from third parties would be able to an individual customer in the dataset and obtain detailed information on their energy consumption patterns.

2.2 Goal

The goal is to aggregate the features X and energy consumption profiles Y such that an adversary can no longer infer detailed information about an individual customer's energy consumption even if they obtained the modified data. In our case, the modified data is kept on secure servers and not released publicly; the aggregation therefore acts as an additional barrier to an attack, but not the sole barrier.

In line with the 15/15 rule, the set of features X needs to be aggregated to groups of at least 15 customers each. This is equivalent to modifying the data in X and Y such that for each group, all rows in X are identical, and all energy consumption profiles Y are changed to the average of that group. This measure satisfies the first half (the first "15") of the 15/15 rule, which is the focus of this paper. We discuss how our approach can be extended for compliance with the second half in the discussion section.

2.3 Approach overview

We modify predictors X such that they form groups of 15 identical rows or more. We do so in two steps: (1) cluster the N rows in X into groups of 15 or larger; and (2) homogenize the data within those groups so that each row is identical. The goal is to accomplish the desired outcome while introducing as little noise and therefore losing as little information contained in the original data as possible.

As a result of the fact that X without Y can be assumed to be public knowledge, we can operate on X directly. Once the groups in X have been defined, our data collaborator, SVCE, can then share the privacy-sensitive data Y, averaged for each group in X.

We present two different approaches to accomplish the first step. The first is a constrained k-means algorithm, which is performs well in terms of information loss but is computationally very expensive. The second is a custom algorithm designed for our purpose that is far less expensive at a slight penalty in terms of information loss. We'll call this custom algorithm k-unique-nn (k-unique-nearest neighbors).

2.4 Clustering

Most regular clustering algorithms, such as k-means, are not well equipped for the first step since they can result in clusters of any size. Recent work has proposed a restricted k-means algorithm, which imposes a minimum group size [13]. To minimize the loss of information, we set the number of groups (N_K) to the floor division between the number of rows in the data (N) and the minimum number of rows per group (K): $N_K = \lfloor N/K \rfloor$.

We apply the restricted k-means algorithm to a min-max normalized copy data such that the maximum value of every column is 1, and the minimum is 0. We use a Python/C++ implementation of the restricted k-means method [14].

The restricted k-means algorithm is computationally expensive. This is amplified in our case by the fact that we are looking to cluster data into many groups, with each group containing only a small number of items. This is different from most applications of clustering, where data is sought to be clustered into a smaller number of groups with more data points within each group.

Here, we propose an alternative, substantially faster approach to cluster the data into evenly sized groups. We evaluate the loss of information of this faster approach against the restricted k-means algorithm as a benchmark.

This algorithm, *k-unique*-nn, searches for k nearest neighbors of each data point, but in a way that each point is only assigned as a neighbor once. It is therefore related to, but different from unsupervised k-nearest neighbor clustering. To start, for each row i, we calculate the square difference between the value in each column j and that column's mean, summed across all K columns:

$$\sigma_i^2 = \sum\nolimits_{j\in K}\left(x_{norm,i,j} - \overline{X_{norm,J}}\right)^2$$

We also define Ξ to be the set of rows not yet processed. Initially, $\Xi = 1 \ldots N$ (all rows in X). The algorithm then works as follows:

1. Pick row $i \in \Xi$ whose value σ_i^2 is largest among all rows in Ξ.
2. Calculate the difference between row i and all rows in Ξ, using the min-max normalized data, summed across all columns j:

$$\Delta_{i,\xi}^2 = \sum\nolimits_{j\in K}\left(x_{norm,i,j} - x_{norm,\xi,j}\right)^2 \;\forall\, \xi \in \Xi$$

3. Select the G rows among Ξ with the smallest difference $\Delta_{i,\xi}^2$ (one of them will row i itself, since $\Delta_{i,i}^2 = 0$) and assign them to a cluster. Then remove these rows from Ξ, and start with step 1.
4. Once there are fewer than $2 \times G$ rows left in the dataset, all remaining rows are assigned to the last cluster.

2.5 *Homogenizing Clusters*

Once the clusters have been defined such that each cluster contains at least 15 rows, we iterate over each of those clusters and homogenize each column such that each column within each cluster contains only one unique value. As a result, each row in X belonging to a given cluster will be identical.

To do so, we set each value of each column j in a given cluster q with rows Θ_q (the set of rows that belong to cluster q) to the value that is closest to the average value of that column and cluster:

$$X_{\Theta_q,j} = x_{z,j} \;\; \text{s.t.} \; \min_{z\in\Theta_q}\left(\left\|x_{z,j} - \overline{X_{\Theta_q,J}}\right\|^2\right) \forall\, \mathrm{q}$$

We use the value closest to the average, rather than the average itself, to avoid the introduction of new values not present in the original data where this might not make sense. For example, for integer-like columns such the construction year, the desired outcome is a cluster value of 2015 or 2016, rather than 2015.5, even if the median is 2015.5.

2.6 *Variance-weighted clustering*

The clustering in both algorithms is based on normalized data, meaning that each row has similar variance. However, we might want certain features to be weighted more (i.e., we want certain features to modified less, or less often) than others.

For example, our data contains four binary columns, indicating whether a building is a single-family residential, multi-family residential, commercial, or other property. Clustering buildings of different types together should be avoided. Similarly, features that we anticipate having a stronger effect in the analysis might need to be weighted more, such that the total amount of error introduced in the homogenization step is inversely proportional to the importance of each feature.

To this end, we create a weighted version of the minmax-normalized data, where each column j is multiplied by a weight w_j. The default weight is 1.0. For the binary columns indicating building types, the weight is set to an arbitrarily large number (e.g., 100) to prevent different building types to be clustered together completely (in the restricted k-means approach) or in all but the final cluster (in our custom algorithm).

2.7 *Measuring Performance*

The goal is to keep the difference between the original set of predictors, X, and the modified data (where each cluster has been homogenized) to a minimum. To evaluate this difference, we calculate the average difference across all $N \times K$ values in the dataset. The change for each individual value, $x_{i,j}$, is measured as the change in that value relative to the scale of the corresponding columns (the difference between the minimum and the maximum value of that column in the original data):

$$\delta\,[\%] \; = \frac{100}{J}\sum_{j}\frac{\sum_i\left(x_{i,j,\text{before}} - x_{i,j,\text{after}}\right)^2}{\sum_i\left(x_{i,j,\text{before}} - \overline{X_{J,\text{before}}}\right)^2}$$

A value of $\delta = 5.0\%$ therefore means that the variance of the difference between the values before and after the modifications is 5/100th of variance in the original data. A value of $\delta = 100\%$ indicates complete loss of information (replacing the entire column with its mean value). A value of $\delta = 0\%$ indicates no change (all values remained identical).

3. RESULTS

3.1 Performance comparison of the two approaches

We find that our custom algorithm is about 1,000 times faster than the restricted k-means clustering approach (Table 1). This makes it feasible to be used on the full dataset with 170,592 rows.

Table 1: Comparison of computational efficiency and information loss between the restricted k-means clustering method (r-km) and our custom algorithm (both for group size G=15). Performance is indicated for the full dataset (N=170,592) for the custom algorithm only, and for two random subsets (all algorithms). Computing time is based on Python/numpy implementations running on an Intel Core i5-8279U. Weighted data is used.

		N		
Metric	**Algorithm**	**5,000**	**20,000**	**170,592**
Computing time	r-km	185s	1.2h	~200h
	k-unique-nn	0.3s	4.8s	376s
Information loss (δ)	r-km	14.0%	10.9%	?
	k-unique-nn	15.2%	11.7%	6.0%

When applied to the full dataset, the custom algorithm causes an information loss of 6%, as defined in the previous section, to anonymize the data such that there are always at least 15 rows with matching properties (Table 1). This loss increases with decreasing size of the data. For those smaller sizes, for which we can run the restricted k-means clustering approach as a benchmark, we observe about a 10% increase in the loss of information incurred by our faster custom algorithm than the benchmark approach (11.7% vs 10.9%). As a result, our k-unique-km algorithm can feasibly be used in place of the more precise r-km algorithm with a modest penalty in information loss, but a substantial increase in computational efficiency.

3.2 Weighting the data

Using weighted data leads to an increase in total information loss, but a decrease the information loss of the columns that are weighted more highly (Figure 2). In our example, the information loss in building square footage is decreased from 10.4% to 1.37% because of weighting it more, while the data in other columns that remained at the default weight, such as year built, are subject to an increase in the loss of information.

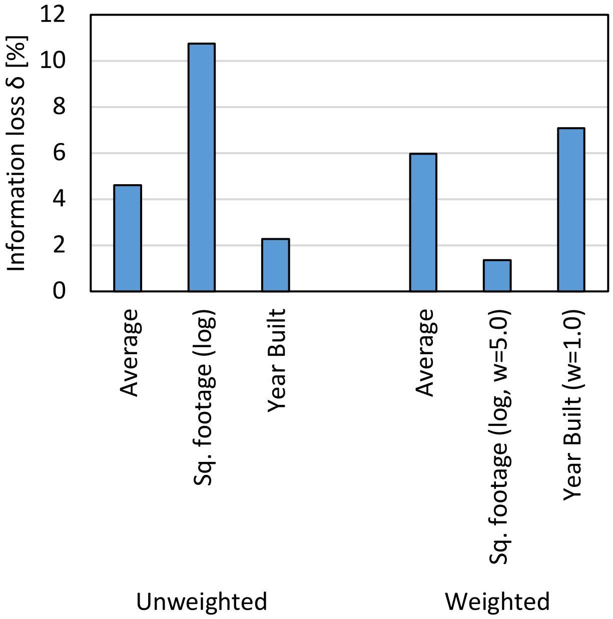

Figure 1: Comparison of information loss between unweighted and weighted data for all columns ("average") as well as two individual columns, one of which is weighted higher and one of which is weighted lower.

Introducing weights is not a zero-sum game: it increases the total amount of information loss across all columns (from 4.6% to 6.0% in this case). Yet, this may still lead to an increase in accuracy of the final model applied to the data if the weights are chosen properly (for example, proportionally to each feature's effect size in a normalized linear regression model). To do so, prior knowledge about the importance of each column is required. This knowledge can be obtained from previous analyses, or by performing the data aggregation twice: a first time without weights for a preliminary analysis, and a second time with informed weights.

3.3 Modifying the group cluster size

The loss of information increases super-linearly with the log of increasing cluster size (Figure 2). As a result, increasing the cluster size from the default of 15 to 50, for example, would increase the loss of information from 6.0% to 10.3% (using the column weights). Therefore, a minimum cluster size higher than about 25 would not be recommended in this case. The log of the computational effort decreases linearly with the log of increasing cluster size, indicating a monomial relationship.

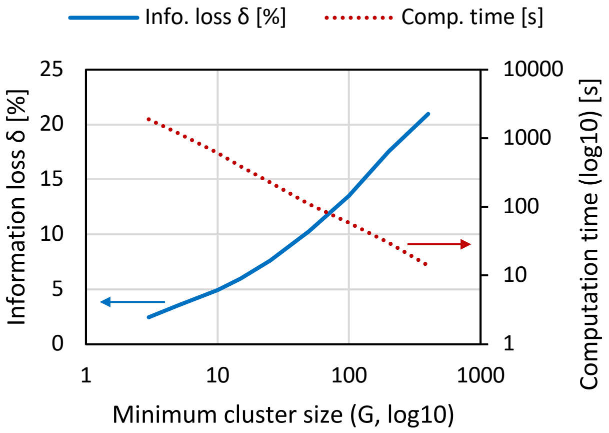

Figure 2: Relationship between information loss (δ) and computation time and minimum cluster size (group size) G. The default is $G = 15$.

3.4 Addressing the maximum contribution clause

As previously noted, this work focuses on the clustering aspect of our policy context, stating that data must be aggregated to groups of a specific minimum size. In our case, compliance with the second aspect—stating that no one individual customer's load can exceed 15 percent of the group's energy consumption—was not required by our collaborator due to additional data security measures that were put in place.

If necessary, however, a simple extension could be made to accommodate this second part: first, we would raise the minimum cluster size (e.g. from 15 to 25), incurring an increase in loss of information from 6.0% to 7.6% (see Figure 2). This makes it unlikely that an individual customer of any cluster contributes more than 15% to the total energy consumption of that cluster (which has an average contribution of 4.0%), since the clustering is based on building features that jointly predict a relatively large share of the variance in annual energy consumption. Then, the few individual customers whose energy use does contribute more than 15% are dropped. Those customers can be considered outliers for energy modeling purposes (since their energy consumption is far higher than what we'd expect for the given cluster), reducing the drawback of having to drop some information from the data for many applications. If fewer than 15 customers remain after iteratively dropping customers whose contribution to the updated total is over 15%, the cluster exhibits an unusually high amount of variance in terms of energy consumption and is dropped from the data entirely.

For more accurate compliance with the second part of the 15/15 rule, the algorithm could be modified to consider that part while defining the clusters. The main disadvantage of this approach, however, is that the algorithm, which now requires information on each customer's energy consumption (Y) at the time of execution, will need to be executed by the data owners, rather than the research collaborators.

DISCUSION

In this work, we introduce an approach to modify data such that individual rows are not uniquely identifiable based on their features by clustering data and then homogenizing each cluster. This approach is tailored to compliance with the "15/15 rule," but may apply to other situations where data needs to be anonymized to preserve privacy as well.

We present two different approaches to produce the clustered data, one based the computationally expensive restricted k-means clustering method, and k-unique-nn, an algorithm developed for this work. The latter is several orders of magnitude faster while causing a 10% (relative) increase in information loss. The k-unique-nn algorithm scales O(n2) with increasing sample size n, which is inferior scaling behavior than the restricted clustering method. However, at dataset sizes so large that the restricted k-means clustering algorithm would catch up with the k-unique-nn algorithm in terms of speed, it should usually be feasible to split the dataset into chunks and process each chunk separately with a minimal decrease in information loss.

In our case, energy consumption data is aggregated across each cluster by the data owner once the clusters have been defined. Therefore, individual energy consumption profiles are going to be smoothened, and demand spikes caused by individual customers will be less salient in the final data. This is an inherent problem to aggregating energy data by averaging or summing it over a given group or cluster. The extent to which this is an issue will depend on the intended use of the aggregated data.

In the context of data privacy, differential privacy has been receiving an increasing amount of attention [5], [8]. It allows to publicly share information about a dataset by describing characteristics of groups within that dataset while withholding information about individuals. As a result, an adversary with auxiliary information is unable to infer whether one particular person or entity is part of the corresponding group.

This goal is somewhat different to our case. First, in our case, the protected information is the energy consumption; other building properties (such as square footage, age, type, etc.), and their combination, are assumed to be public knowledge. Second, our goal was not to prevent anyone from ascertaining that a particular customer is part of a given group, but rather to protect that customer's energy consumption profile.

Nonetheless, there are differential privacy approaches to one-time publishing of non-overlapping counts of data, such as sharing energy consumption profiles along with contextual data [9], [15]. The general idea is similar to our approach: first, the data is grouped; second, aggregate information is released for each of those groups. The difference is that both steps would be carried out in line with differential privacy principles, rather than the 15/15 rule.

An implementation based on differential privacy could ascertain a quantifiable amount of privacy protection of individual customers in a way that the 15/15 rule cannot. It could also alleviate some of the drawbacks of the 15/15 rule, especially if geographical information (such as the zipcode location) is to be preserved [4]. The results presented here can serve as a reference, both in terms of loss of information as well as in terms of computational effort required, for possible future implementations based on differential privacy.

Even once policies based on differential privacy are in place, however, a potential challenge is the burden of implementation: Our collaboration with a relatively small entity was made possible by the fact that most of the effort to design the anonymization strategy (defining the clusters) could be carried out by us, using data that can be assumed to be public information. A potentially costly involvement of a third-party curator was not necessary.

In conclusion, our work can help accelerate energy data sharing to enable urban energy modeling and inform decarbonization pathways. It is particularly suitable where policies correspond to the 15/15 rule or similar principles, providing a straight-forward and computationally inexpensive approach. It can also serve as a baseline for different specific implementations, including those relying on differential privacy.

ACKNOWLEDGEMENTS

We acknowledge and thank the Stanford TomKat Center for Sustainable Energy, the Stanford Center for Integrated Facility Engineering (CIFE), the Swiss National Science Foundation and the U.S. National Science Foundation Award # 1941695 for their funding of this work. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Swiss National Science Foundation and/or the U.S. National Science Foundation. We also thank Thomas Dougherty for the discussion regarding differential privacy.

REFERENCES

[1] W. Li *et al.*, "Modeling urban building energy use: A review of modeling approaches and procedures," *Energy*, vol. 141, pp. 2445–2457, 2017.

[2] A. Nutkiewicz, Z. Yang, and R. K. Jain, "Data-driven Urban Energy Simulation (DUE-S): A framework for integrating engineering simulation and machine learning methods in a multi-scale urban energy modeling work fl ow," *Appl. Energy*, vol. 225, no. June, pp. 1176–1189, 2018.

[3] C. F. Reinhart and C. Cerezo Davila, "Urban building energy modeling - A review of a nascent field," *Build. Environ.*, vol. 97, pp. 196–202, 2016.

[4] K. Crandall, "Rethinking Energy Data Access," 2019.

[5] T. Lamm and E. N. Elkind, "Data Access for a Decarbonized Grid: Policy Solutions to Improve Energy Data Access and Drive the Clean and Resilient Grid of the Future," Berkeley, CA, USA, 2021.

[6] C. McChalicher, "An Avalanche of Energy Data: The Rise of Energy Data Sharing," *Northeast Energy Efficiency Partnerships*, 19-May-2019.

[7] C. Véliz and P. Grunewald, "Protecting data privacy is key to a smart energy future," *Nat. Energy*, vol. 3, no. 9, pp. 702–704, 2018.

[8] M. Young, M.-A. Paré, and H. Bergmann, "Differential Privacy for Expanding Access to Building Energy Data," *ACEEE Summer Study Energy Effic. Build. Proc.*, 2020.

[9] G. Kellaris and S. Papadopoulos, "Practical differential privacy via grouping and smoothing," *Proc. VLDB Endow.*, vol. 6, no. 5, pp. 301–312, 2013.

[10] D. Lee and D. J. Hess, "Data privacy and residential smart meters: Comparative analysis and harmonization potential," *Util. Policy*, vol. 70, p. 101188, 2021.

[11] Ministry of Housing Communites and Local Government, "Making better use of energy performance of buildings data: Data Privacy Impact Assessment (DPIA)," London, 2019.

[12] V. S. Iyengar, "Transforming data to satisfy privacy constraints," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 279–288, 2002.

[13] P. Bradley, K. Bennett, and A. Demiriz, "Constrained k-means clustering," *Tech. Rep.*, p. 9, 2000.

[14] J. Levy-Kramer and M. Klaber, "k-means-constrained," *GitHub Repository*, 2022. [Online]. Available: https://github.com/joshlk/k-means-constrained. [Accessed: 06-May-2022].

[15] T. Zhu, G. Li, W. Zhou, and P. S. Yu, "Differentially Private Data Publishing and Analysis: A Survey," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 8, pp. 1619–1638, 2017.