PNNL-23786

Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

Commercial Building Tenant Energy Usage Data Aggregation and Privacy

October 2014

OV Livingston
TC Pulsipher
DM Anderson
N Wang

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from the
Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062;
ph: (865) 576-8401
fax: (865) 576-5728
email: reports@adonis.osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
email: orders@ntis.gov <http://www.ntis.gov/about/form.aspx>
Online ordering: http://www.ntis.gov


This document was printed on recycled paper.
(8/2010)

PNNL-23786

Commercial Building Tenant Energy Usage Data Aggregation and Privacy

OV Livingston
TC Pulsipher
DM Anderson
N Wang

October 2014

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99352

Summary

A growing number of building owners are benchmarking their building energy use. This requires the building owner to acquire monthly whole-building energy usage information, which can be challenging for buildings in which individual tenants have their own utility meters and accounts with the utility. While utilities can supply anonymized consumption data, this alone does not afford sufficient privacy protection for individual tenants.

Some utilities and utility regulators have turned to aggregation of customer energy use data (CEUD) as a way to give building owners whole-building energy usage data while protecting customer privacy. Meter profile aggregation adds a layer of protection that decreases the risk of revealing CEUD as the number of meters aggregated increases. The report statistically characterizes the similarity between individual energy usage patterns and whole-building totals at various levels of meter aggregation.

Meter data aggregation poses a tradeoff between the risk of tenant identification and the number of buildings which would be covered under different aggregation thresholds to streamline data access. As more and more meters are aggregated, fewer and fewer buildings are eligible for CEUD reporting – lessening the value of the data. The fewer the meters that are aggregated, the greater the number buildings that become eligible for reporting. The fewer the meters that are aggregated, the easier the matching of individual CEUD from the reported aggregated total.

With these concepts in mind, the goal of this study is to establish a quantitative approach for providing practitioners, such as utilities, public utility commissions, and other policy-makers with a defensible aggregation threshold selection method, which will protect tenant privacy and ensure data on the greatest number of buildings can be reported. Our research did not identify existing studies reporting empirical evidence for determining appropriate multi-meter building energy data aggregation thresholds. This study performed applied statistical analysis of actual CEUD supplied by utilities specifically for this study under nondisclosure agreements. Using actual utility data helped inform the relationship between the meter aggregation levels and proportion of individual CEUD that could be estimated from the aggregated total building profile. The report offers an applied statistical analysis of meter aggregation thresholds that can be easily understood by utilities, and explains a methodology for aggregation threshold selection.

An important focus of this study is protecting tenant privacy while utilizing the building-level energy consumption data available to utilities. Therefore, the study estimated the proportion of individual CEUD found to be similar to the aggregated building consumption profile at various levels of meter aggregation. This analysis permits utilities to find the minimum meter count in multi-meter buildings for reporting aggregated monthly energy data at the building level (e.g., monthly energy consumption data for buildings with 2, 3, 4, 5, or more tenants aggregated to the building total) without compromising tenant privacy or requiring disclosure agreements from individual tenants.

To inform the method for selecting an aggregation threshold, we reviewed current literature on anonymization (i.e., encrypting or removing personally identifiable information from data), differential privacy (i.e., measure of risk to privacy from database participation), and other privacy-protecting techniques, and decided to focus this analysis on understanding the variability in building energy consumption profiles, variability in meter profiles, and the relationship between the degree of variability

and the proportion of individual CEUD that potentially could be estimated from building totals. Basic statistical diagnostic techniques, including the use of k-means cluster analysis and descriptive statistics, were used to establish a baseline approach for threshold selection that would be both easy to implement for practitioners working for utilities and statistically robust, without requiring advanced procedures.

CEUD referred to throughout the report are simply the pattern of monthly electricity or natural gas consumption reported from anonymized meter billing data obtained through the voluntary participation, under strict nondisclosure agreements, of six utilities from geographically and climatically diverse areas of the country. Together the six utility billing data sets include the monthly consumption profile of nearly 715,000 anonymized non-residential meters, representing about 129,000 individual buildings (see Table S.1). The average building meter profile (ABMP) is simply the summation of the consumption across all meters associated with that building, divided by the number of meters. The similarity of individual CEUD to ABMP becomes a key metric from which the analysis of aggregation thresholds is derived.

Table S.1. Anonymized utility meter datasets provided for analysis

Provider	Meters	Buildings
A	34,208	11,597
B	52,893	16,066
C	63,091	13,352
D	106,791	23,469
E	400,382	47,011
F	57,242	17,318
Totals	714,607	128,813

PNNL compared the percentage of meters found to be similar to their average building profile at each building meter count with the reporting eligibility for buildings across the six utility service areas. Figure S.1 illustrates the proportion of meters similar to their average building profile for the six participating utilities. In other words, this reveals the percentage of cases in which the building total profile and knowledge of the number of tenants or meters can assist in estimating the energy use of a specific tenant, thereby compromising that tenant's privacy. It should be noted that the meter profile matching is one step removed from tenant reidentification. Associating a particular curve with an individual tenant can only proceed after an energy consumption profile is successfully matched, based on the building total energy consumption profile. The report only addresses CEUD matching or the estimation of the probability of consumption profile similarity, and explicitly does not address extending CEUD matching to actual tenant identification. In addition, this analysis does not apply to single-entity buildings, where aggregation will lead to inadvertent revealing the entity's CEUD, irrespective of the aggregation threshold and number of meters in the building.

In Figure S.1, the horizontal axis depicts the meters per building and the vertical axis indicates the proportion of meter profiles that are statistically similar to ABMP. The figure indicates that the proportion of individual meters resembling ABMP follows a similar trend across all 6 utilities for both the 3-meter and 4-meter buildings, ranging between 25-33 percent for 3-meter buildings and 20-25 percent for 4-meter buildings. However, if a building has three tenants and one tenant moves out, individual CEUD matching would be more likely for the remaining two tenants, rising to that of 2-meter buildings.

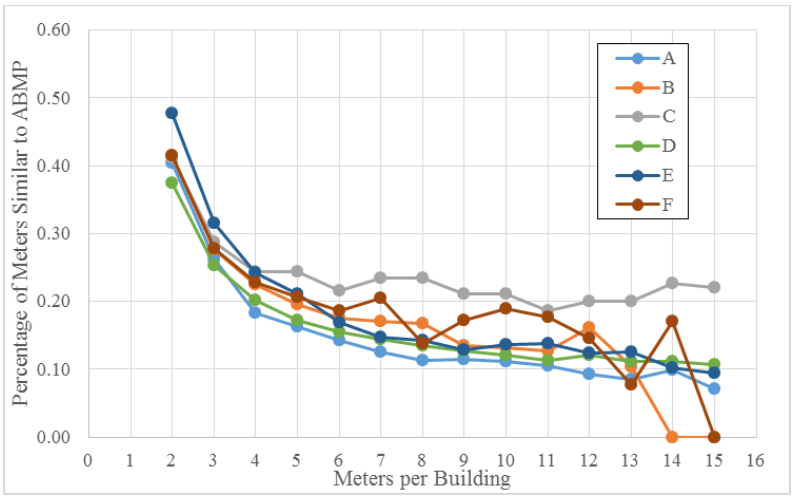

Figure S.1. Percentage of meter profiles similar to their building average by count category.

Moving up in the number of meters per building (out along the x-axis), the spread between utilities increases and profile similarity becomes less likely at a decreasing rate. Across the six utilities, at 4 meters per building, individual consumption profile similarity is likely for 20-25 percent of the meters in multi-meter buildings.

Next, we compare the percentage of meters that are similar to the ABMP with the number of buildings eligible for reporting. The number of buildings that would be eligible for reporting under each aggregation level is an important consideration. As the aggregation threshold increases, the proportion of individual CEUD resembling ABMP declines and the number of eligible buildings also decreases. Figure S.2 illustrates the percentage of buildings eligible for reporting under each aggregation threshold.

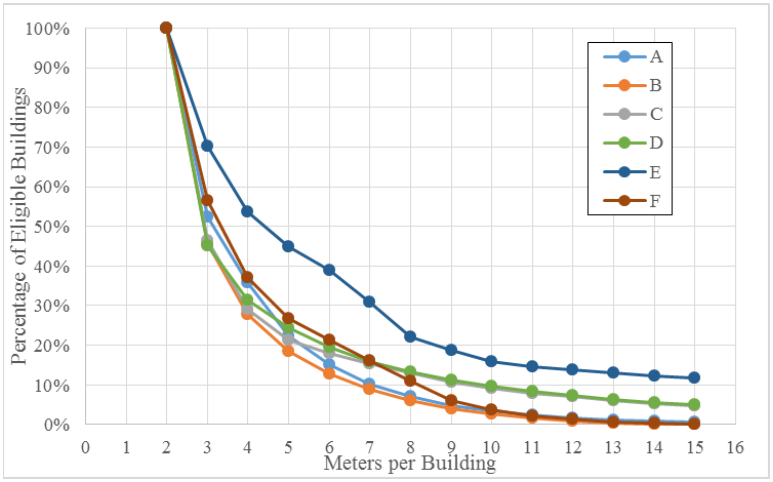

Figure S.2. Building reporting eligibility by building meter count category.

In Figure S.2, the x-axis depicts the number of meters per building (e.g., buildings with 2 meters, buildings with 3 meters, buildings with 4 meters, etc.), and the y-axis indicates the percentage of muli-meter buildings eligible for reporting. The figure, together with the data in Table S.2, illustrates that eligibility is decling faster than the percentage of meters resembling their ABMP (Figure S.1).

Table S.2. Tradeoff between aggregaton threshold and reporting eligibility.

Threshold (# of meters)	Percentage of Meter Profiles Similar to Their Building Profile (%)					
	A	B	C	D	E	F
2	40	42	41	37	48	42
3	26	28	29	25	31	28
4	18	22	24	20	24	23
5	16	20	24	17	21	21
6	14	17	22	16	17	19
7	13	17	23	14	15	20
8	11	17	23	13	14	14
9	12	14	21	13	13	17
10	11	13	21	12	14	19
11	10	13	19	11	14	18
12	9	16	20	12	12	15
13	9	10	20	11	13	8
14	10	-	23	11	10	17
15	7	-	22	11	9	-

Threshold (# of meters)	Percentage of Multi-Meter Buildings Coverage (%)					
	A	B	C	D	E	F
2	100	100	100	100	100	100
3	52	46	47	45	70	56
4	36	28	29	32	54	37
5	22	18	21	25	45	27
6	15	13	18	19	39	21
7	10	9	15	16	31	16
8	7	6	13	13	22	11
9	5	4	11	11	19	6
10	3	3	9	10	16	4
11	2	2	8	8	15	2
12	2	1	7	7	14	1.3
13	1.2	0.3	6	6	13	0.7
14	0.8	-	5	6	12	0.3
15	0.6	-	5	5	12	-

For example, utility F shows an approximately 2 percentage point decline in probability of CEUD matching by moving from 4-meter buildings to 5-meter buildings (in Figure S.1), but that shift reduces the number of eligible buildings by over 10 percentage points – going from 37 percent of multi-meter buildings being eligible, down to 27 percent (in Figure S.2). This simply reflects the difference in the number of 4-meter buildings compared to the number of 5-meter buildings in that utility's data set. We would expect fewer and fewer buildings at increasing numbers of meters per building.

Table S.2 shows the proportion of meters that are similar to their building profile and the percentage of buildings eligible for reporting under each aggregation threshold across 6 analyzed utilities.

For example, if for utility B the threshold is set at 4, the proportion of meters resembling ABMP is about 22%, meaning that roughly one out of five meter profiles could be estimated from the building total profile by dividing it by the number of meters. 28% of buildings have at least 4 meters, meaning that if an aggregation threshold of 4 is applied, 28% of multi-meter buildings would be covered by an aggregation threshold set at 4 meters.

If the threshold were set at 5, the proportion of meters that are similar to their average building profile would drop by 2 percentage points to 20%. However, the eligibility rate for this aggregation level drops by 10 percentage points to 18%. Increasing the aggregation threshold from 4 to 5 drops 2 percentage points in proportion of meters that are similar to their average building profile, but loses 10 percentage points in coverage.

As the aggregation threshold is increased from 5 to 6 for utility B, the percentage of meters that are similar to their average building profile drops by about 3 percentage points (from 20% to 17% as observed in the upper half of Table S.2), while the loss of eligibility is 5 percentage points (from 18% to 13% as observed in the bottom half of Table S.2). If a decrease in percentage of meters that resemble their average building profile can be interpreted as a degeree of protection provided by each aggregation threshold, then the gain in protection by increasing the threshold from 5 to 6 is exceeded by the loss in buildings elegible for reporting.

Similarly, by increasing the aggregation threshold for utility B even further, from 6 to 7, the percentage of meters that are similar to the building average drops by less than one percentage point. This less than one percentage point gain in protection comes at the expense of losing another 6 percentage points in the number of buildings elegible for reporting.

Policy-makers selecting an aggregation threshold may wish to consider not only the percentage of meters resembling their building average at each threshold, but also the tradeoff between the incremental gain in protection compared to the loss in building coverage resulting from increasing the threshold any further. Acknowledging this tradeoff, we have provided the data here to assist in selecting an appropriate threshold.

Acronyms and Abbreviations

ABMP	average building meter profile
CBECS	Commercial Buildings Energy Consumption Survey
CPUC	California Public Utilities Commission
DOE	U.S. Department of Energy
EDA	exploratory data analysis
EFF	Electronic Frontier Foundation
ESPM	ENERGY STAR Portfolio Manager
IQR	inner quartile range
NG	natural gas
PII	personally identifiable information
RECS	Residential Energy Consumption Survey

Contents

Figures

Tables

1.0 Introduction

A growing number of building owners are adopting the practice of regularly benchmarking their building energy use to assist them in energy management. This practice allows tracking a building's energy performance over time and allows a building owner to identify buildings that may have unusually high energy use and be good candidates for saving money through improved energy efficiency. This requires the building owner to acquire monthly whole-building energy usage information, which can be challenging for buildings in which individual tenants have their own utility meters and accounts with the utility. Some utilities and utility regulators have turned to aggregation of customer data as a way to give building owners the whole-building energy usage data while protecting customer privacy.

In multi-meter buildings, two important concepts inform the provision of customer energy-use data (CEUD)[1]. First, tenant profile aggregation refers to the summation of monthly energy meter profiles to form the monthly building total profile across the meters associated with each building. Monthly individual meter profiles are used as a proxy for tenant monthly energy consumption profiles in this analysis. Use of a proxy is necessary because utilities track energy usage at the meter level, and not at the tenant level. The data on the tenant level are generally not available. The report refers to the grouping of multi-meter buildings according to the number of meters present (e.g., 4-meter buildings, 5-meter buildings, 6-meter buildings, etc.). The report also refers to meter aggregation thresholds. The meter aggregation threshold is the minimum number of meters in a building above which the building monthly energy profile can be considered for aggregation.

Second is the concept of aggregation coverage. Depending on the aggregation threshold, only a subset of a utility's population of meters will be covered by the threshold. The relationship between aggregation level and reporting eligibility is that the more meters an individual building has, the fewer of similar such buildings there are – thus fewer to report in the data. The population of 4-meter buildings will typically be larger than the population of 5-meter buildings, and so on. Therefore, if we raise the aggregation threshold, say from 4 to 5 meters, we lose a nontrivial number of buildings because the buildings having 4 meters would drop out of the selection for reporting.

Figure 1 illustrates these concepts. The study examines the trade-offs in reporting eligibility at each meter count category (i.e., buildings with 2, 3, 4, 5, or more meters). For example, a building with 8 meters would not be classified as 2 groupings of 4 meters – it could be classified only as one 8-meter building. To enable meter profile aggregation, the CEUD from the utilities link each meter to its building, but no other building information is provided. Thus, there is no means to determine key characteristics such as floor space, building type, operating schedule, etc., from the utility-supplied data.

[1] The term customer energy use data (CEUD) is used in this report as described by the voluntary code of conduct under development by the utility industry, and facilitated by DOE. Privacy Voluntary Code of Conduct, facilitated by the United States Department of Energy's Office of Electricity Delivery and Energy Reliability and the Federal Smart Grid Task Force. Draft as of 8/12/2014.

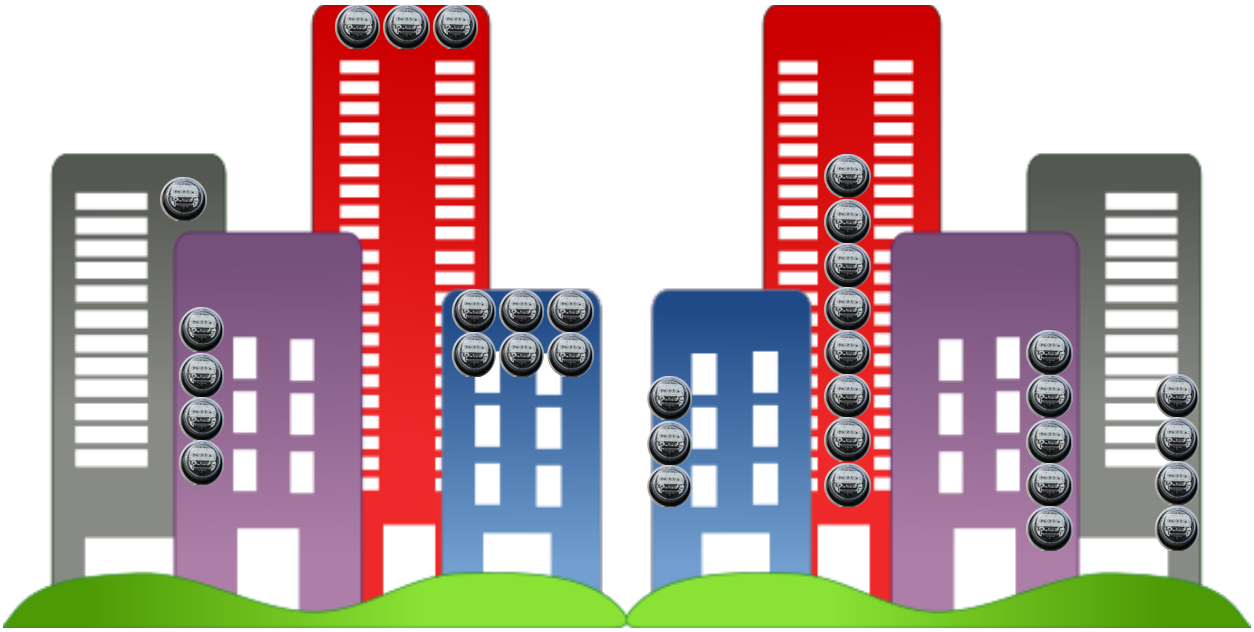

Figure 1. Buildings categorized by number of meters.

Although larger buildings might tend to have more meters and smaller buildings less meters, this is not always the case. A large building can have one meter representing one tenant, while a relatively small building could have several meters representing several tenants. We also recognize that in some multi-meter buildings, one of the meters may also be the owner's meter, which could have the chillers and the heat for the whole building with only plug load on the tenant's meters. This remains as the limitation of the analysis as qualitative information about the meters is not normally captured in the utilities' account records. Total energy consumption for the building is the summation of all the meters at that building. The study examines the monthly energy consumption (electricity and/or gas) at the building level.

The risk of matching individual CEUD is determined by statistically comparing the individual meter profiles of each building to that building's average building meter profile (ABMP). The ABMP is simply the total monthly consumption profile for the building, divided by the number of meters at that building. The report estimates the proportion of individual meter profiles (referred to as CEUD) that cluster with the ABMP. This is distinct from the concept of reidentification, noted in the draft voluntary code of conduct mentioned earlier, which combines the CEUD analysis with data from other sources to identify a specific tenant.

Based on our research findings, four-meter buildings are the first meter aggregation level not subject to simple deduction techniques for estimating individual CEUD. This issue is discussed in more detail in Section 6, Turnover. Establishing a lower threshold would make estimates of individual CEUD relatively simple, while higher threshold would significantly diminish the set of buildings eligible for reporting. For example, if a building has three tenants and one tenant moves out, the risk of matching CEUD increases for the remaining two tenants, rising to that of the 2-meter building. Figure 2 shows the hypothesized relationships that the study attempts to quantify.

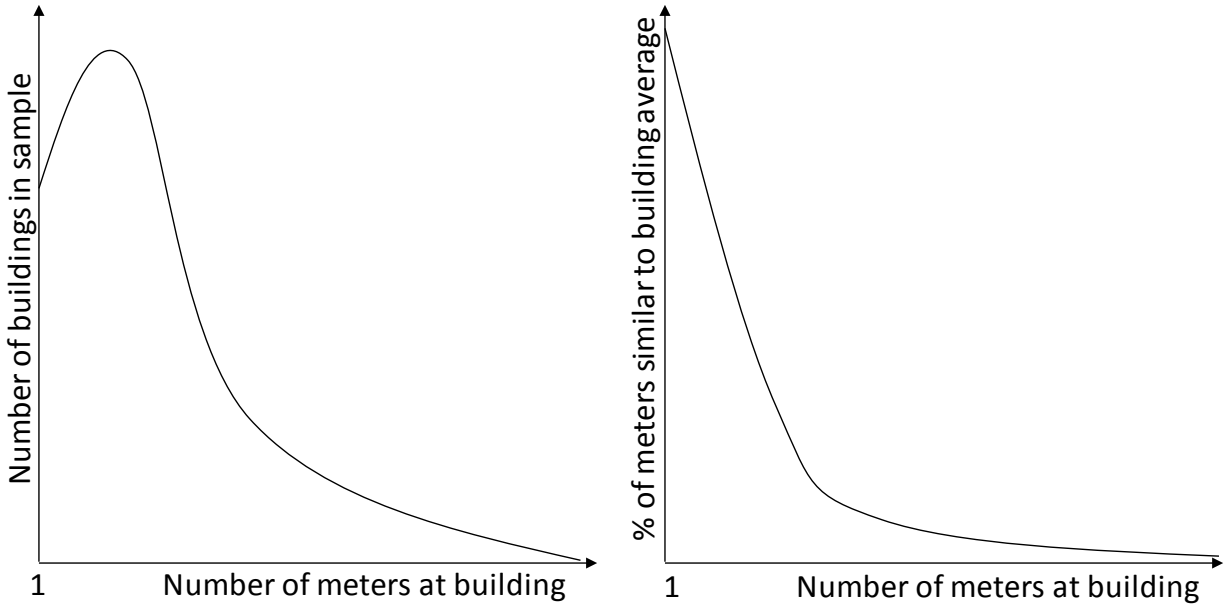

Figure 2. Hypothesized relationships between the number of meters at a building and expected energy consumption reporting eligibility.

The left graph illustrates that as the number of meters at a building increases, it is expected that there would be fewer and fewer buildings in the sample for any single utility. The right graph indicates that as the number of meters aggregated increases, the proportion of CEUD that resembles the average building profile diminishes rapidly, until at some threshold, it levels out and the added protection from further increasing the aggregation threshold is negligible. Both figures suggest that as the level of meter aggregation increases, there will be fewer and fewer buildings eligible for energy consumption aggregation.

With these concepts in mind, the goal of this study is to establish a quantitative approach for providing practitioners, such as utilities, public utility commissions, and other policy-makers with a defensible approach to aggregation threshold selection that protects tenant privacy without excluding so many buildings that its value becomes severely limited.

Our initial research did not identify existing studies reporting empirical evidence for determining appropriate multi-meter building energy data aggregation thresholds. As a result, PNNL was tasked to undertake statistical analysis that could help stakeholders address these issues and understand the implications of selecting different aggregation thresholds.

Several utilities and public utility commissions (PUCs) have made concerted efforts to research the issues of disclosure and methods to ensure data privacy. The most recent effort is the comprehensive report by the Working Group for the Energy Data Center phase of the California Public Utilities Commission (CPUC) Rulemaking 08-12-009.[2]

[2] California Public Utilities Commission. Decision Adopting Rules to Provide Access to Energy Usage and Usage-Related Data While Protecting Privacy of Personal Data. Rulemaking 08-12-009, May 1, 2014. http://docs.cpuc.ca.gov/PublishedDocs/Published/G000/M090/K845/90845985.PDF

While there exists a body of literature that discusses various methods for data aggregation and techniques for making the data anonymous, the emphasis is typically given to the following aspects:

1. Potential concerns with data disclosure at a finer time interval (e.g., hourly data).
2. Review of past disclosure cases in which identities of the reported subjects have been compromised within a very short amount of time from the disclosure, even when personally identifiable information (PII) was not included in the reported information.
3. Potential alternatives to the most widely used aggregation methods and techniques for making the data anonymous (e.g., a data cube with structured queries).
4. Theoretical research on privacy metrics, data anonymization, and randomization algorithms that are methodological in nature.

As will be discussed in detail in Section 3, existing technical literature lacks the very aspect that utilities are seeking, namely a specific recommendation on an easily implementable and justified approach to aggregation rules.[3]

Section 2.0 of this report discusses the current data use cases under consideration within the utility industry. Section 3.0 examines the available literature on data privacy and aggregation methods. Section 4.0 presents an overview of the analytical methodology. Section 5.0 discusses the data and explains the most relevant analytical results. Section 6.0 reviews participant turnover. Section 7.0 contains results and conclusions, and a bibliography is in Section 8.0. Appendix A includes additional comparison of the EUI estimates available from CBECs with the EUIs for the data set that contained some information on building characteristics.

[3] Currently used rules are summarized in section 2, Table 2.1, Summary of Known Aggregation Rules and Thresholds.

2.0 Meter Data Use Cases

This section describes current cases under consideration by the utility industry regarding the provision and use of CEUD. Reviewing these use cases places this study in the context of the current thinking regarding the release of CEUD.

This study is focused on the case of using data for energy benchmarking with ENERGY STAR Portfolio Manager (ESPM), but it also looks at other potential use cases for monthly energy consumption data. ESPM is the most commonly used tool for building energy use benchmarking and disclosure. Initially, the task was intended to focus on releasing monthly totals of whole-building energy consumption specifically to the building owner/manager for the purpose of submitting the data to ESPM. After preliminary discussions with some of the utilities, it became apparent that there is a much broader set of data disclosure issues beyond the data submission to ESPM. In addition, stakeholder interviews confirmed that various utilities use different aggregation rules, with thresholds for data disclosure being selected based on the consensus of the parties developing the disclosure protocols.

DOE engaged several industry experts to advise this study. Discussions with these experts identified several important issues. The risks of sharing data with the building owner are different from those of sharing it with other parties because the building owners/managers possess situational information—they already know the tenants and handle tenant-sensitive information as part of the lease transactions. In addition, building owners and managers usually have physical access to individual tenant meters, as meters often are located in common areas.

A clear distinction should be made between reporting monthly consumption totals to the building owner or manager specifically for the purpose of energy benchmarking and energy management and providing data to them with no constraints as to the further disclosure of the data to other parties. For example, release of aggregated building-level monthly energy consumption to the building owner by the utility is all the utility can control. To facilitate data access for benchmarking, while assuring tenant privacy, the building owner shares the burden of tenant data protection. If the building owner decides to use energy data for multiple purposes, it is beyond the control of the utility.

Given the fact that building owners often have access to these same data, even without a utility participating in the data flow, it is necessary to clearly define the use cases as follows:

- Case 1: Utilities reporting monthly CEUD to building owners or managers only for the ESPM submission and energy management or transmitting the data directly to a building owner's ESPM account
- Case 2: Monthly CEUD being provided by the utilities to outside parties that do not have any special right to the data
- Case 3: Monthly CEUD being provided for multiple buildings aggregated into reporting "groups" at the city block, neighborhood, or zip code level

The basis for these use cases is a recent CPUC decision[4] which contains a larger list of use cases and relevant recommendations from which the three cases above were derived. The primary difference between the CPUC list and the list derived for this analysis is that the CPUC list considered cases with PII and small time-interval data being requested by a third party. We focused only on the monthly energy consumption data and considered data release to building owners or directly to their ESPM accounts (Case 1) rather than general unconstrained release to third parties and the public at the higher level of aggregation beyond a building (Case 2 and Case 3). Our list took CPUC User Case 7 as a very narrow subcase and extended it to include a set of possible data interests, but with a focus only on monthly total energy consumption data in multi-tenant buildings.

Case 1 represents the lowest risk case regardless of the aggregation limit. For Case 1, the risks of having a utility sharing individual CEUD with the building owner may be better mitigated by requiring the owner not to use it for purposes other than energy management or to share the information with other parties unless required by law.

It should be noted that there is some misunderstanding that data from ENERGY STAR Portfolio Manager can be accessed in a way similar to Commercial Buildings Energy Consumption Survey (CBECS) or Residential Energy Consumption Survey (RECS) microdata. To clarify, only ESPM "account owners" have access to building data within their ESPM accounts. Moreover, data submitted via ESPM accounts are not used to calculate or update the benchmark; individual building data are only compared against the CBECS-based benchmark. Note that Case 1 does not provide either the building owner or manager with much more additional information beyond that to which the building owner/manager already has access (meters), that has already been provided by the tenants as part of the property-management process (e.g., lease contract with PII in multifamily buildings), or that can be observed directly by the owner/manager onsite as part of day-to-day operation or occupancy of a building.

These conditions are consistent with the definitions of privacy, privacy breach, and uninformative principle in Machanavajjhala (2008) for several reasons. First, building owners already know who their tenants are, so "reidentification" of the tenant by the owner does not necessarily constitute privacy breach. We acknowledge that there are substantial privacy rules and practices in place in the states that cover utility data privacy and access. Simply being able to read the meter is not a sufficient rationale to dismiss the privacy issues associated with estimating individual tenant CEUD. However access to the tenant information is what distinguishes building owners from the general public. This distinction allows the disclosure to be classified such that preserves the uninformative principle of Machanavajjhala (2008), which requires that the published data should give an attacker very little additional information beyond background knowledge.

Second, entry and exit of tenants would only help building owners estimate the monthly energy consumption attributable to a tenant rather than reveal any personal patterns of hourly or daily activities. Third, inference from the monthly consumption totals to the specific end-uses has been an open subject in the energy efficiency literature for several decades, yet no proven methods exist that would help the building owner unambiguously establish the relationship between tenant monthly energy consumption and tenant's specific energy usage preferences. Fourth, the utilities track the energy usage at the meter, not at tenant level.

[4] California Public Utilities Commission. Decision Adopting Rules to Provide Access to Energy Usage and Usage-Related Data While Protecting Privacy of Personal Data. Rulemaking 08-12-009, May 1, 2014. http://docs.cpuc.ca.gov/PublishedDocs/Published/G000/M090/K845/90845985.PDF

Case 1 also highlights the importance of having appropriate nondisclosure constraints apply to building owners/managers because they can have either access or an ability to obtain tenant information beyond monthly energy consumption data, even without the utility being part of the process. Breach of Case 1 also leads to Case 2, where CEUD are released to outside parties who have no specific rights to the data. The difference is that Case 1 implies a much smaller data volume, assuming owners/managers would receive data for their respective buildings only. Releasing CEUD in one large volume allows anonymization and randomization of the building-level data such that identification of any particular building has to be accomplished before attempting estimate individual CEUD. When CEUD are collected and released in large volumes, it is more likely to contain sufficient diversity (many types of buildings, various operating regimes, various sizes and occupancy levels, etc.) at the building level to prevent identification of individual buildings in the data set, thus increasing the challenge to discern individual tenants within buildings. Data collected and disclosed in smaller increments does not allow for this additional level of protection. For example, the identification of outlier profiles is relatively simpler to accomplish on a small data set than a large and diverse data set.

Case 3 implies a much broader scope of data disclosure/release. During the stakeholder interviews, utility representatives indicated that there is an increased interest from various parties in obtaining energy consumption data coupled with the building characteristics and sometimes owner/tenant demographic information for residential buildings and business-related information for commercial or industrial buildings. There is a wide range of reasons for these requests, starting from energy analysts and building scientists and engineers who are pursuing energy efficiency analysis to green building interests who are seeking more data to understand and influence trends and tendencies in energy consumption at a local level (e.g., down to the neighborhood level).

Some of the utilities tried to accommodate this growing trend and streamlining building owner access to data by proposing a rule along the lines of 15/15 or 4/80. For example, the 4/80 aggregation rule requires that at least 4 customers be aggregated in the reported total with contribution from any single aggregated customer not exceeding 80% of the total. Although this is a fairly simple type of rule that appears easy to implement, it can have some undesirable characteristics. With a 15/15 rule for example, distributions of multi-tenant and multi-meter buildings for several different utility samples indicate that less than 2% of the buildings have 15 meters or more. Therefore, a 15/15 rule may not be a feasible threshold for tenant/meter data aggregation, as it only provides this streamlined approach to less than 2% of multi-tenant buildings.

The 15/15 rule is just one example of existing rules and practices. Current aggregation practices differ significantly among utilities. Table 1 is a list of known aggregation rules and thresholds for a sample of utilities. Lack of specific practical recommendations in the existing literature forced the utilities to select aggregation thresholds and rules based on internal consensus without pursuing broader justification for each selected threshold.

Table 1. Summary of Known Aggregation Rules and Thresholds

Utility	Aggregation Threshold[(a)]	Reference
Austin Energy (TX)[(b)]	4 (4/80 rule)[5]	Institute for Market Transformation – Utilities' Guide to Data Access for Building Benchmarking (March 1, 2013)
Avista Utilities (WA)	No threshold	Leona Doege, Avista Utilities, via CPUC.
Colorado PUC (CO)	15 (15/15 rule)	Code of Colorado Regulations (CCR) 723-3 Part 3. Under Revision.
Commonwealth Edison (IL)	4	Webinar with Kevin Bricknell, "Energy Usage Data System," July 2013
ConEdison (NY)	No threshold	ConEdison http://www.coned.com/energyefficiency/PDF/FAQ-Aggregated-Consumption.pdf
Pepco (Washington, D.C.)	5	Building Electricity Consumption Data Request Form, http://www.pepco.com/business/services/consumptionrequestform/
Puget Sound Energy (WA)	5	Presentation by Chris Thompson, "Energy Data and Benchmarking," Energy Efficient Buildings Hub Regional Data Management Working Group Meeting, October 25, 2012.
Seattle City Light	No threshold	

(a) Aggregation Threshold for Sharing Whole-Building Monthly Data with Building Owners/ Managers Without Explicit Tenant Consent

(b) Applies only to commercial buildings.

Source: RFEE_2013_energy_data_access_map-proposed_discussion_framework_051313, Local Government Sustainable Energy Coalition, May 13, 2013. Provided by CPUC.

Case 3 includes an even broader use case aggregation of multiple buildings into reporting "groups" at the city block, neighborhood, or zip code level. It also includes monthly "group" energy consumption data provided to the general public. An important aspect of properly defining this use case includes establishing the characteristics that allow building energy consumption to be grouped in a way that is meaningful for energy analysts, but still ensures reasonable data privacy. This case can be addressed either through a data cube (i.e., no nested queries allowed to dissect the group aggregates) or by providing monthly energy consumption summary tables. Although this case also requires establishing appropriate aggregation characteristics and limits, and increases reporting burden on the utilities, it may result in a level of information release that is generic enough to mask individual observations appropriately, while providing useful information to analysts, efficiency advocates, and the general public.

Note that Case 3 is not intended to provide representative results that can be generalized to the state, census division, or other higher level. It will require a common understanding that the presented data form a report on energy consumption for a narrowly defined group of buildings, and do not constitute a national or regional representative sample - and therefore should not be used as such.

[5] Similar to other naïve aggregation rules: at least four members should be included in the group, with contribution from each not to exceed a prespecified level.

3.0 Literature Review

3.1 Background

A common characteristic of the literature covering aggregation is its lack of theoretical justification and specific recommendations as to the aggregation techniques and thresholds beyond what are commonly referred to as "naïve aggregation rules." Naïve aggregation rules are rules similar to the 15/15 rule discussed previously, which requires that the data should be summed across at least 15 customers with individual contributions not to exceed 15% of the total. The most recent discussion of this rule appears in the *CPUC Data Access Working Group Report*. This report is by far the most comprehensive outline of issues with existing rules, challenges associated with establishing methods for new rules, and the contradicting views of various parties regarding the definitions, methods, and implications of various user cases. In addition to the comprehensive coverage of the data disclosure issues, the CPUC report contains references to several studies and research reports that were provided to support the Working Group. The following is a list of the most relevant references accompanied by a brief summary:

1. Danezis G. 2013. Privacy Technology Options for Protecting and Processing Utility Readings. Microsoft Research. Cambridge, Massachusetts Available at: http://research.microsoft.com/en-us/projects/privacy_in_metering/privacytechnologyoptionsforsmartmetering.pdf

This study contains an overview of issues with data disclosure, a summary of potential threats, a discussion of the problems with naïve aggregation rules, and a list of potential solutions for how to protect data and customers from potential data breaches or reidentification. The privacy technology options proposed by the author are robust in comparison with existing methods of anonymization and aggregation include allowing a limited list of authorized and overseen researchers samples of the anonymized data, allowing access to only aggregates and statistics on the data (as opposed to microdata itself), or requiring a data export and user authorization mechanism for those cases when microdata are required. While this technical report provides some of the ideas on mitigating data breaches or reidentification risks, it does not contain specific recommendations as to specific aggregation schemes or thresholds that can be applied immediately to deal with monthly energy consumption data in multi-tenant buildings.

2. Dwork C and M Hardt. 2013. Privacy Preserving Data Analysis for the CPUC Energy Data Center EFF "Technical Issues" memorandum to working group participants, CPUC Rulemaking 08-12-009 (Phase III Energy Data Center), April 1, 2013.

The second study proposes two methods to accommodate data disclosure without compromising privacy. The first proposed method is a data cube. A data cube prevents the analyst from seeing the database entries. Only the summary data are presented. "For a given set of attributes, a data cube shows, for every possible setting of the given attributes, how many records in the database matched the particular setting." Instead of providing an analyst with the counts on actual data, another option is for the summaries from the data cube to contain only approximate answers, which differ from the original by some error. The third option is to allow the analyst to work with "synthetic data," which would match the original data on all data cube cells with the incorporation of the approximation error. The difference between the second and the third methods is that for data approximation in the second method, the original data are called and then error is added to the presented result. In the third method, synthetic data with approximation errors are produced as a substitute, which then are queried by the analyst.

Another approach discussed in this technical report is an interactive query system. By inserting an additional piece of software between the database and the analyst, the analyst is prevented from seeing the actual database. The only option to interact with the database is through the intermediate software that attempts to jointly interpret all of the analyst's queries and determine whether they constitute a privacy risk. Based on the determination, the distortion is added by the software to the displayed results.

3. Blasco B and J Byren. 2013. Legal Considerations for Smart Grid Energy Data Sharing. Electronic Frontier Foundation (EFF). Legal and technical memo. http://docs.cpuc.ca.gov/PublishedDocs/Efile/G000/M064/K670/64670678.PDF

4. Blasco B and J. Byren. 2013. Technical Issues with Anonymization and Aggregation of Detailed Energy Usage Data as Methods for Protecting Customer Privacy. Electronic Frontier Foundation (EFF). http://docs.cpuc.ca.gov/PublishedDocs/Efile/G000/M064/K670/64670678.PDF

The third and fourth technical pieces are memoranda from the Electronic Frontier Foundation (EFF). They contain an overview of legal statements that, in the authors' opinions, provide the basis for treating energy consumption data as covered information under the CPUC privacy rules, and therefore should be subject to prior customer consent before the data can be shared with third parties. Note: a distinction differentiating building owners or managers from other third parties is not made.

The EFF memoranda and the Data Access Working Group report contain explicit examples of how both the data cube and interactive queries can be used to identify customers. Emphasis is given to the fact that none of the individual data cube requests or interactive queries may present a privacy threat, but when intelligently constructed, they ultimately lead to unintended customer identification.

In addition, the Working Group requested additional information from technical experts regarding the feasibility of the data cube application. Technical experts noted that no "off-the-shelf" data cube software was available; therefore, utilities would have to carry the burden of software development, support, and customization and that, if any utility decides to pursue this endeavor, it will put them on the very frontier of software research and development in this field.

There is a conclusion in the report that deserves special attention. The report draws attention to the following aspect, "… the Draft Report appears to take for granted that data blurring or other technical approaches to confound re-identification are the only set of solutions. Another solution would be simply to require third-parties seeking customer level energy usage data to execute a contract, under penalty or appropriate recourse that forecloses any efforts to re-identify customers. By making CEUD estimation a violation of the contractual terms, such a requirement could mean that usage data cannot be reasonably used to identify or re-identify a customer, given the consequences violators would face." Although the second statement does point to a long-established solution for minimizing identification attempts by a party seeking the data, it does very little to protect privacy in case of a data breach. An extension of this discussion on privacy concerns and approaches to data protection is also available in Borgeson (2013).

5. Doran K, F Barnes, P Pasrich, and E Quinn. 2010. Report on Consumer Privacy and the Smart Grid. *Smart Grid Deployment in Colorado: Challenges and Opportunities*. University of Colorado, Boulder, Colorado. Available at: https://www.smartgrid.gov/sites/default/files/doc/files/Smart_Grid_Deployment_in_Colorado_Challenges_Opportunities_201003.pdf

This report contains an overview of potential concerns related to energy consumption data becoming a commodity, specifically highly detailed information that can be extracted from raw data collected through advanced metering for the smart grid. Doran et.al point out that the amount of information that is collected by the electric utilities is growing rapidly. Meanwhile, they see potential for an increasing market incentive to tailor these data as a commercially viable commodity. They express concern that existing privacy protection imposed by state statutes may well be inadequate to prevent such behavior. The suggested solutions are to 1) require an opt-in prior to sale or disclosure of information, or 2) apply the Connecticut Department of Public Utility Control's definition of protected "customer information" as "customer-specific information which the electric distribution company or its predecessor electric company acquired or developed in the course of providing electric distribution services and includes, but is not limited to, *information that relates to the quantity, time of use, type and destination of electric service, information contained in electric service bills*." This essentially means that monthly consumption data would be treated as "customer information;" therefore, the utility would be required to "receive prior affirmative, written customer consent."

6. Solove D J, M Rotenberg, and PM Schwartz. 2006. Information Privacy (2006). Discussing the European Union (EU) Data Protection Directive of 1995, Directive 95/46/EC. http://ec.europa.eu/justice_home/fsj/privacy/law/index_en.htm).

After summarizing the EU requirements for data protection based on the European Union Data Protection Directive of 1995, the authors provide a set of recommendations for defining the operating parameters in the context of energy data protection. Some of the representative items are included below:

- Metering and energy usage data should be considered the property of the customer, regardless of whether these data are kept by the customer, utility, or demand-response service provider.
- Electrical and gas corporations, as well as demand-response providers, should be prohibited from sharing customers' energy usage information with third parties unless the customer expressly authorizes the disclosure in writing (i.e., agrees to opt-in).
- Personal information should be defined as "… any information that identifies or describes a family, household, or residence."

This effectively would also treat monthly energy consumption totals for the residential sector as personal data.

3.2 Literature Findings

The review of relevant publications revealed that, in response to rising concerns about advanced metering and implications of implementing other smart grid enabling technologies, the most recent literature is predominantly focused on protecting data with small time intervals (second, minutes, and hours). There is a large set of valid concerns expressed by those seeking to protect customer privacy, as well as parties attempting to find the most optimal set of policies that would allow a reasonable level of data disclosure without unduly compromising customer privacy. Many utilities are facing a growing number of requests for energy-use data without having a conclusive set of specific rules to guide the release of data. Based on conclusions in several relevant technical studies, data anonymization, differential privacy techniques, cryptographic solutions, and aggregation could serve as means to prevent the identification of energy usage of individual customers. However, specific implementation of the

means to protect privacy is case-dependent, as it is strongly dictated by the objectives of the analysis, the nature of the requested data, and the role of the parties requesting access.

We anticipate that, for some of the potential data requests, aggregation by itself may not be an applicable or sufficient privacy protection method as demonstrated in detail in Machanavajjhala (2008) and discussed in Borgeson (2013). However, for Cases 1 and 2 defined in this report, easily implementable and justifiable aggregation rules could be an adequate approach for providing a sufficient level of privacy protection. Current aggregation literature is limited, and most of the theoretical work is devoted to computational aspects of randomization, anonymization, and differential privacy algorithms. This report is an attempt to bridge the gap by conducting an empirical analysis of the identification risks based on the utility-provided monthly energy consumption data at the meter and building level.

The next section of this report discusses the logic underlying the analysis. Sections 4 and 5 describe the methodology and results for the six data sets individually and collectively.

4.0 Methodology

An objective of this study is to analyze the degree to which individual CEUD are similar to average building profile and as a result, can be estimated based on comparison to the ABMP. The emphasis is on finding a suitable approach to selecting the minimum threshold for aggregating CEUD at the building level, such that building monthly total energy consumption can be provided without compromising tenant privacy or obtaining disclosure agreements from individual tenants.

Assessment of risk associated with estimating individual CEUD at different levels of aggregation consists of two main components. The first component determines how likely the CEUD matching might be; the second deals with assessing the consequences. The product of these two components defines risk in traditional risk analysis. This study focuses on providing quantification of the percentage of individual CEUD that is close enough to ABMP to be estimated by dividing the building total by the number of meters at a building.

Quantification or monetization of the consequences in this particular context is an intractable problem. The consequences are case-specific and may vary from minor irritation of a residential tenant who does not want to be bothered with targeted marketing materials (even by a public program with the objective to improve energy efficiency), to inadvertently revealing the business sensitive information of a commercial customer with the subsequent potential for tangible financial losses. For example, a tenant that has a level of energy consumption that is on the extreme side as compared to similar tenants or buildings—something of the magnitude that corresponds to having highly energy-intensive 24-hour operation—may indicate inefficiency. As a result, business sensitive information may be revealed. Alternatively, it may entail not just tangible business consequences, but also law enforcement implications. Because of the high variability in what the potential consequences may be, this analysis leaves out the intractable problem of quantifying or monetizing the consequences of tenant identification. Instead, the analysis focuses specifically on understanding the probability of CEUD estimation at different meter aggregation levels.

No formal definition or metric exists for the probability of CEUD matching that could be immediately applied to the context of the aggregation analysis in this report. The utility data facilitate analysis of whether individual consumption profiles could be estimated based on the aggregate profile for a building. Specifically, we have relied upon the percentage of individual meters that is statistically similar to the average building profile, or ABMP, as an indication of how likely individual consumption profile can be estimated based on total building consumtion divided by the number of meters. It should be noted that energy ***profile*** matching is one step removed from ***tenant*** reidentification.

As was discussed in the previous sections, there is no off-the-shelf model, statistical procedure, or test that provides a concise answer to the question of optimal profile aggregation. The literature that addresses aggregation as a means of preventing identification is limited, while research on anonymization and randomization techniques is getting more attention. The review of the anonymization and differential diversity or differential privacy studies discussed in Section 3 led to several key observations:

1. Within the context of the use cases discussed in Section 2, the anonymization of buildings is not an option as ESPM requires building-specific information along with the energy consumption data to be entered either by the building owner or submitted directly by the utility.

2. If the disclosure is for purposes beyond ESPM, removing all of the quasi-identifiers (i.e., the building characteristics that potentially can be used in combination with the data with other data sets to identify a specific building) is not always an option, as that will render the data set useless for energy efficiency research.
3. Regardless of the aggregation threshold level, only building totals, not subgroup totals, should be reported. For example, if selected aggregation threshold is 5 and there are 10 meters in the building, the total should be reported for all 10 meters, not a subtotal for two groups of 5 meters. Otherwise if repeated or nested querying for various groups of 5 is allowed, it enables a composition attack.
4. Even if individual monthly consumption totals became known, they would be the meter totals, not always the tenant totals. Since the relationship between meters and tenants is not always a one-to-one link, this automatically incorporates an additional degree of separation.[6] An exception to this are single-entity buildings and premises where building or site totals should not be released based solely on meeting the aggregation threshold, irrespective of its level.
5. Again, even in the case that a tenant is successfully identified and linked to the data on the monthly total energy consumption, the current techniques of relating monthly consumption back to individual lifestyles and preferences are limited in their ability to accurately estimate consumption shares across end-uses. For example, Kolter et al. (2010) proposed an allocation method with the best procedure correctly classifying only about 55% of energy use starting with hourly data across nearly 600 homes and over 10,000 appliances.
6. Most importantly, if meter-level monthly profiles within multi-meter buildings possess at least some of the characteristics that are desired for anonymized data in the first place, then aggregating the meter-level profiles can be an adequate measure for protecting privacy.

The last observation was based on Machanavajjhala (2008), who laid out clear principles for privacy protection and illustrated each one of them with a specific example. Although his discussion is focused on anonymization techniques, several aspects are valid in the context of the use cases considered in this report. Since these aspects motivated our choice of analysis methodology, they are discussed below.

Upon initial examination of the data, it became apparent that the factors that influence identification of individual meter profiles have to do primarily with the variability of monthly energy profiles between meters within a building, between meters within buildings with the same number of meters, as well as the typical annual profiles observed across various buildings. The initial expectation was that the more uniform the profiles are, the easier it would be to isolate central tendencies that can guide estimation of individual profiles. Simply put, if the meter profiles look a lot like building average profiles and the magnitudes of the profiles are uniform, then individual meter profiles can be guessed from dividing the building total by the number of meters, which would constitute a homogeneity attack. Therefore, a high degree of homogeneity, or similarity of profile shapes and magnitudes, increases the risk of matching individual CEUD.

On the other hand, consistent with the principles of anonymity, specifically k-anonymity, a certain degree of homogeneity could be interpreted as desirable. For example, a data set is considered to be k-

[6] For the purposes of this analysis, we assume that meters and tenants are correlated, but the degree of correlation is not known. Using meter monthly energy consumption as a proxy for the tenant monthly consumption is the limitation of this analysis, as tenant-level data is not available. This substitution should not impact the applicability of the results.

anonymous if there are at least *k* profiles that cannot be distinguished from each other, and therefore, the individual profiles cannot be easily identified by linking attacks (cross-mapping with other databases based on the building and tenant characteristics). Note that in privacy protection literature, any attempt to identify an individual entry or narrow it down to specific sensitive data is termed as an "attack," regardless of the intention behind that attempt.

At the same time, Machanavajjhala (2008) also shows that this type of anonymity by itself is not sufficient. There should be a little diversity in the sensitive attributes. Thus, individual meter profiles should vary some to ensure an adequate level of privacy protection. However, at the same time they cannot vary so much that they become distinct to the point of being uniquely distinguishable. This balance is hard to quantify without having any readily available practical measure of exactly what degree of variability in the monthly meter profiles is acceptable.

The most applicable discussion from Machanavajjhala (2008) is that of the uninformative principle, which requires that the published data should give an attacker very little additional information beyond background knowledge. The principle is quantified as the difference of prior and posterior beliefs of an attacker within a Bayesian framework.[7] In our context this would mean that an attacker would obtain only very limited incremental information about the tenants energy consumption beyond what can be inferred from national surveys such as RECS and CBECS, city and county parcel data, phone books, or other public data sources.

Note that Machanavajjhala (2008) discusses all these aspects (k-anonymity, homogeneity, sufficient diversity of the data and uninformative principle) in the context of releasing unaggregated data, which would correspond to meter-level totals. We are pursuing the analysis of aggregation at the building level, but if the underlying meter-level data possess these characteristics to at least to some degree, we anticipate the following:

a. Aggregation could serve as adequate means of protecting individual meter/tenant monthly energy consumption.
b. Similarity of individual meter profiles to their building average at various aggregation thresholds is an indicator of how likely an individual meter profile can be estimated based on the building total energy consumption profile.

Therefore, the analysis should focus on understanding the variability in the shape and magnitude of the monthly meter profiles. If the variability is high, only a very small fraction of the meter profiles will resemble their building average. If the variability is low, a large percentage of meters will be similar to their building average. Large percentage of meters similar to the building average means a large percentage of meter profiles can be estimated based on the information about building total profile and the number of meters.

Figure 3 provides an illustrative example of meter profiles for a 5-meter building. It depicts the 5 individual monthly meter profiles, along with the profile of the building (aggregate of all 5 meters). Consumption is plotted in terms of the proportion of annual consumption by month. The basic methods outlined in the report simply evaluate how similar each individual profile is to the ABMP.

[7] Box, G. E. P. and Tiao, G. C. (1973) *Bayesian Inference in Statistical Analysis*, Wiley, ISBN 0-471-57428-7

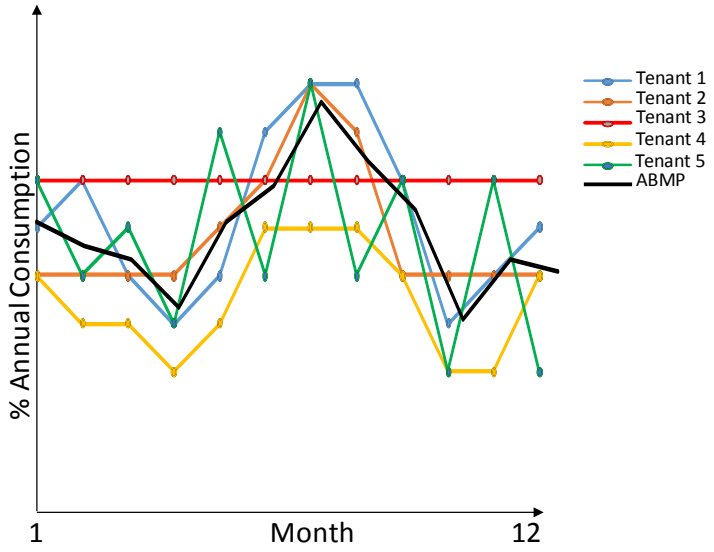

Figure 3. Illustrative meter profiles for a 5-meter building.

An exploratory data analysis (EDA) was performed in the statistical software *R* on the monthly building consumption data to understand how proportion of meters resembling their ABMP changes for each grouping. Groupings refer to the categories of buildings/accounts with different numbers of meters. For example, a grouping of 2 refers to all two-meter accounts. This allows us to understand average characteristics of meters and their similarity across 2 meter-accounts, 3 meter-accounts and so on. Comparing the average proportion of meters that resemble their ABMP for each grouping indicates what fraction of meters can be estimated from the building total for each of the aggregation thresholds.

Various descriptive statistics, data diagnostics, and graphical displays were analyzed in examining the hypothesis that an "attacker" could effectively guess the meter consumption profiles upon seeing the building consumption profile. Exploratory data analysis (EDA) tools employed in this study included k-means cluster analysis as well as descriptive statistics including, but not limited to, correlation, range, and standard-deviation estimation across several cross-sections of data—all used to describe the within- and between-group variability. These are generally well known statistical methods[8]; therefore, we are omitting their description from the report.

The most relevant results emerged from clustering meter profiles with ABMP. Clustering groups the meter profiles with similar/correlated meter profiles such that differences between the groups are maximized and the differences within groups are minimized. Average percentage of meters that clusters with their ABMP is used as an estimator for proportion of meters that are similar to their building average. Since clustering indicates that meters are similar, but does not provide information on the degree or direction of similarity, we also examined correlations between individual meter profiles and ABMP, as well as the ratio of annual meter energy consumption to ABMP annual energy consumption.

[8] Trevor Hastie, Robert Tibshirani, and Jerome Friedman (2001). The Elements of Statistical Learning. *Springer Series in Statistics Springer New York Inc., New York, NY, USA, (2001)*

5.0 Data

This section describes the data and explains the most relevant analysis results across all data sets. In some cases, data sets that contained both gas and electricity data by building, data across gas and electricity meters were first analyzed separately, but then converted to common units and rolled up to the building level to gain a better understanding of the full picture. Findings are reported as the summary of the generalized result across all analyzed utilities.

Six utilities from geographically and climatically diverse regions of the country provided anonymized data to support this analysis. Meters were mapped to accounts by utilities based on the billing records. Once that relationship was established in a data set, all auxiliary information enabling that mapping was removed prior to data transfer. No customer PII (names, phone numbers, etc.) or building addresses were provided; utilities were specifically requested to remove all PII from the data set. The names of the participating utilities and sample data are subject to nondisclosure agreement. Therefore, only summary statistics and comparative results are included in the discussion. Table 2 summarizes data set sizes across six participating utilities.

Table 2. Anonymized utility meter datasets provided for analysis

Provider	Meters	Buildings
A	34,208	11,597
B	52,893	16,066
C	63,091	13,352
D	106,791	23,469
E	400,382	47,011
F	57,242	17,318
Totals	714,607	128,813

Single-meter instances were removed from the analysis. Another source of complexity in tenant-meter relationship is the master-metered buildings, where there are multiple tenants and one meter that the owner controls. These master-metered buildings present themselves in data as single-meter buildings. Therefore, they are also excluded from the analysis. This was done to enable the required simplifying assumption that one meter equals one tenant. Therefore, single-meter buildings are treated as a proxy for single-entity or single-tenant buildings, and are excluded from the analysis. Meters with less than 9 months of data were also removed from the analysis. Note that aggregation does not apply to single-entity buildings as disclosing building total in this case reveals that entity's CEUD.

Aggregation of meters into subgroups of meters within a building is not allowed, as manipulating subgroup composition from query to query allows for profile estimation within the group via composition attack. For example, if the aggregation threshold is 6 and there are 12 tenants in the buildings, the aggregated profile is comprised by summing up all 12 individual tenant/meter profiles into one total, as opposed to having two aggregate profiles for two groups of 6.

Figure 4 illustrates building-level monthly energy-use profiles for the sample (log scale). The intent is to identify predominant building profiles within the data set. Figure 4 shows that the majority of the building-level profiles across participating utilities are either mostly flat or bell-shaped (shown with two red lines). The seemingly abnormal behavior in the bottom third of the plot (zigzag decreases and increases) is simply an artifact of using a log scale for display purposes.

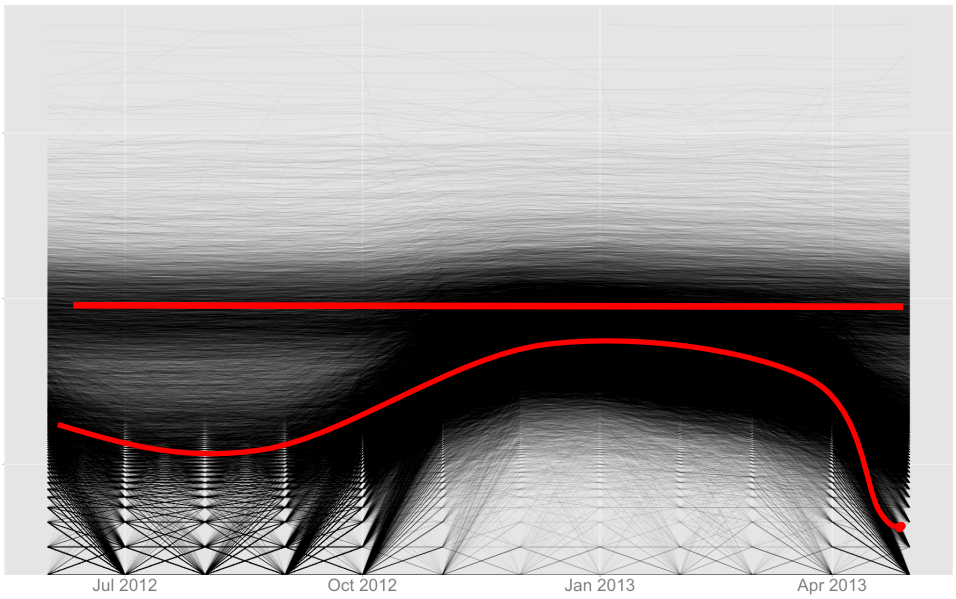

Figure 4. Predominant building profile shapes.

Total building profile divided by the number of meters is the ABMP. Clustering meter profiles with the ABMPs allows understanding the degree of similarity between the meter profiles and the building average.

Boxplots in Figure 5 show one of the most relevant results of the cluster analysis: the percentage of the meter profiles that fall in the same cluster as the corresponding ABMP. This analysis was done separately for each data set. Figure 5 shows a sample result for one of the participating utilities. The percentage of meters that cluster together with the ABMP is used to indicate how likely it is that an individual meter profile can be estimated from the building profile simply by dividing the building monthly totals by the number of meters. We are unaware of any previous work defining a metric for the probability of CEUD matching that could be immediately applied to the analysis in this report. Therefore, we develop a specific practical definition and attach a metric that could provide a meaningful quantification for utilities and other stakeholders. The percentage of individual meter profiles that cluster with their respective ABMP is such a metric.

In Figure 5, the middle line in the box plot indicates the median. The box represents the inner quartile range (IQR), which is the distance between the first and the third quartiles (25% and 75%). The upper whisker extends from the third quartile (75%) to the highest value that is within 1.5 x IQR. The lower whisker usually extends from the first quartile (25%) to the lowest value within 1.5 x IQR. Data that fall outside of the whisker range are plotted as points. Buildings are grouped based on the number of meters (x-axis).

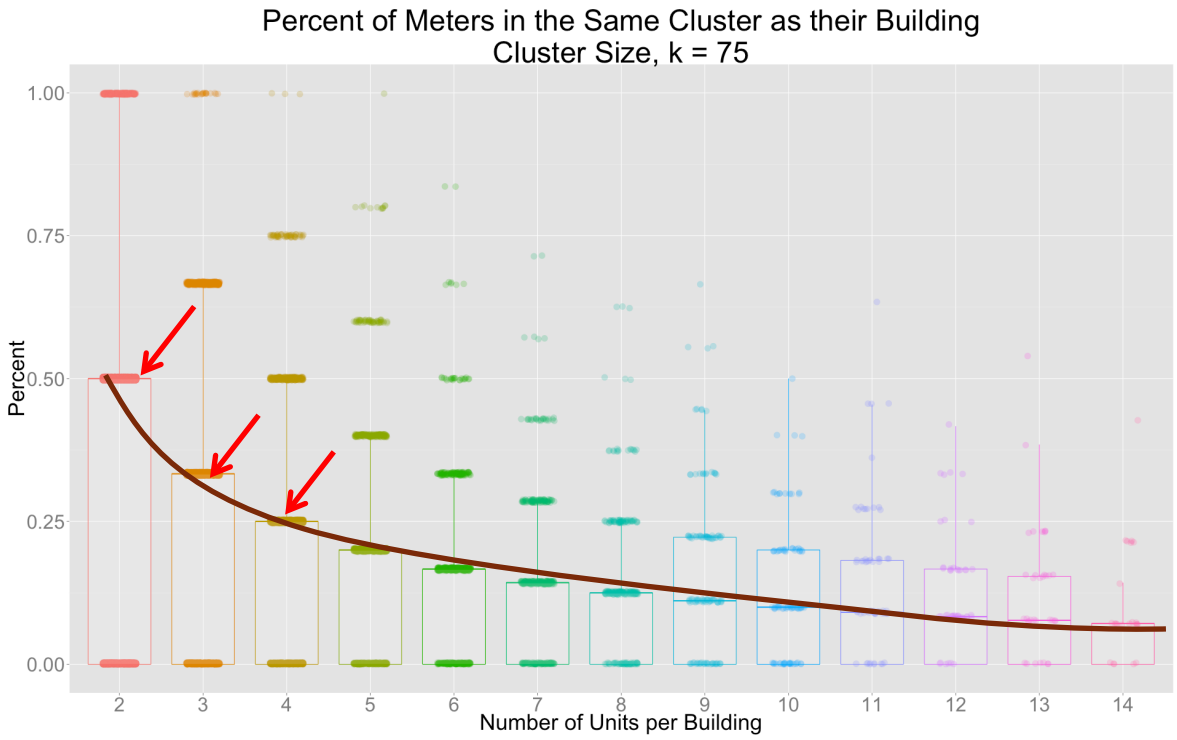

Figure 5. Percent of normalized profiles clustered together with their ABMP (one representative utility).

Let us first consider the boxplot for 2-meter buildings (the very first boxplot on the left) and then generalize the interpretation for the remaining cases. In 2-meter buildings, there are only three possible outcomes: 1) neither of the 2-meter profiles clusters with the ABMP, 2) only 1 of the meters clusters with the ABMP, or 3) both of the meters cluster with the ABMP. The median (50% of the 2-meter buildings) at 0.5 means that for 50% of the 2-meter buildings, 1 out of 2 meters cluster with their ABMP.

For a 3-meter building (second boxplot in Figure 5), the median is at about 0.3, which means that in 50% of the buildings roughly 1 out of 3 meters resembles the ABMP. In other words, the shape of the monthly consumption profile for 1 meter can be guessed from the overall building profile via division by the number of meters. Absent additional building information, this does not tell you which of the meters can be guessed in this manner, and it does not reveal any specific information about the remaining 2 meters.

In 4-meter buildings, the median of the boxplot is at 0.25, which implies that in 50% of cases, 1 out of the 4 individual meter profiles in 4-meter buildings clusters together with its ABMP. The median for 5-meter buildings is at 0.20, implying that in 50% of the 5-meter buildings, 1 out of the 5 individual meter profiles clusters with the ABMP. Connecting the boxplot medians in Figure 5 forms a curve that shows this decrease in the same manner for 6-, 7-, 8-, and 9-meter cases, out to 14 meters.

The desired metric, average percentage of meters that cluster with their ABMP, for 4- and 5-meter buildings is calculated in the same fashion as explained in the example with the 3-meter buildings. This desired metric represents the portion of meters similar to their building average for each aggregation threshold. It indicates what portion of meters can be estimated from the building total for each of the aggregation thresholds. This analysis step was performed separately for each of the data sets. Summary of the results across all participating utilities is presented in Section 7, Results and Conclusion.

6.0 Turnover

An issue that was not explicitly addressed in the previous discussion is that of tenant turnover. Analysis results show that aggregation can provide adequate privacy protection if monthly meter and building profiles possess certain desirable characteristics. There are some cases where aggregation is a necessary, but not sufficient, condition for protecting privacy. Tenant turnover is one of them.

As tenants move in and out, the composition of the reported set changes. While information on the tenant turnover is not as easily obtained for the multifamily buildings, the tenant turnover in the non-residential sector is assumed to be in the public domain. This constitutes background information or instance information, depending on how it is acquired.

This leaves monthly energy consumption totals open to composition attacks[9]. The source of vulnerability is not the aggregation threshold, but the rate and nature of the turnover combined with the background or instance-background knowledge.

For example, if a building has three tenants and one tenant moves out, monthly consumption of that tenant can be easily estimated by comparing the two totals for the months before and after the move. Moreover, the risk of profile matching increases for the remaining two tenants. Similarly, if there are three tenants and one more tenant moves in, that additional portion of the energy consumption in the following month would be a close estimate of that new tenant's energy consumption. The probability of identification does not increase for the other three tenants, but the new one is uniquely identified.

If the turnover rate is high, the effectiveness of the composition attack decreases. Establishing a quantifiable measure of the sufficient turnover based on its impact on the probability of profile matching is not a tractable task, as there are too many variables that can affect the outcome: the proportion of the impacted meter of the building total profile, the shape of the building total profile, or the timing of the move (e.g., small retail during the shoulder season or a high-rise office when a change cannot be distinguished from the weather response of the overall building). Cryptographic solutions akin to homomorphic encryption are recommended in the literature as a promising technique for preventing consumption profile matching via composition attack in this context. Homomorphic encryption masks the sensitive information in such a way that mathematical operations and statistical analysis are still possible, but the information itself cannot be viewed. A cryptographic solution not only would address the issue of profile matching under turnover, but will also significantly decrease the risk of profile matching for the rest of the tenants in the building. While differential privacy and encryption techniques are not part of this scope, we recognize their potential in resolving privacy concerns surrounding energy data disclosure.

[9] Composition attack is based on tracking an individual across disparate data publications using quasi-identifying attributes or across independent anonymized releases of the same data publication to breach privacy.

7.0 Results and Conclusion

The objective of this study is to analyze and report how likely it would be for individual CEUD to be estimated based on the ABMP of monthly energy consumption. The analysis was performed for various levels of meter aggregation to inform the selection of a minimum threshold for aggregating CEUD without compromising tenant privacy or obtaining disclosure agreements from individual tenants.

DOE tasked PNNL to undertake empirical analysis that could help utilities address the data aggregation issues associated with public release of building-level energy consumption data by utilities to building owners for the purpose of energy efficiency benchmarking and energy management, or to other entities who might use aggregated data for research purposes. Utilities want to understand the degree to which aggregation can enable improved energy management while protecting their customers' privacy.

PNNL analyzed data from six participating utilities who, under nondisclosure agreements, supplied anonymized monthly electricity and or gas consumption data by meter. Taken together the six utility billing data sets include the monthly consumption profile of nearly 715,000 non-residential meters, representing about 129,000 individual commercial buildings from geographically and climatically diverse regions of the country.

Table 3. Anonymized utility meter data sets provided for analysis.

Provider	Meters	Buildings
A	34,208	11,597
B	52,893	16,066
C	63,091	13,352
D	106,791	23,469
E	400,382	47,011
F	57,242	17,318
Totals	714,607	128,813

No formal definition or metric exists for the probability of profile similarity or CEUD matching that could be immediately applied to the context of the aggregation analysis in this report. The utility data facilitate analysis of whether individual consumption profiles could be estimated based on the aggregate profile for a building. It should be noted that energy ***profile*** matching is one step removed from ***tenant*** reidentification. Tenant reidentification can only proceed after an individual meter consumption profile is successfully matched, and relies on data collection apart from what utilities provide. For this study, we have relied upon the percentage of individual meters that is statistically similar to the average building profile, or ABMP, as a means to individual consumption profile estimation.

As discussed in Section 4, the percentage of meter profiles clustering with their ABMP was analyzed by first combining meter-level profiles with the ABMP for each building (building total profile divided by the number of meters), then performing the k-means clustering. The percentage of meters clustering with their respective ABMP indicates the fraction of meter profiles that can be identified based on total building consumtion divided by the number of meters. A summary result of the meter profiles that are similar to their building average profile across the six participating utilities is provided in Figure 6.

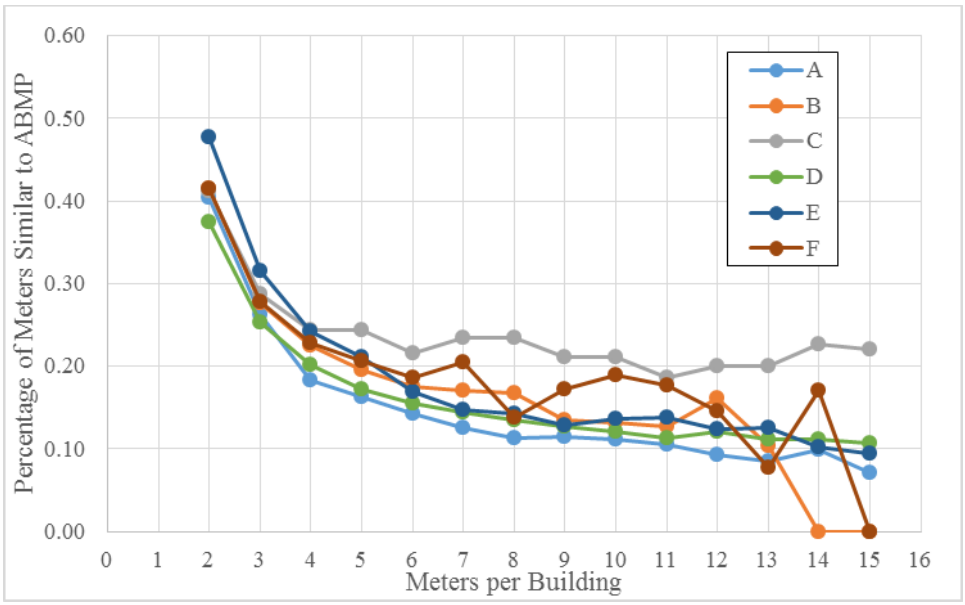

Figure 6. Proportion of individual meters similar to average building meter profile relative to number of meters per building.

Figure 6 illustrates how the proportion of individual individual meters resembling ABMP changes as the number of meters being aggregated increases. The x-axis depicts the number of meters per building (e.g., buildings with 2 meters, buildings with 3 meters, buildings with 4 meters, etc.) and the y-axis indicates the percentage of profiles that resemble the ABMP of their building. The figure illustrates that these relationships are similar across all 6 utilities for both the 3-meter and 4-meter building categories, ranging between 25-33 percent for 3-meter buildings and 20-25 percent for 4-meter buildings.

Moving up in meters per building (out along the x-axis), the spread between utilities increases and the proportion of meters resembling ABMP drops at a decreasing rate. Across utilities, in the case of four meters per building, the proportion of meters that are similar to ABMP is at or below 25 percent, encompassing all subsequent building meter counts within each utility.

Next, we must compare the similarity between meters and their ABMP with the number of buildings eligible for reporting. The relationship between aggregation level and reporting eligibility is that the more meters an individual building has, the fewer of such similar buildings there are – thus fewer to report in the data. The population of 4-meter buildings will typically be larger than the population of 5-meter buildings, and so on. Therefore, if we move up in meter aggregation, say from 4 to 5 meters, we lose a nontrivial number of buildings, because only 5+ meter buildings would be selected for reporting.

Table 4 shows the proportion of meters that are similar to their building profile and the percentage of buildings eligible for reporting under each aggregation threshold.

Table 4. Tradeoff between aggregation threshold and reporting eligibility.

Threshold (# of meters)	Percentage of Meter Profiles Similar to Their Building Profile (%)					
	A	B	C	D	E	F
2	40	42	41	37	48	42
3	26	28	29	25	31	28
4	18	22	24	20	24	23
5	16	20	24	17	21	21
6	14	17	22	16	17	19
7	13	17	23	14	15	20
8	11	17	23	13	14	14
9	12	14	21	13	13	17
10	11	13	21	12	14	19
11	10	13	19	11	14	18
12	9	16	20	12	12	15
13	9	10	20	11	13	8
14	10	-	23	11	10	17
15	7	-	22	11	9	-
Threshold (# of meters)	Percentage of Multi-Meter Buildings Coverage (%)					
	A	B	C	D	E	F
2	100	100	100	100	100	100
3	52	46	47	45	70	56
4	36	28	29	32	54	37
5	22	18	21	25	45	27
6	15	13	18	19	39	21
7	10	9	15	16	31	16
8	7	6	13	13	22	11
9	5	4	11	11	19	6
10	3	3	9	10	16	4
11	2	2	8	8	15	2
12	2	1	7	7	14	1.3
13	1.2	0.3	6	6	13	0.7
14	0.8	-	5	6	12	0.3
15	0.6	-	5	5	12	-

For example, if for utility B the threshold is set at 4, the proportion of meters resembling ABMP is about 22%, meaning that roughly one out of five meter profiles could be estimated from the building total profile by dividing it by the number of meters. The eligibility percentage is 28%, meaning that if an aggregation threshold of 4 is applied, 28% of multi-meter buildings would be eligible for reporting under this aggregation level.

Note that aggregation threshold of 4 provides an upper bound for the percentage of meters that resemble their average building profile across the six utilities analyzed (24%) and it is the first aggregation level that is not subject to immediate decomposition due to turnover discussed in Section 6.

For utility B, if the threshold were set at 5, the proportion of meters that are similar to their average building profile would drop by 2 percentage points to 20%. However, the data reporting eligibility rate for this aggregation level drops by 10 percentage points to 18%. Increasing the aggregation threshold from 4 to 5 drops 2 percentage points in proportion of meters that are similar to their average building profile, but loses 10 percentage points in coverage.

Figure 7 contains a summary of these incremental gains in percentage of meters that resemble their average building profile (left panel) vs loss in coverage (right panel) across six utilities:

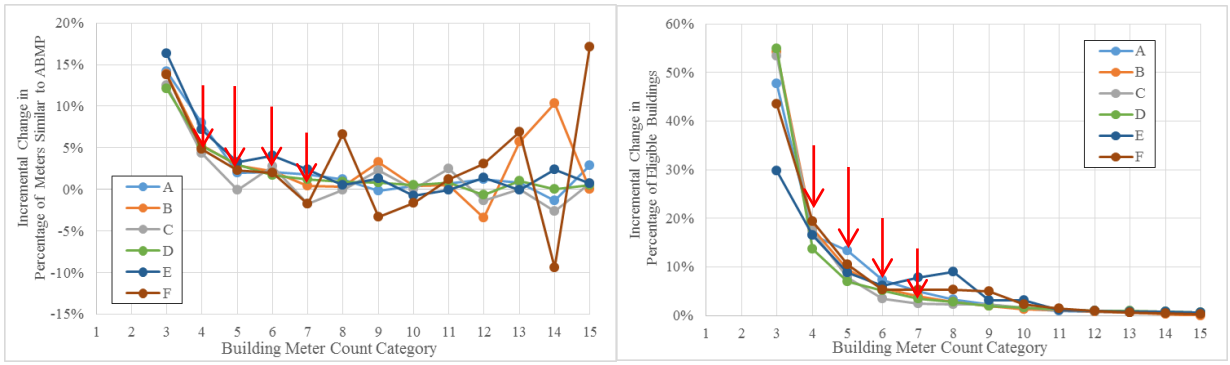

Figure 7. Marginal changes in percentage of meters resembling their building profile and building eligibility.

As aggregation threshold is increased from 5 to 6 for utility B, the percentage of meters that are similar to their average building profile drops by about 3 percentage points as shown on the left panel of Figure 7 (or from 20% to 17% as observed in the upper half of Table 4), while the loss of eligibility constitutes 5 percentage point as shown in the right panel of Figure 7 (or from 18% to 13% as observed in the bottom half of Table 4). If a decrease in percentage of meters that resemble their average building profile can be interpreted as a degeree of protection provided by each aggregation threshold, then gain in protection by increasing the threshold from 5 to 6 is exceeded by the loss of eligibility.

Similarly, increasing aggregation threshold for utility B even further, from 6 to 7, the percentage of meters that are similar to the building average drops by less than one percentage point. This less than one perrcentage point gain in protection comes at the expense of losing another 6 percentage points in building eligibility.

We understand that in choosing the aggregation thresholds decision makers may want to consider both the percentage of meters similar to ABMP and percentage of meters covered under a specific threshold. The tradeoff between incremental gain in protection vs loss in building coverage resulting from increasing the threshold, which is analyzed in this report, intends to inform the decision-making on the important matter of streamlining energy data access.

8.0 Bibliography

Abowd JM and SD Woodcock. 2001. Disclosure limitation in longitudinal linked data. Pages 215–277 in *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*. North-Holland, Amsterdam.

Aggarwal G, T Feder, K Kenthapadi, R Motwani, R Panigrahy, D Thomas, and A Zhu. 2005. Approximation algorithms for k-anonymity. *Journal of Privacy Technology* , pp.67-78.

Barak B, K Chaudhuri, C Dwork, S Kale, F McSherry, and K Talwar. 2007. Privacy, accuracy and consistency too: A holistic solution to contingency table release. In *Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (pp. 273-282). ACM.

Bayardo RJ and R Agrawal. 2005. Data privacy through optimal k-anonymization. In *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on* (pp. 217-228). IEEE.

Blasco B and J Byren. 2013. *Legal Considerations for Smart Grid Energy Data Sharing*. Legal and technical memo, Electronic Frontier Foundation (EFF). Available at: http://docs.cpuc.ca.gov/PublishedDocs/Efile/G000/M064/K670/64670678.PDF.

Blasco B and J Byren. 2013. *Technical Issues with Anonymization and Aggregation of Detailed Energy Usage Data as Methods for Protecting Customer Privacy*. Electronic Frontier Foundation, Available at: http://docs.cpuc.ca.gov/PublishedDocs/Efile/G000/M064/K670/64670678.PDF.

Borgeson SD. 2013. Public interest uses of smart meter data. Energy and Resources Group, University of California at Berkeley, Berkeley, California.

Brewer RS and PM Johnson. 2010. WattDepot: An Open Source Software Ecosystem for Enterprise-Scale Energy Data Collection, Storage, Analysis, and Visualization. Smart Grid Communications (SmartGridComm), In *Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on* (pp. 91-95). IEEE.

California Public Utilities Commission (CPUC). 2014. Decision Adopting Rules to Provide Access to Energy Usage and Usage-Related Data While Protecting Privacy of Personal Data. Rulemaking 08-12-009, May 1, 2014.

Chen B, L Chen, R Ramakrishnan, and DR Musicant. 2006. Learning from aggregate views. In *Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on* (pp. 3-3). IEEE.

Chin F. 1986. Security problems on inference control for sum, max, and min queries. *Journal of the ACM* 33(3):451–464.

Danezis G. 2013. *Privacy Technology Options for Protecting and Processing Utility Readings*. Microsoft Research, Cambridge, Massachusetts. Available at: http://research.microsoft.com/en-us/projects/privacy_in_metering/privacytechnologyoptionsforsmartmetering.pdf

Doran K, F Barnes, P Pasrich, and E Quinn. 2010. Report on Consumer Privacy and the Smart Grid. *Smart Grid Deployment in Colorado: Challenges and Opportunities*. University of Colorado, Boulder, Colorado. Available at: https://www.smartgrid.gov/sites/default/files/doc/files/Smart_Grid_Deployment_in_Colorado_Challenges_Opportunities_201003.pdf

Dwork C. 2006. Differential privacy. Differential privacy. In *Automata, languages and programming* (pp. 1-12). Springer Berlin Heidelberg.

Dwork C and M. Hardt. 2013. Privacy Preserving Data Analysis for the CPUC Energy Data Center. EFF "Technical Issues" memorandum to Working Group participants, CPUC Rulemaking 08-12-009 (Phase III Energy Data Center), April 1, 2013. Appendix to Decision Adopting Rules to Provide Access to Energy Usage and Usage-Related Data While Protecting Privacy of Personal Data. Rulemaking 08-12-009, May 1, 2014.

Efthymiou C and G Kalogridis. 2010. Smart Grid Privacy via Anonymization of Smart Metering Data. Smart Grid Communications (SmartGridComm). In *Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on* (pp. 238-243). IEEE.

Erkin, Z., Troncoso-Pastoriza, J. R., Lagendijk, R. L., & Perez-Gonzalez, F. 2013. Privacy-preserving data aggregation in smart metering systems: an overview. Signal Processing Magazine, IEEE 30(2):75−86.

Li, F., Luo, B., and Liu, P. 2010. Secure Information Aggregation for Smart Grids Using Homomorphic Encryption. In *Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on* (pp. 327-332). IEEE.

Garfinkel S. 2001. Database Nation: the Death of Privacy in the 21st Century. O'Reilly & Associates, Inc., Sebastopol, California.

Jia, W., Zhu, H., Cao, Z., Dong, X., & Xiao, C. 2013. Human-Factor-Aware Privacy-Preserving Aggregation in Smart Grid. Systems Journal, IEEE PP(99):1−10.

Shin K. and J Zhan. 2007. A Verification Scheme for Data Aggregation in Data Mining. In *Granular Computing, 2007. GRC 2007. IEEE International Conference on* (pp. 374-374). IEEE.

Kolter JZ, S Batra, and AY Ng. 2010. Energy disaggregation via discriminative sparse coding. In *Advances in Neural Information Processing Systems* (pp. 1153-1161).

Lambert D. 1993. Measures of disclosure risk. *Journal of Official Statistics* , *9*, 313-313.

Li N, T Li, and S Venkatasubramanian. 2007. t-closeness: Privacy beyond k-anonymity and L-diversity. In IEEE International Conference on Data Engineering. (Vol. 7, pp. 106-115).

Machanavajjhala A, D Kifer, J Gehrke, and M Venkitasubramaniam. 2007. L-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data* 1(1). http://www.cs.cornell.edu/~vmuthu/research/ldiversity.pdf

Machanavajjhala A. 2008. Defining and Enforcing Privacy in Data Sharing. Computer Science, Doctoral dissertation. Cornell University, Ithaca, New York. http://ecommons.library.cornell.edu/bitstream/1813/11192/1/thesis-ashwin.pdf

Martin D, D Kifer, A Machanavajjhala, J Gehrke, and J Halpern. 2007. Worst case background knowledge for privacy preserving data publishing. In Proceedings of the IEEE International Conference on Data Engineering. *ICDE 2007. IEEE 23rd International Conference on* (pp. 126-135). IEEE.

Reiter JP. 2005. Estimating risks of identification disclosure for microdata. *Journal of the American Statistical Association* 100:1103–1113.

Solove D J, M Rotenberg, and PM Schwartz. 2006. Information Privacy. Discussing the European Union (EU) Data Protection Directive of 1995, Directive 95/46/EC. Available at: http://ec.europa.eu/justice/data-protection/index_en.htm

Sweeney L. 2002. K-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* 10(5):557–570.

He, W., Liu, X., Nguyen, H., & Nahrstedt, K. 2009. A Cluster-Based Protocol to Enforce Integrity and Preserve Privacy in Data Aggregation. Distributed Computing Systems Workshops, 2009. International Conference on Distributed Computing Systems Workshops '09. 29th IEEE International Conference (pp. 14-19). IEEE.

Zhang Q, N Koudas, D Srivastava, and T Yu. 2007. Aggregate query answering on anonymized tables. In Proceedings of the IEEE International Conference on Data Engineering, ICDE 2007. IEEE 23rd International Conference on , pp. 116 – 125.

Appendix A

Overview of Building Attribute and Energy Consumption Data Sources

Appendix A

Overview of Building Attribute and Energy Consumption Data Sources

The uninformative principle for data disclosure discussed in Machanavajjhala (2008) requires that the published data should give an attacker very little additional information beyond background knowledge. The principle is quantified as the difference of prior and posterior beliefs of an attacker within a Bayesian framework[10]. In the context of monthly energy consumption, this would mean that an attacker would not learn much about the tenant's energy consumption at the individual level beyond what can be inferred from national surveys such as RECS and CBECS, city and county parcel data, phone books or other public data sources.

This Appendix contains a discussion of the data on buildings and energy consumption that is already publically available. The goal is to demonstrate, that when monthly energy consumption data at the building level is released by the utility, while it does lower the cost of profile breach, it does not violate uninformative principle. We compare average gross EUIs from the analyzed data set with the CBECS summary tables for gross EUI to demonstrate the idea.

While the target for estimation in this context is energy consumption rather than EUI. We also assume that the goal of monthly energy profile estimation is not just to learn the absolute NG or electricity consumption, but rather make inferences about the energy use preferences as compared to the group of peers. Therefore EUI per square foot is an adequate substitute for this comparison, because the two values are interchangeable via a rough estimate of the floor area.

There are several sources of information that enable identification of the gross floorspace of any building. First, the most obvious and easily accessible source is the parcel data available via the tax assessor's webpage for any county in the U.S. Second, some universities undertook an effort to compile the county tax records into searchable state-level databases. For example, parcel data for the State of Washington and the Portland metropolitan area from the University of Washington[11] can be used to a) obtain information on the age and floorspace of a building, b) construct distributions around the number of meters in the multiple-meter or multiple tenant buildings. Third, ParcelPoint, a patented technology developed by CoreLogic, contains parcel data for the whole U.S. by county for public purchase. ParcelPoint crafted the data by combining tax assessor data by county for the whole US.

Note, that these data, along with other information on building attributes, is already available to the building owner or energy manager. Tenant PII is also available, because it is provided as part of the property-management process (e.g., lease contract with PII in multifamily buildings). Moreover, building owners and energy managers have direct physical access to the meters, as they are often located in common areas or areas open for access to building owners and energy managers. We acknowledge that there are substantial privacy rules and practices in place in the states that cover utility data privacy and access. Simply being able to read the meter is not a sufficient rationale to dismiss the privacy issues

[10] Box, G. E. P. and Tiao, G. C. (1973) *Bayesian Inference in Statistical Analysis*, Wiley, ISBN 0-471-57428-7

[11] http://depts.washington.edu/wagis/projects/parcels/

associated with tenant reidentification. However access to the tenant information is what distinguishes building owners from the general public. This distinction allows the disclosure of total building energy consumption data to building owners such that the uninformative principle of Machanavajjhala (2008) is preserved. The risks of sharing aggregated energy consumption data with the building owner are different from those of sharing it with other parties because the building owners/managers possess situational information.

Where other outside parties are concerned, the natural question is what other energy consumption data is available in the public domain, which could allow forming what in Bayesian framework is referred to as prior belief at the level that supports the uninformative principle, i.e. the data released by utility is only incremental in comparison with already available energy consumption data in national surveys, such as CBECS, for example.

CBECS Tables C.9 and C.9A[12], which contain the summary of gross energy intensity by Census Division for the sum of major fuels for non-mall buildings and for all buildings, are used as a benchmark for comparison. The relevant portion of the summary is included in Table A.1 below. The first column shows a building attribute by category, the second column shows total consumption in trillion BTUs, the third column shows total floorspace for the buildings that fall into the corresponding category, and the fourth column shows gross EUI, measured in thousand BTU/sq ft, for the Pacific Census Division.

Table A.1. CBECS summary for gross energy intensity for sum of major fuels, 2003

Attribute	Energy Intensity for non-mall buildings(kBtu/ square foot)	Energy Intensity for all buildings (kBtu/ square foot)
All Buildings*	69.4	71.6
Building Floorspace (Square Feet)		
1,001 to 5,000	73.0	73.0
5,001 to 10,000	92.9	95.1
10,001 to 25,000	58.4	62.4
25,001 to 50,000	57.2	57.5
50,001 to 100,000	63.6	71.2
100,001 to 200,000	73.8	78.0
200,001 to 500,000	69.0	69.6
Over 500,000	Q	Q
Principal Building Activity		
Education	74.3	74.3
Food Sales	Q	Q
Food Service	Q	Q
Health Care	177.7	177.7
Inpatient	Q	Q
Outpatient	Q	Q
Lodging	71.8	71.8
Retail (Other Than Mall).....	52.8	73.4
Office	65.1	65.1
Public Assembly	Q	Q
Public Order and Safety	Q	Q

[12] Source: CBECS 2003, Table C9. Consumption and Gross Energy Intensity by Census Division for Sum of Major Fuels for Non-Mall Buildings, 2003: Part 3.

Attribute	Energy Intensity for non-mall buildings(kBtu/ square foot)	Energy Intensity for all buildings (kBtu/ square foot)
Religious Worship	Q	Q
Service	Q	Q
Warehouse and Storage	23.3	23.3
Other	Q	Q
Vacant	Q	Q
Year Constructed		
Before 1920	Q	Q
1920 to 1945	59.2	59.7
1946 to 1959	55.5	56.3
1960 to 1969	60.0	60.3
1970 to 1979	73.5	79.7
1980 to 1989	83.2	81.2
1990 to 1999	83.3	85.0
2000 to 2003	48.9	52.7
Number of Floors		
One	66.1	-
Two	61.6	-
Three	59.4	-
Four to Nine	106.4	-
Ten or More	Q	-
Predominant Exterior Wall Material		
Brick, Stone or Stucco	79.3	-
Concrete (Block or Poured)	77.4	-
Concrete Panels	46.1	-
Siding or Shingles	88.0	-
Metal Panels	48.9	-
Window Glass	Q	-
Other	Q	-
No One Major Type	Q	-
		-

Gross EUI derived based on electricity and natural gas data analyzed earlier in this report are summarized in Table A.2 below.

Table A.2. Summary for gross energy intensity, analyzed data set.

Attribute	Average	Standard Deviation	Standard Error	Relative Standard Error
All Buildings	106	298	1	1.27%
Building Floorspace (Square Feet)				
less than 1000.....	379	1,098	39	10.33%
1,001 to 5,000	148	359	3	1.78%
5,001 to 10,000	82	178	2	2.03%
10,001 to 25,000	59	122	1	2.04%
25,001 to 50,000	66	193	3	4.28%
50,001 to 100,000	83	307	6	7.06%
100,001 to 200,000	81	153	4	5.36%
200,001 to 500,000	85	285	13	15.49%

Attribute	Average	Standard Deviation	Standard Error	Relative Standard Error
Over 500,000	60	130	13	22.22%
Principal Building Activity				
Education	*118	394	40	*34.0%
Health Care	*175	1,413	61	*35.2%
Hospitality/Lodging	80	279	8	9.9%
Retail	157	296	2	1.2%
Office	53	94	1	1.6%
Public Assembly	81	571	16	19.4%
Public Order and Safety	58	26	2	4.2%
Warehouse and Storage	53	238	3	5.3%
Flex	68	142	3	4.4%
Year Constructed				
Before 1920	116	404	3	2.59%
1920 to 1945	96	195	3	3.09%
1946 to 1959	87	230	3	3.90%
1960 to 1969	99	204	3	2.97%
1970 to 1979	105	238	3	3.11%
1980 to 1989	88	183	2	2.66%
1990 to 1999	123	205	4	2.89%
2000 to 2003	109	207	4	3.59%
Number of Floors				
1	120	343	2	1.5%
2	65	120	1	1.8%
3	69	117	2	3.4%
4-9	73	114	3	4.3%
10 or more	80	151	10	12.09%
Predominant Exterior Wall Material				
Masonry.....	83	194	2	2.62%
Metal.....	58	209	6	9.97%
Reinforced Concrete.....	95	269	4	4.16%
Steel.....	88	141	5	5.66%
Wood Frame.....	104	269	3	3.19%
Unspecified.....	116	335	2	1.72%
*RSE>30%				

Comparison of the last column in Table A.1 and the first column in Table A.2 (for the estimates with the reasonable RSE) shows that while the gross EUIs for the analyzed data set are higher, the breakdown of EUIs based on the building square footage is follows a similar trend. Figure A.1 illustrates the comparison of CBECS EUI with the data set-derived EUI.

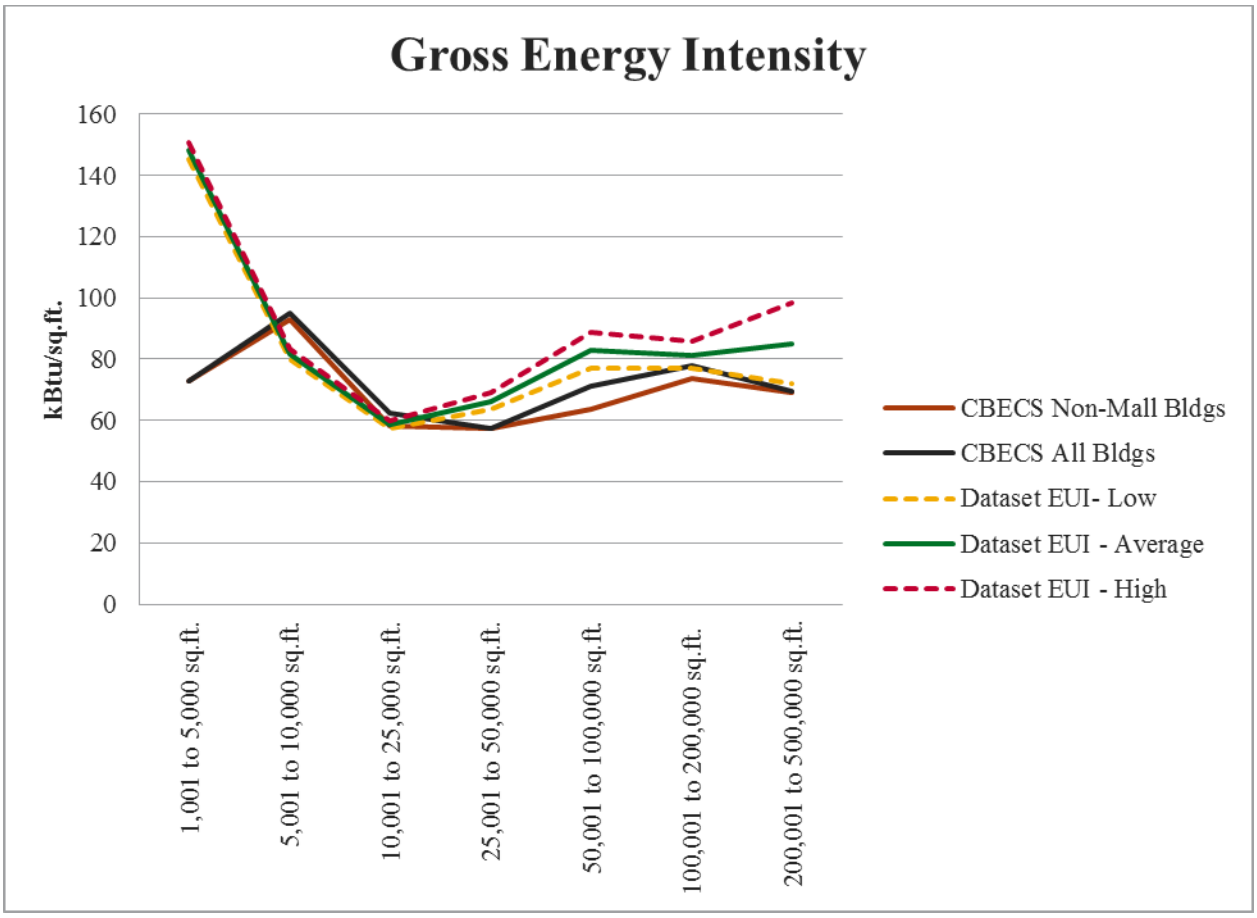

Figure A.1. Comparison of CBECS EUI with the data set-derived EUI.

Average EUI information at the gross building level, available from a national survey, while following a similar trend does not fall within the bounds that would allow it to be classified as statistically indistinguishable from the data set EUI across all of the building size categories. In general, it falls within 15% of the estimate except for the smallest and largest building categories.

On the other hand an argument can be made that it implies EUIs from the CBECS data can be used to support a background information attack. It should be noted that the sensitive information in the monthly total energy consumption discussed in this report lies not in estimating the average EUI by fuel type for the overall building, but rather finding individual relative EUIs for tenant's energy consumption, which would reveal individual tenant energy use preferences and behavioral patterns.

First, identification of individual tenant's EUIs is possible only to the extent that the individual meter/tenant monthly profile is identifiable, which is discussed in the main body of the report.

Second, in general, utilities collect monthly consumption information at the meter level, not the tenant level, and the tracking is done separately for electricity and natural gas. While we argue that estimation of the individual meter/tenant contribution in the building total electricity or natural gas monthly profile can be limited by aggregation, natural gas and electricity monthly meter totals for the same building are inherently dependent. True short-term substitutability between the two fuels is very limited. Separation

of the fuels in the reporting might make little difference for the probability of profile similarity or profile matching. While the report treats gas and electricity as separate data sets, we understand that such data can be easily combined at the building or account level to arrive at the total monthly EUI.

Third, the estimation that attempts to find individual tenant/meter EUIs by apportioning total building EUI based on the occupied square footage, would only provide an average crude estimate for all of the tenants in all of the buildings sharing the same characteristics. In other words, it does not reveal specific tenants energy preferences in a specific building, but rather provides an estimate of average consumption. Any kind of average estimates smooth out those fluctuations in the energy consumption that can be attributed to the individual tenant's energy-use preferences and emphasizes the trends that are usually attributed to the overall building response to weather and seasonality.

Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by **Battelle** *Since 1965*

U.S. DEPARTMENT OF
ENERGY

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99352
1-888-375-PNNL (7665)

www.pnnl.gov