PNNL-SA-106581

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Commercial Building Tenant Energy Usage Data Aggregation and Privacy: Technical Appendix

November 2014

OV Livingston
TC Pulsipher
DM Anderson

U.S. DEPARTMENT OF ENERGY

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-SA-106581

Commercial Building Tenant Energy Usage Data Aggregation and Privacy: Technical Appendix

OV Livingston
TC Pulsipher
DM Anderson

November 2014

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

Pacific Northwest National Laboratory
Richland, Washington 99352

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

This technical appendix accompanies report PNNL–23786 "Commercial Building Tenant Energy Usage Data Aggregation and Privacy". The objective is to provide background information on the methods utilized in the statistical analysis of the aggregation thresholds.

The goal of performing the same analysis on different datasets across multiple utilities is to determine whether the results and conclusions share any similarities across utilities in distinct geographic and climatic regions. The specific intent is to find a universal method for comparing the aggregation thresholds, illustrate how it is applied, and explain how the analysis results inform the choice of the threshold. While we understand that the datasets are region-specific, and each dataset may produce a slightly different result, the objective is to identify if there are any emerging trends despite the differences in the datasets.

No customer PII (names, phone numbers, etc.) or building addresses were provided; utilities were specifically requested to remove all PII from the dataset. 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.

Single-meter instances were removed 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. This analysis is focused on the aggregation of the tenants/meters within multi-tenant buildings to form the building total monthly energy consumption profile.

Aggregation of meters into subgroups of meters within a building is not allowed, as manipulating subgroup composition from query to query allows for re-identification 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. Aggregation of buildings into broader groups as defined in Section 2.0 of the main report is outside the scope.

While analysis of only one dataset is described in detail in this Appendix, the analysis flow is identical for all of the datasets analyzed in the report. Generalized results are reported in the main body of the study as well.

2.0 Analysis

Exploratory data analysis (EDA) tools employed in this study relied on unsupervised machine learning, specifically 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 generally well known statistical methods are described below in the context of this effort.

2.1 K-means Cluster Analysis

Hastie et.al. (2001)[1] provide a simple explanation of k-means clustering. The idea is to group the buildings with similar/correlated buildings such that differences between the groups are maximized and the differences within groups are minimized. Another way of saying this is to maximize the between-group variability, or distance, while simultaneously minimizing the within-group variation. Other clustering methods exist, though k-means is widely applied, and avoids the pitfalls of hierarchical methods that do not allow for a repartitioning of the cluster assignments.

The number of clusters, k, is determined *a priori* in k-means cluster analysis. Numerous metrics have been developed to determine the appropriate number of naturally occurring clusters. This analysis uses the "elbow method" for diagnosing the optimal number of clusters k. The elbow refers to the balance point between minimizing the total within-cluster variation while simultaneously minimizing the number of clusters.

K-means clustering will not return a specific building's profile as a representative profile for a cluster. Instead, k-means clustering finds the center, calculated as the mathematical average, for each cluster. Useful information extracted from EDA and clustering should identify outliers, provide understanding of the relationship between meter-level data and aggregated building level data, and help us understand meters and buildings data that should be tested against the aggregation strategies to determine a more robust algorithm/methodology for aggregation.

Another advantage of k-means clustering over other clustering procedures is that it can help determine the number of natural clusters (groups) in the data. For this report, we clustered both original and normalized profiles multiple times in an attempt to understand the variability between building profiles and to understand the similarity of the meter profiles to their building profiles.

2.2 Statistical Analysis

Correlation between the building consumption profile and the consumption profiles of individual meters in that building may inform as to the variability/uniqueness of each meter's profile. The initial expectation is that the majority of meters in a building exhibit the same behavior in their profiles as that of their aggregate, the building profile. The greater the number of meters in the building, the more likely this assumption does not hold.

[1] 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)*

The maximum correlation (minimum distance) suggests that a particular meter is most like the aggregate profile, whereas the minimum correlation (maximum distance) indicates the meter least like the aggregate profile. This analysis was performed on each category and visually displayed to give an indication of the variability of the relationships between meters and their aggregate by account size (number of meters).

To understand the degree of similarity between both shape and magnitude profiles across different dimensions, other general ad hoc statistical techniques were applied to the data. Analysis was performed on both normalized and original (untransformed) profiles in order to separate variability in shapes from variability in magnitudes. The higher is the homogeneity, the easier it is to estimate an individual meter profile from the building total.

Out of all the statistical methods that were applied to the meter and building monthly energy consumption data, the most relevant and easily interpretable results were obtained from

1. Investigating building profile variability
   - Clustering building profiles to determine the optimal number of groups and typical annual profile shapes
   - Analyzing the relationship between building attributes and cluster composition
2. Comparing the typical building profiles with typical meter profiles
   - Clustering with average building meter profile (ABMP)
   - Correlations between individual meter profiles and ABMP
   - Ratio of individual meter profiles and ABMP

The next section of the Appendix contains a more detailed description of the data, discussion of the observed variability, and figures of relevant results and their implications for estimating the individual meter profiles. Recommendations regarding the aggregation threshold for the explored data are provided in the conclusions.

3.0 Data and Results – Dataset 1

While analysis of only one dataset is described in detail in this Appendix, the analysis flow is identical for all of the datasets analyzed in the report. Generalized results are reported in the main body of the study as well. Dataset 1 contained information on both natural gas meter profiles and electricity meter profiles for June 2012-May 2013 along with the clear indication of which meters belonged together in one building. Dataset 1 is the only dataset out of 6 that comprised the actual metering data at the building level, not utility billing accounts. After matching the building data to the meter data and removing all the single-meter instances from the dataset, the subset with NG meter data (Dataset 1 – NG) included over 17,000 buildings spanning more than 57,000 meter profiles, while the subset with electricity meter data (Dataset 1 – Electricity) included approximately 9,600 buildings spanning over 26,000 meter profiles.

Since this is one of the few datasets that contained both NG and Electricity data by building, data across NG 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. Results here are reported for a Natural Gas subset only to explain the analytical approach. The main report contains the summary of the generalized result across both subsets. Generalized results across 6 analyzed utilities are reported in the main body of the study.

The subset with NG meter data included 17,318 buildings spanning 57,242 meter profiles for June 2012-May 2013. Buildings with 15 or more NG meters were excluded from the analysis as they represent only 2% of the multi-meter buildings[1] in the coverage area of the utility according to the building data count. The results discussion will focus on the buildings with up to 12 units, because the sample data for buildings with 12 or more units is limited as well.

Sampled building count by number of meters in the analyzed subsample (Dataset 1 - NG) is presented in Table 3.1 below.

Table 3.1. Sample Size by Category

Category	Number of buildings	Number of meters
2	8447	16894
3	3721	11163
4	2006	8024
5	1145	5725
6	700	4200
7	456	3192
8	325	2600
9	187	1683
10	115	1150
11	84	924
12	60	720
13	41	533
14	31	434
15 - excluded	22	330
Total count	17318	57242

3.1 Building profile variability

3.1.1 Optimal Number of Clusters for Building Profiles

Figure 3.1 illustrates building-level monthly energy use profiles for the sample (log scale). The intent is to identify predominant building profile shapes within the dataset. Figure 3.1 shows that majority of the NG building-level profiles for this utility 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.

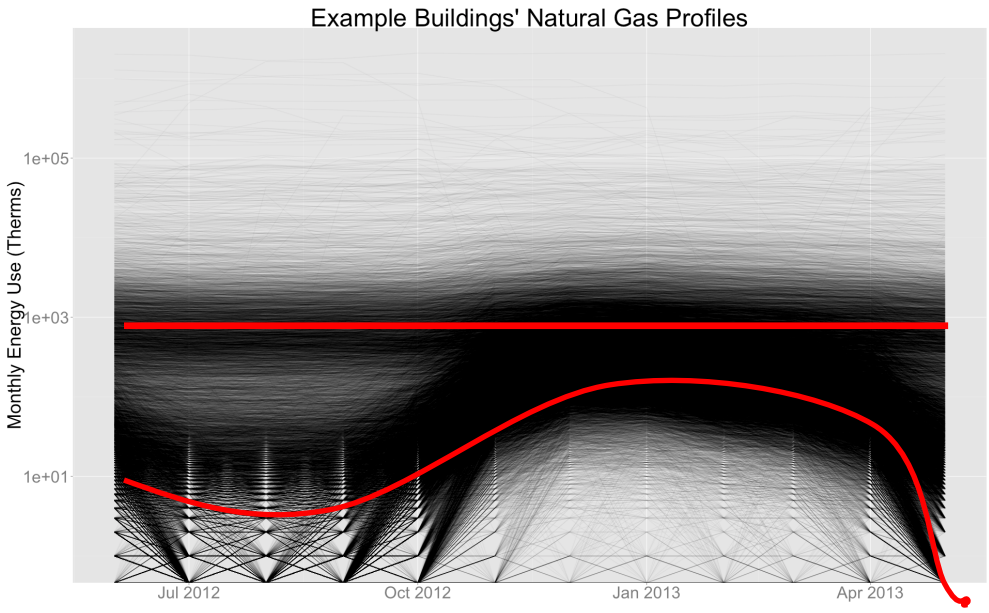

Figure 3.1. Natural Gas 12 Month Profiles for Dataset 1

One of the method objectives is to understand variability of profile shapes and profile magnitudes. To separate the variability of shapes from the variability of magnitudes the analysis of normalized profiles is also performed in parallel. Normalization is done by subtracting the profile mean, estimated as annual average, from each of the 12 monthly values and then dividing the difference by the standard deviation. This normalization suppresses magnitudes and allows focusing on variability in shapes.

Normalized profiles are presented in Figure 3.2. Note that the normalized profiles are centered at zero and their value indicates how many standard deviations the observation is above or below the mean. Two distinct shapes are shown in different colors.

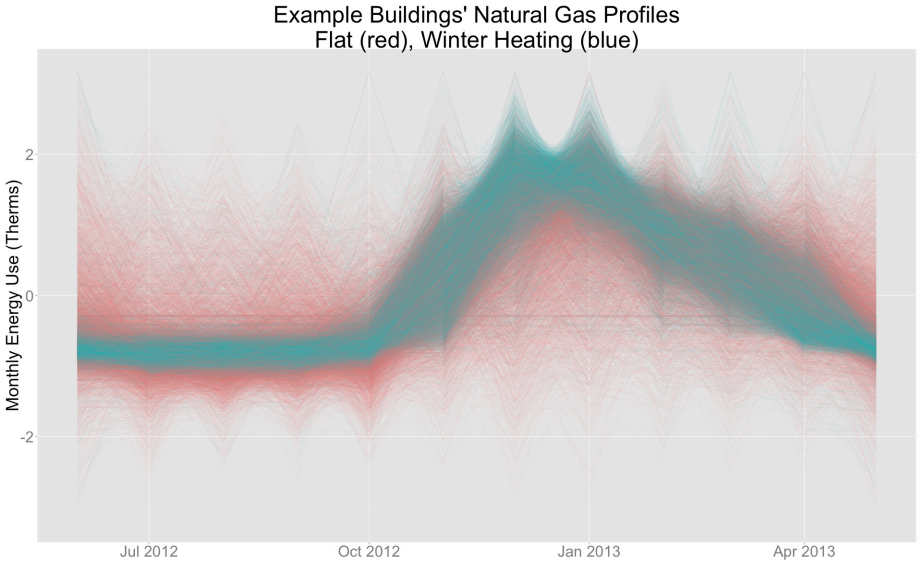

Figure 3.2. Natural Gas 12 Month Profiles for Dataset 1, Normalized

To answer the question whether there are essentially two typical profiles (flat and bell-shaped), a cluster analysis was performed on both original and normalized profiles. K-means clustering takes into account distance between profiles when partitioning the profiles into groups. Thus not just the shape but also the magnitudes dictate how the profiles are partitioned. To determine the number of naturally occurring groupings, which represent distinct profile shapes in this data, it is necessary to perform clustering on the normalized profiles as well.

The K-means procedure requires an analyst to specify the number of partitions or clusters. This is something of a drawback of k-means cluster analysis: in trying to find out what typical profile shapes look like, an analyst has to tell an algorithm how many of them are there in the data to start with, i.e. how many partitions or bundles the data should be split into. Then by comparing the distance between the profiles, the algorithm decides which profiles belong together and which one of the groups they fall into.

There are multiple methods for choosing the optimal or natural number of clusters. Figure 3.3 shows a diagnostic plot that is used as one of the standard methods to determine the optimal number of clusters. The horizontal axis shows the number of clusters. The vertical axis shows the total within the sum of squared distances between the members of the cluster. The latter is a measure of how tight the clusters are for each selected number of clusters, as explained below.

Within sum of squared distances is a standard measure of cluster tightness which represents the degree of similarity between cluster members. *Within sum of squared distances* is calculated as a distance between the cluster members (12-month profiles that fall in the grouping) and the cluster center. The cluster center is the mathematical average of all the members in the cluster.

The smaller the *within sum of squares* is, the tighter the cluster is. In general, the clusters get tighter as the number of clusters increases. The curve in Figure 3.3, often called an "elbow plot," captures this relationship. To determine the optimal number of clusters, the researcher looks for the elbow in the curve, nominally the area of the curve where increasing the number of clusters results in a diminishing reduction of the total *within sums of squares*. This metric suggests that for the NG dataset with 17,318 buildings, k = 100 is the optimal number of clusters in unnormalized profiles.

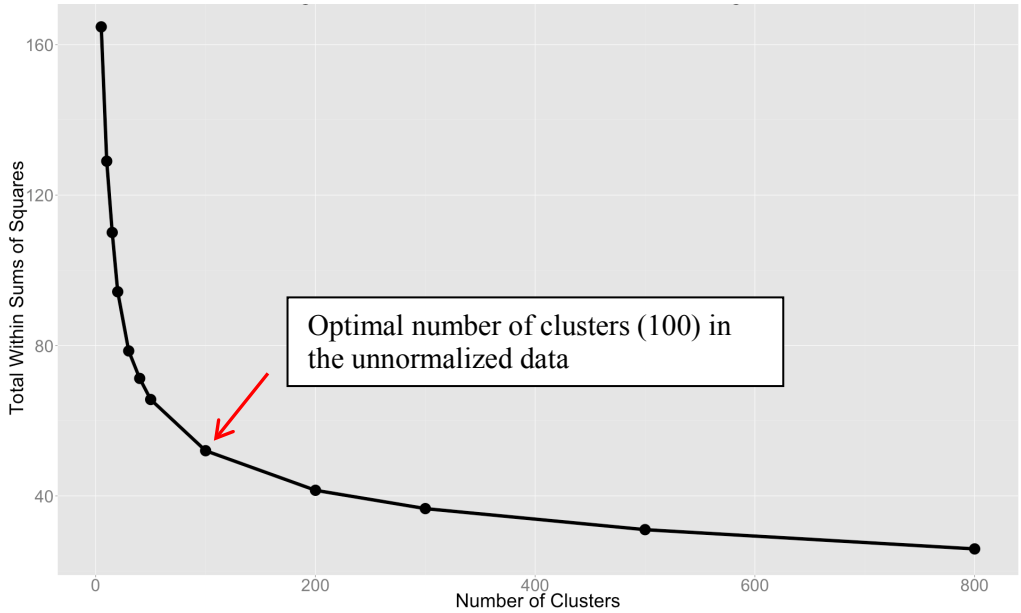

Figure 3.3. Elbow plot of the building NG profile clusters, K-means, original data

Figure 3.4 shows the elbow plot for the normalized profiles. The optimal number of clusters for normalized data is k = 50. Given the two basic building profile shapes that are observed in the data as illustrated in Figure 3.1 (flat and bell-shaped), this large number of natural clusters in the normalized data implies that because of the variation in building profile shapes the generalization at the level of the two basic shapes will result in crude oversimplification.

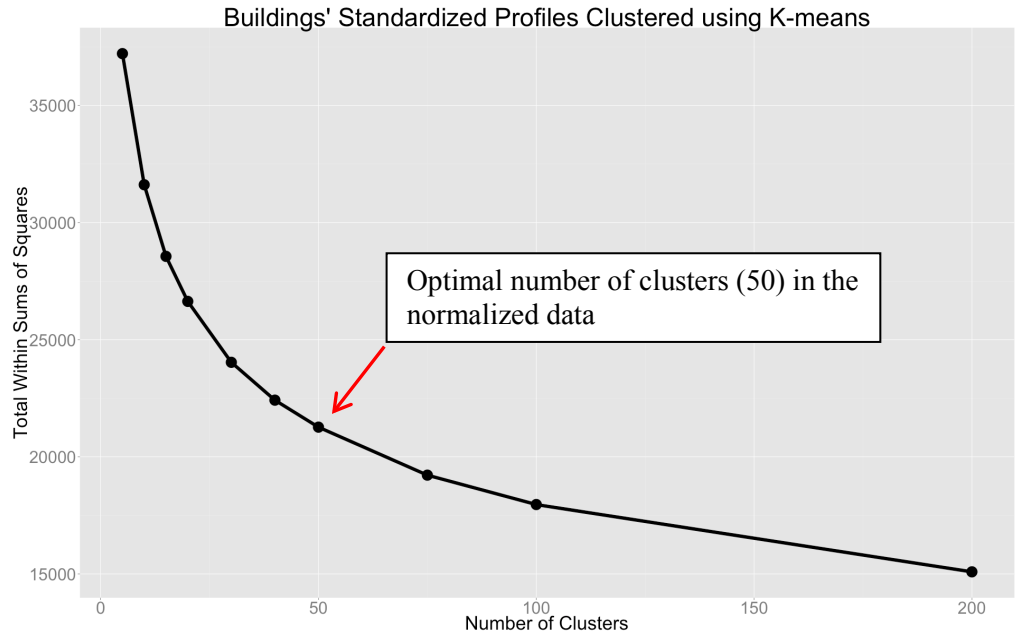

Figure 3.4. Elbow plot of the building NG profile clusters, K-means, normalized data

3.1.2 Cluster Sizes for the Optimal Number of Clusters

Figure 3.5 and Figure 3.6 show the breakdown of the optimal clusters for the original and normalized data by size (k=100 and k =50, respectively). The horizontal axis shows the number of buildings in a cluster, the vertical axis (height of the bar) shows how many clusters of that size are present under optimal clustering.

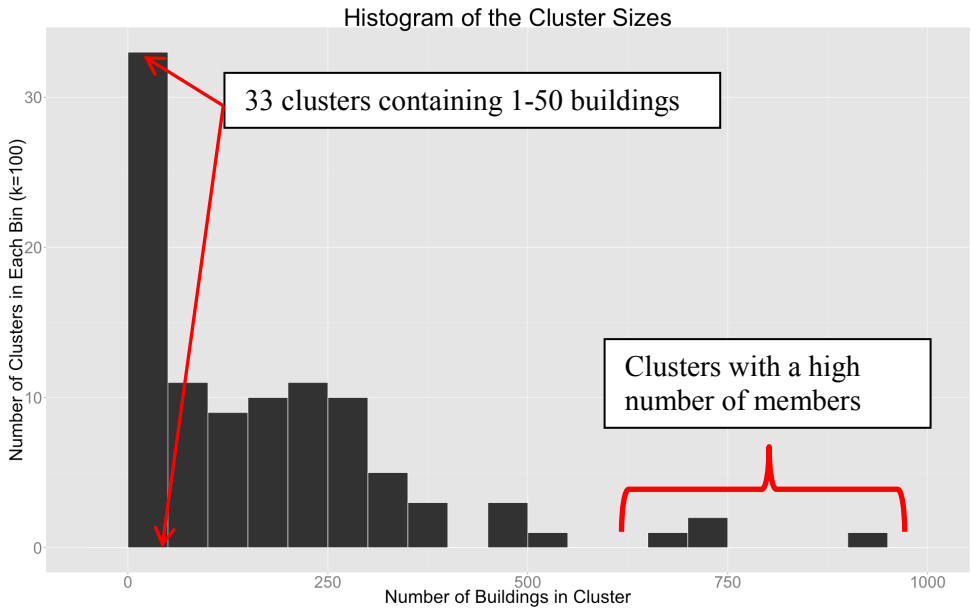

Figure 3.5. Histogram of Cluster Sizes for Original NG Data, Dataset 1

As shown in Figure 3.5, there is a large amount of smaller clusters in the original unnormalized data. For example, the first bar in Figure 3.5 indicates that out of 100 optimal clusters there are 33 clusters that include 50 buildings or less, the second bar shows there are 11 clusters with 50-100 buildings, 9 clusters with 100 -150 buildings, 10 clusters with 150-200 buildings and so on. Note that there are quite a few larger clusters, specifically several clusters with 650-700 and 700-750 members, as well as one with 900-950 members. In the context of this analysis, the clusters with the relatively high number of members represent the dominant building profile types in the unnormalized data. This point is explained in more detail within the Cluster Composition section.

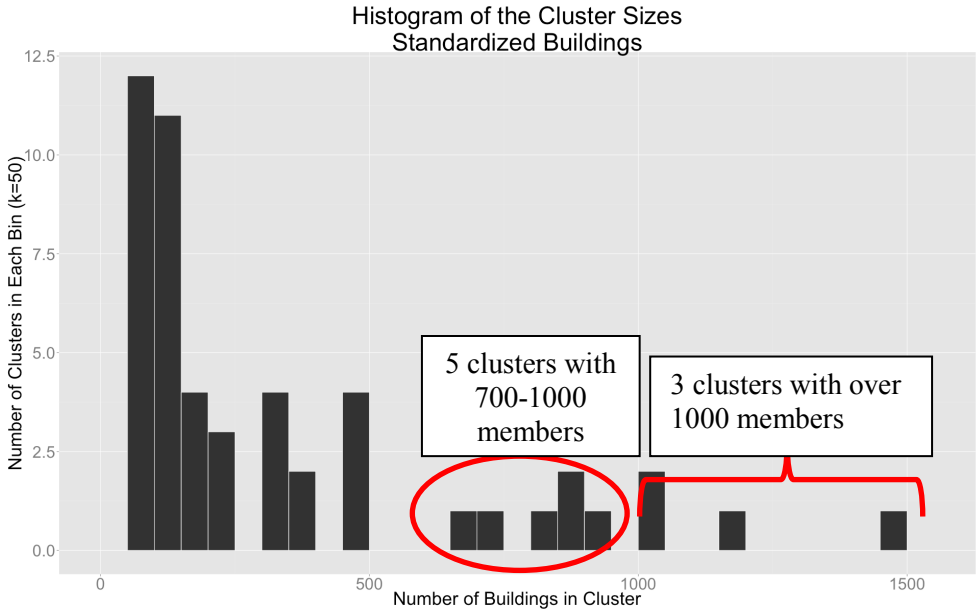

Figure 3.6. Histogram of cluster sizes for normalized NG data, dataset 1

Figure 3.6 shows the same cluster size breakdown, but for normalized data. When the profiles are normalized (i.e. magnitudes are suppressed to give more emphasis to the differences in shapes), a significant portion of the buildings end up in clusters with 700-1000 members (5 bars circled in red in the middle of Figure 3.6), and clusters with over 1000 members (the last three bars on the right). These 8 clusters with high number of members represent dominant profile shapes in the normalized data. The remainder of the profiles are grouped together in the clusters with less than 500 members (first 7 bars on the left of Figure 3.6).

3.1.3 Cluster Composition

Further review of the clusters indicates that besides the clusters with a high number of members, which represent dominant profiles, there are quite a few small clusters. Presence of unique profiles in the data explains the large number of small clusters. Figure 3.7 and Figure 3.8 depict cluster composition for original and normalized data. Cluster number is indicated above each individual graph. Circles on Figure 3.7 and Figure 3.8 denote clusters with the large number of members, i.e., the dominant profiles in each set.

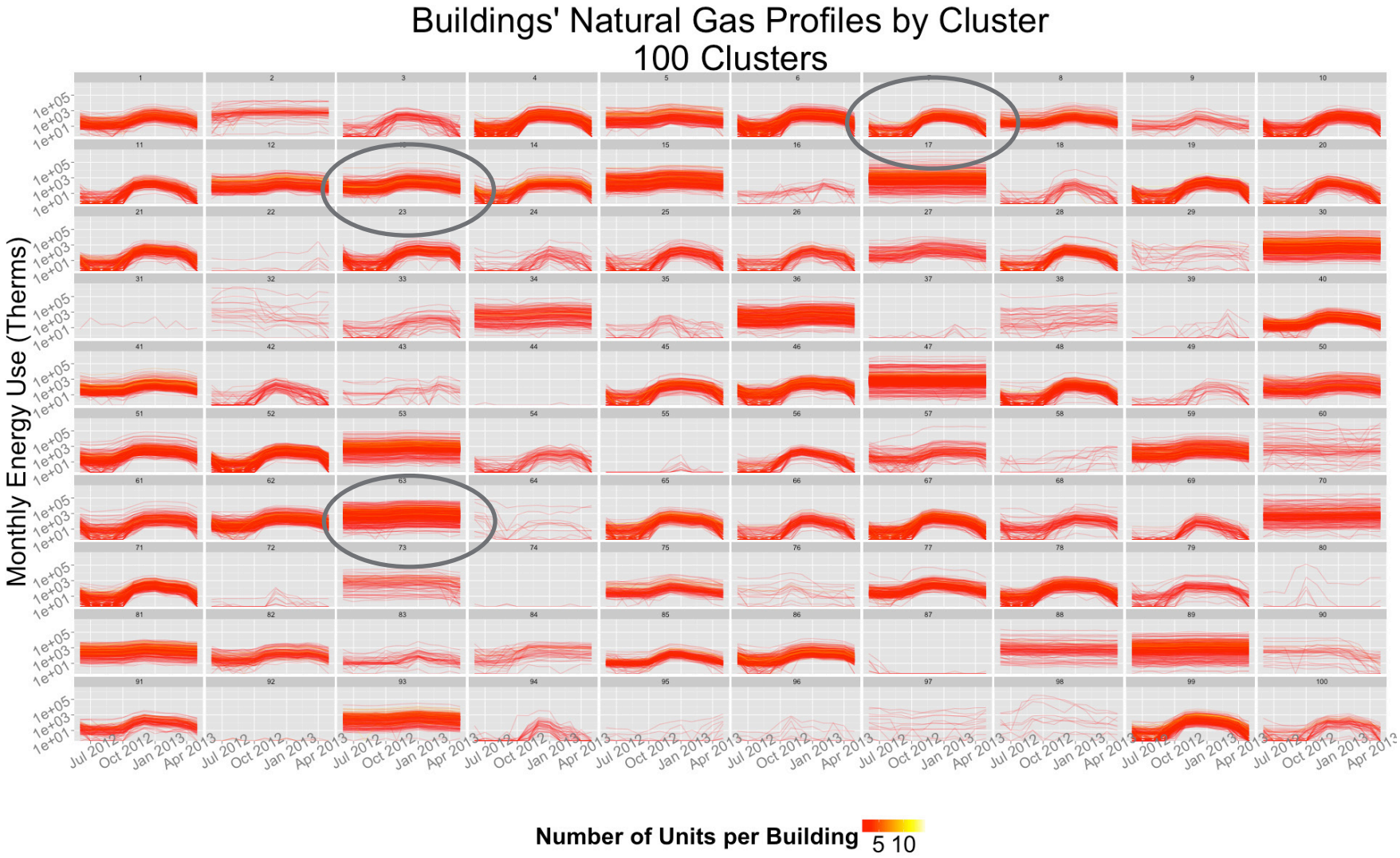

Figure 3.7. Original Profiles Plotted by Cluster

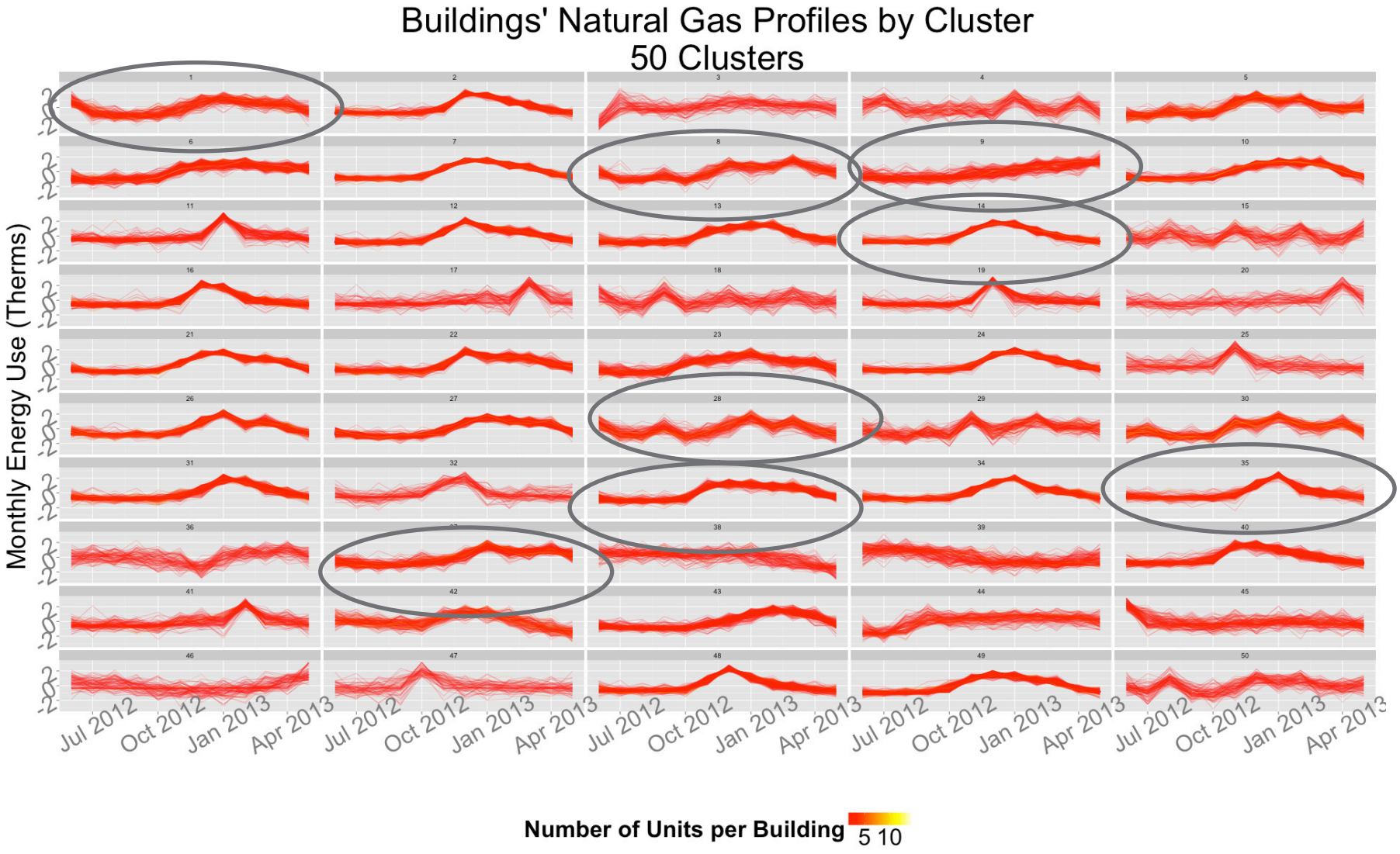

Figure 3.8. Normalized Profiles Plotted by Cluster

Cluster composition for original data (Figure 3.7) shows that while there are only a few dominant building profile shapes (flat and bell-shaped profiles captured by the clusters with the highest number of members), mid-size and large clusters are formed based on minor variations of these dominant profile types. Black ovals on Figure 3.7 denote dominant profile types represented by the largest clusters (3 largest clusters from Figure 3.5).

Cluster composition for the normalized data (Figure 3.8) emphasizes the most dominant profile shapes captured by 8 clusters with the highest number of members depicted in Figure 3.6.

One key insight revealed by the cluster composition analysis is that buildings with the same number of meters are not clustered together. For example, the two-meter buildings are spread throughout all the clusters. An argument can be made that it is the climate/weather and building type that drives the clustering. While climate defines whether the profile is predominantly flat or bell-curved, it does not seem to be the case that building type strongly correlates with the number of meters in a building or cluster membership.

To summarize, when clustering is done on the original data, where both shape and magnitude affect how the building profiles are bundled together, the natural number of clusters is k = 100. When the impact of the magnitudes on clustering is suppressed via normalization, the natural number of clusters drops to 50. This could serve as an indication that, along with variability in shape, there is a desirable degree of variability in the magnitude of the building profiles.

The degree of variability at the building level, not just at the meter level, is important for the following reason. There is a handful of dominant building profile types in the dataset. If the buildings of the same type/size with the same number of meters/tenants are highly homogeneous, successful estimation of tenant profiles in one building from the group can be used to compromise tenant data in the rest of the buildings with the similar profile type and building characteristics. The higher is the energy profile variability at the building level, the less is the likelihood of this type of attacks (homogeneity attacks) being successful. However, extremely high variability is not beneficial either, as it translates to higher distinguishability of buildings within the dataset.

Since there is an indication of desirable degree of variability in building profiles and cluster membership, in both shape and magnitude (i.e., buildings with the same number of tenants/meters are not bundled together), estimation of individual tenant profile would be more difficult. Therefore aggregation may provide an adequate degree of protection for this data: if reidentification of meters occurs for one building, variability in shapes and magnitudes of building profiles will make it harder to generalize that information to the other buildings.

While this aspect might not be of immediate relevance in the discussion of estimating tenant or meter-level profile within a specific building, it becomes more important within the context of turnover discussed in the main report.

3.2 Meter Profile Variability

The main metric of interest is the percentage of meter profiles that are similar to their overall building profile. This metric indicates how easily any individual meter profile can be guessed or estimated from dividing the building profile by the number of meters in the building. The latter, namely building profile divided by the number of meters in the building, is denoted as the average building meter profile (ABMP). This is the primary mode of attack considered in the analysis: guessing a tenant/meter monthly energy profile from dividing the building profile by the number of meters in the building (i.e. based on the similarity of individual meter profile to ABMP).

Cluster analysis is one of the methods used here to determine whether the individual meters within a building resemble the overall building profile and, as a result, the ABMP. If the resemblance is high, a large portion of individual meter profiles clusters together with the corresponding ABMPs.

By design, cluster analysis captures the variability in shape and magnitude between individual meter profiles and ABMPs. A more detailed examination of correlation statistics and analysis of the ratio of the ABMP to the individual meter profiles elucidate further the degree of similarity between overall building and meter profiles.

3.2.1 Clustering of Meter Profiles with ABMP

Figure 3.9 shows the elbow plot for clustering together the actual meter profiles and the ABMPs. This plot suggests k = 20 is the optimal number of clusters for original profiles before normalization. This means that once the ABMPs (17,318) and actual meter profiles (57,242) are pooled together, the overall set of 74,560 monthly profiles has 20 natural groupings. Breaking data into distinct groups any further does not result in a reduction of the total *within sum of squares*. The natural number of clusters in the normalized data is 75 as shown in Figure 3.10.

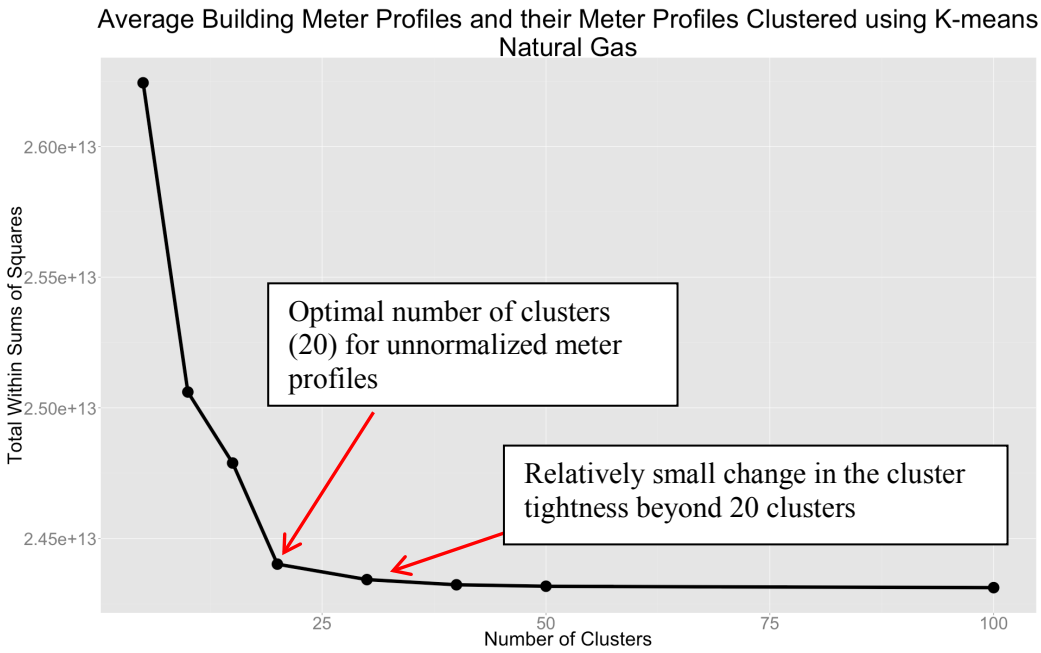

Figure 3.9. Cluster Analysis Based on Meter and ABMP, Original Meter Profiles (k = 20)

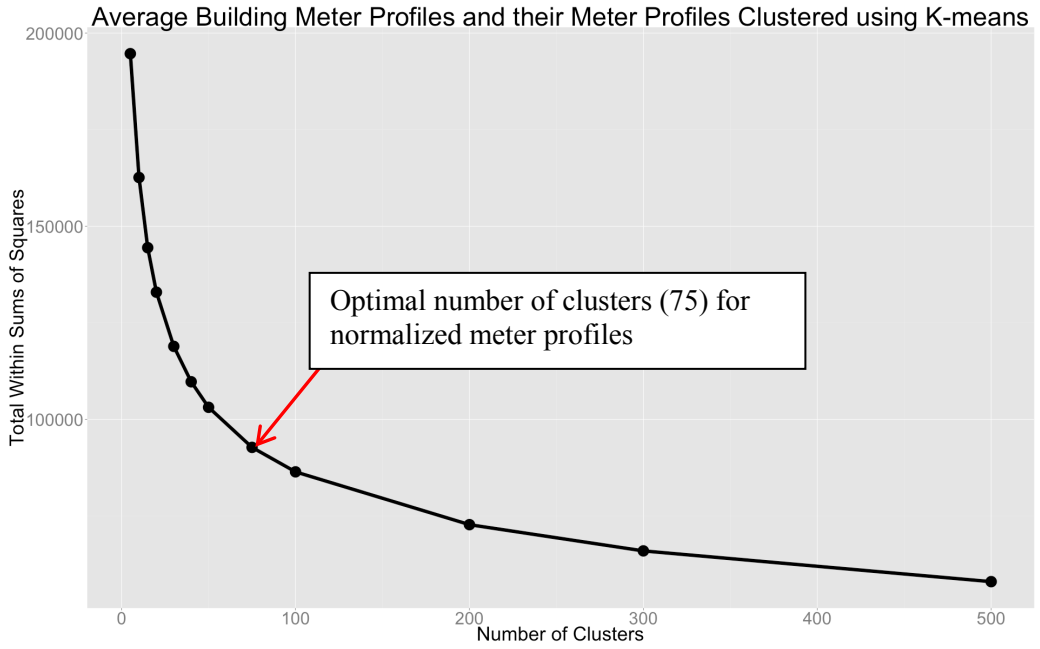

Figure 3.10. Cluster Analysis Based on Meter and ABMP, Normalized Meter Profiles (k=75)

It may appear that having 20 optimal groupings for 74,560 profiles immediately suggests that a large number of meter profiles are similar to the ABMP. But the comparison of cluster composition for unnormalized and normalized pooled data indicates that the variability in profile magnitudes between the 20 optimal groups is high. When the impact of magnitude on clustering is suppressed via normalization, the clustering algorithm is more sensitive to changes in the shape, thus picking up 75 types of profile shape in the data.

Cluster count by size for original and normalized data is shown in Figure 3.11 and Figure 3.13. Cluster composition in original and normalized data is presented in Figure 3.12 and Figure 3.14. The clusters that contain the largest number of profiles, i.e. dominant profile shapes, are marked.

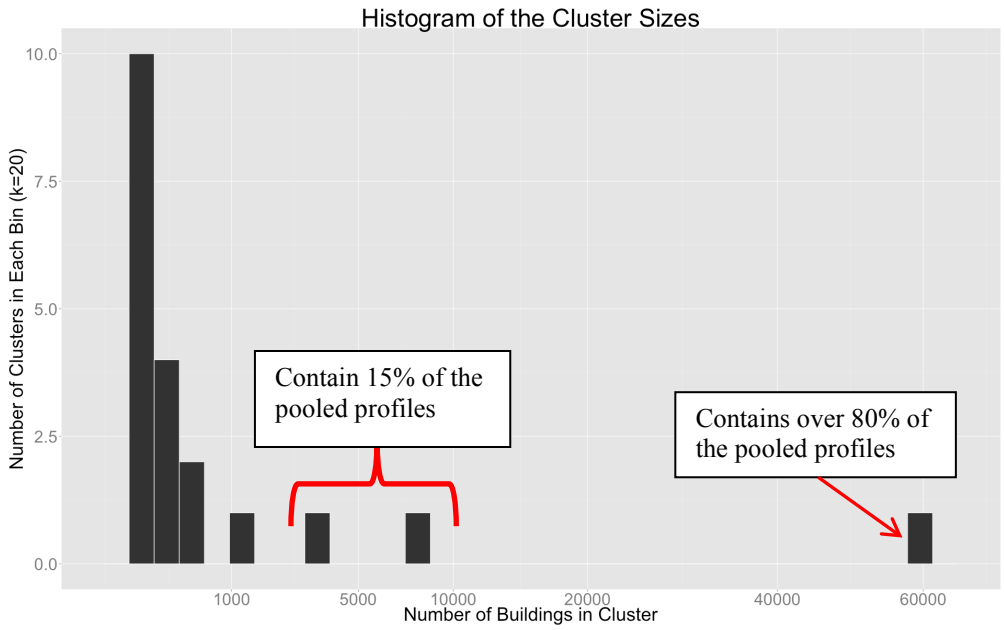

Figure 3.11. Histogram for the Optimal Number of Clusters, Unnormalized Data

Note that in unnormalized data (Figure 3.11) Cluster 2 (which corresponds to the last bar on the right of Figure 3.11) contains 60K out of 74.5K profiles, or over 80% percent of the pooled dataset. Two clusters in the middle of Figure 3.11 contain 3222 and 8484 profiles each, which accounts for over 11K profiles, or another 15% of the pooled profiles. These three clusters together (2, 16 and 18) contain over 95% of the unnormalized profiles.

Cluster composition for unnormalized data with k = 20 (plotted on a log scale) shown in Figure 3.12 confirms that magnitude indeed dominates the clustering, which is observed from the scale difference for each one of the cluster plots. There are several clusters that contain a small number of distinct profiles (for example, Clusters 5, 6, 7, 10 and 17). Several different clusters seem to contain profiles of identical shape, but their magnitudes are significantly different (for example, Clusters 12 and 15).

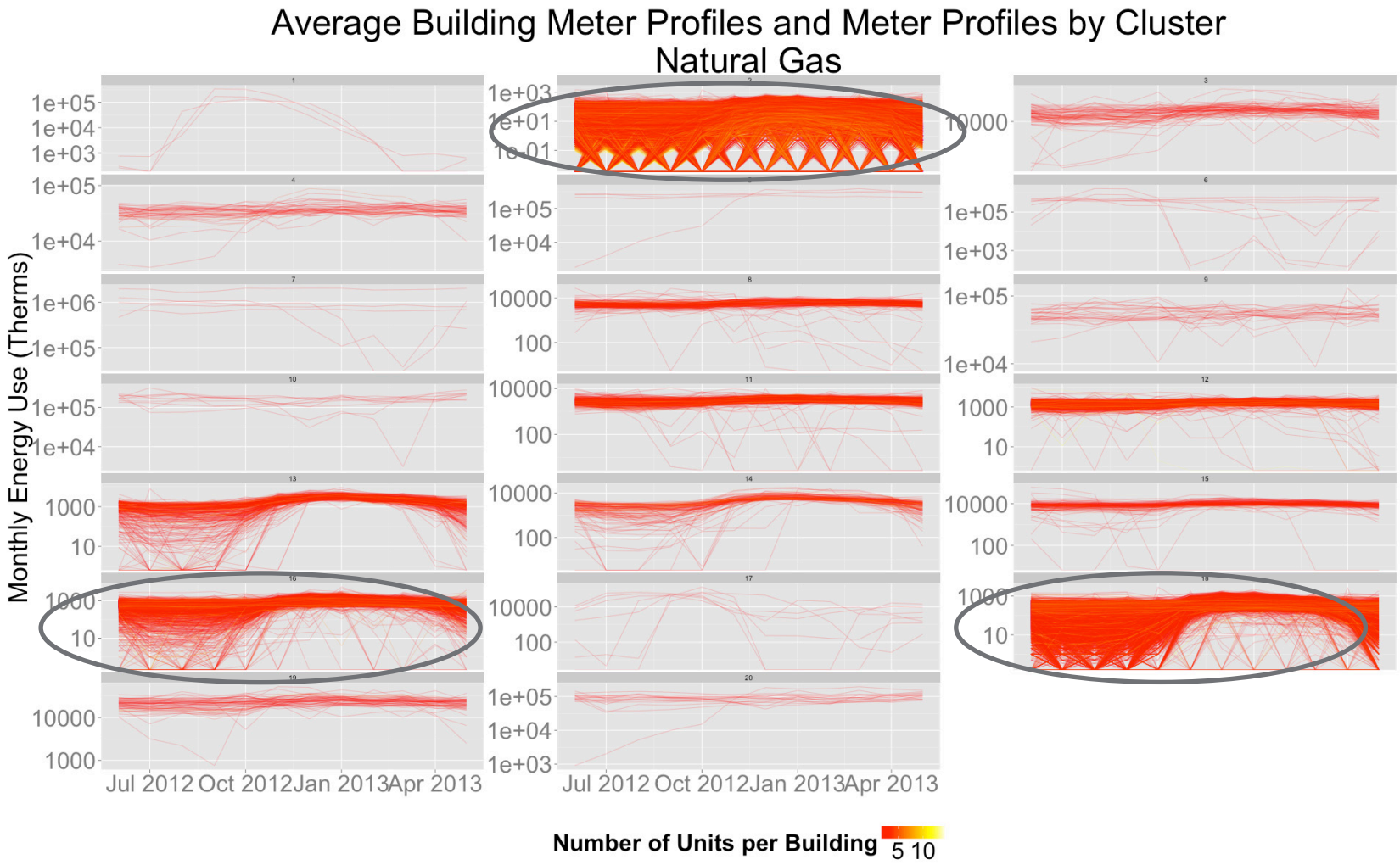

Figure 3.12. ABMP and their Meter Profiles Plotted by Cluster (k=20), Unnormalized Profiles

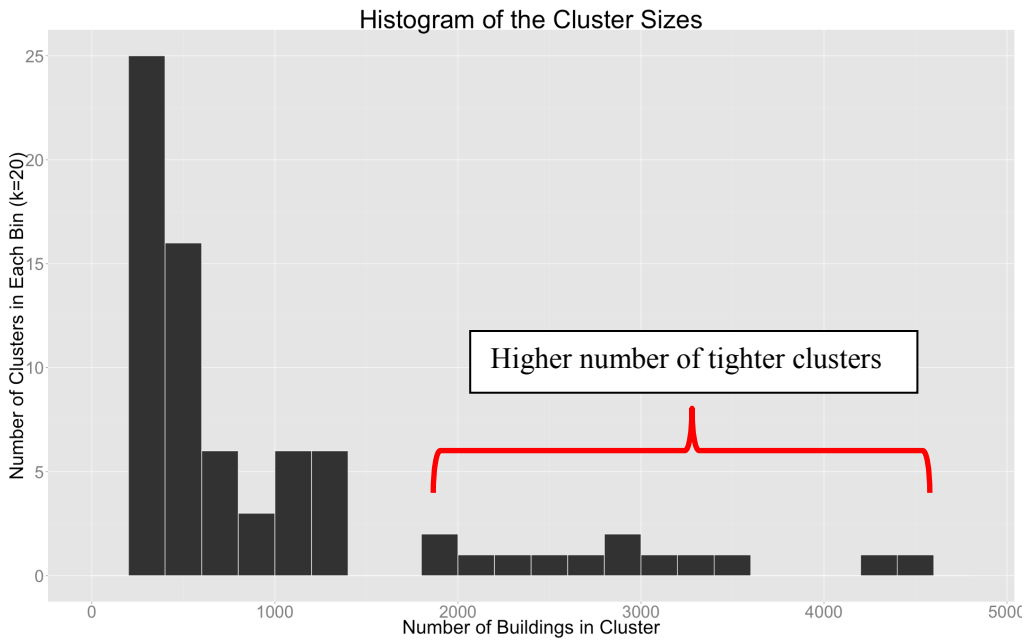

Figure 3.13. Histogram for the Optimal Number of Clusters, Normalized Data

The membership count for optimal clusters in the normalized profiles is more even and clusters are tighter. The largest clusters contain 4303 and 4532 profiles (the two last bars on Figure 3.13, which correspond to Clusters 51 and 18 in Figure 3.14). The smallest cluster contains 220 profiles (first bar in Figure 3.13, which corresponds to Cluster 12 in Figure 3.14).

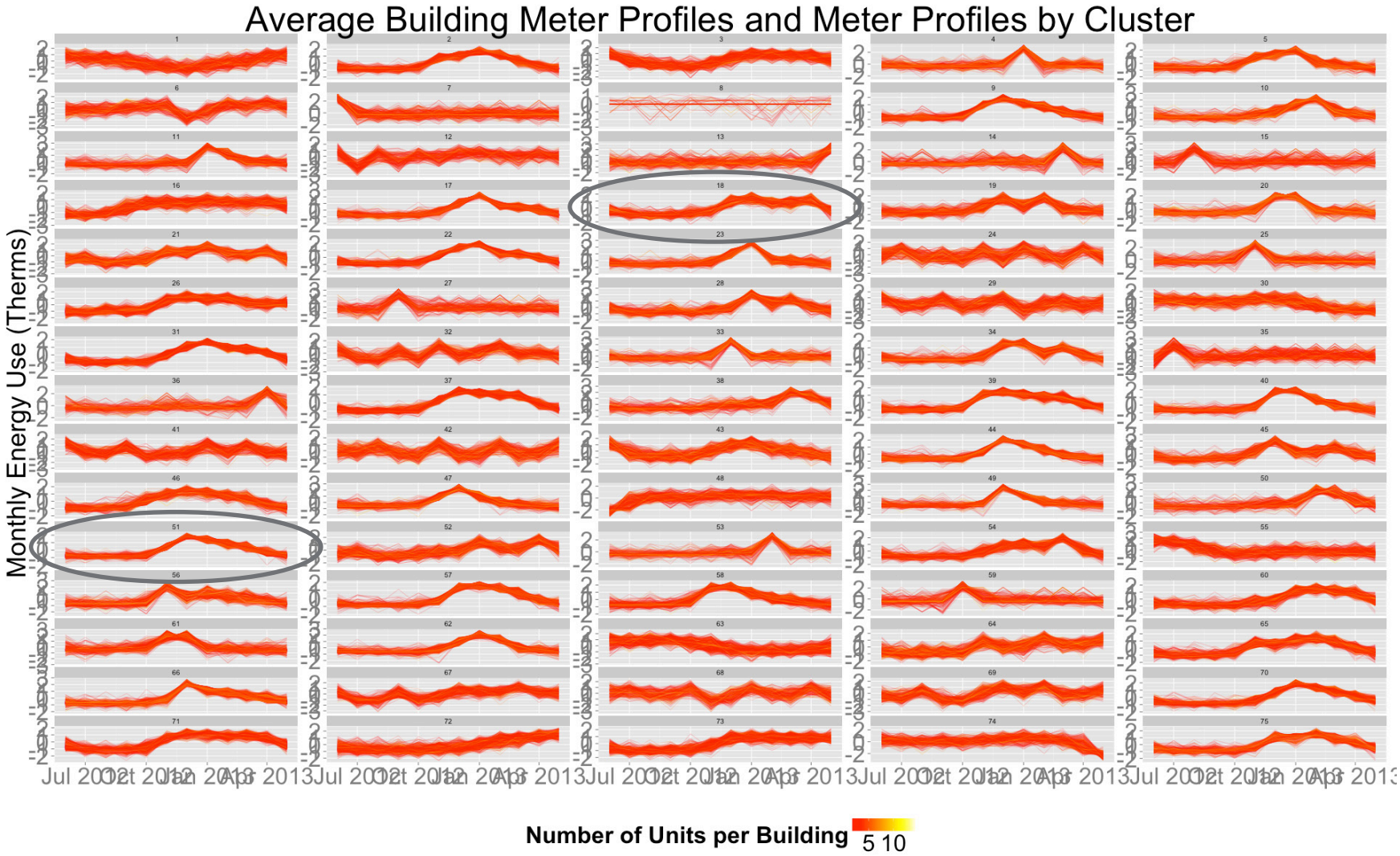

Figure 3.14. ABMP and their Meter Profiles Plotted by Cluster (k=75), Normalized Profiles

Cluster composition for normalized data shows the shape of the 75 dominant profile types in the data (Figure 3.14). The y-axis in Figure 3.14 indicates the number of standard deviations from the mean of the original profile, where original profiles indicate monthly natural gas consumption in therms.

Figure 3.14 also indicates that meters from buildings with the same number of meters, on both the low and high ends of the count, are distributed across several clusters. For example, profiles from buildings with three meters are included in almost every cluster. This outcome confirms that individual meter profiles and ABMP within any specific building count category (eg, all 5 meter buildings) are not homogenous to the point that they would predominantly bundle together. This lowers the risk of homogeneity attack within a building-count category, as well as within any given building.

This is consistent with the results depicted in Figure 3.7, where building profiles for the buildings with the same number of meters do not cluster together either. Note this does not hold for unnormalized data because the current partition bundles over 80% of the profiles in the same cluster due to magnitude dominating the clustering.

Figure 3.15 demonstrates that each cluster in unnormalized raw data is formed using primarily, and possibly only, the magnitude of the annual NG consumption. Number of meters per building is shown on the right of Figure 3.15. Cluster number and order are unimportant here, whereas the lack of horizontal overlap of these distributions suggests cluster composition is based solely on magnitude of consumption. Compare this result to that of Figure 3.14.

Based on this discrepancy in the cluster composition and concentration of unnormalized profiles in one cluster, it would be inappropriate to use unnormalized raw profiles for quantifying the similarity of the individual meter profiles with their respective ABMP. In order to understand how different each meter profile is from its building profile, the clustering to group profiles in this data should be based on shape and not magnitude. Therefore, the clustering analysis should be performed on the normalized profiles instead.

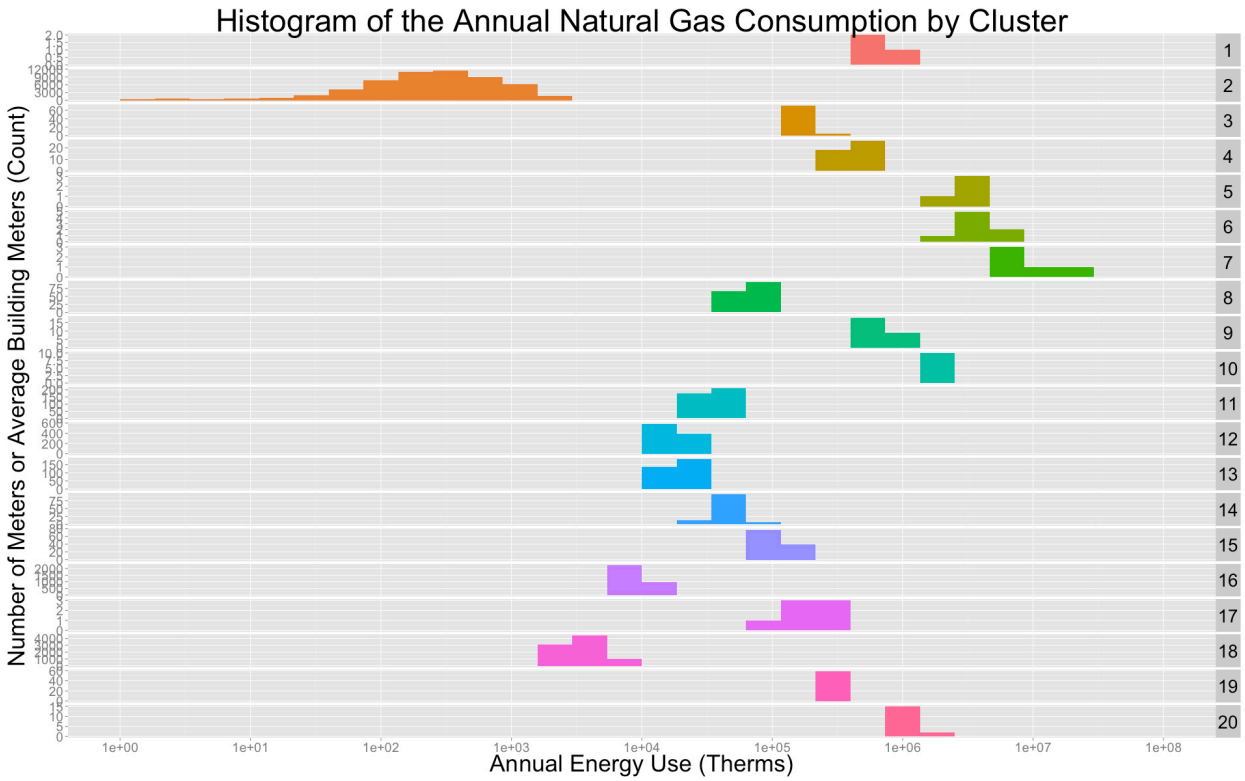

Figure 3.15. Comparison of Annual Meter Energy Use across Clusters in Unnormalized Data

Boxplots in Figure 3.16 show one of the most relevant results of the cluster analysis: the percentage of the normalized meter profiles that fall in the same cluster as the corresponding ABMP. The percentage of meters that cluster together with the ABMP is used to measure the similarity of profiles to the respective building average. It is utilized as a proxy to indicate how likely individual meter profiles can be estimated from the building profile simply by dividing the building monthly totals by the number of meters. There is no formal definition or a metric for the probability of reidentification that could be immediately applied to the context of the analysis in this report. Therefore we are attempting to develop a specific practical definition and attach a metric that could provide a quantification meaningful for utilities and other stakeholders. The percentage of individual meter profiles that cluster with their respective their ABMP is such a metric.

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 falls outside of the whisker range is plotted as points. Buildings are grouped based on the number of meters which is shown on the bottom of the chart.

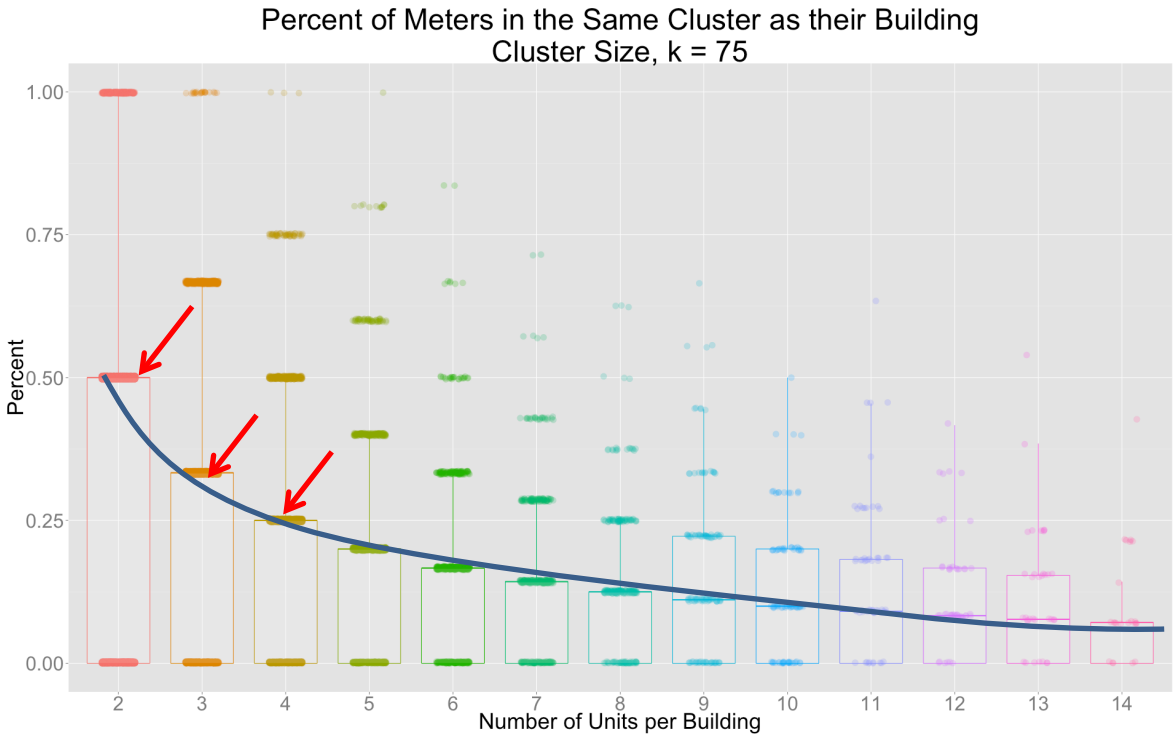

Figure 3.16. Percent of Normalized Meter Profiles Clustered Together with their ABMP for k=75

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

There is an alternative way to look at the clustering results. In 27% of 2-unit buildings none of the profiles cluster with the ABMP. In 66% of the 2-unit buildings only one out of two meters (one but not the other) cluster with their ABMP, and approximately in 6% of 2-unit buildings both of the meters cluster with the AMBP. The average percentage of meters that cluster with their ABMP for buildings with 2 units can be calculated as 27% * (0/2) + 66% * (1/2) + 6% * (2/2) = 0 + 33% + 6% = 39%.

For a three-meter building (second boxplot in Figure 3.16), the median is at about 0.3, which means that in 50% of the buildings roughly one out of three meters resembles the ABMP. In other words, the shape of the monthly consumption profile for one meter can be guessed from the overall building profile via division by the number of meters. If more detailed information about the physical building and occupied square footage is not available, not only does this not tell you which one of the meters can be guessed in this manner, but it does not reveal any specific information about the remaining two meters.

For example, from the mere dimensionality of the problem (three meters), if the shape of the profile for one meter can be approximated as 1/3 of the overall building profile, then the remaining two meters are underidentified in statistical terms. Simply put, when only one out of three profiles is estimated, it does not reveal the shape or the magnitude of the other two profiles. The number of unknowns exceeds

the number of known parameters, i.e., there are more variables than equations. This idea is illustrated in Figure 3.17.

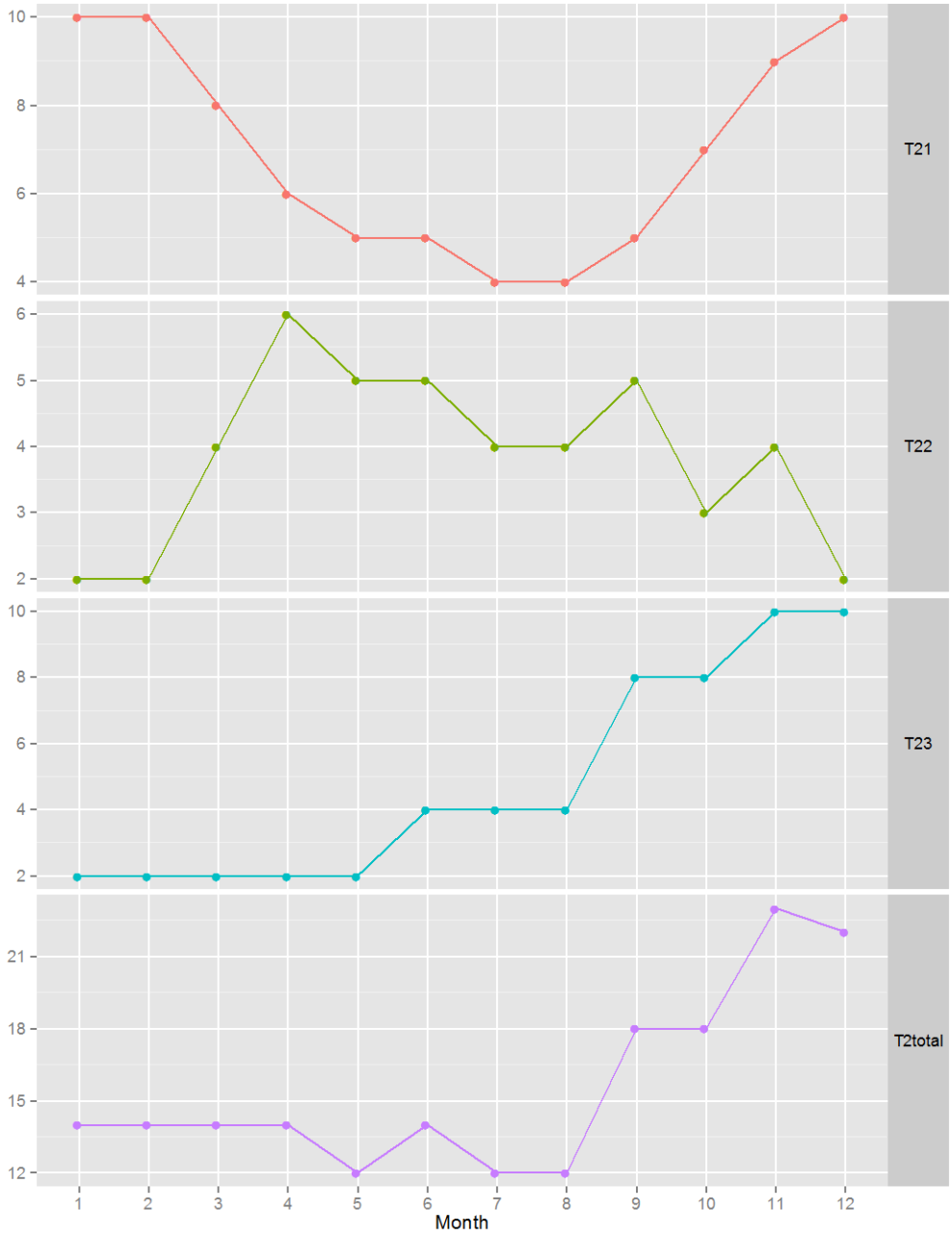

Figure 3.17. Example for a Hypothetical 3-Meter Building

In this example, the profiles T21, T22 and T23 are hypothetical meter profiles for a three-meter building (N=3). T2 total is the sum of T21, T22 and T23, i.e. the building total profile. Let's assume that profile T23 was reidentified, since it closely follows the shape of the total building profile. If we take 12 monthly consumption totals for the building (T2Total, the bottom section of the graph), subtract 1/N from each of the monthly values (of our supposedly identified meter profile T23), then we know that the remaining N − 1/N have to be allocated between two profiles (which do not resemble the building total

profile in this case, although that is not known to an attacker, who was able to identify only T23). Only the sum of monthly totals for the remaining two profiles (T21 and T22) is known to the attacker. Multiple linear combinations of profiles can produce the same total profile as the sum of T21 and T22. As the number of meters per building increases, accurate estimation of individual meters becomes progressively harder, because the degree of what is known in statistics as underidentification gets higher—an increasing number of linear combinations can add up to the same total profile.

Going back to Figure 3.16, the alternative metric, average percentage of meters that are similar to their ABMP for 3-meter buildings, is approximately 24%. In 37% of the 3-meter buildings none of the meters cluster with the ABMP, for 50% of the buildings only one out of 3 meter profiles clusters with the ABMP (but the other 2 meter profiles do not). In 10% of the buildings 2 meter profiles cluster with the ABMP, while in only 1% of the 3-meter buildings all meter profiles cluster with ABMP. Taking weighted average 37% * (0/3) + 50% * (1/3) + 10% * (2/3) + 1% * (3/3) produces the estimate of the average percentage of meters that are similar to their respective ABMPs (24.3%) across 3-unit buildings.

For four-meter buildings, the median of the boxplot is at 0.25, which implies that in 50% of cases, one out of the four individual meter profiles in four-meter buildings clusters together with its ABMP. The median for five-meter buildings is at 0.20, implying that in 50% of the five-meter buildings, one out of the five individual meter profiles clusters with the ABMP. Connecting the boxplot medians in Figure 3.16 forms a curve that shows this decrease for six-, seven-, eight- and nine-meter cases.

The alternative metric, average percentage of meters that resemble their ABMP, for four- and 5-meter buildings is calculated in the same fashion as explained in the example with the 3-meter buildings.

Table 3.2 summarizes the weigted average percentage of meters that cluster together with their ABMP.

Table 3.2. Average Percentage of Meters Similar to ABMP under Analyzed Aggregation Thresholds

Threshold (# of meters)	Percentage of meters clustering with ABMP
2	39%
3	24%
4	19%
5	17%
6	13%
7	13%
8	12%
9	12%
10	11%
11	12%
12	11%
13	11%
14	8%

The average percentage of meters similar to ABMP for 4-meter buildings is 19%. Starting at four meters, percentage of meters clustering with ABMP for each consecutive threshold drops first by about 3 percentage points for two levels, and then continues to drop by less than 1% for the next two levels. Thus the percentage of meters similar to ABMP under a four-meter aggregation rule encompasses the

percentage of meters similar to ABMP under any subsequent aggregation levels (i.e. the similarity of meters to the ABMP does not increase with higher number of meters in this dataset).

Note that because of how clustering works, reducing the number of clusters in the data makes it easier to achieve a higher percentage of meter profiles clustering together with their building profiles. As the target number of clusters gets smaller, the clusters themselves become wider, i.e., the degree of similarity required to fall into a particular cluster is more forgiving, each cluster increases the within-cluster variation, and as a result, more dissimilar profiles are clustered together. Reducing the number of clusters will overstate the percentage of meters clustering with their building, and subsequently overestimate the probability of being able to deduce an individual meter from the overall account or building profile.

Alternatively, increasing the number of clusters results in smaller/tighter clusters, restricting the degree of similarity required for profiles to bundle together. As a result, fewer meters will cluster with their building profile, thus underestimating the probability of guessing an individual meter profile from the building average.

Figure 3.18 illustrates this change in cluster composition and the percentage of meter profiles that can cluster with their respective ABMP. Change in percentage of meters that cluster together with ABMP is caused by the changes in the number of target clusters and, as a result, cluster tightness.

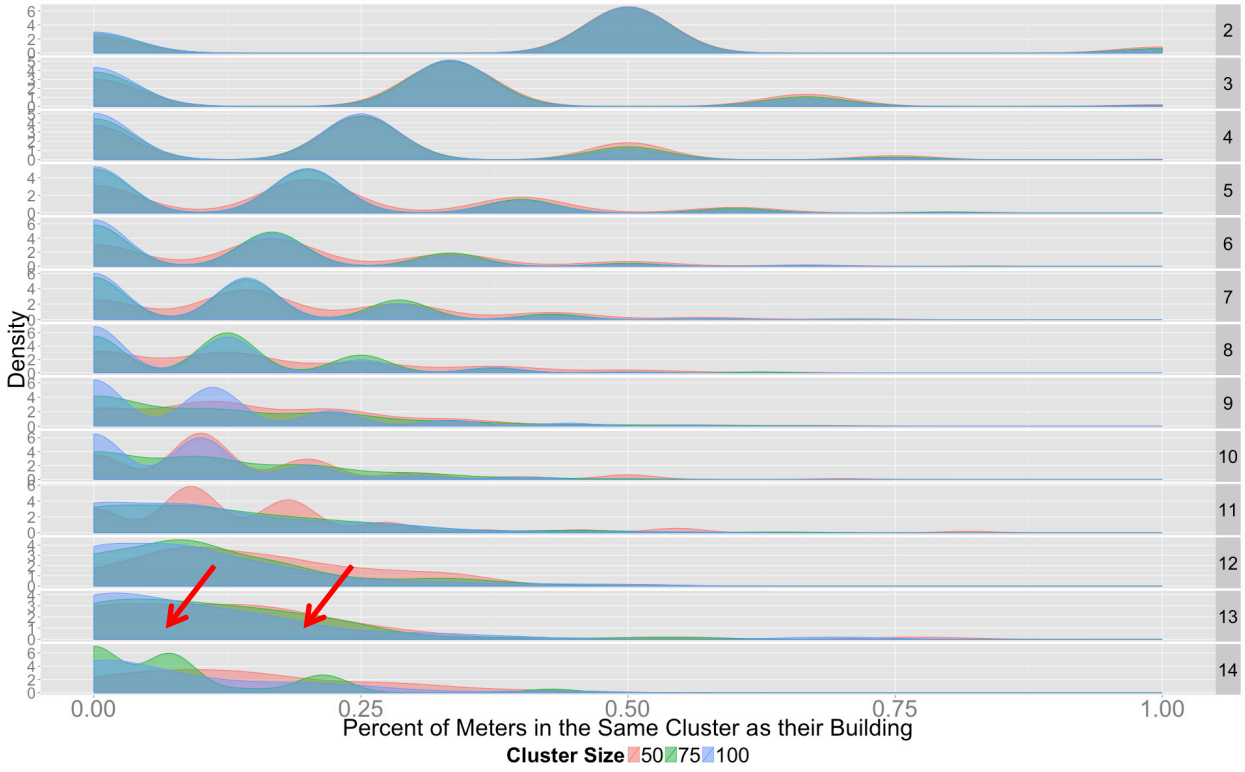

Figure 3.18. Percentage of Meters in the Same Cluster as their ABMP (k=50, 75, 100)

Figure 3.18 shows the distribution around the percentage of the profiles that cluster with their ABMP for three different cases. The blue distribution shows how meter profiles and ABMP profiles are broken down into 100 clusters, green shows the same for 75 clusters, and orange for 50 clusters. For buildings

with the number of meters not exceeding eight, the cluster composition does not shift much from case to case (k = 50, k = 75 or k = 100). For buildings with nine and more meters the difference is more easily observable. For example, consider the buildings with 12 meters. As the number of clusters increases from 50 to 100, the probability density function shifts to the left (from the orange probability density plot to the blue probability density plot shown by the arrows), and, as a result, the median of the displayed probability distribution function shifts to the left, indicating that the percentage of meters that clusters with ABMP for 50% of the buildings decreases. The average percentage of meters similar to the ABMP decreases as well.

Therefore this analysis of percentage of meters clustering with their ABMP as the proxy for the probability of identification is valid only in the context of the optimal number of clusters, as informed by the elbow plot in Figure 3.10.

The percentage of meters clustering with their ABMP tells us what fraction of the meters are similar to the ABMP without any specific characterization of what that similarity looks like. To understand the degree and direction of the relationship, analysis of the correlation between ABMP and individual profiles is included below. To understand the relationship between the magnitudes of individual meter profiles and ABMP, the ratios of annual meter total consumption to total ABMP consumption are described in the section that follows after.

3.2.2 Correlations between Individual Meter Profiles and their ABMP

Correlation between individual meter profiles and their ABMP illustrates the degree and direction of linear relationship between an individual meter profile and the corresponding building profile. It provides additional information on the similarity between profiles, and, as a result, also informs how easily the individual meter can be backed out from the overall building monthly total.

Figure 3.19 is a box and whisker plot of the correlation between the building profile and its meter profiles by each building size (number of meters per building).

As expected, individual meter profiles are highly correlated with their ABMPs with the median being in the 0.9-0.95 interval. Overall, there does not appear to be an effect due to the number of meters per building. Meter profiles in 4-meter buildings are not significantly more correlated with their ABMPs than meter profiles in the buildings with 8 meters are correlated with their respective ABMPs. Rather the figure shows consistency, especially in the median correlation (the middle line of the box).

The IQRs and the medians for three- and four-meter buildings are nearly identical. The medians (and IQRs) for the rest of the boxes are also very similar. The significance of this figure is that meter profiles in four-meter buildings are no less correlated with their ABMP than buildings in three-, six-, seven-, or eight-meter buildings. In fact for buildings with five to nine meters, the distributions of correlation coefficients are indistinguishable from each other. The distributions of correlation coefficients for buildings with four to nine meters are very similar.

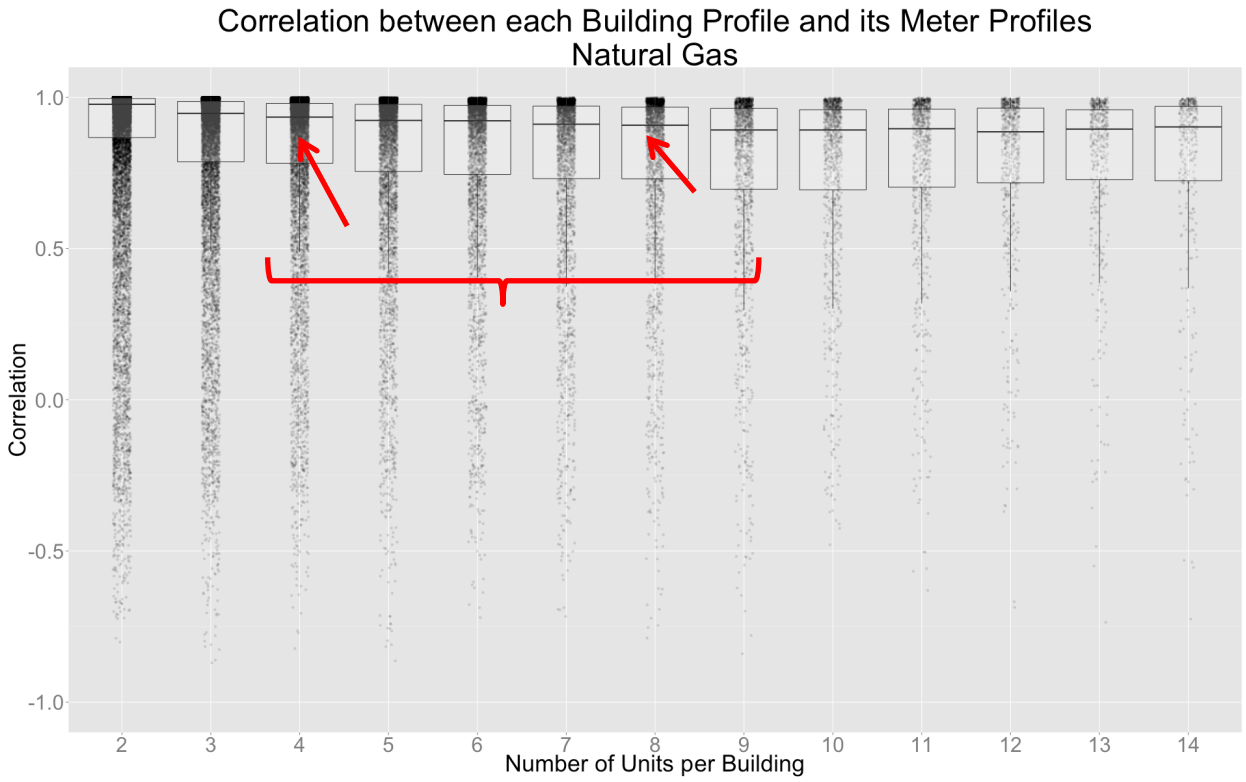

Figure 3.19. Boxplot of Correlation between Building and ABMP

3.2.3 Ratio of Individual Meter Annual Consumption and ABMP

Figure 3.20 and Figure 3.21 show the distributions (in density and boxplot form, respectively) of the ratio of individual meter annual consumption to average building meter annual consumption. This ratio describes the degree of variability in the magnitudes of the individual meter profiles as compared to the ABMP. Buildings are grouped based on the number of meters, which are shown on the right side of the chart. The dotted vertical line is located at a ratio value of one. The values to the left of that line represent meter annual totals that are less than ABMP annual total. The values to far right show the presence of meters with the annual consumption much higher than that of the ABMP. This is indicative of the buildings where a few meters dominate the total profile of the building.

The points outside of the boxplot whisker are traditionally interpreted as outliers. In the context of this analysis, the buildings with meter ratios so far to the right that they fall off the right whisker are the primary candidates for being removed from the reporting irrespective of the aggregation threshold. This is the group of meters/meters where the magnitude of one individual meter profile is high to the point of being uniquely distinguishable.

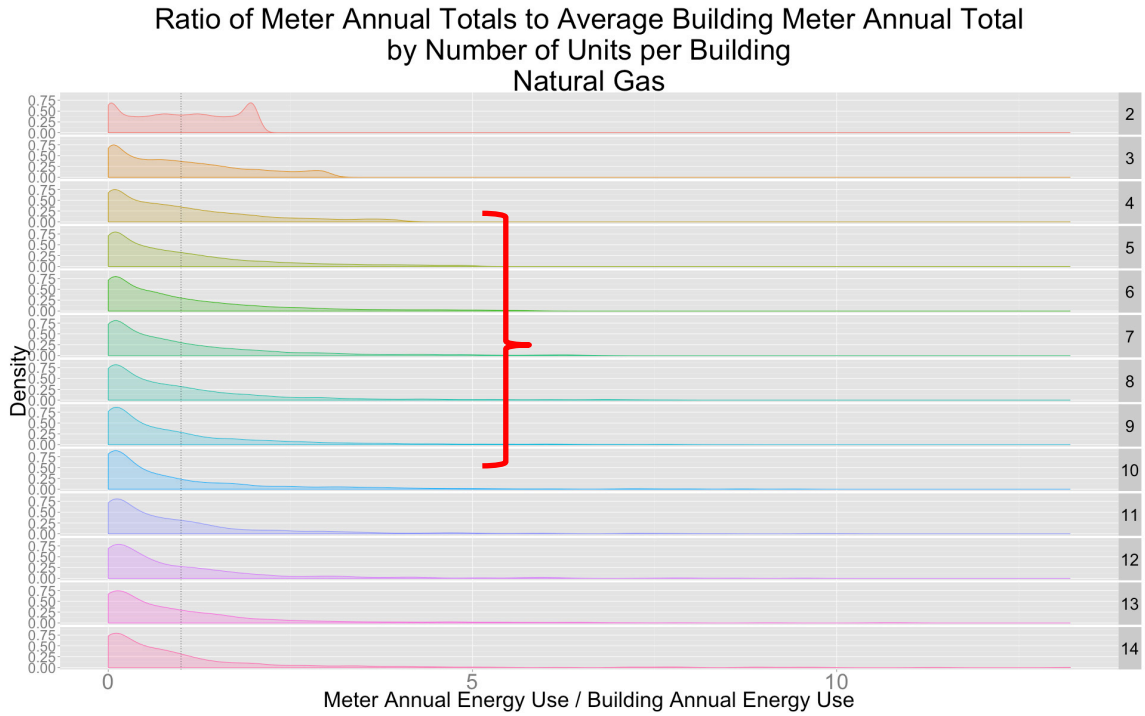

Figure 3.20. Distribution of the Ratio of Meter Annual Consumption to the ABMP Annual Total

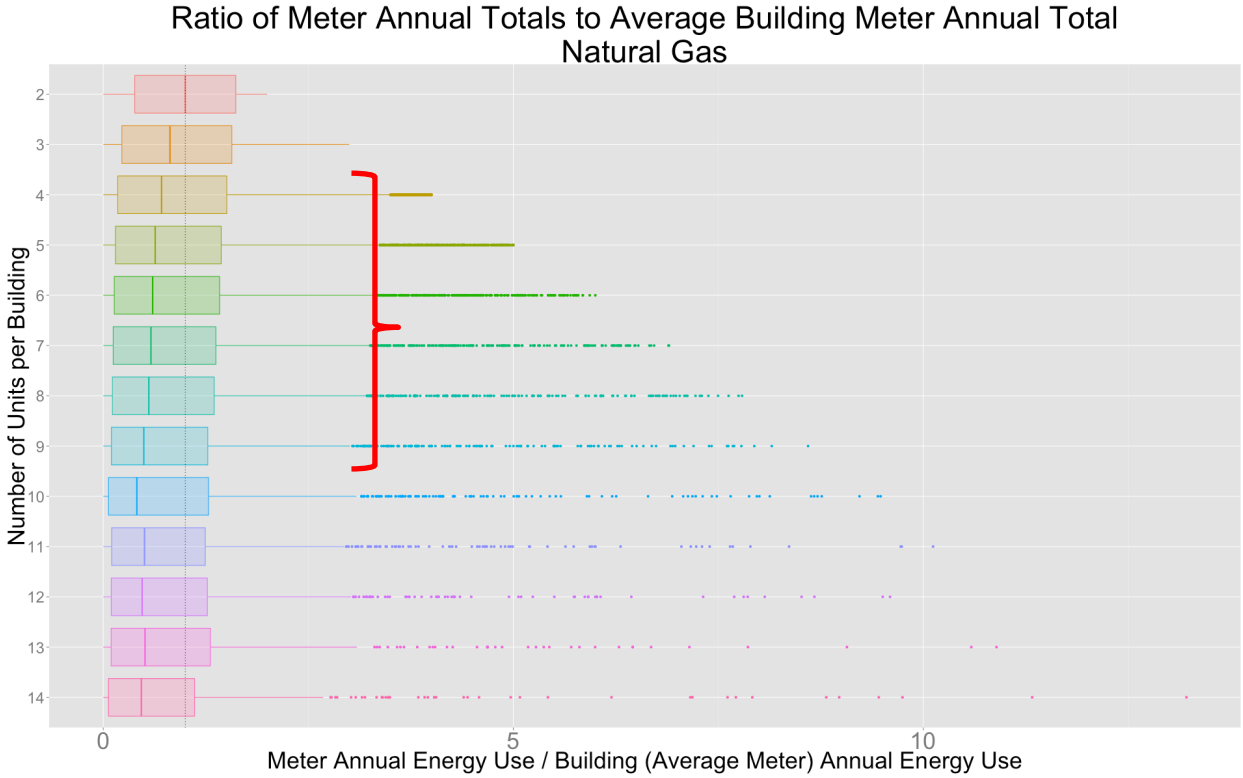

Figure 3.21. Boxplot of the Ratio of Meter Annual Consumption to the ABMP Annual Total

Mass concentrated at 1 (the vertical line) would indicate that annual consumption of meters is very close to the annual total of the ABMP. If the normalized profiles cluster closely together (i.e., are similar in shape to the ABMP) and distribution of the annual totals is narrowly amassed at 1 (i.e. similar in magnitude to ABMP), then individual meter profiles are homogeneous within their respective building groups to the point that they can be backed out fairly easily from the building total profile simply by dividing the building total profile by the number of meters.

As can be observed from Figure 3.20, the distributions do have some differences across groups of buildings with different numbers of meters. But overall the differences are not large, and distributions of this ratio appear quite consistent, especially between 4-meter and 12-meter buildings.

There is a large concentration on the left side of each distribution which indicates that there are many meters that are small in the profile magnitude as compared to their respective ABMPs. A fairly wide distribution with a long tail indicates that there is a significant number of meter profiles with relatively high magnitudes. This means that within the same building size (number of meters) there is a large number of meters with consumption much lower than that of the ABMP, and there is more than one meter with consumption higher than the ABMP.

Having a reasonable range in the ratios along with a good spread of the mass on both sides of the value of 1 indicates that magnitudes of the meter profiles have a desirable balance between the variability of the profile magnitudes (range of the distributions) and dense representation along the range (shape of the distribution).

The high degree of similarity in distributions for four- to nine-meter buildings is most easily observable in Figure 3.21. It shows that distributions of meter magnitudes as compared to building total are similar across these categories. To summarize, based on clustering analysis, only 25% of the meters in four-meter buildings cluster with their ABMP (i.e., the shape for one out of four profiles can be roughly estimated by dividing the building profile by the number of meters), and distribution of the ratio of profile magnitude to the ABMP is wide, thick and has a median of 0.5. Four meters is a first threshold for aggregation that is not subject to immediate decomposition due to turnover. Increasing the threshold from four to five does not lead to a dramatic change in the percentage of meters that are similar to ABMP.

4.0 Conclusions for Dataset 1 – Natural Gas

Analysis of variability in building profiles, meter profile cluster analysis, analysis of meter profile correlations and magnitude ratios relative to ABMP jointly indicate that there is a desirable degree of variability in both shapes and magnitudes of individual meter profiles, even within the buildings with the same number of meters. In addition, cluster membership analysis shows that accurately guessing all individual meter profiles from the building total is unlikely.

Not only do buildings with the same number of meters not cluster together in this dataset, but also individual meters from buildings with the same number of meters do not either a) cluster narrowly together with meters from “similar-sized” buildings, or b) cluster with the ABMPs from the corresponding building. In other words, the variability in shapes and magnitudes of individual meter

profiles within same-size buildings is such that there is only a small portion of meters from, e.g., four-meter buildings, that looks like the ABMP for that building (19%).

The percentage of normalized individial meter profiles that clustered together with their respective ABMP was used as a proxy for how likely individual meter profile can be estimated from the building average. On average, only 19% of the meters in four-meter buildings cluster with their ABMP (i.e., shape, but not magnitude, can be identified by dividing the building profile by the number of meters). That percentage decreases even further for buildings with a higher number of meters.

Variability across buildings with three-, four, and five or more meters is high enough that even in the unlikely case that all meter-level monthly totals are guessed correctly (e.g., any one four-meter building), that guess or its mechanics will be of limited use when applied to another four-meter building.

The percentage of meters similar to ABMP across various aggregation thresholds in this dataset is summarized in Table 4.1.

Table 4.1. Percentage of Meters Similar to ABMP under Analyzed Aggregation Thresholds

Threshold (# of meters)	Pecentage of Meters Similar to ABMP	Percentage of Multi-Meter Buildings by Category	Multi-Meter Buildings Coverage
2	39%	48.8%	100.0%
3	25%	21.5%	51.2%
4	19%	11.6%	29.7%
5	17%	6.6%	18.2%
6	13%	4.0%	11.5%
7	13%	2.6%	7.5%
8	12%	1.9%	4.9%
9	12%	1.1%	3.0%
10	11%	0.7%	1.9%
11	12%	0.5%	1.2%
12	11%	0.3%	0.8%
13	11%	0.2%	0.4%
14	8%	0.2%	0.2%

The third column in Table 4.1 shows the percentage of multi-meter buildings that fall into each category in the analyzed dataset. For example, two-meter buildings comprise almost 49% of all multi-meter buildings in this dataset. The fourth column shows the percentage of multi-meter buildings that would be eligible for reporting if the aggregation threshold was established at the level shown in the first column. For example, if the threshold is set at 3, 51% of multi-meter buildings are eligible for reporting under this aggregation rule, since two-meter buildings are automatically excluded. If the aggregation rule were to be moved up one level (to four meters), all two-meter and three-meter buildings would be excluded from the reporting, resulting in 30% coverage.

If the aggregation threshold was increased even further, to five meters, the incremental change in the percentage of meters clustering with ABMP is small. But the coverage drops from 29% of the multi-meter buildings to 18%.

Clustering results at the building and meter level, analysis of correlation and magnitude ratios consistently showed similarity in the results for four-, five- and six-meter buildings. The incremental

change in the percentage of meters similar to ABMP is small (from 19% down to 17% - moving the threshold from four to five, and then from 17% to 13% moving from 5 to 6). The loss in coverage from increasing the aggregation threshold from 4 to 5 and then from 5 to 6 is about 8-10% of the multi-meter building.

This tradeoff between the percentage of meters similar to their respective building average compared to reporting eligibility guided the comparison of candidate aggregation thresholds.

Comparison of the candidate aggregation thresholds for this dataset shows that

a) Incremental change in the probability of profile matching is 2 percentage points for increasing the aggregation threshold from 4 to 5 meteres, less than 4 percentage points for going from 5 to 6, and then drops by less than one percentage point for all subsequent candidate thresholds

b) The drop in the number of buildings that would be eligible for reporting for each subsequent aggregation threshold after four meters is significantly higher than the change in percentage of meters similar to ABMP (over 10 percentage points in going from 4 to 5, 7 percentage points in going from 5 to 6, and another 4 percentage points in going from 6 to 7);

c) While clustering meters with their ABMP produces the main result of this analysis (percentage of meters that resemble their ABMP and, as a result, can be estimated from ABMP), clustering does not indicate the direction of degree of similarity. Auxiliary results (analysis of correlation and analysis of ratios between total annual meter energy and ABMP total) show a consistent trend. Correlation coefficients for building count categories show that while the correlation between individual meters and their ABMP is high, it is no different statistically for 4-meter buildings than it is for 5, 6, ,7 and 8 meter buildings (Figure 3.19). Similarly, ratios of total annual meter energy consumption to their ABMP is not drastically different for 4-, 5-, 6-, 7 and 8-meter buildings (Figure 3.20).

d) Jointly this suggests using the tradeoff between the probability of consumption profile matching and reporting eligibility as the primary criteria for the aggregation threshold selection.

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