Characterizing Energy Use in New York City Commercial and Multifamily Buildings

David Hsu, University of Pennsylvania

ABSTRACT

Many cities have recently developed and passed benchmarking laws (also known as energy disclosure, information, or transparency laws) to make energy information available for individual buildings. This building-level micro-data can be used to better design and develop municipal policies for energy efficiency.

Based on benchmarking data collected by the City of New York, this paper presents a comprehensive portrait of energy use in New York City commercial and multifamily buildings, including the distribution and concentration of energy uses and building types within the overall population. First, this paper describes energy and building characteristics for the overall population of over 10,000 buildings, comprising over 1.5 billion square feet. Second, using model-based clustering methods, this paper then identifies key clusters of energy use and building characteristics in the multifamily sector. The paper finds that these clustering methods describe sub-groups in the population in intuitive ways. Third, using multivariate regression, the identified clusters are then used to improve predictions of building energy use and the targeting of prospective efficiency efforts. These methods and results should be of interest to energy and policy researchers in New York and other cities.

Introduction

To date, five major cities – New York, San Francisco, Seattle, Austin, and Washington, D.C. – have begun to collect building-level energy data through benchmarking laws, and more are expected to follow (Burr et al., 2011). Lack of such information has been long identified as a major potential barrier to increased investment in energy efficiency (see wide literature ranging from Blumstein et al., 1980 to Allcott and Greenstone, 2012). The primary intention of these laws are to transform the market for energy efficiency by making information about existing building performance available to owners, prospective buyers and tenants.

A secondary benefit of these disclosure policies is that they enable the gathering of data about energy performance, specific to the municipal scale. This new micro-data can be used to better develop and target energy efficiency policies at the local level, where many new policies are being implemented. When matched to existing sources of information about building characteristics (such as from property tax assessor databases), and cleaned, this disclosure data can become a unique new source of data about energy use in a large number of buildings. For example, of the five cities, the City of New York has the largest portfolio of buildings, with over 10,000 buildings and over 1.5 billion square feet benchmarked, and very high rates of compliance (over 60%). This is a very large, comprehensive dataset of building energy performance for a single city's building population. In contrast, the 2003 CBECS, contains 4,859 buildings for all fifty states, and 761 buildings in the entire Northeast region. The richness of this data allows

many new possible opportunities in understanding how energy use varies between fuel types, buildings, and other underlying variables.

This paper seeks to demonstrate the utility of benchmarking data by carrying out three main tasks. First, because benchmarking data is relatively new, this paper describes energy and building characteristics for the overall population of multifamily and commercial buildings in New York City, including summaries of building energy use and greenhouse gas emissions by building type. Second, model-based clustering is introduced to the building energy modeling literature in order to classify buildings in a systematic way, based on multiple, dissimilar characteristics. Third, these classification methods are then combined with multivariate regression models in order to improve the modeling of various fuel uses for individual buildings.

Data Cleaning and Summary

The City of New York gathered this dataset under NYC Local Law 84 (henceforth referred to as the City, NYC data, and LL84, respectively). LL84 requires owners of all commercial and multifamily buildings over 50,000 square feet to submit information to the U.S. Environmental Protection Agency's EnergyStar Portfolio Manager tool in order to obtain a benchmarking rating. This includes over 100 fields about the overall building size, the distribution of space uses within the building, fuel use, and some factors influencing energy use, as well as other derived quantities, such as source (primary) energy use and greenhouse gas emissions, which are calculated using regional or national conversion factors. The City then joined the NYC data to the City's Primary Land Use Tax Lot Output (PLUTO) database, which contains over 100 fields describing building and tax lot characteristics. After joining these two datasets, the City then removed any uniquely identifying information about specific properties, such as building identifying numbers, owner names, or addresses.

The initial dataset comprised 10,016 buildings and 1.59 billion square feet of commercial space. Since the data is self-reported and this is the first year of the benchmarking program, extensive steps were taken to clean the data. 626 buildings were removed since they were outside city boundaries, duplicate entries, or minor building types with less than 10 buildings total in the dataset. 934 buildings that had no energy use and/or floor space data were removed, since that was the principal area of interest. 240 buildings with energy use intensity (EUI) values below 10 or above 1000 kBTU per square foot were also removed from the total dataset. Finally, for each facility type, the top and bottom 5% of remaining EUI values were removed in order to eliminate extreme values. After all of these steps, this left 7,357 buildings, or 73.5% of the original dataset.

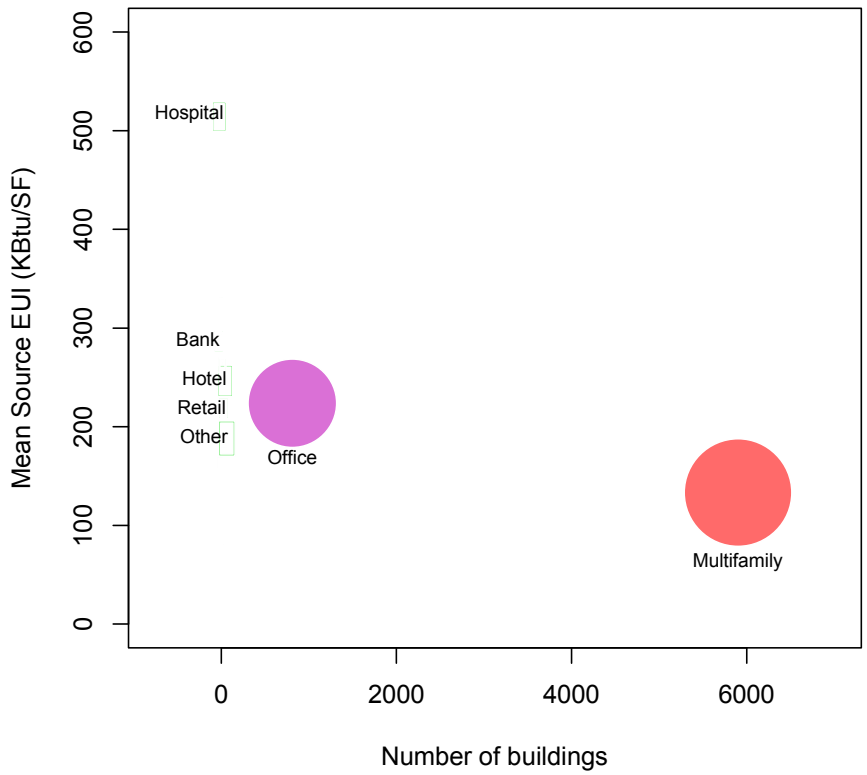
Once checked and cleaned, Table 1 shows the size of the NYC data set relative to other data sources (such as CBECS and RECS), and key features of energy use within the New York City building population. Figure 1 shows how the relationship between EUIs, square footage and total energy use varies among facility types. Table 2 summarizes aggregate metrics for the main facility types. Depending on goals and available policy instruments, city governments may focus on different metrics to set their policy goals. Table 2 also shows that the multifamily sector in New York City is a significant portion of the number of buildings and square footage, but is a proportionally smaller share of the total energy use and greenhouse gas emissions. If the City were to focus on the interventions in the least number of buildings, it would perhaps focus on the other and office types. However, in order to achieve New York City's ambitious goals to reduce greenhouse gas emissions, the table also shows that the City will probably need to devote attention to all of the building types.

Table 1. Number of Buildings and EUI Quartiles by Facility Type After Cleaning

	No. of Buildings		NYC	EUI Quartiles				
Facility Type	**EIA**	**NYC**	**MSF**	**0%**	**25%**	**50%**	**75%**	**100%**
Multifamily	319	5900	781.5	51.7	109.2	132.1	157.1	225.4
Office	155	811	283.8	95.1	169.6	212.8	268.5	424.9
Other	17	132	34.2	33.6	77.2	145.9	282.4	681.3
Hotel	27	108	27.2	129.1	200.9	246.7	289.9	397.7
Warehouse	65	80	9.7	23.6	45.0	71.7	110.5	227.8
Retail	76	67	13.1	73.8	162.3	196.9	269.4	499.5
K-12 School	0	52	4.2	71.7	142.8	192.9	222.4	306.9
Residence Halls	0	43	5.2	101.5	170.7	234.1	337.1	355.3
Senior Care	0	43	5.7	119	217.7	267.5	313.6	450
Hospital	18	37	11.4	316.7	455.6	483.5	572.2	711.6
TOTAL	677	7357	1196.1	23.6	112.4	139.2	173.3	879.8

Only categories with more than 20 buildings are shown. EIA figures from 2009 RECS for all multifamily (apartments with 2+ units) in New England, and otherwise from the 2003 CBECS for the Northeast Census Region, for all square footages; zeroes entered where facility types are ambiguously matched to NYC data.

Figure 1. Chart of Energy Use by Building Sector



Area of the circles indicates the total amount of energy consumed by sector, plotted against the number of buildings (x-axis) and the mean EUI in each facility type (y-axis).

Table 2. Total Summaries by Facility Type

	Multifamily	Office	Other	Total
TOTALS:				
Number of Buildings	5900	811	646	7357
Floor Area (MSF)	781.5	283.8	130.7	1196.1
Total Energy (Ktherms)	107.5	72.4	34.9	214.7
GHG Emissions (Mmtcde)	5.4	2.5	1.3	9.3
PERCENTAGES:				
Number of Buildings	80.2	11.0	8.8	100.0
Floor Area (MSF)	65.3	23.7	10.9	100.0
Total Energy (Ktherms)	50.0	33.7	16.2	100.0
GHG Emissions (Mmtcde)	58.6	27.3	14.1	100.0

Model-Based Clustering

Clustering is the process of identifying similar or coherent groups within a multivariate dataset. This is also often referred to as classification or segmentation, and has been used throughout the natural and social sciences in order to identify sub-groups for further study. Well-known clustering methods include hierarchical agglomerative clustering, iterative partitioning, and *k*-means clustering; however, many of these algorithms rely upon ad hoc assumptions and are not formally related to statistical theory. More recent work in classification, however, has been based on formal probability models (see Bock, 1996, for a survey).

Model-based clustering is a systematic approach to finding similar groups (see Banfield and Raftery, 1993, Fraley and Raftery, 2002, for a more detailed explanation of the statistical formulation). Instead of specifying a particular criteria or model, a finite mixture model is used to incorporate multiple models with differing functional forms and parametric assumptions, and then to select the best fitting statistical model in order to classify the data. Bayesian methods are then used to estimate the parameters, calculate simultaneously which model fits the data best and how many clusters explain the data the best, and the error attached to particular classifications. Model-based clustering is implemented using the MCLUST package in the R statistical programming language, freely available under license from the University of Washington (see Fraley and Raftery, 2010, for implementation, manual and license).

Model-based clustering was applied to the multifamily sector – the largest facility type group in the NYC data comprising 80.2% of the total dataset – in order to further sub-divide and classify buildings in this sector. We develop a classification procedure based on six different metrics:

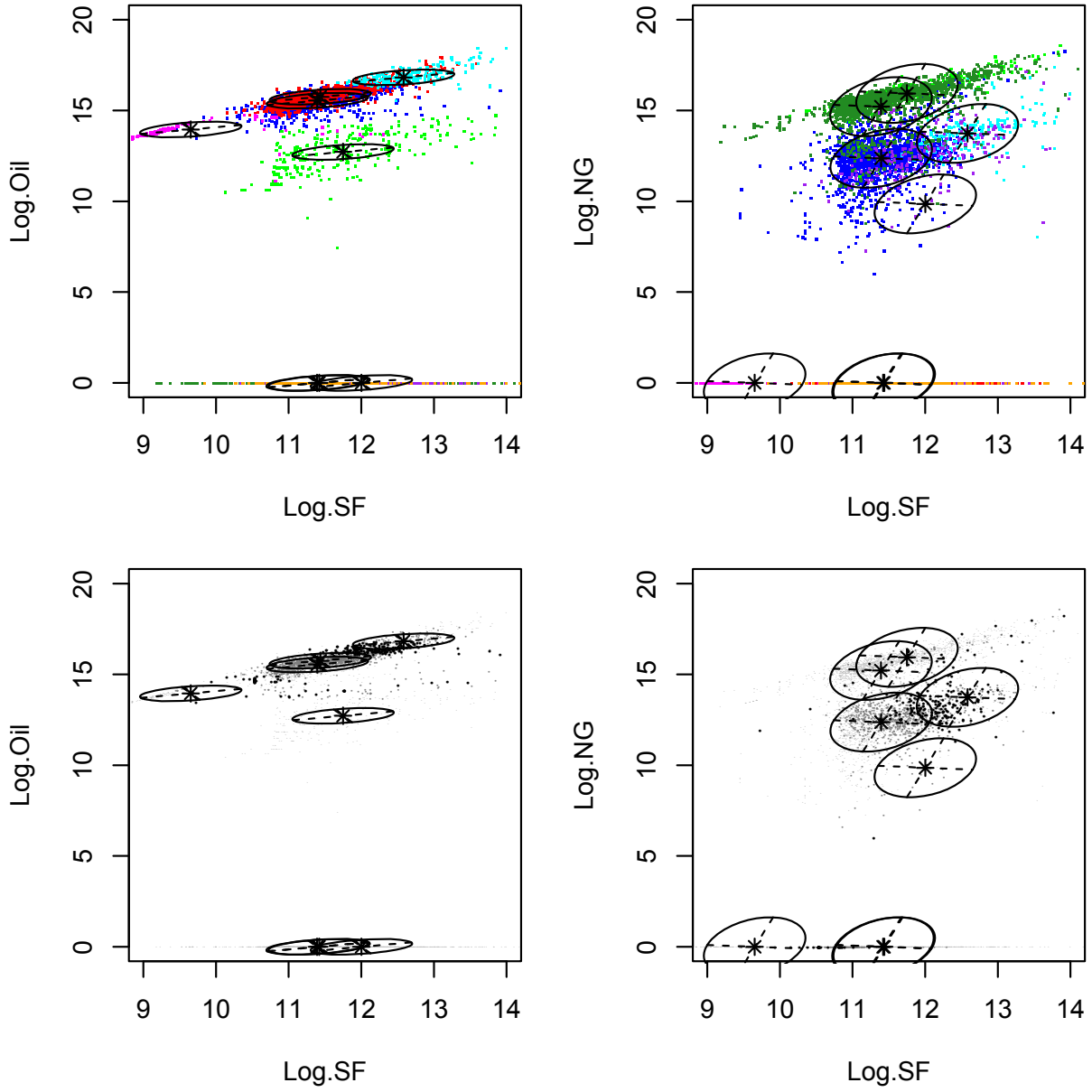
- four energy types: electricity, natural gas, steam, and all oil types (see footnotes)
- building age
- building size

where only the energy types can be compared using the same physical concept and units (kBTU). All oil types (numbers 2, 4, 5, and 6) collapsed into one category, since 95% of all buildings have just one oil type. Building age and size are measured in units of years and square footage.

Figure 2 identifies eight distinct clusters, along with classification uncertainties for each of the points identified in each of these plots. Clusters are summarized in Table 3, which shows the aggregated descriptive statistics for the main characteristics on which the buildings have been clustered, and Table 4, which show how the clusters differ with respect to their use of different energy sources. Some of the findings of the clustering method are summarized here, along with policy implications:

- Clusters 1 and 2 comprise more than half of all multifamily buildings. They have roughly the same average sizes (117.9 versus 101.1 thousand SF, or KSF), but clustering is able to distinguish them based on age (76.2 versus 58.2 years of age, respectively). Either group could be a promising candidate for retrofits, depending on renovation and retro-commissioning history.
- Cluster 5 is a steam-dominated cluster, which is relatively rare among the overall building population. It is a relatively small percentage of the overall square footage (8.2%) but a very high relative percentage of the greenhouse gas emissions (19.9%). Targeting this group of buildings with specific energy efficiency policies may yield disproportionate benefits, unless this apparent relationship is simply the spurious consequence of how national conversion factors have been applied to steam production.
- Clusters 1 and 6 use mostly natural gas, and are only distinguishable by the small percentage of energy used from oil in cluster 6. These clusters produce relatively less GHG emissions than their total area and use of energy, as derived from EPA's national conversion factors, and therefore may be a lower priority effort for efficiency measures, or may require a different type of intervention.
- Clusters 2, 3, 4 and 8 use mostly oil, and are distinguished only by their existing share of natural gas use. These buildings are to varying degrees possible candidates for fuel switching. However, while clusters 2 and 4 are both comprised of oil-dominated buildings, they have very different average sizes and ages (101.1 versus 360.1 KSF, and 76.2 versus 49.3 years old, respectively). These two clusters may also require different incentive or program structures in order to achieve optimum investment in energy efficiency.
- All clusters seem to have roughly 20-30% of their total energy use in electricity, except cluster 7, which is composed of 235 buildings that report 100% electricity use. Since these are multifamily buildings, it is possible that these buildings only reported common area lighting.

Figure 2. Clusters Based on Energy Use, Building Age and Size



In the upper row, ellipses indicate multivariate cluster densities in two-dimensional plots, and colors indicate classification into groups. Only two 2-dimensional plots are shown out of a possible thirty. In the lower row, size of circle indicates the uncertainty of assignment for each data point to a cluster. The clustering algorithm clearly distinguishes between different size buildings with particular energy systems. By inspection of all six dimensions, age and electricity use do not differentiate most of the clusters greatly; use of steam and oil rarely appear in the same building.

Table 3. Multifamily Building Clusters by Aggregate Statistics

Cluster	Number	MSF	E	NG	S	O	TotEN	GHG
1	1863	219.7	5028.2	12520.2	0.0	0.0	17768.1	1141.9
2	1588	160.6	3088.0	1166.3	0.0	9963.6	14378.5	1123.8
3	790	89.2	1571.1	0.0	0.0	6604.3	8264.6	659.7
4	208	74.9	1755.5	431.0	37.4	5171.5	7470.2	592.0
5	273	56.0	1535.7	202.9	2837.9	0.0	4632.5	978.1
6	291	50.8	1076.3	3680.8	15.1	153.0	4976.0	307.2
7	235	30.3	1092.6	0.0	0.0	0.0	1122.9	96.3
8	155	3.7	58.5	0.0	0.0	202.3	264.5	15.8

Clusters are ordered by total area in each cluster (SF, descending). Energy use is coded by type: electricity (E), natural gas (NG), steam (S) and all oil types (O). Clustering based on building age, size (MSF), four main energy types (MBtu) and greenhouse gases (Kmtcde).

Table 4. Multifamily Building Clusters by Average Statistics

Cluster	Age	KSF	%E	%NG	%S	%O	%SF	%EN	%GHG
1	58.2	117.9	28.6	71.3	0.0	0.0	32.1	30.2	23.2
2	76.2	101.1	21.7	8.2	0.0	70.1	23.4	24.4	22.9
3	73.9	112.9	19.2	0.0	0.0	80.8	13.0	14.1	13.4
4	49.3	360.1	23.7	5.8	0.5	69.9	10.9	12.7	12.0
5	56.5	205.3	33.6	4.4	62.0	0.0	8.2	7.9	19.9
6	59.1	174.7	21.9	74.7	0.3	3.1	7.4	8.5	6.2
7	61.3	128.8	100.0	0.0	0.0	0.0	4.4	1.9	2.0
8	71.1	23.9	22.4	0.0	0.0	77.6	0.5	0.5	0.3

Clusters are ordered by total area in each cluster (SF, descending). Energy codes are the same as above. Average statistics for each building in the cluster is indicated in the middle six columns, and each cluster as a percent of the overall building stock is indicated in the right three columns.

Seemingly-Unrelated Regressions (SUR)

Now that we have identified particular clusters of energy use, we now seek to relate them to underlying building characteristics. Due to the various possible combinations of end uses, systems and fuel types, we would expect some fuel types to be complementary and others to be substitutable. For example, we expect electricity use to be correlated with other energy uses, but in an apartment building that obtains steam for heat, we would not expect it to have an overlapping system such as oil for heating. Determining which building has what kind of system is very important for modeling building energy use.

Zellner (1962)'s seemingly unrelated regression (SUR) technique allows the flexible specification and simultaneous solution of systems of equations. The name comes from the fact that individual equations can appear to be unrelated, but are related by the structure of the data, either through covariance or explicit constraints on the regression model (such as the tendency to have certain combinations of energy systems). More detailed explanation of the formulation and assumptions for SUR can be found in Wooldridge (2002). The open-source 'systemfit' package in the R statistical programming language is used to fit the simultaneous equations (Henningsen and Hamann, 2011).

Table 5 shows the model fits, using many of the available fields from Portfolio Manager as covariates in the regression model. When we initially try to predict the use of various kinds of energy for multifamily buildings using general building features in Portfolio Manager – such as number of units, dishwashers, type of heat, age of building and so on – we can predict electricity use fairly well, but our predictions are fairly poor for all other energy types such as natural gas, steam, and all oil types; the first column, under M1, reports the R^2 for each of the individual equations.

We first add in percentage of energy derived from each energy source as an independent variable. The results, reported under column M2, show a significant improvement in modeling results. The model identifies all heat types better, with significantly higher R^2 for predicting natural gas, steam, and oil.

Finally, we add in the results of our model-based clustering. The third column of Table 5, under M3, shows that when we input the identified clusters into a regression model, we greatly improve our predictive modeling of building energy use for natural gas and oil. Model-based clustering contributes a significant improvement in results because the clustering model takes into account the difference in magnitude of various energy sources. In contrast, using dummy variables would only take into account correlation and collinearity between various energy sources, without taking into account their relative magnitude as well.

This has been a very preliminary application of clustering and the SUR regression model to building energy data. An analysis-of-variance (ANOVA) model would allow better explanation of the improvement from adding model-based clusters (model M3) to the previous model (M1). Future work should allow the regression model to incorporate the classification uncertainty with the regression model, and allow a better understanding of various model fits to the data.

Table 5. Model Fits from SUR

	Model R^2		
Energy Use	**M1**	**M2**	**M3**
Electricity	0.819	0.863	0.821
Natural Gas	0.318	0.454	0.553
Steam	0.129	0.555	0.526
All Oil	0.226	0.400	0.582

Model M1 uses many relevant building typology fields from Portfolio Manager, but no energy information. Model M2 adds percentage of various energy sources as covariates, in addition to the covariates in model M1. Model M3 uses model-based clusters in addition to all covariates in model M1.

Conclusions

This paper has sought to describe the NYC data, and to apply two new statistical techniques to model the data. The richness of the NYC data allows descriptive statistics and exploratory data analysis in order to compare between pre-defined categories, such as facility types as defined by the EPA. Model-based clustering is an attractive alternative method to identify particular clusters of interest where no previous categories exist, purely based on observed energy characteristics. Finally, the SUR model may be a useful way to model the co-dependence of different energy types within particular groups of buildings. All three techniques are being further developed on this dataset for use in other cities and research settings.

This paper presented a basic summary of building types and energy uses in Table 1 and Figure 1, along with a breakdown of area, energy use and GHG emissions in Table 2. This paper further found that the main multifamily sector can be subdivided into eight main subgroups, distinguishable by their heating fuel types (natural gas, oil, and steam), size and age. For example, clusters 2 and 4 may require differently structured incentives for efficiency, given their very different average size and age of building, while a special program may be necessary for cluster 5, which is comprised of relatively large, steam-dominated buildings. When these clustering methods are combined with SUR models, they lead to a marked improvement in regression results to predict each of the heating fuel uses, from the low R^2 range of 0.129 to 0.318, to the much more respectable R^2 range of 0.553 to 0.582. These methods and findings could be further applied to other sectors of interest.

This paper contains only two initial analyses of the benchmarking data collected by the City of New York. With a second year of data collection in 2012, and detailed audit information beginning in 2013, the author expects that there will be many further opportunities for energy and policy analysis of the NYC data. This analysis should also serve as an example for other cities to show how to use their benchmarking data. Other possible applications of these include predicting the likelihood of fuel-switching in a given building or geographic area; planning energy distribution infrastructure; and further developing policies, codes and ordinances that govern the design of building energy systems.

Acknowledgements and Disclaimer

This material is based upon work supported by the Energy Efficient Buildings HUB, an energy innovation HUB sponsored by the Department of Energy under Award Number DE-EE0004261.

References

Allcott, H. & M. Greenstone. 2012. "Is There an Energy Gfficiency gap?" *Journal of Economic Perspectives* 26 (1), 3–28.

Banfield, J. & A. Raftery. 1993. "Model-based Gaussian and Non-Gaussian Clustering." *Biometrics*, 803 – 821.

Blumstein, C., B. Krieg, L. Schipper, & C. York. 1980. "Overcoming Social and Institutional Barriers to Energy Conservation." *Energy* 5 (4), 355–372.

Bock, H. 1996. "Probabilistic Models in Cluster Analysis." *Computational Statistics & Data Analysis* 23 (1), 5–28.

Burr, A., C. Keicher, & D. Leipziger. 2011. *Building energy transparency: A framework for implementing U.S. commercial energy rating & disclosure policy*. Technical report, Institute for Market Transformation.

Fraley, C. & A. Raftery. 2010. *MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering*. Technical Report 504, Department of Statistics, University of Washington.

Fraley, C. & A. E. Raftery. 2002. "Model-Based Clustering, Discriminant Analysis, and Density Estimation." *Journal of the American Statistical Association* 97 (458), 611–631.

Henningsen, A. & J. D. Hamann. 2011. *systemfit: a package for estimating systems of simultaneous equations in R*. http://www.systemfit.org/.

Wooldridge, J. M. 2002. *Econometric Analysis of Cross Section and Panel Data*. MIT Press.

Zellner, A. 1962. "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias." *Journal of the American Statistical Association* 57 (298), 348–368.