Applied Energy 160 (2015) 153–163

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy

Comparison of integrated clustering methods for accurate and stable prediction of building energy consumption data

David Hsu

Department of Urban Studies & Planning, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

HIGHLIGHTS

- Summary of recent work identifying sub-groups of energy consumption in buildings.
- Clusterwise (or latent class) regression gives superior prediction accuracy.
- *K*-means gives more stable clusters when the correct number of clusters is chosen.
- A tradeoff between prediction accuracy and cluster stability seems to exist.

ARTICLE INFO

Article history:
Received 20 January 2015
Received in revised form 26 August 2015
Accepted 30 August 2015

Keywords:
Cluster-wise regression
Buildings
Energy consumption
Prediction accuracy
Cluster stability
Latent class regression

ABSTRACT

Clustering methods are often used to model energy consumption for two reasons. First, clustering is often used to process data and to improve the predictive accuracy of subsequent energy models. Second, stable clusters that are reproducible with respect to non-essential changes can be used to group, target, and interpret observed subjects. However, it is well known that clustering methods are highly sensitive to the choice of algorithms and variables. This can lead to misleading assessments of predictive accuracy and mis-interpretation of clusters in policymaking.

This paper therefore introduces two methods to the modeling of energy consumption in buildings: clusterwise regression, also known as latent class regression, which integrates clustering and regression simultaneously; and cluster validation methods to measure stability. Using a large dataset of multifamily buildings in New York City, clusterwise regression is compared to common two-stage algorithms that use *K*-means and model-based clustering with linear regression. Predictive accuracy is evaluated using 20-fold cross validation, and the stability of the perturbed clusters is measured using the Jaccard coefficient. These results show that there seems to be an inherent tradeoff between prediction accuracy and cluster stability. This paper concludes by discussing which clustering methods may be appropriate for different analytical purposes.

1. Introduction

Buildings have become a major focus of energy policy worldwide, because they constitute nearly 40% of all worldwide primary energy consumption and associated greenhouse gas emissions [1,2]. Many different policy initiatives have been recently proposed that are intended to affect building energy consumption. In order for policymakers to design and target policies to reduce building energy consumption effectively, it is necessary to develop ways to find relevant sub-groups in the overall population using methods that are stable, consistent, and statistically-valid.

However, buildings may be grouped in many different ways, because they are complex, multi-dimensional, and heterogeneous objects. In addition, the overall population of buildings may be composed of sub-groups, and appropriate groupings may vary considerably at different scales, such as at the urban, metropolitan, regional, or national level. These scales often represent particular jurisdictions that implement policies and regulations on buildings.

Sub-groups in the overall population of buildings can be found or defined in many possible ways: a large group of papers is reviewed below which seek to do this in the building energy consumption literature. This paper critiques a particularly popular approach, which uses quantitative clustering methods as the first of a two-stage process: that is, as a pre-processing step to divide the overall data into smaller groups, which are then subsequently modeled using either physics-based simulation or statistical regression models. While this approach almost always improves subsequent modeling because it allows separate and different

E-mail address: ydh@mit.edu

http://dx.doi.org/10.1016/j.apenergy.2015.08.126
0306-2619/© 2015 The Author. Published by Elsevier Ltd.

models to be fit to each cluster, this approach may also ignore statistical uncertainties in the clustering step, which leads to over-fitting and/or over-confidence in the results in the second analysis stage. Specifically, it is well-known in the statistical literature that clustering methods are highly sensitive to the choice of method and variables, initial assumptions, cleaning steps taken, the distribution of the data, and that clustering results have significant statistical uncertainties. This is why one scholar of clustering methods describes it as "one of the most fundamental modes of understanding and learning", and yet goes on to say that "in spite of the fact that *K*-means was proposed over 50 years ago and thousands of clustering algorithms have been published since then, *K*-means is still widely used. This speaks to the difficulty in designing a general purpose clustering algorithm and the ill-posed problem of clustering." [3, page 651].

This paper therefore introduces two methods to the building energy consumption literature. First, clusterwise regression (also known as latent class regression) is a statistically-valid technique that integrates classification and regression simultaneously. Second, cluster validation metrics measure the stability of clusters when they are subjected to small perturbations, such as adding noise, bootstrapping, or taking subsets. These methods are likely to be useful in other areas of energy modeling and analysis that are applied to large, heterogeneous populations, and that also rely upon clustering or partitioning observed behavior into different groups. Finding stable and valid clusters is necessary in order to apply and target policies consistently.

These methods improve the modeling of energy consumption in buildings in two ways: first, the integrated approach of clusterwise regression simultaneously optimizes for prediction accuracy and explanatory groupings in a statistically-valid approach. Second, it will be shown that clusterwise regression achieves significantly superior prediction accuracy over the competing two-stage approaches that use *K*-means and model-based clustering in the initial step. However, since the clusters found through clusterwise regression are found to be *less* stable than those found in the two-stage processes with respect to small perturbations, this highlights a fundamental and perhaps unavoidable tradeoff between cluster stability and prediction accuracy.

This rest of this paper is organized as follows. Section 2 reviews the extensive literature that uses clustering to predict building energy consumption, as well as some of the statistical caveats associated with clustering methods. Section 3 then reviews the statistical theory of clusterwise regression using model-based clusterings, as well as the appropriate metrics for cluster validation and stability. Section 4 describes a comprehensive dataset of building energy consumption in a large and highly diverse population of almost 4000 New York City multifamily buildings, and Section 5 presents the results of the analysis and discusses the relative advantages and disadvantages of clusterwise regression over two stage approaches. Section 6 concludes the paper by discussing the implications of the results for energy modeling and analysis, and policies targeted at particular subgroups of buildings.

2. Related work

Clustering methods have been used widely throughout the energy consumption literature. A number of articles in this journal have used clustering to extract similar groups out of overall population data: examples include searching for groups composed of similar energy consumers, load or generation profiles, building or site feasibility [4–9]. Since the overall literature that uses clustering for energy analysis is quite large, this review will focus on building energy consumption as a particular example to illustrate how these clustering methods, which are often thought of as unsupervised learning, are often used in predictive analyses.

Similar to other areas of energy modeling, a wide variety of quantitative methods have been used to describe variation within overall populations of buildings. Building sectors are often analyzed in terms of archetypes, which are based on a variety of approaches, such as expert knowledge [10]; as key sectors of aggregated energy consumption [11,12], or simply as the result of ad hoc decisions to stratify the overall population. Other methods, such as principal components analysis, principal components regression, partial least squares, and self-organizing maps have all been used to describe the key dimensions or linear combinations that describe the variation in buildings, either for exploratory factor analysis [13], parameter investigations [14], or to provide customized recommendations [15]. Decision trees and their extensions, such as classification and regression trees (CARTs) and random forests, have also been applied [16,17].

For buildings, however, clustering has by far been the most popular approach to identify sub-groups in the overall population. Clustering methods used include *K*-means, hierarchical, model-based, fuzzy, or other clustering approaches, with *K*-means as the most popular. Examples include using clustering methods to summarize the key clusters for subsequent simulation analysis [18,19], to assess clusters for particular behaviors and opportunities [20,21], or to identify key patterns from high-frequency data [22–26].

An increasingly popular approach is to use clustering methods as a pre-processing step for subsequent models. Examples include, but are not limited to, using archetypes to justify subsequent regression analysis of aggregate residential energy consumption [27]; to find segments for a complex 'grey-box' model [19]; and to apply subsequent multivariate analysis to measure the operating performance of particular systems and building types [4,28,29].

However, in the statistical literature it is well-known that initial choices in clustering methods can give drastically different results. Depending on the overall goals, choice of algorithm, variables selected, initial assumptions, and the natural shape of the data, clustering results can vary dramatically [3,30,31]. Hennig [32] points out a number of possible problems with clusterings, even if they are stable. To take a simple example, *K*-means clustering assumes and subsequently finds a specific number of clusters, but when applied to homogeneous data, this algorithm will still find the assumed number of clusters even if they are essentially meaningless. In addition, stable clusterings may still be meaningless if they fail to distinguish useful subsets of the overall data. Humans can still sometimes identify meaningful patterns that computers cannot. Finally, clustering algorithms taken to the extreme, such as in hierarchical clustering with many branches, may find that each data point belongs to its own cluster, which is also useless.

This review and this overall paper are therefore intended to raise awareness of the potential problems that need to be considered when using clustering. These issues are often overlooked because of the belief that clustering is an unsupervised learning problem, in which there may be different clusters for different purposes, and therefore there is no one 'true' clustering that exists within the data. However, in many energy analyses and particularly in the previous work described above, cluster analysis is clearly intended to identify heterogeneous sub-groups in order to improve subsequent prediction. Fig. 1 illustrates in a flowchart-style diagram how common approaches in the literature often integrate clustering and prediction. At the top, clustering and prediction are often two important and inter-related activities. Key considerations are the choice of the number of clusters, assignment to clusters, and model selection for accurate prediction. The large arrows at left describe common approaches or algorithmic steps,

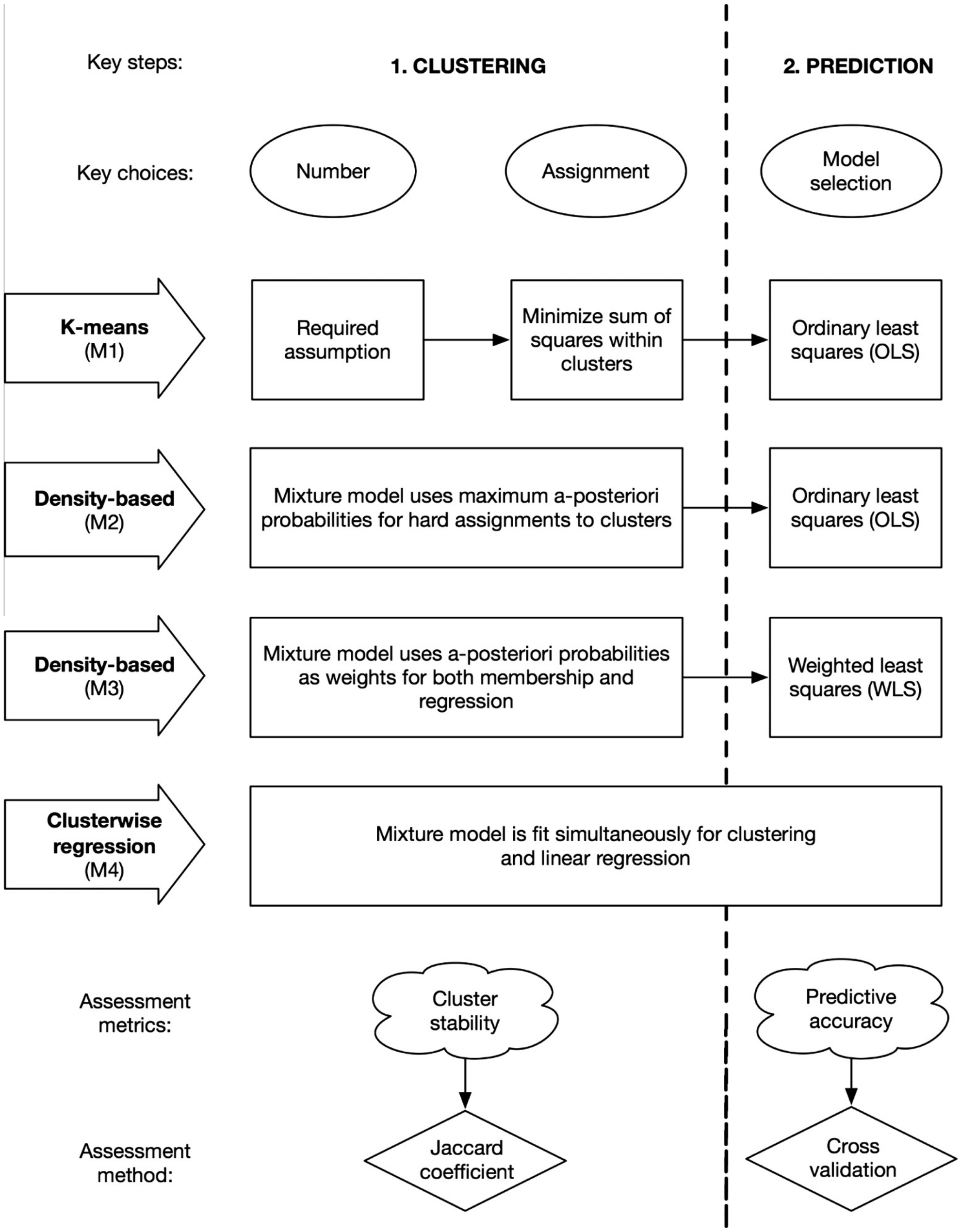
Key steps:
1. CLUSTERING
2. PREDICTION
Key choices:
Number
Assignment
Model selection
K-means (M1)
Required assumption
Minimize sum of squares within clusters
Ordinary least squares (OLS)
Density-based (M2)
Mixture model uses maximum a-posteriori probabilities for hard assignments to clusters
Ordinary least squares (OLS)
Density-based (M3)
Mixture model uses a-posteriori probabilities as weights for both membership and regression
Weighted least squares (WLS)
Clusterwise regression (M4)
Mixture model is fit simultaneously for clustering and linear regression
Assessment metrics:
Cluster stability
Predictive accuracy
Assessment method:
Jaccard coefficient
Cross validation

Fig. 1. Diagram of clustering and modeling processes. Reading from left to right, and then top to bottom, building energy consumption data has often been modeled in a two-stage process, first clustering and then modeling. Key choices in ovals are number of clusters, assignment to clusters, and model selection. Large arrows indicate common approaches and steps: combining *K*-means (M1) and model-based clusters (M2) with ordinary least squares regression. A-posteriori probabilities can also be used as weights in subsequent weighted linear regression (M3) or integrating all steps in latent-class regression (M4). Clouds and diamonds at the bottom indicate key metrics of cluster stability and predictive accuracy, measured by the Jaccard coefficient and cross validation.

which are to use *K*-means or a model-based clustering to pre-process the data, and then to model the data using ordinary least squares regression (OLS). It is also possible to use model-based clustering to give probabilities of membership in each cluster, and also to use these weights in subsequent regression. Clusterwise or latent class regression will be described further below, and offers a statistically-principled method to integrate clustering and prediction simultaneously. Finally, the validity of clustering and prediction can be tested using various assessment metrics and methods such as the Jaccard coefficient and cross validation respectively.

The next section, Section 3, will discuss how the various clustering and modeling techniques are applied and Section 5 evaluates these the implementation of these various approaches on a common dataset.

3. Theory & calculations

The theory of *K*-means and model-based clustering are well-known [3,33], and have been widely implemented in the area of modeling building energy consumption, as the literature review in the previous section demonstrated. The theory and calculations presented in this section will therefore be limited to the less familiar and new techniques introduced by this paper to energy analysis. Statistical theory for clusterwise regression and cluster validation are introduced in Sections 3.1 and 3.2, respectively, and Sections 3.3 and 3.4 describe how these theories are implemented in calculation steps and software.

3.1. Clusterwise regression

Clusterwise regression as a combination of cluster analysis and regression is usually first credited to the work of Späth [34–36] and Cron [37]. Clusterwise regression generally sets up an objective function that simultaneously allows for both clustering and regression, and then uses the expectation–maximization (EM) algorithm [38] to find the optimal parameters to minimize this objective function. In this model, the EM algorithm is used to find the optimal parameters to capture the heterogeneity in the regression coefficients and variances in the data, but within each cluster the functional relationship between explanatory and dependent variables remains the same. This section will outline the approach and important steps; a more detailed treatment of clusterwise regression is provided in Leisch [39] and Grün and Leisch [40].

Starting with the standard linear regression problem:

$$y = \beta x + \epsilon \tag{1}$$

where the vector of responses y is of length N, the vector of coefficients β is of length P, the explanatory predictors are the matrix x of $N \times P$ dimension, and the error term ϵ is independently and identically distributed (i.i.d) normally with zero mean and a constant variance of σ^2, that is $\epsilon_i \sim N(0, \sigma^2)$.

The key elements of clusterwise regression are to create finite mixture models for each of K classes with the specific form:

$$h(y|x,\psi) = \sum_{k=1}^{K} \pi_k f_k(y|x,\theta_k) \tag{2}$$

where the conditional density h is a function of y conditioned on x and the vector of all parameters, $\psi = (\pi_1, \ldots, \pi_K, \theta'_1, \ldots, \theta'_K)'$, π_k is the prior probability of y belonging to component k with density function f_k and the parameter vector θ_k for each component. Prior probabilities are of course constrained to be greater than zero and sum to 1.

The posterior probability P that the observation (x, y) belongs to class j among the K classes is given by:

$$P(j|x,y,\psi) = \frac{\pi_j f_j(y|x,\theta_j)}{\sum_k \pi_k f_k(y|x,\theta_k)} \tag{3}$$

The data is clustered by either assigning each observation to the class with maximum posterior probability, or else simulating cluster assignments. The functions $f_j(y|x,\theta_j)$ are referred to as the mixture components, and can be implemented with different statistical distributions. Gaussian response is the distribution typically used in linear regression, as in Eq. (1). The choice of distribution f_j results in different models of clustering: the univariate normal distribution gives latent class regression, the exponential distribution gives generalized linear models, and the multivariate normal distribution without covariates in the components gives model-based clustering [41].

A log-likelihood function for a sample of N observations is given by:

$$\log L = \sum_{n=1}^{N} \log h(y_n|x_n,\psi) = \sum_{n=1}^{N} \log\left(\sum_{k=1}^{K} \pi_k f_k(y_n|x_n,\theta_k)\right) \tag{4}$$

The parameters are estimated by maximizing the log-likelihood function subject to constraints on the mixing proportions and by applying the following two steps of the EM algorithm iteratively, where the cluster assignments are considered as missing data:

1. The a-posteriori probabilities are determined (E-step);
2. the prior weights and the component specific parameters are determined based on these a-posteriori weights (M-step).

These steps are repeated until a particular criteria is reached, such as if the log-likelihood improvement reaches a particular threshold, the parameters converge, or a maximum number of iterations is reached. The Bayesian Information Criterion (BIC) is used to penalize the log-likelihood in terms of the number of parameters required and to choose the suitable number of mixture components. Multiple restarts can be used to find avoid local maxima and to find the global maximum.

Clusterwise regression and model-based clusterings have particular advantages compared to the more popular heuristic algorithms such as K-means. First, clusterwise regression offers a statistically-valid theory and approach to integrate clustering and prediction, since this integrated approach takes into account uncertainty in the clustering process when simultaneously fitting the regression models. In contrast, clustering using K-means does not pass along any information about uncertainties in clustering to the subsequent ordinary least squares (OLS) regression. Second, statistical theory can also be used to obtain key parameters, such as the appropriate choice of number of clusters for the data [33], that otherwise are required as an assumption in heuristic algorithms such as K-means (although an extensive literature on how to make this assumption is discussed in Tibshirani et al. [42]). Third, mixture-based models are robust when extended to new data: Hennig [43] points out that heuristic methods such as K-means with fixed numbers of clusters and without trimming can be spoiled by adding a single outlier, while mixture models that determine the appropriate number of clusters based on the entire statistical distributions are comparatively more robust.

However, clusterwise regression does not avoid all of the problems of other clustering methods. Brusco et al. [44] find that while clusterwise regression can explain a high proportion of explained variance, much of the improvement comes from the clustering process and not necessarily from attributing response to different variables within each cluster. Also, the statistical significance of particular predictors within each cluster may not be accurate because a particular model criterion has to be chosen, which must be informed by the data itself [45].

This paper therefore evaluates clusterwise regression both in terms of overall predictive accuracy and cluster stability. A standard 20-fold cross validation using sampled training and test data from the overall dataset is used to assess the overall accuracy of predictions [46].

3.2. Cluster stability & validation

The stability and validation of clusters have been extensively studied, particularly in the context of high-dimensional genomic data. The Jaccard similarity coefficient is often used to measure the similarity of clusters [32,43]. This approach resamples or perturbs the data in various ways, compares the resulting new clusters with the original clustering, and calculates the mean similarity within clusters. Where the set of observations x_n can be initially mapped into two subsets $\{C_1, \ldots, C_S\}$ and $\{D_1, \ldots, D_T\}$ with the number of subsets as S and T, respectively, the Jaccard coefficient γ for any two subsets C_s and D_t is defined as:

$$\gamma(C_s, D_t) = \frac{|C_s \cap D_t|}{|C_s \cup D_t|}, \quad C_s, D_t \subseteq x_n \tag{5}$$

which is the size of the interaction of two subsets divided by the size of their union. The Jaccard coefficient ranges between 0 and 1, where higher values indicate greater similarity of groupings.

Induced clusters are created by resampling methods such as bootstrapping, subsets, or adding noise to the sampled distribution. Since the induced clusters are meant to have relatively small changes relative to the original data, for each of the clustering methods, clusters are then matched by maximizing the number of common, non-modified, observations within each cluster. The clusterwise Jaccard coefficient is then calculated across all of the samples as the mean similarity measure between clusters in the original dataset and clusters in the resampled dataset.

3.3. Modeling and assessment steps

Implementation of these methods, as well as comparing existing methods, required initial data cleaning and some necessary assumptions in order to carry out the different modeling processes and to make fair comparisons. For example, variables selected for the clustering and modeling processes were selected through a lengthy regularization process described in a previous paper [47]. There is unfortunately very little guidance in the statistical literature on variable or feature selection for clusterwise regression, so the best variables were chosen for predictions across the entire "global" dataset.

Fig. 1 shows how the two main processes of clustering and prediction are handled in different modeling processes. It is important to note that two separate calculations are performed, respectively, to assess first prediction accuracy and then second, clustering stability.

1. Prediction accuracy is calculated using cross validation. "Folds" composed of a training and test set are created by randomly sampling the data with replacement. By repeating our calculations on each training set and then comparing our predictions with the test set, the cross-validated mean-squared-error (CVMSE) can be calculated. Twenty (20) folds, i.e., 95% training and 5% test sets, were selected to avoid too small training sets in some clusters and to best avoid rank deficiency in making subsequent predictions; stratified sampling was not used to avoid making any clustering assumptions a priori. Each of the main approaches, denoted by the large arrows on the left-hand side of Fig. 1, are calculated in the following steps:
   (a) K-means (M1): assume $K = 4$ or $K = 9$, to compare with the subsequent model-based and cluster-wise regression approaches. An OLS model is then trained within each cluster and then used to predict the outcome values for the remaining test data in each cluster, and to calculate the CVMSE across folds.
   (b) Model-based clustering (M2): the best fitting mixture model is selected by minimizing the Bayesian Information Criterion (BIC) [48]. Different clusters are identified from the training data and then these cluster memberships are assigned using hard assignment by maximum a-posteriori probabilities for the remaining test data [49]. An OLS model is then trained within each cluster and then used to predict the outcome values for the remaining test data in each cluster, and to calculate the CVMSE across folds.
   (c) Model-based clustering (M3): same as above, but cluster memberships are described by a-posteriori probabilities, which are then passed to a weighted linear regression (WLS).
   (d) Clusterwise regression model (M4): integrates clustering and prediction as described in Section 3.1, and the best fitting mixture model is also selected by minimizing the BIC.
2. Clustering stability is first observed and then calculated:
   (a) cross-validation for predictive accuracy creates some variation in the number of clusters, because the clusters are formed on different training sets, i.e., 95% of the whole dataset. This is briefly discussed qualitatively with the other results in Section 5.2; and
   (b) by perturbing the data using noise, subsetting, or bootstrapping, the Jaccard coefficient measures stability between similar clusters after small perturbations [32].

3.4. Software

All of the data cleaning and integration was performed in the `R` environment for statistical computing and graphics [50]. The algorithm for K-means clustering is in the base `stats` package of `R`. Initial variable selection and regularization was carried out using the `glmnet` package [51]. The algorithm for model-based clustering is the `mclust` package [48]. Clusterwise regression is computed using the `flexmix` package [39,40]. For the predictive analysis and cross validation, extending the cluster identification from training data to test data using the various clustering models required the `clue` package [49]. Cluster stability for the cluster found in K-means, model-based, and clusterwise regression are all calculated using the `clusterboot` method in the `fpc` package [32,43].

4. Data

The data analyzed in this paper has been described extensively in a previous paper [47], so only specific changes in the dataset will be described here.

Large buildings represent 48% of all primary energy use in New York City (compared to 17% transportation and 35% from small buildings), and multifamily and office buildings represent 87% of all gross floor area of all large buildings [52]. The datasets were assembled from multiple data sources, including the U.S. Environmental Protection Agency's Portfolio Manager, the City of New York's Primary Land Use Tax Lot Output (PLUTO) database, real estate and financial information from the CoStar Group, and census tract level information from the U.S. Census, including information from the 2010 decadal Census, the 2011 American Housing Survey, and the 2013 American Community Survey (ACS). After cleaning, the datasets for the analysis have 3902 multifamily housing buildings and more than 250 possible continuous predictors.

The outcome variable to be modeled for each building is the center-normalized logarithm of total metered energy, so all results will be interpreted in terms of standard deviations from the mean. Site energy was modeled since all of the buildings are in one city and therefore all share the same conversion coefficients from source (primary) energy.

Table 1
Descriptive statistics for key variables and interactions. Abbreviations: PM = U.S. Environmental Protection Agency Portfolio Manager data, PLUTO = City of New York Primary Land Use and Tax Output, ACS = U.S. Census American Community Survey data. All ACS data describes the surrounding census tracts. Interactions for each variable are formed by multiplying the predictors for each observation.

Variable	Data (units)	Min	25%	50%	75%	Max
Log Total Site Energy	PM (kbtu)	0.3	4.2	4.3	4.5	11.4
% Electricity	PM (0–1)	0.0	0.2	0.2	0.3	1.0
% Steam	PM (0–1)	0.00	0.0	0.0	0.0	1.0
Log Tax Value	PLUTO ($)	11.6	14.1	14.9	15.8	18.2
% Built in 80s	ACS (0–1)	0.0	0.0	0.0	0.1	0.8
% HH Electric Heat	ACS (0–1)	0.0	0.0	0.1	0.2	0.7
% Electricity: % Steam	Interaction	0.0	0.0	0.0	0.0	0.3
% Electricity: Log Tax Value	Interaction	0.0	2.3	3.2	5.0	16.8
% Steam: % Built in 80s	Interaction	0.0	0.0	0.0	0.0	0.8

Independent variables or predictors for both clusterwise regression, and the two-stage clustering and regression procedure, were selected using regularization [47], and also center-normalized. Descriptive statistics for the outcome and predictors are presented in Table 1.

5. Results & discussion

The results will be discussed in terms of the two main assessment methods, prediction accuracy and cluster stability, and also in terms of the implications for policy.

Table 2
Cross-validated Mean Squared Error (CVMSE) for all observations. Prediction error is calculated as the CVMSE of the error between the observed data and the predicted data in 5% test sets, based on 20-folds.

Method		CVMSE	
		Mean	SD
K-means, *K* = 4	M1	4.99	16.72
K-means, *K* = 9	M1	0.49	0.28
Model-based + OLS	M2	0.48	0.26
Model-based + WLS	M3	0.48	0.26
Clusterwise	M4	0.30	0.15

5.1. Prediction accuracy

Typical linear regression approaches applied to an overall population, such as used by the EPA's Energy Star model, have relatively low out-of-sample predictive accuracy. Measured in terms of the standardized outcome, these models typically only predict the energy consumption intensity of a building within one or two standard deviations. When a model using regularization was applied to the entire population, the 10-fold CVMSE of a single linear regression model applied to the entire dataset was roughly 0.40–0.46 [47]. This is roughly on par with the results of the 20-fold cross-validation analysis applied in this paper, summarized in Table 2. The two-stage approaches that combine *K*-means (with *K* = 9) or the model-based clustering with OLS and WLS approaches all give similar CVMSE of approximately 0.48–0.49, with small standard deviations ranging from 0.26 to 0.28. The clusterwise approach gives a slightly better CVMSE of 0.30 with a standard deviation of 0.15. However, the two-stage *K*-means analysis with *K* = 4 gives extremely poor CVMSE, with a mean of 4.99 and a standard deviation of 16.72. This last result does not initially seem correct until the errors introduced by the clustering process are examined more closely.

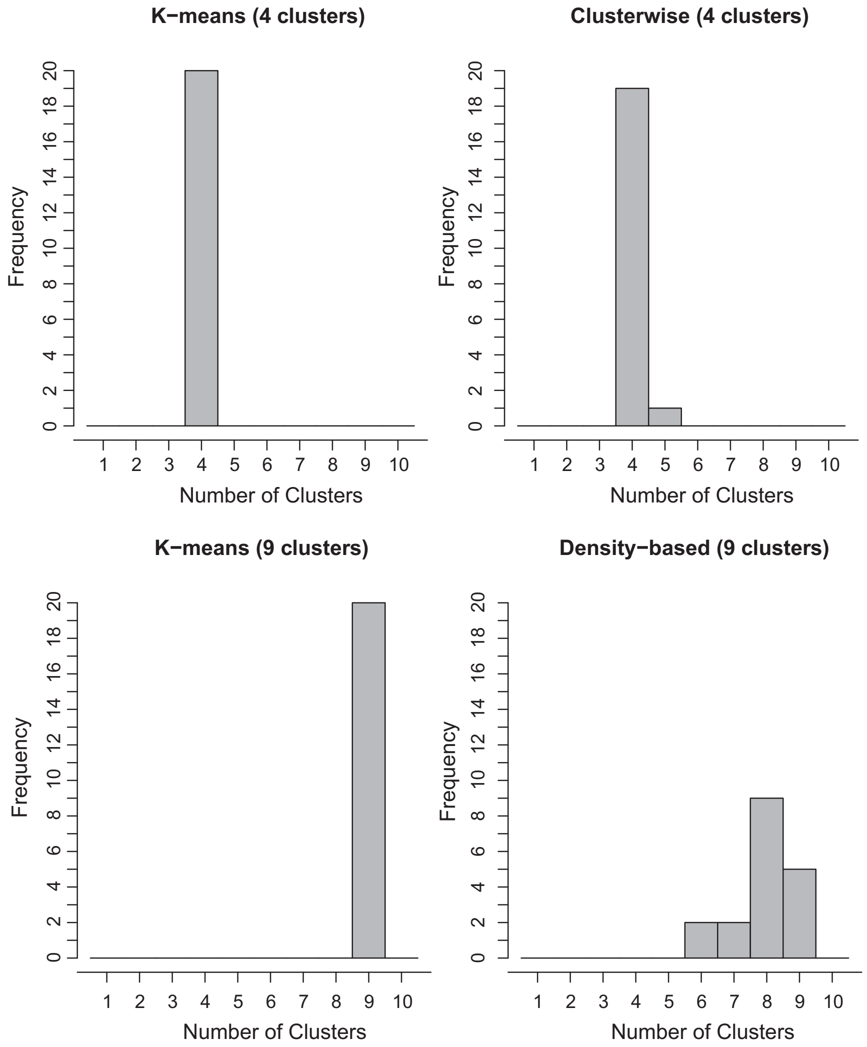

Fig. 2. Histograms of number of clusters found in 20-fold cross validation. *K*-means approaches in the left-hand column automatically find the assumed number of clusters in each fold, while in the right-hand column, the clusterwise and model-based approaches find different number of clusters depending in each fold, because the input 95% training data varies by fold.

It is possible to examine the error introduced by the specific clusters across folds because many of the clusters were found to be fairly distinct in their composition. When clustering is performed with *K*-means, the number of clusters assumed automatically results in finding that number of clusters. In Fig. 2, the left-hand column for the *K*-means approaches shows that all 20 folds return 4 and 9 clusters, as assumed. However, in the clusterwise and model-based approaches, the number of clusters can vary by fold because each clustering is based on a different 95% training dataset. As the figure shows, clusterwise regression consistently finds 4 clusters nineteen out of twenty times, while the number of clusters found by model-based regression ranges between 6 and 9 over the twenty folds. In general, while the identifying labels for clusters can be permuted without loss of information, if there are different numbers of clusters in each fold then it is difficult to compare information for individual clusters.

However, the clusterings found in this dataset are fairly distinct. In Table 3, the number of observations in each cluster averaged across folds appear to be distinctly separated because the standard deviations in the number of observations per cluster are much lower than the associated means; that is, across the folds in cross-validation, similarly-sized clusters repeatedly appear and do not seem to mix or overlap with one another. The right-hand columns show that the mean number of observations per cluster are separated by more than the calculated standard deviation in observations per cluster, with most clusterings finding one large cluster and other distinct clusters. It is generally easier to compare the largest clusters that have a high percentage of total observations, but it may be difficult to compare the clusters with higher identifying numbers from clusterwise regression or the model-based clusterings, because the CVMSE may be skewed by folds with different numbers of clusters.

Table 3
Cross-validated prediction error and composition by cluster. Columns show the method used, cluster IDs, the mean and standard deviation of the prediction CVMSE for 20-fold cross validation, and the size of each cluster across folds.

Method	Cluster ID	CVMSE		Obs. in cluster		
		Mean	SD	Number	SD	% tot
K-Means (M1, *K* = 4)	1	0.36	0.19	2325.9	69.6	59
	2	17.53	60.43	1086.2	20.2	28
	3	2.32	4.20	345.0	27.2	9
	4	1.17	1.41	172.0	27.8	4
K-Means (M1, *K* = 9)	1	0.35	0.36	1360.6	140.2	35
	2	0.35	0.26	872.5	47.7	22
	3	0.21	0.12	613.5	68.6	16
	4	0.34	0.32	300.4	28.9	8
	5	0.84	1.28	261.5	25.2	7
	6	1.03	0.83	209.2	31.1	5
	7	0.95	1.29	159.5	33.6	4
	8	3.03	6.66	123.2	48.0	3
	9	0.60	1.39	28.8	10.0	1
Density + OLS (M2)	1	0.36	0.24	1870.2	256.4	47
	2	0.26	0.22	954.6	235.8	24
	3	0.58	0.50	401.3	172.2	10
	4	2.30	4.81	220.2	39.2	6
	5	0.67	0.58	177.5	27.2	4
	6	0.64	0.80	141.9	34.6	4
	7	0.83	1.07	122.8	31.6	3
	8	0.11	0.11	63.9	37.3	2
	9	0.14	0.06	36.7	20.0	1
Density + WLS (M3)	1	0.36	0.24	1870.2	256.4	47
	2	0.26	0.22	954.6	235.8	24
	3	0.58	0.50	401.3	172.2	10
	4	2.25	4.72	220.2	39.2	6
	5	0.67	0.58	177.5	27.2	4
	6	0.64	0.80	141.9	34.6	4
	7	1.00	1.37	122.8	31.6	3
	8	0.12	0.10	63.9	37.3	2
	9	0.14	0.06	36.7	20.0	1
Clusterwise (M4)	1	0.06	0.03	2517.6	238.8	63
	2	0.09	0.08	838.3	258.0	21
	3	0.32	0.98	414.8	171.8	10
	4	5.44	4.14	153.5	10.7	4

Examining the prediction error by column is presented in the middle columns of Table 3. In the case of *K*-means with *K* = 4, the largest cluster (ID = 1) which comprises 59% of the observations, has a relatively low mean CVMSE of 0.36 and a standard deviation of 0.19. Many of the errors in this analysis stem from the very large errors in cluster 2, which is 28% of the dataset. In contrast, when we set *K* = 9, the errors in most of the clusters drop, meaning that there is a small subset of observations which are either poorly clustered or modeled.

In contrast to the *K*-means results, the model-based approaches combined with OLS and WLS, designated as M2 and M3, do not give very different results from one another, which was surprising to the author. The a-posteriori probabilities obtained from the model-based clustering have a mean and standard deviation of 0.70 and 0.21, respectively, which seems to indicate fairly strong assignment to clusters whether using hard assignment or weighting. However, the predictive accuracy of the model-based clustering methods M2 and M3 remain virtually the same and on par with *K*-means.

Clusterwise regression (M4), however, clearly has much lower CVMSE readings for three out of its four clusters, and the outliers (ID = 4) comprise only 4% of the data in this analysis.

Fig. 3 plots this same data and may be more intuitively interpreted. Reading from top left and counter-clockwise, it can be seen that *K*-means with *K* = 4 has a large concentration of observations in the first cluster (ID = 1, with 59% of the observations) with low mean and variation in the CVMSE, but the plot shows a clear cluster of outliers and another large cluster (ID = 2, with 28% of the observations) predicts poorly enough that it is beyond the limits of the plot. Below, *K*-means with *K* = 9 shows that the majority of the clusters have decent predictive accuracy relatively stable in terms of the mean and standard deviation in CVMSE, but that there are 4–5 clusters that have very poor predictions. Similarly, model-based clustering has a large number of clusters with poor predictive accuracy as measured by the mean and standard deviation in CVMSE. However, clusterwise regression clearly has the largest number of observations near the origin, meaning that 94% of the observations in the first three clusters (IDs = 1, 2 and 3) have a CVMSE of less than 0.32.

The key determinants of energy consumption found by clusterwise regression are reported in Table 4, and show clearly that different factors influence the behavior of each subgroup. This information could be used to target interventions or policies to each cluster. However, the significance patterns also vary by group, but can only be used for exploratory data analysis, since the particular reported model was selected based upon criteria from the initial model runs and so *p*-values do not accurately reflect the null hypothesis that the coefficients and effects are actually zero [45,39].

5.2. Cluster stability

Cluster stability was measured or validated in two ways. First, in the predictive accuracy calculation above, when the folds were sampled, the different methods found different numbers of clusters: Fig. 2 showed this and the variability of the number of clusters depending on the folds was discussed above. Second, cluster stability metrics based on the Jaccard coefficient and the bootstrapping, subset, and noise schemes are calculated and presented in Table 5 and Fig. 4 for all of the clustering methods.

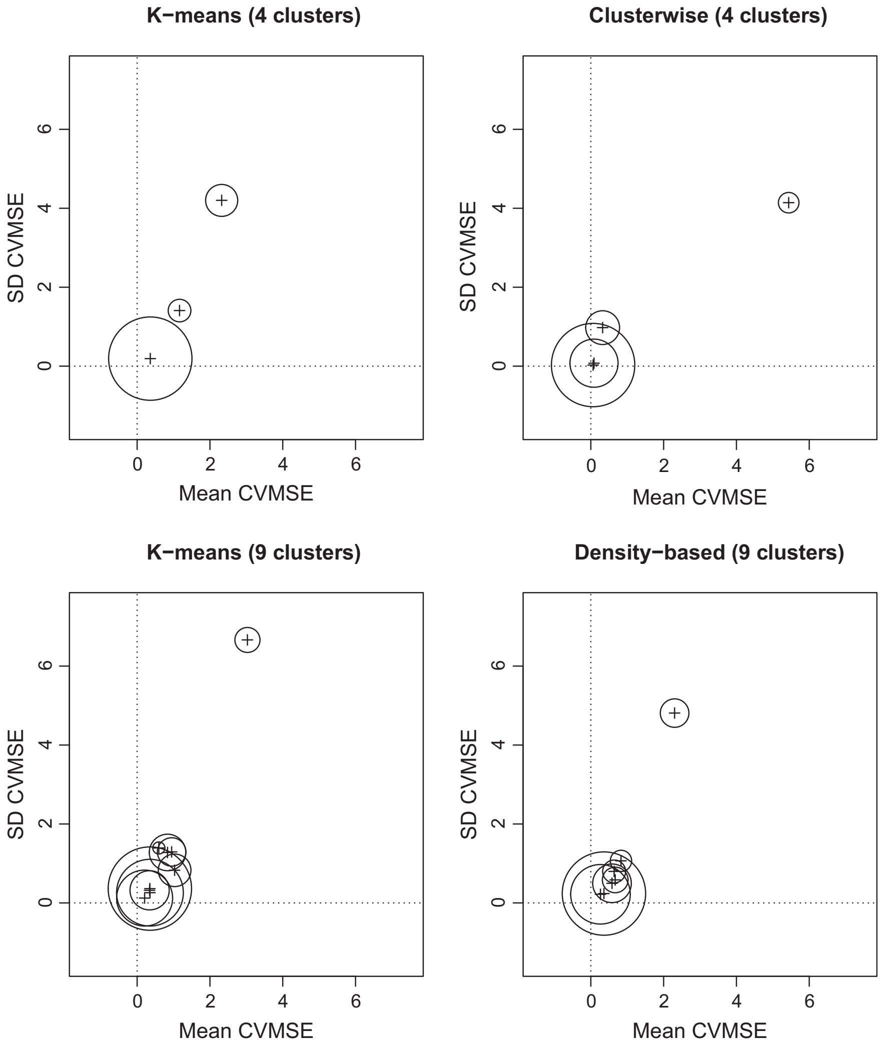

Fig. 3. Plot of prediction accuracies by size of cluster for each method (plot of Table 3). Horizontal axis is the mean cross-validated mean-squared error (CVMSE) from 20-fold cross-validation, the vertical axis is the standard deviation of CVMSE, and the size of the circles are proportional to the average number of observations placed in each cluster. Crosses indicate the center of each circle, and the dotted lines indicate zero CVMSE mean and standard deviation, respectively, so best predictions are as close to the intersection point (0, 0) as possible. One cluster in the K-means ($K = 4$) plot cannot be plotted because of large errors.

Table 4
Clusterwise regression coefficients for each cluster. As discussed in text, statistical significance cannot be reported because of the assumptions used to select the optimal mixture model fit.

Variable	Component			
	1	2	3	4
(Intercept)	0.03	−0.42	0.10	−0.04
Pct Electricity	−3.25	−4.00	0.04	−1.62
Pct District Steam	0.29	0.84	1.12	−0.18
Log Assessed Total	−0.36	−0.40	0.15	−0.06
Pct Yr. Built in 1980s	0.02	0.34	0.06	−0.02
Pct Household Heating Fuel Electric	−0.01	0.15	0.02	0.08
Pct Electricity × Pct District Steam	−0.25	−0.66	−1.06	0.20
Pct Electricity × Log Assessed Total	3.31	3.33	−0.37	1.03
Pct District Steam × Pct Yr. Built in 1980s	−0.00	0.12	−0.02	−0.00

The Jaccard coefficient can be interpreted [43] as representing a dissolved cluster for γ less than or equal to 0.5, a pattern between 0.6 and 0.75, a valid cluster between 0.75 and 0.85, and "highly stable" when above 0.85. The graphic shows that only K-means with $K = 4$ results in four highly stable clusters across all three perturbation methods. All of the other methods have relatively unstable Jaccard coefficients for most of their clusters, and there is no discernible pattern of stability.

5.3. Limitations of the analysis

Although the focus of this paper was largely methodological, assessment of prediction accuracy and cluster stability was carried out on a single large dataset from a particular city, so any conclusions about the effectiveness of these methods may be hard to generalize to other dissimilar data. Furthermore, this paper compares methods that sometimes generate different number of clusters depending on sampling, so the comparisons of predictive accuracy and cluster stability may not be valid where sampling results in very different clusterings. Finally, this analysis explains the difficulties in using clustering as a pre-processing step for predictive analysis, but as many scholars have noted, clusterings can have many other uses and interpretations.

5.4. Policy implications

In summary, clusterwise regression gives extremely accurate predictions for a large section of the population but relatively

Table 5
Jaccard coefficient as a measure of cluster stability. Grouped by method and number of clusters in descending order of cluster size.

Cluster method	Sampling method	Cluster number (K)								
		1	2	3	4	5	6	7	8	9
K-means (M1, K = 4)	Bootstrap	0.98	0.90	0.93	0.93					
	Subset	0.99	1.00	0.97	0.99					
	Noise	0.99	0.97	0.97	0.97					
Clusterwise (M4)	Bootstrap	0.20	0.85	0.30	0.65					
	Subset	0.16	0.81	0.18	0.58					
	Noise	0.16	0.65	0.27	0.55					
K-means (M1, K = 9)	Bootstrap	0.89	0.98	0.78	0.22	0.61	0.54	0.73	0.59	0.81
	Subset	0.84	0.98	0.72	0.24	0.59	0.52	0.66	0.56	0.78
	Noise	0.81	0.99	0.83	0.11	0.58	0.61	0.61	0.58	0.78
Density (M2, M3)	Bootstrap	0.19	0.16	0.55	0.82	0.62	0.54	0.38	1.00	
	Subset	0.25	0.10	0.52	0.79	0.61	0.41	0.33	0.99	
	Noise	0.09	0.12	0.38	0.73	0.51	0.43	0.23	0.97	

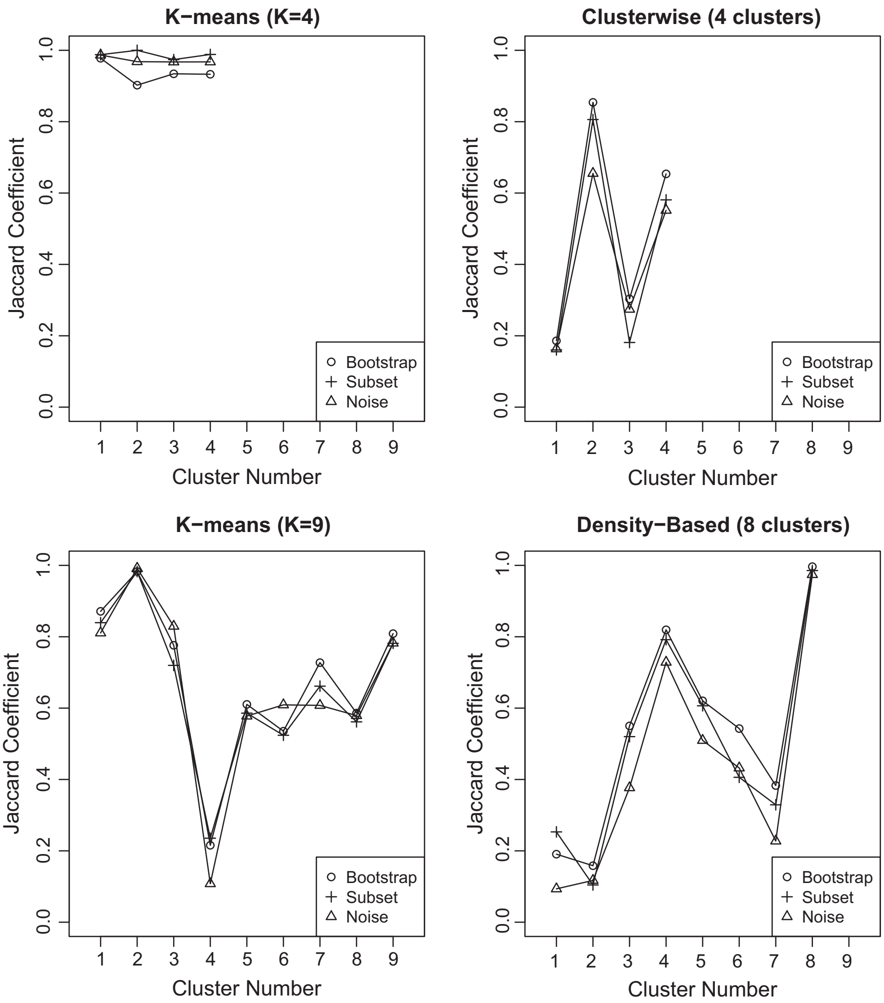

Fig. 4. Plot of cluster stability by method (plot of Table 5, ordered as in Table 3). The stability of the clusters is measured by the Jaccard coefficient with three different perturbation methods: bootstrapping, subsetting, and noise. Ideally, the measures would be consistently high for all three methods as represented by lines. Reading from top left clockwise: K-means with K = 4 is very stable with respect to method and for all clusters; clusterwise regression is relatively stable with respect to method but not cluster; K-means with K = 7 is stable with all methods but stability varies by cluster; and model-based clustering seems to give relatively unstable clusters, with stability fluctuating significantly by both cluster and perturbation method.

unstable clusters, K-means gives more stable clusters but relatively poor predictions, and model-based clustering does neither poorly or spectacularly. These results are useful for different policy scenarios in particular ways. For building analysis, where accurate prediction of energy consumption is an important goal, clusterwise regression can be used to group relatively similar buildings for comparison. Similarly, benchmarking policies require accurate measurement of buildings relative to one another, so it may be enough to use the majority of the data and to discard clear outliers. However, for energy efficiency programs that may require buildings to be consistently targeted on a year-over-year basis, it may be better to use *K*-means clustering, since these clusterings will remain relatively the same as long as the data retains the same shape or structure going forward.

The results of this paper indicate that there seems to be a fundamental tradeoff between the prediction accuracy exhibited by clusterwise regression and the cluster stability of *K*-means clustering. This may be particular to the structure of this data and how well particular strategies capture existing clusters in the data, but this also seems similar to the tradeoff between model flexibility and more accurate predictions. While it is possible in *some circumstances* to find methods that achieve an optimum balance between prediction accuracy and cluster stability – these results indicate that a fundamental tradeoff still appears to exist between these two goals. The necessity to choose between these two goals is an important methodological insight for policymakers and program designers, since some policies need to be consistently applied and some policies need to provide accurate prediction in order to remain credible.

6. Conclusion

This paper presented some of the clustering methods that are used in the building energy consumption literature. Clusterwise regression and cluster validation methods were introduced to the literature on modeling building energy consumption in order to address the separate issues of predictive accuracy and cluster stability. Clusterwise regression was then compared to the commonly-used two-stage processes of *K*-means and model-based clustering as applied to a large dataset of New York City multifamily buildings. Measures of prediction accuracy indicate that clusterwise regression gives extremely accurate predictions but somewhat unstable clusters, while *K*-means gives more stable clusters but very poor predictions in some clusters, which may not be identifiable at the outset of the analysis. Clustering methods should be chosen appropriately for particular use cases depending on the conflicting goals of prediction accuracy and cluster stability.

Acknowledgements

Thank you to the editor and reviewers for their detailed comments which significantly clarified and strengthened the paper. Thanks to the City of New York Office of Long-Term Planning and Sustainability for providing energy disclosure data, to Hong Hu and Ting Meng for comments on initial drafts, and to Albert Han and Daniel Suh for assistance in data assembly and cleaning. Thanks to Bettina Grün, Christian Hennig, and Kurt Hornik for their assistance in implementing their software packages and interpreting the results. This research was funded by the U.S. Department of Energy through the Consortium for Building Energy Innovation, Grant No. DE-EE0004261. The funding agency had no role in the design of the study, including the analysis or interpretation of data, writing, editing, or the decision to submit for publication. All errors belong to the author alone.

References

[1] Pérez-Lombard L, Ortiz J, Pout C. A review on buildings energy consumption information. Energy Build 2008;40:394–8.
[2] Urge-Vorsatz D, Novikova A. Potentials and costs of carbon dioxide mitigation in the world's buildings. Energy Policy 2008;36:642–61.
[3] Jain AK. Data clustering: 50 years beyond *K*-means. Pattern Recogn Lett 2010;31:651–66.
[4] Baker KJ, Rylatt RM. Improving the prediction of UK domestic energy-demand using annual consumption-data. Appl Energy 2008;85:475–82.
[5] Baringo L, Conejo AJ. Correlated wind-power production and electric load scenarios for investment decisions. Appl Energy 2013;101:475–82.
[6] Kabalci E. Development of a feasibility prediction tool for solar power plant installation analyses. Appl Energy 2011;88:4078–86.
[7] Li R, Wang Z, Gu C, Li F, Wu H. A novel time-of-use tariff design based on Gaussian Mixture Model. Appl Energy 2015. http://dx.doi.org/10.1016/j.apenergy.2015.02.063.
[8] McLoughlin F, Duffy A, Conlon M. A clustering approach to domestic electricity load profile characterisation using smart metering data. Appl Energy 2015;141:190–9.
[9] Rhodes JD, Cole WJ, Upshaw CR, Edgar TF, Webber ME. Clustering analysis of residential electricity demand profiles. Appl Energy 2014;135:461–71.
[10] Theodoridou I, Papadopoulos AM, Hegger M. A typological classification of the Greek residential building stock. Energy Build 2011;43:2779–87.
[11] Salat S. Energy loads, CO_2 emissions and building stocks: morphologies, typologies, energy systems and behaviour. Build Res Inf 2009;37:598–609.
[12] Swan LG, Ugursal VI. Modeling of end-use energy consumption in the residential sector: a review of modeling techniques. Renew Sustain Energy Rev 2009;13:1819–35.
[13] Guerra Santin O. Behavioural Patterns and User Profiles related to energy consumption for heating. Energy Build 2011;43:2662–72.
[14] Olofsson T, Andersson S, Sjögren JU. Building energy parameter investigations based on multivariate analysis. Energy Build 2009;41:71–80.
[15] Räsänen T, Ruuskanen J, Kolehmainen M. Reducing energy consumption by using self-organizing maps to create more personalized electricity use information. Appl Energy 2008;85:830–40.
[16] Yu Z, Haghighat F, Fung BCM, Yoshino H. A decision tree method for building energy demand modeling. Energy Build 2010;42:1637–46.
[17] Tooke TR, Coops NC, Webster J. Predicting building ages from LiDAR data with random forests for building energy modeling. Energy Build 2014;68(Part A):603–10.
[18] Santamouris M, Mihalakakou G, Patargias P, Gaitani N, Sfakianaki K, Papaglastra M, et al. Using intelligent clustering techniques to classify the energy performance of school buildings. Energy Build 2007;39:45–51.
[19] Booth A, Choudhary R, Spiegelhalter D. Handling uncertainty in housing stock models. Build Environ 2012;48:35–47.
[20] Yu Z, Fung BCM, Haghighat F, Yoshino H, Morofsky E. A systematic procedure to study the influence of occupant behavior on building energy consumption. Energy Build 2011;43:1409–17.
[21] Petcharat S, Chungpaibulpatana S, Rakkwamsuk P. Assessment of potential energy saving using cluster analysis: a case study of lighting systems in buildings. Energy Build 2012;52:145–52.
[22] Seem JE. Pattern recognition algorithm for determining days of the week with similar energy consumption profiles. Energy Build 2005;37:127–39.
[23] Domínguez-Muñoz F, Cejudo-López JM, Carrillo-Andrés A, Gallardo-Salazar M. Selection of typical demand days for CHP optimization. Energy Build 2011;43:3036–43.
[24] Kiluk S. Algorithmic acquisition of diagnostic patterns in district heating billing system. Appl Energy 2012;91:146–55.
[25] Lam JC, Wan KKW, Cheung KL. An analysis of climatic influences on chiller plant electricity consumption. Appl Energy 2009;86:933–40.
[26] Räsänen T, Voukantsis D, Niska H, Karatzas K, Kolehmainen M. Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data. Appl Energy 2010;87:3538–45.
[27] Famuyibo AA, Duffy A, Strachan P. Developing archetypes for domestic dwellings – an Irish case study. Energy Build 2012;50:150–7.
[28] Yu FW, Chan KT. Using cluster and multivariate analyses to appraise the operating performance of a chiller system serving an institutional building. Energy Build 2012;44:104–13.
[29] Gaitani N, Lehmann C, Santamouris M, Mihalakakou G, Patargias P. Using principal component and cluster analysis in the heating evaluation of the school building sector. Appl Energy 2010;87:2079–86.
[30] Steinley D, Brusco MJ. Selection of variables in cluster analysis: an empirical comparison of eight procedures. Psychometrika 2008;73:125–44.
[31] von Luxburg U, Williamson B, Guyon I. Clustering: science or art? J Mach Learn Res 2012;27:65–80.
[32] Hennig C. Cluster-wise assessment of cluster stability. Comput Statist Data Anal 2007;52:258–71.
[33] Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. J Am Statist Assoc 2002;97:611–31.
[34] Späth H. Algorithm 39 clusterwise linear regression. Computing 1979;22:367–73.
[35] Späth H. A fast algorithm for clusterwise linear regression. Computing 1982;29:175–81.
[36] Späth H. Mathematical algorithms for linear regression. Academic Press Professional, Inc.; 1992. <http://dl.acm.org/citation.cfm?id=SERIES9042.128660>.
[37] DeSarbo WS, Cron WL. A maximum likelihood methodology for clusterwise linear regression. J Classif 1988;5:249–82.
[38] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Statist Soc Ser B (Methodological) 1977;39:1–38.
[39] Leisch F. FlexMix: a general framework for finite mixture models and latent class regression in R; 2003. <http://epub.wu.ac.at/712/>.
[40] Grün B, Leisch F. others, FlexMix version 2: finite mixtures with concomitant variables and varying and constant parameters. J Stat Softw 2008;28:1–35.
[41] Banfield J, Raftery A. Model-based Gaussian and non-Gaussian clustering. Biometrics 1993:803–21.

[42] Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. J R Statist Soc: Ser B (Stat Meth) 2001;63: 411–23.

[43] Hennig C. Dissolution point and isolation robustness: robustness criteria for general cluster analysis methods. J Multivar Anal 2008;99:1154–76.

[44] Brusco MJ, Cradit JD, Steinley D, Fox GL. Cautionary remarks on the use of clusterwise regression. Multivar Behav Res 2008;43:29–49.

[45] Harrell F. Regression modeling strategies. Springer-Verlag New York Inc; 2001.

[46] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. Springer; 2009.

[47] Hsu D. Identifying key variables and interactions in statistical models of building energy consumption using regularization. Energy 2015;83:144–55.

[48] Fraley C, Raftery A, Murphy TB, Scrucca L. MCLUST Version 4 for R: normal mixture modeling and model-based clustering. Technical Report 597, Department of Statistics, University of Washington; June 2012. <http://www.stat.washington.edu/research/reports/2006/tr504.pdf>.

[49] Hornik K. A CLUE for CLUster ensembles. J Statist Softw 14.

[50] {R Development Core Team}. R: a language and environment for statistical computing; 2013. <http://www.R-project.org>.

[51] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Statist Softw 2010;33:1–22.

[52] {City of New York}, New York City local law 84 benchmarking report. Tech. rep., City of New York Office of Long-Term Planning and Sustainability; September 2013. <http://on.nyc.gov/Mi5w7K>.