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/

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,