CEE322: Data Analytics for Urban Systems

Lab 3: PCA and Clustering

Author: Davyd Tamrazov

Due date: Oct 12, 2020

Stanford University

1 Introduction

The following report aims to use electricity consumption data collected from smart meters in the buildings to determine specific consumption patterns and trends. In particular, a yearly time-series data for one building is used to identify daily electricity load shapes that would potentially enable better energy reduction recommendations. One of the questions this analysis aims to answer is how many clusters would provide an accurate and insightful outlook of the daily consumption throughout the year. Further, a second dataset containing summary of the electricity consumption for different end uses is utilized to perform principal component analysis. Such analysis allows for better understanding the dimensionality of the electricity consumption and dependencies between different end uses. As a primary goal, it aims to answer how many components can describe the majority of variance in the data.

2 Dataset

2.1 Dataset 1: Electricity Consumption Time-series

This dataset contains a time-series data with 35040 entries, each corresponding to the raw electricity consumption in kWh recorded at 15 minute interval over a year. Date or year of the recordings is not provided, meaning that potentially useful information about the day of the week for each entry is unknown. This time-series data is summarized in three ways in Table 1. Yearly consumption row describes data characteristics across all data points (35040 records); while daily total and peak consumption characteristics correspond to the data summarized over each day (365 records). No pre-processing was required for this dataset, as data appears to be cleaned by the distributor.

Descriptor	*Mean*	*St. Dev.*	*Min*	*Median*	*Max*
Yearly electricity consumption, kWh	0.203	0.309	0.000	0.090	4.260
Daily total electricity consumption, kWh	19.49	12.82	1.970	18.33	84.00
Daily peak electricity consumption, kWh	1.215	0.787	0.040	1.250	4.260

Table 1: *Summary of the electricity consumption time-series data*

2.2 Dataset 2: Consumption of Electricity by End Use

The second dataset provided has 5215 entries, each summarizing the consumption of electricity for a specific building with 26 features. Identification features include descriptors of the building location, footage, and utility uses. Further, electricity consumption is subdivided into a number different end uses. This analysis focuses only on a subset of 9 electricity consumption end uses for each building, as summarized in Table 2.

Descriptor	*Mean*	*St. Dev.*	*Min*	*Median*	*Max*
Electric heating use, 10^3Btu	2.46E+05	1.38E+06	0.00E+00	0.00E+00	3.37E+07
Electric cooling use, 10^3Btu	1.01E+06	4.30E+06	0.00E+00	5.39E+04	1.41E+08
Electric water heating use, 10^3Btu	1.27E+05	6.68E+05	0.00E+00	0.00E+00	1.23E+07
Electric lighting use, 10^3Btu	2.78E+06	8.02E+06	2.40E+01	2.59E+05	1.21E+08
Electric cooking use, 10^3Btu	2.77E+04	1.20E+05	0.00E+00	0.00E+00	3.02E+06
Electric refrigeration use, 10^3Btu	2.80E+05	1.09E+06	0.00E+00	4.29E+04	4.43E+07
Electric office equipment use, 10^3Btu	1.42E+05	1.00E+06	0.00E+00	7.72E+03	4.69E+07
Electric computer use, 10^3Btu	3.83E+05	1.68E+06	0.00E+00	1.45E+04	4.44E+07
Electric miscellaneous use, 10^3Btu	8.29E+05	2.38E+06	9.90E+01	9.09E+04	5.96E+07

Table 2: *Summary of the considered electricity consumption end use categories*

As a part of the pre-processing, it was identified that some of the buildings did not have any data available for the specific categories of the electricity end use. Thus, entries containing NaN values were removed from the dataset, resulting in the loss of 2.1% of the observations.

3 Analysis

3.1 Daily Electricity Consumption Clustering

To start with, from Table 1 it may be observed that the total and peak daily consumptions vary significantly. Thus, it is important to identify a set of the typical daily consumption shapes that would reduce this variation and accurately summarize yearly electricity use of the building. To achieve this, several experiments were run with K-means algorithm to cluster the time-series data, that was reshaped into 365 observations each with 96 features corresponding to 15 min intervals throughout the day. It was found that due to the high dimensionality of the features, K-means algorithm requires a large number of iterations and initializations in order to converge to a stable and optimal solution. As such, for the purpose of this report, the clustering results were obtained using 10000 iterations and 1000 initializations.

Further, with the combination of the elbow test and visual assessment, it was determined that 6 load shapes result in a set of the representative shapes, balancing size of each cluster and interpretability of the results. These shapes are summarized in Figure 7. It is worth noting that normalization of the data was not necessary in this case since all of the data points have the same range, scale, and units. Nevertheless, by clustering data normalized by the total daily load, it was found that the obtained shapes provide much less insight into the daily consumption (see Appendix 4).

From Figure 7, it may be observed that the algorithm resulted in visibly distinct shapes, identifying days with unusual consumption as well as days with very particular trends. This plot illustrates that, for the considered household, more than two thirds of the days throughout the year have relatively low electricity consumption, as seen in clusters 3 and 4. On the other hand, cluster 5 consists of only 7 days and could be considered an outlier, as it is unique in a way that is contains days with the non-zero electricity consumption at night. Cluster 2 can be interpreted as having daily shapes with the distinctive double peak, in the morning and in the evening – a common electricity load shape across all households. Finally, clusters 1 and 6 represent significantly higher daily loads and combine some days where electricity consumption either increases or decreases throughout the day respectively. From the number of days in each of these clusters, it is clear that such high loads are not prevalent during the year.

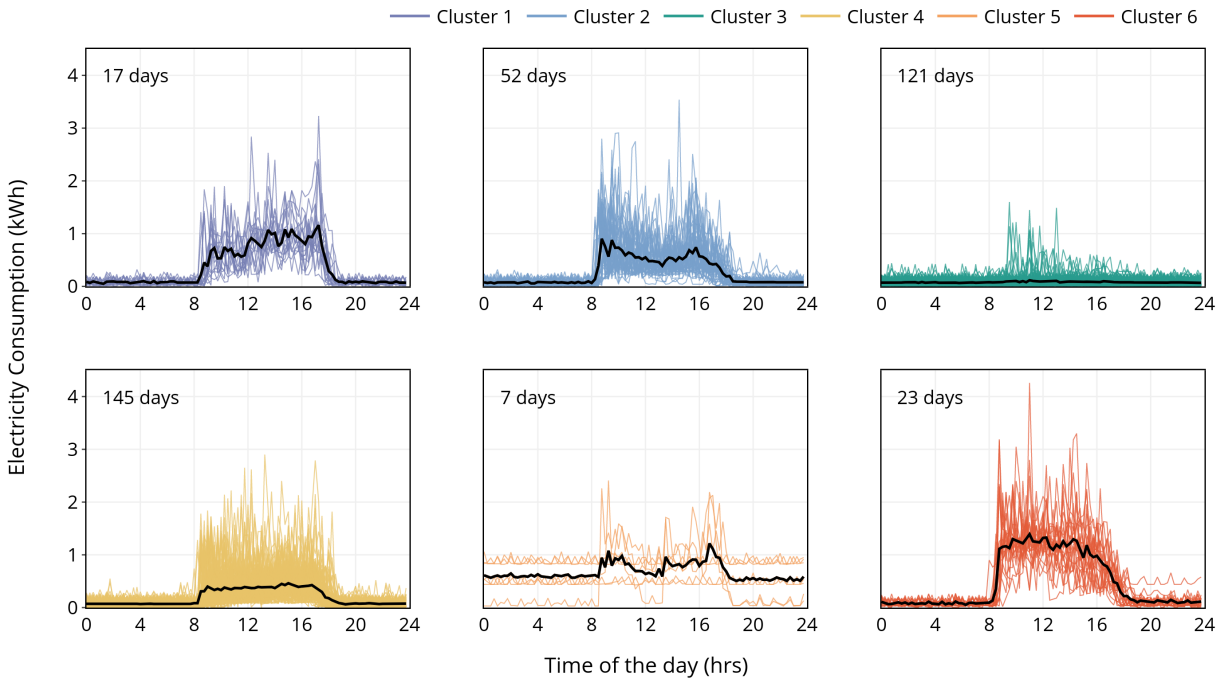

Figure 1: *Representative daily electricity consumption shapes obtained with K-means clustering*

In order to further understand these load shapes, a histogram was generated showing the number of days in the week contained in each cluster, as illustrated in Figure 2. Note that since no information was provided on actual days of the week corresponding to each measurement, each day was inferred by taking daily measurements at 7 day interval. From this plot it may be noticed that days 1,2, and 3 are primarily assigned to cluster 3; on the other hand, the rest of the days are concentrated in cluster 4 and less so in cluster 2. This allows for some interesting hypothesis to be raised, despite the lack of knowledge about the actual days represented. For example, it could be hypothesized that cluster 4 is the typical daily load shape throughout the year and that days 3 to 7, that are concentrated in this cluster, correspond to the electricity consumption during weekdays. From this it follows that days 1 and 2, that are grouped in cluster 3, could correspond to weekends and that this household is not typically used throughout the weekends with the load shape staying close to zero during the day.

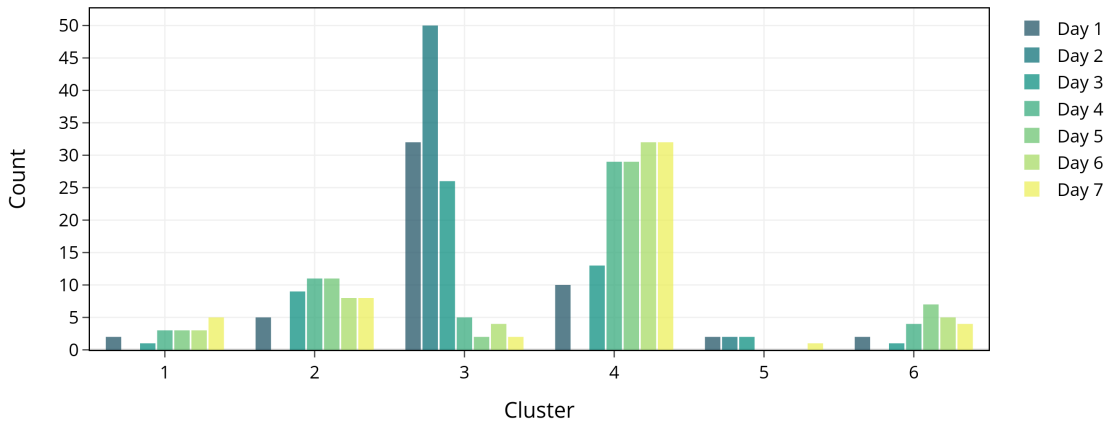

Figure 2: *Distribution of days of the week in each cluster*

Another way to interpret the clustering results is by looking at the distribution of clusters throughout the year, as shown on Figure 3. Assuming that the recording started with the first day of the year, this graph reveals that clusters 1, 5, and 6, previously identified as having higher daily loads, only tend to appear during the first 80 days and last 70 days of the time-series. This very nicely aligns with an assumption that those days correspond to the winter season, meaning that higher loads in these clusters could be justified by higher heating requirements during colder weather. Conversely, clusters 3 and 4 are prevalent throughout the rest of the year, what could indicate warmer temperatures in the location of the household.

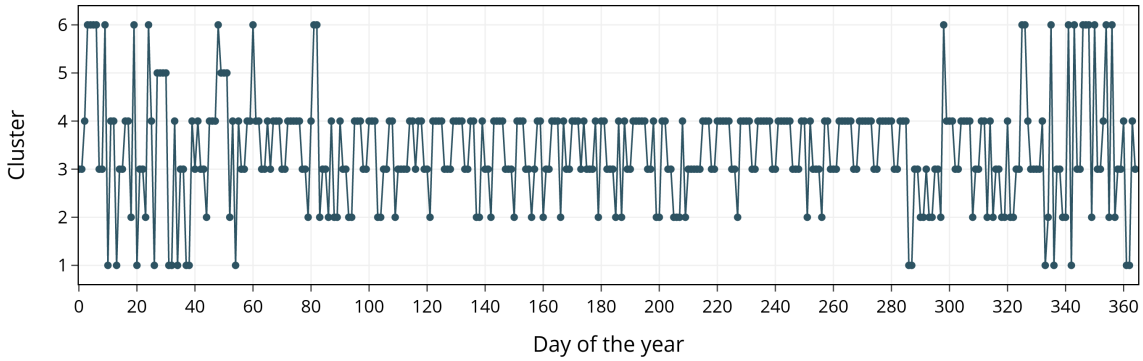

Figure 3: *Appearance of clusters throughout the year*

3.2 Principal Component Analysis

Looking at a more broader picture of electricity consumption across multiple households, it is important to understand the dimensionality of the problem. As such, principal component analysis (PCA) was performed on the second dataset including variables described in Table 2. From this table, it may be noticed that, although each feature is of the same unit, there is significant difference in magnitude, varying by up to a factor of 100. As a result, in order to remove biases introduced by scale difference, each of the features was standardized, meaning data was transformed in a way that mean and variance is 0 and 1 respectively. The resulting weights for each component signify the correlation of the principal component with a particular feature, what is summarized in a form of a

color-coded in Figure 4. Meanwhile, proportion of variance explained by each principal component is shown in Figure 5.

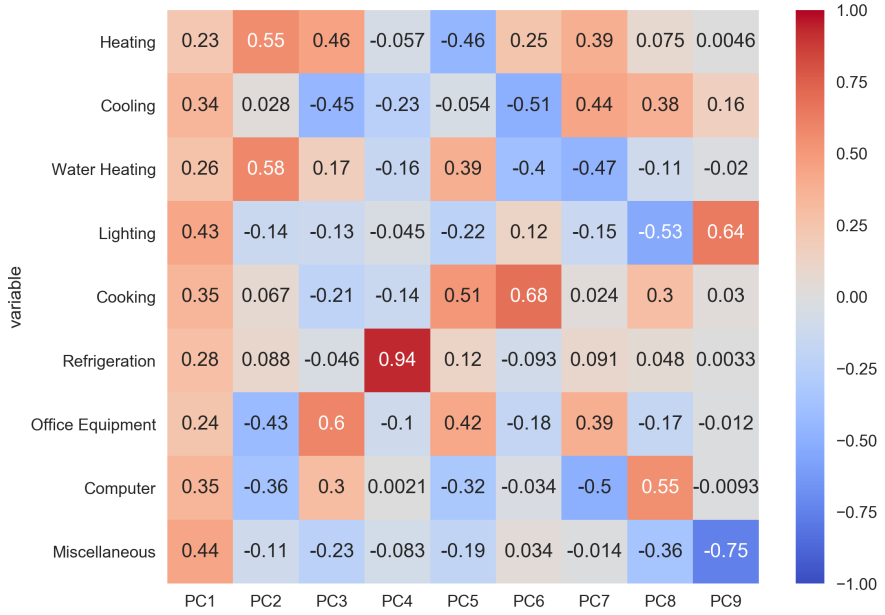

Figure 4: *Matrix showing weights for each principal component*

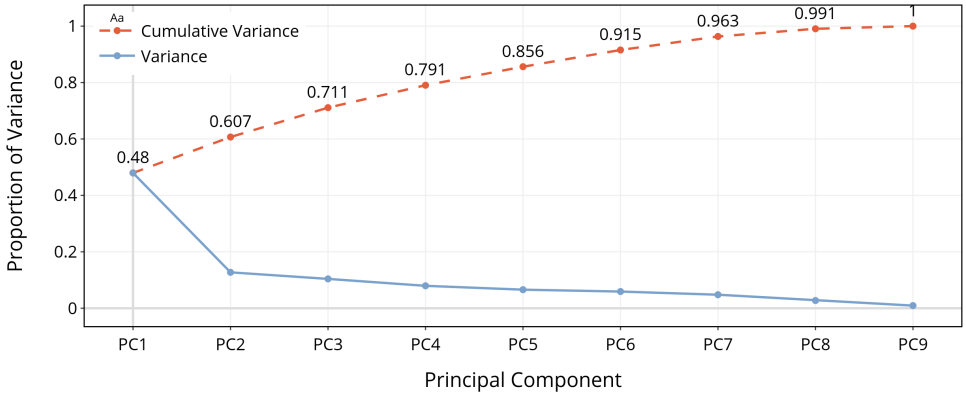

Figure 5: *Proportion of variance explained by each principal component*

The above plot shows that the first principal component (PC1) explains 48% of variance, while the matrix identifies that all of the electricity end uses contribute to that. Additionally, the correlations are positive and very similar what means that all features are positively correlated with PC1. This can be explained by the hypothesis that larger buildings tend to have larger electricity consumption in all of the categories, implying that if one gets larger, the rest follow the trend. Thus, this principle component could be representative of the square footage of the building.

Next, the second principal component (PC2) varies with the higher electric heating and water heating uses as well as lower office equipment and computer uses. This could be viewed as a measure of how residential versus office is the building type. Office buildings tend to require more computer and office equipment loads, while heating and water heating is not as much of a concern

for the offices. The variance explained by the first two components is 60.7%, what is not sufficient to represent the data accurately. In order to achieve a more significant percentage of variance explained ($> 80\%$), five principal components are required. Furthermore, if one would want to explain almost all of the variance ($> 95\%$), seven principal components would be needed, underlining that the electricity use is a highly multi-dimensional problem.

Additionally, the above findings can be corroborated with the bi-plot, shown in Figure 6. Firstly, from the scatter plot on the right, it is evident that a considerable portion of variance is explained by the two components; however, further dispersion can be observed to happen along diagonal axes. Moreover, looking at the left plot, dependencies between features can be inferred. As such, it is clear that heating and water heating are highly associated, what could be potentially explained by a hypothesis that household heating and water heating requirements are equally correlated with colder weathers. Similarly, refrigeration, cooking, and cooling form another cluster of highly correlated features. It could be hypothesized that larger households tend to have higher loads in all three categories. Using the same logic, association between lighting and miscellaneous electricity uses could be justified. Finally, the association between office equipment and computer is self-evident as larger offices tend to have proportionately larger uses in both categories.

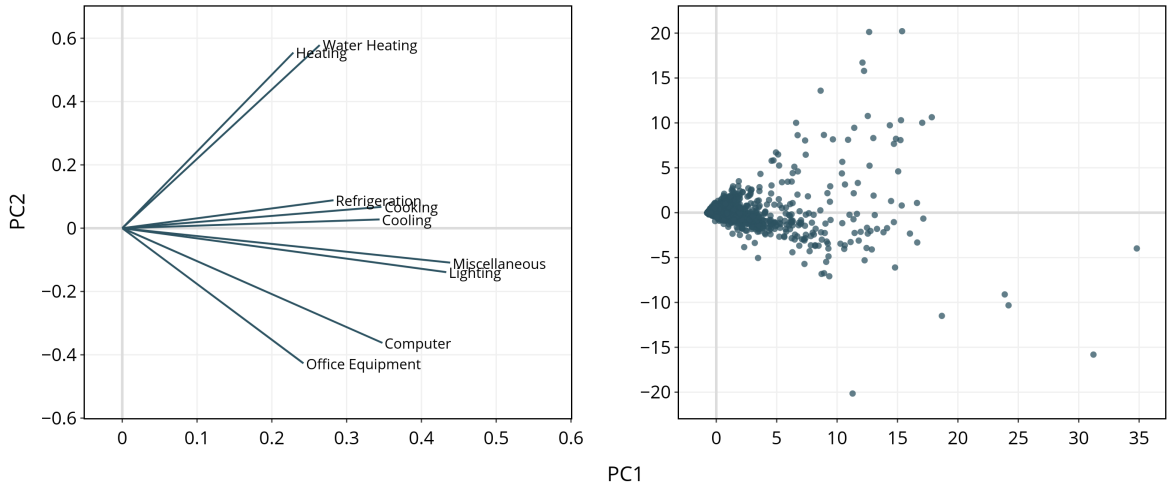

Figure 6: *PCA bi-plot showing transformed data points on PC1 and PC2 axes*

4 Conclusion

To summarize, K-means clustering was used to identify six representative daily electricity consumption load shapes. Each was further interpreted by looking at the dependency of the load shape on the day of the week as well as time of the year. Further, principal component analysis was performed identifying that five principal components are required to explain at least 80% of the variance. The weights of the first two components were explained in detail, while additional dependencies between variables were identified and interpreted with the bi-plot. Overall, it was concluded that electricity consumption of the building is a highly dimensional problem and it is crucial not to overlook the importance of each variable during dimensionality reduction.

Appendix 1

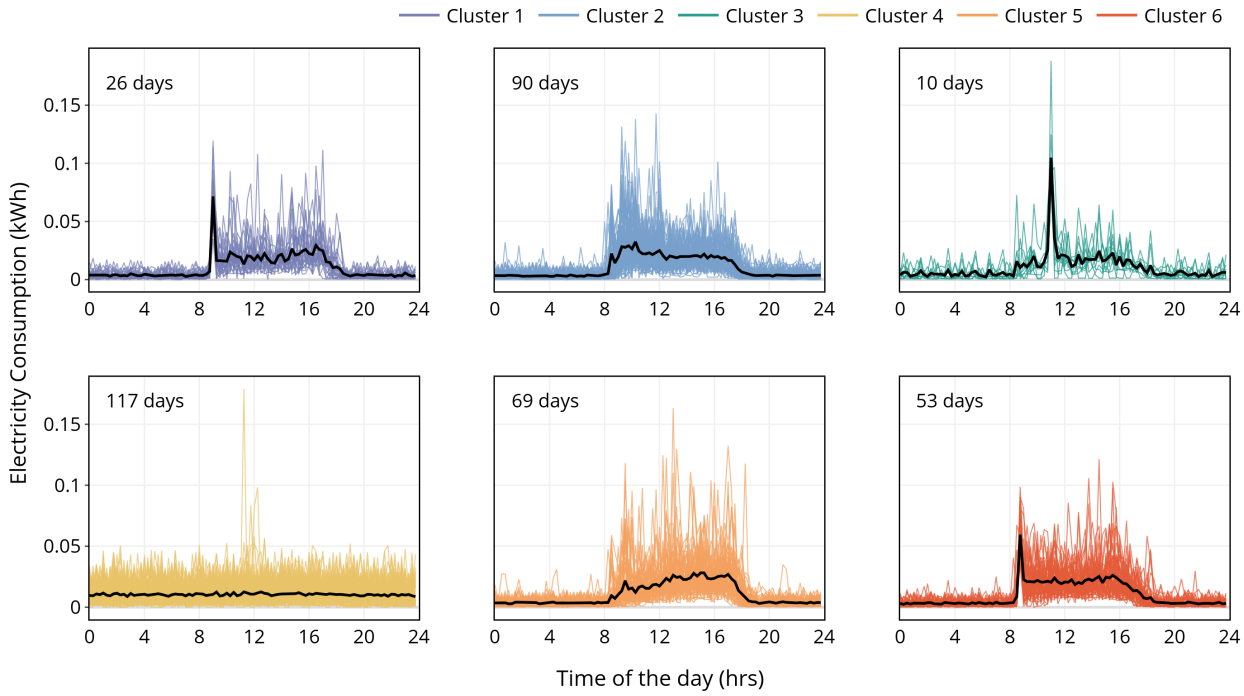

Figure 7: *Representative daily electricity consumption shapes obtained with K-means clustering on normalized data*