# K-Shape clustering analysis on buildings energy consumption collected by GAIA platform

Leonisio Schepis 1533794

July 2 2018

A report submitted in fulfillment of the requirements for the class of Seminars in Advanced Topic in Computer Science of the Master of Science in Engineering in Computer Science of Sapienza University of Rome.

# 1 Introduction

The aim of the project is the evaluation of the K-Shape technique, as clustering algorithm, for the data about the buildings energy consumption gathered by the GAIA platform. The analysis is made on data collected from May 31 2017 to May 31 2018. The project is divided in two parts:

1. Clustering buildings among each others based on the yearly consumption of energy.
2. Clustering daily consumption routines of the building with the most complete collection of data during the proposed time interval.

In this report is described the whole decision-making process and therefore, no mid results, or plots, are provided.

# 2 First Task: Clustering of buildings

Data collected from sensors are easily subject to noises. These noises can be due to sensor failures in which it can retrieve very unlikely measures, or periods in which the sensor are off for any reason. It is clear that all these factors are not helping the analysis of the data. Therefore, the first analysis performed during the project was understanding if the K-Shape technique is, in some way, influenced by these noises. Actually, the very first concern was about the period in which sensors are shut down: since measurements during one year contain a large number of zero-measures, the technique could cluster together buildings sharing zero-periods. Actually, this is not true because, buildings with very large zero-periods were clustered together with those with smaller zero-period: it is like the zero-periods are less-weighted wrt nonzero-periods. In order to evaluate the clustering the used score function are:

1. Silhouette Index[1].
2. Dunn index[2].
3. Davies-Boulding index[3].
4. Calinski-Harabasz index[4].

## 2.1 Cleaning

Even if it seems that noises does not influence badly the clustering, trying to improve the quality of the dataset could not provide worse results and therefore three kinds of improvements are proposed:

1. Remove all timestamps in which we have a zero in a majority of the buildings.

2. After the first step, remove buildings with almost only zero-values, if there exists.

3. Compute median and max, cutting all values which are too far from the median, which means that probably they are faulted measures.

"Too far" means that the measure is larger than the average between the median and the maximum. However, if the maximum is, in some way, near the median there will be a very small loss of information. Instead, if the maximum is very far from the mean (probably faulted), this policy will avoid its bad influence during the clustering. The median could perform better than the mean, because the mean is too sensitive to very large terms such as faulted measures. After the dataset improvements the quality scores of the clusters got some enhancements.

### 2.2 Final Analysis

Summarizing, the number of identified clusters is 3. The K-shape implementation [5] used during the project, even if you want to cluster timeseries in 4 clusters, can retrieve empty clusters. In my opinion, this is some kind of suggestion: perhaps the number of clusters can be reduced. So, after many trial-and-errors the best clustering is achieved with k = 3.

## 3 Second Task: Clustering of daily routines

During this task two main problems came up:

1. Since we have a large dataset and no idea about how many clusters we can have inside it, we should compute something supporting this decision.

2. Days with all zero-values are meaningless and therefore they can be removed.

To solve the first problem the idea is to use a method based on the Elbow Method[6]. Therefore, all clusters from k = 2 to k = 300 are computed plotting the related MSEs, but the MSEs, despite the expectations, have not a monotone decreasing behaviour which means that is not easy to apply the Elbow method. Taking into account all the local minima in the neighboorhood at most at distance 2 the expected decreasing behaviour can be approximated. Therefore, from a visual analysis of the plot the suitable candidate number of clusters can be chosen.

### 3.1 Final Analysis

The number of clusters that came up for building "144242" during the project is 16. The k-shape algorithm left one of those empty and therefore the actual number of clusters is 15. Plotting the clusters many of them are, perhaps, meaningless or at least their meaning is not so clear from the plot. But what can be noticed is that there are at least 7-8 clearly distinct clusters among them.

## References


[1] P. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20(1987) 5365.

[2] J.C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, J. Cybern. 3 (1973) 3257

[3] D.L. Davies, D.W. Bouldin, A clustering separation measure, IEEE Trans. Pattern Anal. Mach. Intell. 1 (1979) 224227.

[4] T. Calinski, J. Harabasz, A dendrite method for cluster analysis, Commun. Stat. 3 (1974) 127.

[5] Mic92 https://github.com/Mic92/kshape

[6] Junjing Yang, et al., "k-Shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement", Energy and Buildings, Volume 146, 2017, Pages 27-37.