Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 11, n. 1, 2025.

Trabalho apresentado no XLIII CNMAC, Centro de Convenções do Armação Resort - Porto de Galinhas - PE, 2024

**Proceeding Series of the Brazilian Society of Computational and Applied Mathematics**

# Clustering Energy Consumption Time Series Using Singular Spectrum Analysis and K-means Clustering

Hans Rolan E. Mersch Fernandez [1], Diego H. Stalder[2], Carlos Sauer[3], Félix Morales[4]
Facultad de Ingeniería, Universidad Nacional de Asunción, Paraguay

The widespread adoption of Smart Metering Systems (SMS) by electrical suppliers has revolutionized energy consumption data collection. SMS not only enables capacity measurement but also unlocks valuable insights from the vast amount of data collected at short intervals from residences and industries. This detailed data empowers energy companies to gain a deeper understanding of their customers' energy consumption behavior. One crucial task within this field is clustering households based on their load demand profiles. This clustering, achieved through smart meter data, allows for the identification of groups of users with similar consumption characteristics. These user groups can then be used to generate representative demand profiles that closely approximate the demand for specific regions or even the entire system. While numerous clustering methods exist, establishing the validity and accuracy of the results obtained through these methods remains an important area of research [1].

This work analyzes residential energy consumption data from a 2021 pilot experiment by Paraguay's National Energy Administration (ANDE). The data, collected via Smart Metering Systems (SMS) from 122 households across various regions for one month, suffers from inconsistencies due to unsynchronized SMS time zones. This significantly impacts the segmentation and characterization of customer consumption curves, as observed in preliminary attempts using Principal Component Analysis (PCA) and K-means for feature extraction and clustering, respectively. Both algorithms rely on accurate time-based data, as demonstrated in Figure 1.

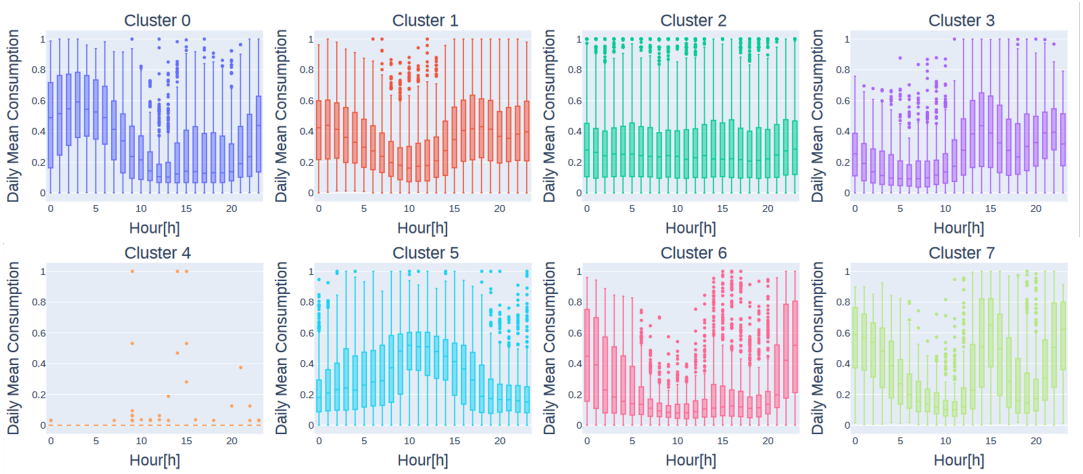


Figure 1: Daily mean consumption profiles for metropolitan residential clients (normalized). Source: From the authors

We propose Singular Spectrum Analysis (SSA) to address the time series inconsistencies caused by unsynchronized SMS data. SSA excels at decomposing time series into interpretable components (trend, periodicity, noise) even with significant noise. This allows us to extract robust features for subsequent clustering, overcoming the limitations of raw time series data.

[1] hmersch@fiuna.edu.py
[2] dstalder@ing.una.py
[3] csauer@ing.una.py
[4] felixmorales@fiuna.edu.py

Let $\{x_t\}_{t=1}^{T}$ be a time series of length $T$, and let $X$ be the consumption profile obtained by embedding the time series into a matrix:

$$X = \begin{bmatrix} x_1 & x_2 & \cdots & x_L \\ x_2 & x_3 & \cdots & x_{L+1} \\ \vdots & \vdots & \ddots & \vdots \\ x_K & x_{K+1} & \cdots & x_T \end{bmatrix} \tag{1}$$

where $L$ is the window length, $K = T - L + 1$, and $T$ is the length of the time series. Singular Spectrum Analysis consists of the following steps: (i) Decomposition: Compute the singular value decomposition (SVD) of $X = U\Sigma V^T$; (ii) Reconstruction: Form the components of the decomposition: $\{C_i\}_{i=1}^{L} = \{u_i \sigma_i v_i^T\}_{i=1}^{L}$; (iii) Grouping: Group the components into sets $\{G_j\}_{j=1}^{M}$, where each set represents a particular type of signal (e.g., trend, periodicity, noise); (iv) Reconstruction: Reconstruct the original time series using selected sets of components: $\hat{x}_t = \sum_{j=1}^{M} G_j(t)$. The resulting reconstructed time series $\{\hat{x}_t\}_{t=1}^{T}$ provides insight into the underlying structures of the original time series.

For clustering the data, we employ the K-means algorithm, an iterative algorithm that assigns data points to the closest cluster centroid and updates centroids based on assigned point means. This process minimizes the within-cluster sum of squares, leading to well-defined clusters.

Gap statistics is used to determine the optimal number of clusters [3]. It compares the within-cluster dispersion observed in the data to an expected null distribution. The gap statistic for a given number of clusters $k$ can be defined as $\text{Gap}_n(k) = E_n^*\{\log(W_k)\} - \log(W_k)$, where $W_k$ is the sum of the squared distances of each point to the centroid of its cluster, and $E_n^*\{\log(W_k)\}$ is the expected value of $\log(W_k)$ under a reference sample from the null distribution. The $k$ with the highest gap statistic is considered optimal.

Clustering validation indices (CVIs), such as the silhouette and Calinski-Harabasz scores, are commonly used to evaluate clustering algorithms. In the case of Figure 1, these scores are 0.19 and 22.33, respectively. However, the variance among these indices can complicate evaluation. To enhance the stability and validation of clusters, clustering bootstrap is applied [2]. This involves randomly resampling and partitioning raw data for each household into $p$ partitions. This validation technique resulted in a 55.35% of matches for $p = 2$ and 47.18% for $p = 3$. Implementing SSA to reconstruct data and eliminate noise allows us to improve these metrics up to 66.29% for $p = 2$ and 57.95% for $p = 3$. This highlights the effectiveness of clustering bootstrap in validating cluster stability and coherence.

By identifying distinct consumer groups, this approach facilitates the development of targeted energy management strategies. These strategies can encompass optimized pricing models, improved demand forecasting, and infrastructure upgrades.

# References

[1] F. Morales et al. "Analysis of Electric Energy Consumption Profiles Using a Machine Learning Approach: A Paraguayan Case Study". In: **Electronics** 11.2 (2022), p. 267. DOI: 10.3390/electronics11020267.

[2] C. A. Field and A. H. Welsh. "Bootstrapping clustered data". In: **Journal of the Royal Statistical Society Series B: Statistical Methodology** 69.3 (2007), pp. 369–390. DOI: 10.1111/j.1467-9868.2007.00593.x.

[3] R. Tibshirani, T. Hastie, and G. Walther. "Estimating the number of clusters in a data set via the gap statistic". In: **Journal of the Royal Statistical Society: Series B (Statistical Methodology)** 63.2 (2001), pp. 411–423. DOI: 10.1111/1467-9868.00293.