Clustering Synthetic Time Series Using Dimensionality Reduction and K-Means

Universidad Politécnica de Madrid — Escuela Técnica Superior de Ingenieros Informáticos

Ottavia Biagi, Leandro Duarte, Emanuele Alberti
Máster Universitario en Inteligencia Artificial
Asignatura: Data Mining
Universidad Politécnica de Madrid (UPM)

***Abstract*—This report presents the application of classical data mining techniques to synthetic time series using the DMonTS tool. Two domains were considered: FAANG stock-like trajectories and City Hotel ADR (Average Daily Rate). Synthetic series were generated using configurations calibrated from real data. After normalization, dimensionality reduction was applied using Fourier coefficients and Piecewise Aggregate Approximation (PAA). Multiple K-Means clustering experiments (K=2,3,4) were performed to assess whether clusters reflect the original domains and internal seasonal structures.**

I. Introduction

This work implements a complete time-series clustering pipeline using the Herramienta de Data Mining para Series Temporales (DMonTS). Two distinct domains were analysed: (i) FAANG stock-like series and (ii) City Hotel ADR series. For each domain, 15 synthetic time series of length 100 were created with the Time Series Random Generator, using parameters calibrated from real data.

The City Hotel ADR dataset first required cleaning and aggregation before it could be used as a reference. The raw data contained booking-level records with heterogeneous attributes and several missing or extreme values. We restricted the analysis to non-cancelled bookings, removed anomalous ADR values (e.g., outliers above 800 EUR), converted dates into a monthly index and aggregated ADR by month. The resulting series exhibits a clear annual seasonality (period 12), a moderate upward trend and low variability, which we reproduced in the synthetic generator through a cosine periodic component, a trend term and controlled noise.

In contrast, the FAANG reference segment shows a strong downward trend, a pronounced mid-series valley and moderate volatility with no detectable periodicity. These characteristics were used to parameterise the decreasing– trend, valley–event and noise modules of the generator.

The overarching goal is to assess whether dimensionality-reduced representations, specifically Fourier coefficients and Piecewise Aggregate Approximation (PAA), allow K-Means clustering to correctly recover the domain structure and reveal internal seasonal patterns. The pipeline follows the main stages of the Knowledge Discovery Process studied in the course: domain understanding, data cleaning and preprocessing, dimensionality reduction, pattern discovery via clustering and evaluation.

II. Data Generation

Synthetic data were generated using domain-specific configurations derived from an extensive analysis of real datasets (FAANG closing prices and Hotel ADR). Real data inspection showed:

- FAANG: downward trend, mid-series valley event, moderate volatility.
- City Hotel ADR: strong seasonal periodicity (12-month cycle), slight upward trend.

These characteristics were encoded in the generator configurations.

A. FAANG Synthetic Series

Fig. 1 shows an example FAANG-like series created with: decreasing trend, valley recovery, and stochastic noise.

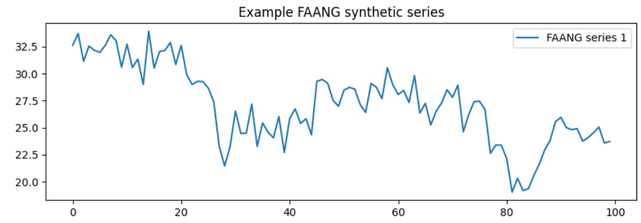

Fig. 1: Example FAANG synthetic series.

B. City Hotel ADR Synthetic Series

City hotel series include cosine periodicity with period 12 and low-variance noise.

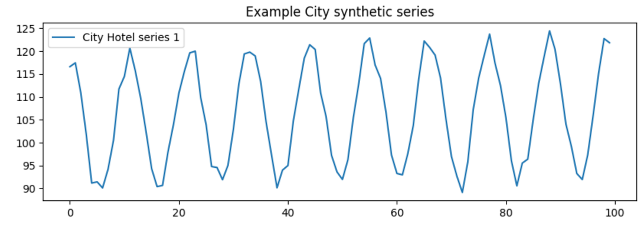

Fig. 2: Example City Hotel ADR synthetic series.

III. Methodology

All experiments followed the same DMonTS pipeline:

1) Load the combined CSV with all 30 series.
2) Apply normalization to $[0, 1]$.
3) Apply dimensionality reduction:
   - Fourier (8 or 16 coefficients)
   - PAA (8 or 16 segments)
4) Apply K-Means clustering with $K = 2, 3, 4$.
5) Compare cluster assignments with domain labels.

IV. Dimensionality Reduction

A. Fourier Transform

Fourier coefficients capture dominant frequency structures.

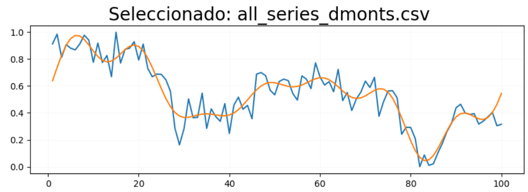

Fig. 3: Fourier representation with 8 coefficients.

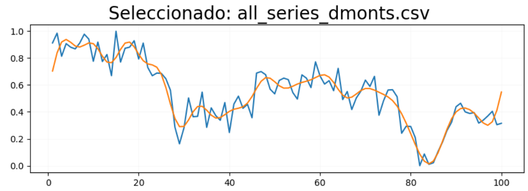

Fig. 4: Fourier representation with 16 coefficients.

B. Piecewise Aggregate Approximation (PAA)

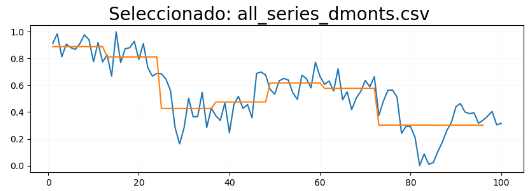

Fig. 5: PAA segmentation with 8 segments.

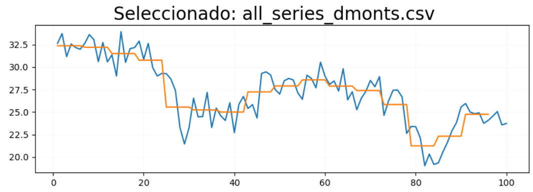

Fig. 6: PAA segmentation with 16 segments.

V. Clustering Experiments

A. Experiment 1 — K = 2 (Domain Separation)

This test checks whether clustering separates FAANG from Hotel ADR.

TABLE I: K=2 clustering (Fourier 8)

Series	Domain	Cluster
1–15	FAANG	1
16–30	Hotel	2

Result: perfect domain separation.

B. Experiment 2 — K = 3

TABLE II: Cluster distribution for K=3

Cluster	FAANG Count	Hotel Count
1	15	0
2	0	10
3	0	5

Hotel series split into high-season vs low-season subclusters.

C. Experiment 3 — K = 4

TABLE III: Cluster distribution for K=4

Cluster	FAANG Count	Hotel Count
1	15	0
2	0	7
3	0	3
4	0	5

Hotel ADR forms three seasonal clusters; FAANG remains homogeneous.

VI. Discussion

Fourier and PAA both successfully separate the two domains at $K = 2$, which confirms that even highly compressed representations preserve the key structural differences between FAANG and Hotel ADR time series. When K is increased, differences between the two representations and between the domains become more evident. Fourier captures long-range structure and highlights spectral differences associated with hotel seasonality, whereas PAA focuses on local averages and is slightly more sensitive to shape noise when $K = 4$. In all configurations, FAANG series remain in a single compact cluster because they share similar trend and volatility patterns, while Hotel ADR series naturally split into multiple clusters that correspond to different seasonal regimes.

Overall, these results show that the choice of dimensionality reduction does not affect domain-level separability, but it does influence how intra-domain variability is exposed. Fourier tends to emphasise global behaviour and smooth transitions between seasons, whereas PAA provides a more piecewise view of the signal, making it easier to isolate short-term changes but also to fragment the series when the number of clusters is large.

VII. Conclusion

The experiments demonstrate that synthetic time series from two distinct domains can be effectively clustered after dimensionality reduction using simple, interpretable techniques. With $K = 2$, all tested configurations (Fourier with 8 and 16 coefficients, PAA with 8 and 16 segments) recover the true domain labels with perfect accuracy, indicating that the global structure of the series is preserved even in very low-dimensional spaces. Among the tested representations, Fourier with 16 coefficients provided the most stable and discriminative behaviour across different values of K.

When K is increased to 3 and 4, the clustering structure remains consistent: FAANG series form a single homogeneous cluster, while Hotel ADR series split into meaningful seasonal subgroups. This suggests that the synthetic data generation process correctly encoded domain-specific properties and that the combination of normalisation, Fourier/PAA features and K-Means is sufficient to reveal both inter-domain and intra-domain patterns. Future work could extend this analysis to real, non-synthetic data and to alternative distance measures such as DTW or to other clustering algorithms, in order to assess how robust these findings are beyond the controlled synthetic setting.