Mitigation of Thermal Storage Sizing Errors in Thermal Aggregated Multi-Energy System Optimization

Lorenzo Organtini[a], Gabriele Leoncini[a], Mathieu Vallée[a] and Antoine Piguet[a]

[a] *Univ. Grenoble Alpes, CEA, Liten, Campus Ines, 73375 Le Bourget du Lac, France ,*
gabriele.leoncini@yahoo.it

Abstract:

The optimization of the design of Multi-Energy Systems (MES) relies on Mixed-Integer Linear Programming (MILP) models that are computationally intractable over long time horizons. To address this complexity, temporal aggregation techniques are used to reduce the time resolution. However, classical clustering techniques based on Euclidean distance struggle in high-dimensional spaces and are often unable to capture relevant temporal structures, leading to inaccurate storage sizing.

This work evaluates methods to more accurately capture meaningful temporal structures in the time series data. First, we evaluate the impact of Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP) to preserve the variance or the non-linear structure of the data respectively. Second, we modify the distance metric itself using Dynamic Time Warping (DTW) to account for temporal shifts.

The resulting typical periods are evaluated on a case study of the Cambridge district in Grenoble (France). They are compared to the baseline methods of Euclidean K-Means and K-Medoids. Next, they are evaluated using an aggregated optimization model, in which all operating variables are aggregated while storage dynamics remain modeled at full temporal resolution through a reconstruction mapping that links each original hour to its corresponding representative period.

Results indicate that the methods proposed offer a more robust dimensioning accuracy of the storage technologies, particularly for the Thermal Energy Storage (TES) component. When averaging the TES sizing error across all tested numbers of typical periods, K-Medoids using DTW or UMAP pre-processing exhibited the highest fidelity to the reference non-clustered model. DTW achieves a 49.5% reduction in TES sizing error relative to K-Means and a 40.9% reduction relative to K-Medoids, while UMAP achieves a 43.8% and 34.3% reduction, respectively. Meanwhile, K-Medoids using PCA pre-processing performs similarly to the baseline models, yielding only a 13.5% reduction relative to K-Means and a 1.2% increase relative to K-Medoids. These findings confirm that an accurate feature extraction can enhance the optimization quality.

Keywords:

Temporal Aggregation; Multi-Energy System Optimization; Storage Sizing.

1. Introduction

The climate targets set by the European Union aim for at least a 55% reduction in greenhouse gas emissions by 2030 compared to the 1990 levels, and “climate neutrality” by 2050 [1]. The design of local energy systems, which are the set of technologies supplying energy to end-users, is key in achieving reliable and cost-effective decarbonization. These infrastructures rely on several energy carriers, such as gas and biomass, to meet the demands of services of different nature, like electricity and heat. Conversion technologies link energy carriers by transforming one form of energy into another, enabling the system to exploit synergies across energy vectors and increase overall efficiency. Moreover, the storage of energy adds flexibility by allowing decision-makers to decouple production and demand, enabling financial strategies to reduce operating costs.

Renewable energy sources, such as photovoltaic panels, wind turbines, and biomass units, need to be integrated into Multi-Energy Systems (MES) to facilitate their decarbonation. However, since the production of weather-dependent renewable sources cannot be controlled (they are non-dispatchable), additional storage capacity must be installed to account for the extra flexibility required. Coupling multiple energy sources with storage systems enables the exploitation of synergies across technology, minimizing the total cost of the system. Sizing such infrastructure requires models that account for operations over long time horizons, typically spanning decades, in order to properly evaluate the investment decisions. At the same time, these operations are modeled at an hourly resolution, resulting in a very large number of time steps.

A common approach is Mixed-Integer Linear Programming (MILP), where decision variables are divided into sizing variables, to determine the capacities of the devices to install, and operating variables, to manage energy production [2] and storage operations. When modeled at hourly resolution over long time-horizons, the number of operating variables becomes large and, combined with binary variables, such as those used to prevent simultaneous charging and discharging of storage, the result is often a computationally intractable MILP. Temporal aggregation techniques are widely adopted to reduce the full high-resolution time series through a reduced set of typical periods, such as representative days or weeks. However, the transition from a full-scale model to a reduced (or aggregated) one introduces approximation errors [2-5], which impact the sizing of the MES. While classical clustering techniques based on Euclidean distance preserve information about total demand or duration curves, they often fail to capture more complex and non-linear temporal structures [2], such as shifted or distorted patterns. This particularly affects storage technologies, whose optimal sizing depends on inter-period dynamics and the representation of extreme behaviors.

The objective of this paper is to reduce the optimization error in Multi-Energy Systems thermal storage sizing by enhancing the feature extraction process of temporal aggregation, tested on synthetic data of the Cambridge district in Grenoble, France. First, dimensionality reduction techniques are introduced prior to clustering. Principal Component Analysis (PCA) is used to project the data onto a reduced space spanned by its principal directions, capturing the dominant patterns of variability in the dataset. Otherwise, Uniform Manifold Approximation and Projection (UMAP) is employed to retain non-linear relationships between periods. Second, the distance metric used for clustering is modified with Dynamic Time Warping (DTW), allowing temporal shifts within daily profiles to be accounted. These approaches are compared to classical Euclidean K-Means and K-Medoids clustering.

1.1. State of the art

Standard clustering algorithms that operate directly in the raw feature space are well explored in the literature. The most widely adopted are K-Means and K-Medoids, both partitional algorithms that minimize the sum of squared distance between candidate periods and their cluster representatives [2]. K-Means represents each cluster by its centroid, which smooths out fluctuations in the reconstructed time series, whereas K-Medoids selects actual periods from the dataset, preserving realistic daily variations [3]. One of the earliest applications is by Domínguez-Muñoz et al. [6], who used K-Medoids to determine typical demand days for cogeneration systems.

Prior to clustering, preprocessing the input time series is a necessary step that has been addressed in various ways across the literature [2]. The full time series is first divided into candidate periods of equal length, typically days, which are normalized to account for the difference in magnitude of the different input time series that compose each period (referred to as attributes). The min-max normalization is the most frequently adopted choice [2,3], but Z-normalization has been used for data following approximately normal distributions with different spreads [2,5]. An additional preprocessing step to address the information loss during clustering is to apply dimensionality reduction techniques. Almaimouni et al. [7] apply PCA prior to K-Mean on a matrix of net load profiles, to retain the dominant patterns of variability, evaluating the resulting typical days through a rolling-horizon method. However, [8] shows that, for investment planning problems, clustering that preserves only the statistical features of input data does not guarantee optimal investment decisions. More generally, as a linear method, PCA may discard non-linear relationships between periods that are relevant for system design. Non-linear alternatives such as UMAP [9], which preserves the local neighborhood structure of high-dimensional data in a low-dimensional embedding, have recently been applied to electricity load pattern analysis [10] with improved clustering accuracy.

Finally, Dynamic Time Warping was used as a metric distance to measure similarity between time series that share a similar shape but are shifted in time. In [11], the authors apply it with hierarchical clustering to select representative days for a capacity expansion model, showing that duration curves for wind and solar energy are better captured. However, comparing it with other clustering methods on operational optimization problems, [5] shows that its performance is problem dependent.

2. Methodology

This section presents the methodology followed in this work, whose main steps are summarized in Figure 1. Starting from the full resolution time series, periods are first normalized and then clustered into representative days using different combinations of dimensionality reduction techniques (PCA, UMAP) and distance metrics (Euclidean, DTW). The resulting typical periods are fed into an aggregated MILP model whose optimal sizing is compared against the full resolution benchmark. The section is organized accordingly: Section 2.1 describes the temporal aggregation framework, Section 2.2 details the dimensionality reduction and clustering methods, Section 2.3 presents the MILP formulation, and Section 2.4 defines the evaluation metrics used to assess aggregation quality.

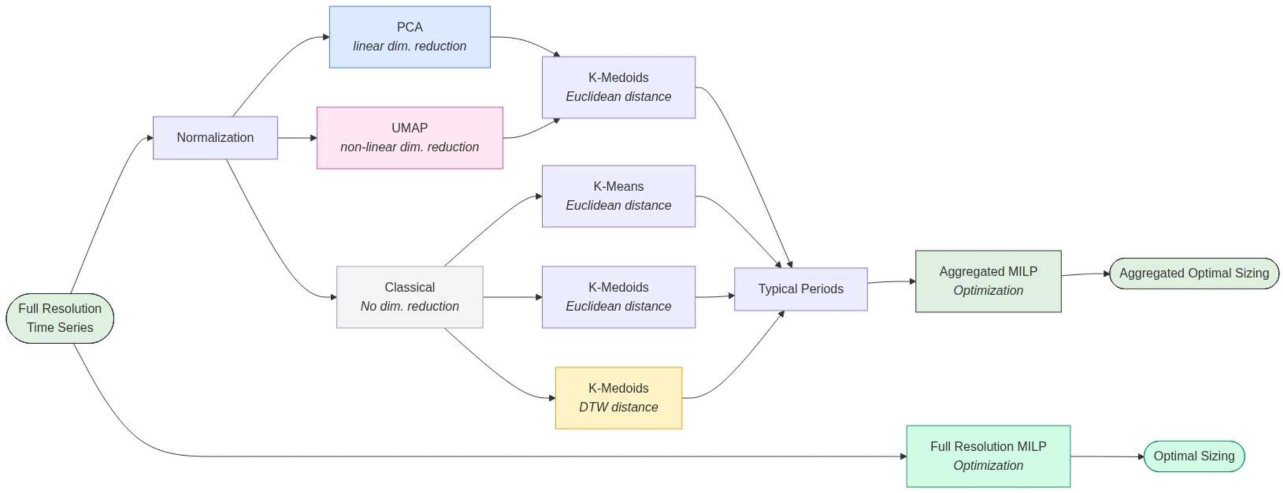


Figure. 1. *Overview of the temporal aggregation and optimization pipeline.*

2.1. Temporal aggregation framework

Temporal aggregation aims at reducing the number of time steps of the original dataset by dividing the original time series into periods of the same length, applying clustering to find representatives, and replacing the original periods with the smaller set of either synthetic or real periods found [2]. To find the representative days, the steps shown in Figure 2 are followed.

Figure. 2. *General steps for clustering time series.*

Our preprocessing step consists in dividing the original time series into periods, normalizing them to account for the difference in magnitude of the attributes, to ensure that no single attribute dominates the distance calculation in the clustering, and then applying either PCA or UMAP.

In the clustering step, either K-Means or K-Medoids is used to find representative periods. Dynamic Time Warping is used as a similarity measure instead of the Euclidean distance when no dimensionality reduction technique is applied to the data, as the aim of the method is to cluster together shifted time series.

Once representative days are found, extreme periods are added to ensure that the energy infrastructure is sized correctly for critical conditions. They are included in the set of representative periods either with a weight of 1 (method 'append'), or by considering them as cluster representatives themselves (method 'new cluster center') [2]. For all methods, the same set of extreme periods are included, as they are defined over the original time series, before preprocessing.

Finally, a rescaling factor is applied so that the weighted mean of the representatives matches the original one. In this work, the rescaling factor leaves the extremes unchanged, although their contribution is considered in the calculation. This implies that rescaled values are allowed to exceed the original maximum values.

The resulting set of representative days is then used to build the aggregated optimization model. Costs are then correctly scaled by the weights of the representative periods to ensure that the total cost of the MES is comparable to the one of the full temporal resolution models.

2.2. Feature extraction methods

The dataset is represented as a flattened matrix $\boldsymbol{X} \epsilon \mathbb{R}^{N \times (H \cdot A)}$, where N is the total number of periods to select, H is the number of hours in each period, and A is the number of time series composing the dataset, called attributes In this work we consider typical days, so $H = 24$, attributes consist of heat and electricity demands, and wind and solar generation, so $A = 4$, and $N = 365$, as we consider one year of data.

By performing PCA, the original data is transformed into a new coordinate system where the axes (principal directions) are ordered by the amount of variance they capture from the data. The number of components is chosen by selecting the smallest number such that the cumulative explained variance ratio (CVR) exceeds 0.95, ensuring that the reduced dataset retains at least 95% of the total variance of the original one. In our case, it corresponds to 27 components as shown in Figure 3, reducing the dimensions by a factor of 3.55.

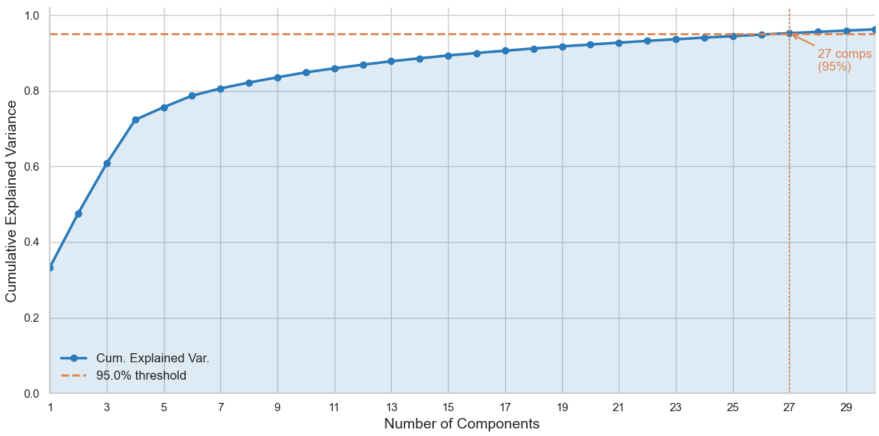


Figure. 3. *Cumulative Explained Variance Ratio over the number of principle directions.*

Alternatively, UMAP is used to project the data into a two-dimensional space that preserves the local neighborhood structure of the original data, thus capturing non-linear dependencies between periods. The resulting space can be visualized on a plane, allowing us to identify the underlying structure of the data. From Figure 4, 6 observable clusters are present.

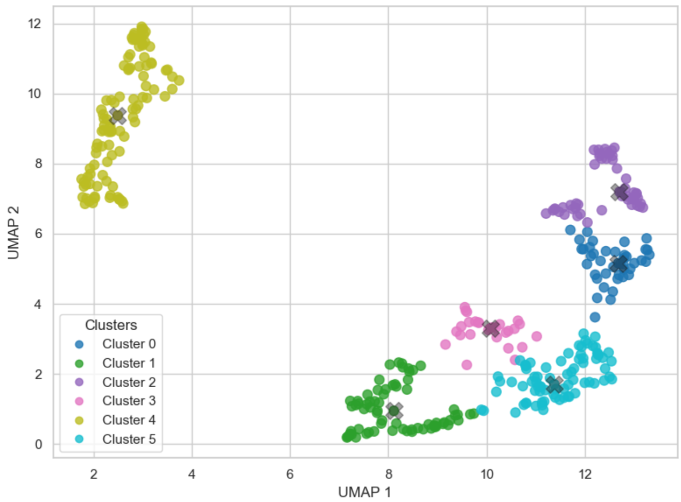


Figure. 4. *Clusters shown on a 2-dimensional space found using UMAP.*

For both the PCA and UMAP methods, we perform the clustering on the reduced space. The caveat of this method is that UMAP is generally not reversible. Therefore, since we want to use real periods as representatives (medoids) we compute the nearest neighbor of the cluster centroids in the reduced space.

Another possibility is to capture temporal shifts by modifying the distance metric used for clustering. Dynamic Time Warping (DTW) is a similarity measure that allows temporal shifts within daily profiles to be accounted for. However, this distance is only applied for non-reduced periods, as temporal shifts can be found only in the original space. Using these feature extraction methods, the set of representative periods are found.

2.3. Multi-energy system optimization model

The design of a Multi-Energy System (MES) is formulated as Mixed-Integer Linear Program (MILP) that simultaneously optimize two categories of decision variables: investment variables, which determine the size of each component of the system; and operational variables, which manage the energy production and storage charging and discharging, defined for each hourly time step $t \in \{0,1,\dots,8759\}$.

The model minimizes the total system cost, including both the initial investment ($CAPEX$) and operational expenditures ($OPEX$), while ensuring that energy demand is met. The optimization is performed over a representative year with hourly resolution, which is assumed to repeat over an investment horizon of 20 years and discounted to account for the time value of money. Indeed, since operational variables are defined at each time step, a MILP formulated on the full investment horizon would have several million variables, including many binary variables associated with storage constraints, making the problem computationally intractable [2]. The formulation used models only in terms of power and energy flows, without including other physical elements. We ignore temperature and pressure for the heat network, making the model more abstract but simpler and linear. The demands and the power generation data are used to determine the sizing of the infrastructure and its operations simultaneously to meet the district's demand.

A key part of the model is the representation of storage technologies, as unlike other components, their variables at time t depend on their values at time $t-1$. In particular, the storage energy level at time t is a function of the level at time $t-1$ and the charge and discharge flows, creating an inter-temporal coupling that should be preserved when performing temporal aggregation techniques. Equation (1) shows this constraint for the thermal storage:

$$(1 + k_{loss}^{TES} \cdot \Delta t) \cdot E_{t+1}^{TES} = E_t^{TES} + \Delta t \cdot \left(\eta_{charge}^{TES} \cdot \phi_{t+1}^{TES,in} - \frac{1}{\eta_{discharge}^{TES}} \cdot \phi_{t+1}^{TES,out} \right) \quad \forall t \in [0, \ldots, 8759). \tag{1}$$

Here, η_{charge}^{TES}, $\eta_{discharge}^{TES}$ are the efficiency factors of the storage (set at 0.95), k_{loss}^{TES} is the percentage of energy lost (set at $1.04 \cdot 10^{-4}$), $\phi_i^{TES,in}$, $\phi_i^{TES,out}$ are respectively the charge and discharge heat flow of the storage at time step i in kW, and E_i^{TES} is the energy level at time step i in kWh. The energy level at time $t = 0$ is set to 0 kWh. This model is later referred to as full temporal resolution model or deterministic model and is used as the baseline to test temporal aggregation models.

A particular challenge of temporal aggregation lies in the modeling of storage technologies, as representative periods are not coupled together. In [4] the authors present three MILP formulations to address this issue. Following the article's second proposal ($M1$ model), the full resolution model is adapted to calculate costs from a set of representative days. While sizing variables remain untouched, operating variables represent a typical time step, rather than an actual point in time. Hence, costs are scaled using the weight of each representative day. However, because energy levels of storage technologies depend on the previous time step, they are modelled at full temporal resolution, using aggregated variables. Therefore, a mapping σ is created to link each original hourly index (from 0 to 8759) to a tuple (p, h), where p is the index of the representative period and h is the intra-hour of that period:

$$\sigma: \{0,1, \ldots, T-1\} \rightarrow \{0,1, \ldots, P-1\} \times \{0,1, \ldots, H-1\}. \tag{2}$$

With this mapping, it is possible to have the storage at the full temporal resolution, while using only aggregated variables. For example, the energy level of storage at time t is then given by:

$$(1 + k_{loss}^{TES} \cdot \Delta t) \cdot E_{t+1}^{TES} = E_t^{TES} + \Delta t \cdot \left(\eta_{charge}^{TES} \cdot \phi_{\sigma(t+1)}^{TES,in} - \frac{1}{\eta_{discharge}^{TES}} \cdot \phi_{\sigma(t+1)}^{TES,out} \right) \quad \forall t \in [0, \ldots, 8759). \tag{3}$$

As a consequence of this coupling, storage technologies are the components most sensitive to the choice of representative periods. Since their operation depends on the sequence and duration of charging and discharging events, their optimal sizing is driven by the temporal structure of the data. When representative periods fail to capture important temporal patterns, particularly when similarity is assessed through pointwise comparisons that ignore shifts or distortions in time, the aggregated model misjudges the required storage capacity. This makes storage capacity an important source of error introduced by temporal aggregation, and for this reason is selected as the key indicator of aggregation quality.

To improve the optimization error, we investigate feature extraction techniques that aim to better preserve the temporal structure of the data.

2.4. Evaluation metrics

While statical measures like the Sum of Squared Errors are useful to evaluate the approximation of the real time series, they are not enough to assess model performance [12]. Since the goal is to obtain a system design that is close to the optimal one, a more useful measure is the optimization error, quantifying the difference between the full temporal resolution and the aggregated model. One of the most common metrics used in the literature is to evaluate temporal aggregation techniques is the optimization error on total cost [2-5], defined as the relative deviation between the aggregated and the full resolution solution:

$$\varepsilon_{tot} = \frac{\hat{c}_{tot} - c_{tot}}{c_{tot}}, \tag{4}$$

where $\hat{c}_{tot}$ is the total cost obtained from the aggregated model, and c_{tot} is the reference value from the model solved at full temporal resolution. However, this metric alone fails to capture differences in internal cost structure of the system, which deviate significantly from the reference solution. Indeed, similar total costs can result from different energy mixes and, as a result, an aggregated solution that matches the total cost may still correspond to a system configuration that is different from the reference one, resulting in poor performances over the full time horizon. In particular, the component whose sizing deviates the most from the reference solution, is the thermal energy storage because of its large capacity. Improving the error made on this component, while not necessarily improving the error made on total cost, helps in better estimating the sizing of the other components. Hence, the clustering methodologies proposed are also compared using the optimization error on the thermal storage, which is defined as:

$$\varepsilon_{TES} = \frac{\hat{c}_{TES} - c_{TES}}{c_{TES}}, \tag{5}$$

where $\hat{c}_{TES}$ is the value obtained from the aggregated model, and c_{TES} is the corresponding value from the deterministic model. Therefore, improving the accuracy of storage sizing, while not necessarily reducing the error on total cost, leads to better designs that will perform more closely to the base model. For this reason, the methods proposed in this work are compared using the relative optimization error on the thermal energy storage.

3. Case Study Description

The MES considered for the case study is based on the Cambridge district in Grenoble, France [13]. The system has two energy vectors: heat and electricity. The interest is in accurately sizing the infrastructure for the renewable sources, the electrical and thermal storage, and a heat pump. All decentralized components within the neighborhood, such as photovoltaic, heat pumps, storage, and buildings, are aggregated into single representative units. Figure 5 shows a schematic of the MES of the Cambridge district in Grenoble, which includes a gas boiler (GAS), a cogeneration biomass unit (CHP), a heat pump (HP), photovoltaic panels (PV), wind turbines (WT), and a thermal (TES) and electrical (EES) energy storage.

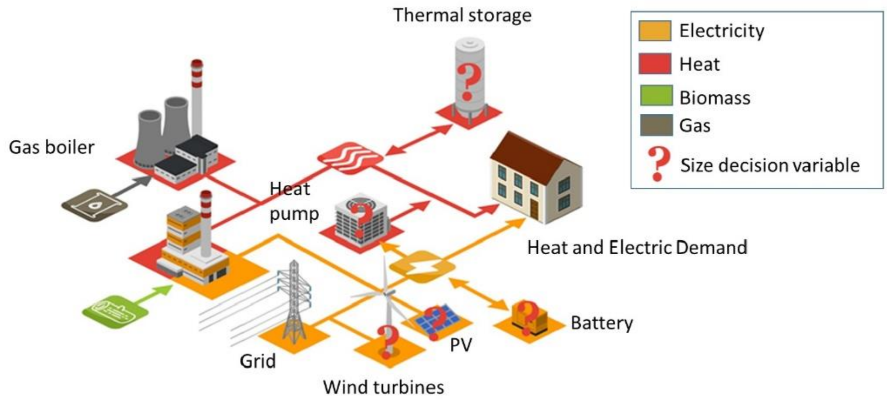


Figure. 5. *The MES of the Cambridge System.*

The data available for this system consists in hourly observations of the heat demand (D_t^{th} in kW), the electrical demand (D_t^{el} in kW), the ratio of actual power production and the theoretical maximum for solar (CF_t^{PV}) and wind (CF_t^{WT}) generation, and the cost of buying electricity from the external grid (c_t^{el} in EUR/kWh). This last attribute is not considered in clustering as it follows a two-tier structure. The analysis focuses on 2019 to have a tractable full-resolution MILP model [13] to use as a benchmark for the aggregation.

Firstly, using the full annual representation, we compute the deterministic size of components, that will be used as a baseline. Then, using the representative periods shown in the previous section, we calculate the size of components on the aggregated version of the MILP.

4. Results

4.1. Representative periods analysis

Using the elbow method on the Sum of Squared Errors (SSE), we decided to select 9 typical days for K-Means and K-Medoids. For PCA clustering, and DTW clustering, the same number of days were taken while, for UMAP clustering, 6 days were picked as confirmed by the visual inspection.

As shown in Figure 6, for K-Means, many representative days concentrate on the lower range of heat demand, corresponding to summer and spring, while fewer days capture the high-demand winter profiles. Moreover, weekend profiles for electricity demand are ignored, while solar and wind generation profiles are well represented in terms of seasonal patterns but appear smoother than the original data. This is inherent to the clustering method, which constructs centroids by averaging all days within a cluster rather than selecting actual days from the dataset.

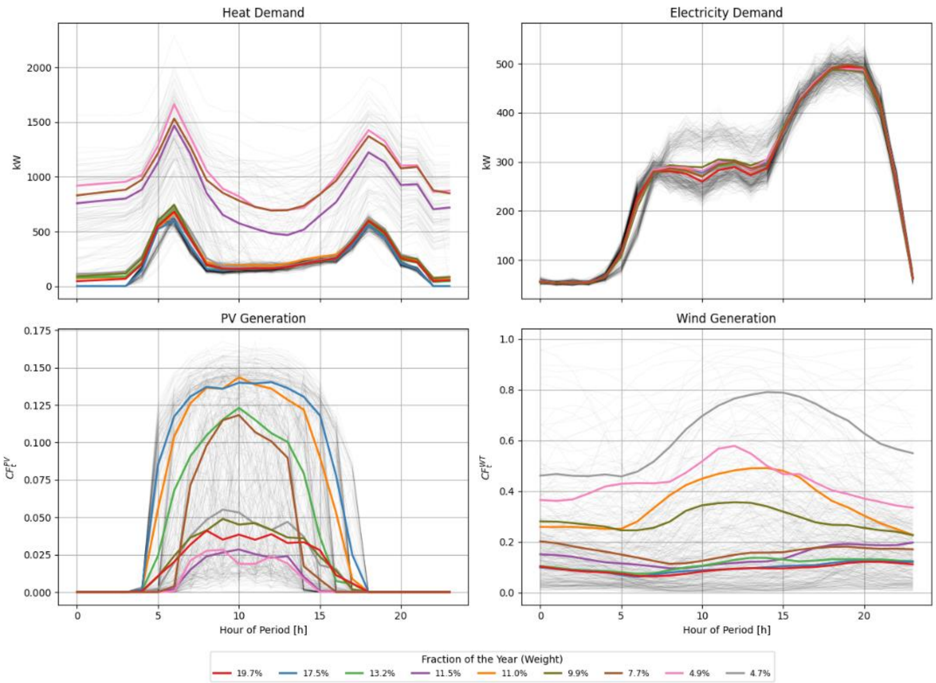


Figure. 6. *The typical days found using K-Means.*

Similar results, shown in Figure 7, are found when using K-Medoids for heat demand and PV generation. However, since medoids are actual days of the dataset, the attributes retain the noise rather than being smoothed out. In electricity demand, a high-demand profile is now selected.

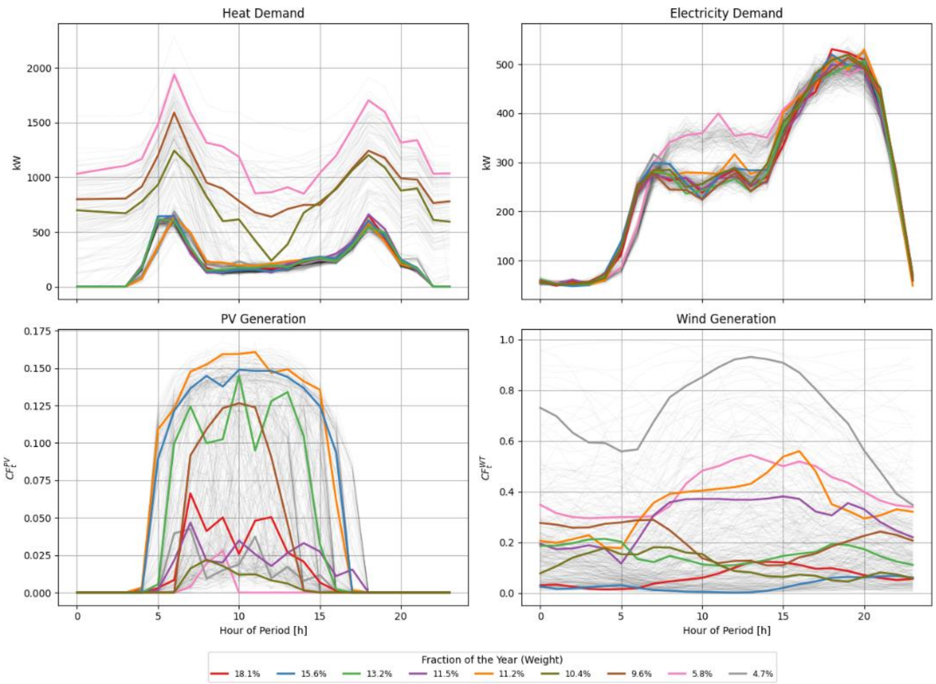


Figure. 7. *The typical days found using K-Medoids.*

Applying PCA as a preprocessing step produces a different set of representatives compared to standard K-Medoids. The weight distribution is more concentrated, with two clusters representing nearly 45% of the year, while the remaining days carry smaller weights, as shown in Figure 8.

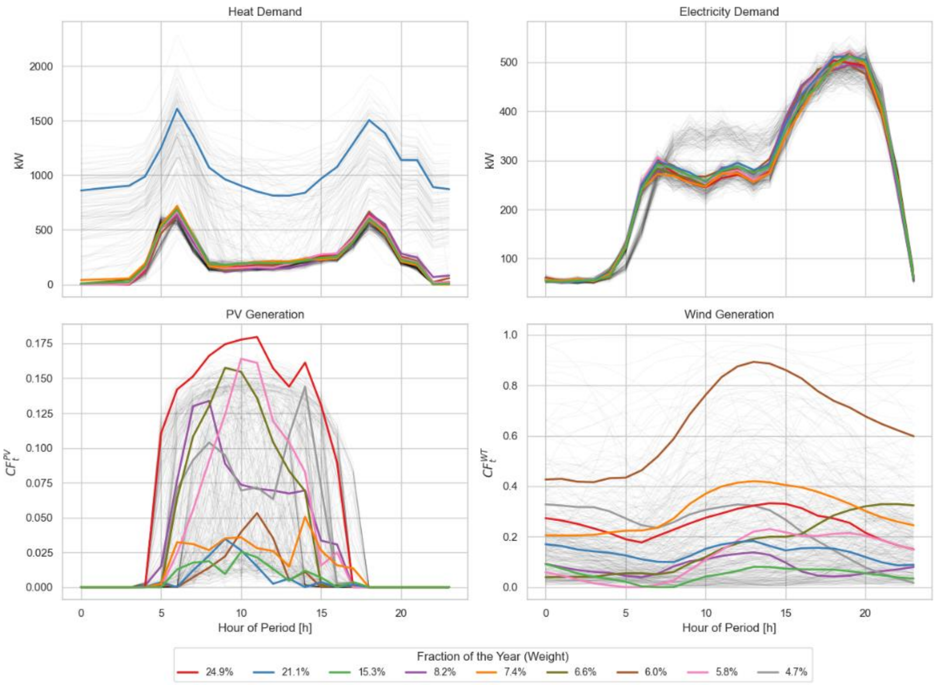


Figure. 8. *The typical days found using PCA clustering.*

In Figure 9, DTW-based clustering groups days that share a similar shape regardless of temporal shifts. This is particularly visible in solar generation, where the seasonal lengthening of the production curve is captured more effectively. However, DTW tends to produce representatives that are very similar in shape, differing mostly in magnitude.

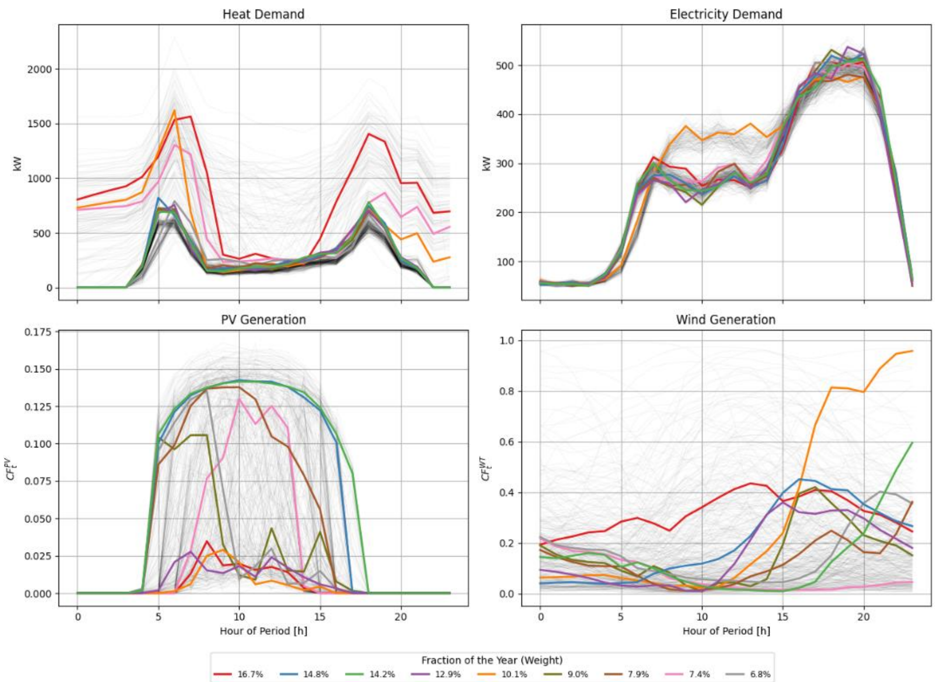


Figure. 9. *The typical days found using DTW clustering.*

For UMAP, Figure 10 shows the corresponding six representative days. Compared to the previous methods, UMAP achieves a clearer separation of distinct temporal patterns with fewer representative days.

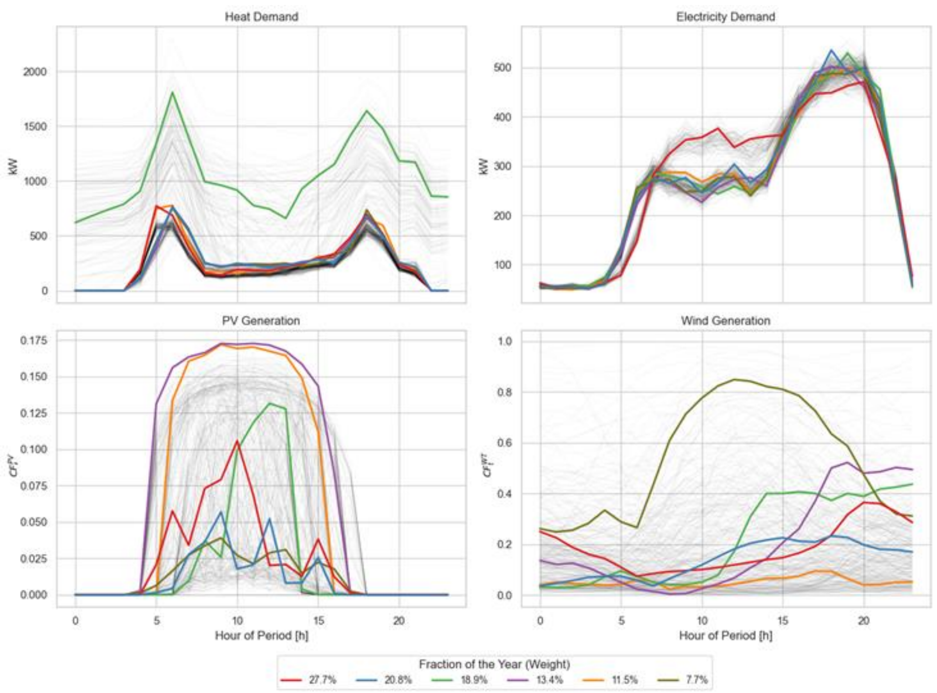
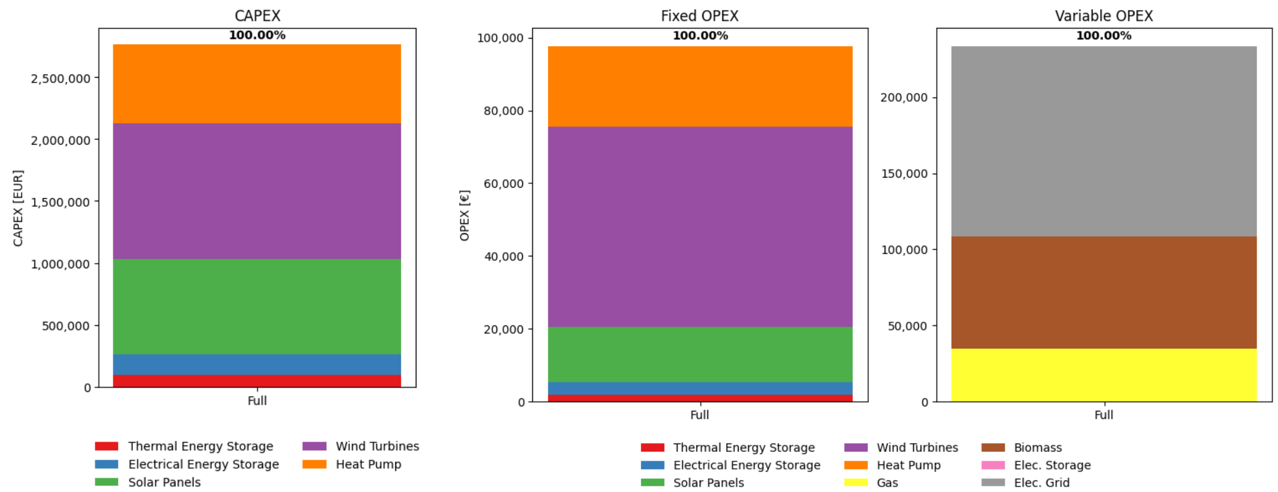


Figure. 10. *The typical days found using UMAP clustering.*

4.2. Impact on storage sizing

The model solved at full temporal resolution yields a minimum total system cost of 6,517,969 €. Figure 11 presents the cost breakdown into capital and operational expenditure. On the investment side, renewable generation (PV and wind) represents the largest part, while the two storage technologies account for similar capital costs. However, looking at the optimal size of each component, reported in Table 1, the thermal energy storage is significantly larger than the electrical storage, reflecting its dominant role in the system. Operation expenditures are reported for a single reference year and are therefore not scaled by the 20-year horizon.



a) b)

Figure. 11. *Financial costs for the deterministic model.*

Table 1. *Optimal values of components.*

Technology	Full Resolution Capacity
EES Capacity [kWh]	1,397.61
TES Capacity [kWh]	6,490.92
HP Capacity [kW]	703.44
Multiplier PV [kW]	3,054.15
Multiplier WT [kW]	1,100.02

Using the aggregated formulation of the Cambridge model, the typical days found serve as inputs to compute the cost of the system. Figure 12 shows the error made in the total cost of the MES for each clustering method, as a function of the number of representative days. As the number of days increases, all methods converge toward the reference total cost, as the aggregated time series becomes a more accurate approximation of the original data.

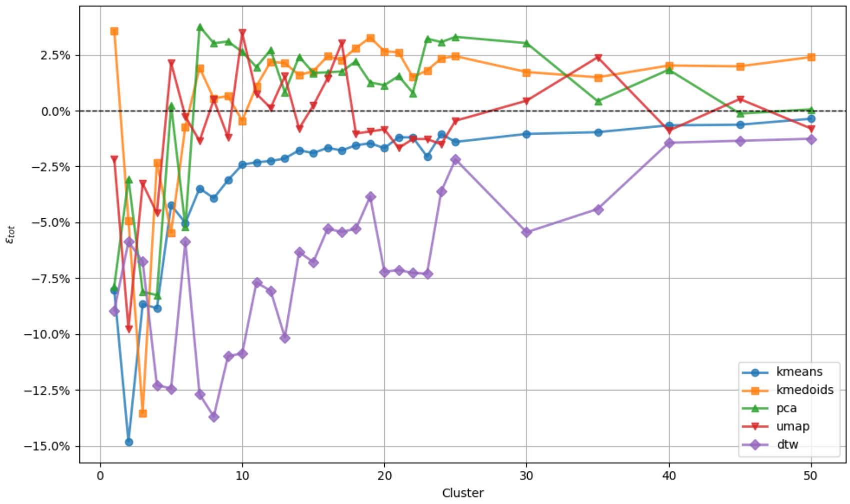


Figure. 12. *Optimization error on total cost.*

DTW clustering exhibits the largest error, making it appear as the worst-performing method when considering total cost alone. However, this conclusion is misleading, as total cost does not fully capture the quality of the aggregation. Similarly to Figure 12, Figure 13 shows the error made on TES sizing in function of the number of representative days used. Here, DTW and UMAP-based clustering achieve the lowest deviations, while the other methods have larger errors of comparable magnitude to one another.

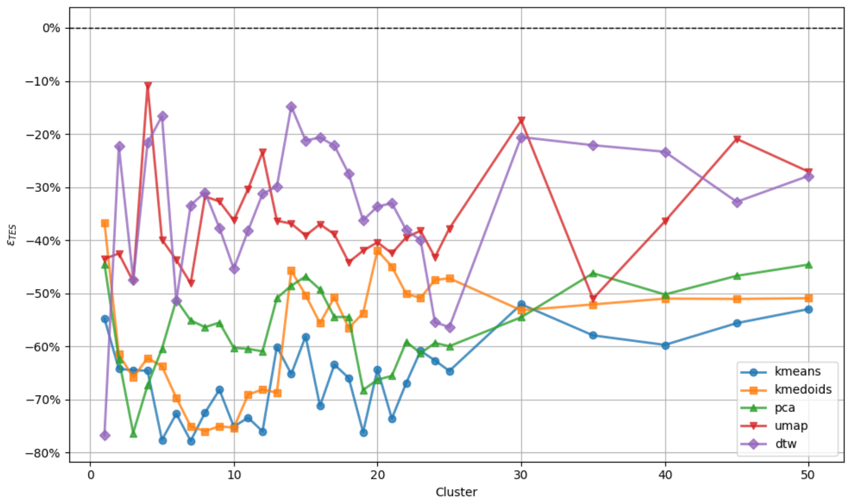


Figure. 13. *Optimization error on the TES total cost.*

Indeed, Table 2 summarizes these results by reporting the weighted average of the TES cost across all tested number of days, confirming the better performance of the DTW and UMAP methods.

Table 2. *Weighted average error of the TES sizing cost.*

Clustering Method	Weighted average error
K-Means	-62.92%
K-Medoids	-53.78%
K-Medoids + PCA	-54.43%
K-Medoids + UMAP	-35.34%
K-Medoids + DTW	-31.18%

5. Discussions

The results presented show that the choice of clustering method significantly affects the sizing of individual components, particularly the thermal energy storage. Its sizing depends on the mismatch between production and demand, with the MES producing and storing energy when it is cost-effective and releasing it during high-demands periods. Correctly estimating the TES capacity therefore requires representative that preserve the timing of such mismatches.

Euclidean distance, used by K-Means and K-Medoids, compares profiles on a point-by-point basis. As a result, they tend to group days that are similar in magnitude rather than in temporal structure. DTW, by contrast, groups days that share a similar shape regardless of temporal shifts. This property is particularly relevant for solar and heat demand profiles, where seasonal shifts in peak timing would otherwise inflate the Euclidean distance between structurally similar days. Consequently, DTW-based representatives better preserve the patterns that determine the TES capacity. Similarly, UMAP constructs a low-dimensional embedding that preserves the local topological structure of the data, capturing nonlinear relationships between daily profiles that pointwise metrics cannot detect.

The underperformance of PCA can be attributed to the fact that the method projects the data onto linear directions of maximum variance, which do not necessarily correspond to the most relevant features for TES sizing.

6. Conclusions

This work investigated the impact of feature extraction techniques on the accuracy of temporal aggregation for Multi-Energy System optimization. Three approaches were proposed and evaluated against the baseline K-Means and K-Medoids methods: (i) Principal Component Analysis (PCA), applied as a linear dimensionality reduction prior to clustering to retain at least 95% of the data variance; (ii) Uniform Manifold Approximation and Projection (UMAP), a non-linear embedding technique that preserves local neighborhood relationships in a two-dimensional latent space; and (iii) Dynamic Time Warping (DTW), an alternative distance metric allowing temporal shifts between daily profiles to be accounted for during K-Medoids clustering. All methods were tested on a real-world case study of the Cambridge district in Grenoble, France, consisting of a multi-energy system with heat and electricity vectors, renewable generation, a heat pump, and both thermal and electrical energy storage.

The results demonstrate that the choice of feature extraction technique has a significant impact on the accuracy of thermal storage sizing. When averaging the TES sizing error across all tested numbers of typical days, K-Medoids with DTW achieved the best overall performance, reducing the weighted average error by 49.5% relative to K-Means and 40.9% relative to K-Medoids. K-Medoids with UMAP pre-processing showed comparable gains, with reductions of 43.8% and 34.3% respectively. In contrast, PCA pre-processing yielded only marginal improvements, reducing the error by 13.5% relative to K-Means while performing slightly worse than standard K-Medoids. These results suggest that the ability to capture non-linear patterns in the data is the key factor driving the accuracy of storage cost.

This work was limited to a single case study, and the impact of data uncertainty was not addressed. Future work could explore additional algorithms, such as density-based approaches like DBSCAN, and enhance the representative periods with engineered features, for example the period's variability, to better capture trends in the data. Finally, extending the analysis to scenario reduction techniques or iterative models that use the result of the optimization to adjust the clustering could improve the accuracy and robustness of system design.

Acknowledgments

The authors thank the project PEPR-TASE: Hymes. Référence: ANR : 22-PETA-0002.

References

[1] European Parliament, Council of the European Union. Regulation (EU) 2021/1119 Establishing the Framework for Achieving Climate Neutrality (European Climate Law). Off J Eur Union 2021;L243:1-17.

[2] Hoffmann M., Kotzur L., Stolten D., Robinius M., A review on time series aggregation methods for energy system models. Energies 2020;13(3):641. doi:10.3390/en13030641.

[3] Kotzur L., Markewitz P., Robinius M., Stolten D., Impact of different time series aggregation methods on optimal energy system design. Renew Energy 2018;117:474-87. doi:10.1016/j.renene.2017.10.017.

[4] Gabrielli P., Gazzani M., Martelli E., Mazzotti M., Optimal design of multi-energy systems with seasonal storage. Appl Energy 2018;219:408-24. doi:10.1016/j.apenergy.2017.07.142.

[5] Teichgraeber H., Brandt A.R., Clustering methods to find representative periods for the optimization of energy systems: An initial framework and comparison. Appl Energy 2019;239:1283-93. doi:10.1016/j.apenergy.2019.02.012.

[6] Domínguez-Muñoz F., Cejudo-López J.M., Carrillo-Andrés A., Gallardo-Salazar M., Selection of typical demand days for CHP optimization, Energy and Buildings, 2011.

[7] Almaimouni A., Ademola-Idowu A., Kutz J.N., Negash A., Kirschen D., Selecting and Evaluating Representative Days for Generation Expansion Planning, 20th Power Systems Computation Conference (PSCC), 2018.

[8] M. Sun, F. Teng, X. Zhang, G. Strbac and D. Pudjianto, "Data-Driven Representative Day Selection for Investment Decisions: A Cost-Oriented Approach," in *IEEE Transactions on Power Systems*, vol. 34, no. 4, pp. 2925-2936, July 2019, doi: 10.1109/TPWRS.2019.2892619.

[9] McInnes et al., (2018). UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software, 3(29), 861, https://doi.org/10.21105/joss.00861.

[10] Wang J., Guo L., Zhang C., et al., An analysis method for residential electricity consumption behavior based on UMAP-CRITIC feature optimization and SSA-assisted clustering, Energy Reports, 2023

[11] Y. Liu, R. Sioshansi and A. J. Conejo, "Hierarchical Clustering to Find Representative Operating Periods for Capacity-Expansion Modeling," in *IEEE Transactions on Power Systems*, vol. 33, no. 3, pp. 3029-3039, May 2018, doi: 10.1109/TPWRS.2017.2746379.

[12] Teichgraeber H., Brandt A.R., Time-series aggregation for the optimization of energy systems: Goals, challenges, approaches, and opportunities. Renew Sustain Energy Rev 2022;157:111984. doi:10.1016/j.rser.2021.111984.

[13] Fitó J., Vallée M., Ruby A., Cuisinier E., Robustness of district heating versus electricity-driven energy system at district level: A multi-objective optimization study. Smart Energy 2022;6:100073. doi:10.1016/j.segy.2022.100073.