Impact of Considering Artificial Worst-Case Scenarios Within Clustering Algorithms

A case study through three newly adapted clustering algorithms

Roman Petar Luka Novosel[1]

Supervisor(s): Germán Morales España[1]**, Maaike Elgersma**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 21, 2026

Name of the student: Roman Petar Luka Novosel
Final project course: CSE3000 Research Project
Thesis committee: Germán Morales España, Maaike Elgersma, Jasmijn Baaijens

Abstract

Planning a long-term energy system relies on models that simulate system operation over many years at an hourly level, which is computationally expensive. A common remedy is temporal aggregation: grouping similar time periods and representing each group by one typical period to shrink the dataset the model must process. This speeds up the computation but tends to average away rare yet demanding conditions, such as days with high energy demand and little energy availability. These extreme periods, however, often determine how much capacity the system requires. This paper introduces three adaptations of widely used clustering algorithms that deliberately embed synthetic worst-case periods into the clustering process, ensuring the representative periods do not ignore the most demanding conditions. We evaluate them against four standard baselines (K-Means, K-Medoids, K-Medoids WC (worst-case), and Hull clustering) by measuring how closely each method's investment decisions match those of a benchmark model that uses the full, unaggregated data: a gap we call relative regret. The standard methods often require a large number of representative periods to approach the benchmark, whereas the proposed worst-case method WCA-K-Means reaches near-benchmark decisions with far fewer periods. By capturing the conditions that drive capacity needs without partitioning the data into excessive detail, it represents a full year with a much smaller dataset, giving planners results that closely match a full-resolution model while substantially reducing the computational cost of solving the energy model.

1 Introduction

The European Union is planning to decarbonise its energy systems by 2050 [1], i.e. substitute the current polluting energy sources with renewable and more environmentally friendly sources. This includes, but is not limited to, solar energy, wind energy, and grid-scale battery storage that can store energy when demand is lower than the supply. Stakeholders need to understand what they should invest in, to ensure the energy transition does not compromise the energy grids. The European Union, for instance, expects investments to reach up to 1.2 trillion euros by 2040 [2]. Thus, while underinvestment directly compromises grid reliability, even minor overestimations already cause massive unnecessary capital-intensive expenditure for stakeholders [3].

To address this problem, planners can use past data to predict the required capacities for future energy systems. A major limitation of the data is that it is too large for modern computers to make calculations from in reasonable time [4]. This has led to extensive research into data clustering, where researchers aggregate similar operational days into single representative periods [5; 6; 7]. Fundamentally, this aggregation process operates along a Pareto trade-off curve: reducing the number of clusters drastically shrinks computational runtime, but it inherently diminishes optimisation accuracy by smoothing out data variability. Preserving these extreme operational boundaries represents a major challenge along the Pareto-front, as capturing peak-day information is essential for valid investment conclusions, yet doing so typically incurs a heavy computational or structural accuracy penalty [5]. Nonetheless, the inclusion of worst-case scenarios does ensure that the method produces a system closer to the optimal design. This is a trade-off between computational time and accuracy, which Hoffmann et al. [8] mathematically formalised by demonstrating how the strategic structuring of typical periods can maximise model accuracy for any given computational time limit.

Current research, however, has not yet been able to create clustering algorithms that balance the inclusion of worst-case scenarios without producing overly conservative upper bounds that lead to system over-investment. For instance, Fazlollahi et al. [5] have proposed the idea of creating clusters of days using K-Means and a set of constraints that minimise the total distance of each data point to the centre, but at the same time maximise the quality of those points. Fazlollahi et al. remove the extreme points from the data, and only after the clustering, manually add extreme cases as 1-sized clusters again. Similarly, Dominguez-Munoz et al. remove extreme days before the clustering process, and re-add them afterwards as isolated clusters [9]. In contrast, Scott et al. argue that while manual addition of extreme cases offers distinct advantages, relying solely on simple extrema (i.e. minimum or maximum values) fails to capture the nuances that drive energy system behaviour. Specifically, the authors demonstrate that rapid changes in net load are more critical for model accuracy. By pre-selecting days with large ramps as initial medoids before the clustering process begins, they achieve results that significantly outperform standard random initialisation methods [10].

Crucially, the aforementioned literature relies on post-processing to force the inclusion of peak periods. These existing algorithms lack a dynamic mechanism to internally adjust cluster formations toward extreme-case scenarios within the clustering process itself. The literature shows that the construction of artificial worst-case days can support this process. Teichgraeber et al., for instance, mention that the creation of 'virtual days' (i.e. synthetic days combining the maximum demand with the minimum availability of each energy source) can support the design of systems that rely on storage, as less availability of energy sources yields the need for backup energy [11]. However, no research has properly investigated the creation of artificial worst-case data within the clustering process.

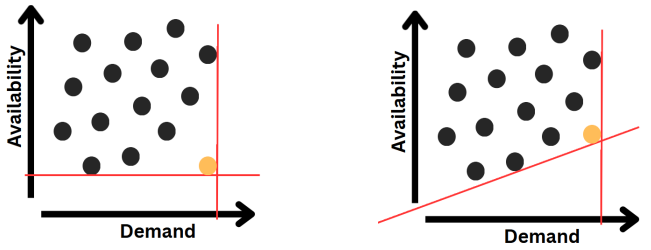
This paper proposes and compares three adapted clustering algorithms — worst-case adaptive (WCA) versions of K-Medoids, K-Means, and Hierarchical clustering (hereafter WCA-K-Medoids, WCA-K-Means, and WCA-Hierarchical, respectively) — that account for worst-case days **within** the clustering process. Analysing how existing clustering methods can incorporate artificial worst-case representative days, and how this affects performance, reveals improvements for worst-case bounded energy systems. Specifically, this work

aims to answer the following research question: how does the embedding of artificial worst-case representative days within clustering algorithms impact the accuracy of energy system model solutions, and the associated computational time? The theoretical contributions of the new algorithms are the following:

1. **Endogenous Worst-Case Integration:** This work proposes a worst-case adapted framework that embeds artificial worst-case periods directly within the iterative loops of standard clustering algorithms, allowing extreme events to actively shape cluster convergence.
2. **Pareto-Optimal Boundary Updates:** By utilising non-dominated sorting inside the aggregation phase, the algorithms preserve critical, multi-objective operational trade-offs that traditional methods typically smooth over.
3. **Near-Optimal Solution Convergence at Reduced Cluster Counts:** The WCA-K-Means framework reconstructs both the optimal investment portfolio and Loss of Load profile using significantly fewer representative periods than standard baselines, demonstrating that endogenous worst-case integration reduces the computational cost of energy system modelling without sacrificing solution quality.

Finally, this paper highlights WCA-K-Means in more detail due to its better performance compared to WCA-K-Medoids and WCA-Hierarchical clustering.

2 Worst-case representative periods



(a) Pessimistic worst-case point creation

(b) Elgersma's improved worst-case point creation

Figure 1: Graphs showing different ways of creating a worst-case point from a set of points

Before embedding worst-case boundaries inside an iterative clustering framework, we first define how to mathematically construct individual worst-case scenarios. Traditional extreme-point aggregation often leads to over-conservative profiles, requiring a more nuanced approach to capturing system stress without inducing over-investment. To create a worst-case hour from a set of hours, the first instinct is to select the maximum demand of all points, and the minimum availability of all energy sources. Intuitively this feels safe, since it assumes the worst of every source at once. This, however, creates an unnecessarily pessimistic hour [12] as shown in Figure 1a, since it pairs two independently chosen absolute extremes that may never have actually occurred together. Instead, as shown in (1), (2), and (3) and visually in Figure 1b, one can create less extreme worst-case points. The key idea is that $\gamma_{g,r}$ is computed as a ratio within each individual hour, before the minimum across hours is taken, so it reflects a real combination of demand and availability that genuinely occurred at one specific hour, rather than an artificial pairing assembled from two unrelated extremes. By scaling this observed worst-case ratio against the period's peak demand, the method captures realistic system stress without generating artificially severe conditions.

$$D_r = \max_{t \in T_r} D_t \quad \forall r \in R, \tag{1}$$

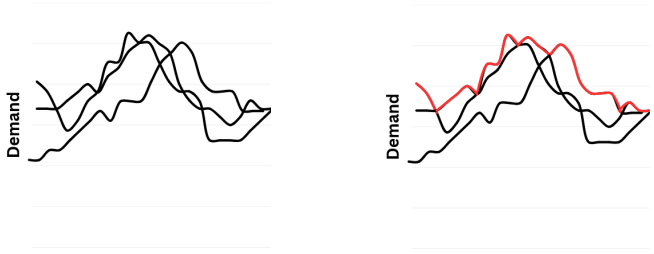
$$\gamma_{g,r} = \min_{t \in T_r} \frac{A_{g,t}}{D_t} \quad \forall g \in G, r \in R. \tag{2}$$

$$A_{g,r} = D_r \cdot \gamma_{g,r} \quad \forall g \in G, r \in R. \tag{3}$$

In these equations, R represents the set of representative periods, $t \in T_r$ indexes the individual time steps within a given period r, and $g \in G$ indexes the sources of energy. D_t represents the historical electricity demand at a specific time step, whereas D_r captures the peak demand for the entire representative period by selecting the maximum D_t value across all its time steps. $A_{g,t}$ denotes the raw availability factor of an energy source in a given time step. Dividing by D_t then yields a time-dependent availability-to-demand ratio. $\gamma_{g,r}$ isolates the minimum (worst-case) of these ratios across the entire period. Finally, $A_{g,r}$ represents the scaled representative availability for generator g during period r, calculated by multiplying the peak demand D_r by the critical ratio $\gamma_{g,r}$ to ensure that this construction conservatively preserves the most constrained operational hour.

Extension to Full-Day Worst-Case Artificial Points

Based on Elgersma et al. [12], this paper extends the method to allow for the creation of a full 24-hour worst-case data point.



(a) Unmarked graph

(b) Marked graph showing the upper envelope in red

Figure 2: Graphs showing how to create an upper envelope from three days of demand

Per hour of a 24-hour day, the algorithm selects the maximum demand. Repeating this 24 times for the full day then creates an upper envelope. This is shown visually in Figure 2b. Next, we find the minimum ratios of availability and demand during a specific hour in the same way by repeating

(2) 24 times, but now by creating a lower envelope. Then, applying (3) per hour constructs the final full day of artificial worst-case hours. The result is a synthetic day that likely never occurred, yet one that captures the shape of the most demanding conditions the system must withstand.

WCA-K-Medoids through Non-Dominated Sorting

The proposed WCA-K-Medoids algorithm shifts the update focus from the entire set of cluster medoids to a targeted subset located on the system boundaries identified via non-dominated sorting as first demonstrated by [13]. These boundary layers are the Pareto-fronts, capturing the specific medoids that exhibit the highest power demands and the lowest renewable availabilities. By framing this selection as a multi-objective optimisation problem, the algorithm preserves critical operational trade-offs that standard clustering methods typically smooth over. Namely, a scenario with high demand and high availability of energy is not necessarily less or more extreme than a day with low demand and low availability of energy.

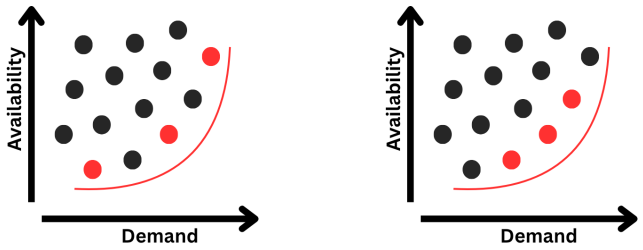
A well-known drawback of non-dominated sorting is its susceptibility to the curse of dimensionality, where the distinctiveness and resolution of the Pareto-fronts deteriorate significantly as the data dimensionality increases [14]. To circumvent this limitation, the algorithm compresses the dimensional space by mapping each 24-hour profile to its daily peak demand and minimum energy availability, which then define the operational layers while reducing the dimensionality by a factor of 24.

Candidate Medoid Selection

The specific implementation of the Worst-Case Adaptive (WCA) K-Medoids algorithm that this paper presents updates a subset of the total points bounded by $n = \lceil 0.25 \cdot K \rceil$, where K denotes the total number of medoids. A structural challenge arises when a given non-dominated front contains c candidate points such that $c > n$, requiring a strategy to select a subset of the candidate points. In traditional multi-objective evolutionary literature, researchers typically resolve this issue using the Crowding Distance metric as proposed in the NSGA-II framework [15]. The Crowding Distance mechanism prioritises a uniform distribution of points to maintain diversity across the entire front, which frequently results in discarding the most extreme outlying solutions in favour of more central, intermediate points.

However, filtering out these boundary points directly contradicts the primary objective of worst-case clustering. To mitigate this, WCA-K-Medoids deviates from standard NSGA-II selection by replacing the Crowding Distance criterion with a distance metric. The algorithm explicitly prioritises the selection of points that minimise the Euclidean distance to the absolute theoretical worst-case data point, defined by maximum demand and zero renewable availability as shown in Figure 3.

This results in a selection of points that best represent the worst-case scenarios, while common data remains adequately represented. Finally, the algorithm incorporates an injected worst-case data point into each of the selected clusters, with a weight proportional to the layer the medoid is located on.



(a) Three points selected through Crowding Distance

(b) Three points selected through taking the most extreme cases

Figure 3: Graphs showing different ways of selecting a subset of points from a Pareto-front

However, the algorithm only updates the selected clusters every τ iterations, with $\tau > 1$, as injecting points every iteration would lead to slow convergence. We set the weight to $25/i$, with i being the layer the medoid is located on after non-dominated sorting. We set τ at 4 in this case study, with both values found by empirical tuning, showing an optimal balance between preserving critical edge-case extremes and maintaining overall cluster representativeness.

Furthermore, due to the injection of artificial worst-case points into the clusters, WCA-K-Medoids has the possibility of selecting an artificial point as a final representative day. While this is not guaranteed to happen, it is important to understand that it is nonetheless possible.

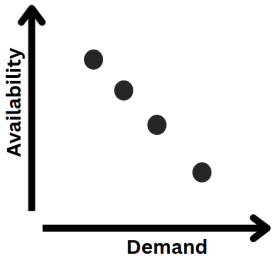
WCA-K-Means

Unlike WCA-K-Medoids, the WCA-K-Means clustering algorithm does not inject new data points. Instead, it focuses solely on updating the centroid values to worst-case scenarios, while preserving the original pool of data points. At the same time, the algorithm updates only $\lceil 0.25 \cdot K \rceil$ cluster centroids toward a worst-case point. Through the use of non-dominated sorting, the algorithm updates centroids on deeper layers to a blend of their cluster mean and the worst-case profile, scaled by a factor $\alpha = 1/L$ where L is the Pareto-layer. This strongly anchors outer boundary clusters to operational limits while allowing interior clusters to smoothly decay to preserve data density.

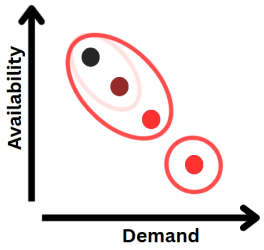
WCA-Hierarchical

Finally, this paper updates the Hierarchical clustering algorithm to better represent worst-case data. Similarly to K-Medoids, hierarchical clustering normally creates a lower bound of the actual energy system due to the averaging out of the data points. The adjusted version, however, aims at creating a worst-case bound for the energy system. Namely, whereas normal merging of two clusters results in creating a new middle point based on the weighted average of the two merged middle points, the new algorithm creates a worst-case artificial point from the merged clusters, and the newly formed cluster will only be represented by that new worst-case point.

Of the three methods, this is the most uncompromising realisation of the worst-case principle, since every merge com-



(a) Points before hierarchical clustering



(b) Points after worst-case hierarchical clustering

Figure 4: Graphs highlighting worst-case clustering through WCA-Hierarchical clustering with two final clusters

mits fully to the worst-case representation rather than blending it with an average. As shown in Figure 4, an issue is the fact that data points can rapidly obtain a final representative period that severely misrepresents a large share of the underlying data points due to the worst-case representative 'shifting' away from the original data points.

Tulipa Energy and Clustering

To solve energy system models, this study utilises Tulipa [16], an open-source framework written in Julia. Tulipa formulations characterise the energy infrastructure as a network of nodes and links, utilising linear programming to minimise total discounted investment and operational system costs. By optimising asset capacities, the model identifies the most cost-effective strategy to satisfy regional energy demands across the designated representative time periods. This work uses TulipaClustering.jl for temporal dimensionality reduction, an open-source Julia package specifically designed for aggregating time-series data within energy system optimisation workflows.

3 Clustering quality assessment

To evaluate the performance and scalability of the proposed methodology, we implemented and tested the clustering algorithms within the case study framework. We designed all numerical experiments to ensure reproducibility and to provide a fair comparison between the newly proposed clustering methods and the reference methods.

We compare each newly implemented clustering algorithm to its original implementation, except for WCA-Hierarchical clustering due to no default implementation existing in Tulipa. For that reason, we compare it to Hull clustering. Hull clustering can be viewed as taking a rubber band, wrapping it around all data points, and seeing at which data points it snaps tight. These data points become the representative days.

To ensure a fair comparison, we execute each algorithm five times per cluster size using distinct random seeds ($s \in 1, 2, \ldots, 5$), except for WCA-Hierarchical and Hull clustering due to their deterministic nature. We present the final results as the average across these five runs, with the accompanying ribbon denoting the absolute minimum and maximum performance boundaries.

We evaluate the algorithms across a range of k clusters, where $k \in [1, 1095]$, sampling k at 50 strictly increasing intervals following an exponential distribution. We evaluate the clustering performance a priori and present the a posteriori results of solving the energy model afterwards.

Evaluation Metrics

We evaluate the performance of the clustering configurations a priori using the Sum of Squared Errors (SSE) and the Silhouette Score, defined respectively below, alongside the total computational execution time:

$$\text{SSE} = \sum_{k=1}^{K} \sum_{x \in C_k} \|x - \mu_k\|^2 \tag{4}$$

$$\text{Silhouette Score} = \frac{1}{N} \sum_{i=1}^{N} \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{5}$$

Where (4) measures the total compactness of the clusters relative to their centres μ_k, and (5) computes the mean separation-to-cohesion ratio across all N data points, with $a(i)$ being the average intra-cluster distance from point i to all other points within its assigned cluster, and $b(i)$ representing the mean distance from point i to all points in the nearest neighbouring cluster.

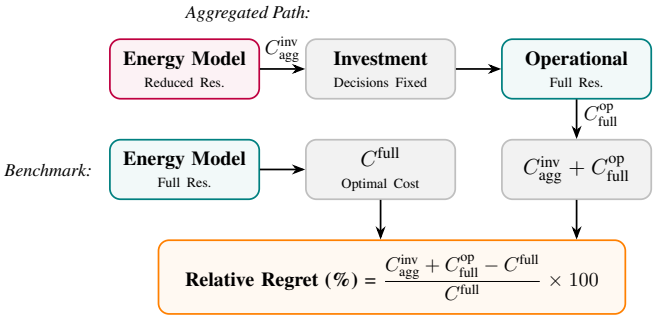


Figure 5: Calculation of relative regret. The aggregated path solves the reduced-resolution model to obtain investment costs, fixes those investments, and re-solves operations at full resolution; the benchmark path solves the full-resolution model directly for C^{full}. Relative regret is the percentage gap between the two [17].

After running the energy model, we measure the regret. As shown in Figure 5, we calculate the regret by first solving the model using the full dataset. This results in C^{full}, the optimal total costs. Afterwards, we solve the energy model using the reduced dataset, resulting in C^{inv}_{agg}, the investment costs. By fixing the investment costs, and resolving the energy model with the full dataset, we obtain the operational costs C^{op}_{full}. Together, they give the relative regret of the clustered dataset.

Finally, we calculate the Loss of Load (LoL), representing the annual unserved energy volume. Notably, even an optimal investment plan can yield a non-zero Loss of Load, even though this might sound counter-intuitive. If the system were to handle 100% of the hours, however, the investments into sustainable resources would be immense, while a

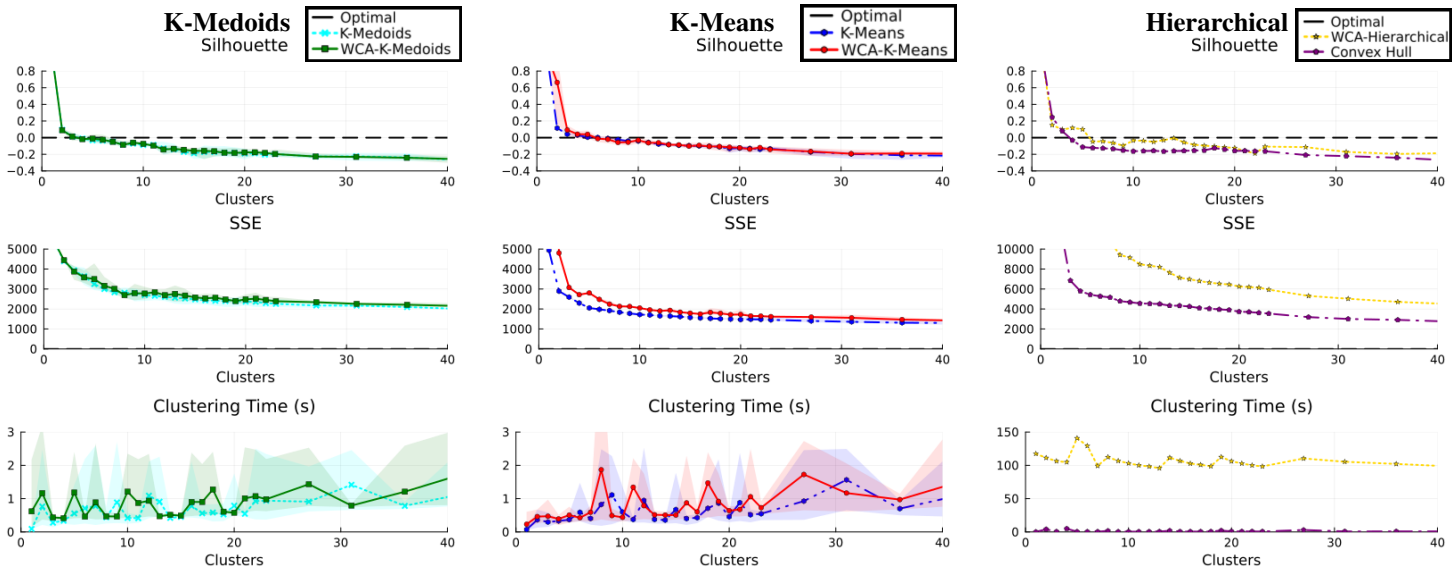


Figure 6: A priori clustering-quality metrics up until 40 clusters for the three methods, including their error margins. Note that the computational time and SSE score for Hierarchical clustering are shown on a separate scale due to its larger magnitude.

slightly lower uptime would be significantly cheaper. During periods of unserved demand, utilities may pay large industrial consumers such as data centres to temporarily reduce consumption, or they may implement critical peak pricing [18]. Consequently, the capacity expansion model heavily penalises any Loss of Load.

4 Case Study

To evaluate the empirical performance of the proposed Worst-Case Adaptive (WCA) clustering frameworks, this section presents a comprehensive case study utilising the Tutorial 9 dataset provided by the Tulipa Energy framework [19], with the system shown in Figure 7. Rather than relying on a static, single-year historical profile, this benchmark dataset structures three distinct weather and demand years as independent operational scenarios. This multi-year scenario framework provides a rigorous testing environment for temporal dimensionality reduction, as it forces the clustering algorithms to navigate both intra-annual volatility and inter-annual structural variations.

We divide the analysis into two sequential phases. First, we evaluate the a priori clustering performance using traditional statistical metrics, before solving the energy model. However, because the literature demonstrates that *a priori* data-handling metrics do not always correlate directly with the results of the solved energy model [20], we subsequently present a detailed *a posteriori* evaluation. In Figure 9, we further compare WCA-K-Means against Tulipa's worst-case baseline K-Medoids WC (K-Medoids, but with a global worst-case point added afterwards with a weight of 10% of the total weight) and Hull clustering. We restrict the plots to a range of up to 40 clusters to emphasise the primary differences. We defer the full-scale graphs to Appendix B.

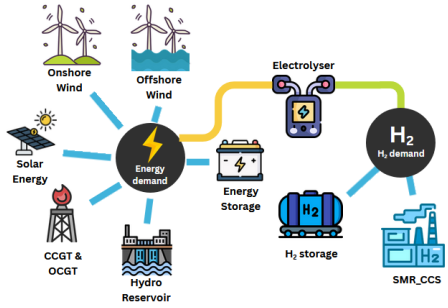


Figure 7: The Tutorial 9 energy system from Tulipa, comprising renewable generation, conventional and low-carbon thermal generation, an electrolyser and hydrogen storage pathway, and battery and hydro reservoir storage, all serving electricity and hydrogen demand.

A Priori Evaluation

As presented in Figure 6, the new adaptations of the basic clustering algorithms have a mixed impact on the a priori performance. WCA-K-Medoids has an SSE and Silhouette Score almost identical to its original implementation. This shows that injecting a small number of worst-case points has minimal effect on the cluster boundaries: the medoids absorb the extremes, while the bulk of common days remain equally well represented by their clusters.

For WCA-K-Means, its SSE is worse, while the Silhouette Score is similar. This behaviour is expected, as it follows from what each metric measures. SSE sums the squared distance between each day and its assigned centroid. Because WCA-K-Means replaces the distance minimising cluster mean with a synthetic worst-case profile for 25% of the clusters, the resulting metric inevitably rises. The Silhouette Score instead reflects only the partition, namely how close

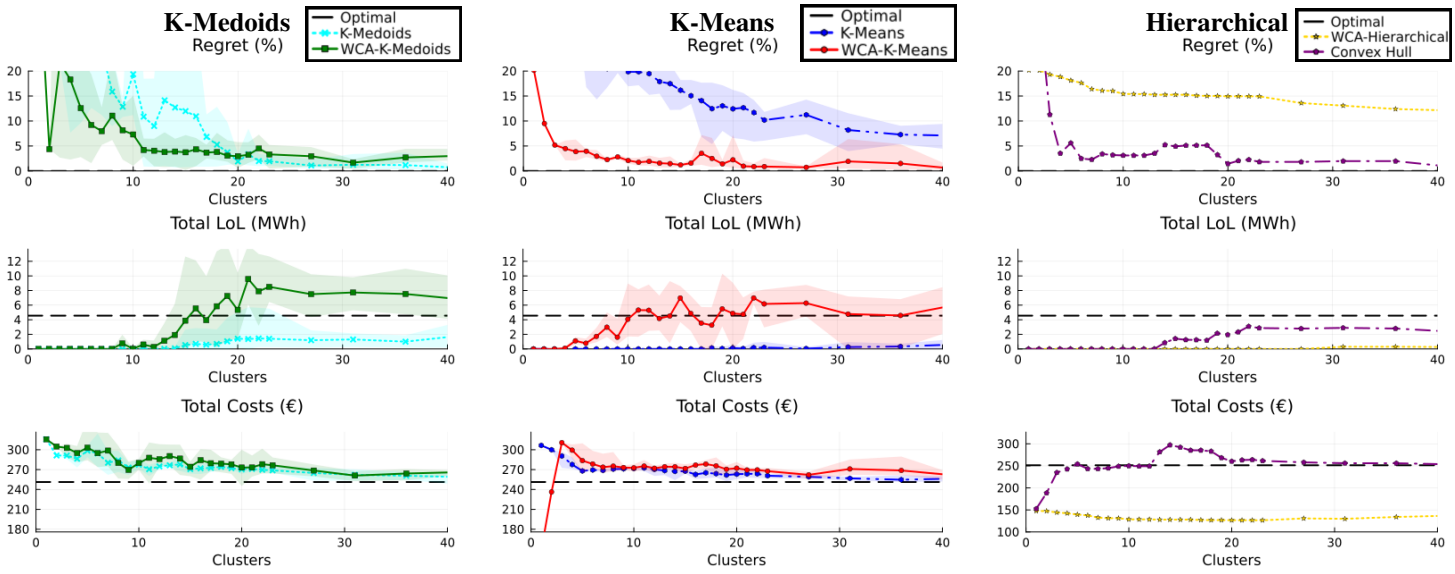


Figure 8: A posteriori optimisation results up until 40 clusters for the three methods, including their error margins. For some scenarios, the error range is so large that it covers the full graph. Each column's legend is overlaid on its top panel.

each day is to its own cluster compared to the nearest other cluster, independent of which specific point represents each cluster. While worst-case centroids can indirectly reshape the partition through the iterative assignment step, only the worst quartile of clusters is affected, and the bulk of the data continues to be assigned in a similar way to standard K-Means.

While WCA-K-Medoids and WCA-K-Means are not strong outliers, WCA-Hierarchical clustering performs poorly on the SSE metric. As mentioned and shown previously in Figure 4, worst-case adapted Hierarchical clustering is prone to linking common, non-extreme data points to extreme case scenarios. As a result, the SSE performs poorly compared to all other methods, as the algorithm assigns data points to clusters they are not closely related to. However, the Silhouette Score is comparable to Hull clustering, because the algorithm merges the closest clusters, maintaining separation between clusters.

Although the worst-case adaptations for K-Means and K-Medoids increase clustering runtime in later stages relative to their baselines as shown in Appendix B, the difference in runtimes when few clusters are required remains modest. WCA-Hierarchical clustering, however, performs poorly, especially when fewer clusters are required due to the algorithm's bottom-up approach. This is an issue, as the goal of worst-case clustering is to avoid infeasibility of the energy system with as few representative periods as possible.

All in all, this mix of a priori results and computational times does not mean the algorithms fail. Instead, they are the direct result of a deliberate distortion needed to capture extreme-case scenarios. Sacrificing traditional cluster compactness and speed is a necessary trade-off: by distorting the dataset to preserve peak demand and low renewable availability, the algorithms increase the likelihood that the resulting system is dimensioned for grid stress events.

A Posteriori Evaluation

We present the a posteriori results in Figure 8, revealing a stark contrast to the earlier a priori clustering metrics. It is evident that most algorithms exhibit fluctuating performance due to their inherent tendency to become trapped in local optima rather than converging to the global optimum. Nonetheless, while the adapted algorithms originally scored comparably in compactness (SSE and Silhouette Score), they demonstrate vastly different performance after running the full energy model. Only for WCA-Hierarchical clustering do the results correlate with the a priori metrics: the algorithm performs poorly, as its regret decreases more slowly, the Loss of Load is too low for too long, and WCA-Hierarchical severely underestimates the total costs. All in all, WCA-Hierarchical clustering in its current form is not a proper replacement for worst-case clustering due to its inherent chaining effect.

Conversely, for WCA-K-Means and WCA-K-Medoids, the results differ significantly from their traditional implementations. WCA-K-Medoids outperforms its original implementation in terms of regret up until 20 clusters, but does overestimate the Loss of Load. This reversal for WCA-K-Medoids in terms of regret is expected rather than anomalous. Standard K-Medoids improves steadily as the number of clusters grows, naturally capturing rare high-demand, low-renewable days without any special intervention, closing the gap with the benchmark. WCA-K-Medoids, by contrast, carries a structural bias: the injected worst-case points displace medoids from their true cluster centres, systematically overestimating system stress. Below roughly 20 clusters, this bias is beneficial because capturing extreme periods within a small number of representative days is what reduces regret. Beyond that point, standard K-Medoids closes the coverage gap on its own, while the worst-case distortion persists, causing the curves to cross.

Although WCA-K-Medoids exhibits lower regret during

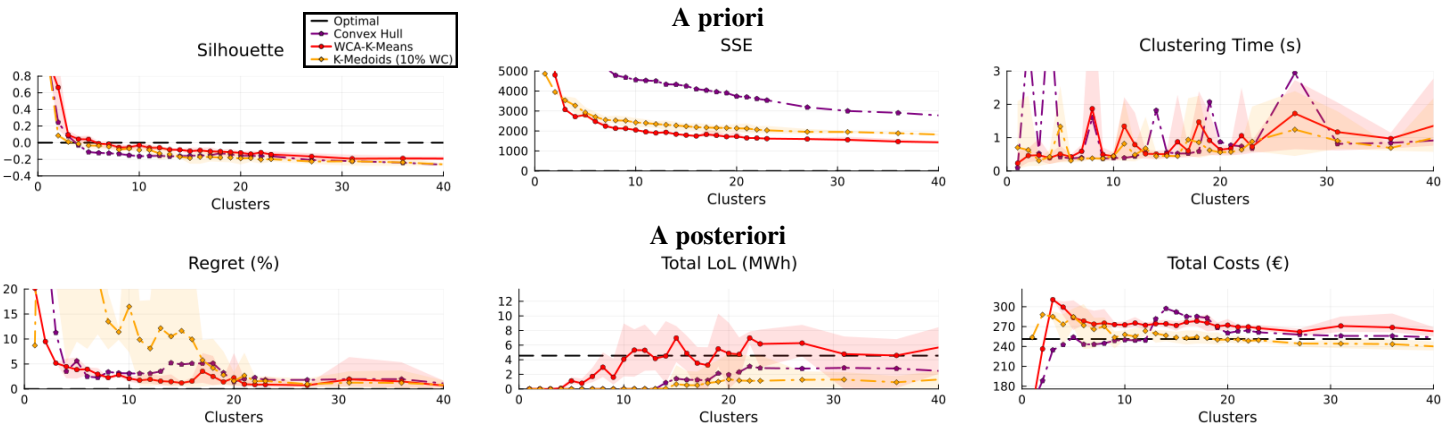


Figure 9: Direct comparison of WCA-K-Means, Hull clustering, and K-Medoids WC clustering up until 40 clusters across a priori metrics (Silhouette, SSE, solve time) and a posteriori results (regret, total Loss of Load, investment cost).

one value of clusters, this trend quickly reverses, a behaviour attributable to an early stochastic convergence toward the global optimum. Throughout the remaining configurations, WCA-K-Means consistently outperforms WCA-K-Medoids. The structural superiority of WCA-K-Means over WCA-K-Medoids lies in how it navigates continuous space. WCA-K-Medoids constrains itself to selecting an actual data point as its cluster centre, and thus can struggle to fully shift its medoids to a true representative worst-case point. WCA-K-Means operates in a continuous, multi-dimensional vector space. By blending the worst quartile of clusters toward an artificial worst-case Pareto profile rather than the plain mean, WCA-K-Means mathematically constructs a synthetic representative profile that smoothly captures extremes.

Furthermore, as illustrated in Figure 9, WCA-K-Means significantly outperforms the K-Medoids WC baseline in terms of regret up until approximately 20 clusters. While Hull clustering exhibits slightly lower regret at the lowest cluster configurations, its regret quickly rebounds because its strict reliance on absolute geometric boundaries becomes overly sensitive to unrepresentative outliers as the cluster count expands. Beyond 20 clusters, the performance curves converge as standard representation gaps naturally close.

However, a low regret alone does not guarantee an optimal investment plan; the resulting Loss of Load must also resemble the benchmark system. WCA-K-Means predicts the optimal Loss of Load at a remarkably low number of clusters, whereas WCA-K-Medoids consistently overestimates it, and the two baselines underestimate it across a wide range of cluster counts. This divergence directly traces back to their profile construction mechanisms:

- **WCA-K-Means** represents each stressed cluster with a smooth, synthetic profile that is mathematically guaranteed to be at least as demanding as the cluster's true peak stress when residing on the first Pareto-front. This produces an appropriately conservative, stable capacity plan with near-optimal Loss of Load, avoiding underinvestment.
- **K-Medoids WC** reaches a lower, cheaper-looking cost prediction by relying on a single, discrete historical day to stand in for each extreme. This structural reliance makes it highly likely that the selected day understates the true combined severity of demand and resource scarcity. Consequently, it designs a grid tailored to absolute outliers while misrepresenting common operating conditions, leading to capacity deficits and underestimated Loss of Load.
- **Hull clustering** faces a different structural limitation: rather than selecting a single historical day, it constructs extreme periods from geometric boundary vertices. While this captures absolute peak bounds well at low cluster counts, it lacks the representative everyday operating context of a clustered mean, tailoring the grid to absolute outliers and causing the observed lower Loss of Load.

Finally, WCA-K-Means's overestimation of the total costs reflects a deliberate bias rather than a modelling flaw, building resilience that the baseline strategies occasionally underestimate.

Comparative Analysis of WCA-K-Means and Standard K-Means

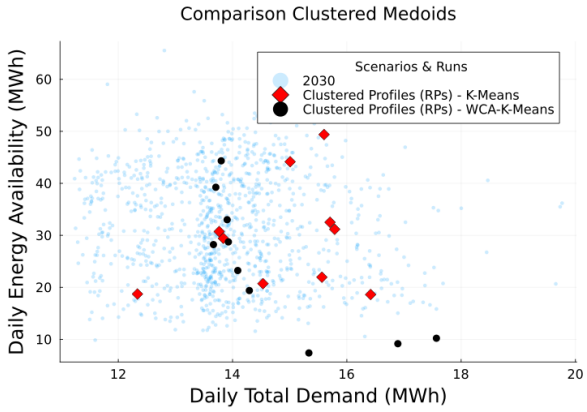


Figure 10: Centroids for Standard K-Means (Red) and WCA-K-Means (Black).

Comparing the top-performing algorithm, WCA-K-Means, with its baseline counterpart highlights the distinct impact of the Pareto-front adaptations. Figure 10 illustrates the baseline K-Means centroids for a 10-cluster configuration alongside their shifted positions under WCA-K-Means. WCA-K-Means creates three artificial profiles to represent worst-case conditions while preserving the standard data distribution across the remaining clusters. Consequently, it maintains a compact aggregation of typical operational periods, whereas baseline K-Means distributes representative points in a more unstructured manner.

Capacity Investment Comparison: Benchmark and Clustered

To further understand whether WCA-K-Means creates a realistic investment plan, we can compare the optimal investment assets and the clustered investments. Figure 11 compares the optimal assets to the averaged assets that the different aggregated datasets predict. These investments are close to identical. Nonetheless, only WCA-K-Means opts for the construction of Open Cycle Gas Turbines (OCGTs). The inclusion of more extreme days with higher demand and lower availability causes the creation of OCGTs as a backup. While they are expensive to build, they can be utilised in scenarios with little natural availability of renewable resources.

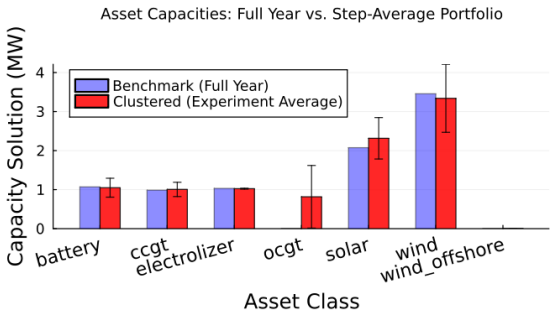


Figure 11: A comparison between the investments made by running the energy model with the full dataset, and the average investments made by the different aggregated datasets for WCA-K-Means with the standard deviation shown.

Other K-Means Variations

Besides the final idea of utilising non-dominated sorting for worst-case clustering, we instead constructed the layers based on the minimum Euclidean distance to the theoretical absolute worst-case point. We assigned each resulting layer exactly one centroid, paired with the updating of that centroid with a progressively decreasing α value. While this distance-based variant achieved comparable results, its performance was slightly inferior to the standard WCA-K-Means. This deficit occurs because the energy system model treats multi-dimensional edge cases along the Pareto-front as operationally significant, demonstrating that simply utilising the distance to the absolute worst-case point inadequately captures these distinct multi-objective extremes.

5 Conclusion

This study proposes three worst-case adapted (WCA) clustering algorithms for worst-case artificial point aggregation, and evaluates the performance of each. Our main empirical finding is that the newly proposed Worst-Case Adaptive (WCA) K-Means framework outperforms alternative WCA variants and has mixed performance compared to existing clustering algorithms. Evaluating WCA-K-Means within the Tulipa energy model framework shows it can closely reconstruct the system's true Loss of Load and optimal capacity expansion portfolios using fewer representative periods, but this advantage decreases when the problem requires more clusters. By reconstructing near-optimal investment and reliability outcomes with fewer representative periods, this framework offers a practical approach to mitigating the capital-intensive expenditures [3] that threaten the EU's projected €1.2 trillion energy transition [2].

The operational success of WCA-K-Means stems from its multi-objective design. Rather than relying on simple distance-based minimisation, the algorithm updates centroids to a synthetic worst-case scenario exclusively for clusters that reside along the multi-dimensional Pareto-fronts (characterised by high demand and low renewable resource availability). Unlike the static post-processing approaches of Fazlollahi et al. [5] and Dominguez-Munoz et al. [9], which manually append extreme days as isolated clusters, our updating mechanism dynamically preserves these critical operational boundaries within the clustering process itself. This targeted adjustment ensures that the respective cluster centroids accurately represent standard historical trends, while the algorithm preserves critical scenarios as well.

When we benchmark the three WCA variants against standard K-Means, K-Medoids, K-Medoids WC, and Hull clustering using purely *a priori* metrics, the proposed algorithms exhibit equal or diminished cluster compactness alongside slower convergence rates. However, their *a posteriori* outcomes vindicate these structural compromises, except for WCA-Hierarchical clustering. While standard reduction techniques inherently smooth out extreme operational days, risking severe capacity under-investment, the WCA frameworks yield resilient infrastructure plans. Ultimately, these findings strongly reinforce the consensus in the literature that *a priori* data-handling metrics are often poor predictors of actual *a posteriori* model performance [20], making worst-case clustering essential for robust solving of energy systems.

Future Work

While WCA-K-Means shows signs of improvement over existing clustering algorithms, this difference diminishes when the problem requires more clusters. Similarly, WCA-K-Means can converge to a local optimum, suboptimally creating an investment plan. Therefore, researchers need to further investigate the implications of alternative initialisation heuristics and parallelised optimisation techniques to mitigate local optima convergence while reducing the overall computational overhead. Additionally, conducting more case studies on diverse datasets is necessary to verify the algorithms' generalisability.

References

[1] European Commission. Commission staff working paper: Impact assessment, accompanying the document energy roadmap 2050. Technical Report SEC(2011) 1565 final, European Commission, Brussels, 2011.

[2] European Commission. Proposal for a Regulation of the European Parliament and of the Council on guidelines for trans-European energy infrastructure, amending Regulations (EU) 2019/942, (EU) 2019/943 and (EU) 2024/1789 and repealing Regulation (EU) 2022/869. COM(2025) 1006 final, CELEX: 52025PC1006, December 2025. Accessed: 2026-06-09.

[3] Nestor A. Sepulveda, Jesse D. Jenkins, Fernando J. de Sisternes, and Richard K. Lester. The role of firm low-carbon energy resources in deep decarbonization of electric power systems. *Joule*, 2(11):2403–2420, 2018.

[4] Leander Kotzur, Peter Markewitz, Martin Robinius, and Detlef Stolten. Impact of different time series aggregation methods on optimal energy system design. *Renewable Energy*, 117:474–487, 2018.

[5] Samira Fazlollahi, Stephane Laurent Bungener, Pierre Mandel, Gwenaelle Becker, and François Maréchal. Multi-objectives, multi-period optimization of district energy systems: I. selection of typical operating periods. *Computers & Chemical Engineering*, 65:54–66, 2014.

[6] Milad Riyahi and Alvaro Gutiérrez Martín. Optimizing capacity expansion modeling with a novel hierarchical clustering and systematic elbow method: A case study on power and storage units in Spain. *Applied Energy*, 353:122099, 2024.

[7] Peng Du, Fenglian Li, and Jianli Shao. Multi-agent reinforcement learning clustering algorithm based on silhouette coefficient. *Neurocomputing*, 596:127901, 2024.

[8] Maximilian Hoffmann, Leander Kotzur, and Detlef Stolten. The pareto-optimal temporal aggregation of energy system models. *Applied Energy*, 315:119029, 2022.

[9] Fernando Domínguez-Munoz, José M. Cejudo-López, Antonio Carrillo-Andrés, and Manuel Gallardo-Salazar. Selection of typical demand days for chp optimization. *Energy and Buildings*, 43:92–101, 2011.

[10] Ian J. Scott, Pedro M. S. Carvalho, Audun Botterud, and Carlos A. Silva. Clustering representative days for power systems generation expansion planning: Capturing the effects of variable renewables and energy storage. *Applied Energy*, 253:113599, 2019.

[11] Holger Teichgraeber, Constantin P. Lindenmeyer, Nils Baumgärtner, Leander Kotzur, Detlef Stolten, Martin Robinius, André Bardow, and Adam R. Brandt. Extreme events in time series aggregation: A case study for optimal residential energy supply systems. *Applied Energy*, 262:114423, 2020.

[12] Maaike Elgersma, Luca Santosuosso, Sonja Wogrin, Germán Morales-España, Mathijs de Weerdt, Greg Neustroev, and Lotte Kremer. Obtaining upper bounds for gep with storage fast: Performance guarantees for tsa based on the worst case. Working draft, 2026.

[13] N. Srinivas and Kalyanmoy Deb. Multiobjective optimization using nondominated sorting in genetic algorithms. *Evolutionary Computation*, 2(3):221–248, 1994.

[14] Qun Yu, Qiang Zheng, Baiwei Feng, Zuyuan Liu, and Haichao Chang. Objective reduction method based on K-means clustering and its application on hull form optimization. *Ocean Engineering*, 341:125736, 2026.

[15] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197, 2002.

[16] Diego A. Tejada-Arango, Germán Morales-España, Lauren Clisby, Ni Wang, Abel S. Siqueira, Ali Subayu, Laurent Soucasse, and Zhi Gao. Tulipa energy model: Mathematical formulation. arXiv preprint arXiv:2309.07711 [math.OC], 2023.

[17] Sven Butzelaar. Extreme-preserving hierarchical clustering for automated temporal partitioning in energy system optimization. Master's thesis, Delft University of Technology, Delft, Netherlands, 2026. Supervisors: Dr. Germán Morales-España and Maaike Elgersma. Jointly conducted with TNO.

[18] S. Lo Piano and S. T. Smith. Energy demand and its temporal flexibility: Approaches, criticalities and ways forward. *Renewable and Sustainable Energy Reviews*, 160:112249, 2022.

[19] TulipaEnergyModel.jl Developers. Obz dataset. https://github.com/TulipaEnergy/TulipaEnergyModel.jl/tree/main/docs/src/data/obz, 2026.

[20] Maximilian Hoffmann, Jan Priesmann, Lars Nolting, Aaron Praktiknjo, Leander Kotzur, and Detlef Stolten. Typical periods or typical time steps? a multi-model analysis to determine the optimal temporal aggregation for energy system models. *Applied Energy*, 304:117825, 2021.

[21] Greetjekoffie. Tulipaclustering.jl. https://github.com/Greetjekoffie/TulipaClustering.jl, 2026.

A Responsible Research

We wrote this paper with strict ethical, transparency, and reproducibility requirements in mind. To guarantee full reproducibility, we have made the entire modelling pipeline, data cleaning routines, and raw results open-source. Readers can access all corresponding data and code from the project's public repository [21].

Computational Environment

We conducted all computational experiments on an HP laptop equipped with an Intel i7 CPU, 16GB of RAM, and a Windows operating system. We implemented the algorithms entirely within the Julia ecosystem using Julia v1.12.6. The custom temporal aggregation routines used the Tulipa Clustering framework (v0.5.2), and we executed the investment simulations using the Tulipa Energy Model framework (v0.21.0). We used Julia's native `Project.toml` and `Manifest.toml` files to lock the package setup, ensuring the code runs in the exact same environment.

Case Study and Data Sources

The validation of the methodology relies on a high-dimensional, multi-regional case study. The dataset comprises hourly operational profiles drawn directly from the Tulipa Energy Model repository [19] and contains detailed historical data tracking onshore/offshore wind, solar levels, and more. We also archive the exact dataset used for the baseline benchmark simulations [21].

Data Integrity and Algorithmic Provenance

To prevent data manipulation bias and ensure an objective comparative evaluation, we performed no heuristic filtering, manual smoothing, or arbitrary outlier removal. The only change made to a parameter of the dataset is the increased cost for the energy not served variable, as in the original data this value was too low. This resulted in an unrealistically high Loss of Load. We changed the value from 0.18 to 3. Furthermore, the algorithms generate the synthetic worst-case data points using deterministic methods, ensuring they produce the same results every time.

Global Socio-Ethical Implications

Beyond the algorithmic dimensions, optimising energy system models carries ethical responsibilities toward the global population. Energy planning decisions directly dictate grid reliability, thereby shaping the socioeconomic well-being of societies. Underestimating worst-case operational conditions in highly renewable systems risks catastrophic grid collapses that burden the world's population. Ultimately, the research in this paper aims to safeguard these systems against failure, ensuring that future power grids are robust.

Use of Generative Artificial Intelligence

In alignment with transparency standards, we utilised AI tools exclusively in a supportive way during the creation of this work. We used AI lightly to assist with language editing, syntax smoothing, and grammar checks. Coding-wise, AI use covered only standard boilerplate code structures, plot generation, and basic debugging support within the Julia scripting pipeline. Crucially, we created all core algorithmic innovations without the use of AI.

B Full plots

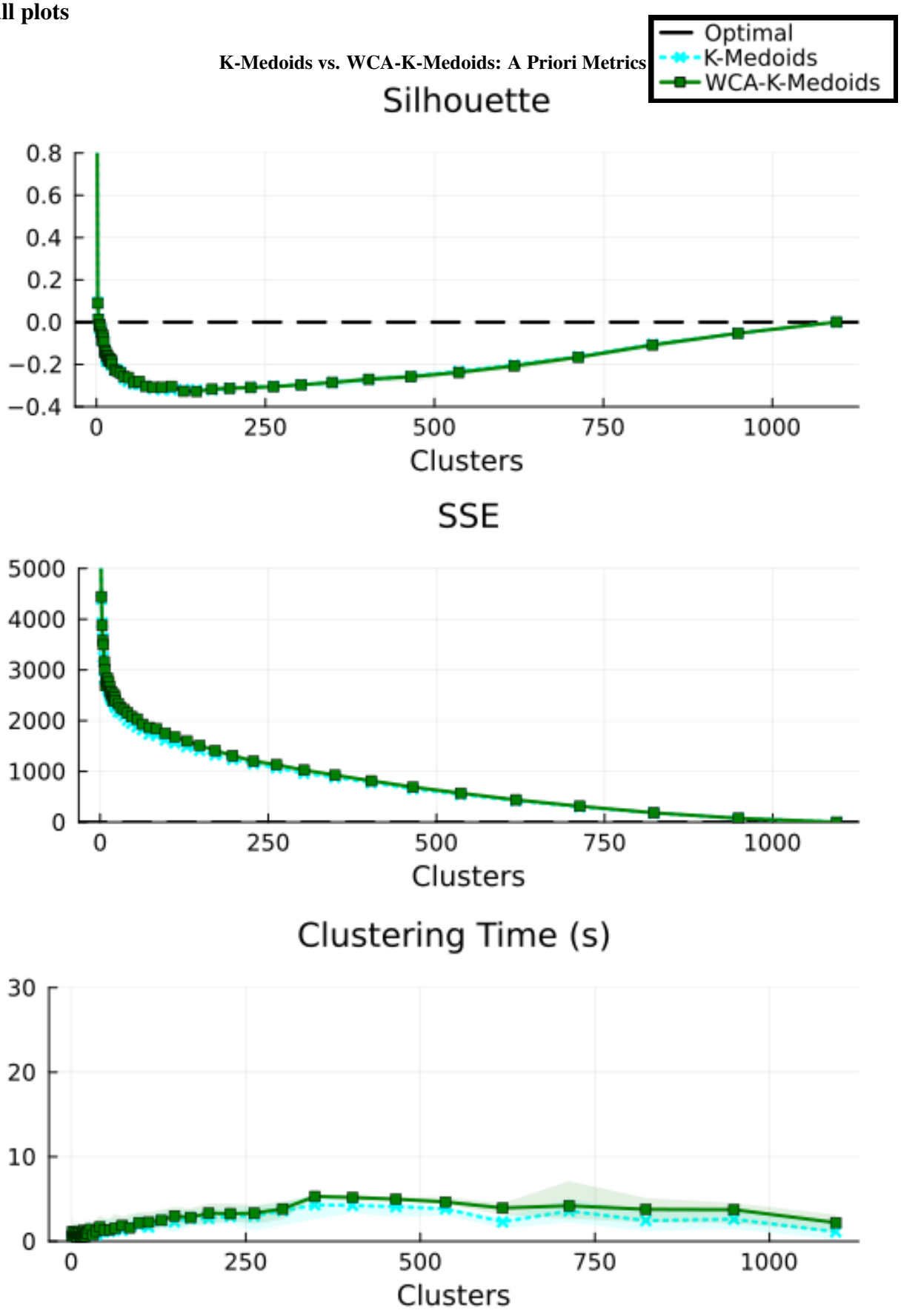


Figure 12: Full-range a priori evaluation metrics (Silhouette, SSE, Clustering Time) for K-Medoids and WCA-K-Medoids across the complete cluster range $k \in [1, 1095]$, including error margins across five random seeds.

Figure 13: Full-range a posteriori evaluation metrics (Regret, total LoL, total costs) for K-Medoids and WCA-K-Medoids across the complete cluster range $k \in [1, 1095]$, including error margins across five random seeds.

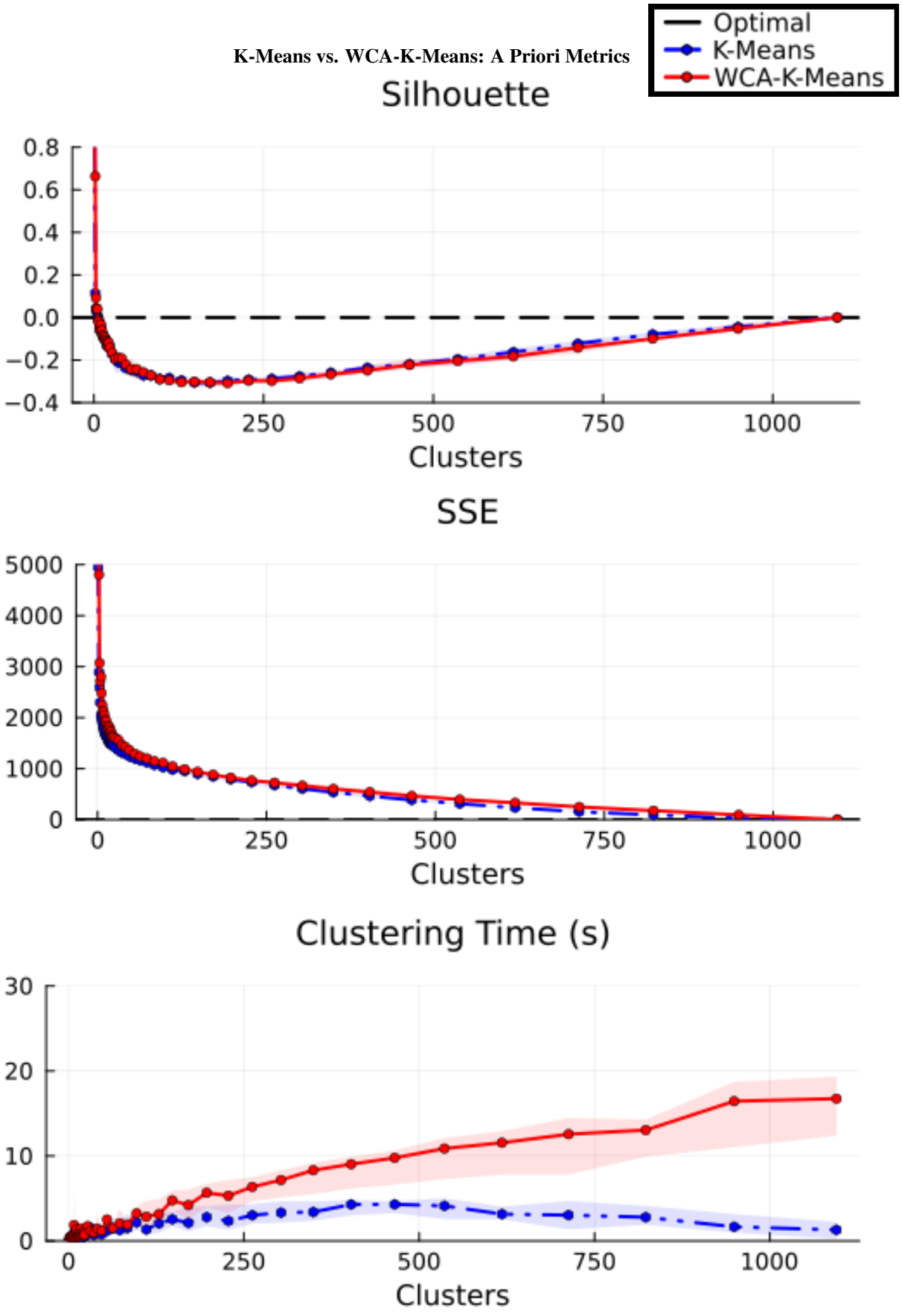


Figure 14: Full-range a priori evaluation metrics (Silhouette, SSE, Clustering Time) for K-Means and WCA-K-Means across the complete cluster range $k \in [1, 1095]$, including error margins across five random seeds.

K-Means vs. WCA-K-Means: A Posteriori Performance

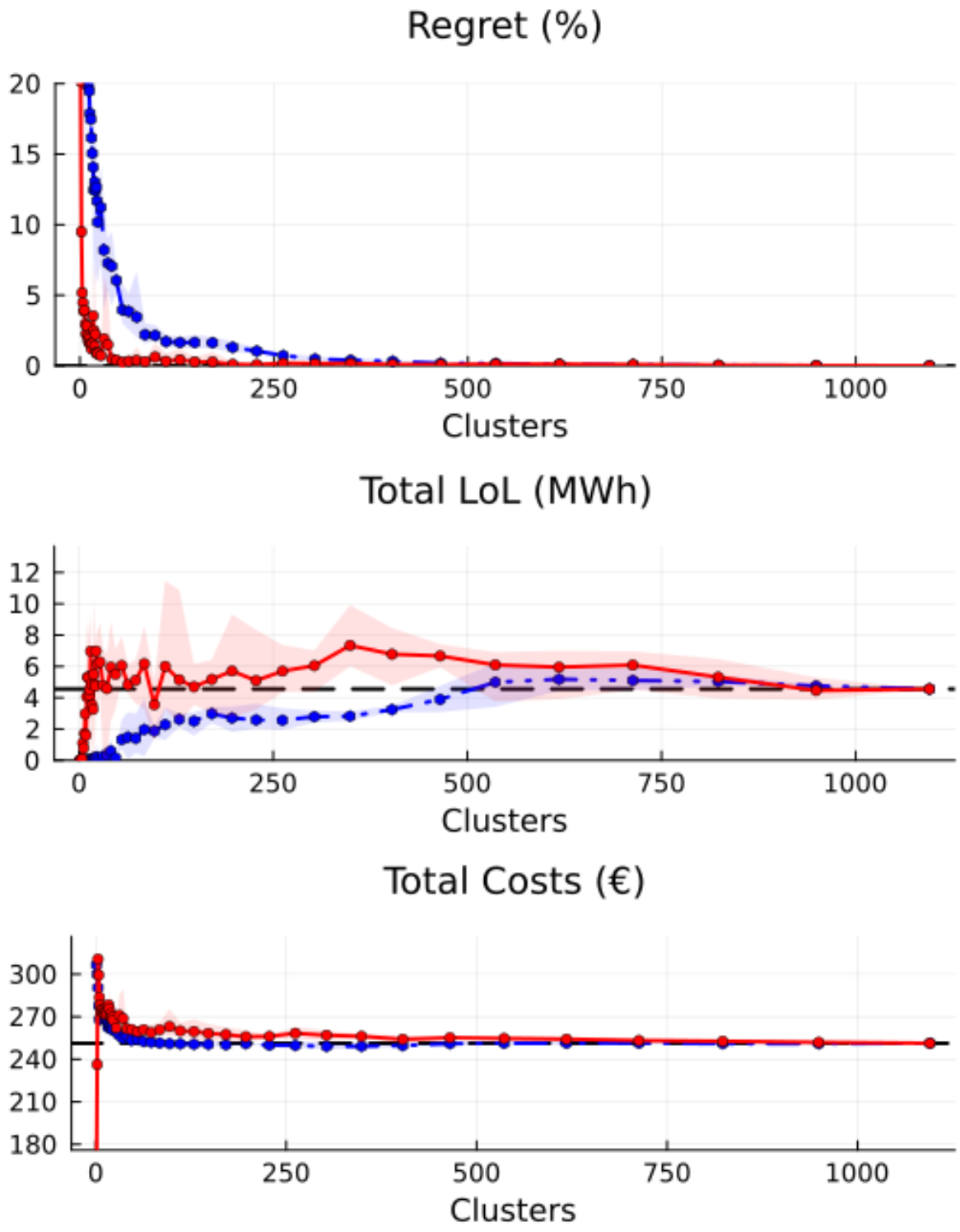


Figure 15: Full-range a posteriori evaluation metrics (Regret, total LoL, total costs) for K-Means and WCA-K-Means across the complete cluster range $k \in [1, 1095]$, including error margins across five random seeds.

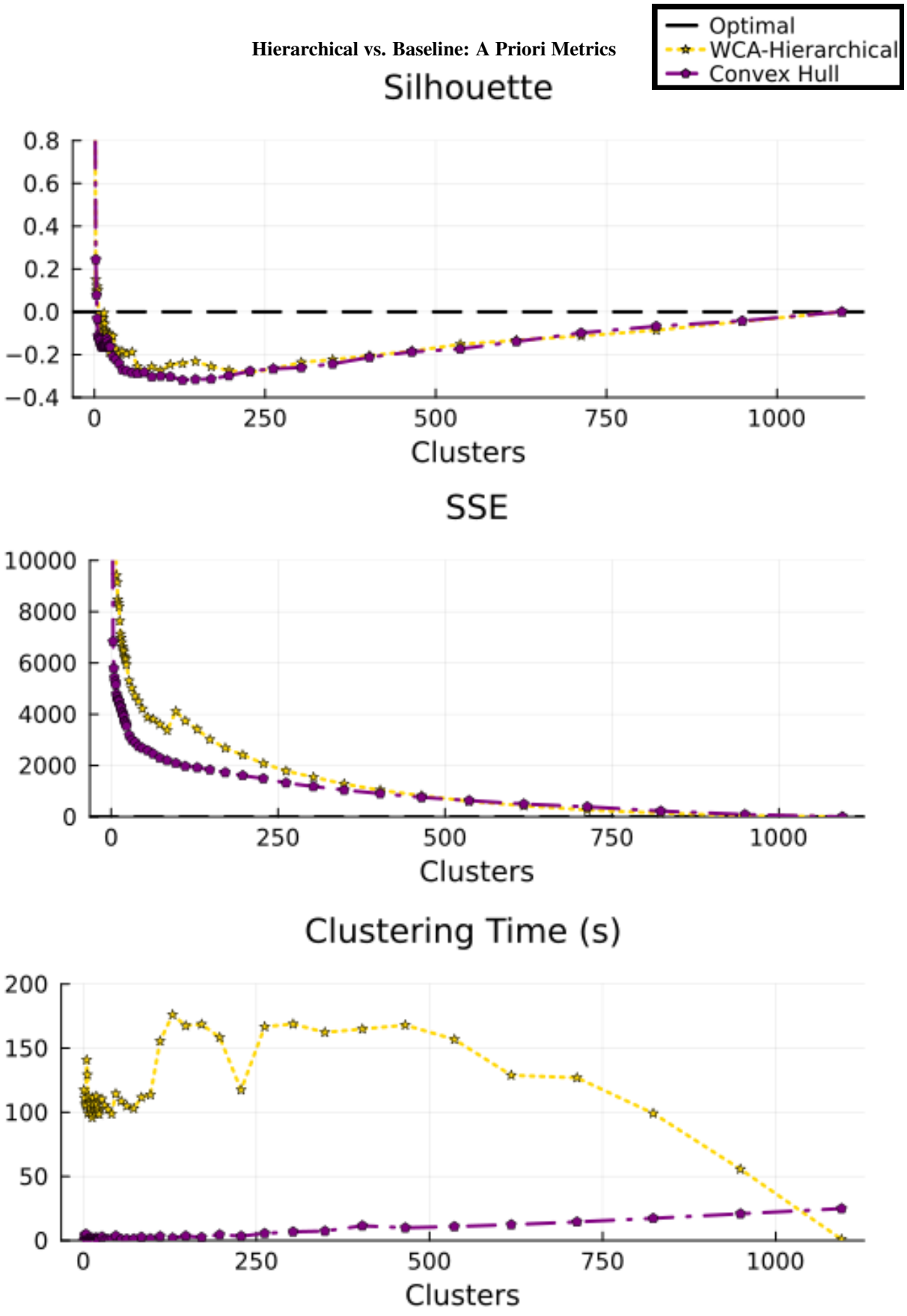


Figure 16: Full-range a priori evaluation metrics (Silhouette, SSE, Clustering Time) for Hierarchical clustering and Hull clustering across the complete cluster range $k \in [1, 1095]$, including error margins across five random seeds.

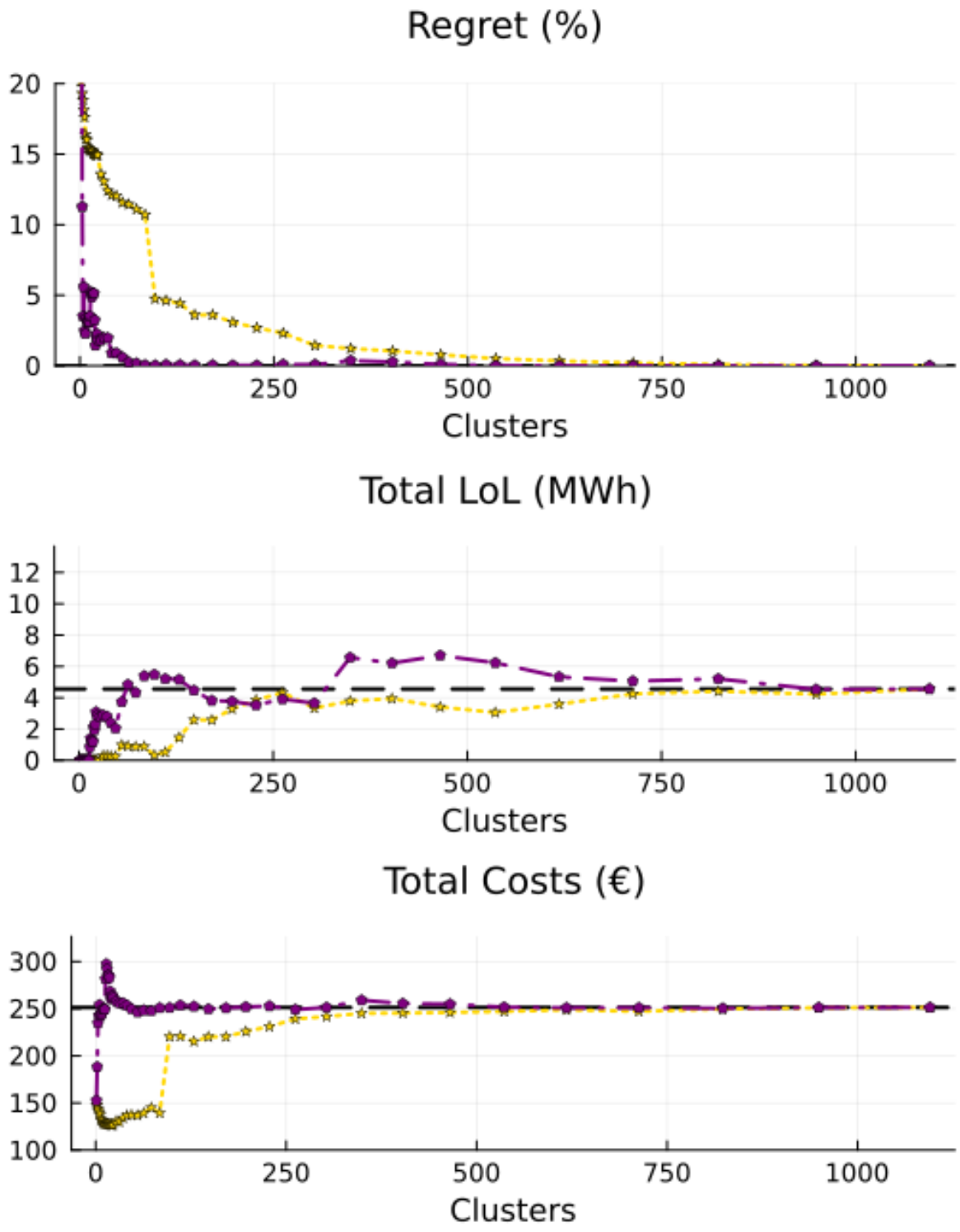


Figure 17: Full-range a posteriori evaluation metrics (Regret, total LoL, total costs) for Hierarchical clustering and Hull clustering across the complete cluster range $k \in [1, 1095]$, including error margins across five random seeds.

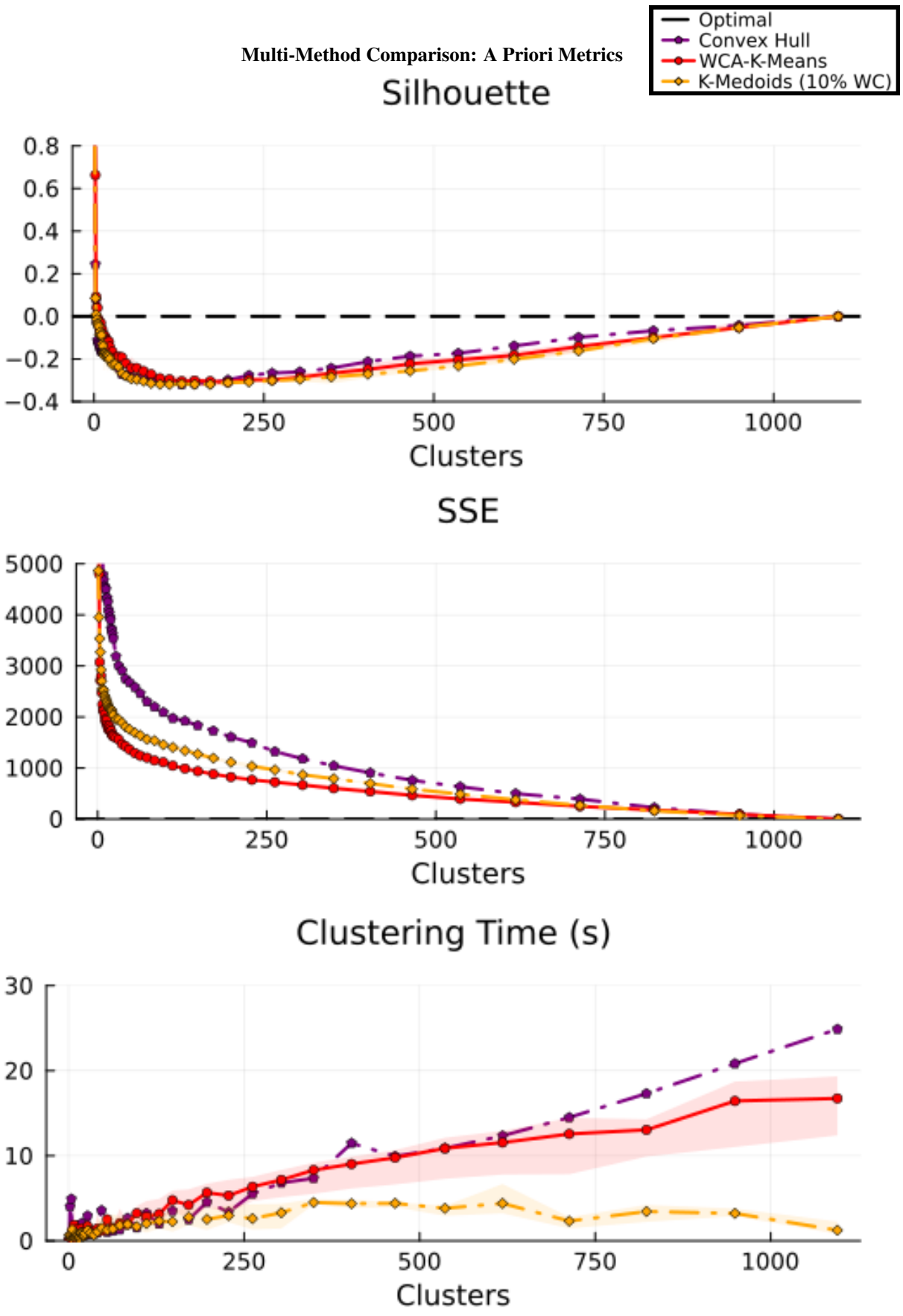


Figure 18: Full-range a priori evaluation metrics (Silhouette, SSE, Clustering Time) for WCA-K-Means, K-Medoids WC, and Hull clustering across the complete cluster range $k \in [1, 1095]$, including error margins across five random seeds.

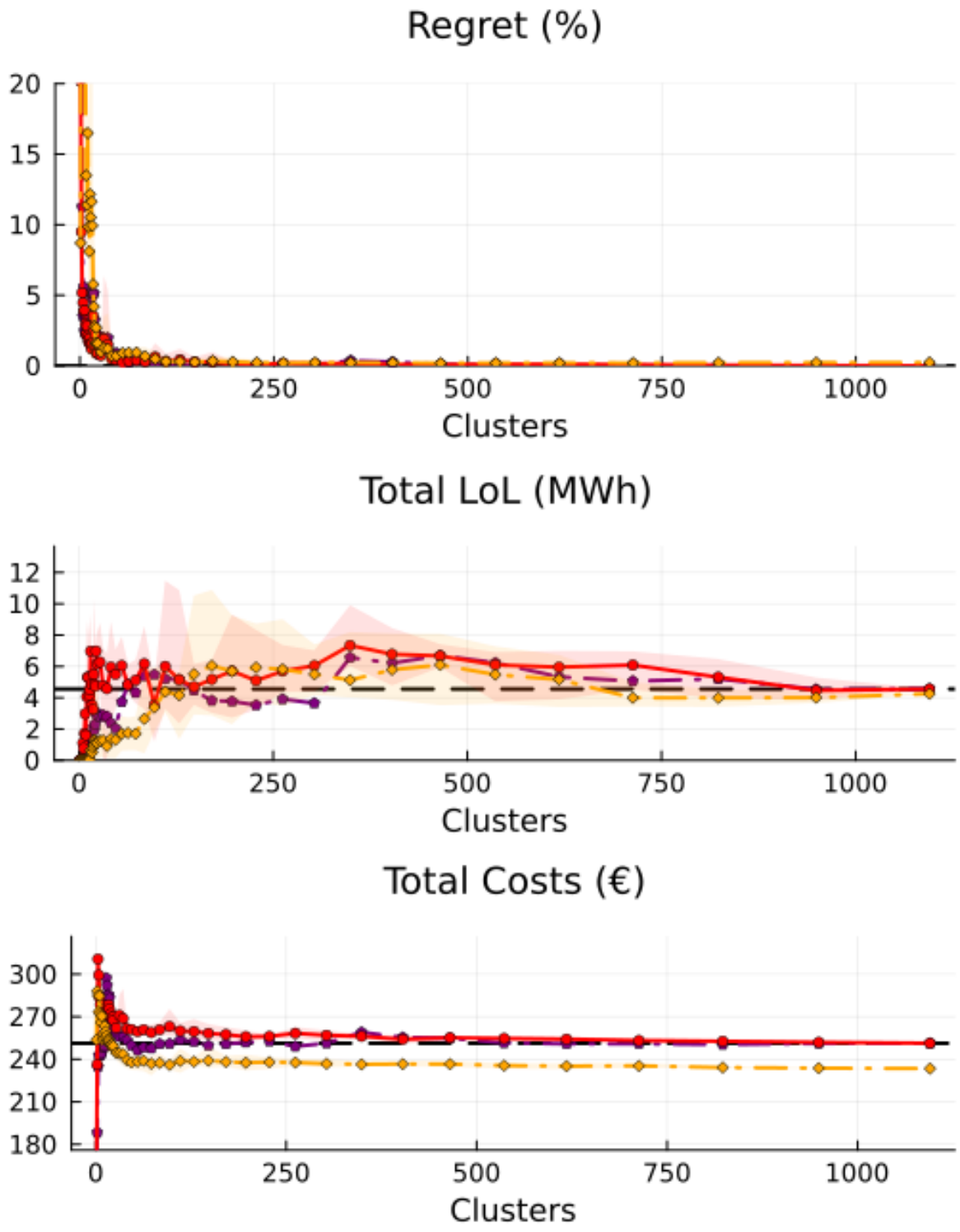


Figure 19: Full-range a posteriori evaluation metrics (Regret, total LoL, total costs) for WCA-K-Means, K-Medoids WC, and Hull clustering across the complete cluster range $k \in [1, 1095]$, including error margins across five random seeds.