On the Selection of Intermediate Length Representative Periods for Capacity Expansion

Osten Anderson, Nanpeng Yu
Department of Electrical and Computer Engineering
University of California, Riverside
Riverside, CA

Konstantinos Oikonomou, Di Wu
System Evaluation Group
Pacific Northwest National Laboratory
Richland, WA

***Abstract*—As the decarbonization of power systems accelerates, there has been increasing interest in capacity expansion models for their role in guiding this transition. Representative period selection is an important component of capacity expansion modeling, enabling computational tractability of optimization while ensuring fidelity between the representative periods and the full year. However, little attention has been devoted to selecting representative periods longer than a single day. This prevents the capacity expansion model from directly simulating interday energy sharing, which is of key importance as energy generation becomes more variable and storage more important. To this end, we propose a novel method for selecting representative periods of any length. The method is validated using a capacity expansion model and production cost model based on California's decarbonization goals. We demonstrate that the representative period length has a substantial impact in the results of the capacity expansion investment plan.**

***Index Terms*—Capacity expansion planning, representative period selection, production cost modeling.**

I. Introduction

A. Background & Problem Statement

Decarbonization of power grids has been identified as a critical component of the response to climate change, and many governments have adopted laws to this end. For example, the California State Legislature passed SB100 in 2018 and set the target of 60% by 2030 and 100% by 2045 of retail electricity sales from renewable sources [1]. This requires substantial investment in green technology, especially in renewable generation and energy storage, with the value of investment and operation over this timeline on the order of 100s of billions of US dollars. As a result, effective planning of the investment rollout is of critical importance for minimizing cost, ensuring reliable operation, and meeting policy regulations.

Models which enable this planning are known as capacity expansion models (CEMs), and are related to the task of generation expansion planning. These models have been used for decades, but have been subject to additional attention in recent years for their role in guiding the transition to lower-carbon, high-renewable grids.

CEMs seek to optimize generation and transmission investment strategies, and typically model two timescales: annual investment decisions and hourly operation decisions. As a result of the hourly timescale, CEMs can quickly become intractable if all 8760 hours in a year are modeled. To address this, CEMs will typically utilize representative periods instead of all 8760 hours. The main goal in selecting these representative periods is to maximize the similarity between the annual behavior and the surrogate representation, while achieving a sufficient reduction in the associated computational load.

B. Related Work & Paper Contribution

The problem of representative period selection in CEMs has received much attention in the literature. The majority of the work on this topic has been built around the framework of time series clustering. The authors in [2] compare a variety of clustering methods. The authors in [3] present a comparison of clustering and downsampling approaches. In [4], a clustering method is proposed which requires that each cluster consist of a contiguous set of days, while a downsampling-based approach is proposed in [5]. These clustering-based approaches all share the common drawback that they are only suitable for the selection of representative days, and not periods of multiple days in length, as will be discussed in detail later.

In addition, several works have focused on modeling the full year contiguously to allow for the tracking of interday and long-term energy storage by reducing the modeling frequency [6], [7]. The key drawback to these works is that they lose the sequentiality of hours, and thus cannot model important inter-hour details, such as ramping in power plants.

The references cited above primarily concentrate on the selection of representative days and offer general assertions regarding the algorithm's ability to choose periods of varying lengths, such as a representative day or week. However, let us consider the case where the desired period is of an intermediate length, such as 3 days. The time series clustering framework upon which the majority of representative period selection algorithms are built, requires the full time series to be divided into subsequences. Clustering these subsequences becomes highly dependent on the starting point. In particular, the load exhibits significant differences in both shape and magnitude between weekdays and weekends. For example, the Euclidean distance of a Friday-Saturday-Sunday subsequence to a Saturday-Sunday-Monday subsequence would likely be large because the loads of Friday would be compared to Saturday and the loads of Sunday would be compared to Monday. On the other hand, clustering of overlapping subsequences, obtained by sliding a window across the full time

series with a stride shorter than the subsequence length, has been established to return essentially random results [8]. Even in the case of representative weeks, there are considerable drawbacks. Generally speaking, capacity expansion problems tend to reduce the annual temporal coverage to roughly 10% or less. In the case of days, this permits 37 days, but in the case of weeks, this permits only 5 weeks. Intuitively, one would expect that representing a full year is more difficult given 5 choices than 37 choices.

Considering the limitations of existing clustering methods, along with the need to capture interday variability in CEMs subject to high levels of energy storage and renewables penetration, the selection of representative periods longer than one day becomes imperative. Indeed, enabling interday sharing of energy through storage modeling is a crucial, yet often ignored, aspect that would become particularly valuable during days of low generation from renewable resources.

While the selection of a representative period length is fundamentally an experimental design decision, there exists a noticeable gap in research when it comes to effectively choosing a period longer than a day but shorter than a week. To bridge this knowledge gap, we introduce a novel snippet algorithm specifically designed for selecting representative periods that extend beyond a single day. By comparing subsequences instead of full sequences, the proposed snippet algorithm is able to select representative periods of arbitrary length from complex datasets. The proposed algorithm draws significant inspiration from [9]; however, we have made several tailored adjustments to accommodate the unique domain to which our proposed algorithm is applied.

The remainder of this paper is structured as follows. Section II provides an overview of the original time series snippets algorithm, its differences from the proposed algorithm, and the proposed algorithm itself. Section III discusses the numerical study setup and results. Section IV presents the conclusions.

II. Technical Method

A. Overview of time series snippets

The discussion of time series snippets below provides a concise overview of the algorithm that served as the inspiration for the proposed method in Section II-B. The time series snippets algorithm is built on top of the matrix profile distance (MPdist) [10] measure, which is in turn built on top of the matrix profile [11]. MPdist compares two-time series and considers them to be similar if they have similar subsequences. At its most basic level, the distance is the j-th smallest Euclidean distance between subsequences. More specifically, the goal of time series snippets is to select, from a time series T with length t, subsequences of length s that best generalize the full-time series. First, the full-time series is separated into non-overlapping subsequences S_i with $i \in [0, t/s - 1]$. Each of these subsequences then has an MPdist profile $MPdist_i$ compared to the full-time series. If each $MPdist_i$ were plotted, the goal would be to select the k profiles that minimize the area under the curve of the combined profiles, as shown in Fig. 1. To select these representative subsequences, a greedy algorithm is proposed, choosing the subsequence that gives the greatest reduction in the cumulative distance in each iteration.

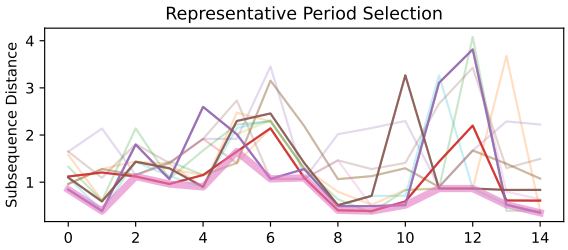

Fig. 1. Visualization of distance profiles and minimization of area under the curve of selected profiles.

The key contribution of these matrix profile-related methods is that they scale well to extremely long time series. The problem of representative period selection for power system planning typically considers one year of data at hourly frequency, for a time series of length 8760, which is extremely short in that context. Further, we don't need to calculate the distance measure for every subsequence, as we can exploit the known daily periodicity of our time series. For this reason, we can calculate a distance measure similar to the $MPdist$ without relying on algorithms related to the matrix profile. This also enables us to utilize overlapping subsequences as S_i. Finally, the problem size allows us to select the representative snippets through convex optimization rather than relying on a greedy algorithm.

B. Proposed Method

Let $T = \{T[0], \dots T[h], \dots, T[t-1]\}$ represent the yearly multivariate time series of length t and $T[h]$ be the tuple of measurements at hour $h \in [0, t-1]$. This tuple typically encompasses load, solar generation, and wind generation information; however, the proposed method remains agnostic to the input features, providing adaptability in the analysis. Let also $S = \{S_0, \dots S_j, \dots S_{m-1}\}$ be the set of subsequences, and u the stride of the window that generates subsequences of length s. The subsequence S_j is then defined as:

$$S_j = T[j \cdot u : j \cdot u + s] \tag{1}$$

There will be $m = \frac{t-s}{u} + 1$ total subsequences, thus $j \in [0, m-1]$. These will be the candidate subsequences used for selecting representative periods. Similarly, we can define non-overlapping subsequences of T, which we will refer to as segments. There will be $n = \frac{t}{u}$ such segments, thus $i \in [0, n-1]$. The segment T_i is then defined as:

$$T_i = T[i \cdot u : (i+1) \cdot u] \tag{2}$$

A visual representation of the definitions can be seen in Fig. 2. For clarity, i will be reserved to index the time series segments T_i and j to index the subsequences S_j.

Inspired by the MPdist, let $\mathcal{D}$ be a matrix of distances with shape $n \times m$. For our case, we will assume u and s are chosen such that n, m, and s/u are integers. We also have domain

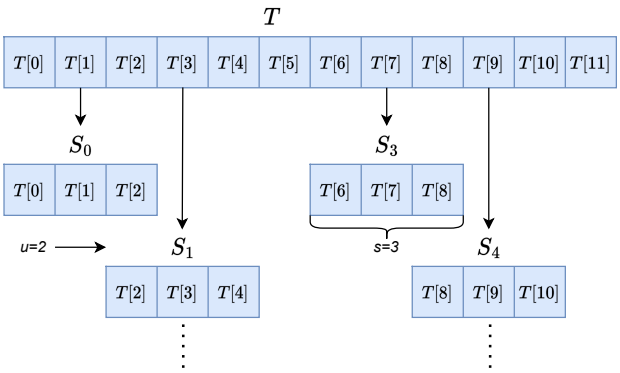

Fig. 2. Visualization of an example time series T and subsequences S_j with $u = 2$, $s = 3$.

knowledge of periodicity. Each of the features, especially load and solar generation, have strong 24-hour cycles. It is unlikely that an afternoon subsequence from one day would be similar to a nighttime subsequence from another day. Further, this is physically meaningless in the context of capacity expansion. For this reason, whereas the MPdist compares subsequences for every timestep, we apply stride $u = 24$ in calculating the distances. In essence, this compares each day in the subsequence to each day in the full time series, and assigns a distance correspondingly.

$$\mathcal{D}_{i,j} = \min_{x} ||S_j[x \cdot u : (x+1) \cdot u] - T_i||, x \in \left[0, \frac{s}{u}\right) \quad (3)$$

The goal is then to find a subset of those candidate days which best captures the patterns for the year as a whole. This goal is the same as the one visualized in Fig. 1. Time series snippets were originally proposed with a greedy algorithm that iteratively selects the subsequence which minimizes the cumulative sum of distances to the full time series, necessitated by the long time series the algorithm was designed for. Because our time series is rather short, we can instead formulate this problem as a mixed integer linear program and find the solution using any suitable optimization solver.

$$\begin{aligned}
&\min_{\alpha} \sum_{i=0}^{n-1} dist_i \quad (4)\\
&s.t. \sum_{j=0}^{m-1} \alpha_j = k\\
&dist_i = \sum_{j=0}^{m-1} md_{i,j} \cdot \ \mathcal{D}_{i,j}, \ \forall i \in [0, n)\\
&\sum_{j=0}^{m-1} md_{i,j} = 1, \ \forall i \in [0, n)\\
&md_{i,j} \leq \alpha_j, \ \forall i \in [0, n), \ \forall j \in [0, m)\\
&dist \in \mathbb{R}^n, \ \ md \in [0,1]^{n \times m}, \ \ \alpha \in [0,1]^m,
\end{aligned}$$

where k is the desired number of representative periods; α_j is a binary indicator selecting subsequence S_j as a representative period; $dist_i$ is the minimum distance between the selected representative periods to T_i; and $md_{i,j}$ is a binary indicator signifying that subsequence S_j has the smallest distance to day T_i. Within CEM, representative periods are typically weighted by the amount of the year that they account for. The weights associated with each representative period are a function of $md_{i,j}$, and can be written then as:

$$w_j = \sum_{i=0}^{n-1} md_{i,j}/(s/u), \quad (5)$$

where s/u ensures the weights sum to 365.

III. Experimental Validation

A. Experimental Setup

To the author's knowledge, no paper has made a dedicated attempt to address the sampling of intermediate-length representative periods in capacity expansion planning. This absence poses a challenge when comparing the proposed method with widely used state-of-the-art approaches. We will compare the performance of the proposed algorithm to a popular method for selecting representative days, and show that our algorithm is at least comparable with the state-of-the-art for this task. The proposed method will also be used to compare single-day planning to multi-day planning. The goal of this comparison is to show the value in simulating representative periods longer than one day, particularly in sizing energy storage. Our goal is not necessarily to show the optimality of a particular representative period length, but rather to demonstrate the differences between period lengths on investment plans and operational cost.

The following general experimental design will be used to validate the proposed method. First, the representative days are selected and used within the CEM. Then, the investment decisions are fixed, and the model is solved again as a production cost model. The production cost model (PCM) is ran for the full year in two-week stages, and results from this will be referred to as fullspace results. The choice of two weeks is somewhat arbitrary, with the key being that this length is considerably longer than each of the candidate representative period lengths to avoid giving bias towards any particular length.

B. Simulation Models

The PCM and CEM used are both zonal models of the Western Interconnection, primarily focused on California. Both utilize a MILP adaptation of the formulation and data of E3's RESOLVE decarbonization model [12]. Further discussion of the decarbonization model including full formulations of the objective and constraints are available in [13]. The principle goal of both models is minimizing the cost of serving load. The PCM focuses on operational decisions, including the scheduling of power plants, to minimize operating costs while satisfying operating and reliability constraints (6). Operating constraints maintain safe resource limits (e.g., power plant

capacities), while reliability constraints guarantee zonal power balance and ancillary services for secure power supply.

$$\begin{aligned} \min \quad & C_{gen} \\ \text{s.t.} \quad & \text{Operating constraints} \\ & \text{Reliability constraints} \end{aligned} \tag{6}$$

The CEM essentially adds an additional level to the PCM by allowing investment in additional generation capacity. It solves operational and investment decisions to minimize the total cost of generation C_{gen}, maintenance C_{maint} and investment C_{inv} (7). Investment decisions include both the addition of new resource capacity and retirement of existing gas generators. Policy constraints are integral to a decarbonization model, and include emissions limits and renewable portfolio standards.

$$\begin{aligned} \min \quad & C_{inv} + C_{maint} + C_{gen} \\ \text{s.t.} \quad & \text{Policy constraints} \\ & \text{Operation constraints} \\ & \text{Reliability constraints} \end{aligned} \tag{7}$$

C. Results and Discussions

First, we compare our method to a popular approach for representative day selection: k-means clustering using load, wind, and solar profiles with medoid cluster center representation used in the CEM. We use our proposed algorithm with subsequence length $s = 24$ to select 21 representative days. Our algorithm has a total cost (investment, maintenance, fullspace operations) of 14.560 billion US dollars as compared to 14.611 billion US dollars. [2] compared many clustering techniques for power system planning, and established that there are not clear patterns on which technique is best, and many have comparable performance. With this in mind, we can suggest that even for representative day selection, our proposed method is at least comparable with one of the most commonly used representative day selection approaches.

With the validity of the proposed method established, we now seek to defend the motivation behind selection of longer representative periods. This will be explored via investment, fullspace operation cost, emissions, and the investment portfolio, considering representative periods of 1 to 5 days, i.e. with subsequence length $s \in [24, 48, 72, 96, 120]$. In each case, periods are adjusted to model a total of 35 or 36 days, nearly 10% of the annual days.

Fullspace emissions are shown in Fig. 3. None of the scenarios meet fullspace emissions limits. The primary driver of emissions is investment in renewable generation and energy storage. Once these technologies are purchased, their use incurs no additional operational costs in the unit commitment model. However, the PCM is myopic in terms of emissions, and may take actions which lead to lower costs but higher emissions, such as export generation from thermal units in CAISO. As the fullspace model is run in discontinuous segments for reasons of computational tractability, it is impossible to effectively enforce emissions limits. It is difficult to say which, if any, of these fleets would be able to satisfy the

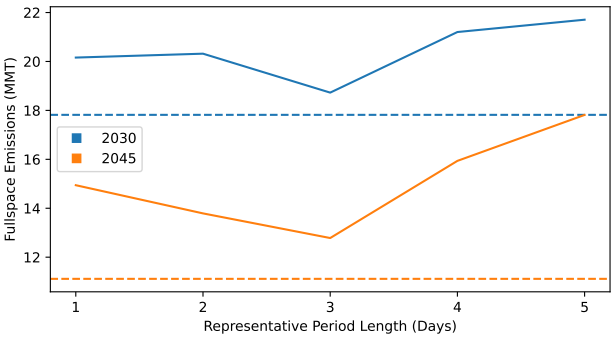

Fig. 3. Fullspace emissions

emissions limits. Still, it is notable that 3-day representative periods present the lowest emissions, and longer, and thus fewer, periods have substantially higher emissions. This suggests that by modeling an intermediate-length period, interday energy storage can be leveraged to lower emissions. However, as the length of period grows, the number of periods must shrink. Without a sufficient number of periods, it is difficult to select periods which represent the annual behavior sufficiently well. This is illustrated by the higher emissions in the 4 and 5-day cases, which sample 9 and 7 periods respectively.

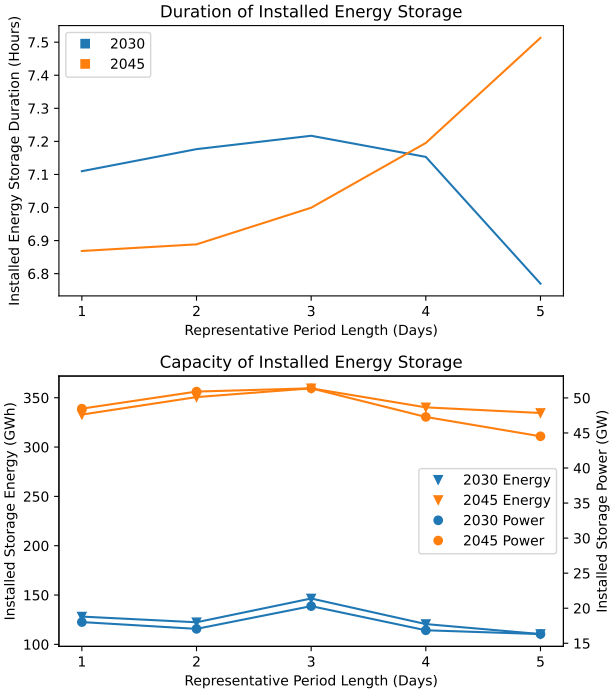

Fig. 4. Impact of representative period length on duration and capacity of installed energy storage.

The duration of installed storage, and installed power and energy capacity of storage as a function of representative period length for 2030 and 2045 are shown in Fig. 4. With regards to storage duration, the key takeaway is that increasing

the length of representative period allows for utilization of storage for interday energy sharing, and the duration increases for lengths between 1 and 3 days. However, the tradeoff between number and length of representative seems to impact the ability of the surrogate days to effectively represent the full year, leading to less predictable effects with lengths over 3 days. With regards to the power and energy capacity of storage, a similar pattern is evident. Between lengths of 1 and 3 days, the installed capacities generally increase, and then begin to decrease again. This result is in line with the emissions result.

Fig. 5 shows the cost by year for each scenario. As one would expect from the emissions violations visualization, $d = 4$ and $d = 5$ have the lowest overall cost due to less build of renewable technologies. Most notable is that the $d = 3$ result is very close to the $d = 1$ result despite larger investment. This suggests that by representing longer periods, it is better able to capture the fullspace value of interday energy sharing. Thus, the cost of additional investment is offset by lower operating costs. Specifically, the total costs for 2030 are 1.1% higher but have 7.1% lower emissions.

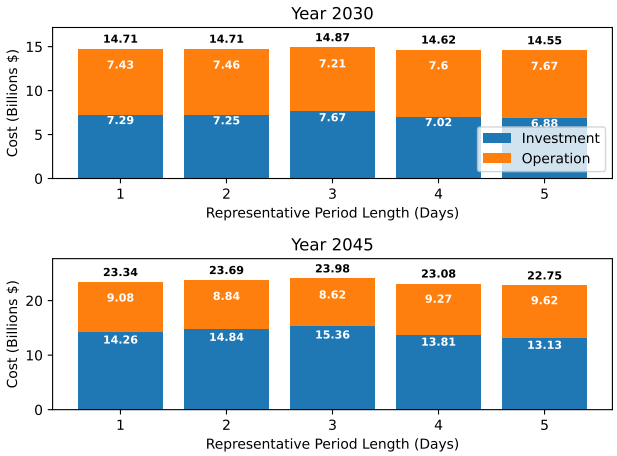

Fig. 5. Investment and operation costs of differing representative periods

Fig. 6 shows an elbow plot of the objective function of (4). Intuitively, for a given number of total modeled days $k \times s/u$, the objective is best for more, shorter representative periods. The gap between the lines is larger at the lower total modeled days and begins to converge at higher. This characteristic explains why, for a fixed total modeled days, the representation degrades with higher period length.

IV. Conclusion

In this paper, we proposed a novel algorithm for selecting representative periods. The algorithm is particularly directed towards selecting periods longer than a single day, and is well suited for planning in systems with high penetration of variable renewable energy and reliance on energy storage. The method chooses representative days which minimize a distance measure to the timeseries of the full year. The proposed method was validated on a CEM based on California's decarbonization

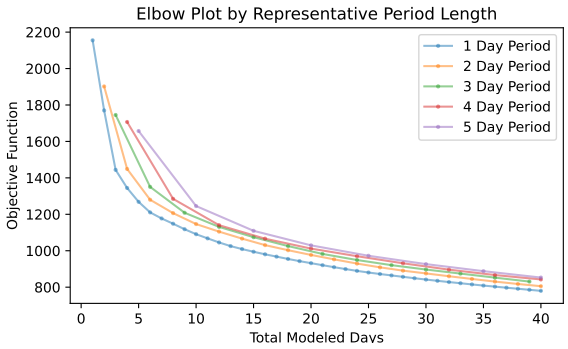

Fig. 6. Elbow plot of objective function (4) at different representative period lengths.

targets. The proposed method is competitive with the state-of-the-art for representative day selection, and we demonstrate the impact of representative period length on investment strategy.

References

[1] L. Gill, A. Gutierrez, and T. Weeks, "2021 SB 100 joint agency report: Achieving 100 percent clean electricity in California: An initial assessment," tech. rep., California Energy Commission, 2021.

[2] H. Teichgraeber and A. R. Brandt, "Clustering methods to find representative periods for the optimization of energy systems: An initial framework and comparison," *Appl. Energy*, vol. 239, pp. 1283–1293, 2019.

[3] S. Pfenninger, "Dealing with multiple decades of hourly wind and PV time series in energy models: A comparison of methods to reduce time resolution and the planning implications of inter-annual variability," *Appl. Energy*, vol. 197, pp. 1–13, 2017.

[4] W. W. Tso, C. D. Demirhan, C. F. Heuberger, J. B. Powell, and E. N. Pistikopoulos, "A hierarchical clustering decomposition algorithm for optimizing renewable power systems with storage," *Appl. Energy*, vol. 270, p. 115190, 2020.

[5] Y. Liu, R. Sioshansi, and A. J. Conejo, "Hierarchical clustering to find representative operating periods for capacity-expansion modeling," *IEEE Trans. Power Syst.*, vol. 33, no. 3, pp. 3029–3039, 2018.

[6] D. A. Tejada-Arango, M. Domeshek, S. Wogrin, and E. Centeno, "Enhanced representative days and system states modeling for energy storage investment analysis," *IEEE Trans. Power Syst.*, vol. 33, no. 6, pp. 6534–6544, 2018.

[7] S. Pineda and J. M. Morales, "Chronological time-period clustering for optimal capacity expansion planning with storage," *IEEE Trans. Power Syst.*, vol. 33, no. 6, pp. 7162–7170, 2018.

[8] E. Keogh, J. Lin, and W. Truppel, "Clustering of time series subsequences is meaningless: implications for previous and future research," in *Third IEEE International Conference on Data Mining*, pp. 115–122, 2003.

[9] S. Imani, F. Madrid, W. Ding, S. Crouter, and E. Keogh, "Matrix profile XIII: Time series snippets: A new primitive for time series data mining," in *2018 IEEE International Conference on Big Knowledge (ICBK)*, pp. 382–389, 2018.

[10] S. Gharghabi, S. Imani, A. Bagnall, A. Darvishzadeh, and E. Keogh, "Matrix profile XII: Mpdist: A novel time series distance measure to allow data mining in more challenging scenarios," in *2018 IEEE International Conference on Data Mining (ICDM)*, pp. 965–970, 2018.

[11] C.-C. M. Yeh, Y. Zhu, L. Ulanova, N. Begum, Y. Ding, H. A. Dau, D. F. Silva, A. Mueen, and E. Keogh, "Matrix profile I: All pairs similarity joins for time series: A unifying view that includes motifs, discords and shapelets," in *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pp. 1317–1322, 2016.

[12] Energy + Environmental Economics, "Inputs & Assumptions: 2019-2020 Integrated Resource Planning ," 2019.

[13] O. Anderson, N. Yu, and M. Bragin, "Optimize deep decarbonization pathways in california with power system planning using surrogate level-based lagrangian relaxation," *arXiv preprint arXiv:2309.07202*, 2023.