Assessing Space-Time Variability in Solar Irradiance Processes.

Nicolas Coloma

December 18, 2023

Abstract

We propose a method to address spatio-temporal processes that present phase variability, such as solar irradiance. Our approach consists in utilizing Dynamic Time Warping (DTW) and a 'Best-Shift' algorithm for curve alignment, with the ultimate goal of mitigating the impact of phase variation. We compare clustering results obtained without any alignment to those achieved after applying DTW or the Best-Shift algorithm, revealing enhanced clustering accuracy. Moreover, DTW proves versatile in uncovering spatial and temporal correlations, offering valuable insights into solar irradiance dynamics. These findings reflect the potential of our method as an exploratory data analysis tool, advancing the understanding and prediction of spatio-temporal processes with the presence of phase variation.

KEYWORDS: DYNAMIC TIME WARPING, SOLAR IRRADIANCE, CLUSTERING, SPACE-TIME PROCESSES, PHASE VARIABILITY.

1 Introduction

In the United states alone, the electric power generation from all non-renewable sources produced 1.65 billion metric tons of CO_2 emissions in 2022, which is about 0.86 pounds of CO2 emissions per kWh [1]. That is why there has been a need for aggresive actions to enhance the effectiveness of the power grid while minimizing its environmental harm.

When talking about the power grid, particularly with regard to renewable sources, significant efforts have been directed towards improving output reliability and addressing variability challenges. Among renewable sources, solar energy is widely utilized; however, its modeling remains challenging. This is due to the fact that, for example, it can become victim of external factors such as cloud coverage which introduces significant variability, even for closely located solar panels.

A key objective for scientists is gaining insight into the solar irradiance process and clustering has proven to be a valuable method for understanding and classifying data effectively, see for example [4, 8]. Ideally, geographically close sites should exhibit similar irradiance data. However, the variable nature of the data itself makes traditional clustering methods less accurate.

To mitigate this external variation and focus only on the intrinsic variations of the process, we propose a two-step method for assessing the similarity of time series. Firstly, we implement curve alignment, leveraging Dynamic Time Warping (DTW) [7, 12] and our 'Best-shift' algorithm, which will be explained in section 2. This will enable us to establish a cost measure between different sites, that takes into account phase variation, facilitating clustering to identify similar site groups. Furthermore, we aim to explore the temporal evolution of these relationships and clusters, leveraging data spanning multiple months, and validating the accuracy of our clusters using the Adjusted Rand Index (ARI).

Our project investigates two main aspects. First of all, we explore the efficiency of clustering techniques before and after applying curve alignment techniques. Additionally, we examine whether there is a spatial pattern in the data and whether this pattern varies during different periods of the day.

The paper is structured as follows: Section 1 serves as an introduction. In Section 2, we detail our dataset and include relevant mathematical background and methods. Section 3 presents the results, and Section 4 concludes with future directions.

2 Methods

2.1 Data

We'll be obtaining data of Global Horizontal Irradiance (GHI) from the National Renewable Energy Laboratory (NREL)[11], collected by a pyranometer at one-second intervals and organized by months. We will be focusing on a dataset of 17 solar panels located at the Inouye airport in Oahu, Hawaii, from June 1, 2010 till August 31, 2010. The sites are labeled DH1, DH2, . . ., DH11, AP1, AP3, AP4, . . ., AP7. Figure 1 shows the graphical location of the sites:

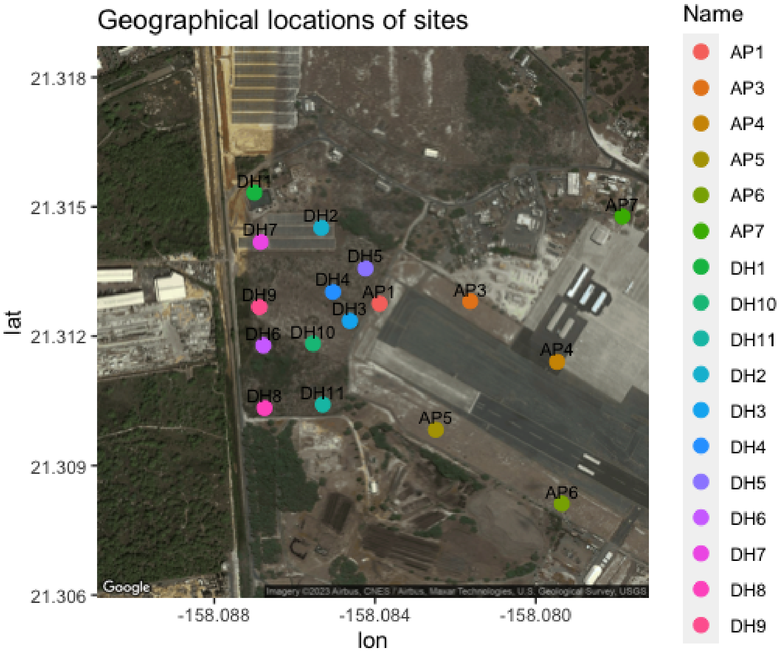


Figure 1: Satellite image of 17 solar panels in Oahu, Hawaii obtained using Google maps API.

We are interested in GHI since it measures direct sunlight as well as sunlight reflected from other surfaces. The interesting thing about the data is that since they provide measurements every second, the data captures the sudden changes in irradiance and is therefore more variable. As we are only interested in times at which there are significant readings of GHI, we will collect data between the times at which GHI $> 80[\frac{W}{m^2}]$ for the first and last time in the day, following [2]. Figure 2 shows GHI data for one day.

Now, denote by $Y(t)$ the GHI process at time $t \in \mathbb{R}$ and $X(t)$ be the corresponding amount of GHI that is expected under clear conditions, usually denoted as clear sky irradiance. When we model irradiance data, we are usually interested in the clear sky index variable (CSI) [14], which is the log-ratio between the measured GHI and the clear sky irradiance, i.e.,

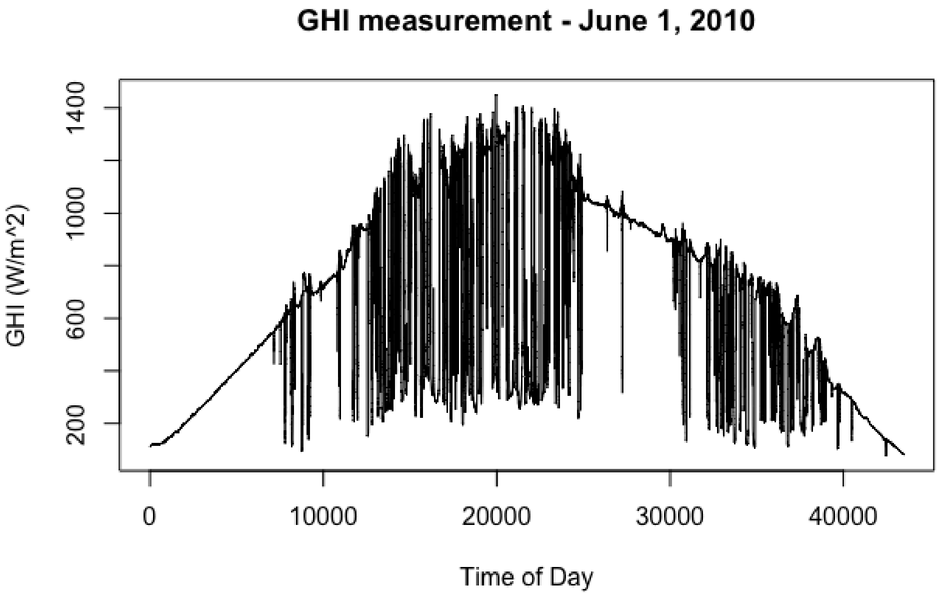
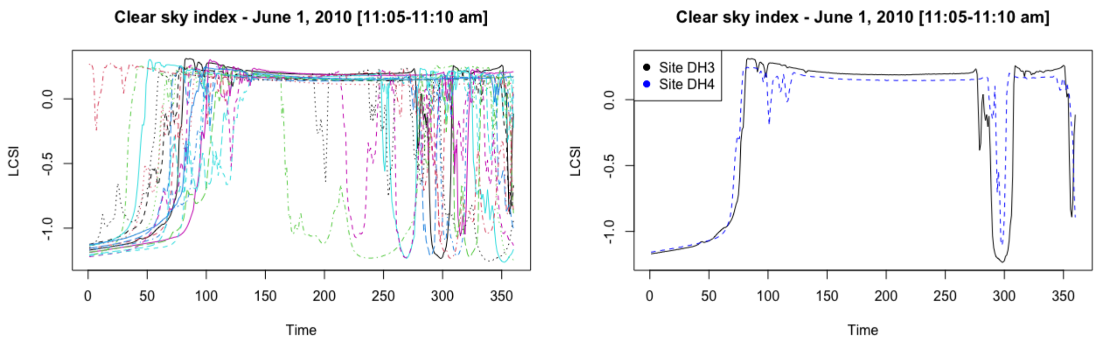


Figure 2: GHI measurements for June 1, 2010 at 1[s] scale for site DH3.

$$Z(t) = \log\left(\frac{Y(t)}{X(t)}\right)$$

where $Z(t)$ is the clear sky index at time t.

As previously discussed, working with this dataset presents a significant challenge due to the influence of external factors on solar panels, making it challenging to conduct an accurate analysis of each panel's energy reception. To illustrate, Figure 3a depicts the clear sky irradiance (LCSI) during a 5-minute interval on June 1, 2010, across all 17 sites. Additionally, Figure 3b highlights the difference on LCSI data between sites DH3 and DH4, which, as indicated in Figure 1, are relatively close geographically. The evident time lag observed in Figure 3b could be attributed to various factors, such as the passage of a cloud affecting one site before the other. We refer to this time lag as phase variation and this is the behaviour we would like to control by aligning curves before comparing them.



(a) LCSI measurements on a 5 minute interval for all 17 sites.

(b) LCSI measurements on a 5 minute interval for sites DH3 and DH4.

Figure 3: LCSI measurements on June 1, 2010 from 11:05 to 11:10 am.

2.2 Dynamic Time Warping

We aim to quantify the dissimilarity between our time series when there is a lag present. Based on some EDA, we saw that the data collected presented time deformations, that is why, we aim to align two time series before comparing them. This is where we use Dynamic Time Warping (DTW). Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, where $n \in \mathbb{N}$, and $\mathbf{y} = (y_1, y_2, \ldots, y_m)$, with $m \in \mathbb{N}$, represent two time series. In the scenario where $m = n$, implying equal lengths for both time series, the difference between the two can be computed using the $\mathbb{L}_2$ distance, defined as

$$||\mathbf{x} - \mathbf{y}|| = \left(\sum_{i=1}^{n} (x_i - y_i)^2 \right)^{\frac{1}{2}}.$$

However, using this distance for quantifying dissimilarity encounters a limitation due to the high variability observed in our data, particulary, the presence of time deformation, or phase variability. Dynamic Time Warping was first introduced by Sakoe and Chiba in 1978 [9] as a tool for speech pattern recognition. Its primary application lies in comparing time series that exhibit time deformations.

For instance, a process that exhibits this behaviour is speech, with individuals taking different durations to pronounce specific words. Similarly, when it comes to growth of an individual, each person follows a distinct growth profile. Overall, DTW is an algorithm designed to stretch, compress, and deform time to align series as closely as possible.

Formally, given two time series $\mathbf{x}$ and $\mathbf{y}$, DTW searchs for a realingment of the time series that minimizes a cost function between them [7]. It's worth noting that the cost function's choice depends on the application; in our case, we opt for the $\mathbb{L}_2$ distance.

That is, we are trying to minimize,

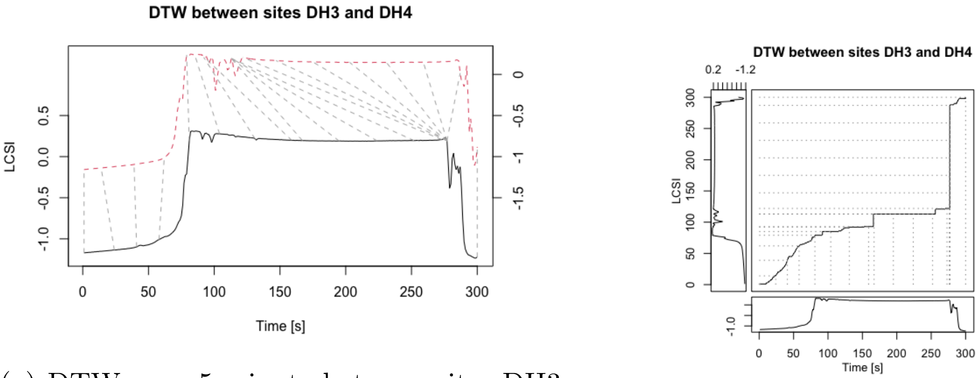
$$\mathrm{DTW}(x, y) = \min_{\pi \in \mathbf{A}} \left(\sum_{(i,j) \in \pi} |x_i - y_j|^2 \right)^{\frac{1}{2}} \tag{1}$$

where the minimization occurs across all admissible paths π, where $\mathbf{A}$ represents the set of all admissible paths. Three conditions must be satisfied for π to be considered admissible. Firstly, each index in a sequence must be matched with one or more indices from the other sequence. Secondly, the initial and final points of a path must coincide, meaning the first index of the sequences and the final elements must match. Lastly, the mapping of indices must be monotonically increasing to prevent 'going back in time'.

A valid path can be also be thought as a mapping between sequences $(1, \ldots, n)$ and $(1, \ldots, m)$, corresponding to a path in an $n \times m$ matrix from entry $(1, 1)$ to entry (n, m). When at entry (i, j), movement is restricted to entries $(i + 1, j)$, $(i, j + 1)$, or $(i + 1, j + 1)$ [12].

Figure 4a visually illustrates the effect of DTW on two 5-minute measurements from sites DH3 and DH4. Additionally, Figure 4b shows the best path obtained to align these series. Given the proximity of the sites and the visual similarity observed in the time series, we would expect a low DTW cost.

The main drawback of using DTW lies on how computationally expensive it is [9], specially when one day of data has lenght 39,600 (our time series range from 7am to 6pm). To mitigate this computational burden, we opt to partition our day into 5-minute intervals. Consequently, we will apply DTW to time series of shorter length, specifically 300, effectively reducing the overall running time of the algorithm.



(a) DTW on a 5 minute between sites DH3 and DH4.

(b) Matrix visualization of optimal path.

Figure 4: Dynamic time warping between sites DH3 and DH4 on June 1, 2010 from 11:05 to 11:10 am.

2.3 Best-shift Algorithm

The concept behind the implemented best-shift algorithm is similar with DTW. Our objective is to 'deform' time to achieve the optimal alignment of our curves, where optimal alignment minimizes the $\mathbb{L}_2$ distance between two time series. The main difference lies in our constraint on time deformations, limiting them to mere translations up to a certain constant τ. In other words, we seek a minimizer for the equation:

$$\min_{|k|\leq\tau} ||x(t) - y(t+k)|| = \min_{|k|\leq\tau} \left(\sum_{i=1}^{n} |x_i - y_{i+k}|^2 \right)^{\frac{1}{2}}.$$

In essence, we are exploring shifts of one of the time series, allowing at most τ seconds to the right or left until we achieve the minimal distance. While the concept is straightforward, the effectiveness of the best-shift algorithm lies in its ability to eliminate time shifts, at least locally, when our data is divided into 5-minute intervals. For this particular dataset, we are allowing the translations to be up to 20% of the lenght of the time series. Since we are working with 5-minute intervals, $\tau = 60[s]$.

2.4 Space-time variability

Exploratory data analysis (EDA) serves as a key stage in data mining, and visualizing the data or its inherent structure constitutes a significant step toward a more profound understanding of our data. In the domain of spatial statistics, a crucial initial step involves characterizing the joint spatio-temporal dependece structure, elucidating how the process correlates in space. This step is pivotal not only for visualizing data but also for optimal prediction of spatio-temporal processes, also known as kriging [3, 13].

We are interested in seeing how our process varies as a function of time and space. This can be achieved with the semivariogram, which, for a spatio-temporal process $Z(\mathbf{s};t)$, is defined for position s_i and s_k at times t_j and t_l as follows [13] :

$$\gamma(s_i, s_k; t_j, t_l) = \frac{1}{2}\mathrm{Var}(Z(s_i;t_j) - Z(s_k;t_l)) \tag{2}$$

if the spatial structure only dependes on the the differences between space and time, equation 2 simplifies to:

$$\gamma(\mathbf{h};\tau) = \frac{1}{2}\mathrm{Var}(Z(\mathbf{s}+\mathbf{h};t+\tau) - Z(\mathbf{s};t)).$$

With our particular data, in addition to it being variable due external factors, there is also large variance between sites. What we propose here is to use the $\mathbb{L}_2$ difference between two sites as an approximation for this quantity. The concept is to eliminate, or mitigate, the phase variation, allowing us to measure the difference between two sites as a function of their geographical distance. We will use this to reveal spatial structure within our data.

In more formal terms, let $\mathbf{x}$ and $\mathbf{y}$ again denote our two time series. Let $s_i = (\mathrm{lon}_i, \mathrm{lat}_i)$ be the longitude and latitude coordinates of the i-th site for time series $\mathbf{x}$. Then, $x(s_i)$ represents the log-clear sky index time series at location i. Lastly, we define a function Γ that captures this difference, that is,

$$\Gamma(s_i, s_j, ||\cdot||) = \Gamma(||x(s_i) - x(s_j)||)$$

where the metric can be chosen from the set $\{\mathbb{L}_2, \mathrm{DTW}, \mathrm{Shift}\}$. For instance, if we choose DTW, the function becomes $\Gamma(s_i, s_j, \mathrm{DTW}) = \mathrm{DTW}(x(s_i), x(s_j))$.

2.5 Method

Our project will be conducted as follows. Initially, we will generate a dissimilarity/cost matrix by computing the $\mathbb{L}_2$ difference between curves. Subsequently, we will repeat this process but after applying DTW and our 'best-shift' to account for the phase variation.

Following this, we will employ the dissimilarity matrix to perform hierarchical clustering on our 17 sites. Ideally, the outcomes should align with a hierarchical clustering approach that considers the geographical locations of our sites. To gauge the efficacy of our clustering technique, we will utilize the adjusted Rand index.

The Adjusted Rand Index is a score that helps us assess the similarity between two distinct clusters, particularly demonstrating robustness when comparing two different hierarchical clusterings [6]. The value ranges from 0 to 1, where a result of 0 indicates random assignments between the two clusters, and a result of 1 signifies identical cluster outcomes [10].

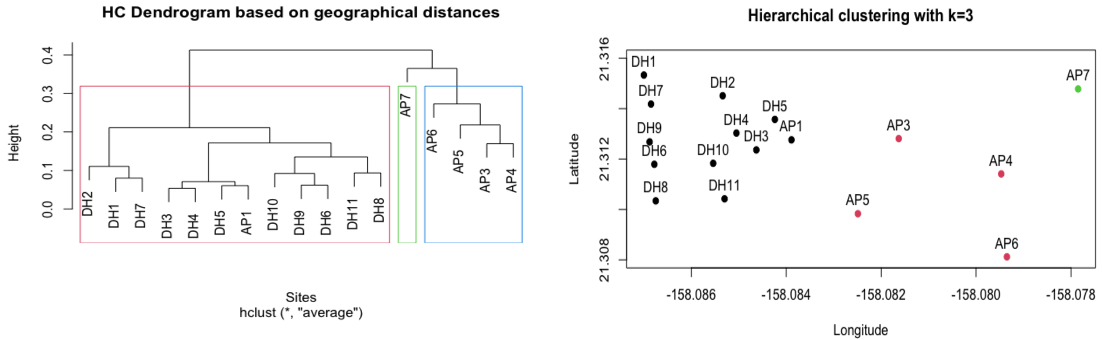
3 Results

3.1 Clustering results

We performed hierarchical clustering based on geographical distances between the sites. Figure 5 shows the obtained clustering of the sites. Hereafter, we will refer to this cluster as our **reference clustering**.

We can also get some insight from the clustering results. We see there is a clear distinction between the groups {DH1-DH11, AP1 } and { AP3-AP6 }. Notably, Site AP17 stands isolated from the rest, a observation evident in Figure 1.

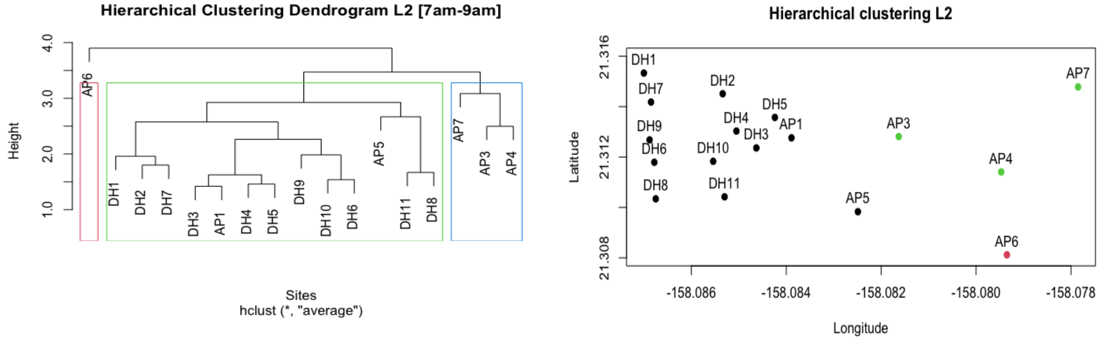
We then performed hierarchical clustering using each one of our 3 distances and for different time periods in the day. Figure 6 represents an instance of clustering obtained using the $\mathbb{L}_2$ difference between 7 am and 9 am without any time stretching.



(a) Hierarchical clustering based on geographical distances.

(b) Visualization of clustering when k=3.

Figure 5: Hierarchical clustering for the 17 sites based on geographical distances.



(a) Hierarchical clustering based on $\mathbb{L}_2$ distance.

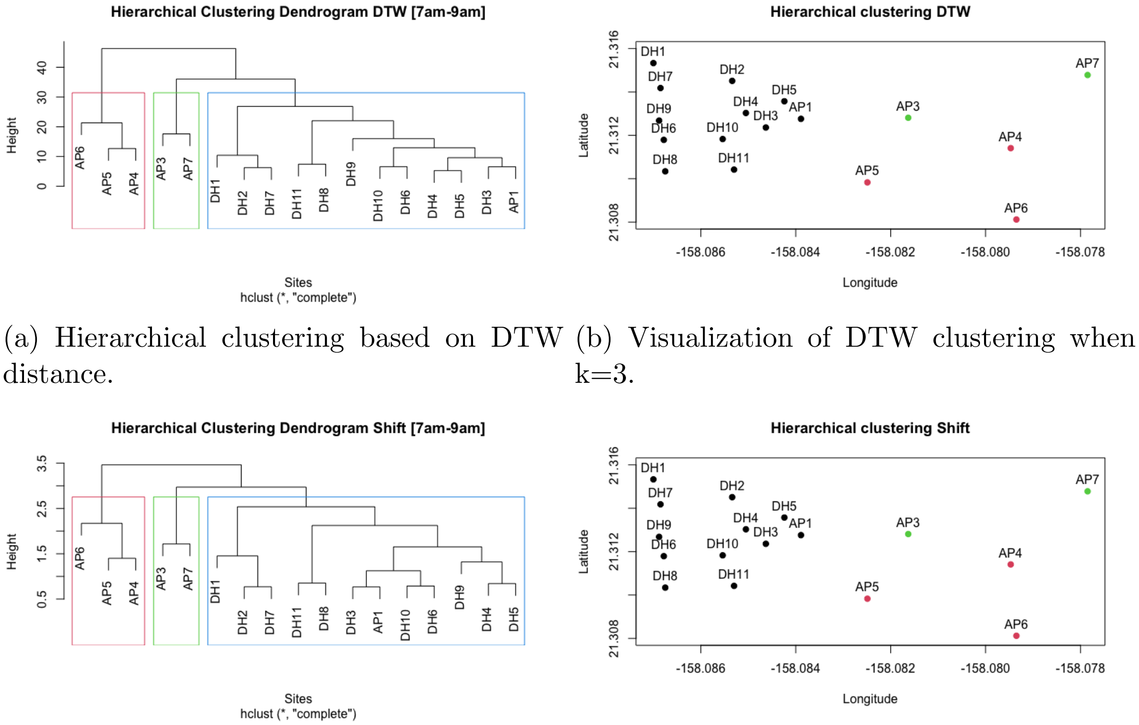
(b) Visualization of clustering when k=3.

Figure 6: Hierarchical clustering for the 17 sites based on $\mathbb{L}_2$ distance.

Notably, the grouping differs, with Site AP6 now isolated, and Site AP5 belonging to a different cluster among the most clear ones. This deviation from our reference clustering is reflected in a Rand Index score of 0.72. Subsequently, we applied DTW and our shift algorithm and Figure 7 shows the results for the time period 7am-9am.

It is interesing to see both methods effectively distinguish the group {DH1-DH11, AP1 }. And even though there is some variation within the clusters itself, they both yield a consistent Rand Index score of 0.94, an improvement from the clustering presented in Figure 6.

Our full clustering analysis considered all three 'metrics' across different time periods of the day. Table 1 presents the Adjusted Rand Index for each of these hierarchical clustering



(a) Hierarchical clustering based on DTW distance.

(b) Visualization of DTW clustering when k=3.

(c) Hierarchical clustering based on Shift distance.

(d) Visualization of Shift clustering when k=3.

Figure 7: Hierarchical clusterings for 7am-9am using DTW and Shift.

processes.

RandIndex	7am-9am	9am-12pm	12pm-3pm	3-pm-6pm
$\mathbb{L}_2$	0.72	0.78	0.78	0.78
DTW	**0.94**	**1.00**	**1.00**	**0.94**
SHIFT	**0.94**	0.94	0.94	**0.94**

Table 1: Adjusted Rand Index for hierarchical clustering, the bold entry shows which method shown more accuracy when compared with the reference clustering.

It is noteworthy that accounting for phase variation, either through DTW or shift, enhances the clustering classification for the sites with respect to the reference clustering.

3.2 Spatial Dependence

We calculated $\Gamma(s_i, s_j, ||\cdot||)$ for all three 'metrics' across different time periods of the day. We plot this costs against the geogrpahical distances between sites. The results are shown in Figures 8-11.

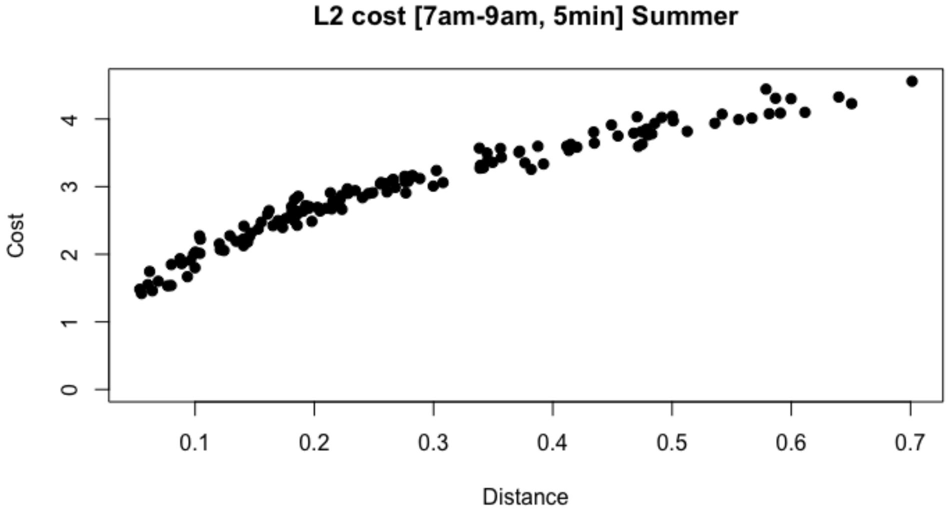


Figure 8: $\Gamma(s_i, s_j, \mathbb{L}_2)$ for 7am-9am.

As we can see, there is a spatial pattern hidden in the data and it agrees with out intuition. The further away two sites are, the more uncorrelated they become. We can also observe how this structure changes over time. A possible next step, but outside of the scope of this

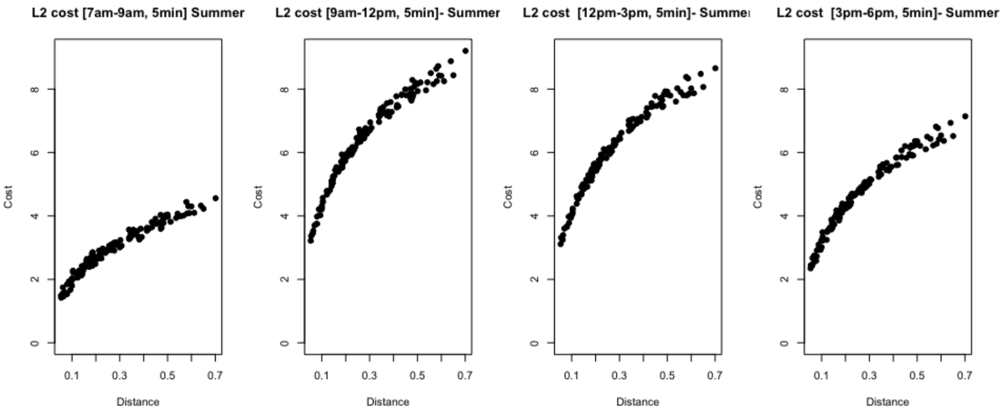


Figure 9: $\Gamma(s_i, s_j, \mathbb{L}_2)$.

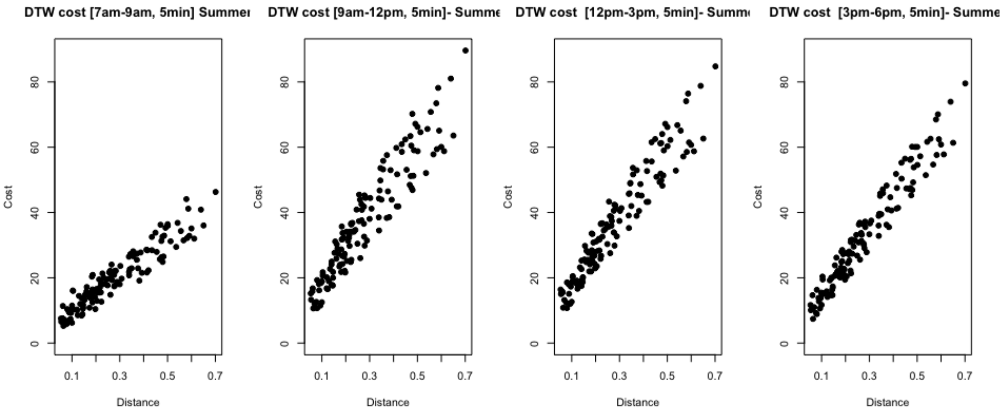


Figure 10: $\Gamma(s_i, s_j, \text{DTW})$.

project, is to fit a variogram model to this points cloud. Once we are able to fit a model we can start gaining information such as how variable the irradiance was in a particular day, or how far apart two solar panels have to be to have some sort of dependence.

4 Conclusions

Spatio-temporal processes often encounter challenges from external factors and phase variability. In this study, we demonstrated the effectiveness of time alignment techniques, specifically DTW and the Best-Shift algorithm, in improving clustering accuracy within our dataset.

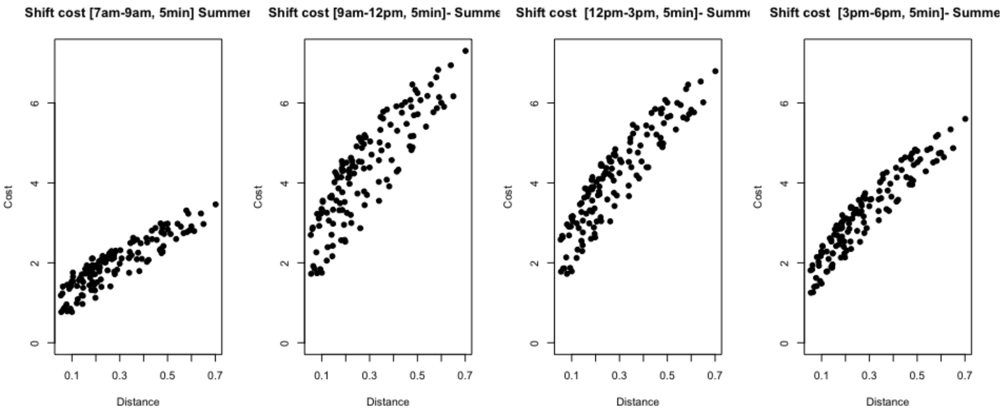


Figure 11: $\Gamma(s_i, s_j, \text{Shift})$.

We used DTW and our Best-Shift algorithm not only to enhance our clustering accuracy, but also to explore spatial structures within our dataset. The results, as showcased in Table 1, indicate a notable improvement in clusterings when accounting for phase variation.

Moreover, DTW emerged as a versatile tool, helping in the analysis of spatially and temporally correlated data. The technique allowed us to capture variability in solar irradiance over both space and time, providing valuable insights into the underlying processes. That is why we propose the use of DTW or Best-shift algoritmh as a exploratory data analysis technique (EDA), since it would be a step to clean our data and get rid of the phase variation before performing further analysis.

Contrary to expectations, our findings reveal higher similarity values in clustering during the middle of the day, possibly attributed to increased irradiance (GHI). This contradicts the intuition that clustering results are better under low variability, since our Γ function results showed that there is more variability in the middle of the day.

Despite the positive results obtained in this project, it is important to mention certain limitations. One key consideration is the choice of dividing each day into 5-minute intervals. While this temporal resolution allowed us to capture local differences between sites and allowed our algorithm to run faster, the decision to use this specific interval length was somewhat arbitrary. Exploring alternative interval durations could provide valuable insights into the sensitivity and robustness of our approach.

Another limitation is our reference clustering. The decision to compare our clustering results to a reference clustering based on geographical proximity aligns with the expectation that spatially close solar panels would exhibit similar behaviors. However, it is important to note that the choice of the reference clustering could affect the results.

It is also important to note that while these methods enhanced our results, there isn't a one-size-fits-all solution for capturing all structures in spatio-temporal data with phase variation [5].

As future directions go, we could fit a semivariogram model to the point cloud obtained from our Γ function, and obtain an estimation of the underlying variances of the process, which we can later do to perform kriging, i.e, prediction of the log-clear sky index on a site in which we don't have a solar panel.

The variogram also highlights variations between sites throughout different periods of the day. We can see that, in the middle of the day when GHI values are expected to be high, the cost is higher. Therefore, there seems to be some correlation between the variation of the process and the expected solar irradiance. We could use the insights from our Γ function and the clustering results to guide decisions in determining suitable locations for introducing new solar panels, strategically optimizing solar energy capture.

In summary, our method can be used as a exploratory data analysis tool for enhancing spatio-temporal clustering with processes prone to phase variation. Furthermore, it also unveils valuable insights into the dynamics of solar irradiance. And even though there are still many questions to answer, the potential applications for advancing both exploratory data analysis and predictive modeling in the realm of spatio-temporal processes are promising.

5 Video presentation

The url link of the video is:

`https://youtu.be/yN0jG32CI2g`

References

[1] U.S. Energy Information Administration. How much carbon dioxide is produced per kilowatthour of u.s. electricity generation?, 2023. Accessed: 10 December 2023.

[2] C. Berry, W. Kleiber, and B. Hodge. Subordinated gaussian processes for solar irradiance. *Environmetrics*, 34(6), 2023.

[3] N. Cressie and C. Wikle. *Statistics for Spatio-Temporal Data*. John Wiley & Sons, 2011.

[4] A. Di Piazza, M. Di Piazza, A. Ragusa, and G. Vitale. Environmental data processing by clustering methods for energy forecast and planning. *Renewable Energy*, 36(3):1063–1074, 2011.

[5] J. Marron, J. Ramsay, L. Sangalli, and A. Srivastava. Functional data analysis of amplitude and phase variation. *Statistical Science*, pages 468–484, 2015.

[6] G. Milligan and M. Cooper. A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate behavioral research*, 21(4):441–458, 1986.

[7] M. Müller. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84, 2007.

[8] C. Peruchena, J. García-Barberena, M. Guisado, and M. Gastón. A clustering approach for the analysis of solar energy yields: A case study for concentrating solar thermal power plants. In *AIP Conference Proceedings*, volume 1734. AIP Publishing, 2016.

[9] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49, 1978.

[10] J. Santos and M. Embrechts. On the use of the adjusted rand index as a metric for evaluating supervised classification. In *International conference on artificial neural networks*, pages 175–184. Springer, 2009.

[11] M. Sengupta and A. Afshin. Oahu solar measurement grid (1-year archive): 1-second solar irradiance; oahu, hawaii (data). Dataset, National Renewable Energy Laboratory (NREL), 2014.

[12] P. Senin. Dynamic time warping algorithm review. *Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA*, 855(1-23):40, 2008.

[13] C. Wikle, A. Zammit-Mangion, and N. Cressie. *Spatio-temporal Statistics with R*. Chapman & Hall, USA, 2019.

[14] W. Zhang, W. Kleiber, A. Florita, B. Hodge, and B. Mather. Modeling and simulation of high-frequency solar irradiance. *IEEE Journal of Photovoltaics*, 9(1):124–131, 2018.