Short Research Article

Persistent Homology in Solar Production

ABSTRACT:

Persistent homology, an algebraic topology-based mathematical framework, presents an innovative method for capturing and characterizing the inherent topological features present in time series datasets. The research aims to evaluate the efficacy of features derived from persistent homology in enhancing the accuracy and interpretability of classification models. This investigation contributes to the expanding convergence of topology and time series analysis, providing valuable insights into the potential of persistent homology for extracting information from temporal data. The study specifically focuses on the analysis of region-wise solar production data obtained from India for the year 2022. The examination of this data is conducted using R-Software, and the resulting topological properties are represented through persistent diagrams.

Keywords: Topological data analysis, Persistent diagram, p_value

INTRODUCTION:

Recently, there has been a significant expansion of Topological Data Analysis (TDA). This topic is ripe to become a mainstay of data analytics, with applications in 3D modeling[10], hospitals[8], and aircraft[9]. Persistent homology is one of the instruments at the topological data analyst's disposal. Using a dataset of all sizes at once, this program records the emergence (birth) and disappearance (death) of topological features inside a series of topological spaces. It is possible to compile and store these birth-death periods within a multi set called a persistence diagram, however deep learning and machine learning algorithms of today cannot work with these diagrams. Several initiatives have been started as a result of this disparity to summarize persistence diagrams in a fashion that is consistent with deep learning and machine learning while preserving topological information. Persistent entropy[1], persistent images, and persistent landscapes[2] are illustrations for summaries. In that paper, they demonstrated how this approach could provide not only some popular summaries, including persistent homolgy and persistence barcode. Some work has been done recently on classifying time series using TDA methods[4].

The escalating demand arising from renewable energy sources has heightened the significance of effective monitoring and methodologies in the realm of solar energy production. Time series data, representing the temporal evolution of solar energy output, poses unique challenges in its analysis and interpretation due to its intricate patterns and dynamic nature. Traditional time series analysis methods often struggle to capture the nuanced topological structures inherent in these datasets.

This research delves into the application of algebraic topology, specifically persistent homology, as an innovative and promising approach to unraveling the intricate dynamics of solar production time series. Algebraic topology provides a mathematical framework for understanding data shape and structure, and persistent homology, as a tool derived from this field, proves particularly adept at capturing and quantifying persistent features such as loops, voids, and connected components.

The primary objective of this study is to explore the potential of persistent homology in extracting meaningful topological signatures from solar production time series data. By representing the temporal data as point clouds, we aim to leverage the ability of persistent homology to reveal underlying structures that traditional methods might overlook. The research is motivated by the need for advanced analytical tools capable of enhancing our understanding of the complex dynamics inherent in solar energy production.

In addition to exploring the theoretical foundations of persistent homology, this study undertakes a practical investigation by applying persistent homology to the classification of solar production states. By evaluating the efficacy of persistent homology-based features, we aim to demonstrate its potential in improving the accuracy and interpretability of classification models.

By this work, we hope to further the rapidly developing field of time series analysis and topology by providing a fresh viewpoint on the comprehension and categorization of time-varying events. The implications of this work extend beyond the domain of solar energy, showcasing the broader applicability of topology in unraveling complex temporal patterns and enhancing decision-making processes in various fields.

2. PRELIMINARIES

2.1. Persistent homology

An invariant topological space under continuous deformations is the subject of the classic mathematical topic of homology. It can be inferred that a space's homology provides valuable insights into its topological configuration. The k-th Betti numbers, or the number of k-dimensional holes in X, are commonly counted using the k-th homology group $H_k(X)$ of a space. $H_0(X)$, for instance, counts the connected components; $H_1(X)$ counts loops; $H_2(X)$ counts voids or "air pockets," and so on.

First described in [6], persistent homology is a potent technique from the topological data analysis discipline that monitors homology changes throughout a filtration.

2.2. Filtration

A filtration of a space X is an increasing sequence of spaces $\emptyset = X_0 \subset X_1 \subset \ldots \subset X_n = X$. One may obtain homology groups for each X_i . Because of the subset relations, we can track the changes of homology groups. This process is what we call persistent homology. For a time series with corresponding continuous piecewise linear function f_s, we can obtain a filtration by considering sublevel sets $f_s^{-1}((-\infty, x]) = \{y \in R) \mid f_s(y) \in (-\infty, x]\}$. The inclusion $X_t \subset X_s$ induces a map $g_{t,s}^k$: $H_k(X_t) \rightarrow H_k(X_s)$ between the homology groups. We note that homology in this sequence only changes at critical values of f [7].

We say a homology class α is born at b if we have $\alpha \in H_k\chi_b$ and $\alpha \notin Im\, g_{b-1,b}^k$. We say that α born at b dies at d, d ≥ b if $g_{b,d-1}^k(\alpha) \notin Img_{b-1,d-1}^k$, but $g_{b,d}^k(\alpha) \in Img_{b-1,d}^k$, i.e. if it merges with a previous class. The rank $\beta_{b,d}^k = rank\, Img_{b,d}^k$ for d ≥ b form the persistent Betti numbers of the filtration. These persistent Betti numbers count the number of classes that were born at or before b and are still alive at d. Inclusion-exclusion allows us to count exactly the number $\mu_{b,d}^k$ of classes born at b and die at d by $\mu_{b,d}^k = \beta_{b,d-1}^k - \beta_{b-1,d-1}^k + \beta_{b-1,d}^k - \beta_{b,d}^k$. The k-th persistence diagram, or just diagram, $P_k(f)$ associated to the filtering function f of a space X is a multi-set, that is a set of points with multiplicity, of birth-death pairs (b, d) with multiplicity $\mu_{b,d}^k$ along with the diagonal points (b, b) each with infinite multiplicity. To shorten notation, we will often represent persistence diagrams with the letter D.

2.3. Wasserstein distance

The stability theorem [5] for persistence diagrams, which states that if f_1, f_2 are functions that have finitely many critical values then $W_\infty(P_K(f_1), (P_K(f_2)) \leq || f_1 - f_2 ||_\infty$, where $W_\infty(P_K(f_1), (P_K(f_2))$ is known as the Wasserstein ∞-metric. According to this stability theorem, a minor change in the original space will also result in a small change in the diagrams. Diagrams need to be summarized in some way because machine learning cannot easily work with them.

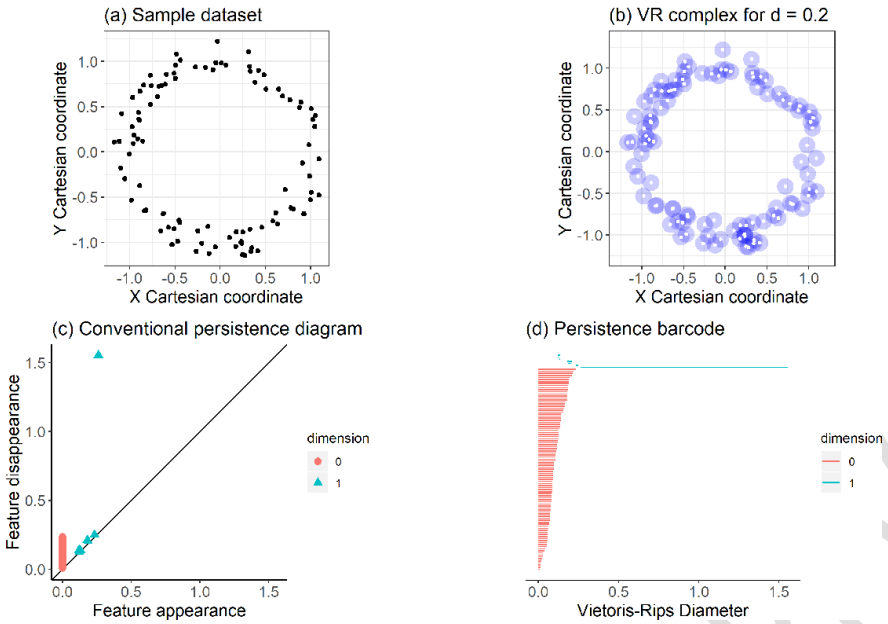

Figure 1: Persistent Diagram, persistent Barcode to visualize the persistent homology

2.4. Time series data

A time series is a collection of measurements or observations that are gathered and kept track of over an extended period of time. The sequence in which the observations are made matters since each one in the series relates to a particular moment in time. Many disciplines, including finance, economics, meteorology, and engineering, frequently deal with time series data.

A few essential ideas concerning time series data

2.4.1.Time Points: There are regular and irregular intervals at which time series data are recorded. The recorded moments for the observations are represented by the time points.

The Time Series Element

- Trend: An enduring pattern or inclination in the information.
- Recurring patterns or cycles with a set duration that are known as seasonality are frequently associated with certain calendar-based events or seasons.
- Designs that repeat, but with different cycle lengths are called cyclic designs.
- Random (Residual): Variations that are unpredictable or random in relation to seasonality and the trend.

Examples, stock prices, Temperature Data**,** Economic Indicators, Web Traffic, Temperature data, Power consumption, Health care, weather data, sales data..etc

Comparing Time series with Persistent homology

In order to compare time series data using persistent homology, topological features must first be extracted from the time series, and the persistence diagrams corresponding to those features must next be examined. An algebraic topology technique called persistent homology is used to find and measure structures or patterns in data that hold true at various scales.

A broad outline of how persistent homology might be used to compare time series data:

1. Set up your time series data by performing data preprocessing. Make that the format is correct and that it accurately depicts how a variable changes over time.

2. Topological Feature Extraction: First, create a simplicial complex by filtering the embedded time series. Alpha, Vietoris-Rips, and Čech complexes are examples of frequently used filtrations.

3. Determine Persistent Homology: The homological characteristics of the simplicial complex can be determined by applying persistent homology techniques. Find any topological properties that remain consistent at multiple scales, such as loops, voids, or connected components.

4. Persistence Diagrams: Use persistence diagrams to illustrate the persistent homology findings.

Every point in the diagram denotes a topological feature, and its coordinates show the feature's birth and death scales.

6. Compare Persistent Diagrams: Examine persistence diagrams between various time series quantitatively. Persistence diagram dissimilarity can be measured using metrics like bottleneck distance or Wasserstein distance.

7. Statistical Testing: Determine the statistical significance of the observed variations in persistence diagrams by conducting statistical tests. For this, permutation tests and bootstrap techniques are frequently employed.

8. Visualization: To comprehend the topological characteristics that vary throughout time series, visualize persistence diagrams.

3. APPLICATION: SOLAR ENERGY PRODUCTION DATA

Recent advancements in computation have allowed us to compute topological invariants of a Point Cloud Data. For this purpose, we use a programming language known as R-language. The topological features of solar energy production is studied using R. The data is taken from https://posoco.in [12] which is a authentic source and is being maintained by Grid controller of India limited, formely known as power system operation corporation limited. The data consists of solar energy production in India regiowise between the 01 Jan 2022 and 31 Dec 2022(datewise). Only the North, South, West, East region data is taken for this study.

3.1.Data preprocessing:

Data has 365 daily data points for solar energy, and there are no missing values. For better understanding of the data, the data for each date is scaled. Next, the Euclidean distance matrix which serves as the data's default distance (point cloud) is found. The data is then divided into four groups namely North, South, West and East.

3.2. Persistent homology:

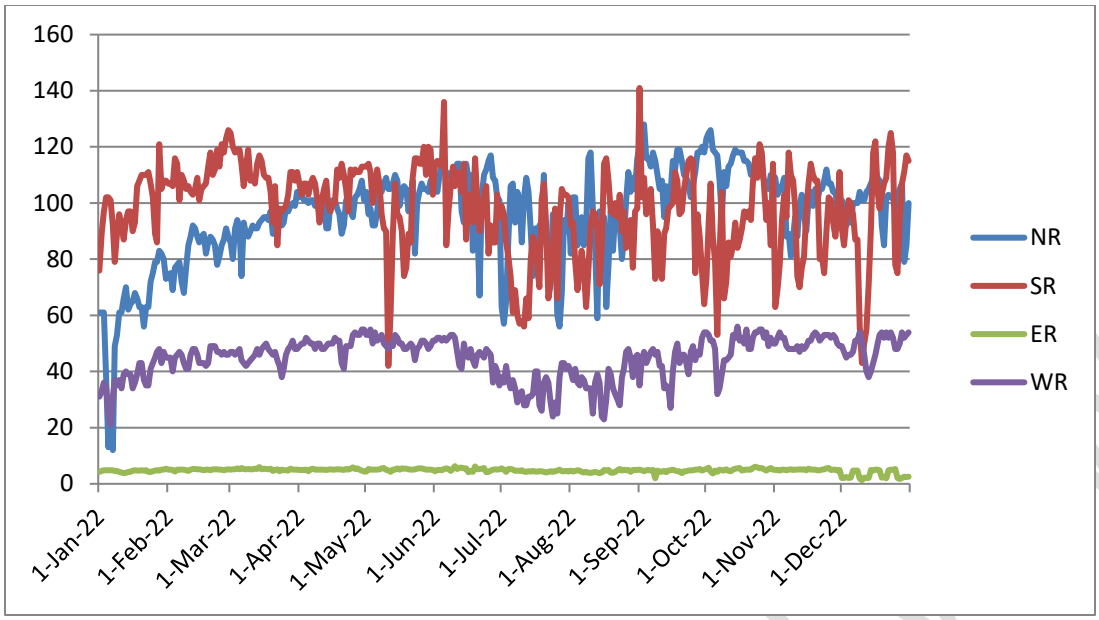

Figure 2: distribution of the region wise data

The Homology groups that arise and die at progressively greater distances are created. At every distance closed balls with half radius is used to calculate b from the distance. In the event that an older topology and one that is younger merge at a certain distance, the younger would be born and the elder would die. The persistent diagram, indicated by D, represents the outcome of the set of points from the point cloud that represent the birth and death of homology groups. Here, the process referred to as connected components exhibits only the 0-th homology.

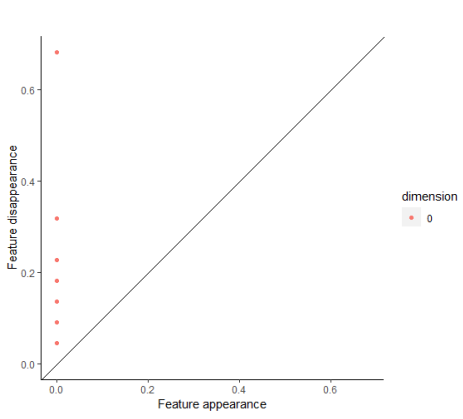

Figure3: North Region- Persistent Diagram

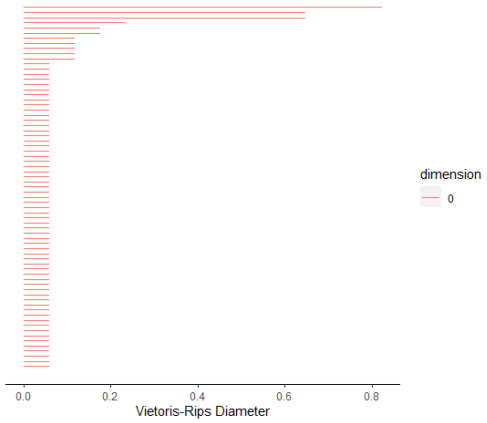

Figure4: North Region-Bar code

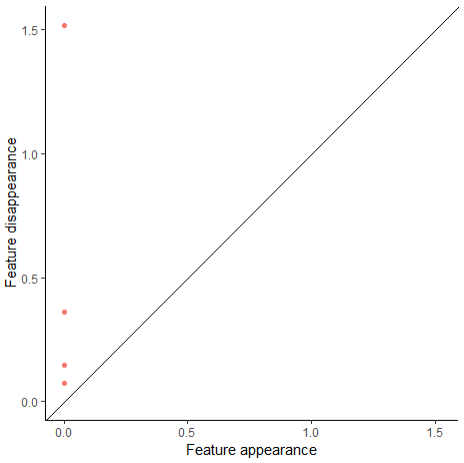

Figure5: South Region- persistent Diagram

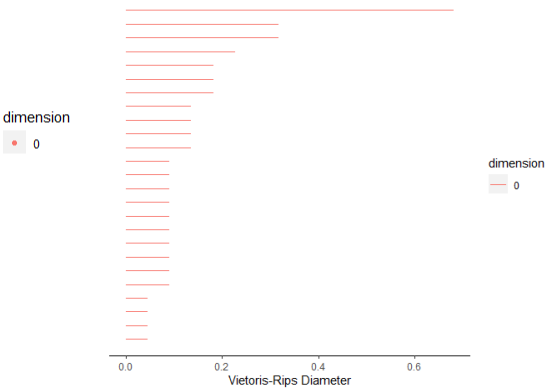

Figure6: South Region – Bar code

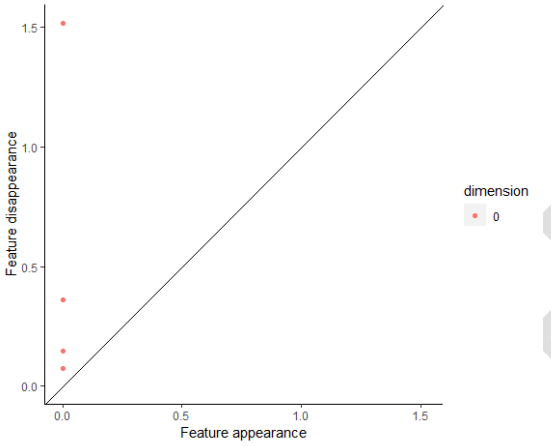

Figure7: East Region- Persistent diagram

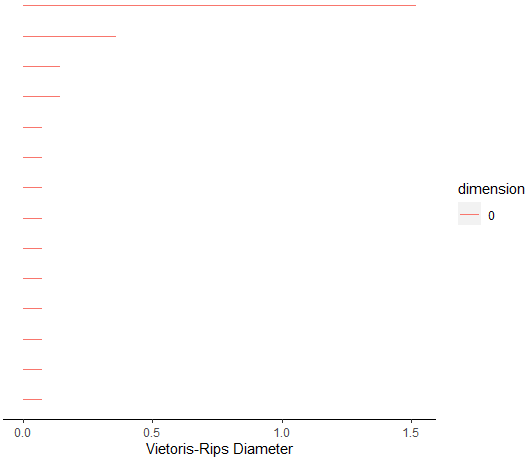

Figure 8: East region- Barcode

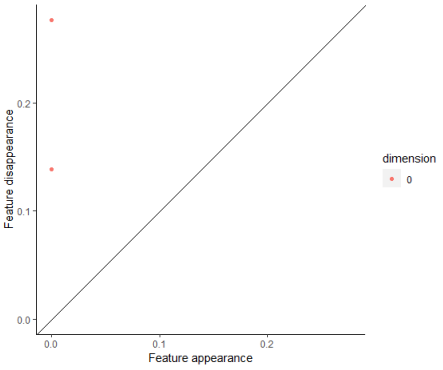

Figure9: West region- Persistent Diagram

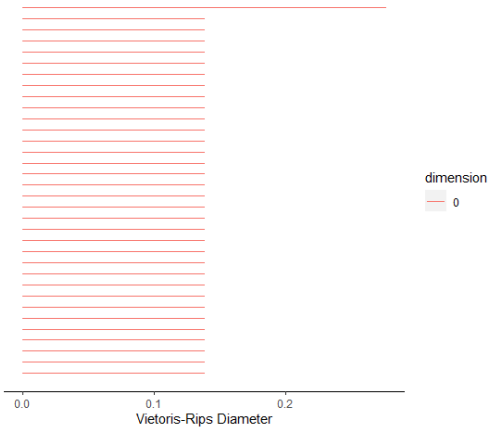

Figure10: West region- Barcode

The persistent diagram is additionally displayed as a barcode diagram. Here, a Euclidean distance is employed during the filtering process. The dispersed nature of the points, which is also depicted as red lines within the barcode diagram, is shown by the persistence of the red dots in Figure 3-10, which indicate the birth and death of the 0-th homology groups.

3.3. Statistical Approach:

Table 1 : Compute the distance between the Persistent homology features ,which is given in below:

Region	Distance
North- South	4.04148
North-East	5.407368
North- West	5.848798
South-East	2.595279
South- West	3.082397
East- West	4.904337

Lower distances typically correspond to higher similarity between characteristics that exhibit persistent homology. It explains these findings as follows:

- Of the pairs, SR and ER have the lowest distance (about 2.595279), indicating that their persistent homology traits are more similar than those of the other pairs.
- The distance between the two pairs, SR and WR, is roughly 3.082397, suggesting a moderate resemblance.
- There is around 4.04148 separating the pair NR, SR.
- At about 4.904337, the pair of ER, WR possesses the greatest distance.
- These distances offer a numerical representation of how different or similar two sets of persistent homology traits are.

3.4. Permutation Test:

The more comparable persistent homology features with the South-East regions in the data sets are examined. A permutation test, sometimes referred to as a randomized or re-randomization test, is a non-parametric technique used in statistical hypothesis testing to determine the statistically significant nature of an observed effect. Analyzing if the observed data differs noticeably compared to what would be predicted by chance alone is the fundamental goal of a permutation test. [10]

Formulate the hypothesis:
• Null hypothesis (H_0): There exists no discernible variation in solar production between the East and South regions.
• Alternative hypothesis (H_1): There exists a substantial discrepancy in solar production between the East and South regions.

A permutation test for nonparametric statistical inference of persistent homology in topological data analysis using the R function (permutation_test) is performed. To sum up, 1.083004 is the observed test statistic. The p-value for the permuted test statistics distribution is 0.48. A p-value over the standard significance level of 0.05 may indicate inadequate data to rule out the null hypothesis.

CONCLUSION

An investigation of the persistent diagrams shows that the data have distinct number of topological features. This is represented by presence of significant Connected components (0-th homology) in persistent diagram of data. This is also proved using p-value, a measure that is used to quantify a significant difference of solar production between South and East Region.

REFERENCES

1. *Atienza, N., Gonzalez-Diaz, R., and Soriano-Trigueros, M, A new entropy based summary function for topological data analysis,* Electronic Notes in Discrete Mathematics*, 68(2018) , 113-118.*
2. *Bubenik, P, Statistical topological data analysis using persistence landscapes,* The Journal of Machine Learning Research*, 16(1)(2015),77-102 .*
3. *Chung, Y.-M., and Lawson, A, Persistence curves: A canonical framework for summarizing persistence diagrams,(2019).*
4. *Chung, Y.-M., and Lawson, A, A Persistent Homology Approach to Time Series Classification,* arXiv:2003.06462 [stat.ME]*, (2020).*
5. *Cohen-Steiner, D., Edelsbrunner, H., and Harer, J, Stability of persistence diagrams,* Discrete & Computational Geometry*, 37(1) (2007), 103-120.*
6. *Edelsbrunner, H., Letscher, D., and Zomorodian, A, Topological persistence and simplification, In* Proceedings 41st Annual Symposium on Foundations of Computer Science*, (2000), 454-463.*
7. *Edelsbrunner, H., and Harer, J,* Computational Topology: An Introduction*. Miscellaneous Books. American Mathematical Society, (2010).*
8. *Li, L., Cheng, W.-Y., Glicksberg, B. S., Gottesman, O., Tamler, R., Chen, R. and Dudley, J. T, Identification of type 2 diabetes subgroups through topological analysis of patient similarity,* Science translational medicine*, 7(311) (2015), 311ra1740-311ra174.*
9. *Li, M. Z., Ryerson, M. S., and Balakrishnan, H. Topological data analysis for aviation applications,* Transportation Research Part E: Logistics and Transportation Review*, 128, (2019), 149-174.*
10. *Robinson, A., and Turner, K, Hypothesis testing for topological data analysis,* Journal of Applied and Computational Topology**,** *1(2) (2017), 241-261.*
11. *Turner, K., Mukherjee, S., and Boyer, D. M, Persistent homology transform for modeling shapes and surfaces,* Information and Inference: A Journal of the IMA*, 3(4) (2014), 310-344.*
12. *https://posoco.in – solar energy data.*