A Method for Measuring the Non-Smoothness of Time Series Data: Dirichlet Mean Energy Function

Lianchao Wang  
China University of Mining and Technology (Beijing)

Yijin Chen  (✉ Y.J.CHEN@263.NET )  
China University of Mining and Technology (Beijing)

Wenhui Song  
China University of Mining and Technology (Beijing)

Hanghang Xu  
China University of Mining and Technology (Beijing)

---

Article

Keywords:

Posted Date: April 10th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2769394/v1

License: ☎️ This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

This paper proposes an effective method for measuring the non-smoothness of time series data: Dirichlet mean energy function. The method expresses the time series data as an n-dimensional vector based on its own properties, and then abstracts the time series model as a chain graph model based on directed graph theory. The incidence matrix of the time series data is established based on the constructed chain graph model, and the Dirichlet mean energy function is defined in the form of matrix function. The contribution of this paper is to propose an effective mathematical tool for measuring the non-smoothness of time series data based on graph theory and matrix theory. In future work, we will further validate the validity of this tool in more application areas and extend this method to high-dimensional time series data.

Introduction

In recent years, with the rise of mobile devices and the Internet of Things, the use of sensors has surged, leading to an exponential growth in the generation of time series data. Along with this trend, there is an increasing demand for analyzing and processing time series data. Time series data analysis is an important branch in the field of data analysis, which mainly focuses on the analysis, modeling, prediction, and other aspects of time series data. Time series data can be applied in many fields such as finance, economics, weather forecasting, signal processing, and so on. However, time series data often exhibit non-smoothness, which can affect the accuracy and reliability of data analysis and modeling. Therefore, measuring the non-smoothness of time series data has become an important issue in time series analysis.

The non-smoothness of time series data is typically manifested as the instability of statistical characteristics of the data over time. For example, the mean, variance, auto correlation, and other statistical characteristics of the data may vary greatly at different moments. In time series data analysis, we typically use some common indicators and methods to measure the non-smoothness of time series data, to better understand the properties and characteristics of time series data. One of the most commonly used indicators is the Sample Autocovariance Function (SACF), which can be used to measure the covariance of time series data at different moments, thus helping us to assess the smoothness of time series data. If the time series data is smooth, then its SACF at different moments should be very similar, otherwise non-smoothness exists. In addition, Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) are also used to measure the non-smoothness of time series data. ACF and PACF can help us determine the correlation of time series data at different moments, further indicating the smoothness of time series data. If the values of ACF and PACF fluctuate greatly over time, then non-smoothness may exist. In addition to statistical indicators, we can also use some graphical methods to measure the non-smoothness of time series data, such as ACF Plot, PACF Plot, Stationarity Plot, etc. These graphical methods can help us more intuitively judge the non-smoothness of
time series data. In general, the non-smoothness of time series data is an important issue in time series data analysis. Measuring the non-smoothness of time series data can help us better understand the properties and characteristics of time series data and choose appropriate models and analysis methods. In practical applications, we can choose suitable indicators and methods based on specific analysis purposes and data characteristics to ensure the accuracy and reliability of analysis and modeling.

Based on graph theory and matrix theory, this paper proposes a Dirichlet mean energy function method for measuring the non-smoothness of time series data. This method measures the non-smoothness of time series data with different lengths and sampling interval by giving a standardized metric. The typhoon wind speed data collected by the sensor is used to validate the validity of the proposed method.

The structure of the remaining parts of this paper is as follows: Section 2 provides a detailed description of the graph model for time series data and presents the Dirichlet mean energy function as a mathematical tool for measuring non-smoothness of time series data. In Section 3, we analyze and validate the validity of the proposed Dirichlet mean energy function in measuring this non-smoothness using time series data collected by sensors. Section 4 presents the conclusion of this paper and provides prospects for future research work.

Methodology

Definition of time series data

Time series data refers to a sequence of data collected or recorded in chronological order, usually at fixed time intervals (such as 1 second, 10 minutes, 24 hours, 7 days, 1 year). Therefore, time series data can be analyzed and processed as discrete time data. The main characteristic of time series data is its continuity related to time, with data at each moment being correlated with data at the moment before and after it, presenting two significant features: non-stationarity and fluctuating amplitude with time.

Therefore, when dealing with time series data, the unique characteristics of the time dimension, such as periodicity and trend, must be considered. In addition, time series data usually contain factors such as noise, missing values, and outliers, which can affect the quality and accuracy of the data and require effective data preprocessing and cleaning. In summary, time series data is a very important type of data that has important implications for research and applications.

Vectorized Representation

Time series data refers to the measurement of an observed object over a period. Assuming the observed object is \( S \) and the measurement at a certain moment is \( X_i \), then the time series data for a certain period \( T \) can be expressed as \( \{ X_i, i = 1, 2, \ldots, n \} \). It is evident that the generation of such time series data has a strict time order, meaning that the sequence of time series measurements for a certain period \( T \) starts with a time-ordered measurement \( X_i \) and ends with a time-ordered measurement \( X_{i+1} \), which is determined by the dynamic attributes of the underlying system's temporal changes. According to the definition of vector, a vector is an ordered list of finite length numbers, usually written as a vertical array
enclosed in square brackets or parentheses. Therefore, it is natural to vectorize time series data according to its inherent properties, that is, to use vectors to express time series data. Eq. (1) is the vectorized representation of time series data, which is expressed as follows.

\[
X = \begin{bmatrix}
X_1 \\
X_2 \\
\vdots \\
X_n
\end{bmatrix}
\]

1

Construction of graph model

Directed graph

A directed graph is a network consisting of a set of nodes connected by directed edges, which can be used to express relationships between objects. The directed graph can be expressed as a set of nodes numbered 1, 2, \ldots, \(n\) and directed edges numbered 1, 2, \ldots, \(m\). Each edge connects one node to another node, and these two nodes are said to be connected or adjacent. Directed graphs are usually drawn using circles or dots to represent nodes and arrows to represent directed edges. A directed graph can be represented using an incidence matrix \(n \times m\), which is defined as:

\[
M_{ij} = \begin{cases}
1 & \text{edge points to node} \\
-1 & \text{edge points from node} \\
0 & \text{otherwise}
\end{cases}
\]

The j-th column of the incidence matrix \(M\) is associated with the j-th edge in the digraph, where the index of two non-zero elements provide information about the connected nodes. The i-th row of the incidence matrix \(M\) corresponds to the i-th node in directed graph and its non-zero elements indicates which edges are connected to the node, that is, whether these edges point to or point from the node.

Chain graph model

In directed graph, time series data \(\{X_i, i = 1, 2, \ldots, n\}\) generated within a certain period \(T\) can be represented using \(n\) nodes, where adjacent nodes are connected by directed edges. The direction of the edge between nodes \(X_i\) and \(X_{i+1}\) strictly follows a certain order, that is, the edge is directed from node \(X_i\) to node \(X_{i+1}\), but not vice versa. Thus, a chain graph model can be naturally constructed with \(n\) nodes and \(n - 1\) directed edges to represent the time series data \(\{X_i, i = 1, 2, \ldots, n\}\) within the period \(T\); the chain graph model is shown in Fig. 1.
Where $n$ nodes represent the time series data values in the data set $\{X_i, i = 1, 2, \ldots, n\}$, and $n - 1$ directed edges connect these nodes in strict order. Accordingly, the raw time series data is abstractly modeled and represented by constructing the chain graph model.

### Vector difference

### Incidence matrix

If the incidence matrix of time series data $\{X_i, i = 1, 2, \ldots, n\}$ is denoted as matrix $M$, based on the definition of directed graphs and the chain graph model constructed, we can easily derive the explicit expression of matrix $M$ which is given as follows:

$$
M = \begin{bmatrix}
-1 & 0 & 0 & 0 \\
1 & -1 & 0 & 0 \\
0 & 1 & \ddots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ddots & -1 & 0 \\
0 & 0 & 1 & -1 \\
0 & 0 & 0 & 1
\end{bmatrix}
$$

Where the $i$-th row of the incidence matrix $M$ represents the $i$-th node in the chain graph model constructed, and similarly, the $j$-th column represents the $j$-th directed edge in the same chain graph model. The value of each element $M_{ij}$ in the incidence matrix is consistent with the definition given in Eq. (2). In this way, we obtain the incidence matrix of time series data $\{X_i, i = 1, 2, \ldots, n\}$.

### Difference matrix

From the correlation matrix $M$, we can derive a $(n - 1) \times n$ difference matrix $D$. The explicit expression for the difference matrix $D$ is defined as follows:

$$
D = \begin{bmatrix}
-1 & 1 & 0 & \cdots & 0 & 0 & 0 \\
0 & -1 & 1 & \cdots & 0 & 0 & 0 \\
\ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\
0 & 0 & 0 & \cdots & -1 & 1 & 0 \\
0 & 0 & 0 & \cdots & 0 & -1 & 1
\end{bmatrix}
$$
Obviously, the difference matrix $D$ defined in this way is the transpose matrix of the incidence matrix $M$, and it is also a sparse matrix. The diagonal elements of the difference matrix $D$ are either 1 or -1, and the elements outside the diagonal are not shown and are all 0.

The product of the time series data vector $X$ by the difference matrix $D$ can yield an $(n-1)$-dimensional vector representing the difference between two consecutive components of the time series measurement $X_i$. Eq. (5) presents the explicit expression for the product of the difference matrix $D$ and the time series data vector $X$.

$$DX = \begin{bmatrix} X_2 - X_1 \\ X_3 - X_2 \\ \vdots \\ X_n - X_{n-1} \end{bmatrix}$$

In the next section, we will present the explicit expression of the Dirichlet mean energy function used to measure the non-smoothness of time series data.

**Dirichlet mean energy function**

In many applications, the graph can be used to represent networks where some physical quantities flow on the network, such as water, electricity, heat, or traffic flow. A graph is very useful when the focus is on some physical quantity that takes values at each vertex or node of the graph. Let $v$ be an $n$-dimensional vector, which is usually interpreted as potentials, with $v_i$ being the potential of vertex $i$ in the graph. The product of the transpose matrix of the incidence matrix (i.e., difference matrix) and the vector $v$, denoted as $u = M^T v$, can be given a simple interpretation about the potential differences across the edges. Specifically, the $m$-dimensional vector $u$ gives the potential differences across the edges: $u_j = v_l - v_k$, where $u_j$ is the potential difference from vertex $v_k$ to vertex $v_l$. Here, we abstract vertex $v_i$ as our time series measurement at a certain moment, and $u$ is the observed difference between two time series measurements. When the $m$-dimensional vector $u$ is small, it implies that the differences between the time series measurements of vertices are small, and its quantified metric is a function of $u$. Then, dividing the resulting function by the total number of elements in vector $v$ yields the Dirichlet mean energy function, which is expressed as:

$$D(v) = \frac{1}{n} ||M^T v||^2$$

In the Eq. (6), $D(v)$ represents the Dirichlet mean energy to be solved, $M^T$ is the transpose matrix of the incidence matrix of time series data, and vector $v$ represents the time series data within a certain period.
This equation is known as the Dirichlet mean energy corresponding to the graph, and can also be expressed as:

\[ D(v) = \frac{1}{n} \sum_{(v_k, v_l)} (v_k - v_l)^2 \]

The Dirichlet mean energy function can effectively metric the non-smoothness of set differences between nodes. A set differences between nodes with small Dirichlet mean energy can be considered as changes in measurements that are relatively smooth on the graph. Conversely, a set of differences with large Dirichlet mean energy can be regarded as non-smoothness or rough.

For a chain graph model constructed from time series data, the expression of the incidence matrix and Dirichlet mean energy function are very concise. The incidence matrix is the matrix \( M \), and the expression for its Dirichlet mean energy function is:

\[ D(X) = \frac{1}{n} \| DX \|^2 = \frac{1}{n} \left[ (X_2 - X_1)^2 + (X_3 - X_2)^2 + \cdots + (X_n - X_{n-1})^2 \right] \]

In the Eq. (8), \( D(X) \) represents the sum of squares of the continuous element differences in vector \( X \), and \( D \) represents the difference matrix, which is the transpose of the incidence matrix of time series data. Since the vector \( X \) here represents a set of time series data, the Dirichlet mean energy function can naturally effectively metric the non-smoothness of time series data.

**Experiment**

For the proposed Dirichlet mean energy function, the non-smoothness of the metric time series data was verified using both the raw observed typhoon wind speed data and the smoothed typhoon wind speed data.

**Experiment data**

To validate the validity of measuring non-smoothness of time series data using the Dirichlet mean energy function, we collected typhoon wind speed data in the East China Sea region during two periods. The data for period one was collected from August 22, 2022, 20:00 to August 24, 2022, 08:00, using four buoy observation systems (No.01, No.06, No.07, and No.09), with a sampling interval of 30 minutes and 73 observations collected for each system. The data for period two was collected from August 22, 2022, 00:00 to August 23, 2022, 24:00, also using four buoy observation systems (No.10, No.12, No.14, and No.20), with a different sampling interval of 10 minutes and 288 observations collected for each system. The time series data for periods one and two are shown in Fig. 2 and Fig. 3, respectively.
Measure of non-smoothness of time series data

Non-smoothness metric for raw time series data

By calculating the Dirichlet mean energy function of typhoon wind speed data from four buoy observation systems in period one and period two, the results show that the Dirichlet mean energy of the four buoy observation systems in period one are 1.354, 0.457, 0.561, and 0.865, among which the Dirichlet mean energy of the No.01 buoy observation system is the largest, while that of the No.06 buoy observation system is the smallest. This indicates that the non-smoothness of typhoon wind speed data from the No.01 buoy observation system is significantly higher than the other three, while the non-smoothness of the No.06 buoy observation system is the weakest. The Dirichlet mean energy of the four buoy observation systems in period two are 0.347, 0.295, 0.317, and 0.346, among which the Dirichlet mean energy of the No.10 buoy observation system is the largest, while that of the No.12 buoy observation system is the smallest. This indicates that the non-smoothness of typhoon wind speed data from the No.10 buoy observation system is significantly higher than the other three, while the non-smoothness of the No.12 buoy observation system is the weakest. The Dirichlet mean energy of typhoon wind speed data from the four buoy observation systems in period one and period two are shown in Fig. 4 and Fig. 5, respectively.

Non-smoothness metric after smoothing of time series data

In addition, two methods of exponential smoothing and moving average were used to smooth the typhoon wind speed data of the No.20 buoy observation system, respectively, to further validate the validity of the Dirichlet mean energy function in measuring the non-smoothness of time series data. The time series data graphs after exponential smoothing and moving average are shown in Figs. 6 and 7, respectively.

From the previous section, the Dirichlet mean energy of the raw observation data of No. 20 buoy was 0.346, and after exponential smoothing and moving average processing, its corresponding Dirichlet mean energy significantly decreased. Among them, the Dirichlet mean energy after single exponential smoothing and double exponential smoothing are 0.167 and 0.098, respectively. The Dirichlet mean energy after single moving average and double moving average are 0.090 and 0.057, respectively. From the smoothed Dirichlet mean energy, compared with exponential smoothing, moving average has a stronger smoothing effect on the raw observation data. The Dirichlet mean energy of typhoon wind speed data observed by the No.20 buoy after exponential smoothing and moving average are shown in Fig. 8 and Fig. 9, respectively.

Results
For time series data with different lengths and sampling intervals, the Dirichlet mean energy function can provide a standardized metric value, and effectively measure its non-smoothness from a numerical perspective. In addition, to demonstrate the effectiveness of this method in measuring the non-smoothness of time series data, we used exponential smoothing and moving average smoothing methods to perform primary and secondary smoothing on the same observation data. After exponential smoothing and moving average processing, the non-smoothness of time series data decreases, and the corresponding Dirichlet mean energy also significantly decreases. This indicates that the Dirichlet mean energy function can effectively measure the non-smoothness of time series data. Furthermore, the Dirichlet mean energy function can also be used to compare and measure the smoothness of different smoothing methods on the raw observation time series data, which can also be confirmed by experiments.

**Conclusion And Future Work**

This paper proposes an effective method for measuring the non-smoothness of time series data. The fundamental purpose of this work is to abstractly express the time series data and their relationships before and after using graph theory, construct a chain graph model for time series data, and establish the incidence matrix of time series data using matrix theory based on the chain graph model. Finally, the paper presents an approach to accurately measuring the non-smoothness of time series data: Dirichlet mean energy function.

Future work involves applying the Dirichlet mean energy function to more diverse fields to further verify its effectiveness in measuring non-smoothness of time series data, as well as extending the application of the Dirichlet mean energy function to high-dimensional time series data.

**Declarations**

**Acknowledgments**

This research was supported by the Fundamental Research Funds for the Central Universities under Grant 2010YD06

**Author Contributions**

L. C. Wang and Y. J. Chen contributed to the study's conception and design. L. C. Wang designed the experiments, analyzed the data, and wrote the first draft. Y. J. Chen proposed key ideas and gave suggestions for the manuscript. W. H. Song participated in the data curation. H. H. Xu gave suggestions for modifications to the manuscript. All authors have read and agreed to the published version of the manuscript.

**Conflicts of Interest**
The authors declare no competing interest.

**Data Availability Statement**


**References**


**Figures**

![Figure 1](image.png)

**Figure 1**

A chain graph model for time series data
Figure 2

Time series data of typhoon wind speed acquired from buoy observation system No.01, No.06, No.07 and No.09, a time series represent by a vector $X$ of length 73.
Figure 3

Time series data of typhoon wind speed acquired from buoy observation system No.10, No.12, No.14 and No.20, a time series represent by a vector X of length 288.
Figure 4

Dirichlet mean energy of typhoon wind speed data of buoy No.01, No.06, No.07 and No.09.
Figure 5

Dirichlet mean energy of typhoon wind speed data of buoy No.10, No.12, No.14 and No.20.
Figure 6

The raw typhoon wind speed time of Buoy No.20 observation system and its two exponentially smoothed (SES and DES), with three vectors of length 288.
Figure 7

The raw typhoon wind speed data of Buoy No.20 observation system and its two moving averages (SMA and DMA), with three vectors of length 288.
Figure 8

Dirichlet mean energy of typhoon wind speed data from the No.20 buoy observation system after single exponential smoothing and double exponential smoothing.
Figure 9

Dirichlet mean energy of typhoon wind speed data from the No. 20 buoy observation system after single moving average and double moving average.