Proceedings Paper DOI: 10.58190/icat.2023.50

PROCEEDINGS OF

INTERNATIONAL CONFERENCE ON ADVANCED TECHNOLOGIES

https://proceedings.icatsconf.org/

11th International Conference on Advanced Technologies (ICAT'23), Istanbul-Turkiye, August 17-19, 2023.

Dissimilarity Metric Score Estimation for Time Series with Missing Values

Hatice Altınok[1], Ahmet Bursalı[1], Sevim Açıksöz[1], Ekin Can Erkuş[1]

[1] *Intelligent Application Department DC, Huawei Technologies Turkey R&D Center Istanbul, Turkey*

hatice.altinok1@huawei.com, ORCID: 0000-0003-3822-0108

ahmet.bursali@huawei.com, ORCID: 0000-0001-7050-769X

sevim.aciksoz1@huawei.com, ORCID: 0000-0003-3953-3245

ekin.can.erkus2@huawei.com, ORCID: 0000-0002-2445-5929

***Abstract*— Missing values in time series data pose significant challenges for data modeling and further analyses. Interpolation methods are often used to fill in the missing values in the data, however, they may cause extra computational complexities and may make the analysis not suitable for real-time operations. Hereby, considering this, this paper focuses on the problem of estimating the dissimilarity metric score for time series data with missing values without interpolating the data. Hereby, we propose an approach to estimate the dissimilarity metric scores without utilizing the imputation methods. Our pro- posed algorithm utilizes a basic, but effective statistical model composed of statistical moments of a time series window to estimate the dissimilarity score of the respective window without applying the interpolation methods. Correlation between the proposed approach scores and the Euclidean dissimilarity metric scores on a benchmark dataset is computed for the most commonly used interpolation methods. To observe the dissimilarity values, several different missing value rates were selected to randomly erase the samples with that ratio from the data. The experimental results show that our proposed method provides comparable correlation results with some dissimilarity measures especially with spline interpolation by creating a correlation coefficient value of 0.819. Hence, the application of such a basic approach to estimating the dissimilarity values without applying interpolation or dissimilarity calculations to observe the time-varying data behavior can be used to reduce the computational complexity in real-time applications.**

Keywords: **Dissimilarity, data imputation, missing value, score estimation, time series**

I. INTRODUCTION

When comparing time series data based on their temporal patterns, dissimilarity metrics assess of how similar or dissimilar two time series are to one another [1]. Decision-making based on temporal patterns is possible through the analysis of time series data, which allows for the extraction of insightful information. Measuring how closely or distantly two time series are similar or dissimilar is a key component of time series analysis and is essential for processes like clustering and anomaly detection, pattern recognition, and data mining [2]. Due to the variety of applications and data characteristics, various types of dissimilarity metrics have been proposed in the literature including, Euclidean distance, square Euclidean distance, Chebyshev distance, and city block distance [3]. The best metric to use depends on the specific use case and the characteristics of the data because each metric has advantages and disadvantages [4].

Due to sensor malfunctions, personal errors, or imperfect data collection techniques, time series data tend to have missing values [5]. For a time series analysis to be accurate, missing values must be dealt with because they can have an impact on the reliability and accuracy of subsequent analyses. By estimating missing values based on nearby or grouped observed data points, interpolation techniques are frequently used to fill in missing values [6]. Even though interpolation has its uses, it can occasionally introduce unnatural patterns that alter the time series' true underlying properties [7]. Also, most importantly for real-time analyses, when working with large datasets, interpolation can also be computationally expensive [8].

The chosen interpolation method affects how well the subsequent analyses perform. There are a few interpolation methods that are most frequently used and selected to be used in this study, such as a quick and easy technique called linear interpolation which draws a straight line between two adjacent data points, or by enabling more intricate fits to the data, polynomial interpolation offers greater flexibility [9]. The spline interpolation method is also used in this study which shows a balance between simplicity and flexibility [10]. Finally, Piecewise Cubic Hermite Interpolation (PCHIP), one of the interpolation techniques employed in this study, entails fitting a polynomial of cubic Hermite form to the provided data points by preserving the monotonicity between adjacent points [11].

Computational complexity becomes a big issue in real-

time analysis scenarios. Data processing needs to be quick and effective for real-time operating systems to adhere to strict system properties [12]. Real-time analysis performance may be hampered by the computational overhead that interpolation introduces. Interpolation for missing value imputation in real-time systems is impractical in some circumstances because the computational cost of interpolation may even exceed the resources available [13].

As a result, the main objective of this study is to suggest statistical method for estimating dissimilarity scores in time series data that contains missing values, particularly within moving windows. Our strategy aims to fill the knowledge gap regarding dissimilarity metric estimation when dealing with missing data, especially in real-time analysis scenarios. We aim to reduce the computational complexity associated with handling missing values in real-time operating systems by building a simple statistical model that directly estimates dissimilarity scores without relying on interpolation. We show that our proposed approach correlates well with the dissimilarity scores on interpolated datasets through extensive experiments and comparisons with existing approaches.

II. METHODS

A. *Interpolation Methods*

1) Linear Interpolation: A quick and effective way to estimate values between known data points is to fit a straight line between two adjacent data points using linear interpolation [14]. The formula for linear interpolation is: given two data samples (x_i, y_i) and (x_{i+1}, y_{i+1})

$$y = x_i + \frac{(x - x_i)(y_{i+1} - y_i)}{(x_{i+1} - x_i)} \tag{1}$$

Simplicity and fast computation are its advantages, and it is mostly used in some quick studies for basic estimation of missing data in time series.

2) Polynomial Interpolation: A degree n polynomial is fitted through $n + 1$ data points using polynomial interpolation. An equation involving polynomial interpolation has the following general form:

$$P_n(\mathrm{x}) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 \tag{2}$$

where a_i are the coefficients to be determined using interpolation conditions. Polynomial interpolation can suffer from overfitting, particularly if the polynomials are of higher order. Using custom polynomials enables flexibility to capture intricate data patterns when the degree is carefully chosen [15].

3) Spline Interpolation: Spline interpolation breaks up the data into more manageable chunks and applies a low-degree polynomial to each chunk to ensure smoothness and continuity. The most typical type of interpolation uses cubic splines, where each segment is represented by a cubic polynomial [14][16]. The cubic spline interpolation equation has the following general form:

$$\begin{matrix} S_i(\mathrm{x}) \text{ if } x_i \le x \le x_{i+1} \\ S_{n-1}(\mathrm{x}) \text{ if } x = x_n \end{matrix} \tag{3}$$

where $S_i(x)$ is the cubic polynomial for segment *i*. Spline interpolation avoids overfitting with a smooth, continuous interpolating curve.

4) PCHIP (Piecewise Cubic Hermite Interpolation): A variation of cubic spline interpolation known as PCHIP ensures monotonicity between adjacent data points. For each segment, a cubic Hermite polynomial that passes through all of the data points and preserves monotonicity is fitted. The PCHIP interpolation formula is as follows when given the two data points (x_i, y_i) and (x_{i+1}, y_{i+1}):

$$y = y_i + h_i \left[0.5 f_i + \frac{(h_i - 1)}{6} f_{i+1} \right] \tag{4}$$

where $h_i = x - x_i$, and $f_i = \frac{(y_{i+1} - y_i)}{(x_{i+1} - x_i)}$.

It provides monotonicity preservation, and interpolating time series data with assured monotonicity may specifically be essential for prediction studies [17].

B. *Euclidean Distance*

The Euclidean distance calculates the straight-line separation between two points in a Euclidean space. It is simple to use, popular, and adaptable to different kinds of data [18]. The following is the formula for the Euclidean distance between two vectors x = (x_1, x_2, . . . , x_n) and y = (y_1, y_2, . . . , y_n):

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{5}$$

C. *Experimental Setup*

This study focuses on estimating the "dissimilarity metric score" without interpolation on time series data with missing values. The experiments are performed using the Python programming language [19] and the Pandas library [20].

1) Dataset: An univariate time series data called Yosemite Temperature Dataset [21] is used for the experiments. The dataset contains more than 18,000 rows and 2 variables: Time and temperature. The data contains daily temperatures in

Yosemite National Park measured at 5-minute intervals between 2017-05-01 and 2017-07-05 dates.

2) Data Preprocessing: Missing values in the original dataset were removed and not used in the analysis. In the original dataset without missing values, 1%, 2%, 5%, 10%, and 20% random samples were deleted to perform independent analyses. Using a sliding window approach, the window size was set to 10% of the dataset (n=1800).

3) Model: Missing data in each window were filled by applying Linear, Spline, Polynomial, and PCHIP interpolation methods. After filling in the missing data, the Euclidean dissimilarity score was calculated with the values obtained by each method and the original values. Then, using a statistical model based on the statistical moments of the time series window, the dissimilarity score of the window is estimated without interpolation methods. The estimated dissimilarity score is computed using the following formula:

$$est(w) = \left| \left(\frac{\#Missing}{\#Total} \right) \frac{\mu\bar{\mu}.(Range)}{(\#NonMissing)^3} \sigma^2 \right| \qquad (5)$$

where $\#Total$ is total number of samples in a window, $\#NonMissing$ is total number of non-missing samples in a window, μ is mean value of non-missing values in a window, $\bar{\mu}$ is median value of non-missing values in a window, $range$ is difference between maximum and minimum value in a window, σ is second statistical moment of series of non-missing value.

4) Performance evaluation: The performance of our proposed method is evaluated by looking at the Euclidean dissimilarity score obtained for each interpolation method and the Pearson correlation coefficient of our estimation method. We also calculated the running times of each interpolation method.

III. RESULTS AND DISCUSSION

The raw data, a sample with 20% missing values, and the Spline interpolation results are visualized respectively in Figure 1, 2, and 3 for the first 100 rows of the Yosemite Temperature dataset.

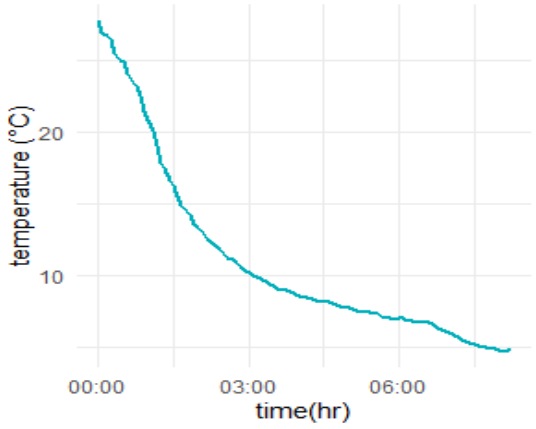

Fig. 1. A sample visualization for the Yosemite Temperature dataset.

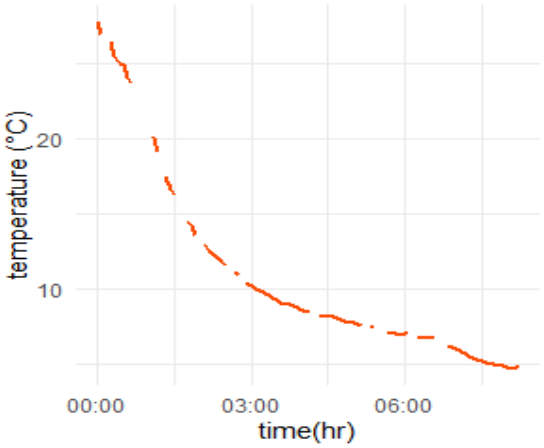

Fig. 2. The remaining data after the 20% random samples were erased from the original set.

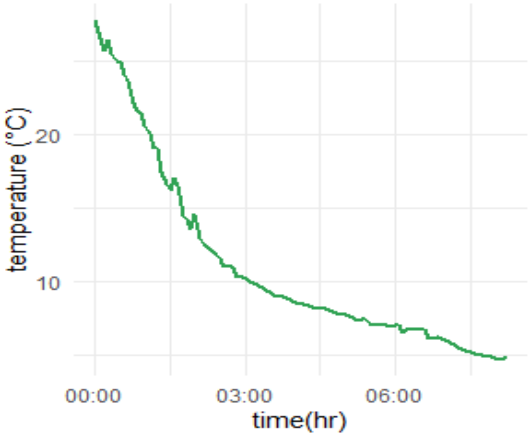

Fig. 3. Spline interpolation results for the 20% missing data sampling.

Table I shows the correlation coefficients between the proposed estimation model and the dissimilarity scores obtained using different interpolation methods and at various missing value ratios.

TABLE I

THE PEARSON CORRELATION COEFFICIENT VALUES BETWEEN THE PROPOSED ESTIMATION MODEL, AND DISSIMILARITY SCORES ON INTERPOLATED DATASET WITH MISSING VALUE RATES.

Missing Value Ratio (%)	**Interpolation Method**			
	Linear	**Spline**	**Polynomial**	**PCHIP**
1	**0.397**	0.164	0.309	0.373
2	**0.399**	0.355	0.381	0.380
5	0.389	**0.659**	0.344	0.398
10	0.347	**0.681**	0.390	0.351
20	0.390	**0.819**	0.470	0.401

Notably, Spline interpolation consistently exhibits the highest correlation coefficients across all missing value ratios. The correlation coefficients reach as high as 0.819 with Spline interpolation, indicating a robust linear relationship

between the proposed estimation model and the dissimilarity scores when utilizing Spline interpolation to fill in missing values.

Table II shows the computation costs in seconds for calculations using different interpolation methods at various missing value rates.

TABLE II

THE ISOLATED COMPUTATION TIME IN SECONDS OF THE IMPUTATION METHODS

Missing Value Ratio (%)	Interpolation Method			
	Linear	**Spline**	**Polynomial**	**PCHIP**
1	**0.025**	0.053	0.185	0.069
2	**0.026**	0.051	0.058	0.032
5	**0.025**	0.053	0.054	0.08
10	**0.026**	0.048	0.047	0.0301
20	**0.027**	0.037	0.039	0.028

According to Table II, the most costly methods in terms of computational complexity are Spline and PCHIP methods, whereas the fastest method is linear interpolation, as expected.

IV. CONCLUSIONS

In this paper, it is aimed to reduce the computational cost of data imputation by using a new proposed method instead of the methods in the literature. Linear, Spline, Polynomial, and PCHIP interpolation methods are used to fill the missing values of each window, and the Euclidean dissimilarity score is calculated. To evaluate the results, correlation between the mentioned interpolation methods and the proposed method is computed.

In order to solve the problem of estimating dissimilarity metric scores for time series data with missing values, our research set out to do so. We suggested a novel statistical method that does not require interpolation and computes dissimilarity scores within moving windows directly. This prevents the introduction of artificial patterns and the compromising of subsequent analyses, while also reducing computational complexity, particularly in real-time operating systems.

We proved the efficacy and efficiency of our method by conducting a thorough evaluation, which allowed us to estimate dissimilarity scores accurately even when dealing with missing values. This is essential for making sure that time series data accurately reflects its true underlying properties and for facilitating reasoned decision-making.

In addition, our study adds to the body of knowledge by providing a useful and effective method for estimating dissimilarity metrics in time series data with missing values. Our algorithm's adaptability and dependability were shown in a variety of use cases, showcasing its potential in practical settings where quick analysis and precise dissimilarity estimation are crucial. Our basic statistical approach can be further extended by incorporating more sophisticated methods or machine learning approaches to improve dissimilarity estimation performance as a basis for further research. Investigating how our algorithm can be applied to various time series data types and its applicability in various domains would also be valuable contributions.

In conclusion, our study advances the discipline of time series analysis by offering a practical method for handling missing values and effectively estimating dissimilarity scores. Hence, the proposed approach can aid in the solution of the time complexity problem of the real-time analyses of time series when dissimilarity metrics are used in data which requires the interpolation methods to be applied for the missing values.

REFERENCES

[1] T. W. Liao, "Clustering of time series data—a survey," *Pattern recognition*, vol. 38, no. 11, pp. 1857–1874, 2005.

[2] J. Han, J. Pei, and H. Tong, *Data mining: concepts and techniques*. Morgan kaufmann, 2022.

[3] S.-H. Cha, "Comprehensive survey on distance/similarity measures between probability density functions," *City*, vol. 1, no. 2, p. 1, 2007.

[4] E. C. Erkuş and V. Purutçuoğlu, "A new collective anomaly detection approach using pitch frequency and dissimilarity: Pitchy anomaly detection (pad)," *Journal of Computational Science*, p. 102084, 2023.

[5] S. A. Imtiaz and S. L. Shah, "Treatment of missing values in process data analysis," *The Canadian Journal of Chemical Engineering*, vol. 86, no. 5, pp. 838–858, 2008.

[6] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, "A survey on missing data in machine learning," *Journal of Big Data*, vol. 8, no. 1, pp. 1–37, 2021.

[7] N. S.-N. Lam, "Spatial interpolation methods: a review," *The American Cartographer*, vol. 10, no. 2, pp. 129–150, 1983.

[8] L. F. Laursen, H. Ólafsdóttir, J. A. Bærentzen, M. S. Hansen, and B. K. Ersbøll, "Registration-based interpolation real-time volume visualization," in *Proceedings of the 28th Spring Conference on Computer Graphics*, 2012, pp. 15–21.

[9] M. A. Azpúrua, E. Páez, J. Rojas-Mora, O. Ventosa, F. Silva, G. Zhang, A. P. Duffy, and R. Jauregui, "A review on the drawbacks and enhancement opportunities of the feature selective validation," *IEEE transactions on electromagnetic compatibility*, vol. 56, no. 4, pp. 800–807, 2014.

[10] P. Benner, S. Gugercin, and K. Willcox, "A survey of projection-based model reduction methods for parametric dynamical systems," *SIAM review*, vol. 57, no. 4, pp. 483–531, 2015.

[11] E. L. Dan, M. Dînşoreanu, and R. C. Mureşan, "Accuracy of six interpolation methods applied on pupil diameter data," in *2020 IEEE international conference on automation, quality and testing, robotics (AQTR)*. IEEE, 2020, pp. 1–5.

[12] R. A. A. Habeeb, F. Nasaruddin, A. Gani, I. A. T. Hashem, E. Ahmed, and M. Imran, "Real-time big data processing for anomaly detection: A survey," *International Journal of Information Management*, vol. 45, pp. 289–307, 2019.

[13] S. Chaturantabut and D. C. Sorensen, "Nonlinear model reduction via discrete empirical interpolation," *SIAM Journal on Scientific Computing*, vol. 32, no. 5, pp. 2737–2764, 2010.

[14] A. Gnauck, "Interpolation and approximation of water quality time series and process identification," *Analytical and bioanalytical chemistry*, vol. 380, pp. 484–492, 2004.

[15] S. Schlegel, N. Korn, and G. Scheuermann, "On the interpolation of data with normally distributed uncertainty for visualization," *IEEE transactions on visualization and computer graphics*, vol. 18, no. 12, pp. 2305–2314, 2012.

[16] P. Thévenaz, T. Blu, and M. Unser, "Interpolation revisited [medical images application]," *IEEE Transactions on medical imaging*, vol. 19, no. 7, pp. 739–758, 2000.

[17] W. Zhu, H. Zhao, and X. Chen, "Improving empirical mode decomposition with an optimized piecewise cubic hermite interpolation method," in *2012 International Conference on Systems and Informatics (ICSAI2012)*, 2012, pp. 1698–1701.

[18] M. R. Berthold and F. Höppner, "On clustering time series using euclidean distance and pearson correlation," *arXiv preprint arXiv:1601.02213*, 2016.

[19] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009.

[20] W. McKinney et al., "Data structures for statistical computing in python," in Proceedings of the 9th Python in Science Conference, vol. 445. Austin, TX, 2010, pp. 51–56.

[21] July 2017. [Online]. Available: github.com/facebook/prophet/blob/main/examples/example yosemite temps.csv