Comparison of Data Interpolation Methods in Time Course Pupil Diameter Data

Mahdieh Farzin Asanjan
Department of Engineering, Middle East Technical University, Ankara, Turkey mahdieh.asanjan@metu.edu.tr

Didem Gökçay

Vilda Purutçuoğlu
Department of Biomedical *Engineering, Middle East Technical University, Ankara, Turkey*
Department of Statistics, Middle East Technical University, Ankara, Turkey
*Corresponding author:
vpurutcu@metu.edu.tr

Fikret Arı
Department of Electrical and *Electronics Engineering, Ankara University, Ankara, Turkey*

Abstract- The missing data problem is one of the main challenges in many datasets. As long as the percentage of loss is under an acceptable range, different methods can be performed in order to fill these unobserved values. In this study the thresholding method, polynomial regression approach, smoothing splines, piecewise linear interpolation and the moving median approaches are used in order to fill the missing data. Among these alternatives, the smoothing spline method typically gives higher accuracy and captures the global feature of the data, whereas, it can eliminate the local changes in the measurements while smoothing. Hereby, in this study, we propose some alternative approaches, called normal ratio and normal ratio weighted with correlation together with modified moving median method in order to fill the missing data. These novel methods are previously applied in meteorological studies where the location of the missing values in a time-course dataset is important.

Keywords- ***Interpolation techniques, normal ratio method, pupil diameter, accuracy measures***

Acknowledgement*: The authors thank to the TÜBİTAK project (no: 117E650) for their support.*

I. INTRODUCTION

In statistical analyses, the interpolation is a type of estimation based on a method of constructing new data points within the range of a discrete set of known data points which have little or no noise. In other words, the interpolation can be defined as an approach of finding intermediate values of a variable from a discrete set of its known values. Moreover, in vast variety of scientific branches like signal processing, heat transfer, climatology and biomedical engineering, since the missing values or erroneous values, which need to be omitted and therefore, become missing values in the further stages of analyses, is a common problem, the interpolation approach has a very wide application.

Accordingly, in biomedical data, as used in this study, several types of interpolation techniques are more preferable [1], [2]. Some of these well-known approaches can be listed as the moving average [3] (i.e., simple moving average, weighted moving average and exponential moving average), smoothing splines [4], and the piecewise linear interpolation methods [5]. In this study, we also propose alternatives namely, the normal ratio and weighted normal ratio methods, which have been used on meteorological data, but have not been applied yet in biomedical studies. We explain our proposal approaches in Section 2. In Section 3, we evaluate the performance of all these methods in time series pupil diameter data. Finally, we conclude our findings and discuss the future work in Section 4.

II. METHODS

In this study, as a novelty, we suggest two alternative methods to fill the missing values. These are the normal ratio and weighted normal ratio, for which, successors are shown in the meteorological datasets.

II.I. Normal Ratio

In literature, the normal ratio method is generally used in meteorological purposes to impute the missing data. In this method, the missing data are found by computing the following expression with the help of weights w which are obtained from the ratios (or differences) of nearby stations.

$$P_x = \frac{1}{n}\sum_{i=0}^{n} \frac{N_x}{N_i} P_i. \quad (1)$$

In the equation (1), P_x denotes the missing value in the time series data, P_i represents the observed values ($i= 1,\ldots,n$) gauged for filling the missing value. Furthermore, N_x implies the mean of the observed values which are measures under the same conditions. Accordingly, N_i describes the mean of the neighborhood observed values which are detected via a window in our analyses similar to N_x, N_i is also computed from the mean of neighborhood observed values which are collected under the same conditions. Finally, n describes the number of these neighborhood observed values to fill the missing values.

II.II. Weighted Normal Ratio

In this study, to adopt the weighted normal ratio, the data are divided into windows with a definite length. Let say 25-neigbor points are taken from both left and right sides of the missing value so that totally 50 observations are used. A missing data point with its neighboring data is named as the target station and the rest of the data in groups of window size, i.e., window defined by ACF, is considered as reference stations. Thus, the corresponding weight for the reference station j defined as

$$w_j = \frac{(n-2)}{1-r_{jt}^2} r_{jt}^2 , \quad (2)$$

where r_{ij} refers to the correlation between the *ith* target and the *jth* reference data while *j*=1,2,…, *N* and *N* is the number of total reference stations. Then, the missing value at the target station is calculated as below:

$$Y_i = \frac{\sum_{j=1}^{N} w_i Y_j}{\sum_{j=1}^{N} w_j} \quad (3)$$

in which Y_i and Y_j denote the estimated *ith* missing value and the *jth* observed reference value (*j=1,2,...,N*), respectively.

III. APPLICATION

III.I. Data Description

In this work a real dataset is used from the pupil diameter changes of 8 users who watch three movies with natural, positive and negative contents. This dataset is collected as a part of the verification procedures of another experiment [6]. In these data, some parts are missing because of the fact that the user blinks and the device cannot measure the diameter of the pupil. Hereby, we implemented the discussed methods in order to estimate those missing data.

III.II. Results

Since our data have lots of noise and also outliers, a user interactive thresholding method is performed to eliminate the outliers. Then, in order to smooth the data, the linear interpolation, smoothing spline method and the polynomial methods are applied. The details of the parametrical settings for the methods described in the earlier section are as follows. In the polynomial method, we take the polynomial degree 9 since it is the highest polynomial degree that can be fitted to the data. The results of these methods on the raw data can be found in Figure1. Looking more closely, it can be seen that in the polynomial degree of 9, only the overall trend of data is obtained and the local variations are ignored. Also the smoothing spline method leads to better results compared to the linear interpolation. Overall, comparing the results of the first three methods, the smooth spline seems to lead to better results and the output is more similar to the input signal and the spikes smoothed better (Fig 1).

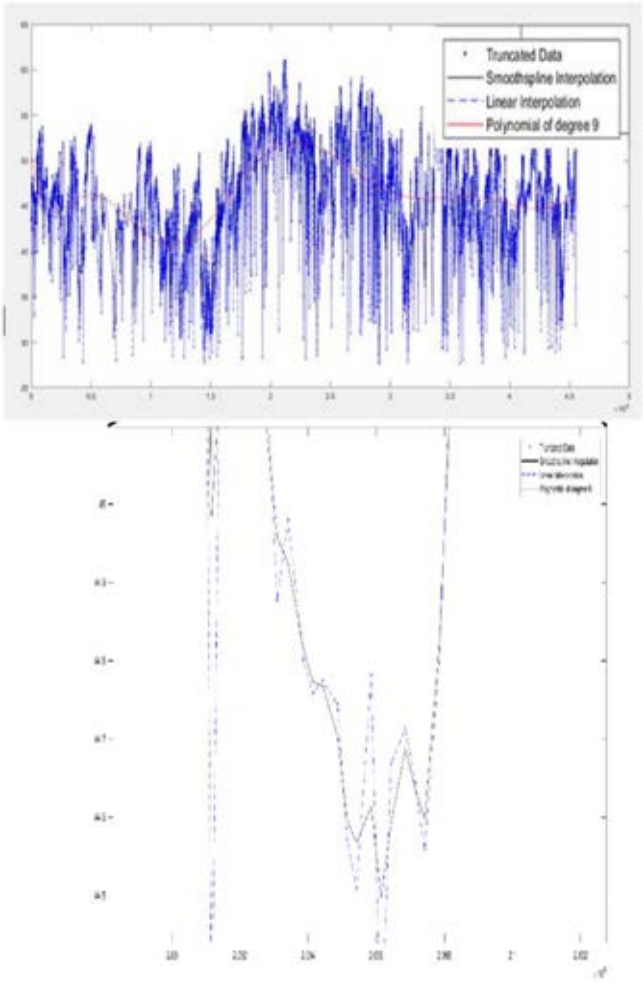

Figure1: Filling incomplete data of a sample observation set of pupil dilation signals with the "smoothed spline", "piecewise linear interpolation" and "polynomial regression" methods after clearing the outlier observation values with the "thresholding technique" (truncated data).

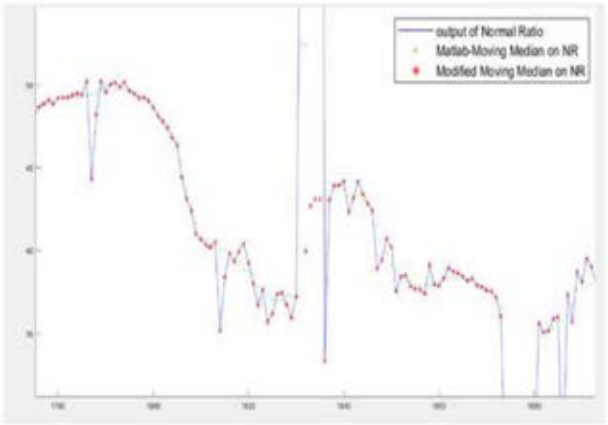

Figure2: The situation in which the pupil dilation signals are discarded with both the "moving median method" and the "modified moving median method" after the "normal ratio method" is filled.

In addition, the normal ratio and the weighted normal ratio methods are also applied to fill in the missing threads and its performance is compared with other methods (Fig 2). In this approach, the proportions (or differences) of the reference neighboring values selected around the missing data are used as weight. While using this method, in this study, the "median movement" method is also modified. That is, outliers in the data are replaced by the value of these adjacent median outliers only if the difference between the median of the neighbors is above a specified threshold. Thus, the outliers are corrected without intervention in the local structure of the signal, thus, without disturbing the original volatility of the data. We consider that of keeping such volatilities in data can be imported to capture the feature selection in the further steps of the analyses.

Table I Absolute mean error (AME) and root mean square error (RMSE) calculated for each imterpolation method.

Method	AME	RMSE
Moving average	0.78	0.83
Exponential moving average	0.43	0.68
Piecewise linear interpolation	0.62	0.79
Smoothing spline	0.9	0.91
Normal ratio	0.51	0.75
Weighted normal ratio	**0.36**	**0.52**

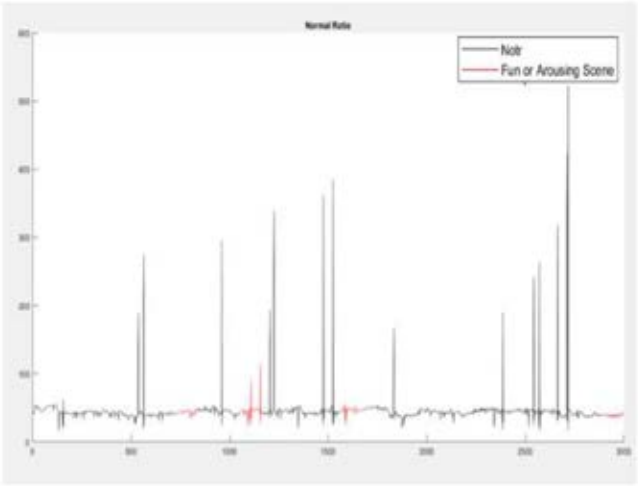

Figure 3: Filling incomplete observations of pupil opening signals using the "normal ratio method"

The results show that the "normal ratio method" captures the local features in the data better when filling in the missing data, and when used in conjunction with the "modified moving median method", it also eliminates the outlier more successfully (Fig 3). For these reasons, we consider that these two new methods have a better performance in both missing data filling and outlier analysis compared to the "regulated spline" method.

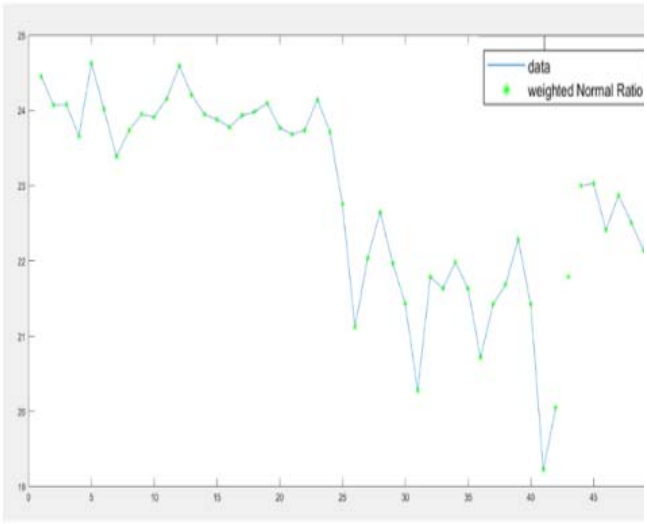

Figure 4: Result of applying weighted normal Ratio on Raw data

On the other hand, in Fig 4, we show the outcomes of the weighted normal ratio on the raw data from the graphs. As it is apparent in this figure, the efficiency of the weighted normal ratio method in predicting the missing values is considerable. While this method applies the other datasets, and the correlation between the target and the reference data as weights, the predicted value becomes more independent than global variations and general trends and hereby, mostly focuses on local trends. Hence, the result of this method obtains more realistic values and can be considered as a promising approach to predict the missing observations.

Finally, in order to compare all methods by evaluating the estimated findings, we accept the part of the observed values as missing data and then, fill these losses via all of the listed five methods. Later, we compute absolute mean error (AME) and root mean square error (RMSE) by comparing the free and estimated values.

The findings of Table I indicates that the estimated values for missing data points are very similar to the true values of the corresponding data points. Having a lower AME/RMSE values presents that the method has lower error to estimate the true value. Overall, the weighted normal ratio, exponential moving average and the normal ratio methods have the lower error valuses compared to that of the other methods for the tested dataset. The smoothing spline method has the largest AME/RMSE values. This outcome implies that this method has a longer error value comparing with the remaing methods to estimate the missing value. The interpolation through the weighted normal ratio proved to be superior to others in terms of errors.

IV. CONCLUSION

In this study, the two new methods have been adapted to be superior for the usage in the biomedical data, after evaluations in corporative analyses. Hence, from the findings it has been observed that once the outliers are eliminated, the smoothing splines approach has a better performance regarding well known alternatives if the data do not have high

variations. On the other side, if the data have fluctuations, the normal ratio and its weighted version can capture these velocities more successful than the spline approach. Therefore, we have concluded that the proposed methods, normal ratio and the weighted normal ratio can be promising alternatives to fulfill the biomedical data too. As an extension of this study, we think that the performance of the proposal approaches can be assessed in distinct biomedical datasets such as EEG, EKG etc. Moreover, the analyses can be extended to get a more comprehensive evaluation by including various scenarios based on Monte Carlo rules under noisy, not noisy or moderately noisy data.

V. REFERENCES

[1] Chang N.F., Chiang C.Y., Chen T.C., Chen L.G., 2011, Cubic Spline Interpolation with Overlapped Window and Data Reuse for On-line Hilbert Huang Transform Biomedical Microprocessor, Conf Proc IEEE Eng Med Biol Soc. 2011;2011:7091-4. doi: 10.1109/IEMBS.2011.6091792.

[2] Chen L., Nguyen□Xuan H., Nguyen□Thoi T., Zeng K.Y. and Wu S.C., 2010, Assessment of Smoothed Point Interpolation Methods for elastic mechanics. Int. J. Numer. Meth. Biomed. Engng., 26: 1635-1655. doi:10.1002/cnm.1251

[3] Khan H., Farooq S., Aslam M., and Khan M., 2018, Exponentially Weighted Moving Average Control Charts for the Process Exponential Ratio Type Estimator, Journal of Probability and Statistics, https://doi.org/10.1155/2018/9413939

[4] Charles A Hall, W.Weston Meyer, Optimal error bounds for cubic spline interpolation, Journal of Approximation Theory, Volume 16, Issue 2, 1976, Pages 105-122, ISSN 0021-9045.

[5] Needham J., 1959, Mathematics and the Sciences of the Heavens and the Earth. Science and Civilisation in China: Volume 3, Cambridge University Press. pp. 147–. ISBN 978-0-521-05801-8.

[6] Gökçay D., Arı F., Avenoğlu B., İleri F. , Erkuş E.C., Balık M., Delikaya A.B., İlerialkan A., Hacıhabiboğlu H., 2019, Preliminary Results in Evaluating the Pleasantness of an Interviewing Candidate Based on Psychophysiological Signals, 15th International Summer Workshop on Multimodal Interfaces, July 8th- August 2nd, Ankara, Turkey