Review of missing values procession methods in time series data

D A Petrusevich

MIREA – Russian Technological University, Prospekt Vernadskogo, 78, Moscow, 119454, Russia

E-mail: petrdenis@mail.ru

Abstract. Missing values is common problem for a lot of time series. Lack of data can be caused with human factor, technical problems non-working measuring stations and so on. Usual methods of handling missing values in time series data suppose that there are models of time series that can make predictions at period one needs to describe. To build them it’s necessary to have data of some time lapse before the period under investigation. Inside of this set of data there shouldn’t be any missing values. So, ordinary approach supposes that there’s a lot of data before the period under question. In this research it’s supposed that missing values can be situated in time series data at any time point. Thus, there’s no whole uninterrupted segment of time series that can be used to train models. Missing values in these time series must be handled first and only after that it’s possible to construct time series mathematical models and make forecasts. At this stage one can evaluate quality of constructed models and whether handled missing values fit known data.

1. Introduction

Support of modern metrological systems often require procession of time series. This type of data usually includes information about sequential measurements of some value under investigation. Often some time values can be wrong or not available. So, problem of missing values procession can be met at any domain of knowledge. Usual approach to time series forecasting supposes construction of mathematical models trained at time series data without missing values. Thorough analysis and construction stages of ARIMA models and other well-known mathematical constructions are available in [1 – 3]. But this approach is inappropriate if there are missing values in time series data. Before the stage of time series construction starts it’s necessary to handle all missing values. In this paper review of methods to deal this problem is presented. Today approaches to this task are based not only on regression and interpolation methods [1 – 3] but they use bootstrap aggregating methods [4], principal component analysis used in multidimensional tasks [5]. Also this problem appears in sight if values of time series are measured with different time steps between them [6]. Models for further analysis of the processed time series models can be found in [7]. Time series procession can be implemented not only in economical mathematical models construction [7] but, for example, in medicine [8] or educational problems research [9]. There are research papers showing that these problems can be solved with game-based methods [10, 11].
2. Methods

Time series values handling methods in this research include polynomial and spline interpolation methods, regression methods, time series construction and neural networks implementation. First of all, the simplest way to handle single missing value is to use values of the closest neighbours. In the formula (1) one can see that absent value $ts(t)$ is obtained from linear expression involving previous value $ts(t-1)$ and the next one $ts(t+1)$:

$$ ts(t) = \frac{ts(t-1) + ts(t+1)}{2}. $$

Another approach that should be mentioned is implementation of the linear regression model trained with use of ordinary least squares method [1–3]. The both methods don’t use any specifical knowledge about time series behaviour and can be used for various tasks. The main problem of this method is construction of single line for the whole dataset. If this method is used locally and uses only local part of time series data its quality decreases. At the same time it can be used to manage a lot of missing values, not only single misses.

Interpolation methods can also be used either with whole time series data or with its part near certain time point under investigation. Polynomial interpolation methods have got well-known problems such as Runge’s phenomenon [12].

To handle such problems one can use spline interpolation methods [1–3] of lower degree than polynomials used in case of interpolation. Spline functions use only local information and they aren’t influenced with data of high variance periods which are situated far from the time point under investigation. Usually cubic splines are implemented. They are free from the Runge’s phenomenon problem. For example, one can read it more thoroughly in [13]. In the experimental part of this work natural cubic splines are used.

Autoregressive models (AR) that are included into complex ARIMA model [1] can also be implemented to use in such cases. Segments of time series without missing values are used as train set to evaluate coefficients in AR or ARIMA model. Missing ones are calculated as forecasts of such models. It’s necessary to have enough data without misses to construct forecasting models. So, if there’s a lot of missed values one have to use cubic splines or other methods.

3. Experiments

Methods enumerated above have been tested at currency exchange rate time series (U.S. dollars and Russian rubles) [14]. There are two types of experiments: procession of single missing values and randomly situated series of misses procession.

3.1. Single missing values procession

Behaviour of the source time series of daily currency exchange rate U.S. dollar / Russian ruble for dates between 2020-01-01 and 2020-04-09 is shown at the figure 1.

![Figure 1. Exchange rate U.S. dollars / Russian rubles time series (single values which are going to be removed are marked with dots).](image-url)
Single values are deleted randomly and they’re situated separately. They’re marked as black dots at the figure 1. Among approximately 250 values there are twelve misses. Missed values are handled with methods presented in the first column of the table 1. Then processed time series is compared to the source one with RMSE metrics [2, 3] shown at the expression (2):

$$RMSE = \sqrt{\frac{\sum_{t} (\tau(t) - ts(t))^2}{N}}.$$  \hspace{1cm} (2)

Here $\tau(t)$ denotes value of the processed time series, $ts(t)$ is value of the source one, $t$ enumerates all handled time points and $N$ is their quantity.

| Method of missing value procession | RMSE |
|-----------------------------------|------|
| Closest neighbours                | 0.52 |
| Linear regression                | 5.67 |
| Linear interpolation             | 0.52 |
| Cubic splines                    | 0.63 |
| Autoregression                   | 0.63 |

Table 1. RMSE of handled currency exchange rate with single misses.

Results are also clearly seen at the figures 2 – 6. If one observes time series with low volatility the best ways to handle single misses are supposed to be “closest neighbours” and cubic splines. They haven’t got Runge’s problem and usually work well. Autoregressive models also show good results but they require a lot of points to evaluate coefficients. The main problem of the linear regression method is clearly seen at the figure 3. There’s single line for all misses. It would be better to construct lines for local surroundings of misses. Basically this idea is foundation of cubic splines.

![Figure 2](image2.png)

**Figure 2.** Exchange rate U.S. dollars / Russian rubles time series handled with “closest neighbours” (1) technique.

![Figure 3](image3.png)

**Figure 3.** Exchange rate U.S. dollars / Russian rubles time series handled with linear regression.
3.2. Procession of missing values series

The same kind of experiment handling series of misses has been made. Number of sequential misses is random but it’s less than 5.

Behaviour of the source time series of daily currency exchange rate U.S. dollar / Russian ruble for dates between 2020-01-01 and 2020-04-09 is shown at the figure 7. Values that are going to be removed are marked with black dots.

Results of missing values procession are shown in the table 2.
Table 2. RMSE of handled currency exchange rate with series of misses.

| Method of missing value procession | RMSE |
|-----------------------------------|------|
| Linear regression                 | 4.36 |
| Linear interpolation              | 0.48 |
| Cubic splines                     | 0.61 |
| Autoregression                    | 0.92 |

Plots of the handled time series are shown at the figures 8 – 11.

Figure 8. Exchange rate U.S. dollars / Russian rubles time series handled with linear regression.

Figure 9. Exchange rate U.S. dollars / Russian rubles time series handled with linear interpolation.

Figure 10. Exchange rate U.S. dollars / Russian rubles time series handled with cubic splines.

Figure 11. Exchange rate U.S. dollars / Russian rubles time series handled with autoregressive models.
Value of RMSE metrics shown in the table 2 shows that linear interpolation, cubic splines and autoregressive models. Again one can see that in case of low volatility the best ways to handle serial misses are supposed to be linear interpolation and cubic splines methods. Autoregressive models can be implemented if series of missing values are far from each other. High volatile time series are going to be observed in future research. Also review of neural network methods is going to be performed.

4. Conclusion
Review of methods that could be used to fill missing values in time series is presented in the paper. In the experimental part the U.S. dollars / Russian rubles exchange rate time series (beginning of 2020) has been investigated. There are methods taking into account nature of data such as autoregressive methods. Also, classical methods (linear regression, linear interpolation and cubic splines) are discussed. Linear regression shows worse results because of lack of training data to evaluate coefficients of such models. Confidence intervals are wide. It would be better to construct them for each single missing value separately and to use only its local surroundings. But this idea is already realized in splines. Neural networks methods are going to be observed and tested in future work. Nowadays this approach is used to handle missing data and make predictions as well as ordinary time series models [15, 16].

According to experiments the best methods to handle single misses are linear interpolation, cubic splines and “nearest neighbours” method (table 1). Further development of autoregressive methods is supposed to be doubtful because there isn’t enough data to evaluate their coefficients.

The same is true in case of series of missing values procession: the best ways to deal with them is to implement linear regression or cubic splines (table 2). Further investigation is going to observe case of high volatile time series and neural networks implementation.

References
[1] Hyndman R J and Athanasopoulos G 2018 *Forecasting: principles and practice* (OTexts) 382
[2] James G, Witten D, Hastie T and Tibshirani R 2015 *An introduction to statistical learning with applications in R* (Springer-Verlag New York) 426 doi: 10.1007/978-1-4614-7138-7
[3] Hastie T and Tibshirani R and Friedman J 2009 *The elements of statistical learning* (Springer-Verlag New York) 533
[4] Andiojaya A and Demirhan H 2019 A bagging algorithm for the imputation of missing values in time series *Expert Systems with Applications* 129 10-26 doi: 10.1016/j.eswa.2019.03.044
[5] Li L, Liu H, Zhou H and Zhang C 2020 Missing data estimation method for time series data in structure health monitoring systems by probability principal component analysis *Advances in Engineering Software* 149 102901 doi: 10.1016/j.advengsoft.2020.102901
[6] Zhang Y, Zhou B, Cai X, Guo W, Ding X and Yuan X 2021 Missing value imputation in multivariate time series with end-to-end generative adversarial networks *Information Sciences* 551 67-82 doi: 10.1016/j.ins.2020.11.035
[7] Andrianova E G, Golovin S A, Zykov S V, Lesko S A and Chukalina E R 2020 Review of modern models and methods of analysis of time series of dynamics of processes in social, economic and socio-technical systems *Russ. Technological J.* (In Russ) 8(4) 7-45 doi: 10.32362/2500-316X-2020-8-4-7-45
[8] Petrushchevich D 2020 Clustering of Covid-19 morbidity cases in Germany *IOP Conference Series: Materials Science and Engineering* 862 042005 doi: 10.1088/1757-899x/862/4/042037
[9] Petrushevich D 2020 The impact of e-learning and social parameters on students’ academic performance *Science for Education Today* 10(6) 143-61 doi: 10.15293/2658-6762.2006.08
[10] Zolkin A V, Lomonosova N V and Petrushevich D A 2020 Gamification as a tool of enhancing teaching and learning effectiveness in higher education: needs analysis *Science for Education Today* 3 127-43 doi: 10.15293/2658-6762.2003.07
[11] Osipova O and Lomonosova N 2019 Application of online courses in the higher education system *Proc. of Int. Multidisciplinary Scientific GeoConf. Surveying Geology and Mining Ecology Management (SGEM)* (Albena, Bulgaria) 19(5,4) 49-54 doi: 10.5593/sgem2019/5.4/S22.007
[12] Runge C 1901 Über empirische Funktionen und die Interpolation zwischen äquidistanten Ordinaten Zeitschrift für Mathematik und Physik 46 224-43

[13] Chen D, Qiao T, Tan H, Li M and Zhang Y 2014 Solving the Problem of Runge Phenomenon by Pseudoinverse Cubic Spline Proc. of 2014 IEEE 17th Int. Conf. on Computational Science and Engineering (CSE) (Los Alamitos, CA, USA) 1226-31 doi: 10.1109/CSE.2014.237

[14] US Dollar/Ruble and Euro/Ruble Exchange Rates and Exchange Trade Indicators Retrieved from: http://www.cbr.ru/eng/hd_base/micex_doc/

[15] Li D, Li L, Li X, Ke Z and Hu Q 2020 Smoothed LSTM-AE: A spatio-temporal deep model for multiple time-series missing imputation Neurocomputing 411 351-63 doi: 10.1016/j.neucom.2020.05.033

[16] Fallah B, Ng K T W, Vu H L and Torabi F 2020 Application of a multi-stage neural network approach for time-series landfill gas modeling with missing data imputation Waste Management 116 66-78 doi: 10.1016/j.wasman.2020.07.034