# Comparing Methods of Imputation for Time Series Missing Values

Renkang Geng, Mingran Li(✉), Mingxu Sun(✉), and Yujie Wang

School of Electrical Engineering, University of Jinan, Jinan 250022, Shandong, China
834331749@qq.com, cse_sunmx@ujn.edu.cn

**Abstract.** Due to the rapid development of modern information engineering, a lot of data are used in machine learning and data cleaning and data mining of the hot research fields, such as a large portion of the data algorithm and related data model are built for complete data set, But in our real life and work, the absence of data exists in a large number of data collection, collation, transmission, storage and other links, it causes many obstacles and difficulties to build a model for complete data. The general way of dealing with missing values for simple delete, that deal with missing value method is a simple convenient but can cause: two aspects of the problem and the inconvenience caused by the original data set to reduce, reduce the reliability of the data, especially in the case of data loss is bigger, can cause a large number of data sets to reduce and missing, This has caused a lot of trouble to our work and research, so we need to find a more efficient and better method than direct deletion. In order to better solve the above problems, we mainly fill in the missing values of time series data, which has become an urgent problem to be solved. In this paper, mean filling, median filling, mode filling, PCA-EM filling and other methods are used to fill traffic data. By comparing these methods, the filling effect of each method is evaluated.

**Keywords:** Missing value · Filling method · Traffic data

## 1 Introduction

As early as the early 20th century, foreign scholars began to study data quality issues, such as data missing and data cleaning. In the late 1940s, the study of data missing problem set off a boom, and experts and scholars proposed various solutions to the missing values. Yates. F, a famous statistician, proposed a method to fill in the missing values because there were too much missing experimental data to complete the data analysis. This method showed a good effect in the different analysis [1]. Then the filling method became a research craze, followed by the mean filling, regression filling, clustering filling, hot card filling, multiple filling and many other classical methods. Entering the 21st century, the processing methods of data missing have become mature, and few new filling ideas have been put forward, most of which are based on the improvement and

This work is supported by Shandong Key R&D Program grant 2019JZZY021005.


S. Wang et al. (Eds.): IoTCare 2021, LNICST 415, pp. 333–340, 2022.
https://doi.org/10.1007/978-3-030-94182-6_24

application of the current field [2, 3]. For example, in 2003, Gebatista et al. analyzed and compared the four missing data processing methods of supervised learning, and proved the performance superiority of the k-nearest neighbour filling algorithm in terms of filling means [4]. In 2018, Zakaria et al. used ambient temperature and humidity monitoring data to evaluate four filling methods (mean filling, regression filling, multiple filling, and nearest neighbour filling). In 2019, Little et al. conducted the foregoing analysis on the latest statistical processing methods of missing data, providing practical application information [5].

The complexity of traffic data, including road network data, road name data, road surface conditions, road name emergencies, weather influence and other factors, under the action of these complex factors, whether the characteristics of traffic data can be accurately extracted and the data analysis and summary, has always been the most important problem in the study of traffic data [6].

The study of traffic data is of great significance to the improvement of urban traffic. In terms of information integration advantages and combination efficiency, it can help us build a better traffic model, reduce traffic congestion and improve traffic safety. However, the lack of data may lead to the deviation of our research results, which is not conducive to the improvement of traffic conditions [7].

For data mining, the existence of missing values causes the following effects: the system loses a lot of useful information; The uncertainty in the system is more significant, the deterministic component contained in the system is more difficult to grasp, and the data containing missing values will make the mining process into chaos, leading to unreliable output. Data mining algorithms themselves are more committed to avoiding the model built by over-fitting data, which makes it difficult for them to handle incomplete data well through their own algorithms. Therefore, missing values need to be derived and filled in by special methods to reduce the gap between data mining algorithms and practical applications.

## 2 Related Work

At present, the processing of missing values is basically divided into three categories: delete, fill, and do not process. According to the absence of data sets and different research contents, different processing methods also show different effects [8]. Although the simple deletion method is easy to operate and fast, it is very easy to cause the loss of valuable information in the data set, and the data distribution is chaotic. Therefore, the application field of this method is very limited. Therefore, we mainly introduce several missing value filling schemes we adopted in this experiment.

Filling methods of missing values can be roughly divided into two categories, namely, statistical methods and machine learning methods. Most statistical methods make assumptions based on the data set itself, and then use the original data set to fill the missing data accordingly [9, 10]. Such methods do not consider the category of the data object itself, and the filling value is often affected by objects of other categories, resulting in poor accuracy of filling results. Common methods include EM filling algorithm, regression analysis, multiple interpolations and so on. Machine learning methods are generally used to classify or cluster missing data sets and then fill them. This kind

of method is rising with the boom of machine learning in recent years, and the representative methods include K nearest neighbor filling, Bayesian network and so on [11, 12].

Mean filling: this method will fill the average value of the data of the existing attribute to the data of the same attribute, this method is simple and convenient, easy to operate and save time and effort, can save cost and time.

Median filling: Similar to mean filling, this method puts together data with observed values for the same attribute and then takes the median to fill in the missing data for the same attribute.

Mode filling. In this experiment, mode filling takes the observed values that appear most in a fixed column to fill in the missing values of other pairs in this column [13].

Probabilistic PCA-EM filling: The first method is a dimension reduction-maximum expectation algorithm. The original data is processed by linear projection technology, and the variance of the data after dimensionality reduction is retained to the maximum extent of the characteristics of the original data [14]. Then, the variance after dimensional reduction is combined with the EM algorithm to calculate the maximum likelihood estimate of the parameter through iteration. Finally, the missing value is filled through the estimated value. This algorithm is particularly efficient in the case of a large amount of data [15, 16].

## 3 Methods

The specific steps of this experiment are as follows: first, read the traffic data and preprocess the data; then, judge the missing values and select the completion method; finally, complete the traffic data and evaluate the complete results. Thereinto, after we read the traffic flow data, we branch the data according to each day, and each column of data corresponds to the traffic flow data of different days in the same period of time. Therefore, each column of data is regarded as the same attribute, and the observed value under the same attribute is used to complete the missing value under the same attribute.

As for the evaluation index of the complete result of this experiment, I adopted the visual analysis method Q-Q diagram. Q-Q diagram refers to a scatter diagram, where Q represents quantile and a probability diagram. The comparison of two data is made by the visualization method. In the figure, the corresponding x coordinate is the real data under the fixed quantile, and the $y$ coordinate is the predicted value (completion value) under the fixed quantile. These two quantiles are the same. If the scatter plot is roughly distributed on both sides of the line $y = x$, then we can determine that the two sets of data are related.

If there is a Q-Q chart, we can judge whether our prediction data are correlated with the complete data. However, we cannot quantitatively describe the relationship between the prediction data and the missing data through the Q-Q chart. Therefore, we adopt several other methods, such as MSE, RMSE, MAE, SMAPE, etc., which can intuitively compare the effect of the complete data. The formula for these indicators is as follows:

$$MSE = \frac{1}{m}\sum_{i=1}^{m}\left(y_i - \hat{y}_i\right)^2$$

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}$$

$$MAE = \frac{1}{m}\sum_{i=1}^{m}\left|(y_i - \hat{y}_i)\right|$$

$$SMAPE = \frac{100\%}{n}\sum_{i=1}^{n}\frac{\left|\hat{y}_i - y_i\right|}{(|\hat{y}_i| + |y_i|)/2}$$

## 4 Results

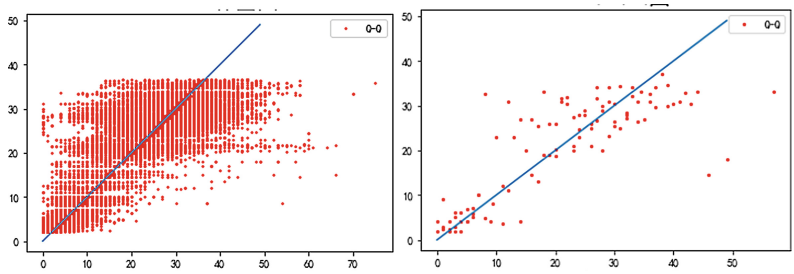


**Fig. 1.** The left figure is the original q-q chart after the mean filling of the data, and the right figure is the q-q chart drawn after the equal spacing sampling of the completed data. The abscissa is the real data, and the ordinate is the completed data. From the figure, we can find that the distribution of the complete data is similar to that of the real data.

**Table 1.** Values of each index under the mean filling

| Index | Value |
|---|---|
| MSE | 58.8 |
| RMSE | 7.7 |
| MAE | 5.4 |
| SMAPE (%) | 33.1 |

After filling with the mean value, we can intuitively see that most points are distributed on both sides of Y = X, so it can reflect that the two sets of data are roughly correlated. When the real traffic flow data is low, most of the points are distributed below the straight line formed by Y = X, and the predicted data of this point is less than the real data. When the real data is 30, our complete data is the most accurate; when the real data exceeds 30, the real data exceeds the predicted data (Figs. 1, 2 and Tables 1, 2).

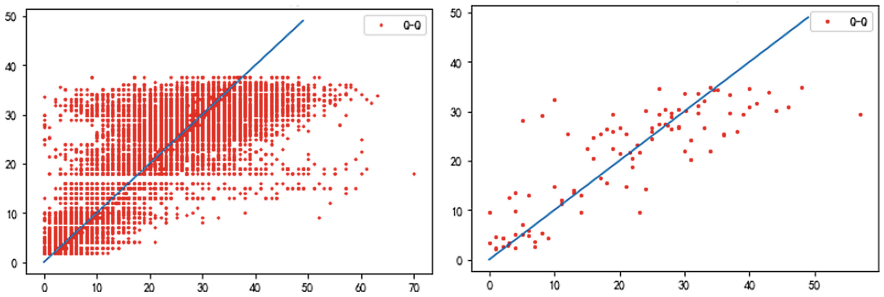


**Fig. 2.** The figure above is the Q-Q diagram after median filling of data, the abscissa is the real data, and the ordinate is the complete data. As can be seen from the figure, the effect of median filling is similar to that of mean filling. When the real data is small, the complete data is larger. When the real data is large, the complete data is small.

**Table 2.** Values of each index under median filling

| Index | Value |
|---|---|
| MSE | 59.8 |
| RMSE | 7.7 |
| MAE | 5.4 |
| SMAPE (%) | 33.3 |

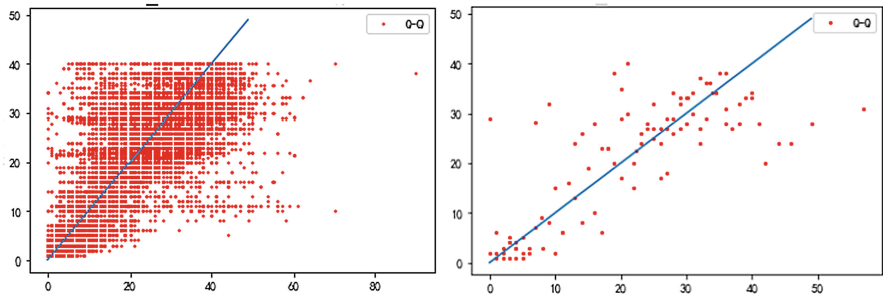


**Fig. 3.** The figure above is the Q-Q graph of data after mode filling. The abscissa is the real data, and the ordinate is the complete data. It can be found in the figure that, compared with mean filling and median filling, the distribution of some points in this case is farther from the base line, indicating that the effect of mode filling is general.

According to the figure and table, we can find that the filling effect of the median is similar to that of mean filling, and the distribution of scattered points is roughly the same as that of mean filling. The filling effect is best when the real data is between 0–10 and about 30. For the traffic flow data used in this study, the accuracy of median filling is slightly lower than that of mean filling. The value of MSE was 59.8, while SMAPE reached 33.3% (Fig. 3).

**Table 3.** Values of each index under mode filling

| Index | Value |
|---|---|
| MSE | 69.9 |
| RMSE | 8.4 |
| MAE | 5.8 |
| SMAPE (%) | 38.2 |

From the data in the Table3, we can see that the effect of mode filling is significantly worse than that of mean filling and median filling. SMAPE is up to 38.2%, about 5 percentage points higher than other filling methods. However, the MSE value reaches 69.9, which is about 10 more than the MSE value of other filling methods, proving that mode filling is not suitable for traffic flow data (Fig. 4).

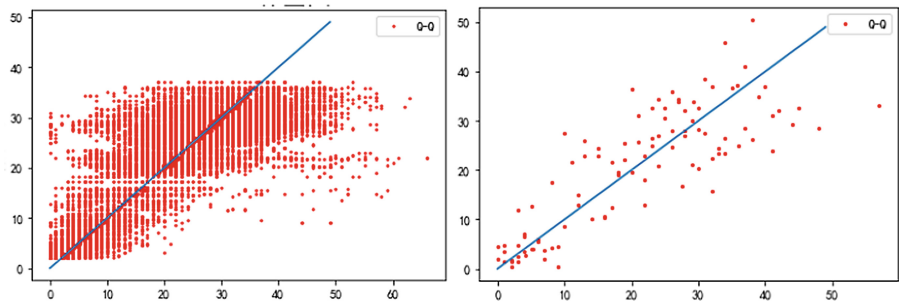


**Fig. 4.** The figure above is the Q-Q diagram after PCA-EM filling of the data. The abscissa is the real data, and the ordinate is the completed data. As can be seen from the figure, no matter what the value of the real data is, the scattered points are always distributed on both sides of the base line, indicating that the filling effect of PCA-EM is good.

**Table 4.** Values of each index under PCA-EM filling

| Index | Value |
|---|---|
| MSE | 56.7 |
| RMSE | 7.5 |
| MAE | 5.4 |
| SMAPE (%) | 33.0 |

As can be seen from the figures in the Table 4, PCA-EM filling is the best filling method among the four methods, with a SMAPE value of 33.0% and an MSE value of 56.7, both of which are the least among the four methods. Therefore, according to the comprehensive analysis of this experiment, for the filling of traffic flow data, the

PCA-EM filling has the best effect, the mean filling and the median filling have similar effects, and the mode filling has the worst effect.

## 5 Conclusion

At present, the research on the problem of missing data has gradually matured, and the processing of missing data is related to various research fields, showing a diversified development. This paper summarizes the research background, causes and types of missing data at home and abroad. All kinds of processing methods are introduced, and filling methods are described in detail. The classical filling methods are compared and summarized, and the latest improved methods of each kind of methods are summarized and compared. At the same time, the commonly used evaluation indexes of data filling effect are introduced from the perspective of parameters and fitting.

## References

1. Fisher, R.A., Yates, F.: Statistical Tables: For Biological, Agricultural and Medical Research. Oliver and Boyd (1938)
2. Ma, L., Sun, B., Li, Z.: Bagging likelihood-based belief decision trees. In: 20th International Conference on Information Fusion (FUSION), Xi'an, China, 1–6 (2017). http://ieeexplore.ieee.org/abstract/document/8009664/
3. Geng, R., Sun, B., Ma, L., Zhao, Q., Shen, T.: Anomaly-aware in sequence data based on MSM-H with EXPoSE. In: 40th Chinese Control Conference (CCC 2021), Shanghai, China (2021)
4. Batista, G.E., Monard, M.C.: An analysis of four missing data treatment methods for supervised learning. Appl. Artif. Intell. **17**(5–6), 519–533 (2003)
5. Sun, B., Cheng, W., Ma, L., Goswami, P.: Anomaly-aware traffic prediction based on automated conditional information fusion. In: International Conference on Information Fusion (FUSION), Cambridge, United Kingdom, pp. 2283–2289. IEEE (2018)
6. Leduc, G.: Road traffic data: collection methods and applications. In: Working Papers on Energy, Transport and Climate Change, vol. 1, no. 55, pp. 1–55 (2008)
7. Sun, B., Cheng, W., Bai, G., Goswami, P.: Correcting and complementing freeway traffic accident data using mahalanobis distance based outlier detection. Tehnicki Vjesnik Tech. Gazette **24**(5), 1597–1607 (2017)
8. Scheffer, J.: Dealing with missing data (2002)
9. Lv, Y., Duan, Y., Kang, W., et al.: Traffic flow prediction with big data: a deep learning approach. IEEE Trans. Intell. Transp. Syst. **16**(2), 865–873 (2014)
10. Ma, L., Sun, B., Han, C.: Learning decision forest from evidential data: the random training set sampling approach. In: 4th International Conference on Systems and Informatics (ICSAI), Hangzhou, China (2017)
11. Little, R.J.A., Rubin, D.B.: Statistical Analysis with Missing Data. Wiley, Hoboken (2019)
12. Sun, B., Cheng, W., Goswami, P., Bai, G.: An overview of parameter and data strategies for K-nearest neighbours based short-term traffic prediction. In: ACM International Conference Proceeding Series 2017, pp. 68–74. ACM (2017)
13. Marlin, B.: Missing Data Problems in Machine Learning (2008)
14. Sun, B., Ma, L., Shen, T., et al.: A robust data-driven method for muti-seasonal and heteroscedastic IoT time series preprocessing. In: Wireless Communications and Mobile Computing (WCMC), p. 6692390 (2021)

15. Yu, L., Snapp, R.R., Ruiz, T., et al.: Probabilistic principal component analysis with expectation maximization (PPCA-EM) facilitates volume classification and estimates the missing data. J. Struct. Biol. **171**(1), 18–30 (2010)
16. Sun, B., Cheng, W., Goswami, P., et al.: Short-term traffic forecasting using self-adjusting k-nearest neighbours. IET Intell. Transp. Syst. **12**(1), 41–48 (2018)