GPS Coordinate Missing Sequence Completion Based on Multivariate LSTM

Sun Xiankun\textsuperscript{1,2}, Hu Xiangyang\textsuperscript{2}, Yin Ling\textsuperscript{2}

1. Shanghai University, Shanghai 200444, China  
2. Shanghai University of Engineering Science, Shanghai 201620, China  
E-mail: 1451819363@qq.com

Abstract: GPS time series completion is of great significance for the analysis and prediction of the motion law of the monitored object. Based on the existing methods of long-term and short-term memory completion, this paper proposes a GPS coordinate time series completion model based on multivariate LSTM. Using the maximum information coefficient method to analyze the GPS coordinate time series data of different observing stations in Shanghai, and find out the data with strong correlation with the target data to be completed, and use these data as input of multivariate LSTM model, using LSTM The model has the characteristics of long-term and short-term memory, and establishes a multivariate LSTM model. By comparing the prediction accuracy with the ARIMA model and the univariate LSTM model, it is proved that the multivariate LSTM has better completion effect on the missing GPS time series.

Keywords: multivariate LSTM; maximum information coefficient; GPS coordinates; time series completion

1 Introduction

Accurate coordinate data of observation points are collected by GPS receivers, and the analysis and research of these data, and then predicting the movement trend of observation points, are widely used in bridge monitoring, building monitoring, mine monitoring and other systems, but in practice In production and life, due to the power failure of the receiver or the impact of natural disasters, the data received by the receiver will be missing, which will affect the prediction of the movement trend of the observation point. In order to reduce this effect, it is necessary to carry out the observation data of the missing part. Completion, due to the long period of the GPS time series itself, high complexity and other factors, making the accurate completion of the missing part of the data becomes more difficult, therefore, the accurate completion of the missing data of the observation point is crucial jobs.

At present, the domestic and international GPS coordinate time series completion methods are mainly based on classical statistical models, such as Lagrange interpolation \cite{1-3}, ARIMA model...
completion method [4-6], Neville interpolation method [7]. Because the classical statistical model relies on the assumptions of time series stability and stability, the stability and accuracy of these completions are greatly reduced with the increase of missing data segments. Therefore, for periodicity and complexity, GPS time series completion does not show a good completion effect. With the rise of deep learning, the deep network represented by Long Short-Term Memory Neural Network (LSTM) [8] has become a popular research direction for time series complementation with its versatility. Yin Ling et al. used the LSTM neural network to complete the GPS coordinate time series [9-13], but this method only performs a simple filtering process on the missing part of the data, which loses the sequence information to some extent. Since the LSTM model requires sufficient context information to learn the mapping of input values to output values, a single input sequence such as Yin Ling cannot provide sufficient context information. This paper explores the introduction of a multivariable LSTM model that can be input in parallel as a model for GPS coordinate time series completion to improve the completion accuracy.

Because GPS receivers located in the same block are affected by many factors such as satellite orbit error, ionosphere, troposphere, multipath effect, crustal movement and other factors in the process of data observation, this makes the observed data The multiple feature indicators are related to each other, that is, the data observed by the receivers located in the same section contain similar temporal and spatial laws.

After correlating the data, the data with strong correlation with the data to be completed is selected as input, sent to the multivariate training model training, and finally the nonlinear relationship between multivariate time series is made by using LSTM network. Dynamic modeling, using SHAO station and surrounding DCMD, SHBS, SHJS station historical data training completion model to complete the missing data of SHAO station U-direction, through the comparison results show the multi-variable based GPS coordinates proposed in this paper The time series model complements the effect better.

2 GPS time series data correlation analysis

The GPS time series data is provided by the China Earthquake Administration's GNSS data product platform. The time series data of the IGS base stations SHAO, DCMD, SHBS, and SHJS stations calculated by Gamit software are 1999-2017, and the 18-year GPS time is checked and downloaded. There are different degrees of missing sequence data. Only the data in 2008 is relatively complete. In view of the experimental data integrity principle, SHAO, DCMD, SHBS, SHJS 2008 E-direction, N-direction, and U-direction are used as a total of 12 data sequences.

There are many factors that affect the completion result of the U-phase data of the SHAO station. If these data are sent to the LSTM model training, it will inevitably make the model learning more difficult. Therefore, it is necessary to SHAO station, DCMD station, SHBS station, SHJS. The data in the three directions of the station ENU is analyzed for correlation, and the data with strong correlation with the U-direction data of the SHAO station to be predicted is selected for use as model training data. In this paper, the maximum information coefficient method [14] is used to select the data input model training with the strongest correlation with the SHA data from the SHAO station, and the low correlation data is discarded to improve the model training efficiency.

The largest information coefficient analysis method is developed on the basis of mutual
information theory. It is mainly used to measure the degree of correlation between two variables. The main idea is to project two variables given to a two-dimensional plane and two variables. The mesh is divided on the scatter plot. According to the approximate probability density distribution of the two variables in the grid, the mutual information values of the two variables can be calculated. After regularization, the maximum value is selected as the maximum information coefficient. This value can be used to measure the correlation between these two variables.

**Definition 1**
Given a finite data set \( D \) of ordered pairs \( <X, Y> \), the sample size is \( m \), dividing the \( x-o-y \) plane into several \( x \times y \) grids \( G \), so that all variables in \( D \) fall into \( G \), defined for:

\[
I^*(X, Y, D, i, j) = \max (X, Y, D|G, i, j)
\]

Equation (1) indicates the maximum mutual information between \( X \) and \( Y \) when \( G \) is divided into \( i \times j \) grids.

**Definition 2**
In data set \( D \) containing two node variables \( X \) and \( Y \), the characteristic matrix of \( X \) and \( Y \) is defined as:

\[
M(X, Y|D)_{i,j} = \frac{I^*(X, Y, D, i, j)}{\log_{2} \min(i, j)}
\]

(2)

**Definition 3**
Assuming that the mesh size of the partition is smaller than \( B(m) \) and \( B(m)=n^{0.6} \), then the maximum information coefficients of the node variables \( X \) and \( Y \) are defined as:

\[
MIC(X, Y, D) = \max_{i \times j < B(m)} [M(X, Y|D)_{i,j}]
\]

(3)

Where \( i \times j < B(n) \) represents the division dimension limit of the grid \( G \). It can be seen from (1) that \( 0 \leq \text{MIC} \leq 1 \). Since the mutual information between random variables is symmetrical, the maximum information coefficient also has symmetry, namely:

\[
\text{MIC}(X, Y|D) = \text{MIC}(Y, X|D)
\]

(4)

The average value of the correlation coefficient between the SHAO station and the DCMD station, the SHJS station, and the SHBS station ENU can be obtained by the equation (4), which are 0.45, 0.83, 0.59, 0.62, 0.65, 0.72, 0.97, 0.68, 0.75, 0.96, 0.65.

![Figure 1. Maximum information coefficient heat map](image-url)
As shown in Fig. 1, the maximum information coefficient heat map of the observation data of each observation station, wherein SHE, SHN, SHU, CDE, CDU, CDN, JSE, JSU, JSN, BSE, BSU, BSN respectively indicate the SHAO direction of the SHAO station, SHAO station N direction, SHAO station U direction, DCMD station E direction, DCMD station U direction, DCMD station N direction, SHJS station E direction, SHJS station U direction, SHJS station N direction, SHBS station E direction, SHBS station U direction, SHBS station N direction. It can be seen from the above results that the data with strong correlation with the SHAO direction of the SHAO station is the U-direction (0.97) of the JSU station and the U-direction (0.96) of the SHBS station. Therefore, the U-direction of the SHAO station, the U-direction of the SHBS station, and the U-direction of the SHJS station are selected. Three sets of data are used for training completion.

3 Multivariate GPS time series completion model
The network structure of the multivariate GPS time series model is divided into three layers: the input layer, the hidden layer, and the output layer. As shown in Fig. 2, the multivariate GPS time series completion model is shown.

Figure 2. Multivariate GPS time series completion model
The input layer mainly includes training set test set partitioning, input data standardization, and conversion to supervised learning. Data standardization refers to converting the input GPS coordinate time series into a time series with a value range of 0-1, mainly to simplify the computation of the model. Then, the data is reshaped into the data input format of the LSTM model standard, that is, the three-dimensional data of samples, time steps, and features. The input layer converts the GPS observation data into supervised learning data, selects t time steps as the training interval, and uses t
data before each time step as the input training data of the current time, and uses the data sample value corresponding to the time as the training sample data. Output values for iterative training.

The hidden layer consists of several LSTM units, which are mainly used to learn and store the information contained in the data. The hidden layer can be reasonably increased and decreased according to the data training effect.

The output layer converts the output data into the same format as the input data, and performs a loss function on the target output data. The weights in the gradient back propagation adjustment formula calculated by the loss function are sent to the hidden layer. The hidden layer updates these weight information into the model and iterates sequentially until the loss function converges. Finally, the output result is inverse normalized and converted into GPS time series data consistent with the input time series.

The network training part uses the theoretical output and the predicted output to calculate the loss function. The Mean Square Error (MSE) is used as the loss function, and the widely used adaptive moment estimation (Adam) algorithm [15] is used as the loss function. Optimizer. With the loss function reaching the minimum as the ultimate goal of training, the data repeatedly iterates through these units and adjusts the weight to reduce the loss function value until convergence. The output layer restores the results to the original data format.

4 experiment

Select 365 u-direction GPS coordinate series of SHAO station in 2008 as the input data of the model. In January 2009, the GPS time series data of the U-direction of the SHAO station is used as the sequence to be completed, that is, the data is missing, and the maximum information coefficient method is used. Analysis of the correlation between the U-view data of the SHAO station and the observation data in other directions, the correlation between the U-direction of the SHBS station, the U-direction of the SHJS station and the U-direction of the SHAO station is strong. Therefore, the experiment selects the U-direction data of the SHAO station in 2008. The SHBS station U-direction data and the SHJS station U-direction data are used as model inputs to complement the January 2009 SHAO station U-direction data. In order to ensure that the experiment is true and effective, the SHAO station U is removed from the data set in January 2009, that is, the data is missing, and the remaining data is divided into training sets and test sets in a ratio of 4:1. After training the model, the data of the missing part is assumed as the input, and the data of the missing part is complemented, and then the completion result is compared with the real data of the missing part to illustrate the completion effect. The experiment selects a solution that completes 30 days, completes 20 days, and completes 10 days to explain the completion effect.

4.1 Experimental environment

The operating system used in the experiment is UBuntu16.04, the programming language python3.7, the algorithm framework is keras2.04; the hardware configuration CPU-Intel i7, memory 8G.

4.2 Experimental results

Select the U-direction data of the SHAO station in 2008 and the most relevant SHBS station U-phase data and SHJS station U-phase data as the model input, and compare the training completion result with the ARIMA model. Figure 5 is a complete complement for 30 days. As a result, Fig. 4 is the result of one completion for 20 days, and Fig. 3 is the result of one completion for 10 days. The orange
curve in the figure is the actual observation of the U-direction of the SHAO station from January 1 to January 30, 2009. The blue dotted line Mul-LSTM Forecast is the complementary value of the multivariate LSTM, and the green dotted line Arima Forecast is the complement of the Arima model. The full value, it can be seen from the figure that the prediction result of the multivariate LSTM is more close to the actual value whether it is completed for 30 days, 20 days or 10 days. The result of Arima's completion is worse than the actual value.

Figure 3. data for one completion 10

Figure 4. data for one completion 20
5. Results evaluation indicators and comparative analysis

In order to facilitate the comparison of the results of the completion, the experiment uses Root Mean Square Error (RMSE) as the complement accuracy. Evaluation criteria, mathematical formulas for root mean square error, such as (10) Shown as follows:

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}
\]  

(5)

Where \( y_i \) and \( \hat{y}_i \) represent the complement value and the true value of the ith data sample, respectively, and \( n \) represents the number of data to be complemented. This paper quantitatively evaluates the fit and complement of the model by calculating the RMSE value of the completion result. Full precision. As in formula (5), the smaller the RMSE, the better the model completion effect.

Table 1 Comparison of the complement accuracy of the three models

| Model          | Number of neurons | batch size | Model parameter of ARIMA \((p, q, d)\) | Completion plan day | RMSE \((m)\) |
|----------------|-------------------|------------|--------------------------------------|---------------------|--------------|
| ARIMA          | -                 | -          | \((1, 1, 1)\)                        | 15                  | 0.0071       |
| ARIMA          | -                 | -          | \((1, 1, 2)\)                        | 30                  | 0.0061       |
| LSTM           | 64                | 10         | -                                    | 15                  | 0.0013       |
| LSTM           | 64                | 15         | -                                    | 30                  | 0.0025       |
| MUL-LSTM       | 32                | 6          | -                                    | 15                  | 0.0012       |
| MUL-LSTM       | 32                | 6          | -                                    | 30                  | 0.0018       |

It can be seen from the comparison results that the multivariate LSTM (this complete model) has the lowest RMSE value and the smallest complement error in the three sets of completion results, regardless of the one-time completion of 15 days or the one-time completion of 30 days. In the experiment, the ARIMA completion model training speed is faster than the multivariate LSTM, but the completion result error is larger. The univariate LSTM completion model is slightly worse than the multivariate LSTM completion model in the one-time completion 15-day scenario, but is significantly worse than the multivariate LSTM completion model in the one-time completion 30-day scenario. It
can be seen that the multivariate LSTM completion model has more advantages in the scheme of complementing multi-day data.

6 Conclusion

Aiming at the problem of insufficient accuracy of GPS coordinate time series complementation, this paper proposes a multi-variable LSTM-based complement method, which not only overcomes the long-term dependence of traditional statistical complement methods, but also effectively improves the completion. Precision. Experiments with different models show that the model is superior to the ARIMA model and the univariate LSTM model in terms of GPS coordinate time series complementation accuracy. The future work of the paper will consider expanding the data set and improving the generalization ability of the model.

References
[1] HE Yujing, YANG Li. A GPS EEG Precise Ephemeris Interpolation Analysis Based on Lagrangian Interpolation Method[J]. Surveying and Mapping Engineering, 2011, 20(5): 60-62.
[2] Liu Jinjian. GPS Satellite Orbital Position Fitting Based on Lagrangian Interpolation Method[J]. Science and Technology Innovation and Productivity, 2018, (7): 11-13.
[3] Xiong Huanhuan, Tian Linya. Sliding Lagrange high and low order combination method in GPS Application of Interpolation in Secret Star Calendar[J]. Site Investigation Science and Technology, 2017(1): 44-46.
[4] Zhao Liang, Lan Xiaoqi, Sheng Jianyue. Application of ARIMA model in satellite clock error prediction[J]. Journal of Water Resources and Architectural Engineering, 2012, 10 (1) : 135-137.
[5] Xu Junyi, Zeng Anmin. Application of ARIMA(0,2,q) Model in Satellite Clock Bias Prediction[J]. Journal of Geodesy and Geodynamics, 2009, (5): 116-118.
[6] Wang, Q, Song, XX. A novel hybridization of nonlinear grey model and linear ARIMA residual correction for forecasting US shale oil production [J]. ENERGY, 2018, (165) 1320-1331.
[7] Chang Yuan, Lin Weihua, Xu Zhanya, et al. Application of sliding Neville interpolation algorithm in GPS precise ephemeris interpolation[J]. Geomatics of Surveying and Mapping, 2017, 42(1): 53-56.
[8] Qin Yao, Li Yong, Wang Shimin. Research on Urban Traffic Congestion Prediction Model with Improved RNN[J]. Electronic world, 2018,(6): 45-46.
[9] Yin Ling, Yin Jingyuan, Sun Xiankun, et al. Neural network completion with missing GPS time series[J]. Journal of Surveying and Mapping Science and Technology, 2018, 35(4): 331-336.
[10] Xu Ning, Xu Changrong. Research on Improved LSTM Deformation Prediction Model[J]. Journal of Jiangxi University of Science and Technology, 2018, 39(5): 45-51.
[11] Yi Lirong, Wang Shaoyu, Yin Lili, et al. Time Series Data Prediction of Industrial Sensors Based on Multivariable LSTM[J]. Intelligent computers and applications, 2018, 8(5): 13-16.
[12] Baek, Y (Baek, Yujin), Kim, HY (Kim, Ha Young). A new forecasting framework for stock market index value with an overfitting prevention LSTM module and a prediction LSTM module[J]. EXPERT SYSTEMS WITH APPLICATIONS 2018, (113): 457-480.
[13] Cao, Kerang; Kim, Hangyung. CNN-LSTM Coupled Model for Prediction of Waterworks Operation Data [J]. Journal of Information Processing Systems. 2018(14): 1508 -1520.
[14] Wang Peng, Zhang Shanceng. Correlation Analysis Method of Delay Data Based on Maximum Information Coefficient[J]. Electronic Measurement Technology, 2015, (9): 112-115.
[15] Yang Guanci, Yang Jing, Li Shaobo, et al. Improved CNN algorithm based on Dropout and ADAM optimizer[J]. Journal of Huazhong University of Science and Technology, 2018, 46(7): 122-127.

Acknowledgments
The successful completion of this paper is inseparable from the help of my family, colleagues and students. Thanks to my family for their support of my work, it is their understanding that makes me I can put a lot of time and energy into my work, thank my colleague Yin Ling for her suggestions during the writing process, and finally for all those who have helped me. Expresssincere gratitude.