# Preprocessing of GPS Coordinate Sequence Based on Singular Spectrum Analysis



# Preprocessing of GPS Coordinate Sequence Based on Singular Spectrum Analysis

**Yingying Ren[1], Lizhen Lian[2*], Jiexian Wang[1], Hu Wang[3]**

[1]College of Surveying and GeoInformatics, Tongji University, Shanghai, 200092, China

[2]CAS Key Laboratory of Planetary Sciences, Shanghai Astronomical Observatory, Chinese Academy of Sciences, Shanghai 20030, China

[3]Chinese Academy of Surveying and Mapping, Beijing, 100830, China

[*]Corresponding author's e-mail: lianlizhen@shao.ac.cn

**Abstract.** Since it is inevitable existence of unknown mutation signals, gross error and data missing in the GPS station coordinates, while the nonlinear signals and misalignment are also neglected by means of piecewise linear detection, thus a novel algorithm called jump determination scheme, which is based on singular spectrum analysis (SSA), is proposed here. The results of real data analysis and simulation tests indicate that the SSA-based pre-processing method has properly detected the gross error and recovered the missing data. Meanwhile, owing to taking into account the correlation in the position time series, it can effectively avoid the influence of gross error on the accurate determination of each uncertain jump. In addition, it can retain the intrinsic characteristics in station coordinates, which has vital significance for the gross error detection and missing data recovery. By this new method, even if the missing rate of data reaches 10% or more, we can eliminate the gross error and then obtain complete data with an accuracy of better than 5 mm and 1 cm for the horizontal components and the height component, respectively.

## 1. Introduction

The continuous GPS station coordinates is one of the basic data in the fields of deformation monitoring, plate motion, etc. At present, most of study on the characteristics of GPS station nonlinear motion have been mainly analysed by use of power spectrum analysis and wavelet analysis to determine the periodic oscillations in time series of coordinate component at some GPS stations. Then, a suitable mathematical model, which involved linear velocity, amplitude and phase of significant periodic terms, etc., was set up and solved in terms of the least squares principle[1]. However, due to the existence of uncertain jump, gross error and the missing in the GPS station coordinates, the intrinsic characteristics such as the estimated station speed in the time series will be seriously affected. Therefore, properly pre-processing the original coordinates, including accurate determination of the occurrence epoch of jump, detection and elimination of gross error and recovery of the missing coordinates with good internality, is the premise to understand the characteristics of the station motion. Singular Spectrum Analysis (SSA) is a sequence processing method which is also based on Principal Component Analysis (PCA). The SSA can extract various components, i.e. trend and periodic terms, and reduce the noise by deconstructing the time series, without any model assumptions. Hence, it has

been widely used in meteorology, oceanography, geography and so on. A large amount of literatures has been focused on the application of this method in time series analysis[2-5].

Jump existing in the time series is a mutagenic signal without any regularity and can severely destroy the intrinsic properties of the sequence. Since the causes of the jump are numerous, and the occurrence epoch and magnitude of the jump are unknown, the accurate determination of the jump has always been an intractable problem. At present, there are few studies on jump determination in China, and they are all limited to relying on piecewise linear least squares detection[6-8], which ignored the periodic information and other signals in the sequence, so the result of jump determination is not ideal. In this paper, a jump determination method based on SSA is proposed. Under the premise of setting the appropriate threshold, an ideal result of jump determination can be obtained. Meanwhile, the application of SSA in gross error detection and missing data recovery is also studied. The results show that SSA has high reliability in terms of pre-processing the time series.

## 2. Principle and method

*2.1The basic principle of SSA*

For a given one-dimensional time series $\boldsymbol{x} = \{x_t, t = 1, \dots, N\}$, the SSA is mainly divided into the following steps[9-11]:

(1) Calculate the Lag self-covariance matrix X of the original sequence (i.e., the trajectory matrix X with a delay amount of 1 based on an embedded dimension L)

$$\boldsymbol{X} = \begin{bmatrix} x_1 & x_2 & \cdots & x_{N-L+1} \\ x_2 & x_3 & \cdots & x_{N-L+2} \\ \vdots & \vdots & \ddots & \vdots \\ x_L & x_{L+1} & \cdots & x_N \end{bmatrix} \tag{1}$$

(2) The singular value decomposition is performed on X to obtain its left singular vector U, eigenvalue matrix S, and right singular vector V, respectively.

$$\boldsymbol{X} = \boldsymbol{USV} \tag{2}$$

$$\boldsymbol{E} = \boldsymbol{U} \tag{3}$$

$$\boldsymbol{a} = \boldsymbol{SV} \text{ or } \boldsymbol{a} = \boldsymbol{E^T X} \tag{4}$$

Wherein, the diagonal element of S is called eigenvalue sorted from large to small. The vector E represents a time empirical orthogonal function (T-EOF) and the vector $\boldsymbol{a}$ represents a temporal principal component (T-PC).

(3) The original sequence can be reconstructed by using T-EOF and its corresponding T-PC. The main components are selected in the reconstruction process to achieve the purpose of information extraction and noise filtering. The calculation formula of the reconstructed components is as follows:

$$\begin{cases} y_i = \frac{1}{M}\sum_{j=1}^{M}\sum_{k\in \boldsymbol{A}} \boldsymbol{a}_{i-j}^{k}\boldsymbol{E}_j^k, & M \le i \le N-M-1 \\ y_i = \frac{1}{i}\sum_{j=1}^{i}\sum_{k\in \boldsymbol{A}} \boldsymbol{a}_{i-j}^{k}\boldsymbol{E}_j^k, & 1 \le i \le M-1 \\ y_i = \frac{1}{N-i+1}\sum_{j=i-N+M}^{M}\sum_{k\in \boldsymbol{A}} \boldsymbol{a}_{i-j}^{k}\boldsymbol{E}_j^k, & N-M+2 \le i \le N \end{cases} \tag{5}$$

*2.2Jump determination method for sliding median based on SSA with IQR criterion*

There are many reasons for the jump occurrence in the GPS station coordinate sequence[12], such as earthquake impact, change of station position, update of receiver or antenna, uncertainty of antenna height measurement, multi-path problem, solution accuracy problem, and reference datum. The jump level caused by various reasons is different, and its determination is affected by the gross error. Therefore, accurately determining the jump in the GPS station coordinates has always been an intractable problem. Although some larger jumps can be first identified manually, it is also necessary to query the relevant site logs to find out the reason for the jump to determine the jump occurrence epoch accurately, and finally perform the time series analysis and fitting.

The jump determination method based on the Singular Spectrum Analysis-Moving Median-Interquartile Range (SSA-MMED-IQR) with the IQR criterion[13] proposed in this paper is an effective way to avoid the impact of the gross error on the jump determination. The SSA-MMED process is mainly divided into the following steps.

(1) Reconstruct the original series by selecting some principal components produced by the SSA. Since the SSA has a certain ability to resist gross error and a small amount of gross error does not affect the properties analysis of the overall data, thus the reconstructed sequence greatly reduces the influence of the gross error, which can avoid the influence of gross error on the jump determination.

(2) Setting a proper sliding window w and a jump detection threshold e, extract two sub-sequences, named sub1 and sub2 with the data length of w/2, which are before and after the reconstructing sequence position i respectively. According to the IQR criterion, we further determine whether there exists gross error in each sub-sequence. If there is no gross error, the difference between the median of sub1 and sub2, which are recorded as med1 and med2 respectively, is further calculated. If their absolute difference is smaller than the detection threshold e, the position i is considered to be the occurrence epoch of a jump.

(3) Perform jump correlation analysis of the NEU time series. Since there is a certain correlation between the jumps of the GPS station coordinates in the NEU direction, in order to further determine the occurrence epoch of each jump, the following criterions are used here: 1 If there exist two or more jumps in the NEU components during a similar period, the jump occurred in the period. 2 If any jump magnitude in the NEU time series is greater than a given threshold during a certain period, there must be jump in the period.

(4) K-means cluster analysis is performed on the occurrence epochs of jump to obtain the specific jump occurrence epochs.

The IQR criterion is a method to measure the concentration and dispersion of a one-dimensional time series sorted from large to small by using a median value and a standardized IQR, both of which are not affected by outliers in the sequence and have the property of a robust estimate. The difference between the high quartile value (the nearest value at three-quarters of the ordered sequence) and the low quartile (the nearest value at the quarter of the ordered sequence), multiplied by the factor of 0.7413, is the standard IQR. According to the IQR criterion, the fractional statistic Z-ratio is listed as follows. When $Z \geq 3$, it is viewed to be an outlier at the confidence level of 99%.

$$Z = \frac{v_i - median(v_{i-w/2,i+w/2})}{standardization\ IQR(v_{i-w/2,i+w/2})} \tag{6}$$

*2.3Gross error detection and missing data recovery method*

The sliding median gross error detection method based on SSA is a non-parametric model and has robustness for gross or outliers. The residual series, the difference between the original series and the reconstructed series obtained by SSA, are used for the gross error detection. According to the IQR criterion, we determine whether there exists gross error in the selected window length w, and then slide the window step by step until covering the entire series. After the first gross error detection is completed and the determined gross error is eliminated, a second gross error detection is performed iteratively until no gross error is detected in the series.

The SSA for Missing (SSA-M) data recovery method[14-15] is proposed by Kondrashov and then applied to GPS coordinates interpolation. The specific steps are: (1)The missing in the original coordinates are marked and their initial value equal to zero (each gross error is regarded as a missing, too), and the coordinate time series is normalized. (2) The SSA is used to decompose the series to obtain PCs and their corresponding RCs. The first RC (R1) is selected to replace the missing values in the original sequence, and iterates to convergence, that is, the inner loop process. (3) After the end of the first inner loop, the missing is reconstructed by adding R2, that is, the missing is recovered by linear superposition of R1 and R2, and then step 2 is repeated until convergence, that is, the outer loop process. With this, the missing data recovery is completed.

## 3. Case analysis and simulation test

*3.1Analysis of station coordinate sequence examples*

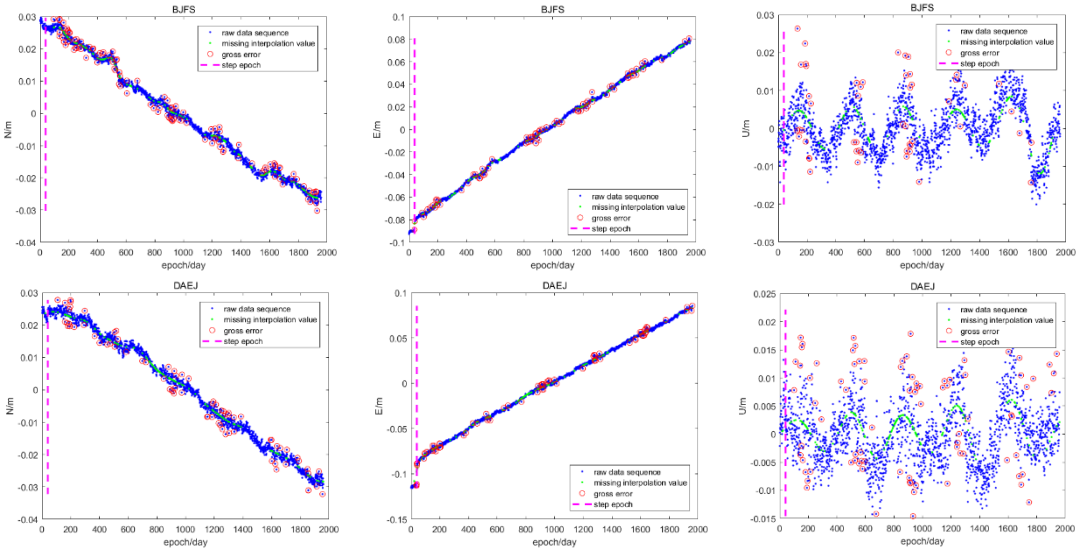
Table 1. Site coordinate sequence processing result.

| site | Jump epoch /day | Jump epoch /ymd | Gross number /NEU | SSA-M inner loop threshold/mm |
|---|---|---|---|---|
| BJFS | 39 | 20110311 | 111, 72, 56 | 0.5 |
| DAEJ | 40 | 20110311 | 76, 75, 91 | 0.5 |
| YNYS | 407, 1067 | 20120312, 20130813 | 29, 11, 55 | 0.5 |

In this paper, the station BJFS, DAEJ and YNYS are taken as examples. The coordinates obtained by GAMIT/GLOBK for time series adjustment are used to perform jump determination, gross error detection and missing data recovery. The statistical results are shown in Table 1. The SSA-MMED-IQR method determined the jump occurrence epoch of BJFS and EADJ on March 11, 2011, which was consistent with the fact that a 311 earthquake (9.0) happened in Japan in 2011. According to the IQR criterion, the number of gross errors detected in the NEU direction of station BJFS and DAEJ are (111, 72, 56) and (76, 75, 96), respectively. The jump occurrence epochs of the station YNYS are March 12, 2012 and August 13, 2013. However, according to the seismic data, a magnitude 5.9 earthquake occurred in the junction between Sichuan and Chongqing on August 31, 2013. Apparently, the jump determination was not accurate, which may be because there is a lack of some data during that period as shown in Figure 2. In other words, the existence of the missing may reduce the accuracy of the jump detection by SSA-MMED-IQR. In addition, there was no earthquake happened on March 12, 2012, the jump occurred at this epoch may be due to other reasons. The number of gross errors detected in the NEU components of the station YNYS are 29, 11 and 55, respectively. The interpolation internal loop threshold is set to 0.5 mm in this work.

Figure 1 represent horizontal and height component time series (the blue scatter) of stations BJFS, DAEJ and YNYS exhibiting significant gross error (the red circle) and jump occurrence epochs (the dashed line) in the time span 2010-2016.

From Table 1 and Figure 1, it can be concluded that the jump determination method based on SSA has accurate results, but it is affected by the jump threshold and the missing data. The recognition of gross error is better, and the accuracy of the missing data recovery is reliable.



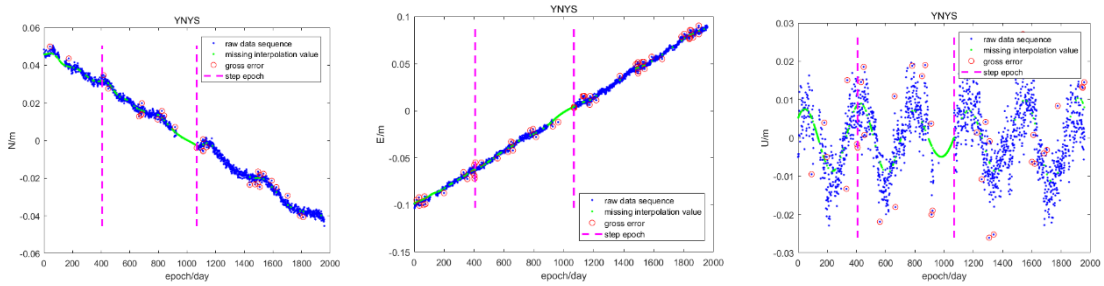



Figure 1. Comparison of pre-processing results of station coordinate sequence

*3.2Station coordinate sequence simulation test*

In order to further verify the accuracy and reliability of the new pre-processing method in the jump determination and the missing data recovery, we take the station BJFS as an example and randomly add some jump terms and the missing in the simulation data based on the station coordinates.

(1) jump number simulation and test analysis

As shown in Table 2, except for the jumps caused by the Japanese 3.11 earthquake in the original sequence of BJFS station coordinates, this experiment adds jump values of +6mm, -7mm and 2cm at different epochs in the NEU direction to perform jump determination.

Table 2. BJFS station coordinate simulation step data mode.

| jump epoch/day | jump sequence | jump value | remark |
|---|---|---|---|
| 39 | E, U | —— | earthquake |
| 400 | N | +6mm | simulation |
| 1000 | E | -7mm | simulation |
| 1400 | U | +2cm | simulation |

According to the above method, the jump in the simulation data is determined. The result is shown in Figure 2. The jumps existing in BJFS station coordinate time series are accurately determined. The jump occurrence epochs are accurately determined at epoch of 39, 400, 1000 and 1400. The SSA-MMED-IQR jump determination method has strong reliability.

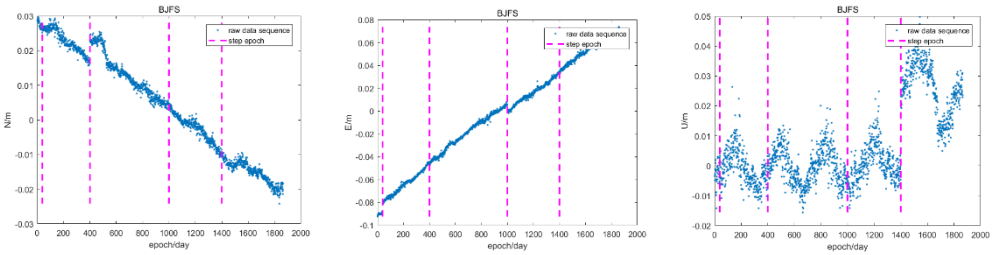



Figure 2. Result of jump determination in the simulation station coordinate sequence

(2) As shown in Table 3, this experiment simulates the missing data segments in 4 different spans as well as the random missing data appearing in the BJFS original coordinate sequence, and the day number of the missing is different. The missing modes are divided into random and continuous, where the maximum day number of consecutive missing is less than 2 months. The total loss rate of data is over 10%.

Table 3. The mode of missing data simulation based on BJFS station coordinate sequence

| missing segment | start time | end time | missing days | maximum missing consecutive days | mode |
|---|---|---|---|---|---|
| 1 | 001 | 1865 | 95 | 15 | random |
| 2 | 471 | 500 | 30 | 30 | continuous |
| 3 | 961 | 1000 | 40 | 40 | continuous |
| 4 | 1551 | 1600 | 50 | 50 | continuous |

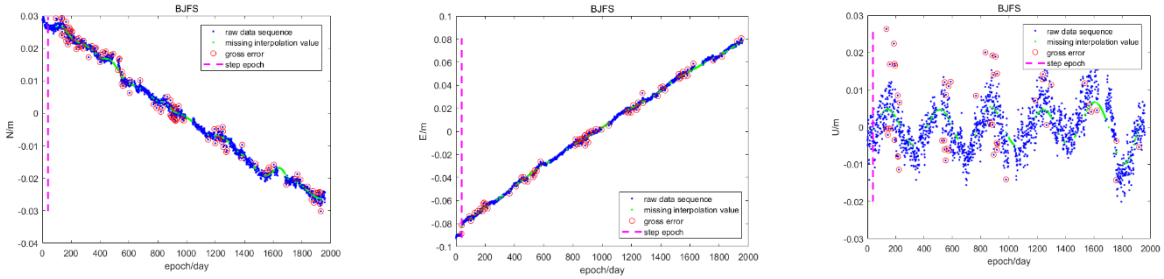



Figure 3. The recovery result of the simulation station sequence with missing data

The simulated missing data is recovered by use of the above method, and the recovery result is shown in Fig. 3. The accuracy of the missing data recovery is shown in Table 4. It can be seen from the Table 4 that since the gross error is eliminated, the accuracy of the missing data recovery is better than 5 mm in the horizontal direction and 1 cm in height direction, respectively. It can be concluded that the SSA-M method in the missing data recovery can achieve result with high precision.

Table 4. BJFS station coordinate sequence interpolation precision.

| direction | verage error / mm | medium error / mm |
|---|---|---|
| N | 1.5 | 1.8 |
| E | 1.6 | 2.2 |
| U | 3.7 | 4.6 |

## 4. Conclusion

Based on the SSA method, a new effective pre-processing method is applied in the GPS coordinate time series, involving jump determination, gross error detection and missing data recovery. By this pre-processing, we can obtain a more accurate initial value as a reference and a clean and complete time series for subsequent time series analysis. Through the analysis of the example and the simulation experiment, the following conclusions are drawn:

(1) The SSA-MMED-IQR jump determination method can avoid the influence of gross error on the jump, and the result of jump determination is relatively accurate with a certain reliability, but it is sensitive to the selected threshold for jump detection. If the threshold is too small, then some false jump terms are included. In addition, some inaccurate jump may be identified due to the existence of missing data.

(2) The SSA-MMED-IQR gross error detection method can better perform gross error detection. Compared with the traditional detection method applied in GPS station coordinate time series, since it does not require any priori model and parameter and can retain the inherent characteristics of the data, thus it can avoid the impact of the difference between the prior model and the real data on the gross error detection.

Owing to retaining the intrinsic characteristics of the data, the accuracy of the result by use of the SSA-M missing data recovery method is high. The simulation experiment indicates that the missing can be recovered with a high-precision when the data missing rate are higher than 10%. Once an appropriate reconstruction order is set, even if abrupt change happened in the time series, the recovery of the missing data can be still performed well. However, if some abrupt changes exist in the period of the missing, the information about abrupt changes cannot be reconstructed.

In addition to the application in pre-processing of GPS coordinates, SSA also plays an important role in subsequent time series analysis, such as the separation of trend component and various periodic components. Based on pure trend component, we can obtain station motion speed with a higher precision.

**References**

[1] Wu Shuguang. Analysis of Coordinate Time Series Characteristics of Regional CORS Station[D]. Wuhan University, 2017.
[2] HUANG Wei, TIAN Lin-ya, SUN Teng-ke, MA Bing-hao. Analysis of periodic characteristics of regional CORS network coordinate time series[J]. Surveying and Spatial Geography Information, 2015,38(09):139-141.
[3] Lu Chenlong. Application of singular spectrum analysis in geodetic time series analysis[D]. Central South University, 2014.
[4] XU Ke-hong, CHENG Peng-fei, WEN Han-jiang. Singular Spectrum Analysis and Wavelet Analysis of Time Series of Sunspot Numbers[J]. Science of Surveying and Mapping, 2007(06):35-38+205.
[5] Dai Haomin, Xu Aiqiang, Sun Weichao.Signal Denoising Method Based on Improved Singular Spectrum Analysis[J]. Journal of Beijing Institute of Technology, 2016,36(07):727-732+759.
[6] Zhang Wang. Extraction of GPS coordinate time series trend items and seasonal item information based on improved singular spectrum analysis method[D]. Southwest Jiaotong University, 2017.
[7] JIANG Weiping, LI Zhao, LIU Hongfei, ZHAO Qian. Analysis of the cause of nonlinear variation of coordinate time series in China regional IGS base station[J]. Chinese Journal of Geophysics, 2013,56(07):2228-2237.
[8] Su Weixing, Zhu Yunlong, Liu Fang, Hu Yuyuan. A Detection Algorithm for Time Series Abnormal Points and Mutation Points[J]. Journal of Computer Research and Development, 2014,51(04):781-788.
[9] Wang Jiexian, Lian Lizhen, Shen Yunzhong. Application of Singular Spectrum Analysis in Sequence Analysis of GPS Station Coordinates Monitoring[J]. Journal of Tongji University(Natural Science), 2013,41(02):282-288.
[10] Jia Yongna. Seismic data interpolation based on improved singular spectrum analysis method [D]. Harbin Institute of Technology, 2014.
[11] Wang Hua. Research on Dynamic Maintenance Theory and Method of National Land Reference Frame [D]. Wuhan University, 2015.
[12] Bruni S, Zerbini S, Raicich F, et al. Detecting discontinuities in GNSS coordinate time series with STARS: case study, the Bologna and Medicina GPS sites[J]. Journal of Geodesy, 2014, 88(12): 1203-1214.
[13] Cheng Pengfei, Cheng Yingyan, Mi Jinzhong. Theory and Practice of the Establishment of National Geodetic Coordinate System [M]. Beijing: Surveying and Mapping Press, 2017.
[14] Wang Xiaoming. Research on Dynamic Characteristics and Nonlinear Modeling Method of CGCS2000 Coordinate Frame[D]. Chinese Academy of Surveying and Mapping, 2013.
[15] Wang Huizan, Zhang Ren, Liu Wei, Wang Guihua, Jin Baogang. An Improved Algorithm for Singular Spectrum Iterative Interpolation and Its Application in Defect Data Recovery[J]. Applied Mathematics and Mechanics, 2008(10):1227-1236.