EURASIP Journal on Advances in Signal Processing

RESEARCH

Open Access

A novel outlier detection method based on Bayesian change point analysis and Hampel identifier for GNSS coordinate time series

Hüseyin Pehlivan[1*]

*Correspondence: hpehlivan@gtu.edu.tr

[1] Department of Geomatics Engineering, Gebze Technical University, Kocaeli, Turkey

Abstract

The identification and removal of outliers in time series are important problems in numerous fields. In this paper, a novel method (BCP-HI) is proposed to enhance the accuracy of outlier detection in GNSS coordinate time series by combining Bayesian change point (BCP) analysis and the Hampel identifier (HI). By using BCP, change points (cps) in the time series are lidentified, and so the time series is divided into subsegments that have properties of a normal distribution. In each of these separated segments, outliers are detected using HI. Each data element identified as an outlier is corrected by a median filter of window size (*w*) to obtain the corrected signal. The BCP-HI method was tested on both simulated and real GNSS coordinate time series. Outliers from three different synthetic test datasets with different sampling frequencies and outlier amplitudes were detected with approximately 98% accuracy after processing. After this process, Signal-to-Noise Ratio (SNR) increased from 0.0084 to 10.8714 dB and Root Mean Square (RMS) decreased from 24 to 23 mm. Similarly, for real GNSS data, approximately 98% accuracy was achieved, with an increase in SNR from 0.0003 to 4.4082 dB and a decrease in RMS from 7.6 to 6.6 mm observed. In addition, the output signals after BCP-HI were examined graphically using Lomb–Scargle periodograms and it was observed that clearer power spectrum distributions emerged. When the input and output signals were examined using the Kolmogorov–Smirnov (KS) test, they were found to be statistically similar. These results indicate that the BCP-HI algorithm effectively removes outliers, and enhances processing accuracy and reliability, and improves signal quality.

Keywords: GNSS data, Outlier, Hampel identifier, Bayesian change point, Time series

1 Introduction

Global Navigation Satellite System (GNSS) measurements have become increasingly popular for determining time-dependent location, amplitude, and velocity changes, particularly in geodetic and geophysical applications. However, when measuring conditions are limited, acquiring low-quality GNSS data is unavoidable [1]. Such data can be influenced by a variety of error causes, such as multipath effects, atmospheric delays, satellite and receiver hardware problems, and other noise sources, resulting in potentially substantial outliers. Outliers can have a significant impact on the accuracy and dependability of GNSS data analysis, preventing accurate conclusions from being obtained. Detecting and eliminating outliers is therefore critical for improving the quality of position time series and properly determining displacement amplitudes and oscillation parameters [2].

Studies on outlier analysis in GNSS time series includes a variety of methods and applications [3–11]. These approaches have been extensively researched and include statistical tests [12–15], filtering strategies [16–18], and time series models [19–24].

While these methods are effective at detecting and removing outliers in time series data, they do have some limits and disadvantages. For example, statistical tests such as Grubbs and Tukey's test are sensitive to data distribution (assumed to be normal distribution) and underlying statistical assumptions, rendering them less effective in the presence of large or focused outliers [25]. Similarly, by smoothing the data, filtering methods such as median filters, moving average filters, and Kalman filters can successfully mitigate the influence of noise and outliers. They may, however, fail to adequately eliminate outliers. However, when the data is excessively noisy or has complicated underlying trends, they may fail to successfully exclude outliers. Furthermore, these methods may cause phase changes in time series data, leading to biased results. While time series models are useful for modeling data trends and seasonality and identifying outliers that do not correspond to the model, they may fail to identify outliers in complex or non-linear trends. Furthermore, when there are a huge number of random outliers, robust statistical outlier detection methods may fail [26, 27].

New methods like as wavelet analysis, machine learning algorithms, artificial neural networks, and Bayesian inference have been developed to address these challenges. These methods show promise in dealing with non-Gaussian and complicated data, as well as improving outlier detection and removal in GNSS coordinate time series data.

Wavelet analysis is an effective method for spotting outliers at various scales. It can successfully discover outliers embedded in the data trend that other approaches may disregard. Wavelet analysis, on the other hand, necessitates the precise selection of the wavelet basis and decomposition stages. For example, [28] presented a wavelet-based method for evaluating GNSS coordinate time series data, while [29] suggested a wavelet-based method to detect noise components and outliers.

Machine learning techniques provide another method for detecting outliers in GNSS data [30] improved the security and accuracy of GNSS-based intelligent transportation systems by using a supervised machine learning classifier [31] used a mixed approach to machine learning to find problems in GNSS data. They mixed a deep neural network that had already been trained with a genetic algorithm to help with feature extraction and classification. For outlier detection in GPS and accelerometer data [32], used convolutional neural networks and long short-term memory networks.

When data exhibits non-Gaussian or complicated connections, Bayesian inference allows for the incorporation of prior knowledge and provides robustness in outlier detection. As a result, Bayesian approaches are now commonly used [33] developed a Bayesian hierarchical model to find outliers in GPS time series data, which uses a mixture of normal distributions to describe the signal and noise components and defines outliers as those with a low probability of belonging to the signal component [34] employed a Bayesian technique to establish the outlier detection threshold.

Traditional approaches for detecting outliers, such as time series clustering algorithms, have also been proposed. For example [35], tried to improve the k-means clustering method in order to lessen the impact of outliers on grouping [36] developed an approach based on the k-means clustering algorithm to improve positioning accuracy.

While these novel methods have the potential to improve outlier detection in GNSS time series data, their performance must be reviewed and compared to existing methods. Also, more study is needed to discover the best effective methods for detecting outliers in various types of GNSS time series data.

There is no published study that explicitly integrates Bayesian changepoint analysis and Hampel descriptors, according to a review of the literature. Separate works on Bayesian changepoint analysis and Hampel descriptors do exist. For example [37], suggest a new method for detecting outliers in GPS coordinate time series data using Bayes' theorem, whereas [34] uses an autoregressive model-based Bayesian detection method [38] apply a Bayesian strategy for detecting cycle-slips in GNSS carrier phase data. In addition, more investigations [39, 40] have explored Bayesian-based techniques.

The Hampel filter outperforms existing outlier detection approaches in terms of reducing the impact of outliers on system identification. For example [41], investigates the effect of outliers on linear and nonlinear system identification and emphasizes the Hampel filter's usefulness [42] assesses the performance of weighted and recursive Hampel filters [43] provides an innovative method for detecting outliers and reducing noise in a dataset that employs the Hampel filter to detect outliers and the Savitzky-Golay filter to minimize noise in the dataset. Similarly [44], compares the performance of Hampel and median filters in detecting undesirable signals and cleaning data effectively. To improve measurement accuracy [45], do outlier analysis with Hampel and moving average descriptors [46] compare moving average, median, Savitzky-Golay, and Hampel filters, particularly in the context of filtering weak GPS signals.

However, an absence of studies combining Bayesian changepoint analysis and Hampel descriptors underlines the need of considering the potential performance increase from merging both methods.

A new method is proposed in this paper that combines the Hampel descriptor with Bayesian changepoint analysis, with the aim of improving the accuracy and reliability of identifying and removing outliers in GNSS coordinate time series data. This integrated approach aims to overcome the limitations of commonly used approaches while capitalizing on the benefits of Bayesian changepoint analysis and Hampel descriptors.

The proposed method offers various advantages by integrating Bayesian changepoint analysis with Hampel identifiers. Bayesian changepoint analysis can uncover outliers contained within the data's underlying trend and detect unanticipated changes in data distribution. In contrast to established measurements such as standard deviation, Hampel descriptors give a robust statistical measure of data distribution that is less sensitive to outliers. This increases the method's effectiveness in identifying outliers that other techniques may miss, particularly large or clustered outliers.

The suggested method would use the strengths of both Bayesian changepoint analysis and Hampel descriptors to improve outlier detection and removal in low-quality GNSS coordinate time series data. The quality and reliability of data processing can be considerably improved by precisely recognizing and removing outliers, allowing for more accurate calculation of displacement amplitudes and oscillation parameters.

Finally, this paper presents a novel method to identify outliers in GNSS coordinate time series data that integrates Bayesian changepoint analysis and Hampel descriptors. The method tries to overcome the restrictions of standard outlier detection methods by combining these two approaches, providing a more efficient and reliable solution for dealing with outliers in GNSS data. Further research and performance evaluations are required to properly analyze the effectiveness of this integrated approach and its potential applications in other forms of GNSS time series.

2 Materials and methods

BCA is applied to a dataset to determine probable change points. It is predicated on the idea that each change point divides the data into two distinct parts, each with a unique Gaussian distribution. The approach seeks to integrate data-based probabilities with prior knowledge to provide the most current and relevant probability distribution [33]. To accomplish this, a preliminary distribution is constructed and the data is modeled using this distribution. The probability that each data point is a change point is then determined, and the probabilities are utilized to locate potential change point locations in the data set.

The Hampel Identifier is a powerful tool for identifying potential outliers in time series data. It uses the median and median absolute deviation to determine the location and distribution of outliers. The adaptive Hampel descriptor decreases false positive and false negative outcomes and allows for more precise detection of outliers.

2.1 Bayesian change point analysis method

BCA is a statistical method that is used to locate changepoints in a dataset [47]. It presupposes the presence of two or more regions with distinct distributional structures in a given time series data [48] and mathematically specifies two or more potential distribution models.

The basic objective of BCA is to identify probable change points at each time point of the $y(t)$ time series, where $y(t) = y(1), y(2),\ldots, y(T)$. In order to do this, at each time point, a binary variable $S(t)$ is defined to indicate the existence or absence of a change. BCA use Bayes' theorem to determine the value of the variable $S(t)$ at each time step [49, 50].

Bayes' theorem can be stated as follows:

$$\begin{aligned} &P\big(S(t) = 1|y(t)\big) \propto P\big(y(t)|S(t) = 1\big)P(S(t) = 1) \\ &\quad = P\big(y(t) + 1 : T|S(t) = 1\big)P\big(y(1) : t|S\_t = 1\big)P(S(t) = 1) \\ &\quad = N\Big(y(t) + 1 : T|\, \mu_2, \sigma_2^2\Big)N\Big(y(1) : t|\mu 1, \sigma_1^2\Big)P(S(t) = 1) \end{aligned} \tag{1}$$

here $P(y(t) \mid S(t) = 1)$ indicates the probability based on the data, while $P(S(t) = 1)$ represents the prior distribution for the changepoint position. The symbol $\propto$ indicates a direct proportionality between these two variables. The normal probability density function is denoted by the function $N()$. The formula depicts the data points after the changepoint as $y(t)+1{:}T$, and the data points before the changepoint as $y(1){:}t$. To eliminate ambiguity regarding the changepoint, the model includes the variable $S(t)$ and assumes that the data follows a normal distribution. In this model, data samples are assumed to have a normal distribution in the region before the change point and a different normal distribution in the region after the change point. $N(\mu_1, {\sigma_1}^2)$ and $N(\mu_2, {\sigma_2}^2)$ are the definitions of the distributions preceding and following the change point, respectively. Here, the first region's mean and variance values are denoted by μ_1 and ${\sigma_1}^2$, while the second region's mean and variance values are represented by μ_2 and ${\sigma_2}^2$. As a result, Bayes' theorem is utilized to compute unique probabilities for each scenario in which each data point is regarded as change point. In this way, this analysis method allows the determination of optimal change point locations by integrating data-driven probability and prior knowledge.

2.2 General framework of the Hampel Identifier

The Hampel Identifier is a method to determine probable outliers in time series data by estimating their position and distribution using the median and median absolute deviation (MAD) [45]. This description uses the median value to predict the data set's location and MAD to estimate the data's standard deviation. To explain the working principle of the Hampel Identifier, let's consider a time series with n elements: x_1, x_2, x_3, ..., x_n. With a window width of w and a specified number of neighboring elements on both sides, denoted by k, the moving window length becomes $2\,k+1$. The following is how the local median $\mathrm{m_i}$ is defined:

$$m_i = \text{median}\left(x_{i-k}, x_{i-k+1}, x_{i-k+2}, \ldots, x_i, \ldots, x_{i+k-2}, x_{i+k-1}, x_{i+k}\right). \tag{2}$$

Furthermore, the scale estimate Si, which represents the median absolute deviation (MADse) of the median estimates, is calculated as:

$$S_i = k \cdot \text{median}\left(\left|x_{i-k} - m_i\right|, \ldots, \left|x_{i+k} - m_i\right|\right) \tag{3}$$

here $k = \left(\frac{1}{\sqrt{2}\, erfc^{-1}\left(\frac{1}{2}\right)}\right) \approx 1.4826$, represents the unbiased estimate of the Gaussian distribution.

The Hampel Identifier is used to determine whether the potential anomaly spots detected are real anomalies. The Hampel function formula for calculating how much a data point differs from other data points can be expressed as follows:

$$H(x) = \begin{cases} x - m, & \text{if } |x - m| \leq k \cdot s \\ s^2 \cdot \text{sign}(x - m), & \text{otherwise} \end{cases}. \tag{4}$$

In here, x is the analyzed data point, m is the moving average of data points, k is a threshold value, and s is the standard deviation of data points. The Hampel function evaluates the distance of x from m and applies different operations depending on the threshold value. If $|x-m|$ is less than or equal to k * s, it returns the difference between x and m, i.e.,, $(x-m)$. This implies that the examined data point x is close to the moving average m or the difference is within an acceptable range (less than k * s), so it is considered as not significantly deviating from other data points. However, if $|x - m|$ is greater than $k * s$, it returns the sign of the difference multiplied by s^2. This operation indicates that x significantly deviates from the moving average m, and it is considered as an outlier. This function is used as a method to identify the outliers in the data points and correct them.

Firstly, the moving average and standard deviation are calculated using the sample data set. The Hampel Identifier method is then used to define any indicated possible anomalies. The adaptive Hampel Identifier approach is used to investigate the identified probable outliers. The adaptive Hampel descriptor finds the mean of each data point and estimates the median absolute deviation of the data inside that window using a window size. The Hampel descriptor parameters are changed based on the characteristics of the identified outliers.

The parameters of the Hampel descriptor are readjusted at this stage to reduce false positive or false negative findings, and outliers are reanalyzed. This procedure increases the sensitivity of the Hampel descriptor, ensuring in more accurate results.

2.3 Theoretical background of the proposed method for outlier detection

In this work, we present the "Bayesian Change Point Analysis and Hampel Identification" (BCP-HI) method for identifying outliers in GNSS data and reliably distinguishing actual anomalies from noise. This method enables the identification of outliers with the HI method among the potential change points determined by BCP Analysis. This ensures that only true anomalies are detected and false positives due to other change points or noise are treated as outliers. Additionally, BCP-HI has an adaptable structure by performing the detection process iteratively, thus increasing the accuracy of the overall outlier detection process compared to other outlier detection methods.

The proposed BCP-HI algorithm can be summarized in seven steps in three main parts: preliminary assessment of outliers, identification of outliers, and refinement of the detection process (as shown in Fig. 1).

2.3.1 The processing steps of the BCP-HI algorithm

1. *Data loading and preprocessing:* The input data is preprocessed before the outlier detection process begins. Noise and outliers are evaluated for their general character by viewing the data. In this stage, drawing a confidence ellipse and histogram can provide insight into the distribution and structural features of the data points. If any values are missing, the interpolation method is used to fill them in while taking into account the sampling interval and length.
2. *Determining initial parameters:* Considering the preliminary analysis results of the input signal data, such as its general character and the density of outliers, approximate values are predicted for the window size (w) and the number of change points (cp).
3. At this stage, the BCP analysis is carried out with the initial cp value. As a consequence of the analysis, the regions in the input data that differ from each other with various distribution structures are examined, and the most likely change spots in the data are determined.

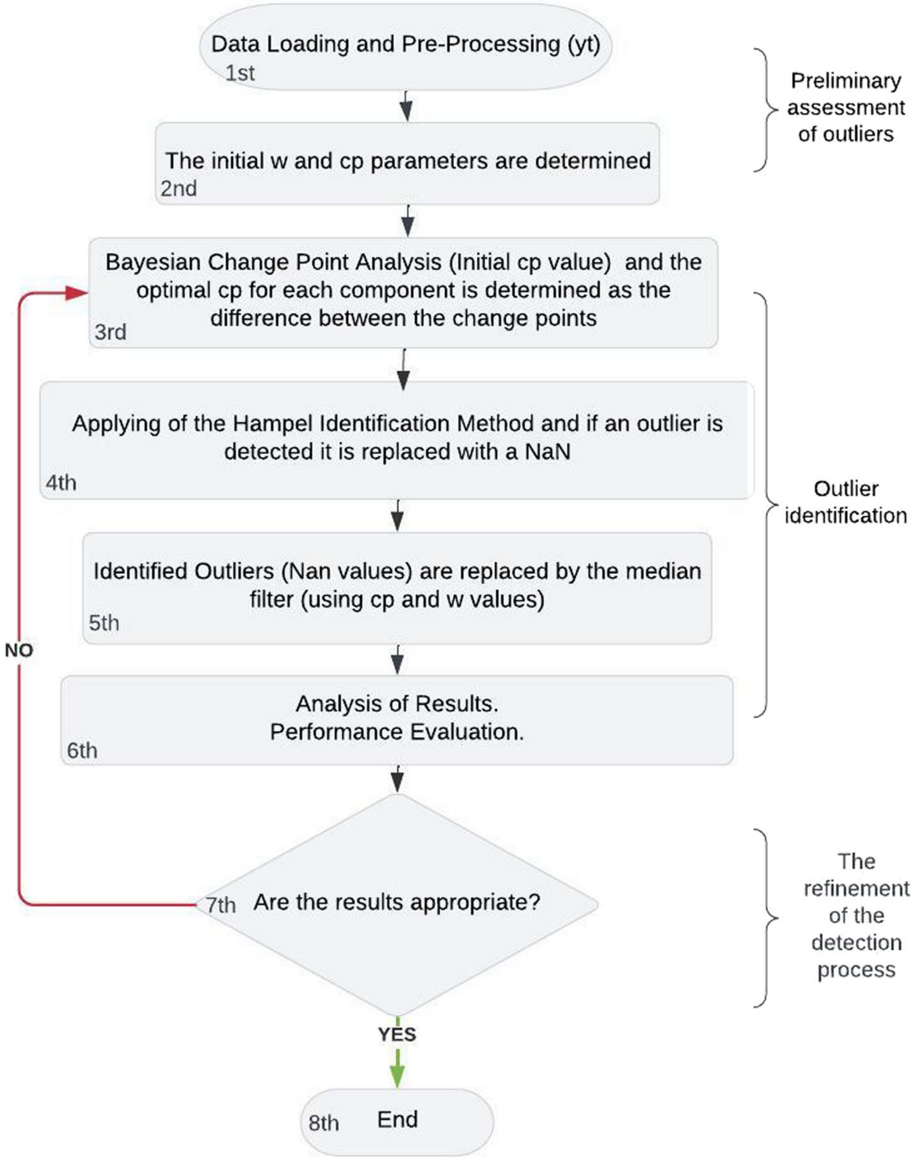
Data Loading and Pre-Processing (yt)
1st
The initial w and cp parameters are determined
2nd
Preliminary assessment of outliers
Bayesian Change Point Analysis (Initial cp value) and the optimal cp for each component is determined as the difference between the change points
3rd
Applying of the Hampel Identification Method and if an outlier is detected it is replaced with a NaN
4th
Identified Outliers (Nan values) are replaced by the median filter (using cp and w values)
5th
Outlier identification
Analysis of Results.
Performance Evaluation.
6th
NO
7th
Are the results appropriate?
The refinement of the detection process
YES
8th
End

Fig. 1 The flow chart of the BCP-HI algorithm

4. *Hampel identifier application:* HI is applied to clusters that are divided into smaller data sets with identified change points. Using the median and median absolute deviation values in each data set, HI takes into account the location and distribution of each data point and identifies potential outliers with the formula (5) given above. The identified outliers are then replaced with 'NaN'.
5. *Replacement of recognized NaN values:* New optimal values are produced instead of each determined NaN value. In this process, a median filter with a window size of *w* is used.

6. *Outlier analysis:* Output data, along with the identified outliers, is visualized, and the results are analyzed, a performance evaluation is made, and the success of the process is examined. The performance evaluation of the output signal can be determined using metrics such as F1 score, SNR, and RMS variations.
7. *Process refinement:* The final step is to improve the process. If the identified outliers are confirmed to represent true anomalies, the parameters of the adaptive HI can be changed to further improve the outlier detection process. The window size of the descriptor can be modified to make it more or less sensitive to changes in the signal, or the outlier detection threshold can be adjusted accordingly.

2.3.2 Outlier identification is carried out in the third, fourth, fifth, and sixth steps

The outlier detection process shown in Fig. 1 was carried out using MATLAB codes (see "Appendix 1"). In the third step, Hampel outlier detection is performed using the change points obtained from the BCP analysis. The process here can be succinctly explained as follows: The data set is divided into sections with change points determined by BCP, and the median and MAD are calculated for each section. Then, in the fourth step, using HI, an outlier check is performed for each data point, and the indices of the identified outliers are added to the 'outlier_indices' vector. See "Appendix Eq. 6" for information on the processing carried out in steps three and four. In the fifth step, the identified outliers are replaced with NaN values (see "Appendix Eq. 7"), and the NaN values in the output signal are filled using a median filter of size '*w*' (see "Appendix Eq. 8").

2.3.3 Performance evaluation of the BCP-HI algorithm and analysis of results are carried out in the seventh step

After separating outlier observations from the input signal, evaluation criteria such as sensitivity (recall), precision, accuracy, and F1 score were used to measure the success of our proposed model (see "Appendix Eqs. 9, 10, 11, 12"). Thus, for each test set, the success of the proposed model in separating outlier data from the input signal and correctly detecting real anomalies in the data was evaluated. Precision: Measures the proportion of correctly identified outliers among all identified outliers. Precision: represents the proportion of true outliers correctly detected among all true outliers. F1 Score: A balanced metric that combines precision and accuracy to evaluate the overall performance of outlier detection (see "Appendix 2"). Additionally, evaluation criteria used to evaluate the performance of the output signal include signal strength (SP), noise power (NP), signal-to-noise ratio (SNR), root mean square (RMS), and percentage outliers (see "Appendix Eqs. 13, 14, 15, 16, 17"). While signal strength and RMS values reflect the general characteristics of the output signal, noise power and SNR values indicate the noise level in the output signal. An accurate calculation of these values can give an idea about the quality of the output signal.

Table 1 Simulated GNSS signals generated for testing

Characteristics of Simulated GNSS data	Sinp1	Sinp2	Sinp3
t (h)	1	1	1
f (Hz)	1	10	100
n	3600	36,000	360,000
Amp. (cm)	1 2 2 1	1 2 2 1	1
SNR (dB)	6	6	6
Outlier Amp.(cm)	10 −10 15	10 −10 15	3 −3 3.5
Outlier index	1000 2000 3000	1000 2000 3000	1000 2000 3000

Table 2 Performance metrics for the BCP-HI model on simulated GNSS signals

Performance metrics of simulated GNSS data	Sig_1 *w/cp: 4/20*		Sig_2 *w/cp: 8/42*		Sig_3 *w/cp: 10/120*	
	Sinp1	Sout1	Sinp2	Sout2	Sinp3	Sout3
f1 score	0.97		0.98		0.98	
Outliers%	4.1%		4.4%		4.3%	
signal power	6	6	11	7	0	0
noise power	6	0	11	6	0	0
SNR (dB)	0.0084	10.8714	4	16.5	0	12.03
RMS	0.024	0.023	0.33	0.3	0.01	0.01

3 Simulated and real GNSS time series analysis

The performance of this method has been tested using simulation and real GNSS time series data. The BCP-HI model was applied to four different simulation datasets provided in Table 1 and the real GNSS time series data provided in Table 3. Each signal with a different sampling frequency, measurement duration, and number of periodic components was subjected to the BCP-HI model, and the accuracy metrics of the resulting output signals were computed and presented in Table 2 and 4.

3.1 Simulated GNSS time series analysis

We can describe a simulated GNSS coordinate time series with randomly mismatched measurements using the following formula:

$$\begin{aligned} x(t) =& A + \Sigma\left(A_1 * \cos\left(2\pi f_1 * t + \varphi_1\right) + A_2 * \cos\left(2\pi f_2 * t + \varphi_2\right)\right. \\ & \left. + \cdots + A_n * \cos\left(2\pi f_n * t + \varphi_n\right)\right) + \varepsilon(t). \end{aligned} \tag{5}$$

In this equation, $x\ (t)$ represents the coordinate value in the time series. A denotes the mean value. A_1, A_2, ..., A_n are the amplitude values of harmonic components. f_1, f_2, ..., f_n are the frequencies of harmonic components. ϕ_1, ϕ_2, ..., ϕ_n are the phase angles of harmonic components. $\varepsilon(t)$ represents the effect of randomly mismatched measurements (noise term).

In this study, simulation data was generated and used by considering harmonic components and noise as the main factors while neglecting the others. Using Eq. (5) and the parameters detailed in Table 1, three data sets (Sinp1, Sinp2, and Sinp3) are generated.

Periodic Movements: These are created using a combination of different amplitudes and periods (ex: 2π, $6\pi/5$, $2\pi/5$, $\pi/5$). Noise: The noise term added to the signal is initially calculated based on a default SNR value and is determined according to the signal power. The noise power is calculated using the variance of the seasonal components of the signal, and random values following a normal distribution with a calculated standard deviation are used to create white noise. Outliers: Outliers are added to the signal by introducing specific amplitude values at certain indices.

Test signals with known parameters, including size, sampling frequency, signal-to-noise ratio, and indices where outliers were added, were generated using the parameters provided in Table 1. Sinp1 simulation data, with a sampling frequency of 1 Hz, 3600 data points, and 6 dB signal-to-noise ratio with added Gaussian white noise, is shown in Fig. 2. Sinp2 simulation data, with a sampling frequency of 10 Hz, 36,000 data points, and 6 dB signal-to-noise ratio with added Gaussian white noise, is shown in Fig. 4. Lastly, Sinp3 simulation data, with a sampling frequency of 100 Hz, 360,000 data points, and 6 dB signal-to-noise ratio with added Gaussian white noise, is shown in Fig. 6.

The three sets of simulated input signals (Sinp1, Sinp2, and Sinp3) were processed for outlier detection and removal using the BCP-HI algorithm, as shown in the process flow diagram in Fig. 1, and the output signals (Sout1, Sout2, and Sout3) were generated.

The BCP-HI algorithm was applied to the Sinp1 signal with parameters $w = 4$ and cp = 20, and outliers were detected. Figure 2 shows the Sinp1 signal and outliers, which are marked by red circles. Figure 3 presents a visual comparison evaluation of the processed Sinp1 signal. Here, we can see the original signal with the red-circled outliers in blue, the cleaned signal with outliers replaced with NaN values in green, and the corrected signal (Sout1) in black after replacing the NaN values with the medians of their nearby values.

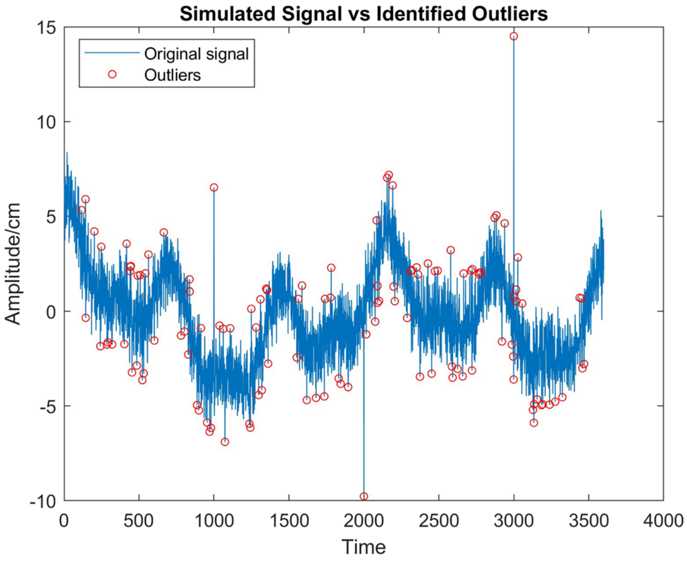

Fig. 2 Sinp1 signal with BCP-HI applied. $w = 4$ and cp = 20

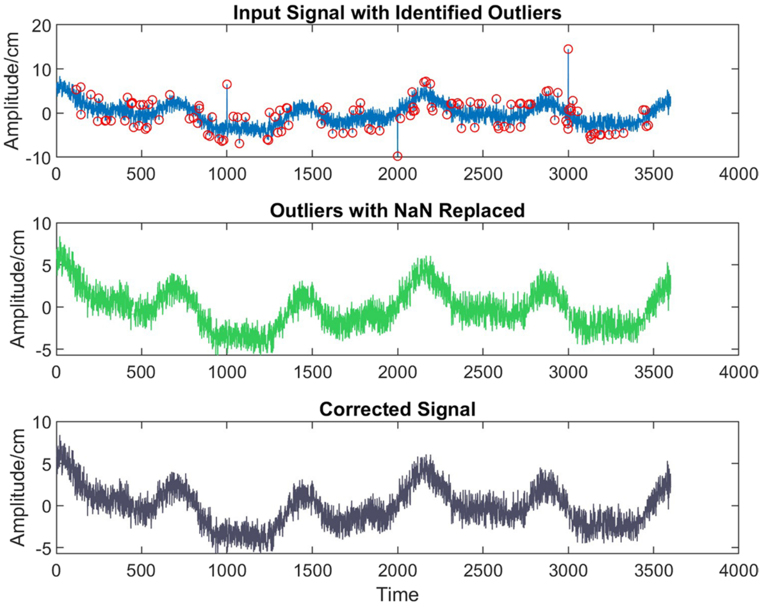

Fig. 3 Simulation data (Sinp1) whose outliers were detected and removed with the BCP_HI method. **a** Outliers identified with $w = 4$ and $\text{cp} = 20$; **b** The case where the identified outlier values were replaced with NaN; **c** The case where NaN values are filtered and replaced with the closest values

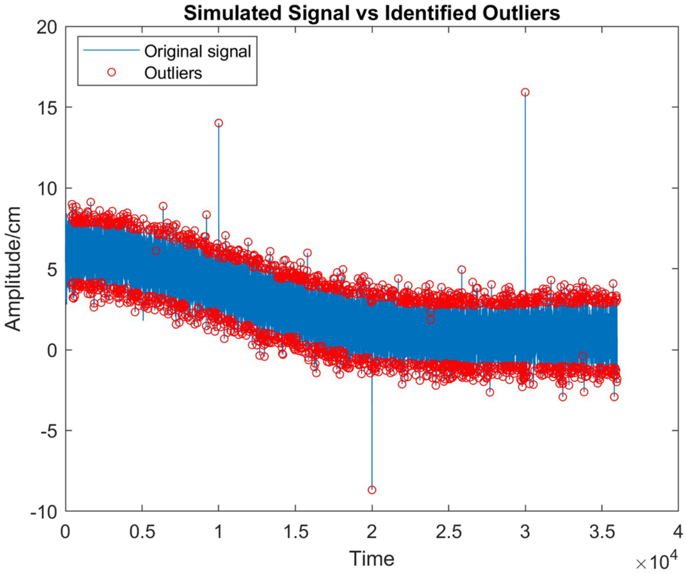

Fig. 4 Sinp1 signal with BCP-HI applied. $w = 8$, $\text{cp} = 42$

Outliers have been identified using the BCP-HI method in the Sinp2 signal with the parameters $w = 8$ and $\text{cp} = 42$. Figure 4 shows the identified outliers, which are marked by red circles, along with the Sinp2 signal. The processed Sinp2 data are compared and evaluated visually in Fig. 5. Here, the original signal with the red-circled

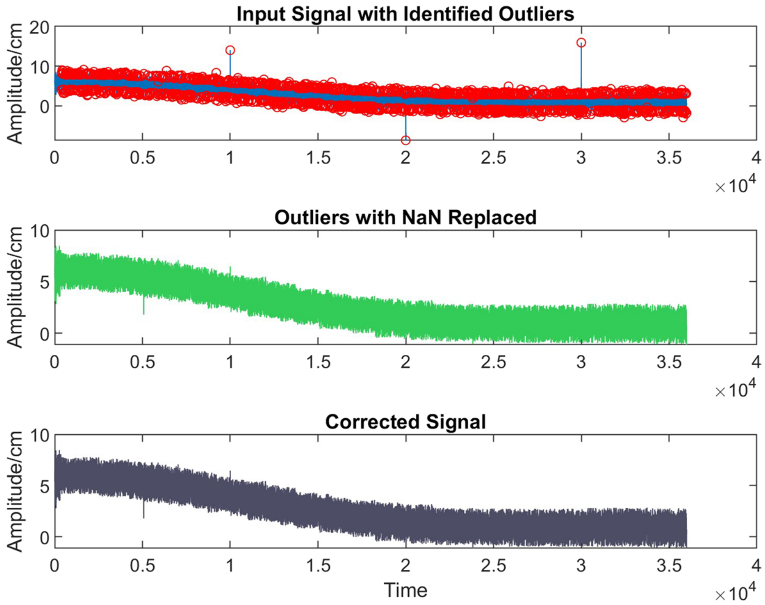

Fig. 5 Simulation data (Sinp2) whose outliers were detected and removed with the BCP_HI method. **a** Outliers identified with $w=8$ and cp = 42; **b** The case where the identified outlier values were replaced with NaN; **c** The case where NaN values are filtered and replaced with the closest values

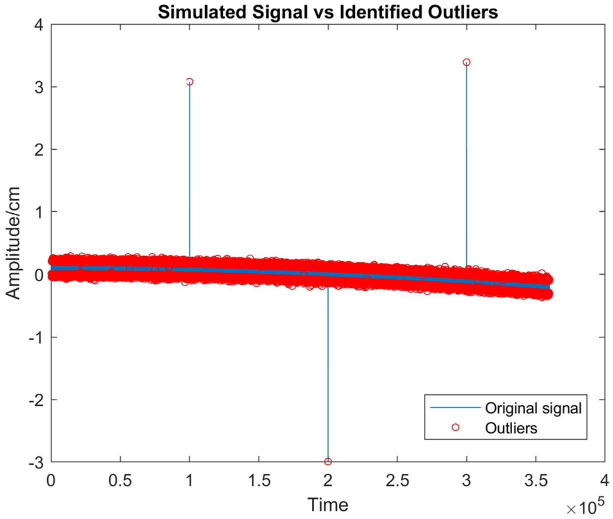

Fig. 6 Sinp3 signal with BCP-HI applied. $w=10$, cp = 120

outliers is shown in blue, and the cleaned signal with the NaN values replaced for the outliers is shown in green, and the corrected signal (Sout2) is shown in black after the NaN values have been replaced with the medians of their near values.

The BCP-HI method was applied to identify outliers in the Sinp3 signal with parameter values of $w=10$ and cp = 120. Figure 6 shows the identified outliers, which are marked by red circles, along with the Sinp3 signal. The processed Sinp3 data are

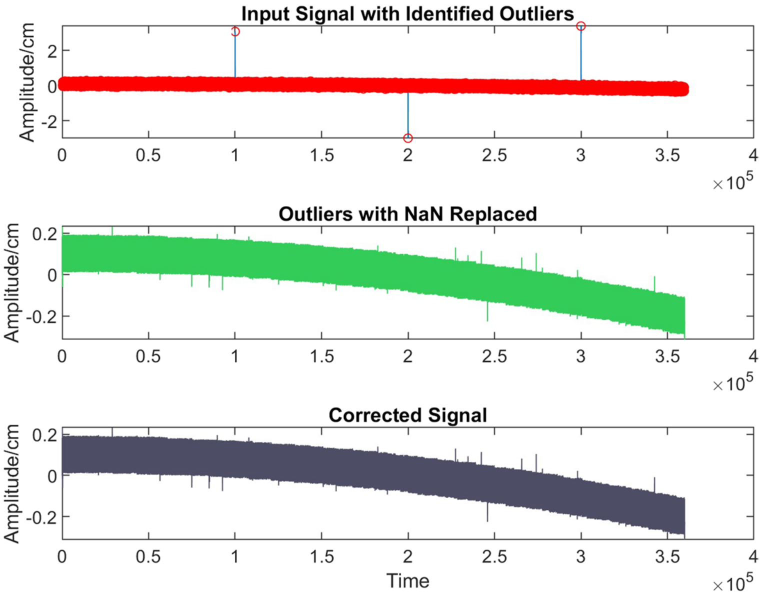

Fig. 7 Simulation data (Sinp3) whose outliers were detected and removed with the BCP_HI method. **a** Outliers identified with $w = 10$ and cp = 120; **b** The case where the identified outlier values were replaced with NaN; **c** The case where NaN values are filtered and replaced with the closest values

com-pared and evaluated visually in Fig. 7. Here, the original signal with the red-circled outliers is shown in blue, the cleaned signal with the NaN values replaced for the out-liers is shown in green, and the corrected signal (Sout3) is shown in black after the NaN values have been replaced with the medians of their near values.

To achieve the best results using the BCP-HI algorithm, it is essential to determine the optimal values for w and cp. These values can be identified through trial and error or through an iterative approach that finds the parameter values where the model exhibits the best performance. For Sinp1, the optimal w value is found to be 4, and the optimal cp value is 20. These values indicate that the model provides the best results and accurately detects outlier values. Similarly, the ideal w and cp values were determined to be 8, 42, and 10, 120 for Sinp2 and Sinp3, respectively.

Table 2 shows the evaluation metrics used to evaluate the signals that were subjected to the BCP-HI method.

The F1-Score value for the input signal Sinp1 being close to 1 show that the model successfully strikes an appropriate balance between sensitivity and precision. According to the overall structure of the input signal and the properties of the outlier values, such as their amplitude and distribution, the captured outlier percentage was determined to be 4.1%. To preserve the overall characteristics of the data while removing the outliers, the algorithm is run iteratively, and the outlier values are consistently determined based on the same criteria.

The SNR value, which was 0.0084 dB for the input signal, has increased to 10.8714 dB for the output signal. The SNR has risen as a result of this increase, which shows that the decrease in signal power compared to noise is higher. The RMS value, which was 0.024 for the input signal, has decreased to 0.023 for the output signal, indicating a re-duction in fluctuations and a smoother signal.

As can be seen from the output signal's higher SNR and lower RMS values compared to the input signal, the BCP-HI method successfully removes outliers from the input signal and improves signal quality.

Similar results can be observed when looking at the performance quantified for the other simulated input signals, Sinp2 and Sinp3. All performance parameters show measurably improved output signals after processing. Figures 3, 5 and 7 show the output signals with the outlier values removed.

3.2 Real GNSS time series analysis

The BCP-HI algorithm has been evaluated using real GNSS data (*X* and *Y* coordinate time series) that included obvious outliers. 7200 location data points were collected on a building during a period of two hours using a Trimble MB-2 OEM GNSS receiver. The sampling frequency was set to 1 Hz, and the height mask angle was set to 15 degrees (Figs. 8, 9, 10). The time length, sampling frequency, number of samples, and SNR of the real GNSS signals (Rinp_x and Rinp_y) used in this study are presented in Table 3.

When the BCP-HI algorithm was applied to the real GNSS time series Rinp_x for outlier detection and removal with parameters $w = 4$ and $\mathrm{cp} = 52$, the output signal (Rout_x) was computed, and outlier values were identified (Fig. 8).

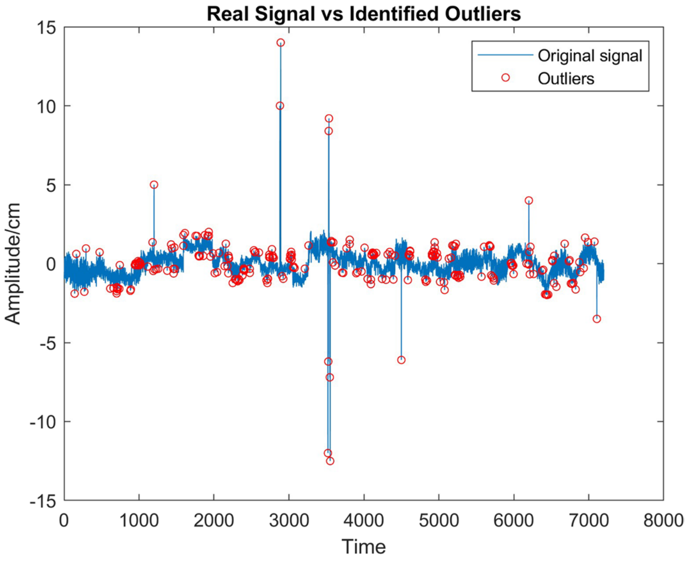

Fig. 8 The Rinp_x data after BCP-HI processing. $w = 4$ and $\mathrm{cp} = 52$

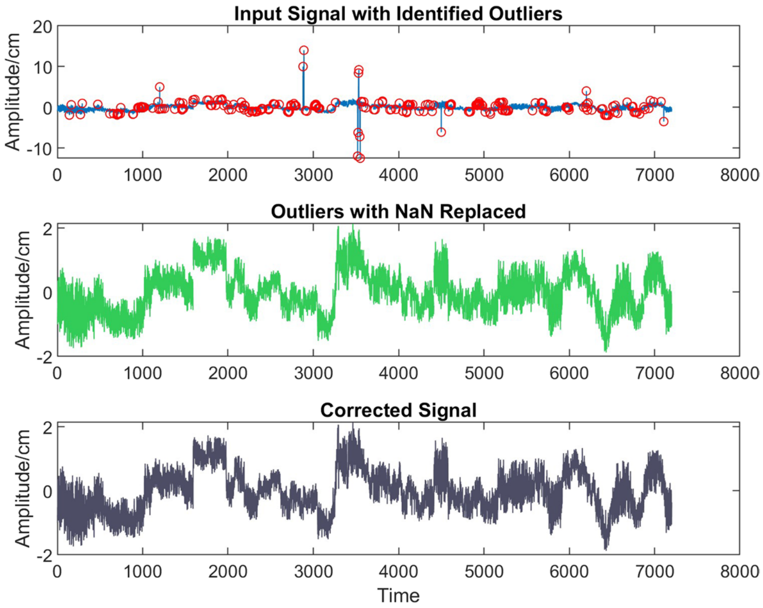

Fig. 9 The real GNSS data (Rinp_x) whose outliers were detected and removed with the BCP_HI method. **a** Outliers identified with $w=4$ and cp = 52; **b** The case where the identified outlier values were replaced with NaN; **c** The case where NaN values are filtered and replaced with the closest values

The identified outliers were assigned NaN values (Fig. 9b), and these NaN values were filled with the median of the nearest values, resulting in the corrected signal (Rout_x) (Fig. 9c).

Similar to this, when the BCP-HI algorithm was applied to the real GNSS time series Rinp_y for outlier detection and removal with parameters $w=4$ and cp = 50, the output signal (Rout_y) was computed, and outlier values were identified (Fig. 10). The identified outliers were assigned NaN values (Fig. 10b), and these NaN values were filled with the median of the nearest values, resulting in the corrected signal (Rout_y) (Fig. 10c).

The optimal values for Rinp_x have been determined to be $w=4$ and cp = 52. These values indicate that the model provided the best results and accurately detected the outlier values. Similarly, for Rinp_y, the optimal values were determined to be $w=4$ and cp = 52. Similar results have been obtained for Rinp_y, where the best values have been determined to be $w=4$ and cp = 52.

The evaluation metrics for the signals processed by the BCP-HI algorithm are presented in Table 4.

When looking at the post-processing performance values shown in Table 4, the F1-Score values for the Sig_x and Sig_y signals are close to 1, which indicates that the BCP-HI model achieves an appropriate balance between sensitivity and precision.

Based on the typical characteristics, magnitude, and distribution of the outlier values, it has been found that the ratio of identified outlier values in the input signals varies from 3.15 to 4.30%. Without changing the basic elements of the data, the selected values for w and cp may effectively identify the most extreme outlier values as well. These percentages can be improved by iteratively adjusting the values of w and cp for the outlier values that undefined (Fig. 11).

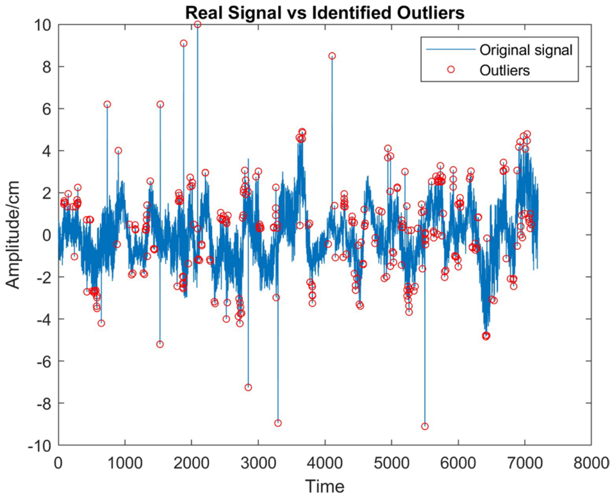

Fig. 10 The Rinp_y data after BCP-HI processing. $w = 4$ and cp $= 50$

Table 3 Real GNSS signals recorded for testing

Performance metrics of real GNSS data	Rinp_x	Rinp_y
t (h)	2	2
f (Hz)	1	1
Samp	7200	7200
SNR (dB)	6	6

While there was a significant increase in the SNR values of the processed x and y signals, there was a visible decrease in the RMS values. The SNR value for Rinp_x was 0.0003 dB, indicating that the noise level was comparable to or slightly higher than the signal strength, while the SNR value for Rout_x increased to 4.4082 dB, indicating that the signal strength was much stronger than the noise. The decrease in the RMS value for Rout_x from 0.0076 to 0.0066 indicates a decrease in the amplitude of the signal after the removal and correction of outliers. The difference between these values is relatively small, indicating that the output signal is still similar to the input signal in terms of overall magnitude. Decreasing RMS values in the output signals means reducing fluctuations and obtaining a smoother signal. In Table 4, it is seen that there are similar improvements in the tests performed with different w and cp values for x and y signals. According to these results, the BCP-HI algorithm proves that it effectively separates the outliers in the input signal, with the output signal having higher SNR and lower RMS values compared to the input signal, thus improving the signal quality.

Table 4 shows that applying BCP-HI with different parameter values to the Sig_x and Sig_y data does not significantly change the findings when comparing the results with different w and cp values. When comparing the results for different w and cp values, it can be observed from Table 4 that applying BCP-HI with different parameter values to

Table 4 Evaluation metrics for real GNSS data

Performance metrics of real GNSS data	Sig_x *w/cp:4/52*		Sig_x *w/cp: 8/100*		Sig_y *w/cp: 4/50*		Sig_y *w/cp: 8/100*	
	Rinpx	Routx	Rinpx	Routx	Rinpy	Routy	Rinpy	Routy
f1 score	0.98		0.98		0.98		0.97	
Outliers%	3.54%		3.15		4.30%		4.18%	
SNR (dB)	0.0003	4.4082	0.0003	4.5246	0.0002	8.9498	0.0002	9.2183
RMS	0.0076	0.0066	0.0076	0.0066	0.0134	0.0126	0.0134	0.0127

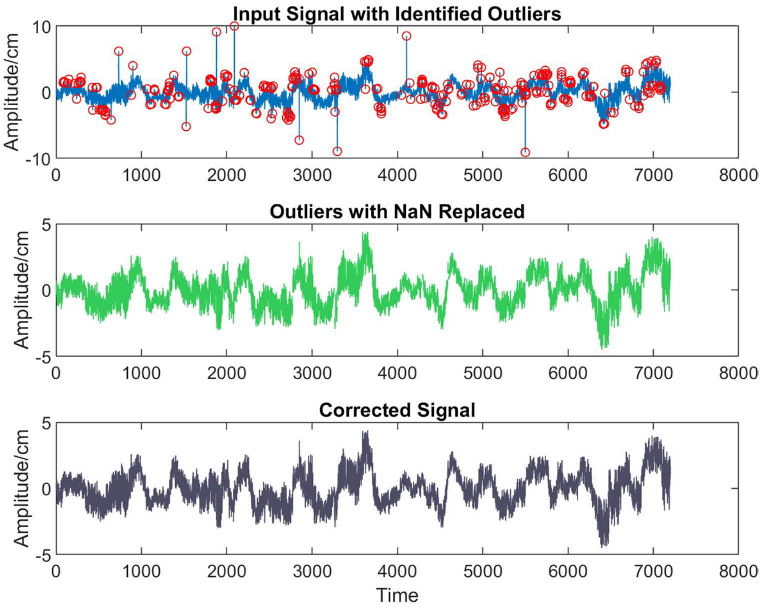

Fig. 11 The real GNSS data (Rinp_y) whose outliers were detected and removed with the BCP_HI method. **a** Outliers identified with *w* = 4 and cp = 50; **b** The case where the identified outlier values were replaced with NaN; **c** The case where NaN values are filtered and replaced with the closest values

Sig_x and Sig_y data does not result in significant changes in the outcomes. When the cp value for Sig_x is increased from 52 to 100, the SNR values change from 4.4082 to 4.5246, respectively, but the RMS value does not change much. The algorithm is able to identify more variations and outliers thanks to the increase in cp from 52 to 100, which leads to a small increase in signal power and SNR. However, it is observed that the changes in these values are relatively small, indicating that the effect of changing the cp parameter on the output signal is limited beyond a certain threshold. This suggests that the most favorable cp value can be adaptively determined based on the signal structure, noise level, and distribution of outliers.

3.3 Results analysis using the Kolmogorov–Smirnov test and Lomb–Scargle periodograms

Using the BCP_HI algorithm are obtained the signal in which the outliers are replaced with NaN and then the corrected signal in which these NaN values are filled with the median of the nearest neighbor elements from the input signal. After applying BCP-HI on real GNSS data, we calculated Lomb–Scargle periodograms to identify the significant frequencies of the input and output signals, compare them, and evaluate the power spectrum distributions of the signals.

Figures 12 and 13 show the power spectral density distributions of the input (Rinp_x and Rinp_y) and post-processing output signals, respectively. These periodograms graphically reveal the changes in the frequency components and power distributions of the signals after BCP-HI processing.

When we examine these periodograms, there appears to be a generally significant flattening in the frequency components of the signals processed with the BCP_HI algorithm. By removing the outliers, peaks that were not evident in the input signal have been emerged more clearly. This means that the signal is cleared of outliers that increase its amplitude.

We employed the Kolmogorov–Smirnov (KS) test to statistically examine the similarity of the power spectral density distributions of the input and output signals in addition to examining it visually. We reasoned that the KS test would assist us in making a more exact assessment of the frequency component similarity or difference between these signals. As an example, the KS test was applied to evaluate the sample distribution, similarities or differences of the real GNSS signal Sig_x presented in Fig. 12 after BCP-HI processing, and the results are shown in Table 5.

Based on the KS test results given in Table 5, we can make the following comments:

1. *Input signal versus NaN-replaced signal* There is a statistically significant difference ($h = 1$). The p value is very low ($p = 0.000001$), which means the final result was obtained with a p value much lower than the threshold of 0.05 (the commonly accepted significance level). This supports the conclusion that there is a statistically significant difference between the original signal and the NaN-modified outlier signal.

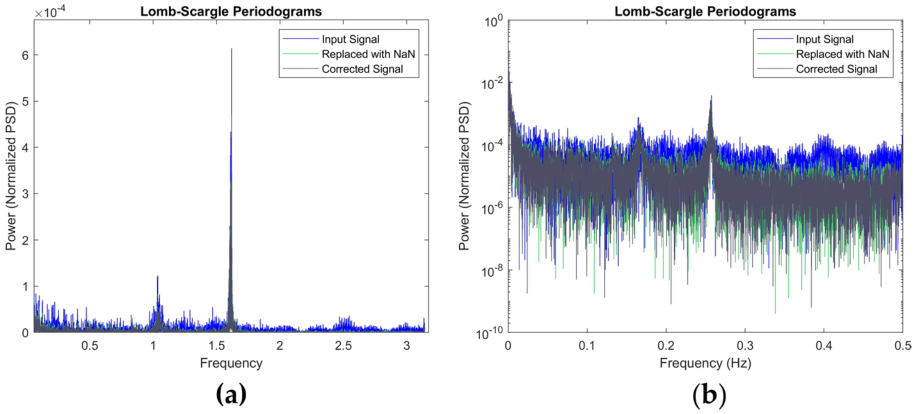

Fig. 12 **a** The normalized spectral power densities of the real GNSS signal Rinp_x, the signal replaced with NaN, and the corrected signal; **b** Different scale views of the same signals

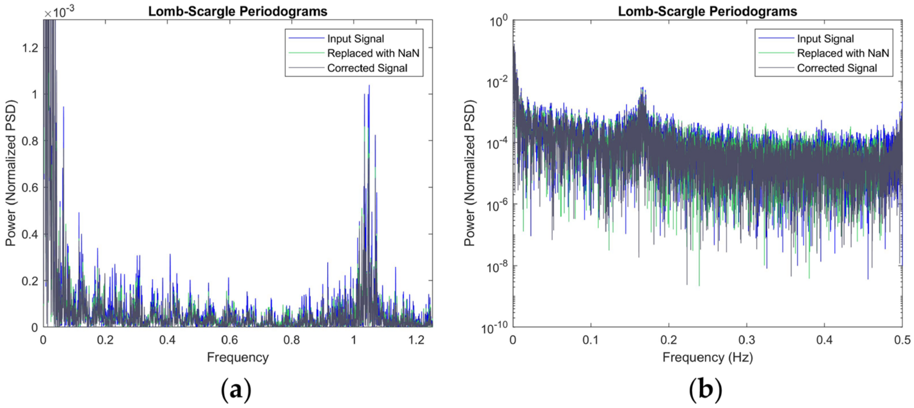

Fig. 13 **a** The normalized spectral power densities of the real GNSS signal Rinp_y, the signal replaced with NaN, and the corrected signal; **b** Different scale views of the same signals

Table 5 Rinp_x için Kolmogorov–Smirnov test results

KS test results	Sig_x *w/cp:4/52*		Sig_y *w/cp:4/50*	
	h	*p*		
Inp. versus NaN-modified	**1**	**0.000001**	**1**	**0**
Inp. versus corrected	0	1	0	1
NaN-modified versus corrected	1	0.000001	1	0

2. *Input signal versus corrected signal* There is no statistically significant difference ($h=0$). The p value is 1, indicating that the final result was obtained with a very high p value, and thus, there is no statistically significant difference between the input signal and the corrected signal.
3. *NaN-replaced signal versus corrected signal* There is a statistically significant difference ($h=1$). The p value is very low ($p=0.000001$), supporting the conclusion that there is a statistically significant difference between the NaN-replaced signal and the corrected signal. In summary, the NaN-replaced signal is statistically different from the original signal. The corrected signal is not statistically different from the original signal. The NaN-replaced signal is statistically different from the corrected signal. These results indicate that there are statistically significant changes in the data during the processing steps of NaN removal and correction.

4 Conclusion and evaluations

The effectiveness of the BCP_HI method we recommend for detecting and removing outliers in GNSS coordinate time series data is demonstrated in this paper. While GNSS data were used to test the method's effectiveness, theoretically, it can be applied to any time series data. The results presented in the graphs and tables above demonstrate that the algorithm achieved high sensitivity and precision in tests conducted on both simulated and real GNSS data.

4.1 Evaluation of simulated GNSS time series

Using the parameters presented in Table 1, Sig_1 (sampled at 1 Hz with 3600 data points), Sig_2 (sampled at 10 Hz with 36,000 data points), and Sig_3 (sampled at 100 Hz with 360,000 data points) were generated. The success of the BCP-HI algorithm was evaluated based on performance metrics for these three data sets as shown in Table 2. Here, f1-score values close to 1 indicate that the model achieves a suitable balance between sensitivity and precision. Outlier values were detected with an approximately 98% success rate in all sets. The percentages of detected outlier values, based on the general characteristics of the three input signals and properties of the outliers such as magnitude and distribution, were determined to be approximately 4.1%, 4.4%, and 4.3%, respectively. When examining the performance metrics measured in the output signals after processing for all three input signals, similar results were observed. For instance, in the case of Sig_1, the SNR value, which was 0.0084 dB for the input signal, increased to 10.8714 dB for the output signal. This increase in SNR indicates that the decrease in signal power is more significant relative to the noise. The RMS value, which was 24 mm for the input signal, decreased to 23 mm for the output signal, indicating a reduction in fluctuations and smoother output signal.

In summary, the BCP-HI method effectively eliminates outlier values from the input signal, as indicated by the higher SNR and lower RMS values in the output signal compared to the initial signal, as demonstrated in Figs. 3, 5 and 7. These results highlight that the BCP-HI model achieves a positive balance between sensitivity and precision, successfully identifying and eliminating outlier values without distorting the original structure of the input signal.

It's important to note here is that the '*w*' and 'cp' values were set to (4/20) for Sig_1, (8/42) for Sig_2, and (10/120) for Sig_3. It has been demonstrated through comparisons with different "w" and "cp" values that the selection of these parameters may affect the output signal's quality. Therefore, it is crucial to take the objectives of the research and the features of the data into account while determining the most appropriate "w" and "cp" values.

4.1.1 Evaluation of Real GNSS Time Series:

As presented in Table 3, performance metrics calculated with different '*w*' and 'cp' values for real GNSS data (Rinp_x and Rinp_y) with a sampling frequency of 1 Hz and 7200 data points are shown in Table 4. According to these results, the f1-score values close to 1 for both X and Y coordinate time series indicate that the model achieves a suitable balance between sensitivity and precision. In both data sets, outlier values could be identified with a 98% success rate.

The X signal was processed with '*w*' and 'cp' values of 4/52 and 8/100, respectively, while the Y signal was processed with '*w*' and 'cp' values of 4/50 and 8/100, respectively. In other words, both parameters were increased by approximately two-fold. Comparisons were made on the results of X and Y signals processed with these different parameters. In this case, the percentage of detected outlier values decreased from 3.54% to 3.15% for X and from 4.30% to 4.18% for Y.

When ‘*w*’ and 52 were used for the *X* input signal, the SNR value increased from 0.0003 dB to 4.4082 dB for the output signal, and when 8 and 100 were used, it increased from 0.0003 dB to 4.5246 dB. For *X* input signal with values of 4 and 50, the SNR value increased from 0.0002 dB to 8.9498 dB for the output signal, and with values of 8 and 100, it increased from 0.0002 dB to 9.2183 dB.

When comparing RMS values, there was no change for the *X* signal, while for the *Y* signal, the RMS value, initially at 13.4 mm, was calculated as 12.6 mm with parameters 4/50 and 12.7 mm with parameters 8/100.

According to these results, using the BCP-HI method to analyze real GNSS data (X, Y) using various ‘*w*’ and ‘cp’ values result in apparent but largely consistent changes in the results. As an example, increasing the ‘cp’ value from 52 to 100 resulted in only slight changes in RMS values while increasing SNR from 4.4082 to 4.5246. This increase in ‘cp’ allowed for the detection of slightly more outlier values, resulting in slightly higher signal power and SNR. These results demonstrate that changing ‘*w*’ and ‘cp’ values can impact signal power, SNR, and RMS, with larger window sizes and higher ‘cp’ values potentially leading to varying outcomes depending on input signal characteristics and outlier structure.

According to the findings, increasing the ‘cp’ value positively influenced output signal quality. This was supported by improvements in signal power, SNR, and RMS values as ‘cp’ increased from 5 to 40. Increased signal power indicates a stronger signal component in the output signal, higher SNR implies improved distinguishability of the desired signal from background noise, and a higher RMS value signifies greater overall amplitude in the output signal. However, it’s essential to recognize that the relationship between ‘cp’ and output signal quality may not be linear and could vary depending on input signal characteristics, the chosen outlier removal and signal smoothing algorithm, and the specific application context. In conclusion, running the BCP-HI algorithm with carefully selected ‘*w*’ and ‘cp’ values can lead to optimal results. If improvements in output signal quality are deemed insufficient after a certain number of iterations, running the process multiple times can yield enhanced results and capture the most logical and suitable outlier values.

Additionally, output signals after BCP-HI processing were visually examined using Lomb–Scargle periodograms, revealing clearer power spectrum distributions. To assess the similarity in power spectrum distribution between the input and output signals, the Kolmogorov–Smirnov test was conducted. The results presented in Table 5 show that the NaN-modified outlier signal was statistically different from the original signal, while the corrected signal was not statistically different from the original signal. These results highlight that the BCP-HI algorithm introduced statistically significant changes in the output signals. The BCP-HI algorithm effectively eliminates outliers, while increasing processing accuracy and reliability and improving signal quality. Testing the BCP-HI algorithm on larger and diverse time series data will contribute to algorithm enhancement and accurate data analysis.

Appendix 1: https://github.com/HuseyinP/Outlier/commit/ac8eafe4a7a76289ee619ee621513807ea800600

```
% function find_outliers                                                        (6)
function outlier_indices = find_outliers(x, cp, w)
    cp_detected = findchangepts(x, 'MaxNumChanges', cp, 'Statistic', 'mean');
    outlier_indices = [];
    for k = 1:length(cp_detected)-1
        segment = x(cp_detected(k)+1:cp_detected(k+1));
        median_segment = median(segment);
        mad_segment = median(abs(segment - median_segment));
        for m = 1:length(segment)
            if abs(segment(m) - median_segment) > 3 * mad_segment
                outlier_indices = [outlier_indices, cp_detected(k) + m];
             end
        end
    end
end
```

```
% replace_outliers_with_nan                                                     (7)
function nan_data = replace_outliers_with_nan(x, outlier_indices)
    nan_data = x;
    nan_data(outlier_indices) = NaN;
end
```

```
% function apply_median_filter                                                  (8)
function sm_data = apply_median_filter(nan_data, w)
    sm_data = nan_data;
    nan_indices = isnan(nan_data);
    nan_locations = find(nan_indices);
    for k = 1:length(nan_locations)
        nan_index = nan_locations(k);
        window_start = max(floor(nan_index - w/2), 1);
        window_end = min(ceil(nan_index + w/2), length(nan_data));
        window_data = nan_data(window_start:window_end);
        median_value = median(window_data, 'omitnan');
        sm_data(nan_index) = median_value;
    end
end
```

Appendix 2

The model's performance is evaluated by calculating the performance metrics, which involve computing the differences between the true class and the predicted class. In the first step, the differences between the true class (input signal) and the predicted class (output signal) are calculated as: diff = abs(true_class − predicted_class); Here, the true class (input signal), predicted class (output signal), and their differences are determined. Using these differences, the performance metrics are calculated as follows:

TP = sum(diff = = 0); (True Positives: when the true and predicted classes are the same).

FP = sum(diff = = 1); (False Positives: when the true class is 0 and predicted class is 1).

TN = sum(diff = = 0); (True Negatives: the true and predicted classes are the same).

FN = sum(diff = = 1); (False Negatives: the true class is 1and predicted class is 0)

$$\text{sensitivity} = \text{TP}/(\text{TP} + \text{FN}); \tag{9}$$

$$\text{precision} = \text{TP}/(\text{TP} + \text{FP}); \tag{10}$$

$$\text{accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}); \tag{11}$$

$$\text{f1\_score} = 2 * \left(\text{precision} * \text{sensitivity}\right)/\left(\text{precision} + \text{sensitivity}\right). \tag{12}$$

Appendix 3

Signal Power: The Signal Power represents how strong the output signal is overall. It is calculated as the mean of the squares of the output data.

$$\text{signal\_power} = \text{mean}\left(\text{out\_data}^2\right). \tag{13}$$

Noise Power: The Noise Power indicates the level of noise in the output signal. It is calculated as the mean of the squares of the differences between the output data and the input data.

$$\text{noise} = \text{out\_data} - \text{inp\_data};$$

$$\text{noise\_power} = \text{mean}(\text{noise}^2) \tag{14}$$

SNR: The Signal-to-Noise Ratio (SNR) represents the ratio between the Signal Power and the Noise Power in the output signal. SNR can be calculated as the logarithm (in dB) of the ratio between the Signal Power and the Noise Power.

$$\text{snr\_db} = 10 * \log 10\left(\text{signal\_power}/\text{noise\_power}\right). \tag{15}$$

RMS: The RMS value indicates how much overall fluctuation the output signal has. It is calculated as the square root of the mean of the squares of the output data.

$$\text{rms\_value} = \text{sqrt}\left(\text{mean}\left(\text{out\_data}^2\right)\right) \tag{16}$$

Outlier Percentage: The improvement rate (*I*) caused by the applied algorithm on the input signal is defined as the percentage of non-outlier values in the total data.

The Outlier Percentage with selected w and cp is calculated as follows:

$$\text{Outliers}\,\% = \frac{N - N_{\text{outliers}}}{N_{\text{total data}}}. \tag{17}$$

Author contribution

This work was prepared by the corresponding author and approved the submitted version.

Funding

This research received no external funding.

Availability of data and materials

The data presented in this study are available on request from the corresponding author.

Declarations

Ethics approval and consent to participate

Not applicable.

Received: 30 September 2023 Accepted: 7 December 2023
Published online: 02 April 2024

References

1. M. Kim, J. Seo, J. Lee, A comprehensive method for GNSS data quality determination to improve ionospheric data analysis. Sensors (Switzerland) (2014). https://doi.org/10.3390/s140814971
2. A. Klos, J. Bogusz, M. Figurski, W. Kosek, On the handling of outliers in the GNSS time series by means of the noise and probability analysis. Int. Assoc. Geod. Sympos. (2016). https://doi.org/10.1007/1345_2015_78
3. S. Hekimoglu, R.C. Erenoglu, D.U. Sanli, B. Erdogan, Detecting configuration weaknesses in geodetic networks. Surv. Rev. **43**(323), 713–730 (2011). https://doi.org/10.1179/003962611X13117748892632
4. S. Hekimoglu, B. Erdogan, Application of median-equation approach for outlier detection in geodetic networks. Boletim de Ciências Geodésicas (2013). https://doi.org/10.1590/s1982-21702013000400002
5. D. Wu, H. Yan, Y. Shen, TSAnalyzer, a GNSS time series analysis software. GPS Solut. **21**, 1389–1394 (2017)
6. X. He, J.P. Montillet, R. Fernandes, M. Bos, K. Yu, X. Hua, W. Jiang, Review of current GPS methodologies for producing accurate time series and their error sources. J. Geodyn. (2017). https://doi.org/10.1016/j.jog.2017.01.004
7. M. Yetkin, Application of robust estimation in geodesy using the harmony search algorithm. J. Spatial Sci. (2018). https://doi.org/10.1080/14498596.2017.1341856
8. G. Blewitt, C. Kreemer, W.C. Hammond, J. Gazeaux, MIDAS robust trend estimator for accurate GPS station velocities without step detection. J. Geophys. Res. Solid Earth **123**(5), 3680–3697 (2018)
9. A. Blázquez-García, A. Conde, U. Mori, J.A. Lozano, A review on outlier/anomaly detection in time series data. ACM Comput. Surv. (2021). https://doi.org/10.1145/3444690
10. F. Zhang, Y. Wang, Y. Gao, A novel method of fault detection and identification in a tightly coupled, ins/gnss-integrated system. Sensors (2021). https://doi.org/10.3390/s21092922
11. R. Karim, M.A.I. Rizvi, M.S. Arefin, A survey on anomaly detection strategies. Lect. Notes Netw. Syst. LNNS (2022). https://doi.org/10.1007/978-3-030-84760-9_25
12. X. Peiliang, Statistical criteria for robust methods. *ITC J.* 1989–1 (1989)
13. Z. Niu, S. Shi, J. Sun, X. He, A survey of outlier detection methodologies and their applications. Lect. Notes Comput. Sci. Includ. Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinform. (2011). https://doi.org/10.1007/978-3-642-23881-9_50
14. M.S. Bos, R.M.S. Fernandes, S.D.P. Williams, L. Bastos, Fast error analysis of continuous GNSS observations with missing data. J. Geod. (2013). https://doi.org/10.1007/s00190-012-0605-0
15. S. Xu, B. Lu, N. Bell, M. Nixon, Outlier detection in dynamic systems with multiple operating points and application to improve industrial flare monitoring. Processes (2017). https://doi.org/10.3390/pr5020028
16. J. Bogusz, A. Klos, On the significance of periodic signals in noise analysis of GPS station coordinates time series. GPS Solut. (2016). https://doi.org/10.1007/s10291-015-0478-9
17. S. Zair, S. le Hégarat-Mascle, E. Seignez, Outlier detection in GNSS pseudo-range/doppler measurements for robust localization. Sensors (Switzerland) (2016). https://doi.org/10.3390/s16040580
18. H. Pehlivan, Frequency analysis of GPS data for structural health monitoring observations. Struct. Eng. Mech. (2018). https://doi.org/10.12989/sem.2018.66.2.185
19. X. He, K. Yu, J.P. Montillet, C. Xiong, T. Lu, S. Zhou, X. Ma, H. Cui, F. Ming, GNSS-TS-NRS: an open-source matlab-based GNSS time series noise reduction software. Remote Sens. (2020). https://doi.org/10.3390/rs12213532
20. D.J. Bartholomew, G.E.P. Box, G.M. Jenkins, Time series analysis forecasting and control. Oper. Res. Q. (1971). https://doi.org/10.2307/3008255
21. X. Luo, M. Mayer, B. Heck, Verification of ARMA identification for modelling temporal correlations of GNSS observations using the ARMASA toolbox. Studia Geophysica et Geodaetica (2011). https://doi.org/10.1007/s11200-011-0033-2
22. X. Luo, M. Mayer, B. Heck, Analysing time series of GNSS residuals by means of AR(I)MA processes. Int. Assoc. Geod. Symposia (2012). https://doi.org/10.1007/978-3-642-22078-4_19
23. A. Khodabandeh, A.R. Amiri-Simkooei, M.A. Sharifi, GPS position time-series analysis based on asymptotic normality of M-estimation. J. Geod. (2012). https://doi.org/10.1007/s00190-011-0489-4
24. S. Hekimoglu, B. Erdogan, R.C. Erenoglu, A new outlier detection method considering outliers as model errors. Exp. Tech. (2015). https://doi.org/10.1111/j.1747-1567.2012.00876.x
25. P. Barba, B. Rosado, J. Ramírez-Zelaya, M. Berrocoso, Comparative analysis of statistical and analytical techniques for the study of GNSS geodetic time series. Eng. Proc. (2021). https://doi.org/10.3390/engproc2021005021
26. C. Chen, L.M. Liu, Forecasting time series with outliers. J. Forecast. (1993). https://doi.org/10.1002/for.3980120103
27. J. Law, F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, W.A. Stahel, Robust statistics—the approach based on influence functions. Statistician (1986). https://doi.org/10.2307/2987975
28. S. Hekimoglu, K.R. Koch, How can reliability of the robust methods be measured? in *Third Turkish German Joint Geodetic Days*, ed. by Altan and Gründing, 1–4 June, Istanbul, Turkey (1999), pp. 179–196
29. D.A. Cucci, L. Voirol, G. Kermarrec, J.P. Montillet, S. Guerrier, The generalized method of wavelet moments with eXogenous inputs: a fast approach for the analysis of GNSS position time series. J. Geod. (2023). https://doi.org/10.1007/s00190-023-01702-8

30. K. Ji, Y. Shen, A wavelet-based outlier detection and noise component analysis for GNSS position time series. Int. Assoc. Geod. Symposia (2023). https://doi.org/10.1007/1345_2020_106
31. L.T. Hsu, GNSS multipath detection using a machine learning approach, in *IEEE Conference on Intelligent Transportation Systems, Proceedings*, ITSC (2018). https://doi.org/10.1109/ITSC.2017.8317700
32. Y. Xia, S. Pan, X. Meng, W. Gao, F. Ye, Q. Zhao, X. Zhao, Anomaly detection for urban vehicle GNSS observation with a hybrid machine learning system. Remote Sens. (2020). https://doi.org/10.3390/rs12060971
33. T. Kieu, B. Yang, C.S. Jensen, Outlier detection for multidimensional time series using deep neural networks. Proc. IEEE Int. Conf. Mob. Data Manag. (2018). https://doi.org/10.1109/MDM.2018.00029
34. B.P. Carlin, A.E. Gelfand, A.F. Smith, Hierarchical Bayesian analysis of changepoint problems. J. R. Stat. Soc. Ser. C (Appl. Stat.) **41**(2), 389–405 (1992). https://doi.org/10.2307/2347570
35. Y. Wang, Q.Q. Zhang, T.Y. Che, Y. Liu, Bayesian outlier-detection method based on autoregressive model for post-fit residuals analysis in GNSS. Zhongguo Guanxing Jishu Xuebao/J. Chin. Inert. Technol. (2016). https://doi.org/10.13695/j.cnki.12-1222/o3.2016.01.009
36. G. Gan, M.K.P. Ng, k-means clustering with outlier removal. Pattern Recognit. Lett. (2017). https://doi.org/10.1016/j.patrec.2017.03.008
37. H. Wang, S. Pan, W. Gao, Y. Xia, C. Ma, Multipath/NLOS detection based on K-means clustering for GNSS/INS tightly coupled system in urban areas. Micromachines (2022). https://doi.org/10.3390/mi13071128
38. F.C. Chan, M. Joerger, S. Khanafseh, B. Pervan, Bayesian fault-tolerant position estimator and integrity risk bound for GNSS navigation. J. Navig. (2014). https://doi.org/10.1017/S0373463314000241
39. G. Zhang, Q. Gui, S. Han, J. Zhao, W. Huang, A Bayesian method of GNSS cycle slips detection based on ARMA model, in *2017 Forum on Cooperative Positioning and Service, CPGPS* (2017). https://doi.org/10.1109/CPGPS.2017.8075128
40. Z. Qianqian, G. Qingming, Bayesian methods for outliers detection in GNSS time series. J. Geod. (2013). https://doi.org/10.1007/s00190-013-0640-5
41. S. Chen, Y. Li, J. Kim, S.W. Kim, Bayesian change point analysis for extreme daily precipitation. Int. J. Climatol. (2017). https://doi.org/10.1002/joc.4904
42. R.K. Pearson, Outliers in process modeling and identification. IEEE Trans. Control Syst. Technol. (2002). https://doi.org/10.1109/87.974338
43. R.K. Pearson, Y. Neuvo, J. Astola, M. Gabbouj, Generalized Hampel filters. Eurasip J. Adv. Signal Process. (2016). https://doi.org/10.1186/s13634-016-0383-6
44. C. Shah, R. Wies, A novel short-term residential load forecasting methodology using two-stage stacked LSTM and Hampel filter. IEEE PES Gen. Meet. (2022). https://doi.org/10.1109/PESGM48719.2022.9917173
45. M. Dagar, N. Mishra, A. Rani, S. Agarwal, J. Yadav, Performance comparison of Hampel and median filters in removing deep brain stimulation artifact. SCI (2018). https://doi.org/10.1007/978-981-10-4555-4_2
46. Z. Yao, J. Xie, Y. Tian, Q. Huang, Using Hampel identifier to eliminate profile-isolated outliers in laser vision measurement. J. Sens. (2019). https://doi.org/10.1155/2019/3823691
47. D. Grzechca, K. Tokarz, K. Paszek, D. Poloczek, Using MEMS sensors to enhance positioning when the GPS signal disappears. Lect. Notes Comput. Sci. Includ. Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinform. (2017). https://doi.org/10.1007/978-3-319-67077-5_25
48. D. Barry, J.A. Hartigan, A Bayesian analysis for change point problems. J. Am. Stat. Assoc. (1993). https://doi.org/10.2307/2290726
49. P. Fearnhead, Exact and efficient Bayesian inference for multiple changepoint problems. Stat. Comput. (2006). https://doi.org/10.1007/s11222-006-8450-8
50. G.T. Wilson, Time series analysis: forecasting and control, 5th Edition, by George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel and Greta M. Ljung, 2015. Published by John Wiley and Sons Inc., Hoboken, New Jersey, pp. 712. ISBN: 978-1-118-67502-1. J. Time Ser. Anal. (2016). https://doi.org/10.1111/jtsa.12194

Publisher's Note