Application of Outlier Detection Methods in GNSS Time Series Analysis

Huynh Dinh Quoc NGUYEN[1)], Quang Ngoc PHAM[2,3)], Vinh Duc TRAN[4)], Quoc Long NGUYEN[2)], Trong Gia NGUYEN[2,3)]*

1) Ho Chi Minh City of Natural Resources and Environment, Ho Chi Minh City, Vietnam; ORCID https://orcid.org/0009-0007-8447-9045
2) Faculty of Geomatics and Land Administration, Hanoi University of Mining and Geology, Hanoi, Vietnam; ORCID https://orcid.org/0009-0006-0765-245X; ORCID https://orcid.org/0009-0003-1616-8625; ORCID https://orcid.org/0000-0002-4792-3684
3) Geodesy and Environment research group, Hanoi University of Mining and Geology, Hanoi, Vietnam
4) Viet Nam's People Naval hydrographic and Oceanographic Department; vinhtduc@gmail.com; ORCID https://orcid.org/0009-0007-3087-8585
* Corresponding author: nguyengiatrong@humg.edu.vn

http://doi.org/10.29227/IM-2024-02-95

Submission date: 09-11-2024 | Review date:02-12-2024

Abstract

In the study of determining vertical displacements of the Earth's crust, GNSS is the technology that enables the highest accuracy in displacement measurement. Moreover, with GNSS time series data, it is possible to identify patterns of displacement over time. An existing issue to address is the detection of outliers and discontinuities within the measurement series. This study investigates outlier detection methods within GNSS time series data to serve the purpose of determining vertical displacements and predicting altitude component values over time. Methods such as IQR, Z-Score, and Percentile were implemented using data from CORS stations named HYEN, QNAM, and CTHO within the VNGEONET network in Vietnam. The data from these stations were initially analyzed using Gamit/Globk software to obtain daily coordinate components of the points. Results from outlier detection and analysis with the Multiple Linear Regression Model indicate that with approximately 2% of measurements identified as outliers, displacement may vary by 0.4mm/year. The LSTM+ICA artificial intelligence model demonstrated excellent performance in prediction with QNAM and CTHO datasets. However, prediction with the LSTM+ICA model raises ongoing research questions, particularly regarding the data collected by the HYEN station.

Keywords: *land vertical movement, plate tectonic, Gamit/Globk, GNSS data analysis, machine learning*

1. Introduction and literature review

The time series of GNSS data not only allows for precise determination of values, but also reveals patterns of Earth's crustal movements. The results in determining vertical crustal displacement are significant not only for geological deformation studies but also for sea level rise research [1]. In cases where land surfaces subside, the impact of typical sea level rise becomes more severe, as seen in the Mekong Delta region [2]. To determine land surface displacements, GNSS technology [3] can be utilized, combined with InSAR [4], integrated with GRACE data [5], or employing precise leveling methods [6].

One significant advantage of GNSS technology in monitoring vertical displacements, specifically, and Earth crust movements, in general, is its capability to determine these displacements with high accuracy [3]. Given the demand for precision in processing GNSS data, software packages like Bernese, Gamit/Globk, Gipsy-Oasis are utilized [7]. When applying GNSS technology for monitoring Earth crust movements, both traditional static relative measurement methods and continuous measurement solutions with CORS station networks can be employed. The advantage of data collected from CORS station networks is the ability to determine displacements over time. Consequently, accurate displacement patterns, including trends and seasonal variations, can be identified when analyzing time series data [8] [9].

Outliers in GNSS data series arise from various causes such as the effects of multipath phenomena, influences from atmospheric layers, or errors caused by hardware and software of the receiver.

There are various methods for outlier detection in general such as Statistical, Distance-based, Density-based, Cluster-based. If employing statistical methods, they can be further categorized into groups such as based on Gaussian distribution, regression analysis, using charts, or Kernel methods [10]. The Bayesian method has been utilized to detect cycle slips when using carrier phase measurements based on adaptive Gibbs sampling [11]. The JN-test method, belonging to the minimum separable bias methods, has been applied to detect outliers in the positioning problem solving process [12]. This method proves effective when the correlation coefficient of the data to be detected is high. Also, aiming to enhance the accuracy of GNSS positioning, a contrario model has been proposed to detect outliers in measurements significantly improving the accuracy of positioning results [13].

The method of Mean Absolute Deviation (MAD) has been applied to analyze GNSS data series over time. The authors experimented with outlier values ranging from 3 to 5 times the standard deviation [14]. The approach using WA_MINQUE has been proposed and shown to be more effective than the traditional LS_MINQUE method [15].

In addition to traditional methods, many research works have applied artificial intelligence in detecting and eliminating outliers. The authors [16] utilized models such as GBDT, LSTM, SVM for outlier detection and elimination. Experimental results showed that applying these methods improved accuracy

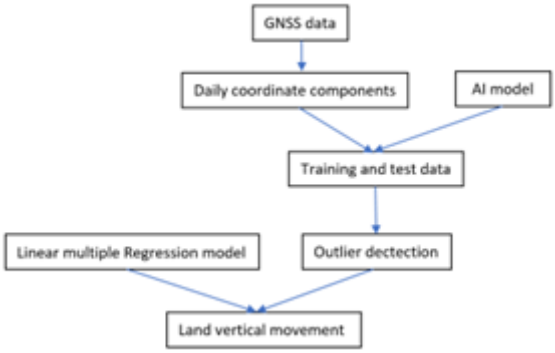

Fig. 1. Method for determining ground displacement with integrated outlier exclusion
Rys. 1. Metoda określania przemieszczenia gruntu ze zintegrowanym wykluczeniem wartości odstających

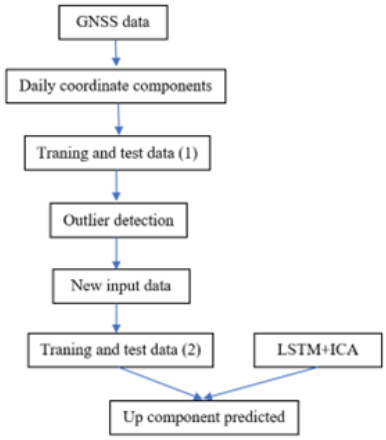

Fig. 2. Prediction of the Up-component value with the LSTM+ICA model
Rys. 2. Przewidywanie wartości składowej górnej za pomocą modelu LSTM+ICA

by 30% compared to traditional methods. The rigorous statistical method of Theta was applied for outlier detection, resulting in a 3.22% increase in accuracy compared to the contrasting artificial intelligence models [17]. The authors [18] surveyed the ability to detect outliers using models such as Isolation Forest, O-C SVM, LOF, with the best performance achieved by the Isolation Forest model. The Multi-Layer Perceptron Neuron Network model was applied to analyze altitude components in GNSS data series over time, with an RMSE result of 0.006 for both the training and testing datasets [19].

The objective of this study is to evaluate the effectiveness of outlier detection methods including Interquartile Range, Z-Score, and Percentile in determine vertical displacements and of the Earth's crust and predict the Up component of the point using artificial intelligence from continuous GNSS measurement data series.

2. Methodology

This study conducts an investigation into the results when applying outlier detection methods to determine the vertical displacement of the Earth's crust and forecast the altitude component of the point using artificial intelligence. The process of applying outlier detection methods in determining the vertical displacement of the Earth's crust is illustrated in Figure 1.

Prediction is a crucial aspect in time series data analysis. In this study, an artificial intelligence model is used to predict the Up-component value, which is the LSTM + ICA function. The process of predicting the Up component of the point using artificial intelligence is illustrated in Figure 2.

2.1. Interquartile Range method (IQR)

The Interquartile Range method (IQR) is a statistical technique used to measure the spread or dispersion of a dataset. IQR is particularly useful for identifying outliers or extreme values in the dataset, with low sensitivity to the presence of outliers, as it relies only on the central part of the distribution to calculate IQR [20]. IQR consists of three values: the first quartile Q_1 (25% of the data), the second quartile Q_2 (50% of the data), and the third quartile Q_3 (75% of the data). The dataset is divided into 4 equal parts with an equal number of data points. IQR is determined from Q_1 and Q_3: $IQR = Q_3 - Q_1$. From there, upper and lower bounds are established to exclude outliers (Figure 3). However, the effectiveness of the IQR method is better when used to identify outliers in datasets with a large number of data points [21].

Furthermore, an adjusted version of the IQR method is proposed to account for the influence of the data point count. The interquartile range is calculated as 4.IQR, denoted as 4IQR[1+0.1log(n/10)], and the upper and lower bounds are recalculated as Q1–k.IQR[1+0.1log(n/10)] and Q3+k.IQR[1+0.1log(n/10)] respectively [21].

2.2. Z-Score method

The Z-Score is a statistical method used to determine the position of a value within a dataset relative to the mean and

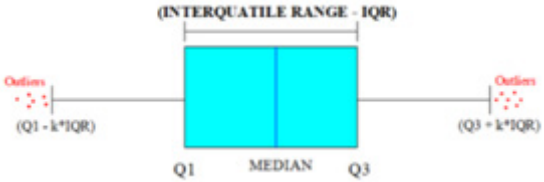

Fig. 3. Chart illustrating the IQR method
Rys. 3. Wykres ilustrujący metodę IQR

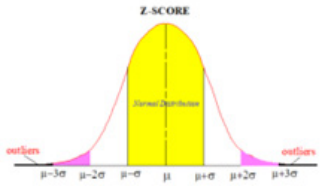

Fig. 4. Chart illustrating the Z-Score method
Rys. 4. Wykres ilustrujący metodę Z-Score

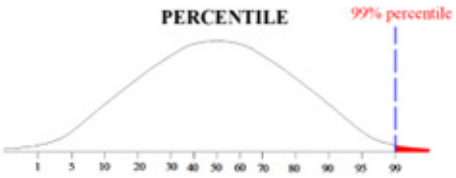

Fig. 5. Chart illustrating the Percentile method
Rys. 5. Wykres ilustrujący metodę percentylową

standard deviation of that dataset. It measures the distance of a value from the mean of the dataset in terms of standard deviations [22].
In which:
x: The value for which Z-Score is to be calculated.
μ: The mean value of the dataset.
σ: The standard deviation of the dataset.

The Z-Score method effectively detects outliers in datasets that follow a normal distribution. However, it is inefficient for datasets that are skewed left or right. Nevertheless, the mean and standard deviation of the dataset can be heavily influenced by one or a few extreme values. Therefore, the median value and the Median Absolute Deviation (MAD) are used in the adjusted Z-Score method [23].

$$MAD = median|x_i - \tilde{x}| \quad (1)$$

In which: $\tilde{x}$he median of the dataset.

2.3. Percentile method

Percentiles are a commonly used algorithm in machine learning for calculating percentages. The Percentiles algorithm, commonly used in statistics, is a descriptive measure of certain percentage values of values lower than it.

When a dataset is unevenly distributed with values skewed to one side, as shown in the diagram below, Percentile methods are used to remove these values. This method removes outliers by considering the ratio between the values within the 90th percentile and removing outliers beyond the 10th percentile.

2.4. Evaluation methods

In evaluating the accuracy of artificial intelligence models, various criteria are considered, including Root Mean Square Error (RMSE), Mean Square Error (MSE), and Mean Absolute Error (MAE). RMSE measures the square root of the average of the squared differences between predicted and actual values, providing insight into the model's overall accuracy. MSE calculates the average of the squared differences, giving a quantitative assessment of prediction errors. MAE computes the average of the absolute differences, offering a clear understanding of the model's average prediction deviation. These metrics collectively aid in assessing the effectiveness and precision of AI models in various applications.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y - \hat{y}| \quad (2)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y - \hat{y})^2 \quad (3)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y - \hat{y})^2} \quad (4)$$

3. GNSS CORS data and analysis

The data used in this study were collected by 3 CORS stations belonging to the VNGEONET network in Vietnam. Information about the data collected at each station is provided in Table 1 below.

The collected data is converted to RINEX format and processed using Gamit/Globk software to ensure the accuracy of the determined daily coordinate components [24, 25] following the procedure described by [26].

During processing, data from permanent IGS stations, precise ephemeris, as well as other necessary auxiliary data, are automatically downloaded by the Gamit/Globk software.

Tab. 1. Information about the collected data

Tab. 1. Informacje o zebranych danych

No.	Station name	Time period	Interval (second)	Receiver / antenna type
1	HYEN	2019/08/10 - 2022/03/18	30	LEICA GR50 / LEIAR25.R4 LEIT
2	QNAM			
3	CTHO			

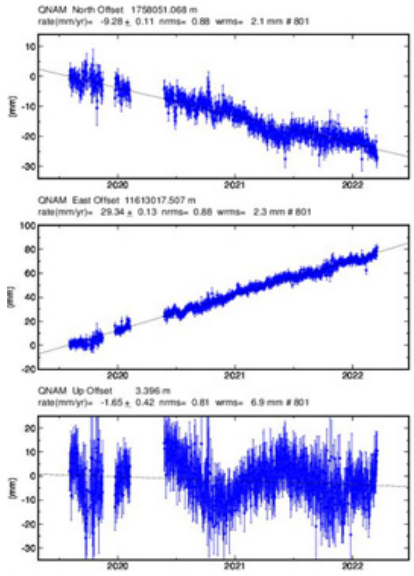

Fig. 6. Daily coordinate components of point QNAM station determined by Gamit/Globk software

Rys. 6. Dobowe składowe współrzędnych punktu stacji QNAM wyznaczone przez oprogramowanie Gamit/Globk

Tab. 2. Statistics of the identified number of outlier measurements

Tab. 2. Statystyki zidentyfikowanej liczby pomiarów odstających

Station	Raw data	Number of outlier measurements		
		IQR	Z-Score	Percentile
HYEN	948	0	0	20
QNAM	801	4	2	16
CTHO	929	0	0	20

Tab. 3. Vertical displacement amounts of points

Tab. 3. Pionowe wartości przemieszczenia punktów

Station	Value of vertical displacement (mm/year)			
	Gamit/Globk	IQR	Z-Score	Percentile
HYEN	-17.59	-13.7	-13.7	-13.3
QNAM	-5.31	-1.7	-1.6	-1.5
CTHO	-14.26	-10.6	-10.6	-10.3

Information about the data from permanent stations is provided in [27].

The result of processing GNSS data is the daily coordinate components of the point, with a graph depicting the coordinate components of point QNAM as shown in Figure 2.

4. Results and discussion

By utilizing the outlier detection process as presented in Figures 1 and 2, the number of outlier measurements can be determined as shown in Table 2.

4.1 Determine vertical displacement value

The outliers identified will be removed from the original dataset to obtain a new input dataset. Subsequently, the multiple linear regression model (formula 5) will be utilized to determine the vertical displacement of the ground in millimeters per year.

$$y = a + bt + c\sin(2\pi t) + d\cos(2\pi t) + e\sin(4\pi t) + f\cos(4\pi t) \quad (5)$$

In formula 4, a, b, c, d, e, f are the parameters to be determined; y is the value of the daily coordinate component. The determined displacement amounts under various scenarios are presented in Table 3.

The results in Table 3 indicate that, with approximately 2% of total measurements identified as outliers, the determined amount of shift may vary at a rate of 0.4mm/year.

To assess the performance of the multiple linear regression model, statistical metrics including R-squared, F-statistic, and Durbin-Watson have been identified. The computed results are statistically summarized in Table 4.

The data in Table 4 indicates that the data from station HYEN is the most suitable, while the data from station QNAM is the least suitable for the Multiple Linear Regression model. This could be attributed to the discontinuity of the QNAM station data throughout the entire observation period. The CTHO station data shows no correlation in the input dataset, whereas the input data of the HYEN station exhibits the highest level of autocorrelation.

Tab. 4. Statistical Characteristics of the Multiple Linear Regression Model
Tab. 4. Charakterystyka statystyczna modelu wielokrotnej regresji liniowej

Station	Model		
	IQR	Z-Score	Percentile
R-squared			
HYEN	0.804	0.804	0.807
QNAM	0.409	0.408	0.399
CTHO	0.702	0.702	0.701
F-statistic			
HYEN	774.6	774.6	776.2
QNAM	109.5	109.2	103.6
CTHO	433.9	433.9	422.7
Durbin-Watson			
HYEN	1.273	1.273	1.277
QNAM	1.502	1.491	1.445
CTHO	1.612	1.612	1.591

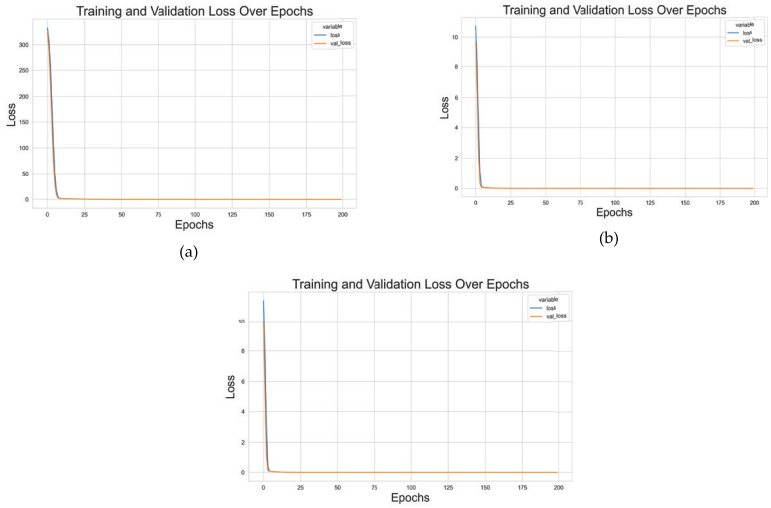

Fig. 7. Training performance of the LSTM+ICA model for forecasting Up component for (a) HYEN, (b) QNAM, (c) CTHO station
Rys. 7. Wydajność treningu modelu LSTM+ICA dla prognozowania składowej Up dla (a) HYEN, (b) QNAM, (c) stacji CTHO

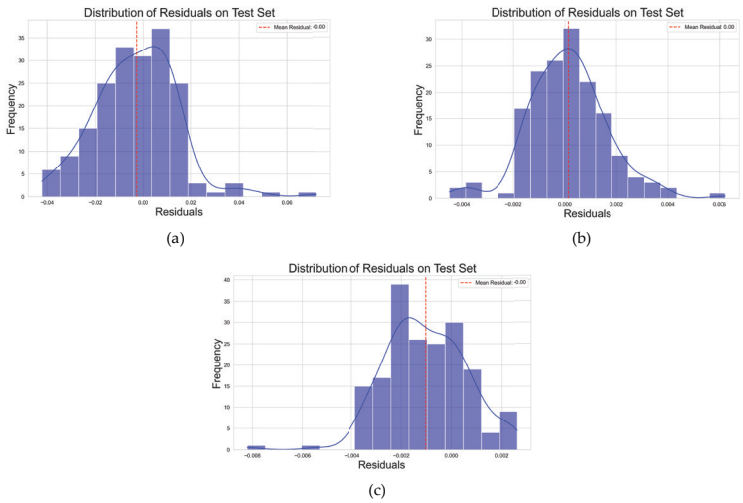

Fig. 8. Distribution of Residuals on Test Set
Rys. 8. Rozkład reszt na zestawie testowym

4.2 Predicting Up component

As mentioned in Figure 2, the LSTM+ICA model has been employed in this study to predict the Up component. The dataset after outlier exclusion is divided into training and testing data with corresponding ratios of 80% and 20%, respectively. During processing, the batch size value is selected as 16, the epoch is set to 200, and the Adam optimization algorithm is used. Figures 7 and 8 below illustrate the prediction results with the LSTM+ICA model for the data of the mentioned stations.

To evaluate the effectiveness of the prediction model, characteristic values have been computed as described in formulas (1), (2), and (3) and are summarized in Table 5.

The data in Table 5 indicates that when analyzing the time series of GNSS results for stations CTHO and QNAM, there is no apparent difference between outlier detection and

Tab. 5. Performance characteristic statistics of the LSTM+ICA model
Tab. 5. Statystyki charakterystyczne wydajności modelu LSTM+ICA

Station	No outlier	Model		
		IQR	Z-Score	Percentile
RMSE (mm)				
HYEN	30.4	28.3		26.9
QNAM	1.7	1.9		1.8
CTHO	2.4	2.0		2.1
MAE (mm)				
HYEN	22.0	15.5		23.1
QNAM	1.1	1.4		1.2
CTHO	1.9	1.5		1.5

non-outlier detection. However, for the case of station HYEN, utilizing the integrated outlier detection solution allows for a 10% improvement in model performance.

Although the data from station QNAM is not continuous, with a 15% (801 out of 948 measurements) missing rate, the proposed model's performance is relatively good compared to existing publications [19]. However, the prediction performance of the LSTM+ICA model is very poor when applied to the dataset of station HYEN. It needs to be investigated whether this is due to the model's unsuitability or the data collected by this station not being properly processed in previous stages.

5. Conclusion

The research highlights the importance of accurately determining vertical displacements and predicting altitude components, particularly in the context of understanding sea level changes and land subsidence or uplift. The research methodology involves evaluating outlier detection methods such as Interquartile Range, Z-Score, and Percentile, alongside utilizing artificial intelligence models like LSTM+ICA for prediction. The process includes comprehensive data analysis and statistical evaluation to assess the effectiveness of these approaches.

The outlier detection results indicate that the Percentile method allows for the detection of outliers with the highest number of detected measurements, reaching approximately 10% of the total input measurements. With the identified number of outlier measurements as mentioned above, the vertical displacement of the Earth's crust at the specified stations varies, with the maximum value being 0.4mm/year. Statistical results show that the CTHO dataset has the lowest level of autocorrelation, while the HYEN dataset has the highest.

When applying the LSTM+ICA model, it demonstrates very good predictive performance for the QNAM and CTHO station datasets, with statistics including RMSE = 2.1 mm, MAE = 1.5 mm, even though the QNAM station data is not continuous (with 15% missing data). However, when applied to predict the HYEN station dataset, the model yields poor results, highlighting the need for further research to address issues related to the high autocorrelation of the HYEN station data.

Acknowledgment

The author group sincerely thanks the Vietnam Department of Survey, Mapping, and Geographic Information (DOSM) for providing the data collected by CORS stations in the VNGEONET network to conduct this paper.

Funding

This work was financially supported by the Ministry of Natural Research and Environment in Vietnam under grant number TNMT.2024.02.07.

Literatura – References

1. Raj, N.J.M., Prediction of sea level with vertical land movement correction using deep learning. 2022. 10(23): p. 4533.
2. Olaf, N., Vấn đề dưới mặt đất-sụt lún đất tại đồng bằng Sông Cửu Long. 2019, Deutsche gesellschaft für internationale zusammenarbeit (GIZ).
3. Uzel, T., et al., Monitoring the tectonic plate movements in Turkey based on the national continuous GNSS network. 2013. 6: p. 3573-3580.
4. Abidin, H., et al., Land subsidence in coastal city of Semarang (Indonesia): characteristics, impacts and causes. 2013. 4(3): p. 226-240.
5. Wang, L., et al., Detecting seasonal and long-term vertical displacement in the North China Plain using GRACE and GPS. Hydrol. Earth Syst. Sci., 2017. 21(6): p. 2905-2922.
6. Kowalczyk, K. and J.J.A.G.e.G. Rapinski, Evaluation of levelling data for use in vertical crustal movements model in Poland. 2013. 10(4): p. 172.
7. Wu, D., H. Yan, and Y. Shen, TSAnalyzer, a GNSS time series analysis software. Gps Solutions, 2017. 21: p. 1389-1394.
8. Goudarzi, M.A., GPS inferred velocity and strain rate fields in eastern Canada. 2016.
9. Klos, A., et al., Modelling the GNSS time series: different approaches to extract seasonal signals. Geodetic time series analysis in earth sciences, 2020: p. 211-237.
10. Smiti, A., A critical overview of outlier detection methods. Computer Science Review, 2020. 38: p. 100306.
11. Qianqian, Z. and G. Qingming, Bayesian methods for outliers detection in GNSS time series. Journal of Geodesy, 2013. 87(7): p. 609-627.
12. Wang, J. and N.L. Knight, New outlier separability test and its application in GNSS positioning. Journal of global positioning systems, 2012. 11(1): p. 46-57.
13. Zair, S., S. Le Hégarat-Mascle, and E. Seignez, Outlier detection in GNSS pseudo-range/Doppler measurements for robust localization. Sensors, 2016. 16(4): p. 580.
14. Klos, A., et al. On the handling of outliers in the GNSS time series by means of the noise and probability analysis. in IAG 150 Years: Proceedings of the IAG Scientific Assembly in Postdam, Germany, 2013. 2016. Springer.
15. Ji, K. and Y. Shen. A Wavelet-Based Outlier Detection and Noise Component Analysis for GNSS Position Time Series. in Beyond 100: The Next Century in Geodesy. 2022. Cham: Springer International Publishing.
16. Gao, W., et al., Modelling and prediction of GNSS time series using GBDT, LSTM and SVM machine learning approaches. Journal of Geodesy, 2022. 96(10): p. 71.
17. Kiani, M., A specifically designed machine learning algorithm for GNSS position time series prediction and its applications in outlier and anomaly detection and earthquake prediction. arXiv preprint arXiv:2006.09067, 2020.
18. Huy, N.Đ. and T.Đ. Trọng, Phát hiện ngoại lai trong chuỗi tọa độ GNSS bằng máy học. Tạp chí Khoa học kỹ thuật Mỏ - Địa chất, 2023. 64(4): p. 22-30.
19. Vân Phong, D., et al., Phân tích chuyển dịch thẳng đứng vỏ Trái đất sử dụng hàm ANN từ kết quả xử lý chuỗi dữ liệu GNSS theo thời gian.
20. Wang, C., J. Caja, and E. Gómez, Comparison of methods for outlier identification in surface characterization. Measurement, 2018. 117: p. 312-325.
21. Barbato, G., et al., Features and performance of some outlier detection methods. Journal of Applied Statistics, 2011. 38(10): p. 2133-2149.
22. Kannan, K.S., K. Manoj, and S. Arumugam, Labeling methods for identifying outliers. International Journal of Statistics and Systems, 2015. 10(2): p. 231-238.
23. Iglewicz, B. and D.C. Hoaglin, Volume 16: how to detect and handle outliers. 1993: Quality Press.
24. Li, Y., Analysis of GAMIT/GLOBK in high-precision GNSS data processing for crustal deformation. Earthquake Research Advances, 2021. 1(3): p. 100028.
25. Rakhimberdieva, M., et al. Processing of GNSS data in Gamit/Globk: On the example of the reference stations of the Uzbekistan network. in E3S Web of Conferences. 2023. EDP Sciences.
26. Savchuk, S., et al., The Seasonal Variations Analysis of Permanent GNSS Station Time Series in the Central-East of Europe. Remote Sensing, 2023. 15(15): p. 3858.
27. Hung, V.T., et al., Contemporary movement of the Earth's crust in the Northwestern Vietnam by continuous GPS data. Vietnam Journal of Earth Sciences, 2020. 42(4): p. 334-350.

Zastosowanie metod wykrywania wartości odstających w analizie szeregów czasowych GNSS

W badaniach nad określaniem pionowych przemieszczeń skorupy ziemskiej GNSS jest technologią, która umożliwia najwyższą dokładność pomiaru przemieszczeń. Co więcej, dzięki danym z szeregów czasowych GNSS możliwe jest zidentyfikowanie wzorców przemieszczeń w czasie. Istniejącą kwestią do rozwiązania jest wykrywanie wartości odstających i nieciągłości w serii pomiarowej. W niniejszym badaniu zbadano metody wykrywania wartości odstających w danych szeregów czasowych GNSS w celu określenia przemieszczeń pionowych i przewidywania wartości składowych wysokości w czasie. Metody takie jak IQR, Z-Score i Percentile zostały zaimplementowane przy użyciu danych ze stacji CORS o nazwach HYEN, QNAM i CTHO w sieci VNGEONET w Wietnamie. Dane z tych stacji zostały wstępnie przeanalizowane przy użyciu oprogramowania Gamit/Globk w celu uzyskania dziennych składowych współrzędnych punktów. Wyniki wykrywania wartości odstających i analizy za pomocą modelu wielokrotnej regresji liniowej wskazują, że przy około 2% pomiarów zidentyfikowanych jako wartości odstające, przemieszczenie może różnić się o 0,4 mm/rok. Model sztucznej inteligencji LSTM+ICA wykazał doskonałą wydajność w przewidywaniu dla zbiorów danych QNAM i CTHO. Jednak przewidywanie za pomocą modelu LSTM+ICA rodzi ciągłe pytania badawcze, szczególnie w odniesieniu do danych zebranych przez stację HYEN.

Słowa kluczowe: *ruch pionowy lądu, tektonika płyt, Gamit/Globk, analiza danych GNSS, uczenie maszynowe*