A Machine Learning Detection of Outliers in InSAR Displacement Time Series

Lukas Kubica[1], Juraj Papco1, Richard Czikhardt[1], Matus Bakon[2,3], Jan Barlak[3], Martin Rovnak[3]

[1]Department of Theoretical Geodesy and Geoinformatics, Faculty of Civil Engineering, Slovak University of Technology in Bratislava, 81005 Bratislava, Slovakia

[2]insar.sk s.r.o., Lesna 35, 08001 Presov, Slovakia; matusbakon@insar.sk

[3]University of Presov, Faculty of Management and Business, Department of Finance, Accounting and Mathematical Methods, Konstantinova 16, 080 01 Presov, Slovakia

Correspondence to: Lukas Kubica[1] (lukas.kubica@stuba.sk)

https://doi.org/10.31490/9788024846026-11

Abstract

Multi-temporal SAR interferometry (InSAR) estimates the displacement time series of coherent radar scatterers. Current InSAR processing approaches often assume the same deformation model for all scatterers within the area of interest. However, this assumption is often wrong, and time series need to be approached individually.

Individual, point-wise approach for large InSAR datasets is limited by high computational demands. The additional problem is imposed by the presence of outliers and phase unwrapping errors, which directly affect the estimation quality.

This work describes the algorithm for (i) estimating and selecting the best displacement model for individual point time series and (ii) detecting outlying measurements in the time series. The InSAR measurement quality of individual scatterers varies, which affects the estimation methods. Therefore, our approach uses a priori variances obtained by the variance components estimation within geodetic InSAR processing.

We present two different approaches for outlier detection and correction in InSAR displacement time series. The first approach uses the conventional statistical methods for individual point-wise outlier detection, such as median absolute deviation (MAD) confidence intervals around the displacement model. The second approach uses machine learning principles to cluster points based on their displacement behaviour as well as the temporal occurrence of outliers. Using clusters instead of individual points allows for more efficient analysis of average time series per cluster and consequent cluster-wise outlier detection, correction, and time-series filtering.

The two approaches have been applied on the Sentinel-1 InSAR time series of a case study from monitoring landslides in Slovakia. The area of interest is affected by characteristic non-linear progression of the movement. Our post-processing procedure parameterized the displacement time series despite the presence of a non-linear motion, thus enabling reliable outlier detection and unwrapping error correction. The validation of the proposed approaches was performed on an existing network of corner reflectors located within the area of interest.

Keywords: InSAR, time series, outliers, phase jumps, machine learning, clustering

MOTIVATION

Implementing InSAR time series analysis based on geodetic estimation theory (Leijen 2014) allows us to obtain information about the displacement of the large number of persistent scatterers with millimetre accuracy over a wide area.

A frequent presence of systematic errors (outliers, phase jumps, etc.) in displacement time-series is dominantly the result of imperfections of the phase unwrapping algorithm as well as imperfections of the atmospheric phase contribution modelling.

Large amounts of data require exploitation of methods which are computationally effective and at the same time allow a certain level of automation. Therefore, the main objective of this paper is to design suitable methods for detection of outlier and correction phase jumps in InSAR displacement time series exploiting conventional statistics methods as well as machine learning algorithms.

METHODOLOGY

In this paper two different post-processing approaches for filtering of InSAR displacement time series are proposed: i) individual, and ii) clustering-based approach. Both approaches use statistical modelling of the time series with selection of suitable mathematical models for deformation estimates. Appropriate model is selected based on hypothesis testing, which verifies the statistical significance of the parameters of the model (Kerkhof et al., 2020). In both cases, the identification of outliers is based on the definition of the median absolute deviation (MAD) confidence interval.

Time series modelling and selection of the appropriate model

The time series are being modelled by a predefined set of selected model types which can be adapted in accordance with the local characteristics of the area of interest (e.g., quadratic model for the landslides). In this work we have used four basic types of deformation models:

- Stationary model
- Linear model or trend
- Seasonal model
- Quadratic model

For reliable detection of outliers, it is necessary to select the model that best describes the deformation recorded for each individual measurement point. The model selection is based on an iterative method, in which the parameters of individual models are tested. The null and alternative hypotheses are defined as:

$$\begin{aligned} H_0&: \quad E(\hat{\theta}_i) = 0 \\ H_1&: \quad E(\hat{\theta}_i) \neq 0 \end{aligned} \tag{1}$$

The test statistics is the value $\hat{T}_i$ calculated as the ration of the estimated parameter $\hat{\theta}_i$ to its standard deviation:

$$\hat{T}_i = \frac{\hat{\theta}_i}{\sqrt{\left\{\hat{\Sigma}_\theta\right\}_{i,i}}} \tag{2}$$

The test statistics has a Student's probability distribution with n-k degrees of freedom and significance level of $\alpha = 0.05$ (5 %).

Determination of the confident interval and detection of outliers

Standard outlier detection approaches are used to determine the confidence interval based on parameters such as mean value μ and standard deviation σ. The confidence interval boundaries are then defined as $\pm 3\sigma$ around the mean value μ. There are 3 major problems while adopting this method (Leys et al. 2013):

- The parameters such as μ and σ are strongly dependent on the presence of outliers in the data set.
- The method strictly assumes a normal probability distribution of the data set.
- The method in unreliable in detecting outliers of a data set with a smaller number of observations.

Based on the above, we decided to use median absolute deviation to determine confidence interval. In accordance with Leys et. al. (2013) the use of the MAD parameter in determining the confidence interval has several advantages in comparison to standard approaches:

- The MAD is a robust parameter, i.e., not dependent on the presence of outliers inside the data set.
- It is independent of the size of the data set.
- It is independent of the probability distribution of the data set.

The MAD was calculated based on the residuals between the estimated model and actual measurements:

$$MAD = median(|v_i - \bar{v}|) \tag{3}$$

where v_i are the time series residues and $\bar{v}$ is the median of the residues calculated from the residual time series as $\bar{v} = median(v_i)$.

Calculated MAD parameter is used to determine the confidence interval of $\pm 3\sigma$, with the difference that the parameter σ was replaced by the value of MAD. For the correct application of this rule, the MAD value must be converted to the same probability distribution as the residuals (normal distribution). The conversion is defined as:

$$MAD = k \cdot median(|v_i - \bar{v}|) \quad \textbf{(4)}$$

where k is the conversion factor. In the case of normal probability distribution, $k = 1.4826$ (Leys, 2013).

Another difference of this approach in comparison with the standard approach is that the confidence interval boundaries of $\pm 3MAD$ are defined from the curve of the function representing the time series model (see Fig. 1) and not from the mean value as in the standard approach.

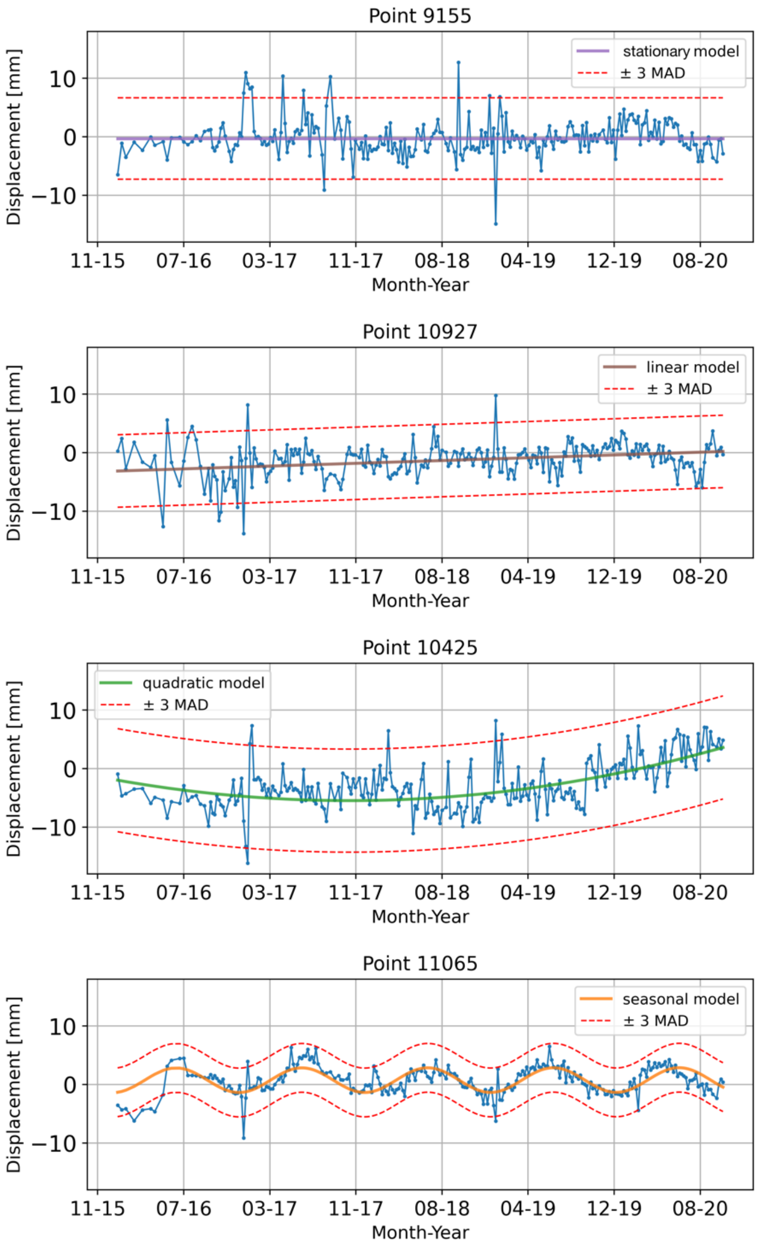

Fig. 1. Four basic types of deformation models with MAD boundaries

Individual approach

The proposed methodology is based on an individual, point-wise approach to estimated displacement time series. From a practical point of view, this means that within the processing chain, each time series are evaluated separately. The steps of the individual approach are shown on a flow chart (see Fig. 2).

The individual approach implements a flexible correction, which ensures a variability of the correction for the given epoch. The correction is determined based on the a priori error defined by the variation components, which are estimated inside the InSAR time series analysis processing chain. The correction value for a given epoch is calculated as:

$$\rho_t = \lambda/2 \ \pm 2.5\sigma_t \tag{5}$$

where λ is wavelength of the radar sensor and σ_t is the value of a variance component for the given epoch.

Clustering approach

The clustering approach uses machine learning algorithms in the form of t-SNE and DBSCAN algorithms. The main objective of this approach is to group the displacement time series based on their similar properties, such as same deformation model, same anomalies or occurrence of outliers in the same epochs. The input data set must be reduced to a pseudo-2-D dimension prior to performing a clustering algorithm. We perform data reduction using the stochastic algorithm t-SNE. After dimensionality reduction, we cluster 2-D dimension using the DBSCAN clustering algorithm. As a result of clustering, the input data set is divided into clustering groups with similar characteristics.

The mean time series is then calculated for each cluster. Estimation of the time series model and suitable model selection is performed only for the mean time series computed for each cluster. Similarly, the detection of outliers is performed for the mean time series. Based on these results, the same correction is applied for the time series allocated in the same group and in specific epochs. A stepwise procedure for clustering approach is shown on flowchart in Figure 2.

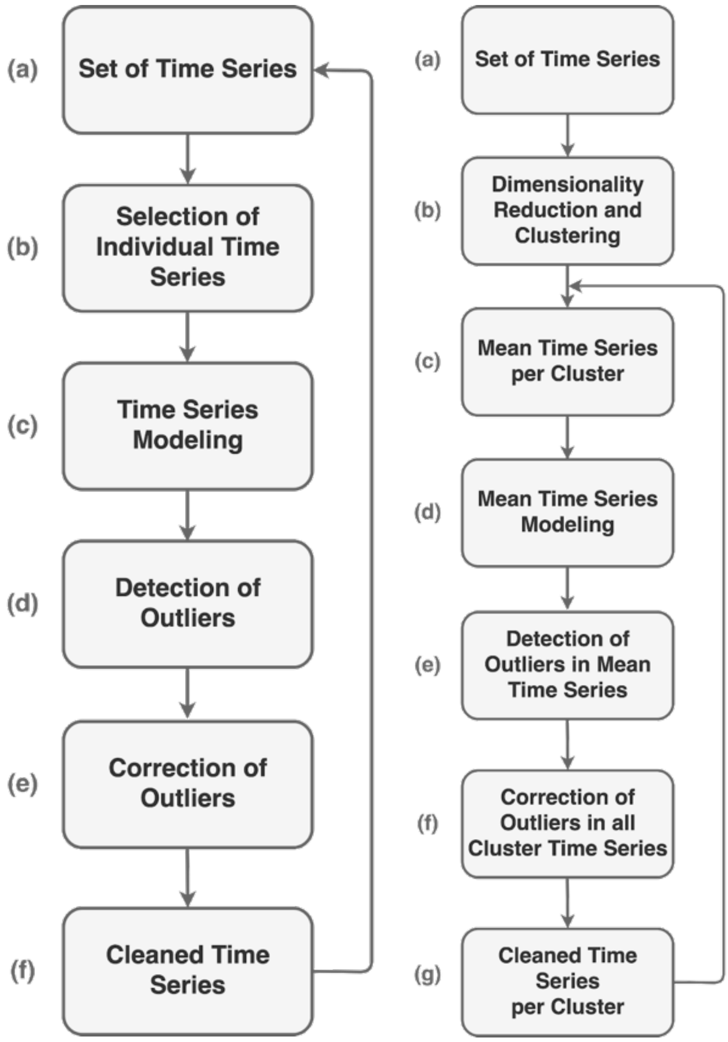

Fig. 2. Flowcharts: Individual approach (left), clustering approach (right)

CASE STUDY FROM MONITORING LANDSLIDES IN SLOVAKIA

Both proposed approaches (individual vs. clustering) for outlier detection and error correction were tested on a case study from the Upper Nitra region and the city of Prievidza, Slovakia. The area of interest is affected by landslides and land subsidence with a characteristic non-linear progression of the movement caused by the geological structure and lignite mining in the past (Czikhardt et al. 2017).

The data are the product of the Multi-Temporal InSAR analysis of Sentinel-1A/B images from 8.1.2016 to 31.10.2020 (254 epochs). The resulting data set contains approximately 6000 persistent scatterers (> 1.5 millions of measurements) visualized in Figure 3.

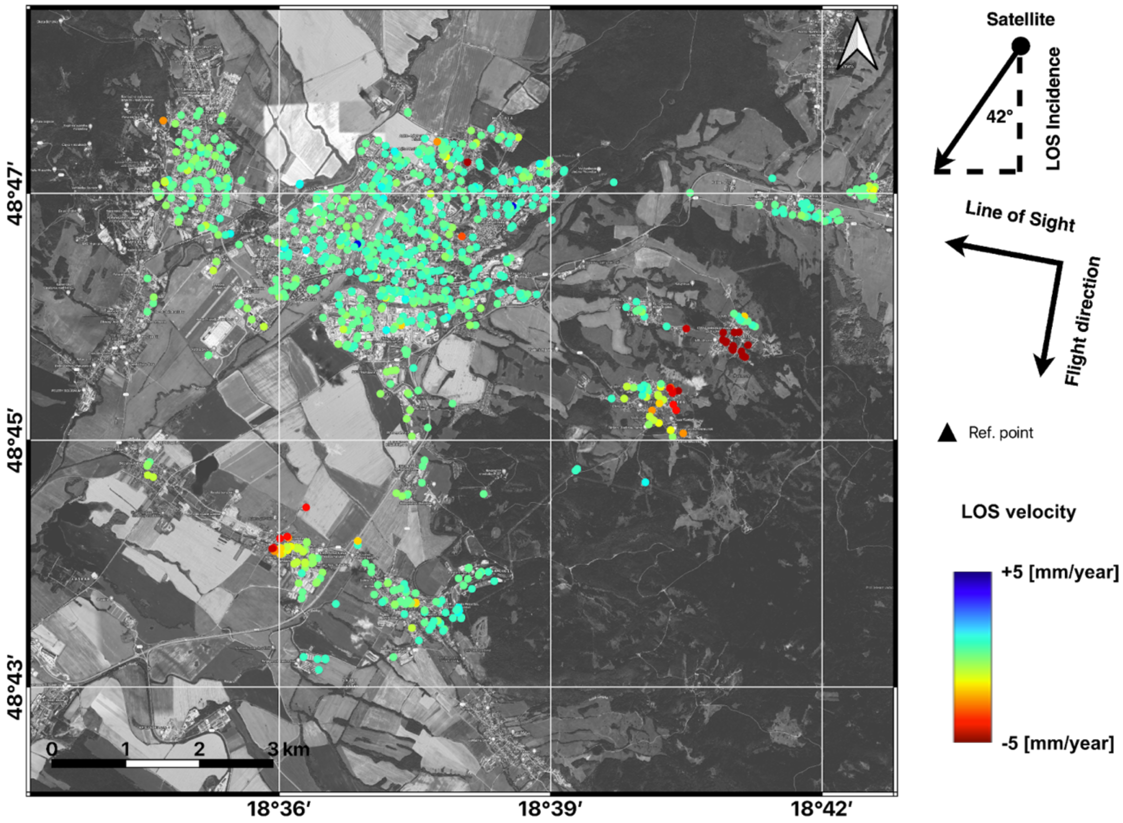

Fig. 3. Linear velocity map

Application of the individual approach:

Table 1 shows a smaller number of implemented corrections in relation to the total number of identified outliers as the individual approach is designed to primarily identify and correct systematic errors caused by phase jumps (Equation 5).

Table 1. Results of Individual approach

Total number of measurements	1 530 858
Number of identified outliers	51 618 (3.37%)
Number of corrections	**8292**

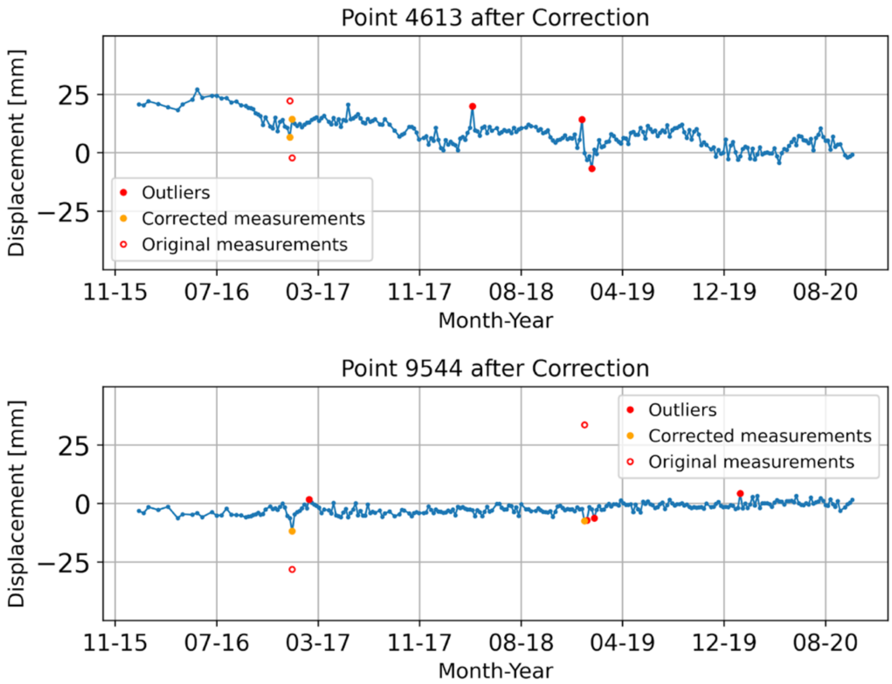

Fig. 4. Time series after application of the individual approach

Application of the clustering approach:

By applying clustering algorithms to the data set, we identify 18 unique groups of the displacement time series with outliers labeled -1 (i.e. points that are not included in any groups).

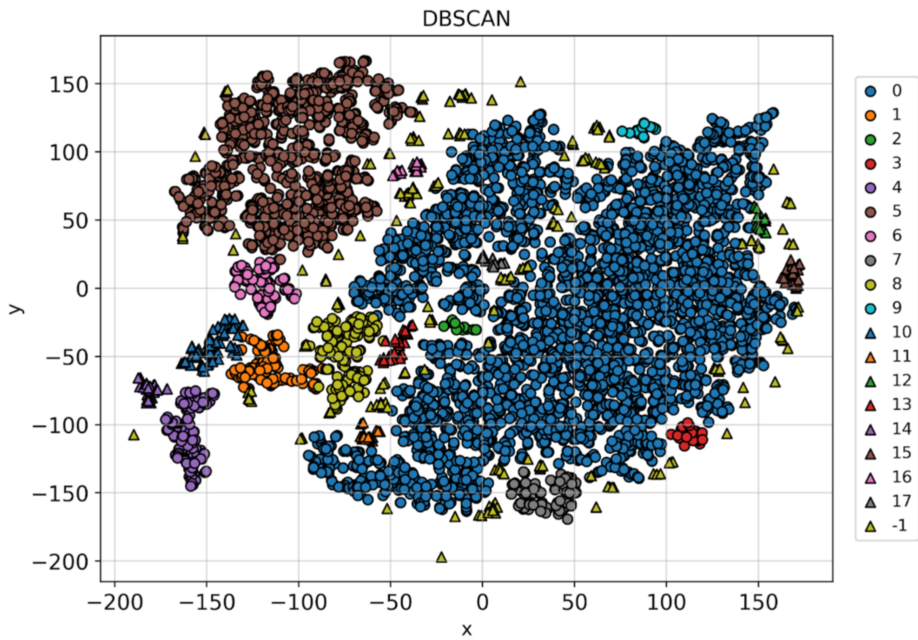

Fig. 5. Pseudo 2-D dimension of the data set after application of the t-SNE algorithm

The mean time series estimated for each group exposes a significant incidence of outliers in the winter months, which is presumably corresponding to the thickness of the show cover. The snow cover significantly reduces the coherence of persistent scatterers and thus weakens the quality of measurements in given epochs.

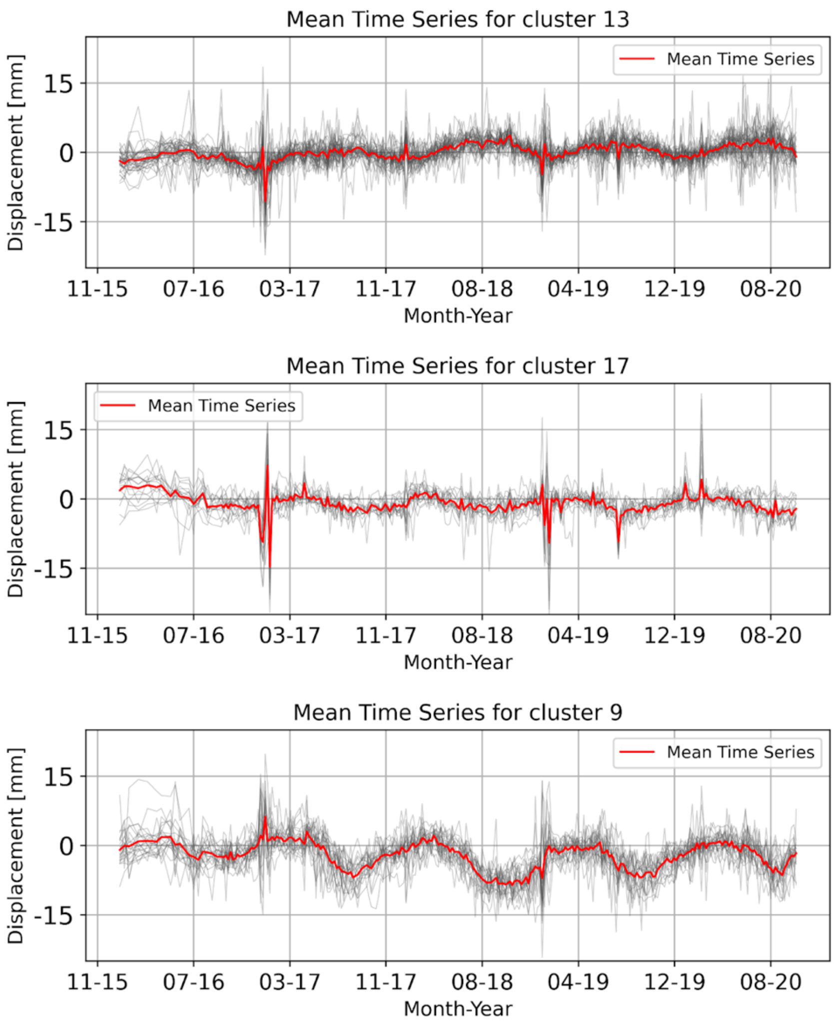

Fig. 6. Mean time series of selected clusters

The mean time series for the same period varies between cluster groups (see. Fig 6) as the clustering is sensitive to the frequent presence of outliers with a similar occurrence in specific epochs.

Table 2. Results of clustering approach

Total number of measurements	1 530 858
Number of identified outliers	31 037 (2.03%)
Number of corrections	**19 588**

An increase in the number of implemented corrections is seen after application of the clustering approach. The higher number of corrections is directly related to the way in which the mean time series correction is designed and implemented. The correction is aimed not only at eliminating phase jumps but also at capturing other types of systematic errors that occur in InSAR time series (e.g., low or partial persistent scatterer coherence).

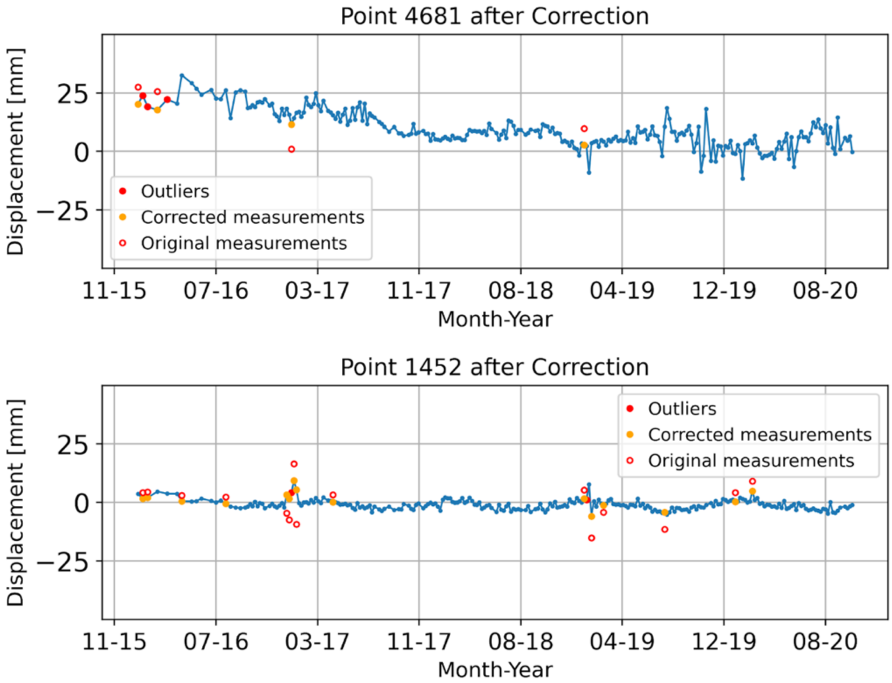

Fig. 7. Time series after application of the clustering approach

SUMMARY

In this work, two different methodologies of outlier detection and error mitigation in InSAR displacement time-series are presented. The individual approach can identify a larger number of outliers, but its ability to correct them is limited due to frequent presence of phase jumps. On the other hand, the clustering approach shows the ability to introduce general corrections for any types of systematic errors. However, exploitation of the mean time series approach, reduces the total number of identified outliers and, outliers which are occurring in epochs that are different from the mean time series epochs remain undetectable.

The proposed approaches are integrable directly into the InSAR processing chains and based on our observations an iterative way of post-processing combining both proposed solutions would presumably provide the most beneficial results.

REFERENCES

Ling Chang and Ramon F. Hanssen, „A Probabilistic Approach for InSAR Time-Series Postprocessing“, IEEE Trans. Geosci. Remote Sens., vol. 54, no. 1, Jan. 2016.

Leijen, F. J. V. (2014), *Persistent Scatterer Interferometry based on geodetic estimation theory*, Delft University of Technology.
URL: *https://doi.org/10.4233/uuid:5dba48d7-ee26-4449-b674-caa8df93e71e*

Leys, C., Ley, C., Klein, O., Bernard, P. & Licata, L. (2013), 'Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median', *Journal of Experimental Social Psychology* pp. 764–766.
URL: *https://doi.org/10.1016/j.jesp.2013.03.013*

Bas van de Kerkhof, Victor Pankratius, Ling Chang, Rob van Swol and Ramon F. Hanssen, „Individual Scatterer Model Learning for Satellite Interferometry“, IEEE Trans. Geosci. Remote Sens., vol. 58, no. 2, Feb. 2020.

Czikhardt, R., Papco, J., Bakon, M., Liscak, P., Ondrejka, P. & Zlocha, M. (2017), 'Ground stability monitoring of undermined and landslide prone areas by means of sentinel-1 multi-temporal insar, case study from slovakia', *Geosciences* 7(3).
URL: *https://www.mdpi.com/2076-3263/7/3/87*