A DATA MINING APPROACH FOR MULTIVARIATE OUTLIER DETECTION IN HETEROGENEOUS 2D POINT CLOUDS: AN APPLICATION TO POST-PROCESSING OF MULTI-TEMPORAL INSAR RESULTS

M. Bakon[1], I. Oliveira[2a], D. Perissin[3], J. Sousa[2b], J. Papco[1]*

[1]Department of Theoretical Geodesy, Slovak University of Technology, Bratislava, Slovakia
[2]UTAD, Vila Real, [a]CITAB, [b]INESC-TEC (formerly INESC Porto), Portugal
[3]School of Civil Engineering, Purdue University, West Lafayette, Indiana, USA
[*]Corresponding author, E-mail: matusbakon@insar.sk

ABSTRACT

Thresholding on coherence is a common practice for identifying the surface scatterers that are less affected by decorrelation noise during post-processing and visualisation of the results from multi-temporal InSAR techniques. Simple selection of the points with coherence greater than a specific value is, however, challenged by the presence of spatial dependence among observations. If the discrepancies in the areas of moderate coherence share similar behaviour, it appears important to take into account their spatial correlation for correct inference. Low coherence areas thus could serve as clear indicators of measurement noise or imperfections in mathematical models. Once exhibiting properties of statistical similarity, they allow for detection of observations that could be considered as outliers and trimmed from the dataset. In this paper we propose an approach based on renowned data mining and exploratory data analysis procedures for mitigating the impact of outlying observations in the final results.

Index Terms— InSAR, data mining, exploratory data analysis, outlier detection, multivariate analysis, DBSCAN, PCA, graph theory, Voronoi diagram, MAD, Jaccard index

1. INTRODUCTION

Multi-temporal InSAR (MTI) technique [1] is successfully applied in measuring of subtle deformations of both natural and man-made objects. The parameters of velocity, height and others, sought as the ultimate MTI estimates, are commonly considered reliable when their ensemble coherence $\in [0, 1]$ is exceeding a certain threshold of, e.g. 0.7 (Fig. 1), and reaches the value of 1. Loss of the coherence is commonly associated with temporal and geometrical decorrelation. Noise from the signal delays caused by the atmospheric disturbances also prevents the interferometric phase from being readable. Beside other reasons for inaccuracies such as sub-pixel positions, sidelobe observations and orbit errors, there are difficulties in resolving non-uniform deformations. Possible scenarios include: non-linear movements such as high-phase gradients (e.g., during landslide activation process or earthquakes), seasonal patterns (e.g., thermal expansion of structures due to temperature changes, dam oscillations related to the water level change) and other displacement-inducing effects, or a combination of more of them. Usually, only the eyes of InSAR experts are searching for the groups of scatterers that are exhibiting similar behaviour, while evaluating their spatial relations and agreement of the estimated parameters within certain surroundings. Experiencing a new era of operational SAR with frequent observations of satellites with enhanced swath coverage (Sentinel-1A), foreseen data boost from constellation missions (Sentinel-1B, TerraSAR-X NG, etc.) and nation-wide monitoring initiatives are making this task more and more complicated. It is therefore of interest to reconsider the practice of imposing simple threshold on ensemble coherence value and to assess its full informative character recognised in a range of thematic mapping applications. Although, lot of advances have been achieved in exploiting low or partially coherent targets [2, 3] all effort in evaluating higher-order products often remains in the hands of end-users, causing common concerns about the reliability of InSAR results by simply looking at the locations of extreme velocities. To limit those concerns and possible misinterpretations, we would like to address the topic of missing concept for finding a statistically significant observations through removing those which appear outlying. In the following, well known statistical procedures, namely Density-based spatial clustering of applications with Noise (DBSCAN), Principal Component Analysis (PCA), Graph Theory Grouping, Voronoi diagram, Mean Absolute Deviation (MAD) and Jaccard index are involved in order to perform outlier detection and removal in MTI results.

TerraSAR-X data were provided by DLR under project ID LAN2833. Sentinel-1 data were provided by ESA under free, full and open data policy adopted for the Copernicus programme. Data have been processed by SARPROZ© using Matlab® and Google Maps™. The work has been supported by the Slovak Grant Agency VEGA under projects No. 1/0714/15 and 1/0462/16 and Portuguese FCT UID/AGR/04033/2013.

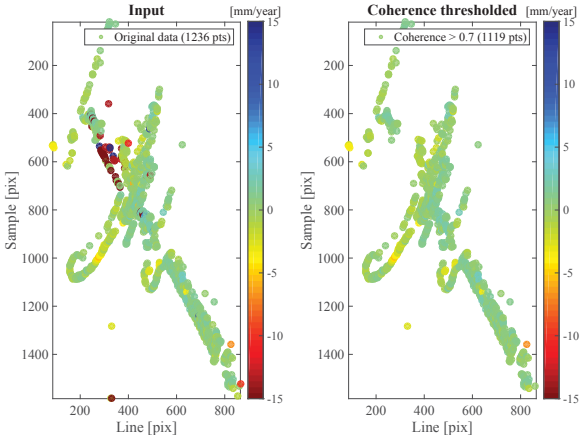

Fig. 1. Velocities before and after imposing a threshold of 0.7 on ensemble coherence value.

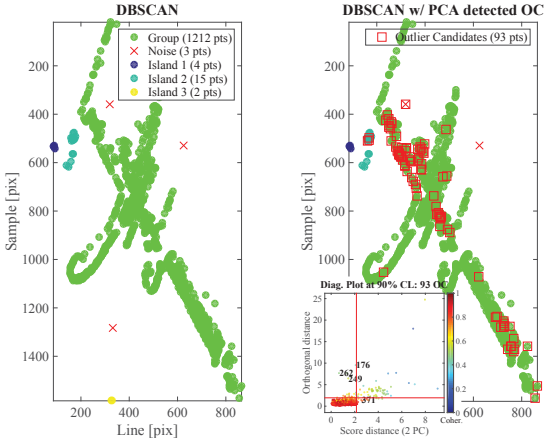

Fig. 2. Clustered data structure subjected to PCA analysis using diagnostic plot.

2. METHODOLOGY

Given a set of persistent scatterers (PS) that have undergone standard MTI processing, we dispose of spatial data with multiple variables declared in every location. These variables define velocity (Fig. 1), height or residual height, their standard deviations, coherence and other parameters that are part of the estimation process assigned to each PS point. The problem of discovering spatial relations among all variables then becomes a subject of multivariate analysis. It has to be noted, that for our approach and the current state of its development we are neglecting the post-processing analysis of deformation time-series, thoroughly studied for example in [4]. As for the time-series analysis, long history of equally sampled observations is often required (which is usually not the case of former SAR missions like ERS, ENVISAT, etc.) our focus was aimed at the designing a set of procedures that will, at the first level, eliminate multivariate outliers exploiting PSs' pointwise variables only. Building upon the six respective techniques that are described in next sections, our implementation is applicable regardless of specifics linked with every location in the world or purpose of its monitoring or could serve as a base for doing so. A description of methods is accompanied by the results from monitoring of one of the biggest Europe's waterworks, Gabcikovo-Nagymaros (part Cunovo), through 52 TerraSAR-X images spanning years 2011 - 2013.

2.1. Density-based spatial clustering of applications with Noise (DBSCAN)

The estimation of PS parameters, such as velocities (Fig. 1), is performed within system of equation utilising several types of networks. The connections (i.e., arcs), formed by using all the PSs in a study area, are often intentionally adjacent in order to decrease the impact of aforementioned systematic errors. In other words, systematic errors are causing groups of PSs separated by a large distance to behave in a different manner from that of being anticipated in mathematical models amongst majority of PS points. One of many possible approaches to search for a location-driven outliers are clustering techniques as a subdivision of data mining methods. For demonstration purposes, we stick with a DBSCAN [5] algorithm, mainly because of its mathematical simplicity and the ability to find clusters of arbitrary sizes and shapes together with detection of noise. As the whole process operates in 2D space, image coordinates are used to define the location of points. Selection of the radius (Eps) in which the points are considered reachable is based on using distance graph employing pairs of PSs from connections network. By plotting distances in ascending order it is possible to detect the knee of such a graph and expose distances that are deviating. Second input parameter required for a DBSCAN is a minimum number of points ($MinPts$) needed to form a cluster. By DBSCAN we can retrieve clusters of points classified as (Fig. 2): GROUP, the core points of the dataset; ISLANDS, to be evaluated in the next steps; NOISE, points to be discarded immediately when confirmed as outliers by PCA or kept when their coherence is greater than a selected minimum. Finally, we get a set of points with a clustered structure (Fig. 2), an advantageous one, providing that points allocated within the same cluster will exhibit the same behaviour, which will be analysed further.

2.2. Principal Component Analysis (PCA)

One of the statistical tools capable of exposing multivariate outliers is Principal Component Analysis (PCA). By mapping a high dimensional space into a low dimensional space, while retaining the maximum variability in terms of the variance-covariance structure, test limits could be applied in order to

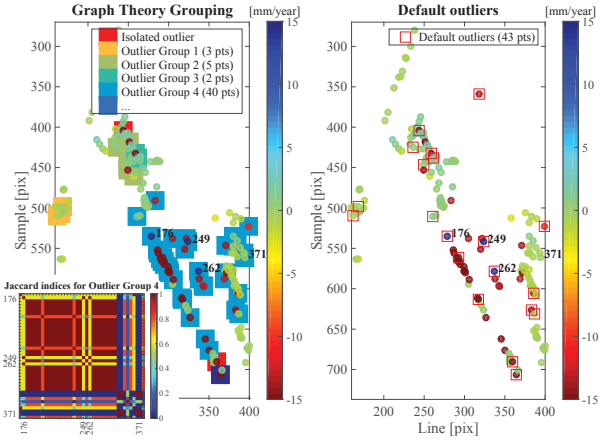

Fig. 3. Zoom at the outliers group creation, matrix of Jaccard indices and propagation to default outliers.

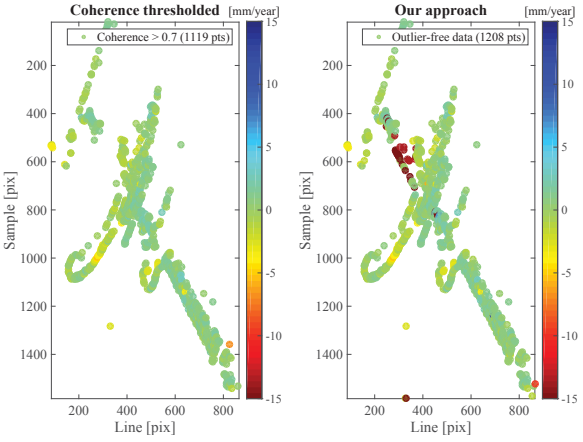

Fig. 4. Final post-analysis results.

determine failing observations. The goal is to detect outlying measurements and, upon PS points classified in the previous DBSCAN step, track the behaviour of all the points allocated within the same cluster in order to answer the question "Why the point has become an outlier?": Is it because the residual height in some location varies significantly in comparison to the points within allocated cluster? Is it because the displacement values varies significantly in comparison to the core points of the dataset? To distinguish between the regular observations and the outliers for multivariate data, we construct a diagnostic plot (Fig. 2) according to [6]. For each observation coloured proportionally to ensemble coherence value, there are score distances and orthogonal distances, to the PCA subspace. To classify the observations, two cutoff lines are drawn, representing statistical confidence level (CL) of, e.g. 90%, for the points being outlier candidates (OC) (Fig. 2) when they are exceeding those limits.

2.3. Decision-making process

Before the final analysis is performed, the outlier candidates (OC) detected by PCA (only), are separated on the basis of finding connected components of undirected graph employing graph theory. Thoroughly explained, those outliers that have not been excluded as noise in the first round of DBSCAN and PCA are separated in the following way: i) outliers without any outlier in the neighbourhood are considered isolated and, ii) outliers with the presence of other outliers in the neighbourhood form an outlier group (Fig. 3). Only the points that share Voronoi adjacency cells and are within Eps radius are taken as neighbouring. This way, tough unfortunately computationally less efficient, the points could have different amount of "natural" neighbours in close surrounding, to the limit of the distance (Eps) that implies boundary of noise. The outlier structure created upon such principles (Fig. 3) is then passed to the algorithm and its performance is tested within allocated DBSCAN clusters. The variables for isolated outliers and grouped outliers are then flagged in accordance with Mean Absolute Deviation (MAD) [7] of their non-outlying neighbours or non-outlying points remaining from the whole cluster, respectively. For isolated outliers, if discrepancy in terms of exceeding the rejection criterion of 2, 2.5 or 3 [7] in any variable haven't been found and the coherence of inspected point is greater than a selected minimum, the point is kept in the dataset. Otherwise, it is recognised as the default outlier (Fig. 3). For outliers in groups, Jaccard similarity coefficient for each pair inside the group is computed (Fig. 3), evaluating final sample sets - the vectors of zeros and ones for each variable in every point, 1 as a flag for variable exceeding rejection criterion again, meaning that point is breaking the rules of inspected outlier group in some parameter. Point pairs with Jaccard index lower than a certain threshold (e.g., 0.6) are seized for the key step of the whole process: minimum coherence value having a final word in deciding whether the point has a bad coherence (lower than a chosen minimum) and would be given away from the dataset, or the coherence is too good (higher than a minimum) for the point to be excluded - such point will be kept in a dataset as prone to be problematic and/or ambiguously integrated to the corresponding outlier group. One could easily grasp the whole concept by following the behaviour of labelled points in Figures 2 and 3. Beside the extreme cases (e.g. when there are more outliers in the cluster than non-outlying ones and MAD statistics would be biased, or the number of points to keep is lower than a minimum number of points needed to form a cluster $MinPts$, etc.), when whole clusters are indicated as default outliers, this key process is responsible for preserving groups of scatterers with similar statistical nature (Fig. 4), even though their coherence is weakened and by the rules of standard thresholding procedure they would be discarded.

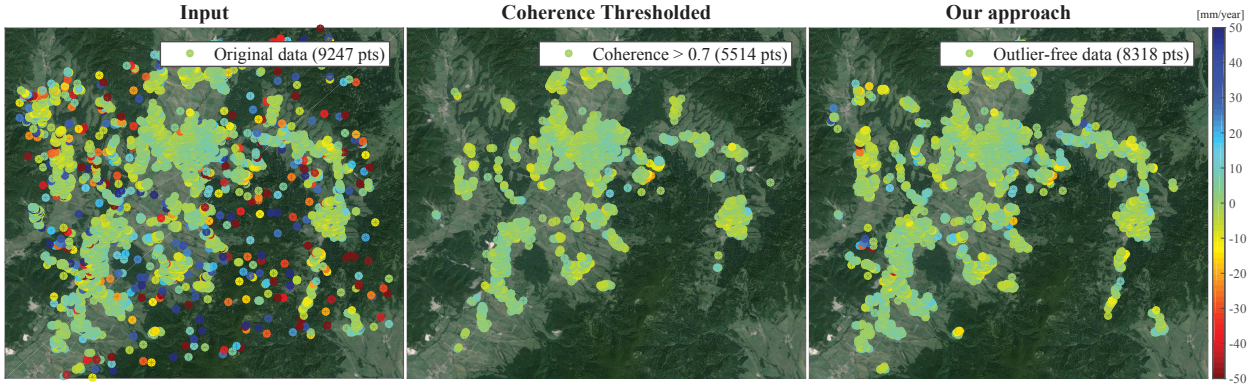

Fig. 5. Deformation map over area of active landslides in Prievidza, Slovakia.

3. RESULTS ON SENTINEL-1 DATA

Relying on the C-band observations of Sentinel-1, we would like to present the performance of our algorithm by the results obtained over the area of active landslides in Prievidza, Slovakia. Performing PSInSAR processing (Fig. 5) on 32 images from Interferometric Wide (IW) swath mode acquired along ascending track 175, we have identified 9247 scatterers. By imposing a standard threshold of 0.7 on ensemble coherence value, this amount decreased dramatically to 5514 PS points. However, applying post-processing analysis following the procedures proposed within this research we get 8318 scatterers, more than half of the amount of standard PS points (Fig. 5), that are exhibiting spatial and/or statistical dependency among themselves as described in Sec. 2. Thanks to it, the problematic areas could be assessed in more detail, as the deformation phenomena of these localities tend to be diminished by the standard thresholding procedure (Fig. 5).

4. CONCLUSION AND FUTURE WORK

This paper presents a novel workflow for detecting outliers in post-processing of Multi-temporal InSAR (MTI) results. Tested upon Sentinel-1 data, this approach has shown its potential in increasing point densities by a half of the total amount of standard PS points. While preserving spatial dependency among low coherent areas, the main benefit of this methodology are enhanced details visible in deformation maps, highlighted zones of scatterers that would require deeper investigation in terms of systematic errors mitigation, replacement of the frequency band, correction of the processing procedures, and others. The platform will help to interpret higher-order MTI products by removing statistically insignificant observations, conserving the full informative character of the whole range of an ensemble coherence value. The multidisciplinary character of the proposed approach allows for modifying the procedures in order to operate with any heterogeneous 2D point clouds of arbitrary high-dimensional variables. Beside the time-series analysis the state-of-art of this approach should focus on making this procedures three-dimensional, fully automatic and capable of predictions.

5. REFERENCES

[1] A. Ferretti, C. Prati, and F. Rocca, "Permanent Scatterers in SAR Interferometry," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 39, no. 1, pp. 8–20, Jan 2001.

[2] A. Ferretti, A. Fumagalli, F. Novali, C. Prati, F. Rocca, and A. Rucci, "A New Algorithm for Processing Interferometric Data-Stacks: SqueeSAR," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 49, no. 9, pp. 3460–3470, Sept 2011.

[3] D. Perissin and T. Wang, "Repeat-Pass SAR Interferometry with Partially Coherent Targets," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 50, no. 1, pp. 271–280, Jan 2012.

[4] L. Chang and R.F. Hanssen, "A Probabilistic Approach for InSAR Time-series Postprocessing," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 54, no. 1, pp. 421–430, Jan 2016.

[5] M. Daszykowski, B. Walczak, and D.L. Massart, "Looking for natural patterns in data – Part 1. Density-based approach," *Chemometrics and Intelligent Laboratory Systems*, vol. 56, no. 2, pp. 83–92, 2001.

[6] M. Hubert, P. J. Rousseeuw, and K. Vanden Branden, "ROBPCA: A New Approach to Robust Principal Component Analysis," *Technometrics*, vol. 47, no. 1, pp. 64–79, Feb. 2005.

[7] Ch. Leys, Ch. Ley, O. Klein, P. Bernard, and L. Licata, "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median," *Journal of Experimental Social Psychology*, vol. 49, no. 4, pp. 764–766, July 2013.