Spatial outlier detection on discrete GNSS velocity fields using robust Mahalanobis-distance-based unsupervised classification

Bálint Magyar1,2 · Ambrus Kenyeres1 · Sándor Tóth1,2 · István Hajdu1 · Roland Horváth1

Received: 10 July 2020 / Accepted: 15 August 2022 / Published online: 24 September 2022
© The Author(s) 2022

Abstract
GNSS velocity field filtering can be identified as a multidimensional unsupervised spatial outlier detection problem. To detect and classify the spatial outliers, we jointly interpret the horizontal and vertical velocity fields with the related standard deviations. We also describe the applied feature engineering process, which represents the underlying problem better than the initial attributes. According to this, we discuss the utilized algorithms and techniques, like the spatial- and non-spatial mapping functions, the k-nearest neighborhood (kNN) technique to retrieve the local environment of each GNSS station, as well as the principal component analysis (PCA) as a dimensionality reduction technique. We also assume that regular velocity field samples containing no outliers come from an approximate multivariate normal distribution (MVN) at the local scale. Regarding this, we evaluate the corresponding sample-wise distance related to model distribution, namely the Mahalanobis distance, with the estimation of the robust covariance matrix derived by the minimum covariant determinant (MCD) algorithm. Subsequently, we introduce the applied binary classification on the values of the derived robust Mahalanobis distances (RMD) which follows the \( \chi^2 \) distribution. We also present three cases of artificially generated, pre-labeled synthetic velocity field datasets to perform cross-validation and comparison of the proposed RMD approach to other classification techniques. According to this, we found that \( k = 12 \) yields > 95% classification accuracy. While the compared methods have a mean classification accuracy of 96.2–99.8%, the advantage of the RMD approach is that it does not require pre-defined labels to indicate regular and outlier samples. We also demonstrate the proposed RMD based filtering process on a real dataset of the EUREF Permanent Network Densification velocity products. The RMD-based approach has been integrated into the EPN Densification as a quality checking algorithm. According to this, we also introduce a co-developed and regularly updated interactive webpage to disseminate the corresponding results.

Keywords EPN Densification · GNSS · Spatial outlier detection · PCA · Mahalanobis distance

Introduction
Dense national and regional GNSS networks provide an outstanding opportunity to detect the kinematics of recent plate motions and crustal deformations (Caporali et al. 2009; Nocquet 2012; Devoti et al. 2017; Kenyeres et al. 2019). The removal of outlier velocity vectors that are apparently different from the local neighborhood due to possible anthropogenic effects or monumentation issues is vital to developing reliable geophysical interpretation. Briefly depicting the variety of available solutions, several approaches exist from the Euler-parameter estimation-related \( \chi^2 \) and F-ratio tests proposed by Stein, Gordon (1984) and Nocquet et al. (2001) through the frame-invariant approach of Xu et al. (2000) to the strain rate pattern-derived solution in Araszkiewicz et al. (2016). The scope of the current research is to find a suitable classification approach that is capable of detecting outliers of GNSS velocity data to support the quality control of the EPN densification products (Kenyeres et al. 2019).

Outlier detection as a classification problem is clearly not trivial, and it significantly depends on the selection approach as described in the classification frameworks of Zhang et al. (2007) and Boukerche et al. (2020). To obtain
a suitable classification, we propose an outlier detection procedure that considers GNSS velocity data only, interpreted within the well-defined classification framework as proposed by Zhang et al. (2007).

We introduce a solution to perform spatial outlier detection with multiple non-spatial attributes. This includes the characterization of the input data and its related attributes and presents the mapping functions between spatial and non-spatial attributes proposed by Lu et al. (2003a, 2003b). We present a feature engineering procedure and related classification approach applied to GNSS data to detect and remove outliers. The proposed solution also delivers the feature extraction process, which transforms the initial attributes into features and presents the related classification algorithm. We highlight the outlier detection capability of the proposed solution compared to existing classification algorithms on an arbitrary generated synthetic dataset. In addition, we also demonstrate the proposed procedure on a real dataset of EPN Densification (Kenyeres et al. 2019).

Problem statement

The characterization of the outlier detection of GNSS velocity field data is twofold. On the one hand, the research scope needs to be investigated as proposed by Zhang et al. (2007). The initial data type, the characteristics of the possible outliers, and the detection approach shall be clarified.

According to this framework, the initial data may be considered as spatially referenced samples with multiple non-spatial attributes that may contain multivariate local outliers, which behave as inconsistent compared to their neighbors (Shekhar et al. 2003; Lu et al. 2003b; Zhang et al. 2007). The input data cannot be seen as a pre-labeled dataset because it does not have any pre-defined attribute, which attribute indicates if the corresponding sample is regular (inlier) or outlier, consequently, the unsupervised learning approach should be applied. We assume that our non-outlier samples are derived from a multivariate normal distribution (MVN) and that the corresponding Mahalanobis distances follow a \( \chi^2_q \) distribution with \( q \) degrees of freedom. This problem can be modeled by parametric methods and outliers which are not derived from the model can be rejected.

On the other hand, it is important to specify the procedure of the classification and describe the processing chain of its input features. Generally, this refers to feature engineering, which consists not only of data preparation but also of creating features that better represent the underlying problem. In the following section, we propose a solution that indicates the main steps of the feature engineering described above and provides the outlier filtered GNSS velocity field.

Proposed solution

As we discussed before, feature engineering is a key aspect of creating processes and features which support outlier detection. To describe such a processing chain, we review the main input attributes, the corresponding mapping functions and the detailed evaluation procedure to obtain the input features of the classification algorithm, called feature extraction. Then we introduce the implemented classification algorithm for outlier detection.

Dataset and pre-processing

We present here the investigated attributes and describe the a priori data-related pre-processing steps. The input dataset includes GNSS station velocities, characterized by spatial and non-spatial attributes. These attributes describe the estimated velocity field expressed in the European Terrestrial Reference Frame 2000 (ETRF2000) with local tangent plate coordinates (LTP). These attributes are as follows:

- **Spatial Point**: represented for a GNSS station as \( x = (\varphi, \lambda) \), such that latitudes \( \varphi \in \Phi \) and longitudes \( \lambda \in \Lambda \) are expressed in the ETRF2000 reference frame and sets of \( \Phi \) and \( \Lambda \) determined by the available dataset. Thus, the set of spatial points can be described as \( X = \{ x_i : 0 \leq i \leq m - 1 \wedge m \in N \} \), where \( m \) equals the number of GNSS stations in the input dataset.

- **A priori attributes**: parameter set that is used in the a priori filtering. These are the length of the time series expressed in weeks, the serial solution number of the corresponding velocity components and the ID of the GNSS station.

- **Velocity related attributes**: represents the northing, easting and up (NEU) velocity components expressed in the local tangent plane (LTP) for the corresponding GNSS station \( x \), such that \( v(x) = (v_N, v_E, v_U) \in R^3 \) for each \( x \) in \( X \). The velocity component related standard deviations, namely \( \sigma(x) = (\sigma_N, \sigma_E, \sigma_U) \in R^3 \), are derived from the full variance–covariance matrix of the CATREF solution (Altamimi et al. 2007). The velocity and uncertainty components are in \( \text{mm/yr} \) unit and expressed in the ETRF2000 frame.

As data pre-processing procedure, we perform data cleaning and filtering based on the a priori information and the corresponding a priori attributes. We keep stations only with time series length at least 3 years as Legrand et al. (2012) and Kenyeres et al. (2019) suggest. The resulting pre-processed dataset will act as the input...
for the feature engineering procedure. The related initial attributes are the velocity components and the corresponding standard deviations for each \( x \) in the set of \( X \).

### Mapping functions

The different functions that describe the spatial attribute mapping to the non-spatial attributes will be defined here. These mapping functions can be interpreted as:

- **Attribute function**: defined as \( e: X \rightarrow \mathbb{R}^d \), where the attribute function \( e \) maps the spatial dataset \( X \) to \( \mathbb{R}^d \) dimensional attribute space (Lu et al. 2003a, 2003b), where \( d \) represents the number of selected six attributes to be investigated. These attributes sets are defined as the velocity and standard deviation attributes registered to each \( x \) in \( X \).

- **Feature function**: defined as \( f: X \rightarrow \mathbb{R}^r \), where the feature function \( f \) maps the spatial dataset \( X \) to a lower dimensional \((d > r)\) feature space. We describe its details in the following Feature extraction subsection.

- **Distance function**: defined as metric \( d_H: X \times X \rightarrow \mathbb{R} \); given \( A, B \in X \), let the corresponding distance between spatial points \( A \) and \( B \) given by the Haversine-metric (Pedregosa et al. 2011; Van Brummelen 2013). Please note that the resulting (great-circle) distances may be affected by a 0.5% error compared to using ellipsoidal metrics.

- **Neighborhoods**: defined as function \( k\text{NN}: X \rightarrow X \), which retrieves the \( k \in N \) nearest neighbors (kNN) of each \( x \) in \( X \) (Lu et al. 2003a) according to the distance metric defined by Distance function (Van Brummelen 2013) using BallTree algorithm (Pedregosa et al. 2011; Mahapatra and Chakraborty 2015). The selection of \( k \in N \) corresponds to the evaluation of selection of the quasi-lowest \( k \) that yields maximum classification accuracy, as we present it in the subsection of Synthetic dataset.

- **Neighborhood function**: defined as function of \( g: X \rightarrow \mathbb{R}^d \), representing a mapping that returns the specific statistical moments (e.g., average or median) of the local spatial environment represented by \( k\text{NN} \) neighbors for each GNSS station for each related attribute (Lu et al. 2003a, 2003b). Both the presented neighborhoods and neighborhood function represents and yields spatial environment of each \( x \) in \( X \).

- **Comparison function**: defined as function of \( h: X \rightarrow \mathbb{R}^d \), which is responsible for implementing the comparison of all features to its local environment, based on the neighborhood function \( (g) \); thus, \( h = e \circ g \) as Lu et al. (2003a, 2003b) suggests. The non-outlier results of the comparison function \( (h(x)) \) are assumed to follow a MVN. Besides, we also assume that the possible outlier samples do not follow MVN and in this context, the outlier detection can be performed via parametric methods.

### Feature extraction

After considering the characteristics of the input dataset and the related mapping functions, we present here the procedure of constructing features that explain the GNSS velocity field from the most informative aspect. This feature extraction requires a complex approach considering the characteristics of the local neighbors as well as indicating the possible outliers in the feature space.

Considering the dataset and the pre-processing steps described in the related subsection, let denote \( X_{\text{CATREF}} = \{ x_i : 0 \leq i \leq m - 1 \land m \in \mathbb{N} \} \), where \( m \) represents the number of samples of the GNSS stations in the CATREF output of the minimum constrained cumulative solution of dense GNSS network measurements (Altamimi et al. 2002; Kenyeres et al. 2019). \( X_{\text{CATREF}} \) also underwent the previously discussed pre-processing step to only consider samples where observation length is 3 years long at least as Blewitt and Lavallée 2002 suggest.

Let us consider \( x_i \), the \( i \)-th element of \( X \), where \( 0 \leq i \leq n - 1 \) and \( n \in \mathbb{N} \), and apply attribute function \( e(x) \), which maps \( x_i \) to the \( \mathbb{R}^d \) attribute space, such that:

\[
y_i = e(x_i) = (y_{i,0}, \cdots, y_{i,j}, \cdots, y_{i,d-1})^T
\]

where \( y_i \) represents the feature vector of the corresponding \( i \)-th \( x \), such that the \( y_j \) component represents the \( j \)-th attribute described in the set of velocity components and standard deviations. Applying (1) for all \( x \) in, we write:

\[
e(X) = (e(x_0), \ldots, e(x_i), \ldots, e(x_{n-1}))^T = (y_g) \in \mathbb{R}^{n \times d},
\]

where the matrix indices are \( 0 \leq i \leq n - 1 \) for rows and \( 0 \leq j \leq d - 1 \) for columns.

To implement the median algorithm proposed by Lu et al. (2003a, 2003b), let us denote:

\[
g_j(x_i) = \text{median}\{ e_j(x) : x \in k\text{NN}(x_i) \}
\]

that is, the neighborhood function \( g_j(x_i) \) retrieves the median of the \( j \)-th features (2) of all \( x \) in the local spatial set. This set is defined by the neighborhood mapping function \( k\text{NN}(x_i) \), representing the \( k \)-nearest neighborhood of the \( i \)-th \( x \) in \( X \) in the Haversine metric.

To obtain results that highlight the effects of the possible outliers, let us consider the evaluation of the comparison function, coinciding with the difference of the attribute (2) and neighborhood functions (Lu et al. 2003a, 2003b) (3):

\[
h(x) = e(x) - g(x)
\]
where $h(x)$ represents the non-spatial feature of each $x$ spatial point in $X$ compared to its local environment. Reasonably, the possible outliers may imply higher $h(x)$ scores compared to the regular/non-outlier samples.

To minimize the effect of the outliers on the feature mean, we perform feature-wise robust centering technique to obtain a quasi-standardized form of (4) via:

$$Y_{ctr} = h(x) - \text{median}(h(x))$$

(5)

where data centering is performed by feature-wise subtraction of feature median, and related scaling is neglected (Pedregosa et al. 2011; Raghav et al. 2020).

Furthermore, we assume that the velocity outliers are due to station monumentation weaknesses and may also be related to data quality (multipath, antenna problems, external noise) issues. None of these error sources are correlated; they contribute to the results independently of each other. The combination process (Kenyeres et al. 2019) may generate a low mathematical correlation between them, but we consider it well negligible compared to the independent station-specific biases.

To remove the correlation between features, we implement an orthogonal linear transformation called Principal Component Analysis (PCA) using Singular Value Decomposition (SVD). In addition to revealing the internal structure of the dataset while finding directions (principal components) of its maximum variances, PCA is also capable of projecting a high-dimensional dataset to a lower-dimensional subspace while keeping most of the information (Jolliffe 2002; Wall et al. 2002; Salem and Hussein 2019).

The implementation of PCA is directly linked to SVD if the related principal components are derived from the covariance matrix of $Y_{ctr}$ (Wall et al. 2002). Let us consider the centered feature matrix $Y_{ctr}$ defined in (5), then define the empirical sample covariance matrix $C \in R^{d \times d}$ of $Y_{ctr}$, which is proportional to the $(1/(n-1)) \cdot Y_{ctr}^T Y_{ctr}$ symmetric matrix. We diagonalize $C$ as $C = W \Lambda W^T$, where $W \in R^{d \times d}$ is a matrix built up by eigenvectors of $C$, such that $W_j$ is the $j$-th eigenvector of the empirical sample covariance matrix (Jolliffe 2002). As well as the diagonal matrix $\Lambda \in R^{d \times d}$, where the diagonal elements $\lambda_j$ of $\Lambda$ are in descending order are the $j$-th eigenvalues of $C$. These eigenvectors can be seen as the principal directions or axes (principal components) of the data, and projections of the data onto these principal axes return the principal component scores, i.e., the transformed variables of the data (Jolliffe 2002; Wall et al. 2002, Salem and Hussein 2019):

$$T_d = Y_{ctr} W$$

(6)

where $T_d \in R^{d \times d}$ represents the resulting principal component scores.

In addition, let us denote the singular value decomposition of $Y_{ctr}$:

$$Y_{ctr} = U \Sigma V^T$$

(7)

where $U \in R^{d \times r}$ can be seen as the left singular vectors and $\Sigma \in R^{r \times r}$ as the rectangular diagonal matrix of the corresponding singular values, while $V \in R^{r \times r}$ as the right singular vectors of $Y_{ctr}$. Full SVD components are evaluated by the divide-and-conquer approach in the LAPACK algorithm (Demmel 1989; Jones et al. 2001; Pedregosa et al. 2011). As linear algebra theorem states, the right singular vectors correspond to the $W$ eigenvectors of matrix $Y_{ctr}^T Y_{ctr}$ (Wall et al. 2002; Datta, 2004). This lead us to substitute (7) into (6), which yields to:

$$T_d = Y_{ctr} \cdot W = U \cdot \Sigma \cdot W^T \cdot W = U \cdot \Sigma$$

(8)

where $T_d \in R^{d \times r}$ equals to the principal component score matrix described by (6).

As Jolliffe (2002) indicates, the variance of the $j$-th principal components can be expressed through the related $j$-th eigenvalue of $C$. Hence the singular values of $Y_{ctr}$ can be represented as the non-negative square roots of eigenvalues of $Y_{ctr}^T Y_{ctr}$ (Datta 2004; Wall et al. 2002), and considering that the eigenvalues of $C$ are proportional to eigenvalues of $Y_{ctr}^T Y_{ctr}$ (Jolliffe, 2002; Salem and Hussein, 2019) allow us to discuss the variance / variance ratio of the SVD derived principal components.

Since (8) involves all principal components, therefore the related explained variance ratio yields $l$ (Jolliffe 2002; Salem and Hussein, 2019). To perform dimensionality reduction and obtain the truncated solution of (8), consider $s = \Sigma = (s_0, \ldots, s_r, \ldots, s_{d-1}) \in R^d$, the vector of singular values, which are the main diagonal elements of $\Sigma$. According to this, we evaluate the first $s$, singular values (Salem and Hussein, 2019) where:

$$\frac{\sum_{j=0}^{r-1} \frac{s_j^2}{j}}{\sum_{j=0}^{d-1} \frac{s_j^2}{j}} \geq 0.98$$

(9)

constraint is satisfied by the first $r < d$ (9). It aims to specify the first $r$ singular values with the corresponding principal components, whose related cumulative ratio of explained variance interprets at least 98% of the full variance according to the investigated feature space.

Thus, the truncated format of (8) with the first $r$ principal components can be generalized as:

$$T_r = U_r \cdot \Sigma_r$$

(10)

where $T_r \in R^{r \times r}$ is the truncated principal score matrix, $U_r \in R^{d \times r}$ is the corresponding left singular vector and $\Sigma_r \in R^{r \times r}$ represents the rectangular diagonal matrix of the
first $r$ largest singular values (Jolliffe 2002; Pedregosa et al. 2011; Wall et al. 2002). The resulting truncated and uncorrelated features describe the investigated features in the most informative aspect.

To perform spatial outlier filtering, let us consider mapping the set of spatial point $x$ in $X$ to the non-spatial feature space described in (10), i.e.:

$$f(x_i) = (f_0(x_i), ..., f_j(x_i), ..., f_{r-1}(x_i))^T, \quad (11)$$

where the feature function $f(x_j)$ maps the $i$-th spatial point to the feature space described in (10) and $f_j$ retrieves the $j$-th feature regarding the indices of $0 \leq i \leq n - 1$ and $0 \leq j \leq r - 1$.

**Classification algorithm**

In this subsection, we present the classification algorithm and the related model of the applied parametric method. According to this we denote the squared Mahalanobis distance of the features described in (11), as Lu et al. (2003a, b), Zhang et al. (2007), Leys et al. (2018) and Etherington (2019) suggest, as:

$$d^2_{\text{Mahalanobis}}(x) = (h(x) - \mu_{\text{MCD}})^T \cdot \Sigma^{-1}_{\text{MCD}} \cdot (h(x) - \mu_{\text{MCD}}) \quad (12)$$

where $\mu_{\text{MCD}} \in \mathbb{R}^r$ and $\Sigma_{\text{MCD}} \in \mathbb{R}^{r \times r}$ are respectively the robust mean vector and the robust covariance matrix estimated by the minimum covariant determinant (MCD) algorithm. Implementation of the MCD-derived robust mean vector and robust covariance matrix is requisite due to the presence of possible outliers, which can significantly affect the sample mean and covariance as well (Rousseeuw and Driessen 1999, Pedregosa et al. 2011; Zhang et al. 2007).

Considering that the squared Mahalanobis distance described by (12) follows a $\chi^2$ distribution with $r$ degrees of freedom, namely, $\chi^2_r (\text{Lu et al. 2003a, b; Zhang et al. 2007})$, we denote the corresponding cumulative density function (CDF) of $d^2_{\text{Mahalanobis}}(x) \sim \chi^2_r$ as $CDF_{\chi^2_r}(r, d^2_{\text{Mahalanobis}}(x))$.

To perform the hypothesis testing, let us consider the null hypothesis, which is that the sample is derived from the test distribution; in our case, this is $\chi^2_r$. Thus, to detect the statistically significant samples, with an observation probability at least as extreme in the $\chi^2_r$ distribution, let us consider $p$-values below a pre-defined significance level ($\alpha = 1 e - 8$). Samples with low $p$-values imply statistical significance, which allows us to reject the null hypothesis within a given significance level. To obtain such $p$-values ($\varphi(x)$), we subtract the $CDF_{\chi^2_r}$ from 1:

$$\varphi(x) = 1 - CDF_{\chi^2_r}(p, d^2_{\text{Mahalanobis}}(x)) \quad (13)$$

where the resulting $p$-value can be utilized during the classification. Thus, the generalized classification function with a given significance level ($\alpha$) can be expressed with (13) as a hypothesis test:

$$clf(x) = \begin{cases} -1 & \text{if } \varphi(x) \leq \alpha \\ 0 & \text{else} \end{cases} \quad (14)$$

where the statistically significant samples with $\varphi(x) \leq \alpha$ indicate the spatial outliers with -1 flags, which do not follow the test distribution and act as outliers. According to the regular samples, Eq. (14) passes with zero flags.

**Application of the proposed solution to a synthetic dataset**

We demonstrate the capabilities of the outlier detection procedure presented above, applying it to a synthetic dataset. The characteristics of the artificially generated velocity field and the evaluation strategy are shown, together with the analysis of the efficiency of the proposed filtering process is discussed.

**Synthetic dataset**

The motivation to present such an artificially created and pre-labeled dataset is to compare our proposed procedure to other methods and estimate its prediction accuracy.

The generated synthetic velocity field is built up from a regular part contaminated with systematic biases and random errors. The regular part is expected to be conserved in the outlier detection procedure, while the errors are supposed to be marked as outliers. Regarding the spatial density, 3 sample sets were generated, sparse, normal and dense cases, each representing fraction of the initial synthetic dataset via stratified K-fold split approach while preserving sample percentage in each pre-defined class (Pedregosa et al., 2011). The presented synthetic datasets are available in Appendix I of the Supplementary Information and visualized in Fig. 1.

**Parameter test**

As we highlighted in the previous Mapping functions subsection, it is requisite to determine the $k$ number of stations to be involved in the $k$NN Neighborhoods mapping function. Here we demonstrate the strategy of evaluation of such $k \in \mathbb{N}$. To obtain the quasi-lowest $k$ that yields maximum or at least 95% classification accuracy, we executed the proposed RMD-based outlier detection over a list of parameter settings for the above three cases. The evaluated accuracy corresponds to the value of the accuracy score implemented.
by Pedregosa et al. (2011). The related results are displayed in Fig. 2:

As Fig. 2, implies that the selection of k = 12 can be seen as a reasonable choice. Hence its classification accuracy yields > 95% at k = 12 in all test cases. On the contrary, selection of k < 12 yields to a steep decrease in accuracy score, while selection k > 12 yields approximately 98.5% classification accuracy at the cost of a larger kNN size.

Comparison between the proposed solution and existing solutions

To rank the performance of the proposed RMD-based outlier filtering approach, we implemented and evaluated several existing supervised classification algorithms (Amancio et al. 2014; Uddin et al. 2019; Varoquaux et al. 2020). Using the synthetic dataset we generated the accuracy scores of the presented cases and compared them to our RMD-based approach. The evaluated accuracy corresponds to the accuracy score implemented by Pedregosa et al. (2011), as same as discussed in the previous subsection. The corresponding results of the RMD-based approach compared to the supervised classification approaches are summarized in Table 1.

Advantages and disadvantages

In this subsection, we discuss the advantages and disadvantages of the proposed solution and the demonstrated supervised techniques as well. These empirical considerations are based on the comparison results, which are presented in the previous subsection.

The theoretical details and limitations of the demonstrated supervised classification approaches are presented in Demsar (2006), Amancio et al. (2014) and Uddin et al. (2019). As Amancio et al. (2014) discuss, the kNN classifier mostly over-performs the other investigated methods, even in higher dimensions. Related to this, Uddin et al. (2019) highlight that Decision Tree (DT), Random Forest (RF) and Support Vector Machine (SVM) classifiers can also have superior prediction accuracy. This corresponds to our empirical results as well, namely, kNN has one of the largest...
accuracy scores, in parallel with DT and RF classifiers. On the other hand, it is important to note that these supervised approaches may require parameter optimization via parameter tuning using grid-search or other parameter optimization techniques. Most importantly, each demonstrated supervised technique requires pre-labeled data for the test dataset of the learning process.

Concerning the advantages and disadvantages of the proposed RMD outlier filtering approach, we emphasize that it does not require any pre-defined labels. Moreover, due to the application of a parametric method and the presented feature engineering procedure, it is capable of detecting outliers with multiple attributes. The classification accuracy score of the RMD-based approach is comparable to KNN, DT and RF classifiers in the three investigated cases. As of its disadvantages are concerned we must mention that using \( \chi^2 \) may be too conservative approach, it may appoint many inlier as outliers, as Hardin and Rocke (2005) discuss. To overcome this problem, we interpreted the outliers in \( \alpha = 1e^{-8} \) significance level. Besides, the proposed solution is only designed for spatial outlier detection, namely to define inlier and outlier clusters from the non-spatial features. Its main advantage is that it does not require pre-defined labels in the input dataset. Also, the comparison of the proposed unsupervised RMD-based approach to the described supervised methods is possible due to the synthetic dataset (Fig. 1) because it has pre-defined labels.

### Application of the proposed solution to real dataset

As a further innovation, we analyzed the current multi-year velocity result of the EPN Densification dataset up to GPS week 2050 to demonstrate the capabilities of the introduced RMD-based classification algorithm on a real dataset as well. The input dataset is available in the Appendix II of the Supplementary Information. This European continental-scale multi-year solution includes 31 national network position and velocity solutions of 26 European GNSS Analysis Centers (ACs), combined with CATREF software (Altamimi et al. 2007) following the processing chain described in Kenyeres et al. (2019). Both horizontal and vertical (UP) velocity field solutions are expressed in the European Terrestrial Reference Frame 2000 (ETRF2000) (Altamimi et al. 2017).

According to this, we illustrate the RMD-based outlier detection technique with the tectonic unit of the Apennine Terrane (Serpelloni et al. 2007, Carminati et al. 2010, Palano et al. 2014 and Minissale et al. 2019). Related tectonic settings are presented and described, as well as are available in vector format in Appendix III of the Supplementary Information. The related horizontal velocity field is visualized in Fig. 3.

After the steps described in the subsection of dataset and preprocessing above, we perform spatial mapping to map the spatial dataset to the attribute space as expressed in (2). To aim for the detection of spatial outliers in the (11) resulting feature space, we implement the median algorithm proposed by Lu et al. (2003a, b), as we discuss in (3) – (4). As the next step, to achieve the principal component scores, we implemented (5)–(8), as Fig. 4 illustrates.

### Table 1

Summary of comparison of classification accuracy scores of the test cases. Abbreviations stands for: RMD – Robust Mahalanobis Distance (proposed solution), KNN – Nearest Neighborhood, Linear SVM – Linear Support Vector Machine, RBF SVM – Radial Basis Function Support Vector Machine, DT – Decision Tree, RF – Random Forest, NNET – Neural Network, ADA – AdaBoost, NB – Native Bayes and QDA – Quadratic Discriminant Analysis. Related parameter sets of the supervised classifiers are derived from Varoquaux et al. (2020)

| Case  | RMD   | KNN   | Linear SVM | RBF SVM | DT    | RF    | NNET  | ADA   | NB    | QDA   |
|-------|-------|-------|------------|---------|-------|-------|-------|-------|-------|-------|
| Sparse| 0.9512| 0.9864| 0.9593     | 0.9892  | 0.9946| 0.9946| 0.9702| 0.9973| 0.9566| 0.9593|
| Normal| 0.9199| 0.9952| 0.9762     | 0.9939  | 0.9966| 0.9946| 0.9790| 0.9993| 0.9572| 0.9627|
| Dense | 0.9650| 0.9973| 0.9823     | 0.9969  | 0.9986| 0.9986| 0.9834| 0.9993| 0.9718| 0.9732|
Corresponding dimensionality reduction can be achieved by implementing (9)–(11), i.e., keeping the first $r$ principal components where the resulting cumulative variance ratio explains at least 98% of the full variance, as Fig. 5 illustrates.

Considering the bottom panel of Fig. 4, the results of these features highlight the possible spatial outliers. By assuming that the regular samples follow $MVN$ and calculating the related squared Mahalanobis distances (12) corresponding to $\chi^2_i$ (Fig. 6), a parametric classification model can be obtained (13)-(14).

According to these results, 15 out of 257 samples can be identified as spatial outliers in the presented Apennine Terrane tectonic unit. Related horizontal velocity field is illustrated in Fig. 7.

As Fig. 7 implies, some of the detected outliers are affected by an error in the horizontal components and/or by the vertical component, as well as the extreme values of related standard deviations. The proposed outlier detection procedure is applied to the current EPN densification solution for each tectonic unit described in Appendix III of the Supplementary Information. The results are available on the co-developed and regularly updated webpage of https://epnd.sgo-penc.hu/velociraptor/.

**Conclusions**

We implemented the presented feature engineering process and related RMD-based classifier to solve the outlier detection of GNSS velocity field data. In order to estimate the performance of the proposed RMD-based outlier detection algorithm, we generated a pre-labeled synthetic dataset with different error contributions and with three different spatial resolutions. We performed a parameter test on the number of the $k$-nearest neighbors ($k$) and we found that the classification procedure yields its $> 95\%$ prediction accuracy with the lowest $k$-value as $k = 12$. We also presented a comparison of the unsupervised RMD-based approach to supervised classifiers. This highlighted that about 94.5% prediction mean accuracy of the RMD technique is 1.7–5.3% less than the accuracy of the tested supervised classifiers, which yielded
96.2–99.8% mean accuracy scores. The main difference is that the investigated set of supervised techniques does require pre-defined outlier labels until the proposed RMD-based approach works without any a priori labeling.

As a further innovation, we integrated the proposed RMD-based method into the EPN Densification products workflow as a quality checking algorithm to remove local outliers. We demonstrated the outlier detection procedure for real data related to the tectonic unit of Apennine Terrane step by step. In addition, we installed a webpage that interactively displays the current EPN Densification solution filtered with the proposed RMD technique. Regarding the scope of further development, it is reasonable to test the RMD-based classification on different datasets (Kreemer et al. 2014), as well as to involve more (even time-domain-related) parameters and attributes of the corresponding multivariate GNSS velocity level data, like non-diagonal elements of the full VC matrix.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10291-022-01323-2.

Acknowledgements Prepared with the professional support of the Doctoral Student Scholarship Program of the Co-operative Doctoral Program of the Ministry for Innovation and Technology from the source of the National Research, Development and Innovation Fund. This research was co-financed by the Hungary-Slovakia-Romania-Ukraine (HUSKROUA) ENI CBC 2014–2020 Programme and is directly linked to the HUSKROUA/1702/8.1/0065 GeoSES project.

The current research is supported by the EPOS IP Project founded by the European Commission under Grant Agreement number #676564 and by the Hungarian Research Fund (OTKA) under contract #K109464. Open access funding is provided by Lechner Nonprofit Ltd. The visualization tasks were performed with Cartopy (MET 2010–2015) and Stamen base maps.

Funding Open access funding provided by Lechner Non-profit Ltd.

Data Availability The datasets generated during and/or analyzed during the current study are available in the Supplementary Information.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing,
adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Altamimi Z, Sillard P, Boucher C (2002) ITRF2000: a new release of the international terrestrial reference frame for earth science applications. J Geophys Res 107:B2214. https://doi.org/10.1029/2001JB000561

Altamimi Z, Métivier L, Rebisz P, Rouby H, Collilieux X (2017) The ITRF2014 plate motion model. Geophys J Int 209:1906–1912. https://doi.org/10.1002/2016G005575

Altamimi Z, Sillard P, Boucher C (2007). CATREF software: combination and analysis of terrestrial reference frames. https://doi.org/10.1002/jgrb.50126

Amancio DR, Comin CH, Casanova D, Travieso G, Bruno OM, Rodriques FA (2014) A systematic comparison of supervised classifiers. PLoS ONE 9(4):e94137. https://doi.org/10.1371/journal.pone.0094137

Araszkiewicz A, Figurski M, Jarosinski M (2016) Erroneous GNSS strain rate patterns and their application to investigate the tectonic credibility of gnsn velocities. Acta Geophys 64:1412–1429. https://doi.org/10.1515/acgeo-2016-005751

Blewitt G, Lavallée D (2002) Effect of annual signals on geodetic velocity. J Geophys Res. https://doi.org/10.1029/2001JB000570

Boukerche A, Lining Z, Omar A (2020) Outlier Detection: Methods, Models, and Classification. ACM Comput Surv 53:1–37. https://doi.org/10.1145/3381028

Caporali A et al (2009) Surface kinematics in the Alpine–Carpathian–Dinaric and Balkan region inferred from a new multi-network GPS combination solution. Tectonophysics 474:295–321. https://doi.org/10.1016/j.tecto.2009.04.035

Carminati E, Lustrino M, Cuffaro M, Doglioni C (2010) Tectonics, magmatism and geodynamics of Italy: What we know and what we imagine. J Virtual Explor 36:1–64. https://doi.org/10.3809/jvirtex.2010.00226

Datta BN (2004) Chapter 3 - Some fundamental tools and concepts from numerical linear algebra. In: Datta BN (2004) Numerical Linear Algebra, In: Datta BN (2004) Numerical Methods for Linear Control Systems – Design and Analysis, ISBN: 978-0-12-203590-6, https://doi.org/10.1016/B978-0-12-203590-6.5X0009

Demmel J (1989) LAPACK: a portable linear algebra library for supercomputers. IEEE Control Systems Society Workshop on Computer-Aided Control System Design. https://doi.org/10.1109/CACSD.1989.69824

Dembars J (2006) Statistical Comparisons of Classifiers over Multiple Data Sets. J Mach Learn Res 7:1–30

Devoti R et al (2017) A Combined Velocity Field of the Mediterranean Region. Ann Geophys. https://doi.org/10.4401/ag-7059

Etherington TR (2019) Mahalanobis distances and ecological niche modelling: correcting the chi-squared probability error. Peer J. https://doi.org/10.7717/peerj.6678

Hardin J, Rocke DM (2005) The distribution of robust distances. J Comput Graph Stat 14(4):929–946. https://doi.org/10.1198/106186005X77685

Jolliffe IT (2002) Principal Component Analysis. Springer-Verlag, New York, New York. https://doi.org/10.1007/b98835.978-0-ISBN-387-95442-4

Jones E, Oliphant E, Peterson P (2001). SciPy: Open Source Scientific Tools for Python, Accessed from: https://scipy.org/.

Kenyeres A et al (2019) Regional integration of long-term national dense GNSS network solutions. GPS Solutions. https://doi.org/10.1007/s10291-019-0902-7

Kreemer C, Blewitt G, Klein EC (2014) A geodetic plate motion and Global Strain Rate Model. Geochem Geophys Geosyst 15:3849–3889. https://doi.org/10.1002/2014GC005407

Legrand J, Begeot N, Bruyninx C, Wöppelmann G, Santamaria-Gómez A, Bouin M-N, Altamimi Z (2012) Comparison of Regional and Global GNSS Positions, Velocities and Residual Time Series. In: Kenyon S, Pacino MC, Marti U (eds) Geodesy for Planet Earth. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 95–103. https://doi.org/10.1007/978-3-642-20338-1_12

Leys C, Klein O, Domínguez I, Ley C (2018) Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance. J Exp Soc Psychol 74:150–156. https://doi.org/10.1016/j.expsov.2017.09.011

Lu C-T, Chen D, Kou Y (2003a) Detecting spatial outliers with multiple attributes. IEEE Trans Appl Indus. https://doi.org/10.1109/TAI.2003a.1250179

Lu C-T, Chen D, Kou Y (2003b) Algorithms for spatial outlier detection. In: Proceedings of IEEE International Conference on Data Mining

Mahapatra RP, Chakraborty PS (2015) Comparative Analysis of Nearest Neighbor Query Processing Techniques. Procedia Computer Science 57:1289–1298. https://doi.org/10.1016/j.procs.2015.07.438

MET (2010–2015). Cartopy: a cartographic python library with a matplotlib interface. git@github.com:SciTools/cartopy.git. 2015–2018. 7b2242e.

Minissale A, Donate A, Proces M, Pizzino L, Giannanco S (2019) Systematic review of geochemical data from thermal springs, gas vents and fumaroles of Southern Italy for geothermal favourability mapping. Earth Sci Rev 188:514–535. https://doi.org/10.1016/j.earscirev.2018.09.008

Nocquet J-M (2012) Present-day kinematics of the Mediterranean: A comprehensive overview of GPS results. Tectonophysics 579:220–242. https://doi.org/10.1016/j.tecto.2012.03.037

Nocquet J-M, Calais E, Altamimi Z, Sillard P, Boucher C (2001) Intraplate deformation in western Europe deduced from an analysis of the International Terrestrial Reference Frame 1997 (ITRF97) velocity field. Journal of Geophysical Research: Solid Earth 106(B6):11239–11257. https://doi.org/10.1029/2000JB000410

Palano M (2014) On the present-day crustal stress, strain-rate fields and mantle anisotropy pattern of Italy. Geophys J Int 200:969–985. https://doi.org/10.1016/j.gji.2014.01.024

Pedregosa F et al (2011) Scikit-learn: Machine Learning in Python. J Mach Learn Res 12:12525–12830

Raghav RV, Lemaitre G, Unterthiner T (2020) Compare the effect of different scalers on data with outliers. Accessed 2020 version in scikit-learn webpage: https://scikit-learn.org/stable/auto_example/preprocessing/plot_all_scaling.html

Rousseuw PJ, Driessen KV (1999) A fast algorithm for the minimum covariance determinant estimator. Technometrics 41(3):212–223

Salem N, Hussein S (2019) Data dimensional reduction and principal components analysis. Procedia Computer Science 163:292–299. https://doi.org/10.1016/j.procs.2019.12.111
Serpelloni E, Vannucci G, Pondrelli S, Argnani A, Casula G, Anzidei M, Baldi P, Gasperini P (2007) Kinematics of the Western Africa-Eurasia plate boundary from focal mechanisms and GPS data. Geophysical Journal International 169:1180–1200. https://doi.org/10.1111/j.1365-246X.2007.03367.x
Shekhar S, Lu C-T, Zhang P (2003) A Unified Approach to Detecting Spatial Outliers. GeoInformatica 7:139–166. https://doi.org/10.1023/A:1023455925009
Stein S, Gordon RG (1984) Statistical tests of additional plate boundaries from plate motion inversions. Earth Planet Sci Lett 69:401–412. https://doi.org/10.1016/0012-821X(84)90198-5
Uddin S, Khan A, Hussain ME (2019) Moni MA (2019) Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak 19:281. https://doi.org/10.1186/s12911-019-1004-8
Van Brummelen G (2013) Heavenly Mathematics: The Forgotten Art of Spherical Trigonometry. Princeton University Press. https://doi.org/10.1515/9781400844807
Varoquaux G, Müller A, Grobler J (2020) Classifier comparison. Accessed 2020 version in scikit-learn webpage at: https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html
Wall M, Rechtsteiner A, Luis R (2002) Singular Value Decomposition and Principal Component Analysis. In A Practical Approach to Microarray Data Analysis. https://doi.org/10.1007/0-306-47815-3_5
Xu P, Shimada S, Fujii Y, Tanaka T (2000) Invariant geodynamical information in geometric geodetic measurements. Geophys J Int 142(2):586–602
Zhang Y, Meratnia N, Havinga PJM (2007) A taxonomy framework for unsupervised outlier detection techniques for multi-type data sets. CTIT Technical Report Series, Paper P-NS/TR-CTIT-07-79, Centre for Telematics and Information Technology (CTIT).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Bálint Magyar Received his Msc geophysicist degree at Eötvös Loránd University in 2016. His research scope in the Satellite Geodetic Observatory of Lechner Non-profit Ltd as geophysicist / GNSS-INSAR researcher includes the geokinematic deformation field analysis of the EPN densification and InSAR automation and processing. His responsibilities include the corresponding application development, numerical implementations and related developer tasks. He is conducting his PhD studies in the Budapest University of Technology and Economics.

Ambrus Kenyeres Graduated as a geophysicist at Eötvös Loránd University and received his Ph.D. at the Technical University of Budapest in 2001. Since 2010, he is the head of the Satellite Geodetic Observatory of the Lechner Non-Profit Ltd at Penc. His main research field is the product-level combination of permanent GNSS networks in the frame of EUREF as WG chair of EPN Densification with the contribution to EPOS GNSS products.

Sándor Tóth He is a qualified land surveyor working at the Lechner Nonprofit Ltd - Satellite Geodetic Observatory and he received his master degree at Budapest University of Technology and Economics in 2021. He has comprehensive expertise in geodesy, surveying and GNSS data analysis. His main research interests include height reference system modernization and GNSS data processing. He is conducting his PhD studies in the Budapest University of Technology and Economics.

István Hajdu Obtained his MSc degree in Geography, with a specialization in Geoinformatics from the University of Szeged in 2017. He has been working at the Lechner Nonprofit Ltd - Satellite Geodetic Observatory since 2019. His main field of interest is open source web application development and spatial data management.

Roland Horváth Received his MSc degree in Geography from the University of Szeged in 2016. Recently, he has been GNSS/InSAR researcher at the Satellite Geodetic Observatory in Lechner Nonprofit Ltd. His work focuses on processing InSAR data, and he is responsible for the exploitation and implementation of the Electronic Corner Reflector (ECR-C) in Hungary.