GPS Solutions (2024) 28:65
https://doi.org/10.1007/s10291-023-01607-1

ORIGINAL ARTICLE

# Machine learning approach for GNSS geodetic velocity estimation

**Seda Özarpacı[1] · Batuhan Kılıç[1] · Onur Can Bayrak[1] · Murat Taşkıran[2] · Uğur Doğan[1] · Michael Floyd[3]**

Received: 3 May 2023 / Accepted: 22 December 2023 / Published online: 25 January 2024
© The Author(s) 2024

**Abstract**
This study aimed to investigate the performance of machine learning (ML) algorithms in determining horizontal velocity at specific points using the current Global Navigation Satellite System (GNSS) velocity field. To achieve this objective, the analysis utilized the most comprehensive velocity field available for Turkey, where 70% of the GNSS velocities was allocated for training the ML algorithms, while the remaining 30% was used for testing. Contrary to the previous research, the significance of considering the tectonic structure within the study area was emphasized at this point. To determine the tectonic structure of the horizontal velocity field in the region, a preliminary clustering procedure was conducted. Subsequently, distinct ML algorithms were trained using velocity fields associated with different tectonic plates. Moreover, to investigate the impact of the tectonic domain, the entire velocity field was also tested using ML algorithms without considering the tectonic structure. Four different ML algorithms, namely, Gradient Boosting Machines (GBM), LightGBM, Random Forest (RF), and eXtreme Gradient Boosting Machines (XGBoost), were employed to estimate the horizontal velocities (east and north components). The findings imply that incorporating the tectonic structure improved the performance of machine learning predictions, as indicated by the GBM algorithm's decreased root-mean-square error values. In addition, when the tectonic structure was taken into account, the accuracy assessment values for the RF and XGBoost algorithms in the east component decreased significantly. In terms of predicting GNSS velocities, the RF algorithm exhibited the lowest root-mean-square error values compared to other algorithms. The horizontal velocity differences between averages of the reference velocity field and the RF velocity estimates are maximum 0.4 mm/yr.

**Keywords** Clustering · Geodetic velocities · GNSS · Machine learning · Random forest

## Introduction

Global plate motion models such as NUVEL-1 or NUVEL-1A (DeMets et al. 1990; 1994) were used before the geodetic velocity field based on the Global Positioning System (GPS) was established to interpret broad crustal movement and tectonic structure. With GPS technology, crustal motions began to be observed directly with a few GPS measurements. In the mid-1990s, 105 survey GPS sites with a few continuous stations measured for 6 years were interpreted to understand the tectonic structure of Turkey (77 sites for Turkey) and the surrounding region (Reilinger et al. 1997). By the early 2000s, there were 119 GPS sites in Turkey, with a total of 190 sites from the Caucasus to Greece in the east–west direction and from the Eurasian plate to the Arabian plate in the north–south direction (McClusky et al. 2000). In the mid-2000s, the number of GPS sites was 165 in Turkey and 433 with a vast area processed and investigated for active tectonics and block modeling (Reilinger et al. 2006). All these data are generally based on survey measurements, and, to increase the number of the stations, researchers densified observations in areas related to specific tectonic structures (Özener et al. 2010; Yavaşoğlu et al. 2011; Tiryakioğlu et al. 2013; Ergintav et al. 2014; Aktuğ et al. 2016). In 2008, the CORS-TR network (Continuously Operating Reference Stations—Turkey) was established homogenously throughout Turkey and northern Cyprus with 146 permanent GNSS stations, now 158 GNSS stations. Finally, in 2023, a study

✉ Seda Özarpacı
ozarpaci@yildiz.edu.tr

[1] Department of Geomatic Engineering, Yildiz Technical University, 34220 Istanbul, Turkey

[2] Department of Electronics and Communication Engineering, Yildiz Technical University, 34220 Istanbul, Turkey

[3] Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

gathered the Turkish National Fundamental GPS Network (TNFGN) of 594 sites, the Turkish Real-Time Kinematic GNSS Network (CORS-TR) of 158 stations, and a few small regional continuous networks, and processed all the data to produce a homogenous and dense GNSS geodetic velocity field (Kurt et al. 2023) with a total of 836 sites.

The geodetic velocity field provides constraints on plate kinematics by explaining how the earth's surface moves and how this is related to tectonic events along plate boundaries (Reilinger and McClusky 2011). By analyzing GNSS velocities, researchers can acquire insight into the deformation and motion of the crust. It should be noted, however, that the velocity field varies between tectonic plates. Interactions between these plates influence the movement and deformation of the Earth's crust, which can result in variations in the observed velocities. In tectonically active regions such as Turkey, velocity variations are significant, because Turkey is situated at the intersection of multiple tectonic plates. The interaction of Eurasian, African, and Arabian lithospheric plates creates a complex tectonic structure in the eastern Mediterranean, of which Turkey is in the middle and actively deforming because of the convergence to the east and subduction to the south–west (Emre et al. 2018). Anatolia is moving at 16–30 mm/yr, increasing from east to west relative to Eurasia, bordered by the dextral North Anatolian Fault (NAF), and sinistral East Anatolian Fault (EAF) and the Dead Sea Fault. The interaction of the Eurasian and African plates specifically includes oceanic lithosphere subduction along the Hellenic and Cyprus Subduction Zones (Fig. 1) and continental extensions such as in the Marmara Sea and Aegean region (Reilinger et al. 2010).

A velocity field derived from GNSS measurements can provide critical information such as moment accumulation rate and strain rate for earthquake hazards (Kurt et al. 2023). In locations with no GNSS stations, conventional interpolation methods such as Kriging modeling have been generally used to estimate geodetic velocities. Recently, Artificial Neural Networks (ANNs), machine learning (ML), and deep learning (DL) techniques belonging to the artificial intelligence (AI) family have become alternatives in this area of geoscience (Konakoglu 2021; Sorkhabi et al. 2022a). While these techniques are widely used across domains, relatively few studies have been conducted for predicting GNSS velocities. Yilmaz and Gullu (2014) evaluated two different types (Back-Propagation Artificial Neural Network and Radial Basis Function Neural Network) of ANN models in order to estimate the velocities of the 125 control points belonging to TNFGN as an alternative tool for the Kriging method in western Turkey. Their results revealed that Back-Propagation Artificial Neural Network is an alternative tool to conventional methods for geodetic station velocity estimation. Konakoglu (2021) comparatively evaluated three different ANN models to estimate the geodetic velocities of 238 TNFGN stations in eastern Turkey and found that the most appropriate model was obtained with the generalized regression neural network. Sorkhabi et al. (2022b) examined the usability of four different DL methods for estimating GPS geodetic velocities at 42 GNSS stations in northwestern Iran. The obtained findings revealed that Deep Boltzmann Machines exhibited superior performance compared to Convolutional Neural Networks, Deep Belief Networks, and Recurrent Neural Networks. To our knowledge, no work has yet explicitly examined and accounted for the tectonic structure of the area the observation sites cover. In addition, no prior research efforts have focused on applying ML methodologies to estimating geodetic velocity.

Within this investigation, we use the presently published GNSS geodetic velocity field and apply supervised ML algorithms, GBM, LightGBM, RF, and XGBoost, to evaluate the efficacy of these techniques in the estimation

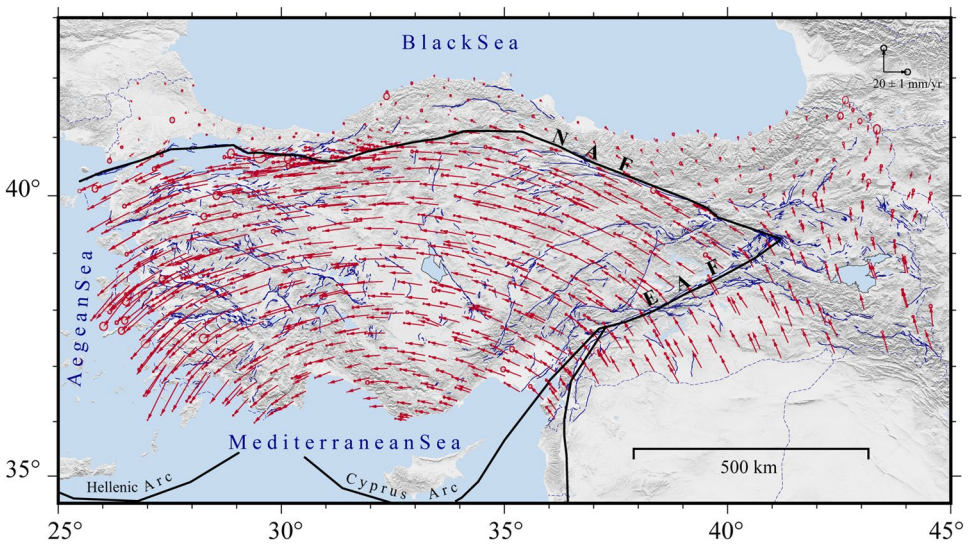


**Fig. 1** Arrows show the current GNSS horizontal velocity field of Turkey with 95% confidence ellipses in the Eurasia-fixed reference frame (Kurt et al. 2023). NAF and EAF stand for North and East Anatolian Faults, respectively. Blue solid lines illustrate active faults (Emre et al. 2013)

of horizontal velocities. In order to demonstrate the importance of accounting for the tectonic structure in the GNSS geodetic velocity field, clustering analysis, known as unsupervised machine learning, was applied to the available velocity field. Clustering analysis for this region has been previously applied (Kilic and Özarpacı, 2022; Özarpacı et al. 2023). However, in these studies, the authors used a sparse GNSS velocity field previously published (Özdemir and Karslıoğlu 2019). In this study, a new and dense GNSS velocity field (Kurt et al. 2023) was used for analysis, and the optimum number of clusters that best fit these data was compared to the literature. Moreover, it was observed that the clustering results are consistent with studies in the literature. Using this velocity field, at first, individual clustering algorithms were performed and comparatively analyzed, and then, Non-Negative Matrix Factorization-based (NMF) consensus clustering was utilized to improve the outputs of the individual clustering ensemble members and to obtain the final clustering outcomes, followed by an evaluation of the fit of the final solutions with the block boundaries. After clustering, all ML algorithms were applied to each cluster separately, and the results were compared with the former ML-based RMSE values to emphasize the effects of primary tectonic structure on the GNSS velocities.

## Data and methodology

The dataset we utilized is the published GNSS geodetic velocity field prepared by Kurt et al. (2023) (see Fig. 1). About 78% of the velocity field analyzed in the research was obtained from campaign data, while 22% was derived from the processing of continuous GNSS stations. The data were gathered between 1992 and 2020, and the daily time for data collection is 7–10 h for the campaign sites and 24 h for the continuous GNSS stations. The velocity field includes data from the Turkish National Fundamental GPS Network (TNFGN) and the Turkish Real-Time Kinematic GNSS Network (TUSAGA-Active). Additionally, data from the Marmara Research Center of Turkey (MAGNET), the Turkish National Permanent GNSS Network (TNPGN), and the regional networks of the water and sewerage administrations of Bursa, Sakarya, and Istanbul were combined with the data to increase the density. The average standard deviations for east and north velocities are 0.22 mm/yr and 0.25 mm/yr, respectively. For a comprehensive understanding of the data used, Kurt et al. (2023) provided further information that may be accessed through their study. Here, we only used this velocity field in our analysis.

The velocity field shown in Fig. 1 is employed with our proposed methodology, which entails two comparative approaches: The first involves ML outcomes derived by neglecting the tectonic configuration of the area under investigation, and the second involves ML results obtained by accounting for the tectonic structure and block boundaries.

In the first stage, the performance of ML-based models was analyzed using the GNSS velocity solution. In the second stage, we included clustering to define the tectonic blocks before ML. For the purpose of clustering analysis, the horizontal velocities of 836 GNSS stations were assessed based on the current GNSS velocity field. Initially, four distinct cluster validity indices, namely, Davies–Bouldin, Elbow, Gap, and Silhouette, were employed to determine the optimum number of clusters (k) that would best fit the data. Subsequently, clustering models were generated from five different clustering techniques, including BIRCH, k-means, mini-batch k-means, HAC, and spectral clustering, to develop the ensemble clustering approach. Then, NMF consensus clustering was utilized to aggregate the outputs of the individual clustering ensemble members and obtain the final clustering outcomes, followed by an evaluation of the fit of the final solutions with the block boundaries. After the first clustering application, some stations along the NAF are assigned to one of the Anatolian clusters, away from that cluster. This can be explained as the plate boundary between Eurasia and Anatolia along the NAF is more dominant as a distinguishing feature than clustering itself (Savage and Simpson 2013; Özarpacı et al. 2023). We cleaned the dataset for these GNSS sites affected by the fault surface trace (1999 Izmit and Düzce earthquake regions). After that, we replicated the identical procedure to derive the ultimate outcomes of ensemble clustering from the data pertaining to the remaining 825 GNSS stations. We determined the block borders and clustered the horizontal velocity data into appropriate blocks. In conclusion, machine learning-based predictors were employed to estimate GNSS geodetic velocities for each cluster. Simultaneously, to estimate the uncertainties of the predicted velocities, model training/testing stages were repeated for standard deviations as well. The performance of the models was assessed using RMSE and mean absolute error (MAE) on the test split data for each respective cluster (Fig. 2).

### Clustering process

In this study, clustering analysis of GNSS velocities was conducted as an initial step to determine the tectonic structure of the region. Clustering analysis of GNSS velocities allows for the identification of coherent groups within the velocity field, providing valuable insights into tectonic processes and deformation patterns. This approach facilitates the identification of distinct blocks with consistent motion, revealing their spatial distribution and behavior. Additionally, clustering aids in the identification of fault system segmentation, block boundaries, and interactions

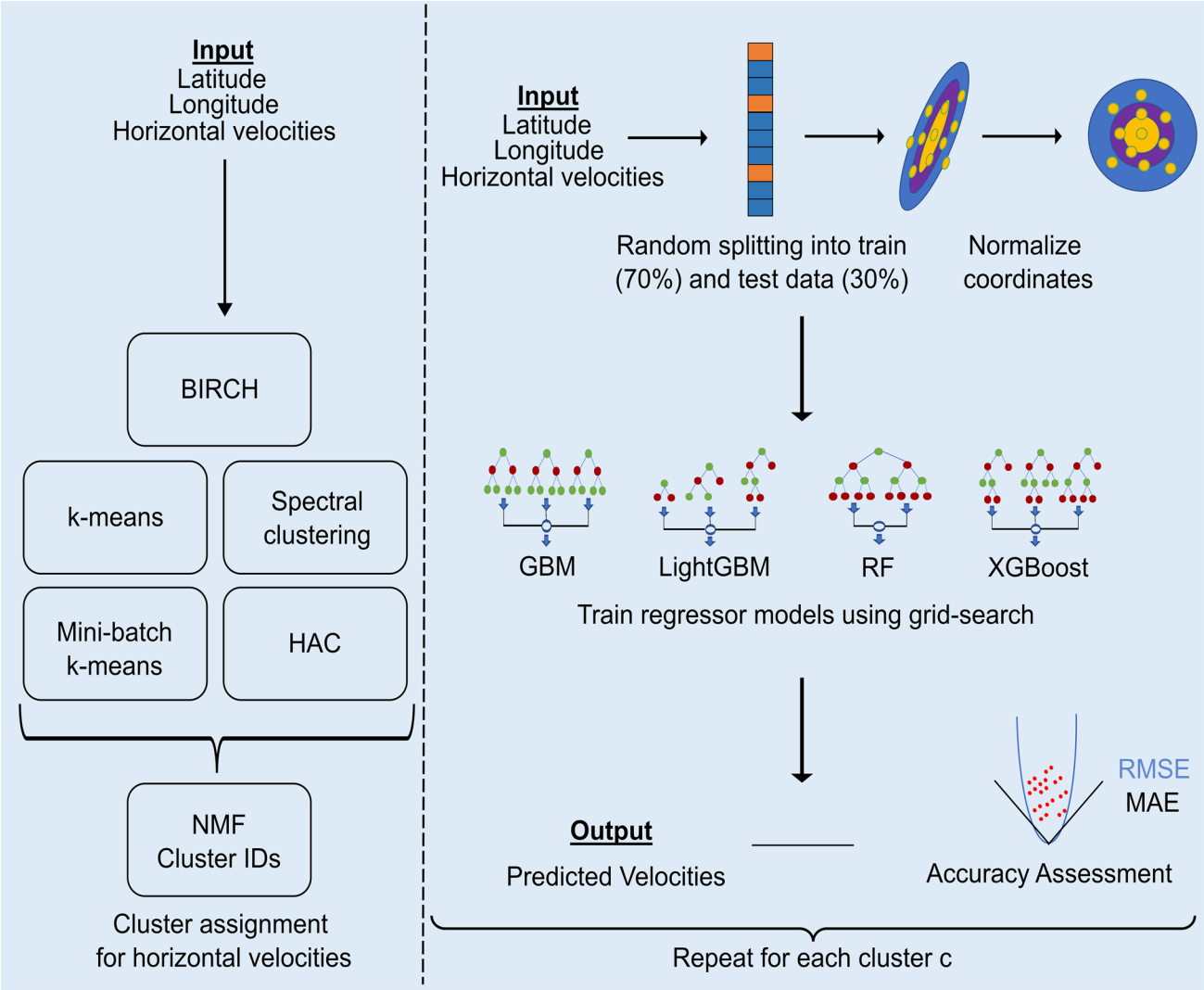


**Fig. 2** Flowchart of the proposed ML-based velocity estimation. Data pre-processing before ML involves the cluster assignment for the velocity field, train–test splitting data, and normalizing the coordinates for data. ML-based models help to estimate velocities for each cluster and compare accuracy assessments with GNSS velocities in the test split

between tectonic units. It also helps delineate localized deformation areas, such as shear zones or regions of strain accumulation, and identifies regions influenced by common driving forces or exhibiting similar kinematic behavior. This procedure helps identify velocity gradients; therefore, ML algorithms process each velocity cluster in individual regions. By performing this, we assume that the ML algorithm will not be affected by the changes in GNSS velocities because of tectonic reactions, and the accuracy of the predicted velocities will be higher. For this purpose, the first step is to determine the optimum number of clusters for the GNSS velocity field. After this step, individual clustering methods are applied for the determined number, and the last step is the ensemble clustering that determines the input velocities of ML algorithms for each cluster. Ensemble clustering helps to eliminate the subjectivity of the individual clustering methods.

## Determining the optimum number of clusters

During the process of clustering data in velocity space, four distinct indices of cluster validity were utilized to determine the most suitable data grouping, estimate the intra-variance (cluster cohesion) and inter-variance (cluster separation). Before executing the data clustering, it is necessary to examine multiple approaches, as opposed to a single method, to obtain the result most compatible with the underlying data structure. The optimum number of clusters for four different cluster validity indices was determined using the Python 3.8 and R programming languages. Table 1 presents

**Table 1** Cluster validity indices results for determining the optimum number of clusters

| | Davies–Bouldin | Elbow | Gap statistic | Silhouette |
|---|---|---|---|---|
| *k, v* | 5, 0.644 | 5, 14,592.830 | 5, 0.517 | 5, 0.513 |

*k* is the optimum number of clusters, and *v* is the best value corresponding to *k* for each cluster validity indices. *v* indicates the relevant value for the optimum number of clusters that best fit the data before clustering the GNSS horizontal velocities

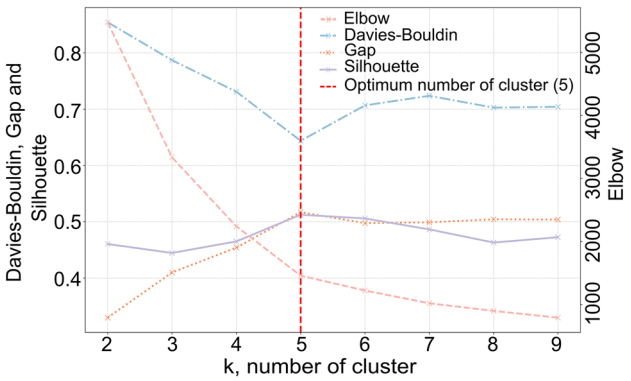


**Fig. 3** Determining the optimal number of clusters that best fit the GNSS horizontal velocities with Davies–Bouldin, Elbow, Gap, and Silhouette methods

the outcomes of the four cluster validity indices utilized to determine the optimal number of clusters that best fit the data ($k=2$ to 10).

Upon reviewing the outcomes presented in Table 1 (also shown in Fig. 3), it is evident that the optimal number of clusters is found to be five for all cluster validity indices. Among $k$ values ranging from 2 to 10, taking into account that the lowest Davies–Bouldin value produces the best results, we still conclude that $k=5$ (0.644) yields the optimum results because the Davies–Bouldin value obtained is closest to 0 (Fig. 3). The Elbow findings generally exhibit a decreasing trend in distortion value ($k=2$–10). However, at a certain value of k, the reduction is gradual, and there is an inflection point ($k=5$) known as the Elbow point. This point represents the optimum cluster number for the Elbow curve (Fig. 3). The Gap statistic algorithm measures the distance between center sites to create intra-cluster observation values and identifies the optimum number of clusters as the longest decline below the reference value (Fig. 3). The optimal number of clusters for this dataset is, therefore, determined to be five. In the Silhouette approach, coherence is established by the average distance among all sites within the same cluster, while cluster separation is determined by the distance to the nearest neighbor. As a result, Silhouette values range from − 1 to 1, and the ideal solution ($k=5$) is obtained as the Silhouette values converge to 1 (Fig. 3).

## Why ensemble clustering?

Each clustering technique groups the dataset from a distinct perspective based on a specific set of criteria, thereby resulting in potentially varying outcomes from different algorithms for the same dataset. In reference to this process, Jain et al. (1999) stated that clustering is a subjective procedure wherein partitioning a given set of data items may differ based on various applications. Therefore, we applied five different clustering methods to evaluate the outcomes in GNSS-based horizontal velocity clustering (Fig. 4a–e).

Upon analyzing the cluster distributions ($k=5$) obtained from five different clustering techniques in Fig. 4a–e, it is evident that all clustering results identify the NAF and EAF, as depicted in Fig. 1, as the borders separating different clusters. As illustrated in Fig. 4, it is noticeable that clustering borders can vary, and the clusters of some GNSS sites can change when different algorithms are employed. At this point, it is necessary to pose the following question. Which is the correct one? Cluster validity indices, which are used to evaluate the quality of clustering outcomes, do not evaluate the outcomes of any clustering algorithm in a totally impartial manner (Vega-Pons and Ruiz-Shulcloper 2011).

Ensemble clustering, also known as the concept of integrating multiple clustering results, has emerged as an alternative method for improving the quality of clustering algorithm outcomes by eliminating the subjectivity. The NMF-based ensemble clustering approach considers the discovery of the median partition as an optimization problem in relation to the cluster ensemble and generates the final solutions to best fit the GNSS geodetic velocity data, as illustrated in Fig. 5a. The separation of the Anatolian block from the Eurasian and Arabian blocks defines NAF and EAF, respectively. Besides, the Aegean block can clearly be seen as a different cluster. The separation of the Anatolian block into two parts, east and west, does not mean that there are tectonic blocks; here, it only means that the velocity difference is enough to create two clusters in that region. In Fig. 5b, one can see five clusters in the velocity field, all in the same colors as each block in Fig. 5a.

## Machine learning approach

Machine learning enables computers to do specific tasks by learning from prior examples, then analyzing new data to get accurate results. Complex problems, especially those involving large datasets and that are difficult for humans to classify, are usually solved using machine learning.

In the following steps, first information is given about the pre-processing and separation of the data as train/test, and in the model training section, detailed information about the machine learning algorithms used in the study is presented.

*Step 1 Train/test split and preprocess*

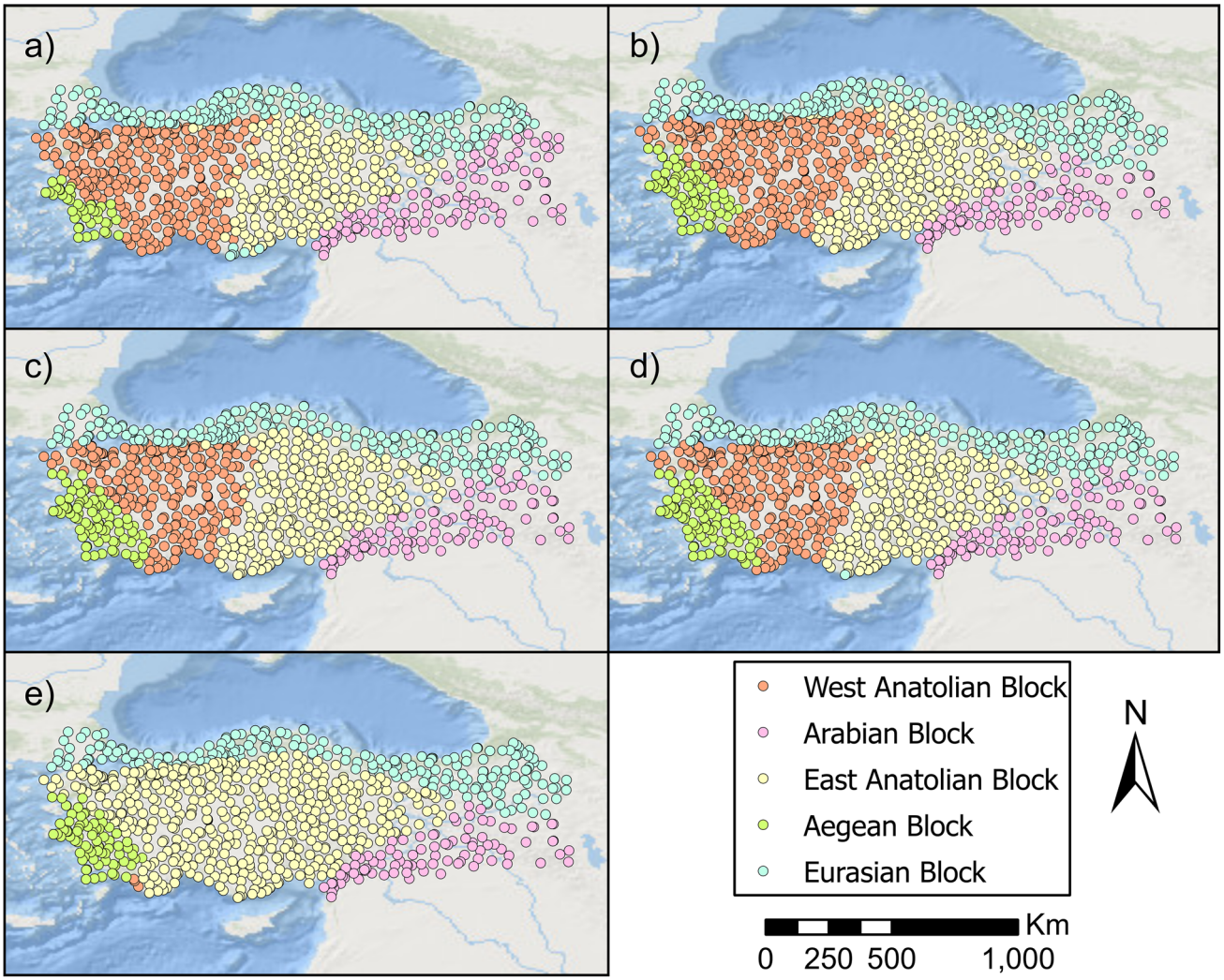


**Fig. 4** Results of single clustering methods for current GNSS geodetic velocity field **a** BIRCH, **b** HAC, **c** k-means, **d** mini-batch k-means, and **e** spectral clustering

In order to examine the performance of the model on unseen samples, among 825 GNSS stations, 70% of training and 30% of test points were randomly selected from each cluster. We selected train/test procedure instead of train/validation/test due to following reasons: (i) limited number of total data, (ii) imbalanced number of data between clusters, and (iii) analyzing model performances under limited train/test data circumstances for usability in daily routine. A total of 576 GNSS velocities and standard deviations of 149, 156, 63, 54, and 154 stations were included in training, whereas a total of 249 GNSS velocities with standard deviations of 64, 68, 27, 24, and 66 stations were involved in testing for clusters 1–5, respectively. The input coordinate pairs of latitude and longitude were subjected to a 0–1 range to avoid over-fitting and stabilize the model training by suppressing the effect of various coordinate pairs.

*Step 2 Model training*

From the literature, it is known that deep learning algorithms require a large amount of training data (Alzubaidi et al. 2021). Due to limited GNSS stations, we picked machine learning-based regression models instead of deep neural networks. Regressors, based on decision trees (Fig. 2), such as RF (Breiman 2001), employ a single decision tree or an ensemble of decision trees to generate predictions. In contrast, boosting-based regressors, such as GBM (Friedman 2001), XGBoost (Chen et al. 2015; Chen and Guestrin 2016), and LightGBM (Ke et al. 2017), employ an ensemble of weak learners for generating predictions. Compared to tree-based models, boosting-based models offer the primary benefit of higher accuracy, particularly for complex and high-dimensional data.

The grid-search technique was utilized for model training, and RMSE and MAE metrics were used for model evaluation. Models were trained and evaluated with the scikit-learn

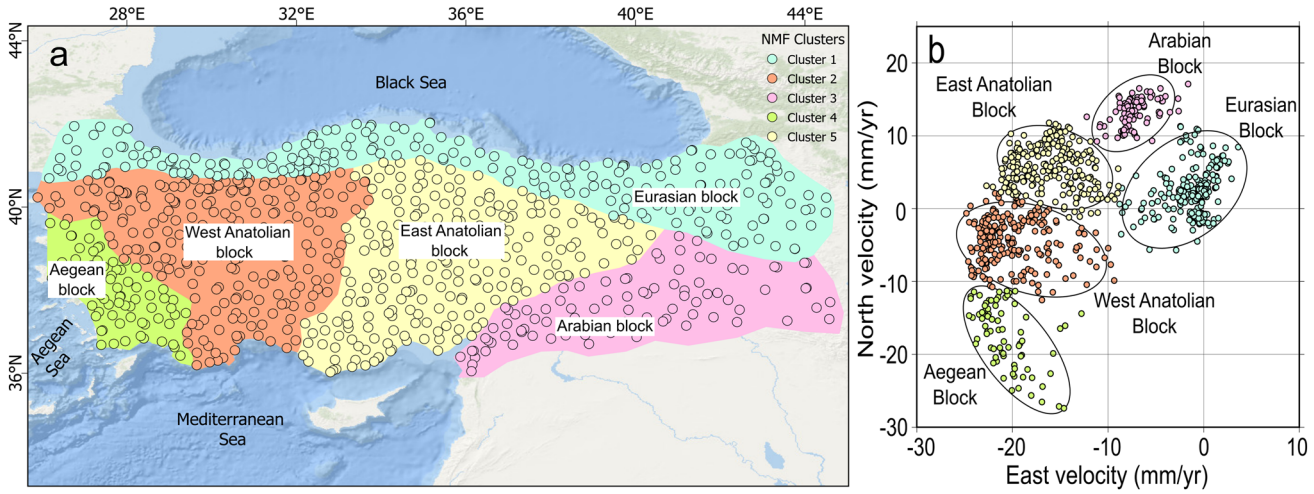


**Fig. 5** Clustering results for current GNSS geodetic velocity field for Turkey **a** NMF consensus clustering results with each color representing a cluster in a geographic space and **b** NMF consensus clustering results with each color representing a cluster in velocity space

(Pedregosa et al. 2011) library in the Python 3.8 programming language.

## Results and discussion

In the circumstances where ignoring the tectonic configuration results in ML-based velocity prediction are demonstrated in Table 2, outperformed predictors are given in bold. The number of train/test splits was 576/249, due to a 0.7/0.3 ratio of 825 GNSS station velocities.

According to Table 2, RF produced the most accurate results for the north velocities, whereas LightGBM outperformed other models in the east velocities without clustering. It is also observed that there are larger errors in the estimation of the velocities with the east component compared to the estimation of the velocities with the north component, especially in the RMSE values. We assume that neglecting the tectonic structure can cause an effect, especially with the Anatolia western escape with respect to the Eurasian plate. We believe that as Anatolia moves toward the west, the east–west velocity component of the growth is not accurately predicted with sufficient sensitivity by the ML algorithms. Therefore, we believe that it is necessary to predefine the velocity changes and evaluate each tectonic block separately.

After defining the tectonic structure with blocks using clustering, velocity prediction via ML algorithms is executed, and the results are illustrated in Table 3. Outperforming predictors are given in bold. The number of test sites was 64, 68, 27, 24, and 66 for clusters 1–5, respectively.

When one examines the RMSE and MAE results in Table 3, the success can be seen for each ML algorithm in each cluster. For instance, for cluster 1, RF has the best error score with 1.01 RMSE and 0.75 MAE for the east velocities. However, for the north velocity component, GBM (0.80 RMSE and 0.68 MAE) was slightly better than RF (0.91 RMSE and 0.72 MAE). Overall results (mean RMSE and MAE scores for each cluster) showed that RF achieved the best performance while LightGBM had the biggest errors.

If the results of neglecting the tectonic structure and considering it are compared, one can see that GBM accuracy assessments are decreasing with the tectonic structure taken into account (Table 2 and Table 3 All (Mean) values). However, LightGBM RMSE and MAE values are increasing with the clustering of velocities. Our results demonstrated that, as the number of training and testing sites increased, LightGBM produced more accurate results due to the generalization ability of regularization and the gradient-based learning approach. On the other hand, when cluster 2 and cluster 3 results are compared, we see that LightGBM produced smaller errors with fewer sites. Therefore, we assume that the variation in the north velocity component in cluster 4, the Aegean block, affected all the RMSE and MAE accuracy assessments, including LightGBM. The Hellenic arc in the southern part of the Aegean Sea and continental extensions

**Table 2** Accuracy assessment of velocity prediction via ML algorithms with neglecting tectonic configuration

| Accuracy Ass | ML algorithms | East | North |
|---|---|---|---|
| RMSE (mm/yr) | GBM | 2.94 | 1.34 |
| | LightGBM | **1.64** | 1.01 |
| | RF | 2.01 | **0.93** |
| | XGBoost | 1.79 | 1.03 |
| MAE (mm/yr) | GBM | 1.12 | 0.87 |
| | LightGBM | **0.84** | 0.77 |
| | RF | 0.99 | **0.75** |
| | XGBoost | 0.89 | 0.78 |

**Table 3** Accuracy assessment of velocity prediction via ML algorithms after clustering

| | | Cluster 1 ($n=64$) | | Cluster 2 ($n=68$) | | Cluster 3 ($n=27$) | | Cluster 4 ($n=24$) | | Cluster 5 ($n=66$) | | All (Mean) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | East | North | East | North | East | North | East | North | East | North | East | North |
| RMSE (mm/yr) | GBM | 1.28 | **0.80** | 0.74 | 1.04 | 2.12 | 0.79 | **0.86** | 2.94 | 1.03 | **0.47** | 1.21 | 1.21 |
| | LightGBM | 1.22 | 1.47 | 2.27 | 1.50 | 1.44 | 1.04 | 3.22 | 11.57 | 3.30 | 0.90 | 2.29 | 3.30 |
| | RF | **1.01** | 0.91 | 0.83 | 1.05 | **1.32** | **0.47** | 0.89 | **2.29** | **0.82** | 0.72 | **0.97** | **1.09** |
| | XGBoost | 1.58 | 0.95 | **0.66** | **1.01** | 1.92 | 0.64 | 0.95 | 2.39 | 1.08 | 0.65 | 1.24 | 1.13 |
| MAE (mm/yr) | GBM | 0.84 | **0.68** | 0.71 | 0.79 | 0.99 | 0.80 | 0.79 | 1.37 | 0.79 | **0.53** | 0.82 | 0.83 |
| | LightGBM | 0.88 | 0.87 | 1.08 | 0.88 | 0.89 | 0.82 | 1.48 | 2.74 | 1.31 | 0.74 | 1.13 | 1.21 |
| | RF | **0.75** | 0.72 | 0.68 | **0.78** | **0.79** | **0.55** | **0.77** | **1.29** | **0.68** | 0.64 | **0.73** | **0.80** |
| | XGBoost | 0.91 | 0.78 | **0.64** | 0.79 | 0.93 | 0.64 | **0.77** | 1.30 | 0.77 | 0.63 | 0.80 | 0.83 |

such as the Sea of Marmara and the Aegean region create a gradient in the velocity field, especially in the north component, that ML algorithms cannot solve.

RF and XGBoost also show a significant decrease in the east component when considering the tectonic structure. However, in the north component, one can see a slight increase in the accuracy assessment values. Also, this could be a reflection of the north component in cluster 4.

Predicting a GNSS geodetic velocity with ML should treat each cluster independently; therefore, the error does not propagate. Considering the tectonic structure, we can say that the predictions made with ML algorithms give more reliable results, especially the RF algorithm providing the most consistent results among the clusters. In Fig. 6, we have used the RF results for illustration only but the GBM and XGBoost results are also suitable for estimating the GNSS velocity components (see Table 3).

Table 4 shows the optimal hyperparameter configurations of the machine learning regressors found by the grid-search technique that produces the lowest RMSE and MAE scores in the corresponding cluster. Subsequently, these hyperparameters were also selected for the estimation of standard deviations. The number of estimators is diverse for all predictors except XGBoost. In general, the learning rates yield 0.1 or 0.7 except for XGBoost and GBM, but lower learning rates achieve more accurate results.

In Fig. 6, the velocity predictions and standard deviations (orange arrows with 95% confidence ellipses) from the RF algorithm are shown with all clusters together in geographic space. Blue arrows illustrate the GNSS geodetic velocity field, and one can see the test results of RF and how it reflects the tectonic structure of the region.

The velocity estimates acquired by ML algorithms are compatible with the reference velocity field. Average east–west component velocity estimates obtained by RF method are − 0.77, − 19.82, − 7.66, − 20.67, and − 15.69 mm/yr, when the reference values are − 0.86, − 19.92, − 7.38, − 20.61, and − 15.56 mm/yr for each cluster, respectively. When the north–south velocity component is examined, the results obtained with the RF

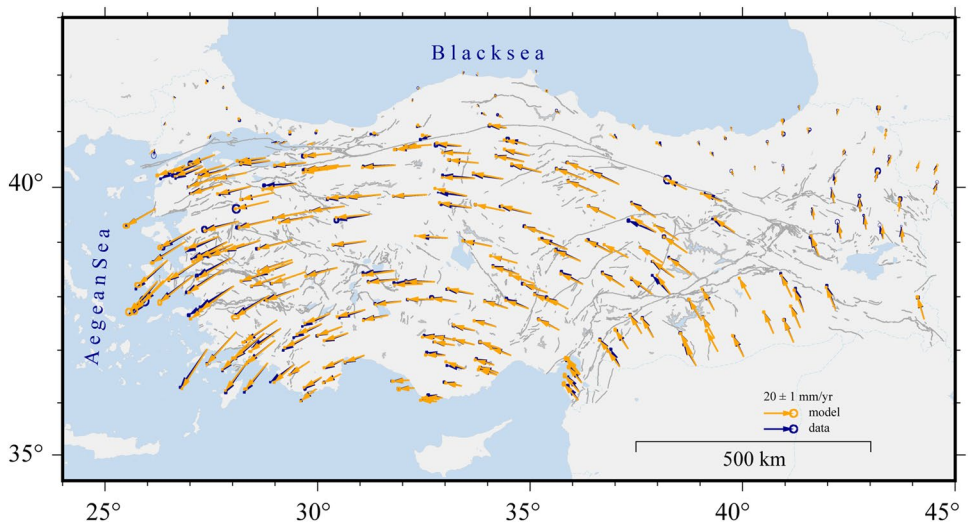


**Fig. 6** RF test results with combination for each cluster shown in geographic space. Blue and orange arrows show the data and the RF results, respectively, with 95% confidence ellipses. Gray solid lines illustrate the active faults (Emre et al. 2013)

**Table 4** Grid-search results for machine learning algorithms

| Predictor | Parameter | Cluster | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 |
| GBM | Number of estimators | 200 | 250 | 100 | 50 | 150 |
| | Learning rate | 0.1 | 0.3 | 0.1 | 0.1 | 0.1 |
| LightGBM | Number of estimators | 150 | 500 | 50 | 100 | 50 |
| | Learning rate | 0.7 | 0.1 | 0.1 | 0.7 | 0.1 |
| RF | Number of estimators | 150 | 100 | 50 | 50 | 200 |
| | Maximum depth | 7 | 8 | 8 | 8 | 8 |
| XGBoost | Number of estimators | 50 | 50 | 50 | 50 | 50 |
| | Learning rate | 0.7 | 0.1 | 0.1 | 0.1 | 0.3 |

algorithm for each cluster are 2.69, − 4.64, 13.16, − 16.99, and 5.26 mm/yr, and the reference north velocity component averages are 2.78, − 4.85, 13.26, − 16.63, and 5.36 mm/yr, respectively. The standard deviations of the velocity values determined by the ML algorithms have also been estimated. When examining the test results obtained with the RF method (Fig. 6), the estimated mean standard deviations for each cluster are found to be 0.22, 0.18, 0.24, 0.41, and 0.20 mm/yr for the east–west component and 0.27, 0.20, 0.28, 0.36, and 0.22 mm/yr for the north–south component, respectively. It is observed that the obtained velocity differences between the model results and the data are maximum 0.2 mm/yr except for the 3rd and the 4th clusters. It is observed that the decrease in the amount of data assigned to clusters has a negative impact on the results as previously explained. The test data values per cluster, as shown in Table 3, dropping below 30 in the 3rd and 4th clusters, widen the difference between velocity estimates and reference velocities. Also, standard deviation values, except for the 4th cluster, are in good agreement with the input data. The lower accuracy in the 4th cluster is thought to be due to the low number of data points in the cluster. Besides, the results of the Hellenic arc extension in the region disturb and affect the north–south velocity component. We assume when estimating GNSS velocities by ML algorithms, especially in tectonic regions where velocity gradient is large, should be done very carefully and enough data should be provided.

In velocity space, the data and the RF predictions for the GNSS velocities are shown in Fig. 7 with residual histograms and distributions. Figure 7a shows the east velocity field with predictions, and Fig. 7b shows the east velocity residuals, or in other words, the differences between the data and the predictions or model are in the normal distribution. Figure 7c shows the north velocity field data on the one axis and RF predictions on the other axes. Figure 7d shows the residuals of north velocities and the normal distribution curve. In Fig. 7a and c, all the circles are around the x = y line, which means that the differences are very small, and residual histograms have proven this (Fig. 7b and d).

## Conclusion

This research explored the application of ML algorithms to estimate horizontal GNSS velocities in an active tectonic region. The dataset was split into 70% for training the ML models and 30% for testing their performance. An essential aspect of this study was the consideration of the tectonic structure within the study area, achieved through a clustering procedure.

In tectonic regions, ensemble clustering techniques can be applied to identify tectonic structures when estimating velocities using ML algorithms. Here, we used four methods to determine the optimum number of clusters, and we found five with all methods for the velocity field used. Afterwards, we used five different individual clustering algorithms to identify the five cluster borders before ensemble clustering. NMF consensus clustering technique is used to eliminate the subjective results of individual clustering techniques and create a robust, stabile, and novel result. As a result, the clusters are compatible with the velocity gradients and the plate boundaries that create transform faults such as NAF and EAF. The each obtained cluster is tested with the ML algorithms separately.

We prove that the horizontal GNSS velocities can be estimated using ML algorithms. However, ignoring the tectonic structure while estimating velocities can lead to incorrect results. Therefore, when working in tectonically active regions, considering the tectonic structure ensures the generation of more accurate results. Our results indicated that incorporating the tectonic structure significantly enhanced the ML predictions, particularly with the GBM algorithm, which exhibited reduced RMSE values. Taking into account, the tectonic structure also led to improved

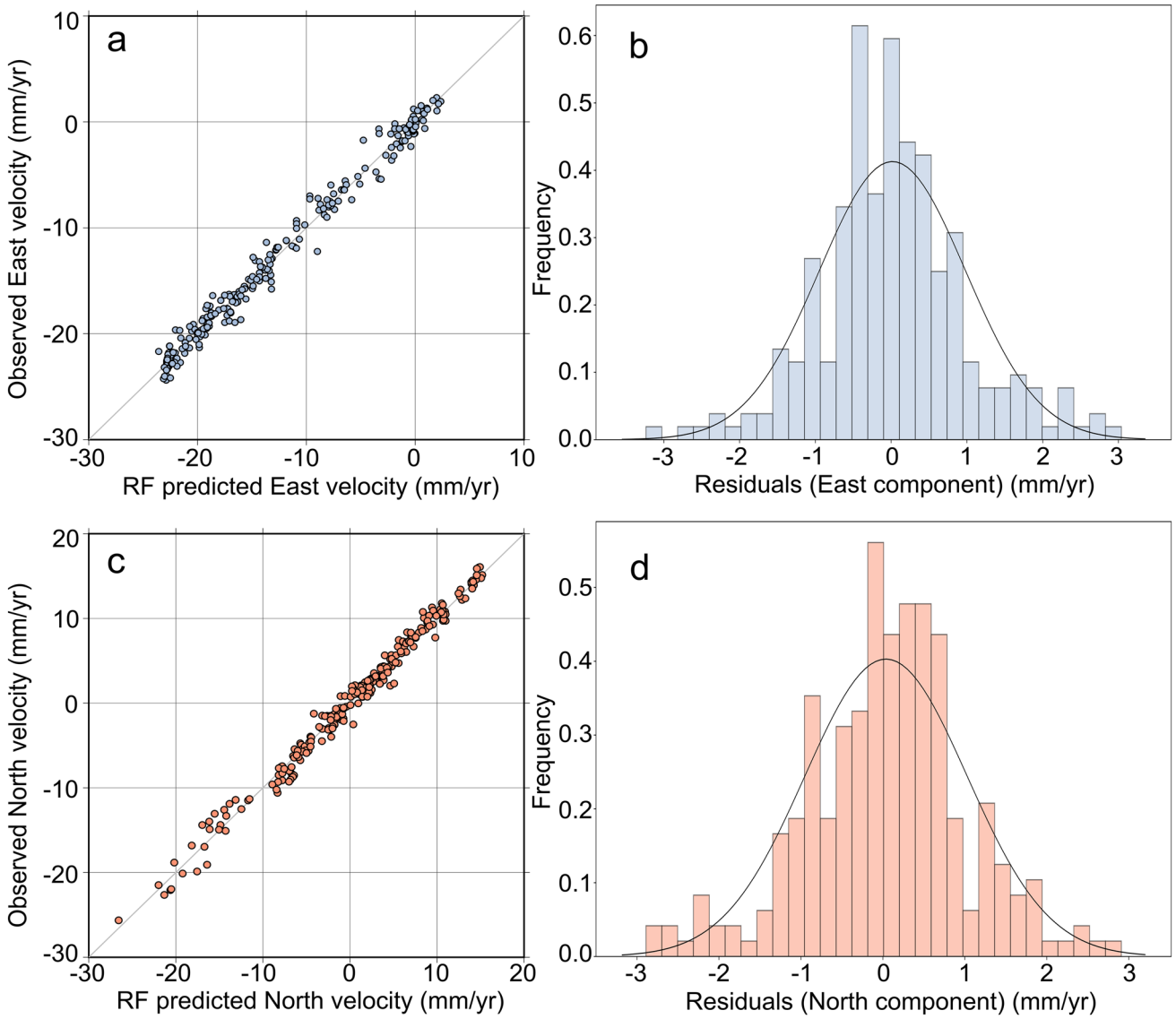


**Fig. 7** Scatter graph of the predicted and observed geodetic velocity values and histogram distributions of the residual **a** East component data and model values in velocity space, **b** east component data and model differences in histogram, **c** north component data and model values in velocity space, and **d** north component data and model differences in histogram

accuracy assessment values for the RF and XGBoost algorithms in the east component.

The RF algorithm demonstrated superior performance in predicting GNSS velocities, displaying the lowest root-mean-square error values among the ML algorithms. The results show that the velocity estimates and the reference velocity differences are not more than 0.4 mm/yr. The maximum differences are in the clusters with the lowest number of GNSS stations. Besides, the highest difference is in the north component of 4th cluster. We assume that the extension caused by the Hellenic arc is affecting the north velocity component in the region. The findings underscore the importance of accounting for the tectonic structure when utilizing ML algorithms for horizontal velocity estimation.

This study contributes valuable insights into the application of ML algorithms for geodetic velocity field predictions and emphasizes the significance of considering tectonic factors in such analyses. The results provide useful information for geodynamic studies and fault displacement analyses. Future research could focus on refining ML algorithms and integrating more advanced techniques to further enhance the accuracy of velocity predictions in tectonically active regions like Turkey. In conclusion, ML algorithms present promising opportunities to advance

geodetic velocity field predictions and deepen our understanding of tectonic movements.

In the future, new studies will carried out with various algorithms and test the results in tectonic and nontectonic regions.

**Acknowledgements** We would like to thank our editor Alfred Leick and two anonymous reviewers for constructive and thorough reviews that substantially improved this paper. We used GMT 6 for the map figures (Wessel et al. 2019).

**Author contributions** SÖ, BK, OCB, and MT wrote the main manuscript text and prepared the figures; UD gave the concepts of the manuscript; and MT, UD, and MF reviewed and edited the manuscript.

**Funding** Open access funding provided by the Scientific and Technological Research Council of Türkiye (TÜBİTAK). This research received no external funding.

**Data availability** The data can be found at https://zenodo.org/record/7916022#.ZFp3C3ZByUl with the data fields of longitude, latitude, east velocity, north velocity, the standard deviation of east component, the standard deviation of north component, coefficient of correlation, up velocity, standard deviation of up component, and station names.

## Declarations

**Conflict of interest** We declare that the authors have no competing interest.

**Ethical approval and consent to participate** This research did not require ethical approval or informed consent for participant involvement.

**Consent for publication** All authors have consented to the publication of this research.

**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

Aktuğ B, Ozener H, Dogru A, Sabuncu A, Turgut B, Halicioglu K, Yilmaz O, Havazli E (2016) Slip rates and seismic potential on the East Anatolian fault system using an improved GPS velocity field. J Geodyn 94:1–12. https://doi.org/10.1016/j.jog.2016.01.001

Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, Santamaría J, Fadhel MA, Al-Amidie M, Farhan L (2021) Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data 8:1–74. https://doi.org/10.1186/s40537-021-00444-8

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

DeMets C, Gordon RG, Argus DF, Stein S (1990) Current plate motions. Geophys J Int 101(2):425–478. https://doi.org/10.1111/j.1365-246X.1990.tb06579.x

DeMets C, Gordon RG, Argus DF, Stein S (1994) Effect of recent revisions to the geomagnetic reversal time scale on estimates of current plate motions. Geophys Res Lett 21(20):2191–2194. https://doi.org/10.1029/94GL02118

Emre Ö, Duman TY, Özalp S, Şaroğlu F, Olgun Ş, Elmacı H, Çan T (2018) Active fault database of Turkey. Bull Earthq Eng 16(8):3229–3275. https://doi.org/10.1007/s10518-016-0041-2

Ergintav S, Reilinger RE, Çakmak R, Floyd M, Cakir Z, Doğan U, King RW, McClusky S, Özener H (2014) Istanbul's earthquake hot spots: Geodetic constraints on strain accumulation along faults in the Marmara seismic gap. Geophys Res Lett 41(16):5783–5788. https://doi.org/10.1002/2014GL060985

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29(5):1189–1232

Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM Comput Surv 31(3):264–323

Kilic B, Özarpacı S (2022) Ensemble clustering in GPS velocities: a case study of Turkey. Appl Sci 12(24):12636. https://doi.org/10.3390/app122412636

Konakoglu B (2021) Prediction of geodetic point velocity using MLPNN, GRNN, and RBFNN models: a comparative study. Acta Geod Geophys 56(2):271–291. https://doi.org/10.1007/s40328-021-00336-6

Kurt Aİ, Özbakir AD, Cingöz A, Ergintav S, Doğan U, Özarpaci S (2023) Contemporary velocity field for Turkey inferred from combination of a dense network of long term GNSS observations. Turk J Earth Sci 32:275–293. https://doi.org/10.55730/1300-0985.1844

McClusky S et al (2000) Global Positioning System constraints on plate kinematics and dynamics in the eastern Mediterranean and Caucasus. J Geophys Res Solid Earth 105(B3):5695–5719. https://doi.org/10.1029/1999JB900351

Özarpacı S, Kılıç B, Bayrak OC, Özdemir A, Yılmaz Y, Floyd M (2023) Comparative analysis of the optimum cluster number determination algorithms in clustering GPS velocities. Geophys J Int 232(1):70–80. https://doi.org/10.1093/gji/ggac326

Özdemir S, Karslıoğlu MO (2019) Soft clustering of GPS velocities from a homogeneous permanent network in Turkey. J Geod 93(8):1171–1195. https://doi.org/10.1007/s00190-019-01235-z

Özener H, Arpat E, Ergintav S, Dogru A, Cakmak R, Turgut B, Dogan U (2010) Kinematics of the eastern part of the North Anatolian fault zone. J Geodyn 49(3–4):141–150. https://doi.org/10.1016/j.jog.2010.01.003

Pedregosa F et al (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830

Reilinger RE et al (2006) GPS constraints on continental deformation in the Africa-Arabia-Eurasia continental collision zone and implications for the dynamics of plate interactions. J Geophys Res Solid Earth. https://doi.org/10.1029/2005JB004051

Reilinger R, McClusky S (2011) Nubia–Arabia–Eurasia plate motions and the dynamics of Mediterranean and Middle East tectonics. Geophys J Int 186(3):971–979. https://doi.org/10.1111/j.1365-246X.2011.05133.x

Reilinger RE, McClusky SC, Oral MB, King RW, Toksoz MN, Barka AA, Kinik I, Lenk O, Sanli I (1997) Global positioning system measurements of present-day crustal movements in the Arabia-Africa-Eurasia plate collision zone. J Geophys Res Solid Earth 102(B5):9983–9999. https://doi.org/10.1029/96JB03736

Reilinger R, McClusky S, Paradissis D, Ergintav S, Vernant P (2010) Geodetic constraints on the tectonic evolution of the Aegean region and strain accumulation along the Hellenic subduction zone. Tectonophysics 488(1–4):22–30. https://doi.org/10.1016/j.tecto.2009.05.027

Savage JC, Simpson RW (2013) Clustering of GPS velocities in the Mojave block, southeastern California. J Geophys Res Solid Earth 118(4):1747–1759. https://doi.org/10.1029/2012JB009699

Sorkhabi OM, Alizadeh SMS, Shahdost FT, Heravi HM (2022a) Deep learning of GPS geodetic velocity. J Asian Earth Sci X(7):100095. https://doi.org/10.1016/j.jaesx.2022.100095

Sorkhabi OM, Milani M, Seyed Alizadeh SM (2022b) Investigating the efficiency of deep learning methods in estimating GPS geodetic velocity. Earth Space Sci 9(10):e2021EA002202. https://doi.org/10.1029/2021EA002202

Tiryakioğlu İ, Floyd M, Erdoğan S, Gülal E, Ergintav S, McClusky S, Reilinger R (2013) GPS constraints on active deformation in the Isparta angle region of SW Turkey. Geophys J Int 195(3):1455–1463. https://doi.org/10.1093/gji/ggt323

Vega-Pons S, Ruiz-Shulcloper J (2011) A survey of clustering ensemble algorithms. Int J Pattern Recognit Artif Intell 25(03):337–372

Wessel P, Luis JF, Uieda L, Scharroo R, Wobbe F, Smith WHF, Tian D (2019) The generic mapping tools version 6. Geochem Geophys Geosyst 20:5556–5564. https://doi.org/10.1029/2019GC008515

Yavaşoğlu H, Tarı E, Tüysüz O, Çakır Z, Ergintav S (2011) Determining and modeling tectonic movements along the central part of the North Anatolian Fault (Turkey) using geodetic measurements. J Geodyn 51(5):339–343. https://doi.org/10.1016/j.jog.2010.07.003

Yilmaz M, Gullu M (2014) A comparative study for the estimation of geodetic point velocity by artificial neural networks. J Earth Syst Sci 123:791–808. https://doi.org/10.1007/s12040-014-0411-6

Chen T, Guestrin C (2016) Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp 785–794

Chen T, et al. (2015) Xgboost: extreme gradient boosting. R package version 0.4–2, 1(4):1–4

Emre Ö, Duman TY, Özalp S, Elmacı H, Olgun Ş, Şaroglu F (2013) "Açıklamalı Türkiye Diri Fay Haritası Ölçek 1/1.125.000", (Map originally in Turkish) Maden Tetkik ve Arama Genel Müdürlüğü Özel Yayın Serisi 30. Special Publication Series" of MTA-30, 30, ISBN: 978-605-5310-56-1

Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY (2017) Lightgbm: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst 30

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

**Seda Özarpacı** received her B.S. and M.Sc. degrees from Istanbul Technical University and her Ph.D. in Geomatic Engineering from Yildiz Technical University. She works as an Assistant Professor at Yildiz Technical University. Her primary research interests include GNSS observations, tectonics, crustal deformation and creep behavior of earth crustal.

**Batuhan Kılıç** received his B.S. degree from Bulent Ecevit University in 2013, M.Sc. degree from Yildiz Technical University in 2017 and is currently pursuing his Ph.D. in Geomatic Engineering at Yildiz Technical University and working at the same department as a research assistant. His main research interests focus on online geocoding approaches, volunteered geographic information, and algorithmic foundations of geographic information systems.

**Onur Can Bayrak** received his M.Sc. degree from Yildiz Technical University of Istanbul, Turkey, in 2020. He currently works as a research assistant in the Geomatics Engineering Department at the same university. His research interests involve applied machine learning and deep learning for image classification and 3D semantic segmentation.

**Murat Taşkıran** received B.Sc. (2013) and M.Sc. (2016) degrees in Electronics and Communication Engineering, from Yildiz Technical University (YTU), Istanbul, Turkey. Since 2023, he has been working as an Assistant Professor in Department of Electronics and Communication Engineering in YTU. His research interests are in image processing, neural networks, time-series analysis and randomness analysis.

**Uğur Doğan** is a Professor at the Yildiz Technical University in Geomatic Engineering Department. His research interests are GNSS observations, geodetic problems, deformation analysis, crustal deformations, and gravity analysis.

**Michael Floyd** is a Research Scientist in the Department of Earth, Atmospheric and Planetary Sciences at the Massachusetts Institute of Technology in Cambridge, Massachusetts, USA. His current research interests involve using GNSS observations for the study of solid Earth phenomena, including tectonics, earthquakes, and geothermal fields, and he is one of the developers and maintainers of the GAMIT/GLOBK GNSS processing software suite.