Contemporary GPS Velocity Field for Turkey Inferred from Combination of a Dense Network of Long Term

Ali Deger Özbakır

September 2021

6 A Statistical Approach for Clustering GPS Horizontal Velocity Vectors

a More descriptive title: clustering outlines the major tectonic plates

6.1 Motivation

1. The Anatolia – Aegean domain represents a distributed deformation zone sub-divided into large plate-like units. First-order features of this deformation have been analyzed in earlier GPS studies (e.g. McClusky et al., 2000) and detailed with the studies carried out in the following years (Reilinger et al., 2006). However,

   Give a precise description

   > *[f]ull resolution of velocity gradients requires spatial sampling of about 0.25 of the locking depth (Smith and Sandwell 2003) which is typically lower than the available GPS station spacing.*
   > —(Haines et al., 2015)

   Smith & Sandwell reference does not contain this information.

   For example, assuming 20km maximum locking depth for North Anatolian Fault, requires 5 km of inter-station spacing. Therefore, continuous acquisition of the GPS data is needed. During the last decade, accuracy of velocity solutions were improved, and the expansion of continuous networks increased the spatial resolution.

2. Link the adjacent paragraphs by mentioning the relevant synthesis studies.

3. The posterior velocity solutions obtained from individual campaigns can be combined with the 7-parameter Helmert transform, however this combination is far from perfect. The basis for creating a precise and an accurate velocity field is uniform processing of the time series recorded both in the campaign and fixed GPS stations, at once and for a single reference system(e.g. Nocquet, 2012). Although the data density has increased X-fold since Reilinger et al. (2006), this crucial calculation has not been made. 6-fold?

4. In this study, we processed the time series of 800+ stations up to 15+ years and present the most accurate velocity field. This data can better constrain the kinematics of continental deformation in Anatolia - Aegean domain.

6.2 How do we typically model the continental deformation: block models

5. Continental deformation is often described as the interaction of rigid blocks. This block approach (or block model) is basically the plate tectonics principles applied to continental deformation. Accordingly, the continental lithosphere is rigid and the deformation is localized along narrow fault zones. The (kinematic) formulation is expressed by inversion of rotation vectors per block, which minimizes the observed velocities. Especially, with the advent of space geodetic techniques block modelling have been successful in calculating the slip-rates on major active faults —a major input for seismic hazard assessment.

With the addition of a new measurement, internal deformation is better quantified for slowly deforming regions. Hence, these might be sub-divided into smaller blocks –one reason for many different block realizations.

6. Some preliminary assumptions are made for this approach to work: negligible internal deformation within and complete closure of predefined blocks. During the course of modelling, block geometries are modified to achieve an improved fit with the observations. Often, however, blocks exhibit significant internal deformation. Hence, the assumptions of rigid blocks is an oversimplification . Furthermore, block boundaries may not be reflected in the geological or seismological record. As such, there has not been an agreement on the choice of block geometries (cf. block geometries for the Aegean region in Nyst and Thatcher (2004); Reilinger et al. (2006); Floyd et al. (2010)). Therefore, slip-rate estimates can significantly vary in different studies.

7. In essence, block modelling partitions many GPS-derived velocity vectors into few distinct groups of blocks with single rotation vector. Simply put, the aim is to identify similar vectors, whose underlying blocks are not *a priori* known. Simpson et al. (2012) addresses the initial subjective choice of blocks and proposes an objective, algorithmic approach: clustering.

Paraphrase and link with the preceding paragraph. Also this info might be useful later to introduce/motivate the need for gap statistics, silhouette or other measures

> *Clustering is a type of unsupervised learning algorithm, which are usually applied to data that does not contain any label information. Clustering algorithms, partition data into distinct groups of similar items. [...] The goal is to split up the data in such a way that points within a single cluster are very similar and points in different clusters are different. As the real assignment of observations is not known a priori, often the only way to evaluate the result of an unsupervised algorithm is to inspect it manually. As a consequence, unsupervised algorithms are used often in an exploratory setting (Müller and Guido, 2016)*

6.3 Previous Work

8. The clustering of GPS-derived velocity vectors provides an objective and intuitive starting point for the analysis of block-like deformation and yields statistically significant number of blocks. Cluster analysis is not a new method, it has applications in various discipline, such as psychology, bio-informatics, and data analysis, but the application is rather recent for continental deformation (Simpson et al., 2012; Savage and Simpson, 2013a,b; Liu et al., 2015; Savage and Wells, 2015; Thatcher et al., 2016; Savage, 2018; Takahashi et al., 2019; Özdemir and Karslıoğlu, 2019)

Better start with contrasting partitional and agglomerative. Then introduce k-means and HAC. k-medoids have been used in all the prev. work, except Takahashi and Simpson

9. There are two popular clustering algorithms: k-means and Hierarchical Agglomerative Clustering (HAC). k-means finds cluster centers for a given number of clusters, minimizing the within-cluster dispersion. Its easy implementation, linear execution time scaling with the number of

observations, and guaranteed convergence renders it one of the most popular clustering algorithms. k-means perform poorly when clusters are of varying sizes. Furthermore, clustering is sensitive to data outliers and the initial sorting of the data. Therefore pre-processing and repeat experiments need to be done.

see google ML doc

10. HAC builds a hierarchy of clusters without having fixed number of cluster. HAC begins with each observation being its own cluster, then sequentially joins near observations (or clusters) until all the observations are grouped into a single cluster. As pairwise distances are calculated at each step, the algorithm is slower compared to k-means and scales with $\mathcal{O}(n^2)$. However, "agglomerative clustering can provide a whole hierarchy of possible partitions of the data, which can be easily inspected via dendrograms." (Müller and Guido, 2016). This is essentially a linkage graph with inter-cluster distances in the y-axis. Then inspection of the dendrogram or variance-reduction measures give indication for the optimal number of clusters. HAC methods work best with low dimensional data.

11. Simpson et al. (2012) applied hierarchical agglomerative clustering approach to GPS velocity vectors in San Francisco Bay region. The resulting clusters and geologically determined fault-bounded blocks are in good agreement. As the displacements in the Bay Area are predominantly translational, station locations are not taken into account.

Move those as particular examples of the aforementioned pop. algorithms, §5-6.

12. Savage and Simpson (2013a,b) applied k-medoids clustering and improved the method by considering station locations on the clustering. After initial clustering of velocities, Euler vectors for each cluster is calculated. Then iteratively, each observation is assigned to that cluster, for which its Euler vector best describes.

13. In this study we used the HAC algorithm, because Anatolia... Takahashi et al. (2019) applied the HAC technique to Taiwan, because GPS observations could be grouped in a hierarchy of tectonic units deforming under the boundary forces of two major plates. We have a similar justification. Furthermore, it is more intuitive to investigate the dendrogram and understand the similarity of different clustering.

14. With each new GPS-data augmented to the existing database, researchers traditionally construct a block model. The motivation is to understand the added value of more data points, both in time and space.

6.4 Clustering

15. Two key components of clustering is the determination of similarity (or proximity) and joining method . There are various choices for both: Simpson et al. (2012) used squared Euclidean metric with centroid linkage, whereas Takahashi et al. (2019) Euclidean metric with centroid linkage.

see linkage tutorial

Most publ. use the Ward method. Simpson et al. (2012) used the "centroid" linkage. Scipy 'centroid' linkage requires the Euclidean metric. Ward, however requires squared Euclidean metric.

16. Data cleaning.

17. dendrogram, how to read

18. velocity space vs. map view

6.5 How many clusters?

19. Gap Analysis, Silhouette number, inconsistency, variance reduction, ...

> *$d_{ii'}$, denote the distance between observations i nd i'. The most common choice for $d_{ii'}$ is the squared Euclidean distance $\sum_{ij}(x_{ij} - x_{i'j})^2$. [...] Sum of pairwise distances for all points in cluster r:*
>
> $$D_r = \sum_{i,i' \in C_r} d_{ii'}$$
>
> *[and the disperson]*
>
> $$W_k = \sum_r^k \frac{1}{2n_r} D_r$$
>
> *So, if the distance d is the squared Euclidean distance, then W_k is the pooled within-cluster sum of squares around the cluster means (the factor 2 makes this work exactly). The sample size n is suppressed in this notation. (Tibshirani et al., 2001)*
>
> ALİ: I use the simplified expression instead of using pairwise distances. This should significantly speed up the dispersion calculations. But somehow centroid linkage in scipy implementation only works for Euclidean distance. Then I have to revert to the slower function.

6.6 Sphericity of the Earth

20. Simpson et al. (2012) method works if GPS observations are distant from their corresponding Euler poles.

References

M. A. Floyd, H. Billiris, D. Paradissis, G. Veis, A. Avallone, P. Briole, S. McClusky, J.-M. Nocquet, K. Palamartchouk, B. Parsons, and P. C. England. A new velocity field for greece: Implications for the kinematics and dynamics of the aegean. *Journal of Geophysical Research: Solid Earth*, 115 (B10), 2010. doi: https://doi.org/10.1029/2009JB007040. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2009JB007040.

A. Haines, L. Dimitrova, L. M. Wallace, and C. A. Williams. *Enhanced surface imaging of crustal deformation: Obtaining tectonic force fields using GPS data*. Springer, 2015.

S. Liu, Z. Shen, and R. Bürgmann. Recovery of secular deformation field of mojave shear zone in southern california from historical terrestrial and gps measurements. *Journal of Geophysical Research: Solid Earth*, 120(5): 3965–3990, 2015. doi: https://doi.org/10.1002/2015JB011941. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015JB011941.

S. McClusky, S. Balassanian, A. Barka, C. Demir, S. Ergintav, I. Georgiev, O. Gurkan, M. Hamburger, K. Hurst, H. Kahle, K. Kastens, G. Kekelidze, R. King, V. Kotzev, O. Lenk, S. Mahmoud, A. Mishin, M. Nadariya, A. Ouzounis, D. Paradissis, Y. Peter, M. Prilepin, R. Reilinger, I. Sanli, H. Seeger, A. Tealeb, M. N. Toksöz, and G. Veis. Global positioning system constraints on plate kinematics and dynamics in the eastern mediterranean and caucasus. *Journal of Geophysical Research: Solid Earth*, 105(B3):5695–5719, 2000. doi: https://doi.org/10.1029/1999JB900351.

A. Müller and S. Guido. *Introduction to machine learning with Python: a guide for data scientists.* " O'Reilly Media, Inc.", 2016.

J.-M. Nocquet. Present-day kinematics of the mediterranean: A comprehensive overview of gps results. *Tectonophysics*, 579:220–242, 2012.

M. Nyst and W. Thatcher. New constraints on the active tectonic deformation of the aegean. *Journal of Geophysical Research: Solid Earth*, 109(B11), 2004. doi: https://doi.org/10.1029/2003JB002830. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2003JB002830.

S. Özdemir and M. O. Karslıoğlu. Soft clustering of gps velocities from a homogeneous permanent network in turkey. *Journal of Geodesy*, 93(8):1171–1195, 2019. doi: 10.1007/s00190-019-01235-z.

R. Reilinger, S. McClusky, P. Vernant, S. Lawrence, S. Ergintav, R. Cakmak, H. Ozener, F. Kadirov, I. Guliev, R. Stepanyan, M. Nadariya, G. Hahubia, S. Mahmoud, K. Sakr, A. ArRajehi, D. Paradissis, A. Al-Aydrus, M. Prilepin, T. Guseva, E. Evren, A. Dmitrotsa, S. V. Filikov, F. Gomez, R. Al-Ghazzi, and G. Karam. Gps constraints on continental deformation in the africa-arabia-eurasia continental collision zone and implications for the dynamics of plate interactions. *Journal of Geophysical Research: Solid Earth*, 111(B5), 2006. doi: https://doi.org/10.1029/2005JB004051. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2005JB004051.

J. C. Savage. Euler-vector clustering of gps velocities defines microplate geometry in southwest japan. *Journal of Geophysical Research: Solid Earth*, 123(2): 1954–1968, 2018. doi: https://doi.org/10.1002/2017JB014874. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2017JB014874.

J. C. Savage and R. W. Simpson. Clustering of gps velocities in the mojave block, southeastern california. *Journal of Geophysical Research: Solid Earth*, 118(4):1747–1759, 2013a. doi: 10.1029/2012JB009699.

J. C. Savage and R. W. Simpson. Clustering of velocities in a gps network spanning the sierra nevada block, the northern walker lane belt, and the central nevada seismic belt, california-nevada. *Journal of Geophysical Research: Solid Earth*, 118(9):4937–4947, 2013b. doi: 10.1002/jgrb.50340,.

J. C. Savage and R. E. Wells. Identifying block structure in the pacific northwest, usa. *J. Geophys. Res.*, 120:7905–7916, 2015. doi: 10.1002/2015JB012277.

R. W. Simpson, W. Thatcher, and J. C. Savage. Using cluster analysis to organize and explore regional gps velocities. *Geophysical Research Letters*, 39(18), 2012. doi: 10.1029/2012GL052755.

A. Takahashi, M. Hashimoto, J. Hu, K. Takeuchi, M. Tsai, and Y. Fukahata. Hierarchical cluster analysis of dense gps data and examination of the nature of the clusters associated with regional tectonics in taiwan. *Journal of Geophysical Research: Solid Earth*, 124(5):5174–5191, 2019. doi: https:

//doi.org/10.1029/2018JB016995. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018JB016995.

W. Thatcher, J. C. Savage, and R. W. Simpson. The eastern california shear zone as the northward extension of the southern san andreas fault. *Journal of Geophysical Research: Solid Earth*, 121(4):2904–2914, 2016. doi: https://doi.org/10.1002/2015JB012678. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015JB012678.

R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.