Recognizing Spatial Seismicity Patterns in Earthquake Data Using Unsupervised Machine Learning Techniques

Seyed Naser Hashemi *

School of Earth Sciences, Damghan University, Damghan, Iran, hashemi@du.ac.ir

KEY WORDS: Pattern recognition, Earthquake hazard, Cluster analysis, Earthquake prediction, K-means clustering

ABSTRACT:

In this study an unsupervised pattern recognition approach was used to identify spatial seismicity patterns in areas around the Kazerun and Kare- Bas fault systems, located in southern Iran. For this purpose, 5,546 earthquake events (with $M \geq 2.5$) recorded between 2006 and February 2024 were extracted from the Iranian Seismological Center catalog. Then, the study region, extending from 50.5°–52.5°E and 28°–31°N, was divided into a 0.2° × 0.2° grid net, resulting in 225 cells, of which 182 contained earthquake data. After that, for each cell, three quantitative seismicity parameters were computed: (1) number of earthquakes, (2) maximum event magnitude, and (3) mean focal depth of earthquakes. The k-means clustering technique was then applied to the obtained dataset using Euclidean distance as the similarity metric, and also with the number of clusters (k) varied between 3 and 5 to generate different seismicity zoning maps. The seismicity pattern maps, obtained in this study, revealed spatially coherent zones corresponding to areas with distinct seismicity behavior, reflecting variations in both frequency and focal depth of earthquakes. The results obtained demonstrate that the k-means clustering technique, as an unsupervised learning technique, can effectively distinguish spatially significant seismicity zones and that this technique can provide a quantitative basis for data-driven seismicity zoning of active regions. In addition, this approach can be easily extended to other tectonically and seismically active regions or refined using higher-resolution grid cells and additional seismicity parameters and variables to improve pattern resolution and predictive capability.

1. INTRODUCTION

The occurrence of earthquakes in a region, viewed as a spatio-temporal process, is a complex phenomenon. Artificial intelligence–based pattern recognition techniques can effectively assist researchers in gaining a clearer understanding of this complex pattern. An improved understanding of this pattern contributes to a deeper comprehension of earthquake occurrence processes and may also be regarded as a step toward earthquake prediction.

During recent decades, extensive earthquake catalogs have been compiled, containing detailed records of epicentral locations, magnitudes, depths, and occurrence times of earthquakes. The huge volume and high complexity of these datasets exceed the capacity of manual analysis alone. On the other hand, AI-based techniques offer the capability to process these large datasets efficiently and consistently, allowing for faster and more precise analysis compared to traditional manual interpretation methods. Additionally, these new methods can reveal subtle correlations and hidden structures and patterns that might otherwise remain undetected. In fact, conventional rule-based algorithms, which rely mainly on predefined criteria, often lack the flexibility needed to capture the nonlinear nature of earthquake occurrence processes. In contrast, modern pattern recognition techniques possess advanced feature extraction capabilities, enabling them to outperform traditional automated earthquake monitoring systems in many scenarios (Taulli, 2019).

In recent years, machine learning (ML) techniques, including deep learning (DL) methods, have experienced rapid advancement across numerous scientific disciplines, including earthquake seismology. These methods have contributed significantly to improving the analysis, interpretation, and classification of earthquake data, resulting in a growing amount of scientific research and practical applications in this field (Kubo et al., 2024). These approaches are especially useful for identifying significant spatio-temporal patterns and structures hidden within seismicity datasets. They also facilitate exploratory analysis, allowing researchers to uncover relationships and patterns directly from raw data without relying solely on predefined physical models (Mousavi and Beroza, 2023). Among these methods, unsupervised learning techniques—like k-means clustering and principal component analysis—are particularly valuable as they can uncover natural clusters and concealed patterns in datasets without prior labeling. These techniques allow for the identification of inherent data structures, which are crucial for comprehending seismicity trends. In Fig. 1, different types and approaches of unsupervised machine learning techniques for analyzing data are shown.

The possibility of applying the k-means clustering technique for seismicity zoning is considered in this research. In this study, the k-means clustering, as an unsupervised machine learning method, was applied to analyze earthquake data and to find seismicity pattern maps. Accordingly, the application of k-means clustering, as an unsupervised machine learning technique, in finding patterns in earthquake data around the Kazerun and Kare-Bas faults (southern Iran) has been investigated.

* Corresponding author

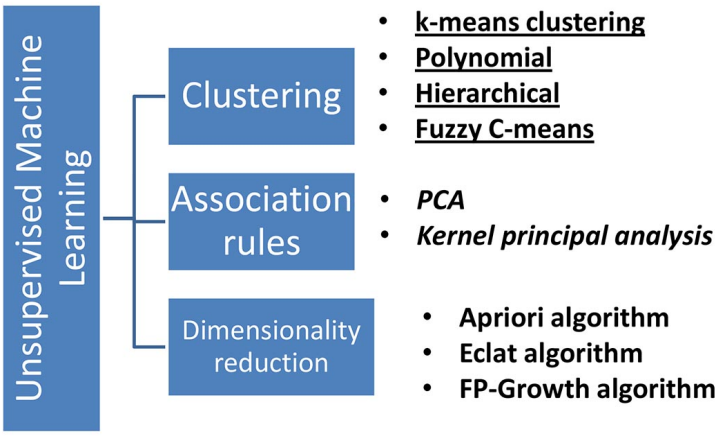

Figure 1. Different types of methods and approaches of unsupervised machine learning techniques

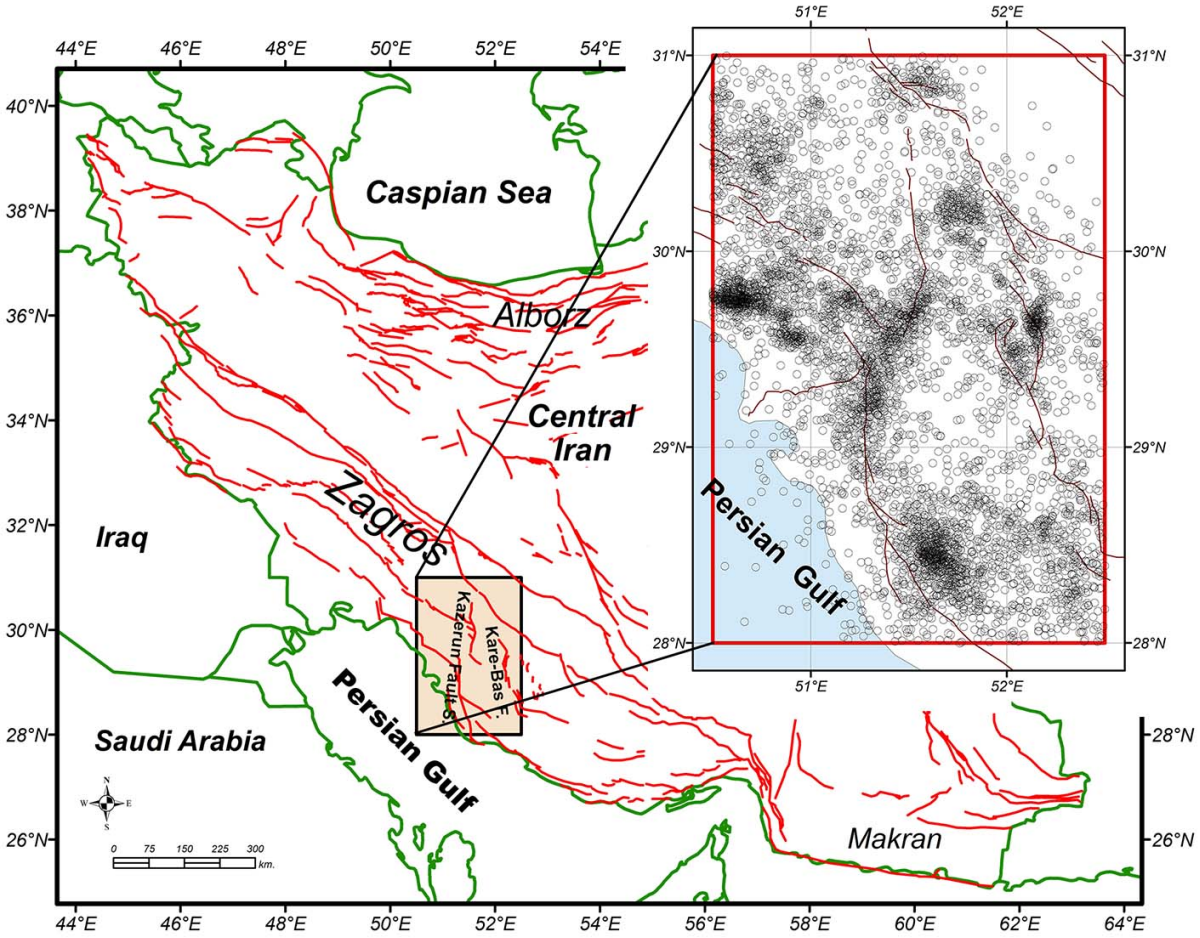

Figure 2. The location of the study region and the seismicity map of the region showing the location of earthquake epicenters

The Kazerun and Kare-Bas fault systems are situated within the central part of the Zagros Mountains of Iran, one of the most tectonically active regions in the Middle East. These fault systems are considered two seismically active basement faults of Iran (Fig. 2). Figure 2 simply shows the location of the study region in Iran and the seismicity map of the areas surrounding the Kazerun and Kare-Bas fault system, showing the epicentral distribution of earthquakes in this region. During the past decades, many destructive earthquakes have occurred in this region. These fault systems include several roughly north-south-trending right-lateral strike-slip segments (Authemayou et al., 2005).

Understanding the spatial pattern and clustering behavior of earthquake epicenters along these fault lines is therefore essential for improving earthquake risk assessment and advancing our knowledge of regional tectonic processes operating around these fault systems.

2. METHODOLOGY

Cluster analysis is a type of unsupervised pattern recognition technique that organizes data into meaningful subgroups based on inherent similarities among observations or variables. Its primary objective is to divide a dataset into distinct categories, known as clusters (or groups), so that observations within the same cluster are more similar, while observations belonging to different clusters are more dissimilar. In fact, this technique aims to maximize similarity within clusters and maximize differences between clusters, thereby revealing natural structures or patterns present in the data without relying on predefined labels (Hastie et al., 2017).

Among the various clustering techniques, k-means clustering is an applied unsupervised machine learning algorithm that is specifically designed to partition a dataset into a predetermined number of clusters, where this number is selected in advance by the researcher or analyst. The algorithm assigns each observation to one of these clusters based on similarity, ensuring that observations grouped together share common characteristics or features. This results in high similarity among members of the same cluster (high intra-cluster similarity) and low similarity between members of different clusters (low inter-cluster similarity). Each cluster is characterized by a central point, called the centroid, which represents the average position of all observations belonging to that cluster. This centroid serves as a reference point for assigning and updating cluster membership during the clustering process (for more details, see Everitt et al., 2011; Kaushik and Mathur, 2014; Han et al., 2023).

The main idea underlying the k-means algorithm is the minimization of variability within clusters, commonly referred to in literature as within-cluster variation. This is achieved by iteratively assigning observations to clusters and recalculating the cluster centroids until the total internal variation is reduced as much as possible. Different versions of the k-means algorithm have been developed, among which the Hartigan and Wong algorithm (1979) is one of the most commonly used standard implementations by researchers. This method simply measures within-cluster variation by calculating the sum of squared Euclidean distances between each observation and the centroid of its given cluster, thereby ensuring that cluster members remain closely grouped around their central point. Accordingly, the sum of squared Euclidean distance is computed as follows:

$$W\left(C_k\right)=\sum_{x_i\in C_k}\left(x_i-\mu_k\right)^2 \qquad (1)$$

where x_i denotes a data point belonging to the cluster C_k and μ_k is the mean value of the points assigned to the cluster C_k.

Each case (or cell here) (x_i) is assigned to a given cluster such that the sum of squares (SS) distance of the cells to their assigned cluster centers μ_k is a minimum. Accordingly, we define the value of total within-cluster variation as follows:

$$TWSS=\sum_{k=1}^{k}W\left(C_k\right)=\sum_{k=1}^{k}\sum_{x_i\in C_k}\left(x_i-\mu_k\right)^2 \qquad (2)$$

This value can be used as a reliable metric to evaluate how tightly grouped or clustered the data points (here, cells) are within each cluster or group. It really reflects the overall internal cohesion of the clusters by measuring the distance between each data point and the center of its assigned cluster. Lower values of this metric indicate that the cases are located closer to their respective group centers, which suggests that these extracted groups are more homogeneous and well-defined. Therefore, an effective and acceptable grouping result in this analysis is characterized by minimizing this metric, as it signifies higher cluster compactness and also better clustering performance in general.

In this work, the k-means clustering method, was employed to model and analyze seismicity data around the Kazerun and Kare-Bas fault systems in our study area. The purpose of applying this technique here was to recognize spatial patterns in seismicity activity and to generate earthquake activity maps that illustrate the distribution and grouping of earthquake events in this area. This clustering-based analysis seems to provide valuable insights into the structural behavior and seismicity-related characteristics of the region.

3. DATA ANALYSIS AND RESULTS

In order to do this investigation, earthquake data corresponding to the study region were obtained from the Iranian Seismological Center database, covering the time interval from January 2006 through February 2024. Accordingly, only events with magnitudes of 2.5 or greater were considered in order to ensure the reliability and completeness of the input dataset. The study area in this analysis is bounded by longitudes 50.5°E to 52.5°E and latitudes 28°N to 31°N. The geographical position of the selected region, together with the spatial distribution of earthquake epicenters, is presented in Fig. 2. As illustrated in Fig. 2, a total of 5,546 earthquakes occurred within these boundaries during the specified period, providing the basis for the spatial analysis. Furthermore, the general workflow adopted in this study for identifying spatial patterns of seismicity in different numbers of clusters using the k-means clustering method is summarized in the flowchart shown in Fig. 3.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-W8-2025-343-2026 | © Author(s) 2026. CC BY 4.0 License.

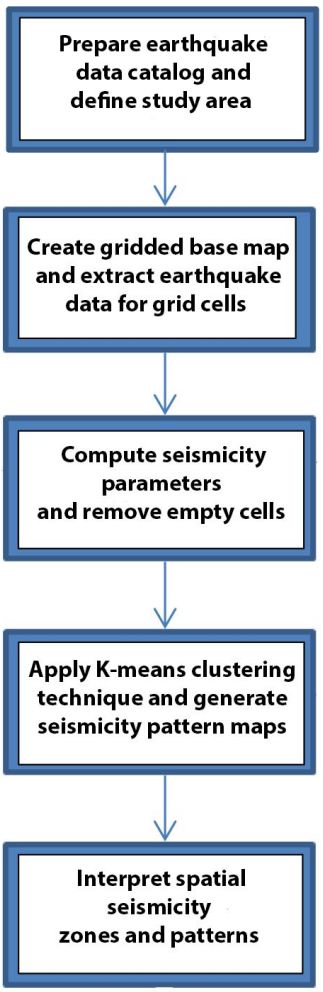

Figure 3. Flowchart of the proposed methodology for recognizing spatial seismicity pattern maps using k-means clustering technique.

For the purpose of systematic spatial analysis, the entire study region was subdivided into a regular grid net composed of cells with dimensions 0.2° × 0.2°. This gridded framework served as the reference spatial unit for organizing and analyzing the seismicity data. Earthquake events falling within each grid cell were extracted from the earthquake catalog and assigned accordingly. Then, grid cells with no recorded earthquake data were excluded from further analysis, since they did not contribute useful information regarding earthquake activity. For each of the remaining cells, several important seismicity indicators were calculated. These parameters included (1) the total number of earthquakes recorded within the cell, (2) the maximum magnitude observed among those events, and (3) the mean focal depth of earthquakes in that cell. These quantitative variables were then spatially mapped to illustrate their spatial variation across the study area. The spatial distributions of these three seismicity parameters are displayed in Fig. 4.

Following the preparation of the spatial dataset and calculation of the relevant parameters, the grid cells were grouped using the k-means clustering algorithm. This unsupervised classification technique partitions the dataset into clusters based on similarities among the selected attributes. In this study, the Euclidean distance metric was applied as the similarity criterion to measure the difference between data points and cluster centers. Based on this distance measure, grid cells with similar seismic characteristics were grouped together, resulting in distinct seismicity zones. By applying the clustering procedure with different numbers of clusters, multiple seismic zonation maps were generated for the study area.

The resulting seismicity pattern maps derived from the k-means clustering analysis are presented in Fig. 5 for different number of 3, 4, and 5 zones. These maps illustrate the spatial classification of the region into several zones, each characterized by internally similar seismic behavior. As shown in Fig. 5, cells belonging to the same cluster exhibit comparable values in terms of earthquake frequency, magnitude, and focal depth, thereby forming coherent seismicity zones. These maps provide a clear and effective representation of the spatial variability of seismic activity within the region. Consequently, the identified zones can be used as a basis for further interpretation, comparison, and evaluation of seismic hazard characteristics across the study area.

4. CONCLUSIONS

This research provided spatial seismicity pattern maps showing earthquake activity for the area surrounding the Kazerun and Kare-Bas fault systems by applying the k-means clustering method to earthquake datasets. Through the examination and pattern recognition of earthquake events, the analysis demonstrated that unsupervised machine learning approaches can be considered effective tools for identifying underlying spatial and temporal seismicity patterns that may not be clearly visible through conventional and traditional analytical techniques. These methods enable the detection of meaningful structures and patterns within seismicity data, offering valuable insights into how seismicity data are distributed across the study region.

The findings of this work represent an initial step toward the systematic and quantitative identification of seismicity patterns in regions that are prone to earthquakes. The methodology employed here in this study is not limited to the specific study area; rather, it provides a simple framework that can be well applied to other seismically active zones worldwide. Furthermore, by improving the quality of input earthquake data, incorporating additional relevant parameters and variables, and utilizing higher-resolution grids, future research can further enhance the accuracy and realism of the resulting seismicity pattern maps. Such refinements would allow for more detailed and reliable interpretations of earthquake behavior in seismically active regions.

In conclusion, the analytical approach presented in this study offers a simple and practical means of detecting and interpreting seismicity patterns across different geographical scales. Its adaptability makes it suitable for application in a wide range of tectonically active regions, contributing to a deeper understanding of seismic processes and supporting more informed assessments of earthquake risk in various parts of the world.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-W8-2025-343-2026 | © Author(s) 2026. CC BY 4.0 License.

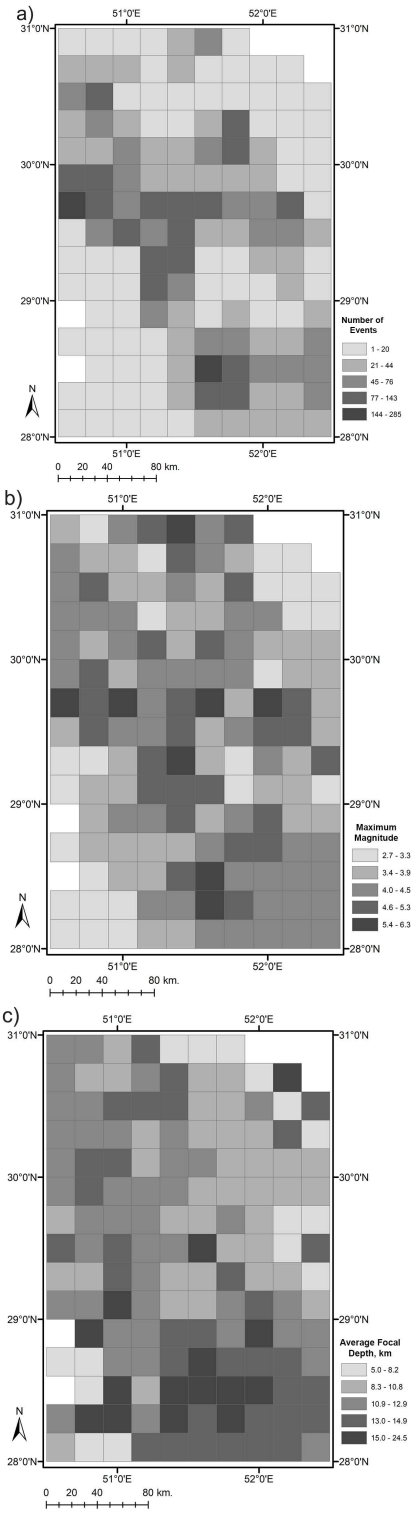

Figure 4. Gridded zoning maps of the region under study, showing the spatial variation of three quantitative attributes; a) number of earthquakes, b) maximum magnitude of events, and c) average focal depth of earthquakes.

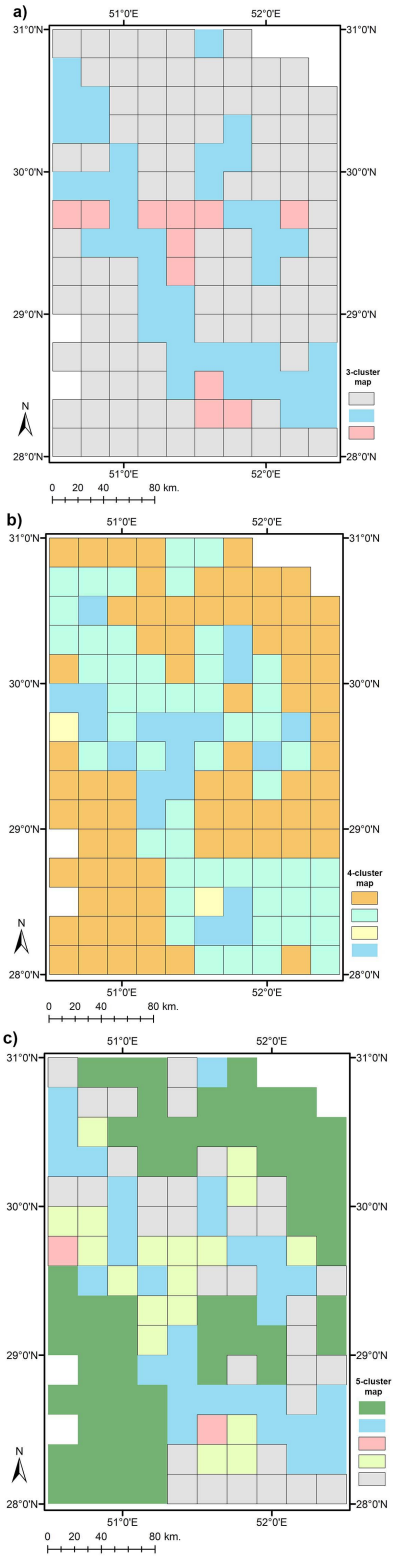

Figure 5. Seismicity pattern recognition maps of the region under study generated by k-means analysis of earthquake data for three different number of zones; a) 3-cluster map, b) 4-cluster map, and c) 5-cluster map.

ACKNOWLEDGEMENTS

This research was partially supported by the Research Council of Damghan University. I would like to sincerely thank the Institute of Geophysics at the University of Tehran for supplying the seismicity data utilized in this research.

REFERENCES

Authemayou, C., Bellier, O., Chardon, D., Malekzade, Z., Abbassi, M., 2005: Role of the Kazerun Fault System in Active Deformation of the Zagros Fold-and-Thrust Belt (Iran). Geoscience 337, 539-545.

Everitt, B.S., Landau, S., Leese, M., Stahl, M., 2011: Cluster Analysis, fifth ed. John Wiley & Sons Inc., West Sussex.

Han, J., Pei, J., Tong, H., 2023: Data Mining: Concepts and Techniques. fourth ed. Morgan Kaufmann Publishers.

Hartigan, J.A., Wong, M.A., 1979: Algorithm AS 136: A K-Means Clustering Algorithm. Journal of the Royal Statistical Society (Series C) 28(1), 100-108.

Hastie, T., Tibshirani, R., Friedman, J., 2017: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. Springer, Berlin.

Kaushik, M., Mathur, B., 2014: Comparative study of K-means and hierarchical clustering techniques. International Journal of Software & Hardware Research in Engineering 2(6), 93–98.

Kubo, H., Naoi, M., Kano, M, 2024: Recent advances in earthquake seismology using machine learning. Earth, Planets and Space 76(36), https://doi.org/10.1186/s40623-024-01982-0.

Mousavi, S.M., Beroza, G.C., 2023: Machine learning in earthquake seismology. Annual Review of Earth and Planetary Sciences 51(1), 105-129.

Taulli, T., 2019: Artificial Intelligence Basics: A Non-Technical Introduction. Apress Berkeley, Monrovia, CA.