Global Clustering of Seismic Event Data: Geophysical Mantle Strata and Eruption Prediction

Charles Hoots
charles.hoots@colorado.edu
University of Colorado, Boulder

Sai Meghashyeam Vangeepuram
sava9298@colorado.edu
University of Colorado, Boulder

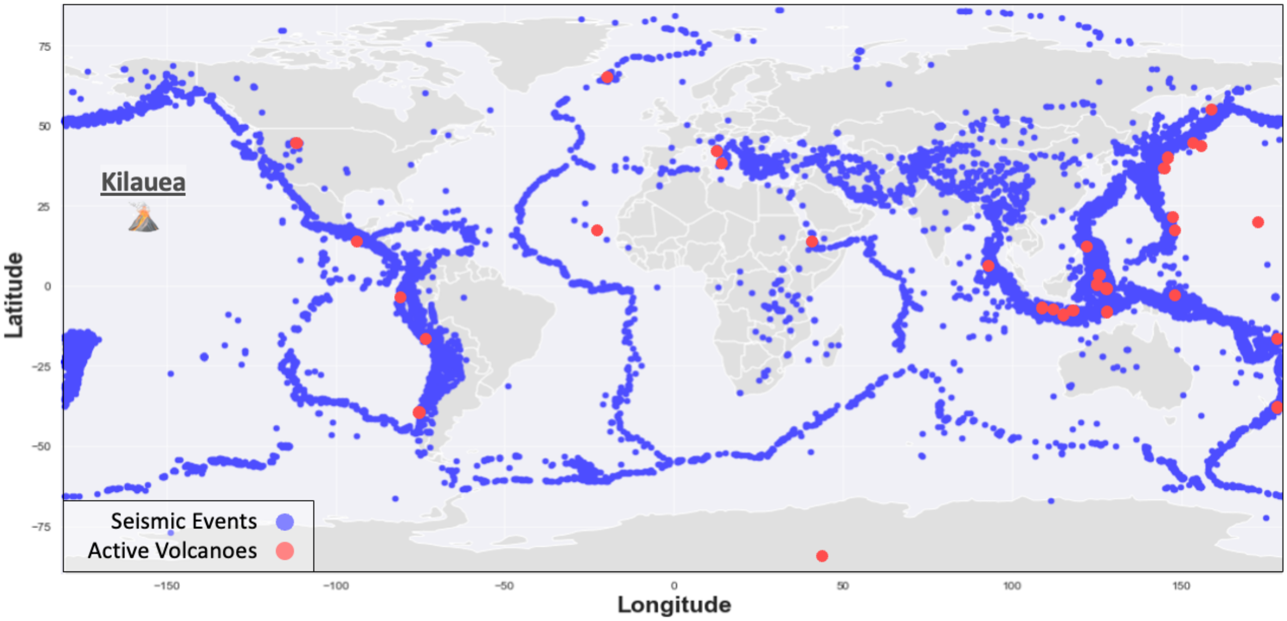

Figure 1: Global map of 20,853 seismic events (blue) and all major volcanic eruptions (red) from 1965 to 2018.

ABSTRACT

Using IRIS event catalogs from 1965-2016, we have compiled an unsupervised training set of 20,853 events for a global cluster classification model as well as a pre- and post- volcanic eruption event classifier using only two weeks of supervised training data on the big island of Hawaii. The global cluster model demonstrates discrimination of event clusters both in depth, thickness, density, and locations consistent with well-established seismic velocity discontinuities at global modeling scales such as AKI and IASP91. Variance sensitivity in the global model to random partitions of the training set up to 40% peaks at 1.75%. The Hawaii pre-/post- eruption event classifier demonstrates an equal or over-determined rank from data to model output. We believe this results in the very high score returns in the Hawaii model which suggests a lack of versatility in the method. Consequently, the Hawaii model needs need more supervised training sets to validate our overall goal in classifying an impending eruption event beyond our single case studies. Component analysis suggest event depth and magnitude to be a superior predictor of classifications over latitude longitude which was confirmed by scores in the final models training dimensions.

CSCI-5502, Boulder, Colorado, December-08-2022

https://doi.org/10.1145/globalseisclustering

ACM Reference Format:
Charles Hoots and Sai Meghashyeam Vangeepuram. 2022. Global Clustering of Seismic Event Data: Geophysical Mantle Strata and Eruption Prediction. In *Proceedings of **Data Mining Project (CSCI-5502)***. Boulder, CO, USA, 8 pages. https://doi.org/10.1145/globalseisclustering

1 INTRODUCTION

175 billion dollars (US) is spent every year across the globe in the mitigation or recovery from earthquake damage [5]. Entire arms of government and industry in all first-world countries are dedicated to using this cost in any currently known application spaces for studying earthquake phenomena and engineering mitigating strategies/infrastructure to the hazards associated with them.

This study demonstrates a computational and quickly deployable proof of concept that penetrates this unmet need in the protection industries of property and people in areas prone to seismic hazards. Our model represents a scale-able clustering algorithm for classifying a seismic event as either leading to an impending volcanic eruption or occurring after one. When deployed in an online training application design, we believe our algorithm will sustain predictive capabilities while it is continuously updated on the latest telemetries from a seismic monitoring array. The use case consists of providing our trained model event data in real time that it can classify as containing geophysical characteristics indicative of an impending volcanic eruption in the local terrane.

At a broader scale, to understand how to mitigate seismic hazards it is instrumental that we characterize the subsurface and underlying mantle where these seismic events are born. In addition to a predictive clustering algorithm, we direct the same model and methods to large-N decadal global seismic data sets. The goal in this application is to have the clustering algorithm find common behaviors in the event metadata indicative of discontinuities in lithosphere and upper mantle detectable by seismic instruments. We demonstrate this by showing with completely unsupervised learning and a total lack of apriori geophysical constraint on the solutions, our model can discriminate major seismic discontinuities that would otherwise only be detectable through computationally demanding processing of available metadata and its derivatives.

2 BACKGROUND

On April 30, 2018, a dike intrusion in the East Rift Zone (ERZ) marked the beginning of intense volcanic and seismic activity at Kilauea volcano, Hawaii [9]. On May 3, 2018, the dike intrusion triggered a major outpouring of lava in the ERZ. More than 1 cubic-km of lava was erupted, destroying hundreds of properties and putting thousands of civilians at risk [9]. Subsequently, a series of caldera collapses, explosions, and seismic events at the summit caldera occurred progressively and lasted for 3 months until August 2018 [2].

The Kilauea collapse sequence is the best-documented sequence in the world with the largest comprehensive data set [10]. This data set includes a multiparameter monitoring network, includ-ing ground deformation measurements with borehole tilt-meters, Interferometric Synthetic Aperture Radar (InSAR), Global Positioning System (GPS), Light Detection and Ranging (LiDar), and Global Navigation Satellite System (GNSS) [9]. The high-resolution volcano earthquake catalog provides a unique opportunity to better understand the spatial and temporal characteristics of seismicity during different stages of volcano deformation [10].

3 DATA

Our data sets consist of 53 years of global seismic records, from 1965 to 2018, and the dates and contemporaneous seismic events of all volcanic eruptions (fig. 2 over the same time frame, resulting in 23,853 unique samples across 4 dimensions (latitude, longitude, depth, and event magnitude (fig. 3).

Exploratory data analysis (EDA) illustrates an asynchronous and multi-modal distribution across both latitude and longitude with little persistant correlation between them (fig. 3). Depth and magnitude are both single-mode distributions with a highly persistent positive correlation between them (fig. 3). As a general rule, with greater depth into the subsurface we get larger magnitude earthquakes [12]. Figure 4 demonstrates the inter-relationships between depth and magnitude and their correlations behave with the density of the event data.

Here we see the component analysis of these two dimensions from the data set with respect to eachother produce new and more subtle density distributions. Depth now behaves along a bi-modal distribution where event density decreases synchronously with the simultaneous increase of both depth and magnitude along these two modes. This, coupled with the poor causal correlation between latitude and longitude, would suggest that using depth/magnitude together and removing lat/lon from the training sets provides the best method of training our clustering models in a manner that is the most consistent with distinguishable and persistent physical trends in the data.

4 METHOD

Our method approaches consisted of trying Neural Networks to impliment an SGD-based classifier, Knearest neighbors, Kmeans, and a density-based spatial clustering of applications with noise (DBSCAN) [3, 8, 13, 14]. Of all these methods, we decided to develop our models using DBSCAN for the global clustering model and Kmeans for the Hawaii predictive clustering model. The other alternatives did not produce model results distinct or superior to these two methods enough to warrant their computational demand during training, specifically with the neural network approach.

4.1 K-means

Our Hawai'i predictive clustering model uses the K-means algorithm to produce a binary classifier. K-means clustering is an unsupervised machine learning algorithm that is used to group data into clusters. Unsupervised learning is a type of machine learning where the algorithm is given a dataset that does not have any labels or categories, and the algorithm must discover the underlying structure of the data on its own. In the case of K-means clustering, the goal is to divide the data into groups (or clusters) in a way that maximizes the similarity within each cluster and minimizes the similarity between different clusters.

The K-means algorithm begins by randomly selecting a number of data points (called centroids) that will serve as the starting points for each cluster. The number of centroids is determined by the user and is specified by the parameter "k", which is why the algorithm is called "K-means". The algorithm then assigns each data point to the cluster with the closest centroid, based on a measure of similarity such as Euclidean distance. Once all data points have been assigned to a cluster, the algorithm recalculates the centroid of each cluster based on the data points assigned to it. This process is repeated until the centroids no longer move or change, at which point the clusters are considered to have converged.

One of the key advantages of the K-means algorithm is that it is relatively simple and easy to implement. It is also relatively fast and efficient, especially for large datasets, and can produce good results for a wide range of data types and applications. For example, K-means clustering can be used for data visualization, outlier detection, and market segmentation.A potential drawback of the K-means algorithm is that it relies on the selection of the initial centroids, which can have a significant impact on the final clusters that are produced. To overcome this, the algorithm can be run multiple times with different initial centroids, and the results can be compared to determine which configuration produces the best clusters.

We specify the number of clusters to generate. In the case of the Hawai'i model application the desired result is a purely binary classifier (pre- or post-eruption), thus our k=2. We initialize the

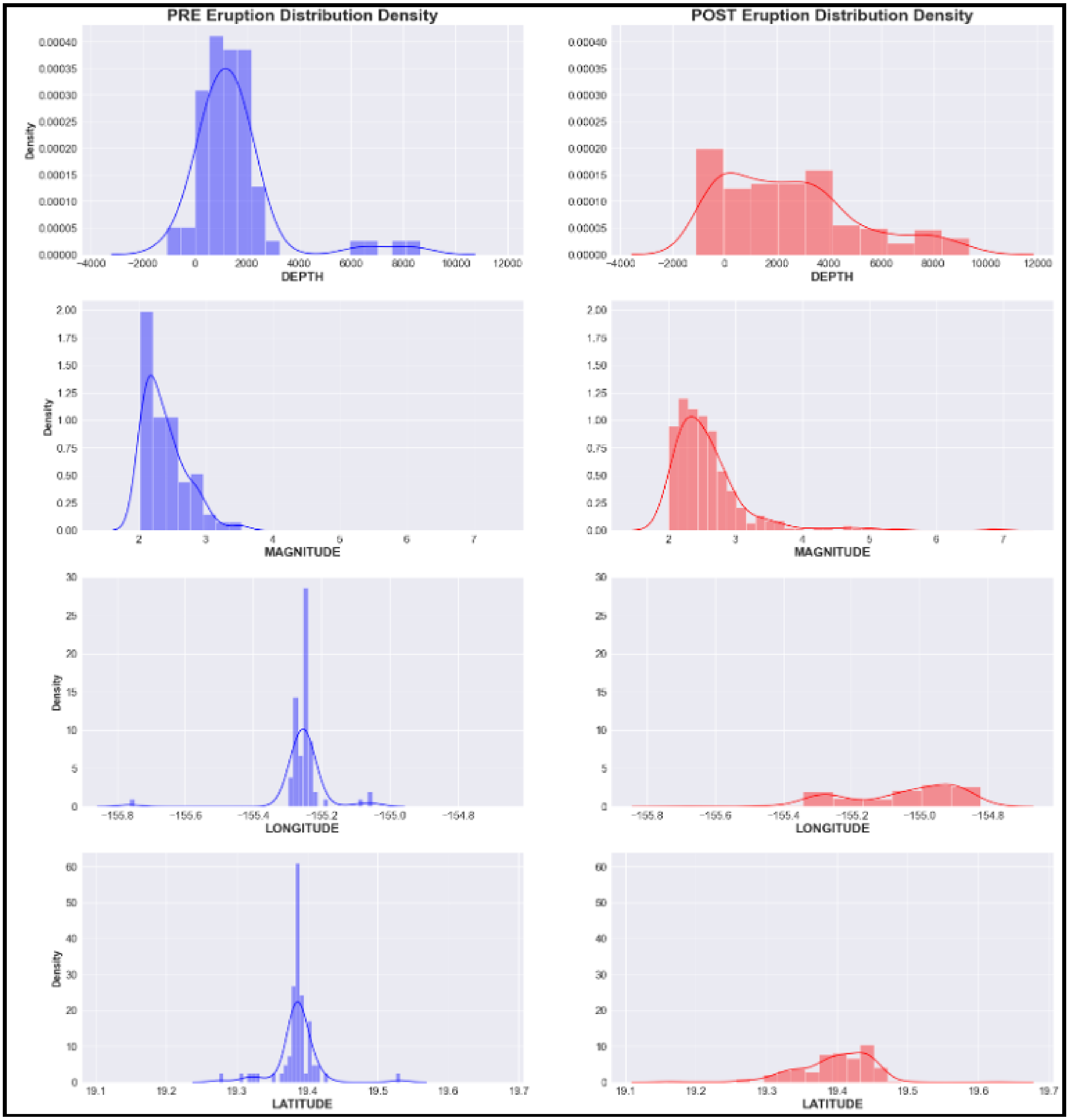

Figure 2: Hawai'i (left) pre- and (right) post- the 2018 Eruption seismic event latitude, longitude, depth, and magnitude distributions. The pre- and post- eruption distributions are windowed to the two weeks leading up to and the two weeks following the 2018 eruption event

centroids of each cluster by randomly selecting "k" data points from the dataset. These data points will serve as the starting points for each cluster. We then calculate the (Euclidian) distance between each data point and the centroids of each cluster, assigning each data point to the cluster with the closest centroid. Once all data points have been assigned to a cluster, we recalculate the new mean

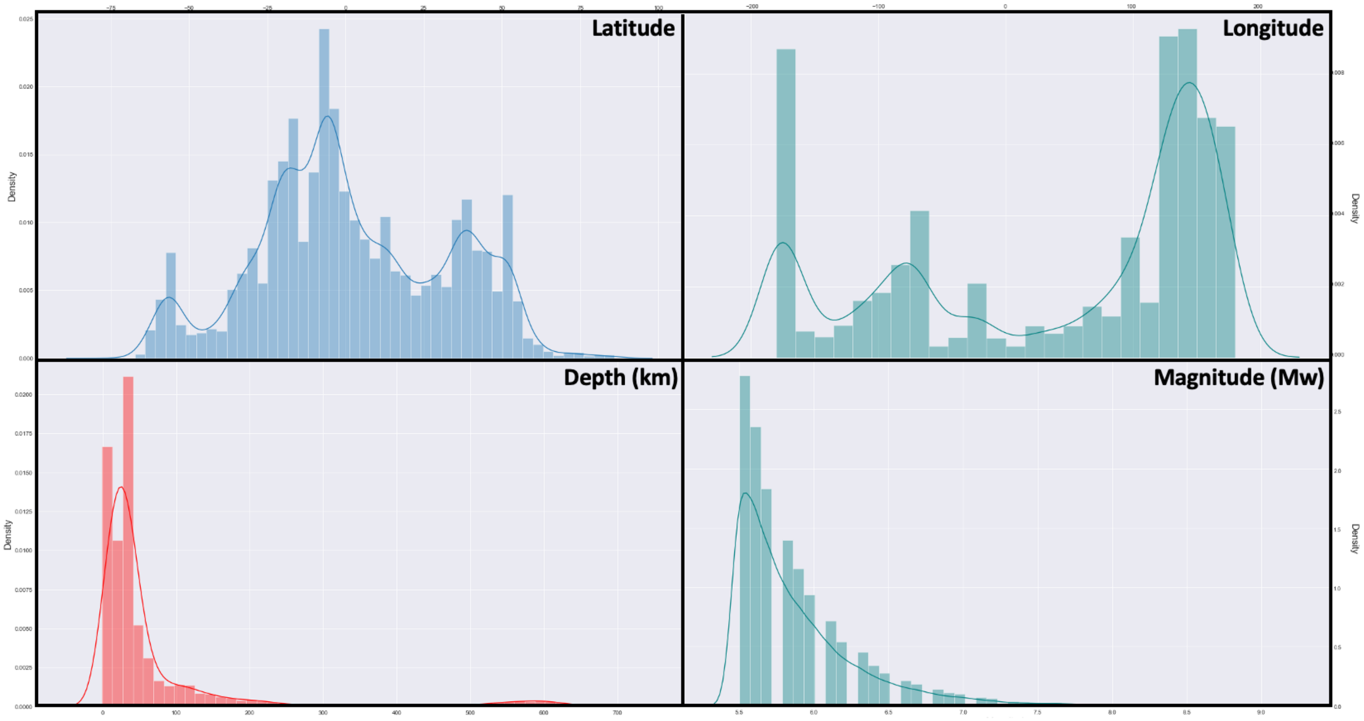

Figure 3: Global seismic event latitude, longitude, depth, and magnitude distributions.

centroid of each cluster based on the data points assigned to it. We then repeated the process up to this point until the clusters populations have converged in data-spatial distribution.

We then evaluated the quality of the clusters by using an ensemble bootstrap method on the silhouette coefficient across each split. This provided an accuracy score that reflect the overall level of a model being robust, meaningful, and well-defined.

4.2 DBSCAN

For the global seismic event clustering model, we implemented the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) clustering algorithm. DBSCAN is a density-based clustering algorithm that is used to group data into clusters. Unlike other clustering algorithms such as K-means, which use a predefined number of clusters and a measure of similarity to group data points, DBSCAN uses the density of data points in the space to determine the clusters [13].

The DBSCAN algorithm has two key parameters: the maximum distance, λ that a data point can be from its nearest neighbor to be associated with a cluster and the minimum number of data points p needed to be within that distance of a point for that point to be considered a member of the cluster. The algorithm iteratively redefines cluster center candidates from the 23,853 data points as a clusters sphere of influence is continuously re-evaluated to include points that are within λ distance of the at least p neighbor points with respect to each member of the cluster. This is iterated until every data point in the training set is collected by a cluster [4, 13].

One of the key advantages of the DBSCAN algorithm is that it does not require the user to specify the number of clusters in advance. Instead, the algorithm automatically determines the number of clusters based on the density of the data points in the space. This makes DBSCAN a good choice for datasets that may have variable numbers of clusters or complex cluster shapes. Another advantage of DBSCAN is that it is able to identify outliers or noise in the data, which are data points that do not belong to any cluster. These points are considered to be part of the cluster "noise" and are not assigned to a cluster. This can be useful for identifying and removing anomalous data points from the dataset [13].

We finish by quantifying the uncertainty in our model using an ensemble bootstrap method with a scoring function that is primarily indicative of the unsupervised training sets sensitivity to the removal random subsets. This we found to be the Euclidian between each point and its current cluster center after every fit over 1000 individual test-train splits. This is a superior metric to sum of squared distances (SSD) or the silhouette coefficient as volume and data variance of depth and magnitude is wide enough that we need to down-filter overall single sample dependence for a robust, meaningful, and well-defined model when complimentary ground-truth is not possible for the dataset [15, 16].

5 RESULTS

We illustrate the accuracy score for the Hawai'i model results (fig. 5) in a Receiver Operating Characteristic (ROC) curve of showing the frequency magnitude in true- and false- positives in the predictions made by the classifier at increasing probabilistic thresholds 6.

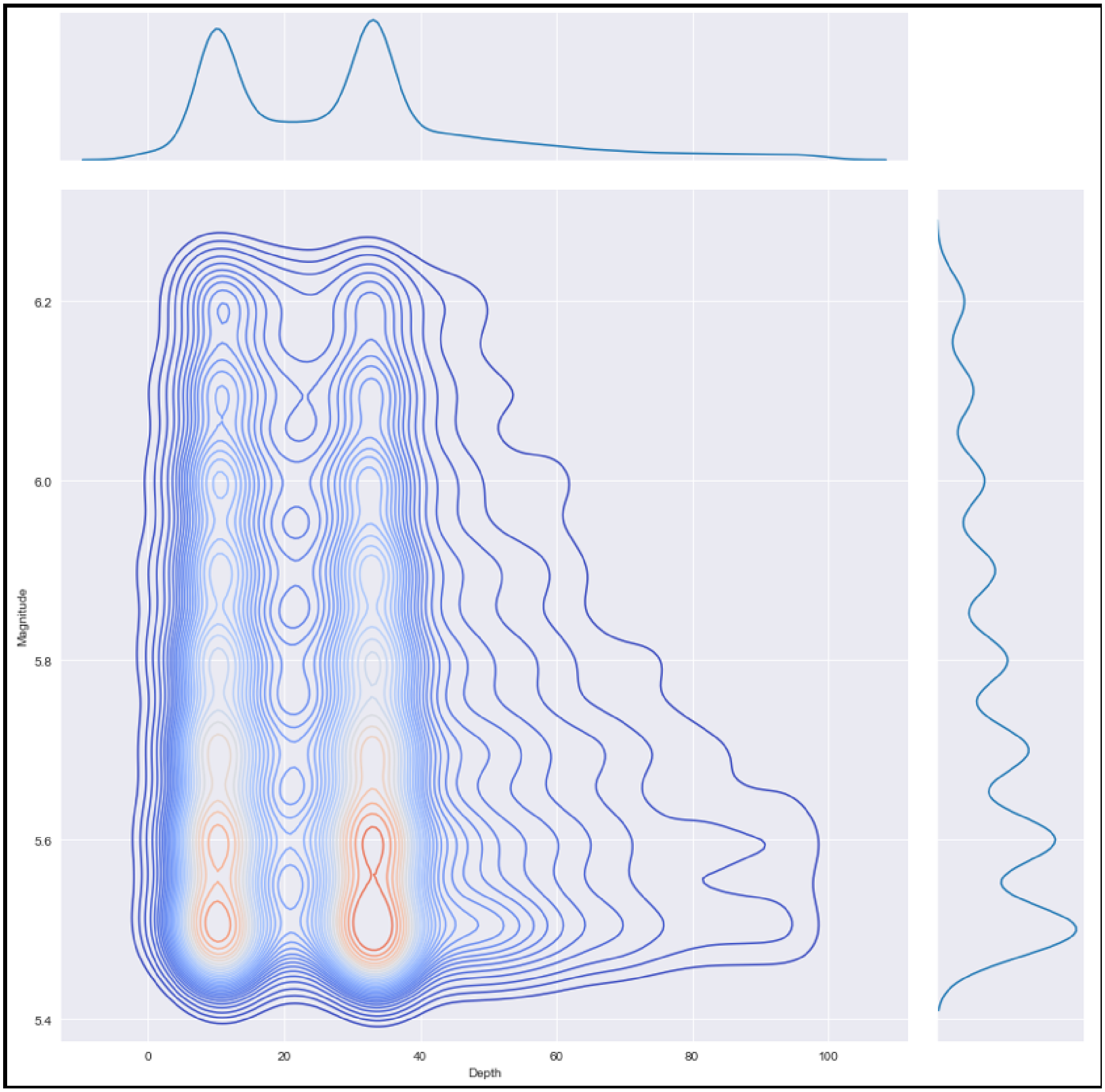

Figure 4: Deconvolved depth and magnitude event density distributions wrt to eachother in the global data set. Each isoline is an increase in density by 5%. The subset figures at the top and right are event density in terms of strictly depth or magnitude, respectively. A clear bimodal distribution of event density if governed by event depth and a tapering multi-modal density distribution is demonstrated with increasing magnitude. This tapering behavior seen in Mw coupled with the bi-modal behavior seen in depth provide an effective indicator for our classification models.

Given all but two samples (fig. 6-Confusion Matrix) were correctly classified by the model, we see the ROC curve demonstrates the model as a near perfect classifier for this eruption case.

As described in 4.2, since the global seismic clustering model has no true labels to evaluate actual accuracy residuals, we comprised with a measure of model stability and independence from high-frequecny or high variance sample anomalies. Shown in fig. 7 is this variance stability across a model ensemble size of 100 test-train splits each with a normally distributed random split proportion with split partitions at 20-50% of the total data volume. The variance oscillation across the entire model ensemble averages less then 0.3%. Coupled with consistency with other published models on seismic discontinuities (6) this would indicate a stable evaluation for the global clustering model.

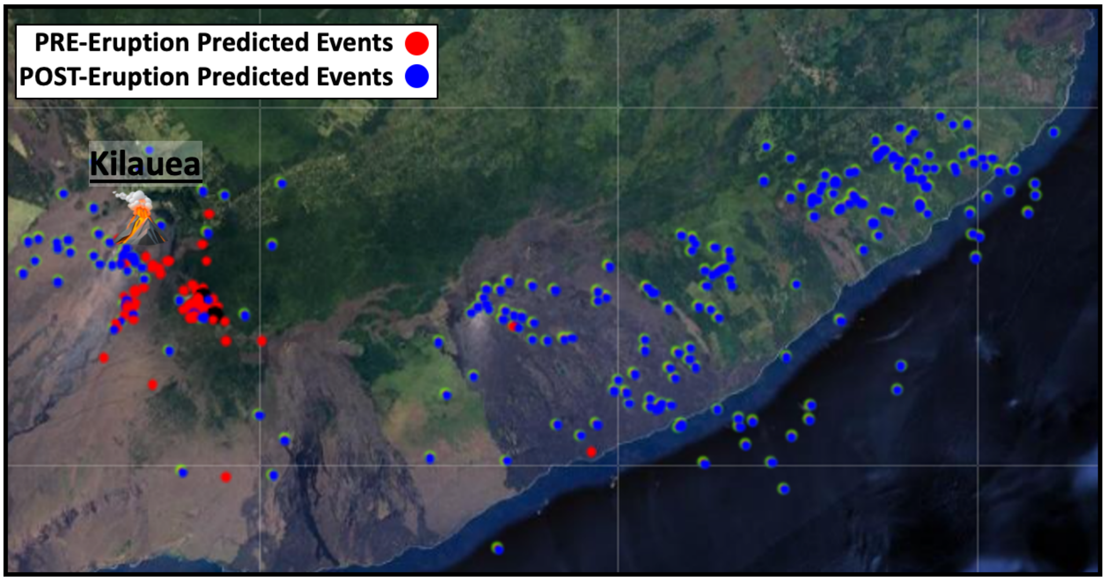
PRE-Eruption Predicted Events ●
POST-Eruption Predicted Events ●
Kilauea

Figure 5: Hawaii 2018 Eruption cluster classification model. Model was trained on 2-weeks worth of labeled event data before and after the 2018 eruption. Shown here are the model output labels for pre-eruption events (red) and post-eruption events (blue). All but two events (black) were correctly classified by the model.

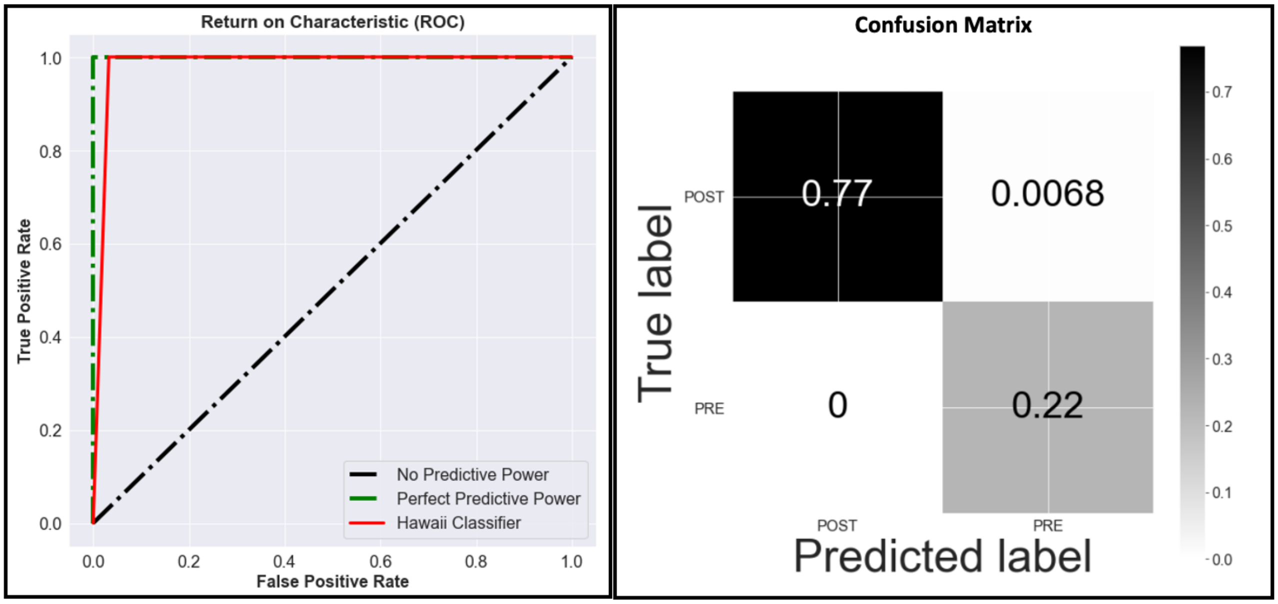

Figure 6: Left, ROC curve for the Hawaii binary classifier cluster model. A perfect model (green) will have FPR=0,TPR=1.0 for all probability thresholds. A completely ineffective model will have FPR=TPR for all thresholds. Right, confusion matrix displaying the distribution of the cluster model to correctly predict the two clusters, pre- and post- eruption

6 DISCUSSION

Figure 8 is the Global Seismic Event Clustering model in a map profile view. There is a highly pronounced lateral homogeneity at sea level greatly indicating a single geophysical distinction in the training data for events at very shallow depths.

Figure 9 is shows the model in a depth profile and is arguably much more interpret-able as it very strongly follows multiple published studies on seismic discontinuities in the lithosphere and upper mantle, e.g. [6, 11]. Here we see clustering model distinguishing the top layer consistent with average lithospheric thickness everyone on the globe [11] between 30km in oceanic basins

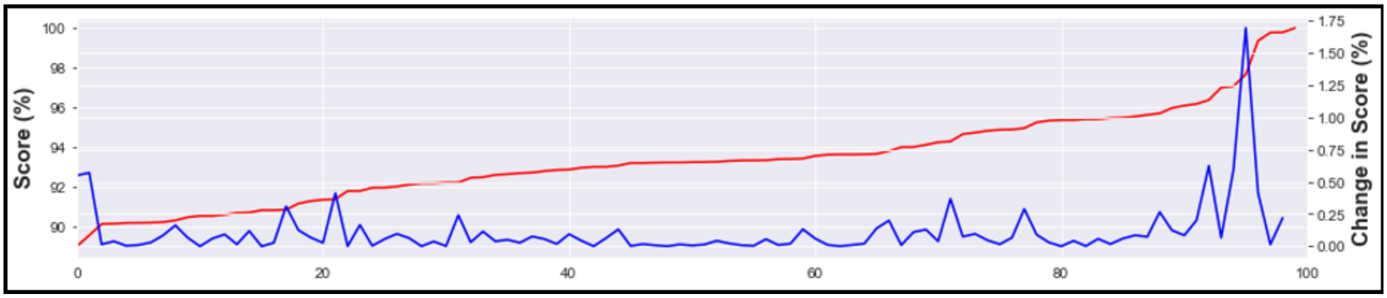

Figure 7: Global cluster ensemble modeling normalized variance of scores (red) and their changes between subsequent tests (blue). The scoring function is a predicted labels distance to the center of its predicted cluster. Ensemble model size includes 100 test-train splits in random proportions from 20-50% of the data volume. For unsupervised cluster models, a stable population of training sets will suggest the overall data set is mostly insensitive to test-train splits when the change in (sorted) variances is very small.

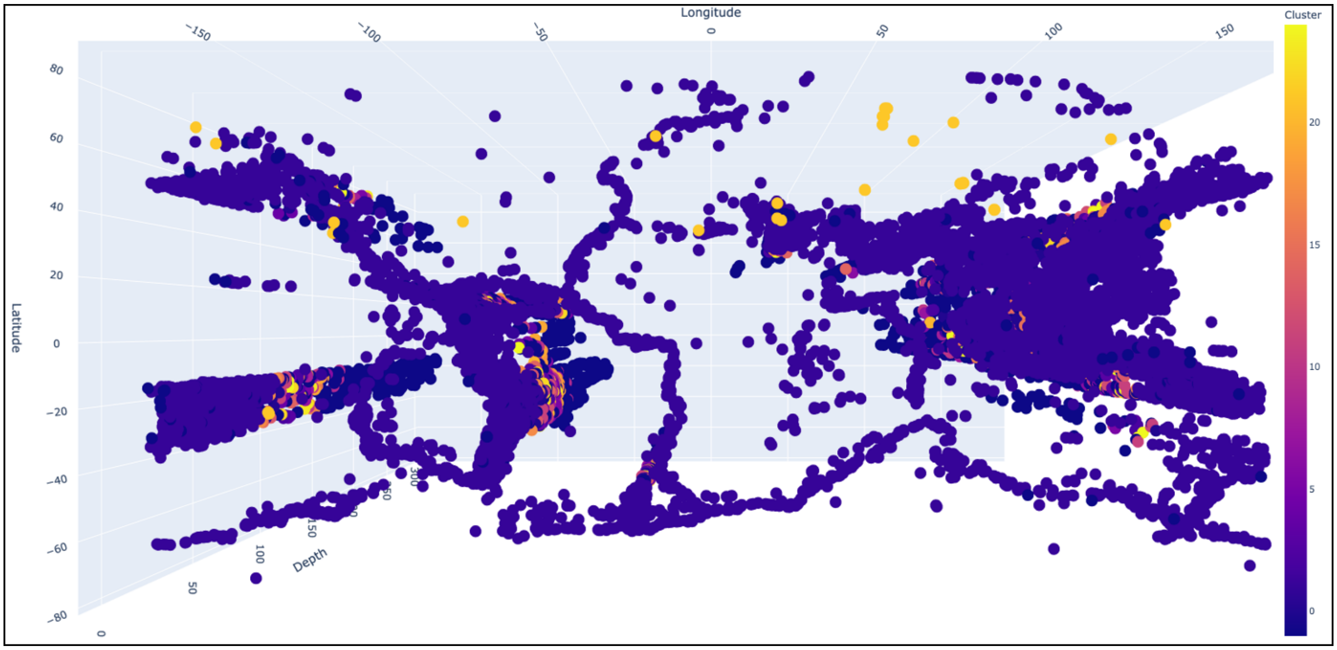

Figure 8: Global clustering model in map profile.

to as high as 65km thickness in the Himalayas. The laterally homogenous but vertically heterogenous distinct set of vertically thin cluster layers all occur across the average depth of the most pronounced and greatest change in upper mantle rheology on Earth, the Lithosphere-Asthenosphere Boundary (LAB). This seismic discontinuity the model effectively identifies in clustering is the base of the continental lithosphere that transitions into the upper mantle. The vertically distinct clusters that dive into the upper mantle 250km all occur in the model at the same positions of subduction margins where oceanic lithosphere is assimilated into the upper mantle, ie [11].

7 CONCLUSION & FUTURE WORK

We have exploited both large and small data sets using DBSCAN and K-means clustering techniques to produce a global clustering model of major seismic discontinuities in the Earth's lithosphere and upper mantle and a predictive classifier for events that lead to an impending eruption on the big island of Hawai'i, respectively. While we hope this represents an effective demonstration on the proof of concept that event metadata can be used in imaging constraints of geophysical mantle strata, a more robust treatment of uncertainty quantification and likely the graduation to a computationally expensive neural network is the next approach in refining the method demonstrated. Likewise, while the Hawai'i predictive model has a near perfect return on accuracy residual, it may indicate a potential lack of versatility and failure of the model to predict

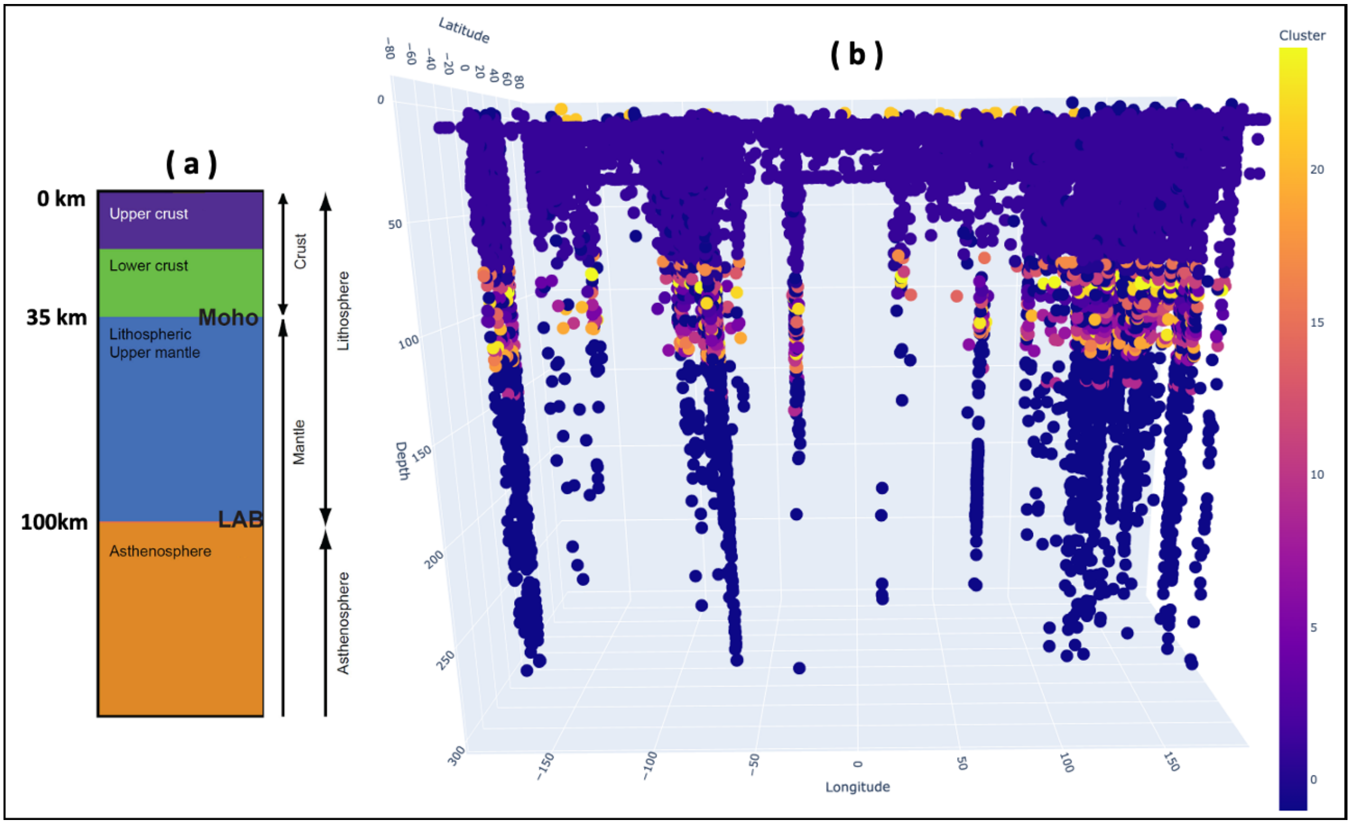

Figure 9: Global clustering model in depth profile. Subset figure on the left is a conceptual map of seismic discontinuities from [11].

distinct distributions of new training or high σ point anomalies requested from the classifier. This is likely the greatest weakness in our proof of concept that we will need to mitigate through a thorough application of the methods demonstrated here on a far larger ensemble set of eruption case studies beyond just the 2018 eruption of Kilauea.

REFERENCES

[1] Keiiti Aki and Paul G. Richards. *Quantitative Seismology*. W.H. Freeman and Co., 2nd edition, 1980.

[2] Kyle R Anderson, Donald A Swanson, Christina A Neal, Ingrid A Johanson, Matthew R Patrick, Brian Shiro, Weston A Thelen, Asta Miklius, Peter F Cervelli, Emily K Montgomery-Brown, et al. The 2018 summit eruption and caldera collapse at kilauea volcano, hawaii. volume 2018, page V41B–03, 2018.

[3] K. L. Du. Clustering: A neural network approach. *Neural Networks*, 23:89–107, 1 2010.

[4] Anil K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31:651–666, 2010.

[5] Aditi Kharb, Sandesh Bhandari, Maria Moitinho de Almeida, Rafael Castro Delgado, Pedro Arcos González, and Sandy Tubeuf. Valuing human impact of natural disasters: A review of methods. *International Journal of Environmental Research and Public Health*, 19, 2022.

[6] R. Kind, X. Yuan, J. Mechie, and F. Sodoudi. Structure of the upper mantle in the north-western and central united states from usarray s-receiver functions. *Solid Earth Discussions*, 7:1025–1057, 2015.

[7] Qingkai Kong, Daniel T Trugman, Zachary E Ross, Michael J Bianco, Brendan J Meade, and Peter Gerstoft. Machine learning in seismology: Turning data into insights. *Seismological Research Letters*, 90:3–14, 11 2018.

[8] Aristidis Likas, Nikos Vlassis, and Jakob J Verbeek. The global k-means clustering algorithm, 2003.

[9] C A Neal, S R Brantley, L Antolik, J L Babb, M Burgess, K Calles, M Cappos, J C Chang, S Conway, L Desmither, P Dotray, T Elias, P Fukunaga, S Fuke, I A Johanson, K Kamibayashi, J Kauahikaua, R L Lee, S Pekalib, A Miklius, W Million, C J Moniz, P A Nadeau, P Okubo, C Parcheta, M R Patrick, B Shiro, D A Swanson, W Tollett, F Trusdell, E F Younger, M H Zoeller, E K Montgomery-Brown, K R Anderson, M P Poland, J L Ball, J Bard, M Coombs, H R Dietterich, C Kern, W A Thelen, P F Cervelli, T Orr, B F Houghton, C Gansecki, R Hazlett, P Lundgren, A K Diefenbach, A H Lerner, G Waite, P Kelly, L Clor, C Werner, K Mulliken, G Fisher, and D Damby. The 2018 rift eruption and summit collapse of kx12b;lauea volcano. *Science*, 363:367–374, 2019.

[10] Angie D. Ortega-Romo and Xiaowei Chen. Spatiotemporal clustering of seismicity during the 2018 kilauea volcanic eruption. *Geophysical Research Letters*, 48, 4 2021.

[11] Jeroen Ritsema. Global seismic structure maps. *Ritsema, Jeroen. "Global seismic structure maps." SPECIAL PAPERS-GEOLOGICAL SOCIETY OF AMERICA*, 11:388, 2005.

[12] Peter M. Shearer. *Introduction to Seismology*. Cambridge University Press, 2009.

[13] Jianbing Shen, Xiaopeng Hao, Zhiyuan Liang, Yu Liu, Wenguan Wang, and Ling Shao. Real-time superpixel segmentation by dbscan clustering algorithm. *IEEE Transactions on Image Processing*, 25:5933–5942, 12 2016.

[14] Bing Shi, Lixin Han, and Hong Yan. Adaptive clustering algorithm based on knn and density. *Pattern Recognition Letters*, 104:37–44, 3 2018.

[15] Qi Wang, David D. Jackson, and Jiancang Zhuang. Missing links in earthquake clustering models. *Geophysical Research Letters*, 37:1–5, 2010.

[16] Ilya Zaliapin and Yehuda Ben-Zion. Earthquake clusters in southern california i: Identification and stability. *Journal of Geophysical Research: Solid Earth*, 118:2847–2864, 2013.