Aftershock Analysis of The Darfield, New Zealand Earthquake (2010)

1st Surafel Tilahun
Department of Computer and Mathematical Science
Auckland university of Technology
Auckland, New Zealand

July 1, 2020

Abstract

Earthquake is a sudden vibration of earth surface that happens due to the release of internal energy to the surface of the earth. Predicting earthquake appears to be a challenging task. However, we can attempt to minimize the consequence and destructions resulting from earthquake by conducting aftershock analysis. In this paper, we analyze the seismic activities of Darfield earthquake that occurred in Christchurch, New Zealand (2010). We implement a combination of unsupervised and supervised machine learning algorithm to find where in space, the epicentre of the earthquake in relation to the clustering activity lies. SOM, K-means and agglomerative hierarchical clustering algorithm were adopted, along with one supervised machine learning algorithm (Naive Bayes). The result from our study shows that K-means algorithm outperforms other clustering techniques. Utilizing the clusters secured by K-means algorithm, we apply Naive Bayes algorithm to classify magnitudes of Darfield earthquake by converting the clusters into classes. Finally, we asses the classes that are produced from Naive Bayes algorithm to detect the epicentre of Darfield earthquake. Our results reveal that class 4 contains the main shock (epicenter) of Darfield earthquake, lying at the intersection of 172.1679 longitudes and -43.52731 attitude. The outcome of this study can be exploited to explore the fault line. This is advantageous to insight if these aftershocks could result in triggering another similar earthquake near this area.

0.1 Introduction

Earthquake is one of the most catastrophic natural disasters that is triggered due to unexpected release of energy in the earth's crust. The strain in the earth's crust can be released at any point of time without warning in areas all around the world. [6] stated that the earth experiences to the minimum of 5000 earthquakes every year, although most of these earthquakes are trivial and not noticeable without scientific experiments. The consequence of earthquakes can be devastating when they strike in built-up areas. The collective impact of earthquake includes damage to buildings, loss of lives and several infrastructures, whilst they tend to cause less damage if they strike in open and remote areas.

A point where the rock starts to fracture and the earthquake is initiated inside the crust is known as focus. The point directly above this place on the surface of the earth is called epicentre, also known as main shock. Aftershock is the seismic activities related to the main shock. The point where this main shock occurs is called fault, which represented as a thin block area that separates the blocks of the earth's crust. When earthquakes happen to strike on fault, one side of the fault slides against the other side of the fault, possibly causing volcano or tsunami.

Even though predicting the forthcoming earthquakes strike (main-shock) is an arduous task to perform, it is undoubtedly feasible to insight the post-strike scenario in order to find spaciotemporal patterns of seismic activities, and this process is known as aftershock analysis. According to [18], aftershock provides information about the association between the earth's crust and substantial magnitude, shedding light on it is becoming the most common research topic in recent time. Aftershock is substantial to gain a spaciotemporal understanding of critical parameters of aftershock activity, which are magnitude (intensity) and depth (i.e. at which depth measured from the earth's surface) at which the shock occurred.

This paper considers the Darfield earthquake (DE) that occurred in Canterbury, New Zealand on 4 September 2010. [4] reported that although the shock did not result in loss of lives, it nevertheless caused a considerable amount of damages to buildings and substructures of the city. It was the most damaging earthquake in New Zealand history, since the earthquake that happened in Hawkes Bay in 1931. It has been reported that DE showed the most considerable intensity of 7.2 and the epicentre was 37 km west of Christchurch, which is 47.1 km far from the town of Darfield.

[3] mentioned that earthquake clustering is an essential analysis technique to specify the frequency of seismic activities between small magnitude cutoff and large magnitude that is close to 8. Hence, the aim here is to track the magnitude and depth of DE on longitude and latitude (gain spatial understanding) searching for the epicentre through both supervised as well as unsupervised machine learning algorithms. We proposed three unsupervised machine learning algorithms to cluster magnitude and depth of earthquake on latitude and longitude. The three algorithms that are applied to mine the underline parameters are a self-organized map (SOM), K-means clustering and agglomerate hierarchical clustering algorithm. For the sake of comparison, initially, we implemented SOM so that we can determine the potential number of clusters where the feature should be grouped and we keep it same across the three types of algorithms. Their performance is compared on the basis of visualization and clustering accuracy

measurements, namely the average sum of square error and cluster silhouette measure. After finding the best clustering algorithm, the clusters are converted into classes and supervised learning algorithm (Naive Bayes classifier) is employed in order to find the space where the largest shock occurred in latitude and longitude. This paper addresses a research question of **where in space (latitude and longitude) does the highest DE lie?** We assume answering this research question will help us to comprehend the Darfield seismic activity and guide during decision-making process. The results from this paper also can be used for disaster mitigation and evacuation because if these clusters of aftershock are on fault, they might cause another earthquake abruptly.

There exists a considerable body of research concentrating on aftershock analysis. [15] performed earthquake clustering analysis on the earthquake that occurred in India, between January 2005 and December 2015. Furthermore, [14] analyzed Bengkulu earthquake that occurred on 12 September 2007 with a strength of 7.9. [19] performed aftershock analysis of Gorkha-Dolakha (Central Nepal) earthquake doublet in 2015. Also, [20] presented agglomerative clustering algorithm to analyze the local cluster structure in the data space of seismic events for the two types of catalogs. Yet another paper by [26] investigated and reported Pakistan seismic activities with K-means clustering algorithm approach. This paper presents a comparison seismic clustering analysis and makes additional contribution towards this particular area.

0.2 Data

The data is derived from earthquake catalog data, from GNS Science New Zealand and can be found at [24]. The data comprises 22 variables. It data is preprocessed in order to extract the relevant variables from the data set and of these 22 variables, only four are considered for this paper, latitude, longitude, magnitude and depth.

Figure 1 shows the correlation among the corresponding variables. There is almost zero correlation between latitude and magnitude, and longitude and magnitude. Hence, the magnitude of the DE does not seem to have relationship with the locations. However, it tends to have positive relationship with depth of DE, even though the the correlation coefficient is less than zero. Magnitude and depth are independent of the space location. Hence, we presume these features are independent, which means they are less likely to be clustered together. However, magnitude and depth can be clustered individually.

In this study, we compiled aftershock occurred between 171.3503 and 173.1961 on longitude and -43.56247 and -42.6680 on latitude. The maximum and minimum magnitude values are 7.2 and 1.737, respectively.

Figure 2 depicts the DE that lies within the range of latitude and longitude mentioned above. From the map plot, it is clear to see that there is considerable intensity between the latitude of -43.7 and -43.5; hence, we must expect the epicentre to fall in within this range. The magnitudes outside of this range on latitude are considered to be outliers.

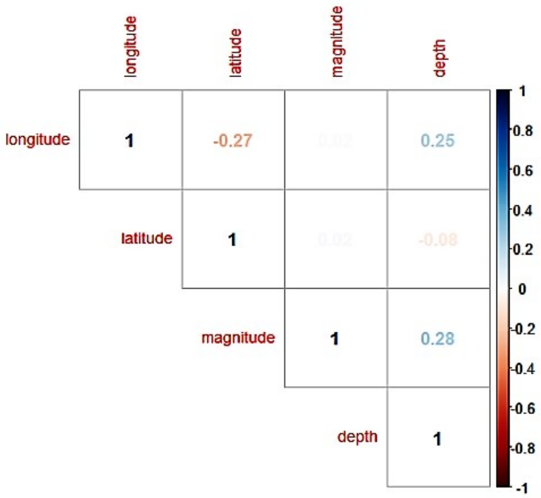

Figure 1: Correlation between latitude, longitude, magnitude and depth.

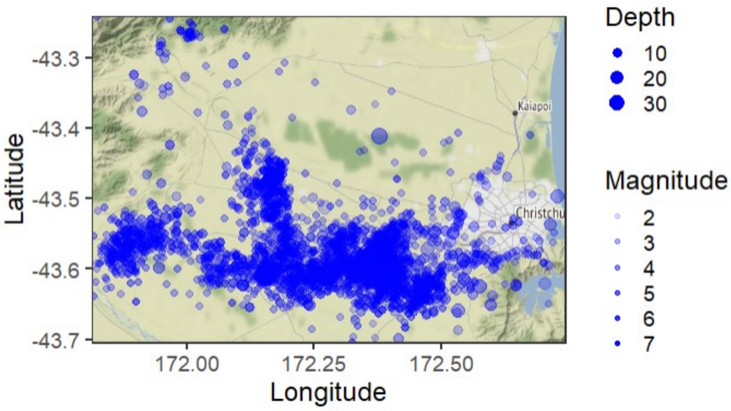

Figure 2: DE and related seismic activities, Christchurch 2010.

0.3 Mining algorithms

In this section, we provide a brief description of the mining algorithms that have been considered in this paper. Three unsupervised and one unsupervised machine learning algorithms are regarded, namely K-means, SOM, agglomerate hierarchical and Naive Bayes.

0.3.1 K-means

May be one of the most widley adopted unsupervised machine learning algorithms to perform clustering [9]. K-means clustering involves dividing the data space into K prototype. At first, arbitrary position of the K-prototypes are allocated before pooling data into K groups using nearest neighbourhood allocation. Then, the algorithm works iteratively to find the centroid of the K clusters based of feature similarity then provides labels for the training data. while this powerful clustering model is implemented, it still bears few disadvantages in terms of clustering performance.

Disadvantages

Disadvantages of K-means clustering is that it requires the users intervention in order to choose the proper number of clusters. Another drawback of this algorithm is that the centroids are arbitrarily and simultaneously as a point in space. That is, centeroids for each cluster have to be recalculated, which results in affecting the efficiency of the algorithm, while dealing with large data set. Furthermore,[8] suggested that K-means does not perform well while it is employed in data set with clusters of random shape or different sizes.

Advantage

The main benefits of K-means clustering is:

- Easy and simple to implement [7].
- It is fast.

0.3.2 SOM

SOM is a family of artificial neural network , that is categorized as unsupervised machine learning algorithm [2]. It is an essential algorithm to visualize nonlinear relations of large dimensional data set by mapping it into two-dimensional plot. SOM comprises two-dimensional grid that represent number of neurons. These neurons are mostly arranged in two ways; hexagonal or rectangular. The distance between these neurons is crucial for the learning process and the prototype vector is related to them, which is a vector that has same dimension as the input data set. This prototype vector approximates a subset of sample vector. In the context of SOM, features or number of columns in the data set are considered as input dimension, ideally greater than 2, while the number of lattices is called output dimension.

During the training process arbitrary sample vectors are selected and weight is assigned to them, then the selected sample vectors are allocated to the most similar prototype vector, known as best matching unit (BMU). The learning process continuously adjusts the codebook vector to match the samples. In other words, the input dimensions that are relatively close in dimensional space are mapped in the grid that are close to each other in the SOM lattice. This step is achieved by an iterative process so that the training algorithm updates the codebook. Eventually, the model creates K number of prototypes according to the association among the input space. The result is well represented in u-matrix plot that indicates the distance between all neighbouring nodes with regards to colour coding [5]. U-matrix serves to determine what potential clustering there are in the SOM lattice it self .

The main advantage of SOM is it is easily interpreted and understood. It is also beneficial for dimentionality reduction.

The disadvantage of SOM is that the analyst must specify the number of clusters. According to [11], the number of neurons can be determined by the following equation.

$$M = 5 * \sqrt{(N)},$$

where N is number of observations (entities) and M is number of number of neurons. However, it is claimed to be a hot topic that require further study. Further problem with SOM is that it demands sufficiently large data so that it can yield meaningful clusters [1]. Another shortcoming of SOM is that there is no certain rule of thumb for determination of learning rate and neighbouring function.

0.3.3 Hierarchical

The aim of hierarchical clustering is that not to obtain a single clustering of the data, instead hierarchy of clusters that is likely to disclose interesting pattern that is disguised in the data. The most widely implemented hierarchical clustering technique is agglomerate clustering [20].

Initially, this technique starts by treating the individual data points as its own cluster, before it gradually connects associated samples to form new clusters [10]. This process is achieved by accessing proximity matrix (used to depict distance between clusters). The hierarchy of connections observed in this step is usually visualized with dendogram and it must be cut based on analysts interest in order to separate into cluster groups. The dendogram that is obtained after cut represents distance between clusters and subclusters.

The advantage of this clustering technique is easy to interpreted, understand and implement. It is also fast clusterer. Furthermore, the result is ordered in hierarchical order, which make the algorithm more informative. Therefore, it is easier to choose the number of clusters by relying on the dendogram.

On the other hand, the restraint of this technique is that most hierarchical clustering do not cope up with missing data [21]. The author also quoted that while using such technique, it is necessary to specify both the distance metric as well as the linkage criteria and there is not sufficient theoretical basis for such decisions.

0.3.4 Naive Bayes

Naive Bayes is a type of supervised machine learning algorithm. It is one of most frequently used classifier that works well with large data set [17]. Naive Bayes is a simple classifier algorithm that relies on theorem of probability known as Bayes' theorem or conditional theorem and strong (naive) independence assumptions among attributes [12] and [23]. It performs well for categorical input variables compared to numerical variables, otherwise in the case of numerical input variables, a normal distribution or bell curve is assumed [16]. However, if Bayes is considered to cluster contentious input variables, the features must be converted into a nominal or categorical variable. One simple way to achieve this is by binning the entities or creating different subsets of entities.

Naive Bayes is capable of calculating the probability of different output classes given our hypothesis, under the assumption of all features are independent. Nonetheless, it is not true all the time because some features might not be independent at all.

The merit of Naive Bayes is that it is fast, accurate and less sensitive to missing values and even data imbalance problems [22].

0.4 Accuracy Metrics

This sections is in consists of description of the accuracy metrics we have implemented to assess the quality of clustering algorithms. To distinguish between these three algorithms, silhouette and average sum of square (SSE) were used.

0.4.1 Silhouette

Silhouettes is an accuracy metric that provides a graphical representation of how well the data point was clustered while it was allocated to a specific cluster [25]. It also enables us to define the quality of clusters in terms of number (silhouette score) that lies in the ranges [-1,1]. A closer silhouette score to one indicates a better clustering quality of the model. Silhouette can be implemented in two different ways. One is to acquire an optimal number of clusters during K-means clustering analysis, while the second one is to compare the quality of multiple clustering algorithms. That is, while different clustering algorithms are applied with same number of clusters, their quality is benchmarked with silhouette score (higher the value better is the model). The first step of silhouette is finding the distance between a data point and other data points in a same cluster. Then, it calculates the distance between from initially observed distance value to other data points in other clusters then divide it by the maximum value of these values to normalize the result. The down side of this method, however is that as the data size is larger, it gets expensive in time because it has to go through each data points in searching for the distance among them.

0.4.2 Sum of Square Error (SSE)

SSE is another guidance metric to choose the number of clusters to apply during the segmentation of our data. This method count on calculating the distance between a data point and other data points with in the same cluster. SSE can be implemented in two different ways to asses the quality of clusters. In other words, the error with in cluster (SSEw) and error between cluster (SSEb). It is mostly used to determine the number of clusters for K-means algorithm (find the trade of between SSEw and SSEb clusters) this significant point is known as ”knee” [13]. While this method is easy and relatively faster, it faces some restrictions in regards to producing accurate result. Put another way, while the number of clusters increase, SSEw become smaller because it assumes each data point as its own cluster thus it might be miss leading for obtaining smaller SSE.

0.5 Result

In this this section, we will illustrate some experimental results that are obtained from the three clustering algorithms(SOM, K-means and hierarchical), along with the result from classifier algorithm (Naive Bayes). We will also include discussions of the results for each algorithm.

0.5.1 SOM

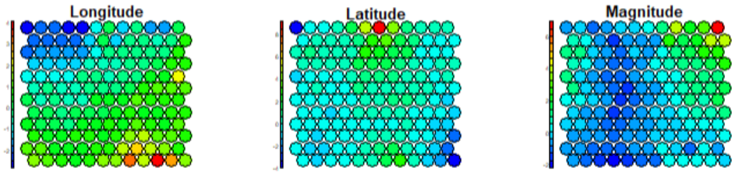

Figure 3: Component plane heatmaps for latitude, longitude and magnitude.

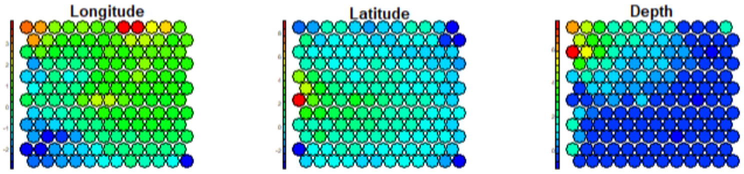

Figure 4: Component plane heatmaps for latitude, longitude and depth (LLD).

We applied SOM with grid size of 12*12 [11]. Figure 3 and figure 4 are component plane heatmaps for each feature that fed into SOM. The figures show the relationship among the corresponding variables. The similarity of heatmap in different component planes denotes association or correlation of variables. Figure 3 illustrates the relationship between longitude, latitude and magnitude (LLM). There is no strong similarity among these components authenticating our assumption in section 3 above. In addition, figure 4 also exhibits week correlation between latitude, longitude and depth (LLD). This suggests that magnitude and depth have negligible correlation with latitude and longitude.

Figure 5 is U-matrix and it shows the distance between each node encoded with different colours according to their neighbour nodes distance. Light blue colour in the U-matrix means small distance between neurons, while dark blue portray larger distance.In block A, it is shown that there are four possible number of clusters represented

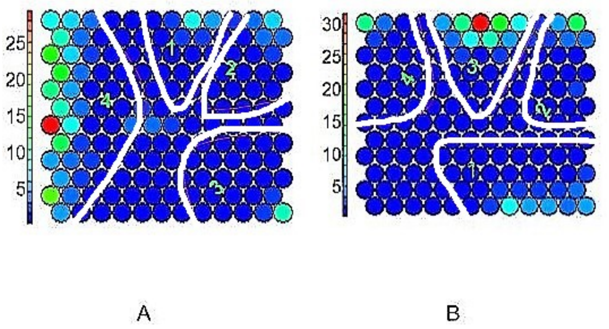

Figure 5: U-matrix of magnitude and depth: A is u-matrix of LLD and B is u-matrix of latitude, longitude and magnitude (LLM).

with a light blue colours (marked with white lines). Similarly, block B suggests four clusters in LLM.

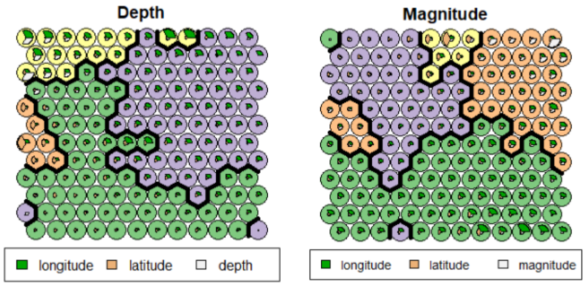

Figure 6: Clusters of LLM and LLD by SOM: Left block is LLD and right block is LLM.

With the help of hierarchical clustering, we extracted the associations represented in the U-matrices (figure 5) to produce meaningful clustering visualization as shown in figure 6. Figure 6 resembles the U-matrix shown in figure 5. This plot is ”fan diagram” of the weight vectors. As it is shown in the diagram, The weight vector in each node is represented by three different colours (green,brown and white). Hierarchical clustering uses the vector codes in figure 5 to inspect the association between each weight vectors and create clusters according to the number of clusters we assign to it. We used 4 clusters so that it groups this vector codes in four groups segments and we presented the outcome of this in figure 6. SOM clustering is not too far off from our speculation made above in figure 5. It shows that most on the clusters coloured light blue in figure 5 seem to be part of each cluster in figure 6.

We extracted the vector codes from SOM clusters in figure 6 and visualized the clusters in 3d diagram. The clusters by SOM is shown in figure 7. As per this figure, SOM algorithm does not seem to cluster LLM and LLD well because the clusters

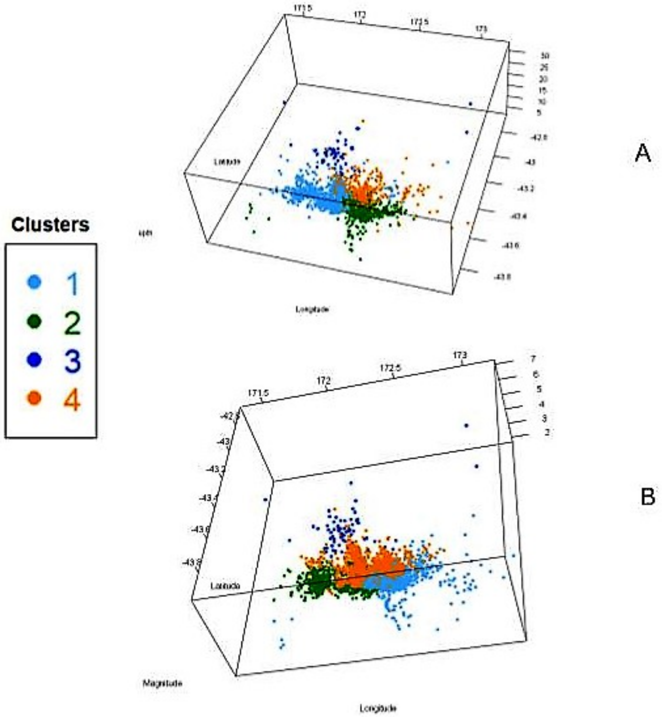

Figure 7: Clusters of LLM and LLD on 3d plot: A is clusters of LLD by SOM and B is clusters of LLM by SOM.

are spread across latitude and longitude. Despite the fact, the correlation between magnitude and space location or depth and space location is worth nothing. However, SOM is clustering these features on different locations in space.

In addition to the 3D plot in figure 7, our results from the silhouette score and SSE also demonstrates that SOM is not producing an optimal segmentation. SOM is producing a silhouette score of -0.04 for LLM and 0.3 for LLD. As it is shown in table 1, SSE of SOM for both LLM and LLD cluster is too large (137.2896 and 3459.3579 respectively), which is consistent with silhouette score.

0.5.2 Agglomerative Hierarchical

We have implemented agglomerative hierarchical clustering algorithm to cluster the underlined features and the result is presented in dendogram shown in figure 9.

This clustering aspects of this algorithm is more apparent in figure 10 and from the diagram, it clear to see that it is performing better than SOM because we can see four well separated segments in figure 10, A and B. It is clustering depth and magnitude and forming different layers of these clusters as we are initially expecting to get because magnitude and depth correlated with their own only.

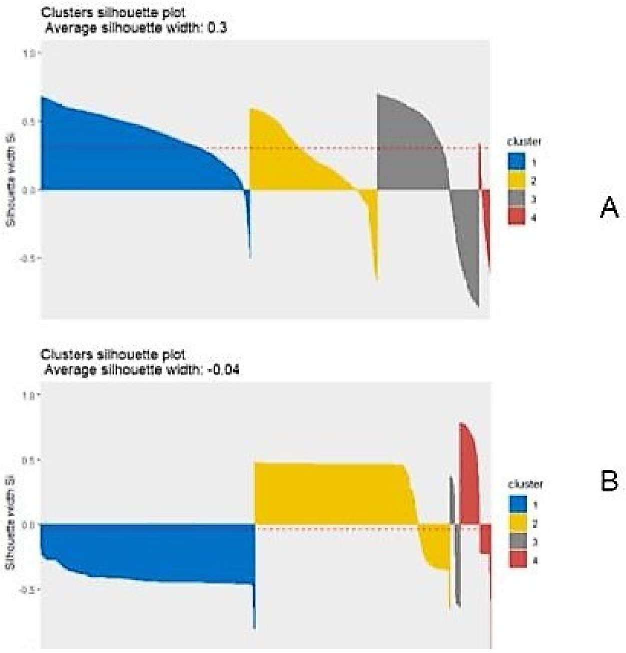

Figure 8: Silhouette score from SOM algorithm: A is silhouette score of LLD cluster by SOM and B is silhouette score of LLM cluster by SOM.

Figure 11 is the silhouette score of agglomerative hierarchical clustering of LLD and LLM. The silhouette score of this algorithm for LLD and LLM is 0.87 and 0.49 respectively, which ascertain our 3D visualization in figure 10. Likewise, from table 1, it can be seen that this algorithm has SSE value of 31.0380 for LLM and 957.2778 for LLD. This gives clearly better results than SOM.

0.5.3 K-means

The result of K-means clustering is shown in figure 13. The clustering result of LLD and LLM is presented in figure 13. It terms of 3D visualization, it is performing good, almost as good as aglomerative hierarchical clustering algorithm.

Table 1: SSE of SOM,K-means and Hierarchical clustering algorithms.

Algorithm	Magnitude	Depth
SOM	137.2896	3459.3579
K-means	27.0305	657.4757
Hierarchical	31.0380	957.2778

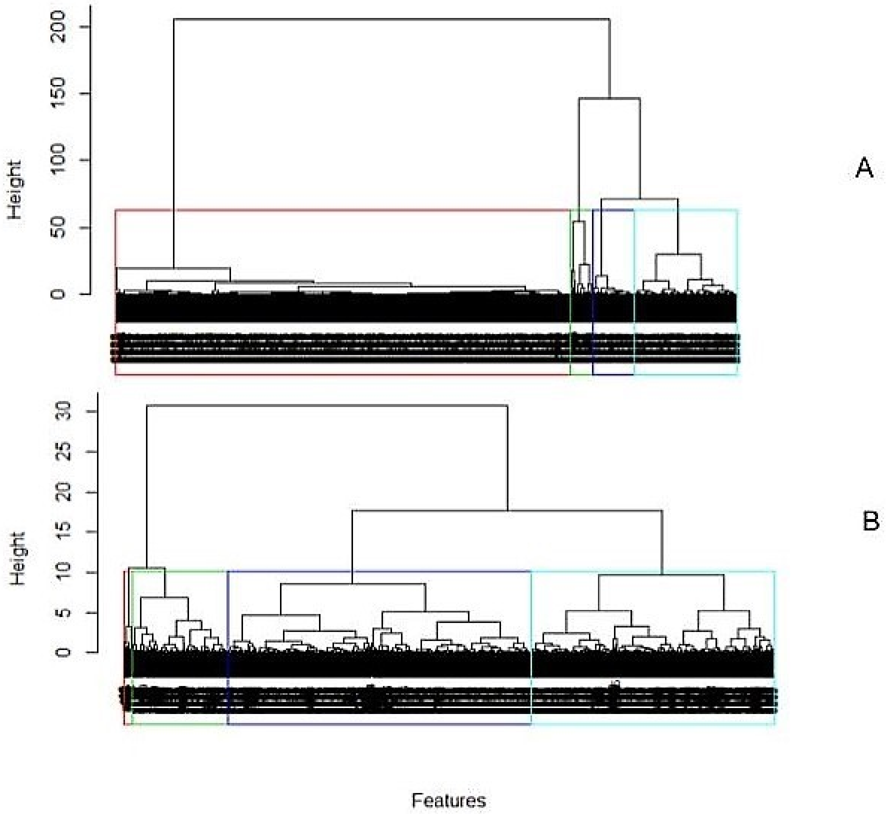

Figure 9: Dendogram of hierarchical clustering of LLD and LLM: A is dendogram of LLD and B is dendogram of LLM.

K-means algorithm is giving a silhouette score of 0.9 for LLD clustering and 0.53 for LLM clustering, giving highest score than the rest of the algorithms. Furthermore, SSE of K-means provide evidence for our silhouette score with a smallest value of 27 for LLM and 657.4757 for LLD (table 1).

Surprisingly, from these results, we can see that the k-means algorithm is outperforming the other two algorithms (SOM and hierarchical clustering). On the other hand, SOM is performing poorly for this particular data set. We assume this is due to the distance in the feature space, and SOM strongly relies on predefined distance in feature space. Hence, the vector codes taken from the two-dimensional space of SOM (figure 6) are maybe happen to be far away in feature space, although they seem close to each other in topology proximity. Besides, SOM is recommended for a large data set with a strong correlation among the features, which is not the case here. In spite of that there is insignificant correlation among the underlined features; consequently, it is forcing the algorithm to produce poor segmentation. Conversely, K-means finds the centroids of the clusters interactively and less likely to rely on predefined distance

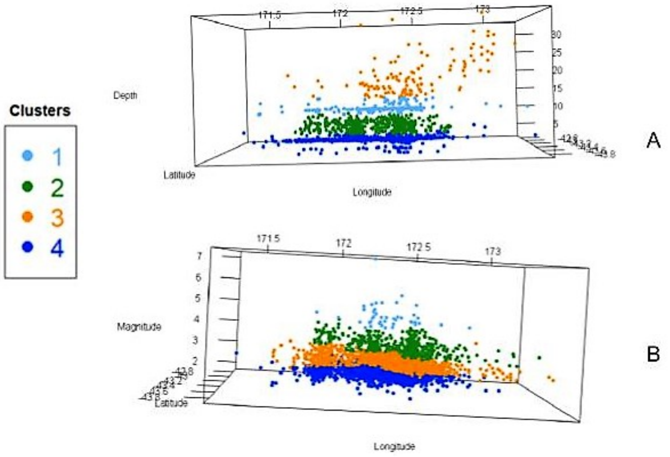

Figure 10: Clustering by aglomerative hierarchical clustering: A is clusters of LLD by hierarchical clustering and B is clusters of LLM by hierarchical clustering

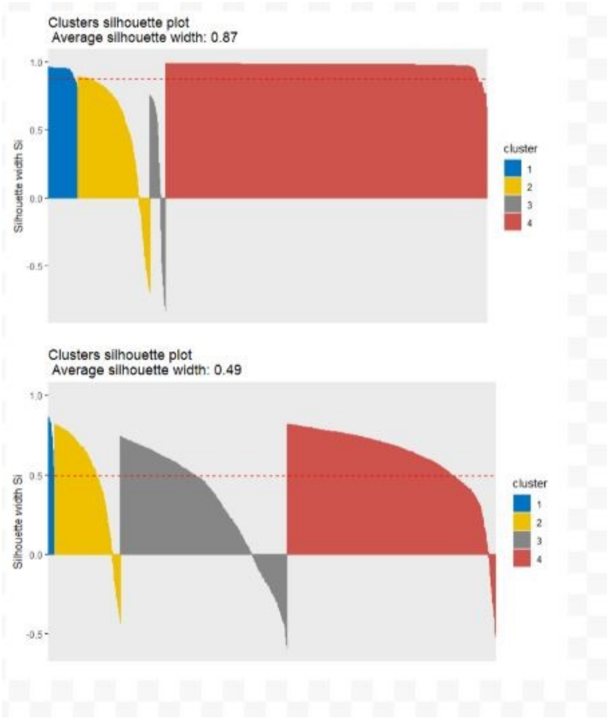

Figure 11: Silhouette score from hierarchical algorithm: A is silhouette score of LLD and B is the silhouette score of LLM.

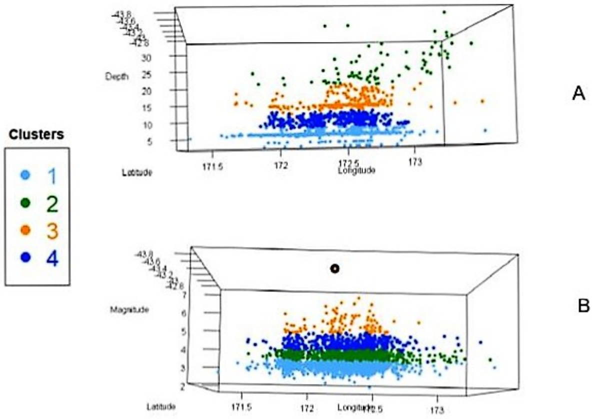

Figure 12: Clusters of LLM and LLD on 3d plot:A is clusters for LLD by K-means and B is clusters of LLM by K-means.

in feature space; hence it finds the better centroid point in feature space that minimizes the sum of square error. As a result, even though, there is week correlation among the LLM and LLD, it is able to find the correlation of magnitude and depth with their own and cluster these features individually, without counting on latitude and longitude (figure 12).

Then, utilizing the clusters obtained from this rigorous algorithm (K-means), we performed classification by converting the clusters into labels. Then, we applied Naive Bayes classification algorithm to find the epicenter of the magnitude.

Figure 14 reports the frequency of magnitude in each class, obtained by Naive Bayes classifier. As the figure portrays, the magnitude is approximately uniformly distributed in class 1 (the seismic activity is almost similar), while it is skewed to the left in class 2 (strong exponentially increasing seismic activity). The magnitude appear to exponentially decrease in class 3 and 4 due to the epicenter. Class 3 and 4 contain the strongest seismic activities. Therefore, from this figure, we can see that the **epicenter of the DE is located in cluster 4**. This is also illustrated in figure 16. It created layer upon layer of clusters of magnitude. This result substantiates our assumption of insignificance of correlation between magnitude, latitude and longitude.

Table 2: Prediction of Naive Bayes algorithm.

Actual			
1	2	3	4
1548	73	0	0
0	1014	0	10
0	0	105	1
0	9	1	529

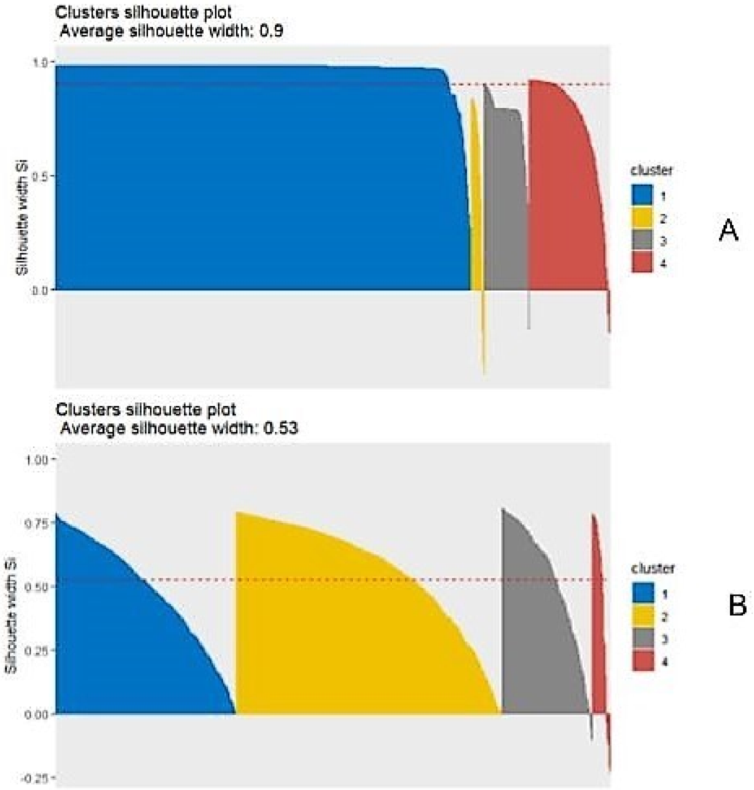

Figure 13: Silhoutte score of K-means clustering.

Table 2 comprises the predicted classes by Naive Bayes algorithm. The prediction accuracy that is acquired from Naive Bayes was 0.9714286, which is remarkable enough to make inference about the classifications of magnitude.

We have finally visualized cluster 4 that encompasses the epicenter (main shock) (figure 16). The point where the arrow indicating is where the epicenter of DE is located. It lies on **172.1679 longitude and -43.52731 on latitude**. As detailed in the figure, the size of the circles signify the depth of the respective magnitude, while the color represents the magnitude or intensity of the shock. The shock tends to cause major destruction to the buildings around this place (magnitude of ≥ 5).

0.6 Conclusion

In conclusion, this paper investigates a data set of Darfield earthquake, that occurred in Christchurch, South Island's east coast of New Zealand. We performed comparison clustering analysis among three different algorithms (SOM, K-means and agglomerative hierarchical), coupled with one classification algorithm (Naive Bayes). The aim

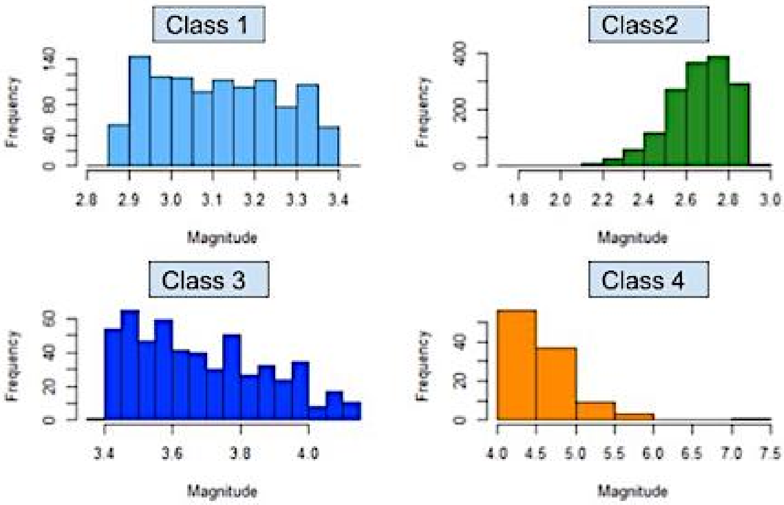

Figure 14: Frequency of magnitude in each class, classified by Naive Bayes classifier.

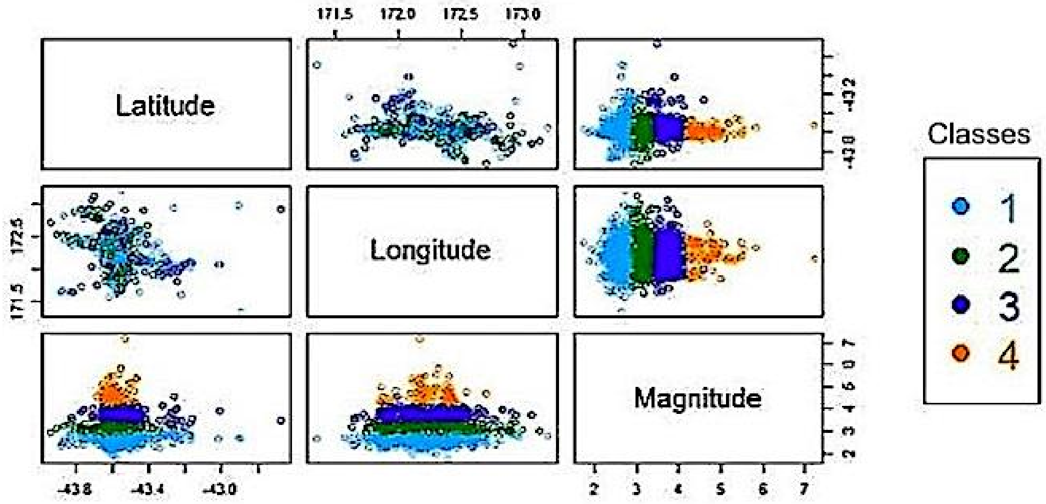

Figure 15: Classifications of LLM by Naive Bayes algorithm.

here was to find the cluster the contains the epicentre then specify where in space (latitude and longitude) this epicentre lies. Our result has led us to conclude that K-means is the winner algorithm in terms of clustering both magnitudes as well as depth with smallest SSE values of 27.0305 and silhouette score of 0.53 for LLM and smallest SSE of 657.4757 and highest silhouette score of 0.9 for LLD. On this basis, we conclude that having implemented multiple complicated clustering algorithms and disregarding

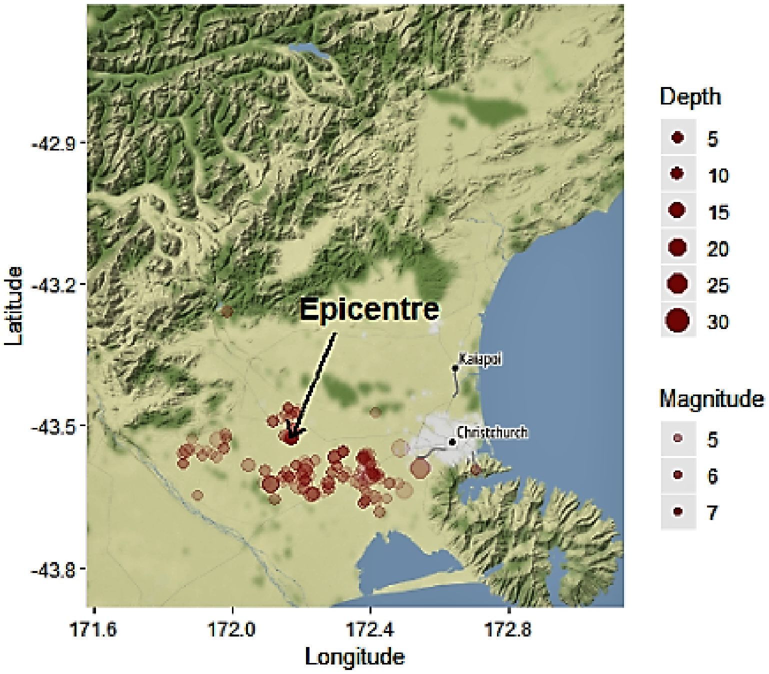

Figure 16: Epicenter and related seismic activities in class 4.

K-means algorithm might lead to miss interpretation of the results as K-means might still provide a better clustering accuracy than others.

According to our results, class four is the segment that contains the epicentre (main shock) of DE. That is, the epicentre falls on 172.1679 longitudes and -43.52731 latitudes, within the range of latitude we initially expected. The magnitude in this cluster tends to cause several damages to the city, while the magnitudes in other cluster have relatively smaller intensity (≤ 5 and are less likely to cause severe damage to the subsidiaries of the city. This area is also likely to be prone to another shock because these shocks are may be lying on fault. And, when earthquakes occur on fault, they tend to cause another earthquake sooner or later. Hence, the authors recommend it is worth staying alerted while living in this particular areas. One thing not to forget to mention, however, is that the shock and magnitude of DE are not randomly distributed across latitude and longitude. Instead, the level of intensity varied within the certain ranges of latitude and longitude considered in this paper, forming different layers of intensity, thus made it difficult to view the clusters from the top (on 2-dimensional plot). This aspect of the cluster is only visible in 3-dimensional plots.

One limitation of this paper is that we performed a mere spatial analysis for this particular data set, instead of spatiotemporal analysis (did not take into account time). Another limitation of this work is that we are making inferences relying on results obtained from only three clustering algorithms. The best algorithm appears to be a K-means clustering algorithm, even though it has a silhouette score of 0.53, which is not strong enough. We suppose other types of unsupervised algorithm might improve this accuracy.

Regardless, future research could continue to explore a spatiotemporal analysis of this particular data set and track depth and magnitude along with time. Additionally, future research should certainly further test whether there are potential unsupervised algorithms that can be employed in this particular data set and produce a better clustering result than K-means clustering.

Bibliography

[1] Kevin Pang. *Self-organizing Maps*. Collage. 2003. URL: https://www.cs.hmc.edu/~kpang/nn/som.html.

[2] Georg Palzlbauer. "Survey and comparison of quality measures for self-organizing maps". In: *Proceedings of the Fifth Workshop on Data Analysis (WDA04)* (Jan. 2004).

[3] David A Yuen et al. "Visualization of Earthquake Clusters over Multi-dimensional Space." In: Jan. 2009, pp. 2347–71.

[4] GNS. "Darfield Earthquake / Canterbury quake / Recent Events / Natural Hazards and Risks / Our Science / Home - GNS Science". In: (2010). URL: https://www.gns.cri.nz/Home/Our-Science/Natural-Hazards-and-Risks/Recent-Events/Canterbury-quake/Darfield-Earthquake.

[5] 2) Sarlin P. (1 and T. (1) Eklund. *Fuzzy clustering of the self-organizing map: Some applications on financial time series.* Vol. 6731 LNCS. Lecture Notes in Computer Science. 2011. ISBN: 9783642215650. URL: http://ezproxy.aut.ac.nz/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edselc&AN=edselc.2-52.0-79959298608&site=eds-live.

[6] Anne Rooney. *Earthquake!*. Nature's fury. Encyclopedia Britannica, 2012. ISBN: 1615356487. URL: http://ezproxy.aut.ac.nz/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=cat05020a&AN=aut.b2688088x&site=eds-live.

[7] "Unsupervised Clustering Method for the Capacited Vehicle Routing Problem." In: *2012 IEEE Ninth Electronics, Robotics and Automotive Mechanics Conference, Electronics, Robotics and Automotive Mechanics Conference (CERMA), 2012 IEEE Ninth, Electronics, Robotics and Automotive Mechanics Conference* (2012), p. 211. ISSN: 978-1-4673-5096-9. URL: http://ezproxy.aut.ac.nz/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edseee&AN=edseee.6524580&site=eds-live.

[8] Karim Mahmoud, Nagia Ghanem, and Mohamed Ismail. “Unsupervised Video Summarization via Dynamic Modeling-Based Hierarchical Clustering.” In: *2013 12th International Conference on Machine Learning and Applications, Machine Learning and Applications (ICMLA), 2013 12th International Conference on, Machine Learning and Applications, Fourth International Conference on* (2013), p. 303. ISSN: 978-0-7695-5144-9. URL: `http://ezproxy.aut.ac.nz/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edseee&AN=edseee.6786125&site=eds-live`.

[9] Matthew Kyan. *Unsupervised learning : a dynamic approach.* IEEE Press Series on Computational Intelligence. John Wiley Sons, 2014. ISBN: 1118875230. URL: `http://ezproxy.aut.ac.nz/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=cat05020a&AN=aut.b13571461&site=eds-live`.

[10] Matthew Kyan. *Unsupervised learning : a dynamic approach.* IEEE Press Series on Computational Intelligence. John Wiley Sons, 2014. ISBN: 1118875230. URL: `http://ezproxy.aut.ac.nz/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=cat05020a&AN=aut.b13571461&site=eds-live`.

[11] Jing Tian, Michael Azarian, and Michael Pecht. “Using self-organizing maps for anomaly detection in hyperspectral imagery”. In: *Proceedings, IEEE Aerospace Conference* (2014). DOI: `10.1109/aero.2002.1035291`. URL: `file:///C:/Users/tjm8021/Downloads/AnomalyDetectionUsingSelf-OrganizingMaps-BasedK-NearestNeig....pdf`.

[12] Wang, Li. “Feature weighted naive Bayes algorithm for information retrieval of enterprise systems.” In: *Enterprise Information Systems* 8.1 (2014), pp. 107–120. ISSN: 17517575. URL: `http://ezproxy.aut.ac.nz/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=93009191&site=eds-live`.

[13] Tippaya Thinsungnoena et al. “The Clustering Validity with Silhouette and Sum of Squared Errors”. In: *Proceedings of the 3rd International Conference on Industrial Application Engineering* (2015). URL: `https://pdfs.semanticscholar.org/8785/b45c92622ebbbffee055aec198190c621b00.pdf`.

[14] Rajanish Kamat and Rajani Kamath. “Earthquake Cluster Analysis: K-Means Approach”. In: *Journal of Chemical and Pharmaceutical Sciences* 10 (Jan. 2017).

[15] Pepi Novinanti, Dyah Setyorini, and Ulfasari Rafflesia. “K-Means cluster analysis in earthquake epicenter clustering”. In: *International Journal of Advances in Intelligent Informatics* 3 (July 2017), p. 81. DOI: `10.26555/ijain.v3i2.100`.

[16] Ritesh. *Naive Bayes Classifier - Fun and Easy Machine Learning*. Youtube. 2017. URL: `https://www.youtube.com/watch?v=CPqOCI0ahss`.

[17] Mofizur Rahman, Chowdhury Afroze, Lameya Refath, Naznin Sultana Shawon, Nafin. "Iterative Feature Selection Using Information Gain and Naive Bayes for Document Classification." In: *2018 21st International Conference of Computer and Information Technology (ICCIT), Computer and Information Technology (ICCIT), 2018 21st International Conference of* (2018), p. 1. ISSN: 978-1-5386-9242-4. URL: http://ezproxy.aut.ac.nz/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edseee&AN=edseee.8631971&site=eds-live.

[18] Dilli Ram Thapa et al. "Aftershock analysis of the 2015 Gorkha-Dolakha (Central Nepal) earthquake doublet". In: *Heliyon* 4.7 (2018), e00678. ISSN: 2405-8440. DOI: https://doi.org/10.1016/j.heliyon.2018.e00678. URL: http://www.sciencedirect.com/science/article/pii/S240584401732772X.

[19] Dilli Ram Thapa et al. "Aftershock analysis of the 2015 Gorkha-Dolakha (Central Nepal) earthquake doublet". In: *Heliyon* 4.7 (2018), e00678. ISSN: 2405-8440. DOI: https://doi.org/10.1016/j.heliyon.2018.e00678. URL: http://www.sciencedirect.com/science/article/pii/S240584401732772X.

[20] Abdullateef Balogun et al. "Performance Analysis of Selected Clustering Techniques for Software Defects Prediction". In: (June 2019), pp. 30–42.

[21] Tim Bock. *What are the Strengths and Weaknesses of Hierarchical Clustering?* 2019. URL: https://www.displayr.com/strengths-weaknesses-hierarchical-clustering/.

[22] Bala Deshpande. *3 challenges with Naive Bayes classifiers and how to overcome*. 2019. URL: http://www.simafore.com/blog/3-challenges-with-naive-bayes-classifiers-and-how-to-overcome.

[23] "Naive Bayes Classifier-Assisted Least Loaded Routing for Circuit-Switched Networks." In: *IEEE Access, Access, IEEE* (2019), p. 11854. ISSN: 2169-3536. URL: http://ezproxy.aut.ac.nz/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edseee&AN=edseee.8610171&site=eds-live.

[24] GNS Science. *GNS Science*. New Zealand's leading provider of Earth, geoscience, isotope research, and consultancy services. URL: https://www.gns.cri.nz.

[25] L Lovmar et al. "Silhouette scores for assessment of SNP genotype clusters." In: *BMC GENOMICS* 6 (n.d.). ISSN: 14712164. URL: http://ezproxy.aut.ac.nz/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edswsc&AN=000228002500001&site=eds-live.

[26] Khaista Rehman, Paul W. Burton, and Graeme A. Weatherill. "K-means cluster analysis and seismicity partitioning for Pakistan." In: *JOURNAL OF SEISMOLOGY* 18.3 (n.d.), pp. 401–419. ISSN: 13834649. URL: `http://ezproxy.aut.ac.nz/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edswsc&AN=000337087600005&site=eds-live`.

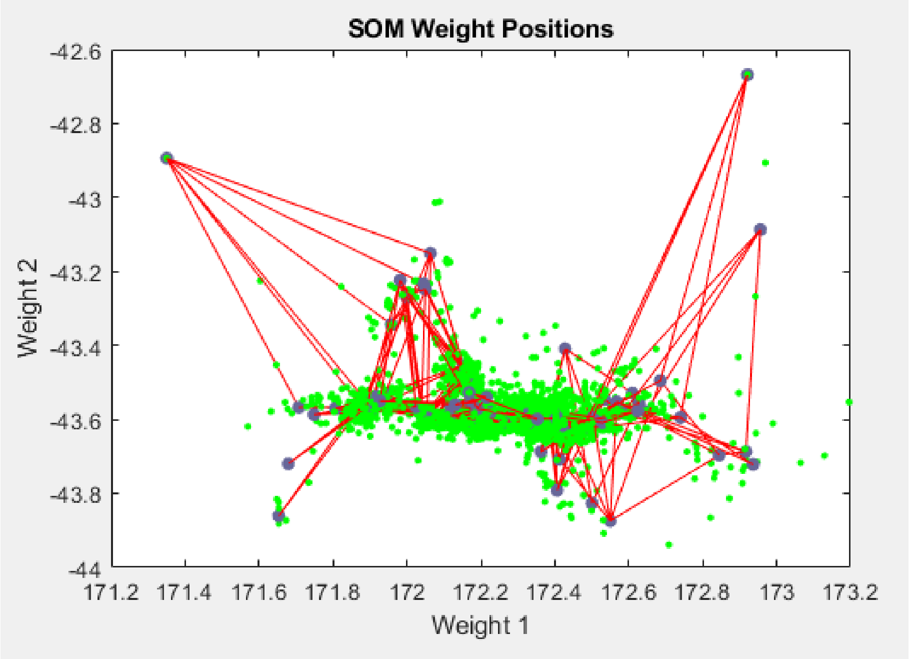

Figure 17: SOM weight positions of LLM.