Datamining and Machine Learning Assignment 2

# Introduction

The objective of the data mining assignment is to conduct knowledge discovery on earthquake aftershock activity from data acquired following the 2010 Darfield Earthquake in New Zealand South Island to establish patterns of magnitude and depth in relation to geographical location.

The task involves unsupervised machine learning and descriptive mining to establish the paters patterns in the form of meaningful clusters within multidimensional feature space i.e. longitude, latitude + magnitude or longitude, latitude + depth. The idea is to establish association rules between location of highest magnitude and their distributions in the surrounding area. Temporal monitoring although vitally important in tracking the progress and evolution aftershocks is not taken into consideration for this assignment.

3 algorithms are considered for use to achieve the knowledge discovery:

* Kohonen Self-Organising Map (SOM) / Self-Organising Feature Map (SOFM)
* Density-based spatial clustering of applications with noise
* K-means clustering

All 3 Algorithms are then evaluated by observing the metrics:

* Sum of Squares Error (SSE)
* Cluster Silhouette
* Computation Time

A winner algorithm is chosen among the 3. Afterwards a follow up predictive classification analysis is done on the resulting clusters to identify the locations of the largest shocks. Achieving this task involves a mix of both visual comparisons and quantitative analysis to support the results produced by the data mining.

# Literature Review

Earthquakes are natural phenomena where the surface of the Earth experiences violent shaking following the collision of continental plates that make up parts of the Earths lithosphere. Depending on the magnitude/ seismicity of these events their impact can range from hardly detectable to capable of toppling buildings and causing severe damage to cities and although they are a cause of much human suffering and loss, they are a natural part of living on planet Earth.

The primary source of the seismic activity of an earthquake is the release of strain energy from the focus also known as the hypocentre. As two continental plates meet each other faults which can be thought of as cracks begin to build up in the crust up to a point where it is eventually released. This release of energy produces propagating waves that travel to the surface causing vibrations and thus shaking effects. The surface point directly above this hypocentre focus is called the epicentre and is the point of highest magnitude of the Earthquake because it is the place where these waves first reach the surface in the shortest distance (straight line). [earthquake paper]

The Epicentre an example of which seen in Figure 1 is given a significant focus in this data mining operation as we wish to predict the location distributions of point of highest magnitude aftershocks that arise from major seismic events.

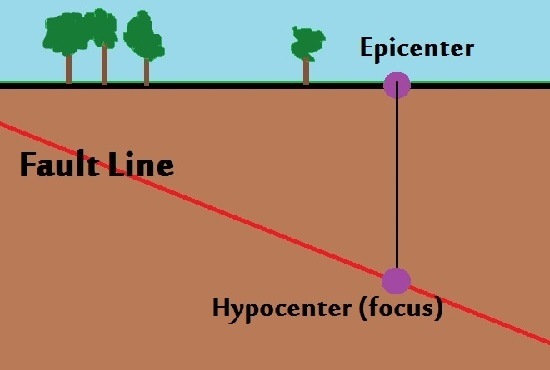


Figure 1: Epicenter and Hypocenter of Earthquake

Because of their origin from colliding tectonic plates, earthquakes happen more frequently in selected parts of the world. By observing a simple mapping of Tectonic Plates in Figure 2 it can be observed that New Zealand is located at the meeting points of Pacific and Australian plates thus making it a region prone to earthquakes and aftershocks that follow the main event. [earthquake paper]

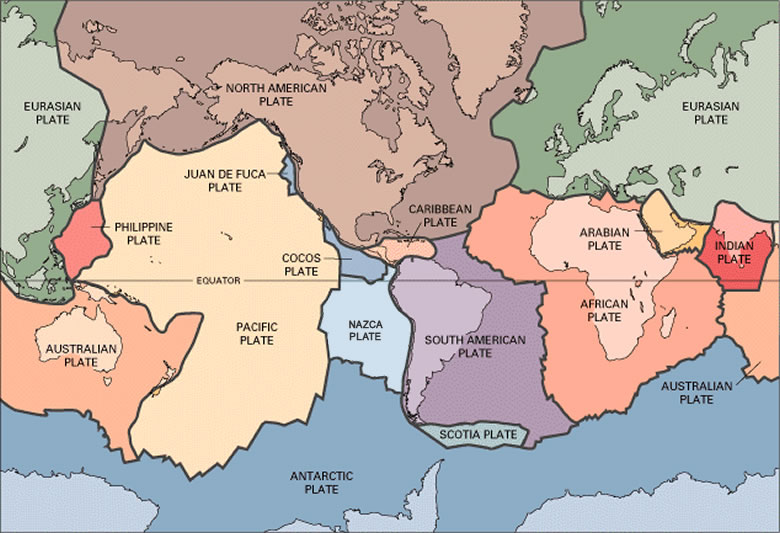


Figure 2:

## Data Mining Earthquakes with Clustering

Clustering analysis is the task of creating groups of data related to an area of research by recognizing patterns of based on similar traits. Good clustering means the data instances within a cluster are of a maximum similarity to each other compared to data in other clusters. Different approaches and rules can be applied to cluster data however, they tend to vary based on the goal/ domain of knowledge of the data being clustered and the objective trying to be achieved.

Clustering is a form of unsupervised machine learning because the aim is to find a structure within the data instead of predicting a certain class or value. The structure of the clusters establishes an underlying pattern of which insight can be gained. The discovered structures can then be targeted to solve a problem. In the case of this task it is the minimization of harm and the reduction of property damage from natural disasters.

Clustering approach has previously been done on Earthquake aftershock data from ref [2] [3] [4] where the data analysed is often multi-dimensional and requires decomposition and reduction to effectively cluster. The approaches taken often combine several clustering algorithms to find relationships in the multiple spatial and non-spatial attributes and to track the progress of points across multiple clusters to establish and understanding of areas of greatest intensity and prone to danger.

Since clustering can be done with many techniques (and the usefulness of the clusters themselves are in the eye of the beholder) they are categorized based their approach/ model. Some popular approaches include:

* Centroid models based on central mean vector
* Density models based on regions of data concentration of instances in feature space
* Hierarchical models based on connectivity

There is no one model fits all for clustering and results always vary based on the dataset. Therefore, it is better to try multiple approaches to find the ideal solution. This assignment focuses on applying a preliminary neural modal to provide the vital parameter (k number of clusters) for follow up clustering with Centroid and Density based modes.

# Data

The data mined for this assignment was recorded for 1 month following the 2010 Darfield Earthquake from the 3rd of September 2010 to the 30th of September 2010. The data is multi-dimensional because involving spatial coordinates of longitude and latitude also involved values of magnitude and depth. Additionally, the first instance of the data is the epicentre of the seismic event (point of highest magnitude). The location of the epicentre near Darfield can be scouted from Google maps.

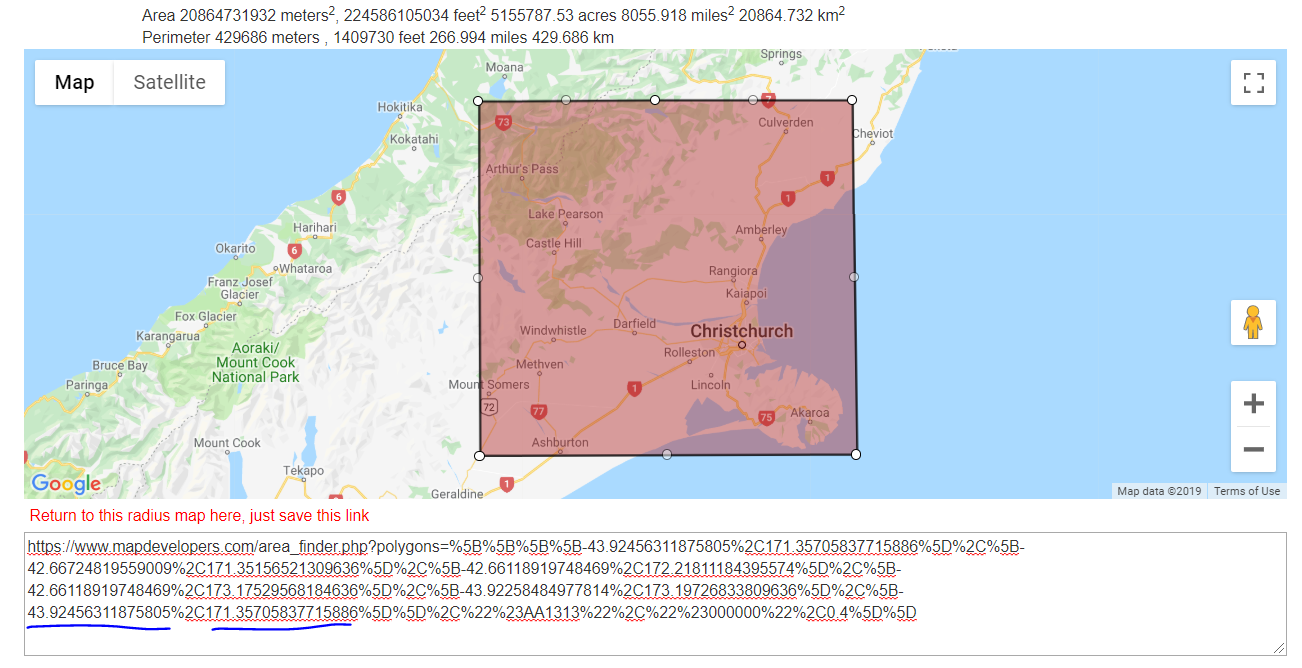
## 

Figure 3: Location of Epicentre in Geospatial Feature Space

It is worth noting that the data comprises of 3290 instances and although is a high dimensional dataset, it is not in any way considered “large” or “Big” it is actually typical of these types of studies [4] therefore most algorithms that are applied should have no problems in computation time.

## Size of Area of Study

To get a better understanding, <https://www.mapdevelopers.com/area_finder.php> is a useful tool that can be used to find the feature space boundaries. This involves finding the location of the maximum and minimum longitude and latitude {code snippet 1} and using the app to define the area subtended by those points subtended. It is found that the area of the feature space encompasses an area of 20864.732 km2



longitude

latitude

Figure 4: Area of Study

# Pre-processing

## Trimming of Data Table

The original dataset provided contained 22 attributes. Many of the values such as used phase and station count, azimuthal gap and origin error are geoscience domain knowledge attributes that may or may not play a role in influencing clustering results once dimensionality reduction has been carried out. However, for this study the attributes are parsed into the only 4 that gives us a significant spatial understanding of aftershocks: longitude, latitude, magnitude, depth (LLMD). All 3 algorithms are applied on 2 subsets of even these 4 parameters i.e. Longitude Latitude Magnitude (LLM) or Longitude Latitude Depth (LLD). These parameters are also taken in [3]

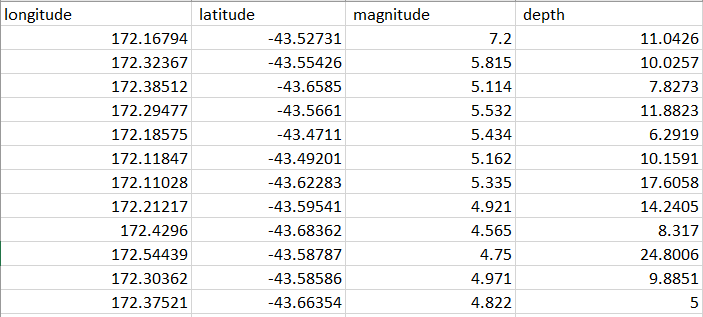


Figure 5: Sub-table of Longitude Latitude and Magnitude and Depth

The data is then loaded into python and converted into Nd array. Nd array data type is chosen because it is faster, consumes less memory and most of our models interact best with this data structure.

# Pre-visualizing

Before any clustering algorithm is applied, visualization of geospatial feature space is done to get a better understanding of any obvious patterns in spatial distribution the dat. The plot is seen in Figure

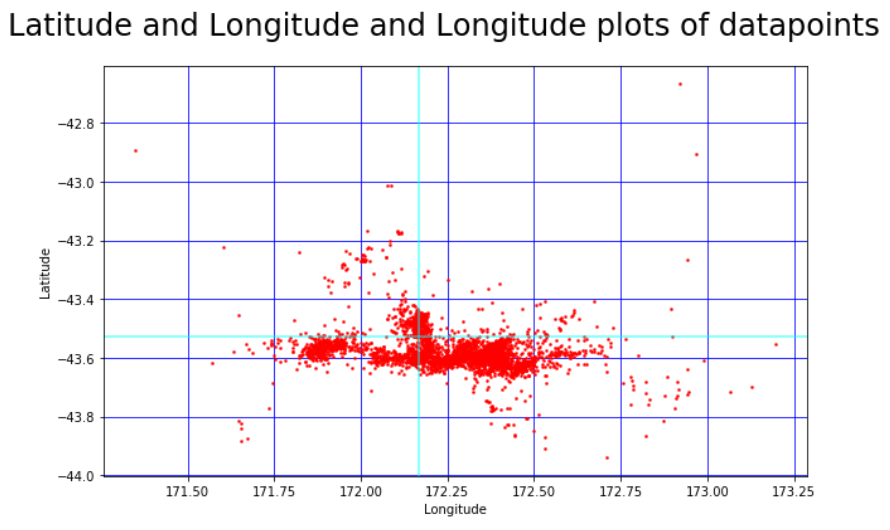


Figure 6: Spatial Locations and Distribution of Earthquake aftershocks

From this plot we can start to see some interesting details we see the presence of outliers (green circles) that may play havoc with certain clustering algorithms and should be removed, and we also notice the central grouping (Blue Oval) and some structure within it (yellow ovals) relative to the location of the epicentre (cyan lines). We observe some scattered secondary outliers as well (grey circles).

Hypothesis: From this plot we can expect our most significant clusters to emerge from the yellow oval shapes within the blue oval while perhaps encompassing some scatterings of grey circle instances.

This is the first step towards any clustering activity; simply eyeballing simple plots of the dataset and applying the neural networks available to all humans, however since the data is multi-dimensional involving magnitude and depth more advanced techniques are required to gain further insight

# Algorithm and Mining

## Self-Organizing Feature Map

Kohonen Self-Organizing Feature Map or SOM is a type of feed forward Artificial Neural Network that maps n-higher dimensional input vectors to a 2-D grid of output neurons [5]. The two main differences between SOM and ANNs are; SOM Applies Competitive Learning Approach over error correction methods such as back propagation and SOM has no hidden layers. It is single grid of output neurons that compete for representation that best fits the input data.

SOM can be used in exploratory data mining phase on unlabelled data to find the ideal number of clusters via visual interpretation of the Unified Distance Matrix and in the case of labelled data can be used for dimensionality reduction of multi variable datasets.

### How it works

1. Size of the Map is defined with each node initialized at a random weight
2. A sample random vector xi is chosen from the data where xi = [Xi, Yi, Di, Mi]
3. Every node weight is compared with the input vector to find which weight is most similar by calculating Euclidean distance. Smallest distance wins and is assigned as Best Matching Unit (BMU). This is known as the competitive phase
4. Neighbourhood function is called to update the nodes surrounding the BMU making them closer to the input vector. The closer nodes to the BMU the more the weights are altered to become the like sample vector. The farther away nodes are less altered. This is called the Cooperative phase
5. Step 2 is repeated for N number of iterations until the feature map stops changing or until iteration limit is reached.

### Advantages vs Disadvantages

|  |  |
| --- | --- |
| Advantages | Disadvantages |
| * Simplifies higher dimensional Spaces. Preserves topological relationships in multidimensional space by mapping them to a group of neurons. * Visual interpretation for ease of interpretation * Simple training with minimal parameter tuning (map size). | * Difficult to be scaled for evolving datasets. * Does not provide a generative model. Cannot provide insight on how to produce the sample data. * Not as effective when data is mixed. (Numeric + Categorical + Nominal) |

SOM has many applications such as image processing

The SOM implementation used in this mining assignment is SOMPY for python 2.7x by Vahid Moosavi aka sevamoo [6]. The only parameter that requires tuning is the map. There is no perfect number of clusters especially for real world datasets. So, for this experiment we will iterate from small map size, to ideal, and a large size. However, based on [5] a general rule of thumb for a SOM of nodes mxn the total area ≈ 5 \* sqrt(n\_instances) in order to begin to get get any significant information from the visualization.

### Visualization and Interpretation

Once the SOM has been trained many visualizations that can be analysed such as shown in Figure : these range from 2D mapview of each Feature, hitmaps and the Umatrix. The visualization that is most of interest is the U-matrix (Unified Distance Matrix).

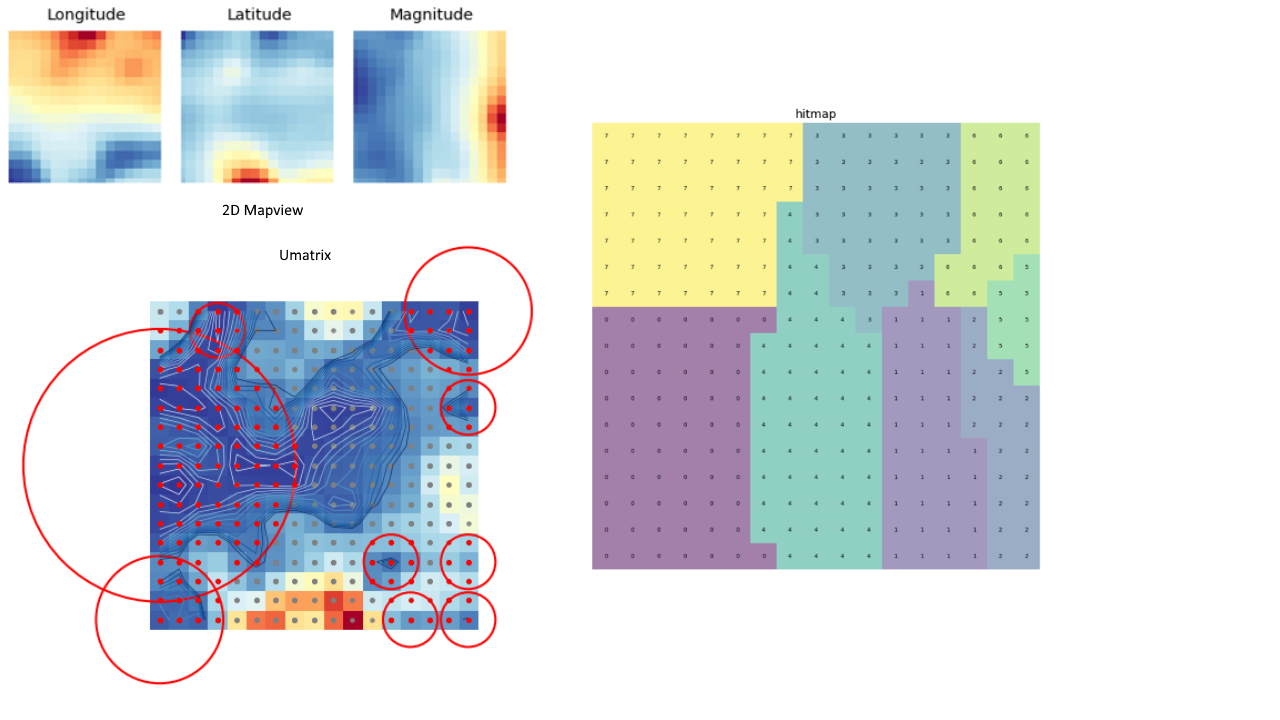


Figure 7: SOM visualizations 2D Mapview, Hit Map, Coloursed U-Matrix

The U-matrix depicts the Euclidean distance between codebook vectors of neighbouring neurons where lighter colours depict closely spaced (clustered) nodes and dark colours depict wider separation between groups of nodes and boundaries between clusters. A colour representation is available but moving forward grayscale is chosen for easier distinction. Colour representation may be hard to distinguish densely packed nodes since compared tor grayscale.

### Longitude Latitude Magnitude

To establish the relationship between Location (Longitude/ Latitude) and magnitude of quake SOM of multiple size area are compared.

|  |  |  |  |
| --- | --- | --- | --- |
| Map size | 6x6 | 17x17 | 50x50 |
| U-matrix |  |  |  |
| N clusters discovered | 1 ~ inconclusive | 4 | 5 |
| Training Time (s) | 0.253000 | 0.801000 | 17.24500 |
| Quantization Error | 0.746467 | 0.320556 | 0.138652 |

Based on the U-matrix 4 clusters from 17 x 17 is enough. SOMPY allows us to assign cluster\_labels to the datapoints. Based on the cluster labels we can visualize the 4 different clusters. The distribution with respect to epicentre is noted and the cluster that contains the epicentre is also noted:

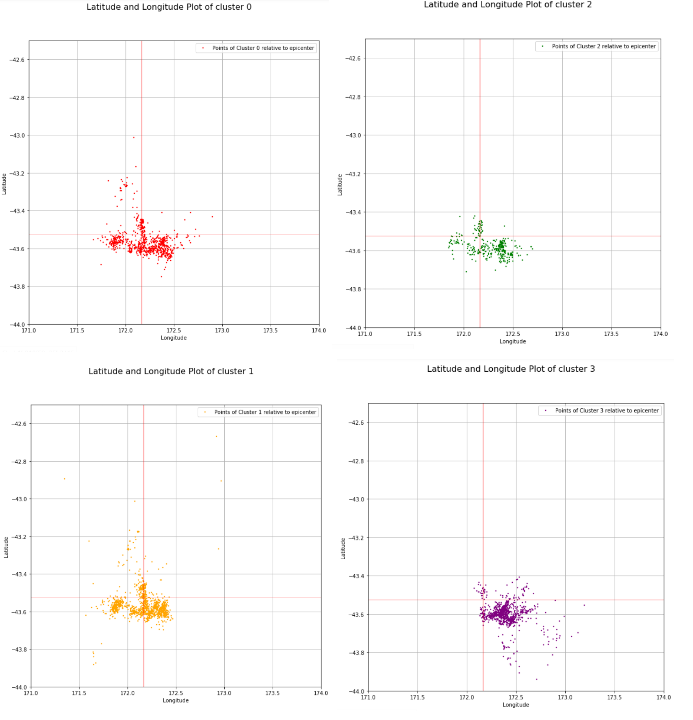


Figure 8: Distribution of SOM clustered data. Epicentre belongs to cluster 2 Green Plot

### Longitude Latitude Depth

Clustering is repeated with Longitude Latitude and Depth (LLD) parameters

|  |  |  |  |
| --- | --- | --- | --- |
| Map size | 6x6 | 17x17 | 50x50 |
| U-matrix |  |  |  |
| N clusters discovered | 2? Also inconclusive | 4 | 4 |
| Training Time (s) | 0.272000 | 0.835000 | 17.192000 |
| Quantization Error | 0.561950 | 0.217925 | 0.080829 |

Again, 4 clusters are deemed appropriate the cluster labels are derived from the 17x17 SOM. The distribution with respect to epicentre is noted and the cluster that contains the epicentre is also noted:

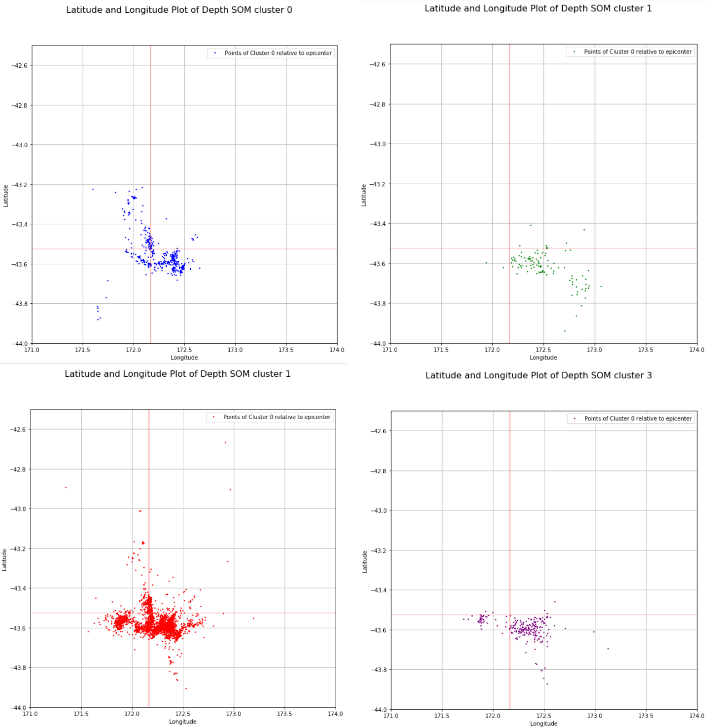


Figure 9 Distribution of SOM clustered Longitude Latitude Depth data. Epicentre belongs to cluster 0

### Interpretation

Based on the SOM U-matrix we select 4 clusters as the number of clusters. As the exploratory phase this factor will play a role in our later cluster algorithms. From the SOM visualization in magnitude we can see that cluster 0 and cluster 1 follow a similar shape with our hypothesis shape while cluster 2 which contains the epicentre almost follows the same outline however with lesser density. The mean magnitude the epicentre cluster is higher than the mean magnitude of the other clusters. It is noted that Cluster 3 has a drift towards the south east despite having less instances.

The same can be said for depth but with some differences. Cluster 0 and cluster 1 (epicentre cluster 0 less so) approximate the hypothesis shape. Cluster 1 and cluster 3 drift towards the south east.

## Density Based Spatial Clustering of Applications with Noise (DBSCAN)

Density Based Spatial Clustering of Applications with Noise (DBSCAN) is an unsupervised learning method that functions on the basis that regions of densely datapoints belong to a cluster and sparser populated regions are noise. It has previously been used for clustering earthquake data as seen in [2] and [4] and is increasing in popularity.

A DBSCAN implementation is considered for this dataset because it may be able to reveal interesting non-convex cluster shapes in the data that more accurately model certain aftershock distributions.

### How it works

1. Identify core points, border points and noise points
   1. Core point: has more than minimum points to make a cluster (minPts) within radius (eps)
   2. Border point: does not have more than minPts within eps but is within eps of a core point
   3. Noise point: any point that does not have minpts within eps and is not within eps of a core point. None of the above
2. All the Noise points are removed
3. For core points not assigned to a cluster:
   1. Create new cluster and add all points that are density connected to it. Density connected points are points with paths of edges available to it.
4. Assign border points to the nearest core point cluster

### Advantages and Disadvantages

|  |  |
| --- | --- |
| Advantage | Disadvantages |
| * Able to discover clusters with varying interesting shapes (non-convex clusters) * Great handling of outliers * Not reliant on index (K) to define number of clusters. | * Not good for datasets with varying density. * High computational complexity and long implementation * Requires fine tuning of hyperparameters minimum points to make a cluster (minPts) and minimum radius (eps) for maximum effectiveness |

### Balancing Epsilon and minimum Points

Epsilon (eps) is radius of circle for a cluster point. The way of determining eps is via KNN distance graph and the appropriate elbow point is approximated. A KNN distance graph is plotted in Figure . The distance of the first sharp change of the ideal epsilon. The hyperparameters are balanced for 4 clusters with no outliers. Minimum Points is set as 1.

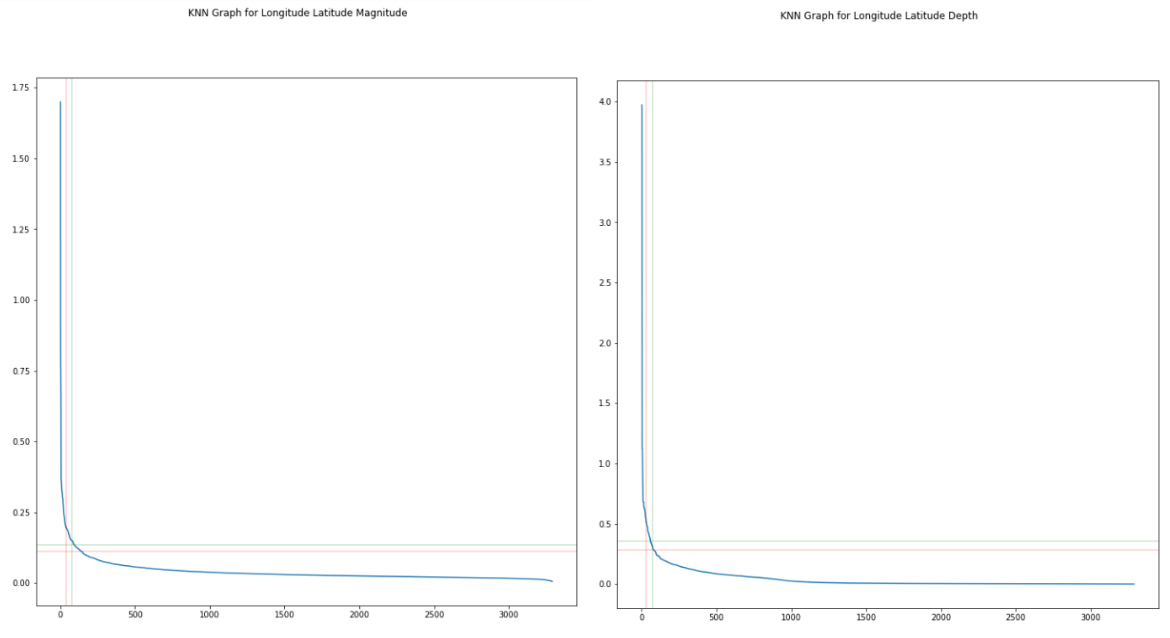


Figure 10 KNN Graph used to approximate for epsilon

|  |  |  |  |
| --- | --- | --- | --- |
| Plot | Epsilon | minPts | Clustering Time (s) |
| Longitude Latitude Magnitude | 0.4 | 1 | 0.307484 |
| Longitude Latitude Depth | 1.1 | 1 | 0.2648351 |

### Results and Implementation

Unfortunately, a properly tuned DBSCAN was unable to be acquired as through most attempts the epicentre ended up belonging to noise. With the listed parameters an asymmetric clustering was acquired. See Evaluation

## K Means Clustering

K means is a centroid based clustering model the uses a mean vector of parameters approach to clustering. It is well known for its many advantages. However, the main disadvantage of the K Means clustering is that it must be supplied with the k number of clusters to group the data into which is arbitrary. This is a challenge [7] that hinders knowledge discovery of finding the ideal number of clusters in feature space.

### How it works

1. Initialize k random centroids within the feature space. Every instance is assigned to the nearest based on squared Euclidean distance.
2. The positions of each centroid are recomputed using the mean of all instances associated with it.
3. Iterations of recalculations are continued until there is very little change in their values (centroids have stabilised)

### Advantages and Disadvantages

|  |  |
| --- | --- |
| Advantage | Disadvantages |
| * Easy to implement * Reduced Time complexity and scales well to larger data * Can be used for multi-dimensional data | * Must be supplied the value of k. Lacks insight/ Knowledge discovery. * Ideal value of k is difficult to predict and requires many tests (KNN distance, SOM, cluster silhouette) [7] * Changing results on different runs of algorithm |

### Longitude Latitude Magnitude

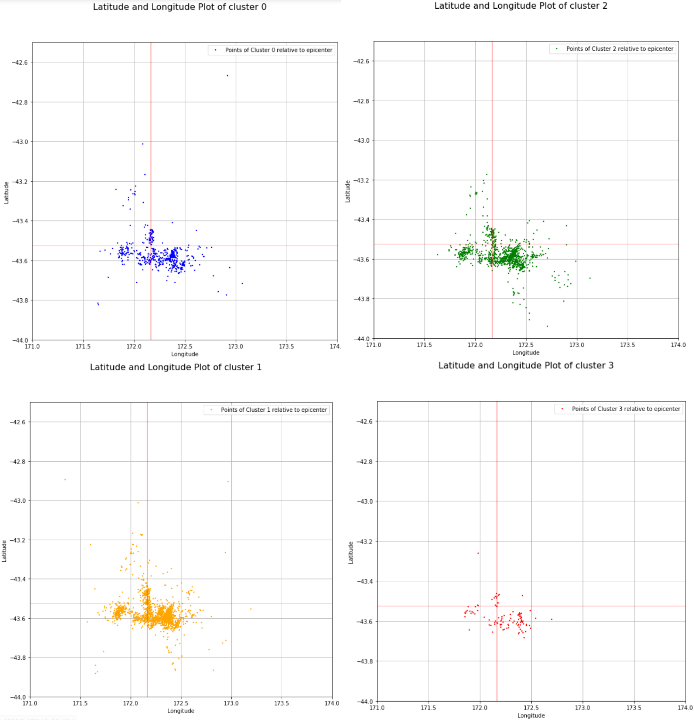


Figure 11 Distribution of KMEAN clustered Longitude Latitude Magnitude data. Epicentre belongs to cluster 3 Red Plot

### Longitude Latitude Depth

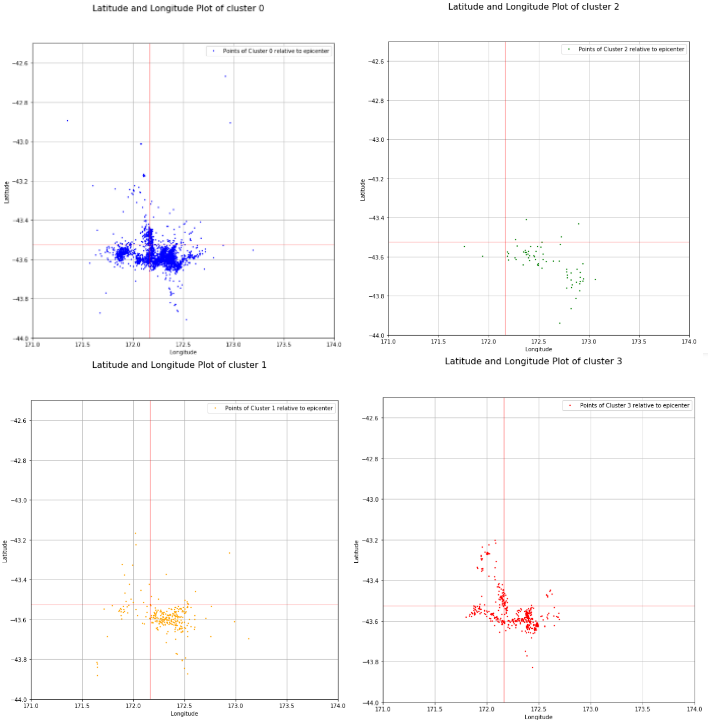


Figure 12 Distribution of KMEAN clustered Longitude Latitude Depth data. Epicentre belongs to cluster 1 orange plot

### Results and Interpretation

|  |  |
| --- | --- |
| Dataset | Clustering Time (s) |
| Longitude Latitude Magnitude | 0.109669 |
| Longitude Latitude Depth | 0.034909 |

K means performed exceptionally well in terms of execution time. The clusters it produced also look very fitting. For the spatial magnitude plot, all 4 produce a rough outline of the hypothesis plot with the epicentre belonging to the 3 cluster being the sparsest cluster.

For the spatial depth plot some variations start to appear. Cluster 0 and cluster 3 follow some the expected outline with cluster 3 less so. Cluster 1 (epicentre cluster) and cluster 2 show some drift towards the South East.

A pattern is starting to emerge.

# Evaluation

Two criteria are used to measure effectiveness of classification algorithms

## Average Sum of Squares Over All Clusters (SST)

Sum of Squares Error is the sum of the squared differences between each observation and the group mean. SSE metric quantifies the variation within a cluster. Since each cluster aims to discover similarity within a group, a lower SST would be better measure of accurate classification.

The SST formula is given as:



Where xi is the value of the ith instance and x̄ is the mean of that observation. SSE is calculated for every value Longitude, Latitude, and Magnitude/ Depth for each cluster and then averaged over total number of clusters which in this case = 4. However, only Magnitude/ Depth values are vital for evaluation.

### Longitude Latitude Magnitude

|  |  |  |  |
| --- | --- | --- | --- |
|  | Longitude Avg. SSE | Latitude Avg. SSE | Magnitude Avg. SSE |
| SOM | 26.05557535271597 | 4.399816953732415 | 40.40626656591273 |
| DBSCAN | 33.462030564312556 | 4.352618325125251 | 33.462030564312556 |
| KMEAN | 33.58748005614066 | 4.662553213345645 | 25.60493051691906 |

### Longitude Latitude Depth

|  |  |  |  |
| --- | --- | --- | --- |
|  | Longitude Avg, SSE | Latitude Avg. SSE | Depth |
| SOM | 31.617036306164493 | 4.522670832400999 | 927.8120068660448 |
| DBSCAN | 33.28472226150159 | 4.5371757027688115 | 8399.9292939727 |
| KMEANS | 31.639744392592966 | 4.5371757027688115 | 642.7419084731628 |

## Interpretation

From the Averaged Sum of Square Error (SST) we can see that K means consistently has the lowest error across all algorithms. Which makes it the winner in SST score.

Why is KMEANS the best? The application of massive data reduction to centroids proves effective in data mining earthquake aftershocks because

## Clustering Cluster Silhouette

Cluster Silhouette (CS) is another way of verifying accuracy of clustered data. CS is a graphical representation of the level of conformity of the objects to the cluster they belong to compared to other clusters they are not a part of. The average silhouette score ranges from -1 to 1 where values closer to 1 represents clusters of well-matched objects to the cluster and lower values represents objects more matched to neighbouring clusters. Silhouette score is a measure of belongingness of objects within their cluster.

### Longitude Latitude Magnitude

|  |  |  |
| --- | --- | --- |
| Clusterer | Avg. Silhouette Score | Cluster Silhouette |
| SOM | 0.270558661 |  |
| DBSCAN | 0.449124619 |  |
| KMEANS | 0.360174640 |  |

### Longitude Latitude Depth

|  |  |  |
| --- | --- | --- |
| Clusterer | Avg Silhouette Score | Cluster Silhouette |
| SOM | 0.764172552 |  |
| DBSCAN | 0.85422649 |  |
| KMEANS | 0.804693545 |  |

## Interpretation

From the cluster silhouette it is discovered that DBSCAN scored the highest average silhouette score. However, the clustering algorithm was wrongly tuned and has produced a single large cluster that has “eclipsed” all other clusters. The measure of conformity to a cluster is bound to be high when the cluster includes 99.8 / 99.9% of all instances. So for this evaluation we reject DBSCAN results. The comparison is mainly between SOM and KMEANS.

Based Average Silhouette score results. KMEANS clustering come out on top again.

# Predictive Analysis

Following the clustering K means clustering is selected as the best performing algorithm. Categorical values (A,B,C,D) are assigned to its clusters (0,1,2,3). A predictive classification analysis using WEKA is done to search for the location of the largest aftershocks i.e. “the worst places to live around Darfield”.

The data is loaded into WEKA and the clusters are converted into classes using the addCluster filter.

## J 48

J 48 is an implementation of c4.5 algorithm. It is an algorithm that mimics the human decision-making process. It is also very reliable and easy to understand like how k means is simple and easy to understand, however may suffer from overfitting and some low accuracy results. J48 algorithm is chosen because of the ease of interpertating of its results which allows for the logic behind the cluster labels decoded. This technique is used to find the predicted cluster.

It will allow us to see all the categories of magnitude

### Results

|  |  |
| --- | --- |
| Classifier: | J48 Decision Tree |
| Model Parameters |  |
| Binary Splits | False |
| Location of Epicentre | Cluster 1 |
| Decision Tree |  |

From the evaluation metrics in WEKA we use the decision tree structure. Observing that the Epicentre falls into Cluster 1 a mapping of all cluster 1 points is produced seen in FIgure

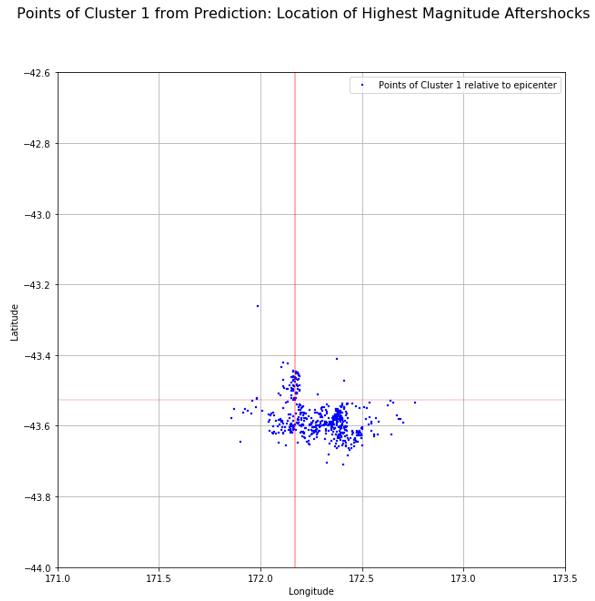


Figure 13 Predicted Location of Highest Magnitude Shocks

## Interpretation

From the classification Analysis we can gain the insight that The Earthquake prone regions are moving towards the East. So, to borrow an approach from statistics of Null Hypothesis.

# Conclusion

The data analysis revealed a drift of seismicity towards the south east for some reason. Towards Christchurch.

This type of data mining (spatial) is only the ground work for the serious task of earthquake data mining. A more impactful study would involve considering the evolution of clusters as a time series and re evaluating the effectiveness of each algorithm. All 3 algorithms will no doubt come up with their own advantages and disadvantages in this case.

To contribute personal background knowledge on the subject matter of Earthquake datamining. Very Long Baseline Interferometry VLBI is the practice of combining observations from radio telescopes of the arrival of the same signal at different times while observing a very bright source like object in the sky for example: Quasi Stellar Objects aka Quasars. By triangulating the distance of the two telescopes to the object, the distance between telescopes known as the baseline (B) can be measured to within a few cm of accuracy for baselines of up 1000 km [8]. This can be used to measure movement of plate tectonics over time. The same plate tectonics that causes the Earthquake. Perhaps an additional parameter related to the plate tectonic movement can be contributed to the dataset and the mining process via this method. One that will improve the accuracy and predictions of the data mining.

One that might save a lot of lives.

# References

[1] Darfield earthquake aftershocks: temporal evolution of the aftershock sequence, faulting and stress

[2] Savaş, Cihan & Yildiz, Mehmet & Eken, Süleyman & İkibaş, Cevat & Sayar, Ahmet. (2018). Clustering Earthquake Data: Identifying Spatial Patterns From Non-Spatial Attributes. 10.4018/978-1-5225-7519-1.

[3] Kamat, Rajanish & Kamath, Rajani. (2017). Earthquake Cluster Analysis: K-Means Approach. Journal of Chemical and Pharmaceutical Sciences. 10.

[4] Sanja Scitovski. (2018). A density-based clustering algorithm for earthquake zoning, Computers & Geosciences

[5] J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map," in IEEE Transactions on Neural Networks, vol. 11, no. 3, pp. 586-600, May 2000.

[6] Vahid Moosavi SOMpy: <https://github.com/sevamoo/SOMPY>

[7] Novinanti, Pepi & Setyorini, Dyah & Rafflesia, Ulfasari. (2017). K-Means cluster analysis in earthquake epicenter clustering. International Journal of Advances in Intelligent Informatics. 3. 81. 10.26555/ijain.v3i2.100.

[8] Peter F. MacDoran. (1974). Radio interferometry for international study of the earthquake mechanism. Acta Astronautica. Volume 1, Issues 11–12, Pages 1427-1444,

# Code Resource

Github: <https://github.com/coderXmachina2/adventuresInMachineLearning/tree/master/ass2>

\*apologies for the messy directory, I haven’t gotten about to arranging the files yet but