Principles of Data Science Project Report
Rory Hurley

Causing the Quakes:
Seismicity and Recent Trends in Induced Earthquakes

Computational notebook link:
https://smcse.city.ac.uk/student/aczd067/inm430/INM430_Rory-Hurley_Final-Submission.html

1. Introduction

In spring 2011, two tremors were felt in Lancashire, UK. This type of activity is unusual in England. There are no active geological faults, and the strongest earthquake occurred over 80 years ago causing little damage. These events were caused by 'Fracking' - the process of pumping chemical liquid at high pressure into the Earth's crust and extracting natural gas as it is forced out. These chemicals are toxic, dangerous to ecology and could potentially contaminate water sources. One effect is seismic movement and earthquakes. Fracking has been on the rise in the UK after becoming popular in the USA, and as a result of the Lancashire tremors, fracking activity was quickly halted due to environmental concerns and public anger.

In October 2018, after a seven-year hiatus, fracking activity has had a resurgence in the UK. However, the public is still very sceptical on about process, and protests have been coordinated outside fracking sites resulting clashes between protesters and the police. It is clear that fracking is a contentious social, environmental and economic issue in modern day Britain. Its supporters claim it is necessary following depletion of traditional fossil fuels and rising price of crude oil. Its critics link it to pollution of groundwater stores and induced earthquakes in the USA. Combining seismic, underground injection and social media data, this project will focus on the contemporary issues around seismicity and fracking today; risk classification, spatio-temporal patterns, relationship to underground fluid injection and the social attitudes towards the practice.

1.2. Research Questions

1. Can machine learning be used to model and classify areas according to earthquake risk?
2. What have been the spatio-temporal seismic patterns in the USA since 1975 to present day?
3. Can seismic activity be predicted using underground fluid injection data?
4. Can public opinion toward fracking be effectively quantified using social media data?

1.3. Research Objectives

1. Apply RFM modelling and K-means clustering in order to segment USA counties according to earthquake risk. Observe patterns, trends and model performance.
2. Use publicly available earthquake data to explore spatial and temporal trends in seismic activity, and identify any anomalous or interesting patterns if present.
3. Combine earthquake and UIC data in order to perform regression analysis to predict seismic activity using underground fluid injection data.
4. Mine Twitter data and perform sentiment analysis on tweets related to fracking.

2. Analysis Strategy & Plan

Using seismic data from the USA, groundwater injection (UIC) data from Oklahoma and social media data, the project will start by combining RFM modelling and k-means clustering to classify US counties according to earthquake risk. The output will be explored alongside spatio-temporal analysis to detect anomalous seismic activity. Using information from these outputs, earthquake events and UIC activity will be modelled using linear regression to predict seismic activity from UIC data, before twitter data is analysed to assess social attitudes towards the practice.

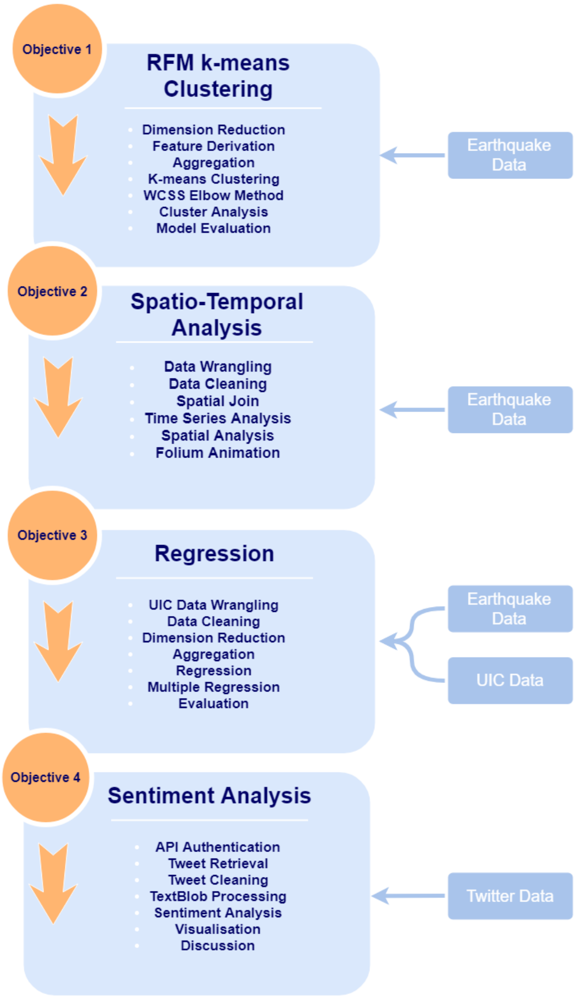

Figure 1- Analysis Workflow

3. Analysis

https://smcse.city.ac.uk/student/aczd067/inm430/INM430_Rory-Hurley_Final-Submission.html

4. Findings and Reflections

Research Objective 1 was to see if counties in the USA could be segmented into different risk categories using RFM modelling and k-means clustering. It was found that after log-normalisation was applied to the data, there was much variation between counties. Some had high values for frequency and magnitude but low recency values, indicating seismic activity over a long time and still active recently. On the other end of the scale, some counties scored high on recency, but low on magnitude and frequency, indicating 'Low Risk' earthquake counties, where there may have been a rogue earthquake in the past. K-means clustering was applied to the normalised RFM model. After iterating the model fifteen times, the optimal number of clusters was 4, as the within cluster sum of squares metric WCSS was reduced minimised. The resulting cluster plot is shown on Figure 1.

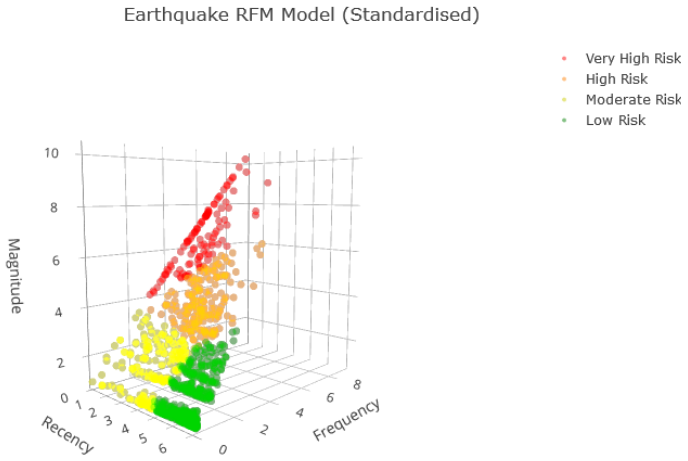

Figure 2 - RFM k-means clustering of US Counties.

These results (Figure 3) show that the model performed well with regards to the evaluation requirements at the start of part 3.1.2.4. It is shown that the model exceeded expectations, as 84.8% of 2018 earthquakes occurred in 'Very High Risk' counties. The 'High Risk' county requirement was not met as this was 7.88% rather than 10% - however, this is attributable to the stronger than expected performance of the 'Very High Risk'. The remaining categories performed well as desired. 1.05% of earthquakes occurred in unclassified counties. This is a slight limitation in the model, and suggests that any counties that are not classifies after RFM clustering should be assigned to 'Low Risk' counties.

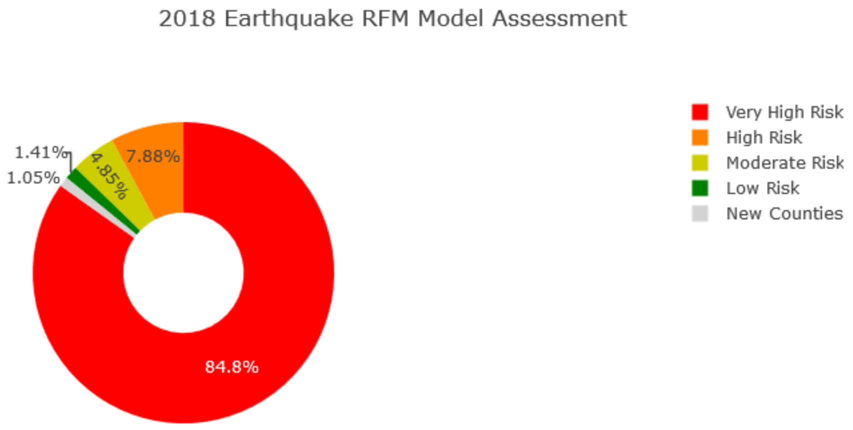

Figure 3 RFM k-means model earthquake model evaluation.

The performance of this model shows that RFM modelling adopted from marketing, coupled with machine learning can provide a good means of assessing seismic risk. This is relevant for many modern day domain such as insurance pricing, local planning and property.

The findings from accomplishing Objective 1 gave some insight into addressing Research Question 2. Spatial analysis of the RFM clustering results indicated that there certain areas that were more prone to seismic activity over time. These included large parts of west USA and pockets around Oklahoma and some smaller clusters near Tennessee (Figure 4).

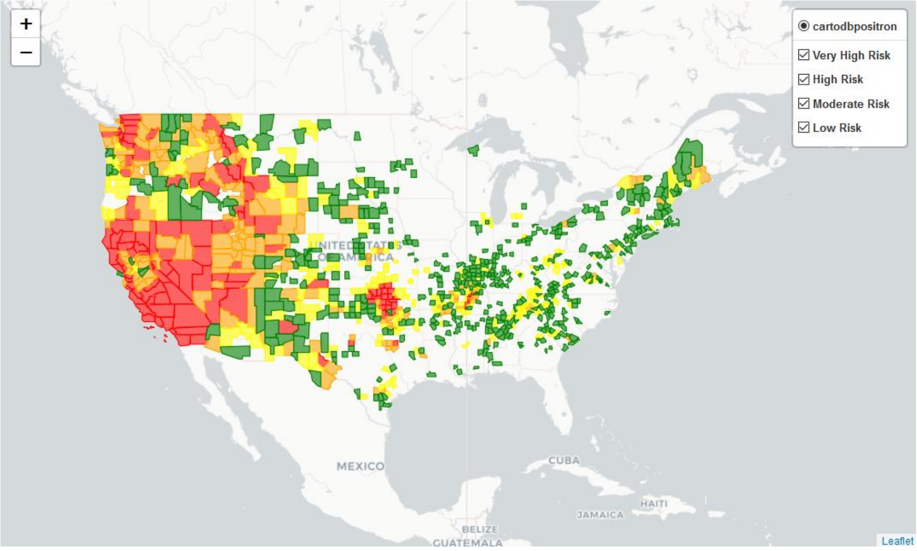

Figure 4 - Spatial distribution of counties classified using the RFM k-means model.

Time series analysis showed that annual earthquake occurrences were quite variable across the USA. Regional analysis found that west USA, located along the active San Andreas Fault, drove most of the USA's seismic activity, with other regions only contributing around 50 per year. This pattern was consistent until 2010, where there was a rise in the number of south-western earthquakes. The number of earthquakes in the southeast outstrips that of the West in 2014 and becomes the main driver of earthquakes. This was due to high numbers of earthquakes in Oklahoma; the number of which peaked at 2,880 in 2015. Findings from the spatio-temporal analysis and RFM modelling led to Oklahoma becoming the focus of further investigation.

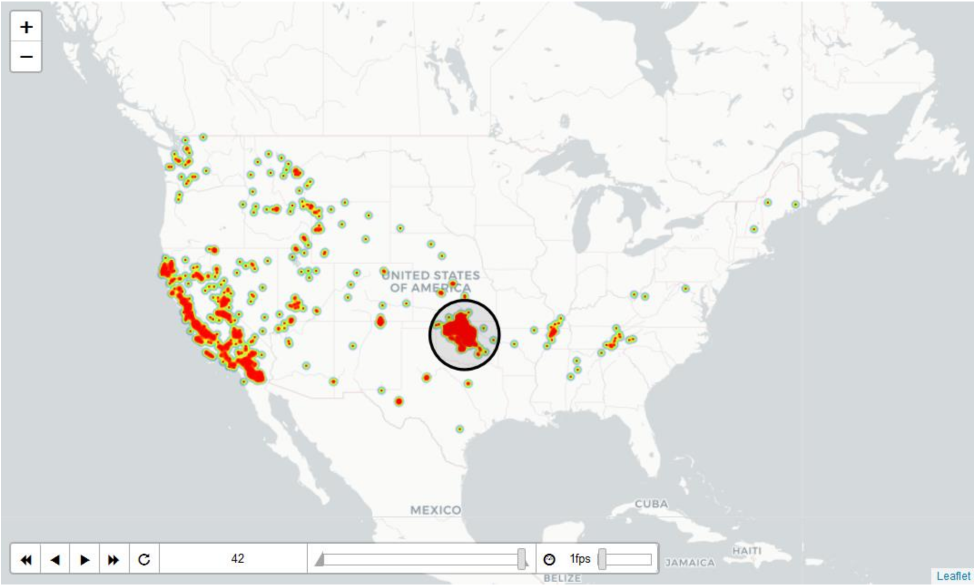

Figure 5 - Recent evidence of seismic activity in Oklahoma.

From domain research, it is was deemed that the large amount of recent seismic events could be attributed to high levels of UIC activity. The focus of the project then moved towards investigating whether there was a relationship between UIC activity and earthquakes. The time series analysis showed that seismic activity was responsive to two aspects of UIC activity, volume and pressure). Both rose steadily from 2011 onwards before rising sharply from 2014 to 2016 (Figure 6), before declining after 2016. Seismic activity rose later, but sharply, from 2014, before falling post-2016 like UIC activity. The similar patterns observed in the time series indicated that there was further cause for investigation.

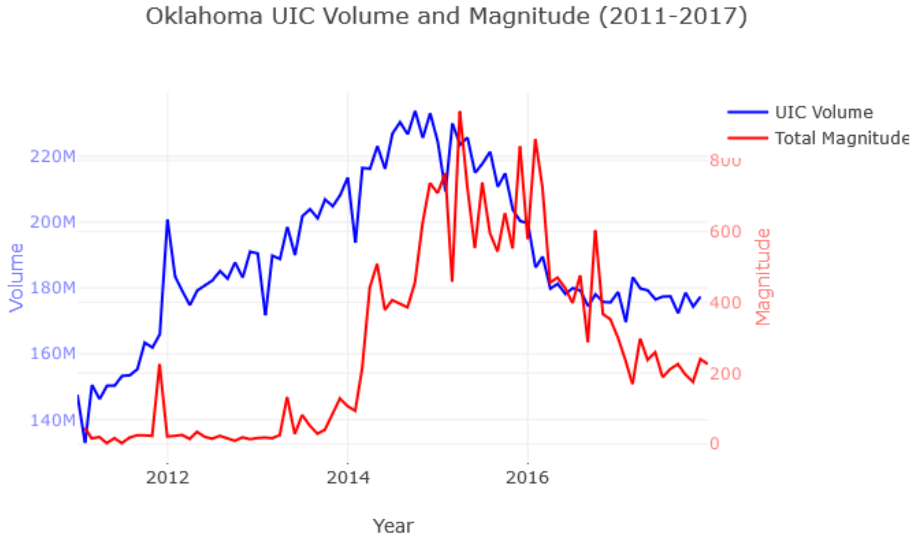

Figure 6- Relationship between montly UIC injection volume and seismic activity in Oklahoma.

For Objective 3, regression analysis was then used to assess whether the relationship between the phenomena meant that seismic activity could be predicted from UIC data. The regression analysis found that there appeared to be a positive correlation between UIC injection volume and seismic activity.

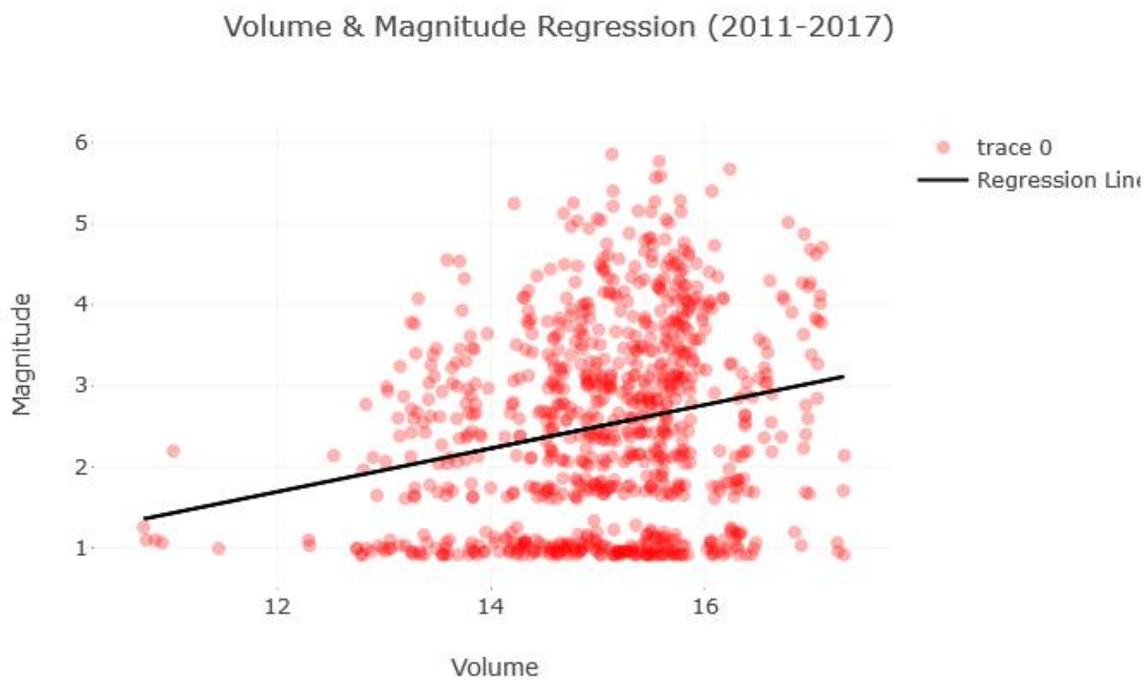

Figure 7- Monthly county UIC injection volume and magnitude regression.

The correlation was weak at 0.26 with R^2 value of 0.048. This was most likely due to the high level of variability in observations as seen in Figure 7. UIC injection pressure and seismic activity were then modelled, which gave a negative correlation of -0.33 and R^2 value of 0.05, again due to the variability of observations. This indicates that UIC injection volume and UIC injection pressure are not robust predictors of seismic activity. Carrying out multiple regression using these independent did not improved the regression model. The model was found to have an average R^2 value of 0.315 after 5-fold cross validation that was statistically significant. These findings indicate that a combination of both underground injection volume and pressure data are not good predictors of seismic activity. This is likely due to the unpredictability of seismic activity, and could be elaborated on further when consulting with seismologists.

Sentiment analysis was carried out on 5,660 tweets containing the hashtag '#fracking'. Using the TextBlob API, tweets were given a score for both sentiment polarity and subjectivity. TextBlob deemed that 4,447 tweets were positive, and 1,213 were negative about fracking. Visualisation of the tweets (Figure 8) shows this. Further investigation showed that there were flaws in the technique, as many tweets were classified incorrectly due to sarcasm. This is a common problem in contemporary natural language processing, and until more sophisticated methods are built, twitter data may not be a useful tool to gauge public opinion on fracking.

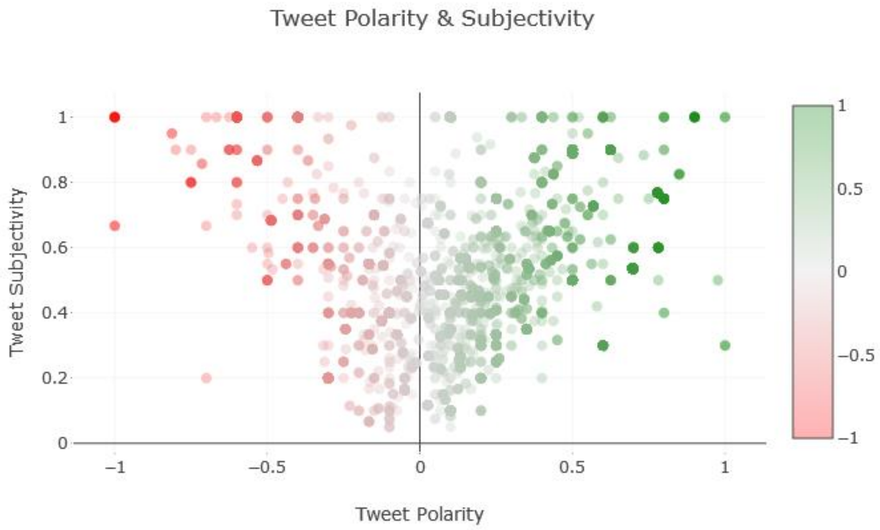

Figure 8- TextBlob sentiment scoring results for tweets containing '#Fracking'.

The project was successful in addressing its research questions and achieving its objectives. The main findings were that RFM modelling can be used with unsupervised learning in order to segment areas according to seismic risk. This model allowed high and low risk areas to be identified, which could be very beneficial for a multitude of domains. It was also found that there is some evidence to believe that there is a relationship between UIC activity and seismic events, however it was proved that UIC volume and pressure data does not allow for robust prediction of seismic activity. Finally, whilst twitter data produced a good collection of public opinion related to fracking, the findings were not conclusive, mainly due to limitations of natural language processing libraries.

References

1. Earthquake Data – USGS https://data.usgs.gov/datacatalog/#fq=dataType%3A(collection%20OR%20non-collection)&q=*%3A*
2. US Shapefiles – United States Census Bureau https://www.census.gov/geo/maps-data/data/cbf/cbf_counties.html
3. UIC Data – Oklahoma Corporation Commission http://www.occeweb.com/og/ogdatafiles2.html
4. Twitter Data – Twitter Standard Search API https://developer.twitter.com/en/docs/tweets/search/overview