# East Coast Seismicity Cluster Analysis in the U.S.A.

Donald H. Scott IV, Dean J. Kroker, Andrea Stiffelman
Lehigh University, dhs217@lehigh.edu, djk216@lehigh.edu, ars216@lehigh.edu

***Abstract* – In an effort to analyze areas of apparent elevated nontectonic seismic action in the North-East United States, we explore applying modern data clustering algorithms to seismic events in the region. In analyzing this data, we hope to identify interesting or unexpected results, as well as to qualify the degree to which seismic activity is clustered into certain zones. Several datasets were used in order to accumulate an accurate catalog of earthquake events and seismic station data. These datasets were plotted in R, visualized via R Shiny, and clustered via geodesic density-based methods.**

*Index Terms* – Clustering, Nontectonic, Seismicity

## INTRODUCTION

Seismic monitoring on the East Coast of the United States progressed significantly from the first recorded events in 1800. In 1972, the first seismometer was installed in South-East Pennsylvania. After additional stations become being installed in 1985, the number of seismic stations increased significantly, to forty-eight stations by 2008. Since then, the introduction of the Advanced National Seismic System (ANSS) has increased the number of seismometers to 246. Within the study region, this caused an increase in both the quality and quantity of data. The improved network has provided a catalog with improved magnitude completeness ($M_c$) which has made apparent specific zones of what appear to be elevated seismic activity.

These regions consist of the faults and "seismic zones" depicted in *Figure 1*. These zones can be labeled as follows:

1. The Greater Ramapo Fault Area
2. The Reading/Lancaster Seismic Zone
3. The Central Virginia Seismic Zone
4. Blue Ridge/Hayesville Fault

This research aggregates numerous catalogs and demonstrates several algorithmic methods to cluster the dataset ignorant of the four manually identified seismic zones.

## EXPERIMENTAL DESIGN

In an effort to simplify data analysis as well as promote an incremental research approach, we developed an easy to manipulate tool which analyzes datasets in a geodesic manner, for a specified time period, in order to allow the researcher to gather statistical and qualitative conclusions.

The specific study parameters for this effort are events between 1800 and 2015, in the region +35.5° to +43.5° Latitude, -84° to -71° Longitude.

IRIS station data and is plotted with respect to latitude and longitude. ANSS and NEIC earthquake data is combined and plotted with respect to latitude and longitude. The application converts given magnitudes to Moment magnitude ($M_w$) and indicates the magnitude of each earthquake on the 2D earthquake plot.

The combination of NEIC and ANSS data allows for preparation of accurate earthquake data through time. Data verification indicates that no station provided data before installation or after closure and that no duplicate events exist.

Geodesic density-based clustering (DBScan) is implemented graphically for cluster analysis over longitude, latitude, and depth. This function contains the ability to modify the reachability minimum number of points (MinPts) and reachability distance (EPS).

Kernel density estimation is implemented across a three-dimensional plane in order to visualize multivariate data through time [1].

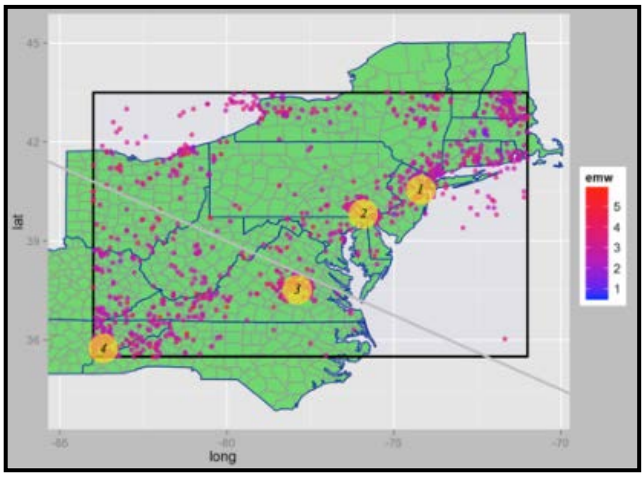


FIGURE 1
AREAS OF INTEREST IN THE NORTHEAST UNITED STATES

## DATA ANALYSIS

Before clustering analysis could be executed, raw data needed to be refined and reorganized from the source earthquake information sets. This process entailed two primary tasks: ensuring catalog completeness and ensuring catalog uniformity via consistent units of measure. In

addition, ancillary steps were taken to ensure catalog integrity as well as to cross-validate source information. Provisional steps included date formatting, filtering of records with parameters outside the region of study and conversion of magnitude measures to a uniform scale.

With regards to the first task, catalog completeness is addressed by joining catalog data from several sources into horizontal catalog entries of date, time, location, and other relevant attributes. It is important to note that as seismometers were installed, depth, magnitude, location, quantity and overall accuracy improved substantially. Prior to 1972, earthquake data is less finite when compared to the new IRIS network. Thus, detected seismicity increased drastically as the network expanded. (*See Figure 2A-D.*)

The second task needed for catalog preparation spoke to the uniformity of earthquake size measures, like magnitude, across datasets. The main step taken in this process was the identification and conversion of magnitude data presented directly from its station source to a scaled magnitude that is consistent with modern ground motion prediction equations. By evaluating whether a station has presented this data as a ML, Mb, Md, Mx, Mh, Mc or Unk type, one can apply the appropriate scaling equation to the Body-Wave magnitude so that its resulting uniform $M_w$ value can be used to develop unbiased estimates of the recurrence of earthquakes. Two methods were used to conduct cluster analysis: *DBScan* and *Kernel Density Estimation*. These methods will be reviewed in depth [2].

## INTERPRETATION

A significant source for interpretation was the set of statistical graphs representing event frequency versus magnitude and depth, as well as cumulative event frequency versus magnitude and time. It is useful to compare magnitude in both frequency and cumulative frequency models since the use of cumulative data makes patterns more easily identifiable. The first derivative of the resulting function highlights changes to the behavior of the function as a function of the independent variable.

The traditional graph, in contrast, shows the number of counts in each bin of a histogram, where empty bins represent spans where no events were recorded. The measure of frequency can be presented in different units including: counts (such as total number of events over some time period), rates (such as the average number of events in some time period), or a probability based on a frequency distribution. In this case, cumulative frequency was used as a count measure against time so that it could be used to estimate seismic activity without background noise [3].

After scaling the earthquake catalog's observations of moment magnitude to a uniform $M_w$ value, as previously described, the graphs of magnitude as a function of frequency are plotted in order to fit a Gutenberg-Richter (G-R) model to our data, which forms the basis for seismic hazard studies and earthquake forecast formularies.

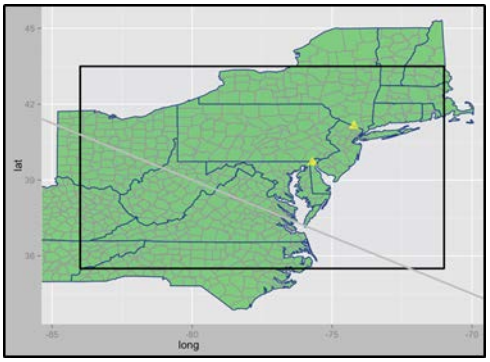


FIGURE 2A
IRIS STATIONS THROUGH TIME: PRE-1975 – 2 STATIONS

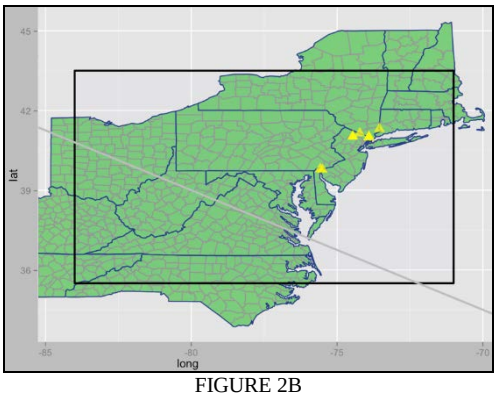


FIGURE 2B
IRIS STATIONS THROUGH TIME: 1975-1985 – 6 STATIONS

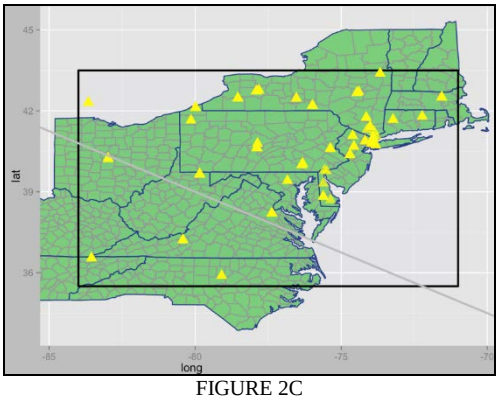


FIGURE 2C
IRIS STATIONS THROUGH TIME: 1985-2008 - 44 STATIONS

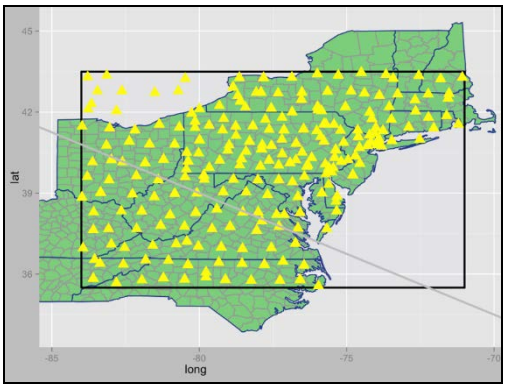


FIGURE 2D
IRIS STATIONS THROUGH TIME: 2008-PRESENT - 234 STATIONS

This model is essential to accurately determining the magnitude completeness. This measure, called, $M_c$, is defined as the lowest magnitude at which one-hundred percent of the earthquakes in a space-time envelope are detected. Precise estimation of this statistic is crucial as a value too high indicates under-sampling, while a value too low *may* be indicative of noise.[1]

In order to properly represent the frequency/magnitude to $M_c$ relationship one must first transform the cumulative frequency axis to a logarithmic scale so the trend can be viewed as log-linear. Since the cumulative frequency is generated by starting at the highest observed magnitude and working back to the lowest (summing the frequencies through the progression) one would expect to see a linear trend-line if $b \approx 1$ in the study area. The frequency-magnitude graph is used as a certification that the calculated parameters are consistent with the data [4]. Taking into consideration the underlying assumption of self-similarity, established by the G-R law, it can be observed that $M_c$ is simply the magnitude increment at which the graph distribution departs from the linear trend in the log-linear plot. (*See figure 3.*)

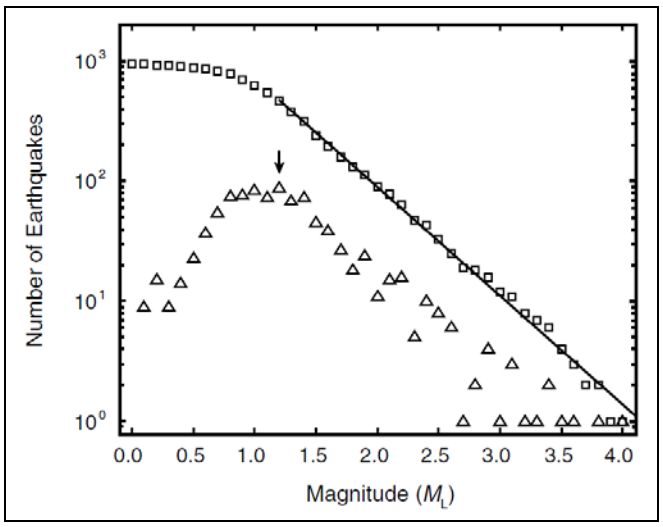


FIGURE 3
IN THIS EXAMPLE GRAPH, PROVIDED BY LOCKRIDGE ET AL [5], THE TRIANGLES REPRESENT THE NUMBER OF EVENTS HAVING OCCURRED AT THAT MAGNITUDE. AS SHOWN, MC IS THE LOCATION IN WHICH THE FREQUENCY BINS ARE MAXIMIZED, AND AT WHICH THE CUMULATIVE TREND-LINE BECOMES NONLINEAR.

In order to assess the relationships identified by the G-R law, a non-parametric regression algorithm, called LOESS, was utilized to fit a curve to the data. Specifically, $M_c$ was calculated using the Maximum-Curvature Method (MAXC). The basis of this technique is to evaluate changes, or possible breaks in the slope, across the frequency-magnitude distribution. The approach of using MAXC to create a LOESS curve is favored since it is straightforward and statistically robust. While it has proven in other experiments to result in a $M_c$ value lower than that calculated by alternate techniques [6], it is reliable with sample sizes of varying length, and is plainly verifiable.

Using the MAXC method $M_c$ was calculated by computing an estimate of the first derivative of the frequency-magnitude curve using finite sums.

$$f'(x) \approx \lim_{h \to 0} \frac{f(x+h)-f(x)}{h} = \frac{\Delta_h[f](x)}{h} \quad (1)$$

Through (1), a simple iterative test can find the maximum value of f(x) with respect to x. This value represents the point of maximum curvature, and is therefore $M_c$.

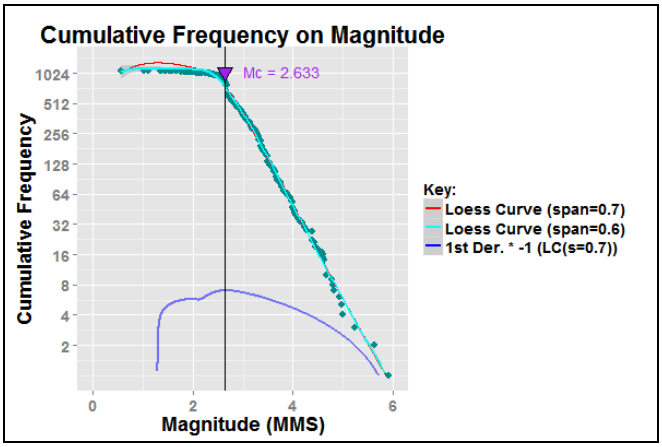


FIGURE 4
VISUAL REPRESENTATION OF THE MAXC METHOD TO FIND $M_c$.

Critical to this process is understanding the inherent statistical scatter associated with the exponential distribution. Frequency magnitude plots record events taken from finite samples. The events in the tail of the distribution are probabilistically less likely, so the difference in recording one event makes a proportionately large difference to the observed counts. By contrast, at low magnitudes where there may be hundreds of events in each bin, the difference in observing a single event has a proportionately small effect.

We need to understand the scatter in the frequency-magnitude data because this is indicative of the amount of scatter we should expect to observe when we look at data drawn from an exponential distribution. These fluctuations, as graphically demonstrated by the frequency-magnitude distribution, raise several points of observation. By taking note of the wide spread in the counts at high magnitudes generated by sampling the exponential distribution directly, it can be observed that, due to sampling effects, this is the statistical noise one should expect to see in earthquake frequency-magnitude data. The spread of the bounds on the cumulative and incremental data, when compared, reveal that the gradient at low magnitudes is the same but the

[1] It is worth noting that noise in an unweighted clustering algorithm would certainly bias results in a material manner, since all events are considered equally.

curves are onset from each other. The cumulative graph shows an apparent roll off at high magnitudes. Roll-off is expected for frequency-magnitude data and is ultimately controlled by the largest possible earthquake in a given region. However, the appearance of roll-off does not necessarily imply that the largest possible earthquake has been observed since a comprehensive sample may take more than a millennium to observe directly. By looking at the intervals plotted on the incremental data graph, it can be confirmed that the sample does indeed lie well within the statistical noise; moreover, the lack of an acutely sharp roll-off demonstrates that it does not lie outside the region of statistical noise. Fortunately, the absence of a disproportionately large earthquake means that a selection bias is not required.

With these observations made, a density based clustering algorithm, DBSCAN[2], was leveraged to generate initial clustering maps of the study area. In *Figure 5* one such graph is shown.

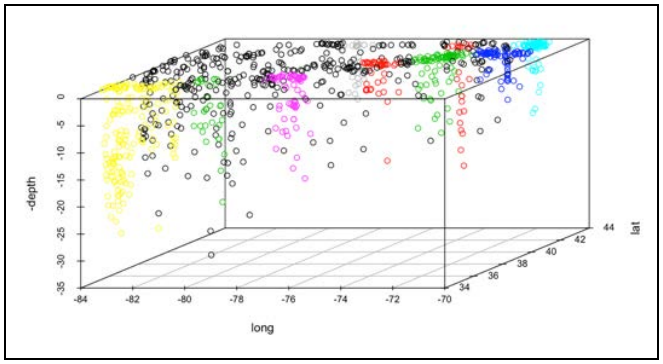


FIGURE 5
GEODISTIC CLUSTERING VIA DBSCAN: 1800-2015
MINPTS: 20, EPS: 43

One of the considerations with using the DBSCAN algorithm is the determination as to what the input parameters should be; i.e. what constitutes a cluster? In order to better visualize the 3D clusters, we felt it was prudent to utilize a new algorithm to discriminate upon the dataset without human bias. This led to the introduction of *Kernel Density Estimation* (KDE). KDE is a non-parametric algorithm to estimate the probability density function of a random variable. In our case, we sought the probability density function of the likelihood of a seismic event.

This algorithm is rooted in function (2), that is, the probability of k factors (of N total) falls within the region $\Re$.

$$\left.\begin{aligned} P &= \int_{\Re} p(x')dx' \cong p(x)V \\ P &\cong \frac{k}{N} \end{aligned}\right\} \Rightarrow p(x) \cong \frac{k}{NV} \qquad (2)$$

Using (2) in conjunction with Parzen Windows [7], a kernel function is defined, which is used to estimate relative probability as well as to generate 3D graphs, like the one shown in *Figures 6-8*.

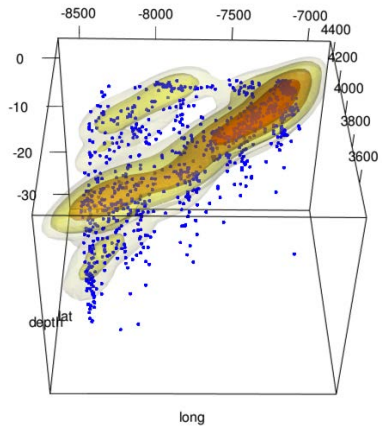


FIGURE 6
TOP-DOWN VIEW OF KDE CLUSTERING: 1800 TO 2015

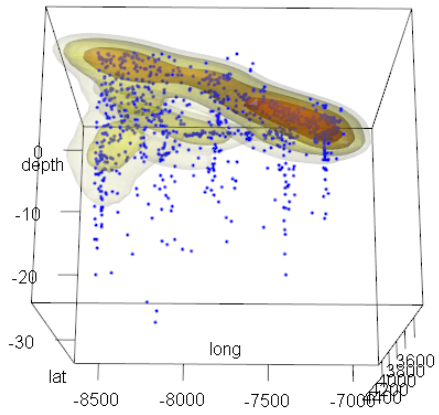


FIGURE 7
HORIZONTAL VIEW OF KDE CLUSTERING: 1800 TO 2015

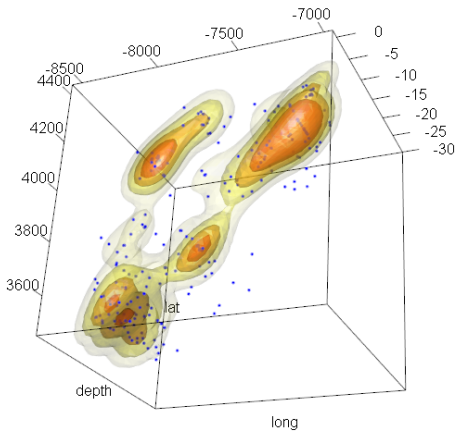


FIGURE 8
KDE CLUSTERING: 2005 TO 2015

When analyzing the KDE renderings, there are two important observations: many events in the historical catalog lacked accurate depth data (and are therefore rendered at a depth of zero) and there are in-fact only four relative clusters of seismic activity. *Figure 8* is most interesting as this figure can largely be considered the most accurate given the

substantial improvements in station coverage as shown in *Figure 2A-D*, as well as the relevant $M_c$ calculations, briefly discussed on page 3.

Another interesting remark is that the cluster in the South-West of the study area appears to have substantially deeper seismic activity than that of other areas in the region. This is likely indicative of differences in lithologic makeup or structure which is causing slippage deeper in the earth's crust. This particular region might be an interesting point of study for future analysis.

## CONCLUSION

To date, there appears to be little academic literature applying data mining techniques to seismic activity in Eastern United States of America. While this particular project discovered no novel or remarkable findings, it served as a building block for future research, through the development of useful tools and visualization methods.

Moving forward, iterative improvements on the approaches and tools used in these endeavor will hopefully serve to advance research in the field.

## ACKNOWLEDGMENT

The authors are grateful this project was sponsored by Dr. Anne Meltzer, with substantial guidance from Lillian Soto-Cordero. We appreciated the time and effort dedicated to helping this project come to fruition.

## REFERENCES

[1] "Kernel density estimate," *kde {ks}*. [Online]. Available: http://www.inside-r.org/packages/cran/ks/docs/kde.

[2] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," 1996, pp. 226–231.

[3] M. Naylor, K. Orfanogiannaki, and D. Harte, "Exploratory data analysis: magnitude, space and time," Oct. 2015.

[4] A. Mignan and J. Woessner, "Understanding Seismicity Catalogs and Their Problems: Estimating the Magnitude of Completeness for Earthquake Catalogs,." CORSSA, 2012.

[5] J. S. Lockridge, M. J. Fouch, and J. R. Arrowsmith, "Seismicity within Arizona during the Deployment of the EarthScope USArray Transportable Array," *Bull. Seismol. Soc. Am.*, vol. 102, no. 4, pp. 1850–1863, Aug. 2012.

[6] J. Woessner, "Assessing the Quality of Earthquake Catalogues: Estimating the Magnitude of Completeness and Its Uncertainty," *Bull. Seismol. Soc. Am.*, vol. 95, pp. 684–698, Apr. 2005.

[7] Sebastian Raschka, "Kernel density estimation via the Parzen–Rosenblatt window method." 19-Jun-2014.

# Documentation

**Installation & Setup**

- Download and install R
- Download and install RStudio
- If Mac: Download and install XQuartz
- Open Terminal
- If you do not have Git installed, follow instructions to install Git.
- Type 'git clone https://github.com/don4of4/intraplate-seismicity.git' without the apostrophes
- Enter Github User and Password

**Execution:**

- Open RStudio
- File/Open File...
- Navigate to app.Rproj in the intraplate-seismicity folder (just added via Github)
- Within RStudio, under Files, click import.R, highlight the entire contents of the file, and press Run
- If needed, open server.R and press Run App

**Updating:**

- Open Terminal
- Navigate to intraplate-seismicity
- type 'git pull' without the apostrophes

**Troubleshooting:**

- 'Error: There is no package called ShinyRGL' – This error is indicative that XQuartz is not installed on Mac. If you just installed it, restart your R session. You should not recieve this error on PC.
- 'Error: object ‘dataset’ not found' – This error is indicative that you have not executed your import.R file. Stop the program execution, highlight contents of import.R, and execute the program.

**Directory Structure:**

```
intraplate-seismicity
├──  data/
├──  docs/
│   ├──  ExceptionReport.docx
├──  app.Rproj
├──  import.R
├──  LICENSE
├──  README.md
├──  server.R
├──  stations_completeness_validation.R
├──  stations_date_validation.R
├──  ui.R
├──  util.R
```

**Additional Guidance:** Be sure to review the repository readme, as well as applicable code comments. If you still cannot find an answer to your problem, please reach out to us.