Alexandra Plassaras
Amp2261
3/14/16

G4071: Midterm Exam:

Due Sunday March 13th (midnight via CourseWorks submission)
Use the following data to answer the questions below. Write a brief response to each questions (no more than a paragraph for each) and accompany that response with an appropriate visual image/map from which your response references. All maps should include titles and legends at a minimum.

Notes from Professor:
There are two points that I would like to correct on the midterm:

1) please perform any interpolation in QGIS. I have attached a lab on doing this in your midterm materials .zip folder. The reason being is that I forgot that you needed a pre-defined grid to interpolate to using the kriging functions. I failed to provide you with that grid.
2) In identifying spatial clusters of points, the L-function may take an unreasonably long time to finish computing. If that is the case... please feel free to use the F or G-function in place of the L-function as support/evidence of clustering.

PART I DATA: Toxic Emissions and Racial Inequality

For this hypothetical research project, you have access to 3 date files and may need to access additional data per instructions in the Labs we have worked through this semester. The three data sets supplied are listed here:

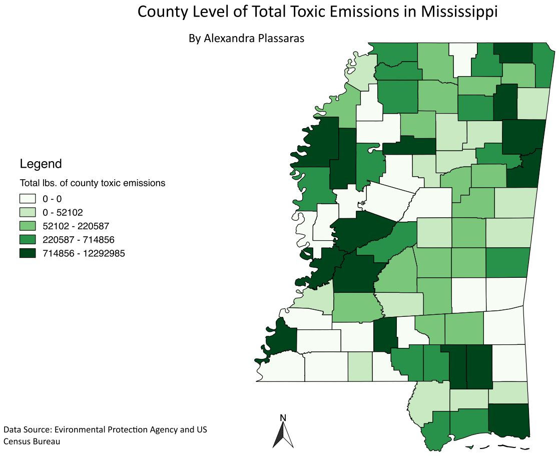
1. **MS_CTY_Centroids_SinglePart.shp** (and accompanying files)
   a. Shapefile of centroids of Mississippi Counties (point file).

2. **MS_Inequality_Data.dbf** (census data containing the county IDs and these variables)
   a. DSEG: County level Black-White residential segregation index, ranging from 0-1 with 1 indicating the highest levels of segregation.
   b. GINI: County level Black-White income inequality, ranging from 0-1 with 1 indicating the highest levels of inequality.
   c. PCTBLACK: The percent of the county population that is categorized as "Black" by the census bureau.

3. **Toxic_Release_Data.dbf** (EPA data containing county IDs and these variables)
   a. TOTAL: Total lbs. of county toxic emissions reported by EPA
   b. AIR: Total lbs. of county air toxic emissions reported by EPA
   c. WATER: Total lbs. of county toxic water emissions reported by EPA
   d. LAND: Total lbs of county toxic land emissions reported by EPA

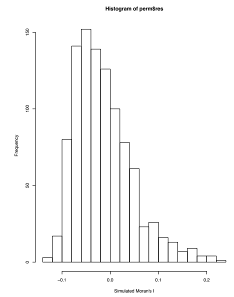
<u>Instructions/Directives:</u> Using the data above and ANY GIS/Spatial Packages you want….

1. **Visualize the <u>county level total toxic emission</u> release data by linking the data in the <u>MS centroids</u> file to a county level polygon coverage for the state of Mississippi and discuss any apparent first order spatial heterogeneity associated with your ideas of clustering, randomness, etc. (subjective interpretation/non-statistical).**

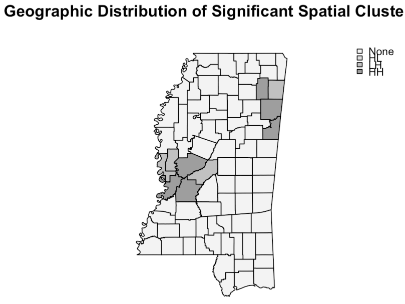

County Level of Total Toxic Emissions in Mississippi
By Alexandra Plassaras
Legend
Total lbs. of county toxic emissions
0 - 0
0 - 52102
52102 - 220587
220587 - 714856
714856 - 12292985
Data Source: Evironmental Protection Agency and US Census Bureau
N

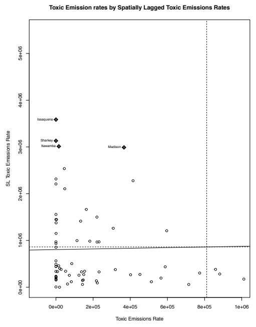
From the image above there appear to be instances of clustering in our data of toxic emissions. We can see that there appear to be potential clustering near Jackson, Greenville, Columbus and Hattesburg. After conducting further analysis using geographically weighted regression and k-riging models we will have a better idea on the clustering patterns of our data.

2. **Using the county level polygon Shapefile, is there global clustering (spatial dependence) in regards to the values of total toxic emission release. Produce both the global scatterplot and the Local Cluster map…. interpret both.**



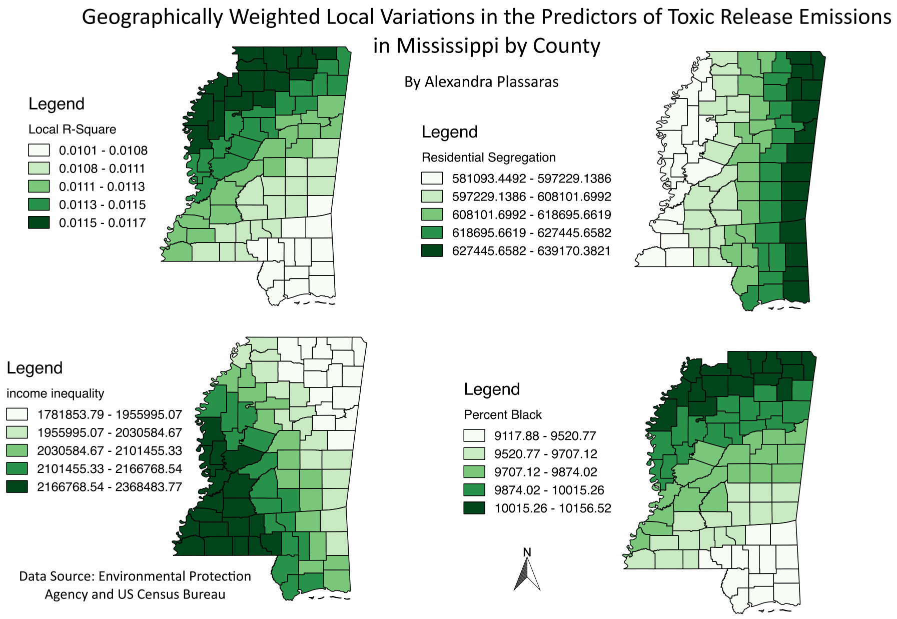
Spatial Distribution of Toxic Emissions
Mississippi County Level
Moran's I = 0.069795038
p-value = 0.08802





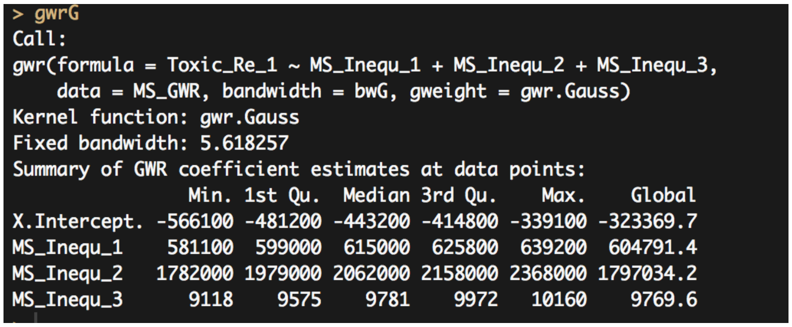
From the above graphs, it appears that there is global clustering in regards to the values of total toxic emission release. More specifically we can see that the regions around Jackson, Mississippi as well as near Starkville and Columbus, Mississippi are showing significant spatial clusters. The toxic emissions rates by spatially lagged toxic emissions rates graph shows that counties like Madison, Sharkey and Issaquena (near Jackson) have higher levels of spatially lagged total emissions. Itawamba county which is located in the Northeast also shows higher levels of spatially lagged total emissions rates. We can also see from the graph above that there is a positive association between an area's total emissions rate and neighbors' total emission rates.

3. **Run a set of Geographically Weighted Regression models using the centroids file in order to obtain the local R-Square and the local slope coefficients associated with the effects of residential segregation, income inequality, and the percent black on the levels of total toxic release emissions by county. Interpret the global regression results using the Median R-Square and slope coefficients for a baseline understanding of average relationships.**


Geographically Weighted Local Variations in the Predictors of Toxic Release Emissions in Mississippi by County

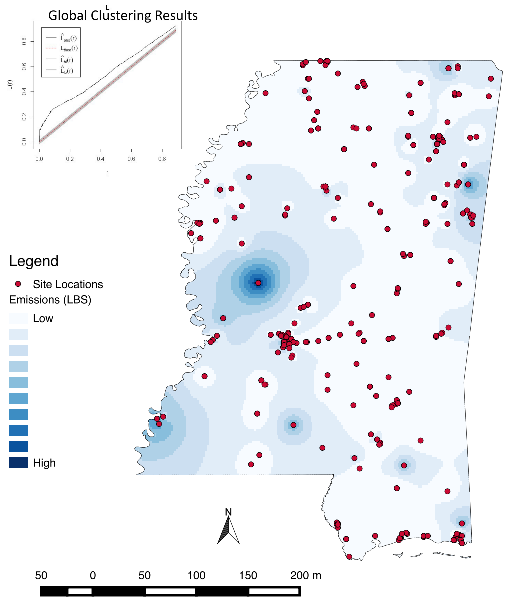
By Alexandra Plassaras

Data Source: Environmental Protection Agency and US Census Bureau


```
> gwrG
Call:
gwr(formula = Toxic_Re_1 ~ MS_Inequ_1 + MS_Inequ_2 + MS_Inequ_3,
    data = MS_GWR, bandwidth = bwG, gweight = gwr.Gauss)
Kernel function: gwr.Gauss
Fixed bandwidth: 5.618257
Summary of GWR coefficient estimates at data points:
                Min. 1st Qu.  Median 3rd Qu.    Max.    Global
X.Intercept. -566100 -481200 -443200 -414800 -339100 -323369.7
MS_Inequ_1    581100  599000  615000  625800  639200  604791.4
MS_Inequ_2   1782000 1979000 2062000 2158000 2368000 1797034.2
MS_Inequ_3      9118    9575    9781    9972   10160    9769.6
```

From the graphs above we can see that there appears to be clustering for all four of the variables – local R-Square and the local slope coefficients associated with the effects of residential segregation, income inequality, and the percent black on the levels of total toxic release emissions by county. Given the slope coefficients it appears that higher levels of residential segregation, income equality and percent black are associated with the levels of total toxic release emissions by county. Income inequality appears to have the highest effect.

4. **Interpolate the GWR results using an ordinary kriging technique and visualize the interpolated results with a "hollow" (no fill) county boundary file (MS polygon file from directive 1 above) overlaid on the interpolated grids. Present these GWR results in a four panel map visualizing the local r-square, local segregation effect, local inequality effect, and the local pct. black.**



Interpolated Emissions in LBS (per EPA),
Mississippi
Alexandra Plassaras

Data Source: Environmental Protection Agency

The above image shows the results of an interpolation analysis of total emissions in pounds throughout the state Mississippi. I attempted to conduct interpolation in order to visualize the local r-square, local segregation effect, local inequality effect, and the local pct. Black from my GWR results but I the only point files that I have are the MS_CTY_Centroids_SinglePart point file which only has the centroid points for each county in Mississippi and the MS_TRI_Emissions which has toxic emissions point data for the whole state. Thus the interpolation above shows higher levels of total emissions in the Jackson region as well as the Northeast region as the results in Question2 also showed. These results appear to show that the high levels of emissions are not associated with many site locations grouped together but instead only a few specific locations which are created the highest rates of toxic emissions.

5. **Interpret the results with a focus on any spatial trends associated with the fit of the model and trends associated with localized variations in slope coefficients obtained.**

As stated above the analysis conducted in questions 1 – 4 there appear to be non-random clusters of county level toxic emission release data. The clusters appear to be in the Northeast in counties such as Itawamba, Lee and Monroe as well as in the West in counties such as Madison, Hinds and Yazoo. After conducting geographically weighted regression analysis it is clear that there are significant levels of clustering and that the levels of total emissions are not randomly distributed. There is also evidence to suggest, as seen in our scatterplot of toxic emissions rates by spatially lagged total emissions rates that neighbors to areas of higher toxic emissions are more likely to have high levels of emissions as well.

PART II DATA. Earthquake Location Analysis

1. **Use the data on earthquake occurrences in the US to examine potential clustering in the locations of quakes.**

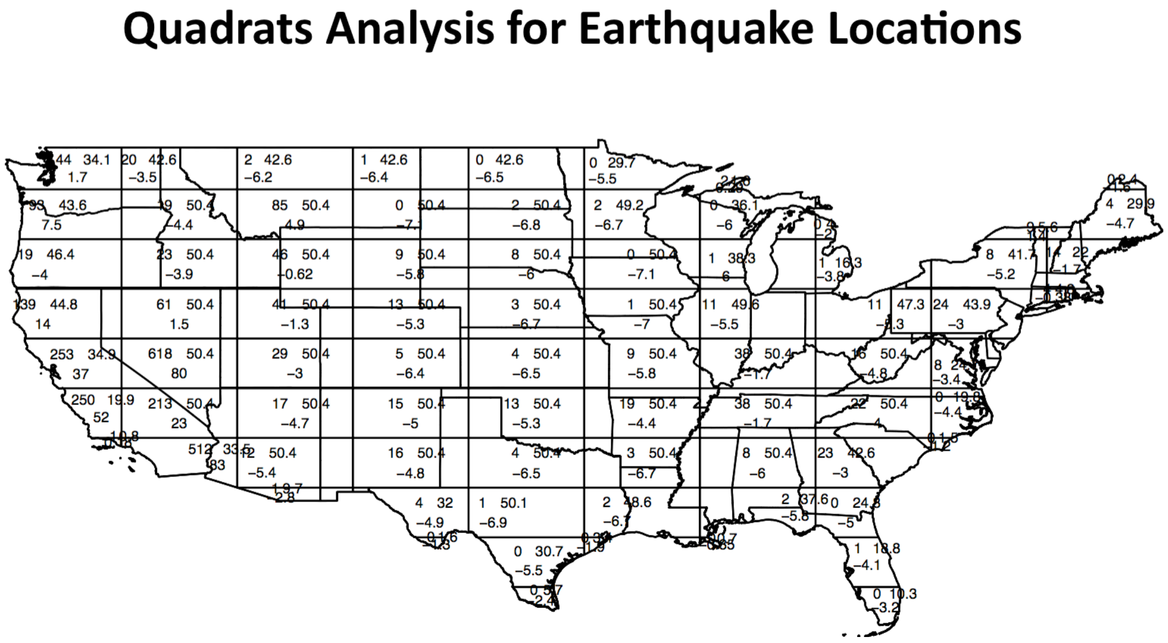
   Be sure to address the issue with three separate analyses and interpret the results with a discussion of differences and/or similarities to answer the question; Does there seem to be clustering in the locations of earthquake occurrences?

   a. **Quadrat analysis**

```
Chi-squared test of CSR using quadrat counts
Pearson X2 statistic

data:  TRI_border
X2 = 19624, df = 77, p-value < 2.2e-16
alternative hypothesis: two.sided

Quadrats: 78 tiles (irregular windows)
```

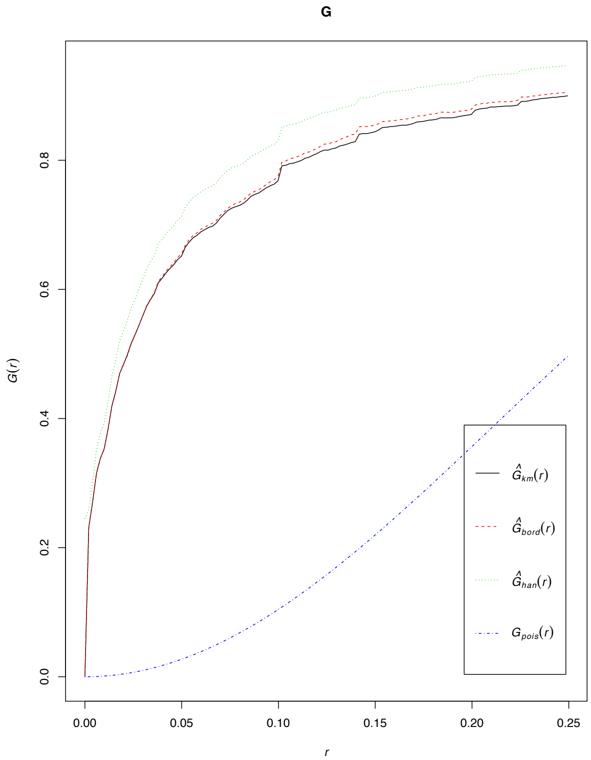


Quadrats Analysis for Earthquake Locations

From my analysis we can see that there is a tendency for events to exhibit a systematic pattern over an area as opposed to being randomly distributed. Using Quadrat analysis we look to measure the intensity in a given area. From our analysis above we can see that our earthquake locations are not randomly distributed and that our findings are statistically significant. For example, between the boarder of California and Nevada we see that λ is 80. Throughout most of California λ ranges from 83-14; all very large λ values. From this we can determine that there is a clustered pattern of earthquakes within the West Coast region (primarily California). This would make sense given than California is home to the San Andreas Fault, the Calaveras Fault, the Hayward Fault and many others. When looking at the rest of the US it appears that λ is quite small and often negative. From this we can deduce that there are random distributions of earthquakes in other regions of the US with negative λ values.

b. Nearest neighbor approach (G function)

```
Function value object (class ‘fv’)
for the function r -> G(r)
.....................................................................
        Math.label      Description
r       r               distance argument r
theo    G[pois](r)      theoretical Poisson G(r)
han     hat(G)[han](r)  Hanisch estimate of G(r)
rs      hat(G)[bord](r) border corrected estimate of G(r)
km      hat(G)[km](r)   Kaplan-Meier estimate of G(r)
hazard  hat(h)[km](r)   Kaplan-Meier estimate of hazard function h(r)
theohaz h[pois](r)      theoretical Poisson hazard function h(r)
.....................................................................
Default plot formula:  .~r
where “.” stands for ‘km’, ‘rs’, ‘han’, ‘theo’
Recommended range of argument r: [0, 0.24946]
Available range of argument r: [0, 1.0218]
```



Using the G-function to conduct nearest neighbor analysis we examine the cumulative frequency distribution of the nearest neighbor distances. Looking at the shape of our G-function we can see that our data is clustered because G increases rapidly at a short distance. We also see that the blue line is what one would expect from complete spatial randomness (CSR). The other lines are what we observe in our data. This plot indicates clustering, a greater proportion of earthquake occurrences have nearest neighbors at each distance than expected under CSR.

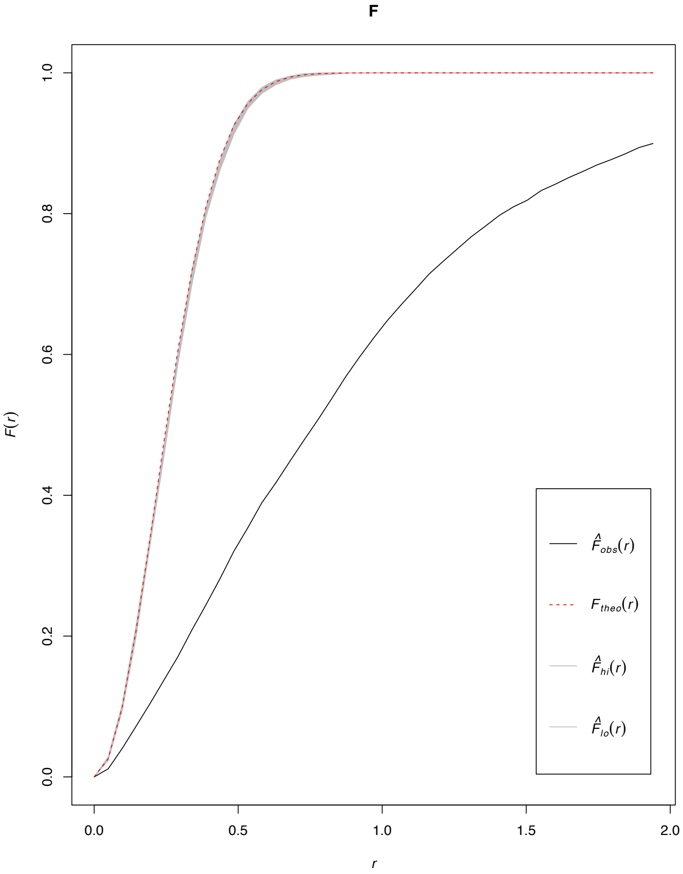
The weakness of the G-function is that it only looks at nearest neighbors. The K-function is a "second order" test in that it looks at the spatial dependence among points. The phrase "second order properties" is synonymous with "spatial dependence." However, in my analysis I was unable to complete the K-function approach analysis so instead I will be using the F function analysis to examine potential clustering in the locations of quakes.

c. Nearest neighbor approach (F function)

```
> r = seq(0,10, by=0.5)
> F <- envelope(TRI_border,Fest,r=r, nsim=10, rank=2)
Error: in Fest(X, r) the successive r values must be finely spaced: given spacing
= 0.5; required spacing <=  0.0485

Pointwise critical envelopes for F(r)
```

```
and observed value for ‘TRI_border’
Edge correction: “km”
Obtained from 10 simulations of CSR
Alternative: two.sided
Significance level of pointwise Monte Carlo test: 2/11 = 0.182
.....................................................................
     Math.label     Description
r    r              distance argument r
obs  hat(F)[obs](r) observed value of F(r) for data pattern
theo F[theo](r)     theoretical value of F(r) for CSR
lo   hat(F)[lo](r)  lower pointwise envelope of F(r) from simulations
hi   hat(F)[hi](r)  upper pointwise envelope of F(r) from simulations
.....................................................................
Default plot formula:  .~r
where “.” stands for ‘obs’, ‘theo’, ‘hi’, ‘lo’
Columns ‘lo’ and ‘hi’ will be plotted as shading (by default)
Recommended range of argument r: [0, 1.94]
Available range of argument r: [0, 9.991]
```

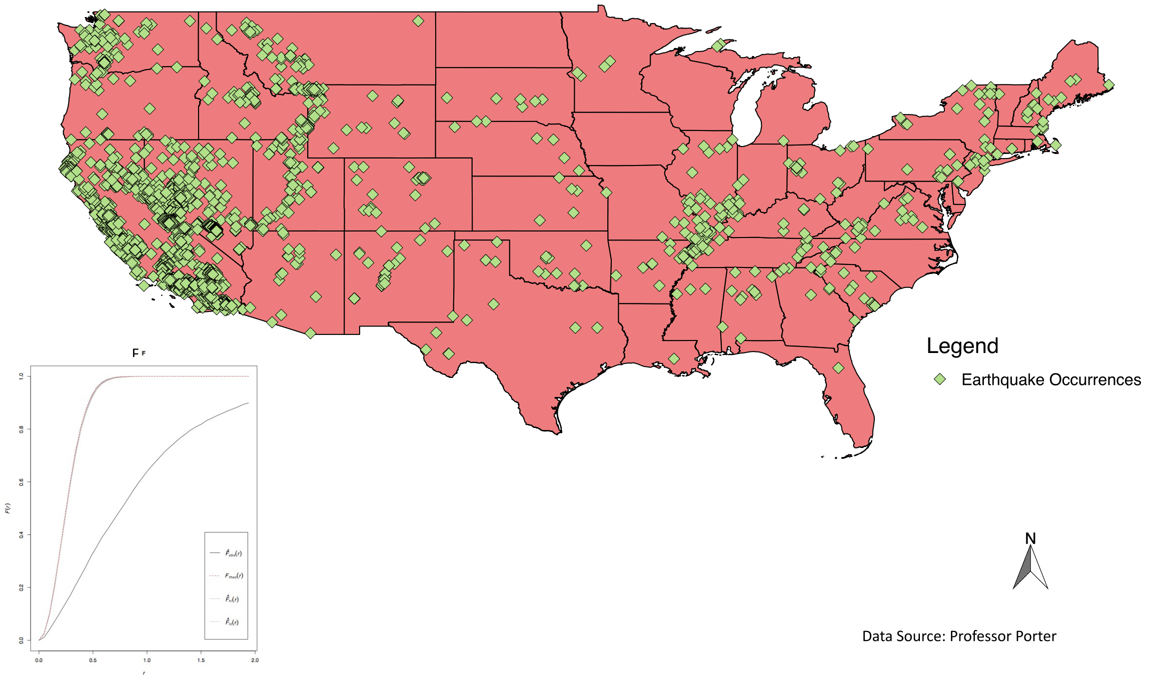


In the above F function nearest neighbor approach we are selecting a sample point location in the study region at random and from there determine the minimum distance from each point to any event in the study area. By looking at the graph above we can see that our observed data fall below the envelop which means that our data is clustered. Given that both the G and F functions only take into account nearest neighbors, it would be better to use the K function as this function uses more points and can provide estimates of spatial dependence over a wider range of scales and assumes isotropy over the region.

2. **Obtain a US State polygon map (lower 48 states only) and overlay the earthquake occurrence locations for visual purposes. Insert the K-function image from directive #1 above in to the map.**

Earthquake Occurances throughout Continental United States

By Alexandra Plassaras



Data Source: Professor Porter

From our analysis in Question 1 Part 2 we saw that our points were clustered using Quadrats analysis, and the G and F functions. In the image above we can see that after mapping our earthquake occurrences on a map there does appear to be clustering in the West Coast region. As I mentioned earlier this is probably very likely due to the fact that in the West Coast is home to many large fault lines.