FATALITY PREDICTION: TORNADOES USA

Cesar Rene Pabon Bernal
Professor Dana Sylvan
Stat 787
City University of New York, Hunter College
Monday, December 16th, 2019

TABLE OF CONTENTS

FATALITY PREDICTION: TORNADOES USA

Cesar Rene Pabon Bernal
Professor Dana Sylvan
Stat 787
City University of New York, Hunter College

ABSTRACT

Since 1950, tornadoes have been recorded and labeled as being amongst the most deadly and financially disrupting natural disasters in the U.S.A. Every year an average of 1200 tornadoes kill up to 60 people, injure 1,500 and cause at least $400 million dollars in economic damage.[1] In this report, we study a geostatistical analysis of tornadoes in the USA from 2007-2018 with EF ratings >2 and predict fatalities at those locations.

INTRODUCTION

Unlike rainfall or temperature, which may be measured by a fixed instrument, tornadoes are short-lived and very unpredictable.[2] A template for the formation of severe tornados can be seen in **figure 1**. Before thunderstorms develop, winds change direction and increase in speed with altitude. Doing so, creates an invisible horizontal spinning effect in the lower atmosphere. Rising air within the thunderstorm tilts the rotation air from horizontal to vertical and forms an arm of rotation 3-10 km wide. It is during the process of rotation that ultimately dictates the severity of a tornado.

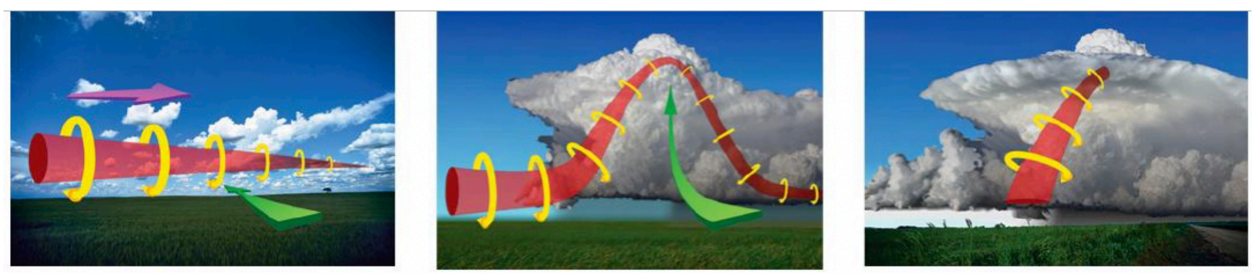

Figure1: Simple template for the formation of a tornado

Table 1 illustrates the Enhancd Fujita (EF) tornado scale system. It classifies the intensity of a tornado in the United States into 6 categories (EF0-EF5) based on the degree of damage to one or more of 28 damage indicators, such as various types of buildings, towers, and trees. Therefore, the higher the EF rating, the more destructive and dangerous a tornado can be. In this report, we focus on tornadoes with an EF rating > 2 and predict fatalities using kriging. Therefore, we state our hypothesis as:

H_O = *The more intense the tornado, the more fatalities are predicted at those locations*

I. DATA

Data was collected from the National Oceanic and Atmospheric Administration (NOAA) national weather service storm prediction center.[3] Python and R were utilized interchangeably throughout sections of the project, especially during the data cleaning and wrangling process. The model utilized for this report consisted of:

- 401 observations from 01/0/200-7-12/31/2018
- Variables; "Mag" : EF rating, "Fat": Fataltlies

[1] https://www.lloyds.com/~/media/Lloyds/Reports/Emerging-Risk-Reports/Tornadoes-final-report.pdf

[2] Wang, Kaiwen, et al. "Assessment of the Public Health Risks and Impact of a Tornado in Funing, China, 23 June 2016: A Retrospective Analysis." *International Journal of Environmental Research and Public Health*, vol. 14, no. 10, Oct. 2017, p. 1201., doi:10.3390/ijerph14101201.

[3] "Storm Prediction Center WCM Page." *Storm Prediction Center*, https://www.spc.noaa.gov/wcm/#data.

TABLE 1: The Enhanced Fujita (EF) Scale

EF Rating	Wind Speed	Damage
EFO	65-85mph, 105-137 km/h	Light damage: Peels surface off some roofs; some damage to gutters or siding; branches broken off trees; shallow-rooted trees pushed over.
EF1	86-110mph, 138-178 km/h	Moderate damage: Roofs severely stripped; mobile homes overturned or badly damaged; loss of exterior doors; windows and other glass broken.
EF2	111-135mph, 179-218 km/h	Considerable damage: Roofs torn off well-constructed houses; foundations of frame homes shifted; mobile homes completely destroyed; large trees snapped or uprooted; light object missiles generated; cars lifted off ground.
EF3	136-165mph, 219-266 km/h	Severe damage: Entire stories of well-constructed houses destroyed; severe damage to large buildings such as shopping malls; trains overturned; trees debarked; heavy cars lifted off the ground and thrown; structures with weak foundations blown away some distance.
EF4	166-200mph, 2 67-322 km/h	Devastating damage: Well-constructed houses and whole frame houses completely levelled; cars thrown and small missiles generated.
EF5	>200mph, >322 km/h	Incredible damage: Strong frame houses leveled off foundation and swept away; auto-mobile sized missiles fly through the air in excess of 100m; steel reinforced concrete structure badly damaged; high-rise buildings have significant structural deformation.

II PLOTS: TORNADO POINTS IN THE USA

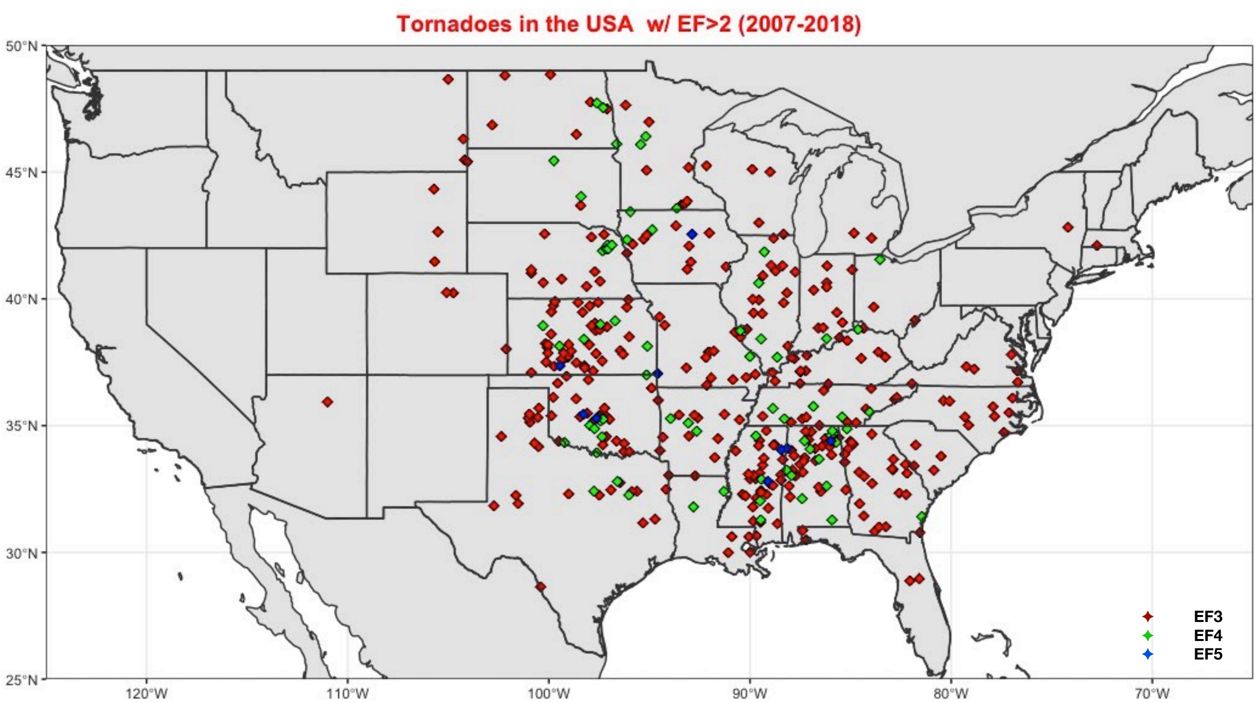

Figure2: Tornadoes in the USA with EF rating >2 (2007-2018)

Figure 2 illustrates geographic coordinates for tornado landings from 2007-2018. The red diamonds represent tornadoes with an EF rating of 3. The green diamonds represent tornadoes with an EF rating of 4. The blue diamonds represent tornadoes with an EF rating of 5. There appears to be a pattern in two sections of the United States. The first spans an area in the southern plains of the central United States and mostly occur during the late spring and occasionally the early fall; often referred to as “Tornado Alley”. The second spans the Gulf Coast area and mostly manifest in the late fall; often referred to as "Dixie Alley”.[4] Based on the above, a fatality cluster appears to be present in the “Dixie Alley” region.

III PLOTS: YEARLY ANALYSIS OF TORNADO FATALITIES IN THE USA & SPATIAL LOCATIONS

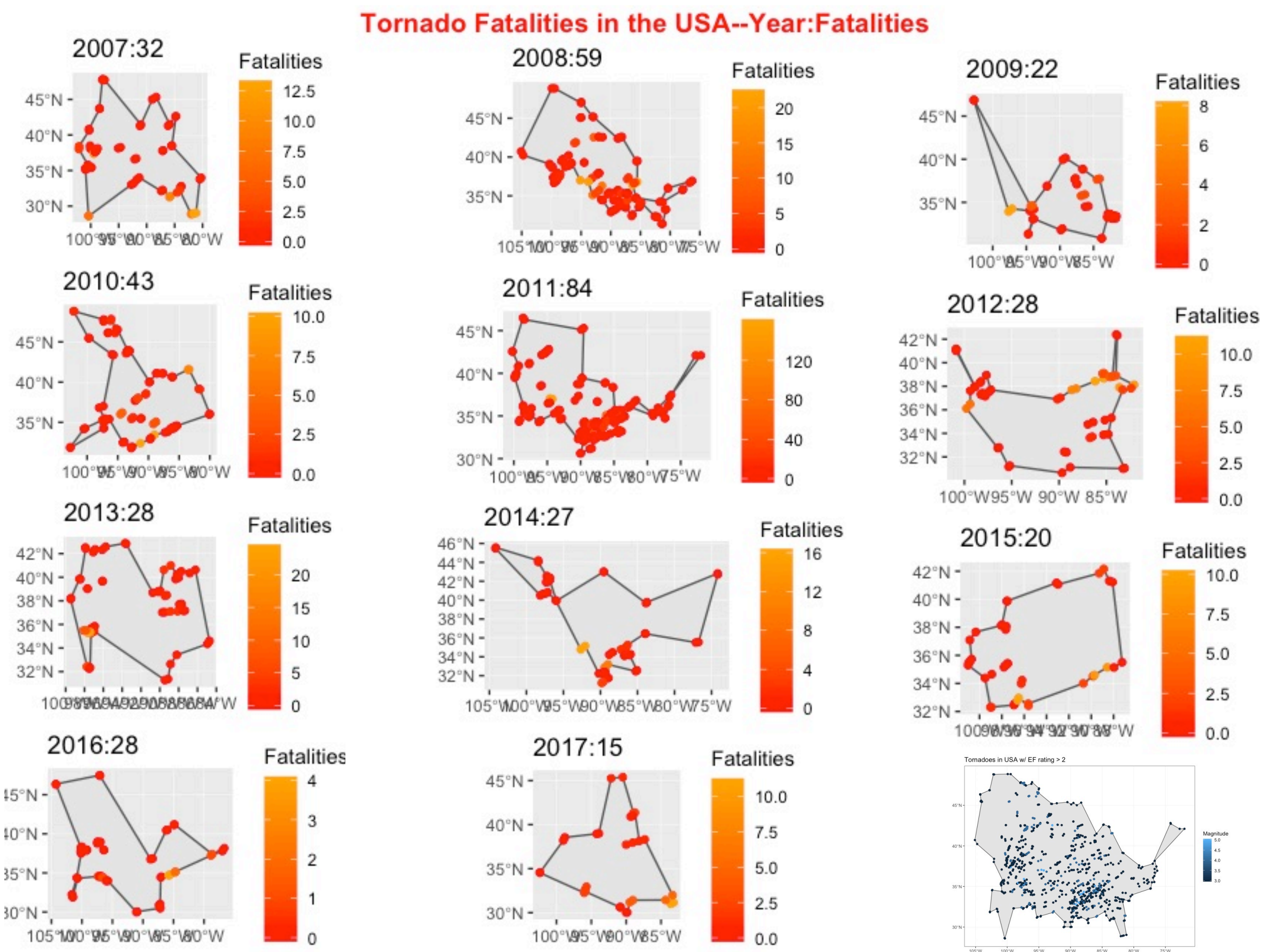

Figure3: Yearly Tornado fatalities in the USA with EF rating >2 (2007-2017)

[4] “Tornado Alley.” *National Climatic Data Center*, https://www.ncdc.noaa.gov/climate-information/extreme-events/us-tornado-climatology/tornado-alley.

Figure 3 illustrates the yearly analysis of tornado fatalities from 2007-2017 (2018 was omitted because only 1 fatality was present). Boundary polygons were created with respective enclosed geographic coordinates for tornado landings per year. A heat bar for each plot represents the number of fatalities. An overall analysis of the plots tell us that some years have randomness but none are completely normally distributed. For the purpose of this report, we will focus on the years that promote random spatial data and positive stationarity (mean & covariance do not vary significantly in space).

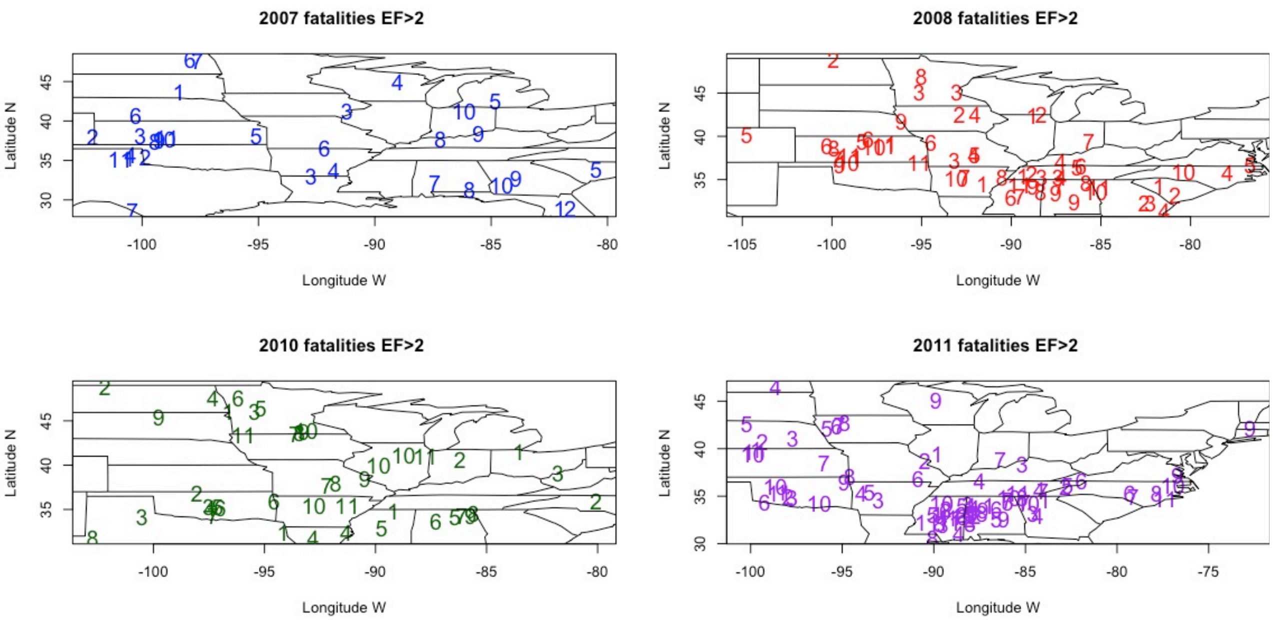

Figure 4: Simple numbered x-y plots for the years that were most attractive

Figure 4 is a simple numbered x-y plots for the years that were most attractive: 2007, 2008, 2011, 2012. Cross-validation selectively provided the years 2007-2008 as the most promising for prediction of fatalities.

METHODS

I. CORRELATION: CLOUD VARIOGRAM AND SAMPLE VARIOGRAM

One method to explore spatial correlation between observations is by plotting a variogram cloud. The cloud itself evaluates all of the possible differences between the observed pairs against their separation distance. We define the specifics of variogram cloud as:

- The differences between the observed fatality pair points : $(Z(s_i) - Z(s_j))^2$ **Eq. 1**
- The respective separation distance: $h_{ij} = ||s_i - s_j||$ **Eq. 2**

We carry the above and define a variogram as:

- $\gamma(h) = \frac{1}{2} E(Z(s) - Z(s+h))^2$ **Eq. 3**

Under the assumption that the variance of *Z* is constant in **equation 3,** we state that the spatial correlation of *Z* does not depend on location of *s* but rather, on *h* (separation distance). Consequently, multiple pairs $\{z(s_i), z(s_j)\}$ that have (nearly) identical separation vectors, $h = s_i - s_j$, can be formed. Ultimately, this permits an estimate of their correlation. In addition, if we assume that isotropy is present (spatial correlation structure is the same in all directions but specifically when the direction of semivariance is independent), then we can interchange the vector *h* with its length $||\ h\ ||$.

The variogram can then estimated from N_h sample data pairs $z(s_i),\ z(s_i + h)$ for a number of distances $\bar{h}_j$.[5] The model estimate is called the sample variogram and defines as:

- $$\hat{\gamma}(\bar{h}_j) = \frac{1}{2N_h}\sum_{i=1}^{N_h}(Z(s_i) - Z(s_i + h))^2, \quad \forall h \in \bar{h}_j \qquad \textbf{Eq. 4}$$

Figure 5 is a variogram cloud plot for fatalities in 2007 and state possible non-stationarity due to low variability. **Figure 6** is the sample variogram for fatalities in 2007 utilizing a tree distance algorithm for characterization of spatial correlation. Since we are evaluating dissimilarity (separation of distance of paired data values) between the distances, I removed the data points in which fatalities "zero" were counted for in order to improve model selection. In addition, a Log-Transformation was done to help reduce low variability and improve prediction for fatalities in 2008.

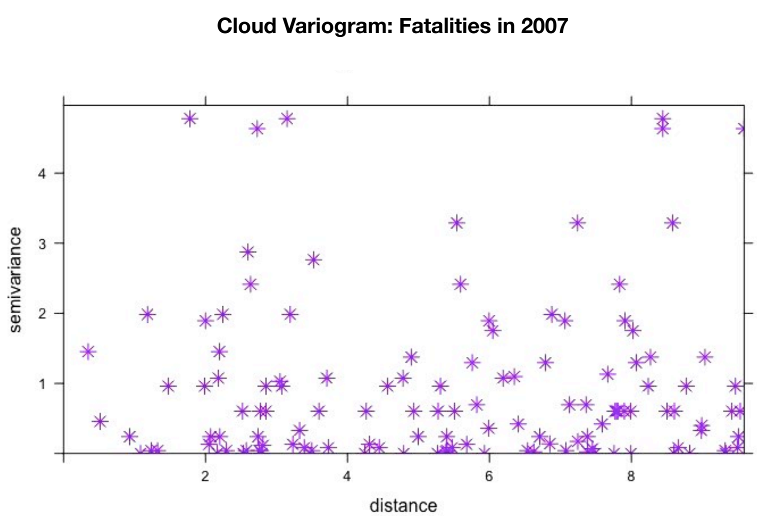

Figure 5: Variogram Cloud Plot for Fatalities in 2007

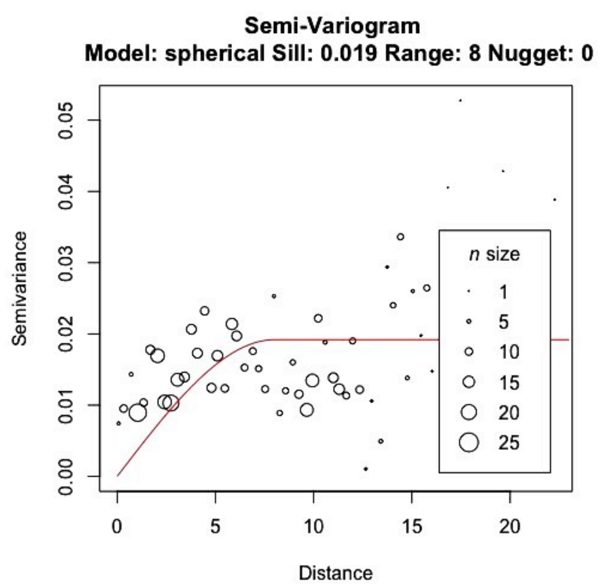

Figure 6: Sample Variogram Plot for Fatalities in 2007

[5] Bivand, Roger S., et al. Applied Spatial Data Analysis with R. Springer New York, 2013.

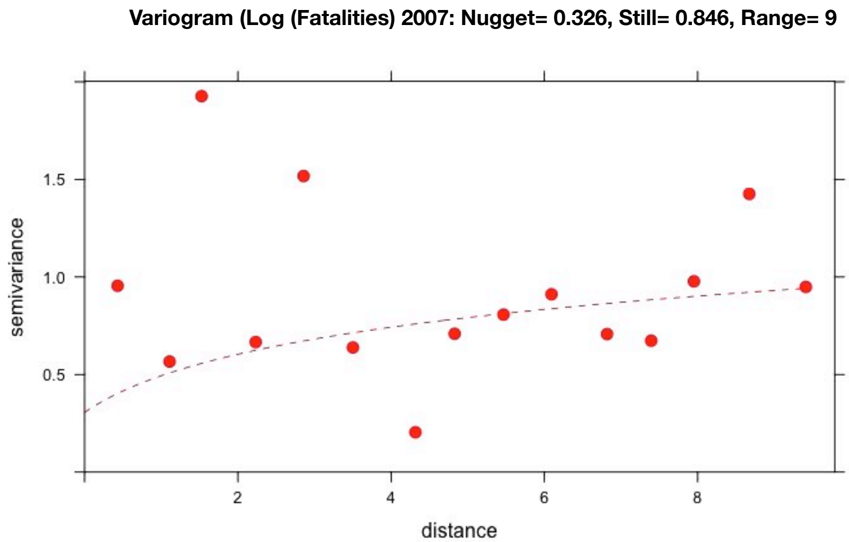

Figure 7: Sample Variogram Plot for Fatalities in 2007 w/ Log-Transformation

II. PREDICTION: LOCATIONS

Figure 8 are the locations of fatalities in the USA in 2008 for tornadoes with an EF rating greater than 2. The number within the circle represent those respective values. Our objective is to use the model found above to appropriately predict fatalities from 2007 to 2008.

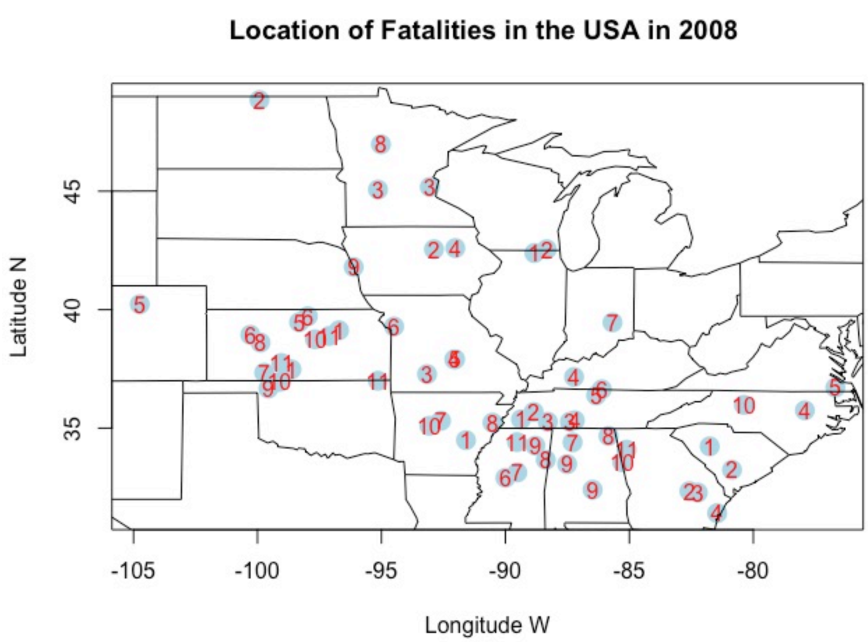

Figure 8: Location of fatalities in 2008

II. PREDICTION: KRIGING

Figure 9 illustrates the kriging prediction of tornado fatalities in 2008. **Plot 9.A** is a 2-dimensional prediction heat map of the geospatial points in **Figure 8.** Using the automatic model kriging function, we are able to see that location fatalities are consistent with where they are expected. There are two clusters of importance in the map. Both occur in the lower part of the Dixie Alley region. **Plot 9.B** illustrates the S.E. error heat map for these points. We conclude that the S.E. is low when the number of fatalities increase and explore this further in testing. The automated ideal variogram for our final model is seen in **Plot 9.A**

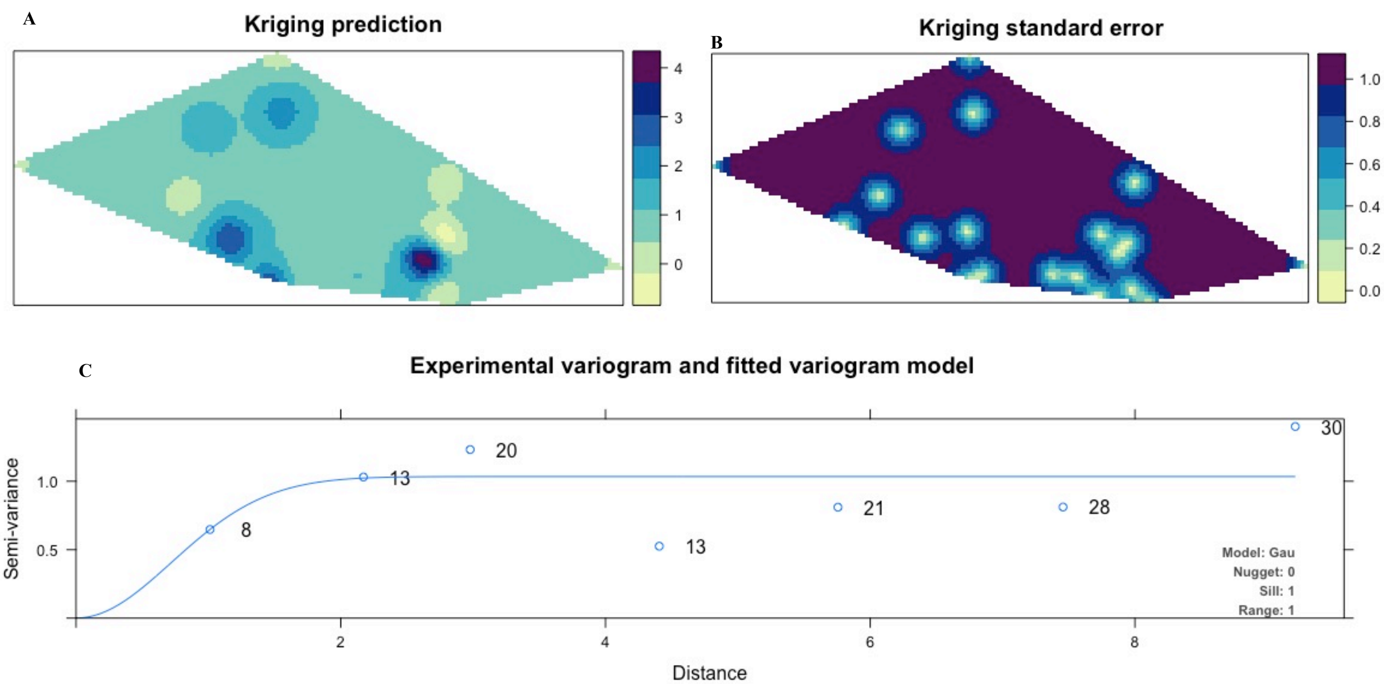

Figure 9: Kriging prediction of tornado fatalities in 2008

II. PREDICTION: TESTING

In order to completely test the error of the model, we compared the actual fatalities caused by tornadoes with an EF rating greater than 2 in 2008 (59) versus those found in the kriging model (39). We conclude an error of 33.898%.

CONCLUSION

Through careful model selection, the years 2007 and 2008 were the most appropriate for analyzing fatalities in the U.S.A. Boundary plots were carefully analyzed and respective variograms, upon data log-transformation, were explored. An automated kriging function provided the best sample variogram model with a Gaussian distribution, a nugget =0, sit =1, and range =1.

As mentioned in the introduction, tornadoes are extremely unpredictable. This ideology carried itself into our prediction with an error of 33.898%. Although we were able to accept our hypothesis, the more intense the tornado, the more fatalities are predicted at those locations, the difficulties surrounding this particular type of data, is apparent. For further studies, we suggest a some other type of prediction of geospatial analysis due to the above.

REFERENCES

1) https://www.lloyds.com/~/media/Lloyds/Reports/Emerging-Risk-Reports/Tornadoes-final-report.pdf
2) Wang, Kaiwen, et al. “Assessment of the Public Health Risks and Impact of a Tornado in Funing, China, 23 June 2016: A Retrospective Analysis.” *International Journal of E*
3) “Storm Prediction Center WCM Page.” *Storm Prediction Center*, https://www.spc.noaa.gov/wcm/#data.
4) “Tornado Alley.” *National Climatic Data Center*, https://www.ncdc.noaa.gov/climate-information/extreme-events/us-tornado-climatology/tornado-alley.
5) Bivand, Roger S., et al. Applied Spatial Data Analysis with R. Springer New York, 2013.

SPECIAL THANKS

Thank you for mentorship and guidance, Professor Sylvan. The last year has been incredibly rewarding and I feel that from your classes, I’ve rounded my completeness as a competitive statistician. Thank you!