SI 671/721 Project Final Report
Analyzing/Predicting Severe Weather From Historical Records

Version 1.0

Jonathan Hartman

Abstract

Severe weather is a significant and increasing concern in the US. In the past four decades, severe weather was responsible for more than $1.5 trillion dollars worth of damage, and in 2017 alone, resulted in more than $300 billion. (Smith, 2018). Modern weather models and forecasting are capable of predicting these events weeks in advance; however arguably more important than the prediction is making sure that resources are in place to mitigate the dangers posed by these events. This paper analyzes a dataset provided by the NOAA in order to determine if it is possible to predict when and where severe weather is most likely based on historical records. I attempt to create a method whereby this dataset can be leveraged to make predictions regarding the probability of similar events occurring in the future. Although specific predictions prove to be problematic, more broad statements about given geographic locations and times are possible to make with reasonable certainty. This sort of information could be used to evaluate the staging of emergency services for disaster relief, to ensure that local municipalities are best prepared for the most likely severe weather incidents.

1 Introduction

As it's inefficient to prepare everyone for everything all of the time, it's important to specialize and be prepared for the most common dangers that appear in a given place at a given time. Thus, firefighters concentrate along the west coast in the dry season and state transportation departments stock up on salt in the winter. But how specific and accurate a prediction can be made about the potential for severe weather given past data? In this paper, I look at a public collection of granular data released by the National Oceanic and Atmospheric Administration (NOAA) covering the past seventy years to see if there's a possibility to predict the probability of weather events based on geographic and temporal factors.

This is a question worth pursuing both from an immediate practical concern - which specific areas of the country are at risk, how much at risk, and when - but also given the potential for observing an increase in the frequency and intensity of these sorts of events as climate change starts to affect seasonal patterns.

In order to approach this task, I created a method to evaluate, based on historical data, where various types of severe weather had occurred in the past and identify which geographic locations were most at risk and at what time of year. This approach, detailed below, uses a combination of matrix-based and geo-spatial analysis in order to create two types of predictions - one based on the probability of a specific event occurring in a geographic area in a given year, and another which attempts to predict specific time-frames for a variety of event types in a given location.

2 Problem Definition and Data

The problem I decided to look at was incidences of severe weather in the US, and if it's possible to identify areas that are more at risk of severe storm event in a given time frame than others. The dataset I used is the NOAA's Severe Weather Data Index (SWDI), which contains observed tornado, hail, and severe wind storms from 1950 through the present. Aside from the clear "tornado alley" in the US midwest, I'm curious to see if I can identify times and locations that are more prone to severe and damaging weather based off of historical data.

In order to determine success, I restricted my analysis to a subset of years - 1950-2015, to determine if any patterns identified in that time frame

hold true in later years. Of course, an innate difficulty in dealing with weather are global effects - general climate change and varying intensity of other global weather phenomenon - e.g. strong El Nino years. Still, I hope to provide a method of predicting weather events with a certainty greater than 50% .

The SWDI data itself consists of a series of large csv files, each containing (for each event) general geographic details, along with Latitude/Longitude start and end points, date and time, state and county data, and general narrative descriptions.

Although data is provided back to 1950, as might be expected, the density of reported events and variability in types of events and the amount of detail provided for each is skewed towards recent dates. From 1950 through 1955 the only recorded events are Tornado, and from 1955 through 1996 only Tornado, Damaging Wind, and Hail events are included. Following NWS directive 10-1605 in 1996, however, 48 separate types of events are identified and included. Of those post-1996 events, about 50% are fairly evenly split between either Thunderstorms or Hail, followed by Flooding, Drought, Winter Storm, and High Wind to make up the next 25% combined.

The set contains 1,514,188 distinct events, with at minimum a date/time and location (lat/log and FIPs county ID) for both the start and end of each event, a unique identifier, and an indicator of the type of event. A majority of entries also include statistics related to property/crop damage and direct/indirect deaths which could be used to estimate the severity of an event in the absence of a provided concrete measure (e.g. tornadoes are all listed along with their assumed intensity on the Enhanced Fujita scale.) There is a complimentary data set from the same source that includes more granular location reporting for a large portion of the main set, which includes additional lat/long coordinates, range and azimuth readings from specific reporting stations which covers about 60% of the main set of events.

The data itself is fairly clean. There do appear to be about 2,000 data entry errors from the 1997 and 1998 sets, where the given latitudes and latitudes are occasionally orders of magnitude outside of realistic values. There is also a lack of regularity when indicating the event type - although the NWS only indicates 48 events, the set includes 74 unique types which needed to be re-coded (e.g. ”Thunderstorm wind/ trees” and ”Thunderstorm winds/Flooding”, for which there were only 5 combined entries were re-coded as ”Thunderstorm Wind”). I am a little concerned after plotting the provided coordinates on a map that there seems to be a bias towards metropolitan areas - presumably a result of where there are more observers/remote sensors to report.

3 Related Work

The Relationship between Severe Weather Warnings, Storm Reports, and Storm Cell Frequency in and around Several Large Metropolitan Areas - Jason Naylor and Aaron Sexton.

This paper attempts to locate the spatial distribution of severe weather as it relates to large metropolitan areas. Specifically, they looked at six large Midwestern cities and analyzed NWS issued storm warnings and observed severe weather events. Their purpose was to determine if the proximity of a large city to a storm event played a role in where severe weather occurred. This study is focused on a small area and a relatively small dataset from that which is available - only events within 20 miles of 6 major cities in the years between 2007 and 2017. (Naylor and Sexton, 2018)

I found this paper immediately useful when data cleaning as they provided an explanation for something I had been noticing in the set - on a few occasions there were several event reports for the same day at nearly the same times but with the exact same latitude/longitude down to the 4th decimal. This appears to be a result of some reporting offices using a ”default” location for nearby events. This paper decided to remove all such duplicates - I opted to leave in results from the same day as long as they described different events. In the end this only affected 11 records.

Their analysis is based on combining this dataset with several others, notably a collection of shapefiles collected from the Iowa Environmental Mesonet indicating the locations of severe storm warnings for 2007-2017. Calculating the centriods of these polygons, they identify the location relative to the city and plot the results on a heat map over the city using a .25x.25 mile grid to identify areas which are more prone to receiving storm warnings. I could probably make something similar with the matrices I create - I just need to look more into what libraries or tools are available to

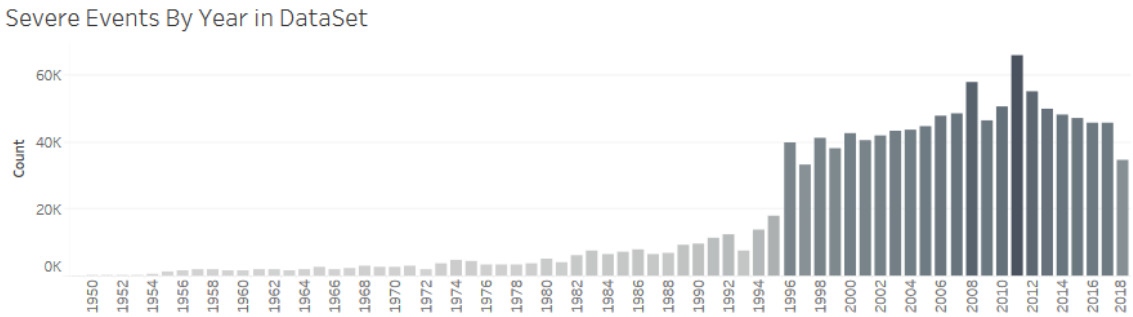

Figure 1: Events in the Dataset by Year

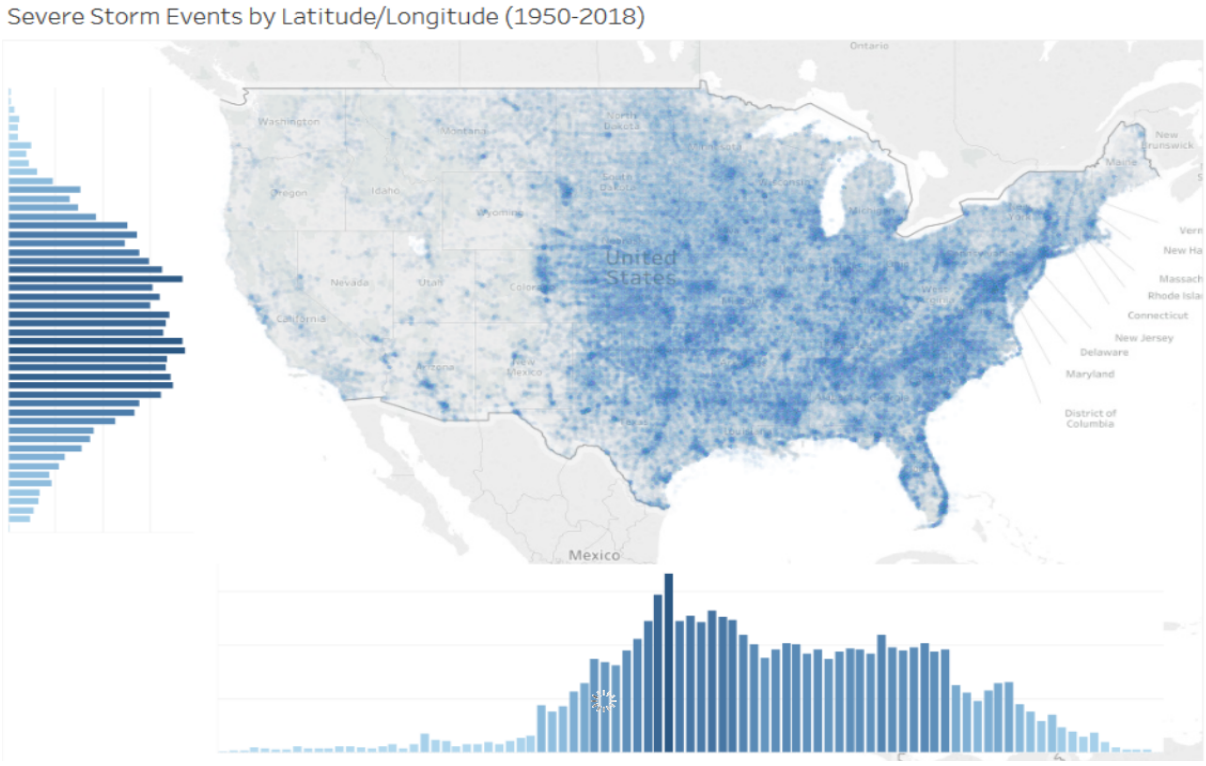

Figure 2: Event Start By Latitude/Longitude

assist with that.

Probabilistic Prediction of Climate Using Multi-Model Ensembles: From Basics to Applications - Palmer et al.

A little in the weeds for me, with a fair bit of climatological lingo that I keep having to look up, but it does contain a section discussing predicting standard (as in not severe) weather by applying the typical climate of an area to the observed weather on a given day to identify when variations occur. That might be a bit of a reach for me to try and pull in past the halfway point of this project, but I did find a data set from the NOAA which includes hourly average temperatures for weather stations around the country from 2000 to present. It might be interesting to try and use time-series analysis to see if there's a pattern in temperature immediately preceding a severe event in the area. (Palmer et al., 2018)

Some comments on the reliability of NOAA's Storm Events Database - Renato P. dos Santos

Not included in the initial paper, but I stumbled on this in the interim and found it useful as I went about getting started. It's not necessarily a scholarly paper and as such I'm not sure I would have included it my first submission anyways, but it was useful for pointing out both a number of outliers in the economic damage columns that likely

resulted from data entry errors (e.g. a flood in a small town in California that appeared to be the most costly natural disaster in world history.) The analysis is relegated to exploring things like event naming conventions, missing values, and data distribution, which is not particularly relevant to my project, but I thought it was appropriate to mention it. (dos Santos, 2016)

Social Media and Severe Weather: Do Tweets Provide a Valid Indicator of Public Attention to Severe Weather Risk Communication?

An adjacent study using this dataset in conjunction with a corpus of twitter posts that mention the word "tornado" and population data from the US census. The author compares geographic locations of severe weather in conjunction with geographic data from twitter to see if there is a correlation between twitter mentions and actual weather events. The study attempts to find a regression model based on spikes in the appearance of twitter mentions of the word "tornado" in order to more accurately predict the occurrence and path of a tornadic event. This stuck out to me first off as an interesting attempt to combine social media data with meteorological, and secondly as perhaps the sort of avenue I could go down as an alternative to my current plan. (Ripberger, 2014)

This paper did not end up being particularly relevant, as their analysis is mostly based on the geographic location and date/time of tornadic events, and using that as a basis to rate the accuracy of their twitter-based predictions. They do make use of a population grid published by Columbia University to estimate population density, which I considered trying to apply that to my own data to see if I could identify areas that are both at high risk of severe weather and high in population, but my matrix doesn't have a similar resolution and I wasn't able to make them match up in time for this report.

4 Methodology

The first step of the process involved a fair bit of data cleaning, performed mostly through pandas (McKinney, 2010–), to remove inconsistencies in the data, correct input errors, and fill out incomplete records. Event types were first standardized according to NWS directive 10-1605 (Murphy, 2018), to result in 55 distinct types of event. These were grouped into five "meta-categories" based on the general type of damaged caused by the event - Wind, Heat, Cold, Water and Other. As some events were logged as simply having occurred on the county level, they were not provided with geographic coordinates. For these events, the FIPS code was used to obtain the geographic centroid of the county in pace of a specific origin. Likewise, some events contained an end data, but no starting date. For these, the starting date was assumed to be the same as the end date[1].

For each type of event, I iterated through each week of the year and performed the following three step process. First, I created a scipy (Jones et al., 2001–) 5800x2760 linked list sparse matrix, each element of which corresponded to a latitude and longitude with two points of precision. I then filtered the dataset for all events which had occurred within two weeks either side of a given date.[2] Iterating through the results, each was assigned a value based upon how many days off from the given date the event occurred, discounted by $1/1 + log(days)$. This value was then added to the matrix dependent on the event type. For events with a path (Tornados, Dust Devils, etc.), I applied Bresenham's line algorithm (Bresenham, 1965) to find all points along the path denoted by the starting and ending latitude/longitude within the matrix. For events without a clear path, but which do not have a specific point of origin (Thunderstorms, Blizzards, etc), points were added to all matrix elements within five rows/columns of the origin, to represent an area roughly ten miles in diameter [3]. Finally, events which were report simply on the county level (Droughts, Extreme Heat, etc.) were assigned a single value based on the county centroid.

Next, the matrix was converted into a Geopandas dataframe of events, with the row and column of each non-zero cell corresponding to an event's latitude and longitude, and with the value of that cell as an extra variable. I then create a Geopandas dataframe from a shapefile of all counties in the continental US. I left-join my dataframe to that one, so that all events from my dataframe that oc-

[1] This may result in some inaccuracy with predictions, as some of these events are types such as "Drought" or "Winter Weather" which might not have as clear a defined starting point as other events.

[2] Any event between 1950 and 2015 inclusive is considered. Events in 2016 and later are withheld for evaluating predictions.

[3] These events typically have multiple reports from several different locations. As such, a five mile radius was considered large enough to account for a single report, whereas the event may have affected several states.

cur within the same county have their event value summed together. The resulting dataframe contains enough data for a plot of the continental US by county, along with a value representing the frequency of a specific event type based on a given date.

This array of values is then provided to the Pysal (Developers, 2014–) library. Pysal Weights are calculated from the counties shapefile using DistanceBand with a threshold of three, and a Gettis-Ord G* z-statistic calculated based on those weights and the values calculated for each county. The resulting array is stored in a csv file, which can be referenced later.[4] These values can be used to locate "hot spots" where a particular kind of event is more likely to occur than anywhere else on a given date. Examples of the plots created by this method are provided in Figure 3 and Figure 4. Clearly visible in these two plots are season peaks of the so called "Hail Alley" near Colorado, Nebraska, and Wyoming, and "Tornado Alley" near Okalhoma, Kansas, and Nebraska.

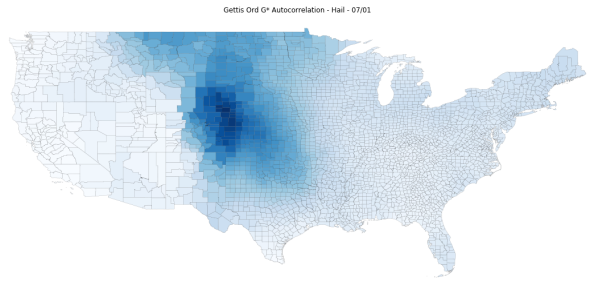

Figure 3: Hail Likelihood by County for July 1st

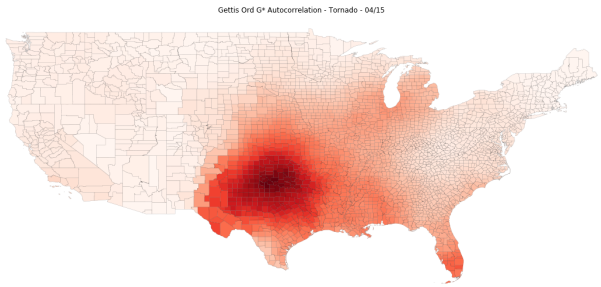

Figure 4: Tornado Likelihood by County for April 15th

In order to make predictions, these values are collected on a per county basis. Plotting them on a time series allows a visual representation of the likelihood of a particular event in a given county when compared to the continental US as a whole. (See examples in Figures 5 and 6). This approach is sort of like taking a series of these plots for an entire year, stacking them on top of one another, and drawing a line vertically though them to pull out the values for a given county throughout a year. By setting a cutoff point, we can estimate that those dates in which the value is above the cutoff indicate that the county is more likely to have a specific weather event than not, with high peaks indicating more strongly that an event will occur here compared to the remainder of the continental US.

A complete Jupyter Notebook containing all of the code used to create this analysis with this method is available at `https://github.com/jonhartm/SI671_Project`.

5 Evaluation and Results

5.1 Final Evaluation

Evaluation is based on two approaches, both based on withholding from the analysis of the dataset of all events which occurred after December 31, 2015.

The first involves randomly selecting dates and locations and determining if the prediction holds true. "Positive" labels are simply randomly selected events from the dataset from 2016 or later. "Negative" labels are randomly selected coordinates, dates, and event types, which have no events of the given type within two weeks of the selected date within two degrees of latitude/longitude. Results from a set of 1,000 predictions, equally split between positive and negative labels are indicated in the confusion matrix in Figure 5.1. A True prediction means that this method determined that there was a greater than 1 in 10 chance that in a year, a given severe event would occur.

Predictions based on Gettis Ord G scores		
	Predicted False	Predicted True
Actual False	955	45
Actual True	377	624

Unsurprisingly, negative labels are extremely easy to predict. For almost any type of severe weather, given a random location, it is far more likely for an event not to happen. False positives are very small in number, which is desirable - any prediction system which may have a bearing on public warnings would want to minimize the number of incorrect warnings provided as a result. Lowering this was a major concern in my interim

[4] Iterating through all fifty-two weeks for each separate event type takes about four hours.

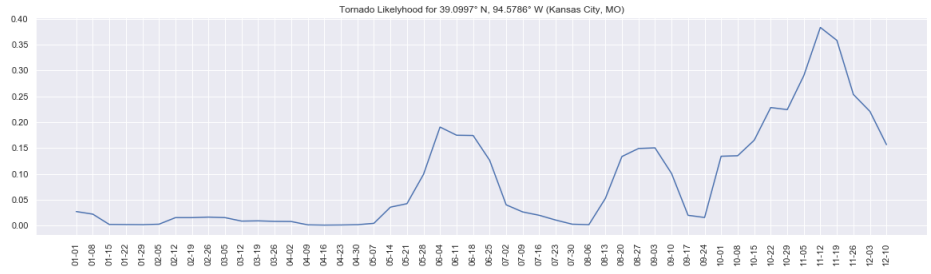

Figure 5: Tornado Likelyhood for Kansas City, MO

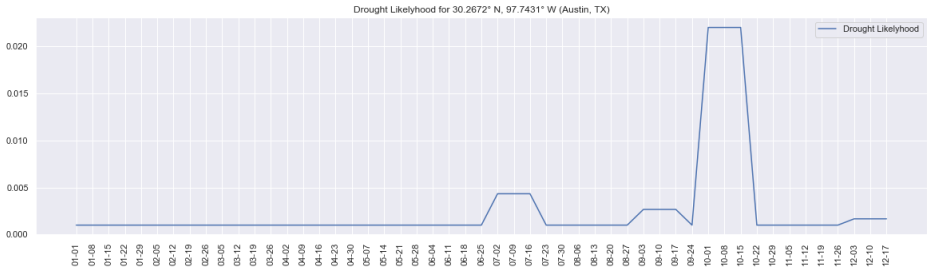

Figure 6: Drought Likelyhood for Austin, TX

report, so I'm glad it's been brought under control. The True-Positive results are a little disappointing. I assumed that making predictions based on specific event types would be a little less accurate, but I thought I would be landing somewhere in the low 70's.

The second approach involves taking a given location and predicting what severe weather events will occur and at what time of year. To do this, I picked a location at random, then iterated through each of the 15 most common event types, pulling the scores my method created and sorting them from high to low. If the scores were below a threshold, it was assumed that the event was not likely at this location. If the score was above the threshold, I took the two dates with the highest score, and pulled all events from any date after 2016 within two weeks of that date. If a severe weather event of the same type was recorded near that date and in that location, it was considered a True prediction. Likewise, if my method predicted that this event was not likely and no event was recorded, that was a true prediction as well. An example of this prediction for a location selected at random is provided in Figure 7 (Bolded Rows are missed predictions). This result is more or less typical of this approach: Several True-False predictions, some True-True predictions, and a few False Positives/False Negatives.

I consider this approach to be a slightly better method of determining accuracy. Although computationally it is about 10 times more difficult than the random True/False predictions, this I feel is more in line with the sort of thing I was trying to do in the first place, which is given a location, what are the likely events and when will they occur. I expected these predictions to be a little worse than the first approach, and so was initially pleasantly surprised when my accuracy seemed to hover around 66%. Like the first approach, there are some free guesses - in the example table, for instance, it's very unlikely that South Central Texas is going to have an Ice Storm in any year.[5] However, positive predictions are barely better than random chance. In part this may just be a result of the test set being restricted to a three year period, meaning that seasonal variations may be accountable for missed predictions.

[5] Although, as I write this, Lubbock, TX, 75 miles NW of Ringgold County, is receiving 8-10 inches of snow

Predictions for Ringgold County, TX			
Event Type	Prediction	Event Date	Truth
Blizzard	False	-/-	False
Extreme Cold	False	-/-	False
Hail	True	7/2/17	True
Heavy Snow	False	-/-	False
Ice Storm	False	-/-	False
Winter Storm	False	-/-	False
Drought	**True**	**5/7**	**False**
Excessive Heat	**True**	**9/21**	**False**
Wildfire	False	-/-	False
Flash Flood	**True**	**10/29**	**False**
Heavy Rain	False	-/-	False
High Wind	False	-/-	False
Hurricane	False	-/-	False
Thunderstorm Wind	True	4/23/17	True
Tornado	True	4/23/17	True

Event Predictions based on Location			
	Predicted False	Predicted True	
Actual False	978	342	*1,320*
Actual True	247	233	*480*
	1,225	*575*	

5.2 Naive Evaluation (interim report)

In practice, I actually started running into the opposite problem I was expecting. I had imagined that events would be so spaced out in time that I would have a hard time getting positive results. In reality, there's so much noise that I've had quite a bit of difficulty in lowering the amount of false positives I get.

The following tables are some of the results from this method - a few hand selected locations based on known weather and a series of locations chosen at random.

Locations With No Severe Weather			
City	Month	Predicted (Prob)	Actual
Los Angeles, CA	Jan	No (3%)	No
San Diego, CA	Jul	No (0%)	No
Simi Valley, CA	Nov	No (0%)	No

Locations With Known Severe Weather			
City	Month	Predicted (Prob)	Actual
Colo Springs, CO	Jul	Yes (99%)	Yes
Assumption, LA	Feb	No (0%)	Yes
Caddo, OK	May	No (47.3%)	Yes
Harris, TX	May	Yes (63.2%)	Yes

Locations/Months Selected At Random			
Lat/Long	Month	Predicted (Prob)	Actual
32.49N/-90.98W	Apr	Yes (99%)	Yes
31.41N/-99.16W	Jun	Yes (63.2%)	Yes
36.62N/-90.91W	May	No (42.1%)	Yes
38.83N/-84.33W	May	Yes (99%)	Yes
27.77N/-82.77W	Jan	No (42.1%)	Yes
47.74N/-122.3W	Apr	No (0%)	No
28.64N/-89.79W	Oct	No (0%)	No
42.15N/-102.8W	Mar	No (21.1%)	No
32.74N/-91.27W	May	Yes (99%)	No
35.86N/-101.9W	Dec	No (5.6%)	No

6 Discussion

6.1 Final Report

Overall I'm a little disappointed with the prediction section of this approach. I had hoped it would be a little easier to make predictions since so many of them are seasonally and geographically dependant. This is revealed in how simple it is to make negative predictions - for some event types we can simply rule out a significant part of the map and calendar year. I considered that this may be a result of the time slices I used (analyzing 1950-2016 and predicting on 2016-2017), and so I re-ran the analysis, this time using 1950-2010 and predicting on 2011-2017, but with similar results.[6]

I'm not sure it's fair to compare these results with my baseline, as the baseline was a simple binary approximation, more suited to determining if my matrix method was a valid approach for transforming the data. That said, I did get similar results in terms of raw accuracy with my second approach, but I attribute a fair amount of that to how simple it would be to just guess "False" for every event. In the table provided for my second approach, for instance, simply predicting False for each event would still result in 12/15 correct. That's not to say I consider this approach a complete failure - I was expecting the second method to be difficult. I think if I had a little more time I might be able to find a better balance between my analysis set and the test set which could improve the predictions. I may also be too restrictive in what I consider a "Correct" prediction. My best results so far come from considering +/- two weeks and a range of 1.5 degrees of latitude/longitude. More tuning with those parameters might result in an improvement as well.

That said, I do think the plots produced by this method are very interesting. Just given a basic familiarity with the sorts of weather in the continental US, seasonal trends like summer storms, hurricane season, and wildfires are immediately apparent. Creating weekly plots for a year and animating them produces some very pleasing visuals - particularly when looking at some of the more widespread events such as winter storms and tornadoes. I think if the goal of this project was to create interpretable and interesting visuals rather than making concrete predictions, this approach is perfectly fine with a little tweaking.

[6] Actually 0.4% worse, but I was unable to perform enough predictions to say it was statistically different.

One of my lingering concerns is my use of the counties as groupings. On the East Coast counties tend to be very small, whereas San Bernadino county alone is larger than several states. Since the data I'm looking at is irrespective of population, using these artificial designations may not always be appropriate, and may account for some of the significant variation that's present in some plots. Performing a PCA analysis on the entirety of the dataset, for instance, indicates that the most significant variation in the dataset is a result of San Bernadino and the surrounding counties, followed by the larger counties in Maine and parts of the Northwest. (Figure 7).

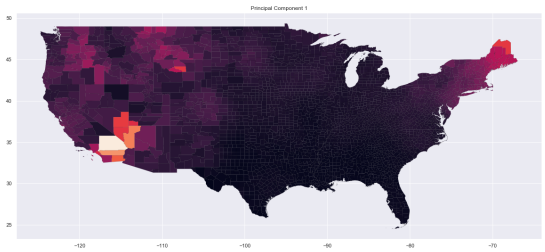

Figure 7: 1st Principle Component - Entire Dataset

6.2 Interim Report

Running 2,000 iterations of the random predictions, I'm about 77% on target - of the remainder, it's split pretty evenly between false positives and false negatives. Though that looks pretty good it isn't phenomenal and I feel like all I'm really getting is seasonal weather patterns, which isn't very interesting. I'm also disappointed with the false negatives I'm getting. It's particularly a problem if your goal is for the public benefit - of the four locations I picked for having known severe weather all had either a tornadic event or a severe hailstorm in the month I was looking at and only one was predicted as positive with much certainty.

I'm also concerned at false positives in occasions like the second to last in the random list, when my method predicts with near certainty that there will be an event and none occurs. There's of course going to be some unpredictability - it is the weather after all - and maybe the prediction would hold true if I looked at a few years following to see if my percentage predictions are at all close. Just out of curiosity, however, I ran a set of 1000 predictions - of those, 14% were very high probability and of those, 25% were incorrect, which is something I was a bit disappointed about.

7 Things left undone

There's so much other data related to this dataset that I didn't include, either because I wasn't sure I would have enough time to perform a proper analysis or because I was worried about having the focus of the project become too broad.

My initial proposal mentioned a secondary dataset related to this one, which had more precise location data related to a subset of these events - latitude/longitude of the reporting station, azimuth direction, and range. I started to use these as a more accurate method of populating the matrix for events with data in both datasets, but it quickly became apparent that I was spending far too much time figuring out how to calculate the approximate shape of an event, so I dropped this approach in favor of simply approximating a radius for these events. I think my estimation works well enough for a student project, but making my matrix more accurate would certainly be the first step I would take beyond where this report ends.

I also think a potential fix to the issue of varied county size throwing off the Gettis-Ord G analysis might be to ignore counties all together and run the same analysis on just a shapefile of a grid that covers the US. Since the analysis is irrespective of population areas, I think this would just be a case of creating the shapefile and substituting it in place of the US counties shapefile. I just ran out of time before I could figure out exactly how to do that.

There is also a dataset from the same source which is used in one of the papers I referenced earlier (Naylor and Sexton, 2018) on a smaller scale, which contains shapefiles representing polygons of every storm warning published by the NWS for the past 10 years. I think an interesting approach to this project would be to look at the overlay of that dataset with the results from this one and see if there were areas which I identified as being of high risk which were not receiving the same level of warning as other areas.

Another related set I considered was focused entirely on damaging hail storms, and which contained incredibly granular detail about every hail-producing event for the past thirty years based off of NEXRAD radar reporting, rather than human reports. This set was significantly larger than the one I used here, focused on a single type of event, and contained a high level of meteorological detail. If I had been working in a team with someone with a meteorological background, that might

have been an interesting set to include. As it was, I didn't feel knowledgeable enough in the terminology used in the set to attempt an analysis.

Climate change would also be an interesting topic to explore as a related topic to this analysis. I'd be interested to see if there was something that could be observed by creating a window to analyze, say 1990-1995, and seeing how the results of that compared with 1991-1996 and so on.

8 Work Plan

8.1 Final Report

The first half of my methodology didn't differ entirely from what I had proposed earlier. The idea of populating a matrix based on latitude and longitude was present in every iteration of what I did. I did end up reducing the size of the matrix by 1/10 - in part to make the calculations faster and in part because two points of precision was as specific as most of the data was. In the end, there was virtually no difference visually between the two.

The major change is with regard to how specific my predictions are. My original intention was to predict based off of a given latitude and longitude, however as my final approach leverages the Gettis-Ord G method, predictions based off of the county level were more natural. I still used the latitude and longitude when selecting points for evaluation, but these are immediately converted into a county. I also originally intended to run my analysis day by day. The computational time for this was clearly going to be far too long, so I opted to instead look at week by week analysis and compute a rolling average to fill in the gaps. This had the result of dampening some of the data, but I determined that to be a reasonable tradeoff.

My first evaluation was based off of a simple binary prediction - did I think a location had severe weather on a given date or not. This is obviously a simple prediction to make, and my naive approach (finding the average values of all matrix elements within a given radius) was pretty much just a seasonality detector. Southern California was almost always negative, whereas Eastern Oklahoma in tornado season was almost always positive. I ended up taking two approaches to this. First, I split the data into "Meta-Types" based on the kind of damage typical of that kind of weather. This mostly worked, but I wasn't particularly happy with the way I had to lump some events together (e.g. Tornadoes and Hurricanes, Hail storms and Blizzards). As I created more and more subgroups, I decided I was almost better off just doing them individually. I left Meta Types in the code just in case, but more or less abandoned them as a predictor in the end.

It's also a bit of a trope in the class at this point, but I did think for a while there might be a way to shoehorn an SVD decomposition into this. As soon as we covered Geospatial mining in class, I decided this was not the way to go.

8.2 Interim Report

The first step will be data cleaning - a constant comment in papers that reference this data set is that the data is in places incomplete or poorly formatted, so getting it to the point that I can reliably load in data will probably be the first hurdle.

Secondly will be coding to load and precompute all of the matrices I have in mind, as well as calculating secondary geographic data from path and range data. At this stage I'll also need to have a method of quickly querying the data and making predictions.

My current analysis feels pretty basic, so I'd really like to look at applying more of the spatial analysis techniques we've discussed so far, as well as looking to see if there's anything time-series analysis can identify when looking at data grouped at the state level.

To that end I'm planning on doing three things moving forward. First, I'd like to take one more shot at my current method and group the weather events into more broad categories (snow, damaging wind, fire, flooding, etc.) and see if I can make more specific predictions - i.e. this place has an X% chance of a snow event, X% chance of a flooding event, etc. Secondly, I'd like to explore some of what we talked about in the most recent class regarding spatial analysis, and specifically Getis-Ord Gi statistics, given that I do have the ability to group these events by county. Thirdly, if I have time, I'd like to see if I can do anything with time-series analysis while looking at data for a specific state or grid square. If all I'm getting at the moment is indeed seasonal patterns, maybe there's something interesting I could find looking more generally about an area's climate.

References

J. E. Bresenham. 1965. Algorithm for computer control of a digital plotter. https://www.cse.

iitb.ac.in/~paragc/teaching/2011/cs475/papers/bresenham_line.pdf.

PySAL Developers. 2014–. Pysal python spatial analysis library. https://pysal.readthedocs.io/en/latest/.

Renato P. dos Santos. 2016. Some comments on the reliability of noaa's storm events database. *CoRR* abs/1606.06973. http://arxiv.org/abs/1606.06973.

Eric Jones, Travis Oliphant, Pearu Peterson, et al. 2001–. SciPy: Open source scientific tools for Python. http://www.scipy.org/.

Wes McKinney. 2010–. pandas: A python data analysis library. http://pandas.sourceforge.net.

John D. Murphy. 2018. National weather service instruction 10-1605. http://www.nws.noaa.gov/directives/sym/pd01016005curr.pdf.

Jason Naylor and Aaron Sexton. 2018. The relationship between severe weather warnings, storm reports, and storm cell frequency in and around several large metropolitan areas. *The American Meteorilogical Society* https://doi.org/10.1175/WAF-D-18-0019.1.

T. N. Palmer, F. J. Doblas-Reyes, R. Hagedorn, and A. Weisheimer. 2018. Probabilistic prediction of climate using multi-model ensembles: From basics to applications. *Philosophical Transactions: Biological Sciences* https://www.jstor.org/stable/30041389.

Joseph T. Ripberger. 2014. An analysis of severe weather data: 2000 - 2015. *The American Meteorilogical Society* https://doi.org/10.1175/WCAS-D-13-00028.1.

Adam B Smith. 2018. 2017 u.s. billion-dollar weather and climate disasters: a historic year in context. https://www.climate.gov/news-features/blogs/beyond-data/2017-us-billion-dollar-weather-and-climate-disasters-historic-year.

A Supplemental Material

NOAA Severe Weather Data Inventory https://www1.ncdc.noaa.gov/pub/data/swdi/database-csv/v2/

NOAA Storm Events Database https://www.ncdc.noaa.gov/stormevents/details.jsp