Predicting Atlantic Hurricanes

Evan Solomonides, Omar Abdul-Rahim, Erinn Liu

ORIE 4740: Statistical Data Mining

Abstract

In recent years, hurricanes have become an increasing problem on the Atlantic Coast. Famous hurricanes, such as Katrina and Maria, still bring terrible memories to those directly impacted or whose loved ones were directly impacted. These hurricanes often cause billions of dollars in damage and kill or displace thousands of people. In order to mitigate this threat, we have developed models to aid in the prediction of Atlantic hurricanes. We found that hurricanes could be predicted with an accuracy of 61% primarily using month and geospatial information as predictors.

I. Introduction

Hurricanes can be life-altering disasters for millions of people. Those affected lose homes, businesses, and livelihoods. Hurricane Katrina, one of the most infamous hurricanes in human history and the costliest in US history, caused over $81 billion in damage and cost the city of New Orleans and national relief organizations over $160 billion for recovery efforts. 800,000 homes were destroyed, leaving people homeless and stranded. Hospitals were shut down, airports were overrun, and the people affected had little recourse.

However, there is a more personal factor that is not captured by the high-level statistics previously mentioned. Many people who were forced to sleep in the Louisiana Superdome can no longer comfortably attend a New Orleans Saints football game without reliving the past trauma of close quarters, unhealthy conditions, and no food. Parents fortunate enough to get a plane ticket were forced to send their children alone to distant relatives for safety. Many who had lived their entire lives in New Orleans lost their entire past to the storm and had to rebuild a new identity elsewhere. Needless to say, hurricanes hurt people deeply, oftentimes for the rest of their lives.

We decided that this cause would be a highly justified use of our efforts if it could in some way help prevent the effects previously outlined. We set out to build models that help predict hurricanes in order to stop their effects on human life by allowing people to take proactive measures to avoid impending hurricanes. We hope that this effort may one day save someone's livelihood.

II. Data

Our dataset was pulled from Kaggle. The title of the overall dataset is "Hurricanes and Typhoons, 1851-2014." From this overall dataset, we chose to use the data from the Atlantic

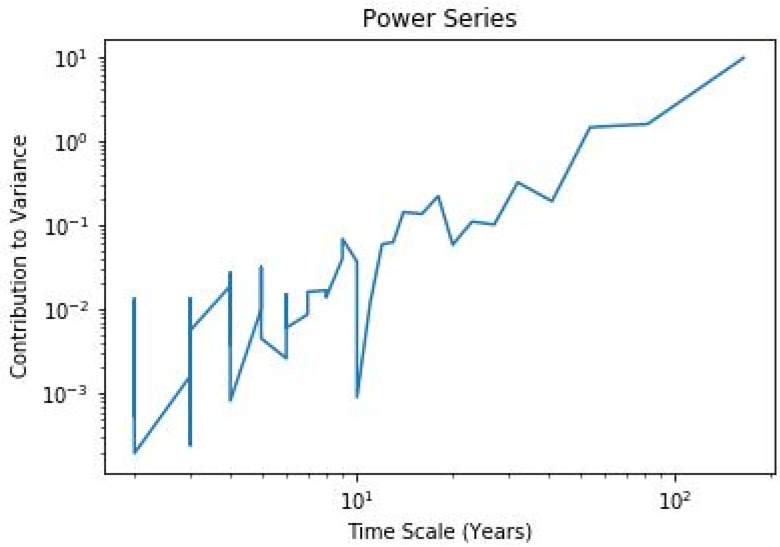
Ocean because it was the most pertinent to our analysis of hurricane impact on the United States. This dataset contains 22 features on 49,105 examples. Explanations of the features can be found in the bibliography. Many of the examples on the set are multiple logs of the same hurricane (e.g. Hurricane Katrina is logged 70 times at various points during its lifespan). Throughout this project, we examined subsets of our Atlantic storms dataset as appropriate; for example, the first log each unique storm was examined to analyze storm origin points.

III. Analysis

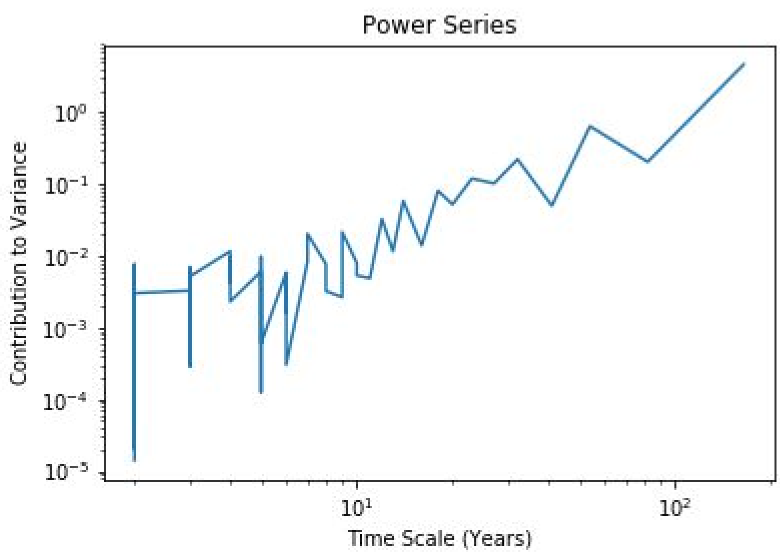
Fourier Series

Methods: A Fourier series was fit to a histogram of the years in which storms occurred, and then later to the first instance of these storms becoming hurricanes. This series was then transformed into a power series and the x-axis converted to the corresponding characteristic wavelength (in time), to show the extent to which any patterns such as “a storm ever X years” held true. See the attached code for the details of how this was done in Python.

Results:



Power series for the year in which storms occurred



Power series for the year in which hurricanes occurred

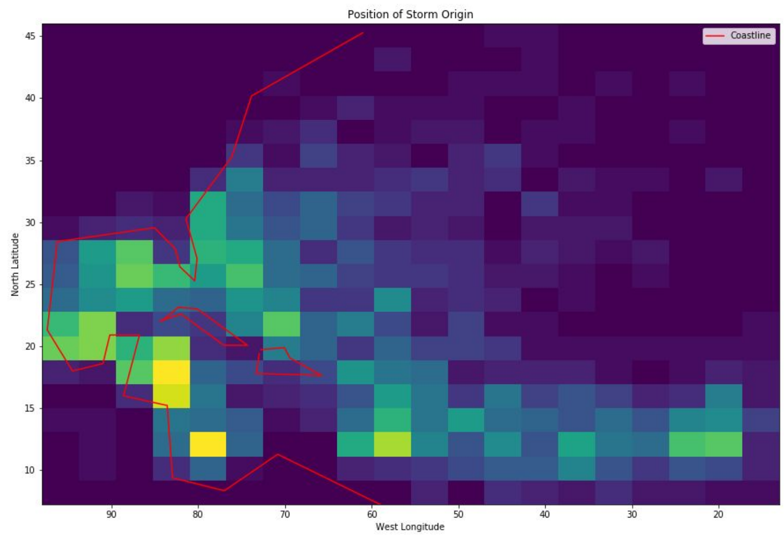
Interpretation: Looking at out Fourier fits on the histograms for both storms and hurricanes, there are no discernable, unexpected local trends. This does not bode particularly well for our ultimate goal of predicting hurricanes. However, the models do not take advantage of the many other predictors available to us, so there is still opportunity. In addition, this lack of local trend is corroborated by the power series of the models. As the time scale increases, so does the contribution to variance. This suggests that the common belief that there is a devastating hurricane every *n* years is incorrect.

One interesting thing to note is that while there are no observable local trends on the model, the global trend is slowly upward. This means that, on average, more storms are occuring per year as time goes on. This observation seems to be in line with the existing prevalent literature on climate change. While more research is required to fit our observation into the broader landscape of climate research, it is reassuring that our model suggests a consistent trend.

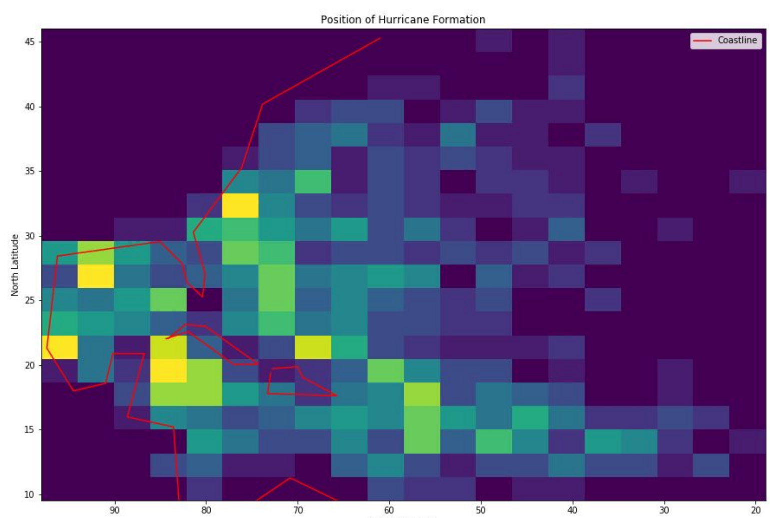
2D Histograms

Methods: The first occurrences, first classifications as hurricanes, and presence of hurricanes were all plotted as 2D histograms, using latitude and longitude as positions on the plot and equal weighting. The coastline of North and Central America was added to give perspective on what the data actually correspond to. The attached code will again demonstrate how this was performed in Python.

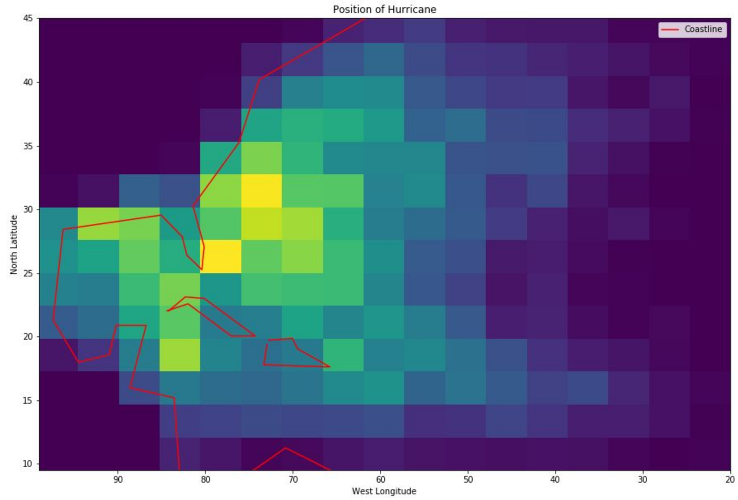
Results:



2D histogram of the first recorded observation of a storm



2D histogram of the first classification of a storm as a hurricane



2D histogram of all recorded observations of hurricanes

Interpretation: The histogram titled “Position of Storm Origin” shows the first log of every unique storm in the dataset, hence displaying the origin points of all the storms and conveying the frequency of storm origination in each area. The most common storm formation points are near [12N, 78W] (near Costa Rica, Panama, and Colombia) and [17N, 83W] (near Honduras, Guatemala, and Belize). These findings are consistent with climate research on storm formation; because of the large bays around those two areas, it is easy for storms to form.

The histogram titled “Position of Hurricane Origin” shows the first log of every unique storm that holds “Hurricane” status, hence displaying the frequency of hurricane formation in each area. The most common points are near [22N, 97W], [27N, 93W], [18N, 84W], and [33N, 77W]. Again, these points are coastal, and these findings are consistent with what we know about storm origin and hurricane formation.

The histogram titled “Position of Hurricane” shows every log of a storm with “Hurricane” status (meaning that a given hurricane could be catalogued multiple times). This gives us the frequency of where hurricanes exist. The most common position for hurricanes to exist is near

[26N, 78W], right off the southeast border of Florida. While this is bad news for Florida, this is good news for us because it is consistent with what we already know about hurricanes.

Constructed CSV

Methods: To make the data more manageable, it was processed from its original form into a CSV with one row per storm, and the first recorded location of that storm. Also stored are a boolean operator for whether or not the storm eventually became a hurricane, and two additional position columns which would allow regression to more accurately interpret the position of a storm. This data was saved, and used in the next section. The code for this process is also in the attached file.

Linear Regression

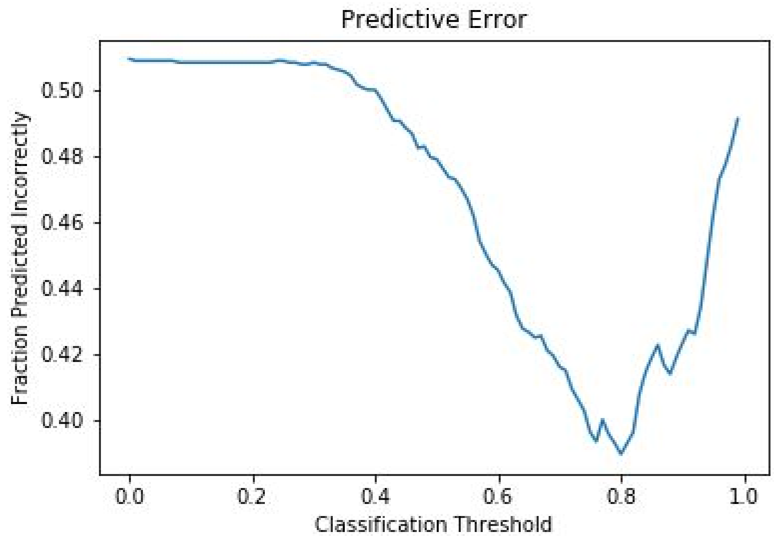
Methods: The CSV generated above was then run through the lm function in R, attempting to predict the boolean column representing whether or not a storm became a hurricane at any point. The exact way this was done can be seen in the attached, very short R file. This function was then analyzed in the following section.

Interpretation: Generally speaking, our model performs quite well. Being able to anticipate the natural world in all its seeming randomness with 61% accuracy is not insignificant. We believe that many meteorological groups and disaster relief organizations would benefit from the use of our model. In fact, given information from the early logs of Hurricane Katrina, our model would have predicted the hurricane, giving those in New Orleans more time to evacuate and FEMA a better chance to get organized to provide relief.

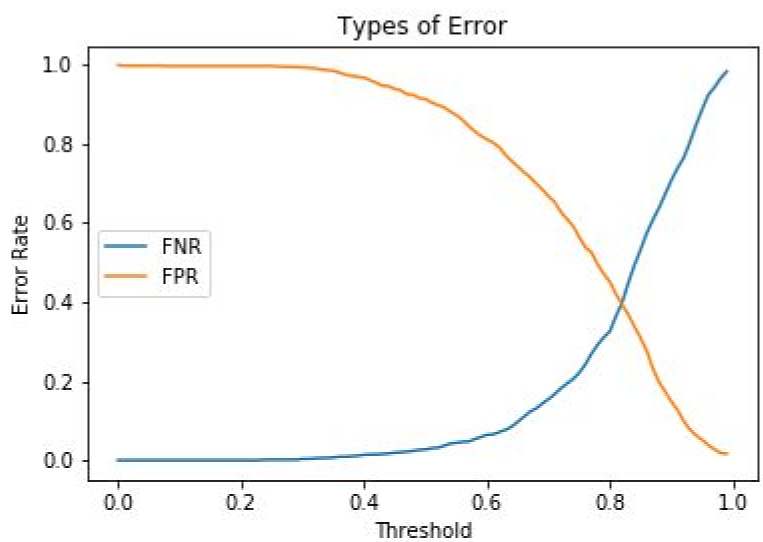
Error Analysis

Methods: The linear function determined above was then manually entered into Python, and compared to the actual data to determine accuracy. Different thresholds for positive classification were investigated, and the corresponding errors were plotted to give a sense of which threshold would give us the desired results.

Results:



The absolute error of our linear predictor function with different thresholds



The false positive and false negative rates for different thresholds

Interpretation: In examining the error of our linear prediction model, we see that the minimum error rate (approximately 0.39) is achieved at a hurricane classification threshold of around 0.8. From a scientific standpoint, this means 0.8 is the optimal threshold to pick for our model. However, examining the error more closely leads to different conclusions for different audiences.

When observing the trends in the rates of false negatives and false positives as the classification threshold increases, it is clear that the minimum overall error occurs at 0.8. However, predicting hurricanes is not a trivial matter. A false negative prediction could mean immense loss for many people, and the false negative rate makes up for almost half of the total error at a threshold of 0.8. So, for the purposes of users like the head a national disaster relief organization, a better classification threshold is 0.6. Though the overall error rate is higher than at 0.8, it is still low enough that the model is useful, and the false negative rate is significantly lower, meaning that the user is more likely to catch all actual hurricanes. We are still careful to keep the rate of false positives low, however, as false positives can needlessly incur large costs. Draining the coffers of disaster relief organizations by having them respond to phantom hurricanes will quickly erode public confidence. Especially given the “it cannot happen to me” mentality that is so pervasive in human psychology, even one false positive can result in people not heeding future accurate warnings. On a smaller day-to-day scale, this is seen when people ignore snow storm warnings after being disappointed by a prior snow storm warning that did not materialize.

IV. Relations to Existing Literature

The ability to predict hurricanes has a clear benefit to society: by being better prepared, communities can take preventive action to reduce the costs of hurricanes. According to the

National Oceanic and Atmospheric Administration, since 1980, the United States has had 40 hurricanes that have each caused at least a billion dollars in damages resulting in cumulative damages of approximately $862 billion (using 2018 inflation-adjusted dollars). To put this in perspective, the interest on national debt is $363 billion; reducing dollars lost to hurricane damage could help the United States wipe out its interest payments on its giant national debt! Furthermore, the reconstruction cost for at-risk homes, including labor and materials, is $1.6 trillion. This is same amount of money that the government collects in income taxes from the entire population.

Interestingly, current events only make hurricane prediction more important. President Donald J. Trump has captured daily headlines as his nationalistic views have influenced his economic decisions. Specifically, his skepticism of trade agreements and proclivity towards escalating trade wars has increased the prices of imports into the United States. For example, American lumber businesses have had a long history of competing with Canadian lumber businesses, alleging that the Canadian government subsidizes the wood industry therefore making American lumber businesses less competitive. However, the net result to the American people is that they are able to purchase lumber at a lower price. The increase in tariffs has caused the cost of wood and therefore the cost of building homes to increase, complicating rebuilding efforts after hurricanes. This has a domino effect; when people are unable to find steady housing and must move from shelter to shelter their ability to find a steady job is hindered.

The current state of hurricane prediction still leaves a lot to be desired. Its lack of precision results in higher costs: areas that need to be evacuated are not always evacuated in time, and areas that do not need to be evacuated are sometimes needlessly evacuated. The National Hurricane Center issues forecasts at various time points, 120 hours, 96 hours, 72 hours, 48 hours, 24 hours, and 12 hours. Naturally the the forecast becomes more accurate as the hurricane becomes closer. Our goal is to help make the prediction at each time period more accurate. Since slight errors in forecasting can yield results hundreds of miles apart, the instructions necessary for a particular community may be completely different. Thus the need to dial in as accurate a forecast as early as possible is clear.

V. Conclusion

In conclusion, we believe the models we built are useful in the prediction of hurricanes. Our Fourier series models revealed information about the state of hurricanes over time since the beginning of the Industrial Revolution, indicating that there is no predictable trend solely due to time at a local level, but that the frequency of hurricanes continues to increase as the climate changes because of human influence. Our histograms showed consistency in our models and our assumptions, and they revealed important geographical areas of focus with regard to the origin and prediction of hurricanes. Our linear model allowed us to predict hurricanes with 61% accuracy at best, and our error analysis showed how different classification thresholds could be used for different prediction purposes. All in all, we believe that we have done analysis and shown results that few without meteorological specialization have considered.

VI. Extensions

While we are pleased with the results of our analysis, there can always be work done to further our findings and refine our work. A big opportunity for future work is based on the data itself. While our data did tell us a lot, it is a set of only 22 features with measurements inside a limited scope. For example, our data includes a plethora of information on wind, but it does not include information on the other half of what basically makes up a hurricane: water. We could be missing some important information potentially captured in data about the temperature, sea level, underlying earth structures, saline level, fish population, or other water-related factors. In addition, our model suggests that climate change has had an impact on the number of hurricanes over time, so data pertaining to climate health could be useful in refining our model.

In addition, our work so far could feasibly be applied to predicting other natural disasters or phenomena. We have seen that our models work reasonably well for predicting hurricanes, and the types of predictors for all natural phenomena are governed by similar natural laws, so it stands to reason that similar statistical approaches to earthquakes, tsunamis, or other natural disasters would be similarly successful.

Bibliography

Bjarnadottir, Sigridur. “A Probabilistic-Based Framework for Impact and Adaptation Assessment of Climate Change on Hurricane Damage Risks and Costs.” *NeuroImage*, Academic Press, 21 Mar. 2011, www.sciencedirect.com/science/article/pii/S0167473011000063.

Dapena, Kara. “The Rising Costs of Hurricanes.” *The Wall Street Journal*, Dow Jones & Company, 29 Sept. 2018, www.wsj.com/articles/the-rising-costs-of-hurricanes-1538222400.

Domonoske, Camila. “Michael Will Cost Insurers Billions, But Won't Overwhelm Industry, Analysts Say.” *NPR*, NPR, 14 Oct. 2018, www.npr.org/2018/10/14/657333509/michael-will-cost-insurers-billions-but-wont-overwhelm-industry-analysts-say.

National Oceanic and Atmospheric Administration. (2016). *Hurricanes and Typhoons, 1851-2014* [Data file and description document]. Retrieved from https://www.kaggle.com/noaa/hurricane-database/version/1

Nelson. “Tariffs to Raise Cost of Rebuilding After Hurricane Florence.” *The New York Times*, The New York Times, 21 Sept. 2018, www.nytimes.com/2018/09/21/business/economy/tariffs-rebuilding-hurricane-florence.html.

Nordhaus, William D. “The Economics of Hurricanes and Implications of Global Warming.” *World Scientific Publishing Company*, 2010, www.worldscientific.com/doi/abs/10.1142/S2010007810000054.

Padgett, Jamie, et al. “Bridge Damage and Repair Costs from Hurricane Katrina.” *ASCE Library*, Journal of Bridge Engineering, 2008, www.ascelibrary.org/doi/full/10.1061/%28ASCE%291084-0702%282008%2913%3A1%286%29.

Samost, A. (n.d.). Hurricane Predictions and Prediction Methods

web.mit.edu/12.000/www/m2010/teams/neworleans1/predicting hurricanes.htm

Whitehead, John C. "One Million Dollars per Mile? The Opportunity Costs of Hurricane Evacuation." *NeuroImage*, Academic Press, 25 Feb. 2004, www.sciencedirect.com/science/article/pii/S0964569104000043.