Hurricane Characterization with Common Inferential Statistics and Machine Learning Techniques

Charlie Edelson

Abstract

This is where the abstract will eventually go. However, currently we only have a place holder.

Keywords: word1, word2, word3

1 Introduction

Hurricanes are an interesting meteorological feature. They occur at geological scales, with relatively slow propagation speeds. Their large size means that geophysical effects that usually are hard to observe are easily trackable. Examples of this include the coriolus force, which causes the spiraling shape of the storms, and large scale heat engines, the driving force behind a hurricane. The above properties make them prime subjects for physical and statistical modeling, especially when their effects on people are considered.

The purpose of this paper is to investigate the properties of named storms in the Atlantic ocean basin using inferential statistics and machine learning techniques. Quadratic and Cubic polynomial regressions will be used to parameterize and model storm windspeed as a function of time, with coefficients binned across storms. ARIMA models will then be fit to each storm, and the most common models will be further investigated. Additionally, the average number of storms per year, λ, will be investigated using Bayesian statistics to get a distribution on probable values of λ. Finally, points in windspeed vs pressure phase space will be clustered using two classical machine learning technique, k means clustering and hierarchical clustering. These results will serve as a starting point for further investigation into storm characterization with statistical methods.

Unisys 2000-2010 Hurricane/Tropical storm data was used throughout this paper[1]. This data consists of timestamped measurements of storm pressure, temperature, windspeed, latitude, and longitude for all major tropical depressions, tropical storms, and hurricanes between 2000 and 2010. Measurements were taken at approximately six hour intervals. Furthermore, storm name is included for all tropical storms and hurricanes.

This data was selected for two reasons. The first is the data is consistent, with few missing time intervals. The second is the data tracks the storms directly. The research does not have to make a judgment call of which local sensors (buoy, ground station, etc.) are best representative of the storm at a given time and location. This increases the repeatability, since the original data does not need to be reinterpreted by each new investigator.

2 Methods

Unnamed storms were removed from the data, as they are not the focus of this study, leaving 157 named storms over the 10 year period.

2.1 Polynomial Regression

To understand the general shape of the windspeed profile, quadratic and cubic regression of the following forms

$$w_i = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 \tag{1}$$

$$w_i = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + \beta_3 t_i^3, \tag{2}$$

where w_i is the windspeed at time t_i, were fit to each storm. The resulting coefficients, $\hat{\beta}_i$, were then binned according to values to create empirical distribution.

2.2 ARIMA Modeling

Each storm was cast into a time series and missing time steps were linearly interpolated. The minimum AIC ARIMA was then computed for $p, q, d \leq 3$. The order tuple was recorded and tallied for each unique occurrence.

2.3 Estimation of λ

Since named storms are an example of a poisson process, an estimate of the shape parameter λ, the average number of storms per year, can be computed using Bayesian analysis. The likelihood function for the evidence D given λ would then be

$$P(D|\lambda) = \frac{\lambda^k e^{-n\lambda}}{k!}, \tag{3}$$

where k is the number of storms observed in n years. It is well known that the gamma distribution, $Gamma(\alpha, \beta)$, is a conjugate prior for the poisson distribution, where α and β are the shape and inverse scale parameter, respectively. Therefore, given a gamma prior, with appropriate initial shape and inverse scale parameters, we can compute the posterior distribution as

$$P(\lambda|D) = Gamma(\alpha + k, \beta + n). \tag{4}$$

2.4 Cluster Analysis of Pressure and Wind Speed

Observed points were plotted in pressure vs wind speed phase space and then clustered using two common machine learning technique: k means clustering and hierarchical clustering. To compare results to the Saffir-Simpson scale[], 6 clusters were selected in both cases.

3 Analysis and Results

All computational analysis was performed using the python programming language with the Pandas, Scipy, NumPy, Scikit-Learn, and Stats-Models libraries[2, 3, 4, 5, 6]. Graphics and visualizations were created using the Matplotlib and Seaborn libraries[7, 8].

3.1 Polynomial Regression

In both the quadratic and cubic regressions the constant coefficient dominates the model (Fig. 1 and 2). This indicates the shape of a storm's wind profile is not representable as a polynomial of low degree. Although a higher degree polynomial could be used, reducing model variance, this would increase model bias. Aside from the risk of overfitting, a higher degree polynomial has no clear interpretation in the given context.

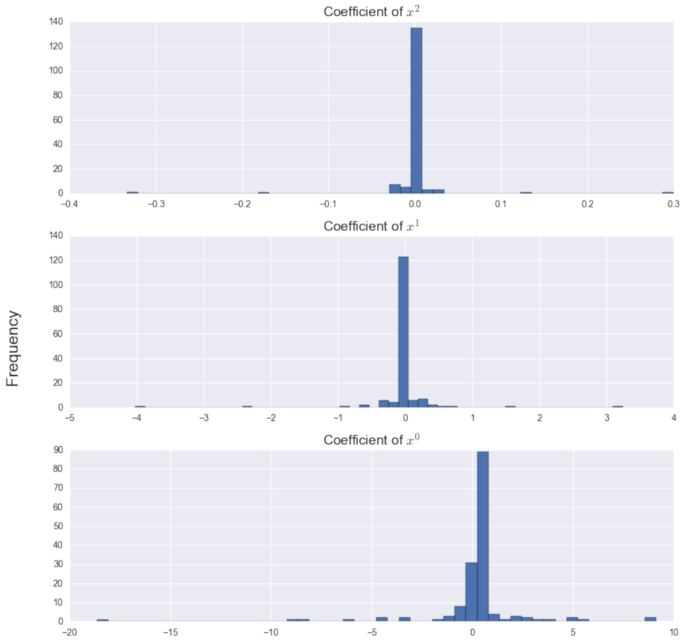

Figure 1: Histograms of the coefficients for the quadratic models fitted to each storm. Notice that $\hat{\beta}_1$ and $\hat{\beta}_2$ are both centered near zero. $\hat{\beta}_0$ has a much larger range, and is shifted slightly in the positive direction, indicating the constant coefficient dominated most models.

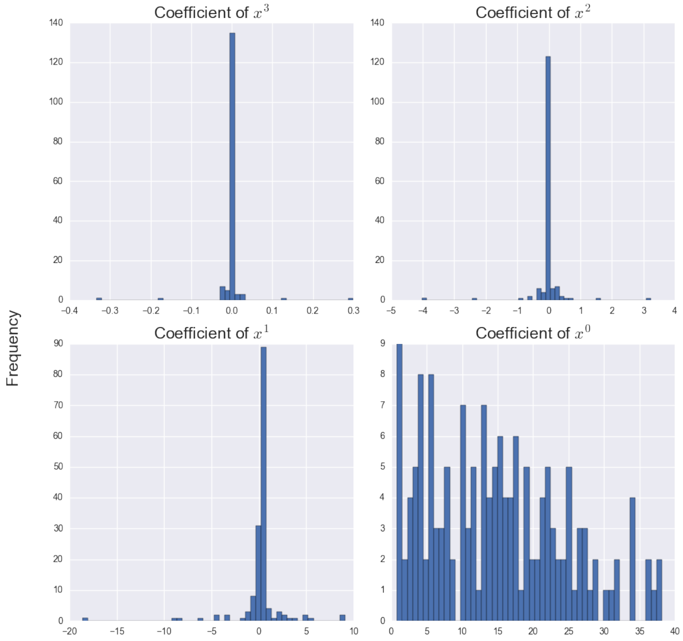

Figure 2: Histograms of the coefficients for the cubic models fitted to each storm. Notice that the first coefficient for x_1, x_2, and x_3 are all centered around zero, while x_0 has a much larger range and many more non-zero values. This indicates – as in the quadratic model – the constant coefficient is dominating the model.

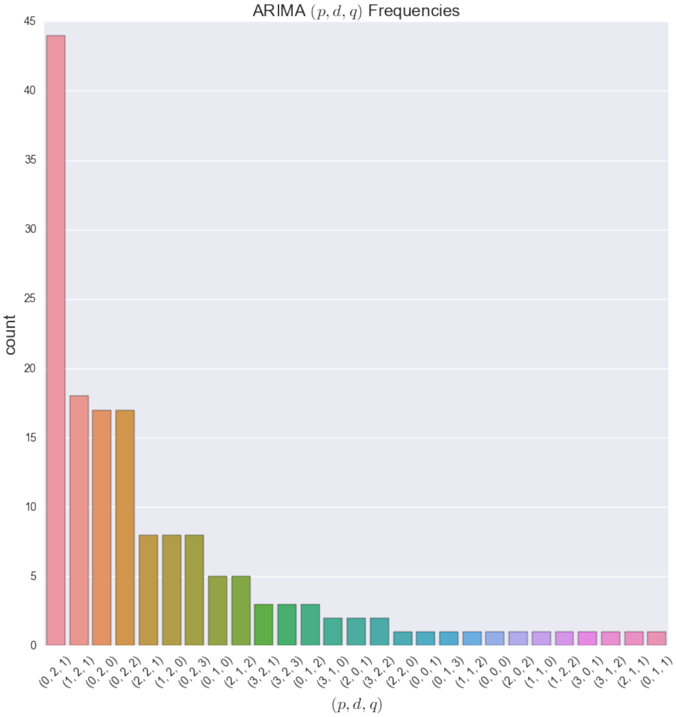

Figure 3: Count plot for the minimum AIC ARIMA coefficient for the 157 named storms. The $(0, 2, 1)$ model appears far more often than any other model. Additionally, a differencing of $d = 2$ appears in the top seven models.

3.2 ARIMA Model

For the 157 storms, an order tuple of $(0, 2, 1)$ showed up more consistently than any other option (Fig. 3). A differencing of $d = 2$ is seen in eight of the ten top models, indicating a quadratic trend in windspeed. Furthermore, $q = 1$ and $p = 0$ suggests that the storm depending on only the previous time step with zero autocorrelation.

3.3 Bayesian Estimation of λ

The prior distribution is given by $Gamma(\alpha = 2, \beta = 1/12)$, and is plotted in Fig. 4. This choice of parameters creates a large wide hump with a mode of 12, which is what the author believe to be the most likely value of λ. The width is indicative of the uncertainty in this estimate.

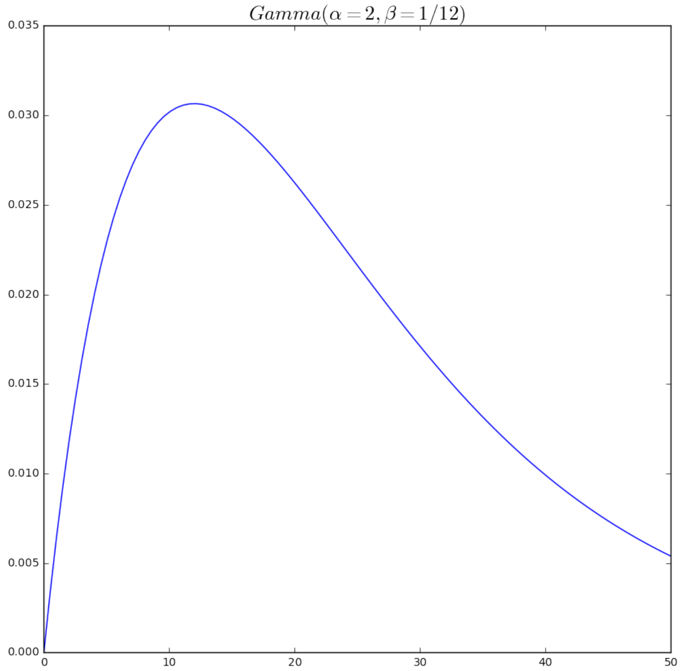

Figure 4: Gamma prior distribution with shape and inverse scale parameters $\alpha = 2$ and $\beta = 1/12$, respectively. These parameters create a wide hump centered over 12, indicating the estimate of $\lambda = 12$ is not very certain.

Since the gamma distribution is the conjugate prior to the poisson distribution, the posterior distribution for λ is found by computing $\alpha' = \alpha + k$ and $\beta' = \beta + n$, where k is the number of storms that occur in n years. For this sample, there were 157 named storms in 10 years, making the posterior distribution of λ $Gamma(\alpha' = 159, \beta' = 10.0833)$. This distribution is shown in Fig. 5. Notice that is has a much sharper peak, and with a mode of approximately 15. The narrow width of the posterior distribution indicates that the values of λ are much more certain after the Bayesian update.

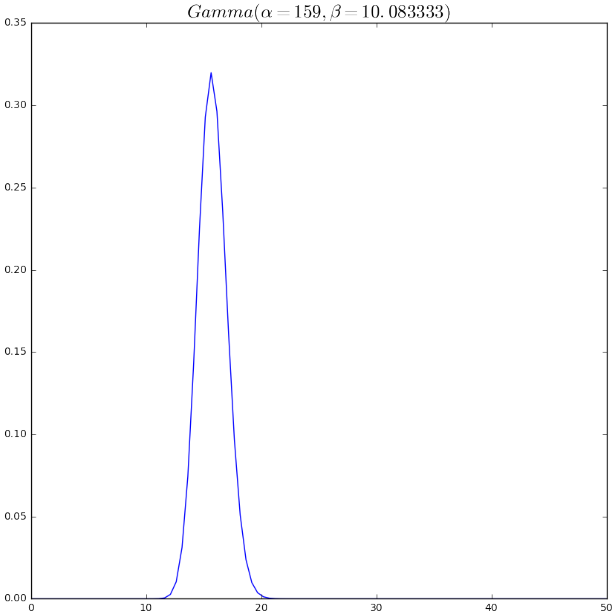

Figure 5: Posterior distribution for λ. Notice that is is a much sharper peak than the prior, and the mode has shifted slightly to the right.

3.4 Cluster Analysis: K-Means and Hierarchical Clustering

The results of both the K-means and hierarchical clustering are shown in Fig. 6. Both clustering algorithms used Ward's loss function and in both cases 6 clusters were selected. The systems are remarkably similar, both exhibiting sharp diagonal dividing lines between most clusters. The major difference lies in the clusters at lower wind speeds with higher pressures. K-means clustering still exhibits the sharp dividing lines, while hierarchical clustering in stead has a wedge shaped second cluster. Additionally, while the k-means clusters exhibit uniform width, the hierarchical clusters are narrower in the low wind speed and high pressure partition of the sample.

Additionally, the blue vertical lines indicate the divides in storm category for the Saffir-Simpson scale. The lines are are compressed to the right side of the wind speed vs pressure phase space. This is because of the scales focus on hurricanes. Interesting, it visually appears that turning the line by 45 degree would line them up with the cluster breaks.

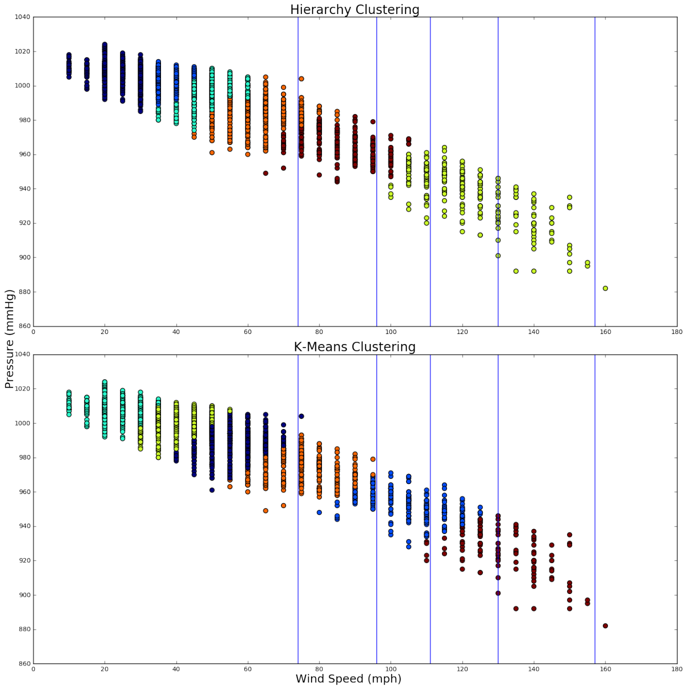

Figure 6: Hierarchical (top) and K-Means (bottom) clustering. The clusters appear fairly similar, and have sharp linear edges. The blue lines indicate the different catagories for the Saffir-Simpson scale.

References

[1] Unisys, “Hurricane/tropical data.” http://weather. unisys.com/hurricane/. Accessed: 04-20-2017.

[2] W. McKinney *et al.*, “Data structures for statistical computing in python,” in *Proceedings of the 9th Python in Science Conference*, vol. 445, pp. 51–56, van der Voort S, Millman J, 2010.

[3] E. Jones, T. Oliphant, P. Peterson, *et al.*, “SciPy: Open source scientific tools for Python,” 2001–. [Online; accessed 04-26-2017].

[4] S. v. d. Walt, S. C. Colbert, and G. Varoquaux, “The numpy array: a structure for efficient numerical computation,” *Computing in Science & Engineering*, vol. 13, no. 2, pp. 22–30, 2011.

[5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[6] J. Seabold and J. Perktold, “Statsmodels: Econometric and statistical modeling with python,” in *Proceedings of the 9th Python in Science Conference*, 2010.

[7] J. D. Hunter, “Matplotlib: A 2d graphics environment,” *Computing In Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[8] M. Waskom, O. Botvinnik, P. Hobson, J. B. Cole, Y. Halchenko, S. Hoyer, A. Miles, T. Augspurger, T. Yarkoni, T. Megies, L. P. Coelho, D. Wehner, cynddl, E. Ziegler, diego0020, Y. V. Zaytsev, T. Hoppe, S. Seabold, P. Cloud, M. Koskinen, K. Meyer, A. Qalieh, and D. Allan, “seaborn: v0.5.0 (november 2014),” Nov. 2014.