Forecasting Attributes of Tropical Cyclones Using Robust Locally Weighted Regression

Ali Yahya
Stanford University

Abstract

Critical sensor data about the state of an active tropical cyclone is often incomplete or inconsistent. We present a forecast model that attempts to capture a fundamental understanding of the relationships between a storm's different numerical attributes. The model can thereby help verify or reconstruct corrupted data about any given tropical cyclone. The forecast model we present is based on a supervised machine learning algorithm known as LOESS which implements robust locally weighted regression in n-dimensional space, where n is the number of attributes of the storm under consideration. The model presented yields good results with error margins of less than 10% for attributes that exhibit a high degree of correlation with other attributes in the model.

1 Introduction

Tropical cyclones are highly structured storm systems that are characterized by numerous thunderstorms, strong winds, and heavy rain that revolve around a large low-pressure center. Because of the high degree of structure in tropical cyclones, their data model is well defined and relatively self-contained. Consequently, forecast models that treat tropical cyclones as systems in isolation—that is, without consideration for collateral or larger scale factors—can be more successful than similar models applied to less structured storm systems.

1.1 Patterns in Hurricane Data Models

At any given moment, any one tropical cyclone and its present behavior can be almost entirely described with values for a relatively small set of attributes. Moreover, previous studies have indicated that said attributes not only represent a reasonably complete snapshot of a tropical cyclone's behavior but also that they are highly correlated with each other [3]. One important subset of those attributes could include: minimum central pressure, maximum wind speeds, latitude, longitude, radius of eyewall, radius of maximum wind speed, radius of outer closed isobar, pressure of outer closed isobar, among others.

In this paper, we analyze data for approximately over 1,000 hurricanes in the Atlantic and Eastern North Pacific Basins. For each hurricane in the dataset, we considered between 10 and 30 six-hourly snapshots of the storms attributes. We point out a number of statistically significant correlations between seven attributes that generally pertain to tropical cyclones [3].

Further, we formalize our observations and present a forecast model, based on locally weighted linear regression, that allows meteorologists to use historical storm data to extrapolate the value of a single corrupted data attribute in the data model of a storm, given correct values for a sufficient number of other attributes.

1.2 Motivation: Extrapolating Sensor Data

Often times, at the most critical moment of a storm's development, sensor data about its state can be incomplete or fraught with inconsistencies. In situations like that, a forecast model that is based on a fundamental understanding of the relationships between a storm's different attributes can help verify or reconstruct the dataset of an ongoing storm.

Additionally, a forecast model that sheds light on the interactions between the different attributes in a storm's data model, enables meteorologists to better predict the side-effects of changes to any one of those attributes. Consequently, a successful forecast model for tropical cyclones may not only improve the accuracy of warnings issued, but also their qualitative preciseness.

For instance, although it's fairly well understood that an increase in minimum central pressure will result in a decrease of maximum wind speeds, it is also important to consider a more subtle effect: A decrease in minimum central pressure will indeed result in lower wind speeds, but because of the law of Conservation of Angular Momentum, the radius around which maximum wind speeds revolve will also increase. Therefore, the resulting storm surge and size of the area of devastation will be different than with lower pressure and higher wind speeds. Understanding exactly how such factors are likely to be different is critical for the effectiveness of a tropical cyclone warning system.

2 Patterns Observed

As a preliminary step to crafting a unified forecast model, we present a number of statistically significant pairwise correlations that were observed from plots of data from *The Tropical Cyclone Extended Best Track Dataset* [4]. The following attributes were tracked every six hours throughout the lifetime of approximately 1,000 tropical cyclones:

- latitude
- longitude
- maximum wind speed
- minimum central pressure
- radius of maximum wind speed
- pressure of outer closed isobar
- radius of outer closed isobar

In the following sub-sections, we present a summary of the four most important pairwise correlations observed.

2.1 Pressure vs. Wind Speeds

The most clearly expressed pairwise relationship in the data was that between tropical cyclones' minimum central pressure and it's corresponding maximum wind speeds. As illustrated by Figure 1, there is a strong negative correlation between the two attributes.

Because the magnitude of the pressure gradient in the atmosphere is the primary driver of wind speeds [1], it is not unexpected that a storm with a lower pressure center would produce higher wind speeds.

2.2 Pressure vs. Max. Wind Speed Radius

A more subtle pairwise interaction (that was alluded to before) is that between tropical cyclones' minimum central pressure and the radius around which the cyclone's maximum wind speeds revolve. Figure 2 illustrates a statistically significant positive correlation between the two attributes.

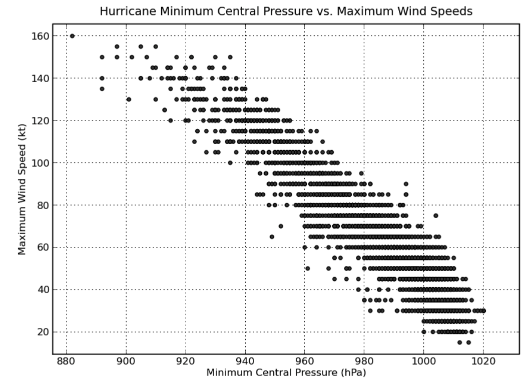

Figure 1: Pressure vs. Wind Speeds

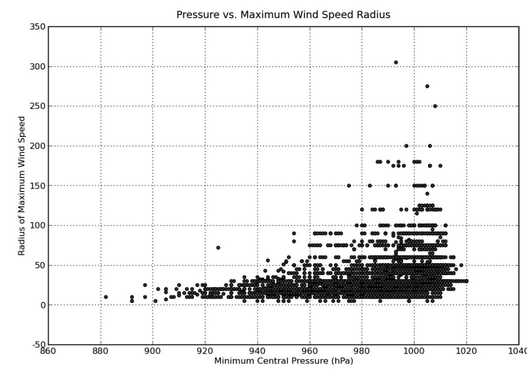

Figure 2: Pressure vs. Max. Wind Speed Radius

Note that, because of the relationship between pressure and wind speed described in the previous section, a higher central pressure is correlated with slower wind speeds. Because of the law of Conservation of Angular Momentum, slower wind speeds are in-turn correlated with larger wind speed radii. Consequently, a tropical cyclone's minimum central pressure is positively correlated with the radius of the storm's maximum wind speeds.

2.3 Maximum Wind Speed vs. Maximum Wind Speed Radius

A direct corollary of the relationship described in the previous subsection, is that a storm's maximum wind speed is negatively correlated with its maximum wind speed radius. The relationship is illustrated by Figure 3. Note that as maximum wind speeds increase, the maximum wind speed radius tends to decrease.

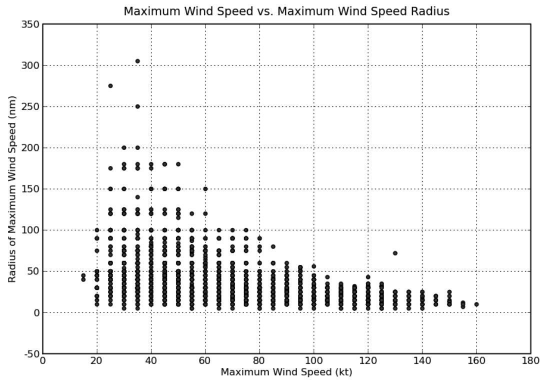

Figure 3: Max. Wind Speed vs. Wind Speed Radius

2.4 Longitude vs. Wind Speed

Finally, Figure 4 reveals that as hurricanes in the Atlantic basin move westward, their maximum wind speeds tend to increase. However, once storms reach a certain longitude (i.e. approximately 90° West), wind speeds dramatically drop. This can be attributed to probable landfall against the Atlantic coast of the North American continent.

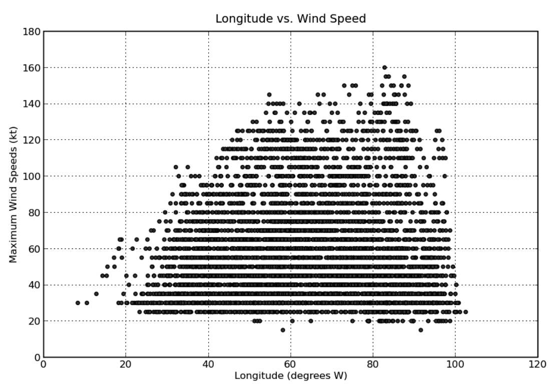

Figure 4: Longitude vs. Max. Wind Speeds

As a tropical cyclone travels over land, its primary source of energy, warm evaporating water, is no longer available. Consequently, the storm weakens and eventually dissipates. Note that the cluster of data points with wind speeds of around 120 knots at longitudes of around 100° occurred over the Gulf of Mexico.

3 Crafting a Numerical Weather Model

We present a numerical model that captures the many-to-many relationships between the different numerical attributes of a tropical cyclone's behavior. The model is based on a supervised machine learning algorithm known as LOESS [2]. The algorithm implements robust locally weighted regression in n-dimensional space, where n is the number of attributes taken into account by the model. In contrast to the manual, pair-wise analysis used to identify a few trends in the previous section, the numerical model succeeds at capturing the relationship between each attribute and every other attribute in the model.

3.1 Defining a Hypothesis Function

The preliminary analyses of storm data presented in the previous section suggest that most interactions between attributes in tropical cyclones can be approximated with a linear function. As a result, the hypothesis function for the forecast model that we present is of the form:

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n \qquad (1)$$

In the expression above, $x_1, \ldots, x_n$ are referred to as the model's features. Each x_i is equal to the value of the i-th known attribute of storm S. The linear model's parameters are given by $\theta_0, \ldots, \theta_n$ where θ_0 specifies the intercept term and θ_1 through θ_n specify the linear function's coefficients. The value of $h_\theta(x)$ is the model's prediction for the value of y, a single attribute of storm S whose value is not known (e.g. was corrupted).

To simplify our notation, we introduce the convention of letting $x_0 = 1$ so that the hypothesis function can be written in matrix vector form:

$$h_\theta(x) = \sum_{i=0}^{n} \theta_i x_i = \theta^T x \qquad (2)$$

Above, n is the number of features (not counting x_0). The subexpression on the right hand side views θ as a vector of the model's parameters, and x as a vector of its features [5].

3.2 Training Data

Predictions made by the numerical model about any particular tropical cyclone are based on a database containing historical data of similar tropical cyclones that have occurred in the past. For this reason, LOESS is a supervised learning algorithm, as it requires training data whose correctness is, in a sense, *supervised*.

A training dataset is composed of a matrix X and a vector $\vec{y}$. Each row in X represents a single datapoint (namely, a single observation of a tropical cyclone) and each column denotes an attribute of the storm (e.g. minimum central pressure). The observed historical values for the attribute that we would like to predict are stored as elements of the vector $\vec{y}$. The elements of vector $\vec{y}$ are known collectively as the labels of the training set.

Note that the i-th label in $\vec{y}$ corresponds to the data point stored in i-th row of the matrix X.

As mentioned before, the training data used for our forecast model comes from *The Tropical Cyclone Extended Best Track Dataset* [4].

3.3 Estimating the Model's Parameters

3.3.1 Defining a Cost Function

The algorithm minimizes the following cost function with respect to a parameter vector θ:

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{m} w^{(i)} \left(\theta^T x^{(i)} - y^{(i)} \right)^2 \qquad (3)$$

Equation 3 is a variation of the familiar least-squares cost function that gives rise to the ordinary least squares regression model [5]. In order to factor in possible nonlinear relationships between attributes in a storm, we introduce a weighing factor, $w^{(i)}$, a multiplier to each term within the sum.

For convenience, let r be equal to $\lfloor \alpha\, n \rfloor$ where $\alpha \in (0, 1)$ is a tuning parameter given to the algorithm, and n is the number of data points observed. Note that, in our case, n is equal to approximately 1,000. The value of $w^{(i)}$ then is given by the expression:

$$w^{(i)} = W \left(\frac{x - x_i}{h_i} \right) \qquad (4)$$

where h_i is a scaling factor that is equal to the r-th smallest number among $|x_i - x_j|$ for $j = 1, \ldots, n$. In equation 4, $W(x)$ is the tri-cube weighting function, given by:

$$W(x) = \begin{cases} (1 - |x|^3)^3, & \text{for } |x| < 1 \\ 0, & \text{for } |x| \geq 1 \end{cases} \qquad (5)$$

Note that the weights depend on the particular query point x at which we are trying to evaluate our hypothesis. Moreover, if $|x^{(i)} - x$ is small, then $w^{(i)}$ is closer to 1 than if $|x^{(i)} - x$ is large. Hence, the cost function given by equation 3 gives a higher weight to data points that are close to the query point x than those that are far away [5].

Finally, if we express cost function in equation 3 in matrix-vector notation we obtain the following expression for J:

$$J(\theta) = \frac{1}{2} (X\theta - \vec{y})^T W (X\theta - \vec{y}) \qquad (6)$$

3.3.2 Minimizing the Cost Function

We minimize the cost function J analytically by taking the derivative with respect to θ of equation 6 above, setting the result equal to zero, and solving for θ.

Thereby, our estimate of our forecast model's parameters θ is given by:

$$\theta = (X^T W X)^{-1} X^T W \vec{y} \qquad (7)$$

The above expression optimally minimizes the cost function J, and therefore provides the best possible estimate for the parameters of our linear model.

Once θ has been computed, it can be plugged back into our hypothesis function (equation 2) in order to perform a prediction for any new data point x whose corresponding y may be unknown.

4 Results

Testing was conducted by performing take-one-out cross validation on the original training dataset. On tests where y was either minimum central pressure or maximum wind speed, the average percent error in the model's predictions was 8.4%. Whenever y was any one of the other, less correlated, attributes, the total average percent error in the model's predictions was 24.7%.

5 Conclusion

The use of locally weighted linear regression as the foundation of a forecast model for tropical cyclones yields acceptable error margins on attribute predictions. The model presented can be used to reconstruct incomplete or inconsistent data about the state of a tropical cyclone. Further, the forecast model sheds light on the interactions between the different attributes in a storm's data model.

References

[1] AHRENS, C. D. *Meteorology Today: An Introduction to Weather, Climate, and the Environment*. Brooks/Cole, Cengage Learning, Belmont, CA, 2007.

[2] CLEVELAND, W. S. Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association 74*, 368 (December 1979).

[3] DEMARIA, M. An Updated Statistical Hurricane Intensity Prediction Scheme (SHIPS) for the Atlantic and Eastern North Pacific Basins. *Journal of Weather and Forecasting 14*, 3 (1999).

[4] DEMUTH, MARK DEMARIA, K. The Tropical Cyclone Extended Best Track Dataset. *Journal of Applied Meteorology 45* (2006), 1573–1581.

[5] NG, A. Supervised Learning. *Stanford University, CS 229 — Machine Learning* (2010), 3–15.