University of South Florida
Scholar Commons

Graduate Theses and Dissertations | Graduate School

April 2016

A Statistical Analysis of Hurricanes in the Atlantic Basin and Sinkholes in Florida

Joy Marie D'andrea

Follow this and additional works at: http://scholarcommons.usf.edu/etd

Part of the Mathematics Commons, and the Statistics and Probability Commons

Scholar Commons Citation

D'andrea, Joy Marie, "A Statistical Analysis of Hurricanes in the Atlantic Basin and Sinkholes in Florida" (2016). *Graduate Theses and Dissertations.*
http://scholarcommons.usf.edu/etd/6077

This Dissertation is brought to you for free and open access by the Graduate School at Scholar Commons. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Scholar Commons. For more information, please contact scholarcommons@usf.edu.

A Statistical Analysis of Hurricanes in the Atlantic Basin & Sinkholes in Florida

by

Joy D'Andrea

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Mathematics &Statistics
College of Arts and Sciences
University of South Florida

Major Professor: Rebecca D. Wooten, Ph.D.
Chris Tsokos, Ph.D.
Gregory McColm, Ph.D.
Dan Chen, Ph.D.

Date of Approval:
April 4, 2016

Keywords: Hurricanes, Sinkholes, Soil types, Logistic Regression, Regression Analysis

Copyright © 2016, Joy D'Andrea

DEDICATION

To my advisor Dr. Rebecca Wooten. Thank you for the wonderful experience and journey into statistics analysis of environmental issues. I will never forget all the long and tedious hours of programming in your office. The preparation and vast amounts of discussion has more than prepared me to go out into the world and achieve success. It has been a great path to follow in your footsteps. Thank you for everything! I cannot wait for our future endeavors to follow from this project.

ACKNOWLEDGEMENTS

First, I would like to thank my advisor Dr. Rebecca Wooten, for taking a chance on me, for the guidance, support, direction, and keeping me on track to finish. You're an amazing person to learn from. You have truly inspired me to become a better writer and statistician. I would never have been in this position if it wasn't for your patience, understanding and continued pushing to make me a strong statistician. I hope we continue to work together in the future.

I would like to thank my committee members for being patient and helpful throughout the dissertation process. Thank you Dr. McColm for being on my committee and reading about material that is not in your field of interest. I greatly appreciate your time and efforts to be a part of this project. Thank you Dr. Shen for willing to be on my committee and providing your assistance. Thank you to Dr. Tsokos for being a great director and providing valuable advice and feedback during this process. I would like to personally thank my dissertation chair Dr. Thomas Becker for taking his time and efforts to preside over the defense process.

I would like to thank my workplace, The University of South Florida Sarasota - Manatee, for their contribution to paying for my dissertation hours and other courses needed to finish this degree. In particular, I would like to thank my boss, the Dean of the College of Arts and Sciences, Dr. Jane Rose. Thank you for always being supportive in the process of continuing my education. You're a phenomenal example to follow. I would like to thank Barbara Melfi, Cindy Kish, Darryl Waddy, and Angie Chromiak - Sears for all of their help with the scheduling, paper work, office transfers, phone calls, assistance with my questions, and having an enormous

amount of patience with me. I would also like to thank Holly Fitch, Amanda Crouch and Dana Arace (from student services) for their continued kindness and collaboration with working with students in my classes. I would like to thank all the administrative staff and people who work in the Math Department office: Vickie, Denise, Mary-Ann, and Francis. Thank you for your assistance of my paper work, printing, coffee, and kind words. I would like to say thanks to one of my best friends Wendy Pogoda, who has always been in my corner and always listens to me. I would like to thank my academic friends Jonathan Burns, George Kimber, Dr. Brendan Nagle, Dr. Boris Shekhtman, and Daviel Leyva for providing humorous and fun times when I needed to unwind. I would like to thank Richard Brower for his patience, kindness, love, and extreme generous efforts to keep me calm. To Casha, thank you for the late night telephone conversations that helped keep me on track. Lastly, I would like to thank my Mother for being there for me when I had my moments of doubt and frustration. I love you so much Mom. Thank you for putting up with me.

TABLE OF CONTENTS

List of Tables v

List Of Figures..... vi

Abstract..... ix

Chapter 1: Motivation, Background & Statistical Methodology..... 1
- Motivation 1
- Background... 2
- Statistical Methodology..... 4
  - Parametric Analysis..... 5
  - Nonparametric Analysis 6
  - Circular Analysis 7
  - Exploratory Factor Analysis..... 8
  - Correlation..... 9
  - Simple Linear Regression..... 10
  - Multiple Linear Regression 11
  - Logistic Regression 12
  - Non – Response Analysis..... 13
  - Forward Selection, Backward Elimination and Subset Analysis 17
  - Survival Analysis..... 18

Chapter 2: Latent Storm Factors and Their Indicators & Non – Response Factor Modeling of Hurricanes 20
- Exploratory Factor Analysis..... 20
  - Non – Response Analysis Model 34
- Exploratory Factor Analysis & Non – Response Modeling on the Florida Keys 36
  - Non – Response Analysis Model 40
- Usefulness & Contributions 41

Chapter 3: Logistic Regression of Hurricanes in the Atlantic Basin from 1990 – 2014..... 43
- Introduction to the Data..... 43
  - Hurricane and Buoy Data 43
  - Compilation of Hurricanes 44
- Variables of Interest 53
  - Description of the Response Variable and Contributing Entities..... 54
    - Buoy Wind Speed 54
    - Buoy Wind Direction..... 54
    - Buoy Pressure 55

Buoy Atmospheric Temperature 55
Buoy Water Temperature 55
General Descriptive Statistics for the Buoy Conditions 55
Binomial Case of Logistic Regression 57
Model Measurement of Accuracy 58
Model Development 59
Multinomial Case of Logistic Regression 63
Model of a Storm Being Present Categorically 66
Exploratory Factor Analysis in Conjunction with Non-Response Analysis 71
Usefulness & Contributions 76

Chapter 4: A Statistical Analysis of Florida Sinkholes 77
Ranking of Soil Types 78
Relationship between the Sinkhole Length and Width 79
Average Diameter of a Sinkhole 85
Probabilities of a Sinkhole Occurring 87
Usefulness & Contributions 90

Chapter 5: Survival Analysis of Florida Sinkholes 91
Probable Kaplan Meier Estimate TTE. 91
Probable Kaplan – Meier Estimate of Soil Type 93
Probability Distribution that Best Characterizes the TTE 96
Association of Covariates of a Sinkhole 97
Developed Cox Ph Model for the Associated Covariates 99
Hazard Ratio for the Associated Covariates 100
Usefulness & Contributions 100

Chapter 6: Future Projects and Works 102

References 104

LIST OF TABLES

Table 1.1: Hurricane Classification Wooten - Tsokos Scale 3

Table 2.1: Factor Loadings of the Factor Indicators 22

Table 2.2: Variances Explained from EFA 23

Table 2.3: Factor Indicator Correlations 24

Table 2.4: Factor Correlations 25

Table 2.5: Factor Loadings 27

Table 2.6: Variances Explained from EFA 28

Table 2.7: Factor Indicator Correlations 29

Table 2.8: Factor Correlations 29

Table 2.9: Factor Loadings of the Factor Indicators – Florida Keys 38

Table 2.10: Variances Explained from EFA 39

Table 2.11: Factor Indicator Correlations 39

Table 2.12: Factor Correlations 40

Table 3.1: General Descriptive Statistics for Buoy Conditions 56

Table 3.2: Mean Values for Buoy Conditions when Storm is Present, Not Present and Overall 57

Table 3.3: Measurement of Accuracy and R – Squared values for the Developed Models 61

Table 3.4: Average Buoy Conditions for Storm Present (Categorically) 64

Table 3.5: Probabilities of a Storm being Present (categorically) using all 5 Developed Models 70

Table 3.6: All Possible Terms 74

Table 3.7: SS Loadings 75

Table 4.1: Soil Type Sand Ranking of Frequency of Occurrence (Mixed or Combined) 79

Table 4.2: Descriptive Statistics for Sinkhole Length and Sinkhole Width 80

Table 4.3: Goodness – of – Fit - Tests for the Best Fit Distributions for the Sinkhole Length 81

Table 4.4: Goodness – of – Fit - Tests for the Best Fit Distributions for the Sinkhole Width 82

Table 4.5: ANOVA for Sinkhole Length and Sinkhole Width 84

Table 4.6: Average Values of Sinkhole Length, Width, Depth, & Slope for the Soil Types 88

Table 5.1: Goodness – of – Fit - Tests for the Best Fit Distributions for the TTE 96

Table 5.2: Calculations from the Cox PH Model 99

LIST OF FIGURES

Figure 1.1: Illustration of the Angle of Separation 16

Figure 1.2: Distance Removed (Height) 17

Figure 2.1: Factor Indicators 21

Figure 2.2: Second EFA Factor Indicators 26

Figure 2.3: Bar Graph of the Variable Month 32

Figure 2.4: Scatterplots of Starting Latitude, Starting Longitude and Month 33

Figure 2.5: Predicted Model using Standard Regression (red) & Non – Response Analysis (black) 35

Figure 2.6: Non – Response Model Fitting 35

Figure 2.7: Variables of Interest – Florida Keys 37

Figure 3.1: Variables from the Unisys Weather Site 44

Figure 3.2: Hurricane Data 1990 – 2014 from Unisys 45

Figure 3.3: Added Headings to the Smaller Hurricane Data Set 46

Figure 3.4: Data Set 3: Extract List of Storms for Given Years and Add Year to List 46

Figure 3.5: Data Set 1 and Data Set 2 merged 47

Figure 3.6: Location of Buoy 42001 48

Figure 3.7: Variables of Interest from the National Buoy Data Center (Buoy 4) 49

Figure 3.8: Additional Variables Included 50

Figure 3.9: Variables of Interest 51

Figure 3.10: Final Included Variables of Interest for Compilation Data Set 52

Figure 3.11: Variables of Interest 53

Figure 3.12: Boxplots of the Average Atmospheric Temperatures (Categorically) 65

Figure 3.13: Boxplots of the Average Water Temperatures (Categorically) 66

Figure 3.14: Data diagram of named storms in the Atlantic Basin 72

Figure 3.15: Measured variables of interest including time shifts in the buoy conditions 73

Figure 3.16: Correlation between the observed and estimated wind speed based on the buoy conditions over the give time delay 76

Figure 4.1: Variables of Interest 77

Figure 4.2: Best Fit Probability Density Function of Sinkhole Length and Width 82

Figure 4.3: Scatterplot of the Sinkhole Length and Width 83

Figure 4.4: Best Fit Probability Distribution for the Diameter of a Sinkhole 86

Figure 5.1: Kaplan – Meier Graph 92

Figure: 5.2: Survival Function of TTE 93

Figure: 5.3: Kaplan – Meier Survival Probabilities of TTE (censoring) 94

Figure 5.4: Kaplan – Meier graph of TTE in the Soil Types 94

Figure 5.5: Kaplan – Meier Survival Probabilities of TTE in Soil Types 95

Figure 5.6: Best Fit Probability Distribution of TTE 97

Figure 5.7: Variables of Interest 98

ABSTRACT

Beaches can provide a natural barrier between the ocean and inland communities, ecosystems, and resources. These environments can move and change in response to winds, waves, and currents. When a hurricane occurs, these changes can be rather large and possibly catastrophic. The high waves and storm surge act together to erode beaches and inundate low-lying lands, putting inland communities at risk. There are thousands of buoys in the Atlantic Basin that record and update data to help predict climate conditions in the state of Florida. The data that was compiled and used into a larger data set came from two different sources. First, the hurricane data for the years 1992 – 2014 came from Unisys Weather site (Atlantic Basin Hurricanes data, last 40 years) and the buoy data has been available from the national buoy center. Using various statistical methods, we will analyze the probability of a storm being present, given conditions at the buoy; determine the probability of a storm being present categorically. There are four different types of sinkholes that exist in Florida and they are: Collapse Sinkholes, Solution Sinkholes, Alluvial Sinkholes, and Raveling Sinkholes. In Florida there are sinkholes that occur, because of the different soil types that are prevalent in certain areas. The data that was used in this study came from the Florida Department of Environmental Protection, Subsidence Incident Reports. The size of the data was 926 with 15 variables. We will present a statistical analysis of a sinkholes length and width relationship, determine the average size of the diameter of a sinkhole, discuss the relationship of sinkhole size depending upon their soil types, and acknowledge the best probable occurrence of when a sinkhole occurs. There will be five research chapters in this dissertation. In Chapter 2, the concept of Exploratory Factor

Analysis and Non-Response Analysis will be introduced, in accordance of analyzing hurricanes. Chapter 3 will also address the topic of hurricanes that have formed from the Atlantic Basin from 1992 – 2014. The discussion of the probability of a storm being present (also categorically) will be addressed. In Chapter 4 a study of sinkholes in Florida will be addressed. In Chapter 5 we will continue our discussion on sinkholes in Florida, but focus on the time to event between the occurrences of the sinkholes. In the last chapter, Chapter 6, we will conclude with a future works and projects that can be created from the foundations of this dissertation.

CHAPTER 1: MOTIVATION, BACKGROUND & STATISTICAL METHODOLOGY

MOTIVATION

The motivation behind the research found in this dissertation was the author's personally experiences with hurricanes and sinkholes. In 1992 Hurricane Andrew made its first landfall in Elliot Key. The author was fifteen years of age when this tragedy occurred. When the hurricane made landfall the author's family had been staying at a shelter and the author found herself separated from her family for over two days. During that time, there was continuous flooding, and she found herself drifting on a hotel door. This event sparked interest in better understanding meteorological events. This type of phenomenon affects the majority of individuals in the United States, especially Florida. There are about 19.89 million people that live in the state of Florida, according to the United States Census Bureau (as of 2014). Hurricane season lasts approximately 6 months. This is from the dates of June 1st until November 30th.

If a person lives in Florida then they should have knowledge of how to prepare for a hurricane and understand other environmental issues such as sinkholes. One reason that the author decided to study sinkholes and perform a statistical analysis on them was because a dear friend of hers died as a result of a sinkhole. This devastation happened in April 2012. The event left the author intrigued as to predicting the probabilities of where sinkholes may occur in different parts of Florida. Sinkholes usually occur most frequently during the spring months and lower in the fall months of the year. Another reason for choosing the topic of hurricanes and sinkholes, is to

analyze them to better understand the subject phenomenon and make better predictions. Florida has the highest frequency of sinkhole occurrences in the United States. The motivation to educate others on the topic of sinkholes in Florida is an ongoing battle. Sinkholes can sometimes happen right under our homes, schools, work places, etc. Preparation and awareness of when and how sinkholes arise should become a part of a Florida citizen's everyday knowledge.

BACKGROUND

Hurricanes have been a topic of interest for over 500 years. Scientists began to better understand hurricanes during the 1800s, with forecasters being able to issue warnings as storms approached Hurricanes remain difficult to predict, especially because they can suddenly intensify in ways that are poorly understood. In this paper, the hurricanes of interest occur from the years 1975 – 2014. There are two ways of hurricane classification; the Saffir – Simpson scale and the Wooten-Tsokos scale developed in 2009. We will be using the Wooten – Tsokos scale in this dissertation.

In the addition the study of hurricanes in the Atlantic Basin, sinkholes are another environmental issue that we will discuss in this dissertation. Sinkholes occur more in Florida than any other state in the United States. "Florida's peninsula is made up of porous carbonate rocks such as limestone that store and help move groundwater" [12]. "Dirt, sand and clay sit on top of the carbonate rock" [12]. "When the dirt, clay or sand gets too heavy for the limestone roof, it can collapse and form a sinkhole" [12]. Sinkholes are caused naturally, however they can be triggered by outside events [12].

Table 1.1: Hurricane Classification Wooten - Tsokos Scale

Type	**Category**	**Pressure (hPa)**	**Wind (knots)**
Tropical Depression/Tropical Storm	**0**	995 – 1010	10 - 42
Hurricane	**1**	972 - 994	43 - 77
Hurricane	**2**	951 - 971	78 - 102
Hurricane	**3**	932 - 950	103 - 122
Hurricane	**4**	911 - 931	123 - 142
Hurricane	**5**	< 911	>143

There are four different types of sinkholes that exist in Florida and they are: **Collapse, Solution, Alluvial,** and **Raveling.** "**Collaps**e sinkholes occur in areas where there are extensive cover materials over a limestone layer" [13]. "When solution creates a hole in the limestone and the limestone roof over the cavern either dissolves or no longer can support the weight of the overlying materials, these cover materials collapse into the cavern, leaving a funnel shaped sinkhole, usually circular in outline" [13].

"If the overlying cover is clastic sediments it is called a cover collapse sink" [13]. "If it is limestone, it is a rock collapse sink" [13]. "It is common that the formation of collapse sinkholes is sudden or even catastrophic" [13]. "This may be a result of human activity, especially those that affect the hydrology of an area" [13]. "**Solution** sinkholes form more slowly and gradually as a result of enlargement of joints by solution" [13]. "Eventually the rocks may settle and the

cover material washes into the cavern in a process called raveling" [13]. "These sinkholes are not as potentially impacted by human activities as are collapse sinkholes" [13]. "**Alluvial** sinkholes are older sinkholes partially or entirely filled with sediments due to subsequent marine deposition or by materials washed in from the sides are called alluvial sinkholes" [13]. "Where the water table is shallow, they are often indicated by ponds, wetlands or cypress domes" [13].

"Raveling sinkholes arise from the above alluvial sinkholes and may become reactivated when the aquifer levels rise or drop" [13]. "The lowering of the aquifer levels creates a loss of buoyant support, increasing the water content of the plug such as happens when the water levels rise, increases the load and decreases the cohesion of the sediments" [13]. "When the sediments are no longer supported, the plug rapidly collapses" [13]. "They are only one of many kinds of karst landforms, which include caves, disappearing streams, springs, and underground drainage systems, all of which occur in Florida" [13].

STATISTICAL METHODOLOGY

In this section we will introduce the statistical methods that were used to analyze and produce the results for the research questions/statements addressed in this dissertation. The statistical methods used in this dissertation were Parametric Analysis, Nonparametric Analysis, Circular Analysis, Exploratory Factor Analysis, Correlation, Simple Linear Regression, Multiple Linear Regression, Logistic Regression, Non-Response Analysis, Forward Selection, Backward Elimination, Subset Analysis, and Survival Analysis.

Parametric Analysis

Parametric Analysis is a branch of statistics which assumes that the data have come from a type of probability distribution and makes inferences about the parameters of the distribution [7]. We will use maximum likelihood estimates to fit various probability distributions and determine the probability distribution that best characterizes the variable of interest. For each continuous numerical measure of our data sets, 65 continuous distributions, will be compared and ranked using the goodness-of-fit tests; **Anderson-Darling, Kolmogorov-Smirnov**, and **Chi-Square**.

The null hypothesis for all such tests is the data fits the desired distribution; and the alternative hypothesis is the data does not fit the desired distribution.

For the **Anderson Darling** Goodness of Fit Test, the test statistic is:

$A^2 = - N - S$, where $S = \sum_{i=1}^{N} \frac{2i-1}{N} \left[\ln\left(F(Y_i) + \ln\left(1 - F(Y_{N+1-i})\right)\right)\right]$, with F as the specified cumulative distribution and Y_i as the ordered data.

For the **Kolmogorov-Smirnov** Goodness of - Fit Test, the test statistic is:

$D = \max_{1 \le i \le N}(F(Y_i) - \frac{i-1}{N}, \frac{i}{N} - F(Y_i))$, with F as the specified cumulative distribution and Y_i as the ordered data.

For **Chi-Square** Goodness of - Fit Test, the test statistic is: $\chi^2 = \frac{(O_1 - E_i)^2}{E_i}$, where the expected value (E_i) of the data based on the assumed distribution, and the observed value (O_i) of the data that is given [Gei]. As in standard hypothesis testing, if the test statistic for each of our above tests is greater than the critical value, then we can reject the null hypothesis and conclude that we

will not have a good fit for the data. Otherwise, we can fail to reject the null hypothesis and conclude that we will have a good fit for the data [7]

Nonparametric Analysis

"Nonparametric statistics are statistics that are not based on parameterized families of probability distributions" [18]. "Nonparametric analysis includes both descriptive and inferential statistics" [18]. "Nonparametric statistics makes no assumptions about the probability distributions of the variables being assessed" [18]. Some examples of nonparametric statistics are the Wilcoxon rank-sum test or the permutation and resampling tests. In this dissertation, we will use the Wilcoxon rank-sum test. The Wilcoxon rank - sum test is used is to test the null hypothesis that the median of a distribution is equal to some value [18]. The procedure for the Wilcoxon rank - sum test is:

1) State the null and alternative hypothesis: $H_0: \tilde{\mu}_1 - \tilde{\mu}_2 = 0$, $H_1: \tilde{\mu}_1 - \tilde{\mu}_2 \neq 0$
2) Order the data values from both samples in a single list arranged from the smallest to largest.
3) In another column, we assign the numbers 1 to N, where N = $n_1 + n_2$. Note that these are the ranks of the observations.
4) Now let W denote the sum of the ranks for the observations from the first population.
5) If there is no difference between the two medians (the null is true), the value of W will be around half the sum of the ranks.
6) Calculate the expected value $E(W) = \frac{n_1(N+1)}{2}$ and the variance $V(W) = \frac{n_1 n_2(N+1)}{12}$
7) Calculate the test statistic given by $z = \frac{W - E(W)}{\sqrt{V(W)}}$ and find the associated p-value using normal approximation.

8) Use the decision rule to reject or fail to reject the null hypothesis

Circular Analysis

"Directional or circular distributions are those measures that have no true zero and any designation of high or low values is arbitrary such as: compass direction, hours of the day, months of the year, and wind direction" [9]. We are given a sample of n angles, where these angles are in degrees. For these angles to analyze directional data, they must first be transformed from polar coordinates to rectangular coordinates. "The mean angle cannot be the sum of the angles divided by the sample size, this is because the mean angle of 359° and 1° (north) would be 180° (south) [9]. Hence, we need to use the following equations" [9]:

$$\bar{y} = \frac{\sum_{i=1}^{n} sin\alpha}{n}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} cos\alpha}{n}$$

$$\bar{r} = \frac{\sqrt{(\sum_{i=1}^{n} sin\alpha)^2 + (\sum_{i=1}^{n} cos\alpha)^2}}{n}$$

The standard deviation is $v = -2\ln(r_1)$.

"The calculated quadrant process is similar to the trigonometric quadrant calculation process. There are four cases to calculate:

1) Where sine is positive and cosine is positive, the mean angle is computed directly.
2) Where sine is positive and cosine is negative, the mean angle = 180 – θ_r.
3) Where sine and cosine are negative, the mean angle = 180 + θ_r.
4) Where sine is negative and cosine is positive, the mean angle = 360 - θ_r." [9]

Exploratory Factor Analysis

Exploratory factor analysis (EFA) is commonly used in the sciences for explaining the variance between several measured variables as a smaller set of latent variables [11]. "Exploratory factor analysis (EFA) is used to determine the number of latent variables that are needed to explain the correlations among a set of observed variables" [11]. "The latent variables are called factors, and the observed variables are referred to as factor indicators" [11]. There will be three basic decision points when using EFA, and they are:

- decide the number of factors,
- choosing an extraction method,
- choosing a rotation method [11].

To perform the first decision point, we have to decide the number of factors, thus we first need to calculate the eigenvalues associated with each factor indicator [11]. "These eigenvalues are produced by a process called principal components analysis (PCA) and represent the variance accounted for by each underlying factor" [11]. "They are not represented by percentages but use itemization scores to total the number of items" [11]. The approach we will use is called the Kaiser-Guttman rule and simply states that the number of factors are equal to the number of factors with eigenvalues greater than 1.0. Next we will discuss the extraction method. "The best evidence in choosing this extraction method is the principal axis factoring with iterated communalities (a.k.a. least squares)" [11]. This extraction method produces factor loadings for every item on every extracted factor [11]. We are interested in our results that will show what is called simple structure, with most items having a large loading on one factor but small loadings on other factors [11]. The last decision point to perform is the rotation method. "Rotation is a

way of maximizing high loadings and minimizing low loadings so that the simplest possible structure is achieved" [11]. The two types of rotation, are orthogonal and oblique. If we use orthogonal, then we are assuming that the factors are uncorrelated with one another [11]. Oblique rotation derives factor loadings based on the assumption that the factors are correlated, and this is probably most likely the case for most measures [11]. In the oblique rotations, we assume that the independent factors are relaxed and the new axes are then free to take any position in the factor space, but the "degree of correlation allowed among factors is, generally small because two highly correlated factors are better interpreted as only one factor" [1].

Correlation

"Correlation analysis is a measure of the relationship or association between two continuous numeric variables that indicates both the direction and degree to which they co-vary with one another, without implying that one is causing the other" [22]. "It refers to the simultaneous change in value of two numerically valued random variables" [22]. The correlation measures the strength of the linear relationship between numerical variables, for instance, the length and width of a sinkhole or the water temperature and atmospheric temperature of a storm being present in the Atlantic Basin. "In these situations the goal is not to use one variable to predict another, but to show the strength of the *linear* relationship that exists between the two numerical variables. Correlation is used to see if linear regression is applicable" [22].

"The strength of linear association between two numerical variables in a population is determined by the correlation coefficient $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$, where σ_x and σ_y are the population standard deviations and σ_{xy} is the population covariance" [22]. "The correlation coefficient $\rho = +/-$, where ρ takes the sign of the slope" [22]. "A ρ value of 1 indicates a perfect positive linear

correlation" [22]. "This happens when the values of both variables increase together and their coordinates on a scatter plot form a straight line" [22]. "A ρ value of -1 indicates a perfect negative linear correlation" [22]. This means when the values of one variable increases while the other variable decreases. Correlation analysis usually measures the extent to which two quantitative variables vary together, including the strength and direction of their relationship [22]. "The strength of the relationship refers to the extent to which one variable predicts the other" [22]. "The direction of the relationship shows whether the two variables vary together directly or inversely" [22]. In a direct relationship, the two variables increase together, whereas in an inverse relationship, one variable tends to decrease while the other increases [22].

Simple Linear Regression

"Simple Linear Regression is the least squares estimator of a linear regression model with a single explanatory variable" [19]. "Simple Linear Regression fits a straight line through a set of points in a way that makes the sum of the squared residuals of the model as small as possible" [19]. This distance can be measured as a value of prediction error, in the sense that it is the discrepancy between the actual value of the response variable and the value predicted by the line. The model under consideration is:

$$y = \beta_0 + \beta_1 x$$

And the observed data is of the form

$$y_i = b_0 + b_1 x_i + \epsilon_i$$

where ϵ is assumed to normally distributed with mean vector 0 and non constant variance. Now f the relationship doesn't have constant variance, the result is that the residuals will reflect this non constant dispersion.

Multiple Linear Regression

Multiple Linear Regression is an extension of simple linear regression. "As a predictive analysis, the multiple linear regression is used to explain the relationship between one continuous dependent variable from two or more independent variables" [14]. The general equation is denoted as: $\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p$,

where $\hat{y}$ is the predicted or expected value of the dependent variable. The predictor (independent) variables are the $x_1, \ldots, x_p$, b_0 is the value of y when all of the independent variables are equal to zero, and the estimated regression coefficients are $b_1, \ldots, b_p$. Note that every regression coefficient represents the change in the dependent variable to a one unit change in the respective independent variable [14]. The Multiple Linear Regression in matrix form is the following:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{p1} \\ 1 & x_{12} & \cdots & x_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \cdots & x_{pn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_p \end{bmatrix}$$

Where X is the design matrix, β is a vector of parameters, ϵ is a error vector, and Y is the response vector. To proceed in finding the normal equations we start by using the following equation to solve for β;

$$X'Y = (X'X)\beta$$

Next, when solving this equation for β, we obtain the least squares solution for $b = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{bmatrix}$.

Lastly, we multiply on the left by the inverse of the matrix $X'X$ and get the following equation for

$$b = (X'X)^{-1}X'Y \text{ [14].}$$

Logistic Regression

"Logistic regression is a type of probabilistic statistical classification model that is used for predicting the outcome of a categorical dependent variable based on one or more predictor variables" [16]. "It estimates the parameters of a qualitative response model" [16]. There will be two levels of logistic regression (binomial and quasinomial or multinomial) used to help address the research statements/questions. In binary logistic regression, the outcome is usually coded as 0 or 1, as this leads to the following;

$$d = \begin{cases} 1, & success \\ 0, & not\ success \end{cases}$$

and the total counts $x = \sum d$, from which we can estimate the relative frequency $\hat{p} = \frac{x}{n}$, and estimate of the probability. Probability can be manipulated to odds.

In logistic regression, there is a logistic transformation of the odds (logit) that will serve as the dependent variable. The odds are denoted as: odds = $\frac{p}{1-p} \in (0, \infty)$

The general model is denoted as

$$logit(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

where p represents the functions parameter as a probability.

Multinomial logistic regression deals with situations where the outcome usually can have three or more possibilities [16]. "In the multinomial logit model we assume that the log - odds of the response follow a linear model $y = logit(p) = \log\left(\frac{p}{1-p}\right) = \alpha_j + x_1\beta_{1j} + \cdots$, where α_j is a constant and β_j is a vector of regression coefficients, for $j = 1, 2, \ldots, J - 1$" [17]. "This model is the similar to a logistic regression model, except that the probability distribution of the response is multinomial instead of binomial and we have $J - 1$ equations instead of one" [17]. "The $J - 1$multinomial logit equations contrast each of categories $1, 2, \ldots, J - 1$ with category J, whereas the single logistic regression equation is a contrast between successes and failures" [17]. "If $J = 2$ the multinomial logit model reduces to the usual logistic regression model" [17]. "We need only J equations to describe a variable with J response categories" [17].

Non – Response Analysis

"Wooten introduced Non-Response Analysis the founding theory in Implicit Regression where Implicit Regression treats the variables implicitly as co-dependent variables and not as an explicit function with dependent/independent variables as in standard regression" [20]. The contribution of this research include an underlying theory to better address co-dependent relationship among measured variables with normal random error, and specifically, detecting constants and inverse relationships with bivariate random error [20].

"Both standard regression and non – response analysis can be used to measure the constant nature of a variable" [20]. "The coefficient of determination, R^2, is the percent of the total sums of squares explained by the mean" [20]. "As the variance approaches $0, R^2 \to 1$, and for uniformly distributed variables, $R^2 \to 0.75$" [20].

In standard regression, we have that the subject response (y) is constant ($\beta = \mu$),

$$y = \beta$$

and that there is random error in the observed data,

$$y_i = \beta + \varepsilon_i.$$

where $E(\varepsilon) = 0$ and $V(\varepsilon) = \sigma_y^2$; and parameter estimate given by

$$\hat{\beta} = \frac{\sum y_i}{n} = \hat{\mu}_y.$$

However, using the non-response model we have that the subject response (y) is a non-zero constant (μ), but instead of minimize the error, rather minimizes the percent error,

$$\frac{y - \mu}{\mu} = \alpha y - 1;$$

or equivalently, modeling

$$\alpha y = 1$$

where the random error that exist is related to the coefficient of variation,(CV); the ratio of standard deviation to the mean over the mean alone

$$\alpha y_i = 1 + \omega_i,$$

where $E(\omega) = 0$ and $V(\omega) = CV_y^2 = \frac{\sigma_y^2}{\mu_y^2}$; and parameter estimate given by

$\hat{\mu}_y = \frac{1}{\hat{\alpha}} = \frac{\sum y^2}{\sum y}$, a self-weighting mean [20]. Both of these point estimates yield a coefficient of determination given by $R^2 = \frac{n\bar{x}^2}{\sum x^2}$.

Non –response analysis can be extended to bivariate and multivariate analysis.

Non-response Analysis is testing the constants coefficient effects on the other terms [21]. Consider the model $z = \beta_0 + \beta_1 x + \beta_2 y + \beta_3 xy + \cdots$ where z is unobserved but assumed to be normally distributed. The alias matrix,A, is such that the expected value of the beta coefficients and the constant coefficient are related as follows

$$E\left(\begin{bmatrix}\beta_1\\ \beta_2\\ \vdots\end{bmatrix}\right) = \begin{bmatrix}\beta_1\\ \beta_2\\ \vdots\end{bmatrix} + A\beta_0$$

measuring the bias the constant intercept has on all the remaining parameters is:

$$A = (X_1'X_1)^{-1}X_1'X_2,$$

where $X_1 = \begin{bmatrix} x_1 & y_1 & x_1y_1 & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ x_n & y_n & x_ny_n & \cdots \end{bmatrix}$ and $X_2 = \begin{bmatrix}1\\ \vdots\\ 1\end{bmatrix}$; that is, this view shows the bias introduced by the constant and is equivalent to testing

$$1 = \alpha_1 x + \alpha_2 y + \alpha_3 xy + \alpha_4 x^2 + \alpha_5 y^2,$$

as $E(\hat{\alpha}) = \alpha$ and and $V(\hat{\alpha}) = \sigma^2(X_1'X_1)^{-1}$.

"In general, non – response analysis can model any functional or non – functional relationship of the form $1 = h_\theta\ (x_1, x_2, \dots, x_p)$, where θ is the set of parameter coefficients" [20]. "This can be further extended to implicit regression, which models relationships of the form $g(x_1, x_2, \dots x_p) = h_\theta\ (x_1, x_2, \dots, x_p)$, where the set of terms in the expressions g and h are

mutually exclusive" [20]. In standard regression, $SST = SSM + SSE$ (SS is the sum of squares, T, M, E are total, model, and error respectively) the angle θ_T (angle of separation) between the M and E in the vector space is 90°. However, as the assumption of independents is not satisfied and the degree of separation, θ_T is not guaranteed to be 90°; hence, we invoke the law of cosines to measure θ_T,

$\theta_T = \arccos\left(\frac{SSM+SSE-SST}{2\sqrt{SSM \times SSE}}\right)$ [20].

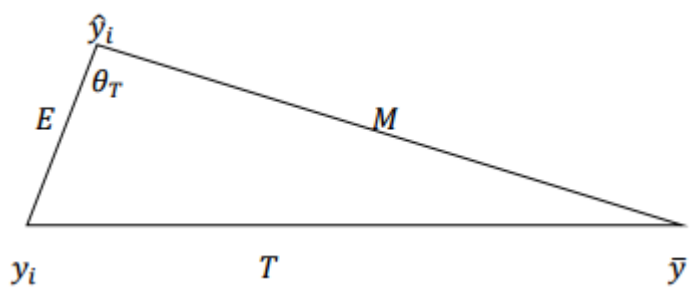

Figure 1.1: Illustration of the Angle of Separation

The height or extent to which the estimates are removed from the data and the mean is given by $h = \hat{E}Sin(\theta_M)$, where $\hat{E} = \sqrt{\frac{SSE}{n}}$. A good model should have an angle close to 90° with height h, close to the ratio $\frac{ME}{T}$; that is, in a right triangle, $ratio = \frac{hT}{ME} = 1$ which be estimated using $ratio = \sqrt{\frac{SST}{SSM}} Sin\theta_M$. The closer this ratio is to one and the closer the degree of separation, θ_T, is to 90°, the better the developed model teases out the true relationship among the measured variables [20].

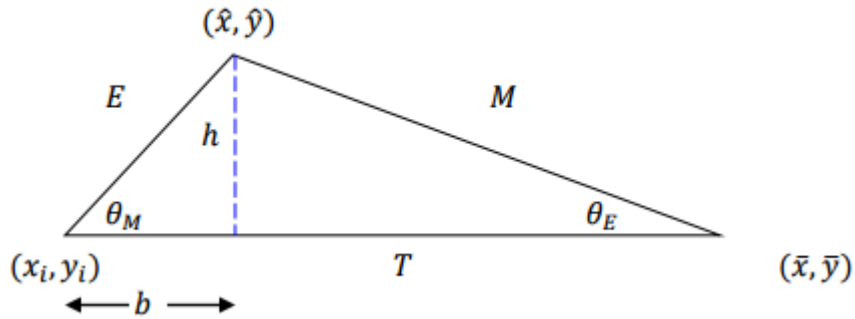

Figure 1.2: Distance Removed (Height)

Forward Selection, Backward Elimination and Subset Analysis

There are three selection methods that one would use in order to develop a statistical model, and they are **Forward selection, Backward selection (or elimination)** and **Subset analysis**. "In forward selection one starts with the best one variable model, where that variable has the highest simple correlation with the response variable, then the second variable is picked that gives the maximum improvement in fit" [4]. "This is revealed by the maximum of partial correlations of all independent variables with the response variable" [4]. "Then keep adding variables until no additions provide adequate reduction in the error mean square as stated by the p-value or the process can be continued until variables are included" [4].

The second selection method is backward selection (or elimination). "In this selection method one starts by considering the full model, which includes all candidate variables" [4]. The variable that contributes least to the model is deleted [4]. The coefficients for the remaining (m - 1) variable model are examined and the variable contributing the least is eliminated [4]. "The process is repeated and then end when all the rest of the variables are contributing at a preset level of significance" [4].

The third selection method is subset analysis. First, all of the models that have one predictor variable are included and checked, then the two models with the highest R^2 are selected. Next, all models with two predictor variables are included and checked, models with the highest R^2 are selected. One will continue to estimate all combinations containing two variables at a time, then three at a time, etc. Then choose a subset that has the most table set of independent variables [4].

Survival Analysis

"Survival analysis is typically defined as a set of methods for analyzing data where the outcome variable is the time until the occurrence of an event of interest" [6]. "The time to event or survival time can be measured in days, weeks, years, etc" [6]. "In survival analysis, subjects are generally followed over a specified time period and the focus is on the time at which the event of interest occurs" [6]. One can estimate two functions that are dependent on time, the survival and hazard functions. "The survival function is denoted as $S(t) = P(T > t)$. This gives, for every time, the probability of surviving or not experiencing the event up to that time" [6]. "The hazard function, $h(t) = -\frac{S\prime(t)}{S(t)}$, gives the potential that the event will occur, per time unit, given that an individual has survived up to the specified time [6]. We will be using the Kaplan Meier method, which is a nonparametric estimator of the survival function, is widely used to estimate and graph survival probabilities as a function of time [6]. "The regression model for the analysis of survival data is the Cox proportional hazards regression model" [6]. "It allows testing for differences in survival times of two or more groups of interest, while allowing to adjust for covariates of interest" [6]. This regression model is a semi parametric model, making fewer assumptions than typical parametric methods but more assumptions than those nonparametric methods [6].

The product limit or PL method of Kaplan and Meier is used to estimate S:

$$\hat{S}(t) = \prod_{t_i < t} 1 - \frac{d_i}{n_i}$$

where t_i is the duration of study at point i, d_i is the

number of sinkholes up to point i and n_i is the number of possible sinkholes at risk just prior to t_i.

The Cox proportional hazards (Cox PH) model fits survival data with covariates z to a hazard function of the form $h(t|z) = h_0(t)\exp\{\beta' z\}$,where β is an unknown vector and $h_0(t)$ is the baseline hazard, which is nonparametric [6]. The analytical model is denoted as:

$$h_i(t) = h_0 \exp\left(\beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_{13} + \beta_4 x_{14} + \beta_5 x_{15}\right)$$

The hazard ratio is denoted as:

$$\text{HR} = \frac{h_i(t)}{h_j(t)} = \frac{h_0(t)\exp\left(\beta_1 x_{1i} + \cdots + \beta_k x_{ik}\right)}{h_0(t)\exp\left(\beta_1 x_{1j} + \cdots + \beta_k x_{jk}\right)}$$

CHAPTER 2: LATENT STORM FACTORS AND THEIR INDICATORS & NON – RESPONSE FACTOR MODELING OF HURRICANES

The Atlantic Basin is our neighbor, so hurricane season is not a new concept to understand. In this chapter, we will investigate the month, day of month, hour of the day, starting latitude and longitude, latitude, longitude, pressure, wind speed and maximum wind speed of the storms in the Atlantic Basin (1975 – 2014) to see if there is a statistically significant reasoning of the formulation of these storms. More specifically, we will interpret the latent storm factors that describe the correlation amongst the Atlantic Basin storm indicators.

EXPLORATORY FACTOR ANALYSIS

In our data set; the observed variables (factor indicators) that we are interested in seeing if there is a correlation between are: **month, day, hour, starting latitude, starting longitude, latitude, longitude, pressure, minimum pressure, wind speed, and maximum wind speed.**

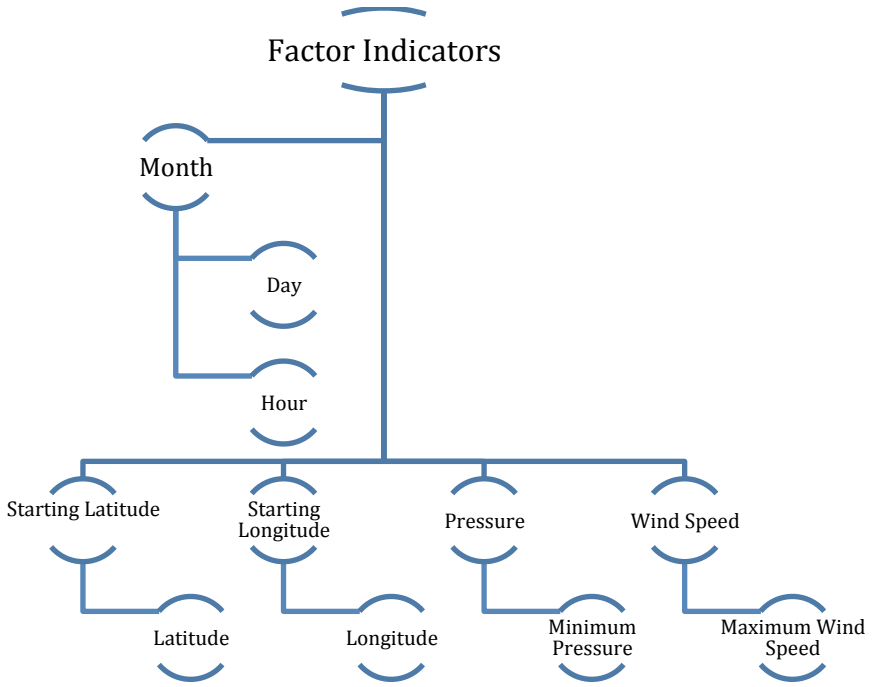

Figure 2.1: Factor Indicators

We are interested in determining the latent storm factors that explain variance and measure the correlation that exist between their respective storm indicators. To decide the number of factors, we first need to calculate the eigenvalues associated to each factor indicator. They are not represented by percentages but scores that total to the number of items. For example; an 11-item scale will theoretically have 11 possible underlying factors, each factor will have an eigenvalue that indicates the amount of variation in the items accounted for by each factor. In our analysis, the first factor has an eigenvalue of 3.0, it accounts for 27% of the variance (3/11=.27). The total of all the eigenvalues is 11, since there are 11 items, so some factors will have smaller eigenvalues. After our calculating, we found that there were three factors with eigenvalues that were greater than 1.0. The first factor had an eigenvalue of 3.0, it accounts for 27% of the variance (3/11=.27). The second factor had an eigenvalue of 2.0, it accounts for 18% of the variance (2/11=.18). The third factor had an eigenvalue of 4.0, it accounts for 36% of the

variance (4/11=.36). The next thing to do is to choose an appropriate extraction method. In the following Table, we show the factor loadings of our factor indicators. The notation of the 3 factors are PA1, PA2, and PA3. The factor indicator minimum pressure was 0 throughout all of the factor loadings, and so it was removed from the following Table. However, since it was one of the original factor indicators, then it will still be used for the eigenvalue itemization process.

Table 2.1: Factor Loadings of the Factor Indicators

	PA1	PA2	PA3
Month	0.83	0.01	0.05
Day	0.81	-0.02	0.01
Hour	0.75	0.03	-0.08
Starting Latitude	-0.07	0.87	0.01
Starting Longitude	0.07	0.94	0.05
Latitude	0.13	0.85	-0.03
Longitude	0.12	0.92	-0.01
Pressure	0.02	0.04	0.77
Wind Speed	0.06	-0.02	0.84
Max Wind Speed	-0.01	0.01	0.82

By looking at our factor loadings, we can begin to assess our factor solution. We can see that month, day, and hour all have high factor loadings beginning with 0.75 on the first factor (PA1). Therefore, we might call this factor PA1, *calendar* and consider it representative of the time of year a storm is present. Similarly, starting latitude, starting longitude, latitude, and longitude

load highly on the second factor (PA2), which we may consider calling this factor *locations.* Notice that latitude and longitude have a lower loading on the second factor (PA2) than starting latitude and longitude, but they had a slight loading on the first factor (PA1). This could suggest that latitude and longitude is less related to *locations* than starting latitude and longitude. Lastly, pressure, wind speed and maximum wind speed all have high factor loadings with a 0.77 on the third factor (PA3).

Thus, we might want to call this factor PA3, *atmospheric*. In the Table below, we can see that each factor had a different accountability of the variance in responses. The first factor *calendar* had a 27% of the variance in responses, the second factor *location*s had an 18% of the variance in responses, and the third factor *atmospheric* had a 36% of the variance in responses. This leads to a factor solution that accounted for 81% of the total variance among the month, day, hour, the starting latitude, starting longitude, latitude, longitude, pressure, wind speed, and the maximum wind speed.

Table 2.2: Variances Explained from EFA

Variances Explained from EFA.	PA1(Calendar)	PA2 (Locations)	PA3(Atmospheric)
Proportion Variance	0.27	0.18	0.36
Cumulative Variance	0.27	0.45	0.81

In Table 2.3, the correlation of the storm factor indicators with factors is 94% in the first factor, 88% in the second factor and 96% in the third factor. This could suggest that there is a

higher correlation within the third factor *atmospheric* for when a storm is present. The multiple R – squared values with factors is much higher for the third factor *atmospheric* than it is for the first factor *calendar*. With a R^2 value of 95% versus a R^2 value of 91%, the factor indicators for the factor *atmospheric,* have a much better explanation of the relationship among the measured variables.

Something else to consider is the minimum correlation of possible factor indicators. For the first factor *calendar*, this value is 0.54, for the second factor *locations*, the value is 0.74, and for the third factor *atmospheric*, the value is 0.87. This means that if we only considered the factor calendar with its factor indicators, than only 54% of any correlation between the factor indicators could be explained. Whereas in the second factor, at least 74% of any correlation between the factor indicators could be explained. Most importantly, in the third factor, 87% of any correlation between the factor indicators can be explained.

Table 2.3: Factor Indicator Correlations

Correlation of Factor Indicators with Factors	0.94	0.88	0.96
Multiple R-Squared with Factors	0.91	0.86	0.95
Minimum Correlation of Possible Factor Indicators	0.54	0.74	0.87

Next, we will choose a rotation method to determine how much our factors are correlated. In the following Table, notice that the three factors are correlated at a value of 0.26. We are looking for a very low correlation value between all of the factors. Thus, a correlation of 0.26

indicates that there may be too many factors. Next, we will go through another EFA process using the first factor's indicators, the starting latitude and longitude from the second factor, and pressure and the wind speed from the third factor to and we need to find the correlation between them.

Table 2.4: Factor Correlations

	PA1 (Calendar)	PA2 (Locations)	PA3 (Atmospheric)
PA1 (Calendar)	1.00	0.26	0.26
PA2 (Locations)	0.26	1.00	0.26
PA3 (Atmospheric)	0.26	0.26	1.00

The new observed variables (factor indicators) that we are interested in seeing if there is a correlation between are: **month, day, hour, starting latitude, starting longitude, pressure and wind speed (denoted as wind), interactions between month and starting latitude and longitude, and the interaction between starting latitude and starting longitude.**

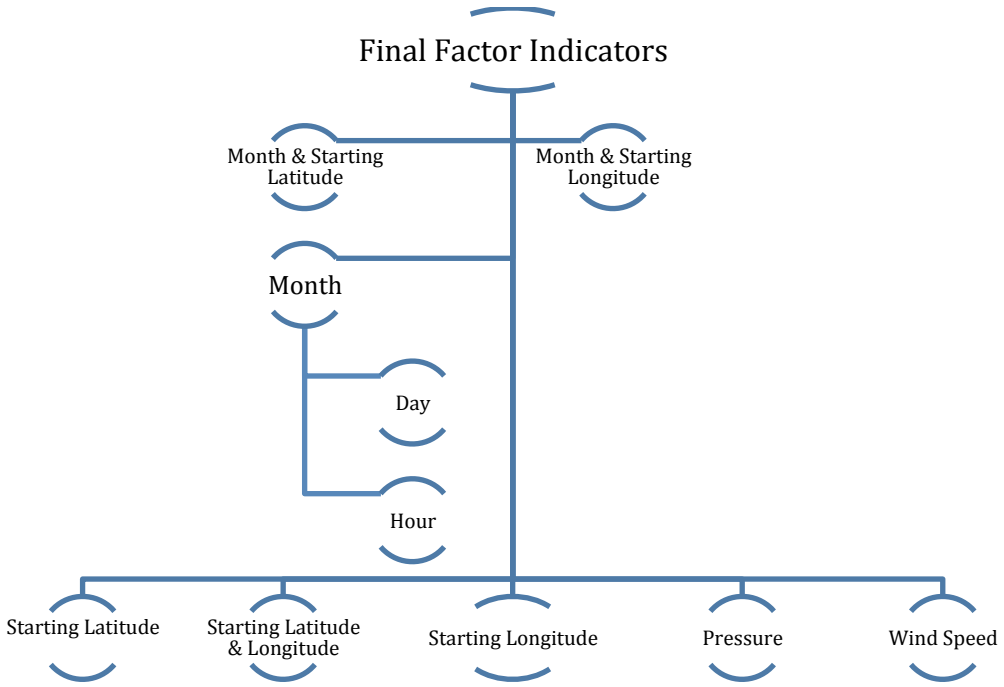

Figure 2.2: Second EFA Factor Indicators

In this second EFA process, we will assume that the appropriate number of factors will be determined to be 2, since the first EFA process with 3 factors had factor indicators that loaded high on some factors and low on another factor. We will use the principal axis factoring for our extraction method. In the following Table, the calculated factor loadings are shown.

Table 2.5: Factor Loadings

	PA1	PA2
Month	0.84	0.03
Day	0.77	0.04
Hour	0.82	0.05
Starting Latitude	-0.06	0.79
Starting Longitude	0.02	0.91
Pressure	0.15	0.93
Wind Speed	-0.05	0.96
Month & Starting Latitude	0.07	0.78
Month & Starting Longitude	0.01	0.77
Starting Latitude, Longitude	-0.05	0.95

Looking at our factor loadings, we can begin to assess our factor solution. We can see that month, day, and hour all have high factor loadings beginning with 0.77 on the first factor (PA1). Therefore, we might call this factor PA1, *calendar* and consider it representative of the time of year a storm is present. Similarly, starting latitude, starting longitude, pressure, wind speed, month & starting latitude, month & starting longitude, and starting latitude & longitude load highly on the second factor (PA2), which we may consider calling this factor *cal-location atmospheric.* Notice that pressure has a lower loading on the second factor (PA2) than starting latitude and longitude, wind speed, month & starting latitude, month & starting longitude, and starting latitude & longitude but it had a slight loading on the first factor (PA1). This could

suggest that pressure is less related to *location atmospheric* than starting latitude and longitude, wind speed, month & starting latitude, month & starting longitude, and starting latitude & longitude. In the Table below, (Table 2.6), we can see that each factor accounted around 48% of the variance in responses, leading to a factor solution that accounted for 94% of the total variance in when a storm is present based off the month, day, and hour of that day, the starting latitude, starting longitude, pressure and the wind speed.

Table 2.6: Variances Explained from EFA

	PA1 (Calendar)	PA2 (Cal-Location Atmospheric)
Proportion Variance	0.48	0.49
Cumulative Variance	0.48	0.97

In Table 2.7, the correlation of the storm factor indicators with factors is 93% in the first factor and 98% in the second factor. This could suggest that there is a higher correlation within the second factor *cal-location atmospheric* for locating when a storm could be present. Notice that the multiple R – squared values with factors is much higher for the second factor *location atmospheric* than it is for the first factor *calendar*. With an R^2 value of 96% versus an R^2 value of 85%, the factor indicators for the factor *cal-location atmospheric,* has a much better explanation of the probable conditions of when a storm is present. Something else to consider is the minimum correlation of possible factor indicators. For the first factor *calendar*, this value is 0.74, and for the second factor *cal-location atmospheric*, the value is 0.88. This means that if we only considered the factor calendar with its factor indicators, than only 78% of any correlation

between the factor indicators could be explained. Whereas in the second factor, at least 86% of any correlation between the factor indicators could be explained.

Table 2.7: Factor Indicator Correlations

Correlation of Factor Indicators with Factors	0.93	0.98
Multiple R-Squared with Factors	0.85	0.96
Minimum Correlation of Possible Factor Indicators	0.74	0.88

Next, we will choose an oblique rotation method to determine how much our factors are correlated. In the following Table, notice that the two factors are correlated at a value of 0.137. This means that the two factors *calendar* and *cal-location atmospheric* are 13.7% correlated. This correlation value is much smaller than the previous EFA process, where the correlation value was 0.26. This smaller correlation value indicates that the two factors *calendar* and *cal-location atmospheric* are better correlated.

Table 2.8: Factor Correlations

	PA1 (Calendar)	PA2 (Cal-Location Atmospheric)
PA1 (Calendar)	1.00	0.137
PA2 (Cal-Location Atmospheric)	0.137	1.00

Based off the factor *calendar,* we can conclude that the month, day, and hour have a fairly large influence as to when a storm is present in the Atlantic Basin. Now looking at the second factor that we referred to as *cal-location atmospheric*, the starting latitude, longitude, pressure, month & starting latitude, month & starting longitude, and starting latitude & longitude and the wind speed are also an influence when a storm is present in the Atlantic Basin. Through the use of EFA, we were able to simplify the situation by looking at variables that could be correlated within groups. By looking at these variables that were correlated, we can detect that in the month of the year and the day of that month, there is a correlation as to when a storm may be present in the Atlantic Basin. The pressure and wind speed were a big part of this higher correlation between the two factors. This is due to the fact from the first EFA analysis that pressure and wind speed were in the factor atmospheric, which had the highest correlations of its factor indicators. The starting latitude and longitude were much better correlated in this second EFA analysis because the latitude and longitude were not as highly correlated in the first EFA analysis. As far as the month & starting latitude, month & starting longitude, and starting latitude & longitude interaction terms, they were loaded high on the second factor also, which could indicate that the month of the year and the starting locations could be correlated in determining when a storm is present in the Atlantic Basin.

In determining the latent storm factor measures, we have verified through the process of EFA that there are two factors that are correlated to describe storm formation indicators. These two factors are *calendar* and *cal-location atmospheric*. The factor *calendar* has the storm indicators of which month, day and hour of that day, can help us determine when a storm may form in the Atlantic Basin. The second factor *cal-location atmospheric* has the storm indicators of the month, possible starting locations, and the pressure and wind speed that a storm may possess in

order to form in the Atlantic Basin. If we wanted to have a model with only two variables, then we would use the two factors *calendar* and *cal-location atmospheric*. The correlation using the factor indicators is 93% for the first factor and 98% for the second factor. Next, we will use our results from the EFA process to build our model. Standard regression will be used to build our model using the factors and their respective factor indicators. The factor indicators that we would include in our future model are **month (x_1), day (x_2), hour (x_3), starting latitude (x_4), starting longitude (x_5), pressure (x_6),** and the response variable is **wind speed.** In standard regression we use the general model of the following:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \varepsilon$$

In this section we will use the analytic model with its variables denoted as:

$$y = \beta_0 \ + \sum_{\forall i,j} \beta_i x_i^{a_i} x_j^{a_j} \ , i \neq j \ a_* \in \{0,1\}$$

There were a total of 16 terms in the above model. Out of these 16 terms only 5 were significantly contributing of at least a 1% significance level. In this first developed model,

$$y = \beta_0 + \ \beta_1 x_1 + \ \beta_6 x_6 + \ \beta_7 x_1 x_4 + \ \beta_8 x_1 x_5 + \ \beta_9 x_4 \, x_5$$

$$\hat{y} = 0.725 - 0.021 x_1 + \ -7.450 x_6 + \ 0.008 x_1 x_4 + \ 5.36 x_1 x_5 + \ 0.005_9 x_4 \, x_5$$

This model shows us that the month, pressure, and the starting locations have a significant impact on where a storm may be present in the Atlantic Basin. In the following Figure, the bar graph for month shows that there are a lot of storms that occur between the months of August and September.

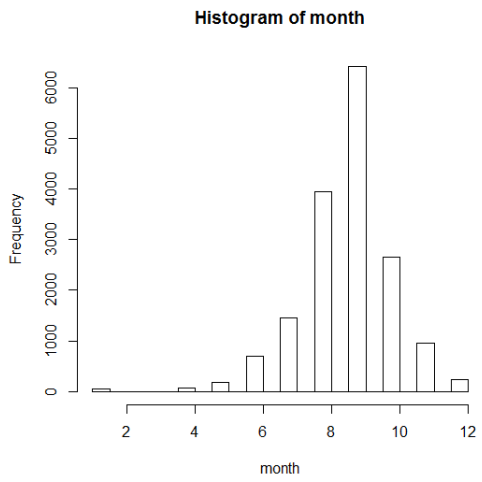

Figure 2.3: Bar Graph of the Variable Month

The model indicates that the starting locations and the month have a correlation as to when a storm may be present in the Atlantic Basin. Notice that between the months of August and September there are more storms than any of the other months. In both of the scatterplots, the starting locations in the Atlantic Basin during the months of August and September are stronger than starting locations for storms in any other month of the year. This can be seen in the following scatterplot.

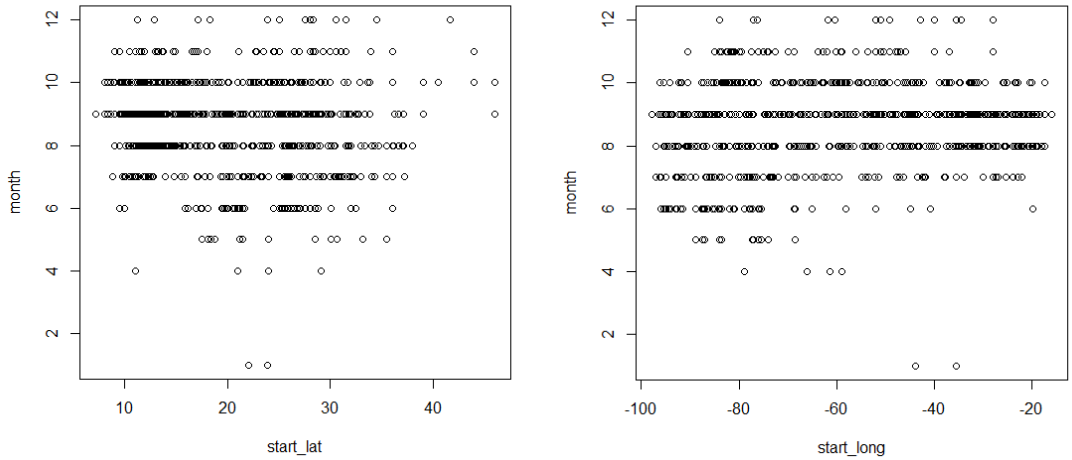

Figure 2.4: Scatterplots of Starting Latitude, Starting Longitude and Month

The above Figure shows that the starting locations will have a significant correlation when a storm may have a higher frequency in between the months of August and September. Although our standard regression model had only 5 explanatory variables, the correlation between the explanatory variables was slightly significant with an $R^2 = 0.82$ and had an adjusted r-squared value of 0.81. From our EFA process we found that the two factors *calendar* and *cal-location atmospheric* were highly correlated with a low oblique rotation value of 0.137. Considering that the first factor *calendar* had the factor indicators month, day, and hour, it can be presumed that since month had the highest factor loading, then month would be the variable that may be kept to be put in future models. Looking back at the second factor *cal-location atmospheric*, all of the factor loadings were high in this factor. Thus it makes sense to keep the variables of starting latitude, starting longitude, pressure, wind speed, and the interactions between month & starting latitude, month & starting longitude, and starting latitude and longitude in future models. Next, we will compare our results with another statistical method known as Non-Response Analysis.

Recall that our second EFA process, we determined that the second factor *cal-location atmospheric* and its factor indicators were highly correlated. From our standard regression analysis in the previous section, we found that the month, the pressure and the locations were the explanatory variables that best described the response variable. Since we have determined that the time between August and September are the best indicators of high volume activity of when a storm is present, then the variable month will also be considered in this non – response analysis comparison between standard regression. However, since the wind speed was the response variable in the standard regression model, then it will not be used in our non-response analysis model as the response variable but as a predictor variable.

Non – Response Analysis Model

In this section, we are interested in determining if there is a correlation between the predictor variables: **month (x_1), starting latitude (x_2), starting longitude (x_3), pressure (x_4), wind speed (y)** and the **interactions of month & starting latitude, month & starting longitude, and starting latitude and longitude.** Consider the following analytic model;

$$1 = \alpha_1 x_1 + \alpha_2 x_4 + \alpha_3 x_1\, x_2 + \alpha_4 x_1 x_3 + \alpha_5 x_2 x_3 + \alpha_6 y$$

This model had every predictor variable result in a 1% level of significance (using standard t-test). Although this model is not the standard regression model, it held a $R^2 = 1$ value. The developed model is;

$$\hat{1} = 0.002x_1 + 0.003x_4 + 0.003x_1x_2 + 0.006x_1x_3 + 0.007x_2x_3 - 0.009y$$

In the following Figure, the predicted model for standard regression and our non – response model is given.

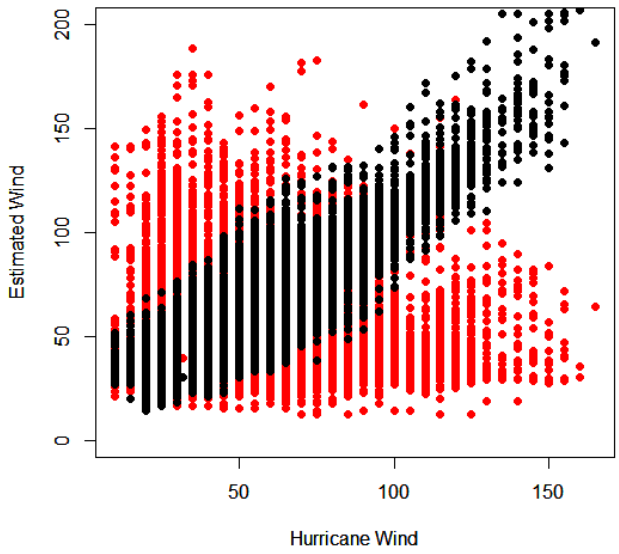

Figure 2.5: Predicted Model using Standard Regression (red) & Non – Response Analysis (black)

Comparatively speaking, our model fits better than the standard regression model that was used in the previous section. In the following Figure, the non – response model is shown, showing the equilibrium in the system.

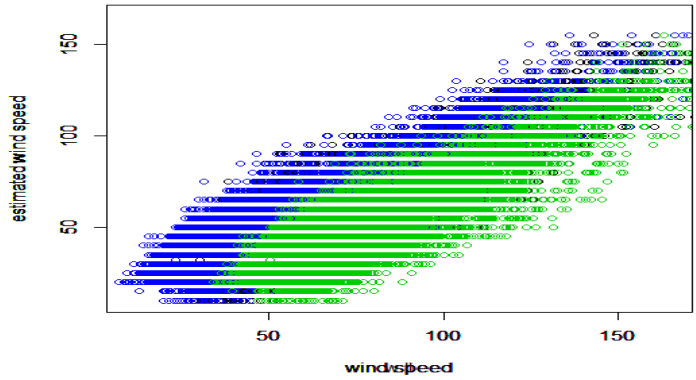

Figure 2.6: Non – Response Model Fitting

Since the assumption of independence is not required, when solving for $\hat{y}$, the error terms are no longer perpendicular to the mean but rather is given by $\theta_T = \arccos\left(\frac{SSM+SSE-SST}{2\sqrt{SSM \times SSE}}\right)$ [20].

In comparison to the standard regression measured angle of 90°, our non – response model had a degree of separation of $\theta_T = 76.4$, including all the terms. The height h, in the non – response model had a value of 0.79. Since h is the distance between the point estimates and the line between the data and the means, then the lower the height the better the model will fit. We can conclude that the non – response model was the best fitted model. In the next section we will investigate a smaller subset of the hurricanes that have hit the Florida Keys in the last 100 years.

EXPLORATORY FACTOR ANALYSIS & NON – RESPONSE MODELING ON THE FLORIDA KEYS.

In this section, we will present a statistical survey of the major hurricanes that have hit the Florida Keys using an exploratory factor analysis approach to constructing a non-response analysis model. In our data set; the observed variables (factor indicators) that we are interested in seeing if there is a correlation between are: year, **month, day, hour, starting latitude, starting longitude, latitude, longitude, pressure,** and **wind speed.**

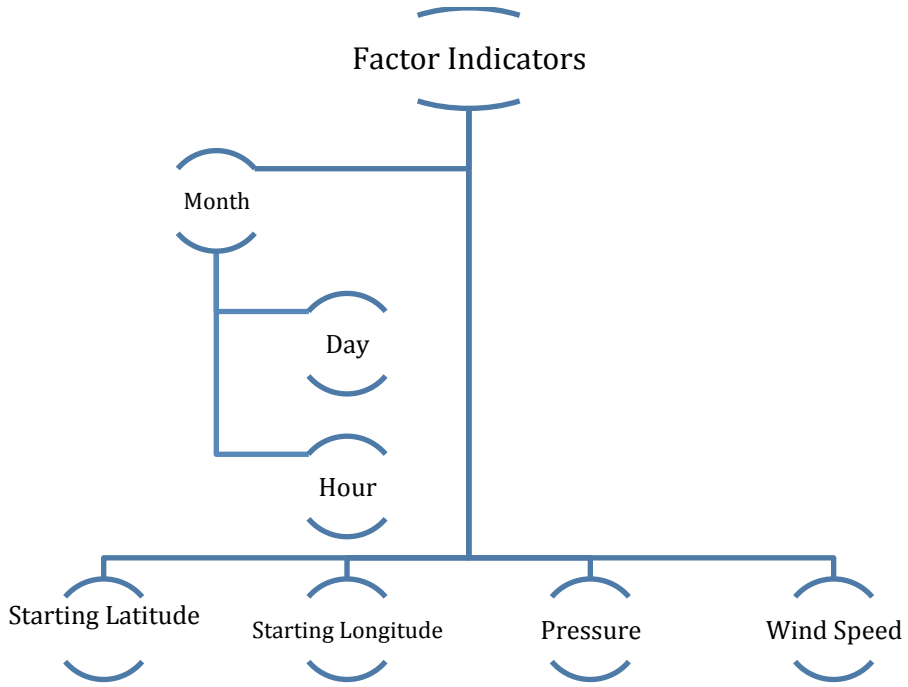

Figure 2.7: Variables of Interest – Florida Keys

Before we begin the analysis of addressing our research questions, let us give a brief history of the catastrophic or disastrous hurricanes that have hit the Florida Keys from 1900 – 2000. In 1919, Key West, FL was hit by the most powerful hurricane in its history at that time. The Labor Day storm hit the Florida Keys in 1935. From 1950 to 2000, the most intense hurricanes to hit the Florida Keys were in 1960, 1965 and 1992. These hurricanes were Hurricane Donna (nicknamed Deadly Donna), Hurricane Betsy (1965), and Hurricane Andrew (1992).

In this section we are only interested in the catastrophic or disastrous hurricanes that have hit the Florida Keys in the last 100 years (1900 – 2000) to see if we can produce a statistical model in helping to produce inferences as to when a super storm may hit the Florida Keys.

In our data set; the observed variables (factor indicators) that we are interested in seeing if there is a correlation between are: **month, day, hour, starting latitude, starting longitude, pressure and wind speed (denoted as wind).** In the following Table, we show the factor loadings of our

factor indicators. In this paper, we will assume that the appropriate number of factors will be determined to be 2, for a more general analysis.

Table 2.9: Factor Loadings of the Factor Indicators – Florida Keys

	PA1	PA2
Month	0.86	0.03
Day	0.91	0.07
Hour	0.84	-0.02
Starting Latitude	-0.06	0.93
Starting Longitude	-0.96	0.92
Pressure	0.06	0.81
Wind Speed	-0.34	0.85

By looking at our factor loadings, we can begin to assess our factor solution. We can see that month, day and hour all have high factor loadings beginning with 0.84 on the first factor (PA1). Therefore, we might call this factor PA1, *calendar* and consider it representative of the time of year a catastrophic or disastrous storm that has hit the Florida Keys. Similarly, starting latitude and longitude, pressure and wind speed, load highly on the second factor (PA2), which we may consider calling this factor environmental. In Table 2.10, we can see that each factor accounted for around 37% of the variance in responses, leading to a factor solution that accounted for 71% of the total variance in when a storm may become catastrophic or disastrous based off the month, day, hour, the starting latitude, starting longitude, pressure and wind speed.

Table 2.10: Variances Explained from EFA

	PA1 (Calendar)	PA2 (Environmental)
Proportion Variance	0.37	0.34
Cumulative Variance	0.37	0.71

In Table 2.11, the correlation of the storm factor indicators with factors is 96% in the first factor, and 89% in the second factor. This could suggest that there is a higher correlation within the first factor *calendar* for when a storm is present. The multiple R – squared values with factors is much higher for the second factor *environmental* than it is for the first factor *calendar*. With a R^2 value of 90% versus a R^2 value of 87%, the factor indicators for the factor *environmental,* have a much better explanation of the relationship among the measured variables. Something else to consider is the minimum correlation of possible factor indicators. For the first factor *calendar*, this value is 0.77, for the second factor *environmental*, the value is 0.83, and for the third factor *atmospheric*, the value is 0.87. This means that if we only considered the factor calendar with its factor indicators, than only 77% of any correlation between the factor indicators could be explained. Whereas in the second factor, at least 83% of any correlation between the factor indicators could be explained.

Table 2.11: Factor Indicator Correlations

Correlation of Factor Indicators with Factors	0.96	0.89
Multiple R-Squared with Factors	0.90	0.87
Minimum Correlation of Possible Factor Indicators	0.77	0.83

Next, we will choose an oblique rotation method to determine how much our factors are correlated. In the following Table, notice that the two factors are correlated at a value of 0.156. This means that the two factors *calendar* and *environmental* are 15.6% correlated. This smaller correlation value indicates that the two factors *calendar* and *environmental* are better correlated.

Table 2.12: Factor Correlations

	PA1 (Calendar)	PA2 (Environmental)
PA1 (Calendar)	1.00	0.156
PA2 (Environmental)	0.156	1.00

Based off the factor *calendar,* we can conclude that the month, day, and hour have a fairly large influence as to when a storm is present in the Atlantic Basin. Now looking at the second factor that we referred to as *environmental*, the starting latitude, longitude, pressure and the wind speed are also an influence when a storm is present in the Atlantic Basin. Through the use of EFA, we were able to simplify the situation by looking at variables that could be correlated within groups.

Non – Response Analysis Model

In this section, we are interested in determining if there is a correlation between the predictor variables: **month (x_1), day (x_2), hour (x_3), starting latitude (x_4), starting longitude (x_5), and pressure (x_6).** Consider the following analytic model;

$$1 = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \alpha_4 x_4 + \alpha_5 x_5 + \alpha_6 x_6$$

This model had every predictor variable result in a 1% level of significance (using standard t-test). Although this model is not the standard regression model, it held a $R^2 = 0.92$ value. This indicates that the model explains 92% all the variability of the response data around its mean. The developed model is;

$$\hat{1} = 0.265x_1 + 0.23x_2 + .0687x_3 + 0.415x_4 - 0.011x_5 + 0.342x_6$$

Since the assumption of independence is not required, when solving for $\hat{x}$ and $\hat{y}$, the error terms are no longer perpendicular to the mean but rather is given by $\theta_T = \arccos\left(\frac{SSM+SSE-SST}{2\sqrt{SSM \times SSE}}\right)$ [20].

Our non – response model had a degree of separation of $\theta_T = 85.1$, including all the terms. The height h, in the non – response model had a value of 0.84. Since h is the distance between the point estimates and the line between the data and the means, then the lower the height the better the model will fit. We can conclude that the non – response model was the best fitted model.

USEFULNESS & CONTRIBUTIONS

The results in this study are useful for numerous reasons, for instance this is the first time that exploratory factor analysis to build a statistical model for hurricane related variables. This is the first time that Non – Response Analysis has been used in conjunction with Exploratory Factor Analysis to develop a statistical model. This analysis shows how well exploratory factor analysis determine the latent storm factors that explain variance and measure the correlation that exist between their respective storm indicators. Non-response analysis was used in this chapter as a comparative theory to standard regression for the statistical modeling of hurricanes in the Atlantic Basin. Furthermore, exploratory factor analysis was used to find the correlated between

the observed variables in both the larger hurricane data set, as well as the smaller Florida Keys data set. These methods combined can be useful to create simple structures for statistical models, including codependent relationships.

CHAPTER 3: LOGISTIC REGRESSION OF HURRICANES IN THE ATLANTIC BASIN FROM 1990 – 2014

INTRODUCTION TO THE DATA

Hurricane and Buoy Data

Big Data refers to any collection of data sets that are large or complex; and that may often become difficult to process with traditional statistical software. The data that was compiled into a larger data set came from two different sources. First, the hurricane data for the years 1990 – 2014 came from Unisys Weather site (Atlantic Basin Hurricanes data) and the second data set came from the National Buoy Center (for the years 1990 – 2014). We will start with the first data set and describe the structure of the hurricanes.

The variables from the Unisys Weather site are: **Year, Month, Day, Hour, Storm, Name, Latitude, Longitude, Wind Speed** (knots), **Pressure** (milliards).

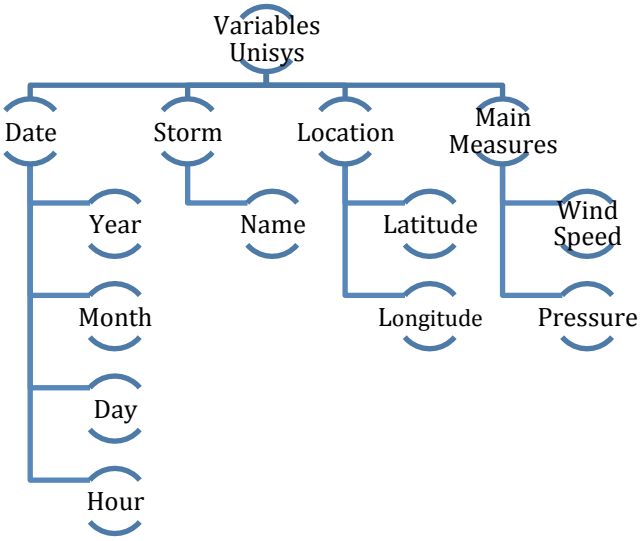

Figure 3.1: Variables from the Unisys Weather Site.

Compilation of Hurricanes

In this case, the files are organized by the decade. The information on these files was space delimited, thus this required that additional steps in creating the data set. The data from the Unisys Weather Site was arranged in decades. One particular statistical software that is capable of reading these types of data files, is the program R. We read in the data by each decade (from the website), then spliced the data from its single column (containing the information) into ten columns and wrote into a CSV file. Next, we created a timeline that would be used to fill in the gaps of the missing data. The timeline variable is denoted as:

$timeline[i] = date[i] - date[0]$, where $date[0]$ is January 1, 1990, and

where i is a particular date.

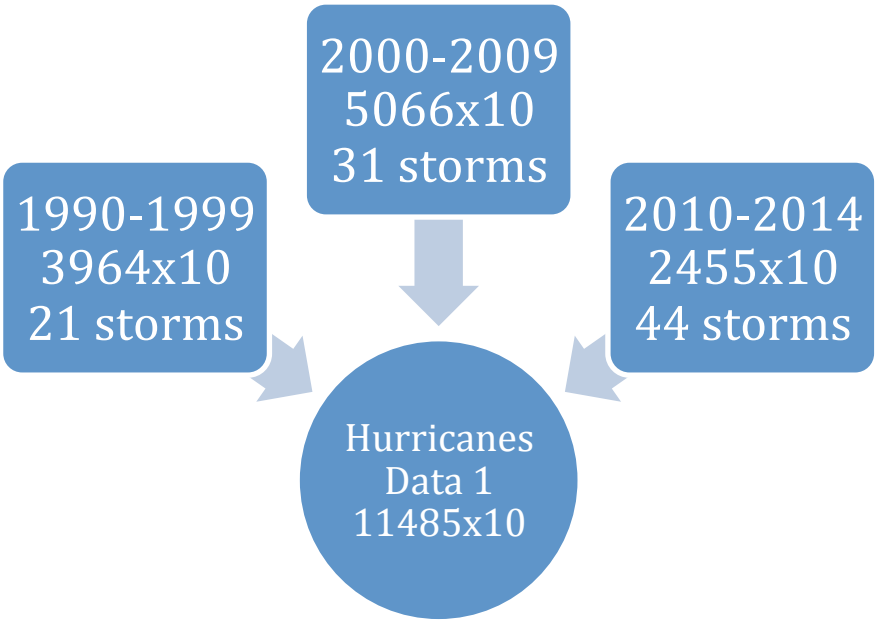

Figure 3.2: Hurricane Data 1990 – 2014 from Unisys

In Data Set 2 (the smaller hurricane file), we read in each storm (or observation) by the year. Then it was compiled and written to a CSV file to be spliced and broken down in Excel.

In Figure 3.3, we open the csv file for the smaller hurricane file while using Excel to splice, we add the headings and pinpoint the variables. In Figure 3.4, we extracted a list of storms for the given years, and added a year to the list (this creates Data Set 3).

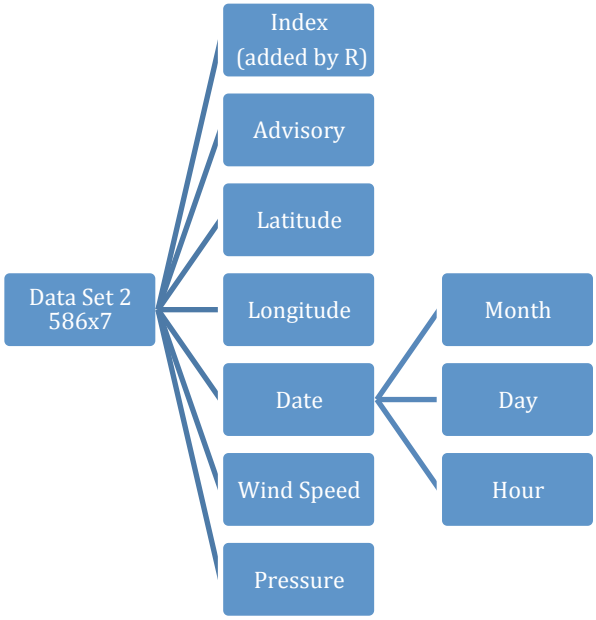

Figure 3.3: Added Headings to the Smaller Hurricane Data Set

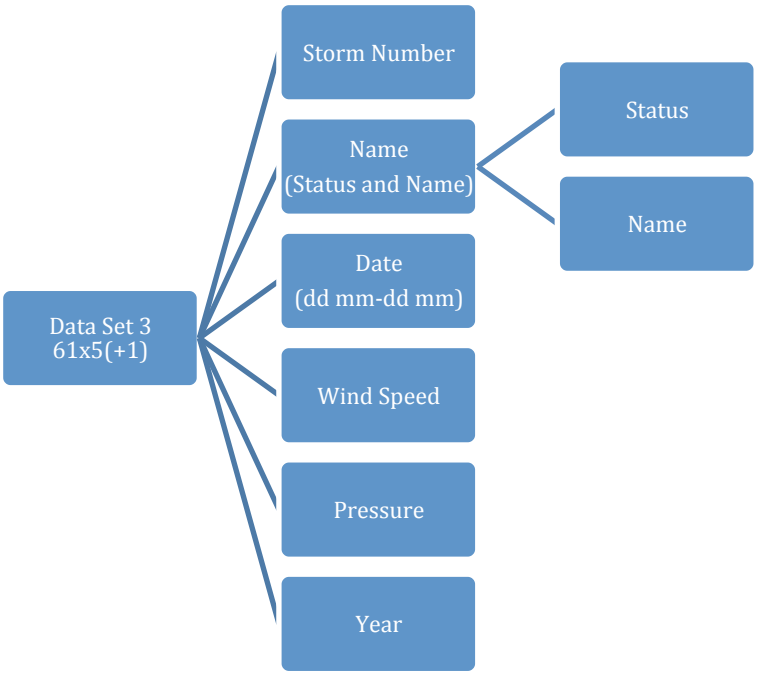

Figure 3.4: Data Set 3: Extract List of Storms for Given Years and Add Year to List

Then we had to prepare the smaller hurricane data file to use vertical lookups to map missing information into this file; namely, year and storm name. First enumerate storms by year and year of storm; this increases the width of the smaller hurricane date file to 14; with dimensions 586 x14 Now we code the smaller hurricane data file and 3 by year and the storm number, then read the name of the storm from the merged hurricane data set into the smaller hurricane data set. Then save the smaller hurricane data file as a CSV file and will have to reformat to have the common variables from large hurricane data (by decade) set from 1990 – 2014.

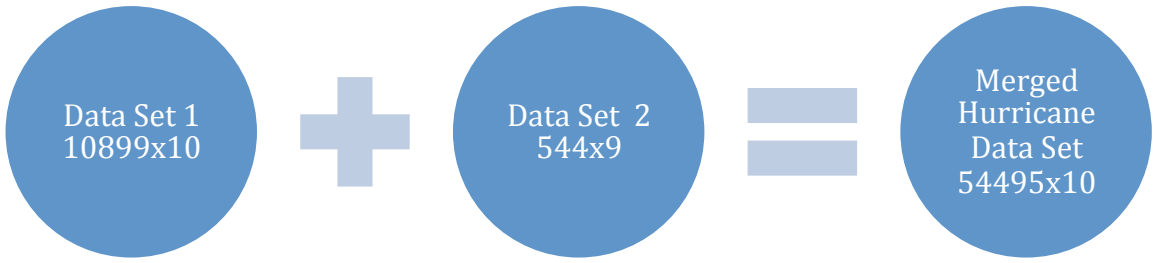

Figure 3.5: Data Set 1 and Data Set 2 merged

The second large data set came from the National Buoy Center. There were originally four buoys of interest; Buoy Data: B1 41040, B2 42036, B3 42056, and B4 42001. The fourth buoy was the first choice to use for merging with hurricane data because of the years of recorded data: 1975 – 2014. Although we only considered the years of 1990 – 2014 of the recorded buoy data for the merging process. This is because before 1990 there was not a lot of updated buoy readings kept for records. In the following Figure, the buoy data that was used in this chapter

came from the buoy numbered 42001. The dark arrow in the image shows where the buoy is in the Gulf of Mexico.

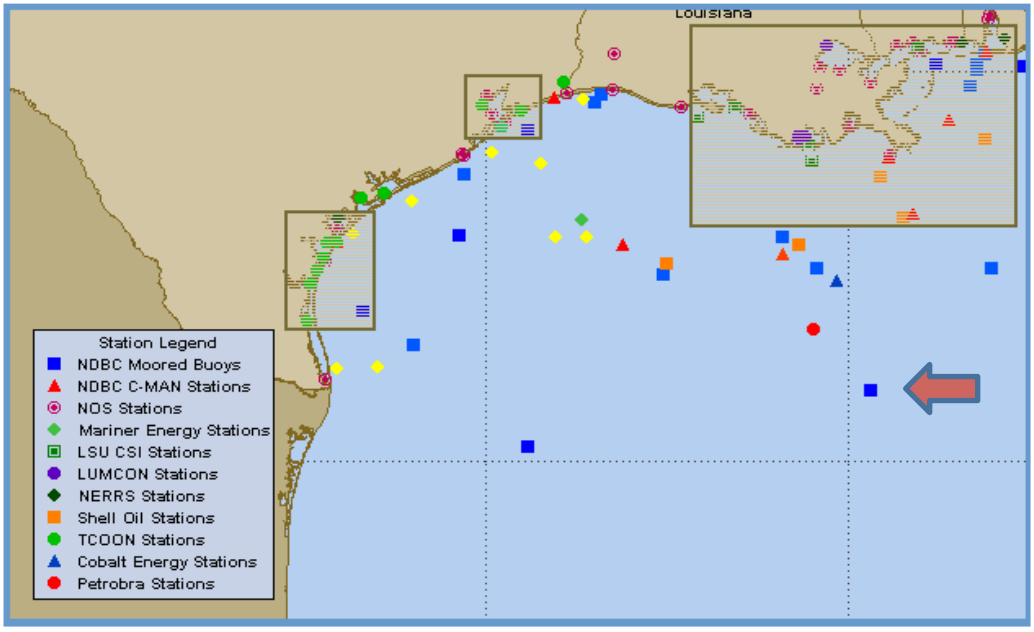

Figure 3.6: Location of Buoy 42001

The variables that came from this data set are: **Year**, **Day, Month, Hour, Buoy Wind Direction, Buoy Wind Speed, Buoy Pressure, Buoy Atmospheric Temperature, and Buoy Water Temperature.**

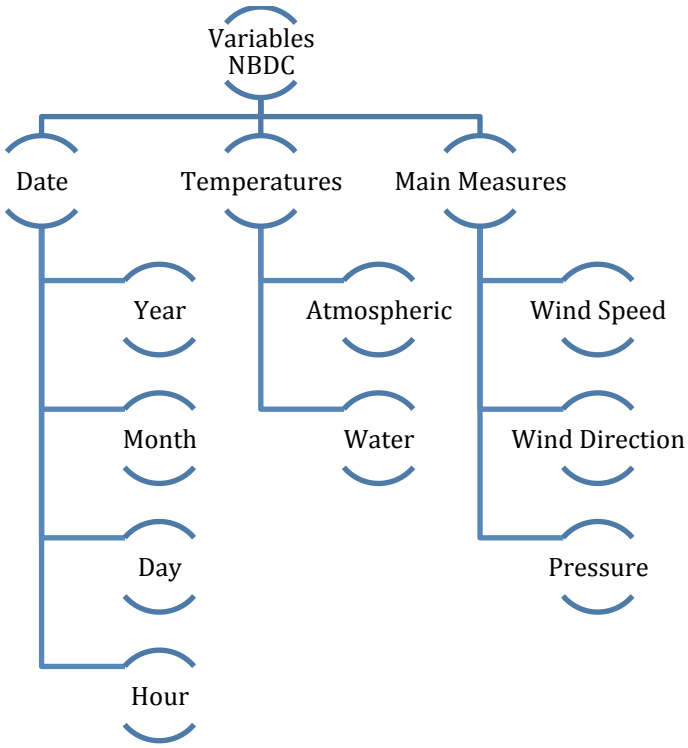

Figure 3.7: Variables of Interest from the National Buoy Data Center (Buoy 4)

Since the buoy data had gaps missing in the wind speed and wind direction, we needed to fill these gaps in order to proceed to achieve our final compilation data set. These gaps were filled using Fourier Series

$$Var = \beta_0 + \beta_1 cos(k \times Timeline) + \beta_2 sin(k \times Timeline) + \beta_3 cos(2k \times Timeline) + \beta_4 sin(2k \times Timeline) + \beta_5 cos(3k \times Timeline) + \beta_6 sin(3k \times Timeline);$$

where $k = \frac{2\pi}{365.25}$ and Var is the name of any variable in the buoy data set that has/had gaps to fill.

The next additional variables added were **Starting Latitude**, **Starting Longitude**, **Maximum Wind Speed**, **Minimum Pressure**, the differential of wind speed **dWS** (the change in wind speed between readings within a storm) $dWS(t) = WS(t) - WS(t-1)$, the differential of wind direction **dWD** (the change in wind direction between readings within a storm) $dWD(t) = WD(t) - WD(t-1)$, the differential of pressure **dP** (the change in pressure between readings within storm) $dP(t) = P(t) - P(t-1)$, the differential of atmospheric temperature **dATMP** (the change in atmospheric temperature between readings within a storm) $dATMP(t) = ATMP(t) - ATMP(t-1)$, and the differential of water temperature **dWTMP** (the change in water temperature between readings within a storm) $dWTMP(t) = WTMP(t) - WTMP(t-1)$. In Figure 3.8, the additional variables included can be seen.

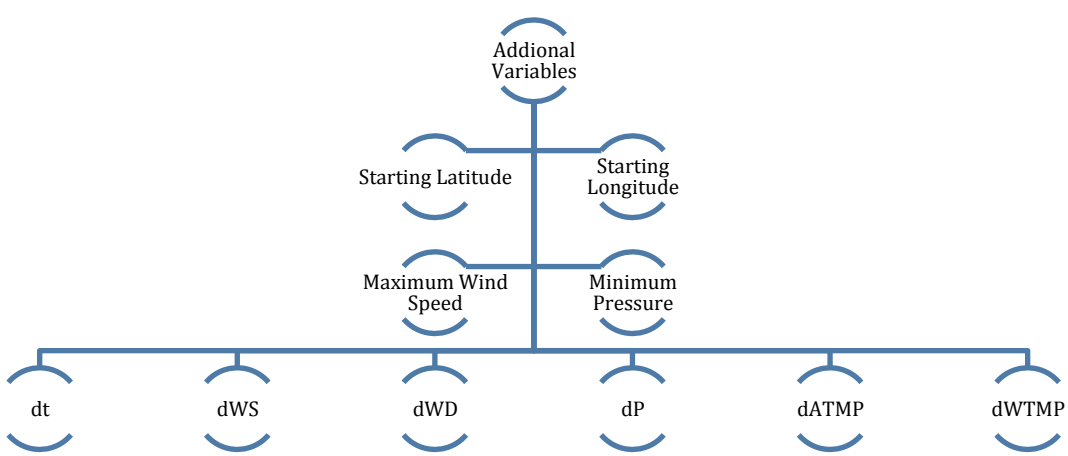

Figure 3.8: Additional Variables Included

Next, we had to create additional variables so we would then have a compilation data set of hurricanes in the Atlantic Basin and buoy data from the Gulf of Mexico. First, we had to sort the dates of the storms within the merged data set. Linear interpolation was used in R on the hurricane data file to show the increased readings on an hourly basis. This is because the date, year, month, day and hour were corrected in EXCEL. This produced a total of 17 variables of interest to be complied with the buoy data set. The included variables of interest are: **Timeline, Year, Month, Day, Hour, Storm, Name, Lat (Latitude), Lon (Longitude), WS (Wind Speed), dWS, Max wind, P (Pressure), dP, Min pres, Start Lat (Latitude), Start Lon (Longitude).**

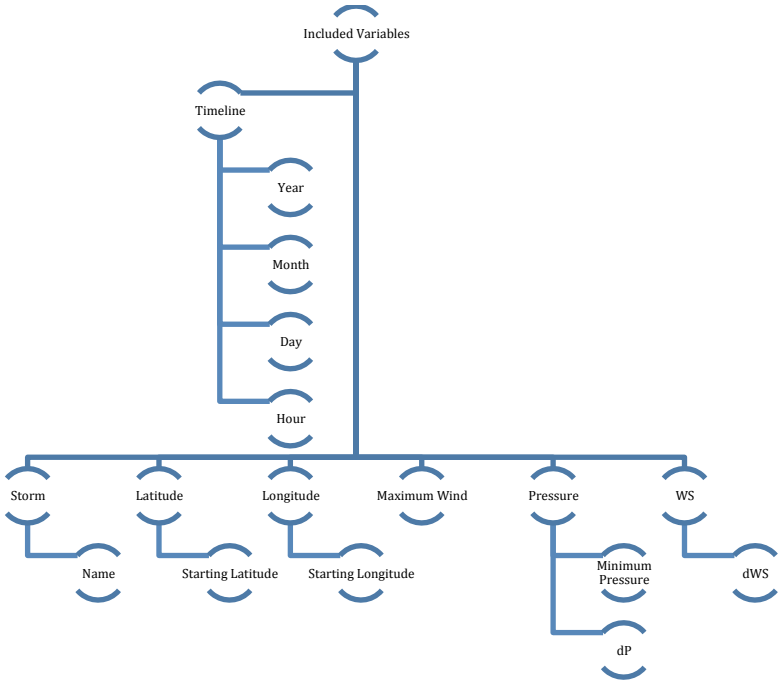

Figure 3.9: Variables of Interest

Then, merged the above variables with the buoy data.

After the gaps were filled in for any variable that had missing data, the final new included variables of interest included: **Year, Month, Day, Hour, Wind Direction, Wind Speed, Pressure, Atmospheric Temperature, Water Temperature, Date, Timeline, dt, dWS, dWD, dP, dATMP, and dWTMP.**

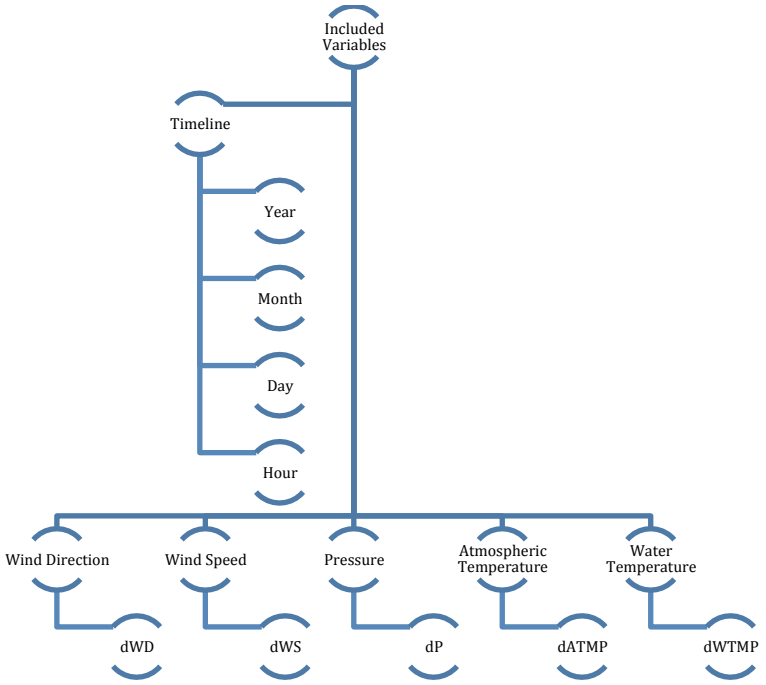

Figure 3.10: Final Included Variables of Interest for Compilation Data Set

As a result, we ended up with 17 variables of interest in the compilation of the data sets. This was a useful way to gather the information to answer our subjective research statements/questions.

VARIABLES OF INTEREST

In this present case study, the two data sets that was compiled into a larger data set came from two different sources. The hurricane data for the years 1990 – 2014 came from Unisys Weather site (Atlantic Basin Hurricanes data) and the buoy data has been available from the National Buoy Center. Next, we developed numerous statistical models to estimate the when a storm was present or not present in the Atlantic Basin. This will enable the distinction of which contributing factors will formulate when a storm is present or not in the Atlantic Basin. In this study we will statistically model the **storm present** as a function of **Buoy Wind Speed (x_1), Buoy Wind Direction (x_2), Buoy Pressure (x_3), Buoy Atmospheric Temperature (x_4), Buoy Water Temperature (x_5), Differential of Buoy Wind Speed (x_6), Differential of Pressure (x_7), Differential of Atmospheric Temperature (x_8), and Differential of Water Temperature (x_9).**

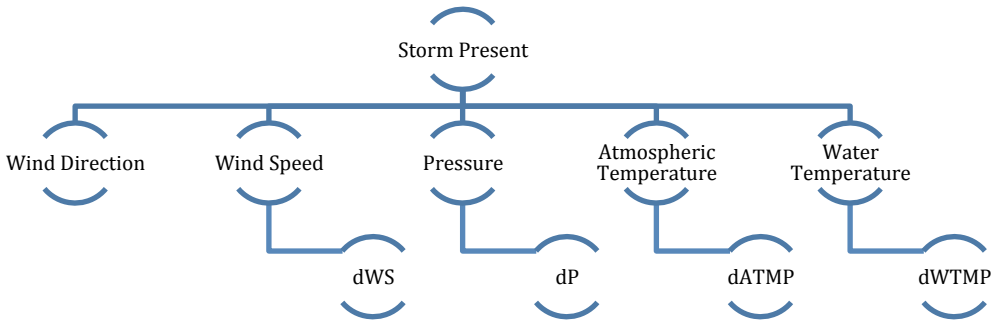

Figure 3.11: Variables of Interest

In this present study, we will address the following questions:

1) Determine the probability of a storm being present in the Atlantic Basin, given the conditions at the buoy.

2) Determine the probability of a storm being present categorically, given the conditions at the buoy.

Description of the Response Variable and Contributing Entities

The relationship between the wind speed, wind direction, pressure, atmospheric temperature, water temperature and their respective differentials were analyzed independently. The formation of a hurricane on any given day is a dichotomous measure in that either there is a storm present or there is no storm present. The atmospheric conditions are the factors that drive such storm formation. Obtaining a better understanding of these factors that drive such a storm formation, we will be able to determine probabilistically characterize the behavior of the phenomenon of interest and statistically model when a storm is present (also categorically) as a function of outlined variables.

Buoy Wind Speed (x_1)

The wind speed is recorded by the buoy in meters per second, (m/s), averaged over an eight-minute period, and then reported hourly [NDBC]. The wind speeds that were measured at this buoy are somewhat small and had a maximum value of 40.1 knots, as seen in Table 3.1.

Buoy Wind Direction (x_2)

The wind direction is the direction at which the wind is blowing. The buoy wind direction was calculated through Circular Analysis (see the statistical methodology chapter for further discussion on circular analysis).

Buoy Pressure (x_3)

The pressure is measured at sea level. In Table 3.1, the average condition for pressure was 1016.5 (considering all possibilities for a storm being present or not). Whereas in Table 3.2, the average condition for pressure was 1014.5 (storm present) and 1017 (storm not present).

Buoy Atmospheric Temperature (x_4)

The atmospheric temperature is measured in degrees Celsius.

Buoy Water Temperature (x_5)

Water Temperature (also known as Sea surface temperature) is a climate and weather measurement that is obtained by buoys [10]. There are different types of instruments that measure the temperature at different depths. Most buoys have sensors located at about 1 meter depth.

General Descriptive Statistics for the Buoy Conditions

In the following Table, the general descriptive statistics of the buoy conditions when they are the average atmospheric conditions. The Table shows that the buoy wind speed has a mean of 6, the buoy wind direction has a mean of 220, whereas the range is 140. The buoy pressure has a mean and a median of 1016.5, the buoy atmospheric temperature has a mean and a median of 25, the buoy water temperature has a mean and median of 26, the differentials all have a median of 0. The pressure has the lowest drop value by 13.8, which means that a storm could be present when the pressure drops 13.8 mb's below its average value. Notice that the buoy atmospheric temperature and the buoy water temperature have a similar mean value, while their ranges are different in a value of 10. The atmospheric temperature and the water temperature have similar maximum values, yet their minimum drop values vary by 10 degrees. This could imply that the

water temperature has a greater affect than the atmospheric temperature when a storm could be present in the Atlantic Basin

Table 3.1: General Descriptive Statistics for Buoy Conditions.

Buoy Conditions	**Mean**	**Median**	**Variance**	**Standard Deviation**	**Min**	**Max**	**Range**
x_1	6	5.8	8.9	2.98	0	40.1	40.1
x_2	220	219	14.32	3.05	112	252	140
x_3	1016.5	1016.5	18.198	4.26	935	1037	101.9
x_4	24.966	25.3	11.973	3.46	9.6	33.3	23.7
x_5	26.417	26.4	7.464	2.732	20.1	33.8	13.7
x_6	-0.0037	0	1.101	1.04	-13.5	12.4	25.9
x_7	-0.002	0	0.277	0.527	-13.8	19.3	33.1
x_8	-0.002	0	0.1315	0.3627	-7.7	3.7	11.4
x_9	0	0	0.151	0.123	-1.7	4.3	6

To gather a better understanding how the average atmospheric conditions, consider the average atmospheric conditions when a storm is present and not present, this can be seen in Table 3.2.

Table 3.2: Mean Values for Buoy Conditions when Storm is Present, Not Present and Overall

Buoy Conditions	Mean Storm Present	Mean Storm Not Present	Mean Overall
x_1	5.58	6.10	6
x_2	214	207	220
x_3	1014.5	1017	1016.5
x_4	28	24.27	24.966
x_5	29	26	26.417
x_6	-0.0021	0	-0.0037
x_7	-0.001	0	-0.002
x_8	-0.003	0	-0.002
x_9	0	0	0

Binomial Case of Logistic Regression

In this section we will address our first research question: Determine the probability of a storm being present in the Atlantic Basin, given the conditions at the buoy.

In the Binomial case, we will start off by using all of the variables of interest. The probabilistic analytic form of a logistic model is denoted as the following:

$y = \frac{e^{-(\beta_0 + \beta_1 x_1 + \cdots \beta_k x_k)}}{1+e^{-(\beta_0 + \beta_1 x_1 + \cdots \beta_k x_k)}}$, where $y = P(d = 1)\ and\ x_1, \ldots, x_k$ are the predictor variables.

Model Measurement of Accuracy

In order to find the probabilities for our binomial and multinomial models, we need to have a valid measurement of accuracy for the models. Having this measurement of accuracy will sustain the most valid model comparatively to another model. The response variable in both models is y. In general, we can have multiple predictor variables in a binomial logistic regression model, however, there are two outcomes; there was a storm present or there wasn't a storm present,

$$d = \begin{cases} 1, & if\ a\ storm\ is\ present \\ 0, & otherwise \end{cases}$$

First, we need to find and estimate for d which is denoted as $\hat{d}$.

$$\hat{d} = \begin{cases} 1, & \hat{p} > 0.5 \\ 0, & \hat{p} < 0.5 \end{cases}$$

The above shows that the two outcomes of the response variable estimates if our model is 50% chance of a storm being present and a 50% chance of a storm not being present. To find our measurement of accuracy we now extend our estimate of $\hat{d}$ to d^*.

$$d^* = \begin{cases} 1, & \hat{d} = d \\ 0, & \hat{d} \neq d \end{cases}$$

The last part of the procedure for finding the measurement of accuracy if to find the ratio denoted as p^*, $p^* = \frac{\sum d^*}{n}$, where n is the total number of outcomes if a storm was present or not, and $\sum d^*$ is the sum of the outcomes when there was storm present. This measurement of accuracy is the proportion of times that our model accurately predicts whether or not a storm is present.

In logistic regression, there is a logistic transformation of the odds (logit) that will serve as the dependent variable. In our first model, we considered every predictor variable and possible combination up to four way interaction.

Model Development

The analytic logistic transformation model that we will be using is

$$y = \beta_0 + \sum_{\forall i,j,k,h} \beta_i x_i^{a_i}\, x_j^{a_j}\, x_k^{a_k} x_h^{a_h}, i \neq j,k,h\; j \neq k,h, k \neq h, a\; \in \{0,1\}.$$

There were a total of 216 terms in the above model. Out of these 216 terms only 17 were significantly contributing of at least a 1% significance level. The differential terms and every possible term associated with a differential term was not significantly contributing of a level of 1% or higher.

In this first analytic model,

$$y = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_1 x_2 + \beta_7 x_3 x_4 +$$

$$\beta_8 x_3 x_5 + \beta_9 x_4\, x_5 + \beta_{10} x_1 x_3 + \beta_{11} x_1 x_4 + \beta_{12} x_1 x_5 + \beta_{13} x_3\, x_4 x_5 + \beta_{14} x_1 x_3 x_4 +$$

$$\beta_{15} x_1 x_3 x_5 + \beta_{16} x_1 x_4 x_5 + \beta_{17} x_1 x_3 x_4 x_5$$

the 17 predictor variables were found to be significantly contributing with an $R^2 = 0.181$. Since logistic regression is similar to regression after the transformation model, we are using the r – squared values for comparison values between models in this chapter. The measurement of accuracy of this first developed model was found to be 58% accurate.

After computing the Maximum Likelihood estimates for our reduced model above, we found that the intercept and each predictor variable except the interaction between buoy wind speed and buoy wind direction (x_1x_2) was significantly contributing by 1, 5 and 10% to the model above. Thus, we dropped this interaction term and considered the developed model without it. The measurement of accuracy of this second developed model was found to be 79% accurate. Now we will consider another model to investigate and draw conclusions from for our research statement.

The new developed model is as follows:

$$\hat{y} = 1.15 - 7.60x_1 - 0.008x_2 - 1.14x_3 - 4.45x_4 - 4.30x_5 + 0.04x_3x_4 + 0.05x_3x_5 + 1.67x_4x_5 + 0.07x_1x_3 + 2.21x_1x_4 + 2.89x_1x_5 - 0.01x_3x_4x_5 - 0.002x_1x_3x_4 - 0.03x_1x_3x_5 - 0.08x_1x_4x_5 + 0.09x_1x_3x_4x_5$$

The above model has 16 predictor variables that are all significantly contributing at the 1% level of significance. Here, using subset analysis, we will consider another smaller model. Consider the following logistic model:

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \beta_5x_5 + \beta_6x_1x_2 + \beta_7x_1x_5 + \beta_8x_4x_5$$

The developed model is:

$$\hat{y} = 92.47 + -5.77x_1 - 0.181x_2 - 0.146x_3 + 1.52x_4 + 2.10x_5 + 0.014x_1x_2 + 0.05x_1x_5 - 0.64x_4x_5$$

In the above developed model each predictor variable was found to be significantly contributing at the 1% level of significance. With the third developed model, every predictor variable was found to be significantly contributing at the 1% level of significance and had an $R^2 = 0.6835$. In

the fourth developed model, every predictor variable was also significantly contributing at the 1% level of significance. However, the $R^2 = 0.7122$. Since the law of parsimony states that entities should not be multiplied needlessly, and the simpler of two competing theories is to be preferred, then the fourth developed model is a better model to use.

In Table 3.3, the measurements of accuracy for the four developed models are shown. The first developed model with 216 variables had an $R^2 = 0.181$ and a measurement of accuracy of 58%, the second developed model had 17 variables with a $R^2 = 0.654$ and a measurement of accuracy of 79%, the third developed model had 16 variables with a $R^2 = 0.6835$, and a measurement of accuracy of 81%. The fourth developed model with 8 variables had a

$R^2 = 0.7122$, and a measurement of accuracy of 84%. Therefore, we can conclude that the fourth developed model is the better model to use in drawing conclusions for our first research question.

Table 3.3: Measurement of Accuracy and R – Squared values for the Developed Models

Developed Models	Measurement of Accuracy	R^2
1 – Full Model	58%	0.181
2 – Second Model with 17 terms	79%	0.654
3 – Third Model with 16 terms	81%	0.6835
4 – Fourth Model with 8 terms	84%	0.7122

Using the fourth developed, we can now predict the probability of a storm being present, given w the conditions at the buoy. To achieve our goal, we will use the following regression equation

$$\hat{p} = \frac{e^{-(92.47+ -5.77x_1 -0.181x_2-0.146x_3+ 1.52x_4+ 2.10x_5 + 0.014x_1x_2+ 0.05x_1x_5- 0.64x_4x_5\ )}}{1+e^{-}(92.47+ -5.77x_1 -0.181x_2-0.146x_3+ 1.52x_4+ 2.10x_5 + 0.014x_1x_2+ 0.05x_1x_5- 0.64x_4x_5\ )}$$

Where (calculated buoy conditions) is the calculated number of the output, from inputting specific buoy conditions. Recall that the fourth developed model was

$$\hat{y} =$$

$$92.47 + -5.77x_1 - 0.181x_2 - 0.146x_3 + 1.52x_4 + 2.10x_5 + 0.014x_1x_2 + 0.05x_1x_5 - 0.64x_4x_5.$$

Now, inputting specific buoy conditions into the fourth developed model, will produce the calculated buoy conditions that we will need to input into our regression equation.

Consider the three situations outlined in Table 3.2, where the average atmospheric conditions are given. In case 1, we can estimate the probability of a storm being present, using the overall standard atmospheric conditions; when the wind speed is 6, wind direction is 220, pressure is 1016.5, atmospheric temperature is 25, and the water temperature is 26. Thus when the overall average atmospheric conditions are used $\hat{p} = 0.68$, there is a 68% chance that there is a storm present in the Atlantic Basin. In case 2, we will consider using the average buoy conditions for when a storm is present. The values we will use for the buoy conditions are: wind speed is 5.58, wind direction 214, pressure 1014.5 mb, atmospheric temperature 28, water temperature 29. In this example, our probability is $\hat{p}$= 0.71.

This means when we consider the average atmospheric conditions for when a storm is present, there is a 71% chance that there is a storm present in the Atlantic Basin. Now in case 3, we will consider the average atmospheric conditions when a storm is not present. The specific

values we will use for the buoy conditions are: wind speed is 6.10, wind direction 207, pressure 1017, atmospheric temperature 24.27, and water temperature 26. In this example our probability is

$\hat{p}$ = 0.62. This means when we consider the average atmospheric conditions for when a storm is not present, there is a 62% chance that there is a storm present in the Atlantic Basin. Now let us consider a fourth case when the when the wind speed is 15, wind direction is 214, pressure is 1000 mb, atmospheric temperature is 28, and the water temperature is 30. In this example, our probability is $\hat{p}$= 0.93. Thus, when the wind speed is 15, and the mean wind direction is 214, and there is a drop in pressure by 13.8 (rounded to 14 mb's), and the atmospheric temperature is it's mean value while the water temperature is higher than its mean value, then there is a 93% chance of a storm being present in the Atlantic Basin. What this means is that when there is a significant increase in the wind speed and there is significant drop in the pressure and the water temperature is higher than its mean value, then there is a greater chance of a storm occurring in the Atlantic Basin.

Multinomial Case of Logistic Regression

In this section we will present our second research question: Determine the probability of a storm being present categorically, given the conditions at the buoy. Since we are considering a storm being present categorically, then we will use multinomial logistic regression to address our second research question. In the multinomial logit model we assume that the log-odds of the response follow a linear model of the logistic transformation of the odds (logit) that will serve as the dependent variable. In the multinomial logit model we assume that the log - odds of the response follow a linear model $y_j = logit(p_j) = \log\left(\frac{p_j}{1-p_j}\right) = \alpha_j + \beta_{1j}x_1 + \cdots + \beta_{pj}x_p$, To

find our probabilities, we will use the probabilistic analytic form of a logistic model is denoted as the following:

$$p = \frac{e^{-(\alpha_j + \beta_{1j} x_1 + \cdots + \beta_{pj} x_p)}}{1 + e^{-(\alpha_j + \beta_{1j} x_1 + \cdots + \beta_{pj} x_p)}}.$$

The following Table 3.4 shows the average buoy conditions during tropical storms, where the rows indicate 0 to 5 (the severity of the storm, i.e., the values of 1 to 5 represent category 1 to 5 storms and a value of 0 represents a tropical depression or no storm).

Table 3.4: Average Buoy Conditions for Storm Present (Categorically)

Categories	x_1	x_2	x_3	x_4	x_5
0	6	150	1016.6	25	26.2
1	5.92	171	1014	27.5	28.7
2	6.64	180	1012	28	29.18
3	5.89	188	1013	28	29.03
4	5.75	214	1014.4	27.8	28.95
5	5.06	220	1014.3	29	30

From Table 3.4, the average buoy conditions for the atmospheric temperature and water temperature show results that as a hurricane gets stronger and higher categorically, their temperatures go from 25 to 29 and 26.2 to 30. In the following Figure, the atmospheric temperatures indicate that as a storm is present categorically in the Atlantic Basin, the average temperature of 25 shows that a tropical depression or storm may be present. The higher the

atmospheric temperature rises, then the higher the category of a hurricane being present in the Atlantic Basin.

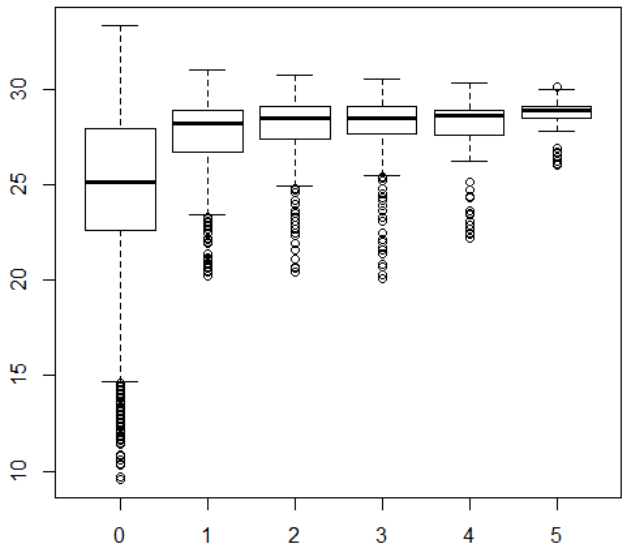

Figure 3.12: Boxplots of the Average Atmospheric Temperatures (Categorically)

Next, a boxplot by category of the average water temperatures are shown in the following Figure. Notice that there are no outliers shown when there is no hurricane present. The average is 26, which is what the combined average was from Table 3.1. As the water temperature averages go higher, the higher the category of a hurricane. In the first research question, we found that when the water temperature was 30, then there was a higher probability of a storm being present in the Atlantic Basin.

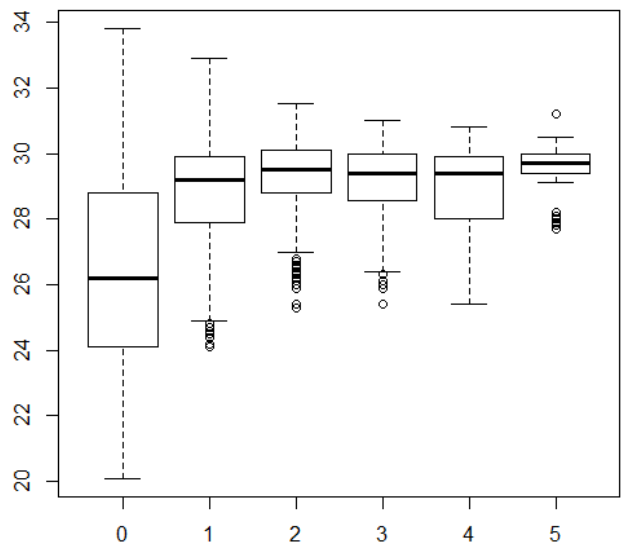

Figure 3.13: Boxplots of the Average Water Temperatures (Categorically)

Model of a Storm Being Present Categorically

Since the second developed model had a measurement of accuracy of 79%, we will first consider that model in the multinomial logistic regression case to address the second research question. Although the response variable will now be storm present with categorical outcomes to represent hurricanes of categorically from 0 to 5. Recall that the second developed model was of the form:

$$y_J = \alpha_J + \beta_{1j}x_1 + \beta_{2j}x_2 + \beta_{3j}x_3 + \beta_{4j}x_4 + \beta_{5j}x_5 + \beta_{6j}x_1x_2 + \beta_{7j}x_3x_4 + \beta_{8j}x_3x_5 + \beta_{9j}x_4x_5 + \beta_{10j}x_1x_3 + \beta_{11j}x_1x_4 + \beta_{12j}x_1x_5 + \beta_{13j}x_3x_4x_5 + \beta_{14j}x_1x_3x_4 + \beta_{15j}x_1x_3x_5 + \beta_{16j}x_1x_4x_5 + \beta_{17j}x_1x_3x_4x_5$$

Since there are 5 categories of a storm being present (cat 1 through cat 5), then there will be 5 different developed models that arise from our multinomial base model. First we will discuss the two most interesting and significant of the developed models for cat 1 and cat 2 hurricanes, because their measurement of accuracy is over 85%. The developed model for determining when a storm is present (for cat 1) is:

$$\begin{aligned}\hat{y}_1 = &-0.91 + 3.24x_1 - 0.182x_2 - 0.09x_3 - 41.26\,x_4 + 1.43x_5 + 0.007x_1x_2 + 0.04x_3x_4 \\ &- 0.004x_3x_5 + 1.65x_4\,x_5 - 0.01x_1x_3 + 1.45x_1x_4 - 0.11x_1x_5 - 0.01x_3\,x_4x_5 \\ &- 0.02x_1x_3x_4 - 2.78x_1x_3x_5 + 0.05x_1x_4x_5 - 0.03x_1x_3x_4x_5\end{aligned}$$

This developed model had a measurement of accuracy of 94%. We can predict the probability of a storm being present (categorically), using the buoy atmospheric conditions of the average values of a hurricane cat 1. This means when the wind speed is 6, wind direction is 171, pressure is 1014, atmospheric temperature is 27.5, and the water temperature is 28.7.

This probability is $\hat{p}$ = 0.99. This means that when we use the buoy average atmospheric conditions for a category 1 hurricane there is a 99% chance of a category 1 hurricane being present in the Atlantic Basin. This is a high probability for the chances of a category 1 hurricane occurring, given the average buoy conditions for when a storm is present in the Atlantic Basin. Next we will discuss the developed model for a category 2 hurricane.

The developed model for determining when a storm is present (for cat 2) is:

$$\begin{aligned}\hat{y}_2 = &-0.75 + 1.72x_1 - 0.33x_2 - 0.108x_3 - 24.28\,x_4 - 0.94\,x_5 + 0.02x_1x_2 + 0.03x_3x_4 \\ &+ 0.004x_3x_5 + 1.09x_4\,x_5 + 0.05x_1x_3 + 0.08x_1x_4 - 0.46x_1x_5 - 0.01x_3\,x_4x_5 \\ &+ 0.002x_1x_3x_4 + 2.04x_1x_3x_5 + 0.01x_1x_4x_5 - 0.07x_1x_3x_4x_5\end{aligned}$$

This developed model had a measurement of accuracy of 85%. Now using the average atmospheric buoy conditions for a category 2 hurricane; this means when the wind speed is 6.64, wind direction is 180, pressure is 1012, atmospheric temperature is 28, and the water temperature is 29.18. Our obtained probability is $\hat{p}$ = 0.82. Hence, there is an 82% chance of a category 2 hurricane being present in the Atlantic Basin storm when we consider using the buoy average atmospheric conditions for a category 2 hurricane. This is a relatively medium to high probability for the chances of a category 2 hurricane occurring, given the average buoy conditions.

The first two developed models for a storm being present categorically had the highest model measurement of accuracy. Now we will discuss the remaining three developed models for categories 3 through 5. The developed model for determining when a storm is present (for cat 3) is:

$$\begin{aligned}\hat{y}_3 = 97.3 - 7.77x_1 - 0.23x_2 - 0.18x_3 - 2.96\,x_4 + 3.11x_5 + 0.008x_1x_2 + 0.002x_3x_4 \\ - 0.001x_3x_5 + 0.006x_4\,x_5 + 0.001x_1x_3 + 1.23x_1x_4 - 0.08x_1x_5 \\ - 0.01x_3\,x_4x_5 - 0.02x_1x_3x_4 - 0.012x_1x_3x_5 + 0.006x_1x_4x_5 \\ - 0.001x_1x_3x_4x_5\end{aligned}$$

This developed model had a measurement of accuracy of 77%. The developed model for determining when a storm is present (for cat 4) is:

$$\begin{aligned}\hat{y}_4 = 21.6 - 5.6x_1 - 0.30x_2 - 0.06x_3 - 1.13\,x_4 + 1.50x_5 + 0.06x_1x_2 + 0.005x_3x_4 \\ - 0.0002x_3x_5 + 0.007x_4\,x_5 + 0.002x_1x_3 + 0.40x_1x_4 - 0.007x_1x_5 \\ - 0.01x_3\,x_4x_5 - 0.002x_1x_3x_4 - 0.034x_1x_3x_5 + 0.0006x_1x_4x_5 \\ - 0.003x_1x_3x_4x_5\end{aligned}$$

This developed model had a measurement of accuracy of 63%.

The developed model for determining when a storm is present (for cat 5) is:

$$\hat{y}_5 = -39.1 - 5.08x_1 - 0.39x_2 - 0.13x_3 + 2.87\,x_4 - 5.82x_5 + 0.03x_1x_2 + 0.002x_3x_4 - 0.0001x_3x_5 + 0.007x_4\,x_5 + 0.0004x_1x_3 - 0.19x_1x_4 + 0.001x_1x_5 - 0.001x_3\,x_4x_5 - 0.008x_1x_3x_4 - 0.027x_1x_3x_5 + 0.0005x_1x_4x_5 - 0.004x_1x_3x_4x_5$$

This developed model had a measurement of accuracy of 48%.

These three developed models for determining when a storm is present, for categories 3 through 5 had a model measurement under 80%. This is why consideration for the first two developed models were held in higher interest. Something interesting to consider in this analysis is using the buoy average atmospheric conditions for the categorical storms in other models, and determining what their probabilistic significance is. We will consider 5 cases; case 1 will be using the buoy average atmospheric conditions for a category 1 hurricane in the four other models, and case 2 will be using the buoy average atmospheric conditions for a category 2 hurricane in the four other models, case 3 will be using the buoy average atmospheric conditions for a category 3 hurricane in all the models. We will continue this process along with case 4 and case 5 in a similar fashion for category 4 and category 5 hurricanes. The illustration in Table 3.5 will better provide a clear and concise view of the probabilities that were generated using the buoy average atmospheric conditions for all the category hurricanes into the remaining developed models.

Table 3.5: Probabilities of a Storm being Present (categorically) using all 5 Developed Models

Models	**Cat 1**	**Cat 2**	**Cat 3**	**Cat 4**	**Cat 5**
1	0.99	0.92	0.89	0.85	0.77
2	0.97	0.82	0.80	0.60	0.42
3	0.88	0.73	0.51	0.46	0.44
4	0.95	0.91	0.76	0.33	0.25
5	0.87	0.78	0.75	0.52	0.35

Note that in the above Table, the probabilities using the buoy average atmospheric conditions for the last three models who have a lower model measurement of accuracy versus the first two models, have a sufficient larger probabilistic significance in the other models, than their own. For example, using the buoy average atmospheric conditions for the third developed model (cat 3), and substituting those values into the other models, we see that these conditions lead to a higher probability in the developed model for determining a category 1 and category 2 hurricane. It also is similar when considering the buoy average atmospheric conditions for the second, fourth, and fifth developed models. Also, notice that when we consider the buoy average atmospheric conditions for the first developed model (cat 1), the conditions lead to lower probabilities in the other four models.

Exploratory Factor Analysis in Conjunction with Non-Response Analysis

In this section we will further investigate the hurricane and buoy data. We will demonstrate how exploratory factor analysis can be used to determine the distinct factors that house the terms that explain the variance among the co-dependent variables and how non-response analysis can be applied to model the non-functional relationship that exist in a dynamic system. "Moreover, the analysis indicates that there are pumping actions or ebb and flow between the pressure and the water temperature readings near the surface of the water days before a tropical storm forms in the Atlantic Basic and that there is a high correlation between storm conditions and buoy conditions three-four days before a storm forms" [21].

The hurricane data used in this analysis are taken from UNISYS Weather Center from 2000-2009, Figure 3.14 and includes a time stamp, name of the hurricane, location (latitude and longitude) and the main variable of interest **wind speed** and **pressure**.

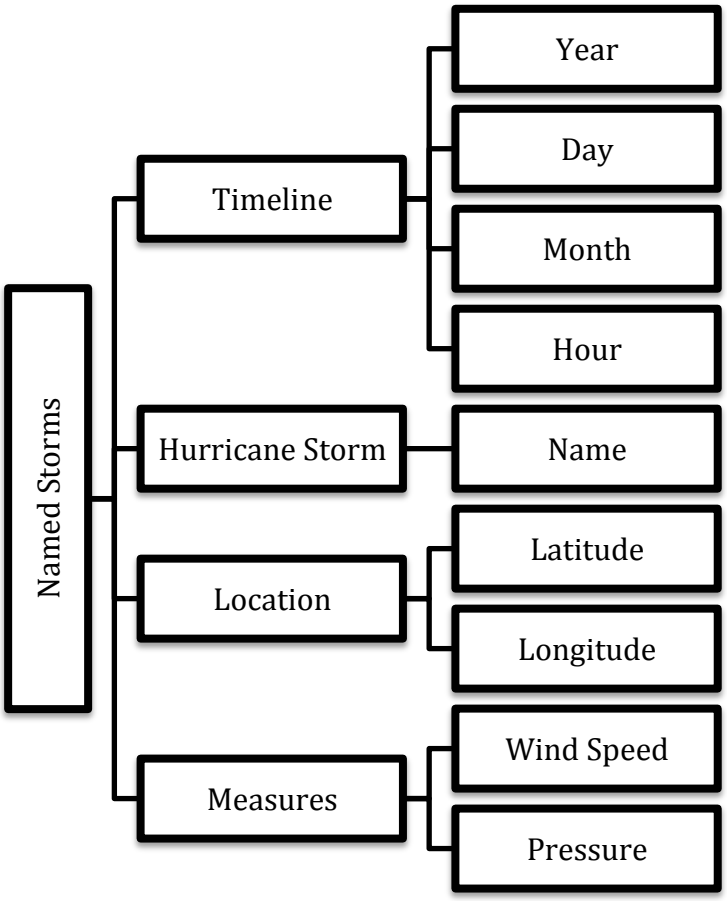

Figure 3.14: Data diagram of named storms in the Atlantic Basin

The second data set in Figure 3.15, from "the National Data Buoy Center containing the **wind speed**, **pressure**, **atmospheric temperature** and **water temperature** where added to the **wind speed** and **pressure** readings from the hurricanes with 36 daily time shifts used to measure the buoy conditions days before the formation of a tropical storm" [21].

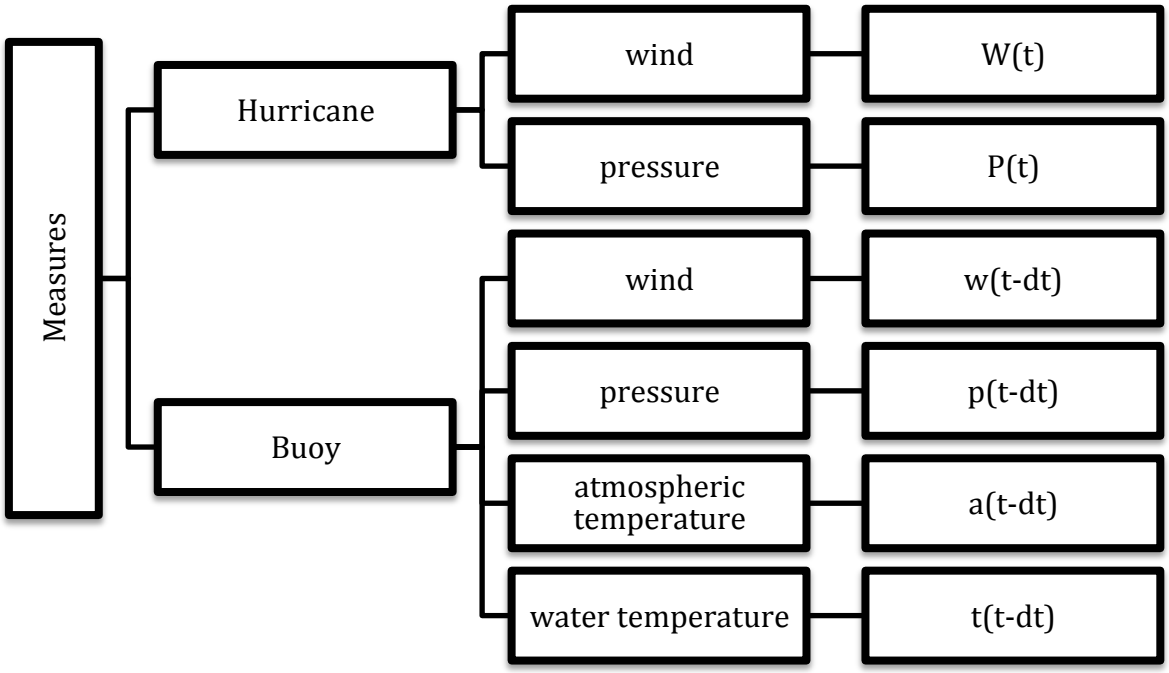

Figure 3.15: Measured variables of interest including time shifts in the buoy conditions.

"The terms to be considered using exploratory factor analysis and non-response analysis includes the following 36 terms: the primary variables, the second degree terms and all first order interaction terms: $\{W, P, w, p, a, t, W^2, WP, Ww, Wp, P^2, \ldots\}$" [21].

All of the factors are listed is in Table 3.6 which sorts the terms into factors. Using exploratory factor analysis, we found four principle components.

Table 3.6: All Possible Terms

	Factor1	Factor2	Factor3	Factor4
W	1			
P	-0.94			
W^2	0.97			
WP	1			
wW	0.7		0.62	
pW	1			
aW	0.98			
tW	0.99			
P^2	-0.94			
pP	-0.92			
a		0.96		
t		0.96		
aP		0.96		
tP		0.95		
ap		0.96		
tp		0.97		
a^2		0.97		
at		0.98		
t^2		0.96		
w			0.97	
wP			0.97	
w^2			0.94	
wp			0.97	
wa			0.98	
wt			0.99	
p		-0.3		0.91
$p2$		-0.3		0.91

Table 3.7 gives the "SS loading weights, the proportion of variance contained in each factor and the cumulative proportions; and indicates that four components (factors) were sufficient, explaining 96% of the variation" [21]. Since there was a SS loading that is less than 1, the fifth

factor was found to be insignificant [21]. The first factor with an SS loading of 9.05 indicates that at least 34% of the variance among the terms exists [21].

Table 3.7: SS Loadings

	Factor1	Factor2	Factor3	Factor4	Factor5
SS loadings	9.05	8.77	6.35	1.84	0.37
Proportion Variance	0.34	0.32	0.24	0.07	0.01
Cumulative Variance	0.34	0.66	0.9	0.96	0.98

In this section, the terms of interest are those variables, interaction and second degree terms belonging to the first principle component and the primary variable of interest is wind speed of a hurricane as related to the pressure of the hurricane and the buoy conditions [21].

Let us consider the non-response model:

$$unity = \alpha_1 W + \alpha_2 P + \alpha_3 W^2 + \alpha_4 P^2 + \alpha_5 Ww + \alpha_6 Wp + \alpha_7 Wa + \alpha_8 Wt + \alpha_9 WP + \alpha_{10} Pp$$

where $unity$ is a column vector of 1 and $\alpha_i's$ are the weights that balance the system [21]. If we want to determine the number of days (dt) before the storms formation that best predicts the intensity of a storm, then using the correlation between W and $\widehat{W}$, we found computed for

$$dt = 1,2, \dots ,36\ days.$$

"The maximum correlation was found to be 0.9882843 when dt is three days; that is the buoy condition three days before the hurricane reading shows the highest correlation with the storm conditions" [21]. The following image shows that there is a sinusoidal relationship in the measured correlations [21].

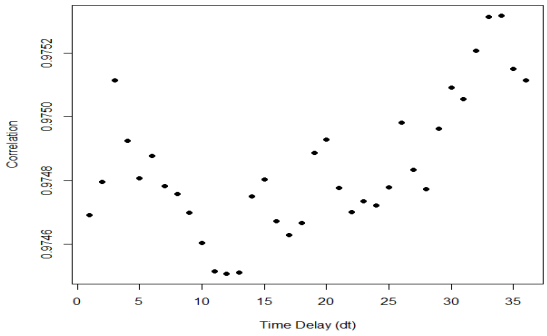

Figure 3.16: Correlation between the observed and estimated wind speed based on the buoy conditions over the give time delay.

"This analysis is useful in the field of meteorology as it allows co-dependent relationships among atmospheric conditions to be expressed implicitly" [21].

Usefulness & Contributions

The findings in this study are important for numerous reasons; this is the very first time that someone has used logistic regression and atmospheric conditions at a given buoy to estimate the probability of a storm being present in the Atlantic Basin. Further extending the binomial regression to the multinomial regression allowed us to better predict when a storm is present categorically in the Atlantic Basin. The comparison of these 5 developed models leads us to further estimate that the probabilities of a category 1 and category 2 hurricane occurring, given the conditions at the buoy. In regards to the last section of this chapter, "this analysis is useful in the field of meteorology as it allows co-dependent relationships among atmospheric conditions to be expressed implicitly" [21]. "The end result of this analysis will be an application which reads the current conditions at the buoy and predict the formation of a tropical storm based on the conditions near the surface of the water" [21].

CHAPTER 4: A STATISTICAL ANALYSIS OF FLORIDA SINKHOLES

In this chapter a statistical study of the sinkholes that have occurred in the state of Florida will be discussed. Sinkholes occur more in Florida than any other state in the nation. In fact, in the city of Tampa, it is known as 'Sinkhole Valley'. From the motivation section, we know that there are four different types of sinkholes and they are **Collapse, Solution, Alluvial,** and **Raveling.**

The data that was used in this study on sinkholes came from the Florida Department of Environmental Protection, Subsidence Incident Reports from 1970 – 2008. The dimensions of the data was 926 with 15 variables. In this study, the variables of interest are: **sinkhole length (x_1)**, **sinkhole width** (x_2), **sinkhole depth** (x_3), **sinkhole slope** (x_4), **diameter** (x_5) and **soil types** (Y).

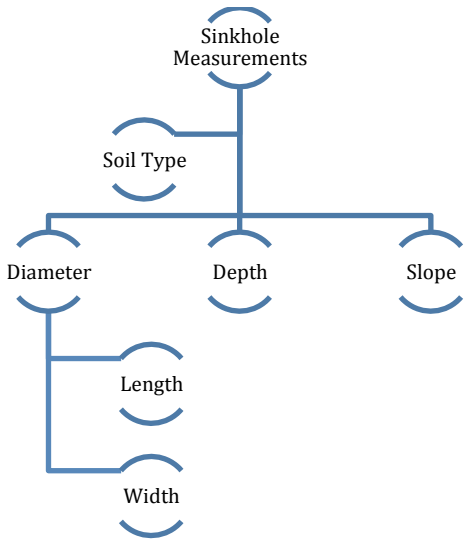

Figure 4.1: Variables of Interest

RANKING OF SOIL TYPES

In this current study the five soil types are sand, unknown, clay, rock, and limestone (which will be referred to as lime). In the following Table 4.1, the soil types are ranked from highest to lowest based upon frequency of occurrence in the last 38 years (1970 – 2008), taking into account of the possible 'mixed' or 'combined' soil types within the data set. The specific soil types that were included in this study that had the largest amount of occurrence within the data were: sand, unknown, clay, rock, and limestone. Since there was many that were different types of sand, they were classified as sand. The unknown soil type was not included within the possible 'mixed' or 'combined' soil types within the data set, but it did have a frequency of occurrence of 296. In Table 4.1, the soil type Sand had the highest ranking of frequency of occurrence and that when examined with the mixed or combined soil types, that it still has the highest frequency of occurrence, when just looking at sand by itself and not the mixed or combined soil types.

When taking into consideration the mixed or combined soil types of sand, we can see that in Table 4.1, that sand/clay had 80 frequencies of occurrence, sand/rock had 15 frequencies of occurrence and sand/lime had 10 frequencies of occurrence. After, examining the data and taking into consideration the mixed or combined soil types of both rock and lime, we can see that Lime will now be ranked third and higher than Rock in terms of frequencies of occurrence.

Table 4.1: Soil Type Ranking of Frequency of Occurrence (Mixed or Combined).

Soil Type	Pure	With Clay	With Rock	With Lime	With Sand	Total
Sand	447	80	15	10		552
Clay	24		1	1	80	106
Lime	11	1	3		10	25
Rock	8	1		3	15	27

The research questions/statements that are to be addressed in this study are:

1) Determine the relationship between a sinkhole's length and width.

2) Determine the probability distribution that best characterizes the diameter and a confidence interval that detects the average diameter of a sinkhole in Florida.

3) Determine the probability a sinkhole on a certain soil type, given the sinkhole length, sinkhole width, depth and slope.

RELATIONSHIP BETWEEN THE SINKHOLE LENGTH AND WIDTH

To address this research question/statement, we will compare the means of the two measures; using parametric analysis to determine if their means are similar. Then we will determine the best fit probability distributions between the sinkhole's length and width; and verify our findings by comparing their medians using non-parametric methods.

In Table 4.2, the descriptive statistics for the sinkhole length (left) and sinkhole width (right) are shown. Notice that the mean for the sinkhole length is 14.129 and the sinkhole width is 12.961 (or approximately 13), this shows that the means of both of these variables are similar).

Table 4.2: Descriptive Statistics for Sinkhole Length and Sinkhole Width

Sample Size	926
Range	349.5
Mean	14.129
Variance	684.18
Std. Deviation	26.157
Coef. of Variation	1.8512
Std. Error	0.85957
Skewness	7.9221
Excess Kurtosis	89.861

Sample Size	926
Range	349.5
Mean	12.961
Variance	638.6
Std. Deviation	25.271
Coef. of Variation	1.9498
Std. Error	0.83044
Skewness	8.638
Excess Kurtosis	103.87

By invoking the central limit theorem, regardless of the data's distribution, as our sample size is 926, the sampling distribution will approach the normal distribution. Therefore to view the relationship between the sinkhole length and width, a standard t - test was used at the 0.05 level of significance for the comparing of means hypothesis test to see if the mean of the sinkhole length is similar as the mean of the sinkhole width. The null hypothesis was that the sinkhole length and the sinkhole width have the same means. The alternative hypothesis was that the

sinkhole length and the sinkhole width have significant differences in their means. The test statistic t = 0.9776, with a p-value of 0.3284 we will fail to reject the null hypothesis. Hence, at a 0.05 level of significance we can conclude that the mean of the sinkhole length and the mean of the sinkhole width are the same.

Next we will compare the sinkhole length and width to see if they have the same distributions. For our two data sets (sinkhole length and width), they will be compared and ranked against 65 continuous distributions, where the goodness-of-fit tests (Anderson-Darling, Kolmogorov-Smirnov, and Chi-Square) was performed. Using Maximum Likelihood Estimates, among the 65 different continuous distributions were taken into account, it was found that the best fit probability distribution for the sinkhole length and width was the Log – Pearson 3. The top 5 best fit distributions for the sinkhole length and width can be seen in Table 4.3 and Table 4.4.

Table 4.3: Goodness – of – Fit - Tests for the Best Fit Distributions for the Sinkhole Length

Sinkhole Length	Anderson - Darling		Kolmogorov-Smirnov		Chi - Square	
Distribution	Statistic	Rank	Statistic	Rank	Statistic	Rank
Log – Pearson 3	2.3991	1	0.05486	1	78.014	15
Frechet(3P)	2.9461	2	0.05321	2	33.406	1
Pearson 5(3P)	2.8815	3	2.8815	8	57.065	8
Dagum	2.8512	4	0.0652	7	59.835	4
Dagum(4P)	2.8639	5	0.0562	4	55.884	10

Table 4.4: Goodness – of – Fit - Tests for the Best Fit Distributions for the Sinkhole Width

Sinkhole Width	Anderson - Darling		Kolmogorov-Smirnov		Chi - Square	
Distribution	Statistic	Rank	Statistic	Rank	Statistic	Rank
Log – Pearson 3	2.4908	1	0.0518	1	33.442	6
Pearson 5(3P)	2.8984	2	0.05359	2	33.786	8
Pearson 6(4P)	2.7688	3	0.0551	7	87.913	17
Frechet(3P)	2.9741	4	0.0545	4	33.733	7
Dagum(4P)	2.6792	5	0.05454	3	84.16	16

In Figure 4.1, the best fit probability distribution Log – Pearson 3, for the sinkhole length and width is given by:

$$f(x) = \frac{1}{\beta\tau(\alpha)}\left(\frac{x-\delta}{\beta}\right)^{\alpha-1} e^{-(x-\delta)/\beta}$$

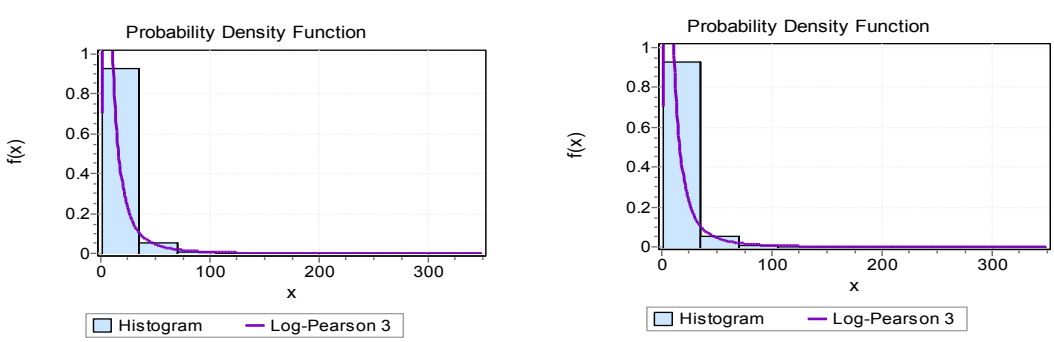

Figure 4.2: Best Fit Probability Density Function of Sinkhole Length and Width

The Figure 4.3 below shows that there is a positive or direct association between the sinkhole length and the sinkhole width; the wider the sinkhole is, the larger the sinkhole (length) will be, and the smaller the sinkhole is, then the less the sinkhole will be in length.

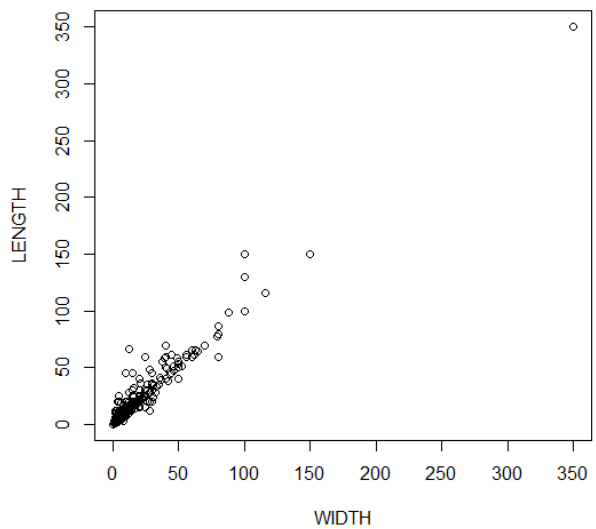

Figure 4.3: Scatterplot of the Sinkhole Length and Width

To further show that there is a relationship between the sinkhole length and width, we will create a simple linear regression model to show their correlation. The sinkhole length (x_1) will be the response variable (y) in the model, with one explanatory variable (sinkhole width). The sinkhole width (x_2) will be denoted as (x). The analytic model is denoted as:

$$y = \beta_0 + \beta_1 x.$$

The ANOVA Table 9 shows that the correlation coefficient is 0.98, which indicates a strong association between the length and width of a sinkhole. The coefficient of determination R^2 = 0.96; this means that the fitted regression equation explains 96% of the variation in y.

Table 4.5: ANOVA for Sinkhole Length and Sinkhole Width

	Estimate	t value	P - value
Intercept	0.929641	5.384	9.26e-08
Width	1.018423	167.430	2e-16
R^2 = 0.96	r = 0.98		

The developed model is: $\hat{y} = 0.93 + 1.02x_2$. In the developed model, the slope is approximately 1, which indicates that the length and width change in tandem.

To verify our findings by comparing their medians using non-parametric methods, we will perform the Wilcoxon signed rank sum non – parametric test. We will assume the data to not have a normal distribution. At a 0.05 significance level, we will decide if the sinkhole length data and sinkhole width have similar medians. Our null hypothesis is that the sinkhole length and the sinkhole width have similar medians, and the alternative hypothesis is that the sinkhole length data and the sinkhole width data have different medians.

Using Wilcoxon rank sum test, with a p – value of 0.1353 at the 0.05 level of significance, we fail to reject the null hypothesis. At the 0.05 level of significance we are certain that the sinkhole length and sinkhole width data have similar medians. The relationship between a sinkholes length and width is that they have similar medians and the same probability distributions. Therefore, width and length will be considered as estimates of the diameter. This brings us to the next hypothesis to be addressed: determine the average diameter of a sinkhole in Florida and the probability distribution that best characterizes the diameter.

AVERAGE DIAMETER OF A SINKHOLE

The second research question/statement to be addressed is to determine the average diameter of a sinkhole in Florida and the probability distribution that best characterizes the diameter. According to St. John's Water Management District in Southwest Florida, most sinkholes have a diameter between 10ft and 12 ft. This may be common in certain counties in Florida, since St. John's Water Management District only covers northeast and east – central Florida counties. Thus, even in those areas of Florida the sinkholes have a diameter between 10ft and 12 ft, this is not true for all of Florida. We will estimate the average diameter of a sinkhole using confidence intervals. First we need to find the best fit probability distribution for the diameter in order to find its parameter estimates to be used in calculating an appropriate confidence interval. For our data set it will be compared and ranked against 65 continuous distributions, where the goodness-of-fit tests (Anderson-Darling, Kolmogorov-Smirnov, and Chi-Square) will be performed. Using Maximum Likelihood Estimates, among the 65 different continuous distributions were taken into account, it was found that the best fit probability distribution for the diameter was the Log Normal distribution.

The Log Normal distribution has two parameters, (μ, σ). The probability distribution for the Log Normal is given by:

$$f(x) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma x} \exp\left(-\frac{[\ln(x) - \mu]1^2}{2\sigma^2}\right), x \in (0, \infty).$$

The best fit probability distribution for the diameter of a sinkhole can be seen in Figure 4.4.

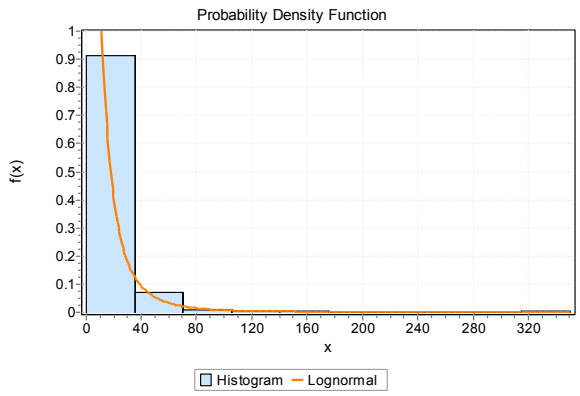

Figure 4.4: Best Fit Probability Distribution for the Diameter of a Sinkhole

The MLE of (μ, σ), for our sample is $\hat{\mu} = \frac{1}{n} \sum_{j=1}^{n} \log(x_j)$

and $\widehat{\sigma^2} = \frac{1}{n} \sum_{j=1}^{n} (\log(x_j) - \hat{\mu})1^2$ then the MLE of the mean is $\hat{\delta} = e^{\hat{\mu}\frac{1}{2} + \frac{\sigma^2}{2}}$. By resampling we obtain a bootstrap sample of $\hat{\delta}$. The maximum likelihood estimators for the parameters are as follows: $\hat{\mu}$= 0.57, $\widehat{\sigma^2}$ = 1.87, $\hat{\delta}$ = 11.02. Using a 95% confidence level, the upper and lower confidence limits are calculated by:

$$\delta'_U = \hat{\mu} + \frac{1}{\sqrt{2\pi}} \int_0^{\infty} e^{-\frac{x^2}{2}} dx\, \sigma$$

$$\delta'_L = \hat{\mu} - \frac{1}{\sqrt{2\pi}} \int_0^{\infty} e^{-\frac{x^2}{2}} dx\, \sigma$$

The upper confidence limit was found to be $\delta'_U = 14.36$ and the lower confidence limit was found to be δ'_L = 10.21. Hence a 95% confidence interval for the diameter of a sinkhole is (10.21, 14.36). We can conclude that at the 95% confidence level, the average diameter of a sinkhole in Florida may be between 10.21 ft and 14.36 ft. Using a 99% confidence level, the

upper and lower confidence limits were found to be be $\delta'_U = 15.78$ and the lower confidence limit was found to be δ'_L = 11.32. Thus a 99% confidence interval for the average diameter of a sinkhole is (11.32, 15.78). We can conclude that at the 99% confidence level, the average diameter of a sinkhole in Florida may be between 11.32 ft and 15.78 ft. Next, we will address the third research question: Determine the probability a sinkhole on a certain soil type, given the sinkhole length, sinkhole width, depth and slope.

PROBABILITIES OF A SINKHOLE OCCURRING

We will use multinomial logistic regression to further address our research statement. In the multinomial logit model we assume that the log-odds of the response follow a linear model of the logistic transformation of the odds (logit) that will serve as the dependent variable

$y_j = logit(p_j) = \log\left(\frac{p_j}{1-p_j}\right) = \alpha_j + \beta_{1j}x_1 + \cdots + \beta_{pj}x_p$, To find our probabilities, we will use the probabilistic analytic form of a logistic model is denoted as the following:

$$p = \frac{e^{-(\alpha_j+\beta_{1j}x_1+\cdots+\beta_{pj}x_p)}}{1+e^{-(\alpha_j+\beta_{1j}x_1+\cdots+\beta_{pj}x_p)}}.$$

In the following Table 4.6 the average values for the sinkhole length, width, depth, and slope conditions on all the soil types are given.

Table 4.6: Average Values of Sinkhole Length, Width, Depth, & Slope for the Soil Types

Variables	**Mean**	**Min**	**Max**
Sinkhole Length x_1	14.13	0.50	350
Sinkhole width x_2	12.96	0.50	350
Sinkhole depth x_3	9.6	0.10	170
Sinkhole slope x_4	79	40	165

The first model that we will consider is a model that has up to four way interaction between the predictor variables. This analytic model is denoted as:

$$y_j = \alpha_j + \beta_{1j}x_1 + \beta_{2j}x_2 + \beta_{3j}x_3 + \beta_{4j}x_4 + \beta_{5j}x_1x_2 + \beta_{6j}x_1x_3 + \beta_{7j}x_1x_4 + \beta_{8j}x_2\,x_3 + \beta_{9j}x_2x_4 + \beta_{9j}x_3x_4 + \beta_{10j}x_1x_2x_3 + \beta_{11j}x_1\,x_2x_4 + \beta_{12j}x_1x_3x_4 + \beta_{13j}x_2x_3\,x_4 + \beta_{14j}x_1x_2x_3x_4$$

After computing the Maximum Likelihood estimates for our model above, we found that the intercept was significantly contributing by 1, 5, and 10%. However, there were predictor variables that were not significantly contributing by the 1, 5, and 10% values. The predictor variables that were significantly contributing were: $x_1, x_2, x_3, x_4, x_1x_3$. Thus, we dropped all the terms that were not significantly contributing and considered the developed model without it. The new model is denoted as: $y_j = \alpha_j + \beta_{1j}x_1 + \beta_{2j}x_2 + \beta_{3j}x_3 + \beta_{4j}x_4 + \beta_{5j}x_1x_2 + \beta_{6j}x_1x_3$,

Since there the soil type sand had the highest frequency of occurrence, then there will be 4 different developed models that arise from our multinomial base model. The most interesting and significant models that we will discuss is the developed model for the soil type sand (combined and mixed) and (with clay), because its measurement of accuracy is over 50%, whereas the developed models for the other soil types are under 20%.

The developed model for determining the probability of a sinkhole in sand (all combined and mixed), given the sinkhole length, sinkhole width, depth and slope:

$$\hat{y}_1 = 3.43 + 0.03x_1 - 0.04x_2 + 0.28x_3 - 0.09x_4 + 0.34x_1x_3.$$

This developed model had a measurement of accuracy of 85%. We can predict the probability of sinkhole occurring in sand, using the average values of all soil types from Table 4.6. This means when x_1 is 14.13, x_2 is 12.96, x_3 is 9.6, x_4 is 79. The probability of a sinkhole occurring in sand is $\hat{p}$ = 0.82. Thus, there is a 82% chance of a sinkhole occurring in sand when the length is 14.13, width is 12.96, depth is 9.6, and slope is 79. This is a relatively high probability for the chances of a sinkhole occurring, given the average conditions for the sinkhole length, sinkhole width, depth and slope.

The developed model for determining the probability of a sinkhole in sand (with clay), given the sinkhole length, sinkhole width, depth and slope:

$$\hat{y}_2 = 2.72 + 0.02x_1 - 0.03x_2 + 0.17x_3 - 0.07x_4 + 0.15x_1x_3.$$

This developed model had a measurement of accuracy of 78%. We can predict the probability of sinkhole occurring in sand, using the average values of all soil types from Table 4.6. This means when x_1 is 14.13, x_2 is 12.96, x_3 is 9.6, x_4 is 79. The probability of a sinkhole occurring in sand mixed with clay is $\hat{p}$ = 0.73.

This means that there is 73% chance that a sinkhole will occur in the mixed soil type of sand and clay, when the length is 14.13, width is 12.96, depth is 9.6, and slope is 79. This is a relatively medium probability for the chances of a sinkhole occurring, given the average conditions for the sinkhole length, sinkhole width, depth and slope.

USEFULNESS & CONTRIBUTIONS

Using parametric and nonparametric statistical methods, we have found the length of a sinkhole is not significantly different as the width of a sinkhole; following the same probability distribution. Simple Linear Regression further shows that length and width can be considered measurements of the diameter, which allows us to fit the probability distribution of the diameter. The probability distribution that was best characterizes the sinkhole diameter can be used to find confidence intervals. Comparing our results from the information from St. John's Water Management District in Southwest Florida, that most sinkholes have a diameter between 10ft and 12 ft, we conclude that we are 95% confident that the average diameter of a sinkhole in Florida may be between 10.21 ft and 14.36 ft. Also, we found at the 99% confidence level, the average diameter of a sinkhole in Florida may be between 11.32 ft and 15.78 ft. This is new knowledge that may help the citizens of Florida better understand the probable size of sinkholes in Florida. The final developed model was the first of its kind to estimate the probability of a sinkhole occurring in a given soil type, as a function of the outlined dimensions. This is useful because it provides a better insight to the relationship among the length, width, depth, slope, and soil type of a sinkhole.

CHAPTER 5: SURVIVAL ANALYSIS OF FLORIDA SINKHOLES

This chapter is an extension of the previous chapter on the occurrence of sinkholes in Florida. We are interested in the time to event between the occurrences of sinkholes in Florida, and evaluating their probable measures. One statistical field that is relevant to understanding and determining time to event occurrences is survival analysis. In this chapter we will we interested in the variable

Time to Event (TTE). TTE (Time to Event) is the measurement of time between the recorded occurrences of sinkholes in Florida. The soil types under consideration are **sand, clay, unknown, lime,** and **rock**.

The research questions to be addressed in this section are:

1) Determine the probable TTE, based upon the Kaplan - Meier estimate.
2) Determine the probable TTE (in soil types), based upon the Kaplan – Meier estimate.
3) Determine the best probability distribution that characterizes the time to event between occurrences of sinkholes.
4) Determine the associative covariates in sinkhole occurrences in Florida.

PROBABLE KAPLAN MEIER ESTIMATE TTE.

We will use a Kaplan - Meier analysis (nonparametric methods), to aid in the assistance of addressing the probable TTE. Using the Kaplan – Meier method, which is a nonparametric

estimator of the survival function, and is widely used to estimate and graph survival probabilities as a function of time, we obtain the following graphs in Figures 5.1 and 5.2.

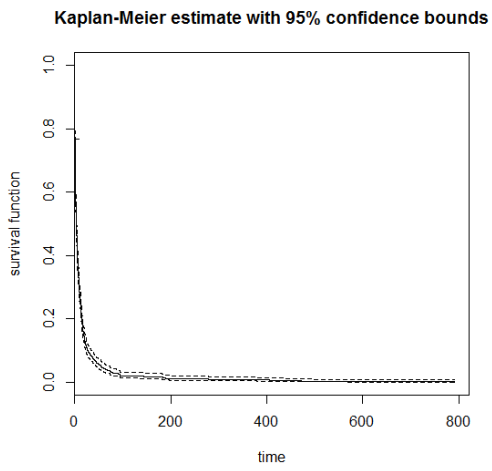

Figure 5.1: Kaplan – Meier Graph

The Figure below shows all 38 years of recorded sinkhole occurrences. When t =1, this means that there was a sinkhole that occurred the previous day. The number of times that this event happened was 107 times in the last 38 years. The Kaplan – Meier point estimate at t = 1 is

$\hat{S}(t = 1) = \frac{683}{790} = 0.86$. This means that there is an 86% chance that we survived a day without a sinkhole occurring, and there is a 14% chance of a sinkhole occurring the next day.

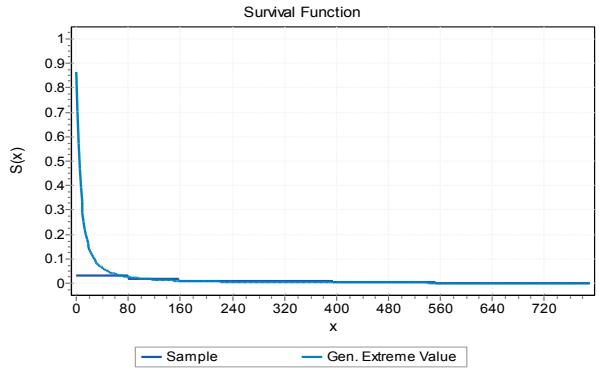

Figure: 5.2: Survival Function of TTE

Notice that as time t gets larger and larger, the likelihood of surviving without a sinkhole decreases, and the probability of a sinkhole increases. Meaning, that as more days are between sinkhole occurrences, then the greater chance another sinkhole occurring. We can conclude that when there is more than 10 days in between a sinkhole occurrence ($t > 10$), then a sinkhole occurrence is highly likely. This leads us to our next research statement, determine the probable TTE (in soil types), based upon the Kaplan – Meier estimate.

PROBABLE KAPLAN – MEIER ESTIMATE OF SOIL TYPE

In this section we will be looking at when a sinkhole has occurred in a certain soil type. In the following Figure, the KM graph represent the TTE probabilities that have occurred in a soil type.

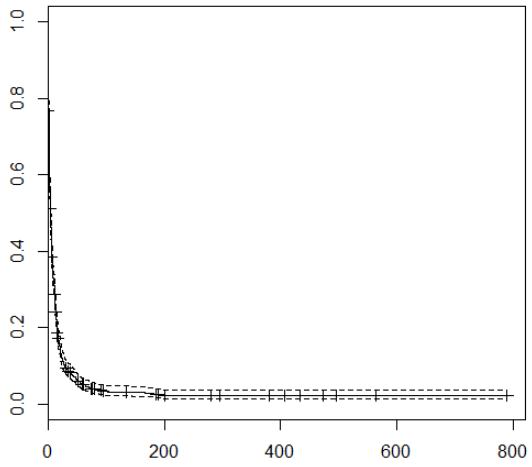

Figure: 5.3: Kaplan – Meier Survival Probabilities of TTE (censoring)

In Figure 5.4, the Kaplan – Meier graphs of the survival probabilities of TTE in the different soil types is shown.

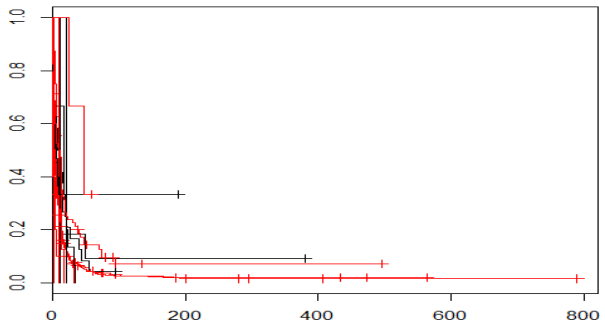

Figure 5.4: Kaplan – Meier graph of TTE in the Soil Types

In Figure 5.5 the histogram for TTE is shown, along with the Kaplan – Meier graphs for when a sinkhole has occurred in a certain soil type. From the previous chapter we found that the soil type sand had the largest occurrence of sinkholes in Florida. Therefore, the TTE in the soil types that are of interest are the TTE in sand. In Figure 5.5, second KM graph for the soil type sand is shown, where the survival function probabilities are shown to range from 0 to 0.71. The Kaplan – Meier point estimate at t = 1 is $\hat{S}(t = 1) = \frac{317}{729} = 0.43$. This means that there is a 43% chance that we survived a day without a sinkhole occurring in sand, and there is a 57% chance of a sinkhole occurring in sand the next day. As t becomes larger, the greater chances there are of a sinkhole occurring within the soil type sand. In regards to censoring the TTE (yearly), preliminary studies show us that around 3 months in the TTE data, there is a seasonal effect as to when sinkholes occur more often in the year. In the future, we will also consider using Poisson processes.

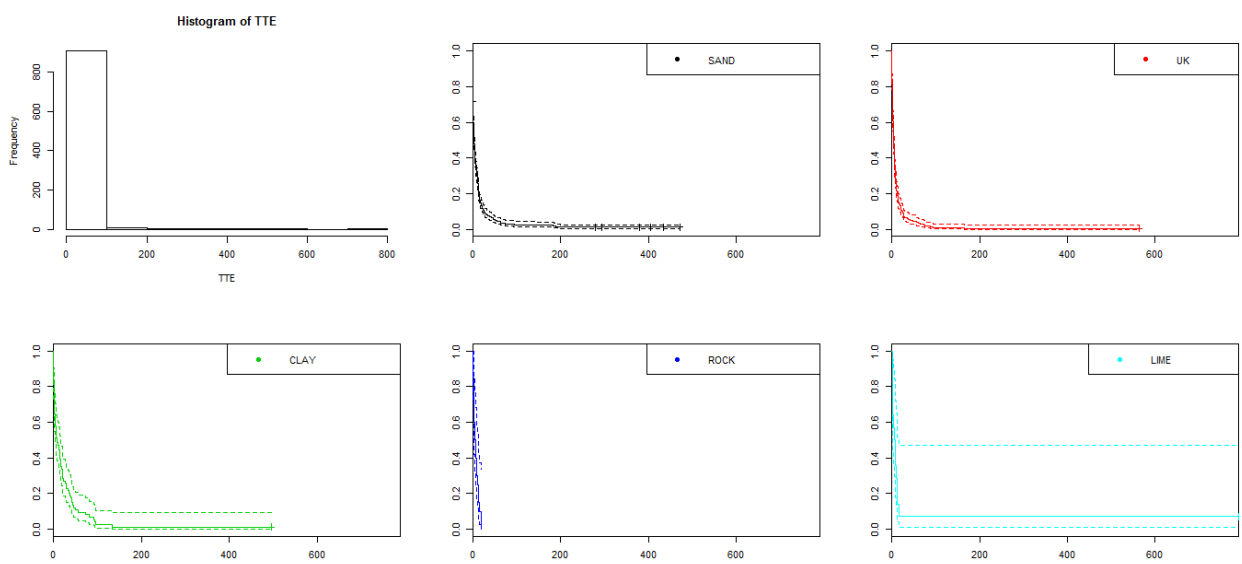

Figure 5.5: Kaplan – Meier Survival Probabilities of TTE in Soil Types

PROBABILITY DISTRIBUTION THAT BEST CHARACTERIZES THE TTE

The next research question/statement to be tested is to determine the probability distribution that best characterizes the time to event between occurrences of sinkholes. The variables of interest in this research question are TTE (Time to Event). To address this research question/statement we will use parametric analysis. For our two data set, it was compared and ranked against 65 continuous distributions, where the goodness-of-fit tests (Anderson-Darling, Kolmogorov-Smirnov, and Chi-Square) was performed. Using Maximum Likelihood Estimates, among the 65 different continuous distributions that were taken into account, it was found that the best fit probability distribution the TTE was the *Fréchet* distribution (from the General Extreme Value Distribution, (Table 5.1).

Table 5.1: Goodness – of – Fit - Tests for the Best Fit Distributions for the TTE

	Anderson – Darling		Kolmogorov-Smirnov		Chi - Square	
Distribution	Statistic	Rank	Statistic	Rank	Statistic	Rank
GEV - Fréchet	15.782	1	0.1343	1	33.981	4
Wakeby	13.134	2	0.1346	2	33.243	1
General Pareto	13.134	3	0.1346	3	33.243	2
General Logistic	15.628	4	0.1410	4	33.352	3
Inverse Gaussian	186.63	5	0.2376	8	237.36	6

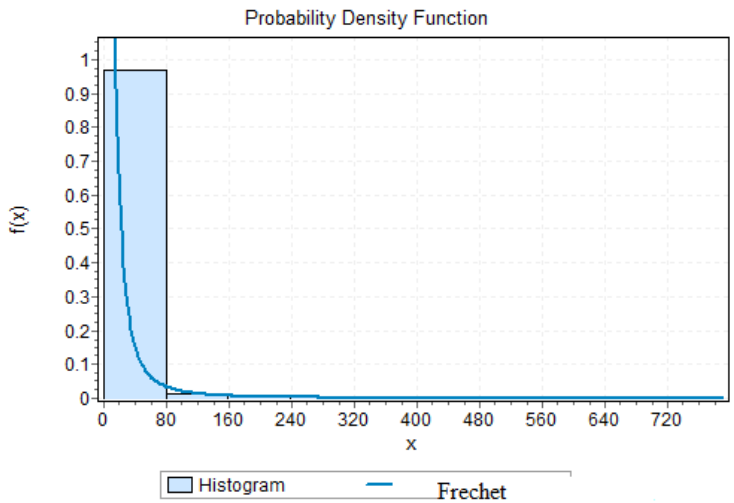

Figure 5.6: Best Fit Probability Distribution of TTE

Hence, we can conclude that the best fit probability distribution associated with time to event between occurrences of sinkholes in Florida is the General Extreme Value Fréchet distribution.

ASSOCIATION OF COVARIATES OF A SINKHOLE

In this section we will use the Semi Parametric Method of the Cox Proportional Hazards Regression analysis. From the previous chapter ranked the soil types based on their depth of the number of occurrences of sinkholes, therefore we will use depth and soil types as covariates in this section. The variables of interest in this section are **depth (x_{11}), lime (x_{12}), sand (x_{13}), rock (x_{14}),** and **clay (x_{15}).**

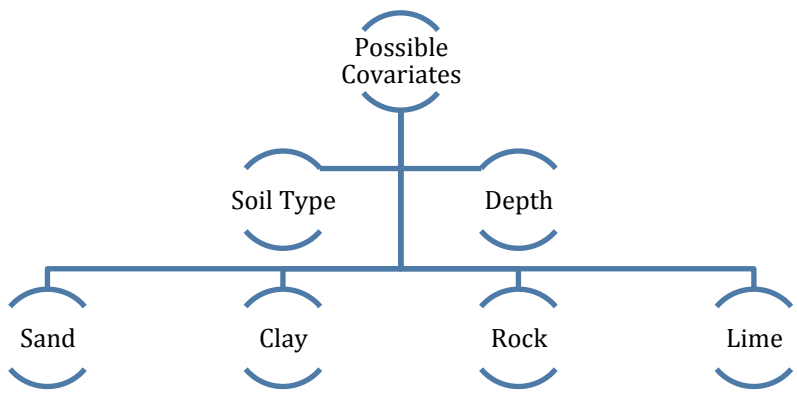

Figure 5.7: Variables of Interest

The analytical model is denoted as:

$$h_i(t) = h_0 \exp\left(\beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_{13} + \beta_4 x_{14} + \beta_5 x_{15}\right)$$

The hazard ratio is denoted as:

$$\text{HR} = \frac{h_i(t)}{h_j(t)} = \frac{h_0(t)\exp\left(\beta_1 x_{1i} + \cdots + \beta_k x_{ik}\right)}{h_0(t)\exp\left(\beta_1 x_{1j} + \cdots + \beta_k x_{jk}\right)}$$

Primary interest lies in estimating the parameter β using the partial likelihood:

$$\text{L}(\beta) = \prod_{i=1}^{D} \frac{\exp\left[\beta' z_i\right]}{\sum_{j \in R_{ti}} \exp\left\{\beta' z_j\right\}}$$

"The MLE $\hat{\beta}$(a vector) is asymptotically N(β, I^{-1}), where I represents the Fisher information" [6]. A local test examined a subset of the elements of β, testing the claim that depth does not

depend upon the soil type (null). The alternative is that depth does depend upon the soil type. Two covariates were used to help address this research statement, such as depth and soil type.

In the following Table, estimates of the β_k, including standard errors and p-values for each test, an estimate of the risk ratio and its confidence interval, plus the p-values for likelihood ratio, Wald, and score tests for the global null are shown.

Table 5.2: Calculations from the Cox PH Model

Covariates	Coefficients	Exp Coefficients	SECoefficients	z	p
Depth	-0.008	0.991	0.002	-3.371	0.007
Lime	-0.018	0.981	0.309	-0.062	0.954
Sand	0.4221	1.525	0.253	1.674	0.096
Rock	0.397	1.489	0.125	3.182	0.001
Clay	0.410	1.508	0.131	3.134	0.001
LRT	26.5 on 5 df				0.000
Wald Test	24.7 on 5 df				0.000

Developed Cox Ph Model for the Associated Covariates

The developed Cox model is as follows:

$$\hat{h}_1(t) = 2\exp(-0.008x_{11} - 0.018x_{12} + 0.4221x_{13} + 0.397x_{14} + 0.410x_{15}).$$

From Table 5.2, we can see that the parameter β was calculated using the partial LRT, and its value is 26.5, with a p - value of 0. This means that since L(β) = 26.5, with a p - value of 0.00, then we reject the null hypothesis. Also, since the Wald Test was performed and it's statistic is 24.7 with a p – value of 0.00, we reject the null hypothesis. This means that depth does depend upon the soil type of the sinkhole. Also, since depth is covariate 1, it will have an effect on the soil type (which is covariate 2). Now, we will discuss the interesting hazard ratios over the value of 1. Notice that the hazard ratio for sand and clay are the highest.

Hazard Ratio for the Associated Covariates

In this section we will discuss the hazard ratios for the soil type sand, rock, and clay and their associated covariate depth. The hazard ratio for sand is HR = 1.525. This means that for every unit increase in depth (feet), the likelihood that a sinkhole occurs in sand increases by a factor of 0.525. The hazard ratio for clay is HR= 1.502. This means that for every unit increase in depth (feet), the likelihood that a sinkhole occurs in clay increases by a factor of 0.502. The hazard ratio for rock is HR= 0.489. This means that for every unit increase in depth (feet), the likelihood that a sinkhole occurs in rock increases by a factor of 1.489. The hazard ratio for lime is HR = 0.981. This means that the unit increase in depth will not be as profound as it was in the other soil types. This make sense since lime had a very low frequency of sinkhole occurrences.

USEFULNESS & CONTRIBUTIONS

The findings in this chapter are useful in predicting the probable time to event between sinkhole occurrences. The developed Kaplan - Meier model is useful in determining the probable time between sinkhole events, and also as a function of soil types. For instance, we can predict the

probability that a sinkhole will occur tomorrow, the next day, etc. The Cox Ph model is useful in predicting the probable time to event between sinkhole occurrences taking into account the associated covariates, such as depth and soil types. This also allows the hazard ratios to be computed, which determines the increase in the likelihood of the sinkhole occurrence over time.

One contribution to the field of Applied Statistics in Environmental Studies is the application of the Cox Ph model to sinkholes, which has not been found in any literature review. This can be useful to many citizens of the state of Florida because they will have better understanding of when the probable time to event of sinkholes occurs in either sand or clay. The analysis shows that the longer the time to event between events, then the deeper the sinkhole. In conclusion, we hope to look further into this research to detect what time of day these sinkholes have occurred during the season.

CHAPTER 6: FUTURE PROJECTS AND WORKS

In the future one of the projects that the author would like to invest time into is the creation of a hurricane tracking application (for smart phones). The goal is that this app will be able to provide substantial statistical information on where a hurricane may land in the state of Florida. For the past two years, there has been countless big data sets merged and complied by the author and her mentor Dr. Rebecca Wooten that would be used in the creation of this app. So far, the implementation of these big data sets has provided useful information on how to handle data that is messy or missing information. In the summer of 2016, another project that the author is interested in pursuing is the writing of a book with the topic of statistics and fitness. Since the author is a credentialed personal trainer and has taught aerobics for over 15 years, then the collaborative ideas of utilizing statistics and fitness is very exciting. In the aspect of further studying environmental issues, the author has always had a passion for analyzing turtle nests and building a video gram tracking app device that will show when hatchings have escaped or become prey for other animals. Specifically the app will be called TNT: Turtle Nesting Tracking App for online viewing or Android, Smartphones.

Browser compatibility: google chrome, Firefox, internet explorer.

Purpose:

TNT is a descriptive - qualitative data based program designed to bring awareness of the Florida sea turtle nesting trends in certain counties. Currently, there are five species of sea turtles that inhabit Florida's beaches. The counties of interest with the largest sea turtle nesting sites are:

Sarasota, Charlotte and Collier. This app will provide data collected over the last five years on sea turtle nesting, hatching, and false crawls.

Usefulness: 1) To inform citizens in the counties of Sarasota, Charlotte and Collier about their areas turtle nesting trends.

2) To educate and enlighten others of the ongoing struggles of sea turtle survival.

3) To spark interest in creating new solutions to help these counties have larger sea turtle survival rates.

4) To provide viable information to recruit new volunteers.

In fall 2016, another future project that the author is interested in working on is analyzing coral reef data and building a statistical model that will help predict the population growth trends between the coral reefs in the Florida Keys.

REFERENCES

[1] Abdi, Herve. "Factor Rotations in Factor Anaylsis." University of Texas at Dallas. January 2016. Retrieved from http://ftp.utdallas.edu/~herve/Abdi-rotations-pretty.pdf

[2] Andersen, Erling B. *Discrete Statistical Models with Social Science Applications*. North Holland, 1980.

[3] Alves, I. "Extreme Value Distributions." Ceaul , Deio. Faculty of Sciences, University of Lisbon. Retrieved from FragaAlves.pdf

[4] Casella ,G., Berger, R. *Statistical Inference*. Second Edition. Duxbury Thomson Learning. Pgs. 324 - 495. 2002.

[5] Duda, R. *Pattern Classification*. John Wiley & Sons, Pg. 101, Nov 9, 2012

[6] Elandt-Johnson, Regina; Johnson, Norman. *Survival Models and Data Analysis*. New York: John Wiley & Sons, 1999.

[7] Geisser, S; Johnson,W.M. *Modes of Parametric Statistical Inference.* New York: John Wiley & Sons, 2006

[8] Machos, G. "The Story of the Power and Fury of Hurricane Andrew." August 2014. Retrieved from http://www.hurricaneville.com/andrew.html.

[9] Marr, P. "Directional Circular Statistics". Shippensburg University. November 2015. Retrieved from http://webspace.ship.edu/pgmarr

[10] Parker, P. "Worst Hurricanes to Hit Florida in the Past Century." August 2014. Retrieved from http://pparker.org/hurricanes/hurricane_history.htm

[11] Quick, J. "R Tutorial Series: Exploratory Factor Analysis." January 2016. Retrieved from http://rtutorialseries.blogspot.com/2011/10/r-tutorial-series-exploratory-factor.html

[12] Schneider, M. "Florida's Porous Peninsula Leads to Sinkholes". August 13, 2013. August 2014. Retrieved from http://www.cnsnews.com/news/article/floridas-porous-peninsula-leads-sinkholes

[13] Sharron, H. "Types of Florida Sinkholes." Santa Fe College. August 2014. Retrieved from http://dept.sfcollege.edu/natsci/physsci/jean.klein/Cave/cave12.htm

[14] Statistics Solutions. "Multiple Linear Regression". July 2014. Retrieved from http://www.statisticssolutions.com/what-is-multiple-linear-regression/

[15] Stigler, S. *The Epic Story of Maximum Likelihood.* Statistical Science Vol 22, No.4, 598 - 620, 2007

[16] Strickland, J. *Operations Research Using Open-Source Tools.* Lulu Publishing. Pgs. 505-515. 2015.

[17] Sohel,R., Habshah,M., Sarkar,S.K. "Determinants of Desire for Children: A Multinomial Logistic Regression Approach". LifeScience Journal, Vol 10, No. 2. 1-8, 2013.

[18] Wikipedia. "Non-ParametricStatistics". March 2015. Retrieved from https://en.wikipedia.org/wiki/Nonparametricstatistics

[19] Wikipedia. "Simple Linear Regression". July 2014. Retrieved from https://en.wikipedia.org/wiki/Simplelinearregression

[20] Wooten, R. D., Baah K, & D'Andrea, Joy. "Implicit Regression: Detecting Constants and Inverse Relationships with Bivariate Random Error", arXiv.org – in affiliation with Cornell University Library, eprint arXiv:1603.07948

[21] Wooten, R. D., D'Andrea, Joy. "Modeling Hurricanes using Principle Component Analysis in Conjunction with Non-Response Analysis", arXiv.org – in affiliation with Cornell University Library, eprint arXiv:1512.05307

[22] Yan, X. *Linear Regression Analysis: Theory and Computing*. World Scientific. Pgs. 291 - 300. 2009