FORECASTING THE VALUE OF SINGLE FAMILY HOMES AS A RESULT OF HURRICANE FLORENCE

An Undergraduate Research Scholars Thesis

by

CHRISTOPHER JAMES WALKER

Submitted to the Undergraduate Research Scholars program at
Texas A&M University
in partial fulfillment of the requirements for the designation as an

UNDERGRADUATE RESEARCH SCHOLAR

Approved by Research Advisor: Dr. Michelle Meyer

May 2019

Major: Urban and Regional Planning

TABLE OF CONTENTS

ABSTRACT

Forecasting the Value of Single Family Homes as a Result of Hurricane Florence

Christopher James Walker
Department of Landscape Architecture and Urban Planning
Texas A&M University

Research Advisor: Dr. Michelle Meyer
Department of Landscape Architecture and Urban Planning
Texas A&M University

The purpose of this study is to investigate and forecast the impact of hurricane Florence on home prices in North Carolina. Quantitative data was gathered from real estate comparison websites such as Zillow and Redfin, while qualitative data from news reports and FEMA disaster declarations were used to better analyze the impact of hurricane Florence. Using both qualitative and quantitative sources, an analysis was conducted to compare historical home values with home values immediately following hurricane Florence. Using methodologies from past studies, forecasts were made for each impacted zip code that estimate the median home value one year following hurricane Florence to better understand the rate at which different areas recover. Charts and maps were used to visualize historical trends as well as forecasts. By conducting this data analysis study, similar studies on focused on other hurricanes can be used to compare the impact of Hurricane Florence to other hurricane events. Conducting this research serves to enhance the growing literature on the economic impact of hurricanes.

CHAPTER I

LITERATURE REVIEW

As this research project is primarily focused on predicting home prices, using well developed and applicable financial modeling techniques is critical for the success of the forecasts described herein. Furthermore, using effective and efficient data storage techniques also has an impact upon the scale and speed at which forecasts can be made. For these reasons, substantial research has been conducted to understand which forecasting models and data storage methods would serve this project most effectively.

As previously mentioned, the forecasting models used to generate home value forecasts can have a significant impact on the validity of the results produced. As a result, a comparison of common machine learning, statistical, and classical models was conducted from a variety of literature sources. Machine learning, as the name implies, refers to developing methods to have machines (computers) develop models that can be used to categorize and predict data by learning from historical data. Simplistic machine learning models blur the lines between classical statistics and machine learning. Models that fall into the simplistic category may include linear regression while more advanced models including neural networks would fall into a more complex category of machine learning. That said, using a more complex model does not always yield more accurate results. As is the case with many things in life, there are decreasing marginal returns to added complexity.

An article written by computer science professors from all over the world provides a high-level overview of current machine learning models. This article titled “An Empirical Comparison of Machine Learning Models for Time Series Forecasting” begins by introducing the history of three modern machine learning models and how they compare to classical statistical models. While the authors mention that statistics shares many attributes with machine learning, there are distinct differences (Ahmed et al., 2010). In their discussion, the authors introduce neural-networks, k-nearest neighbor models, and regression trees. Furthermore, the authors discuss how changes in the preprocessing of data impacts the outcome. Regardless of preprocessing, the authors found neural-networks, a more complex machine learning model, to be the most effective out of the models tested. The authors conclude that while neural-networks were most effective model, the type of data used to test the models was strictly business and economics related. Other models may be more effective for different disciplines such as chemistry, physics, or engineering.

One machine learning model that was not mentioned in the previous article goes by the name of Support Vector Machines (SVMs). SVMs are gaining popularity for financial forecasting but have traditionally been used for classification problems including handwriting classification and image classification in Google images. Researchers from MIT present an interesting case as to why SVMs may be effective when working with financial data (Cao and Francis, 2003). They compare SVMs for time series forecasting to several other established models in the space such as neural networks. Overall, the researchers find the unique characteristics of SVMs to be promising for financial data forecasting. Despite this, the researchers say that the parameters used to develop a model for time series forecasting may be more sensitive than neural network counterparts and may require greater fine-tuning of the model as to not overfit the data.

As important as the machine learning model is, all models must be housed. Whether it is a simple regression in Excel or a neural network in Python, the models are only effective if they are usable and accessible. For the purposes of this research, Python and Scikit-learn will be the primary tools used for data analysis. Python is a general-purpose programming language used for a variety of applications. From web development to data science, someone is using Python to do it. One particularly good aspect of Python is its wide support for data science libraries including Scikit-learn. Data in Python is stored in lists, tuples, dictionaries, arrays, and most importantly dataframes. Dataframes come from a library called Pandas (McKinney, 2010). Much like a spreadsheet or database, data in Pandas dataframes have columns and rows. Often, tasks commonly performed in spreadsheets can be performed faster and more accurately in a Pandas dataframe. For this reason, dataframes are often the data structure of choice for statistical computing. Data used in this research project is imported into a Pandas dataframe and analyzed using multiple linear regression in Scikit-learn.

Scikit-learn is a popular open source machine learning library for Python. It features many of the most popular machine learning models including k-nearest neighbors, multiple linear regression, SVMs, and neural networks. Scikit-learn provides the flexibility, documentation, and community support to utilize many models rapidly (Pedregosa, 2011). For this reason, Scikit-learn is critical for this research as it provides the tools to try different forecasting models when predicting the prices of homes impacted by hurricane Florence at an aggregate and individual level.

CHAPTER II

METHODS

Data Preparation

Any data science project is only as good as the data used in the analysis. For this reason, data must be collected, cleaned, and analyzed properly to derive accurate findings and results. In the case of this project, data was collected from Zillow Research, a division of the Zillow Group that publishes aggregated data on the median home price in zip codes across the United States. To build and accurate forecasting model, a list of several past hurricane events was collected along with the zip codes they impacted. Table 1 showcases the hurricane events that were used in this analysis in addition to the number of zip codes impacted by that storm.

Table 1. Past Hurricane Events

Hurricane	Max Wind	Category	Month	State	Zip Codes
Ike	145	4	2008-9	TX	245
Isaac	80	1	2012-8	LA	102
Irene	120	3	2011-8	NC	93
Matthew	165	5	2016-9	FL	118
Harvey	130	4	2017-08	TX	261
Sandy	115	3	2012-10	NJ/NY	902
Irma	180	5	2017-8	FL	676

Once a list of past hurricane events was assembled, a review of official FEMA report for each hurricane events was completed. These reports indicate which counties received aid from FEMA following each hurricane. Once the review was completed, the median home price data for each zip code within the impacted counties two months before, one month before, during the event, one month after, two months after, three months after, four months after, and one year after each hurricane event was pulled. Data from Zillow Research was provided in spreadsheet format. This spreadsheet included all median home prices for nearly all United States zip codes for every month since April of 1996. This data allows for the construction of a model that uses historical home price data (two months before to four months after) to predict the median home price one year following the hurricane event in each zip code. Table 2 indicates the columnar structure of the dataset, displays the purpose of each column, and provides an example data point from a zip code impacted by hurricane Ike.

Table 2. An excerpt from the aggregate home price dataset

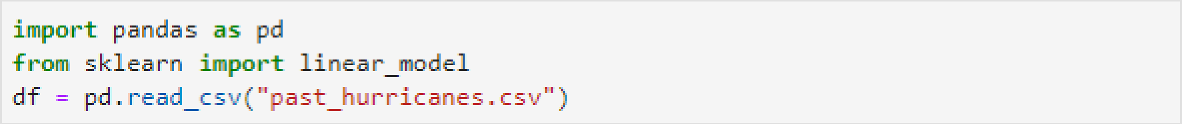
Column	**Purpose of Data**	**Example Data**
Zip Code	Unique Identifier	77494
Hurricane	Hurricane Identifier	Ike
Category	Hurricane Identifier	4
Hurricane Month	Hurricane Identifier	2008-09
State	Location Identifier	Texas
County	Location Identifier	Harris County
Two Months Before	Input Dimension	$261,500
One Month Before	Input Dimension	$260,400
Month of The Hurricane	Input Dimension	$260,000
One Month After	Input Dimension	$259,200
Two Months After	Input Dimension	$258,700
Three Months After	Input Dimension	$258,600
Four Months After	Input Dimension	$259,200
One Year After	Output – What is being predicted	$262,300

Once data had been assembled in as comma separated values (CSV) for seven hurricane events in 2087 zip codes spanning nearly a decade a forecasting model using multiple linear regression was built in Scikit-learn. Traditional linear regression makes use of statistical formulas to create a function represents the relationship between two dimensions. Because traditional regressions compare the relationship between two dimensions, they can be represented on a two-dimensional graph. Multiple linear regression, however, is different. The model in this project makes use of seven input dimensions and one output dimensions for a total of eight dimensions. It is impossible to represent eight dimensions within a single visualization.

For this reason, all visualizations of the model represent *real* outputs of the model as opposed to visualizing all *possible* outputs of the model.

Analysis

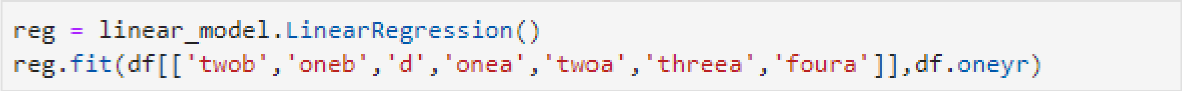
To create a model in Scikit-learn using Python, a CSV of training data must first be loaded into a Python script as a Pandas dataframe object. Using Pandas' read CSV function, a CSV was imported which contained data regarding past hurricane events into Python as shown in Figure 1.



```
import pandas as pd
from sklearn import linear_model
df = pd.read_csv("past_hurricanes.csv")
```

Figure 1. Importing a CSV for analysis into Python

Once the training data has been imported as a dataframe object, dataframe columns are selected and fitted to te model. In the case of this prediction model, the seven columns which represent historical median home price data were included as the input dimensions and the column which represents the median home price one year after the event was included as the output dimension. Because the training data includes a column which indicates the median home price in a zip code one year following a hurricane event, Scikit-learn uses regression to understand the relationship between each of the input dimensions and the one output dimension. Scikit-learn can estimate the value of the output dimension when the seven input dimensions are provided. In the case of hurricane Florence, all seven input dimensions were available, so the model can produce an estimate for the home price one year after the hurricane event. Figure 2 displays how training data is fitted to a multiple linear regression model in Python.



```
reg = linear_model.LinearRegression()
reg.fit(df[['twob','oneb','d','onea','twoa','threea','foura']],df.oneyr)
```

Figure 2. Fitting training data to a multiple linear regression model in Python

The first seven dimensions (*'twob', 'oneb', 'd', 'onea', 'twoa', 'threea', 'foura'*) in brackets refer to the input dimensions (columns) while the last dimension (*df.oneyr*) represents the dimension the model is trying to predict. When the data is fitted, Scikit-learn develops coefficients for each input dimension that when combined with an intercept form the model. This model (equation) that can be used to predict the median home price in zip codes impacted by hurricane Florence.

Scikit-learn provides a simple method to determine the coefficients and intercept of a fitted model. In this research, the model is called by the variable *reg* which is shorthand for regression. Using the *coef_* and *intercept_* methods, Python will output the coefficients in the order they were listed in while fitting the model, as well as the intercept. The model can be represented by the following equation:

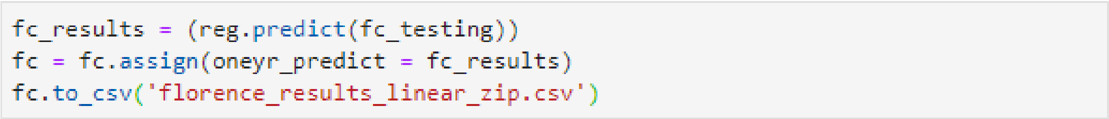
$$Predicted\ Median\ Home\ Value = (twob * -1.66) + (oneb * -5.31) + (d * 6.78) + (onea * -3.45) + (twoa * -1.15) + (threea * -0.35) + (foura * 2.84) + 1719.01$$

Following the creation of the model, data was run from hurricane Florence through the model to generate predictions. Input data from hurricane Florence looks very similar to the training data used to create the model except for the column which indicates median home prices one year after the hurricane event is blank. Table 3 indicates how hurricane Florence input data is formatted.

Table 3. An excerpt from the hurricane Florence data

Column	Purpose of Data	Example Data
Zip Code	Unique Identifier	27713
Hurricane	Hurricane Identifier	Florence
Category	Hurricane Identifier	4
Hurricane Month	Hurricane Identifier	2008-09
State	Location Identifier	North Carolina
County	Location Identifier	Durham County
Two Months Before	Input Dimension	$243,500
One Month Before	Input Dimension	$244,900
Month of The Hurricane	Input Dimension	$246,800
One Month After	Input Dimension	$249,400
Two Months After	Input Dimension	$252,300
Three Months After	Input Dimension	$255,200
Four Months After	Input Dimension	$257,200
One Year After	Output – What is being predicted	NULL

Using Scikit-learn, Python automatically applies the forecasting model to each of the rows in the hurricane Florence input dataset. Python first reads the necessary data columns, in this case the seven input dimensions, and creates an array of forecasted values. The order of the elements in this array corresponds to the order of rows inputted into the forecasting model. The array can then be applied to the input dataset as a new column as shown in Figure 3.



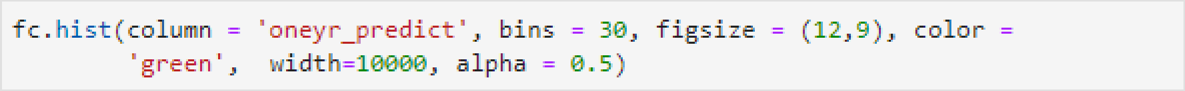
```
fc_results = (reg.predict(fc_testing))
fc = fc.assign(oneyr_predict = fc_results)
fc.to_csv('florence_results_linear_zip.csv')
```

Figure 3. Applying the forecasting model to the Florence dataset and creating a new CSV

Visualization

With the newly created dataset in place, visualizations of the data can be created to draw conclusions of the study. Given the geographic nature of the study, choropleth style maps lend themselves to zip code-oriented data. When visualizing the results of the model, three key elements were emphasized: the median price of homes in each zip code, the dollar changes in median home value one year after the hurricane event, and most importantly, the percent change in median home value one year after the hurricane event.

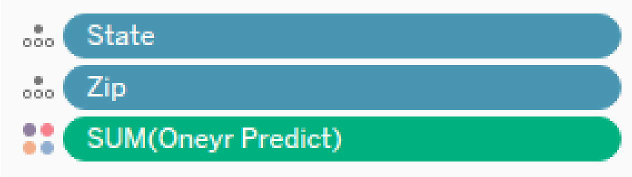
Using Python and Tableau, visualizations were created to highlight these distinct elements. Histograms are the first style of visualization used to represent this dataset. They are used to show the distribution of datasets and can be created in using the commands in Figure 4.



```
fc.hist(column = 'oneyr_predict', bins = 30, figsize = (12,9), color =
        'green',  width=10000, alpha = 0.5)
```

Figure 4. Creating a histogram of median home prices one year after hurricane Florence

The second style of visualization utilizes spatial data joined with columnar data to create filled color maps. They serve to represent values across a geographic area in localized distinct regions. The choropleth maps in this report can be created in Tableau using two spatial dimensions and one numerical metric as shown in Figure 5.



State
Zip
SUM(Oneyr Predict)

Figure 5. The marks for a choropleth map in Tableau

Visualizations serve to represent the findings of a dataset in a concise and digestible fashion. By utilizing a mixture of visualizations, the contents of the Florence dataset can be better understood.

CHAPTER III

RESULTS

As mentioned in the visualization section of Chapter II, the results of the forecast focus on three key topics: the forecasted median home price, the absolute change in median home price, and the percent change in median home price. To visualize the results, choropleth maps and histograms are used to visualize the forecasts geographically and as a distribution.

Median Home Values

When analyzing the median home price of over one hundred zip codes, it is useful to visualize the distribution as a histogram in combination with summary statistics. Figure 6 displays the distribution of the forecasted median home prices.

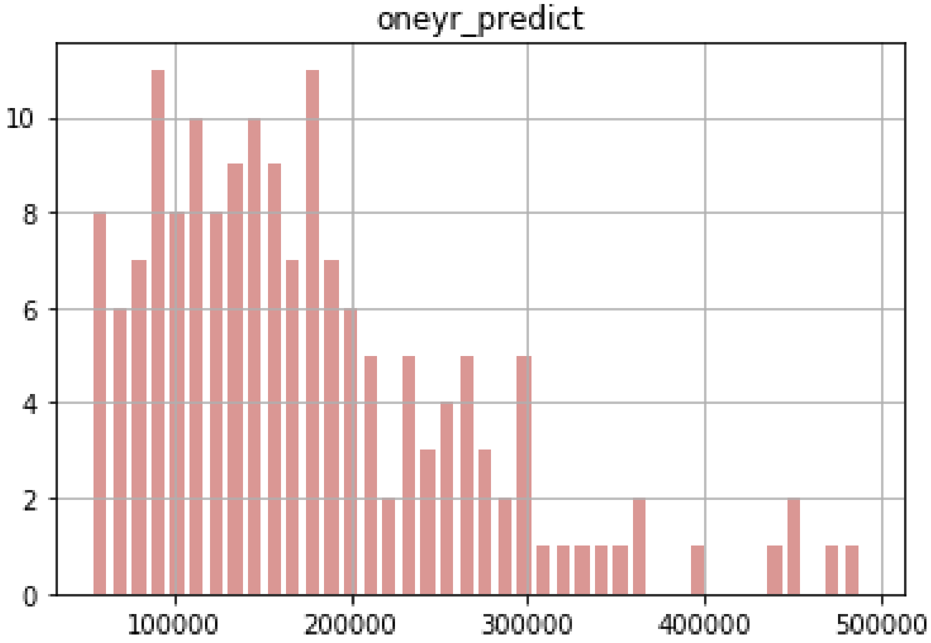


Figure 6. The distribution of forecasted median home prices

Figure 6 reveals a distribution skewed to the right. Of the 164 records in the dataset, the mean median home value is $177,532.51 with a standard deviation of $90721.24. Using three standard deviations from the mean as a measure to determine outliers, only two zip codes fall into this category with a predicted median home price greater than $449,696.11. These zip codes fall in Orange and Carteret counties.

When represented on a map, the distribution can be represented geographically highlighting areas with very high median home prices and very low median home prices. Areas with high concentrations of high median home prices include Myrtle Beach and Chapel Hill, North Carolina. In the gaps within the map, data from Trulia was not available.

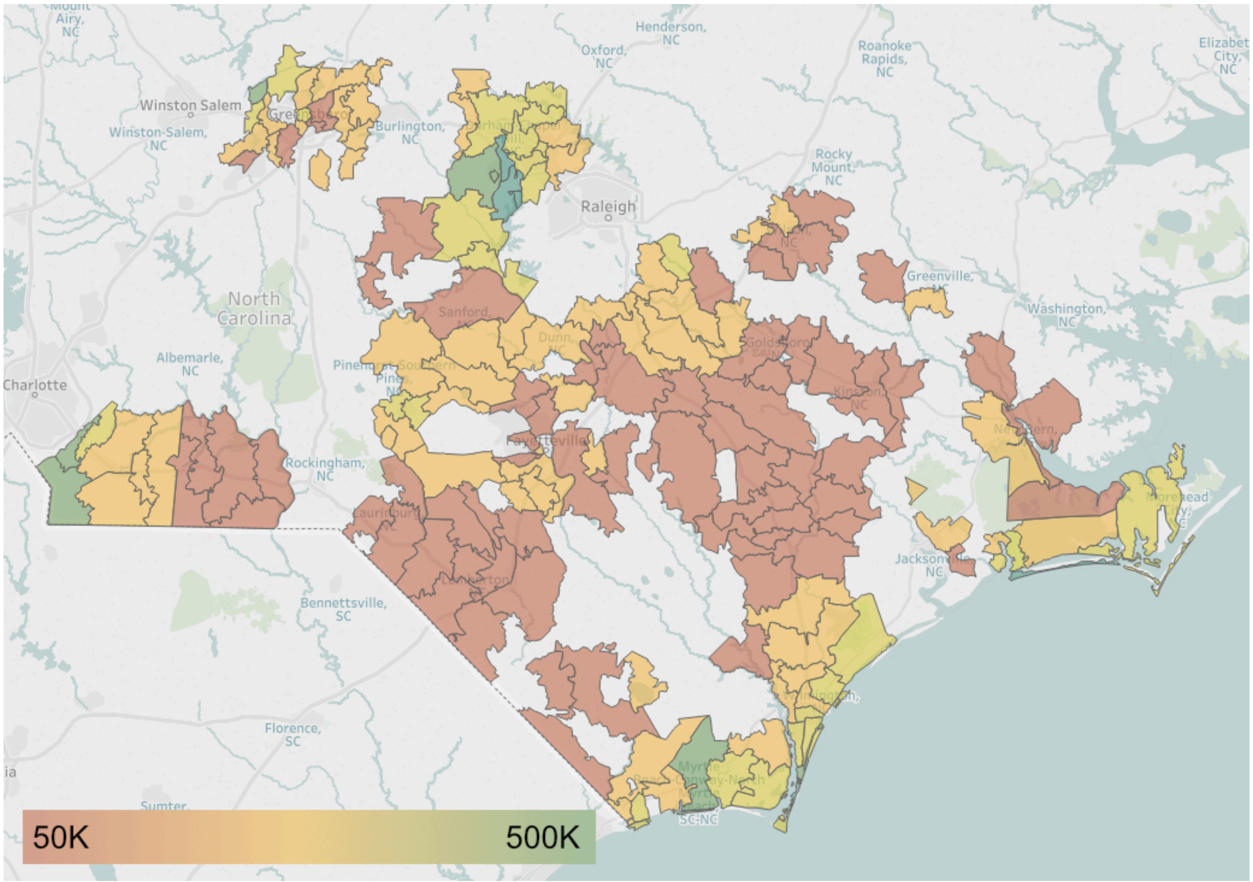


Figure 7. A choropleth map representing the distribution of forecasted median home prices

Absolute Median Home Value Change

While understanding the predicted median home value in each zip code is important, analyzing how the median home values are expected to change one year after the hurricane event provides additional insight. To complete this task, Pandas was utilized in Python to create an additional column of data. This column is calculated by subtracting the predicted value from the recorded median home value during the month of the hurricane event. Figure 8 shows the distribution of median home value changes.

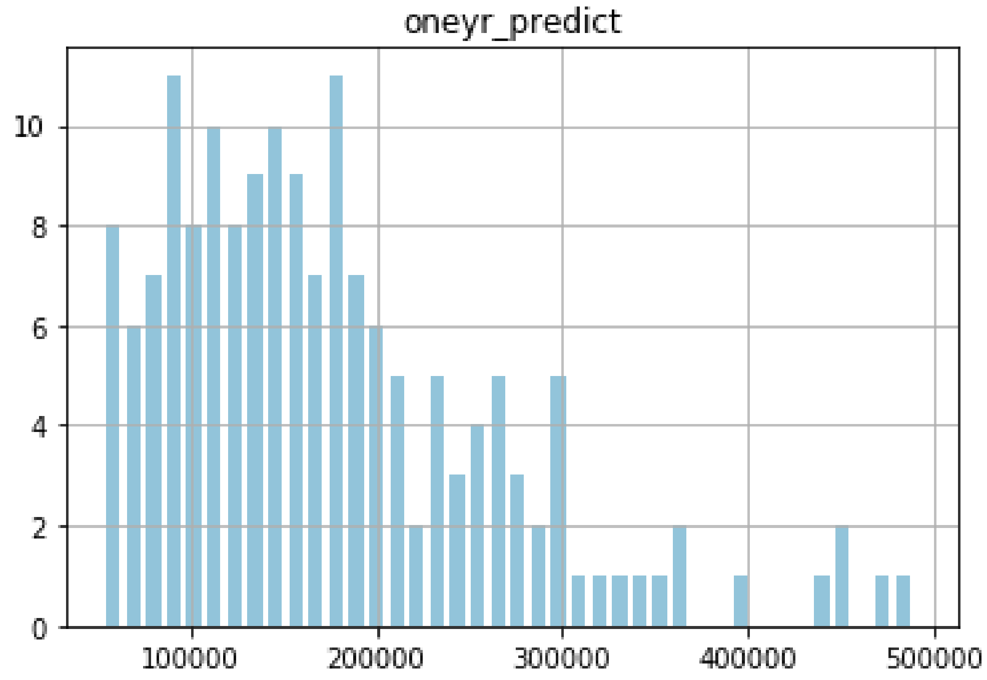


Figure 8. The distribution of median home value changes.

Compared to the distribution of median home prices, the distribution of median home value changes is less skewed. Still consisting of 164 records, the mean is $11,811.76 with a standard deviation of $8,014.53. Only one zip code in Wayne County, NC falls outside of three standard deviations from the mean at a predicted decline in median home value of -$21,992.40.

Figure 9 is a choropleth map that visually indicates the distribution of price changes in North Carolina. Areas shown in grey are near the mean change of $11,811.76 while areas in blue exhibit higher price increases. Only two zip codes are expecting large decreases in median home value and are shown in orange.

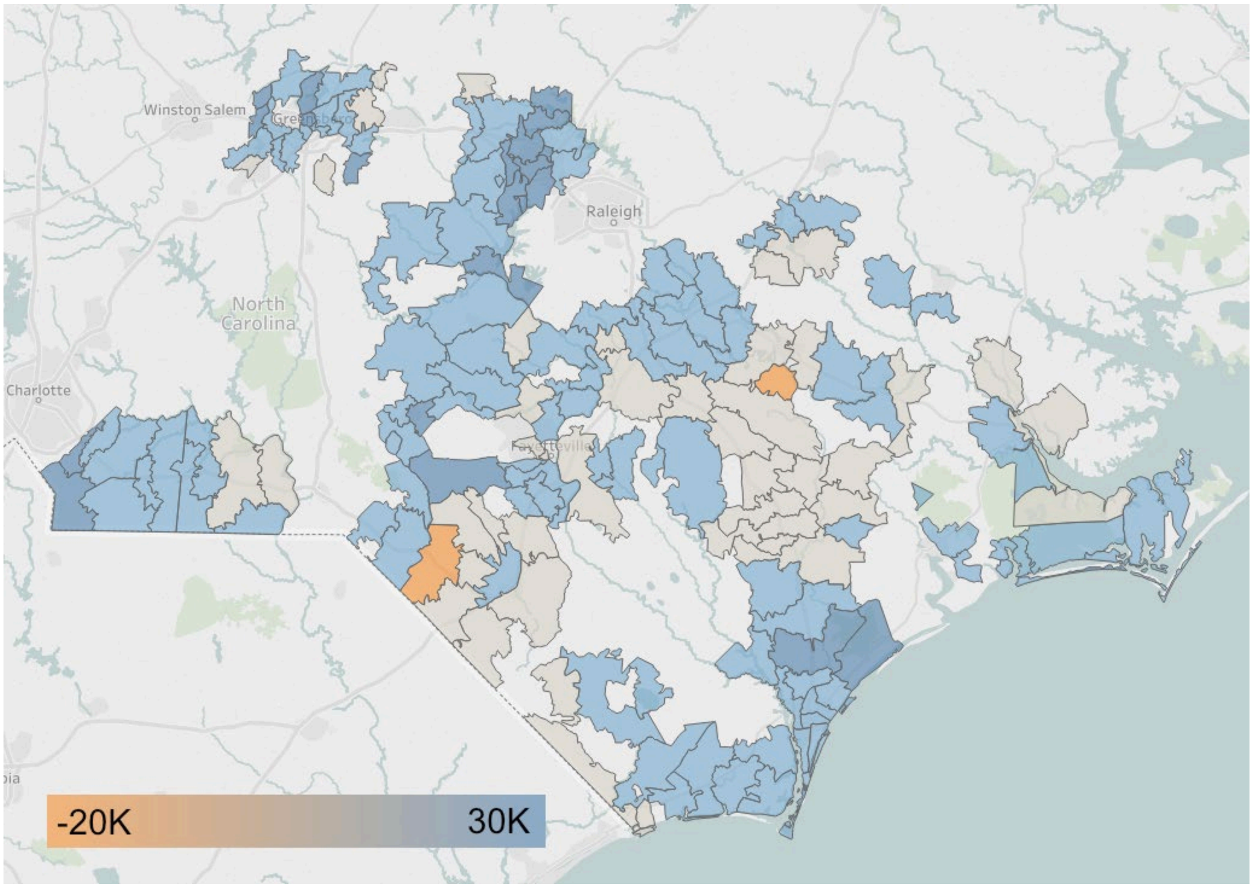


Figure 9. A choropleth map showing the distribution of price changes.

Percent Change of Median Home Values

The final area of focus deals with understanding the percent change of median home values from the month the hurricane occurred to the predicted values one year after the hurricane event. This is calculated by dividing the predicted value by the value during the month of the hurricane and subtracting one. The results are shown in the distribution present in Figure 10.

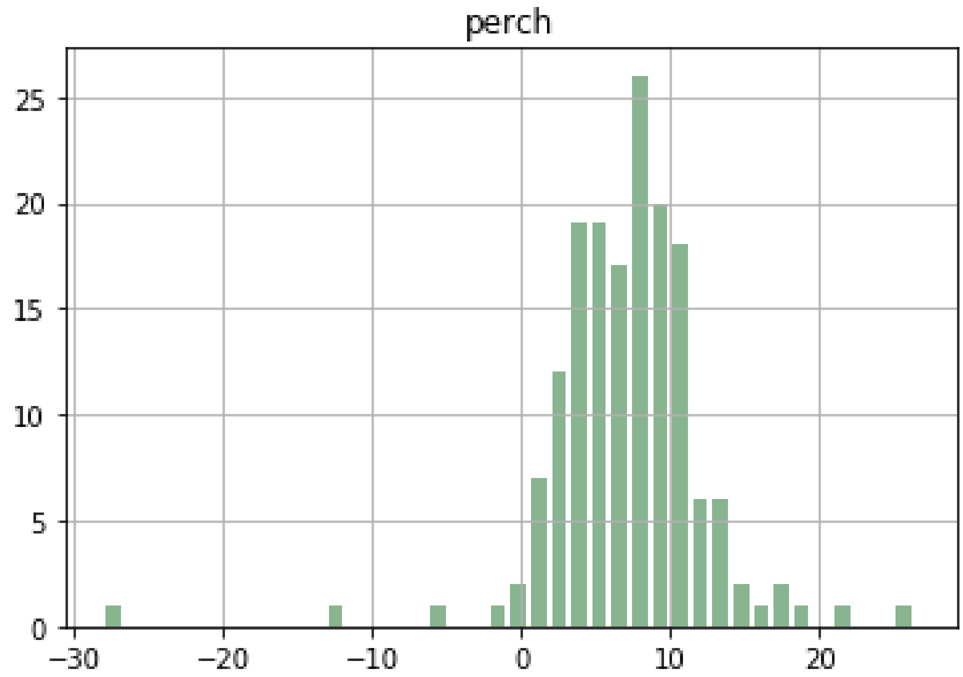


Figure 10. The distribution of percent change values in each zip code

Of the 164 records, this distribution has a mean of 7.28% and a standard deviation of 5.27%. With a minimum value of -27.87% and a maximum value of 26.48%, only one zip code falls outside of three standard deviations of the mean. This zip code has a value of -27.87% and resides in Wayne County, North Carolina. Once again, a choropleth map is shown in Figure 9 to visually represents the expected price change one year following the hurricane event.

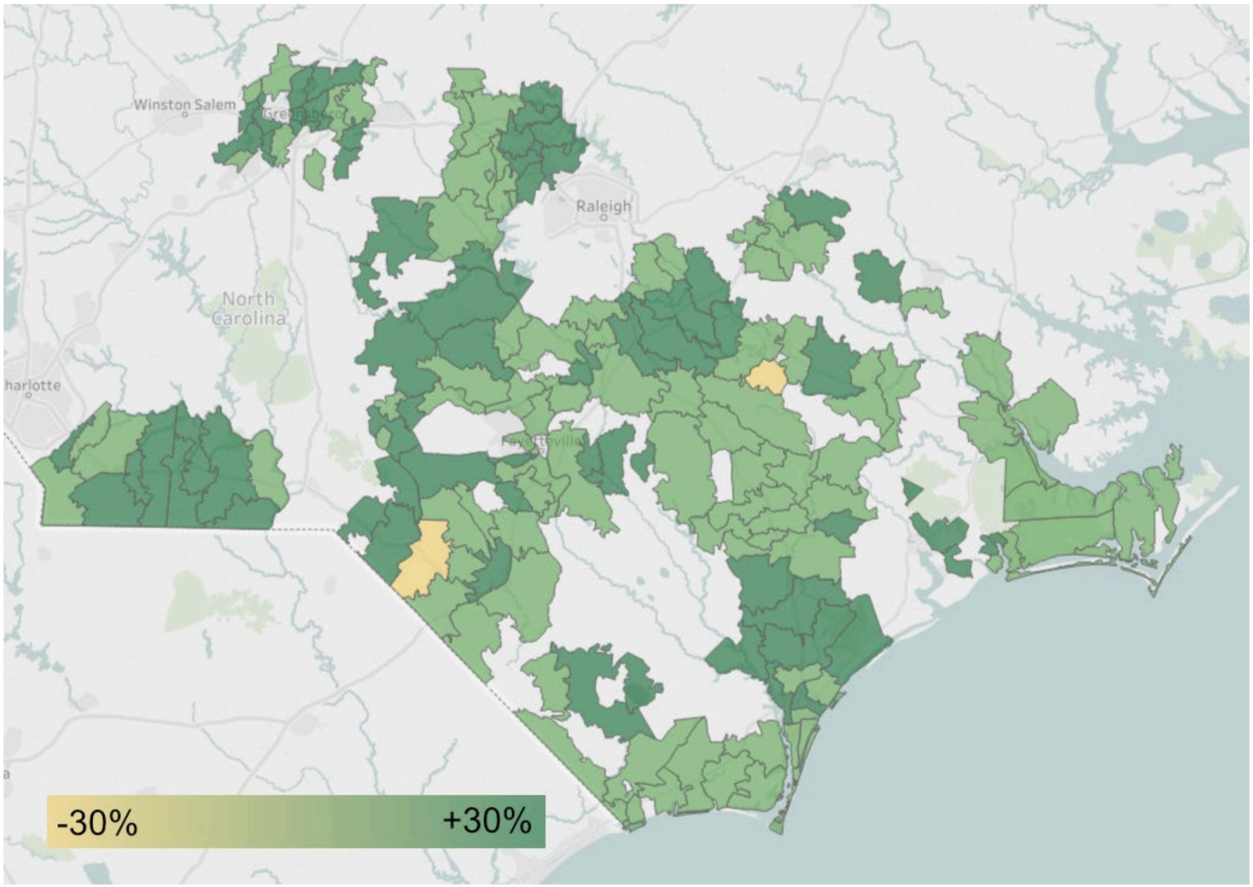


Figure 9. A choropleth map showing the distribution of predicted home values as a percentage

CHAPTER IV

CONCLUSION

Limitations

While the methods used in this research to generate median home value forecasts are reinforced by numerous trials from other researchers, all forecasting models must accept inherit limitations. Home valuation and the economy is an infinitely complicated system of decision makers in a network of buyers and sellers. Similarly, natural disasters present unique scenarios that even the most advanced weather prediction methods cannot forecast with complete certainty. Therefore, the complexity of the economy and weather systems make forecasting median home values with absolute certainty is impossible. Additionally, the forecasts presented in this research rely on second-hand data from other hurricanes. For this reason, all forecasts should be consumed with a critically and should be used to drive future natural disaster research at scale.

Finally, the data used to generate the models only utilizes seven fiscal input dimensions to generate one fiscal output dimension. This relatively small number of input dimensions limits the complexity of the forecasts. More advanced machine learning models can account for an increased number of input dimensions including wind speed, flood levels, flood plain boundaries, and local regulations impacting home construction quality.

Data Interpretation

As previously mentioned, the average percent change in median home value one year following hurricane Florence is 7.28% with a standard deviation of 5.27%. Figure 10 shows a stacked histogram which compares the average one-year median home value change since 1996 to the distribution of forecasted percent change values in each zip code one year after hurricane

Florence. The average percent change for the 164 impacted zip codes since 1996 is 5.64% with a standard deviation of 8.08%. This indicates that, according to the forecasting model, zip codes impacted by Hurricane Florence should expect slightly above average (1.64% higher) median home value growth in the 12 months following Hurricane Florence.

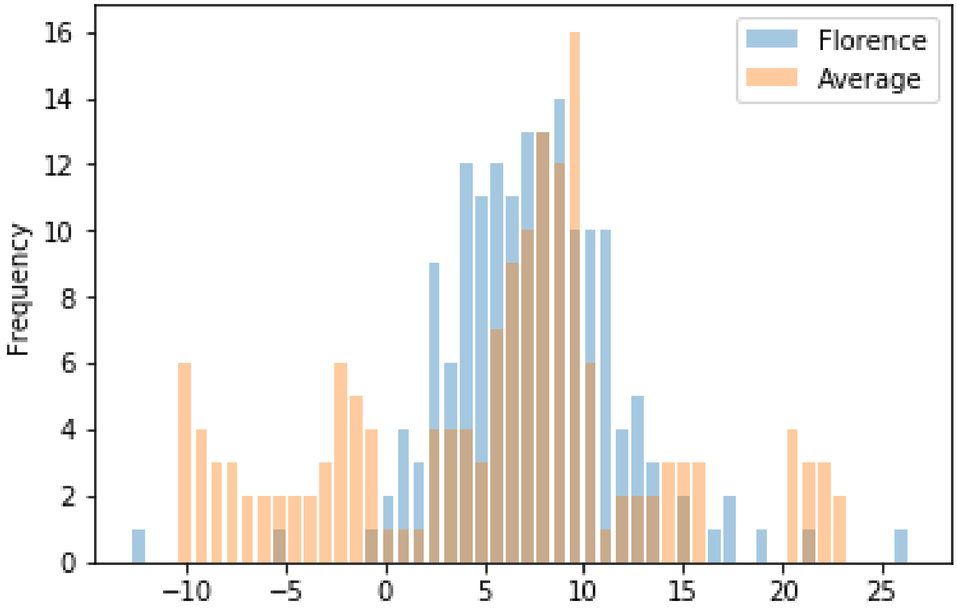


Figure 10. A comparison of the forecasted percent changes to the average percent changes

Future Research

This research can be expanded upon in several key ways. Firstly, including more input dimensions related to the hurricane event itself allows for the creation of more complex forecasts. More advanced models could provide greater insight to decision makers in the allocation of recovery funds, policy making, and more.

Additional considerations for future research include more advanced data visualization that provides greater insight about the hurricane event itself. More complex data visualizations could overlay wind, flood, and debris damage data to compare high disaster impact areas with home value predictions.

REFERENCES

Ahmed, Nesreen, et al. "An Empirical Comparison of Machine Learning Models for Time Series Forecasting." Econometric Reviews 29.5-6 (2010): 594-621.

Burrus Jr, Robert. T., et al. "Catastrophic Risk, Homeowner Response, and Wealth-Maximizing Wind Damage Mitigation." Financial Services Review 11.4 (2002): 327.

Cao, L. J., and Francis. E. H. Tay. "Support Vector Machine with Adaptive Parameters in Financial Time Series Forecasting." IEEE Transactions on Neural Networks 14.6 (2003): 1506-18.

Cavallo, Educardo A., et al."The Economics of Natural Disasters." IDB Working Paper Series 124 (2011): 1-50.

Crawford, Gordon, and Michael Fratantoni. "Assessing the Forecasting Performance of RegimeSwitching, Arima and Garch Models of House Prices." Real Estate Economics 31.2 (2003): 223-43.

Cutter, Susan L. "Social Vulnerability to Environmental Hazards." Social Science Quarterly 84.2 (2003): 242-61.

Dua, Pami, and David Smyth. "Forecasting Us Home Sales Using Bvar Models and Survey Data on Households' Buying Attitudes for Homes." Journal of Forecasting 14.3 (1995): 217-27.

Eves, Chris. "The Long-Term Impact of Flooding on Residential Property Values." Property Management (London) 20.4 (2002): 214-27.

Flanagan, Barry, et al. "A Social Vulnerability Index for Disaster Management."Journal of Homeland Security and Emergency Management" 8.1 (2011).

Gunes, A. Ertug., and Jacob P. Kovel. "Using GIS in Emergency Management Operations." Journal of Urban Planning and Development 126.3 (2000): 136-49.

Hallstrom, Daniel G., and V. Kerry Smith. "Market Responses to Hurricanes." Journal of Environmental Economics and Management 50.3 (2005): 541-61.

Highfield, Wesley, et al. "Mitigation Planning." Journal of Planning Education and Research 34.3 (2014): 287-300.

Kia, Masoud, et al. "An Artificial Neural Network Model for Flood Simulation Using Gis: Johor River Basin, Malaysia." Environmental Earth Sciences 67.1 (2012): 251-64. 9

Kunreuther, Howard. "Mitigating Disaster Losses through Insurance." Journal of Risk and Uncertainty 12.2-3 (1996): 171-87.

McKinney, Wes. "Data Structures for Statistical Computing in Python." Python in Science Conference 9 (2010): 51-57.

Murphy, Anthony. "The Impact of Hurricanes on Housing Prices: Evidence from US Coastal Cities." Federal Reserve Bank of Dallas (2010): 31.

Park, Byeonghwa, and Jae Bae. "Using Machine Learning Algorithms for Housing Price Prediction: The Case of Fairfax County, Virginia Housing Data." Expert Systems with Applications 42.6 (2015): 2928-34.

Peacock, Walter, et al. "Inequities in Long-Term Housing Recovery after Disasters." Journal of the American Planning Association 80.4 (2014): 356-71.

Pedregosa, F. "Scikit-Learn: Machine Learning in Python." The Journal of Machine Learning Research 12.Oct (2011): 2825.

Pryce, Gwilym, et al. "The Impact of Floods on House Prices: An Imperfect Information Approach with Myopia and Amnesia." Housing Studies 26.2 (2011): 259-79.

Rathfon, Dana, et al. "Quantitative Assessment of Post-Disaster Housing Recovery: A Case Study of Punta Gorda, Florida, after Hurricane Charley." Disasters 37.2 (2013): 333-55.

Spielman, Seth E. "Patterns and Causes of Uncertainty in the American Community Survey." Applied Geography 46 (2014): 147-57.

Van Zandt, Shannon, et al. "Mapping Social Vulnerability to Enhance Housing and Neighborhood Resilience." Housing Policy Debate 22.1 (2012): 29-55.

Vigdor, Jacob. "The Economic Aftermath of Hurricane Katrina." Journal of Economic Perspectives 22.4 (2008): 135-54.

Zhai, Alice, and Jonathan Jiang. "Dependence of Us Hurricane Economic Loss on Maximum Wind Speed and Storm Size." Environmental Research Letters 9.6 (2014): 064019.

Zhang, Peter. "Time Series Forecasting Using a Hybrid Arima and Neural Network Model." Neurocomputing 50 (2001): 159-75.

Zhang, Yang, and Walter Peacock. "Planning for Housing Recovery? Lessons Learned from Hurricane Andrew." Journal of the American Planning Association 76.1 (2009): 5-24.