Using Geographic Interdependencies to Predict Damage and Recovery Time for Parcels in Galveston after Hurricane Ike

Christian Gullberg*, Paul McElroy†, Hari Ramanan‡, Brian Vancil§, Lei Wang¶
Department of Electric Engineering and Computer Science
School of Engineering
University of Kansas
Lawrence, KS 66046
*Email: christian.gullberg@ku.edu
†Email: pcm@ku.edu
‡Email: hramanan@ku.edu
§Email: brian.vancil@ku.edu
¶Email:l290w868@ku.edu

***Abstract*—Studying the impact that a natural disaster has on a community is vital to developing measures that can mitigate the damage caused by future storms. In this project, we analyze economic damage assessments and recovery times for residential structures in Galveston after Hurricane Ike in 2008 using machine learning in order to determine if these outcomes may be better predicted and their contributing factors better inferred from socioeconomic and geographic features. After creating new features in the dataset based on geographic interdependencies, we trained a number of machine learning models on the data for prediction and inference purposes. Of the models chosen for this experiment, random forest was the most accurate in predicting the amount of damage accrued by a given damaged property. Similarly, for predicting the recovery time of damaged property, we found a random forest to be the most accurate model. We found that the variable that most contributed to a property's relative damage, and thus its recovery time, was its location.**

I. Introduction

This research project aims to answer the following question: How can geographic interdependencies be measured and used to explain damage caused by, and progress in recovery following Hurricane Ike in Galveston? This project extends on previous research studies done in the wake of natural disasters that also aimed to analyze and predict property damage in light of hurricanes. However, this project is decidedly less vested in the impact of socioeconomic factors on housing recovery. Rather, it deals with the proximity of useful assets in the Galveston region to the parcels that were devastated by the Hurricane. A geographic interdependency can be described as a relationship between two objects in close spatial proximity where, for example, physical damage to one object could cause correlated damage to the other [1]. Thus, this work analyzes the locations of public utilities and recovery resources in the Galveston area and searches for interdependencies that may exist with Galveston's residential parcels.

There have been many studies conducted that perform analysis on the state of a real estate market in the wake a catastrophic natural disaster [2]. Often, these studies aim to find a correlation between the economic damage to a specific parcel of land, or group of parcels to any number of variables. One such study is that of Sara Hamideh who, in 2016, determined that there was a relationship between a property's high levels of damage and its proportion of minority or low-income residents. In attempting to model and predict property recovery time in the wake of Hurricane Ike, found that race and income both had "significant implications" for access to critical resources in the aftermath of the storm [3]. A similar study to Hamideh's is Wesley Highfield's 2014 mitigation planning analysis, which aimed to assess the contributions of social vulnerability as well as structural vulnerability in causing structural damage. Highfield also found that properties in areas with high concentrations of Hispanics and African Americans, defined as socially vulnerable groups, experienced more damage than properties in other areas [4]. Our study uses a similar definition of damage in economic terms as the relative change in assessed structure value from pre-disaster to post-disaster. We also aim to predict recovery time, defined as the number of years before a structure regains its pre-disaster value.

II. Background

A. 1900 Galveston Hurricane

In September 1900, Galveston, Texas, was hit by the deadliest hurricane to ever make landfall in the U.S [5]. The 1900 Galveston Hurricane took thousands of lives in a then-booming town and also remains one of the costliest natural disasters in American history. The silver lining to this storm was the town's ability to learn from the

storm and rebuild the city in a way that would protect it from future such catastrophes. A seawall was built on the coastline, sand was pumped under buildings to elevate them above sea level, and when the next major storm hit 15 years later, the death toll and financial damage to the town were comparably much smaller [5].

A major motivation for undertaking this project is to mimic the learning efforts shown by survivors of the 1900 Hurricane. Modern technology has allowed the Galveston community to compile gigabytes of data on the damage caused to properties in the area, and the researchers will aim to find patterns of damage based on the locations of properties, and their proximity to landmarks and utilities in the Galveston area. It is the hope of those conducting this project that we can learn from the effects Hurricane Ike and properly prepare and defend Galveston and its people the next time such a storm hits the area.

B. Effects of Hurricane Ike in Southeastern Texas

Early in the morning of September 13, 2008, Hurricane Ike made landfall on the Gulf Coast of Texas, which includes the city of Galveston and its surrounding counties. Despite then-Governor Rick Perry announcing a mandatory evacuation for the region some days before, 84 people were killed by the storm and damage to the region was estimated at nearly 38 billion in 2008 dollars [2].

In the years since Hurricane Ike, gigabytes of data has been collected about the state of Galveston and its surrounding counties to determine if anything can be changed before the next major storm. This data includes damage inflicted to most properties in the region, as well as their restoration times. This goal of this project is to find geographic predictors that indicate damage and recovery time after a natural disaster.

III. Methods

A. Provenance of Data

Data for this project was assembled by Elaina J. Sutley from an earlier project from at least two sources. The first source is land parcels tracked by the Galveston Central Appraisal District, which are available online as ESRI shape files, a geospatial file format used natively by ArcGIS [6]. Besides encoding the detailed shapes and locations of each parcel, the shape files provide the value of any improvements on the property (such as buildings) and the area in acres as well as a number of features we did not use in modeling, like value of the land, total value of the parcel, owner, legal address (situs), site names, assessed entities, area in acres, land use codes, page, neighborhood, exemptions, and flags. Some of these unused features may be interesting to consider, but we either did not have access to the data for each year (in the case of the land and total values) or did not have a data dictionary for the code meanings. The second source is tract-level census data from the year 2000 U.S. Census [7], which provides tract code, block group number, block group identifier, median household income, racial/ethnic composition, multifamily/renter-occupied/occupied property rates, and percent of the local population in poverty, with no telephone, with no high school diploma, unemployed, or that speak English not well.

Although the Galveston Central Appraisal District provides shape files online, only the current version of the current year's data is readily available. We had only 2008 and 2017 shape files, but to investigate damage and recovery following Hurricane Ike necessitates access to data on the assessed value of parcel improvements from each year from 2008 until the present. We did have access to summaries for the intervening years. In all, we had access to the following:

1) A spreadsheet summary of parcel records for 2008 to 2015, including situs (legal address), 2008 owner information, and improved value assessments [6]. Parcels were coded by location (Galveston v. Bolivar), occupation type (renter v. owner), and housing type (multifamily, duplex, or single family). Census tract information about racial and ethnic composition and median household income were joined into the spreadsheet [7]. We retained most of these predictors.
2) A different spreadsheet summary of parcel records for 2016 containing information about nearest electrical substations. We used this only to obtain the assessed value of parcel improvements in 2016.
3) Shape files for parcels in 2017 with assessed values (land, improvements, and total), situs, owner, etc. We used this dataset for the geospatial shapes of the parcels and the assessed value of parcel improvements in 2017.

B. Processing of Data

The data was joined and processed using the Tidyverse ecosystem of packages for the R programming language [8]–[12]. Because the shape files were unavailable for all years, we could not use a spatial join to assemble an improved parcel value for each year for a given 2008 parcel. Instead, we started with the spreadsheet summary for 2008 parcels with 2008-2015 improved values and joined in the 2016 improved values by matching the address (situs) across datasets. We retained all 2008 parcels and included all 2016 data with a matching situs (legal address) and XREF (an identifier stored in the original shape files). Because of address changes in our second dataset, approximately 20% of improvement values are missing for 2016. We then joined in the 2017 parcel data, also by situs and XREF, to obtain a record of structure values (via the assessed improvement values) from 2008 to 2017.

Key to the research questions were the recalculation of damage and recovery time, which were seen to be faulty in the original data source. *Damage D* is operationalized

as the relative loss in assessed structure value from 2008 to 2009:

$$D = \begin{cases} 1 - \frac{\text{value}_{2009}}{\text{value}_{2008}}, & \text{if value}_{2009} \leq \text{value}_{2008} \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

Recovery time T_{recovery} is operationalized as the minimum number of years after 2009 at which point the assessed structure value regains or exceeds its value in 2008:

$$T_{\text{recovery}} = \begin{cases} \texttt{NA}, & \text{if value}_{2008} \leq \text{value}_{2009} \\ 1, & \text{if value}_{2008} \leq \text{value}_{2010} \\ \vdots & \\ 8, & \text{if value}_{2008} \leq \text{value}_{2017} \\ \texttt{NA}, & \text{otherwise.} \end{cases} \quad (2)$$

The final analytic dataset, which includes geographic interdependencies described below, also afforded the creation of a small number of transformations of existing features, such as the ratio of 2008 value to area of the parcel and common logarithms of the approximately log-normal assessed values and parcel areas.

The calculation of damage and recovery time allowed the removal of assessed structure values for the years 2009 to 2017 from the analytic dataset.

C. Feature Creation and Proximity Analysis

One major task of this research project was the creation of features that were not available in the original data set. From the data described above, the group calculated distances from land parcels to various resources that were used to measure and test the effects of geographic interdependencies on these land parcels. Features created for this project are described below.

- Drainage district facilities: Drainage districts in Galveston were established in the early 20th century for the purpose of draining overflowed lands in the region. Their activities include the construction, maintenance, and operation of drainage facilities in the area. We calculated the distance of each parcel to the nearest of three drainage district control centers to find a correlation with the damage accrued by these parcels. We thought the drainage districts had strong correlations with the damage and recovery time due to they could limit the flooding time [13].
- Emergency service district centers: There are two Emergency Service Districts (ESD) in Galveston county. These districts fund the Medical Emergency Serivces as well as volunteer fire departments in the region. We identified the main office for each district and calculated the distance to the nearest district office from each parcel. We thought this could be a good feature because the closer the parcel was to the ESD, the quicker it would receive the assistance after the disaster
- Coastline: We used the flood shapefile and it gave us the estimation of the distribution of the water and land for the city of Galveston and Bolivar Island, we were able to create an outline for the border between land and sea in the region. After calculating the distance from each property to the nearest point on the coastline, we sought to find a correlation with this distance and the damage accrued to each property. Intuitively, we thought the closer a parcel was to the coast the more damage it would suffer.
- Distance to commercial features: After compiling a list of addresses for gas stations, churches, and food stores (grocery stores, convienience stores, big box stores) as well as the singular Galveston Walmart, these addresses were geocoded to their respective latitudes and longitudes. We then calculated the distance of each parcel to the nearest of each type of commercial feature.
  The commercial features were chosen based on the ease with which addresses could be compiled from either Google Maps or from the Yellow Pages and on a guess that they may provide some relief or benefit to those residents closest to them. Churches were chosen because they may provide community benefits and support to families devastated by the hurricane. Gas stations were chosen because easy access to fuel and convienience shops close to homes was hypothesized to be beneficial to nearby neighborhoods. Stores that sell food were selected for similar reasons.

D. Model Creation

For the purposes of predicting how badly damaged property was affected by Hurricane Ike the two primary outcomes analysed have been how damaged the property was in percentage of total value of the property and how many years after being damaged does the property take to recover to its pre-Ike value. To make predictions of the two outcomes requires regression model selection and feature selection.

We looked at several types of models including Gradient Boosting Trees, Random Forests and Linear Regression. Linear Regression was our starting point because of the ease with which it could be set up and trained. The tree based models were used because it was expected they would provide significant accuracy increases over the Linear Regression model.

Gradient Boosting Trees were found to be more accurate over Linear Regression. In addition the ability to analyze the relative influence of predictors on the outcomes allowed us to produce some inferential results. This was used for feature selection for the models predicting time to recovery. For example, whether the property was on Galveston or Bolivar, or was a duplex had a relative influence of zero in any gradient boosting tree model except when they were used as lone predictors.

Random forests were found to have more accurate results than the gradient boosting tree model, however without the ability to view the relative influence of predictors.

Almost all available features were used as predictors except when those features provided data from after the Hurricane. The lone exception was that the total damage percent value was used as a predictor for the time to recovery models because recovery is a years long process.

We excluded data from the Census data which only provided Census-Centric identification information such as tract code, block group number and block group identifier.

The features we included in order of relative influence derived from analysis of the gradient boosting model used for recovery time were: damage, 2008 population density, distance to Walmart, distance to the nearest gas station, distance to the coast, building age in 2008, latitude, distance to nearest drainage district, longitude, distance to the nearest church and food store, the area of the property in square feet, the value of the property in 2008, whether the property was multi-family or single-family, percentage SV, percentage non-hispanic black, percentage household income, percentage non-hispanic white, percentage hispanic, percentage unemployed, percentage occupied, percentage multifamily, percentage non-white, percentage with no telephone, percentage with no high school diploma, percentage renter occupied, percentage in poverty, percentage that speak English not well, and whether the property was renter occupied in 2008. The Census data as well as most of the data is anonymized based on parcel and so individual information for single-family dwellings is not available. For example, whether the property was single family or multi-family was a boolean value for the entire parcel, whereas the Census multi-family percentage was based on the entire neighborhood which included the parcel.

Feature selection for the damage predicting models contained all of the same census features as the models for recovery time, the value in 2008, the area in square feet, the age of the building in 2008 and the latitude and longitude.

IV. Results

A. Exploratory Data Analysis

Figure 1 displays the location of each parcel in the dataset.

Fig. 1. Spatial extent of the parcels along Galveston Island (left) and Bolivar Peninsula (right). Each dot represents a parcel in the dataset. Produced using `ggmap` and GGPlot2 [12], [14].

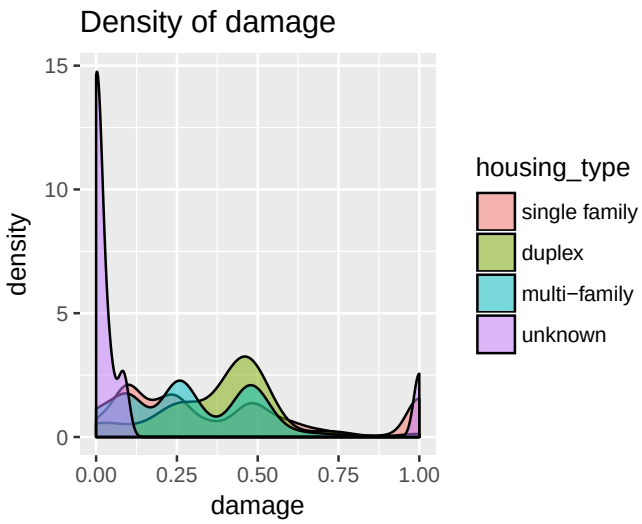

Fig. 2. A smoothed density of damage by housing type shows some differences in damage by housing type. Unfortunately, most of the interesting differences are with the unknown type. Produced using `ggmap` and GGPlot2 [12], [14].

B. Model Results

For the time to recovery the damage predictor was found to be by far the most important. This is based on the relative influence derived from the gradient boosting models.

Not surprisingly, many features that contained spatial information such as latitude and longitude or distance to Walmart were largely interchangeable without largely affecting the mean absolute error. We feared that longitude and latitude would provide too specific of influence over the results so that the models generalizations to other hurricane events would be affected. However, because spatial information was so interchangeable it seems that all spatial information, or rather, information that could be used to derive the layout of Galveston, may be a poor predictor if generalization for other hurricane events is desired.

The largest contributor to total damage was whether the property was on Galveston or on Bolivar. The vast majority of properties on Galveston did not sustain the total damage that occured on Bolivar.

Table I and Table II contain the errors of the trained models over Damage and Recovery Time, respectively. For both cases the random forest regression model performed the best with 1000 trees. In general the gradient boosting

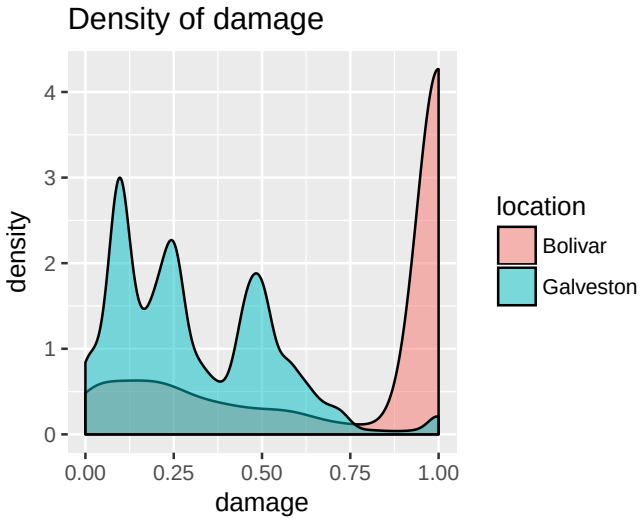

Fig. 3. The smoothed distribution of damage is very different for Galveston Island versus Bolivar Peninsula, which was more heavily damaged. Figure 5 shows the corresponding location-differentiated plot for recovery time. Produced using `ggmap` and GGPlot2 [12], [14].

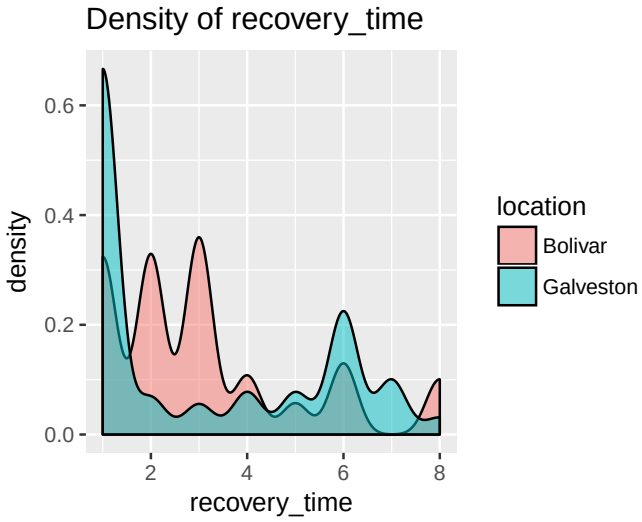

Fig. 5. The smoothed density of recovery time by location shows that recovery time for Galveston Island was distinctly bimodal versus Bolivar Peninsula. Produced using `ggmap` and GGPlot2 [12], [14].

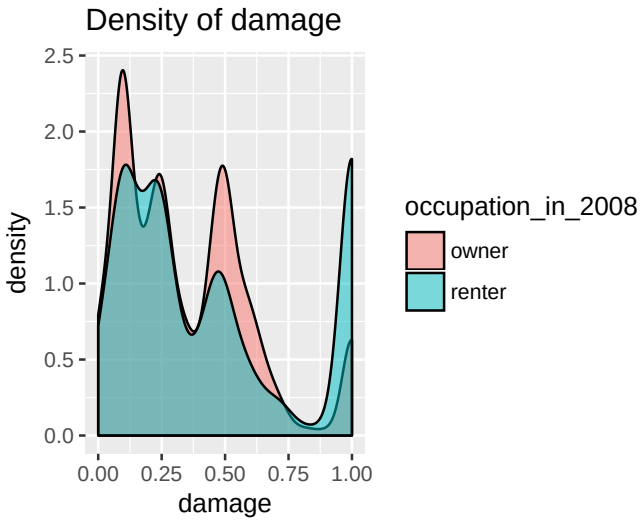

Fig. 4. The smoothed distribution of damage is not dissimilar for owner-occupied and renter-occupied properties, with renter-occupied properties having slightly more heavily damaged parcels. Produced using `ggmap` and GGPlot2 [12], [14].

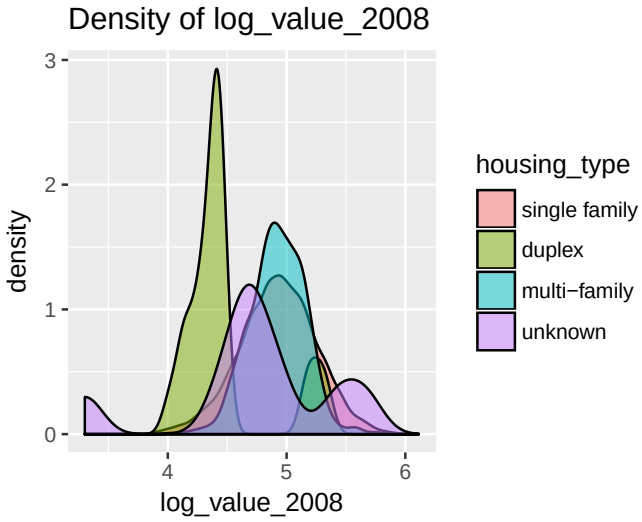

Fig. 6. The assessed improved values in 2008 show differences by housing type, with duplexes having much lower values. Produced using `ggmap` and GGPlot2 [12], [14].

descent tree models performed second best for both. Linear Regression contained a higher amount of error than the other models suggesting that the data is probably not linearly separable. The K-Nearest Neighbor results are special as they were only trained on the longitude and latitude, with the addition of damage for the recovery version. The KNN models were simplified to show how moderately low error results could be obtained using a small number of predictors. This perhaps indicates that data is highly correlated to location with many damaged neighborhoods being tightly clustered together.

We have shown that good predictive results can be obtained using random forest models and gradient boosting trees for both damage and recovery time. These tree models had nearly comparable mean squared error and absolute errors. For KNN a simplified set of predictors easily obtained excellent prediction, exhibiting the second lowest Mean Absolute Value and Median Absolute Error. For all cases the lowest standard deviation in the simple errors obtained were obtained using random forests.

V. Conclusion

The Seawall on Galveston is the single most striking difference between Galveston and Bolivar. Bolivar had

TABLE I
Damage Errors

	Damage			
	Random Forest	Gradient Boost Descent Trees	Linear Regression	KNN
Mean Absolute Value	6.943801	8.127972	14.58948	7.916559
Median Absolute Error	3.052125	4.680142	11.12408	0.04587777
Mean Squared Error	148.9172	164.7472	361.2068	310.4477
Error Standard Deviation	12.2038	12.833	19.00651	17.61458

TABLE II
Recovery Errors

	Recovery Time			
	Random Forest	Gradient Boost Descent Trees	Linear Regression	KNN
Mean Absolute Value	1.588804	1.741542	1.950626	1.694793
Median Absolute Error	1.328132	1.558593	1.842909	1.44
Mean Squared Error	4.067451	4.314288	5.011347	4.493457
Error Standard Deviation	2.016348	2.077299	2.238824	2.119886

near total damage done to each building and we hypothesize that the Seawall as the most unique differentiating feature is responsible for saving Galveston from a higher percentage of houses totally destroyed.

The lack of more specific building and personal information as well as elevation information provided a difficult challenge in finding adequate answers to why certain properties are damaged more than others. However, the time to recovery for a damaged property is clearly going to reflect how damaged that property is. It is because of this that we must question the values we have for total damage done to a property.

We showed that creative use of commercial features such as the distance to gas stations or to churches allow the easy creation of new and valuable features. This was only possible because of the spatial nature of the Galveston Hurricane data. Similar distance features can be created simply from a list of addresses or gps coordinates.

Using ensemble tree methods was shown to be a good approach to predictive modelling for the damaged properties in Galveston and Bolivar. Furthermore, we showed that little information was needed to produce low error using a K-Nearest Neighbor approach, most likely because of the highly clustered nature of the damage.

A. Discussion

More spatial information should be created based off of commercial and social features building off of the relative success of the distance features created for this project. We propose for these newly created features to be selected by generalizability across hurricane events and so that they do not allow the recreation of a layout of the specific event in question.

Some features should be rethought such as the total damage. For example, any damage above zero percent indicated that there must be a time to recovery. This however does not take into account several factors: does the owner intend to recover a small amount of damage? If the damage is total, does this indicate that the property was completely destroyed or that the property's damage, as judged by a property tax assessment, is the total value of the property? In addition, land and property values are not adjusted to account for the real estate collapse and recession in 2008 and is not adjusted for normal inflation.

Acknowledgment

The authors would like to thank Prof. Nicole Beckage and Prof. Elaina J. Sutley, both of KU, for access to the data and assistance with the project.

References

[1] S. M. Rinaldi, J. P. Perenboom, and T. K. Kelly, "Identifying, understanding, and analyzing critical infrastructure interdependencies," *IEEE Control Systems Magazine*, p. 11–25, Dec 2001.

[2] W. contributors, "Effects of hurricane ike in texas — wikipedia, the free encyclopedia," 2017, [Online; accessed 15-December-2017]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Effects_of_Hurricane_Ike_in_Texas&oldid=813534954

[3] S. Hamideh, W. G. Peacock, and S. Van Zandt, *Association of Collegiate Schools of Planning*, 2016.

[4] W. E. Highfield, "Mitigation planning: Why hazard exposure, structural vulnerability, and social vulnerability matter," *Journal of Planning Education and Research*, vol. 34, no. 3, p. 287–300, 2014.

[5] Wikipedia, "1900 galveston hurricane — wikipedia, the free encyclopedia," 2017, [Online; accessed 5-December-2017]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=1900_Galveston_hurricane&oldid=812380345

[6] "Parcels with data." [Online]. Available: http://www.galvestoncad.org/index.php/Shape_Files

[7] U. S. C. Bureau, "2000 census." [Online]. Available: http://factfinder.census.gov

[8] H. Wickham, *tidyverse: Easily Install and Load 'Tidyverse' Packages*, 2017, r package version 1.1.1. [Online]. Available: https://CRAN.R-project.org/package=tidyverse

[9] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2017. [Online]. Available: https://www.R-project.org/

[10] H. Wickham, R. Francois, L. Henry, and K. Müller, *dplyr: A Grammar of Data Manipulation*, 2017, r package version 0.7.2. [Online]. Available: https://CRAN.R-project.org/package=dplyr

[11] H. Wickham and L. Henry, *tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions*, 2017, r package version 0.7.0. [Online]. Available: https://CRAN.R-project.org/package=tidyr

[12] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. [Online]. Available: http://ggplot2.org

[13] "Home - galveston county texas drainage district 1," 2011. [Online]. Available: http://galvestoncountydrainagedistrict1.us/

[14] D. Kahle and H. Wickham, "ggmap: Spatial visualization with ggplot2," *The R Journal*, vol. 5, no. 1, pp. 144–161, 2013. [Online]. Available: http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf