Predicting Critical Infrastructure Vulnerabilities Using Machine Learning Classifiers

Baiherula Abula, David Cross, Aaron Kingery

I. INTRODUCTION

Critical infrastructure resilience against various disruptions is an important, if lesser known, part of city development and maintenance. Recently, the topic of resilience has gained increased attention in many fields due to growing sense of vulnerability. By using models and theories to determine what aspects (or sections) of infrastructure systems are vulnerable during crisis events like hurricanes and terrorist attacks, we can better create a first-response network that will be prepared for when the city faces a real crisis event. Our study aims to determine vulnerable sections in urban critical infrastructure systems, more specifically road networks, and distinguish them from those that are able to withstand crisis events using the power of machine learning classifiers applied to real-world data. This is important for three major reasons. First, our modern critical infrastructure systems are becoming increasingly interdependent due to the rise of digital technology, so they should no longer be treated as isolated objects when designing and modelling. Second, as urban areas around the globe continue to grow, the influx of population poses stress to the infrastructure of the city, which only makes disruptions or breakdown even more unaffordable. Finally, climate change, terrorists attacks, and other crisis events have not only increased in frequency, but also in unpredictability. This makes taking pre-cautionary measures, like vulnerability assessment, a necessary strategy.

As could be seen from below figure, even though streets/roads in certain neighborhood share certain characteristics (like elevation, topography), the spatial distribution of roads and streets don't seem to exhibit any type of pattern or clustering. Therefore it worth researching about the features of roads and streets in order to be able to predict what features tent to lead to flooding, which is a context in which application of common classification algorithms in machine learning couldn't be more suitable.

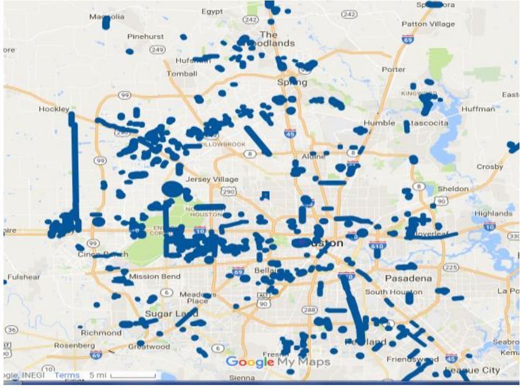

Figure 1: Flooded streets during Hurricane Harvey (Source: Google maps)

II. LITERATURE REVIEW

Modern analyses of flood risk rely on both quantitative and qualitative analyses. Qualitative analyses involve the discovery of features that are highly correlated with flood risk. These features are often obviously correlated with flooding such as the lack of drainage systems or the presence of nearby dams and spillways. A quantitative analysis instead relies upon mathematical models to predict flooding. As this paper proposes a quantitative method for vulnerability assessment of roads, methods with quantitative characteristics have been reviewed in more detail. It came to authors' attention that there are four major category of quantitative methods: (1) Probabilities methods[1]; (2) Analytical methods [2]; (3) Fuzzy inference systems based methods[3]–[5]; (4) Graph theory based methods [6]–[8]. While each of these methods has its own merits and could be suitable under the right context, there exists a significant gap between main-stream civil engineering based vulnerability assessment methods and increasingly powerful machine learning methods, which not only could address the complexity of the problems with large amount of data but also could result in predictions with much higher prediction accuracy. This is the main motivation of this paper. In the context of flood prediction, below are some examples of literature.

The current state of the art analyses primarily utilize graphical information system (GIS) data in order to identify regional flooding[9]–[11]. While this GIS data has been shown effective at prediction of regional flooding, it has not been shown that analyses of this data will yield sufficient results on the scale of individual roadways[12], [13]. Instead, we propose the use of a set of more simple binary qualitative variables that describe the local environment of the roadway such as whether or not the roadway is located in a floodplain or if the roadway is between open fields. This approach is more similar to more classical qualitative approaches, however, our approach is inherently able to deal with huge amount of data set and possibly filter out noise in the data..

While some machine learning methods have been applied to critical infrastructure risk analysis, particularly with regards to telecommunications infrastructure, the field as a whole has not yet explored the breadth of machine learning tools in analysis of risk to critical infrastructure. Current methods focus on regression analysis in order to predict flooding volume at a regional scale. There has been some limited use of classification algorithms in order to analyse critical infrastructure, however, it has been limited in both breadth and depth[10]. Most of these analyses focus on a single classification algorithm and make no mention of any sort of hyper-parameter tuning[13]. Due to this, we use a variety of machine learning classification tools in order to predict roadway flooding, as well as experimenting with different hyper-parameter tuning.

III. PROBLEM FORMULATION

Critical infrastructure (CI) resilience is an often-overlooked aspect of building a city that is nonetheless one of the most important aspects. CI resilience essentially informs city planners how vulnerable an area is to major disaster events, such as flooding, fires, and hurricanes. Areas with low CI resilience tend to not only be affected the most by these disaster events, but also tend to have the most issues when it comes to helping those in need by providing services like police, firefighters, and essential medical care. Therefore, this project presents a preliminary study on how to measure the CI resilience of an area using machine learning methods. By learning which areas of a city are vulnerable to disaster events, we can better shore up defenses and services to make sure that there is minimal destruction in these areas.

IV. PROPOSED SOLUTION

In order to accomplish our goal of determining the vulnerabilities in the critical infrastructure of urban areas, we applied several classifiers commonly found in the realm of machine learning to data collected from Houston Transtar. First, we collected data from Houston Transtar that represents the features of the various urban areas in Houston immediately after the flooding and devastation caused by Hurricane Harvey. These features are numerous and can be broken up into three major types: those that deal with closures due to construction, those that deal with closures due to other reasons such as flooding, and those that deal with traffic accidents. These three groups of data were collected by Houston Transtar between August 24th, 2017 and September 4th 2017, which encapsulates the dates at which the floodwaters from Hurricane Harvey in Houston were the highest. Within each group, there are various features such as the name of the roadway affected, the direction of travel the roadway was going, a brief description of the incident, and other features. In order to use classifiers on this data, we needed to clean and classify the data. To do this, we simply gathered data from Houston Transtar, picked the features that we wanted to use, and made sure to get rid of any null entries and fill in missing entries where we could so that our classifiers did not encounter significant issues.

After cleaning and organizing the data, we used the University of Waikato's "Weka" tool to run several classifiers on the data. This tool is open source under the GNU General Public License. We decided to use this tool because it provides easy access to pre-implemented classification algorithms, meaning that we could spend more time working on the experimentation in the project than working on developing the algorithms themselves, which are already familiar to the machine learning community. The classifiers we chose to use in Weka are k Nearest Neighbor, Multilayer Perceptron (a simple neural network), Random Forest, Logistic Regression, and Naive Bayes. We chose this algorithms after testing many other classifiers to see which has the highest accuracy in predicting critical infrastructure vulnerability. Because the purpose of this project is to provide a way to predict whether an area of a city is vulnerable to floodwaters created by hurricanes, we wanted to find the best classifiers for determining the safety of the roads in question. In this case, we used a basic accuracy measure as a distinction between the classifiers, but we also considered using f-measure to incorporate the possibility of false positives and false negatives. As previously mentioned, we tested many of the classifiers available in Weka before deciding to use these five classifiers.

Finally, after using these classifiers on the data, we determined which classifiers and hyper-parameters are best to be used to predict the vulnerability of the critical infrastructure of an area. The classifier/hyper-parameter combination we chose and the reasoning behind it will be discussed more in the next section.

V. DATA DESCRIPTION

A. Variable Introduction

In order to identify features of streets which has been flooded, we have consulted practitioners, reviewed literature and analyzed the historical news reports. The output variable *y* is Road status, which takes the value of "Closed " due to flood or "Open". The unit of analysis is one section of a road or street. If it is open, the value will take the value of "0", if the road section is closed due to flooding, the value will take ''1". The output variables are as follows:

- X1: flood plain variable - If the road is within flood plain (which includes both 100 year and 500 year flood plain), if it is then value will take "1", otherwise it will take "0".
- X2: dam or floodgate variable - If the road in question is close (within a mile) to a dam or flood gate, then it will take the value of "1", otherwise "0".
- X3: bayous variable - If road intersects or passes through a bayou, then it will take the value of "1", otherwise it will take the value of "0".
- X4: channel variable - If road intersects or passes through a channel, then it will take the value of "1", otherwise it will take the value of "0".
- X5: open field variable - Usually there are two sides of a road, if both sides have open space (no occupied by residential or commercial building blocks), then the value will take 2. One side has , then "1", if no sides of the road has it, then it will take value of "0".

It is worth mentioning that the values for the chosen variables have been manually obtained from various sources and values for some features (i.e. open space) are visual estimation based on the 3D map of the region on Google Earth, which introduces certain level of subjectivity to the result of the observation. Apart from that, considering the numerous factors, as well as their complex dynamic nature, which all could lead to flooding of streets, there could be many other variables like, elevation of the roads, slope configurations, alignment, the type of materials used for pavement, availability and configurations of the drainage underneath of the streets, the sheer volume of rain drops, the topographical characteristics of the roads nearby, which could undermine the vulnerability of roads for the flooding.

B. Data Acquisition

The data for this project has been obtained from multiple sources. First of all, we have obtained the road closure information from Houston Transtar, an organization which mapped the live traffic condition for Harris County during the Hurricane Harvey. We have chosen to work on the roads on the west side of Houston due to the fact that roads in this region is relatively heavily flooded compared to the east and spatial distribution flooding is relatively even. For the road (or street) sections which have not been flooded, we have chosen roads in the vicinity of flooded ones, in order to control for other factors. In total we have identified 75 roads (or street) sections to work on. Second, we have obtained the flood plain map for the region of interest from FEMA and Harris County Flood Control District website. These maps helped us to identify whether or not certain road lies in the flood plain. It worth mentioning that, there are two types of flood plains, 100 -year and 500-year, which have different likelihood of being inundated during heavy downfalls. This study didn't make this distinction for the flood plain feature of roads due to the fact that, it requires more accurate and higher resolution map. Third, a combination of Google Map and Google Earth tools have been used for the identification of the values for the other features.

VI. RESULTS

As previously mentioned, we ran many classifiers on our data, but we ended up deciding to only use the five highest accuracy classifiers. Although we were able to run many classifiers for the project, we were not as fortunate with the data. We only managed to get 75 points of data from Houston Transtar, which we distilled five features from. In an optimal scenario, we would have much more data with more features, but due to the available data for this project, this was not the case. That said, we did manage to get relatively good results considering the data that we do have, with the accuracy on some classifiers rising as high as 84%.

The charts featured below will explain the accuracy levels of different classifiers as we tuned the hyper-parameters for each. First, we will look at the highest accuracy achieved for each classifier:

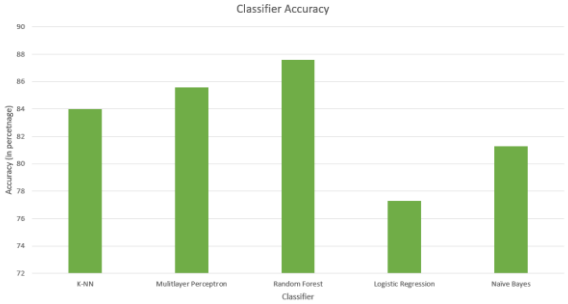

Figure 2: Accuracy of different classifiers compared side-by-side

As shown above, the highest accuracy classifier was the Random Forest. This is likely due to the low number of features and thus limited branching factor of the trees in the forest. The next few graphs will show the change in accuracy as hyper-parameters change for those classifiers that have hyper-parameters to change. First, we'll look at the accuracy of the K-Nearest Neighbor classifier as we tune the hyper-parameter (in this case, the value of K):

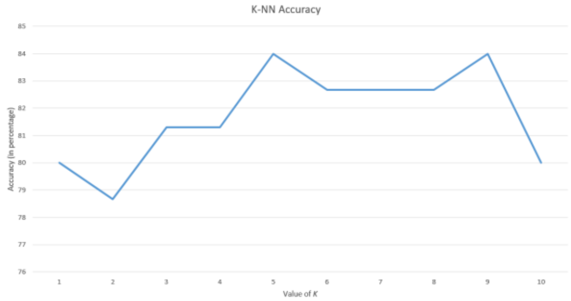

Figure 3: Accuracy of the KNN classifier as the value of K was changed

As the graphs reveals, the highest accuracy was achieved with a K value of 5. As the value of K was further changed, we noticed that accuracy decreased, likely due to overfitting. The other two graphs we will look at are the accuracy of the Multilayer Perceptron classifier as we changed the hyper parameter of learning rate and accuracy of the Random Forest as we changed the max depth hyper parameter.

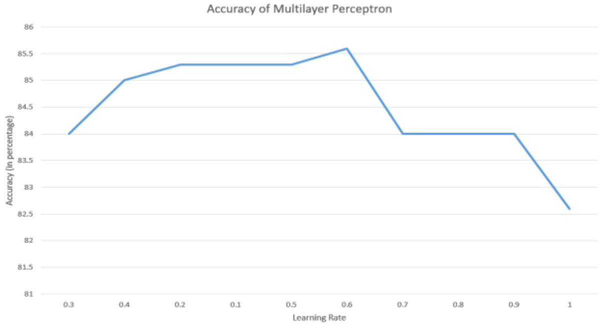

Figure 4: learning rate vs prediction accuracy

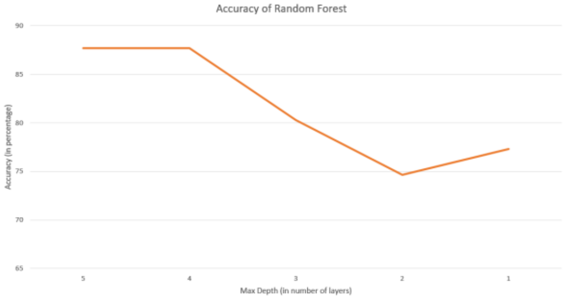

Figure 5: Max-depth vs prediction accuracy

The figure on the left shows that as the learning rate increases, the accuracy decreases for the Multilayer Perceptron, likely due to the fact that the learning rate increasing means that the classifier will make faster, but not necessarily more accurate, decisions. The figure on the right shows that as the max depth of the trees in the random forest decrease, so too does the accuracy, because the classifier cannot be as granular with fewer levels.

VI. CONCLUSION

This study aimed to find a classifier and feature set that could predict the critical infrastructure vulnerability of an area. The results, as shown in the previous section, found that the Random Forest classifier combined with the five features listed gives the highest prediction accuracy for critical infrastructure vulnerability. By testing out a variety of classifiers on our feature set, we have shown that it is possible to predict with a relatively high accuracy how vulnerable a road will be to disaster events.

There are several potential avenues for future works in relation to this study. Perhaps the most obvious choices are the addition of more features and more data. Although 75 data points, as we used in our study, is sufficient to begin to draw conclusions for our specific domain (Hurricane Harvey), it is not sufficient to apply these conclusions to other disaster events. One promising approach would be automation of the data collection, an example of which could be extracting the feature variables from the geographical information of the roads. More data from different disaster events would help provide a more generalized critical infrastructure analysis system. Additionally, more features could be incorporated to test the effects of different factors on the vulnerability of an area. Additional room for expansion includes applying our results to different domains such as terrorist attacks, using a mix of classifiers, and using a deep neural network instead of a simple neural network.

VII. REFERENCES

[1] M. Bruneau *et al.*, “A framework to quantitatively assess and enhance the seismic resilience of communities,” *Earthq. spectra*, vol. 19, no. 4, pp. 733–752, 2003.

[2] S. E. Chang and M. Shinozuka, “Measuring improvements in the disaster resilience of communities,” *Earthq. Spectra*, vol. 20, no. 3, pp. 739–755, 2004.

[3] K. Heaslip, W. Louisell, and J. Collura, “A methodology to evaluate transportation resiliency for regional networks,” 2009.

[4] D. Freckleton, K. Heaslip, W. Louisell, and J. Collura, “Evaluation of transportation network resiliency with consideration for disaster magnitude,” in *91st annual meeting of the transportation research board, Washington, DC*, 2012.

[5] N. U. Serulle, *Transportation network resiliency: A fuzzy systems approach*. Utah State University, 2011.

[6] W. H. Ip and D. Wang, “Resilience and friability of transportation networks: evaluation, analysis and optimization,” *IEEE Syst. J.*, vol. 5, no. 2, pp. 189–198, 2011.

[7] G. Leu, H. Abbass, and N. Curtis, “Resilience of ground transportation networks: a case study on Melbourne,” 2010.

[8] D. King, A. Shalaby, and P. Eng, “Performance Metrics and Analysis of Transit Network Resilience in Toronto,” in *Transportation Research Board 95th Annual Meeting*, 2016, no. 16–2441.

[9] M. S. Tehrany, B. Pradhan, and M. N. Jebur, “Flood susceptibility mapping using a novel ensemble weights-of-evidence and support vector machine models in GIS,” *J. Hydrol.*, vol. 512, pp. 332–343, 2014.

[10] H. Mojaddadi, B. Pradhan, H. Nampak, N. Ahmad, and A. H. bin Ghazali, “Ensemble machine-learning-based geospatial approach for flood risk assessment using multi-sensor remote-sensing data and GIS,” *Geomatics, Nat. Hazards Risk*, pp. 1–23, 2017.

[11] D. T. Bui, T. C. Ho, I. Revhaug, B. Pradhan, and D. B. Nguyen, “Landslide susceptibility mapping along the national road 32 of Vietnam using GIS-based J48 decision tree classifier and its ensembles,” in *Cartography from pole to pole*, Springer, 2014, pp. 303–317.

[12] G. L. Harvey *et al.*, “Qualitative analysis of future flood risk in the Taihu Basin, China,” *J. Flood Risk Manag.*, vol. 2, no. 2, pp. 85–100, 2009.

[13] J. French, R. Mawdsley, T. Fujiyama, and K. Achuthan, “Combining machine learning with computational hydrodynamics for prediction of tidal surge inundation at estuarine ports,” *Procedia IUTAM*, vol. 25, pp. 28–35, 2017.