PUS2020 final: project report

Don’t modify above this line. Everything in blue below this line has to be updated. Everything in gray should be read carefully, followed carefully, and removed from the template

---

SPPA667

Urban Flood Risk and Social Vulnerability Assessment in Wilmington, Delaware: A Comparative Study of Machine Learning Techniques

Kyungmin Lee, github.com/KLeeDE, kmlee@udel.edu

Solo project.

Abstract:
Analysis between flood risk and socioeconomic status has been researched in various nations, which presented the vulnerable people, who belong to groups within society that is either oppressed or more susceptible to harm, are more exposed to the urban flood risk. However, the issue, whether vulnerable people living in the Wilmington city are exposed to flood risk, has not been researched in case of the City of Wilmington in Delaware. Therefore, this research aims to analyze flood risk and socioeconomic and demographic status and find out whether vulnerable people are living in the urban flood risk zone and predict the future flood risk impact based on 2014-2018 census data. This research conducted data correlation analysis, geospatial analysis between flood risk and socioeconomic and demographic data and applied machine learning techniques for prediction. As for the machine learning techniques, this research used random forest classifier, logistic regression classification, and random forest regressor, and compared the model accuracy. As a result, this research finds out the low median income household, people not white, and people living under poverty line are more exposed to the urban flood risk in Wilmington. However, the random forest classifier and the logistic regression classification show too high model accuracy, while the random forest regressor shows too low model accuracy. Although there are significant limitations, this research implies that the vulnerable people are exposed to the flood risk in Wilmington. During the policy-making decision process, the state government may have to consider the vulnerable people and flood risk in the local climate adaptation policy.

Introduction:

Analysis between flood risk and socioeconomic status has been researched in various n ations, which presented the vulnerable people, who belong to groups within society that is either

PUS2020 final: project report

Don’t modify above this line. Everything in blue below this line has to be updated. Everything in gray should be read carefully, followed carefully, and removed from the template

---

oppressed or more susceptible to harm, are more exposed to the uraban flood risk. The urban flood risk is important in terms of the environmental justice and vulnerability to protect minority. It is because the vulnerable people are more exposed to the flood risk and more suceptible to disaster such as the flood (Eisenman et al., 2007; Walker et al., 2011; Maantay et al., 2009), which should be considered during the policy making process. However, urban flood risk and social assessment has not been studied in the urban area in Delaware, where the state is in the east coast area and exposed to the annual flood risk. Therefore, this research aims to conduct geospatial analysis of the urban flood risk and social vulnerability in Wilmington and find out the applicable machine learning techniques for the future studies.

Background:

The flood risk and socioeconomic status has been researched in various nations. The previous studies analyzed the relationship between flood risk and socioeconomic status using machine learning techniques at the urban city level. For example, Eini et al. (2020) conducted hazard and urban flood risk mapping using Maximum Entropy (MaxEnt), and Genetic Algorithm Rule-Set Production (GARP) and evaluated the role of urban districts in Kemanchah city, Iran. Chakraborty et al. (2020) analyzed the social vulnerability to flood hazards in the context of environmental justice in Canada using sampling adequacy and variance test. Darabi et al. (2019) analyzed the urban flood risk mapping using the GARP and Quick Unbiased Efficient Statistical Tree (QUEST) models and compare the machine learning techniques.

However, the City of Wilmington in Delaware has not been researched. Also, it has not been researched based on the urban evidence-based policy. This purpose of this research is to find out the circumstances of the socioeconomic and demographic status is in the flood zone and suggest policy implications. Therefore, this research aims to answer the following question: Are vulnerable people living in the Wilmington city exposed to flood risk?

This research is conducted based on the approach of urban evidence-based policy. The Evidence Based Policy is to make more defensibly policy decisions based on conscientious, explicit, and judicious use of scientific evidence by using science (Big Data and machine learning) as evidence in public policy (Straf et al., 2012). In this way it is useful for policy makers by providing insights into which policy interventions are most likely to lend desirable outcomes (Androutsopoulou et al., 2018).

In this research, a vulnerable person can be defined as someone who belongs to a group within society that is either oppressed or more susceptible to harm, belonging to populations such as children, senior citizens, low-income workers, and asylum seekers. The United Nations Convention on the Rights of the Child defines child as "a human being below the age of 18 years unless under the law applicable to the child, majority is attained earlier" Medicare enrollees aged 65 years.

PUS2020 final: project report

Don't modify above this line. Everything in blue below this line has to be updated. Everything in gray should be read carefully, followed carefully, and removed from the template

---

***Data Description**:*

There are three datasets used in this research. The details are following:

Dataset name	URL	Number of rows	Number of columns	Number of relevant columns	Number of valid rows (not NaN on relevant columns)	Data type for each relevant column
Flood Depth Grid (New Castle County)	https://github.com/KLeeDE/PUS2020_Klee/tree/master/Project/data	18244	10134	10134 *This is a raster image dataset	18244 *This is a raster image dataset	'dep': Array
Census Data Block Group (Delaware)	https://github.com/KLeeDE/PUS2020_Klee/tree/master/Project/data	574	13	3	574	'GEOID': string 'AWATER': integer 'geometry': polygon
Socioeconomic and Demographic (SED) data (New Castle County) *This dataset is cleaned up for analysis	The API key is stored in the google drive. However, if needed, please contact the main author. *Data name: nc_demo	365	8	8	365	'cblockgid': string 'mincome': integer 'per_nonwhite': float 'per_below_povlev': float 'population': integer 'no_school': integer 'under18': integer 'over65': integer

Flood depth grid is the one percent annual chance flood in coastal area at New Castle County (Unit: Feet). Flood hazard is defined by a relation between depth of flooding and the annual chance of inundation greater than that depth. Depth grid is defined by the percent annual chance floods. This is usually only the 1% annual chance flood. This figure is based on the raster data from the Federal Emergency Management Agency (FEMA) in 2014, which is the most recent available data in Delaware.

PUS2020 final: project report

Don’t modify above this line. Everything in blue below this line has to be updated. Everything in gray should be read carefully, followed carefully, and removed from the template

---

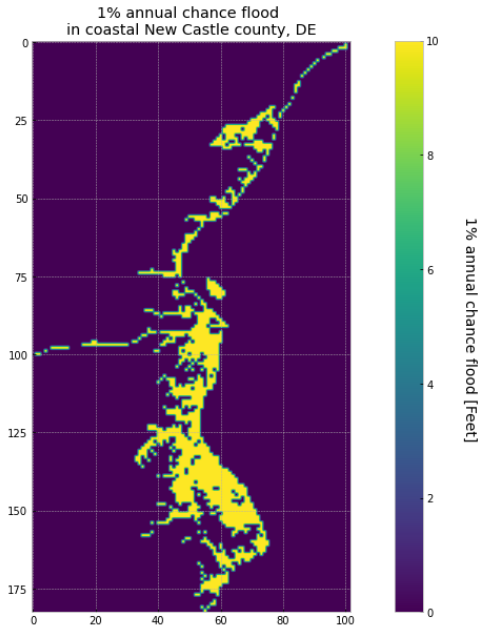

Fig1. Flood Depth Grid

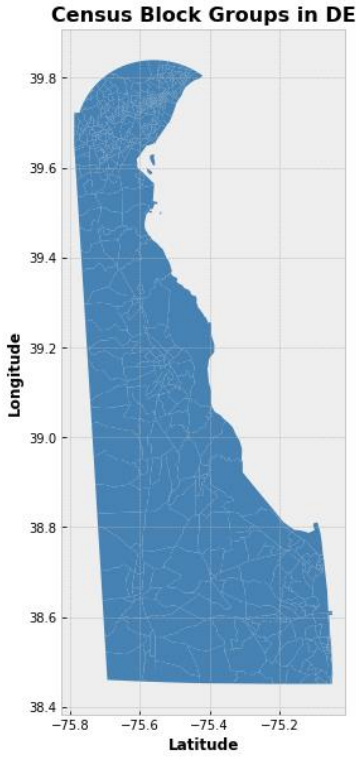

Fig 2. Census Block Groups in Delaware

PUS2020 final: project report

Don’t modify above this line. Everything in blue below this line has to be updated. Everything in gray should be read carefully, followed carefully, and removed from the template

---

	lat	lon	dep	dep_cl	mincome	per_nonwhite	per_below_povlev	population	no_school	no_school_frac	under18	over65	has_water_int	log_dep
count	203.000000	203.000000	203.000000	203.000000	203.000000	203.000000	203.000000	203.000000	203.000000	203.000000	203.000000	203.000000	203.000000	203.000000
mean	39.750172	-75.553137	0.516256	0.516256	55261.177340	57.237379	18.446325	967.147783	727.669951	0.756534	200.970443	135.093596	0.285714	0.066267
std	0.011163	0.016557	1.620714	1.620714	30641.633629	32.408620	17.069917	315.344881	232.317797	0.087995	128.113460	98.604802	0.452871	0.203469
min	39.724676	-75.586413	0.000000	0.000000	10177.000000	3.810976	0.000000	389.000000	280.000000	0.417404	21.000000	12.000000	0.000000	0.000000
25%	39.741115	-75.565193	0.000000	0.000000	32206.000000	20.620621	2.692308	719.500000	576.000000	0.701105	99.000000	64.000000	0.000000	0.000000
50%	39.752075	-75.554583	0.000000	0.000000	47976.000000	66.979362	16.071429	962.000000	745.000000	0.762509	187.000000	108.000000	0.000000	0.000000
75%	39.760295	-75.540436	0.000000	0.000000	69403.500000	88.115942	28.991597	1141.000000	876.000000	0.793226	261.000000	178.000000	1.000000	0.000000
max	39.771255	-75.519216	10.000000	10.000000	162143.000000	100.000000	77.600000	1792.000000	1179.000000	1.022071	673.000000	390.000000	1.000000	1.000000

Fig 3. SED data description

Methodology:

This research conducted data correlation analysis, geospatial analysis between flood risk and socioeconomic and demographic data and applied machine learning techniques for prediction. As for the machine learning techniques, this research used random forest classifier, logistic regression classification, and random forest regressor, and compared the model accuracy. The data exploration is constituted of data summary, distribution, and correlation. The geospatial analysis is done by spatial joining flood depth grid and census block group shapefile, merging census block groups shapefile to the SED data, and sub-setting New Castle County dataset to Wilmington City dataset. For comparison between machine learning techniques, this research used three models: random forest classifier, logistic regression classification, and random forest regressor.

Firstly, this research conducted the random forest classifier and the logistic regerssion classification and aimed to compare two models together. The classification is the problem of predicting a "discrete" class label output. Presence or absence of flood risk is used for target variable. Census block groups are objects. Features are used as estimated median income, percent of population that is not white people, percent of people below the poverty line, fraction of people who is not enrolled in school, estimated total number of children (age under 18), estimated total number of seniors (age over 65), water area (where the fraction of the current water area is over 0.05).

However, the result of the classification models shows 100% model accuracy. It means the target variable is too easy to be predicted. Therefore, this research also conducted the regression model, which is used for the problem of predicting a "continuous" quantity output. Target variable used in regressor model is the 1% annual chance flood risk (multilevel). The object and features are same as classification model.

As a result, the random forest classifier and the logistic regression classification show too high model accuracy, while the random forest regressor shows too low model accuracy. To improve the model accuracy, the model can be checked and solved by adding more data, treating missing and outlier values, feature engineering, feature selection, multiple algorithms, algorithm tuning, ensemble methods (bagging and boosting), and cross validation (Sunil Ray, 2015). In this

research, feature engineering, multiple algorithms, and algorithm tuning are used for improving the model accuracy and comparison between the models.

Deliverable:

Regarding to the data distribution, the flood risk zone in Wilmington city is calculated by census block groups. The 1% annual chance flood is multilevel data set, while the water area is binary dataset. The SED data is estimated from 2014 to 2018 and the data source is from the U.S. Census Bureau. Based on the census block groups, the mean of the median household income is $55,261, the percentage of not white people is 57.24%, and the rate of the total percentage of people under poverty line is 18.45%. The mean of the fraction of people not enrolled in school is 0.76 (75.65%), the total estimate number of children (age under 18) is 201, and the total estimate number of seniors (age over 65) is 135. As a result of the correlation analysis, the correlation between the 1% annual chance flood risk and percentage of people not white, and the percentage of the people below poverty line are highly correlated.

The flood risk status by census block groups in Wilmington, DE. The 1% annual chance flood is multilevel data set, while the water area is binary dataset.

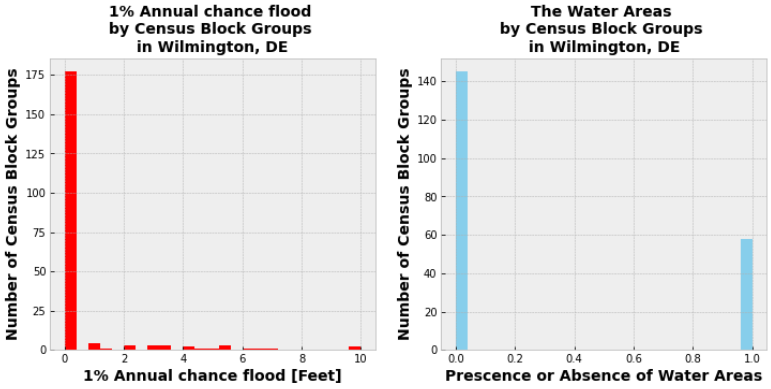

Fig 4-1. The Flood Depth & Water Areas

The data is estimated from 2014 to 2018 and the data source is from the U.S. Census Bureau. Based on the census block groups, the mean of the median household income is $55,261, the percentage of not white people is 57.24%, and the rate of the total percentage of people under poverty line is 18.45%. The mean of the fraction of people not enrolled in school is 0.76 (75.65%), the total estimate number of children (age under 18) is 201, and the total estimate number of seniors (age over 65) is 135.

PUS2020 final: project report

Don’t modify above this line. Everything in blue below this line has to be updated. Everything in gray should be read carefully, followed carefully, and removed from the template

---

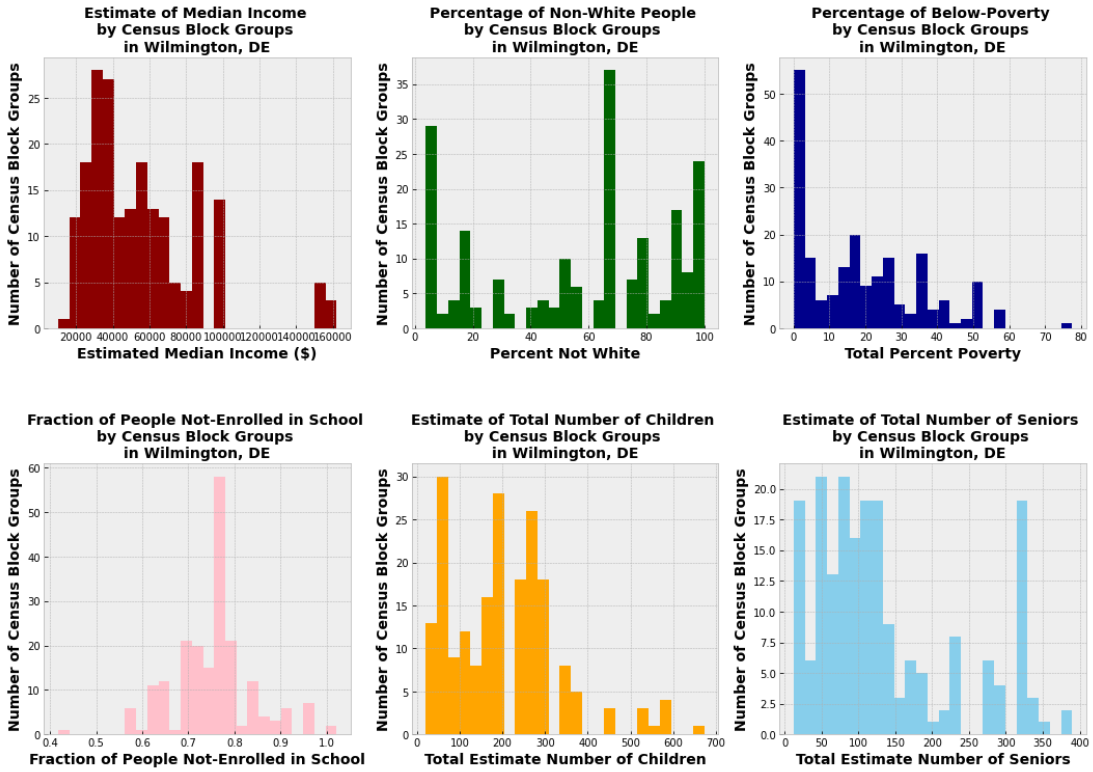

Fig 4-2. The socioeconomic demographic status in Wilmington, DE

The correlation between the 1% annual chance flood risk and percentage of people not white, and the percentage of the people below poverty line are highly correlated.

PUS2020 final: project report

Don’t modify above this line. Everything in blue below this line has to be updated. Everything in gray should be read carefully, followed carefully, and removed from the template

---

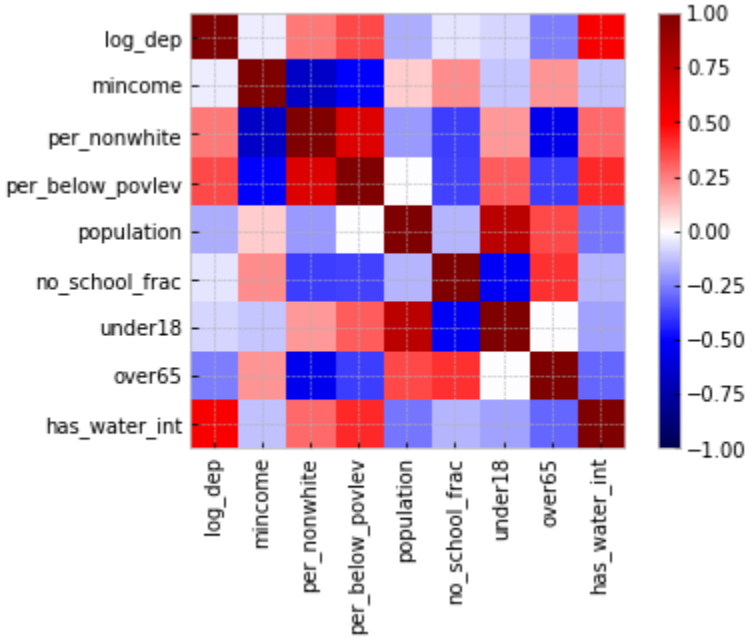

Fig 5. Correlation between features

As for the geospatial analysis, the figures below show only three census block groups are exposed to the flood risk. However, the FEMA data I used in this research did not show whether the flood depth is measured in area around the flood risk zone. The issue related to the data availability is an issue particularly when researching the middle size of the urban city rather than when researching the megacity such as the New York City.

Based on the geospatial analysis, people living in the low median income household, people not white, and people under poverty line are living in the flood hazard zone. However, people not enrolled in school, children (people under 18 years old), and seniors (people over 65 years old) are not highly close to the flood hazard zone in the City of Wilmington, DE.

Based on the geospatial analysis, people living in the low median income household, people not white, and people under poverty line are living in the flood hazard zone in the City of Wilmington, DE.

PUS2020 final: project report

Don’t modify above this line. Everything in blue below this line has to be updated. Everything in gray should be read carefully, followed carefully, and removed from the template

---

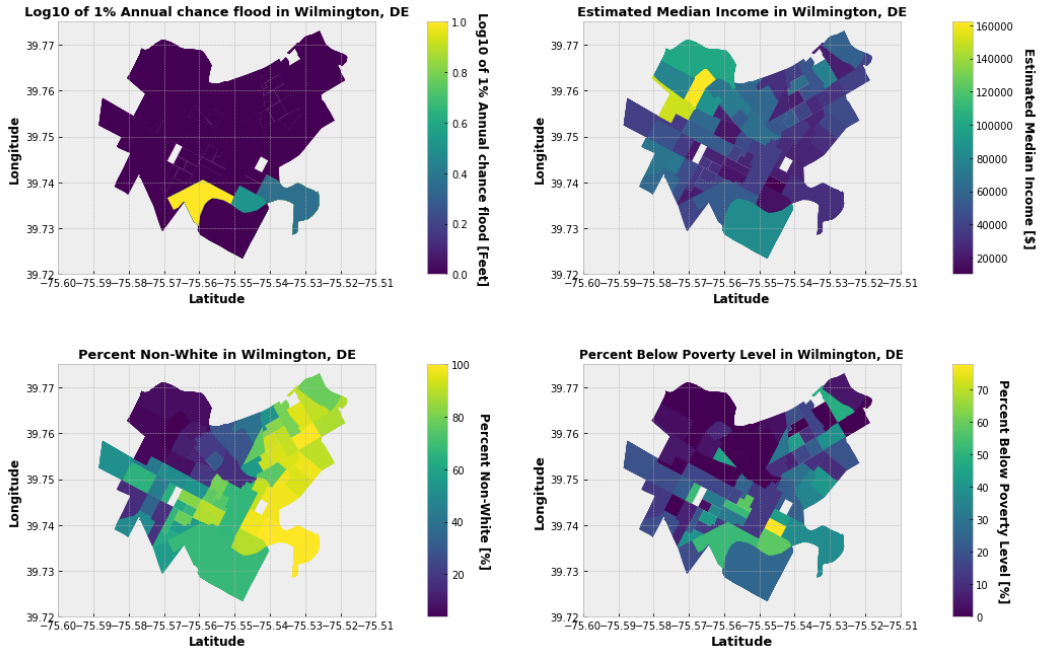

Fig 6-1. Geospatial comparison between flood risk and SED data

Based on the geospatial analysis, people not enrolled in school, children (people under 18 years old), and seniors (people over 65 years old) are not highly close to the flood hazard zone in the City of Wilmington, DE.

Fig 6-2. Geospatial comparison between flood risk and SED data

---

In order to compare the machine learning techniques, this research applied three ways for improving model accuracy: feature engineering, multiple algorithms, and algorithm tuning. Firstly, target variable is changed to log10 scale for normalization. Changing the scale of a variable from original scale to scale between zero and one helps to extract more information from existing data. It may have a higher ability to explain the variance in the training data and giving improved model accuracy. Secondly, this research applied multiple machine learning algorithms such as random forest classifier, logistic regression classification, and random forest regressor. As mentioned above, since the classification models are showing the high model accuracy, this research applied regression model to find the better suited algorithm and compare the model accuracy. Lastly, this research tuned the algorithm. By using the library tool, this research finds out the optimum value for each parameter to improve the accuracy of the model, which makes the model accuracy better. The result is shown as below. However, the random forest regressor model accuracy is so low that it is hard to get the significant result.

This figure shows the flood risk data is not equally distributed. This makes random forest classifier and logistic regression classification did not fit in this research. It makes the target variable too easy to be predicted.

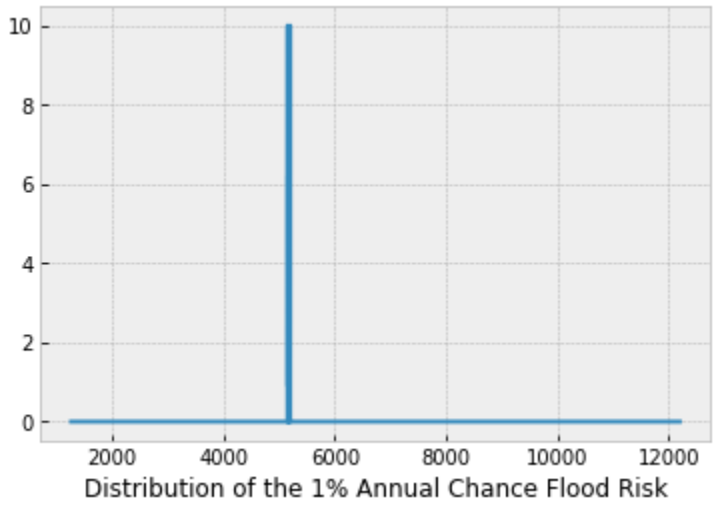

Fig 7. Distribution of the 1% Annual Chance Flood Risk

The data exploration result shows that the flood risk and socioeconomic and demographic data are not equally distributed. The correlation between the 1% annual chance flood risk and percentage of people not white, and the percentage of the people below poverty line are highly correlated.

According to the geospatial analysis result, the vulnerable people is exposed to the flood risk in the Wilmington, Delaware. People in the low median income household, people not white, and people under poverty line are exposed to the flood risk. The geospatial analysis result implies

that the state government officers, policy makers, and researchers may have to consider the vulnerable people and flood risk in the local climate adaptation policy.

However, the random forest classifier and the logistic regression classification show too high model accuracy, while the random forest regressor shows too low model accuracy. This research aims to compare the models by feature engineering, multiple algorithms, and algorithms tuning. The random forest classifier and the logistic regression classification show too high accuracy, which means it is too easy to be classified. Therefore, the model is changed to the random forest regressor and tuned for improving the accuracy. Nevertheless, the result is not significant. The reason of this issue can be occurred because of the quality of the data in that the number of the object in the model is only 203 census block groups.

Overall, in the urban evidence-based policy context, policy makers can consider the analysis result as a factor for the policy intervention during the policy-making decision process. Since the flood risk is highly linked to climate change, which is reported to make sea level rise in the long term, the flood risk is considered in the climate policy nationally and internationally.

There are significant limitations in this research. First, in terms of the middle size of the urban city, it is limited to get diversified dataset of the flood depth grid from the open data source platform. Second, since the accuracy of the RFR machine learning model is very low, the model should be more adjusted. Lastly, the flood risk depth raster data has limitation information of how it is collected. For the future study, the machine learning model should be adjusted to increase model accuracy. To expand the research, the New York City can be researched for the further study.

Link to GitHub repo: https://github.com/KLeeDE/PUS2020_KLee/tree/master/Project/Final

Bibliography:

Androutsopoulou, A., & Charalabidis, Y. (2018, April) A framework for evidence based policy ma king combining big data, dynamic modelling and machine intelligence. In Proceedings of th e 11th International Conference on Theory and Practice of Electronic Governance (pp. 575 -583).

Chakraborty, Liton, Horatiu Rus, Daniel Henstra, Jason Thistlethwaite, and Daniel Scott. 2020. "A Place-Based Socioeconomic Status Index: Measuring Social Vulnerability to Flood Hazards in the Context of Environmental Justice." *International Journal of Disaster Risk Reduction* 43 (November 2019): 101394. https://doi.org/10.1016/j.ijdrr.2019.101394.

Darabi, Hamid, Bahram Choubin, Omid Rahmati, Ali Torabi Haghighi, Biswajeet Pradhan, and Bjørn Kløve. 2019. "Urban Flood Risk Mapping Using the GARP and QUEST Models: A Comparative Study of Machine Learning Techniques." *Journal of Hydrology* 569 (February 2018): 142–54. https://doi.org/10.1016/j.jhydrol.2018.12.002.

Eini, Mohammad, Hesam Seyed Kaboli, Mohsen Rashidian, and Hossien Hedayat. 2020. "Hazard and Vulnerability in Urban Flood Risk Mapping: Machine Learning Techniques and Considering the Role of Urban Districts." *International Journal of Disaster Risk Reduction*

PUS2020 final: project report

Don’t modify above this line. Everything in blue below this line has to be updated. Everything in gray should be read carefully, followed carefully, and removed from the template

---

50 (May): 101687. https://doi.org/10.1016/j.ijdrr.2020.101687.
Eisenman, David P., Kristina M. Cordasco, Steve Asch, Joya F. Golden, and Deborah Glik. "Disaster planning and risk communication with vulnerable communities: lessons from Hurricane
Katrina." *American journal of public health* 97, no. Supplement_1 (2007): S109-S115.
Mosavi, Amir, Pinar Ozturk, and Kwok Wing Chau. 2018. “Flood Prediction Using Machine Learning Models: Literature Review.” Water (Switzerland) 10 (11): 1–40. https://doi.org/10.3390/w10111536.
Rahmati, Omid, Ali Golkarian, Trent Biggs, Saskia Keesstra, Farnoush Mohammadi, and Ioannis N. Daliakopoulos. 2019. “Land Subsidence Hazard Modeling: Machine Learning to Identify Predictors and the Role of Human Activities.” Journal of Environmental Management 236 (February): 466–80. https://doi.org/10.1016/j.jenvman.2019.02.020.
Silva, M. M.G.T. De, and Akiyuki Kawasaki. 2020. “A Local-Scale Analysis to Understand Differences in Socioeconomic Factors Affecting Economic Loss Due to Floods among Different Communities.” International Journal of Disaster Risk Reduction 47: 101526. https://doi.org/10.1016/j.ijdrr.2020.101526.
Straf, M. L., Prewitt, K., & Schwandt, T. A. (2012) Using science as evidence in public policy.
Walker, Gordon, and Kate Burningham. "Flood risk, vulnerability and environmental justice: evidence and evaluation of inequality in a UK context." Critical social policy 31, no. 2 (2011): 216-240.
Maantay, Juliana, and Andrew Maroko. "Mapping urban risk: Flood hazards, race, & environmental justice in New York." Applied Geography 29, no. 1 (2009): 111-124.