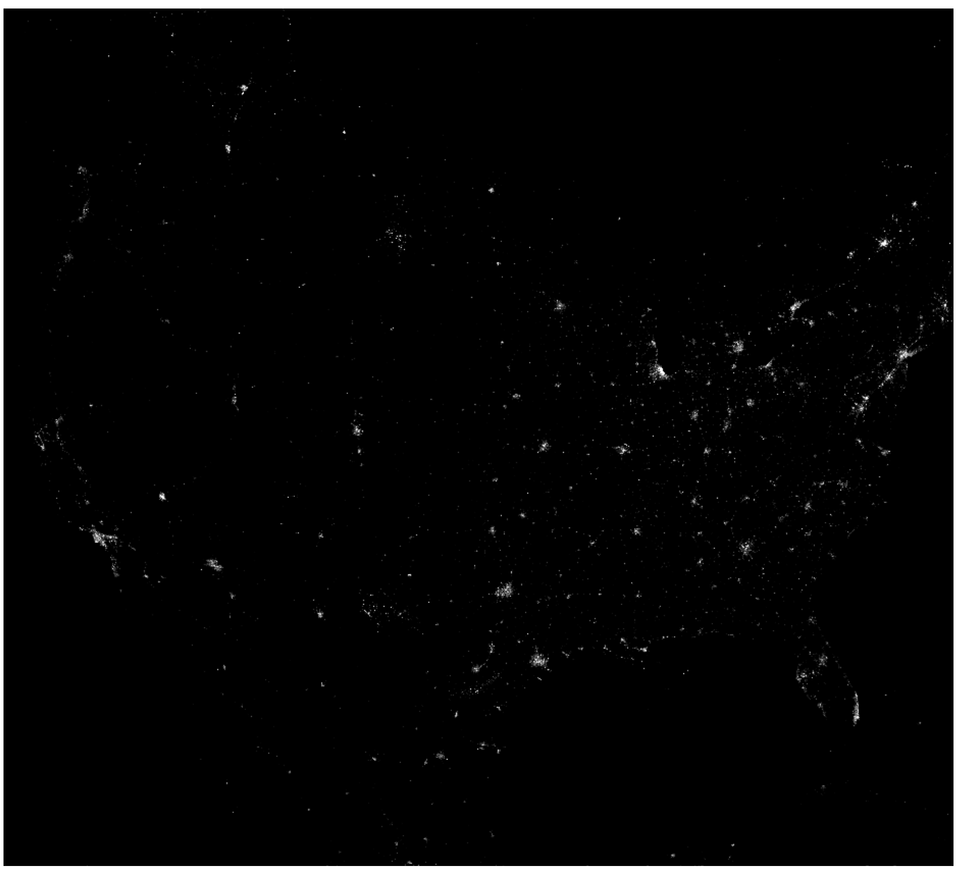
Social Vulnerability in Flood Prone Areas:
Using Satellite Imagery as a Socio-Economic Predictor

Matthew Preisser
Dual Master's Student Environmental and Water Resources Eng. and Public Affairs
GIS for Water Resources Fall 2018


Unprocessed Day Night Band Imagery showing the Outline of the United States

I. INTRODUCTION

a. Motivation

The frequency of extreme weather events, including category 4 and 5 hurricanes, are predicted to double by the end of the 21st century (Bender et al, 2010). Hurricane Harvey, Florence, and Michael are just some of the more recent destructive examples to have made landfall on the U.S. Beyond the environmental damage and the threat to cripple critical infrastructure, extreme weather events have the ability to disrupt food production, spread diseases, and displace populations (Patz et al, 2005). However, there are apparent disparities on how these storms affect different socioeconomic classes. Depending on your access to resources will affect your ability to cope with a natural disaster and often times resources are not distributed evenly among populations. There has been an expressed desire from US Agencies including NASA and USAID, as well as international organizations (ADPC, 2018) to incorporate near real time socioeconomic data into existing flood inundation models. Census data collection processes can lead to inaccuracies due to poor sampling methods. Along with this, census data is collected infrequently and costs taxpayers billions of dollars. The goal of this project is to begin to explore the use of satellite imagery, specifically the Visible Infrared Imaging Radiometer Suite (VIIRS) Day Night Band (DNB) from the Suomi National Polar-orbiting Partnership (S-NPP) satellite to estimate population and to develop a framework for predicting social vulnerability in a flood prone area. This project is a very basic introduction to numerous graduate research proposals currently being reviewed by the National Science Foundation and NASA.

b. Literature Review

There are numerous studies that have been and are currently being conducted to identify correlations between night time satellite imagery and socioeconomic factors including but not limited to population (Sutton et al, 2001), gross domestic product (Doll et al, 2006), poverty (Elvidge et al, 2009), electrical consumption (Chand et al, 2009), and carbon dioxide emissions (Shi et al, 2016). Other studies have also examined the use of other satellite imagery to predict socioeconomic factors such as the correlation between vegetation cover and poverty (Sunderlin, et al, 2008). The intended use of many of these studies is to the specific application of publicly available imagery to benefit rural countries that have little to no access to census data (Zhao et al, 2018). Remote sensing socioeconomic data has the ability to be near real-time while removing human biases in sampling techniques, giving agencies an up-to-date snapshot of a region, in this case right before extreme weather events occur. A key component of this project (and future intended research proposals) is to use open source/public datasets and code, as the produced models should not be excluded or restricted from any organization/agency that desires to use it.

c. Study Area

The study area chosen for the purpose of this project was an area effected by Hurricane Harvey. With the end goal of this research being the application of predicting social vulnerability during natural disasters, this study area is representative of a common extreme weather event that results in high levels of flooding. Figure 1 shows the Disaster-Declared counties in Texas both at the state and federal level. The accompanying figure (Figure 2) shows a rescaled DNB monthly composite image of the visible night time lights. The scale in this image has been adjusted to show relative radiance values in order to more clearly see differences between areas with light and areas without light. Due to computing limitations, analyzing all of these counties would not be possible at one time, so a single sub-basin was selected (Figure 3). This sub-basin directly encompasses the downtown Houston area and Figure 3 shows the census tracts for this location. This specific sub-basin (NHD sub-basin 12040104) as it has seen several severe floods in the past couple of years and is one of the largest metropolitan areas in the country.

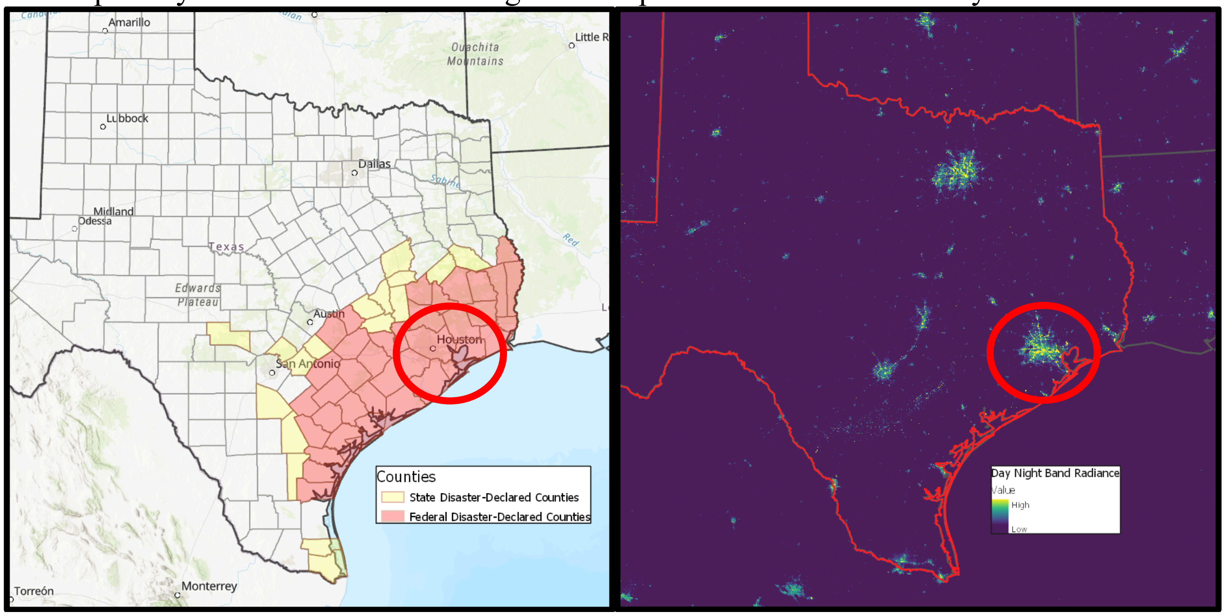


Figure 1 & 2: Disaster declared counties from Hurricane Harvey (left) and rescaled DNB monthly composite (right)

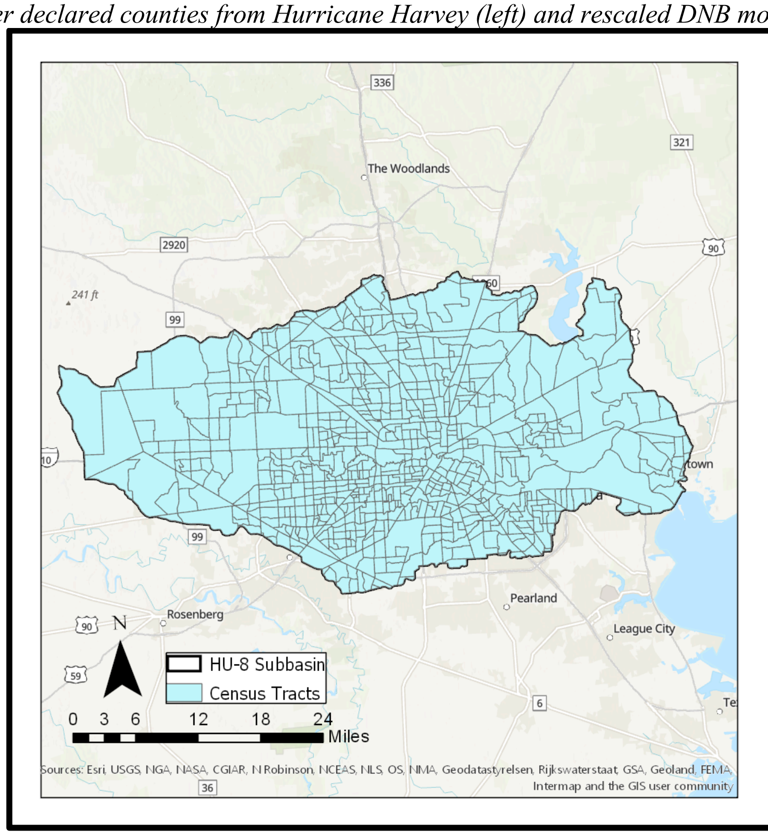


Figure 3: Sub-basin 12040104 (Houston Metropolitan Area) and the Census Tract areas

II. METHODOLOGY

This project can be divided into two separate components: social vulnerability determination and satellite imagery analysis. For the social vulnerability determination, census estimation data will be used to determine what areas of the previous mentioned sub-basin in Houston are more vulnerable to flooding based on a HAND Analysis (height above nearest drainage). This will serve as a snapshot as what census tracts are most vulnerable. A discussion on developing this method into a social vulnerability prediction tool is discussed later on in this report. The second component of this project is to analyze DNB satellite imagery for the state of Texas to determine if similar conclusions that have drawn in current research can be made for the state of Texas (does night time light correlate with population). All data sets were put in the same projection as well (NAD 1983 UTM 15).

a. Data Sources

Three main data sources were used during this project and are listed in Table 1.

Table 1: Data sources and their purposes

Data Source	Purpose
VIIRS DNB October 2018 Monthly Composite	The most recently available monthly composite of nighttime radiance was used as a measure of night time light. Monthly composites take the average cloud free radiance values each night and has 750-meter resolution and are also filtered to exclude stray lights from fires and astronomical anomalies. This data was acquired from the National Centers for Environmental Information Website (formerly known as the National Geophysical Data Center).
NHD Sub-basin HU8: 12040104	As previously mentioned, due to computing limitations, a single NHD Sub-basin was chosen to be analyzed. This sub-basin contains the majority of the downtown Houston area. This geodatabase (including streams, watershed boundary, and a DEM file) were accessed through the United States Geological Survey National Hydrology Database website.
Census Tract Estimation Data (2016 Estimates)	For the purpose of determining social vulnerability, census data at the tract level was collected from the U.S. Census Bureau TIGER/Line website from 2016. The specific data layers that were used include population, median age, median household income, and labor force unemployed.

b. Fuzzy Logic

Fuzzy Logic is a mathematical many-value logic formula first introduced by Lotfi Zadeh in 1965 (Zadeh, 1965). In this type of analysis, data is rescaled to be any real number between 0 and 1, where 0 is completely false and 1 is completely true. This allows for values to be "partially true", compared to some logic models that only allow absolute true/false (values of 0 or 1). A common example used often used to describe fuzzy logic is in height measurement situations and is described in the Table 2.

Table 2: Fuzzy Logic Height and Tallness Example

Height	Fuzzy Logic Value	Output
5’ 0”	0	Not Tall
5’3”	0.25	Not Very Tall
5’6”	0.5	Moderately Tall
5’9”	0.75	Kind of Tall
6’0”	1	Tall

In spatial analysis, raster layer values are similarly rescaled to values between 0 to 1. The distribution of values can take on many forms including (but not limited to) the following:

- Gaussian Distribution: Values from 0-1 are distributed across a Gaussian function
- Linear Distribution: Values from 0-1 take on a linear relationship
- Small: Smaller values are given a higher weight of being true
- Large: Larger values are given a higher weight of being true

The most common application of Fuzzy Logic is in suitability analysis. Multiple layers that represent desired qualities are overlaid in order to determine what locations are the most suitable. For this project the same analysis is applied, but traits that are seen as making a location more vulnerable (i.e. lower income) are used to determine which location has the most factors effecting the population at that point. Once all of the layers are reclassified (the ArcGIS tool is called Fuzzy Membership), they are overlaid on top of each other. (using Fuzzy Overlay tool). The output of the Fuzzy Overlay tool is a new raster layer with values that range from 0 to 1, with 1 still being absolute truth and zero being absolute false.

The Fuzzy Overlay also has different weighting factors depending on what is being looked for in a location. The two most common options are a Fuzzy Membership of ‘AND’ and ‘OR’. These logical statements describe the relationship between the layers being overlaid on one another. For example, a Fuzzy Membership of AND will look for locations that have high truth vales. A Fuzzy Membership of OR will look for locations were one of the layers has a high truth value. There are other types of Fuzzy Membership (product, sum, and gamma) but these are not as common and were not used in this project.

Fuzzy Logic is useful for this type of analysis because of the ability to have partial truth values. A location may not be the most vulnerable, but it may not be perfectly immune to a disaster either, and this method still allows for the identification of those locations. Fuzzy Logic is also advantageous when using raster layers, or layers that do not have strict boundaries. For this project this is not a factor since census tract data is being used, but for future directions of only using satellite data, this is a key advantage. Another benefit to using Fuzzy Logic is that it is able to correct for missing/incorrect data. It does this during the overlay process by correcting for missing data. This again will be an advantage when moving to only using satellite data.

For this project, there were four specific socio-economic layers that are listed in the Table 3 and include the reason why they were included. These are certainly not the only socio-economic factors that determine a populations ability to recover from a natural disaster, and there are many more that could be included.

Table 3: Socio-Economic Layers used in the Fuzzy Analysis with their reason for inclusion

Socio-Economic Layer	**Reason for Inclusion**	**Fuzzy Membership Used**
Population	The more people that live in a tract mean that more resources will be needed to aid in recovery	Large
Age	Older and younger populations are more vulnerable to injury/diseases and require more care	Small and Large
Median Household Income	Financial resources and having the ability to afford food, transportation, and rebuilding costs is a major factor of someone's resiliency during extreme weather events	Small
Unemployment	Not having a job can reduce your ability to recover after a storm and higher unemployment rates in a census tract can be indicative of the social status of those that live there	Large

These were chosen as they are well known factors that are easily quantified and understand by the general population. The Fuzzy Logic Model used in this project assumed that a value of 1 is extremely vulnerable, and subsequently a value of 0 is not vulnerable at all. The specific Fuzzy Membership that was used for these 4 socio-economic layers depended on what they were measuring and is also included in the table. Linear relations were not chosen for any of the layers because it was assumed that these factors do not reduce a person's vulnerability evenly as they increase/decrease. This is best understood when considering median household income: a low income is going to have a much greater impact on someone compared to someone with an average income, and even more compared to someone with many financial assets. A Fuzzy Membership of AND was used when combining the layers. This will then identify what areas that have the highest combined truth values. Specifically, for the age layers (young and old were separated into two different layers) these were combined using a Fuzzy Membership of OR beforehand. An area that has a high or a low population will increase the truth value. The intermediate layers that were generated are also shown in Figure 4, including the HAND analysis.

A HAND analysis was also done (similarly to assignment 5) in that an area that is closer to a stream is more vulnerable to flooding. This was the layer that was used in the Fuzzy Logic model to represent environmental factors linked to vulnerability in relationship to flooding. More layers could be included in future analysis that could also represent environmental hazards (slope to represent runoff or impervious surfaces to represent low infiltration and thus increasing the likelihood of flooding). An analysis on how the HAND method has its limitations in predicting vulnerability to flooding is discussed later on in this report.

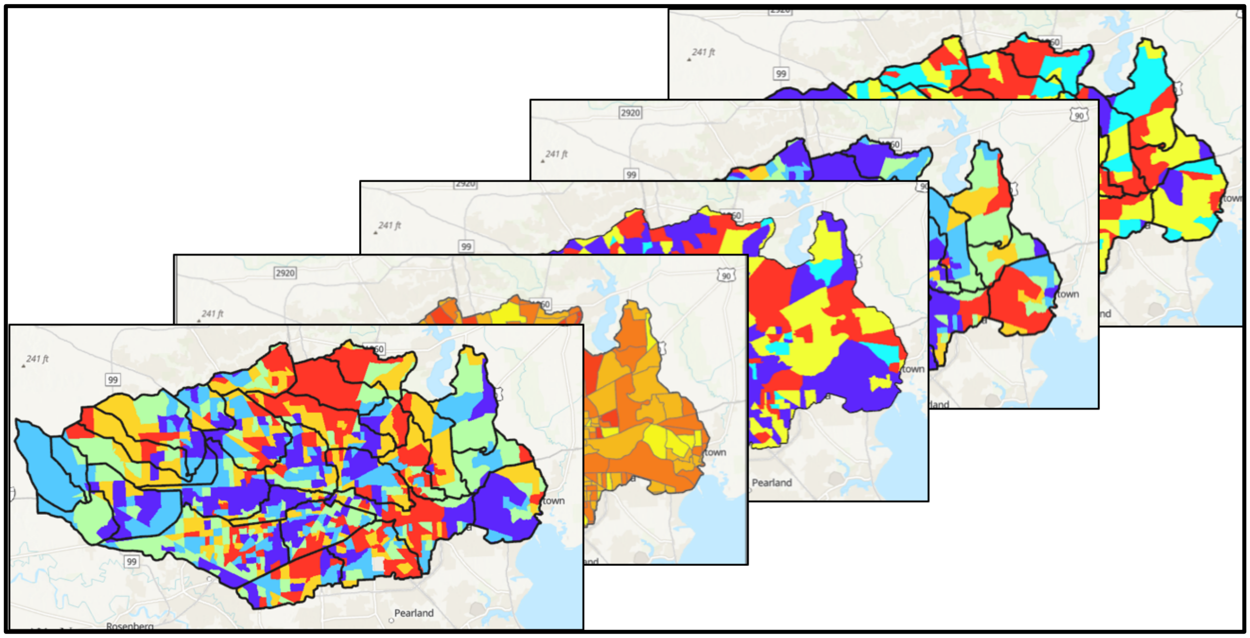


Figure 4: Different Fuzzy Layers that were overlaid to form a single vulnerability image

c. R Studio

The National Oceanic and Atmospheric Administration (NOAA) and NASA have publicly released numerous resources to make their data tools more available and accessible to researchers across the world (Blumenfeld, 2018). These resources have led to numerous GITHUB codes by associated research scientists from these agencies and others at universities and institute across the globe. These resources were adjusted, corrected, and manipulated to function for this specific project in R-Studio. The final R-Script accesses online resources to download the necessary shapefiles, census data, and nighttime lights.

For the sake of this project, the geographical area that was analyzed was metropolitan statistical areas or MSAs. The United States Office of Management and Budget (OMB) defines an MSA as adjacent counties that has an urban core of at least 50,000 people. There are 25 MSAs in Texas (including Texarkana). 82 urban counties make up these MSAs, and account for 88% of the population of Texas (DSHS, 2014).

The R-Script has two primary functions. The first is that it uses Google Application Programming Interface (Google API) to find coordinates for inputted cities (in this case the Texas MSAs) and clip the relevant shapefiles to an area around these cities. It then rescales the DNB raster layer so that each city is on the same color scale. This makes it possible to make visible qualitative comparisons between cities more easily. The code is also capable of graphing the unscaled images of each image, but this was not analyzed. The second primary function is to average the total night time light around each city. This single value can then be graphed against population to examine if there is a correlation. This code can easily be used to analyze any MSA in the United States. Simple adjustments can be made to the code to analyze cities outside of the US to ensure the proper areas (i.e. shapefiles) are being quantified.

III. RESULTS

a. Model Builder

Figures 5 and 6 show the ArcGIS models that were built using ModelBuilder. These will be made publicly available on Hydroshare and the images below are meant to represent their complexity and overall flow (refer to Hydroshare to examine functionality). ModelBuilder was used to expedite the GIS process to complete the HAND Analysis (Figure 5) which becomes an input layer for the Fuzzy Overlay Analysis (Figure 6). With this project only analyzing a specific sub-basin compared to the entire area that might be affected by a storm, by using ModelBuilder, it is now possible to quickly input a new set of input shapefiles and rasters to create new layers. A naming mechanism was also utilized in the HAND analysis so that as the process is repeated new layers are not overwriting previous layers. The repeatability that comes with ModelBuilder allows for future work to build off of this when access is acquired to more efficient and powerful processing equipment.

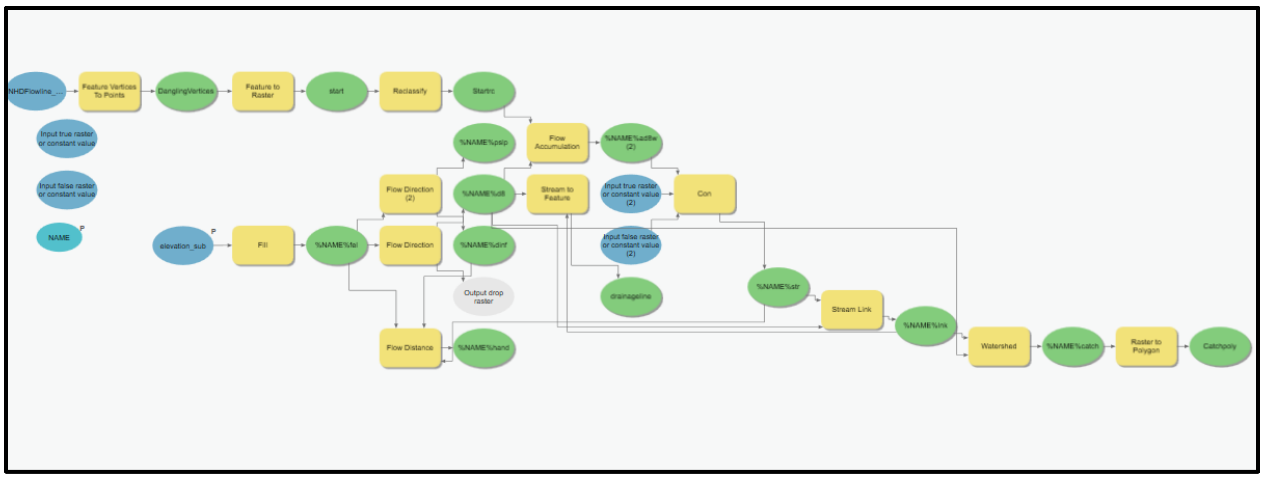


Figure 5: ModelBuilder of HAND Analysis

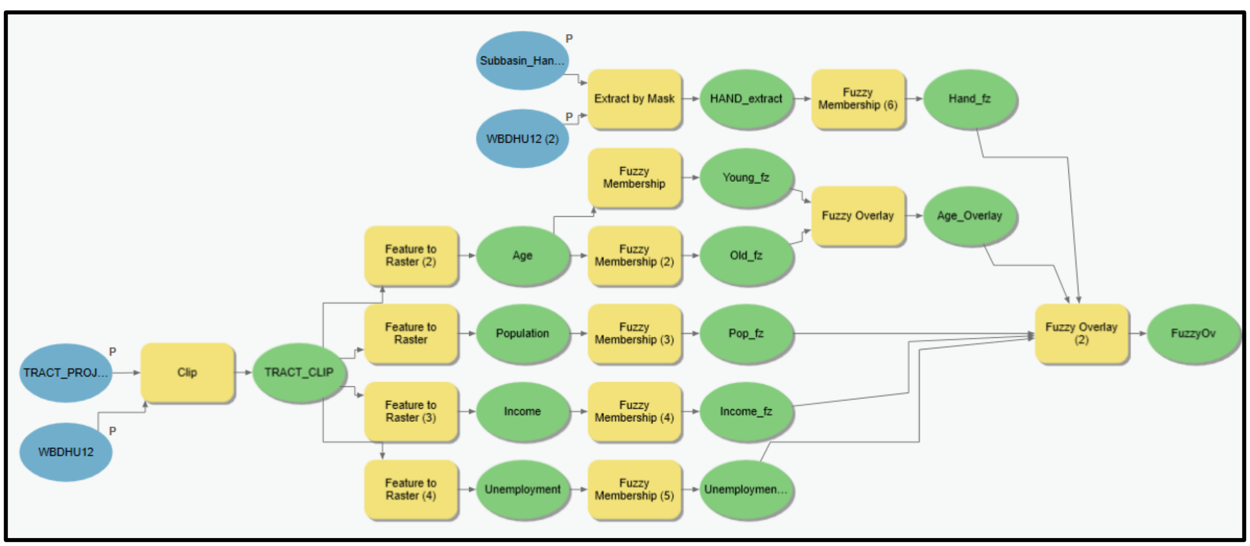


Figure 6: ModelBuilder of Fuzzy Membership and Overlay

b. Social Vulnerability

Figure 7 shows the final Social Vulnerability Map that was created as a result of the ModelBuilder. The census tracts were divided into 5 subsequent categories to rank to compare their vulnerabilities. They were divided so that 20% of all the tracts is in each rank. This was done so that tracts could be compared to each other based on their relative vulnerability. For major storm events such as Hurricane Harvey, it is difficult to compare vulnerability between tracts when everyone is affected. This analysis focuses on potential vulnerability, and which tracts are more likely to struggle when coping with a smaller event. Everyone is vulnerable to some extent when it comes to the threat of flooding, especially in a city like Houston that is predominantly flat and near the coast.

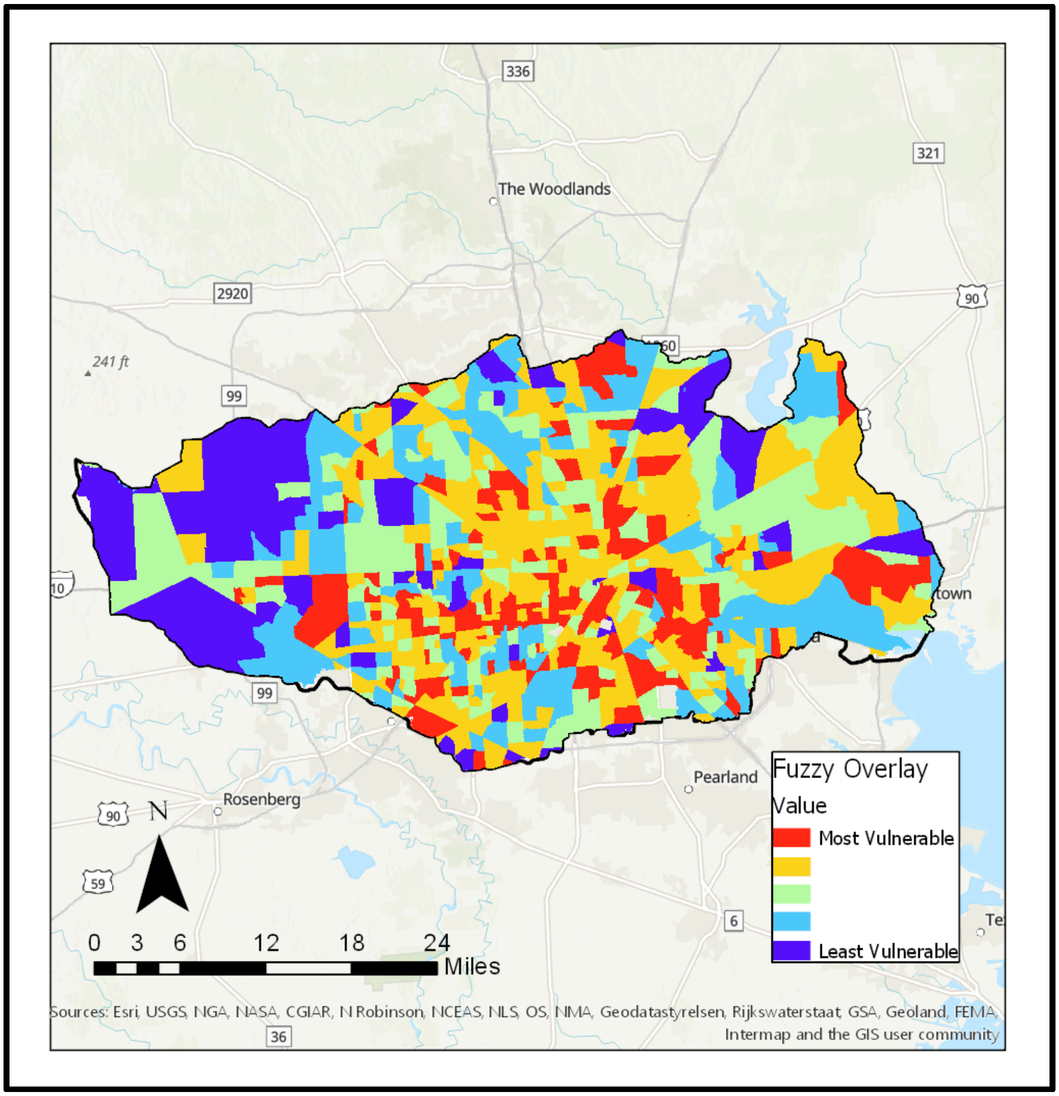


Figure 7: Social Vulnerability to Flooding in the Houston Area

Upon first examination, there appears to be no specific pattern of vulnerability across the sub-basin. There are specific trends and tracts however that represent that is model has potential as a vulnerability indicator. The first indicator is how there are multiple tracts near the western edge of the sub-basin that are purple (least vulnerable). This part of Houston (approaching the Katy, Texas area) is known for being more affluent than other parts of the city. The population in this area tends to have a higher income, is of moderate age, and the majority are gainfully employed. This area is known to be where many families live with parents commuting into the city for their jobs.

To further analyze the accuracy of this model, a tract that was labeled as more vulnerable was selected and compared to other available census data. Tract number 2125 in Harris County has a population density of 892 people per square mile. The median age is 40.6, which is 25% higher than both that of Houston, and Harris county. The Median household income is approximately $30,000, which is just over half of the median income of Harris county. 48% of children (people under the age of 18) live in poverty, with 29% of the entire population below the poverty line (1.5 times higher than Harris County rate). The median value of owner-occupied housing units is half of that for the rest of Harris County. All of these statistics further support that this area would have a higher vulnerability during a natural disaster. While some of these census facts were not specifically used in the Fuzzy Overlay, this analysis shows that the layers that were used are good indicators of vulnerability.

c. R Studio and Population Correlation

Figure 8 shows the scaled DNB image created using the R-Script for the 16 most populous cities in Texas. As mentioned, the images have been resampled so that comparisons can be made between cities. This process is done be collecting random points in the extent of each of the cities being analyzed (20,000 pixels) and the scale bar is adjusted to fit these points. There are many qualitative differences that can be seen between populations. For example, comparing Houston and Austin, there is a noticeable difference in the physical footprint of the city. Based solely on visual inspection, it would appear that Fort Worth is a larger city than Austin, although Austin has 100,000 more people living in the city. These differences in the physical spread of the cities must be considered when correlating night time light with population.

To accurately compare the cities, the total night time light (TNL) values were determined. Total night time light is defined as the sum of the radiance values at each point multiplied by the pixel count (how many times each pixel was measured). This creates a weighted average of the radiance for each of the cities. TNL was determined for each of the 24 MSAs in Texas (Texarkana was not used in this part of the analysis since it is in two different states). The R-Script generated plot is shown in Figure 9. The generated plot is on a logarithmic plot because light intensity increases logarithmically with population. This plot shows a clear linear relationship between log(Population in millions) and log(TNL) (natural log was used).

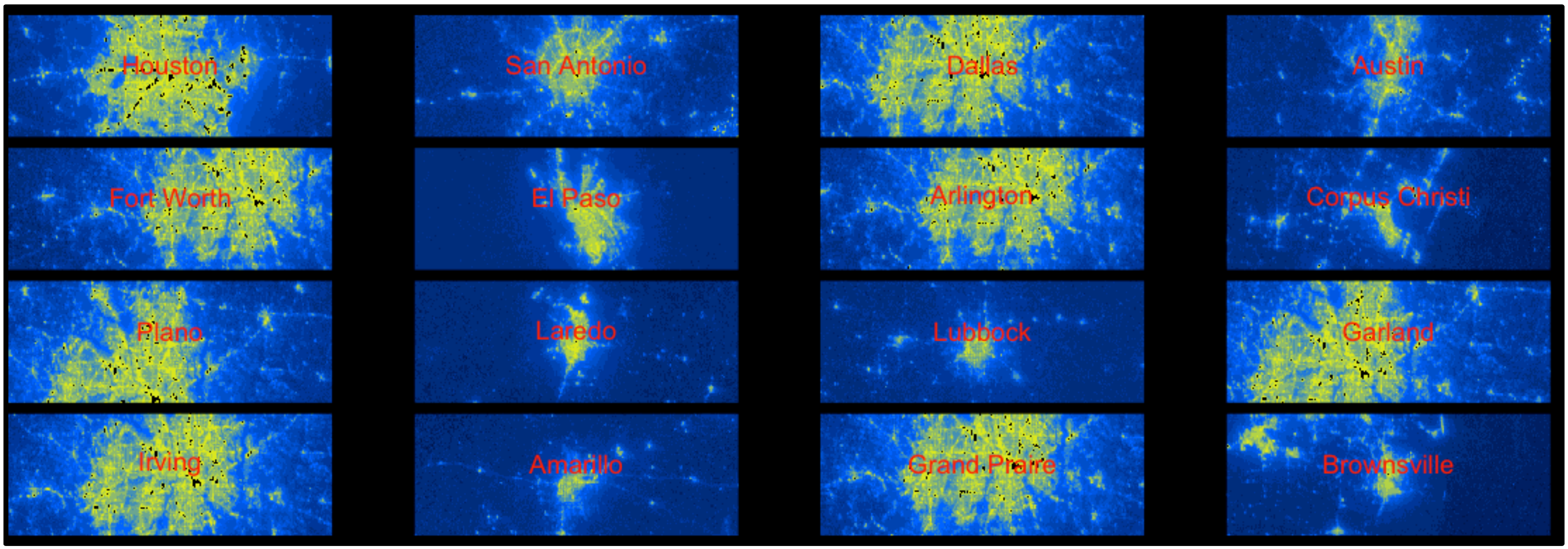


Figure 8: Composite image created in R Studio showing the scaled DNB imagery for the 16 most populous cities in Texas

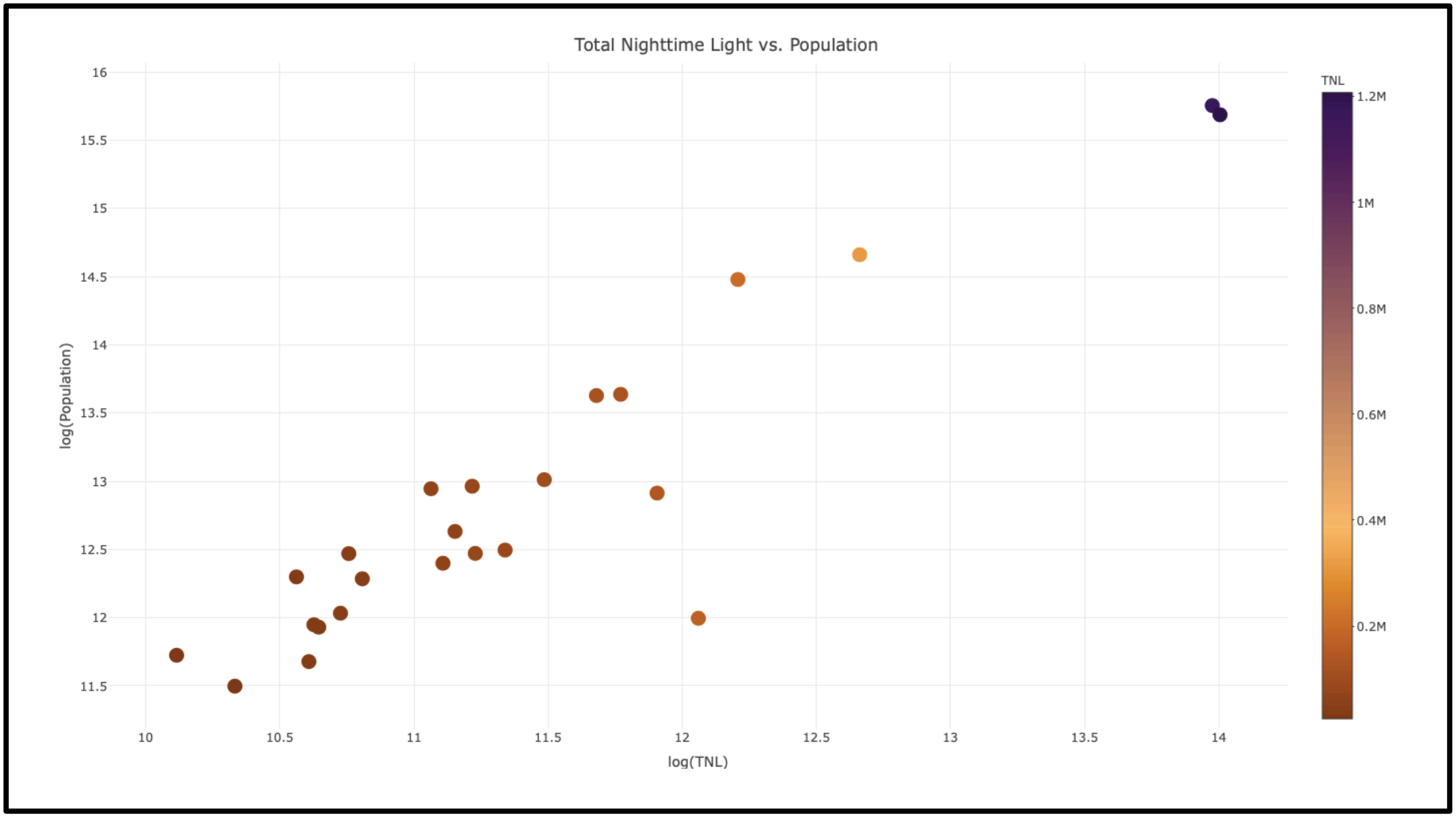


Figure 9: Natural log of Population versus Natural Log of TNL for the 24 Metropolitan Statistical Areas in Texas (excludes Texarkana)

As can be seen in this graph, there is a strong apparent linear relationship. Performing a simple linear regression gives an R^2 value of 0.85. A similar analysis was done using every MSA in the United States (image not shown as it was not a part of this specific analysis) that created a similar R^2 of 0.83. Since the R^2 value is strong for the Texas comparison, it is concluded this relationship between TNL and population stands up for MSAs in Texas McKinney, Texas was identified as the outlier city, which as a noticeably lower population than cities that produce a similar amount of night time light. It is unclear why this city produces as much light as it does. It could be due to expanding industries that produce a lot of night time light, or perhaps because the city is located fairly close to the Dallas-Fort Worth Metroplex (which is only 30 miles south), some of these MSAs might have potential overlapping. This theory is supported with the fact that the Dallas Fort Worth MSA is one of only eleven in the United States that could further be divided into metropolitan divisions because of how the extreme nature or urban sprawl in this area.

IV. CONCLUSIONS AND FUTURE DIRECTION

This project attempted to identify general social vulnerability for an NHD sub-basin that included the Houston metropolitan area. Using Fuzzy Logic, it was identified that different parts of Houston are more vulnerable based on socio-economic factors to flooding, in relationship to the distance they are from a river/stream system. This project also successfully quantified the relationship between total nighttime light and population in the metropolitan statistical areas of Texas.

This project served as a stepping stone as well to further research that would incorporate flood inundation for specific storms. This would in theory create a social vulnerability prediction tool by accessing real time flood data from existing inundation models, such as the National Water Model. Other future components of this research also will incorporate more satellite derived socioeconomic factors to increase the applicability of this model to regions of the world that do not have access to census resources. Another goal of future research is to incorporate an adaptive neural-fuzzy logic inferencing system (ANFIS) into the model. ANFIS networks are a soft computing method commonly used in remote sensing image classifications (Benz et al, 2004). This method will develop statements of the degree to which each cell/pixel will lead to an end result while using a training algorithm to estimate future scenarios (the level of damage after the storm happens), all in a single strategic framework. Predicting at risk communities is critical in shifting how we deal with natural disasters from a response culture, to a preparedness one. This again builds off of the benefits of using fuzzy logic in these kinds of prediction tools.

Future work could also look into analyzing the limitations of satellite derived socio-economic data based on resolution. The spatial resolution of the DNB layers is currently 750 meters. Most analysis that has been done in the past relies on MSA areas, or global population predictions. Examining if the relationship between TNL and Population still exists at different spatial levels (i.e. the county level or the census block level) will shed more light into this area. This will be particularly important for applications in rural areas/countries.

It is also important to note the potential sources of error in this project. The first of these is how each of the fuzzy layers are weighted. For this project, they all carried the same weight when be overlaid on each other. In reality, some of them play a far more critical role than others. For example, population may not be the best indicator of vulnerability. If there are a lot of people in a city in general (such as Houston), a higher density in one area should now play as significant a role in determining potential risk. Other factors that were used in this project might compound on each other to form a weighted bias, which may or may not be beneficial to the model. For example, unemployment and median household income are correlated with each other and using both of these factors might influence the model. However, this could be seen as an advantage or a disadvantage as access to economic resources is often considered a higher indicator of vulnerability than other factors. Future analysis would have to examine the role of each fuzzy layer.

Another potential source of error is during the analysis of the population. Both areas that had a high and a low median age were considered more vulnerable, but when combined these layers might offset each other and, in the end, have no effect on the final outcome. Using a HAND analysis also has its limitations, especially in a city similar to Houston. Houston is known for being a flat city, with only small changes in elevation. HAND would then imply that most areas are equally vulnerable because flood plains would encompass large tracts of land. As previously stated, future research goals hope to address these concerns.

In conclusion, this project successfully built on the concepts we learned in class to incorporate new GIS tools with the specific application to water resources. New models (Fuzzy Logic) and data analysis processes (R Studio) were utilized to meet the goals of this project.

V. RESOURCES

ADPC. Historical Flood Analysis Tool. *SERVIR Mekong.* Server.adpc.net/tools/historical-flood-analysis-tool.

Bender, M., et al. (2010). Modeling Impact of Anthropogenic Warming on the Frequency of Intense Atlantic Hurricanes. *Science.* 327(5964), 454-458.

Benz, U., et al. (2004). Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS-ready information. *Journal of Photogrammetry and Remote Sensing.* 58(3-4), 239-258.

Blumenfeld, J. (2018). Bringing Light to the Night: New VIIRS Nighttime Imagery Available through GIBS. *www.Earthdata.nasa.gov* . Retrieved from: https://earthdata.nasa.gov/viirs-dnb .

Chand, T., Badarinath, K., Elvidge, C., and Tuttle, B. (2009). Spatial characterization of electrical power consumption patterns over India using temporal dmsp-ols night-time satellite data. *International Journal of Remote Sensing*. 30, 647-661.

DSHS. (2014). Texas Metropolitan Status by County. *Texas Department of State Health Services.* Retrieved from: https://www.dshs.texas.gov/chs/info/current-msa.shtm .

Doll, C., Muller, J., and Morley, J. (2006). Mapping regional economic activity from night-time light satellite imagery. *Ecological Economics.* 57, 75-92.

Elvidge, C., Sutton, P., Ghosh, T., Tuttle, B., Baugh, K., Bhaduri, B., and Bright, E. (2009). A global poverty map derived from satellite data. *Computer Geoscience.* 35, 1652-1660.

Patz, J., et al. (2005). Impact of regional climate change on health. *Nature.* 438, 310–317.

Shi, K., Chen, Y., Yu, B., Xu, T., Yang, C., Li, L., Huang, C., Chen, Z., Lui, R., and Wu, J. (2016) Detecting spatiotemporal dyanmics of global electric power consumption using DMSP-OLS stable light data using panel data analysis. *Applied Energy.* 168, 523-533.

Sunderlin, W., Dewi, S., Puntodewo, A., Müller, D., Angelsen, A., and Epprecht, M. (2008). Why Forests are important for global poverty alleviation:a spatial explanation. *Ecology and Society.* 13(2), 1-21.

Sutton, P., Roberts, C., Elvidge, c., and Meij H. (2001). Census from heaven: An estimate of the global human population using night-time imagery. *International Journal of Remote Sensing.* 63, 1303-1313.

Zadeh, L. (1965). Fuzzy Set. *Information and Control.* 8(3), 338-353.

Zhao, X., Yu, B., Liu, Y., Yao, S., Lian, T., Chen, L., Yung, C, Chen, Z., and Wu, J. (2018). NPP-VIIRS DNB Daily Data in Natural Disaster Assessment: Evidence from Selected Case Studies. *Remote Sensing.* 10(1526), 1-25.