Ruoyang Gu

12/13/2021

GEOG498G

Final Project: Highway Segregation in U.S. Cities

In the beginning of 2021, President Biden launched his ambitious and gigantic infrastructure bill, which was recently settled to spend one trillion dollars on U.S. infrastructure. Though the exact scale of this mega spending package is dependent on various political negotiations and compromises. Such large spending unavoidably triggers intensive debates with conflicting interests. Political turmoil aside, one detail in White House’s announcement regarding this project’s goal was to "reconnect neighborhoods cut off by historic investments" by spending "40 percent of the benefits of climate and clean infrastructure investments to disadvantaged communities.". The ‘historic investments’ referred here is the Federal Aid Highway Act of 1956, which was the foundation of American’s interstate system. To make the interstates successfully connects the majority of U.S. cities, city planners and construction directors have to purchase considerable number of lands in urban area. The cost would be unbearable if they chose the shortest or the most optimal routes. Instead, they purchased lands from poorer communities, where people of color were resided. There were also some arguments that interstates and highways were purposed used as barriers to create racial zonings (Archer). Furthermore, original communities were devasted by the new interstates, with hundred and thousands of residents displaced, many once thriving communities were plagued with crimes and drugs. Combined with the practice of discouraging bank loans, mortgages, and insurances in these communities, African American cities were further deteriorated (Kruse).

While the historical memories of these stories were indeed miserable, and there were a lot of accounts to back many claims up, I have yet to see a simple and efficient data analysis on comparing the actual differences between two sides of the interstates. The only statistics listed in new sources were St Paul’s share of black population by census tracts between the I-94, which shows some distinguishing results, but by itself doesn’t tell anything meaningful aside of heavy racial segregation exited. How are other cities segregated by the highways and interstates? At what degree are this segregation significant? Some further research seems necessary. Nevertheless, a topic like this maybe too broad for geographical analysis, as there are too many variables under this roof, such as crime rates, job rates, and many others. Yet find a statistic that is representative for the overall development of the neighborhoods is challenging. Ultimately, I decided to use property value as an index to evaluate the data, as it’s a statistic that can be influenced by various factors, thus making it an ideal parameter for distinguishing the inequal development around the interstates.

In the beginning a very promising method was the spatial regression discontinuity design. Since comparing data before and after the construction of interstates were almost impossible (sadly parcel data back then weren’t available or haven’t been digitalized), and simply comparing data may ignore many other interactions, such as the financial disadvantages for black residents as I mentioned earlier. A more realistic alternative is to use the idea of spatial discontinuity, that is, a cutoff between two regions, whether it be a border, or in our case an interstate. In fact, this spatial method has been frequently used for policy difference across the borders. It shares many similarities with the regression discontinuity design, which is being widely used for many economic and social research that deal with jumps. This method is much more widely used compares to the spatial regression discontinuity designs, and indeed there are much more resource on it. Unfortunately, we could not apply the method in our research as it assumes *geographic homogeneity*. For example, in a case of seeking the relationship between minimal drinking age and the mortality rates young drivers, basic regression discontinuity may be fine if there are sufficient data to leverage other factors with local treatment effect. Therefore, one can use the age of 21 as the cutoff and get some compelling results. But in the case of an air pollution and life expectancy analysis, the geographic factors cannot be ignored. In the case of spatial regression discontinuity design, estimates are localized with high fidelity (after all, it’s really costly to change address). Furthermore, in the case of a geographical border, individuals tends to have more similarities. Nevertheless, while spatial regression discontinuity offers much of the advantage, I was having great difficulties finding the proper way to implement the SRD. The regression discontinuity, on the other hand, can be archived using traditional statsmodels package. Moreover, at that time I was struggling to come up a model, since for my problem there was no definite independent variable in it.

Regardless of my difficulties to find the ways to implement spatial regression designs, finding property value data can be a relatively easy task under the correct state. Yet the biggest problem was the lack of the uniformity between states and counties in term of property parcel data, a topic I would discuss later in this paper. Though Maryland’s data management is probably the best among them all. Unlike many other states, one can easily browse through Maryland’s real property data record quickly to find what he wants. For the raw data, originally, I decided to seek the data from Baltimore city’s Open Baltimore data collection, it soon turned out that their real property information is much more restricted than State of Maryland’s. In the end I used Maryland’s parcel data within the MD iMAP Data Catalog, which contains numerous columns. What makes the dataset worse is despite these hundreds of columns, there are no documentation on details of the datasets, such as what is each column for, date of the data collection, and other miscellaneous information, something NASA and NOAA had done a great job of. For example, as I seek for the total value of properties, I had to cross reference with QGIS and to know that they are stored in the “NFMTTLVL” column. In addition, some extra double check is needed from Maryland Catalog to sure the “NFMTTLVL” is the total value. Due to the inconsistency, some other columns are self-explanatory, such as “New Appraised Land Value” and “New Appraised Improvement Value”. One may get the total property value from the combination of two, but have a standalone column is certainly more convenient. Another problem I encountered during the data processing are the errors. Even though I extracted Baltimore City’s data from dataset using “OWNER CITY == BALTIMORE”, after reexamining there were still some properties lying around in other parts of the state. It turned out there were few mistakes when these properties were inputted. In such case the use of jurisdiction code could be very helpful. I successfully removed them by doing ['JURSCODE'] == 'BACI'. I suspect unlike owner city column, the jurisdiction was automatically assigned rather than inputted, which is why it’s much more reliable than typical address. Lastly, an interstate geographic file was downloaded from the federal website for obvious reasons, making it the border between two supposed clusters.

Even though Maryland’s datasets got its own problem, many other data have been much less organized. One crucial problem I encountered while browsing through them is the absence of distinction between commercial and residential properties. In the case of New Orleans, an analysis was nearly impossible as commercial properties would heavily distract the regionalization results. Therefore, I manually filtered all properties over 2 million values in this analysis, however this attempt did not provide meaningful results. In the end I decided to use the land value to divide the area for the average price per area, which induce to something far more convincing. Some other states such as Georgia’s dataset does not even contain city name and the jurisdiction code. Which yields the problem of how to differentiate Atlanta, a place I would like to analyze from rest of the state. Thankfully, the column “OwnerAddr2” does contain the actual city names in it, however zip codes are within it as well (e.g., ATLANTA GA 30050). Here I used the expression str.replace('\d+', '') to get rid of all the zip codes. Though this method was not the perfect solution, as I later examine some values has a “-“ symbol after the “GA”, possibly due to human input errors. Nevertheless, this expression does provide a satisfactory though not perfect result. One additional issue that needs to be addressed when dealing with cities near water was the some of the shoal islands and water area were categorized into properties. They often have way larger area than any of other geographical features, creating anomalies in the results. In the case of Miami, I had to manually exclude these data from the dataset using QGIS. I was appalled, however, to see the kay components of Salt Lake County’s parcel data are behind a 1200$ paywall for “developers”. Salt Lake City is a city I would really like to do an analysis on, since this is a city that has dominant white populations, compares to Baltimore, A city known as the cluster of African Americans. Through my thorough search all I could find was a free sample that has parcel shapes but without actual data in it. Although I’m not familiar with the regulation in Utah State on data usage but seeing a public institution asking a such large amount of money for a dataset that should be publicly available is always depressing.

The general statistic method I ultimately chose was to using k-means to regionalize areas to see if there’s significant differentiation. It’s a clustering algorithm that partitions the input dataset into k clusters, while each cluster is a representation of adaptive centroid.

Another idea was surfaced during my brainstorming process. Could spatial regression partially explain the topic I’m discovering? After all,