Predictive Modeling for Public Transit Development:
Providing Mass Transportation to Minority Groups During Development Planning

von der Lippe, Michael
msvonderlippe@smcm.edu

Sahdeo, Deonarine
dsahdeo@smcm.edu

DeMay, John
jtdemey@smcm.edu

Kerzner, Alex
ajkerzner@smcm.edu

Badger, Stuart
sbadger@smcm.edu

Stephen, Proctor
ineedyouremail@smcm.edu

St.Mary's College of Maryland
4/28/2017

1 ABSTRACT

Public Transportation is an important major social issue as the movement of people from one place to another allows them access to societal necessities such as proper education, adequate and fulfilling work, as well as many other public services. The allocation of spending on public transportation has shifted towards a focus on suburban movement, with large expansive highways and long railways, which provide little to no advantages for the minority groups within the metropolitan center. This paper justify the problem we will be attempting to solve, provide a predictive model that uses data from three major and distinct cities, in order to estimate the best way to allocate budget during the planning of future urban development projects. The end result of this study will be an application that provides predictions based on our training set, as well as a template method for cleaning up future transportation data in order to strengthen the training set information for more accurate and representative models.

2 JUSTIFICATION

Throughout the twenty-first century, we have seen public transportation play an increasingly important role in developing societies with economic vitality. Mass transit can be broken down into two parts: the movement of people and the movement of goods. On the one hand, multinational corporations like Google and Amazon have studied the transportation of goods, looking for ways to improve that system. On the other hand, not much research has been done into the movement of people despite recent advancements in data science and data analytics. The Civil Rights Project at Harvard performed research to show that "where people live can greatly affect what types of transportation options are available to them to travel to work and to carry out their daily activities." This is supported by the claim that "since 1960, people of color have increasingly populated metropolitan areas. Only 52 of the 100 largest cities have a majority white population, according to the 2000 census data" (Moving to Equity, 7). With this information, we decided that society's need for efficient public transportation may not be fulfilled by our current infrastructure. "People's income levels generally correspond with their ability to own a car and the type of transportation they use" (Moving to Equity, 9). This provoked the thought that efficient public transportation would be able to benefit minorities and may assist in providing equal advantages when attempting to gain access to the working world. As shown in this figure, public transportation use is much higher among minority groups. This means that as we improve public transportation, the impact of those improvements

is much more valuable to the community who uses it. As a team, we began sifting through some of the many available data sets on public transportation. We started analyzing public transportation infrastructure in aspiration of spotting inefficiencies in transportation. We found that there were three monolithic providers of public transport service: taxis, buses, and underground rail systems such as metro and subway. We decided that a taxi service, although public, is for personal consumption, not mass transit. Therefore, the two services we are most interested in are the public bus system and the metro system (subway). We began to ask questions about what could be done to improve the implementation of these two services. The final question that we reached is: depending on the population density, which is more efficient, underground rail or public bus routes? We predict that there will be a trend in our data that will show as population density increases, larger systems of transportation infrastructure, such as metro, will be more cost-effective than smaller systems such as bus routes. Conversely, as population density decreases, we predict that bus routes will be more cost-effective than larger systems of transportation. The goal of our research is to first distinguish the effectiveness between the two systems based on the average annual spending, usage, transportation coverage, and cost for citizens. The second goal of our research is to create a predictive model to properly allocate spending on public transportation infrastructure to minimize annual spending (and in turn the money spent by citizens for public transport), while simultaneously maximizing transportation coverage and hopefully improving public transportation usage. By finding a way to improve public transportation infrastructure, we can provide a better transit system for minority groups which would allow them to achieve higher economic vitality and a more fulfilling social life.

3 SCOPE

To define the scope of our project we conducted research into what constitutes as Public Mass Transportation. Merriam-Webster defines Mass Transit as the following:
"the transportation of large numbers of people by means of buses, subway trains, etc., especially within urban areas; also : the system, vehicles, or facilities engaged in such transportation"
From this definition we gathered that the elements of mass transportation include the following:
1.) How Many People are Being Moved
2.) How These People are Being Moved
3.) How The System is Operated
In order to reduce the scope of this project to a manageable size we did research on these questions in order to properly identify significant factors for each element of mass transportation. The significant factors for the quantity of people moved included the number of people in the metropolitan center, the number of people that are covered by the transportation infrastructure, and the amount of people that use the infrastructure. Significant factors for modes of transportation included Bus and Metro statistics. Consideration was given to taxis however the scope of this project is Mass Transportation, and although taxis are a public service, they provide limited mobility for a large crowd. Other modes of transportation such as faeries and bike shares was given consideration as well, the issue with these modes however is a lack of consistency across other major cities. Significant factors for how this system operated was determined to be budget that pertains to the maintenance and distribution of the service and the infrastructure required for its operation.

4 DEFINING VARIABLES

This section pertains to all relevant factors identified during research of the current transportation infrastructures, and the guidelines we set for proper measurement. Due to a lack of preconceived methodology for analyzing and understanding transportation information, the following definitions are those pertinent to the understanding of our projects goals and measurement specifications.
1.)Transportation Coverage - The total area that a transportation infrastructure operates over a metropolitan center. That is, the effective radius of a transportation stations influence on the movement of people around it.
2.)Effective Ridership - The quantity of people that are transported through the given public transportation.

3.)Goal Ridership - The assumed quantity of people that accounted for that require the use of transportation services using public transportation use statics provided by The Civil Rights Project.
4.)Budget - The money spent on transportation efforts. Inclusive to staff management, construction of infrastructure, and yearly extremities. Exclusive to dept payments, and revenue (Consideration for profits is included during concluding thoughts and suggestions for further research).
a.) Allocation - The way the budget is divided among the two predefined categories of transportation, bus and metro.
5.) Transportation Efficiency - The total coverage of a transportation system and the ratio of effective ridership to goal ridership, in comparison with the budget for the transportation infrastructure. A high transportation efficiency would imply that there is a larger effective coverage area, a significant percentage of effective ridership, and a budget with properly allocated spending to maximize these variables.

5 CURRENT APPROACHES

Little to no work is being conducted in order to remedy the issues that have been presented by this project. Currently urban development is regulated by government policies that take little account for the entire population of an urban complex. Since World War II there has only been a few outstanding policy changes that benefit the minority groups within the United States. "In the 1990s, the primary federal transportation funding law, the Intermodal Surface Transportation Efficiency Act (ISTEA), changed the way funding was allocated and began to erode the long-standing preference for highway funding."(*The Civil Rights Project*, 2). This article goes on to discuss a history of policy changes, and inevitably, all major minority transportation aide laws were subsided in the early 2000s. In London however, a recognition that there was a spatial disparity in access to public transportation for minority groups was recognized and an amendment to policy was made in order to increase the development of bus routes in lower income areas.

5.1 Errors with current approaches

The current issue with transportation development issues is highlighted by *The Civil Rights Project* when discussing who is planning the development of these systems and the factors that those people are concerned about:
"There are significant inequities between bus service, which tends to serve more low-income riders, and rail service, which tends to serve higher-income riders. These inequities pale in comparison to the differences between governmental financial and political support for highway systems and for public transit systems. Many transportation planners and policymakers, concerned primarily with the needs of suburban commuters, have focused on constructing highways and commuter rail lines that do little to serve the needs of minority and low-income communities that depend on public transportation."
This inequities impact are validated by the research into the spending priorities that indicates vast spending on suburban transportation rather than inside the city. One claim that the article from Harvard University made in accordance to this is that "more research examining geographically coded data on spending between cities and other areas would provide a better understanding of how transportation spending patterns impact minority and low-income communities". This geographical is one part of our data collection process that we will discuss in the Data Collections and Methods.

6 DATA COLLECTION

The goal set for data collection was to acquire relevant data on each of the variables that we had defined above for each of the three cities. Findings will be presented by variable and errors or gaps in the data collection process will be noted in the Data Collection Errors subsection. Each variable was gathered from the cities associated database for transportation. For NYC, the Manhattan Transit Authority(MTA) and New York City Transport(NYCT). For DC, the Maryland Transit Authority(MTA). For London, Transport for London(TfL) along with a data store, London Transportation Data Store.

6.1 Data Collected for Transportation Coverage

The data that we gathered to represent the coverage of a given transportation infrastructure was data provided by the infrastructure itself, typically labeled as "Operation Coverage Surface Area". This data proved to be misleading, an explanation will be provided in Data Collection Errors, so we cross checked the supposed coverage by using shape files for the geographical location of each station for the two modes of transportation. We then separated the area within the station location data into a category of high density and the area around the boundaries of these locations. Bus stations that were within the city limits and labeled as high density were assigned a 1km effective coverage radius, that is each station provides a circle of coverage around it with a radius of 1 kilometer, this is in order to account for reasonable accessibility among each station, so that in order to access the service, a patron would have to travel no more than a kilometer without a mode of personal transportation. Metro stations that were within the city limits and labeled as high density were assigned a 2km effective coverage radius, this radius was assigned with similar reasoning to that of High Density Bus Station with a slight distinction that due to metro services being so expansive, and typically used for larger distance, the distance for reasonable accessibility to the service was increased to account for the outcome of using this mode of transportation. Lastly, metro stations that outline the boundaries of transportation coverage area and therefore labeled as low density, were given an effective radius of 30km away from the city center, this value was to account for average commuting distance. Each of these specified classification was provided in order to accurately represent the use of transportation coverage.

6.2 Data Collected for Effective Ridership

The data that we gathered to represent the effective ridership was provided by the associated agencies for each city in the form of monthly reports (US Cities) and quarterly reports (London) as passenger journeys. According to these agencies one passenger journey accounts for one passenger entering and exiting the system. In the US cities this data was presented in a set of data that is labeled Performance Data. For London this data was presented in their yearly reports that take place the first quarter of each year, meaning that data for London is collected on a different schedule than the US data, but as we are provided with dates the data pertained to, shifting the data to match that of the US produced no discontinuities in data.

6.3 Data Collected for Goal Ridership

The data that we gathered to represent the goal ridership is inclusive to census reports on population for each city, and the percentage ownership of a mode of personal transportation presented in *The Civil Rights Project*. This assumed value is also cross checked with a measurement of population density and coverage area in order to ensure accurate measurement. Population density was gathered when analyzing census reports and by taking the product of the population density and effective coverage area, we were able to create an assumed goal ridership that is insignificantly distinct from the goal ridership gathered by taking the total population assumed to not be in ownership of a personal mode of transportation, that is to say the number of people who would rely on public transport for regular mobility needs.

6.4 Data Collected for Budget

The data that we gathered to represent the budget for a transportation infrastructure was provided by the associated agencies for each city. This data was in the form of a yearly spending report for the US cities, and for London this information was provided in the previously mentioned yearly reports. As this is the same document as before with London, shifting the data to match the US produced no discontinuities in data, however the London data for budget had its own challenges that are later discussed in Data Collection Errors.

6.5 Data Collection Errors

As mentioned this subsection covers inconveniences and errors that were found during the data collection process that resulted in changes in

methodology. To begin, the most prevalent error that we encountered during our collection process was data that had little basis. There is a slew of data that is regularly updated and presented by transportation agencies, however this data often lacked meaning or evidence to prove validity. One key offender of this is in regards to transportation coverage. The area that was measured for transportation coverage in the United States cities that we analyzed (NYC and DC) was vague and non-revealing. Bus data had been presented in the area of each bus route. This posed two issues for us. Firstly, that bus routes were being calculated inclusive to overlaps, meaning there was no distinction between two buses operating on the same route and each individual route. Secondly, that metro data was being collected for operational surface coverage was presented in kilometers, simply a distance, and then bus data was collected in kilometers squared, an area. These two issues proved insurmountable so we decided on generating our own method for fairly and accurately measuring coverage. Another difficulty with data collection that we were faced with was that of missing or corrupted data. For example, original attempts to account for ridership within the US cities was turnstile data. This was assumed to be the most accurate and useful data as it was updated most frequently and was a raw count of entrances and exits into and out of the metro infrastructure. However, a large number of these data files were published with missing data due to what appeared to be a glitch in the system, weeks and months of regularly reported data would disappear or sometimes scrambled among different categories, making the data and processing the data impossible. This is what lead to us settling on monthly performance reports to accurately judge the ridership data. The final major concern during data collection was that our London data had a change of scope during our desired analysis period. Meaning that budget and coverage data that was labeled as strictly bus was replaced with data that pertained to "Surface Operations", which was inclusive to taxis and faeries, variables that were excluded from the projects scope. This led us to only including a small subset of data for London to our training set in order to avoid using misleading information.

7 DATA CLEANUP

8 MACHINE LEARNING ALGORITHM

Analysis of our data through the use of Machine Learning took place in two parts. A Classification Algorithm and a Predictive Algorithm. The classification algorithm was used in order to classify points as either being a "high-efficiency spending model" or a "low-efficiency spending model", definitions of these models will be described in the Classification Algorithm subsection. After each point had been classified, the set of "high-efficiency model"s was sent to a Predictive Algorithm that given a Budget and Population, returns a prediction for the optimized spending model for the features of population and budget.

8.1 Classification Algorithm

The classification algorithm made use of three different classification algorithms as well as a voting-classification algorithm that took the weighted probabilities produced by each algorithm to determine the most accurate set of classification. The following list consists of the features provided to each of the classification algorithms along with the justification for that feature. 1.Total Ridership - In order to account for the total number of people being moved by both transportation systems. 2. Bus Budget - In order to account for the amount of money that the bus system was provided. 3. Metro Budget - In order to account for the amount of money that the metro system was provided. As discussed previously, we set out to classify the each entry in our data set by the efficiency of their spending model. The term spending model refers to the ratio of the total budget that is delegated to metro systems, and the ratio of total budget that is delegated to bus systems. A high-efficiency spending model would be a spending model that provides above average movement of people at below average cost. In contrast a low-efficiency spending model would be a model that provides lower than average ridership at above average costs. The following list consists of the the three algorithms that our voting-classification algorithm implemented to find probability, the reason for using that classi-

fier, specifications of certain parameters, as well as an explanation for their assigned weight. 1. Decision Tree Classifier - This classifier was used due to the low feature count of our data in comparison with our higher data point count. We assigned the Decision Tree Classifier a maximum depth of 2 in order to focus on consistency of models. The weight that this classifier was assigned was the lowest of the three as its goal was to provide a mixture of consistency grouping and defining efficiency boundaries. 2. K-Nearest Neighbors Classifier - This classifier was again used due to our low feature count with less of a focus on finding efficiency but instead grouping consistent efficiency models. The neighbors parameter for this algorithm was decidedly six in order to fixate on the most consistent models. This classifier was given the highest weight as it was implemented with the goal of reducing the probability of using 'fluke' models and requiring some consistency between models in order for them to be labeled as efficient. 3. Gaussian Naive Bayes Classifier - This classifier was used in order to properly determine what was efficient and what was inefficient. This algorithm was assigned the middle weight in the voting algorithm as it was required to determine efficiency boundaries along the hyper-plane.

8.2 Predictive Algorithm

9 CONCLUSION

9.1 Further Research Opportunities