UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

IDENTIFYING LATENT DIVERSITY, EQUITY, INCLUSION, AND ACCESSIBILITY (DEIA) INDICATORS FOR MULTIMODAL TRANSPORTATION SYSTEMS.

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE

By

FARIHA NAZNEEN RISTA

Norman, Oklahoma

2023

IDENTIFYING LATENT DIVERSITY, EQUITY, INCLUSION, AND ACCESSIBILITY (DEIA) INDICATORS FOR MULTIMODAL TRANSPORTATION SYSTEMS.

A THESIS APPROVED FOR THE
SCHOOL OF CIVIL ENGINEERING AND ENVIRONMENTAL SCIENCE

BY THE COMMITTEE CONSISTING OF

Dr. Arif Mohaimin Sadri, Chair

Dr. Musharraf Zaman

Dr. Amy Cerato

Dr. Dominique Pittenger

Acknowledgements

I want to express my sincere gratitude to my advisor, Dr. Arif Mohaimin Sadri who has provided me with a great opportunity to pursue this research idea and fostered my academic growth and passion. His insightful feedback significantly enhanced my research skills and professional development. I am also indebted to my esteemed committee members, Dr. Musharraf Zaman, Dr. Amy Cerato, and Dr. Dominique Pittenger, for dedicating their valuable time and expertise to review and guiding my research. Their guidance led me to explore my subject matter from diverse perspectives and delve deeper into its complexities.

I would like to extend my thanks to Dr. Keith Strevett and Dr. Sherri Irvin for their unwavering support and the time they generously invested in aiding me with the successful defense of my thesis. Special recognition goes to my dedicated research colleague and PhD student at the University of Oklahoma, Mr. Khondhaker Al Momin, whose efforts in Twitter data collection (4.1.1) and processing (4.1.2) were invaluable to the project's success. Additionally, I would like to express my sincere gratitude to Ms. Ashley Herndon for her tremendous support throughout my graduate program.

Finally, I want to convey my deepest love and gratitude to my family for their unwavering support and encouragement throughout this endeavor. Their constant belief in me has been a driving force in my pursuit of academic excellence.

Table of Contents

List of Figures

List of Tables

Abstract

Traditional transportation policies unfairly affect marginalized travelers and under-represented groups, limiting their access to social and economic opportunities and contributing to residential segregation. Even though there is growing acknowledgement among policymakers and planners about the need for fair and inclusive transport systems covering all modes, the empirical literature remains inconclusive and lacks sufficient intellectual tools, data sources, and standards to incorporate a system-level perspective and assess the diversity, equity, inclusion, and accessibility (DEIA) indicators for multi-modal transportation systems. Existing approaches rely on time-consuming, labor-intensive, expensive surveys that lack real-time capabilities. In contrast, this research utilizes alternative data sources such as large-scale, open-source social media and street network data to identify latent DEIA indicators for transportation systems using network science theories and advanced data-driven methods.

First, road network data from sources like OpenStreetMap or Google Maps offer a promising avenue to measure the accessibility of social opportunities for marginalized populations, such as bicycle users and transit dependents. This research aims to establish indicators of bike accessibility by utilizing open-source street network data from OpenStreetMap and extracting the bicycle network of 40 cities in the United States. Various macro network parameters (e.g., density, diameter, average path length, circuity, average degree) were calculated for the cities, along with demographic parameters obtained from the American Community Survey 2020 data (e.g., population size, per capita income, percentage of bike users). Statistical regression analysis revealed a significant relationship between accessibility score and certain network and demographic parameters. The regression model can assist planners in identifying the accessibility of the bike network in any given area using network data. The study also presents a systematic intervention method that utilizes the betweenness centrality measure to increase the accessibility of an existing network, with lower centrality nodes found to be more critical for interventions aimed at improving accessibility.

Next, social media data presents a cost-effective and real-time alternative for capturing public opinion on transportation issues, which can serve as an indicator of a transportation system's DEIA. The study leveraged approximately three months' worth of Twitter data (around 1.46 million tweets) from the state of New York to identify key transportation-related DEIA issues discussed by users. Natural language processing techniques were employed to extract transportation DEIA-relevant conversations, followed by the use of a Bidirectional Encoder Representations from Transformers (BERT) model for tweet classification. Socio-demographic information of users was detected using Random Forest machine learning algorithm. Major topics discussed by the users in the sample dataset were public transportation infrastructure, active transportation, ridesharing, accessibility, etc. Finally, a logistic regression model was developed to understand the relationship between users' demographic data and DEIA concerns. This model helps identify specific transportation DEIA issues raised by different marginalized groups, providing valuable insights for urban planners.

In conclusion, this research utilizes an innovative dataset and sophisticated data analysis techniques to introduce a unique method for assessing the diversity, equity, inclusion, and accessibility (DEIA) of transportation systems. This approach yields crucial insights for planners, pinpointing the specific locations and demographic segments most impacted by existing transportation inequities. Furthermore, the study presents a distinct strategy for systematically enhancing network accessibility. Collectively, these findings represent substantial advancements in our comprehension of, and ability to address, DEIA issues within transportation networks.

Chapter 1

INTRODUCTION

Transportation diversity means designing and implementing transportation infrastructure and services that can cater to the needs of individuals from various backgrounds, income levels, and abilities [1]. Equity in transportation means providing fair access to transportation infrastructure and services for all individuals, regardless of their background or circumstances. This can include providing equal access to public transportation in all neighborhoods and providing affordable transportation options [2-4]. Inclusion in transportation means creating an environment where all individuals feel welcome and valued. This can involve designing transportation infrastructure and services that are accessible to people with disabilities such as providing wheelchair ramps and lifts on buses and trains, installing audio and visual announcements on public transportation, and providing accessible parking spaces [5, 6]. Accessibility in transportation means ease of reaching social and economic opportunities such as education, employment, healthcare [7]. Usually, transit or active transportation dependents suffer from lack of accessibility to many opportunities because of traditional car-based transportation infrastructure in the USA.

The transportation system has a vital role in affording people a variety of choices for reaching their destinations, and it profoundly impacts their overall quality of life [8]. Due to an ever-increasing influx of immigration, the United States (U.S.) is becoming more racially and ethnically diverse, needing a paradigm shift in the transportation planning system to solve DEIA of its citizens from all socioeconomic backgrounds [9]. Traditionally, transportation organizations in the U.S. have primarily prioritized safety and mobility as their key areas of focus. Unfortunately, the consideration of DEIA factors has often been neglected in decision-making processes. This oversight has resulted in the perpetuation of socioeconomic imbalances within the transportation system.

Figure 1*: Graphical Representation of residential segregation due to car-based transportation infrastructure development.*

Figure 1 above demonstrates inequal distribution of resources and opportunities that stem from disparities in transportation access between rich and vulnerable communities. The “rich” community is defined by the people who have the ability to own cars and afford housing in areas of well-connected transportation infrastructure. As such, the rich community has better access to various opportunities such as education, employment, healthcare, stations, recreation, grocery stores etc.

On the contrary, the term "vulnerable" community refers to a group of individuals who face financial constraints that prevent them from residing in areas with improved infrastructure or owning a vehicle, individuals with physical limitations that restrict their driving ability, or those who choose not to drive. Thus, they have limited access to all the social and economic opportunities, resulting in poor living conditions, low economic growth, high unemployment rates, social isolation, and long-term social inequalities [7, 9, 10].

From the figure above, the constituent elements of a diverse, equitable, inclusive, and accessible transportation network can be identified. These elements encompass various infrastructure and services, including bike routes, transit routes, medical facilities, grocery and shopping centers, educational and financial institutions, recreational venues, eateries, places of work, fitness centers, high-speed internet connections, and accessible parking spaces. Any locality that lacks these amenities may be classified as a poorly linked area.

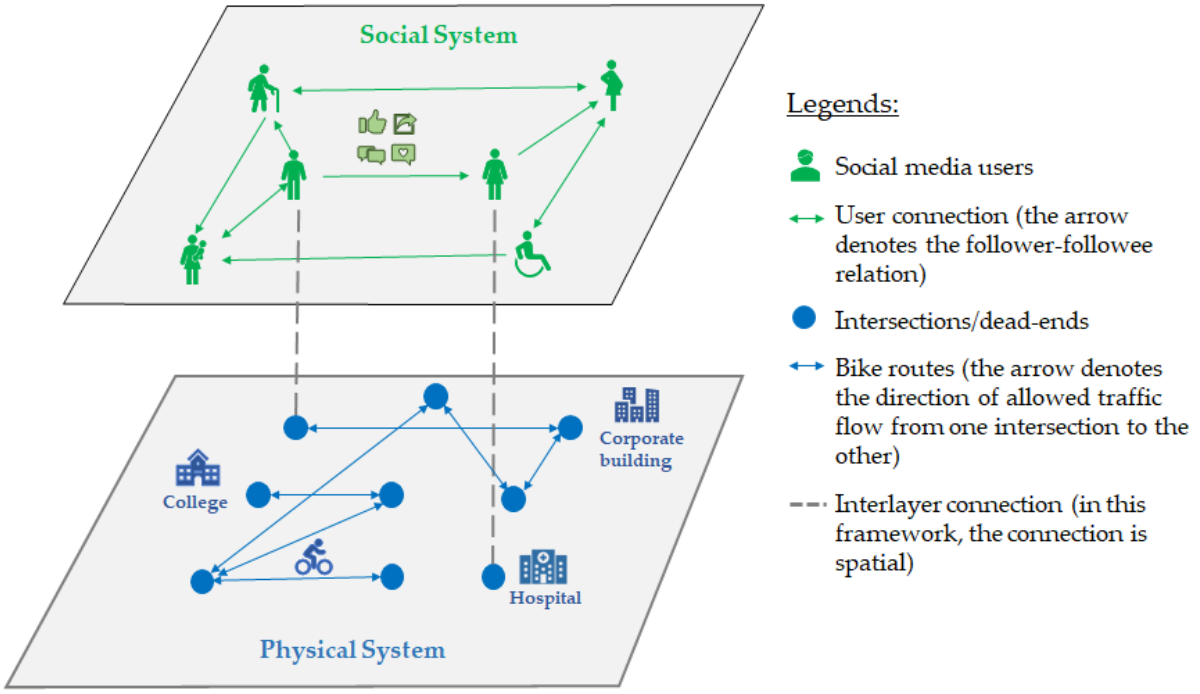

***Figure 2**: An interconnected network formed by two layers, with interlayer links connecting different elements. The physical system layer comprises of the bike road network and social system layer comprises of the social media user network.*

The transportations system can be classified in two layers as shown in **Figure 2**, the physical system layer and the social system layer. In this thesis, for ease of analysis, the author will only consider bike network as the physical system network. Bicycles are classified as active transportation since they rely on human power for propulsion. For individuals who cannot afford to purchase a car, biking represents a practical alternative mode of transportation. Nevertheless, as illustrated **Figure 2**, there are numerous destinations that are not easily accessible via bicycle due to the inadequacy of the infrastructure. This may necessitate bikers

to traverse longer distances on foot or take a longer, indirect route to reach their destination, unlike drivers who have the advantage of a more direct path.

The social system layer in this framework consists of the social media users who live in the area of study. Public opinion expressed on social media platforms represents a valuable resource for analyzing DEIA issues within the study area.

In this framework, the social system layer encompasses the social media users residing within the study area. A significant proportion of these social media (SM) users are individuals who utilize alternate modes of transportation, and they actively express their opinions on SM platforms regarding transportation DEIA. As a result, SM platforms hold considerable potential as a valuable alternative to traditional survey methods for gathering public opinion pertaining to transportation system's DEIA.

Improving opportunities for public transit and active transportation can contribute to a more equitable and healthier transportation for vulnerable populations. Multiple studies have indicated that as alternative modes of travel improve, there is a noticeable shift from car usage to these alternative transportation options [11-14]. Therefore, implementation of a multimodal transportation system can address the diverse needs of both the marginalized community and the rich community.

Numerous scholarly studies have extensively addressed the imperative of establishing an equitable transportation system to cater to the diverse population within the United States [1-15]. Some studies have also proposed user-centric approaches to quantifying and assessing equity at the local level [16, 17]. These studies provide evidence to the policymakers to facilitate the implementation of public and active transportation infrastructures. As such, the US government and policymakers have undertaken several notable policies and initiatives to establish DEIA in transportation. Noteworthy among these are the Executive Order 13985 (January 2021) [18], and the Justice40 policy (January 2021), which aims to direct federal investments towards disadvantaged communities [19]. Furthermore, notable funding programs, such as the Reconnecting Communities Pilot (RCP) program (2021) and the Rebuilding American Infrastructure with Sustainability & Equity (RAISE) program (2022), have been established to support the development of multimodal transportation infrastructure [20, 21]. These initiatives have already initiated numerous projects focused on achieving DEIA objectives.

To this point, an essential question persists regarding standards for quantifying and assessing DEIA in order to effectively identify communities that require utmost attention. In light of this research gap, the primary objective of this study is to introduce a novel measure for estimating DEIA within transportation systems. This measure will leverage a unique and innovative data set, providing valuable insights and contributing to the advancement of DEIA goals in transportation planning and policymaking. To achieve this goal, the study will address the following specific research questions for the physical system layer.

RQ1: Is there any network property that can indicate the level of accessibility of a bike network?

RQ2: Is there any socio-demographic property of travelers strongly influenced by network inaccessibility?

RQ3: Is there a method to identify crucial nodes that needs more attention to increase accessibility of a network?

RQ4: Is there a systematic way of network intervention using graph theory that can make the network more accessible?

For social system layer, the author will address the following research questions.

RQ5: To what extent can social media interactions serve as a viable alternative to traditional surveys in capturing public opinion regarding transportation?

RQ6: Can social media opinions be utilized to identify and characterize vulnerable communities in relation to transportation issues?

The research proposes the following hypotheses:

H1: The accessibility of a road network increases as it becomes more direct.

H2: Increased accessibility within a bicycle network leads to an increased interest of biking among the residents living in the network.

H3: Network components occupying less central positions are more crucial for network intervention aiming at achieving higher accessibility for marginalized travelers (e.g., bikers).

H4: Social media users discuss issues related to transportation DEIA.

H5: Social media data can effectively identify and map the locations of vulnerable communities in terms of transportation accessibility.

Overall, the study utilized novel datasets and cutting-edge data analysis techniques to introduce a unique method for assessing the diversity, equity, inclusion, and accessibility (DEIA) of transportation systems. This approach yields crucial insights for city planners, helping them find the communities most affected by inequitable transportation and a novel strategy for enhancing network accessibility. These findings contribute to significant progress in the understanding of and capacity to tackle DEIA concerns within transportation networks.

Chapter 2

This chapter covers relevant literature on transportation DEIA as well as some cutting-edge analytical techniques that have been applied to transportation research and that will be employed for the analysis of this study. The chapter is divided into four major sections, starting with the methodologies used to measure the DEIA of transportation in the literature, followed by the use of network data and network science in existing transportation research, the use of twitter data in existing transportation literature, and finally the use of machine learning and natural language processing algorithms in twitter data analysis.

2.1 REVIEW OF LITERATURE ON TRANSPORTATION DEIA

In recent years, transportation researchers, including the U.S. DOT, state Department of Transportation, American Society of Civil Engineers, Transportation Research Board, National Academies, and the private sector entities, have shown a growing emphasis on addressing social equity issues in transportation [8, 10, 22-31]. The goal is to alleviate the adverse impacts experienced by travelers from underrepresented communities. While some researchers have underscored the importance of transportation DEIA in reducing residential segregation, there has been limited discussion on how to assess and improve it. Litman provided a summary of transportation equity concepts and methodologies to measure it [15]. Litman suggests that factors like the quality of available transportation alternatives, average trip distances, and trip costs can serve as potential indicators of accessibility, and he proposes conducting a public survey to quantify these indicators [17]. The research team at the University of South Florida's Center for Urban Transportation Research (CUTR) introduced a two-fold strategy for incorporating equity into traditional planning processes. They developed an equity audit tool to identify the transportation needs of the community from the equity perspective, alongside an equity scorecard tool that rates projects based on the outcomes derived from the needs assessment generated by the equity audit tool [32]. However, these tools rely on data collected through public surveys, which can be a time-consuming and costly process. Moreover, by the time the survey is completed, the opinions of the respondents may have evolved or shifted. To overcome these limitations, the incorporation of real-time and cost-effective data sources, such as road network data and social media data, provides a viable solution. This alternative approach mitigates the shortcomings associated with traditional survey-based methods and offers a timelier and dynamic tool for effectively identifying and assessing transportation DEIA.

2.2 NETWORK DATA IN TRANSPORTATION RESEARCH

Street networks aid as the primary structure for urban transportation systems, influencing how people and vehicles move and enhancing the dynamics of urban life [33]. Street network data has been employed in numerous research applications, including the analysis of travel patterns [34, 35], optimization of transit routes [36-38], calculation of the shortest routes and estimated travel times [39-41]. However, no existing research utilizes this data to assess and improve DEIA of transportation systems.

Traditional sources of street network data include municipal and state repositories, expensive commercial datasets, and in the US, the census bureau's TIGER/Line roads shapefiles [42, 43]. An alternative data source is OpenStreetMap, an online collaborative mapping project covering the entire world [44]. OpenStreetMap contains a vast amount of geospatial objects and descriptive tags, including streets, trails, building footprints, land parcels, rivers, power lines, points of interest, and more [41, 45, 46].

Researchers typically obtain street network data from OpenStreetMap using three main

approaches. The first approach involves using the Overpass API to query geospatial features, although the query language can be challenging to use directly [44, 47]. The second approach is to utilize commercial services that download data extracts for specific areas or bounding boxes and provide them to users. However, these services can be expensive, slow, and lack customization, making them less suitable for acquiring data in multiple precisely bounded study sites.

The third method involves using OSMnx, a free and open-source Python package specifically designed for downloading and analyzing street networks from OpenStreetMap [33, 48, 49]. OSMnx is a tool that retrieves street network data from OpenStreetMap. It offers different query options like bounding boxes, addresses, polygons, or place names. You can download networks for driving, walking, or biking, and analyze them for properties like shortest paths, centrality, clustering, and geometric measures. OSMnx corrects the topology, retaining accurate geometry and length of street segments. In short, OSMnx simplifies access to street network data and enables comprehensive analysis. The OSM street data is regarded as reliable for many cities, although there is room for improvements on micro-level details. Nonetheless, the OSM data quality is adequate for large-scale analysis [50].

2.3 REVIEW OF NETWORK SCIENCE LITERATURE

Network science, also known as graph theory, is a widely recognized mathematical tool that has proven very useful in the study and analysis of the features and dynamics of street networks in a multitude of contexts and practical applications [11, 33]. Graph theory provides several metrics that can be employed to evaluate the performance of a network. These metrics include density, average degree, average circuity, centrality, diameter, radius, average path length, and average centrality. Assessing these parameters allows researchers to gain insights into different aspects of complex networks. Consequently, this method has gained significant popularity among scientists, particularly transportation planners, who extensively utilize it to analyze various aspects of road networks.

Researchers utilize network science to model traffic flow as a dynamic process on a network, allowing them to analyze congestion patterns, identify bottlenecks, and propose effective strategies for traffic management and congestion mitigation [51-54]. Graph theory has also been extensively used to evaluate the resilience and robustness of road networks when faced with disruptions such as accidents, disasters, or infrastructure failures [55-58]. By identifying critical nodes or links, researchers can develop strategies to enhance network resilience and ensure efficient recovery from disruptions. Moreover, network science contributes to understanding the intricate relationship between road networks and land use patterns. It enables the integration of various transportation modes, including roads, public transportation, and pedestrian or cycling paths. By considering the interconnections and interactions between these modes, researchers can optimize multi-modal transport systems, enhance connectivity, and promote sustainable transportation options [59-61]. Lastly, by integrating network science with geographic information systems (GIS), researchers perform spatial analyses, visualize road networks, and overlay additional geographic data. This integration enables a deeper understanding of the spatial context and facilitates informed decision-making in urban planning, transportation management, and emergency response [52, 62, 63].

The author wants to make use of this effective tool for network accessibility analysis and methodology development to increase accessibility for vulnerable users of active transportation which has not been explored yet in the past literature.

2.4 TWITTER DATA IN TRANSPORTATION RESEARCH

Several studies have emphasized the importance of leveraging data from social media platforms that capture user behaviors and interactions. Robust empirical research has shown that big data from social media can yield significant insights into public opinion patterns regarding ongoing societal issues [64]. For instance, Twitter data has been employed by numerous researchers to investigate different areas such as service characteristics [65, 66], retweeting activity [67, 68], text classification and event detection [69-73], situational awareness [74, 75], online communication among emergency responders [76, 77], human mobility [78, 79], developing sensor techniques for early awareness [80], and disaster relief efforts [81]. In recent times, transportation researchers have made extensive use of social media data sources, such as Twitter, to study human mobility patterns [82], modeling of activity-pattern [83-86], origin-destination demand estimation [87-91], social influence in activity patterns [92], transit service characteristics [93], travel survey methods [94, 95], and crisis informatics [96], among other research areas.

2.5 MACHINE LEARNING AND NATURAL LANGUAGE PROCESSING IN TWITTER DATA ANALYSIS

Machine learning has found extensive applications in research across various fields. Researchers utilize machine learning algorithms and techniques to analyze complex data, discover patterns, make predictions, and gain insights. It has been applied in areas such as healthcare [97, 98], climate modeling [99, 100], finance [101-103], image recognition [104-107], and recommendation systems, enabling advancements and discoveries that were previously challenging or impossible to achieve with traditional statistical methods. In the field of transportation research, it used to predict travel demand [108-110], optimize routing and scheduling, analyze traffic patterns [111], and develop intelligent transportation systems [112-114]. Thus, it aids in improved decision-making, enhanced efficiency, and the development of innovative transportation solutions.

Natural Language Processing (NLP), is a field of artificial intelligence that focuses on enabling computers to understand, interpret, and generate human language. Natural Language Processing (NLP) is extensively used in research to analyze and extract valuable information from textual data [115-117]. It enables researchers to automate tasks such as sentiment analysis [118, 119], topic modeling [120, 121], information extraction, and language generation, contributing to a deeper understanding of text-based datasets and facilitating more efficient data-driven research.

ML and NLP algorithms are widely used in Twitter data analysis research to extract valuable insights from large volumes of tweets. Researchers employ these algorithms to perform sentiment analysis, topic modeling, user classification, and other tasks [122-127]. By leveraging such techniques, they can gather deeper understanding of public opinion, social dynamics, and real-time events [128].

This study aims at gathering large-scale Twitter data from the state of New York and analyze the tweets related to transportation systems, as discussed by local residents on the platform to identify the transportation DEIA concerns and the demographic relation with such concerns.

Chapter 3

In this chapter, bike network data from forty cities across the US was extracted and subjected to in-depth analysis using network science and statistical regression techniques. The aim was to identify key network parameters that contribute to bike network accessibility. Building upon this understanding, a systematic intervention method was developed to enhance the accessibility of the network. The findings of this study offer a robust tool for planners, enabling them to identify areas where people are suffering from inadequate accessibility, locate crucial road segments requiring network intervention, and pre-assess the effectiveness of interventions on specific road segments. This research presents valuable insights that can empower planners to make informed decisions and facilitate targeted improvements in bike network accessibility.

3.1 DATA ANALYSIS METHODS

Graph theory is a branch of mathematics that deals with the study of graphs. A graph (G) refers to a finite set of nodes or vertices $V = v_1, v_2, \ldots\ldots, v_n$ (where n is the number of nodes) and a finite set of edges (or connections) E. Nodes represent entities or elements, while edges represent connections or relationships between those entities.

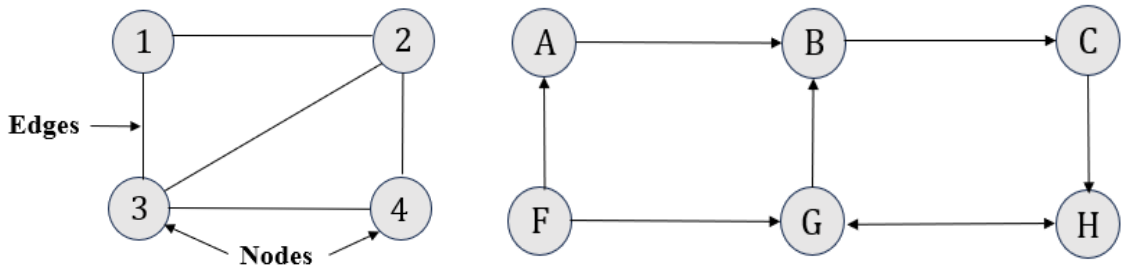

***Figure 3**: An undirected graph (left) and a directed graph (right).*

Figure 3 shows examples of directed and undirected graphs. In an undirected graph, the edges do not have a specific direction associated with them. The relationship between nodes is bidirectional or symmetric. In an undirected graph, such as a social network, if there is an edge between node 1 and node 2, it implies a mutual relationship. For example, if nodes 1 and 2 represent two people and the edge represents friendship, an undirected edge signifies that person 1 considers person 2 as a friend, and vice versa. This symmetrical representation indicates that both individuals have a reciprocal perception of friendship towards each other. In a directed graph, also known as a digraph, the edges have a specific direction or flow associated with them. This means that the relationship between nodes is one-way or asymmetric. The directed edge from node A to node B in **Figure 3** indicates that node A considers node B as a friend, but node B does not consider node A as a friend.

Graph theory provides a framework for analyzing the relationships, connectivity, and properties of these networks. Thus, graph theory has been very helpful to analyze complex networks in the field of computer science communication network [129-134], transportation and urban planning [52, 135, 136], social network analysis[137-139], biology and bioinformatics [140-144], operation research [145-147], physics and chemistry [148-152] and many more. It encompasses understanding the structure, properties, and behavior of graphs, as well as developing algorithms and techniques for solving graph-related problems.

Street networks can be effectively represented and analyzed as graphs in urban planning and transportation studies. Nodes (V) in the graph correspond to intersections or dead ends, representing points where streets intersect and the points where roads end. The edges (E) of the graph are the street segments or road links connecting those intersections and dead ends. An

example is shown in **Figure 4**.

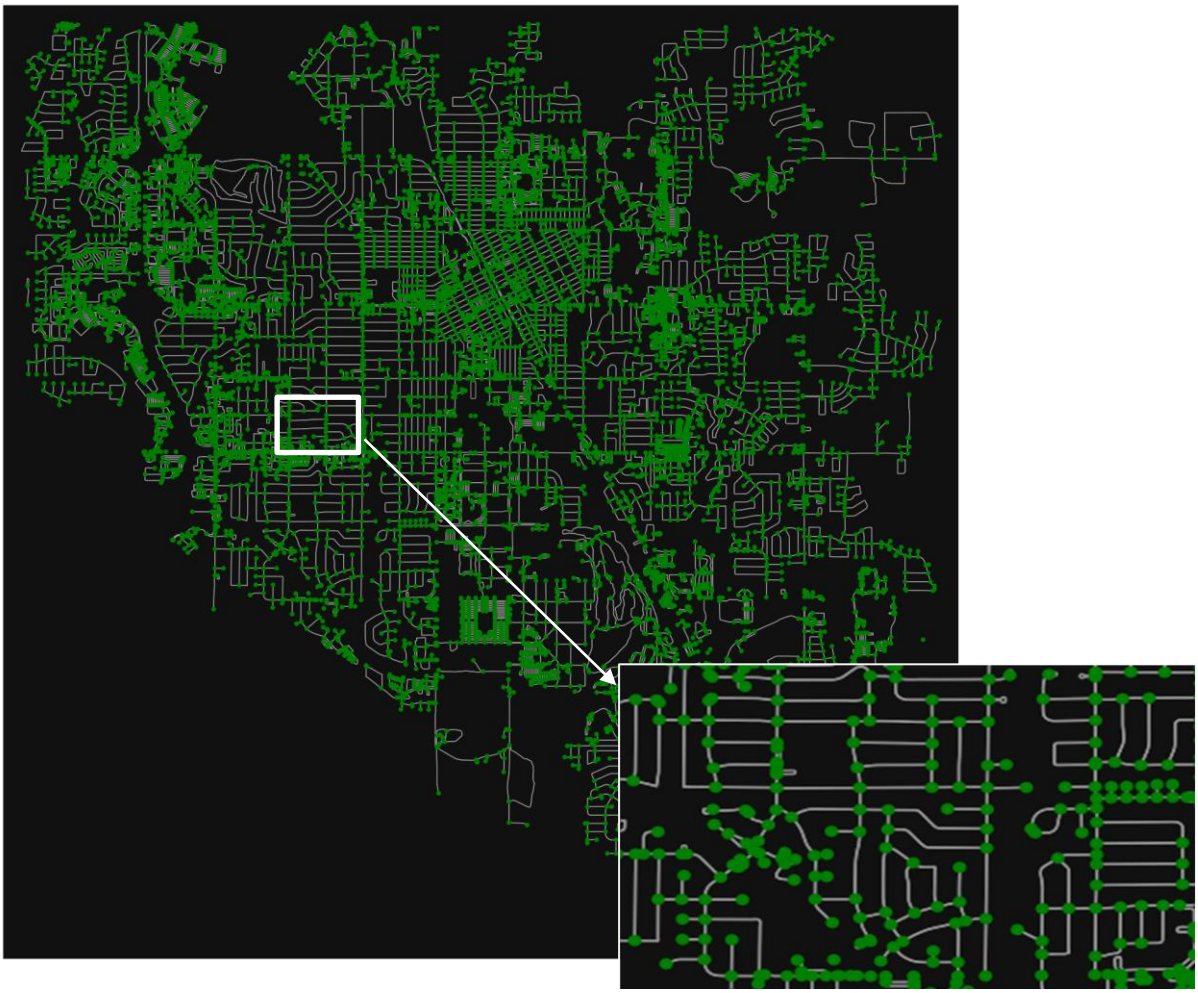

Figure 4: *Graphical representation of road network as a set of nodes and edges*

Figure 5: *Finding shortest route using Dijkstra's algorithm and graph theory.*

Various graph-based measures can be employed to analyze the network structure and characteristics. Additionally, graph algorithms enable the exploration of efficient routes, identification of shortest paths, and evaluation of network connectivity. Algorithms like Dijkstra's algorithm or A* search algorithm can be applied to find optimal routes (**Figure 5**) or estimate travel times within the street network [153-156]. Overall, graph theory provides a powerful tool for studying the spatial organization and functionality of urban transportation systems, aiding in the improvement of urban planning practices.

The characterization of networks in the field of transportation geography and network science involves various methods, as reviewed in [157-159]. In this paper, specific measures have been chosen to analyze and describe networks, and they are discussed below.

3.1.1 Graph Theory and Network Characterization

3.1.1.1 Average Degree

Average degree is a measure used in network analysis to quantify the average number of connections or links that each node has in a network. It provides insights into the overall connectivity and complexity of the network. The mathematical formula to calculate the average degree in a street network is:

$$Average\ Degree\ = \frac{Number\ of\ Edges}{Number\ of\ Nodes} \quad \textbf{(1)}$$

The average degree of a network can have a significant impact on network accessibility. A higher average degree implies a greater number of connections or links between nodes, indicating a more interconnected transportation network. This increased connectivity can enhance accessibility by providing multiple routes or paths for travelers to reach their destinations. It facilitates easier movement and reduces the chances of congestion or bottlenecks in the network. Improving the average degree of a network through infrastructure development, such as constructing new roads or adding public transportation routes, can contribute to enhancing transportation accessibility by providing better connectivity and more efficient travel options for users.

3.1.1.2 Diameter

The diameter of a network refers to the longest shortest path between any two nodes in that network. In other words, it represents the maximum distance between any two points in the network. The diameter is a measure of the network's overall size or extent.

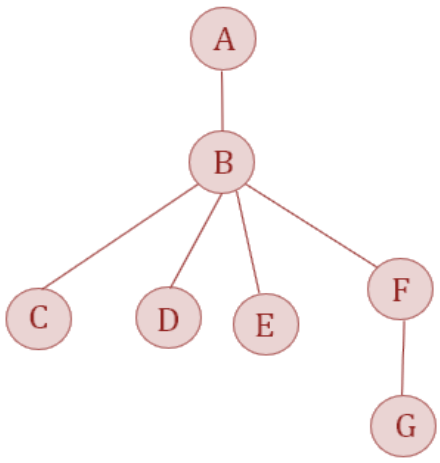

***Figure 6**: A social network example.*

Diameter of the Network in **Figure 6** is 3 (A-B-F-G, E-B-F-G, D-B-F-G, C-B-F-G). The impact of the diameter on transportation accessibility can be significant. The impact of the diameter on transportation accessibility can be significant. A smaller diameter implies shorter travel distances and potentially faster travel times between different locations within the network. This can improve overall accessibility by reducing travel costs, increasing connectivity, and facilitating efficient movement of people, goods, and services. On the other hand, a larger diameter indicates longer distances and potential travel delays, which can negatively impact transportation accessibility. Longer travel distances may discourage commuting, limit connectivity between areas, and hinder accessibility to essential facilities

3.1.1.3 Circuity

Circuity has been extensively used by the researchers in understanding the network structure, connectivity, and urban development [33, 160-167]. Circuity refers to the ratio between the actual path distance traveled by a vehicle and the straight-line or Euclidean distance between the origin and destination points. For an unweighted network,

$$Average\ Circuity = \frac{\sum network\ distances\ between\ all\ origin - destination\ pairs}{\sum Euclidean\ distances\ between\ all\ origin - destination\ pairs}$$

$$C_u = \frac{\sum D_N}{\sum D_E} \quad \textbf{(2)}$$

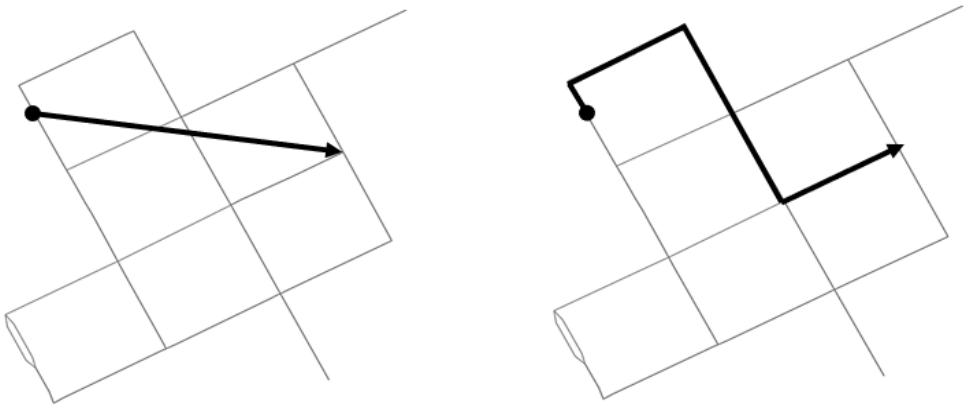

Figure 7: Euclidean distance (left) and network distance (right) in street network.

Thus, the circuity value can never be less than '1'. Higher circuity values indicate a more circuitous route, meaning that the actual distance traveled is greater than the straight-line distance. This can result in increased travel times, inefficiencies in transportation systems, and reduced accessibility between locations. Therefore, circuity considerations are important in designing public transportation systems, pedestrian and cyclist-friendly routes, and efficient routing algorithms for various transportation modes.

3.1.1.4 Density

Density refers to the degree of connectivity or interconnectedness within a network. It quantifies the links present in the network in relation to the total number of possible connections. The mathematical formula for density in a transportation network is:

$$Density = \frac{Number\ of\ Edges}{Maximum\ Possible\ Edges} \quad \textbf{(3)}$$

The graph on left of **Figure 8** had a lower density than the graph on the right. The impact of density on transportation accessibility is significant. A higher network density implies better accessibility within the transportation system. It allows for more direct routes and multiple options for traveling between different locations. This can lead to shorter travel distances, reduced travel times, and improved overall accessibility for individuals using the transportation network.

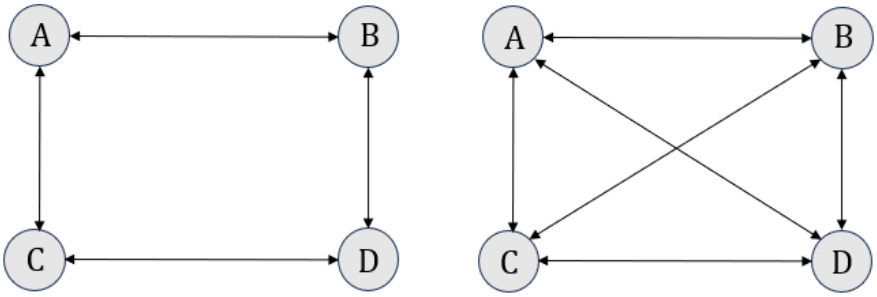

Figure 8: *The existing edges (left) vs the possible edges for the network (right)*

3.1.1.5 Average path length

Average path length is a network measure that quantifies the average distance between pairs of nodes in a network. Mathematical formula to calculate average path length is:

$$l_a = \frac{1}{n(n-1)} \sum d_{ij} \quad \textbf{(4)}$$

Here,

n = number of nodes

d_{ij} = shortest path distance between i and j

It provides insight into the overall efficiency and accessibility of transportation networks. A longer average path length implies that individuals have access to a wider range of destinations within the network. For example, in a transportation network with a larger average path length, there may be multiple routes or connections available to reach different destinations offering more options for travelers to access various facilities, services, and opportunities.

3.1.1.6 Betweenness Centrality

Betweenness centrality is a network centrality measure that quantifies the importance of a node within a network based on its position as a bridge or intermediary between other nodes. It measures the extent to which a node lies on the shortest paths between pairs of other nodes in the network. Nodes with high betweenness centrality have a significant influence on the flow of information, resources, or movement within the network. For transportation network analysis, betweenness centrality measure is considered the most important centrality measure by the researchers. Betweenness centrality of a node i is calculated using the following equation:

$$C_i = \frac{1}{n^2} \sum_{st} \frac{n_{st}^i}{g_{st}} \quad \textbf{(5)}$$

Here, n_{st}^i is 1 if node i is in the shortest path between two nodes s and t in a network, 0 otherwise. g_{st} the total number of shortest paths between nodes s and t and 'n' is the total number of nodes in the network. **Figure 9** displays nodes and links with maximum centrality

of a street network.

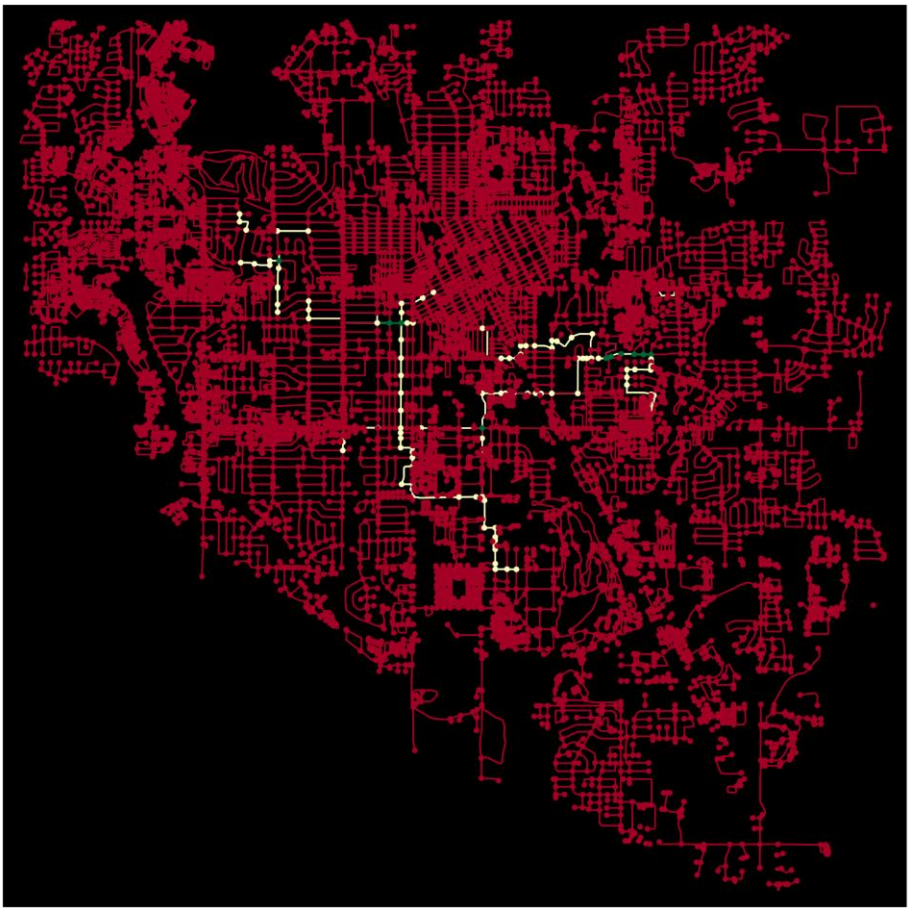

Figure 9*: A street network showing the most central links and nodes (green representing top 5 central links and nodes, yellow representing next most central nodes and links).*

In transportation networks, the literature finds that nodes with high connectivity tend to have a high centrality [168-170]. For example, nodes that are major intersections or hubs, connecting multiple roads or routes, often have high connectivity and are considered to be central nodes. Detecting such nodes is important in urban planning as the disruption of such nodes will have a substantial impact on the overall connectivity and efficiency of the transportation network. There has been substantial research effort to identify critical links and nodes of road network in disaster and risk mitigation [171, 172], improving resilience [173-175], etc. The concept is based on focusing on the nodes or links that aids to increased mobility of the network. However, the concept may not be useful in increasing accessibility of the network.

3.1.1.7 Accessibility

Transportation accessibility refers to the ease with which individuals can reach desired destinations or engage in activities. In the car-based transportation infrastructure of the U.S., often the bicyclists or transit dependents suffer from lack of accessibility as they often have to change mode of transportation or travel a much longer distance due to the circuitous nature of the network to reach a destination. Such inaccessibility results from lack of investment and planning in public and active transportation networks by the government agencies. Especially the poor and disabled suffer the most from such inaccessibility as they do not have the affordability of ability to buy or drive a car. While the recent attention to enhance bicycle network accessibility is appreciable, the method to measure and improve accessibility is still

unestablished. This study focuses on establishing a method to measure bicycle accessibility and improve it using graph theory.

For this purpose, bicycle network data of forty cities was collected. The network properties described in section 3.1.1.1 through 3.1.1.6 were determined. Additionally, relevant demographic data of these cities was collected. The bicycle accessibility score, as outlined in the methodology of section 3.1.2.2, was calculated. Statistical regression analysis was conducted to establish a relationship between the accessibility score and both network and demographic parameters. By deriving insights from the regression equation, a network intervention approach is proposed in section 3.1.4, aiming to improve bicycle accessibility based on the identified factors.

3.1.2 Data Collection

3.1.2.1 Bike Network Data

Bike networks of forty cities (listed in **Table 1**) across the United States were collected from OpenStreetMap using the Python software package OSMnx [48, 176, 177]**.** OSMnx provides a convenient way to extract and manipulate street network data, creating graph objects compatible with the NetworkX package in Python. One example of the extracted bike networks has been shown in **Figure 10**. Subsequently, the obtained networks were analyzed using the OSMnx package to assess several network parameters. These parameters included the number of nodes, number of edges, circuity, average degree, average path length, density, diameter, betweenness centrality, among others.

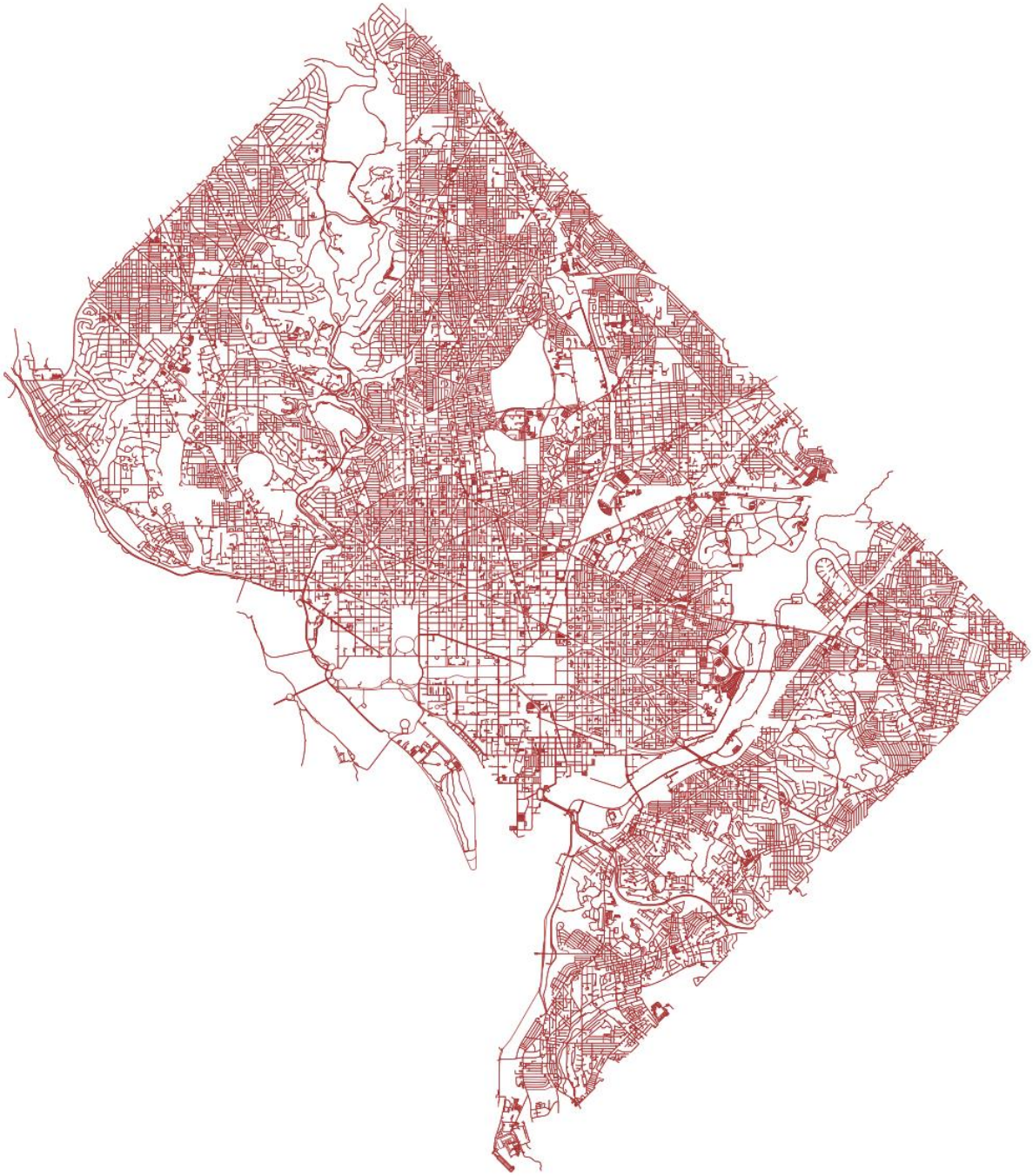

Figure 10: *Bicycle network of Washington city, D.C. (source: OpenStreetMap)*

3.1.2.2 Bike Accessibility Score Data

Accessibility was measured in each study area through the aggregation of two approaches: the traffic stress analysis method [178-182] and the assessment of access to various opportunities within a 30-minute cycling distance [183, 184].

Traffic stress analysis is a methodology used to assess the level of stress or perceived safety experienced by different road users, particularly pedestrians and bicyclists, in relation to the surrounding traffic conditions. It aims to evaluate the potential discomfort, fear, or inconvenience caused by vehicular traffic on non-motorized modes of transportation. The factors considered to determine traffic stress include vehicle speed, traffic volume, road design, presence of dedicated cycling or pedestrian facilities, and interactions with motorized vehicles. Once the traffic stress has been established for all street segments, accessibility score was determined based on reachability to different opportunities. The scoring system ranges from 0 to 100 and takes into account the number of low-stress opportunities available as well as the ratio of low-stress to total destinations within biking distance. The opportunities were divided into six areas, including access to people, employment and educational opportunities, healthcare services, retail establishments, transit alternatives, and recreational attractions. Weights are assigned to represent the relative importance of each destination type within the category. Higher scores were assigned to the initial low-stress destinations through a stepped scale. Beyond those initial destinations, points are prorated up to a maximum of 100 based on the ratio of low-stress to high-stress connections. If a destination type is not reachable by either high- or low-stress means, it is excluded from the calculations for the corresponding city. **Figure 11** displays the bicycle accessibility scores assigned across the study areas.

Table 1: *Study Areas Ranked by Accessibility of Bicycle Network.*

Rank	Place	Rank	Place
1	Minneapolis, MN	21	Miami, FL
2	New York City, NY	22	Oklahoma City, OK
3	Washington City, D.C.	23	Montgomery, AL
4	Denver, CO	24	New Orleans, LA
5	Detroit, MI	25	Newark, NJ
6	Anchorage, AK	26	Tampa, FL
7	Oakland, CA	27	Baton Rouge, LA
8	Tucson, AZ	28	Chesapeake, VA
9	Aurora, CO	29	Honolulu, HI
10	Baltimore, MD	30	Boise City, ID
11	San Diego, CA	31	Jacksonville, FL
12	Phoenix, AZ	32	Indianapolis, IN
13	Boston, MA	33	Fort Wayne, IN
14	Kansas City, MO	34	Dallas, TX
15	Charlotte, NC	35	Fort Worth, TX
16	Jersey City, NJ	36	Greensboro, NC
17	Atlanta, GA	37	Arlington, TX
18	Tulsa, OK	38	Chicago, IL
19	Richmond, VA	39	Wichita, KS
20	Birmingham, AL	40	Buffalo, NY

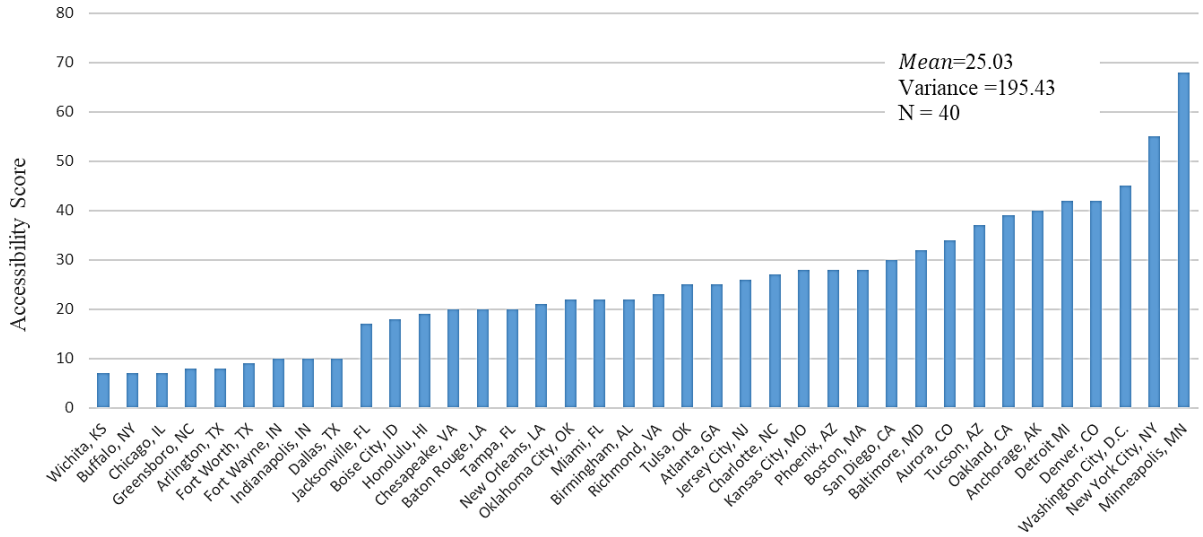

Figure 11: *Bar chart of bicycle accessibility score over study areas.*

3.1.2.3 Socio-Demographic Data

Socio-demographic data for forty study areas was collected from the American Community Survey (ACS) 5-year survey published in the year 2020. The specific variables utilized in this research included population size, per capita income and percentage of population who commutes by bike or motorcycle to work.

3.1.3 Statistical Regression Analysis

Transportation Researchers have extensively employed statistical modeling approaches to predict various aspects of travel behavior [185-190] , network structure [166, 191-194], urban features [195-198] and many more. However, the accessibility of bicycle modes has not yet been adequately modeled in existing literature. By utilizing statistical modeling techniques to analyze bike mode accessibility, considering network structure parameters and demographic information, valuable insights can be gained and a new avenue for measuring transportation accessibility can be explored. This information can be instrumental for city planners and government investors in prioritizing areas requiring improvements. Although several statistical modeling approaches are available, this study will focus on utilizing multiple linear regression (MLR) to predict the accessibility score of bicycle networks in a given area.

3.1.3.1 Multiple Linear Regression with Logarithmic Transformation

In many engineering challenges, it's essential to understand the relations between different variables. Regression analysis is a key statistical technique that researchers frequently use to address this problem. Regression models offer a means to capture and interpret the complex dynamics within a system by fitting mathematical models to empirical data.

Linear regression models utilize linear predictor functions to model the data and estimate output parameters based on the input variables. Multiple linear regression (MLR) is a statistical approach utilized for predicting the outcome of a variable by considering the values of two or more independent variables. It serves as an extension of linear regression. In this technique, the variable that we aim to predict is termed the dependent variable, while the variables used to estimate the value of the dependent variable are referred to as independent or explanatory variables. By analyzing the relationship between these independent variables and the dependent

variable, multiple linear regression provides a framework for making predictions and understanding the influence of various factors on the outcome variable.

The general form of a multiple linear regression model is expressed in equation (**6)**

$$y_i = b_0 + \sum_{j=1}^{n} b_j x_{ij} \quad \textbf{(6)}$$

where y_i represents the model's dependent or predicted variable, x_{ij} refers to the independent input variables, and b_1, $b_{2, \ldots.}$ b_n represent regression coefficients representing the change in y relative to a one-unit change in $x_{i1}, x_{i2}, \ldots\ldots x_{in}$. b_0 is the y-intercept, i.e., the value of y when all x_{ij} values are 0.

These coefficients are determined by fitting the MLR model using ordinary least squares (OLS) regression which aims to minimize the differences between the model's predicted outcome of the dependent variable and the actual values of the dependent variable on the training dataset. This optimization process minimizes the sum of squared vertical deviations between each data point and the regression equation. When a data point lies precisely on the fitted line, the vertical deviation is zero.

The histogram displaying the distribution of accessibility scores in the dataset reveals a right-skewed pattern, as illustrated in **Figure 12**. To address this skewness and achieve a more symmetrical distribution, a logarithmic transformation was applied, resulting in a spread-out histogram that approaches symmetry, as depicted in **Figure 13**. Consequently, multiple linear regression with logarithmic transformation was selected as the preferred modeling approach for our dataset. This choice aligns with the objective of fulfilling the assumption of constant variance in the context of linear modeling.

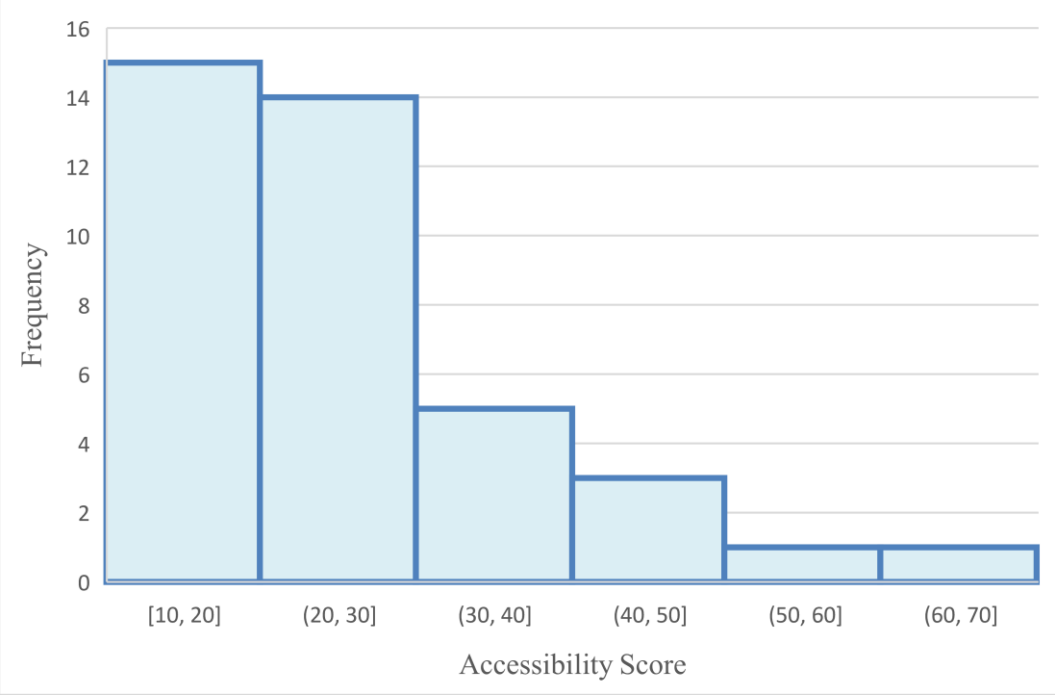

***Figure 12**: Frequency distribution of accessibility scores.*

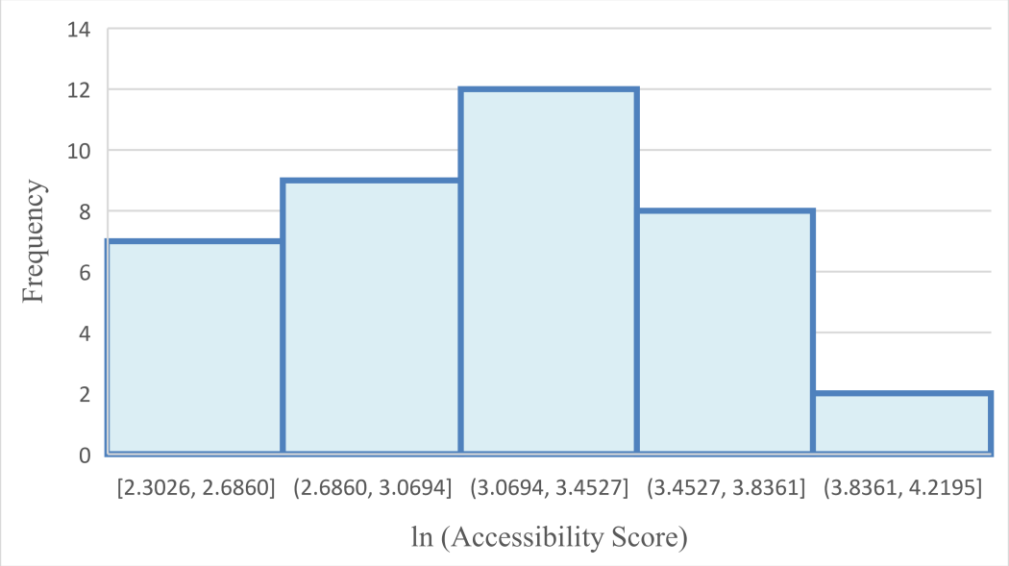

***Figure 13**: Frequency distribution of ln(accessibility scores).*

Therefore, in this study, a multiple linear regression modeling with log-transformed variables was constructed to investigate the relationship between the accessibility score and various network and demographic characteristics.

Logarithmic transformations are one of the most used methods in regression analysis among researchers [199]. It can be done in many ways:

Level- log regression (only the independent variable is transformed):

$$y = b_o + b_1 \ln x_1 + b_2 \ln x_2 + \cdots + b_n \ln x_n \quad \textbf{(7)}$$

Log-level regression (only the dependent variable is transformed):

$$\ln y = b_o + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n \quad \textbf{(8)}$$

Log-log regression (both independent and dependent variable is transformed):

$$\ln y = b_o + b_1 \ln x_1 + b_2 \ln x_2 + \cdots + b_n \ln x_n \quad \textbf{(9)}$$

Besides, the right side of the equations can include a combination of logarithmic terms and non-logarithmic terms. In this specific study, log-log regression will be utilized as the preferred modeling approach where all independent and dependent variables will be log-transformed.

Taking the exponential on both sides, the equation **(9)** simplified as follows:

$$\ln y = \ln e^{b_0} + \ln {x_1}^{b_1} + \ln {x_2}^{b_2} + \cdots + \ln {x_n}^{b_n}$$

$$\ln y = \ln(e^{b_0} {x_1}^{b_1} {x_2}^{b_2} \ldots\ldots {x_n}^{b_n})$$

$$y = e^{b_0} {x_1}^{b_1} {x_2}^{b_2} \ldots\ldots {x_n}^{b_n}$$

e^{b_0} can be expressed as constant c and the equation can be re-expressed as:

$$y = c\, {x_1}^{b_1} {x_2}^{b_2} \ldots\ldots {x_n}^{b_n} \quad \textbf{(10)}$$

Explanation:

If all other explanatory variables of the model in equation (***10***) $x_{i2}, \ldots\ldots x_{in}$ remain unchanged, one-percent change in x_1 would change the y as shown in the following equation:

$$y_{new} = c\ (1.01\ x_1)^{b_1} {x_2}^{b_2} \ldots\ldots {x_n}^{b_n}$$

$$y_{new} = 1.01^{b_1}\ \times \left(c\ {x_1}^{b_1} {x_2}^{b_2} \ldots\ldots {x_n}^{b_n}\right) \quad \textbf{(11)}$$

Subtracting (***10***) from (***11***) ,

$$y_{new} - y = 1.01^{b_1}\ \times \left(c\ {x_1}^{b_1} {x_2}^{b_2} \ldots\ldots {x_n}^{b_n}\right) -\ c\ {x_1}^{b_1} {x_2}^{b_2} \ldots\ldots {x_n}^{b_n}$$

$$y_{new} - y = 1.01^{b_1}\ \times y -\ y$$

$$y_{new} - y = (1.01^{b_1} - 1)\ y$$

$$\frac{y_{new} - y}{y} = 1.01^{b_1} - 1 \quad \textbf{(12)}$$

Thus, 1% change in an independent variable is associated with $(1.01^{b_1} - 1) \times 100$ percent change in dependent variable given that all other variables remain unchanged.

Similarly, x% change in an independent variable is associated with a $\left(\left(1 + \frac{x}{100}\right)^{b_1} - 1\right) \times 100$ percent change in dependent variable given that all other variables remain unchanged.

3.1.3.2 Assumptions of MLR

The MLR relies on several key assumptions to ensure the validity and reliability of the results. The common assumptions associated with multiple linear regression are listed in **Table 2**.

Table 2: *Assumptions of Multiple Linear Regression Modeling*

Topic	**Assumptions**
Independence	The observations used in the regression analysis should be independent of each other. There should be no autocorrelation or systematic patterns in the residuals. The commonly used test to assess independence is the Durbin-Watson test statistic. A Durbin-Watson test statistic value between 1.5 and 2.5 indicates no significant autocorrelation [200-203].
Homoscedasticity	The assumption of homoscedasticity in a regression model states that the residuals are drawn from a population with a constant variance. To assess this assumption, one can plot the standardized residuals against the predicted values. If the spread of the residuals appears to be relatively constant, with no discernible trend or funnel shape, it suggests the presence of homoscedasticity. On the other hand, if the spread of the residuals varies noticeably as the predicted values change, indicating a systematic pattern, it may indicate the violation of the homoscedasticity assumption.

Multicollinearity	The independent variables should be minimally correlated with each other. High correlations between independent variables can lead to instability and unreliable estimates of the regression coefficients. The best method to test for the assumption is the Variance Inflation Factor method. As a rule of thumb, a VIF score below "3" is considered good. As VIF increases, the regression results become lesser reliable.
Multivariate normality (MVN)	Multivariate normality refers to the assumption that the residuals in a regression model follow a normal distribution. Plotting a histogram of the residuals and overlaying a normal curve can visually evaluate how well the residuals align with a normal distribution. It can also be tested using a Normal Probability Plot. This plot involves ordering the residuals and comparing them to theoretical quantiles derived from a standard normal distribution. If the points on the plot follow a straight line, it suggests that the residuals conform to a normal distribution.

3.1.4 Network Intervention

Existing literature has extensively studied the importance of betweenness centrality in improving network resilience to disaster and crisis, sustainability of urban mobility [171-175]. Researchers suggest prioritizing the most central links and nodes, determined through the betweenness centrality measure, while investing in road network improvements. This will strengthen the network's ability to withstand disruptions and ensure the smooth flow of traffic and mobility during crisis. However, such measures may not necessarily enhance network accessibility.

The accessibility of a network is rather affected by the absence of links which leaves the network broken, disconnected or incomplete. Since the study focuses on bicycle accessibility, the author will explain accessibility from a cyclist's perspective. For instance, if there is a highway between an origin and destination, a bicyclist would need to either switch transportation modes or take a longer and circuitous route to reach the destination. The node where the highway begins naturally will have fewer connections than others making it a less central node and bicyclists will avoid such nodes. If the planners do not detect such nodes and create alternative connections, the network cannot be made accessible for the bicyclists. Therefore, instead of prioritizing central nodes, investments aimed at enhancing network accessibility should focus on nodes that are less connected or have lower centrality based on the betweenness centrality measure. This study puts forth a hypothesis suggesting that when planning new bicycle routes or adding bicycle lanes, creating connections with less central nodes will result in a more accessible network compared to creating connections with the most central nodes. The author aims to contribute to establishing a novel method of systematic intervention in the network that aims at enhancing network accessibility.

For that purpose, a small bicycle network covering a 1000-square-meter area in Norman, Oklahoma (**Figure 14**) was extracted from OpenStreetMap using the Python software package OSMnx (circuity = 1.02345, no. of edges = 780, no. of nodes = 276). The network was analyzed using the OSMnx package to assess betweenness centrality of all nodes within the network. **Figure 15** illustrates the identification of the most and least central nodes through the utilization of the betweenness centrality measure.

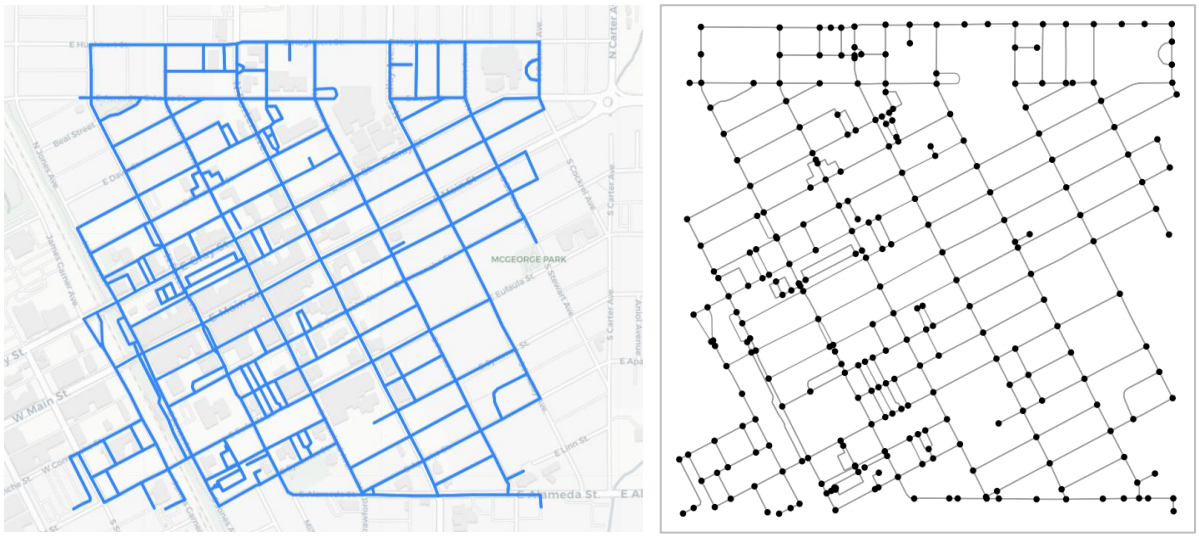

Figure 14: Bicycle network from the Norman, Oklahoma (left) & graphical representation of the network as set of nodes and edges (right)

Based on the centrality values obtained, the two most central nodes and two least central nodes were selected for intervention. Additionally, two random nodes were selected for intervention purposes. A new bike lane (bidirected) was added to each of the selected nodes using OSMnx package. The impact of these interventions on the network has been reported and investigated is section 0.The findings provide valuable insights into the importance of targeted interventions and the potential improvements that can be achieved by systematically enhancing connectivity at critical nodes.

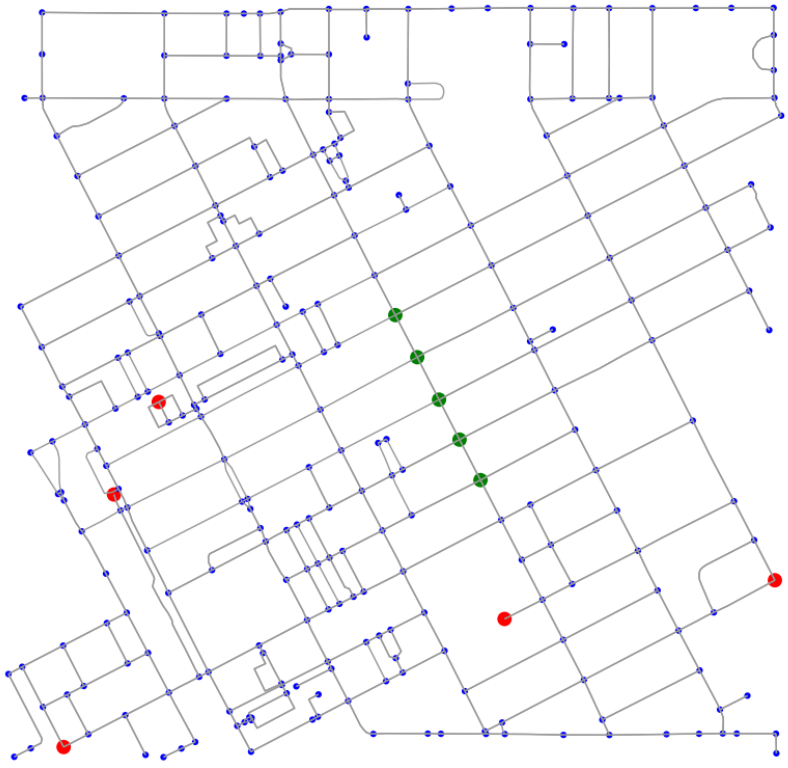

Figure 15: Most central nodes (green) and least central nodes (red) of the network

3.2 RESULTS

This study conducts a comparative analysis of bicycle accessibility in different cities, examining the interplay between macro and micro-level street network measures. The objective is to identify significant relationships between mode accessibility and network parameters, which can inform future network design strategies. Additionally, the research proposes a cost-effective approach to enhance the accessibility of an existing network by leveraging graph theory principles, thereby offering valuable insights for practical network improvements.

3.2.1 Data Description

Table 3: Mean Network Variables by Network Size Quintile.

Variables	1st quintile	2nd quintile	3rd quintile	4th quintile
Number of Nodes	23916	31404	56031	101895
No of Edges	62363	82364	145294	238821
Density	0.0003419	0.0000953	0.0000558	0.0000266
circuity	1.128	1.079	1.081	1.084
Average path length	139	78	76	82
Average Degree	2.59	2.60	2.59	2.60
Diameter	231	272	307	442
Accessibility Score	20	27	33	24
Population	251043	379822	613991	2186211
Per Capita income	33633	36574	38193	35227
% Bike Commute to Work	2.32	3.06	3.17	2.22

Table 3 presents a summary of the network parameters and demographic statistics for the study areas categorized into quintiles based on network size. The quintiles are defined such that quintile 1 includes the 10 smallest cities, quintile 2 includes the next 10 smallest cities, and so on. The network size is determined by the number of population where a higher population indicates a larger network size.

3.2.2 Modeling Accessibility

3.2.2.1 Model Results

MLR regression was utilized to model Accessibility score as the dependent variable and all other variables from **Table 3** as independent variables. Both dependent and independent variables were log-transformed. The final model was determined based on the highest R-square value, and the findings and relationships of this selected model are summarized in **Table 4**.

Table 4: *Dependent Variable ln(Accessibility)*

	Coefficient	*Std. Error*	*Standardized Coefficient*	*t-stat*	*P*
ln(circuity)	-8.921	2.895	-0.964	-3.081	0.004
ln(avg. path length)	1.229	0.464	0.858	2.649	0.013
ln(diameter)	-1.036	0.323	-0.675	-3.207	0.003
ln(percent bike user)	0.477	0.137	0.479	3.479	0.002
ln(per capita income)	0.246	0.306	0.109	0.804	0.427
constant	-3.288	4.075		-0.807	0.426
Adjusted r^2	0.608				
N	40				
Durbin-Watson test stat	1.678				

The explanatory variables that showed statistical significance, with p-values lower than 0.1, were included in the final model. The variables that were found to be statistically significant in explaining accessibility score were circuity, average path length, diameter, and the percentage of bike users. These variables had a significant impact on the accessibility score, indicating that they were important factors to consider.

While per capita income did not exhibit statistical significance in the model, it was included due to its hypothesized positive correlation with accessibility score based on prior literature.

Among the significant variables, circuity had the highest standardized coefficient of -0.964. The standardized coefficient represents the comparative importance of the parameter in the output variable. In this case, it indicates that circuity had the maximum impact on the accessibility score. A coefficient of -0.964 suggests that an increase in circuity is associated with a substantial decrease in the accessibility score.

The adjusted R-squared value of 0.608 indicates that the model provides a good fit to the data. This value represents the proportion of the variance in the accessibility score that can be explained by the included variables. A value of 0.608 suggests that the model explains approximately 60.8% of the variability in the accessibility scores, indicating a relatively strong fit.

3.2.2.2 Assumption Testing

a) Independence

To evaluate the independence of the variables, the Durbin-Watson test statistic [204-206] was computed using SPSS and the results are presented in Table 4. The Durbin-Watson test statistic is derived from the residuals of the model and falls within the range of 0 to 4. The mathematical formula for determining Durbin-Watson test static is shown in equation (***13***).

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} {e_i}^2} \quad \textbf{(13)}$$

Here, 'n' indicates the number of observations and e_i indicates the ith residual from the regression model.

A value of 2 indicates the absence of autocorrelation, while comparison with critical values helps determine the presence of autocorrelation. When the Durbin-Watson test statistic approaches 0, it suggests positive autocorrelation, signifying a positive correlation between residuals at neighboring observation points. Conversely, a value close to 4 indicates negative autocorrelation, implying a negative correlation between adjacent residuals.

In our analysis, a Durbin-Watson test statistic value of 1.678 was obtained, indicating that the data satisfies the assumption of independence. This falls within the range of 1.5 to 2.5 (rule of thumb), which is indicative of no significant autocorrelation. Therefore, we can conclude that the model exhibits independence between the variables as required by the regression analysis.

b) Homoscedasticity

To assess homoscedasticity, one commonly used approach is to examine the scatter plot of standardized residuals. Standardized residuals are calculated by dividing the residuals by their estimated standard deviation, which allows for a more meaningful comparison of residual values across different observations.

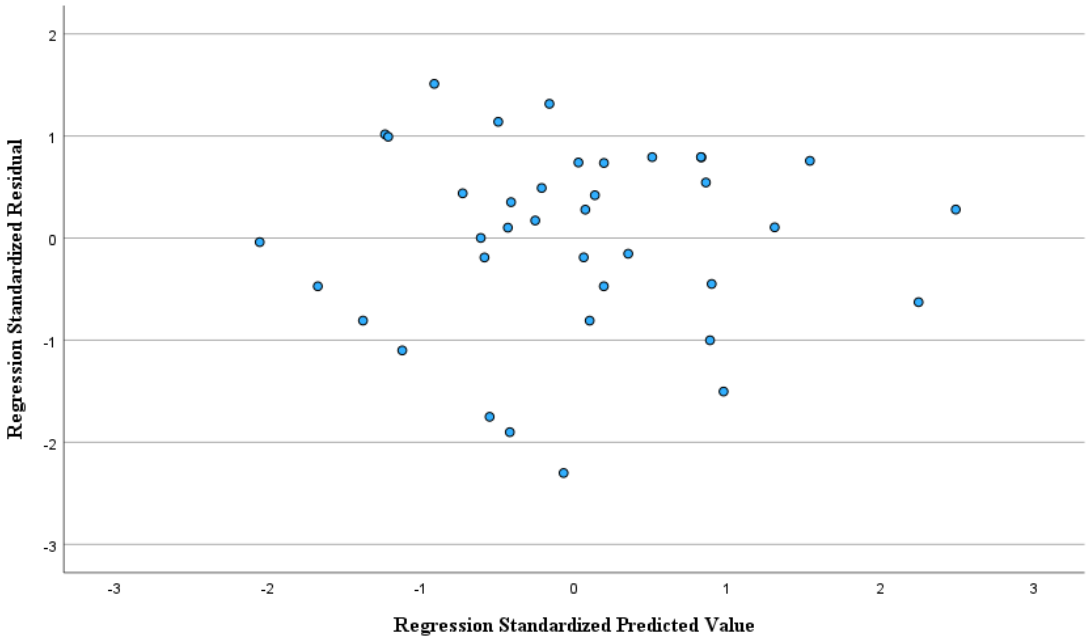

Figure 16:*Scatter plot of standardized residuals against standardized predicted values in regression model.*

If the scatter plot of standardized residuals displays a random and uniform distribution, with no discernible pattern or trend, it provides evidence that the variability of the residuals is constant across the range of predicted values. This supports the assumption of homoscedasticity. Conversely, if the scatter plot exhibits a funnel-like shape, a pattern of increasing or decreasing spread, or any other systematic trend, it suggests the presence of heteroscedasticity, which violates the assumption of constant variance.

In **Figure 16**, the standardized residuals are uniformly scattered without any discernible pattern or trend, which provides empirical support for the assumption of homoscedasticity in this regression model.

c) Multicollinearity

Collinearity statistics for the model parameter were computed using SPSS, and the results are presented in **Table 5**. The table displays the formula for tolerance and variance inflation factor (VIF) estimates.

$$VIF = \frac{1}{1 - {R_i}^2} = \frac{1}{Tolerance} \quad \textbf{(14)}$$

Here, R_i^2 represents the coefficient of determination obtained from regressing the i^{th} variable as the dependent variable and all the other explanatory variables as independent variables.

Tolerance is defined as the reciprocal of the VIF for each predictor variable. Tolerance values range from 0 to 1, with values closer to 1 indicating low levels of multicollinearity and values closer to 0 indicating high levels of multicollinearity. Conversely, a VIF value of 1 indicates no correlation between variables, while VIF values between 1 and 5 suggest moderate correlation, and VIF values above 5 signify high correlation. It is generally recommended to aim for a VIF below 5 [207].

The VIF values obtained from the model show no significant collinearity among the variables. This implies that the predictor variables in the model are not strongly correlated, ensuring the validity and reliability of the regression analysis.

Table 5: *Collinearity statistics of the model parameters.*

Explanatory Variables	*Tolerance*	*VIF*
ln(density)	0.271	3.688
ln(circuity)	0.429	2.331
ln(avg. path length)	0.521	1.92
ln(diameter)	0.285	3.505
ln(percent bike user)	0.667	1.5
ln(per capita income)	0.692	1.445

d) Multivariate Normality

The histogram in **Figure 17** displays the distribution of standardized residuals. For multivariate normality, it is expected that the histogram will demonstrate a symmetric bell-shaped curve, resembling a normal distribution. Deviations from this shape, such as skewness (asymmetry) or kurtosis (peakedness), may indicate non-normality. In this case, the histogram supports the assumption of normality as it exhibits a bell-shaped curve, suggesting that the residuals follow a normal distribution.

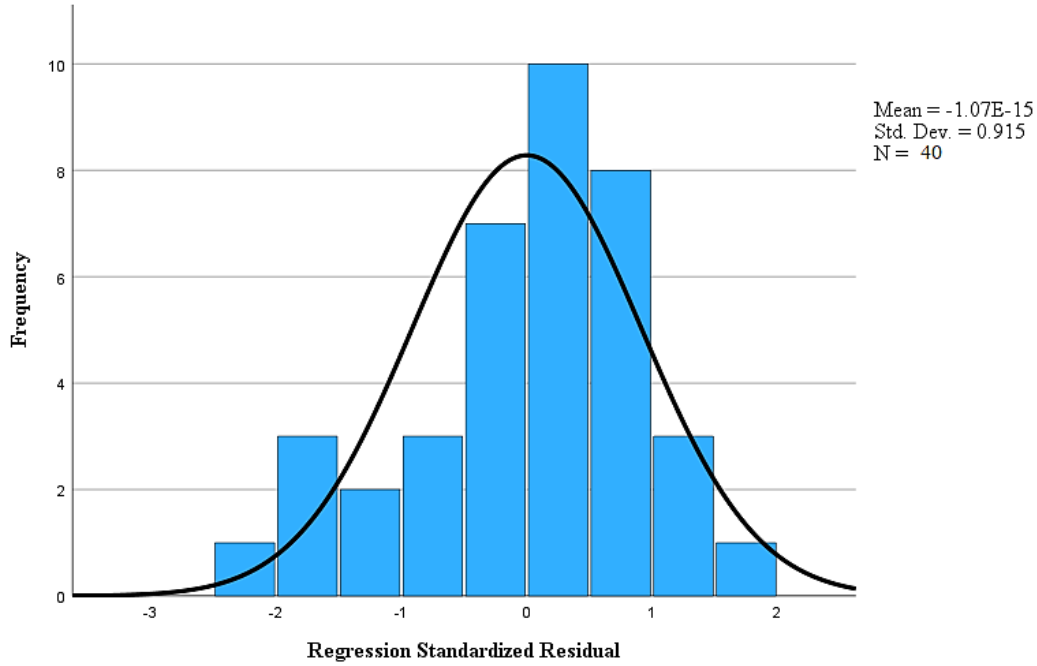

Figure 17: *Histogram of standardized residuals.*

The assumption of multivariate normality can also be tested using the normal probability plot, also known popularly as P-P plot, that compares the ordered standardized residuals against the expected quantiles of a standard normal distribution. When the points in the plot align closely along a straight line, it indicates that the standardized residuals conform to a normal distribution. Departures from the straight line suggest deviations from normality, such as positive or negative skewness when the points curve upward or downward, respectively. Any significant deviations, such as sharp bends or distinct patterns in the plot, may indicate non-normality. ***Figure 18*** visually illustrates the P-P plot, revealing that the points closely align along the straight line, providing evidence for the normality of the residuals.

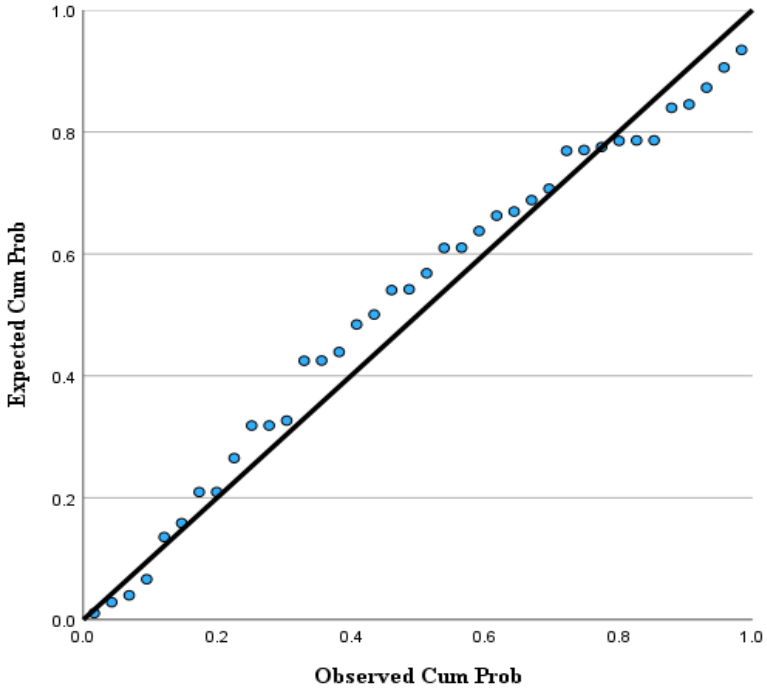

Figure 18: *Normal P-P plot of regression standardized residual.*

3.2.2.3 Model Interpretation

The number of explanatory variables in the selected model output is 5. Thus, the final model can be expressed as follows:

$$y = c\ {x_1}^{-8.291} {x_2}^{1.229} {x_3}^{-1.036} {x_4}^{0.477} {x_5}^{0.246} \quad \textbf{(15)}$$

y = accessibility score

$x_1 = circuity$

$x_2 = average\ path\ length$

$x_3 = diameter$

$x_4 = percent\ bike\ commuter$

$x_5 = per\ capita\ income$

c = 0.037

The significant findings from the model analysis are as follows:

- The relation between the accessibility score and circuity is negative, which aligns with the hypothesis. As the network becomes less circuitous, meaning the routes become more direct, the accessibility increases. This suggests that a more connected network with direct paths enhances accessibility for cyclists.
- The average path length has a positive relation with the accessibility score. This indicates that cyclists have access to a wider range of routes and options to reach their destinations. A greater variety of paths can contribute to increased accessibility by providing cyclists with more choice and flexibility.
- The diameter of the network has a negative relation with the accessibility score. As the diameter decreases, it indicates a more connected network, allowing a shorter path between two locations in the network. A well-connected network with shorter distances between different parts improves accessibility by reducing travel times and increasing connectivity.
- The percentage of bike commuters has a positive relation with the accessibility score. A higher percentage of bike commuters indicates a more accessible bike network.
- Per capita income also has a positive relation with accessibility. This suggests that higher-income individuals have the affordability to live in more accessible network areas, as land prices are generally higher in such locations. This finding highlights the relationship between socioeconomic factors and accessibility, indicating that accessibility might be influenced by economic disparities.
- The interpretation of coefficients is crucial for understanding the impact of each variable on the accessibility score. For example, let's consider circuity. If all other variables remain constant, a 1% increase in circuity would lead to a change in the accessibility score by $100 \times (1.01^{-0.08291} - 1)$ %, which equals a 7.919% decrease in the accessibility score. This means that even a small decrease in circuity can have a significant improvement on accessibility, emphasizing the importance of reducing circuitous routes for improving overall accessibility.

3.2.3 Network Intervention

The study's final model revealed that circuity has the greatest influence on network accessibility. Consequently, circuity was adopted as the indicator of network accessibility in our method. A lower circuity is indicative of higher accessibility in the network.

Figure 19 displays the nodes selected for intervention, including the two most central, two least central, and two random nodes. ***Table 6***provides their respective centrality values and labels assigned. The intervention method involves adding a new bi-directed edge to the existing network without introducing any new node. Six cases were tested, with each case entailing the addition of a new edge with one of the six selected nodes. While the chosen nodes serve as the starting nodes for the new connections, there are multiple potential end nodes already present within the network. ***Figure 29*** to ***Figure 31*** shows the available end nodes considered in all six cases.

Table 6*: Description of Nodes to be Intervened.*

Case	Label	Betweenness Centrality
Most Central Nodes	A	0.28562265
	B	0.274671533
Least Central Nodes	C	0
	D	0
Random Nodes	E	0.052196417
	F	0.034304357

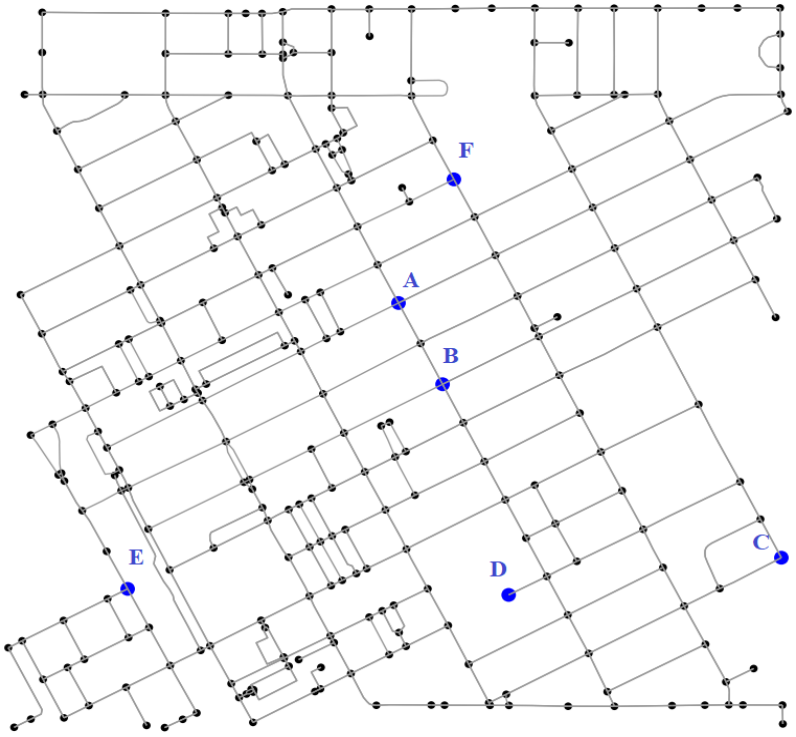

Figure 19: *Nodes to be intervened.*

The network before any intervention had an average circuity value of 1.02345. For each intervention case, the new average circuity value was measured using OSMnx. Given that the minimum circuity value of a network is "1", the following equation was utilized to calculate the percentage change in the circuity value.

$$\%\ Change = \frac{Circuity_{after} - Circuity_{before}}{(Circuity_{before} - 1)} \times 100 \quad \textbf{(16)}$$

The results of the intervention have been listed in **Table 7**. The results have been further summarized graphically in **Figure 20**. It was found that the average circuity value decreased up to 35% when the least central nodes were strengthened. On the other hand, average circuity only decreased by 4.75% when the most central node was strengthened. Strengthening the random nodes showed a decrease of up to 4.62% in average circuity.

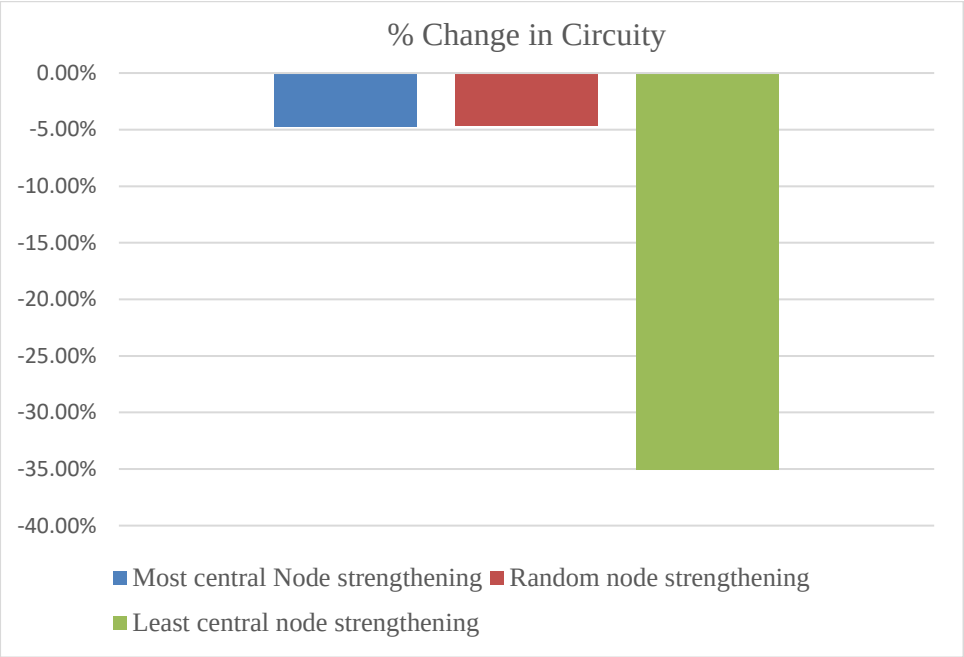

***Figure 20**: Summary of network intervention.*

The key findings obtained from the network intervention analysis are:

- The nodes with low centrality are more critical for network accessibility improvement. In other words, strengthening a low centrality node would more rapidly increase the accessibility of the network.
- Betweenness centrality can serve as a tool to identify critical nodes for accessibility interventions.
- A graph theory framework to increase accessibility has been presented here. This framework provides a systematic approach for planners to analyze the overall effect on accessibility by improving specific nodes within the network. By assessing the impact of interventions on the network's connectivity and accessibility metrics, planners can make informed decisions about which links to target for improvement.

Table 7: *Effect on average circuity upon creating a new connection in the network.*

Intervention Details	*Source Node*	*Target Node*	*Circuity*	*%Change*	*Remarks*
Strengthening the Most Central Nodes	A	1	1.02456	4.75%	
		2	1.02244	-4.30%	
		3	1.02238	-4.57%	
		4	1.02267	-3.33%	
	B	5	1.02234	-4.75%	⇐ Maximum Decrease
		6	1.02453	4.62%	
		7	1.02249	-4.10%	
		8	1.02293	-2.22%	
		9	1.02239	-4.51%	
Strengthening Random Nodes	E	1	1.02325	-0.88%	
		2	1.02325	-0.88%	
	F	3	1.02309	-1.54%	
		4	1.02246	-4.21%	
		5	1.02241	-4.42%	
		6	1.02289	-2.41%	
		7	1.02237	-4.62%	⇐ Maximum Decrease
Strengthening the Least Central Nodes	C	1	1.01522	-35.08%	⇐ Maximum Decrease
		2	1.02243	-4.37%	
		3	1.02236	-4.64%	
		4	1.02047	-12.73%	
	D	5	1.02476	5.57%	
		6	1.02122	-9.53%	
		7	1.02117	-9.73%	
		8	1.02164	-7.72%	
		9	1.02112	-9.93%	

In summary, in this chapter, street network and demographic data were utilized, and the Network Science theory and multilinear statistical regression model were employed to develop a new method for measuring the accessibility score of a bicycle network in a city. Furthermore, a method was proposed, leveraging Network Science concepts, to improve the accessibility of the bike network.

Chapter 4

In this chapter, the author focuses on utilizing Twitter data, to identify and examine transportation DEIA issues. The aim is to understand the relationship between specific DEIA challenges faced by travelers and their demographic or geographic characteristics. To achieve this, a combination of advanced analytical methods was employed, including machine learning, natural language processing algorithms, reverse geocoding, and statistical regression. By analyzing the collected data, several intriguing insights were uncovered. These insights have the potential to assist planners and policymakers in identifying underprivileged populations and areas that require greater attention in order to establish a more equitable and accessible transportation system throughout the United States.

The research questions that will be addressed in this chapter are:

- *To what extent can social media interactions serve as a viable alternative to traditional surveys in capturing public opinion regarding transportation?*
- *Can social media opinions be utilized to identify and characterize vulnerable communities in relation to transportation issues?*

The following hypotheses will be tested through this analysis:

- *Social media users discuss issues related to transportation DEIA.*
- *Social media data can effectively identify and map the locations of vulnerable communities in terms of transportation accessibility.*

4.1 DATA ANALYSIS METHODS

4.1.1 Data Collection and Preparation

4.1.1.1 Twitter Data

Twitter data was collected from New York City (NYC) which consists of five counties─ Bronx, Queens, Manhattan, Brooklyn, Staten Island using the Academic Application Programming Interface (API) [208], which provides the full history of public conversation through a full-archive search endpoint [209]. The data collection process employed the Python programming language, along with relevant Python libraries. Geolocation-based search queries were utilized to retrieve data from the study area (**Figure 21**) from February to April 2020. A total of ~ 2.75 million tweets were collected, originating from ~14.1k unique users.

The tweets obtained from the academic track API contain supplementary details such as user ID, username, profile information, and tweet location. For the analysis conducted in this study, the tweet text, user information, and location information were taken into consideration. To address the inherent ambiguity of tweets, which can arise from non-standard spelling, inconsistent punctuation, and capitalization, additional preprocessing steps were implemented. The purpose of these steps was to extract clean tweet text and usernames suitable for analysis. This involved cleaning the text data and usernames by eliminating noise elements such as HTML tags, character codes, emojis, and stop words. Additionally, the tweets underwent tokenization, a process in which expressions, sentences, paragraphs, or entire text documents were broken down into smaller units, referred to as tokens, which are typically individual words or phrases.

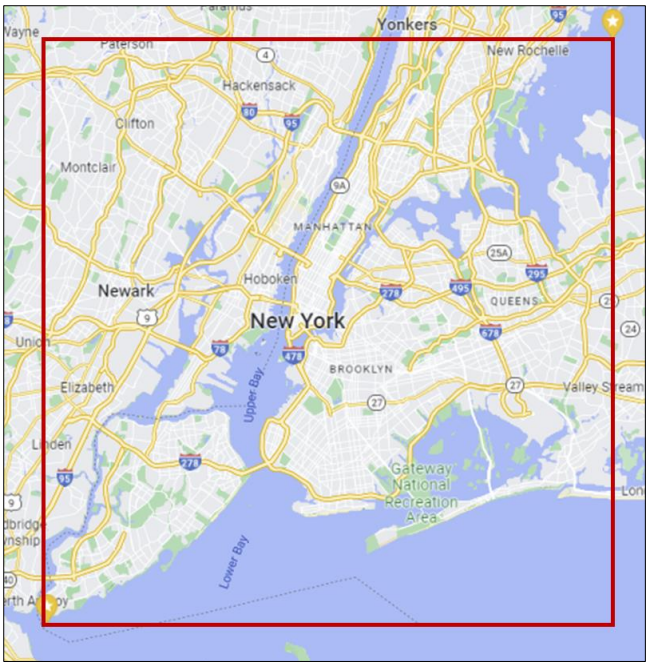

***Figure 21**:Bounding box of the area for tweet collection.*

Although the data was gathered through geolocation-based search queries, it included numerous geotagged tweets that originated from locations outside of New York City. The tweets that have a (latitude, longitude) coordinate outside the bounding box [-74.264667, 40.487217, -73.766128, 40.911357] were removed from the dataset (**Figure 21**). Additionally, the Pandas library in Python was employed to identify and eliminate any duplicates within the dataset, ensuring that all tweets were unique. The study specifically focused on tweets related to transportation DEIA. A tweet's relevance was assessed by identifying specific keywords or tokens within the tweet; further details regarding the steps and significance of relevance filtering can be found here [80]. Relevance filtering was used to find tweets about DEIA using the keyword list below:

DEIA related keywords (40 words)	*"dei", "diversity", "equity", "excluded", "inequity", "inclusion", "unequal", "accessibility", "inaccessible", "inequality", "inequitable", "injustice", "unjust", "justice", "afford", "unaffordable", "affordable", "discriminated", "discrimination", "disability", "disabled", "wheelchair", "ada", "gender", "poor", "women", "disadvantaged", "underserved", "deprived", "underprivileged", "denied", "marginalized", "exclusion", "polarization", "aged", "lowincome", "income", "racism", "race"*

To determine the DEIA relevance of the tweets, a tweet was considered relevant if it contained at least one of the DEIA keywords identified for this study. While this approach may exclude some potentially relevant tweets, it guarantees that all tweets containing these keywords are only included in the filtered dataset for subsequent analysis. At the end of this step, a clean dataset consisting of 37,552 tweets was obtained, which is 1.36% of the original dataset. Demographic Data

Reverse geocoding was applied on the tweets to find the census tract information of the tweets. 431 census tracts were identified from the DEIA related tweets (n=37,552). Each census tract has unique socio-demographic characteristics. To identify the transportation DEIA challenges faced by different demographic groups of people, the following demographic characteristics were collected for each census tract from the 5-year survey of American Community Survey

(ACS) published in the year 2020:

- UPL: Percentage of population living under the poverty limit
- HSE: Percentage of population who completed high school education
- PI: Per Capita Income

Based on the survey respondent's last name, the Census Bureau compiles annual estimates of the racial origin distribution for each county of the US. The estimates are based on the most recent decennial census and population change estimates (including deaths, births, and migration) since then. In this study, last names categorized by different ethnicities were collected from the census bureau to identify the ethnicity of users in the Twitter data applying machine learning algorithms. User's ethnicities were categorized as Asian, Black, Hispanic, and White. The study utilizes the 2010 estimates for this analysis.

4.1.1.2 Social Security Administration Data

The Social Security Administration (SSA) collects names from social security card applications for individuals born in the United States after 1879. In order to analyze the gender demographics of Twitter data, we obtained the first names of the applicants from the SSA website. According to the SSA data, there are 63,152 male names and 37,212 female names included in their records.

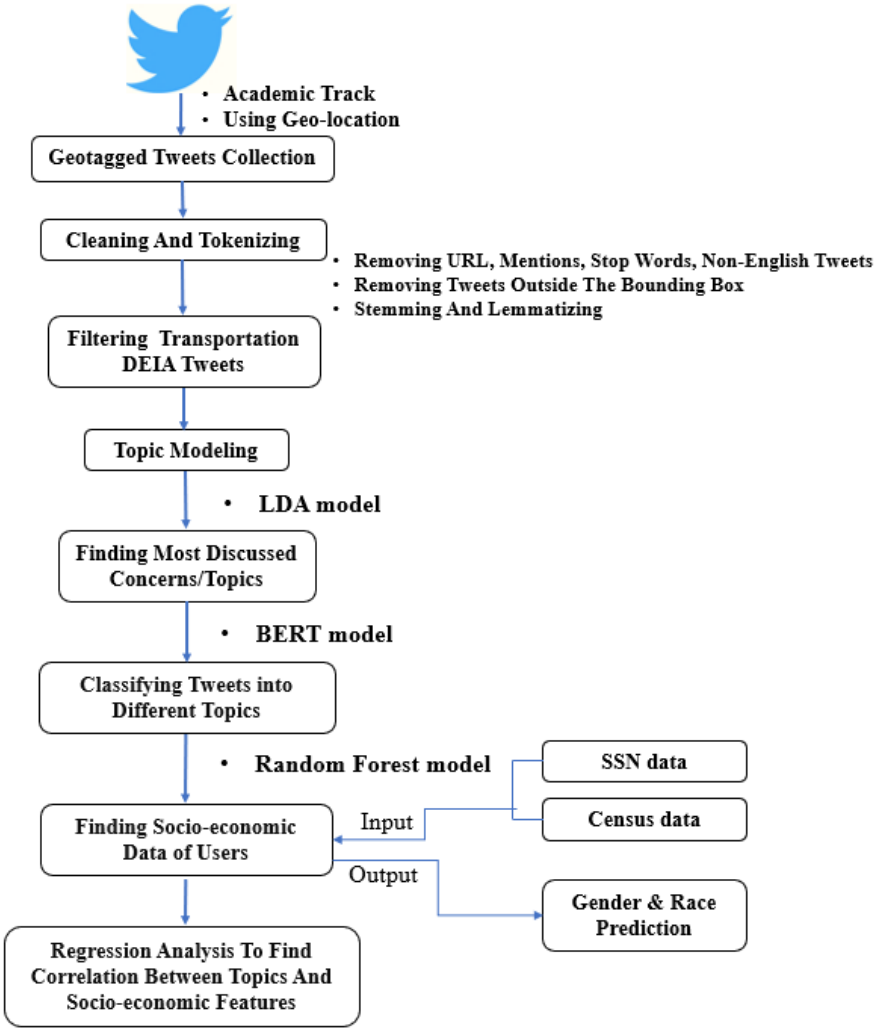

Figure 22: Methodology of the study

4.1.2 Tweet Classification

The topic modeling method was utilized to identify the major topics within the tweets. Once the major topics were determined, the tweets were classified into respective topic categories, and each tweet was assigned a label based on the topic category to which it belonged. The next sections provide a detailed explanation of the procedure.

4.1.3.1 Topic Modeling

Topic modeling is a computational technique used to uncover hidden thematic patterns or topics within a collection of text data. It is a way to automatically analyze and organize large volumes of text by identifying the underlying topics or themes that are prevalent in the data. Among many methods available, the study utilized the Latent Dirichlet Allocation (LDA) model [210]. The reason behind selecting this model is its ability to uncover hidden, unexplored topics within the dataset. Supervised models can only identify topics they have been trained on, whereas LDA, being a generative probabilistic model, assumes that each text exhibits a variable distribution of underlying themes. Each document in LDA is supposed to be a combination of topics, and each topic is assumed to be a combination of words. The following activities are taken by LDA to assign topics to each of the documents:

- It begins by presuming there are K subjects in the document, loops through it, and assigns each word to one of the K topics at random.
- It cycles over each word in each document and computes:
  a) P ($w_j|t_k$): Percentage of assignments to topic t_k across all documents for a specific word w_j
  b) P ($t_k|d_i$): Percentage of words in document d_i that are assigned to topic t_k
- Considering all other words and their topic assignments, reassign topic 'T' to word w_j with probability $p(t_k|d_i)*p(w_j|t_k)$.

The final step is repeated iteratively until a steady condition is achieved, where further changes in topic assignments no longer occur. The topic allocations obtained through this process are then used to compute the topic ratios for each document. The method is illustrated in **Figure 23**.

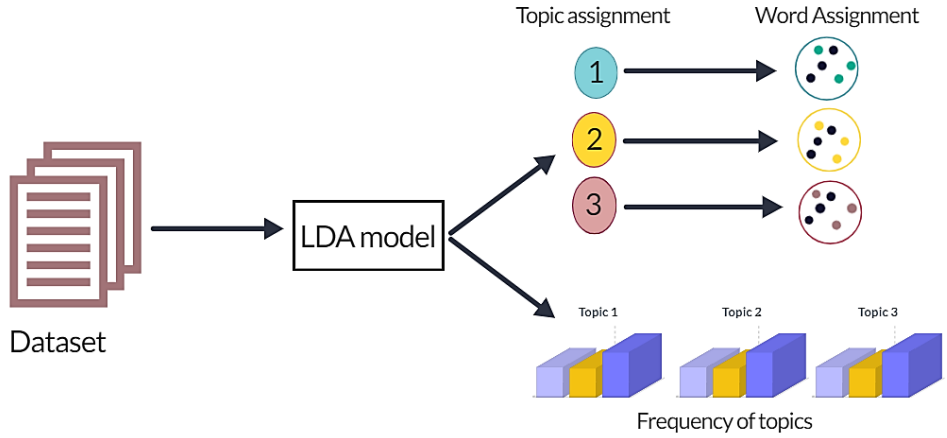

Figure 23: *Conceptual Figure of LDA model.*

The model identified five major topics related to transportation DEIA discussed by the netizens in the dataset (Transit infrastructure, social disparity, active transport, accessibility and ride

sharing) with a coherence score of 0.3926. The details are shared in **Table 8**.

Table 8: *Optimum topics identified in the dataset.*

Topics	**Most probable words in a topic (probability)**
Transit	station (0.131), line (0.112), bus (0.018), train (0.010), busstop (0.009)
Social disparity	rich (0.067), inequity (0.043), income (0.027), equity (0.026), race (0.010)
Accessibility	Wheelchair (0.028), disabled (0.020), woman (0.016), access (0.012), opportunity (0.011)
Active transport	bike (0.016), walking (0.015), blocked (0.014), blockedbikenyc (0.014), vissionzero (0.014)
Ridesharing	ride (0.017), rent (0.017), rideshare (0.011), carpool (0.011), uber (0.012)
Others	virus (0.032), aged (0.028), quarantine (0.026), chinese (0.018), symptom (0.009), incident (0.067), construction (0.013)

4.1.3.2 Data Labeling

Manual labeling in tweet classification refers to the process of assigning predefined categories or labels to individual tweets within a dataset. This is typically done by human annotators who review each tweet and determine which category or label best describes its content. This annotated dataset serves as the training ground for machine learning algorithms, enabling them to learn patterns and associations between the textual content of tweets and their corresponding labels. Manual labeling is crucial for building accurate and reliable machine learning models.

In this study, the tweets were annotated into six categories: ‘Transit infrastructure’, ‘Social disparity’, ‘Active transport’, ‘Accessibility’, ‘Ride sharing’ and 'Others'. 350 tweets from each of the categories were randomly selected with a total of 2100 tweets for manual annotation by two human annotators. To ensure accuracy, the labels were assigned only when both annotators agreed on them, guaranteeing the retrieval of correct labels. Each tweet was assigned only one label from the six available options.

4.1.3.3 BERT modeling

BERT (Bidirectional Encoder Representations from Transformers) classification is a popular Natural Language Processing (NLP) technique that employs deep learning models to solve text-based issues and was developed by Google's AI researchers in 2018 [211]. The earlier models could only read text data unidirectionally, which means that the model can only use the information available from the words that come before it in the sentence. In this context, the model lacks access to future words and relies solely on the preceding context. In comparison to them, BERT's bidirectional approach enables it to understand the context from both sides of each word in a sentence (from left to right and right to left), resulting in a greater comprehension of the sentence structure and meaning. BERT facilitates the accurate categorization of text into predefined classes or categories within the context of text classification. This could involve sentiment analysis (categorizing texts as positive, negative, or neutral), spam detection (categorizing emails as spam or not spam), or any other text-based categorization.

The BERT model was trained with manually annotated tweets (n=2,100). For training, K-fold-cross-validation was used which involves dividing the dataset into k equal-sized subsets, or

folds, where k was set to 10.

The trained model was applied to the dataset to predict the labels. The predicted labels were compared with the actual labels (manually annotated) through the confusion matrix (**Figure 24**). A confusion matrix can reveal the following information:

- True Positives (TP): The number of instances that were correctly predicted as positive by the model.
- True Negatives (TN): The number of instances that were correctly predicted as negative by the model.
- False Positives (FP): The number of instances that were incorrectly predicted as positive by the model.
- False Negatives (FN): The number of instances that were incorrectly predicted as negative by the model.

The commonly used performance metrics to evaluate the prediction performance of the model are precision, recall, and F1 score. Precision was calculated as:

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive} = \frac{TP}{TP+FP} \quad \textbf{(17)}$$

Recall was calculated as:

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative} = \frac{TP}{TP+FN} \quad \textbf{(18)}$$

F1 score was calculated using the following equation:

$$\text{F1 score} = \frac{2*(Precision*Recall)}{(Precision+Recall)} = \frac{TP}{TP+\frac{1}{2}(FP+FN)} \quad \textbf{(19)}$$

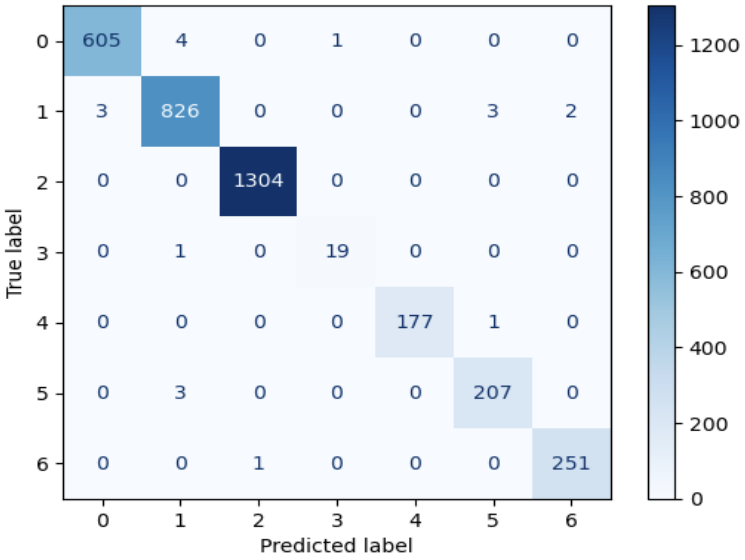

Figure 24: *Confusion matrix showing prediction performance of the training model.*

The F1 score of the model is determined to be 0.9868673809846785, indicating a high level of overall performance. Additionally, the precision value is measured at 0.9868349564997437, signifying the model's ability to accurately identify positive instances among the predicted

positives. Moreover, the recall value is computed as 0.9869055723953398, reflecting the model's capacity to correctly identify positive instances among the actual positives. These evaluation metrics collectively demonstrate the model's strong performance in terms of accuracy, precision, and recall.

4.1.3 Gender and Race Prediction

The study determined the demographic details, such as gender and race, of the individuals posting tweets by referencing social security and census information. This process was conducted simultaneously with the classification of tweets. The resulting demographic data was then utilized in a discrete choice model study to examine its impact on the individuals' DEIA related concerns.

The study utilized the names associated with Social Security Numbers (SSNs) sourced from the Social Security Administration for gender identification. The dataset consisted of 63,152 male and 37,212 female names. Some names appeared on both lists, so these duplicates were removed, leaving a total of 85,736 unique names. Several supervised machine learning techniques such as Random Forest, Naive Bayes, Support Vector Machine, and K-nearest Neighbor were trained, and the best performing random forest model was chosen to predict the gender from the first names of Twitter users.

Regarding the determination of race or ethnicity, this was challenging due to the limited information available on each Twitter user. The study conducted an analysis based on the self-reported last names in the user's profile. We used the data published by the Census Bureau, which provides a breakdown of the racial and ethnic composition of each surname with a population of over 100,000 people in the United States, as recorded in the 2010 Census. Using several supervised machine learning approaches, we compared the users' last names with the 2010 Census data. Support vector machine outperformed all other techniques in this case [212]. For instance, the surname "Smith" is associated with individuals who identify as White 70.9 percent of the time, Black 23.11 percent, Asian 0.5 percent, and Hispanic 2.4 percent.

4.1.4 Discrete Choice Model

The discrete choice framework was pioneered by McFadden (1973) in the field of travel demand analysis [213]. Initially, discrete choice models were primarily used to examine travel mode choice, which involved selecting between options like train, bus, car, or airplane for travel purposes. As the framework evolved, it was also employed to study the choice of travel routes and destinations, as demonstrated by the Ben-Akiva in 1985 [214].

In a discrete choice scenario, a decision maker denoted as "n" is faced with the task of selecting one option from a set of "J" alternatives. Here, the term "alternatives" refers to the various items, actions, or locations that can be chosen, while the word "choice" pertains to the decision made by the decision maker in selecting a specific alternative. Conventionally, the entire range of available options is referred to as the "choice set" or “set of alternatives”.

The decision maker, denoted as "n" derives a certain level of utility (such as profits or satisfaction), labeled as "U_{ni}," from alternative "i" if that particular alternative is chosen. The principle of utility maximization states that the decision maker will choose alternative "i" only if they anticipate deriving more utility from it compared to any other available alternative. Consequently, if the decision maker selects alternative "i" it implies that they expect to obtain less utility from each of the other alternatives:

$$U_{ni} > U_{nj} \; ; (\forall j \neq i) \qquad \textbf{(20)}$$

Only the decision maker possesses knowledge of the utilities. Researchers do not have the knowledge; however, researchers can observe to the "J" alternatives, certain attributes (a_{ni}) of the alternatives, and decision maker attributes (d_n). A representative utility or systematic utility (V) can be established, that links these observed attributes to the decision maker's utility:

$$V_{ni} = V(a_{ni}, d_n) \; \forall i \quad \textbf{(21)}$$

The researchers have an incomplete understanding of utility, so generally $U_{ni} \neq V_{ni}$. To address this issue, the utility can be expressed as the sum of representative utility (V_{ni}) and an unobserved term (ε_{ni}) that encompasses the factors determining utility but remains unobservable to the analyst. This unobserved term is typically treated as random:

$$U_{ni} = V_{ni} + \varepsilon_{ni} \quad \textbf{(22)}$$

The probability of the decision maker (n) selecting alternative (i) is equivalent to the probability of the utility associated with choosing alternative (i) being greater than the utility associated with any other alternative within the choice set.

4.1.5.1 Multinomial Logit Model (MNL)

When the unobserved random utility components (ε_{ni}) follow an independent and identically distributed (IID) extreme value distribution, commonly known as a Gumbel distribution, the Multinomial Logit (MNL) or Conditional Logit (CL) model can be used.

This study used a multinomial logit (MNL) model to analyze public concerns regarding different DEIA concerns. The MNL is a popular type of random utility Discrete Choice model [215]. In this model, an individual (represented by n) selects one choice from discrete alternatives by assessing the associated features J (J= set of DEIA issues) in order to maximize their utility.

The MNL model is constructed based on the assumption that each unobserved term, ε_{ip}, follows an independent and IID extreme value distribution, such as the Gumbel or type 1 extreme value distribution. The likelihood of person 'n' selecting alternative j can be calculated by solving the following mathematical formula:

$$P_{ij} = \frac{e^{V_{ni}}}{\sum_{j=1}^{J} e^{V_{nj}}} \quad \textbf{(23)}$$

In MNL, the representative utility (V_{ni}) is defined as:

$$V_{ni} = \sum_{k=1}^{K} \beta_{ki} X_{kn} \quad \textbf{(24)}$$

In this equation, 'K' denotes the number of predictor variables, X_{kn} represents the value of k^{th} predictor variable for n, of observed explanatory variables associated with choosing a particular alternative, and β represents the parameter for the observed utility.

Combining equation (**23**) and (***24***)

$$P_{ni} = \frac{e^{\sum_{k=1}^{K} \beta_{ki} X_{kn}}}{\sum_{j=1}^{J} e^{\sum_{k=1}^{K} \beta_{kj} X_{kn}}} \quad \textbf{(25)}$$

This formula was used for understanding a correlation between transportation DEIA concern and various socioeconomic characteristics.

4.2 RESULT

4.2.1 Classification Outcome

The BERT classification model successfully predicted the category of each tweet in the test dataset. Each tweet was labeled one of the six topics “transit”, “active transport”, “ride sharing”, “social disparity”, “accessibility” or “others”. The distribution is shown in **Figure 25**. It is interesting to find out that during the study period, NYC people talked most about the transit. This finding can be attributed to the peak of the COVID-19 pandemic, which instilled fear in people, discouraging them from utilizing public transportation and leading to numerous discussions on this subject. Moreover, the temporary suspension of transit services disrupted the mobility of individuals heavily reliant on public transportation. Residents also talked a lot about social disparities, as many individuals faced job losses during the COVID-19 crisis, posing significant challenges to their survival and well-being. The topic of active transportation was frequently discussed. For those who cannot use transit for commuting, active transportation is an affordable alternative. However, biking is challenging to them due to the inequitable bike network. They frequently addressed these topics on twitter.

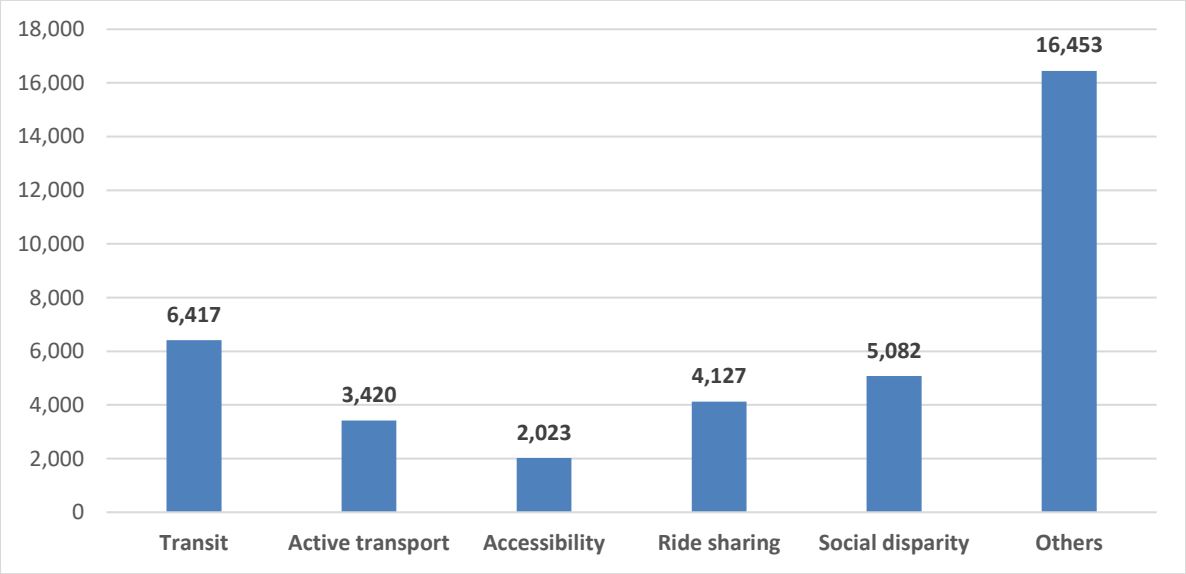

***Figure 25**: Text classification outcome.*

Interestingly the number of male twitter users are almost two times than the female users in the final dataset. Each tweet was labeled either male or female based on the result of Random Forest gender and race detection model [212].

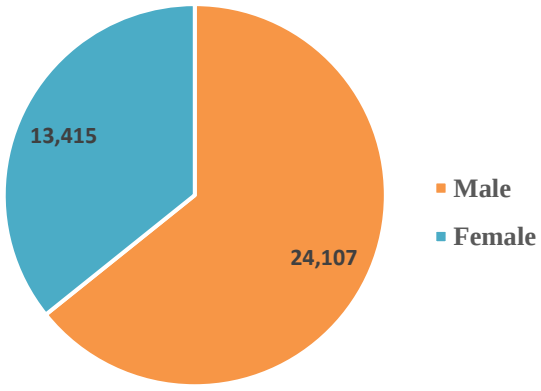

Figure 26: distribution of gender among users.

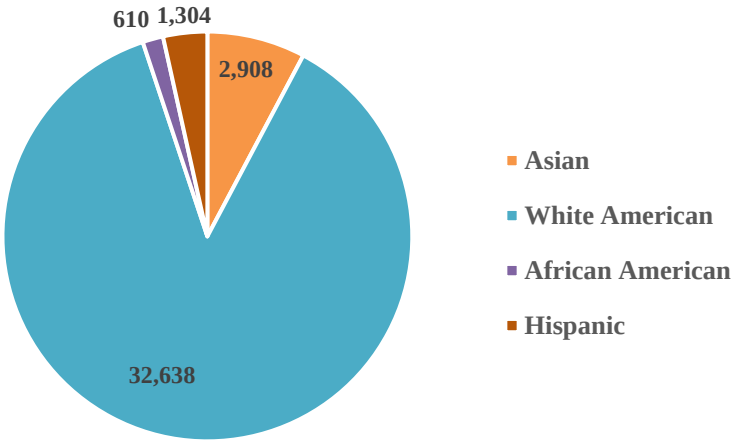

Figure 27: Distribution of race among users.

About 87% of the users were white American in the dataset while the number of African American were least (~2%) The white American users are most vocal about DEIA issues on twitter (**Figure 27**). Maximum transportation DEIA related tweets were generated from Manhattan county whereas the minimum number of tweets were generated from Richmond county. The Manhattan users are more vocal about their equity issues on social media (**Figure 28**).

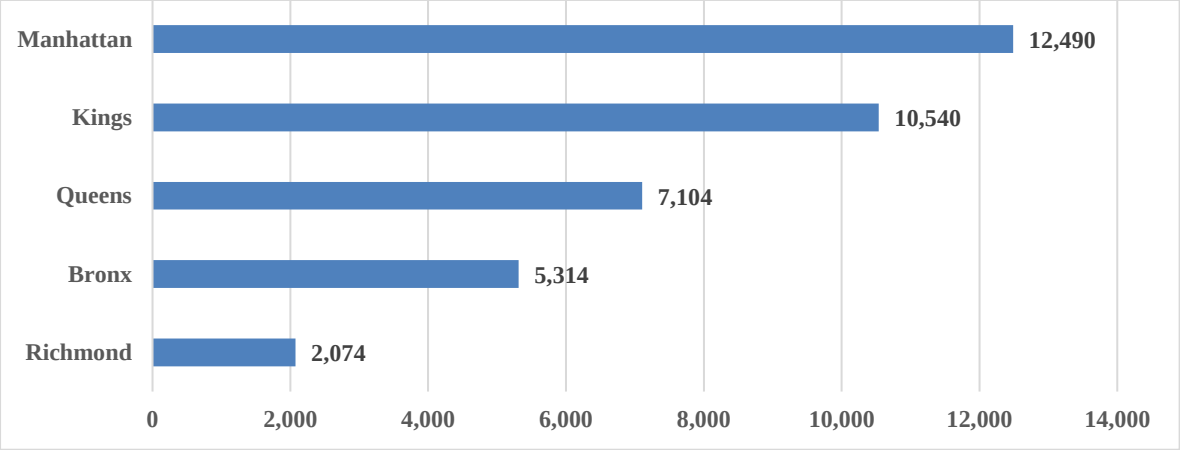

Figure 28: Distribution of tweets among counties.

4.2.2 Modeling Demographic relation with DEIA challenges

The Multinomial Logit (MNL) model was applied to examine the relationship between demographic or geographic factors and the specific challenges related to DEIA faced by individuals. The six topics detected from social media conversations (transit, active transport, ridesharing, accessibility, social disparity, and others) were treated as alternative set of choices. Five demographic variables (gender, race, per capita income, education rate, and poverty rate) and five geographic variables (Manhattan county, Kings county, Queens county, Bronx county, and Richmond county) were included to model how they influenced the selection among the choice set of DEIA challenges.

4.2.2.1 Model Result

The statistical description of variables is shown in **Table *9***. All variables in the model are categorical except the PI, HSE and UPL. The final result of the model in showed in **Table *10***.

Table 9: Descriptive Statistics of Key Variable.

Variable Description		Mean or %	Minimum	Maximum
User's Gender				
	1: Female 0: Male	64%	0	1
User's Race				
	White	87.13%	0	1
	Asian	7.76%	0	1
	Black	1.63%	0	1
	Hispanic	3.48%	0	1
Tweet Location: County				
	Manhattan	33.29%	0	1
	Kings	28.09%	0	1
	Queens	18.93%	0	1
	Bronx	14.16%	0	1
	Richmond	5.53%	0	1
User's Socioeconomic attribute				
PI: Per Capita Income($)		$45,502	$2,758	$354,695
HSE: High School Education Rate(%)		0%	48.56%	100%
UPL: Below Poverty Limit Rate (%)		0%	8.50%	100%
DEIA concerns				
	Transit	17.10%	0	1
	Active Transport	9.11%	0	1
	Accessibility	5.39%	0	1
	Ridesharing	11.00%	0	1
	Disparity	13.54%	0	1
	Others	43.85%	0	1

The p-value or Pr(>|z|), represents the statistical significance of each variable. Values that have a p-value below 0.05 are regarded as statistically significant. The McFadden Pseudo R-squared value for the model is 0.501. This metric provides an estimate of the goodness-of-fit, indicating that approximately 50% of the total variation in the dependent variable is explained by the independent variables included in the model. The Hosmer-Lomeshow goodness of fit test (HL test) is used to evaluate the model's fit. The resulting p-value from the test is 0.392. Since this p-value is greater than the conventional significance level of 0.05 [216], it suggests that the model adequately fits the data.

The coefficients represent the log-odds of the outcome variable for each unit change in the explanatory variable, while holding all other factors constant. A positive sign implies that an increase in the variable is associated with a higher probability of choosing the corresponding topic, while a negative sign suggests a lower probability.

Table 10: Model Result (* means that variable is statistically significant at $\alpha = 0.05$)

Variable	Category	Coeff.	p-value	Significance
Female	Transit	0.914	0.015	*
	Active Transport	0.438	0.782	
	Accessibility	-0.898	0.113	
	Ridesharing	-2.330	0.262	*
	Disparity	0.6119	0.038	
PI: Per Capita Income	Transit	-0.914	0.142	
	Active Transport	-0.115	0.323	
	Accessibility	0.703	0.013	
	Ridesharing	0.17	0.005	*
	Disparity	0.361	0.912	
HSE: High School Education	Transit	-0.527	0.573	
	Active Transport	0.139	0.755	
	Accessibility	-0.031	0.066	
	Ridesharing	0.999	0.037	*
	Disparity	0.679	0.382	
UPL: Under Poverty Limit	Transit	0.089	0.248	
	Active Transport	0.234	0.018	*
	Accessibility	-1.745	0.755	
	Ridesharing	0.219	0.174	
	Disparity	0.146	0.045	*
Asian	Transit	0.724	0.41	
	Active Transport	0.08	0.081	
	Accessibility	0.469	0.045	*
	Ridesharing	-0.048	0.188	
	Disparity	0.157	0.015	*
Tweet location: Richmond County	Transit	0.034	0.351	
	Active Transport	0.063	0.028	*
	Accessibility	-1.104	0.048	
	Ridesharing	0.038	0.102	
	Disparity	0.089	0.38	
Tweet location: Queens County	Transit	0.726	0.032	*
	Active Transport	-0.193	0.205	
	Accessibility	0.05	0.683	
	Ridesharing	0.213	0.581	
	Disparity	0.895	0.174	
Tweet location: Bronx County	Transit	0.914	0.015	
	Active Transport	0.439	0.782	*
	Accessibility	-0.899	0.113	
	Ridesharing	2.331	0.262	*
	Disparity	0.612	0.039	
Number of cases	37,522			
McFadden Pseudo r^2	0.501			
Hosmer-Lomeshow test: p-value	0.392			

The magnitude of the coefficient represents the strength of the relationship. Larger coefficients indicate a more significant impact on the likelihood of selecting a particular DEIA topic. If a coefficient has a p-value below the chosen significance level (e.g., $\alpha = 0.05$), it suggests that the variable has a statistically significant impact on the choice of DEIA topic.

4.2.2.2 Model Interpretation

The choice set for this model comprises six transportation DEIA topics: transit, active transport, ridesharing, accessibility, social disparity, and others. The model was constructed to analyze the relation between ten demographic/geographic factors with the probability of tweeting about one of the six topics in the choice set. The model yields intriguing insights into the relationships between these factors which can help planners to detect the marginalized population and areas with inequitable transportation networks.

Specifically, it was found that females in NYC exhibited a strong inclination to discuss transit and ridesharing topics in their tweets. Furthermore, per capita income displayed a negative association with transit, indicating that higher-income populations tend to discuss transit less frequently. Conversely, there was a statistically significant positive relationship between per capita income and ride sharing, suggesting that individuals with higher incomes tend to tweet more about ride sharing compared to other topics.

The percentage of the population with a high school education exhibited a statistically significant positive relationship with the topic of accessibility, indicating that individuals with higher levels of education are more likely to engage in conversations related to ridesharing and accessibility.

Additionally, the percentage of the population below the poverty limit showed a positive relationship with both active transport and social disparity, implying that lower-income populations rely more on active forms of transportation and are more sensitive to social and economic disparity. Consistent with prior literature, it was observed that Asians tended to express more concerns about accessibility and social disparity [217, 218].

Interestingly, residents of Richmond County were found to engage in more discussions about active transport, while individuals residing in Queens County had a higher tendency to discuss transit topics. This suggests that Queens County may have a less developed public transit infrastructure, while the residents of Richmond County face challenges in cycling due to inequitable cycling infrastructure.

These findings shed light on the intricate connections between demographic/geographic factors and transportation-related topics, providing valuable insights into transportation DEIA issues in NYC.

4.3 SUMMARY

In this chapter, the twitter conversations of NYC citizens that are related to DEIA were analyzed using machine learning and natural language processing algorithms to find out the key transportation DEIA concerns. The concerns were further correlated with demographic (e.g., gender, race, income of twitter users) and geographic factors (e.g., location of twitter users) using statistical model to understand the relation between such factors and the sensitivity to the DEIA issues. The model yields the following key findings.

- Females were more likely to discuss the inequities they faced in transit and ridesharing while traveling, while Asian individuals showed greater sensitivity to accessibility and social disparity issues.

- Socioeconomic factors, including education, income, and poverty, also played a role in influencing travel choices. Discussions about ride sharing were more prevalent among those with higher education and higher income.
- Moreover, people's tweeting behavior concerning DEIA concerns was influenced by their location. For example, residents of Richmond county were more engaged in discussions about active transport, while those in Queens county showed a higher focus on transit-related issues compared to other concerns.

Chapter 5

5.1 CONCLUSION

5.1.1 Summary

The study aimed to analyze bicycle network data from various regions in the USA to assess the impact of graph properties on network accessibility. A statistical regression model was employed to examine the relationship between accessibility scores and graph properties, as well as demographic characteristics as explanatory variables. Building upon the insights gained from the model, a systematic intervention approach was proposed using network science to enhance network accessibility. The findings of this study provide valuable guidance to planners, aiding them in identifying target areas and road segments for improving bike accessibility in the United States.

Another objective of this study was to examine the role of social media platforms, specifically Twitter, in gaining insights into public perceptions and attitudes towards transportation diversity, equity, inclusion, and accessibility (DEIA) indicators. Through the utilization of a deep learning approach, tweets from New York City users between February 2020 and April 2020 were classified based on various DEIA issues. Additionally, a discrete choice model was developed to assess user sensitivity towards different transportation DEIA matters, considering demographic characteristics and geographic locations. The findings offer valuable implications for planners, as the model can aid in identifying target populations and locations for prioritizing efforts towards transportation DEIA development.

5.1.2 Key Findings

Key findings of bike network data analysis:

- Using the Multiple Linear Regression (MLR) model, the study explored how network and demographic factors influence bike network accessibility in the area, revealing some interesting insights.
- Several network parameters, including diameter, circuity, and average path length, are crucial indicators of accessibility. The study found that diameter and circuity have a negative relation to accessibility, whereas average path length showed a positive relation.
- Demographic factors such as percent bike commuters also indicate bike accessibility as found from this study. Higher percentages of bike commuters were associated with increased bike network accessibility, providing planners with valuable information for identifying areas that require attention.
- The research highlighted the importance of reducing circuity to enhance accessibility, with even minor reductions having a substantial impact on overall accessibility.

- Nodes with low centrality emerged as key players in improving network accessibility, emphasizing the need to strengthen these nodes to enhance the bike infrastructure.
- The utilization of betweenness centrality proved instrumental in identifying critical nodes for accessibility interventions, offering valuable guidance for targeted improvements.
- The graph theory framework presented in this study provides planners with a systematic approach to analyze the impact of interventions on network connectivity and accessibility. It serves as a valuable tool for making informed decisions and directing focused efforts to improve the bike infrastructure.

Key findings of the social media data analysis:

- The study uses an unsupervised machine learning algorithm (LDA) to detect the major transportation DEIA issues discussed by the twitter users of NYC during study period. Five topic were detected ‘transit infrastructure’, ‘active transport’, ‘ridesharing’, ‘accessibility’, ‘socio-economic disparity in the optimal model.
- The study develops a Multinomial Logit (MNL) model to analyze the demographic and geographic correlation of individuals with their sensitivity to the transportation DEIA issues.
- Personal characteristics such as race and gender have significant relation with DEIA issues. Females are more likely to talk about transit and ridesharing inequities they face while travelling, while Asian people are more sensitive to accessibility and social disparity issues.
- Socioeconomic factors like education, income, poverty also influence travel choices. Higher education and high income are associated with discussions about ride sharing.
- Location also affects people's tweeting behavior related to DEIA concerns. For example, people in Richmond county discuss more about active transport, while those in Queens county focus more on transit than other issues.

5.1.3 Limitations

The study acknowledges limitations related to the presence of bot-generated tweets and suggests further research to eliminate such tweets using available methods. A bot-generated tweet is a message posted on Twitter that has been automatically generated by a computer program, often lacking human touch and genuine interaction. Conducting surveys among Twitter users and integrating national databases, such as the National Household Travel Survey, could improve the model.

The social media users may not be representative of the opinion of the entire population living in the study area. As a result, the findings may primarily reflect the views and perspectives of specific segments of the population, particularly the younger demographic and those residing in areas with high internet accessibility. Using stratified sampling method can produce a better representation of the community.

Although the author identified key DEIA related topics discussed in the twitter data, it doesn’t capture the positive or negative sentiment of the tweets. By applying sentiment analysis, this limitation can be addressed. Such analysis can provide a better understanding of the DEIA issues of an area.

5.1.3.1 Privacy Concern

During the course of this research, Twitter data was employed to investigate various social phenomena and aspects pertinent to the study's scope. Nevertheless, it is essential to address potential privacy concerns arising from the utilization of such data. While Twitter is a public platform where users voluntarily share information, analyzing their tweets raises ethical considerations. Even though no direct quotes or specific tweet content were used in this study, and all data was anonymized and aggregated, there remains a possibility that some users' identities or sensitive details could be indirectly inferred through patterns or associations in the data. To mitigate such risks and uphold user privacy, stringent measures were taken to ensure the complete anonymization of the data used. Usernames or any other identifiable information were excluded from the analysis to protect the privacy of Twitter users fully. Additionally, this research adheres to ethical guidelines and regulations to respect and preserve the privacy of individuals contributing to the public discourse on social media platforms like Twitter.

5.1.4 Future Directions

Several potential future directions can be explored in the network data analysis presented in this thesis:

- Consideration of other web map platforms: While this study focused on utilizing a specific web map platform (OpenStreetMap), other platforms such as Google Map can be explored.
- Testing alternative statistical regression models: Although this research employed an MLR model, there is room for further investigation and experimentation with alternative models. Testing different regression techniques can provide a comparative analysis, enabling the selection of the most appropriate model for network data analysis.
- Expansion of sample size: Conducting the analysis with a larger sample size, encompassing a greater number of cities, would enhance the accuracy and generalizability of the model.
- Extending analytical scope: While the research focused on bike routes, the methodology applied in this study can be extended to analyze accessibility concerns for other transportation modes, such as buses and trains, by utilizing their respective shapefiles. This expansion of analysis has the potential to provide a comprehensive understanding of transportation equity across various modes, guiding policymakers and urban planners in improving transportation DEIA in the study area.

The future direction for utilizing social media data in analyzing transportation DEIA presents several potential avenues for exploration. These include:

- Expanding data sources: Exploring the use of additional social media platforms, such as Facebook and Uber could provide a more comprehensive understanding of public perception and attitudes towards transportation trends and DEIA issues. However, such data may not be cost-free and publicly accessible.
- Incorporating additional demographic parameters: Testing the inclusion of additional demographic parameters, such as age, alongside existing variables, can further enrich the analysis and provide a more nuanced understanding of the relationship between demographic factors and transportation DEIA.

- Extending geographic scope: Conducting similar analyses in different geographic areas can offer insights into regional variations and shed light on the diverse challenges faced by communities in different locations.
- Addressing bias: Recognizing that social media usage may not represent the entire population, future studies should consider implementing stratified sampling techniques to reduce bias and ensure a more representative sample for analysis.

By exploring these future directions, researchers can enhance the effectiveness and inclusivity of social media and network data analysis in examining transportation DEIA, leading to more robust findings and actionable insights for policymakers and planners.

Appendix

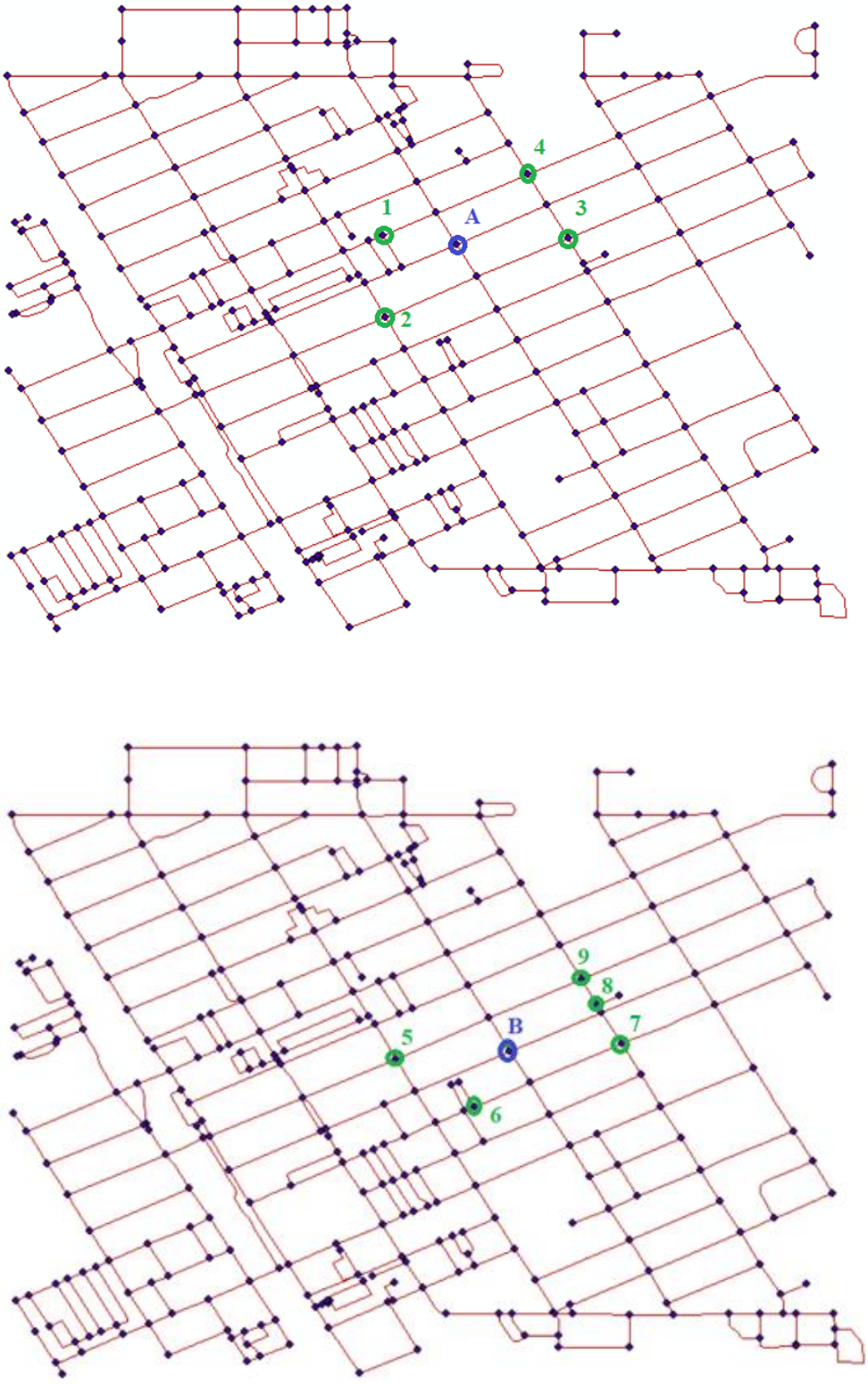

Figure 29: *Source nodes and target nodes for most central nodes intervention.*

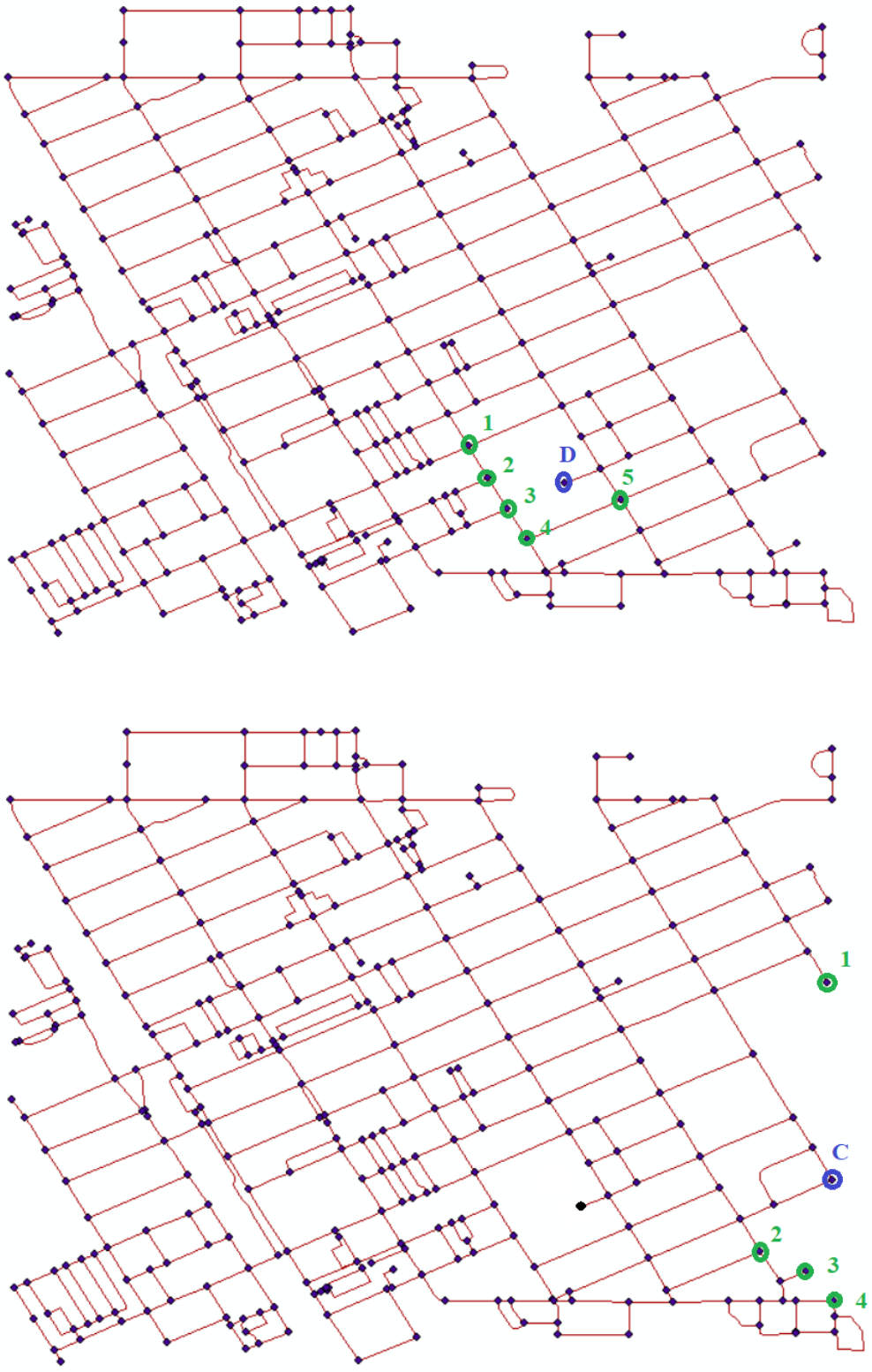

Figure 30: *Source nodes and target nodes for random nodes intervention.*

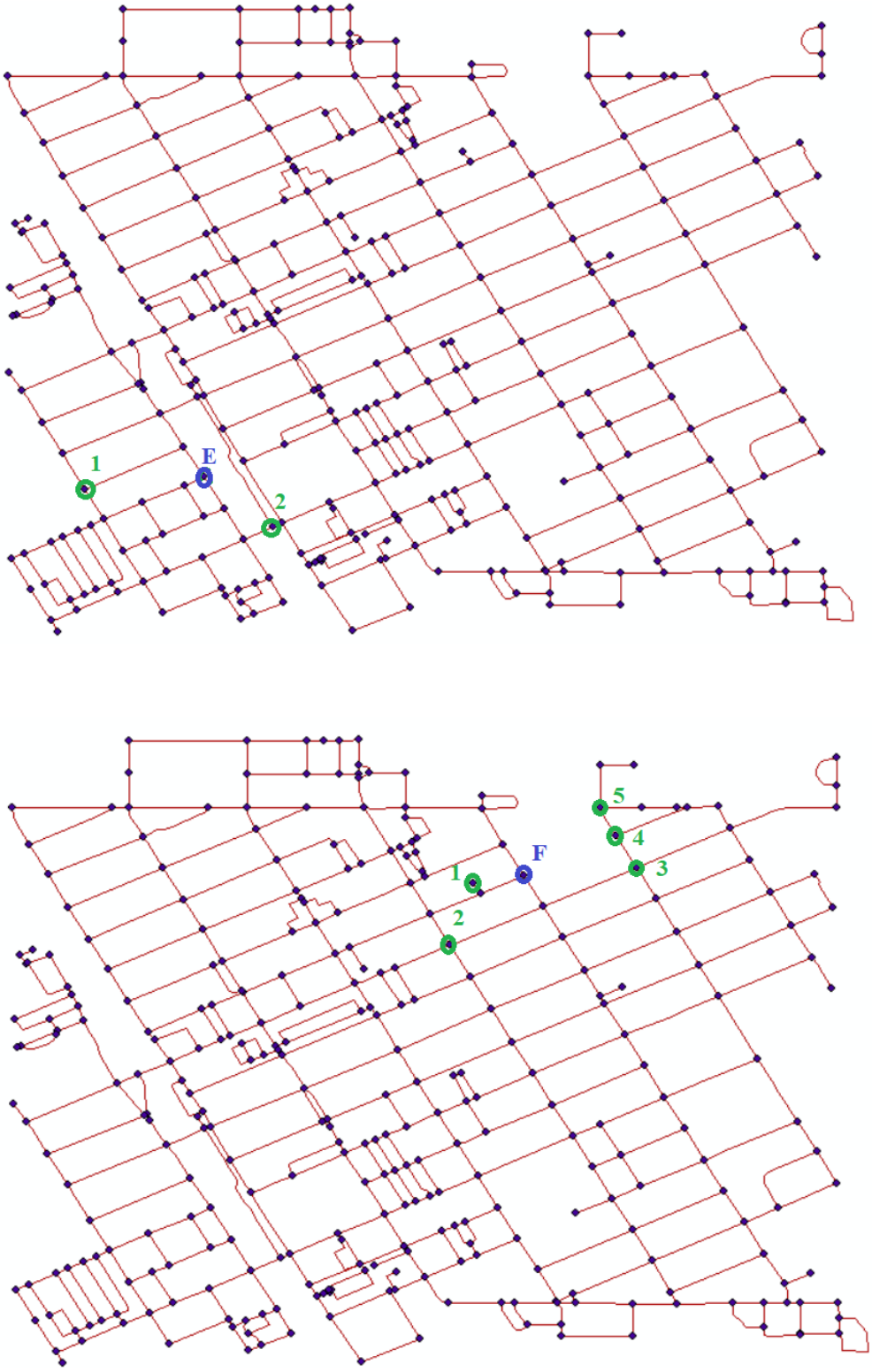

Figure 31: *Source nodes and target nodes for least central nodes intervention.*

References

1. Litman, T., *Evaluating transportation diversity*. 2017: Victoria Transport Policy Institute Victoria, BC, Canada.
2. Bullard, R.D., *Addressing urban transportation equity in the United States.* Fordham Urb. LJ, 2003. **31**: p. 1183.
3. Litman, T., *Evaluating transportation equity.* World Transport Policy & Practice, 2002. **8**(2): p. 50-65.
4. Welch, T.F., *Equity in transport: The distribution of transit access and connectivity among affordable housing units.* Transport policy, 2013. **30**: p. 283-293.
5. Litman, T., *Social inclusion as a transport planning issue in Canada.* 2003.
6. Lucas, K. and A. Musso, *Policies for social inclusion in transportation: An introduction to the special issue.* Case Studies on Transport Policy, 2014. **2**(2): p. 37-40.
7. Mittal, S., T. Yabe, F. Arroyo Arroyo, and S. Ukkusuri, *Linking Poverty-Based Inequalities with Transportation and Accessibility Using Mobility Data: A Case Study of Greater Maputo.* Transportation Research Record, 2023. **2677**(3): p. 668-682.
8. Boisjoly, G. and G.T. Yengoh, *Opening the door to social equity: local and participatory approaches to transportation planning in Montreal.* European transport research review, 2017. **9**(3): p. 1-21.
9. Preston, J. and F. Rajé, *Accessibility, mobility and transport-related social exclusion.* Journal of transport geography, 2007. **15**(3): p. 151-160.
10. Garrett, M. and B. Taylor, *Reconsidering social equity in public transit.* Berkeley Planning Journal, 1999. **13**(1).
11. Dingil, A.E., F. Rupi, and Z. Stasiskiene, *A macroscopic analysis of transport networks: The influence of network design on urban transportation performance.* International Journal of Transport Development and Integration, 2019. **3**(4): p. 331-343.
12. Dingil, A.E., J. Schweizer, F. Rupi, and Z. Stasiskiene, *Updated models of passenger transport related energy consumption of urban areas.* Sustainability, 2019. **11**(15): p. 4060.
13. Ingram, G.K. and Z. Liu, *Determinants of motorization and road provision.* Available at SSRN 569257, 1999.
14. Melo, P.C., D.J. Graham, and S. Canavan, *Effects of road investments on economic output and induced travel demand: evidence for urbanized areas in the United States.* Transportation research record, 2012. **2297**(1): p. 163-171.
15. Litman, T., *Evaluating transportation equity*. 2017: Victoria Transport Policy Institute.
16. Williams, K.M., J.H. Kramer, Y. Keita, L.D. Enomah, and T. Boyd, *Integrating Equity into MPO Project Prioritization*. 2019, Center for Transportation, Equity, Decisions and Dollars (CTEDD)(UTC).
17. Litman, T., *Evaluating transportation economic development impacts*. 2017: Victoria Transport Policy Institute Victoria, BC, Canada.
18. HOUSE, T.W. *Executive Order On Advancing Racial Equity and Support for Underserved Communities Through the Federal Government*. 2021; Available from: https://www.whitehouse.gov/briefing-room/presidential-actions/2021/01/20/executive-order-advancing-racial-equity-and-support-for-underserved-communities-through-the-federal-government/.
19. (USDOT), U.S.D.o.T. *Justice40*. 2021; Available from: https://www.transportation.gov/equity-Justice40.
20. (USDOT), U.S.D.o.T., *Reconnecting Communities Pilot (RCP) Program.* 2022.
21. (USDOT), U.S.D.o.T. *RAISE Discretionary Grants*. 2023; Available from:

https://www.transportation.gov/RAISEgrants.
22. Behbahani, H., S. Nazari, M.J. Kang, and T. Litman, *A conceptual framework to formulate transportation network design problem considering social equity criteria.* Transportation research part A: policy and practice, 2019. **125**: p. 171-183.
23. Blumenberg, E., *Social equity and urban transportation.* The geography of urban transportation, 2017. **332**.
24. Boisjoly, G. and A.M. El-Geneidy, *How to get there? A critical assessment of accessibility objectives and indicators in metropolitan transportation plans.* Transport Policy, 2017. **55**: p. 38-50.
25. Karner, A., *Planning for transportation equity in small regions: Towards meaningful performance assessment.* Transport policy, 2016. **52**: p. 46-54.
26. Lee, R.J., I.N. Sener, and S.N. Jones, *Understanding the role of equity in active transportation planning in the United States.* Transport reviews, 2017. **37**(2): p. 211-226.
27. Levine, J., *Urban transportation and social equity: Transportation-planning paradigms that impede policy reform*, in *Policy, Planning, and People*. 2013, University of Pennsylvania Press. p. 141-160.
28. Litman, T. and M. Brenman, *A new social equity agenda for sustainable transportation*. 2012: Victoria Transport Policy Institute VictoriaCanadá.
29. Managh, K., M.G. Badami, and A.M. El-Geneidy, *Integrating social equity into urban transportation planning: A critical evaluation of equity objectives and measures in transportation plans in North America.* Transport policy, 2015. **37**: p. 167-176.
30. Managh, K. and A. El-Geneidy, *Who benefits from new transportation infrastructure? Using accessibility measures to evaluate social equity in public transport provision*, in *Accessibility Analysis and Transport Planning*. 2012, Edward Elgar Publishing.
31. El-Geneidy, A., D. Levinson, E. Diab, G. Boisjoly, D. Verbich, and C. Loong, *The cost of equity: Assessing transit accessibility and social disparity using total travel cost.* Transportation Research Part A: Policy and Practice, 2016. **91**: p. 302-316.
32. Kristine M. Williams, A.J.K., AICP; Yaye Keita, PhD; Tia Boyd, *TRANSPORTATION EQUITY SCORECARD –A TOOL FOR PROJECT SCREENING AND PRIORITIZATION*. 2020.
33. Boeing, G., *The morphology and circuity of walkable and drivable street networks*. 2019: Springer.
34. Tang, J., W. Bi, F. Liu, and W. Zhang, *Exploring urban travel patterns using density-based clustering with multi-attributes from large-scaled vehicle trajectories.* Physica A: Statistical Mechanics and its Applications, 2021. **561**: p. 125301.
35. Parthasarathi, P., D. Levinson, and H. Hochmair, *Network structure and travel time perception.* PloS one, 2013. **8**(10): p. e77718.
36. Ceder, A., *Public transit planning and operation: Modeling, practice and behavior*. 2016: CRC press.
37. Farahani, R.Z., E. Miandoabchi, W.Y. Szeto, and H. Rashidi, *A review of urban transportation network design problems.* European journal of operational research, 2013. **229**(2): p. 281-302.
38. Guihaire, V. and J.-K. Hao, *Transit network design and scheduling: A global review.* Transportation Research Part A: Policy and Practice, 2008. **42**(10): p. 1251-1273.
39. Goldberg, A.V. and C. Harrelson. *Computing the shortest path: A search meets graph theory*. in *SODA*. 2005.
40. Bast, H., D. Delling, A. Goldberg, M. Müller-Hannemann, T. Pajor, P. Sanders, D. Wagner, and R.F. Werneck, *Route planning in transportation networks.* Algorithm engineering: Selected results and surveys, 2016: p. 19-80.

41. Luxen, D. and C. Vetter. *Real-time routing with OpenStreetMap data*. in *Proceedings of the 19th ACM SIGSPATIAL international conference on advances in geographic information systems*. 2011.
42. Klosterman, R.E. and A.A. Lew, *TIGER products for planning.* Journal of the American Planning Association, 1992. **58**(3): p. 379-385.
43. Carr, L.J., S.I. Dunsiger, and B.H. Marcus, *Walk score™ as a global estimate of neighborhood walkability.* American journal of preventive medicine, 2010. **39**(5): p. 460-463.
44. Mooney, P. and M. Minghini, *A review of OpenStreetMap data.* Mapping and the citizen sensor, 2017: p. 37-59.
45. Haklay, M. and P. Weber, *Openstreetmap: User-generated street maps.* IEEE Pervasive computing, 2008. **7**(4): p. 12-18.
46. Bennett, J., *OpenStreetMap*. 2010: Packt Publishing Ltd.
47. Olbricht, R.M., *Data retrieval for small spatial regions in OpenStreetMap.* OpenStreetMap in GIScience: Experiences, Research, and Applications, 2015: p. 101-122.
48. Boeing, G., *OSMnx: A Python package to work with graph-theoretic OpenStreetMap street networks.* Journal of Open Source Software, 2017. **2**(12).
49. Boeing, G., *A multi-scale analysis of 27,000 urban street networks: Every US city, town, urbanized area, and Zillow neighborhood.* Environment and Planning B: Urban Analytics and City Science, 2020. **47**(4): p. 590-608.
50. Basiri, A., M. Jackson, P. Amirian, A. Pourabdollah, M. Sester, A. Winstanley, T. Moore, and L. Zhang, *Quality assessment of OpenStreetMap data using trajectory mining.* Geo-spatial information science, 2016. **19**(1): p. 56-68.
51. Riedel, T. and U. Brunner, *Traffic control using graph theory*. Control Engineering Practice, 1994. **2**(3): p. 397-404.
52. Derrible, S. and C. Kennedy, *Applications of graph theory and network science to transit network design.* Transport reviews, 2011. **31**(4): p. 495-519.
53. Wang, S., D. Yu, M.-P. Kwan, L. Zheng, H. Miao, and Y. Li, *The impacts of road network density on motor vehicle travel: An empirical study of Chinese cities based on network theory.* Transportation research part A: policy and practice, 2020. **132**: p. 144-156.
54. Ding, R., N. Ujang, H.B. Hamid, M.S.A. Manan, R. Li, S.S.M. Albadareen, A. Nochian, and J. Wu, *Application of complex networks theory in urban traffic network researches.* Networks and Spatial Economics, 2019. **19**: p. 1281-1317.
55. Dunn, S. and S.M. Wilkinson, *Increasing the resilience of air traffic networks using a network graph theory approach.* Transportation Research Part E: Logistics and Transportation Review, 2016. **90**: p. 39-50.
56. Aydin, N.Y., H.S. Duzgun, F. Wenzel, and H.R. Heinimann, *Integration of stress testing with graph theory to assess the resilience of urban road networks under seismic hazards.* Natural Hazards, 2018. **91**: p. 37-68.
57. Meyer, N.K., W. Schwanghart, O. Korup, and F. Nadim, *Roads at risk: traffic detours from debris flows in southern Norway.* Natural hazards and earth system sciences, 2015. **15**(5): p. 985-995.
58. Rousset, L. and C. Ducruet, *Disruptions in spatial networks: a comparative study of major shocks affecting ports and shipping patterns.* Networks and Spatial Economics, 2020. **20**(2): p. 423-447.
59. D'Acci, L. and M. Batty, *The mathematics of urban morphology*. 2019: Springer.
60. Ducruet, C. and L. Beauguitte, *Spatial science and network science: review and outcomes of a complex relationship.* Networks and Spatial Economics, 2014. **14**(3-4):

p. 297-316.
61. Uitermark, J. and M. Van Meeteren, *Geographical network analysis.* Tijdschrift voor economische en sociale geografie, 2021. **112**(4): p. 337-350.
62. Kermanshah, A. and S. Derrible, *Robustness of road systems to extreme flooding: using elements of GIS, travel demand, and network science.* Natural hazards, 2017. **86**: p. 151-164.
63. Batty, M., *Network geography: Relations, interactions, scaling and spatial processes in GIS.* Re-presenting GIS, 2005: p. 149-170.
64. Lazer, D., A.S. Pentland, L. Adamic, S. Aral, A.L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, and M. Gutmann, *Life in the network: the coming age of computational social science.* Science (New York, NY), 2009. **323**(5915): p. 721.
65. Guy, M., P. Earle, C. Ostrum, K. Gruchalla, and S. Horvath. *Integration and dissemination of citizen reported and seismically derived earthquake information via social network technologies*. in *International Symposium on Intelligent Data Analysis*. 2010. Springer.
66. Li, J. and H.R. Rao, *Twitter as a rapid response news service: An exploration in the context of the 2008 China earthquake.* The Electronic Journal of Information Systems in Developing Countries, 2010. **42**.
67. Kogan, M., L. Palen, and K.M. Anderson. *Think local, retweet global: Retweeting by the geographically-vulnerable during Hurricane Sandy*. in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. 2015. ACM.
68. Starbird, K. and L. Palen, *Pass it on?: Retweeting in mass emergency*. 2010: International Community on Information Systems for Crisis Response and Management.
69. Caragea, C., N. McNeese, A. Jaiswal, G. Traylor, H.-W. Kim, P. Mitra, D. Wu, A.H. Tapia, L. Giles, and B.J. Jansen. *Classifying text messages for the haiti earthquake*. in *Proceedings of the 8th international conference on information systems for crisis response and management (ISCRAM2011)*. 2011. Citeseer.
70. Earle, P.S., D.C. Bowden, and M. Guy, *Twitter earthquake detection: earthquake monitoring in a social world.* Annals of Geophysics, 2012. **54**(6).
71. Imran, M., S.M. Elbassuoni, C. Castillo, F. Diaz, and P. Meier, *Extracting information nuggets from disaster-related messages in social media.* Proc. of ISCRAM, Baden-Baden, Germany, 2013.
72. Kumar, S., X. Hu, and H. Liu. *A behavior analytics approach to identifying tweets from crisis regions*. in *Proceedings of the 25th ACM conference on Hypertext and social media*. 2014. ACM.
73. Sakaki, T., M. Okazaki, and Y. Matsuo. *Earthquake shakes twitter users: real-time event detection by social sensors*. in *Proceedings of the 19th international conference on World wide web*. 2010.
74. Power, R., B. Robinson, J. Colton, and M. Cameron. *Emergency situation awareness: Twitter case studies*. in *International Conference on Information Systems for Crisis Response and Management in Mediterranean Countries*. 2014. Springer.
75. Vieweg, S., A.L. Hughes, K. Starbird, and L. Palen. *Microblogging during two natural hazards events: what twitter may contribute to situational awareness*. in *Proceedings of the SIGCHI conference on human factors in computing systems*. 2010. ACM.
76. Hughes, A.L., L.A. St Denis, L. Palen, and K.M. Anderson. *Online public communications by police & fire services during the 2012 Hurricane Sandy*. in *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. 2014. ACM.

77. St Denis, L.A., L. Palen, and K.M. Anderson. *Mastering social media: An analysis of Jefferson county's communications during the 2013 Colorado floods*. in *11th International ISCRAM Conference*. 2014.
78. Wang, Q. and J.E. Taylor, *Quantifying human mobility perturbation and resilience in Hurricane Sandy.* PLoS one, 2014. **9**(11): p. e112608.
79. Wang, Q. and J.E. Taylor, *Resilience of human mobility under the influence of typhoons.* Procedia Engineering, 2015. **118**: p. 942-949.
80. Kryvasheyeu, Y., H. Chen, E. Moro, P. Van Hentenryck, and M. Cebrian, *Performance of social network sensors during Hurricane Sandy.* PLoS one, 2015. **10**(2): p. e0117288.
81. Gao, H., G. Barbier, R. Goolsby, and D. Zeng, *Harnessing the crowdsourcing power of social media for disaster relief*. 2011, DTIC Document.
82. Hasan, S., X. Zhan, and S.V. Ukkusuri. *Understanding urban human activity and mobility patterns using large-scale location-based data from online social media*. in *Proceedings of the 2nd ACM SIGKDD international workshop on urban computing*. 2013. ACM.
83. Hasan, S. and S.V. Ukkusuri, *Urban activity pattern classification using topic models from online geo-location data.* Transportation Research Part C: Emerging Technologies, 2014. **44**: p. 363-381.
84. Hasan, S. and S.V. Ukkusuri, *Location contexts of user check-ins to model urban geo life-style patterns.* PloS one, 2015. **10**(5): p. e0124819.
85. Lee, J.H., A. Davis, and K. Goulias. *Activity Space Estimation with Longitudinal Observations of Social Media Data*. in *Paper submitted for presentation at the 95th Annual Meeting of the Transportation Research Board. Washington, DC*. 2016.
86. Zhao, S. and K. Zhang. *Observing Individual Dynamic Choices of Activity Chains From Location-Based Crowdsourced Data*. in *Transportation Research Board 95th Annual Meeting*. 2016.
87. Cebelak, M.K., *Location-based social networking data: doubly-constrained gravity model origin-destination estimation of the urban travel demand for Austin, TX*. 2013.
88. Chen, Y. and H.S. Mahmassani. *Exploring Activity and Destination Choice Behavior in Two Metropolitan Areas Using Social Networking Data*. in *Transportation Research Board 95th Annual Meeting*. 2016.
89. Jin, P., M. Cebelak, F. Yang, J. Zhang, C. Walton, and B. Ran, *Location-Based Social Networking Data: Exploration into Use of Doubly Constrained Gravity Model for Origin-Destination Estimation.* Transportation Research Record: Journal of the Transportation Research Board, 2014(2430): p. 72-82.
90. Lee, J.H., S. Gao, and K.G. Goulias. *Comparing the Origin-Destination Matrices from Travel Demand Model and Social Media Data*. in *Transportation Research Board 95th Annual Meeting*. 2016.
91. Yang, F., P.J. Jin, X. Wan, R. Li, and B. Ran. *Dynamic origin-destination travel demand estimation using location based social networking data*. in *Transportation Research Board 93rd Annual Meeting*. 2014.
92. Hasan, S., S.V. Ukkusuri, and X. Zhan, *Understanding social influence in activity-location choice and life-style patterns using geo-location data from social media.* Frontiers in ICT, 2016. **3**: p. 10.
93. Collins, C., S. Hasan, and S.V. Ukkusuri, *A novel transit rider satisfaction metric: Rider sentiments measured from online social media data.* Journal of Public Transportation, 2013. **16**(2): p. 2.
94. Abbasi, A., T.H. Rashidi, M. Maghrebi, and S.T. Waller. *Utilising Location Based Social Media in Travel Survey Methods: bringing Twitter data into the play*. in

Proceedings of the 8th ACM SIGSPATIAL International Workshop on Location-Based Social Networks. 2015. ACM.
95. Maghrebi, M., A. Abbasi, T.H. Rashidi, and S.T. Waller. *Complementing Travel Diary Surveys with Twitter Data: Application of Text Mining Techniques on Activity Location, Type and Time*. in *2015 IEEE 18th International Conference on Intelligent Transportation Systems*. 2015. IEEE.
96. Ukkusuri, S., X. Zhan, A.M. Sadri, and Q. Ye, *Use of social media data to explore crisis informatics: Study of 2013 Oklahoma tornado.* Transportation Research Record: Journal of the Transportation Research Board, 2014(2459): p. 110-118.
97. Ahmad, M.A., C. Eckert, and A. Teredesai. *Interpretable machine learning in healthcare*. in *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*. 2018.
98. Shailaja, K., B. Seetharamulu, and M. Jabbar. *Machine learning in healthcare: A review*. in *2018 Second international conference on electronics, communication and aerospace technology (ICECA)*. 2018. IEEE.
99. Yuval, J. and P.A. O'Gorman, *Stable machine-learning parameterization of subgrid processes for climate modeling at a range of resolutions.* Nature communications, 2020. **11**(1): p. 3295.
100. Bochenek, B. and Z. Ustrnul, *Machine learning in weather prediction and climate analyses—applications and perspectives.* Atmosphere, 2022. **13**(2): p. 180.
101. Dixon, M.F., I. Halperin, and P. Bilokon, *Machine learning in finance*. Vol. 1170. 2020: Springer.
102. Culkin, R. and S.R. Das, *Machine learning in finance: the case of deep learning for option pricing.* Journal of Investment Management, 2017. **15**(4): p. 92-100.
103. Renault, T., *Sentiment analysis and machine learning in finance: a comparison of methods and models on one million messages.* Digital Finance, 2020. **2**(1-2): p. 1-13.
104. Wäldchen, J. and P. Mäder, *Machine learning for image based species identification.* Methods in Ecology and Evolution, 2018. **9**(11): p. 2216-2225.
105. Decenciere, E., G. Cazuguel, X. Zhang, G. Thibault, J.-C. Klein, F. Meyer, B. Marcotegui, G. Quellec, M. Lamard, and R. Danno, *TeleOphta: Machine learning and image processing methods for teleophthalmology.* Irbm, 2013. **34**(2): p. 196-203.
106. Jean, N., M. Burke, M. Xie, W.M. Davis, D.B. Lobell, and S. Ermon, *Combining satellite imagery and machine learning to predict poverty.* Science, 2016. **353**(6301): p. 790-794.
107. Wieland, M. and M. Pittore, *Performance evaluation of machine learning algorithms for urban pattern recognition from multi-spectral satellite images.* Remote Sensing, 2014. **6**(4): p. 2912-2939.
108. Hafezi, M.H., L. Liu, and H. Millward, *A time-use activity-pattern recognition model for activity-based travel demand modeling.* Transportation, 2019. **46**: p. 1369-1394.
109. Koushik, A.N., M. Manoj, and N. Nezamuddin, *Machine learning applications in activity-travel behaviour research: a review.* Transport reviews, 2020. **40**(3): p. 288-311.
110. Hagenauer, J. and M. Helbich, *A comparative study of machine learning classifiers for modeling travel mode choice.* Expert Systems with Applications, 2017. **78**: p. 273-282.
111. Pacheco, F., E. Exposito, M. Gineste, C. Baudoin, and J. Aguilar, *Towards the deployment of machine learning solutions in network traffic classification: A systematic survey.* IEEE Communications Surveys & Tutorials, 2018. **21**(2): p. 1988-2014.
112. Pineda-Jaramillo, J.D., *A review of Machine Learning (ML) algorithms used for modeling travel mode choice.* Dyna, 2019. **86**(211): p. 32-41.
113. Zantalis, F., G. Koulouras, S. Karabetsos, and D. Kandris, *A review of machine learning*

and IoT in smart transportation. Future Internet, 2019. **11**(4): p. 94.
114. Boukerche, A. and J. Wang, *Machine learning-based traffic prediction models for intelligent transportation systems.* Computer Networks, 2020. **181**: p. 107530.
115. Socher, R., Y. Bengio, and C.D. Manning, *Deep learning for NLP (without magic)*, in *Tutorial Abstracts of ACL 2012*. 2012. p. 5-5.
116. Jones, K.S., *What is the role of NLP in text retrieval?*, in *Natural language information retrieval*. 1999, Springer. p. 1-24.
117. Srinivasa-Desikan, B., *Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras*. 2018: Packt Publishing Ltd.
118. Groh, G. and J. Hauffa. *Characterizing social relations via nlp-based sentiment analysis*. in *Proceedings of the International AAAI Conference on Web and Social Media*. 2011.
119. Nasukawa, T. and J. Yi. *Sentiment analysis: Capturing favorability using natural language processing*. in *Proceedings of the 2nd international conference on Knowledge capture*. 2003.
120. Kao, A. and S.R. Poteet, *Natural language processing and text mining*. 2007: Springer Science & Business Media.
121. Kumar, S., A.K. Kar, and P.V. Ilavarasan, *Applications of text mining in services management: A systematic literature review.* International Journal of Information Management Data Insights, 2021. **1**(1): p. 100008.
122. Ayo, F.E., O. Folorunso, F.T. Ibharalu, and I.A. Osinuga, *Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions.* Computer Science Review, 2020. **38**: p. 100311.
123. Gautam, G. and D. Yadav. *Sentiment analysis of twitter data using machine learning approaches and semantic analysis*. in *2014 Seventh international conference on contemporary computing (IC3)*. 2014. IEEE.
124. Sahayak, V., V. Shete, and A. Pathan, *Sentiment analysis on twitter data.* International Journal of Innovative Research in Advanced Engineering (IJIRAE), 2015. **2**(1): p. 178-183.
125. Allen, C., M.-H. Tsou, A. Aslam, A. Nagel, and J.-M. Gawron, *Applying GIS and machine learning methods to Twitter data for multiscale surveillance of influenza.* PloS one, 2016. **11**(7): p. e0157734.
126. Mahor, V., R. Rawat, S. Telang, B. Garg, D. Mukhopadhyay, and P. Palimkar. *Machine learning based detection of cyber crime hub analysis using twitter data*. in *2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON)*. 2021. IEEE.
127. Corvey, W.J., S. Vieweg, T. Rood, and M. Palmer. *Twitter in mass emergency: What nlp can contribute*. in *Proceedings of the NAACL HLT 2010 workshop on computational linguistics in a world of social media*. 2010.
128. Reynard, D. and M. Shirgaokar, *Harnessing the power of machine learning: Can Twitter data be useful in guiding resource allocation decisions during a natural disaster?* Transportation research part D: Transport and environment, 2019. **77**: p. 449-463.
129. Deo, N., *Graph theory with applications to engineering and computer science*. 2017: Courier Dover Publications.
130. Pardalos, P.M., F. Rendl, and H. Wolkowicz, *The quadratic assignment problem: A survey and recent developments.* 1994.
131. Judge, T.R. and P. Bryanston-Cross, *A review of phase unwrapping techniques in fringe analysis.* Optics and Lasers in Engineering, 1994. **21**(4): p. 199-239.

132. Lewenstein, M., A. Sanpera, and V. Ahufinger, *Ultracold Atoms in Optical Lattices: Simulating quantum many-body systems*. 2012: OUP Oxford.
133. Kou, Y., S. Lin, and M.P. Fossorier, *Low-density parity-check codes based on finite geometries: a rediscovery and new results.* IEEE Transactions on Information theory, 2001. **47**(7): p. 2711-2736.
134. Nikravesh, P.E., *Computer-aided analysis of mechanical systems*. 1988: Prentice-Hall, Inc.
135. Derrible, S. and C. Kennedy, *Network analysis of world subway systems using updated graph theory.* Transportation Research Record, 2009. **2112**(1): p. 17-25.
136. Erath, A., M. Löchl, and K.W. Axhausen, *Graph-theoretical analysis of the Swiss road and railway networks over time.* Networks and Spatial Economics, 2009. **9**: p. 379-400.
137. Martínez-López, B., A. Perez, and J. Sánchez-Vizcaíno, *Social network analysis. Review of general concepts and use in preventive veterinary medicine.* Transboundary and emerging diseases, 2009. **56**(4): p. 109-120.
138. Chakraborty, A., T. Dutta, S. Mondal, and A. Nath, *Application of graph theory in social media.* International Journal of Computer Sciences and Engineering, 2018. **6**(10): p. 722-729.
139. Takac, L. and M. Zabovsky. *Data analysis in public social networks*. in *International scientific conference and international workshop present day trends of innovations*. 2012.
140. Mason, O. and M. Verwoerd, *Graph theory and networks in biology.* IET systems biology, 2007. **1**(2): p. 89-119.
141. Pavlopoulos, G.A., M. Secrier, C.N. Moschopoulos, T.G. Soldatos, S. Kossida, J. Aerts, R. Schneider, and P.G. Bagos, *Using graph theory to analyze biological networks.* BioData mining, 2011. **4**: p. 1-27.
142. Gao, W., H. Wu, M.K. Siddiqui, and A.Q. Baig, *Study of biological networks using graph theory.* Saudi journal of biological sciences, 2018. **25**(6): p. 1212-1219.
143. Koutrouli, M., E. Karatzas, D. Paez-Espino, and G.A. Pavlopoulos, *A guide to conquer the biological network era using graph theory.* Frontiers in bioengineering and biotechnology, 2020. **8**: p. 34.
144. Sporns, O., *Graph theory methods: applications in brain networks.* Dialogues in Clinical Neuroscience, 2018. **20**(2): p. 111-121.
145. Kitsak, M., A. Ganin, A. Elmokashfi, H. Cui, D.A. Eisenberg, D.L. Alderson, D. Korkin, and I. Linkov, *Finding shortest and nearly shortest path nodes in large substantially incomplete networks by hyperbolic mapping.* Nature Communications, 2023. **14**(1): p. 186.
146. Hong, I., M. Kuby, and A.T. Murray, *A range-restricted recharging station coverage model for drone delivery service planning.* Transportation Research Part C: Emerging Technologies, 2018. **90**: p. 198-212.
147. Kröger, M., *Shortest multiple disconnected path for the analysis of entanglements in two- and three-dimensional polymeric systems.* Computer Physics Communications, 2005. **168**(3): p. 209-232.
148. Schwikowski, B., P. Uetz, and S. Fields, *A network of protein–protein interactions in yeast.* Nature biotechnology, 2000. **18**(12): p. 1257-1261.
149. Walkey, C.D. and W.C. Chan, *Understanding and controlling the interaction of nanomaterials with proteins in a physiological environment.* Chemical Society Reviews, 2012. **41**(7): p. 2780-2799.
150. Martínez, V., F. Berzal, and J.-C. Cubero, *A survey of link prediction in complex networks.* ACM computing surveys (CSUR), 2016. **49**(4): p. 1-33.
151. Ito, T., T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, *A comprehensive*

two-hybrid analysis to explore the yeast protein interactome. Proceedings of the National Academy of Sciences, 2001. **98**(8): p. 4569-4574.
152. Ideker, T., V. Thorsson, J.A. Ranish, R. Christmas, J. Buhler, J.K. Eng, R. Bumgarner, D.R. Goodlett, R. Aebersold, and L. Hood, *Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network.* Science, 2001. **292**(5518): p. 929-934.
153. Broumi, S., A. Bakal, M. Talea, F. Smarandache, and L. Vladareanu. *Applying Dijkstra algorithm for solving neutrosophic shortest path problem*. in *2016 International conference on advanced mechatronic systems (ICAMechS)*. 2016. IEEE.
154. Ojekudo, N.A. and N.P. Akpan, *Anapplication of Dijkstra's Algorithm to shortest route problem.* IOSR Journal of Mathematics (IOSR-JM), 2017. **13**(3): p. 14.
155. Tirastittam, P. and P. Waiyawuththanapoom, *Public transport planning system by dijkstra algorithm: Case study bangkok metropolitan area.* International Journal of Computer and Information Engineering, 2014. **8**(1): p. 54-59.
156. Gbadamosi, O.A. and D.R. Aremu. *Design of a Modified Dijkstra's Algorithm for finding alternate routes for shortest-path problems with huge costs*. in *2020 International Conference in Mathematics, Computer Engineering and Computer Science (ICMCECS)*. 2020. IEEE.
157. Barthélemy, M., *Spatial networks.* Physics reports, 2011. **499**(1-3): p. 1-101.
158. Blanchard, P., *Mathematical analysis of urban spatial networks*. 2009: Springer.
159. Boeing, G., *Street network models and indicators for every urban area in the world.* Geographical analysis, 2022. **54**(3): p. 519-535.
160. Levinson, D. and A. El-Geneidy, *The minimum circuity frontier and the journey to work.* Regional science and urban economics, 2009. **39**(6): p. 732-738.
161. Huang, J. and D.M. Levinson, *Circuity in urban transit networks.* Journal of Transport Geography, 2015. **48**: p. 145-153.
162. Giacomin, D.J. and D.M. Levinson, *Road network circuity in metropolitan areas.* Environment and Planning B: Planning and Design, 2015. **42**(6): p. 1040-1053.
163. Boeing, G., *Urban spatial order: Street network orientation, configuration, and entropy.* Applied Network Science, 2019. **4**(1): p. 1-19.
164. Boeing, G., *Spatial information and the legibility of urban form: Big data in urban morphology.* International Journal of Information Management, 2021. **56**: p. 102013.
165. Bergmann, F.M., S.M. Wagner, and M. Winkenbach, *Integrating first-mile pickup and last-mile delivery on shared vehicle routes for efficient urban e-commerce distribution.* Transportation Research Part B: Methodological, 2020. **131**: p. 26-62.
166. Levinson, D., *Network structure and city size.* PloS one, 2012. **7**(1): p. e29721.
167. Dong, S., H. Wang, A. Mostafavi, and J. Gao, *Robust component: a robustness measure that incorporates access to critical facilities under disruptions.* Journal of the Royal Society Interface, 2019. **16**(157): p. 20190149.
168. Irwin, M.D. and H.L. Hughes, *Centrality and the structure of urban interaction: measures, concepts, and applications.* Social Forces, 1992. **71**(1): p. 17-51.
169. Fleming, D.K. and Y. Hayuth, *Spatial characteristics of transportation hubs: centrality and intermediacy.* Journal of Transport Geography, 1994. **2**(1): p. 3-18.
170. Li, F., H. Jia, Q. Luo, Y. Li, and L. Yang, *Identification of critical links in a large-scale road network considering the traffic flow betweenness index.* PloS one, 2020. **15**(4): p. e0227474.
171. Tian, Y., X. Liu, Z. Li, S. Tang, C. Shang, and L. Wei, *Identification of critical links in urban road network considering cascading failures.* Mathematical Problems in Engineering, 2021. **2021**: p. 1-11.
172. Helderop, E. and T.H. Grubesic, *Flood evacuation and rescue: The identification of*

critical road segments using whole-landscape features. Transportation research interdisciplinary perspectives, 2019. **3**: p. 100022.
173. Vivek, S. and H. Conner, *Urban road network vulnerability and resilience to large-scale attacks.* Safety science, 2022. **147**: p. 105575.
174. Zhang, W. and N. Wang, *Resilience-based risk mitigation for road networks.* Structural Safety, 2016. **62**: p. 57-65.
175. Zhou, Y., J. Wang, and H. Yang, *Resilience of transportation systems: concepts and comprehensive review.* IEEE Transactions on Intelligent Transportation Systems, 2019. **20**(12): p. 4262-4276.
176. Boeing, G., *Urban street network analysis in a computational notebook.* arXiv preprint arXiv:2001.06505, 2020.
177. Boeing, G., *OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks.* Computers, Environment and Urban Systems, 2017. **65**: p. 126-139.
178. Imani, A.F., E.J. Miller, and S. Saxe, *Cycle accessibility and level of traffic stress: A case study of Toronto.* Journal of transport geography, 2019. **80**: p. 102496.
179. Mekuria, M.C., P.G. Furth, and H. Nixon, *Low-stress bicycling and network connectivity.* 2012.
180. Faghih Imani, A., E.J. Miller, and S. Saxe, *Cycle accessibility and level of traffic stress: A case study of Toronto.* Journal of Transport Geography, 2019. **80**: p. 102496.
181. Foth, N., K. Manaugh, and A.M. El-Geneidy, *Towards equitable transit: examining transit accessibility and social need in Toronto, Canada, 1996–2006.* Journal of Transport Geography, 2013. **29**: p. 1-10.
182. Miah Md, M., P. Mattingly Stephen, and K. Hyun Kate, *Evaluation of Bicycle Network Connectivity Using Graph Theory and Level of Traffic Stress.* Journal of Transportation Engineering, Part A: Systems, 2023. **149**(9): p. 04023080.
183. Owen, A. and D.M. Levinson, *Modeling the commute mode share of transit using continuous accessibility to jobs.* Transportation Research Part A: Policy and Practice, 2015. **74**: p. 110-122.
184. Levinson, D.M., *Accessibility and the journey to work.* Journal of Transport Geography, 1998. **6**(1): p. 11-21.
185. Van Acker, V. and F. Witlox, *Car ownership as a mediating variable in car travel behaviour research using a structural equation modelling approach to identify its dual relationship.* Journal of Transport Geography, 2010. **18**(1): p. 65-74.
186. Mokhtarian, P.L. and X. Cao, *Examining the impacts of residential self-selection on travel behavior: A focus on methodologies.* Transportation Research Part B: Methodological, 2008. **42**(3): p. 204-228.
187. Chen, C., H. Gong, and R. Paaswell, *Role of the built environment on mode choice decisions: additional evidence on the impact of density.* Transportation, 2008. **35**: p. 285-299.
188. De Vos, J., P.L. Mokhtarian, T. Schwanen, V. Van Acker, and F. Witlox, *Travel mode choice and travel satisfaction: bridging the gap between decision utility and experienced utility.* Transportation, 2016. **43**: p. 771-796.
189. Cao, X.J., P.L. Mokhtarian, and S.L. Handy, *The relationship between the built environment and nonwork travel: A case study of Northern California.* Transportation Research Part A: Policy and Practice, 2009. **43**(5): p. 548-559.
190. Torun, A.Ö., K. Göçer, D. Yeşiltepe, and G. Argın, *Understanding the role of urban form in explaining transportation and recreational walking among children in a logistic GWR model: A spatial analysis in Istanbul, Turkey.* Journal of Transport Geography, 2020. **82**: p. 102617.

191. Derrible, S., *Network centrality of metro systems.* PloS one, 2012. **7**(7): p. e40575.
192. Barthelemy, M., *The structure and dynamics of cities*. 2016: Cambridge University Press.
193. Özbil, A., D. Yeşiltepe, and G. Argın, *Modeling walkability: The effects of street design, street-network configuration and land-use on pedestrian movement.* A| Z ITU Journal of the Faculty of Architecture, 2015. **12**(3): p. 189-207.
194. Ozbil, A., T. Gurleyen, D. Yesiltepe, and E. Zunbuloglu, *Comparative associations of street network design, streetscape attributes and land-use characteristics on pedestrian flows in peripheral neighbourhoods.* International journal of environmental research and public health, 2019. **16**(10): p. 1846.
195. Kamruzzaman, M., D. Baker, S. Washington, and G. Turrell, *Advance transit oriented development typology: case study in Brisbane, Australia.* Journal of transport geography, 2014. **34**: p. 54-70.
196. Ma, X., J. Zhang, C. Ding, and Y. Wang, *A geographically and temporally weighted regression model to explore the spatiotemporal influence of built environment on transit ridership.* Computers, Environment and Urban Systems, 2018. **70**: p. 113-124.
197. Cottineau, C., E. Hatna, E. Arcaute, and M. Batty, *Diverse cities or the systematic paradox of urban scaling laws.* Computers, environment and urban systems, 2017. **63**: p. 80-94.
198. Molinero, C. and S. Thurner, *How the geometry of cities determines urban scaling laws.* Journal of the Royal Society interface, 2021. **18**(176): p. 20200705.
199. Benoit, K., *Linear regression models with logarithmic transformations.* London School of Economics, London, 2011. **22**(1): p. 23-36.
200. Chouchane, H., M.S. Krol, and A.Y. Hoekstra, *Virtual water trade patterns in relation to environmental and socioeconomic factors: A case study for Tunisia.* Science of the total environment, 2018. **613**: p. 287-297.
201. Genc, O., M. Bayrak, and E. Yaldiz, *Analysis of the effects of GSM bands to the electromagnetic pollution in the RF spectrum.* Progress in Electromagnetics Research, 2010. **101**: p. 17-32.
202. Khan, M.A., *An empirical assessment of service quality of cellular mobile telephone operators in Pakistan.* Asian Social Science, 2010. **6**(10): p. 164.
203. Knoke, T., *Predicting red heartwood formation in beech trees (Fagus sylvatica L.).* Ecological Modelling, 2003. **169**(2-3): p. 295-312.
204. Durbin, J. and G.S. Watson, *TESTING FOR SERIAL CORRELATION IN LEAST SQUARES REGRESSION. I.* Biometrika, 1950. **37**(3-4): p. 409-428.
205. DURBIN, J. and G.S. WATSON, *TESTING FOR SERIAL CORRELATION IN LEAST SQUARES REGRESSION. II.* Biometrika, 1951. **38**(1-2): p. 159-178.
206. Verbeek, M., *A guide to modern econometrics*. 2008: John Wiley & Sons.
207. Craney, T.A. and J.G. Surles, *Model-dependent variance inflation factor cutoff values.* Quality engineering, 2002. **14**(3): p. 391-403.
208. Statt, N. *Twitter is opening up its full tweet archive to academic researchers for free*. 2021 [cited 2022 25 july]; Available from: https://www.theverge.com/2021/1/26/22250203/twitter-academic-research-public-tweet-archive-free-access.
209. Tornes, A. *Product News: Enabling the future of academic research with the Twitter API*. 2021 [cited 2022 25 July]; Available from: https://developer.twitter.com/en/blog/product-news/2021/enabling-the-future-of-academic-research-with-the-twitter-api.
210. Blei, D.M., A.Y. Ng, and M.I. Jordan, *Latent dirichlet allocation.* Journal of machine Learning research, 2003. **3**(Jan): p. 993-1022.

211. Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding.* arXiv preprint arXiv:1810.04805, 2018.
212. Momin, K.A., *Race-Gender-Prediction*. 2022: Github.
213. McFadden, D., *The measurement of urban travel demand.* Journal of public economics, 1974. **3**(4): p. 303-328.
214. Ben-Akiva, M. and M. Bierlaire, *Discrete choice methods and their applications to short term travel decisions*, in *Handbook of transportation science*. 1999, Springer. p. 5-33.
215. Train, K.E., *Discrete choice methods with simulation*. 2009: Cambridge university press.
216. Chertow, G., S. Soroko, E. Paganini, K. Cho, J. Himmelfarb, T. Ikizler, R. Mehta, and P.t.I.C.i.A.R. Disease, *Mortality after acute renal failure: models for prognostic stratification and risk adjustment.* Kidney international, 2006. **70**(6): p. 1120-1126.
217. Lee, R.J., R.A. Madan, J. Kim, E.M. Posadas, and E.Y. Yu, *Disparities in cancer care and the Asian American population.* The oncologist, 2021. **26**(6): p. 453-460.
218. Hall, W.J., M.V. Chapman, K.M. Lee, Y.M. Merino, T.W. Thomas, B.K. Payne, E. Eng, S.H. Day, and T. Coyne-Beasley, *Implicit racial/ethnic bias among health care professionals and its influence on health care outcomes: a systematic review.* American journal of public health, 2015. **105**(12): p. e60-e76.