Quantifying Equity and Equity Biases in Transportation and Data

Samuel Ricord

A dissertation

submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:

Yinhai Wang, Chair

Don MacKenzie

Angela Kitali

Daniel Abramson

Mark Hallenbeck

Program Authorized to Offer Degree:

Civil and Environmental Engineering

University of Washington

Abstract

Quantifying Equity and Equity Biases in Transportation and Data

Samuel Ricord

Chair of the Supervising Committee:
Yinhai Wang, Professor
Department of Civil and Environmental Engineering

Equity is a critical field of study in transportation. The built transportation network does not serve the needs of the population to achieve equal levels of economic vitality and prosperity. Because of these concerns, there has been a recent effort to address these equity issues in the transportation network. This effort coincides with a massive growth in the data available for transportation practitioners, known as big data. The growth of data has led to data-driven decision-making to allow for more effective transportation policies and decisions than was afforded with classical methods. However, as the amount of data available has grown, there is a great concern for understanding the biases within this data. Though significant effort has been spent to mitigate the biases present in transportation datasets, there is little understanding of the equity implications of these biases. Since biases inherently misrepresent certain population segments, there is a possibility that critical populations will be underrepresented in our data sources and therefore our data-driven decision-making, which is directly counter to our goals of increasing equity in transportation research and practice. One of the key pillars in addressing equity in transportation is ensuring that the decision-making process includes rigorous representation of all impacted parties for ongoing transportation projects. Therefore, it is critical that representation also be

maintained when using data in data-driven decision processes. How do we identify and quantify representation for transportation datasets that influence decision making? To understand this issue, we can reframe our idea of data biases as equity biases: an equity bias is any bias in a data source that produces a negative equity outcome by underrepresenting critical populations. Here, critical populations can mean any population of interest, including communities of low income, communities with high poverty levels, or historically disadvantaged communities such as black, indigenous, and peoples of color (BIPOC) communities. This definition focuses on the equity outcomes of data biases as opposed to the precipitating causes of bias.

With this definition, this dissertation presents a methodological framework to address the key challenges relating to equity biases: how do we identify, quantify, and address equity biases and representation such that transportation datasets best serve in data-driven decision-making? This work addresses these issues by utilizing ecological regression to define the representation of datasets to accurately understand which populations are under- and over-represented in transportation datasets. Ecological regression is well suited for this task as, unlike other regression methods, it can address individual level characteristics (such as income, racial demographics, etc.) in aggregate datasets, allowing us to consider these critical equity demographics when assessing the representation of a transportation dataset. This allows us to define representation for different demographic strata, thus allowing for the comparison of representation between datasets. This framework is tested on datasets of tolling data collected from the five tolling facilities in Washington State.

Keywords: Data Equity, Data Bias, Representation

Acknowledgments

I have many people that I would like to acknowledge who helped me to both create this dissertation and conduct my studies throughout graduate school. First and foremost, I would like to thank my advisor, Professor Yinhai Wang, for supporting my research goals from my years as an undergraduate and throughout my graduate studies. I would also like to thank Mark Hallenbeck, who helped me find and work on projects that addressed my passion for transportation equity. Many of the findings from these projects were instrumental to the completion of this dissertation, and for that support I am thankful.

Of equal importance to the technical and academic support I have received is the support I have gotten from my friends and family. I am lucky to have so many people close to me to name outright, so for all those who have known me and supported me through my journey to this point, thank you. I specifically want to thank three people, my parents Patty and TR, and my wife Emily. You three have supported me more than you could know and for that I am and will forever be grateful.

I would also like to thank all of the lab mates and UW student and STAR Lab colleagues I have had through the years who have made working in the UW truly enjoyable. This includes, in no particular order, Yinsheng Kou, Hubert Chen, Ian Nisbet, Ray Huang, Cesar Maia De Souza, Dr. Ziyuan Pu, Dr. John Ash, Roberto Gomez, Dr. Ruimin Ke, Dr. Zhiyong Cui, Warkaa Almustafa, Iman Haji, Dr. Yifan Zhuang, Dr. Meixin Zhu, Jason Chen, Peter Yu, Adam Huang, Pranati Awasthi, Fengze Yang, Arthur Semionov, Yifan Ling, Shuyi Yin, Ollie Wiesner, Goo Jantarathaneewat, Bingzhang Wang, Dennis Tsai, Mehrdad Nasri, Frank Yang, Cole Kopca, and Chenxi Liu.

There have also been several colleagues outside the UW who have made my time and work throughout my Ph.D. extremely enjoyable. Most notable among these people is Hollyanna Littlebull and Dr. Wei Sun. You have made my work both interesting and enjoyable by bringing forward interesting research questions and always being a fun colleague to work with.

I also would be remiss if I did not mention the members of my dissertation committee. Two have already been mentioned above, Chair Yinhai Wang and Member Mark Hallenbeck. I would like to thank the other members of my committee as well, Professor Don MacKenzie, Professor Angela Kitali, and Professor Dan Abramson. I appreciate the time and effort you have put in reading through this tome to help support my graduation.

Finally, I would like to thank the organizations that primarily provided my funding throughout my graduate career. These are the Center for Safety Equity in Transportation (CSET), the Pacific Northwest Transportation Consortium (Pactrans), and the Washington State Department of Transportation (WSDOT). Without these organization's support, I would not have been able to conduct this research that fascinates me and would not have been able to graduate, so thank you.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Introduction to Equity Bias

Many communities across the United States are currently addressing major equity issues in transportation. Often times, the availability, quality, affordability, and safety of transportation options for different groups and communities can vary significantly (Sanchez 2004). There are many causes for these issues including historical policy and design decisions that either explicitly or inadvertently exacerbated divisions within communities. It is important to always strive to provide equity in any transportation project. Equity, in this sense, relates to ensuring that all users have an equal, positive outcome for their transportation needs. While having equity from all perspectives is difficult, this goal should be pursued to the greatest extent possible. This will significantly help communities increase their quality of living by providing safe and reasonable transportation options which will help stimulate economic opportunity for all community members, especially those historically underserved by the transportation system.

Currently in most transportation projects, equity is a primary goal for the engineers and planners working on implementing changes to the transportation network. It is understood that the goal of transportation projects should be to serve the entire community in a way that stimulates growth and vitality instead of only serving a subset of the community. This is directly related to the idea of transportation justice. Transportation justice refers to the removal of systemic barriers that prevent equal outcomes for the population (Fan 2019)(Karner 2020). Obviously, the focus on equity and transportation justice is positive because it ensures that those who have been historically underserved by the transportation network will have more options to get to destinations safely and

efficiently. In contrast, however, much less thought is put into understanding the equity implications of different data sources that play a fundamental role in our transportation decisions. The transportation sector is going through a period where the amount of data collected is growing exponentially, often called the era of big data. Many of these datasets show great promise for changing how practitioners understand the transportation system by providing a richness and depth of information that was previously unattainable. This has led to a growing use of data-driven decision-making, where design and policy decisions are made based on data as opposed to the limited understanding afforded by classical analytical methods. Overall, this trend is positive, as it allows for decision makers to have a more accurate, precise, and complete understanding of the transportation network which overall correlates with 'better' transportation decisions being made (Gamage 2016)(Höchtl 2016). However, there are some drawbacks related to the use of big data for data-driven decision-making. These datasets contain significant amounts of data biases that introduce errors in a dataset that overweigh or over represents parts of the dataset away from the true condition (Delgado-Rodriguez 2004).

One of the many ways biases can present themselves in data is by systemically misrepresenting certain segments of the population the dataset is trying to capture (Griffin 2020). Thus far, there is a gap in practice in the transportation field to connect these types of population misrepresentations to the real-world implications of these errors. Though there are many definitions of equity in the context of transportation, all revolve around two key concepts: first that equity must be inherently comparative such that the situations of different populations of interest are measured against one another, and second that addressing equity revolves around understanding the underlying socio-economic factors that roughly translate to livelihood and 'economic success' for community members in a fair way. In practice, this translates to finding

the segments of the population which are comparatively disadvantaged based on a wide variety of socio-economic factors (or transportation metrics that directly impact socio-economic factors) to distribute various resources more equitably to these communities. Since biases inherently misrepresent certain populations segments, there is the possibility that critical populations will be underrepresented in our data sources and therefore our data-driven decision-making, which is directly counter to our goals of increasing equity in transportation research and practice. In the typical study of data biases, the precipitating causes of the bias are explored to mitigate the biases to the greatest analytical extent. However, this method does not directly consider the equity implications of these biases. To understand this issue, we must reframe our idea of data biases as equity biases: an equity bias is any bias in a data source that produces a negative equity outcome by underrepresenting critical populations. Here, critical populations can mean any population of interest, including communities of low income, communities with high poverty levels, or historically disadvantaged communities such as black, indigenous, and peoples of color (BIPOC) communities. The critical difference between this idea and previous ideas about data bias is that equity biases primarily address primarily the equity outcomes of using data in data-driven decision-making instead of addressing preexisting flaws in the datasets.

1.2 Introduction to Ecological Regression

Ecological regression is a regression technique that incorporates individual level characteristics from aggregate level data (Jackson 2006). This method is also known as ecological inference. This characteristic is extremely powerful in this case because often, studying equity relies on the extraction of individual characteristics (such as income, race, etc.) from aggregated

transportation data. Other regression methods are not able to assess this information with statistical rigor as they rely on the assumption that all individuals in an aggregate group behave the same on all levels of secondary axes. This method originates in the fields of medicine and political science, where the ability to infer individual characteristics from aggregate data has been paramount (Gelman 2016). In the medical field, this was necessary due to the nature of clinical trials, which requires the anonymization and therefore aggregation of clinical data even though the prevalence of a disease is an inherently individual trait. Similarly, in the field of political science, this was used to understand the voting habits of individuals based upon aggregate, anonymous survey data. Though the application of ecological regression has rarely been applied in the field of transportation, it has been done successfully before, specifically relating to transportation equity. One example is a study conducted of variable toll lanes on the I-405 corridor in the Seattle area (Leung 2019). These toll lanes use a variable toll system which increases the price of the toll lane as congestion increases to maintain an optimal flow of vehicles through the toll lane. This study utilized ecological regression to understand the demographics of people using the toll lane and to connect those critical demographics to the distributed benefits of using the toll lanes. It was found that the toll lanes, in this case, were generally equitable. The demographic profile of those who use the toll lane was very similar to the demographic profile for the region, and it was found that people of lower income gained a greater value per trip through the toll lane than those of higher income (Leung 2019). It was found that people of higher income were more likely to use the toll lane regardless of roadways conditions and pay the fee even if their time savings were relatively small. In contrast, lower-income people were more likely to use the toll lane only when the value gained in time saved was highest. This study shows how ecological regression is useful for

analyzing aggregate data when looking for variables that are different based on individual characteristics.

1.3 Research Objective

This research targets the shortfalls found in addressing equity biases for transportation datasets. There is no methodology to quantify and mitigate this issue. This methodological framework will provide transportation practitioners with a way to inform their data-driven decision-making process to best understand biases and maintain equity. This addresses the key research question: how do we identify, quantify, and address equity biases in representation such that transportation datasets best serve in data-driven decision-making? From this question, we can describe several research objectives:

1) Conduct a thorough state-of-the-art and literature review to show the current state of equity practice, big data, and data biases to inform our understanding of equity biases and how to quantify and mitigate them.
2) Define a methodology to quantify and mitigate equity biases utilizing ecological regression.
3) Implement this methodology in a case study utilizing Washington State tolling data to show how it can be implemented in practice to minimize the impact that equity biases have on transportation.

1.4 Dissertation Organization

This dissertation defines a methodology to assess, quantify, and mitigate equity biases in transportation datasets. The organization of this dissertation is shown in Figure 1.1.

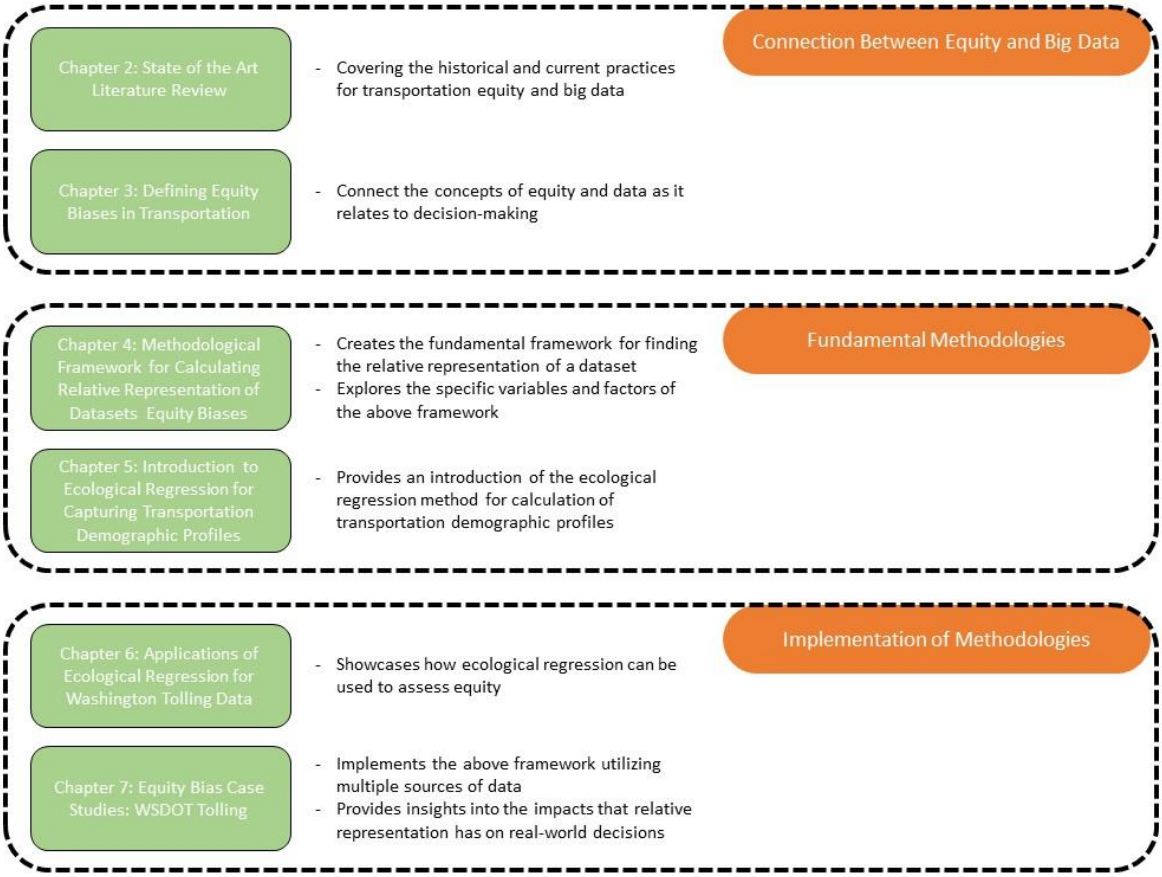

Figure 1.1: Dissertation Organization and Summary of Contributions

Chapter 2: State of the Art Literature Review. This chapter covers the current state of practice in the fields relating to this dissertation. This includes the areas of equity in transportation, equity study in transportation, and big data in transportation.

Chapter 3: Defining Equity Biases in Transportation. This chapter explores and defines in detail equity biases. This includes a thorough review of the current state of practice relating to general data biases as well as the different implications that equity biases have for transportation.

Chapter 4: Methodological Framework for Calculating Relative Representation of Datasets from Equity Biases. This chapter lays out the fundamental concepts of the methodological framework for calculating relative representation. This includes providing insight into the fundamental terms and key concepts needed to calculate relative representation.

Chapter 5: Introduction to Ecological Regression for Capturing Transportation Demographic Profiles. This chapter introduces the fundamental concepts of ecological regression and how it can be applied to determine the demographics of transportation networks. This includes an introduction of the statistical methods behind ecological regression, and exploration of how this can be implemented to calculate a demographic profile in transportation, and a case study highlighting this process in the Central Puget Sound Region.

Chapter 6: Applications of Ecological Regression for Washington Tolling Data. This chapter introduces many of the applications that ecological regression can be used for. This includes conducting equity analysis of toll lanes in the state of Washington through the method of ecological regression.

Chapter 7: Equity Bias Case Study: WSDOT (Washington State Department of Transportation) Tolling Data. This chapter details case studies that utilize the methodological framework created to quantify equity biases, as well as the real-world implications this has on decision making. The data includes WSDOT toll transaction data and Cambridge Systematics vehicle trace data.

Chapter 8: Conclusions, Challenges, and Future Directions. This chapter provides a summary of the findings and contributions of the dissertation, the critical conclusions drawn from this work, and points toward challenges and future research directions to enhance the products herein.

1.5 Dissertation Contributions

The research conducted in this distribution contributes to the field of transportation engineering in many ways. The key contributions are as follows:

1) A thorough understanding of the interactions between equity, big data, and data bias in the field of transportation.
2) A robust methodological framework to quantify and assess equity biases in transportation dataset.
3) An introduction to the method of ecological regression for various transportation equity applications including the definition of transportation demographics for understanding representation equity.
4) A model by which practitioners can understand the equity implications of the datasets being used to make data-informed decisions at all levels.

Chapter 2: State of the Art Literature Review

2.1 Overview

Equity in transportation is an increasingly important field of study for transportation professionals. The transportation system in the United States does not adequately serve the needs of the entire population. With the advent of new data sources, often described as big data, the field of transportation is changing rapidly. This has led to a growing trend to use data as a key consideration for transportation policy and decision making, labeled data-driven decision-making. With that change comes a growing emphasis on addressing equity in transportation. Though there has been a significant increase in understanding the equity implications of the transportation network, there is a significant gap in understanding the equity implications of the datasets that describe the transportation network. This literature gap has developed as the growth of data sources has occurred without a thorough understanding of the biases in these datasets that unequally describe the transportation network. These discrepancies can therefore lead directly to equity concerns that can be called equity biases. To understand the current practices related to equity bias, we first need to understand what role equity has to play in transportation, what role these new data sources play in transportation, and then find where these areas intersect to fully understand what equity implications these data sources have in relation to how this data is being used in the decision making process.

2.2 Contributions and Chapter Structure

In this chapter, we review relevant literature on the topics of equity and big data in transportation. This includes looking at various theoretical definitions of equity, the historical causes of inequity in transportation, exploring current practices relating to equity in transportation, and understanding how big data intersects with equity in transportation. The contribution of this chapter are as follows:

- The historical and current practices for transportation equity are explored such that an understanding of the precipitating causes and ongoing methods to address these causes are understood.
- Different types of data sources causing the growth of big data are explored to examine impact on equity.

The chapter is organized as follows. In section 2.3, we identify several different theories to define equity in transportation. In section 2.4, we describe the historic circumstances that cause the inequities we face today in transportation. In section 2.5, the current practices employed by transportation practitioners to address the equity concerns found in section 2.4 are covered. Section 2.6 covers a thorough review of big data in transportation, and how these new data sources impact equity. Finally, section 2.7 concludes the chapter with a summary and concluding remarks.

2.3 Defining Equity in Transportation

Equity in the transportation sector is one of the most pressing issues faced by communities across the United States, and indeed the world. Transportation is a critical resource that all people rely on for vitality and economic advancement, whether that be commuting to and from a place of work, traveling to and from educational institutions, or collecting food and other goods from stores

and markets to name a few of the many necessities that transportation facilitates. The availability, reliability, efficiency, and safety of transportation options can vary significantly between different groups throughout communities. Therefore, it is of utmost importance that transportation be accessible to all populations to achieve economic vitality and wellbeing. This concept of equity can be framed in many different ways. The basic concept of equity can be framed as "the distribution of benefits and costs over members of society" (Boucher and Kelly 1998)(Miller 1999). This idea has two key components: first that it relates to the impacts felt by individual members of a population or community and second that it is an inherently comparative process where these impacts are related between members of said population.

There are many different ways to define equity in transportation. There are many theories and concepts of how distributions of resources and outcomes are 'just', and therefore many fundamentally different, and sometimes conflicting, ways to begin to conceptualize equity (Lewis 2021). This idea is mirrored in a study by Martens and Golub (2021) which defines different normative standards for addressing equity in transportation into a four rung ladder according to Title VI of the Civil Rights Act. The rungs of the ladder in ascending order are as follows: explicit non-discrimination, Pareto-Plus Improvement, Proportional Equity, and Restorative Justice. Of these four rungs, the final two rungs are generally the concepts that the term 'equity' refers to in practice. Proportional equity refers to a distribution of resources such that the intended outcomes are equally shared among the entire population whereas restorative justice takes this a step further to also include ideas about also accounting for historically denied benefits. This ladder shows both that there are many different ways to interpret 'equity' and the differences between more general distributional definitions of equity compared to definitions of justice that address historical inequities (Martens 2021). These concepts have many implications in transportation, as

transportation serves as a critical function of social life that provides necessary access to goods and services (Pereira 2021). Historical inaccessibility of goods and services can hinder the growth of communities, where the concepts of justice attempt to address these issues.

In many cases it was found that the local implementations of equity had varying results and utilized a wide variety of different methods to define and quantify equity (Lewis 2021). Many current municipalities utilize some form of aggregate equity metrics to define their projects, however these metrics are often insufficient as they hide in-group variation and do not account for different populations needs in access to resources (Martens 2022). These errors can lead to non-ideal distributions of resources when attempting to implement equity. For example, a study in California found that when attempting to increase public transit service Metropolitan Planning Organizations (MPOs) typically attempt to increase transit in their urban cores as this has the greatest aggregate impact on increasing ridership and reducing vehicle miles traveled, among other benefits such as reduced air pollution etc. However, this approach was found to systemically underserve some of the poorest communities in these cities which, due to rising housing costs, have been pushed to the edge of the metropolitan region, which works against the stated goals of addressing equity in these changes (Heyer 2020). There are methodologies to address these kinds of issues. One example is called the robust decision-making approach, which utilizes many different scenarios to identify the most robust plan with regards to equity (Lempert 2020).

There are many examples of this concept being defined specifically for transportation. One such comes from a study conducted by the Center of Transportation Studies from the University of Minnesota, where Fan et. al. (Fan 2019) defines an equitable transportation system in three parts:

- A transportation system that is multimodal, affordable, sustainable, reliable, efficient, safe, and easy to use
- Quality transportation services that allow all populations to reach destinations independently
- A decision-making process that includes community insight and public engagement to reduce long standing economic and socioeconomic differences throughout communities

Another definition of equity in transportation is put forward by Di Ciommo and Shiftan (Di Ciommo). They also define transportation equity in three parts:

- What are the benefits and costs that are being distributed
- What are the population groups that over which these benefits and costs are distributed
- What is the distributive principle that determines if a given distribution is 'acceptable'

These definitions share many items in common, including the key concept that addressing equity is an inherently comparative process. From these examples, we develop a basic definition of equity for transportation, where an equitable transportation network is one which provides efficient, reasonable, and safe transportation to all members of a community such that the opportunity benefit and associated cost outcomes relating to these goals are fairly distributed through said community.

Beyond this definition there are many different ways to categorize equity issues. Broad, social level structures often act to enforce inequity, and there are many ways to frame these problems. Two common terms that are used are horizontal and vertical equity. Horizontal equity refers to the distributions of costs and resources across a population, while vertical equity focuses on the equal distribution of outcomes by imposing greater costs on those with more opportunity, thereby disrupting the systemic effects that enforce inequity (Carleton 2018). These terms originate

in the field of tax policy but can be translated to the realm of transportation. There are some issues with this, most notably in regard to horizontal equity where it is difficult to define distributional costs in the context of transportation. These concepts also relate to the idea of transportation justice, defined as practices aimed at the removal of systemic barriers that inhibit vulnerable population's use of the transportation network (Fan 2019).

2.4 History of Equity and Transportation

In reality the transportation network does not provide equitable service for all. It is well documented that the transportation policies in the United States, especially those implemented since World War II, have had a detrimental effect on minority communities and Black, Indigenous, and Peoples of Color (BIPOC) communities (Sanchez 2004). It wasn't until the 1990's that transportation policy makers, planners, and practitioners began to directly address equity concerns in the transportation sector, but by that point extreme disparities had formed in the effectiveness of the transportation system for different groups throughout the U.S.

Many historical factors and policies have all compounded together to cause these inequities in the post war period. First and foremost, there are significant differences to geographic and demographic distributions of populations throughout the US. In general, most US cities have a much higher level of minorities in the city center while suburbs around cities are persistently white (Sanchez 2003). This phenomenon of white flight to the suburbs and the subsequent decay of inner cities due to racially driven disinvestment has many precipitating causes, including discriminatory housing policies like redlining, racially driven economic subsidies, and auto centric transportation policies that prioritized highway construction through poor neighborhoods (Frey 1979). This level

of geographical and demographic separation has many implications for the transportation network. This separation facilitated an unequal distribution of resources in the transportation network which has been a direct cause for many of the equity issues we face today. The geographic differences these historic policies have caused have become even more complicated with the increase in gentrification. Gentrification is the process by which lower income, usually BIPOC, communities of inner cities are being displaced to the periphery of metropolitan areas by wealthier white communities (Allen 2021). This has exacerbated the transportation inequities faced by these communities as it forces them to live in areas with lower access to transportation overall with significantly less public and active transportation options, higher car dependance, and significantly lower density of amenities.

Many of these issues have been both facilitated by and compounded by the destruction of public transportation and highway construction (accompanied by automobile supremacy) through these times. In the U.S., there was a systematic campaign undertaken by General Motors and the oil lobby to replace rail based public transit operations with bus service (St. Clair, 2009). One famous example of this is the liquidation of the Red Car lines in Los Angeles, a streetcar service that was deposed in favor of gas buses by the General Motors backed National City Lines Bus Company. This destruction of rail public transit represents how public policy failed to support public transit in the U.S. This process exacerbated the transportation strain felt by BIPOC and minority communities, as these communities were already more reliant on public transportation as their primary means of transportation (Garrett 1999). This period also coincided with the extreme expansion of the highways system, starting with the Eisenhower plan in the 1950's but continuing through the 1960's and 1970's. The urban freeways were consistently built through the poorest neighborhoods of cities, displacing residents, and having an extreme negative impact on minorities

and those with lower income (Bullard 2004, Mohl 2004). Apart from the physical implementation of discriminatory infrastructure and policy, there is historically an unequal distribution for low income and minority populations with regards to transportation investments and cost burden. Most of the major funds targeting transportation were used for highway construction, which as shown above disproportionally served the wealthier white people who lived in the suburbs. All of these factors have compounded to present a much higher cost burden for transportation for those of lower income. In 2001, those in the lowest income quartile spent 36% of their income on transportation while those in the highest quartile only spent 14% of their income on transportation expenses (Sanchez 2006). All of the historical factors listed above have compounded to make the transportation system in the U.S. highly unequal which has directly led to many of the inequities we face today. Put simply, the discrepancies in the transportation network and mobility create a lower level of opportunity for minorities and vulnerable populations.

Minorities, BIPOC communities, and people with lower income still face many of these same issues today. There is still a significant cost burden associated with transportation for minorities and low-income communities currently (Cochran 2018). As with the example above, this is caused by a combination of spatial distribution of the communities, spatial distribution of opportunity, and the investment in infrastructure and public transportation connecting the two. This is coupled with a high level of public transit dependency for low income and minority populations. This also corresponds with a higher level of time spent using transportation (Yeganeh 2018). Despite all of these issues, the situation has been improving since the mid-1990's. In 1991, Congress passed the Intermodal Surface Transportation Efficiency Act (ISTEA) (Sanchez 2003). This was the first piece of legislation that actively required states, cities, and other jurisdictions that receive federal money for transportation to consider equity as a key component of policy and

decision making. There were many changes that this legislation brought, including providing funding for non-single occupancy vehicle modes and requiring large scale planning operations through Metropolitan Planning Organizations (MPOs). This was followed by the Transportation Equity Act for the 21st Century (TEA-21) which mandated increased public involvement in planning actions for MPOs and States (Sanchez 2003). These pieces of legislations laid the groundwork for addressing transportation equity that is being utilized today and were followed by other key pieces of legislations which continued the changes set in motion by these two landmark bills, including: the Safe, Accountable, Flexible, Efficient Transportation Equity Act (SAFETEA), the Moving Ahead for Progress in the 21st Century Act (MAP-21), the Fixing America's Surface Transportation Act (FAST Act), and most recently the Infrastructure Investment and Jobs Act of 2021. All of these pieces of legislation have continued to help address the equity concerns present in the transportation network by increasing the diversity of funding available and requiring that equity and engagement be a central pillar of use of the funding.

2.5 Current Study of Equity in Transportation

Given the need to address equity in transportation and the growing emphasis of equity being put into transportation funding and legislature, it has become a key field in both practice and academia. There are many programs that seek to understand and implement equity as a key pillar in transportation. However, though these projects espouse transportation equity, there is no common method or even basic structure for these studies and programs to follow (Fan 2019).

Much effort has been put into documenting and defining different methods of evaluating transportation equity. One of the first examples of this type of study was conducted in 2002, titled

Evaluating Transportation Equity that has been updated periodically since then (Litman 2018). This study presents a practical categorization useful for implementing transportation equity analyses for practitioners. Specifically, it defines different axes that can be used to assess equity in the context of transportation planning and engineering. These categories include the different types of equity (horizontal and vertical), the equity impacts felt by users (economic, transportation performance, etc.), different units of equity measurement (monetary, per person, per travel unit), and different demographic categorizations (racial, income, etc.). To effectively create a measure for equity, all of these different factors need to be understood. Transportation equity can be studied in many valid ways which will show different results. This work highlights the idea that there is no singular way to address equity, which is supported by the myriad of different reasonable approaches to understanding and assessing equity for transportation.

The above literature shows that it is critical that equity in transportation is well understood. Despite this, implementing equity studies, projects, and policies in practice often have limitations and barriers. Many of the issues that occur when public agencies attempt to implement policy to address social equity in transportation are highlighted at the state level. States often struggle to implement transportation justice programs due to an overreliance on longstanding classical analytical methodologies for understanding transportation issues that may not always capture the implications of equity at a desired level (Karner 2020). Rural areas experience this issue even more acutely. Small and rural areas are often left behind in equity evaluation methodologies because most equity analysis methods are aimed at larger metropolitan and urban areas (Karner 2016). This again highlights the limitations that are present for agencies attempting to address equity issues. Furthermore, there is no common method for assessing equity in different communities. When finding and evaluating areas where transportation justice is a primary concern, most jurisdictions

use individualized methods rather than a more standard methodology (Beiler 2017). A direct consequence of this is inconsistency in assessing equity across multiple jurisdictions. To address this issue, Beiler et. al. proposed a method called the Transportation Justice Threshold Index Framework to identify transportation justice areas uniformly. This methodology finds areas disproportionately affected by transportation issues with a scalable, uniform method (Beiler 2017). There have also been several studies that assess new methodologies for assessing equity in the transportation network that attempt to address the limitations and issues described above. Bills et. al. describes a new method to assess transportation equity using activity-based demand modeling. The activity-based demand model is able to generate disaggregate transportation measures which directly relate to demographics and therefore equity issues (Bills 2017). Specifically, this methodology utilizes distributional measures of equity which are a better descriptor of equity implications than simple average measures.

To better understand how transportation projects addressing equity compare, Fan et. al. (Fan 2019) utilizes a multi-criteria typology developed by Elman (Elman 1995). This allows for the categorization of different programs in practice into comparable metrics. This study found that of the 24 equity programs that were evaluated, 10 originated from Metropolitan Planning Organizations (MPOs), 5 from local governments, 5 from non-profits, 3 from transportation agencies, and 2 from federal or state agencies. Additionally, it was found that most programs take a compensatory approach to addressing equity as opposed to a procedural approach to addressing equity. In this case, compensatory equity refers to mitigating inequities that are already present in the transportation system whereas procedural inequities refer to the equal distribution and execution of transportation methods and programs. Finally, these programs are categorized based on the primary activities 'direction' to address equity. The two main categories for this are

described as addressing the already present inequities in the transportation system and addressing general social equity issues through improvements to the transportation system. In this category, there is a somewhat even split with 16 in the former category and 17 in the latter category (some of the projects fit into both categories and thus were counted twice).

These research studies show that there is an extremely broad spectrum of methods to address equity, and yet there is still significant room in the realm of transportation equity to explore. Equity in transportation is a broad topic that touches on many fields outside transportation and thus many of the projects aimed at implementing transportation equity require partnerships with outside groups that carry expertise and ability beyond that of transportation agencies. One example of a study to assess equity of the transportation system was conducted by Leung et. al. This study analyzed the equity of I-405 express toll lanes in the Puget Sound Region of Washington. This study utilized ecological regression to understand the population that uses the toll lanes from census data (Leung 2019). This study found that the toll lanes were generally equitable for both lower income and higher income users per trip as lower income users on average gain greater value per trip than high income users while high income users gain a greater net benefit when using the toll lanes.

2.6 Introduction to Big Data in Transportation

Besides issues revolving around equity in transportation, one of the biggest disruptors in the transportation sector is the growth of transportation data sources. Often described by the term ‘big data’ this phenomenon has seen the breadth, depth, and richness of data sources that describe transportation grow massively (Zhu 2019). This is fundamentally changing how transportation

practitioners and academics are approaching issues in transportation. This phenomenon is not only being felt in the transportation sector, but in all fields. It is estimated that by 2025 there will be over 163 zettabytes of raw data collected every year through the internet and other sources (Reinsel 2017). This can have applications in other fields, such as business, social networking, and medicine as examples (Chen 2012). As transportation systems have evolved over the past several decades and technology has improved, as with these other fields, more data is available than ever before (Qi 2008).

The type and style of data that can be contained in big datasets can hold untold variety in the type and style of data collected. As such it can be difficult to compare some of these datasets together, however all of these disparate datasets can be categorized according to three V's: volume, variety, and velocity (Zikopoulos 2012)(Basche 2011). Volume refers to the sheer amount of data contained in a dataset. This can be described as either a finite volume of data for a complete dataset or as a rate of data collection. In transportation, the volume of data has increased extensively with the advent of new sensors and new methods of collecting data. Variety refers to the different types of data that are connected and related together. Again, in transportation the variety of data that is collected has widened greatly, with the advent of the new technologies and data collection methods mentioned above. The velocity of big data describes how quickly data is generated and can be accessed and utilized by the end user. This again has also increased tremendously in the transportation field with the growth of information technology which has facilitated a dramatic increase in the ability for data to be utilized in a rapid and even real-time manner. With the marked increase in the three Vs for transportation data, there has been a fundamental shift in the use of data to big data analytics models. These models consist of three parts: Data collection, data analytics, and applications (Zhu 2019).

There are many different types of big data that are available in the field of transportation. They all have different uses, advantages, and drawbacks. Table 2.1 below is a list of many of the different types of data that are commonly used in transportation study, how they are collected, and what they are used for, however this list is certainly not exclusive. It is also important to note that some types of data can fit multiple of these categories (Zhu 2019).

Table 2.1: List of Big Data Sources, Collection Methods, and Uses

Sources of Data	**Methods of Data Collection**	**Collected Data**
Smart Card	Transit Smart Car Use	Travel Times, O/D Flows
GPS	Vehicle GPS	Vehicle Position, Density, Speed
Sensor: Roadway Site	Induction Loops, Pneumatic Road Tubes, Piezoelectric Loops, Microwave Radar, LIDAR, Acoustic	Vehicle Speed, Density, Position, Classification
Sensor: Floating Car	Cell Phone Tracking, GPS	Travel Times, O/D Flows
Sensor: Wide Area	Cell Phone Tracking, Vehicle GPS	Travel Times, O/D Flows
Sensor: Video	Video Camera	Vehicle Speed, Density, Position, Classification
Connected and Autonomous Vehicle (CAV)	Vehicle Dynamics Sensors	Vehicle speed, acceleration, safety data
Passive Data Collection	Social Media, Mobile Phone	Travel Times, O/D Flows

All of these different data sources have distinct advantages and disadvantages and can be used in different contexts. Smart card data is one of the primary methods currently for understanding the flow of passengers in transit systems around the globe (Bagchi 2005). This data typically captures the boarding time and location when a passenger enters the public transportation network as they use their card for fare payment, either for bus, rail, or both. Some can also provide origin destination information as they can record when a passenger leaves the public transportation system as well. An example of this kind of data is from the agency Transportation for London, which collects information on the use of the London area smart card (the OYSTER card) which sees data from around 8 million trips per day (Pelletier 2011).

Another source of data that is regularly used in big data analytics for transportation is GPS data. This data is typically used for location tracking in some form. There are a myriad of different sources and ways to collect GPS data, including fleet tracking GPS data, smart/connected vehicle data, vehicle trace data, and many others. These data sources can be used to understand many issues in transportation, including travel mode detection (Gong 2012)(Wang 2017), traffic monitoring (Herrera 2010), and travel delay measurements (Asensio 2009).

A further source of big data for understanding transportation comes from the many sensors that have been deployed. There are numerous types of sensors that are available and in use to collect transportation data. These generally can fall into three categories: roadside detectors, floating car detectors, and wide area detectors (Lopes 2010). Data from roadside detectors refers to data that is collected at a single point on a specific piece of transportation infrastructure, usually roadways. Floating car data refers to mobile sensors that are deployed in vehicles such that they can collect data but are non-static unlike roadside detectors (Huang 2010). Wide area detectors refer to detectors that are static but collect information from many different pieces of infrastructure

at once (Antoniou 2008). An example of this kind of detector would be a sound recording device or decibel meter, where the useful area could cover multiple roadways or pieces of infrastructure at once. Most sensors on the roadway fall into the first category of roadside detectors. There are numerous systems that can be implemented for roadside detection. Some of the classic systems include magnetic induction loops, pneumatic road tubes, piezoelectric loop arrays, and microwave radars. More recent technological advancements have introduced many more types of sensors, including ultrasonic sensors, acoustic sensors, magnetometer detectors, infrared detectors, cellular device detectors, light detection and ranging (LIDAR), and video detection to name a few. One critical sensor type is video collection which can be unlocked using video analytics. Datasets driven by video analytics are powerful because they are often relatively low cost and can provide all manner of important data metrics, including vehicle count, vehicle classification, pedestrian and bicycle detection, incident detection, near miss detection, and many others (Courage 1996).

Another rapidly growing source of big data is connected and autonomous vehicle data (CAV). As vehicle technologies have rapidly improved, vehicles now produce significantly more information about their individual performance than before. All of this data is critical for the operation of the vehicles but has the added benefit that it can be extracted and used by transportation practitioners. A major component of CAV technology is the communication between multiple vehicles (V2V) and between vehicles and the surrounding infrastructure (V2X). Not only will this technology greatly improve the performance of the roadway network (Uhlemann 2015)(Chen 2017), but it can also provide critical insight into the current performance of the network through the real time information on the CAV's vehicle dynamics.

Passive data collection is another rapidly growing source of big data for transportation. Unlike the other sources of data described above that are all actively collected datasets, passively collected data instead uses datasets that were not collected for use in the transportation field but nonetheless provide valuable insight for transportation (Gong 2016)(Kang 2012). There are many kinds of passive big data, including mobile phone data, social media data, internet access data, and human activity data. Passively collected data proves especially useful for studying the mobility and travel patterns of people by extracting useful information from these passively collected datasets (Zeyu 2017)(Liu 2013). The most commonly used passive datasets are those generated by social media platforms such as Twitter, Facebook, and LinkedIn to name a few. The interactions of users on these sites can give critical insight into the attitudes and behaviors of transportation users when processed properly (Gal-Tzur 2014)(Chen 2014).

These sources of big data for transportation should certainly not be considered complete; there are many other sources of big data for transportation than these. However, these do represent the major categories of data that most data sources can fit in. It is important to note that some datasets can span between multiple categories as well; they are not defined to be mutually exclusive. Furthermore, there are some other kinds of big data that do not fall into these categories. One example of this is in infrastructure status data, which measures the conditions of specific pieces of infrastructure. This can be measured using many of these data types, including sensors, CAV data, and passive data (Cottrill 2015).

All of these different sources of big data have advantages for transportation practitioners and academics. The richness of data they provide allows for far greater insight into the transportation world than was previously possible, especially when compared to classical data sources and analytic methodologies (Zannat and Chowdhury 2019). More and more frequently,

big data sets are used to facilitate strategic decision making, phrased as data-driven decision-making (Gamage 2016)(Höchtl 2016). However, most of these datasets are relatively new and there are likely implications for them that are not well known. Of primary interest to this paper are the implications these datasets have on equity and equity studies. Though the richness and depth of these datasets has significantly grown, it is often unclear what populations are represented by these datasets. There is currently no methodology to understand what populations are underrepresented in these datasets. It is easy to imagine how some of these equity biases might be present in these datasets. Take CAV data as an example; this dataset will miss out on a significant portion of the population because these datasets require ownership of a newer car that has CAV technology implemented in it. This requirement to be included in the data will ensure that a significant proportion of the population is missed in this dataset, especially those of lower income who are less likely to own a car at all, let alone a new car with this technology. This dataset then likely misses out on this critical population whose behavior is of great importance to transportation practitioners. To ensure that datasets accurately capture the breadth of the population, it is imperative that transportation practitioners can empirically show how biases are present in the data.

2.7 Chapter Summary

This chapter covers the current state of practice and literature related to equity and big data in transportation. Throughout we 1) Cover the historical and current practices for transportation equity such that an understanding of the precipitating causes and ongoing methods to address these causes are understood and 2) explore different types of data sources that are causing the growth of

big data such that their impact on equity can be examined. We find that there are many historical causes to the inequities we see today in the US transportation network, including the geographic distribution of populations, car centric infrastructure and policy, the destruction of public transit, and the historic disinvestment from BIPOC communities. To address these critical issues, there are many practices in place to address these equity concerns. Finally, we explore the different sources of data that are fundamentally changing how transportation practitioners address the transportation network. From these sources, we see there is the potential for data biases to skew this new understanding, which can cause equity concerns if these data sources systematically undercount historically disadvantaged populations and are being used as decision making tools.

Chapter 3: Defining Equity Biases in Transportation

3.1 Overview

Data biases are an issue that every transportation practitioner needs to be aware of. All data sources will have some levels of inherent bias, and there are many methods to mitigate these biases. However, there is little understanding of the equity implications that these biases have on datasets. Data biases have the potential to systematically underrepresent certain segments of the population. These issues have grown more acute as the amount of transportation data has grown with big data. This has the potential to raise concerns about equity as this systemic difference in representation can lead to historically disadvantaged populations being underrepresented in data. This is especially critical as new data sources are regularly being used for data-driven decision-making, which means that these biases can directly lead to inequitably implemented policies which is contrary to the goals to address equity in transportation.

3.2 Contributions and Chapter Structure

In this chapter, we explore data biases in transportation and the impacts these have on equity. We begin with a review of data biases in transportation, including addressing definitions and methods to mitigate these biases. From this introduction, we define what an ‘equity bias’ is in transportation engineering to inform how these biases impact equity. The contributions of the chapter are as follows:

- We provide a thorough review of data biases in transportation to understand what data biases are and different ways to mitigate them.
- We connect how these biases occur to the potential equity implications they may have to create a definition of equity biases.

The chapter is organized as follows. In section 3.3, we introduce data biases in transportation to define them and review practices relating to addressing and mitigating the biases. Section 3.4 then takes these basic definitions and reframes these concepts to address the equity concerns of these biases, leading to a definition of equity bias and a discussion of the implications this idea has on practices related to data-driven decision-making and equity. Finally, section 3.5 concludes the chapter with a summary and concluding remarks.

3.3 Introduction to Transportation Data Biases

In order to understand the impact of biases with equity implications in any source of transportation data, it is critical to define general data bias in transportation. Data bias is any error in a dataset that mis-weight or mis-represents parts of the dataset away from the true condition (Delgado-Rodriguez 2004). There are many examples of this occurring in the transportation sector. With the growth of big data, the nature of these biases has changed profoundly. Griffin et. al. defines data bias for transportation into four main categories: coverage biases including sampling and non-response bias, measurement bias, social desirability and demographics bias, and aggregation bias (Griffin 2020). There are many causes for these biases and subsequently many techniques available to mitigate these biases. Some of the most common include fusing multiple

data sources together to provide a more complete picture of a transportation problem or weighting data to better represent the ground truth conditions.

Despite all of the above techniques to address biases in transportation data, there is still little connection between how these biases present in transportation related work and how they affect the decisions and outcomes of that work. Equity is one of the subfields in transportation where this disconnect has a profound impact. Many of the biases described above do touch on critical equity considerations, especially those biases that address discrepancies in demographics, but no work has been done to connect these discrepancies to equity outcomes of the data biases present in these transportation datasets. Understanding these challenges is especially critical as these biased datasets are more and more frequently being used for data-driven decision-making. However, there are some examples of this type of work being done in other fields that can be related to the problems present in transportation. One such study highlighted these biases in the disaster response to Hurricane Sandy that struck the Northeast United States in 2012. The dataset used in this study included all tweets that pertained to Hurricane Sandy. It was found that the tweets concentrated around New York City, which is not surprising given its large size and population. However, many of the areas that were affected much more severely than New York City were vastly underrepresented due to several factors including power access and internet access being severely reduced by the storm and the areas being less affluent in general, which correlates with lower smartphone ownership and lower wireless network coverage (Crawford 2013). This directly relates to both the public perception and the media response to this particular natural disaster. In this case, there was a direct equity implication in this response to the hurricane due to biases present in the real time available data as the focus was partially drawn away from the

most devastated and vulnerable communities to focus more on the more affluent and less vulnerable communities of the city.

There are several studies that touch on the biases that occur in specific transportation datasets, even if there are no studies that connect the biases to equity outcomes. Understanding these biases is critical for finding the implications they have on equity. For example, a study was conducted by Bergman and Oskanen to understand cycling routes in Helsinki, Finland. To do so, they used passively collected trace data gathered from fitness mobile apps to understand the travel demand patterns of cyclists. Though they were able to correct this issue in their study, they found that this data introduced a significant amount of bias as it only included a relatively small segment of the population (Bergman 2016). This data predominantly captured those that were cycling regularly, often to commute. This, related to the behaviors of this group of cyclists, skewed the data to not show an accurate depiction of the true cyclist behaviors of Helsinki such that more casual cyclists were under-represented. For further examples highlighting biases see: ((Diao 2016), (Garcia-Albertos 2018), (Griffin 2015), (Schweitzer 2017), (Zhou 2018)). The primary concern with these biases is summed up by Shearmur; “However big the data, Big Data is not about society, but about users and markets” (Shearmur 2015). This quote exemplifies most of the biases relating to big data; they are often sources of data which are not able to describe society as a whole, but only a segment of society. These biases often reflect the historic inequities that are faced by historically disadvantaged communities, exacerbating the equity impacts to these communities. Often it is found that users of the fitness mobile app are wealthier recreational cyclists as opposed to commuters (Kwayu 2021). Hypothetically, if this uncorrected dataset were to be used as cycling mobility data for bicycle infrastructure development policy for Helsinki, the needs of many important populations for the city would be underrepresented and therefore underprioritized. This

phenomenon of incomplete representation of the population directly leads to the equity issues explored herein.

One example of a transportation project that considered the equity outcome of biases present in used datasets is the Streetbump mobile app used in Boston to detect potholes (O'Leary 2013). This app is used as a non-active data collection method, meaning that users of the app do not need to manually enter information as it is instead collected automatically. As an individual drives around Boston, the app uses the phone's accelerometer to detect when the car is jostled significantly, indicating a likely pothole, and records that data point using GPS in a cloud database. This is a very efficient way to collect information on pothole detection and can greatly reduce monitoring costs, however there are significant biases in this dataset. There are several population areas that were found to be underrepresented in this dataset, including those in lower income brackets, the elderly, and youths. This is due to the fact that to use the app, smartphone and car ownership is required; these groups are far less likely to meet all of these requirements and thus are much less likely to be counted. This leads to some roadways not receiving the same level of data collection as others, a direct equity concern as this could lead to an unfair distribution of road maintenance funds and resources. In this example, this is overcome by the app only being available to city employees who have agency directives to equally assess all roadways, ensuring that roadways are more evenly measures throughout the city. This kind of equity mitigation is not feasible for most big data sets in transportation (O'Leary 2013). This example shows why there is a need for the equity outcomes of biases to be understood to effectively address the equity concerns of a transportation dataset.

Another example where policymakers overcame biases related to equity can be found in the development of the Owl Bus in Seoul, South Korea. Founded in 2013, the Owl bus functions

from midnight to 5am with nine routes developed by the Seoul Metropolitan Government. The Seoul Government worked with a major Korean telecom company to gather information on the use of cell phones, supplemented with taxi usage data, during late night hours. This data was leveraged to develop bus routes with the highest potential usage for the new bus system. This methodology, however, greatly favored high-income neighborhoods with higher response rates, potential usage, and potential revenue generation. Recognizing the bias in their data and the negative equity implications of this bias, the Seoul Government made the conscious decision to spread Owl Bus service more widely throughout the city. While the development of Seoul's Owl Bus system is widely hailed as a "win" for Big Data and evidence-based policymaking, one of the core lessons from Seoul is the limitations of Big Data and how cities and policymakers can and should actively work to overcome such equity biases (Hong et al. 2019).

3.4 Defining Equity Biases in Transportation

The examples shown above allow for the idea of data biases that impact equity to be reframed: *equity bias is any bias in a data source that produces a negative equity outcome by underrepresenting historically disadvantaged populations*. The key factor of this definition that differs from the above examples about demographic biases is that it deals directly with the outcomes of the biases present in the datasets, not the initial flaws that precipitated those biases. Many of the examples above can be considered equity biases; especially those that deal with biases due to under- or un-counted demographics. Since they do not accurately represent the entire population of study the full implication for the dataset is unknown, which has an inherent impact on equity, especially if the population left out of the dataset is a key vulnerable or historically

disadvantaged population. These equity biases should be of major concern to transpiration professionals because these biases can present themselves in studies and methodologies that play crucial roles in real world data-driven decision-making. Many of the critical performance metrics that are used by transportation professionals are inherently flawed and can introduce significant equity concerns with their use (Karner 2020). Though these flaws come from a myriad of sources, it is well documented that the biases and small sample sizes of the data used to develop many classical performance metrics causes many of these concerns and bring into question the usefulness of these metrics (Dumbaugh 2014). Though many of these sources of bias highlighted above are equity biases, the definition is inherently broader than any of these individual sources of bias. Equity bias can accrue from any source of bias, and thus cannot be narrowly defined as one type or a category of types; instead, the focus must be held inherently in the equity outcomes of the bias present regardless of the precipitating cause of that bias.

The application of big data is a boon for transportation practitioners who are given a previously unimaginable understanding of their field. The access to the aforementioned data sources presents an unprecedented opportunity for decision makers to make informed and evidence-based decisions. However, the known potential for bias, equity-based or not, in the data has caused hesitation on the parts of local governments to fully utilize this resource (Guendez et al. 2020)(Williamson 2014). Despite that hesitation, the utilization of big data in the public sector will only grow as governments are swayed by the value of these data sources (Gamage 2016)(Höchtl 2016). Knowing that big data will continually be used under the umbrella of 'evidence-based decision making,' it is critical to create strategies to mitigate bias most effectively. There are several potential methods for addressing these issues. The first step to addressing these biases is to quantify them. As the concept for equity bias implies the outcomes of using a dataset

with inherent bias would be quantified in how well it describes the population it is trying to capture. Dealing with the over- and under-counting of populations in data has direct equity implications. Currently there are no proven quantitative methods for measuring equity bias.

Mitigation techniques can also be utilized to reduce the equity bias of different datasets. There are many well studied methods to mitigate biases present in datasets. These can generally be split into two categories; mitigating biases in the dataset before use and modifying the outputs of the data to address biases during or after analysis. This divide is created predominantly when the bias corrections occur, pre or post analysis. There are already many methods to address data biases pre analysis (e.g., see (Griffin 2020)) that could be utilized to target biases that create equity data bias. There are also methods that correct for equity concerns during analysis (e.g., see (Beiler 2017)), however there are no methods that specifically address equity issues derived from the biases present in the dataset.

There are several other ways decision makers can ensure used data does not contain significant amounts of bias, equity or otherwise. There is a wide array of tactics for mitigating selection bias when working with big data, including controlling for confounding factors by using a model based on regression discontinuity design and using large, well-established data sources to establish benchmarks to “gut-check” any results against these benchmarks (Seely-Gant and Frehill 2015). Since most practitioners rely on third party vendors to supply this data, some researchers stress the need for quality assurance checks (Mcardle and Kitchin 2016). Additionally, there is a growing movement to incorporate traditional “small data" sources to corroborate or challenge any findings from big data sources (Seely-Gant and Frehill 2015)(Bollier 2010). There are also several promising methodologies using new technology to address these highlighted issues. For example,

AI methodologies have the potential for reducing implicit biases in data sources collected for various uses (Lin 2020).

Contrary to the shift to data and evidence-based decision making, in *The Promise and Peril of Big Data*, David Bollier argues that perhaps the solution to mitigating biases is to avoid big data or for policymakers to minimize their use of big data altogether: "More data collection doesn't mean more knowledge. It actually means much more confusion, false positives and so on. The challenge is for data holders to become more constrained in what they collect." Bollier also argues that many of the bias-mitigation strategies mentioned above are moot – when researchers manipulate data to remove bias, they introduce their own bias, further compromising the results (Bollier 2010). Despite these reservations, the literature shows the shift to data-driven decision-making is continuing to advance (Gamage 2016)(Höchtl 2016). In the transportation sector, there are too many sources and uses of data for transportation practitioners to stop utilizing these resources. It was found that big data sources provided "at least as good if not better" models/tools for public transportation planning compared to traditional methods. Furthermore, the majority of the studies found that the use of big data sources is significantly cheaper than traditional methods (Zannat and Chowdhury 2019).

Transportation has a long history of creating and perpetuating inequity through misguided planning and engineering that created negative outcomes for minorities and vulnerable populations in the U.S. Transportation practitioners need to understand the limitations of big data to avoid equity biases and work to intentionally make transportation more equitable. Data bias and equity bias are critical issues that must be addressed in transportation practice and research as we continue into a future where the amount of data at our disposal continues to grow and data-driven decision-making becomes the norm for transportation practitioners.

3.5 Chapter Summary

This chapter covers data biases in transportation and the equity impacts these biases have. Throughout we 1) provide a thorough review of data biases in transportation to understand what data biases are and different ways to mitigate them and 2) connect how these biases occur to the potential equity implications they may have to create a definition of equity biases. We find that there are many precipitating causes of data bias in transportation datasets, and many of these biases have equity implications that are unexplored. From these concepts, we develop a definition of equity biases: *equity bias is any bias in a data source that produces a negative equity outcome by underrepresenting historically disadvantaged populations*. This reframes the concept of data bias to focus on the equity outcomes of the bias as opposed to the traditional view of addressing the precipitating causes of bias. This brings into question the usefulness of big data for data-driven decision-making as these equity biases can directly lead to inequities in transportation policy and decision making. Realistically, this means that we must develop a methodology to mitigate the equity biases present in transportation datasets as data-driven decision-making is too effective otherwise to get rid of.

Chapter 4: Methodological Framework for Calculating Relative Representation of Datasets from Equity Biases

4.1 Overview

As new and more abundant transportation data sources become available, transportation decision makers have tried to use these data sources to make more informed decisions. This process is known as data-driven decision-making, where design and policy decisions are made based upon the observed condition of the transportation network through data as opposed to classical analytical methods which rely on inaccurate and ineffective methods to guide transportation decisions (Dumbaugh 2014). Data-driven decision-making addresses these concerns by providing more accurate, precise, and complete understanding of the transportation network which can lead to 'better' transportation decisions (Gamage 2016)(Höchtl 2016)(Zannat and Chowdhury 2019).

Though this is generally considered a positive, data-driven decision-making is not a cure-all for decision makers; at times greater amounts of data can potentially cause more confusion leading to 'worse' decisions (Bollier 2010). While data-driven decision-making does typically provide 'as good if not better' information (Zannat and Chowdhury 2019), there are still many ways in which data-driven decision-making can be improved. Currently in study, there is little understanding of how data biases affect the data-driven decision-making process. As data biases are by definition some form of misrepresentation for data, this misrepresentation could negatively affect the outcome of data-driven decision-making by overweighting the importance of certain demographics in the process, leading to a potentially less equitable outcome (Martens and Golub 2021). Therefore, it is important that representation of the datasets being used is considered to

mitigate the potential of a negative outcome from data-driven decision-making. To achieve this goal, a methodological framework is created to quantify the representation of datasets to provide a comparable metric for decision makers to judge datasets based on their representativeness.

4.2 Contributions and Chapter Structure

In this chapter a methodological framework is presented to quantify relative representation for transportation datasets. We begin with an overall definition of representation that will be used by this framework, followed by the formulation of the framework itself. From this we can show how representation can be used to better understand the quality and decision implications for transportation datasets. The contributions of this chapter are as follows:

- We explore the fundamental terms that must be included in calculations to compare representation.
- We create an implementable framework that combines these terms to find the representation of different datasets.

This chapter is organized as follows: Section 4.3 provides the key definitions and overview for the methodological framework. Section 4.4 showcases the detailed formulation of the framework. Section 4.5 highlights key discussion points about the effectiveness of the framework. Finally, section 4.6 concludes the chapter with a summary and concluding remarks.

4.3 Methodological Framework Overview

The key goal of this methodological framework is to quantify the relative representativeness of a dataset that can be used as a metric to compare multiple data sets. Representation can be defined as the description or portrayal of someone or something in a particular way or as being of a certain nature. When specifically using this term in the context of transportation and data, representation generally refers to the description and portrayal of different demographic groups that are captured in each particular dataset. This definition holds two key implications for the purposes of this study; first that representation is inherently comparative when assessing multiple strata, and second that representation has a direct impact on equity in transportation.

When defining equity in transportation, representation consistently plays a key role. It is widely accepted that representation is critical to achieving equity in the transportation decision making processes (Fan 2019)(Martens and Golub 2021). By adequately representing all parties affected by transportation decisions, especially historically disadvantaged communities, the needs of those parties can be heard and accounted for such that a 'better' outcome can be achieved for those communities. One shortcoming in this area is that in current practice representation is achieved exclusively through community engagement (Fan 2019). While this is certainly a positive outcome, this does not address the impact of representation in the data used for data-driven decision-making, though this type of representation is of similar importance. This is the gap that this methodological framework fills: to quantify the representation of datasets to better inform equitable data-driven decision-making.

Leading from this idea, it is important to define what is equitable with regards to representation. Another way to frame this question is 'what is fair' with regards to representation. To define fairness in representation, we use the ladder of equity to define different levels of equity

and fairness, developed by Martens and Golub (2021). For this discussion, we will focus on the two highest rungs: proportional equity and restorative justice as these levels are widely used to define equity in current practice. Proportional equity refers to a state where no particular group receives a notable difference in treatment, i.e., all benefits and costs are shared equally between all groups. In the context of representation in decision making, this would be interpreted as all groups being represented in the decision-making process proportionately to the community. This clearly shows a level of fairness in that each group is providing input in a relatively equal manner. The other definition of equity that can be used is restorative justice. This refers to a state where groups that have been historically denied benefits must be accounted for such that any existing systemic shortfalls are mitigated. When applied to representation, this idea indicates that at times certain historically disadvantaged groups should be overrepresented such that their needs are amplified to address the disparity seen by these groups. In practice in the transportation field, there is already precedence for this, as many surveys and community outreach campaigns aiming to capture the behaviors of the entire population deliberately oversample disadvantaged communities such as BIPOC communities, communities with high levels of poverty, and communities with low levels of income (Sanchez 2004).

With these definitions in mind, we can define the basic framework for quantifying representation in transportation datasets. This process can be broken down into several steps: 1) define the demographic strata that are to be compared, 2) connect these demographic strata to the transportation datasets of interest, 3) determine the representation of each stratum compared to the overall community in each dataset, and 4) calculate the overall representation of each dataset by weighing and averaging the representation of each stratum such that these terms can be compared between datasets. In this methodology, the demographic strata that are being used are determined

by the user. These could be any demographic the user wishes to showcase, such as income levels or racial demographics. Once these strata have been defined, the representation of each of these strata can be calculated and then weighted and compared to determine an overall representation. The fourth step is critical to determining the level of equity that the framework produces. By using either no weight or by weighing only by strata size, this step will show proportional equity as this will compare the results directly against the demographics of the overall community. However, if weights are used such that the results of disadvantaged strata are favored, then this will showcase, at least partially, proportional justice as the outcomes of these disadvantaged populations carry more weight over others.

4.4 Methodological Framework Implementation

As stated above, the goal of this framework is to find the relative representation of different datasets to compare them together as a metric. For this task, we will use the percentage representation of the dataset as the metric. This will effectively convey 'X percent of the underlying population is represented in this dataset'. It is important to note that this implementation of the framework is not adequate to show a value for absolute representation; instead, it is only intended to showcase the relative representations between different datasets to aid in the data-driven decision-making process. The percent representation shall be presented as P and will be a value between 0 and 1 where 1 represents perfect representation, and 0 represents no representation.

As mentioned in the basic processes in the previous section, the first step is to determine the demographic strata that will be analyzed. This can be any level of demographic split determined

by the user. Income demographics or racial demographics are by far the most commonly used demographic characteristics for this, however any applicable demographic can be used. The strata should be selected based on the specific context of the implementation of this methodology. Due to the situational nature of addressing equity in transportation, most equity solutions require local specifications to successfully address the equity concerns faced by a community (Lewis 2021). Therefore, the determination of the strata must address this issue of locality by selecting the demographic strata that are most applicable in each case of this framework's implementation. Once the strata boundaries are defined, the goal will be to calculate the percentage representation of each stratum, P_x, where x represents the stratum. P_x follows the same convention as P (i.e., it is a percentage between 0 and 1). Once all P_x are calculated, the variation in representation of each stratum can be used to determine the overall representation.

There are several factors that need to be considered when calculating P_x, most of which rely on determining the demographics of the population that is captured in the dataset and comparing that population to the overall population. This is where step 2 of the framework comes into play, where general demographic information must be connected to the transportation data being analyzed because transportation data rarely includes direct demographic information (Karner 2020). This involves developing the demographic profile for those included in the dataset. This demographic profile serves a key role in determining the representation of the dataset. There are many potential methodologies that can be used for this purpose. One example is ecological regression, which can be used to determine the demographic profile characteristics of transportation datasets (this method is used later in this dissertation: see chapter 5 for more details). This is however not the only method that can be used; other methods such as machine learning and artificial intelligence methods could potentially serve this purpose as well (Lin 2020). For this

framework, the exact method to calculate this is less important than ensuring that a consistent method is used; in order for the results to be comparable the same method must be used to eliminate any skewing from utilizing different methods. Again, the main methodology used in this dissertation is ecological regression (see Chapter 5), though others may be suitable for this purpose.

Once the demographic profile for the dataset has been calculated, we can use this to calculate P_x. There are three key factors that must be accounted for to calculate P_x. These factors can all be presented in a similar fashion to P and P_x as how these factors impact the percentage of representation. Once each of these factors is calculated for each stratum x, they can be multiplied together to find P_x for each stratum. This constitutes step 3 of the general process.

The first term that needs to be considered when determining the representation of different transportation datasets is the network representation of the dataset. Network representation generally refers to the differences between the demographic profile of the transportation network calculated in step 2 above and the 'overall' demographic profile of the community of interest. To highlight this idea, imagine a dataset that captures the traffic information for a particular road in the network of a city. In this case, the transportation network that is captured in this dataset is the roadway which is a subset of the overall transportation network. The population that is captured in this dataset will therefore be a subset of the overall population that uses the transportation network and may or may not be different from the overall population. The goal of this term is to measure the difference (or the over- or under-representation) between these populations, because in this example if the use of this road is skewed towards certain demographics, then this dataset will be less representative. This term must be determined at the strata level, as each stratum can have a different magnitude of difference between the data and overall population. Network

representation can be symbolized as R_x, where again x represents each stratum. R_x can be calculated using equation (1):

$$R_x = 1 - |Population_x - Sample_x| \quad (1)$$

$$Where: x\ represents\ each\ stratum$$

This equation will be constrained to the necessary format where the result falls between 0 and 1 and showcases the precent representation for each stratum x. Additionally, the terms of $Population_x$ and $Sample_x$ was chosen to not be squared as these terms do not carry statistical variance and thus gain no information from being squared. Instead the variation can be incorporated into the models or data used to generate these demographic values.

The second key term that needs to be considered to find the percent representation of each stratum is the observability of the users in that particular stratum. In certain cases, some strata may be less 'observable' in a dataset than others. This can also be considered systemic errors, where certain strata have higher rates of error in the data collection process than others. As an example, for this term, consider data collected from cellular phones. There is a key barrier to entry for this dataset; in order to be counted, one must have a cellular capable device. This barrier may not be evenly distributed among different strata. For example, the elderly and those of lower income are less likely to own cell phones (Crawford 2013). Because of this systemic error, these strata are less observable in this dataset, which will reduce their overall representation. Therefore, it is critical that this term be considered when calculating representation of data. This term can be represented as O_x, for each stratum x. This term shall be formulated as the percentage of each stratum that is observable, thus keeping the same convention of the other terms as laying between 0 and 1 with 1 being perfect representation and 0 be no representation.

The final term that is required to determine the overall percent representation of each stratum is the accuracy, or random error, of the dataset. All data may be party to some form of random error; in practice no data collection method is 100% perfect at collecting data. If this error is truly random, this on its own will not impact the representation of a dataset as the errors should be evenly distributed between all strata equally. However, this error can compound the error found by the two previous terms, so therefore must be included in the calculation of P_x. The term that must be used in this framework is the percent accuracy and can be represented as A, which is $1 - data\ error$. This will ensure that the term follows the same convention and the previous terms. Note that this term is the same for all strata.

Once all of these terms have been calculated, they can be used to find the percent representation of each stratum P_x. This can be done through the equation (2):

$$P_x = R_x \cdot O_x \cdot A \quad (2)$$

$Where: R_x\ is\ the\ network\ representation\ or\ each\ stratum\ x$

$O_x\ is\ the\ observability\ of\ users\ in\ stratum\ x$

$A\ is\ the\ data\ accuracy$

These terms are multiplied together because the error for each term will compound together to reduce the overall representation. This will provide a percentage representation for each stratum x. Once these are calculated they can be used to find the overall representation for the entire dataset. To do this, we calculate the weighted average of the values of P_x to find the overall value P. The general form of this equation can be found in equation (3):

$$P = \frac{\sum(P_x \cdot Overall\ Population\ Share_x \cdot weight_x)}{\sum(Overall\ Population\ Share_x \cdot weight_x)} \quad (3)$$

$Where: P_x\ is\ the\ percentage\ representation\ of\ each\ population\ x$

$Overall\ Population\ Share_x\ is\ the\ share\ of\ strata\ x\ from\ the\ general\ populaiton$

$weight_x\ is\ the\ weight\ value\ for\ each\ strata\ x\ assigned\ by\ the\ user$

This equation will calculate the percentage representation for this particular dataset. It is important to note that there are many different ways for the user to implement different weights. For example, the population share of each stratum is not required; this would result in effectively an arithmetic average being calculated regardless of the size of each demographic stratum. Further, different weight values will produce different results. These weights can be assigned by the user to address the local contexts of the application of this framework as implementations of equity need to vary significantly due to the local context of the implementation (Lewis 2021). There are many ways different weights can be assigned, however the general principle that users should follow is to assign higher weights to historically disadvantaged populations, such as lower income or BIPOC communities. This will ensure that this methodology helps to address these historical inequities through restorative justice (Martens and Golub 2021). The magnitude and assignment of these weights must be applied in accordance with the needs of the local community for which this framework is being implemented to ensure that the needs of said community is being met. It is critical that the user be cognizant of the choice of weights to use and understand the implications that those weights have. It is also important that scores only be compared between datasets where the same weighting scheme has been used to ensure they are truly comparable. This will provide a value between 0 and 1, where 1 represents perfect representation and 0 represents no representation. This score can be directly compared to the score of other datasets that used the

same methods for calculating the terms. The dataset with a higher value can be considered to have a higher relative representation.

4.5 Methodological Framework Discussion

This methodological framework has many implications for transportation decision makers. This will allow decision makers for the first time to judge the representation of different datasets. This has a direct impact on the equity implications of these decisions. A more representative dataset will, all else being equal, provide a more equitable outcome when being used for data-driven decision-making (Fan 2019).

There are several key benefits to this methodology that enhance the ability for practitioners to implement equitable solutions for the transportation network. Firstly, this framework gives multiple levels of information that are useful for conveying different ideas. For example, the final output of the framework is a single value of relative representation for each dataset, which is ideal for high level decision makers as it is a single, easy to interpret number that can easily show the critical metric, relative representation, between datasets. This framework also showcases much more in-depth detail as well through the necessary calculation of the underlying demographic profiles of the transportation networks of study. This can provide critical insights into the over- and under-representation of the networks in question, and subsequently the data that describes those networks which is equally as critical.

Throughout this framework, the process where the implications of equity change the most is with the weighting process. As mentioned above, this is the key step where users can overweigh critical strata to increase their impact on the overall representation score. This is a form of

restorative justice, where historically disadvantaged groups can receive greater emphasis to address some of the historic shortfalls in both representation in data and representation in decision making. However, this could also create issues as the end user could weigh against historically disadvantaged groups. Ultimately, this highlights that this framework is a tool for comparing the representation of datasets together and cannot account for users that misinterpret or misuse the framework.

There are several other limitations that this framework holds. This framework was developed as a general guide to determining the relative representation of datasets. Therefore, it is impossible to account for all of the unique datasets that exist in the field of transportation. For some datasets, they may not have the supporting data to calculate all of the required terms to calculate the relative representation or the local context may not support the use of the terms exactly as outlined here. This is an unavoidable aspect of transportation equity, as equity is highly driven by local context such that it is extremely difficult to create universal methods to assess equity (Beiler 2017). This framework can certainly fall victim to this and thus may not be fully implementable in all cases.

Finally, another limitation of this framework is that it can only serve as a comparable metric for data representation and not an absolute metric. Essentially this means that a single value of representation for a dataset on its own does not have meaning. It only has meaning when compared to another value for another dataset. This is again a trait of this being a general framework; calculating an absolute value of representation will require unique methods depending on the characteristics of each dataset that can vary significantly (Griffin 2020). Therefore, in order to create an easily implementable metric, relative terms of representation are used.

4.6 Chapter Summary

In this chapter, we formulate the methodological framework for quantifying representation in datasets. Throughout this chapter we 1) explore the fundamental terms that must be included in calculations to compare representation and 2) create a framework that combines these terms to find the representation of different datasets. To calculate the overall representation of a dataset, these four basic steps must be followed: 1) define the demographic strata that are to be compared, 2) connect these demographic strata to the transportation datasets of interest 3) determine the representation of each stratum compared to the overall community in each dataset and 4) calculate the overall representation of each dataset by weighing and averaging the representation of each stratum such that these terms can be compared between datasets. From these steps, we can create a basic mathematical formulation to implement the framework by calculating the percentage representation of for the dataset. To do this, the percentage representation of each of the defined strata must be found to provide a comparison of the representation of each stratum. This can be calculated through three terms: the network representation, the observability of users, and the data accuracy. From these terms, the percentage representation of each stratum can be calculated. From these values, the overall representation of the dataset can be calculated by computing a weighted average of the value for each stratum. This allows users to account for historically disadvantaged communities. Once calculated, the percentage representation score of a dataset can be directly compared to the scores of other datasets to determine which is most representative. This has a direct implication for data-driven decision-making, where more representative datasets will result in more equitable decisions, all else being equal.

Chapter 5: Introduction to Ecological Regression for Capturing Transportation Demographic Profiles

5.1 Overview

Understanding the demographics of those using the transportation network is critical for any study of equity or equity biases. Equity is inherently comparative, and the demographics of those who do and do not use the transportation network are critical in this regard as the comparison point between these groups. Currently there is no methodology to determine this demographic profile, nor are there easily accessible or universal datasets that capture this information empirically. This chapter presents a methodology utilizing ecological regression to estimate the demographic profile transportation networks. Ecological regression, also called ecological inference, allows for the extraction of individual level characteristics from aggregate data sources. Using other regression methods, this would normally be impossible to calculate with statistical rigor as transportation behavior violates the assumption that all those in aggregate level strata will not systematically behave differently based on other individual level characteristics that transcend those levels of aggregation. This provides a more thorough understanding of the demographic profile of a transportation network.

5.2 Contributions and Chapter Structure

In this chapter, we cover how ecological regression can be used to determine the demographic profiles of transportation networks. We begin with an overview of ecological regression and the

statistical principles that allow it to fulfill this goal. We then show how this can be implemented to determine the demographic profile. Finally, we showcase a case study of this methodology's implementation for the Central Puget Sound Region. The contributions of the chapter are as follows:

- We provide a thorough introduction to ecological regression such that the applications of the methodology are clear.
- We show how this methodology can be used to determine the demographics of transportation networks.

This chapter is organized as follows. Section 5.3 introduces ecological regression, including its initial uses and statistical functionality that makes it fit for the task of determining transportation demographics. Section 5.4 highlights the methodology by which these demographics can be determined with statistical rigor. Section 5.5 shows an example of this methodology being implemented in the central Puget sound region. Finally, Section 5.6 concludes the chapter with a summary and concluding remarks.

5.3 Introduction to Ecological Regression

Ecological regression models allow for the relationship between individual level quantities to be found from aggregate (or ecological) data (Jackson 2006). The key difficulty that ecological regression addresses is that group level exposure may not be the same as the individual exposure within an aggregate region, causing ecological bias (for more discussion see: (Richardson 1987, Greenland 1989, Greenland 1994 and Richardson 2000)). Ecological regression considers the within-area distribution of exposure to give a more accurate estimate of true individual exposure

levels that other regression methods cannot. Take for example an outcome count in area i with population N_i as y_i. Typically, the model of y_i, using as an example a Poisson or binomial regression, utilizes the area-level covariates z_i, which when using binary covariates shows the proportion of the population exposed in the area based upon each covariate. The typical regression models will only measure the relationship between the aggregate exposures and outcomes. Therefore, these types of models are only accurate when each individual in area i has the same covariates z_i, resulting in a perfectly linear model. This however is often not the case, as many real-world scenarios involve individuals in aggregate areas having different covariates for certain key explanatory indicator variables. This deviation from the perfectly linear scenario is the direct cause of the ecological bias experienced by binomial and Poisson regression models (Ricardson 1987).

To address these biases, ecological regression takes advantage of supplementary data sources, specifically individual level data, to address the individual variation within the aggregate area. The model used is Goodman's Method of Bounds regression (Goodman 1959). This can be modeled using Equation (1) below:

$$logit(p_{ij}) = \mu_i + \sum \alpha x_i + \sum \beta z_{ij} + \gamma_{s_{ij}} \qquad (1)$$

$Where: p_{ij}\ is\ the\ total\ risk\ for\ an\ individual\ outcome\ y_{ij}$

$j\ is\ each\ individual$

$i\ is\ each\ area$

$x_{ir}\ is\ each\ group\ level\ covariate$

$z_{ijr}\ is\ each\ individual\ level\ covariate$

$\alpha_r\ is\ each\ grop\ level\ coefficient$

$\beta_r\ is\ each\ individual\ level\ coefficient$

μ_i *is the baseline risk*

$\gamma_{s_{ij}}$ *is the multistrata risk across each strata s*

From this equation, the average risk for an individual in area i can be found by integrating the individual model over the joint within-area distribution of covariates. This is shown in Equation (2) below:

$$p_i = \int p_{ij}(x) f_i(x) dx = E_x(p_{ij}|i) \quad (2)$$

Where: p_i *is the average risk for an individual in area* i

This information is enhanced in quality when corroborated with individual level data within each region. This individual data, which contains data on a subset of individuals in each region, allows for the model to ensure accuracy of the individual covariates (Richardson 1987)(Wakefield 2002)(Prentice 1995). This allows the average risk for each individual to be better calibrated to actual outcomes.

The output of an ecological regression model is inherently probabilistic: generally speaking, it indicates the likelihood that an individual in an aggregate dataset fulfills a given category of interest (Jackson 2006). A key implication is that this method is only statistically valid when the results of the model are re-aggregated together. This means that the model is not accurate in inferring characteristics about specific individuals in a dataset but is only accurate over a relatively large sample.

This method originates in the fields of medicine and political science where the ability to infer individual characteristics from aggregate data has been paramount (Gelman 2016). In the medical field, this was necessary due to the nature of clinical trials, which requires the anonymization and therefore aggregation of clinical data even though the prevalence of a disease is an inherently individual trait. Similarly, in the field of political science, this was used to

understand the voting habits of individuals from differing demographic groups based upon aggregate, anonymous survey data.

Ecological regression is seldom used in the context of transportation, though there are some examples. One such example relates to traffic safety, where a Bayesian ecological regression model was created to analyze road mortality in Europe and Tunisia (Eksler 2008)(Kammoun 2020). Both of these studies utilize ecological regression to isolate the individual level characteristics that lead to traffic collisions as a way to determine risks associated with each geographic region. Both studies found that generally speaking, collision rates decrease as the population density increases, indicating that rural areas carry higher collision and fatality risk than urban areas. It is important to note that these studies utilize Bayesian ecological regression, which is a different method of ecological regression to the Goodman's Method of Bounds regression used here. Ecological regression has already been used in some cases to understand equity issues, such as the representation and distribution of benefits to toll facility users (Leung 2019). Ecological regression allowed this study to connect the benefits gained by using the toll facility to the demographics of those using the toll facility. It found that while the net benefits of using the toll facility were much higher for high income individuals, low-income individuals had a greater net gain per trip, indicating that this toll facility was generally equitable. Another example of ecological regression being used to study equity comes from Baton Rouge, LA, where ecological regression was used to study the socio-demographic and spatial effects on ridership of the transit system (Kuai 2020). Semi-parametric geographically weighted regression is used, which finds that neighborhoods with higher concentrations of non-white minorities, recent immigrants, and carless households most positively influence public transit ridership. This is one example

of many potential equity analysis scenarios which could benefit from a more direct understanding of the critical demographics being studied.

5.4 Using Ecological Regression to Determine Transportation Demographic Profiles

Taking advantage of ecological regression's ability to infer information about the individual level characteristics from aggregate datasets allows it to be utilized in the context of transportation. There are three general datasets that are required to complete this kind of analysis. First is the overall Origin-Destination (O/D) data for the region of study. This data is critical because it informs where the people of the region of study are travelling to and from. This is important because we must know the general travel patterns of transportation users to extricate the subset of users who travel on the network of interest and tie that to the critical demographics of interest. Typically, O/D data takes the form of a set of Traffic Analysis Zones (TAZs) from which the number of trips and modal split to and from each zone is known. The second critical source of data that is required is census data or other area aggregated demographic data. This information conveys the demographics of the region which is needed to determine the demographic profiles of overall transportation users and freeway network users. Finally, individual trip data is required to supplement the general O/D data. This information gives a better depth of detail that can be used to refine the regression model by showing individual level behavior as well as inform the determination of network usage. An example of this kind of data is a Household Travel Survey (HHTS) where a subset of individuals records all trips taken to provide more detailed travel information than that created by O/D models.

To apply this as a binary model to a transportation network, the first step is to define the usage of the transportation network in question. O/D data typically comes in the form of all trips, not specific trips on a network. Therefore, the number of trips using a specific network to and from each TAZ must be estimated if the O/D data does not contain network specific information already. Once the number of network trips is known, the ecological regression models can be built. These models take as the primary covariates the demographic characteristic of interest, with the outcome variable being the number of trips that use the network for those key demographics. The ecological regression model does require the assumption that the behavior across each TAZ and each income level is stable (i.e., the percentage of each income level that uses the network is constant). The model then outputs the odds ratio or the probability that the 'risk' for each demographic stratum to use the network is higher or lower by calculating the exponent of the individual and group coefficients. In this case an odds ratio of one would indicate an even likelihood that a stratum will use the network of study, while an odds ratio above one indicates a higher likelihood that a stratum uses the network, and vice-versa for an odds ratio lower than one. To find the total relative population of each income bracket that uses the freeway, the odds ratio can be multiplied by the overall population for each demographic characteristic of interest for each aggregated area and then further aggregated along that demographic to find the total demographic profile. A sensitivity analysis can then be conducted to understand how impactful the magnitude of the odds ratio is, above or below the value "one". To do this the percentage difference in population for each level of the demographic characteristics of interest is found. This can be calculated from the odds ratios found in the regression model, as the odds ratios directly indicate the representation of each stratum in the transportation network. From the percent change of each demographic stratum, the total

weighted percent change can be calculated to see how dissimilar the network of interest's demographic profile is from the overall demographic profile of the region.

5.5 Demographic Case Study Using Central Puget Sound Freeway Network

A case study was conducted on the freeway network of the central Puget sound region, including King, Pierce, Snohomish, and Kitsap counties, to validate the methodology described above. This area includes many major cities, including Seattle and Tacoma, as well as several urban and ex-urban freeways, most notably, I-5, I-405, SR 167, and SR 520. This region is home to over 1.6 million households. The data used in this study primarily comes from the four-county region MPO, the Puget Sound Regional Council (PSRC). The PSRC provided trip level O/D data, census data, and HHTS data (Puget Sound Regional Council 2022). Each dataset provides a key piece of information that is necessary for building these models; the O/D data provides information on the number of people travelling between each TAZ, the census data provides demographic information of the income distribution across the region, and the HHTS provides more detailed information on individual trips to supplement both the O/D and census data. Several cleaning steps were taken to ensure the data is suitable for the model. This includes aggregating the O/D data from the TAZ level to the census tract level and mitigating discrepancies with the census data. Since TAZ's have a smaller resolution, it is necessary to assign the census data to each TAZ within each census tract (Puget Sound Regional Council 2022). Additionally, based upon the format of the data, household income was split into 10 brackets: less than $10,000, $10,000 to $14,999, $15,000 to $24,999, $25,000 to $34,999, $35,000 to $49,999, $50,000 to $74,999, $75,000 to

$99,999, $100,000 to $149,999, $150,000 to $199,999, and $200,000 or above. The overall demographic profile of the central Puget sound region is included in Figure 5.1.

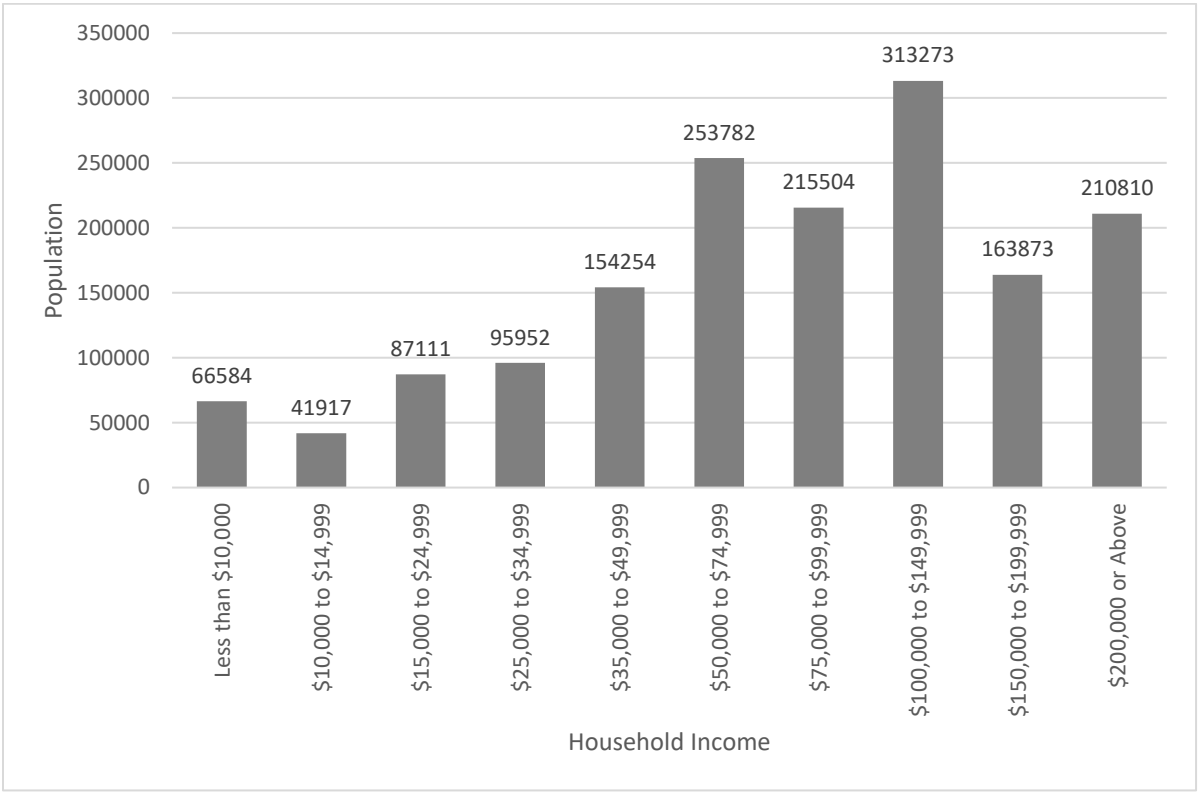

Figure 5.1: Overall Demographic Profile of the Central Puget Sound Region

Following the methodology outlined above, first the trips that utilize the freeway must be determined. Confidence intervals are used on the HHTS data to estimate freeway use based on the speed of each trip. 97.5% confidence is achieved in determining that a given trip did or did not use the freeway. Next, an ecological regression model is built to infer the demographics of the subset of the population that utilized the freeway. The income distribution for each census block group, the primary demographic of interest, was used as the covariates for this model, with the outcome being a probabilistic output along the binary 'did this trip use the freeway or not'. From the odds

ratios indicating this probability, the demographic profile of the freeway network can be aggregated. The models indicate that the demographic profile of freeway users throughout the Central Puget Sound Region was very similar to that of the overall region. Figure 5.2 compares the overall demographic profile to that of the freeway users of the region.

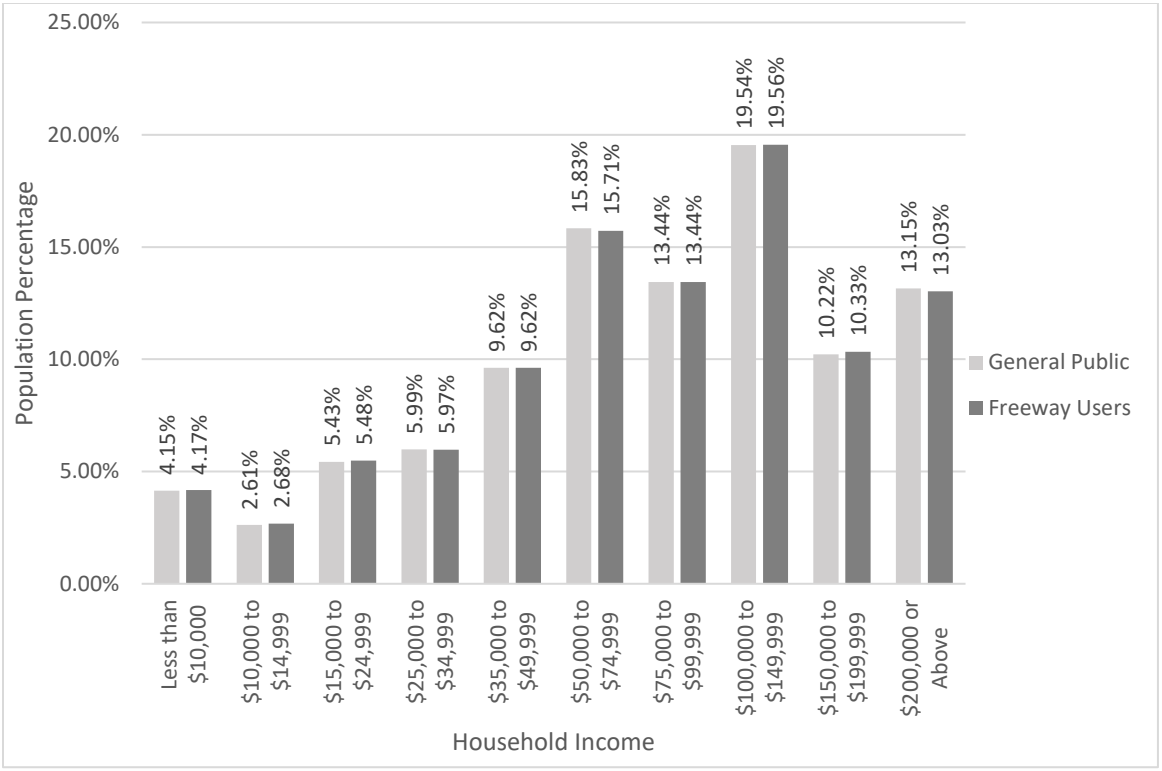

Figure 5.2: Demographic Profile Percentage Comparison of General Public and Freeway Users

Visually, all of the demographics for each income bracket look similar. This is due to our model providing odds ratios that are very close to 1. Because odds ratios are multiplied by the overall population for each income bracket, the population statistics are to be very similar to that of the overall population. The model results are included in Table 5.1.

Table 5.1: Odds Ratios and Ecological Regression Model Results

Income Level	Odds Ratio of the Income Bracket using the Freeway	Statistical Significance using 95% Confidence Interval
Intercept	2.30560	0.0068
Less than $10,000	1.00430	0.0159
$10,000 to $14,999	1.02366	0.0154
$15,000 to $24,999	1.00908	0.0158
$25,000 to $34,999	0.99796	0.0161
$35,000 to $49,999	0.99957	0.0160
$50,000 to $74,999	0.99270	0.0162
$75,000 to $99,999	1.00018	0.0160
$100,000 to $149,999	1.00105	0.0160
$150,000 to $199,999	1.01051	0.0157
$200,000 or Above	0.99114	0.0163
-2 Log Likelihood: 1352545		

Though visually, the proportion of "General Public" and "Freeway Users" of each bracket in Figure 3 appears to be similar, the true relationship between the two should be found. This allows for a more accurate understanding of how much the demographic profile changes between freeway users and the general population. Even if the odds ratios of an ecological regression model are close to 1, the overall change depends on the magnitude of the variable being measured (i.e., an odds ratio applied to a large population will have a greater percentage effect than the same odds ratio applied to a smaller population). Since the total population of this study is relatively large

(over 1 million households), a sensitivity analysis must be conducted to show the percentage change to the demographic profile between the freeway and the overall population. To achieve this, the relative percent change of each stratum is calculated, from which the overall relative weighted percent change for the entire demographic profile is then calculated. The results of the sensitivity analysis are found in Table 5.2.

Table 5.2: Relative Percent Change of the Demographic Profile for Freeway Users

Income Level	Relative Percent Change in Freeway use form General Travel
Less than $10,000	0.4185%
$10,000 to $14,999	2.3542%
$15,000 to $24,999	0.8968%
$25,000 to $34,999	0.2154%
$35,000 to $49,999	0.0536%
$50,000 to $74,999	0.7406%
$75,000 to $99,999	0.0066%
$100,000 to $149,999	0.0943%
$150,000 to $199,999	1.0401%
$200,000 or Above	0.8973%
Total Weighted Percent Change: 0.5066%	

5.6 Chapter Summary

In this chapter, we cover an introduction to the methodology of ecological regression and show how it can be utilized to determine the demographic profile of a transportation network. Throughout we 1) provide a thorough introduction to ecological regression such that the applications of the methodology are clear and 2) show how this methodology can be used to determine the demographics of transportation networks. We follow this discussion with a case study to determine the demographic profile of the Central Puget Sound Region freeway network. From the results of the analysis, the demographic profile of freeway network users is essentially the same as that of the overall population in the Central Puget Sound Region. It was found that there was an overall 0.5066% dissimilarity in the two demographic profiles. Overall, this result is reasonable. Given that freeway use is such a ubiquitous part of the American transportation landscape, it is conceivable that the income breakdown of freeway users is proportional to that of those who live in the same geographic area. More importantly, this shows how ecological regression is a valid methodology for determining the demographic profile of freeway networks. Further, the flexibility of ecological regression allows its application to estimate the income distributions of many other transportation networks and can also be used for covariates other than income, such as race.

Chapter 6: Applications of Ecological Regression for Washington Tolling Data

6.1 Overview

Tolling facilities on roadways are increasingly being used as tools to address the increasing strain on our roadway networks. They can be used to achieve many goals, including congestion mitigation/reduction, revenue generation, and as funding mechanisms for infrastructure projects. Proponents of toll facilities champion tolls as true user-pay facilities where costs incurred are directly born by the facility users (Krol 2017). Opponents of tolling facilities cite potential equity concerns where these incurred costs, including monetary, accessibility, pollution, etc. can impact certain users much more than others (King 2009)(Peseky 2018). Generally, both of these ideas are true, where the progressive- and regressive-ness of the tolling facility can vary greatly based on unique policy decisions and contexts of each facility (Burris 2013)(FHWA 2008). Therefore, it is critical to evaluate the revenue generation policies of each facility to ensure that the benefits and costs are being distributed fairly across the community.

6.2 Contributions and Chapter Structure

In this chapter, ecological regression is used to determine the demographic profile of users based on income of the 5 Washington State Department of Transportation (WSDOT) tolling facilities. We begin with a description of the transaction data and methods used to calculate these demographic profiles, followed with the model results. With this information we can show how

use of the facilities differs both across income demographics and geographically. The contributions of the chapter are as follows:

- We provide a background on tolling transaction data characteristics.
- We utilize ecological regression to determine the demographic profile of users for each toll facility.
- We show how these users are distributed geographically throughout the Central Puget Sound Region.

This chapter is organized as follows. Section 6.3 provides background on the tolling transaction data and supplemental data required to conduct the ecological regression analysis. Section 6.4 highlights the ecological regression model methodology and results. Section 6.5 visualizes the use of the facilities based upon the results of the model. Finally, Section 6.6 concludes the chapter with a summary and concluding remarks.

6.3 Background of Tolling Data

Currently there are 5 toll facilities in the State of Washington: the I-405 Express Toll Lanes (ETLs), the SR 167 High Occupancy Toll lanes (HOT Lanes), the SR 520 tolled floating bridge, the SR 99 tolled downtown Seattle tunnel, and the SR 16 Tacoma Narrows Bridge. All of these facilities are located in the Central Puget Sound Region (also known as the 4-county region) of the state of Washington, largely centered around the Seattle area. Figure 6.1 shows the location of all toll facilities in the region.

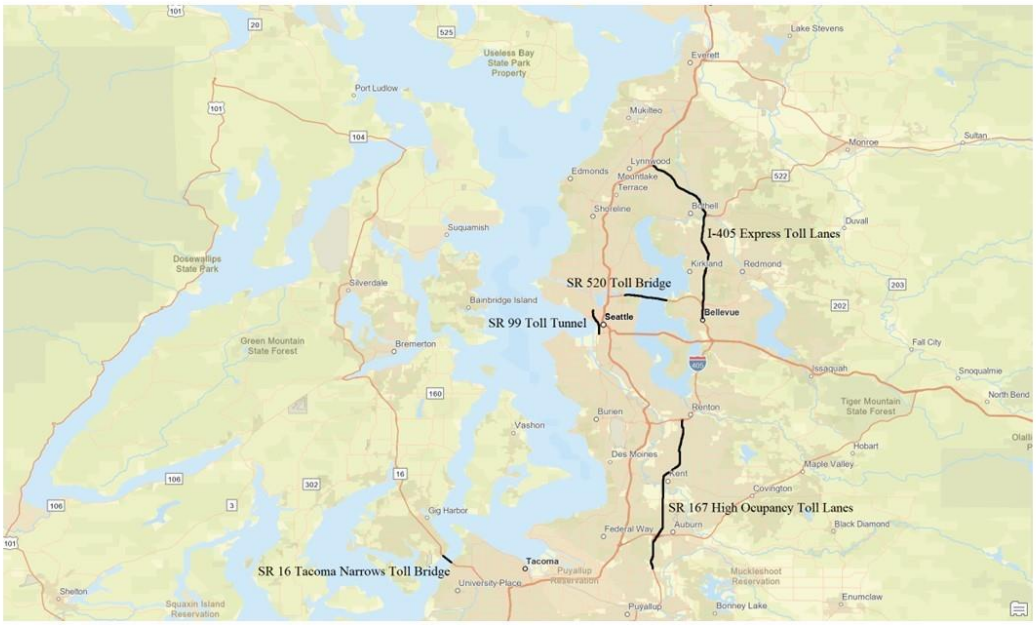

Figure 6.1: WSDOT Toll Facilities

To create an ecological regression model using the toll transaction data between July 2021 and June 2022 to predict the income demographic profile of toll users, we must have demographic data. Census data in the form of the American Community Survey (ACS) from 2020 is used for this objective. This provides information about the household incomes of each census block group (CBG) for the entire state. Figures 6.2 and 6.3 showcase the number of households per income level per census block group. For simplicity, though the census data includes 16 distinct income levels, in graphics figures which showcase differentials in income levels only four brackets will be shown: up to $49,999, from $50,000 to $99,999, from $100,000 to $149,999, and $150,000 and above. The method used to create the color scale for all figures used herein is the Jenks Natural Breaks Method (see ESRI 2022 for more details).

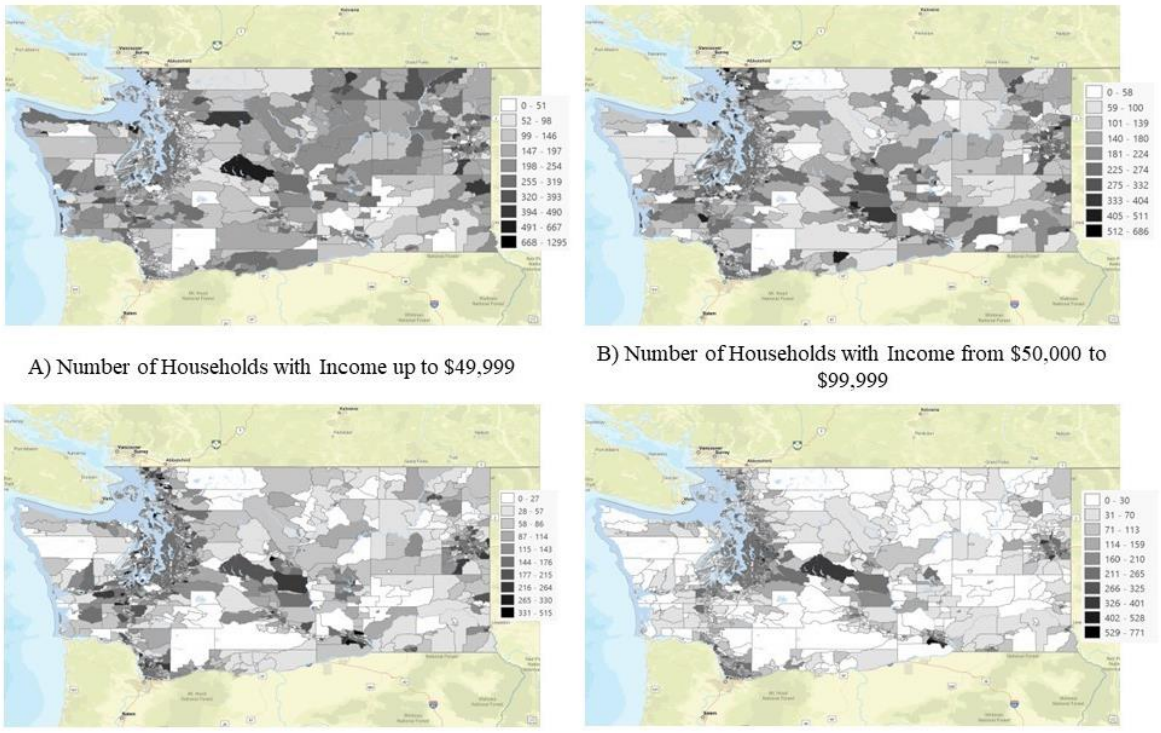

Figure 6.2: Income Brackets for the State of Washington

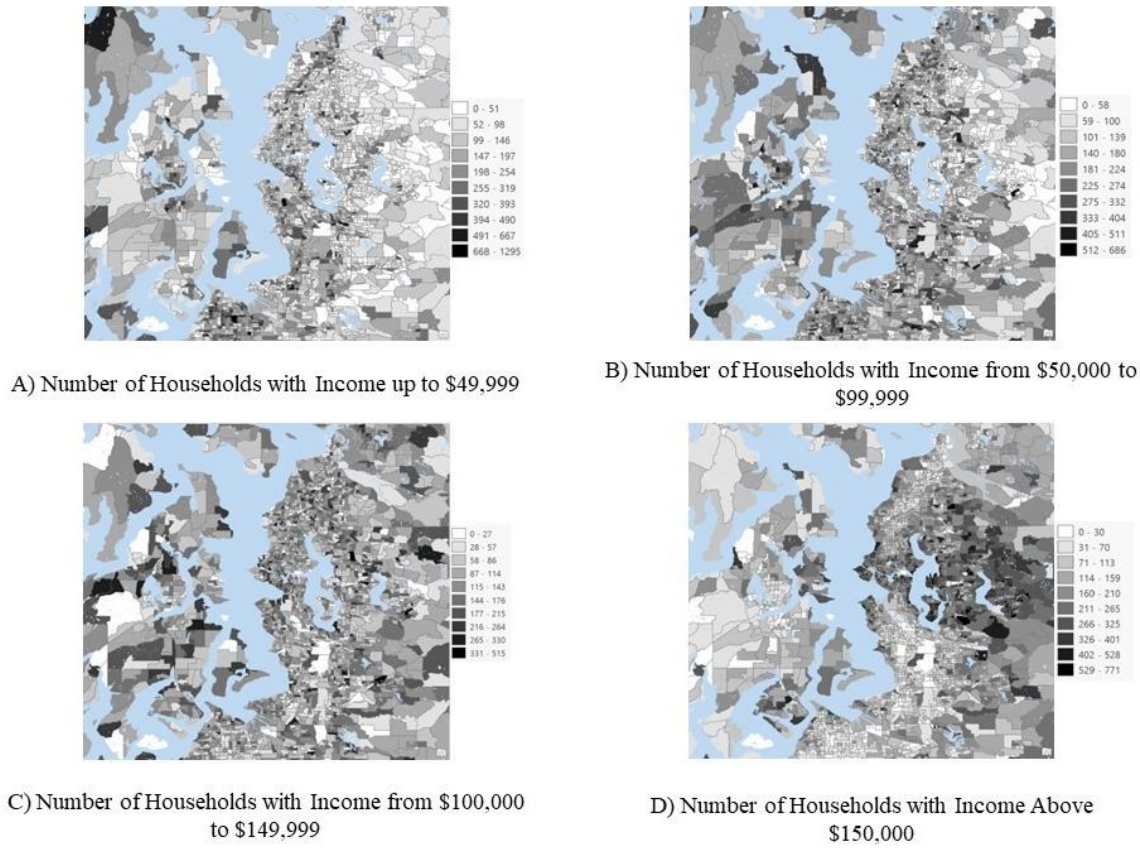

Figure 6.3: Income Brackets for the Central Puget Sound Area

From these figures, we can begin to see general trends in the income distribution for the state. First, generally speaking, CBGs east of the cascade mountains tend to have more lower income households and fewer high-income households than CBGs west of the mountains. This makes sense as east of the cascades tends to be more rural while the west tends to be more urban, which correlates to lower and higher incomes. When looking more closely at the Central Puget Sound Region, the closer to downtown Seattle, downtown Bellevue, or the eastside of Lake Washington, the greater the number of higher income households. There is an especially large concentration of the highest quadrant of income on the eastern side of lake Washington.

We can also look at the concentration of non-commercial accounts (non-commercial accounts were defined as having 6 or less toll tags associated with the account) that utilized the toll facilities at least once between July 2021 and June 2022 and is located in the state and/or Puget Sound Region. Overall, in this timeframe there were 51,239,599 trips made over all of the facilities. Figure 6.4 shows the total number of accounts for the entire state. Figure 6.5 shows a zoomed in view of the number of accounts per CBG in the Central Puget Sound Region. In these graphics, "accounts" includes both Good-To-Go toll accounts maintained with WSDOT and the registered address associated with license plates billed as a result of those vehicles having been identified as using at least one of the WSDOT toll facilities, but not having a Good-To-Go Account.

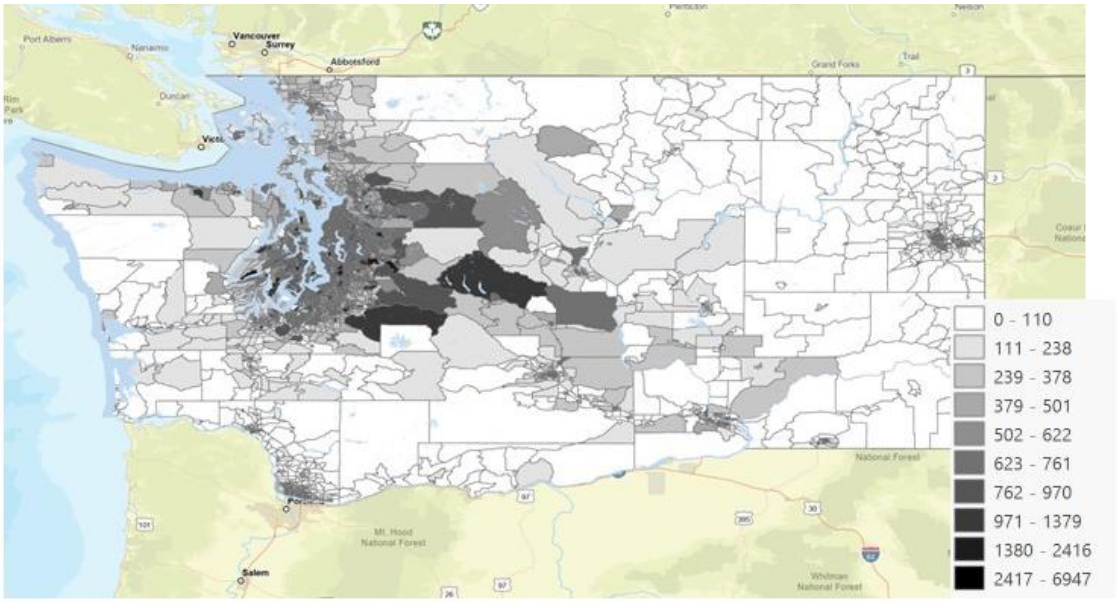

Figure 6.4: Toll accounts for each CBG Across the State

Figure 6.5: Toll Accounts for each CBG in the Central Puget Sound Region

From this figure we can see some important trends. Unsurprisingly, the number of accounts located in the eastern half of the state is low. This is due to the fact that all of the toll facilities are located in the western half of the state, thus making it unlikely that residents in those households will make trips on one of the toll facilities. Therefore, all upcoming figures that depict the use of each individual facility will focus on the Central Puget Sound Region. This will mean that not all trips are visualized, however the number of trips missing from the graphics is relatively small. One important note about these not visualized trips from Eastern Washington is that they include a higher proportion of trips from lower income households due to the fact that these outlying areas from the urban center tend have more lower income and low frequency users (only 1 or 2 uses over the year) of the toll facilities. This intuitively makes sense, as people from outside the metro area are likely to only need to make one or two trips across the tolled road network during the year and are less likely to understand the alternative routes, they could use to avoid the tolls.

Additionally, they may also be more reliant on routing services such as Google Maps which are more likely to route those users across toll facilities due to the calculated time savings. Generally speaking, most of these trips form the eastern portion of the state use either the I-405 Express Toll Lanes or the SR 520 bridge, which again makes sense as these are the facilities most easily accessible form the primary route across the mountains (I-90) that do not have a bypass (SR 18 bypasses the SR 167 toll lane from I-90).

From Figure 6.5 we can also see several apparent anomalies in the distributions of accounts where some areas show unusually high numbers of accounts in outlying areas from the urban core. There are several reasons for these anomalies. First and foremost, most occur in outlying rural areas with small towns. In these areas the CBGs are scaled for rural areas but represent a relative concentration of households. For example, towns such as Issaquah, North Bend, Duvall and Monroe all show areas of concentration of accounts where in these cases, unlike in the urban core where CBGs are more closely scaled based on population, the CBGs are more reflective of the low population density in the immediate vicinity of these towns, leading to what appears to a relatively high concentration of accounts per area, though use patters per household are relatively stable. Another cause of these anomalies is the phenomenon where individuals will register vehicles at a secondary or vacation home to avoid paying vehicle tab fees associated with more urban areas. In these cases, accounts will appear as if they exist in these outlying areas though in reality the primary home of the user is located more within the urban areas. Finally, there is also likely to be higher levels of accounts in outlying areas on the Kitsap Peninsula (left part of Figure 6.5). In order to reach any of the major urban centers (Seattle or Tacoma), users must cross the Tacoma Narrows Toll Bridge with no viable alternative. This will naturally lead to higher levels of utilization

compared to other outlying areas as users are much more likely to use a non-toll alternative when entering the city.

Overall, the general distribution of household income and toll accounts is about what we would expect for the state of Washington with denser concentrations of accounts in CBGs that are close to the five toll facilities and decreasing numbers of accounts as those CBGs get further away from the toll facilities. Multiple CBGs with large numbers of accounts are noticeable centered around the SR 167 toll facility, and on the Kitsap peninsula, north of the Tacoma Narrows bridge. One somewhat surprising geographic area with a high number of accounts is centered on I-90, east and southeast of Issaquah.

6.4 Tolling Ecological Regression Model and Results

In this study, we can utilize ecological regression for determining the demographic profiles of the users of each facility. This is an optimal regression technique to use because of its ability to infer characteristics at an individual level; the choice to use a toll facility is done on an individual level and is determined by individual characteristics of each user. Though there are many characteristics that can influence an individual's choice to use a toll facility, the characteristic that this regression model assesses is household income. All of the toll transaction data we gathered from the use of the facilities only provides information on the geographic location of the residence that pays a toll; we have no direct information connecting any toll payment account with a particular income level. We draw our household income data from the census data described above. Critically, this data is aggregated geographically into CBGs as opposed to being aggregated by income level. This has convenience as it can directly connect with the geographic home locations

of the toll users, but there is no direct connection between each particular user and their income level, since we know that income does influence and individual's choice to use a toll facility beyond simple geographic delimitations (Leung 2019). Therefore, ecological regression allows us to extract the individual level characteristic of income to probabilistically assign income levels of each toll trip based upon the characteristics of the home CBG of the account associated with that trip.

To extract the statistical demographic models for each facility, we must create a separate model for each facility utilizing only trips for that facility as travel behavior may not be similar between all facilities using the basic methodology described in Chapter 5. To build each model, we first must assign each account that used the tolling facility to the income demographic profile of its home CBG. This income profile then becomes the explanatory variables for our model such that it can predict how over- and under-represented each income strata is for the facility. This over- and under-representation can then be applied back to each trip taken on that facility as a probability that a trip was made by each income strata. When aggregated together, this provides a statistically accurate representation of the population that used the tolling facilities. It is important to note that these probabilities have little meaning on a trip level, i.e., it is inaccurate to say that a particular trip from a particular account, and therefore the household income of that account, was likely of a particular stratum. Instead, this provides insight into the overall profile for the facility. Once the number of trips for each income strata is calculated, the percentage of trips taken by each stratum on each facility can be calculated such that the relative use of each facility for each stratum can be directly compared as each facility has widely varying total number of trips. This also allows us to compare the overall demographic profiles of the region to understand differences between the total populace and those that use the toll facilities.

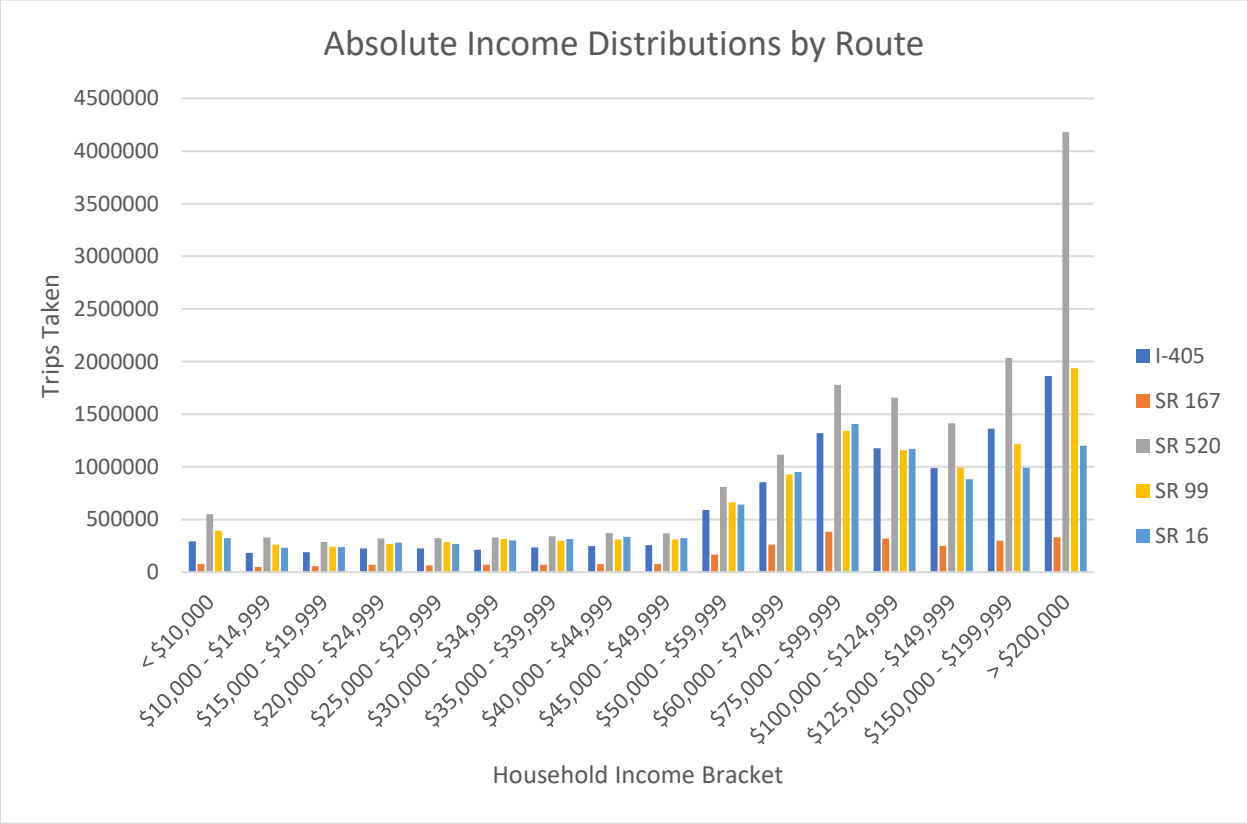

Figure 6.6: Absolute Distribution of Trips Taken on Each Toll Facility by Income

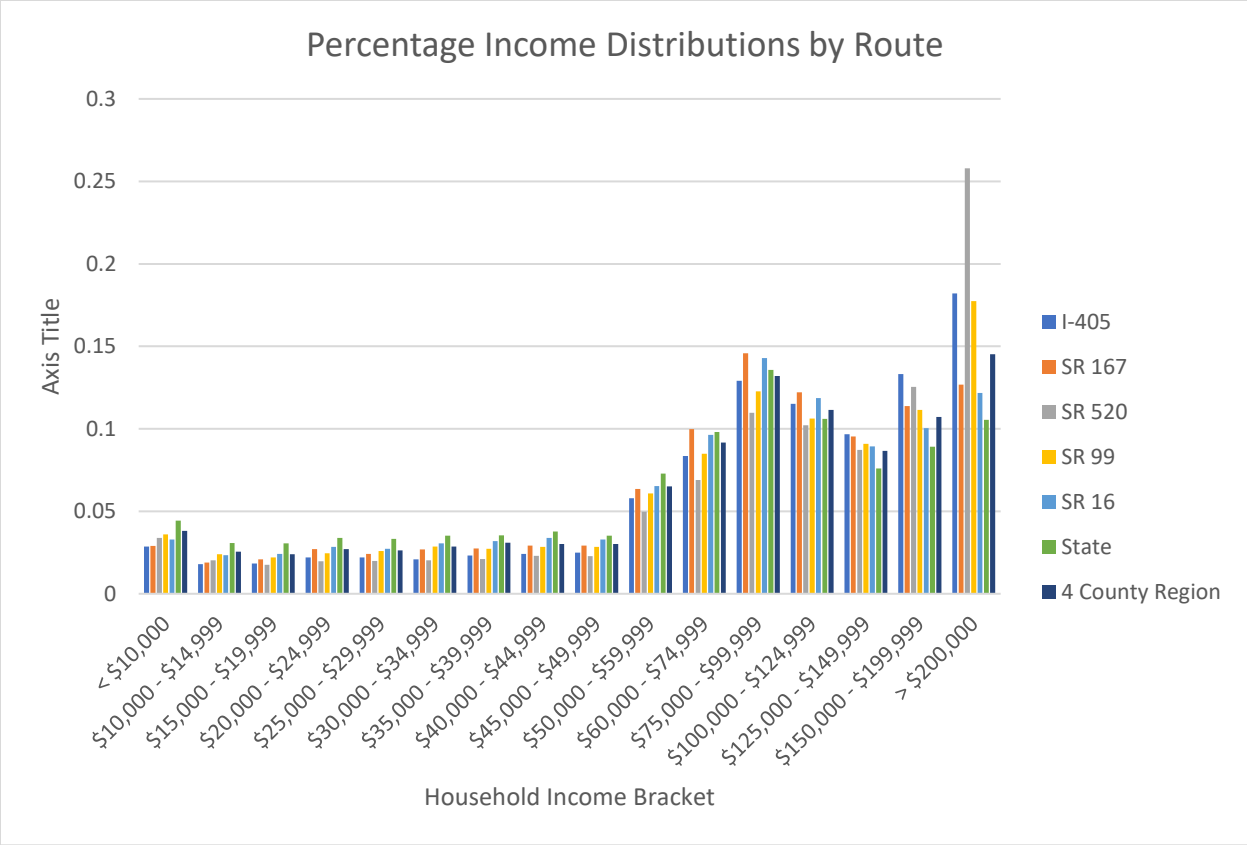

Figure 6.7: Percentage Distribution of Trips Taken on Each Facility by Income

The graphs in Figure 6.6 and Figure 6.7 show the calculated demographic outputs from the ecological regression models, where Figure 6.6 shows the absolute distribution of trips and Figure 6.7 shows these distributions as percentages for comparison purposes. In Figure 6.7, the demographic profile percentages for both the entire state and the 4-county region are included to compare the facility use back to the overall population. From these graphs, we can see several critical trends. Firstly, we can see that all routes follow the same general income distribution pattern of the state and 4-county regions. We do notice some discrepancies though. Firstly, SR 520 seems to skew to higher income levels, while SR 167 and SR 16 tend to skew to slightly lower income levels and SR 99 and I-405 have a middle range (though this does not change the greater

trend mentioned above). These are all slightly above the levels that we see for the state as a whole, but much more in line with the income levels of the 4-county region. These trends are what we expected to see given we know that SR 520 typically has higher income people living around the facility and SR 167 and SR 16 have generally lower income people living around the facility when considering the 4-county region. Furthermore, it is not a surprise that the state has overall lower incomes than all facilities as the eastern part of the state has lower incomes than the west, which has much lower utilization than the 4-county region. Overall, this shows that the model results are reasonable as the results generally follow all expected trends.

6.5 Toll Facility Use Visualizations

From the generated model, we can visualize how trips are distributed throughout the region onto each facility. Additionally, we can assess how these trips vary related to the income of users. Each subsection here addresses the results of one of the facilities.

6.5.1 I-405 Express Toll Lanes

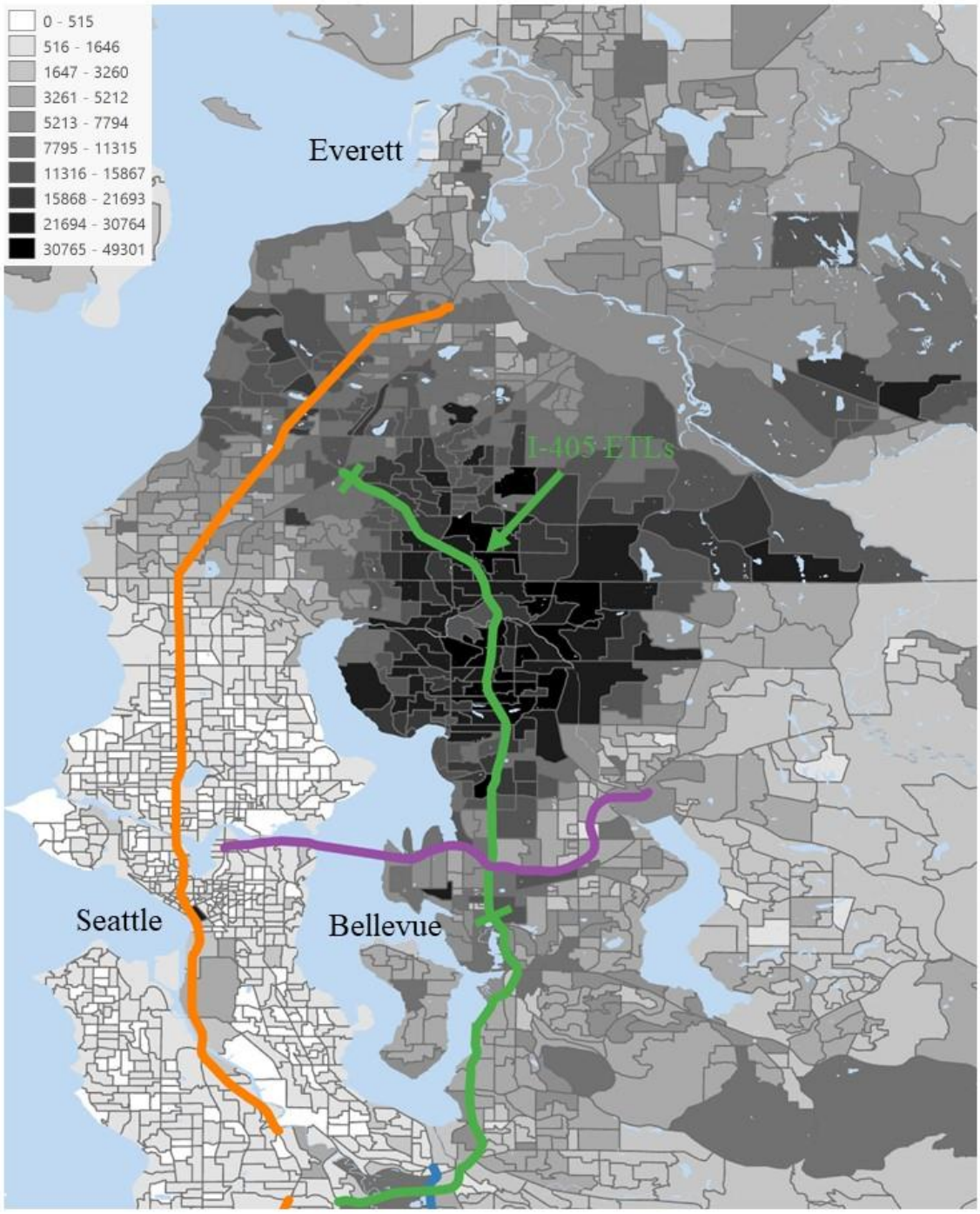

Figure 6.8: Trips Taken on the I-405 ETLs from each CBG

Figure 6.8 shows the number of trips that used the I-405 Express Toll Lanes by household location between July 2021 and June 2022. Unsurprisingly, this facility's primary users originate from along the corridor between Bellevue and Everett (the north section of I-405 and along I-5 from the junction of I-5 and I-405 to Everett). The I-5/I-405 corridor connects residents of the northern Puget Sound communities and the major eastside activity centers. Once on northern I-405, they can choose between I-405's general purpose (GP) lanes and the ETLs, based on many factors, including individual values of time and reliability, among others. Alternative routes for these travelers (mostly I-5 to I-90 or I-5 to SR 520) frequently take longer than using I-405. Figures 6.9 and 6.10 show the geographic areas where the majority of I-405 Express Toll Lane users reside and the income distributions of those CBGs.

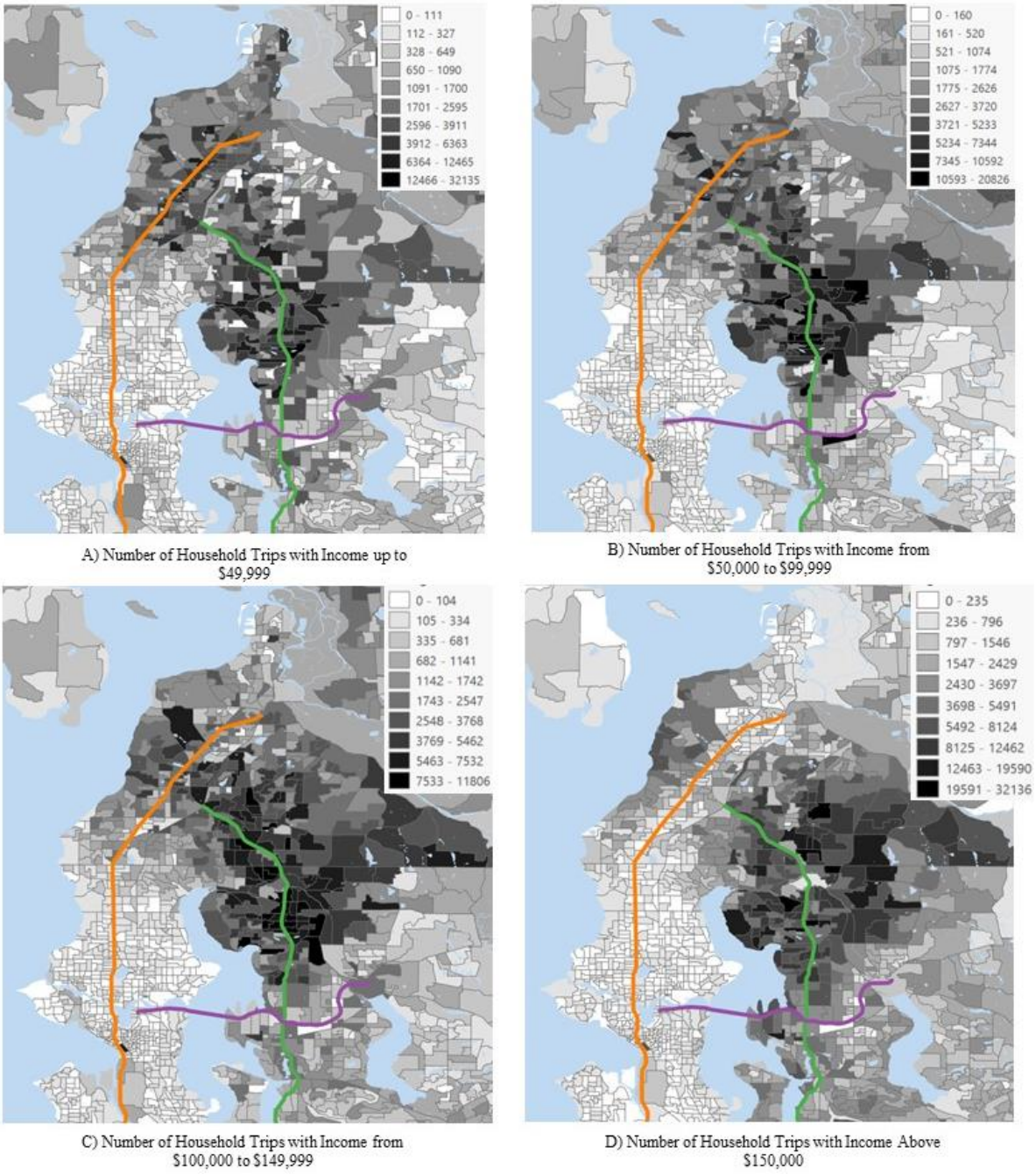

Figure 6.9: Trips Taken on the I-405 ETLs from each CBG Split by Income

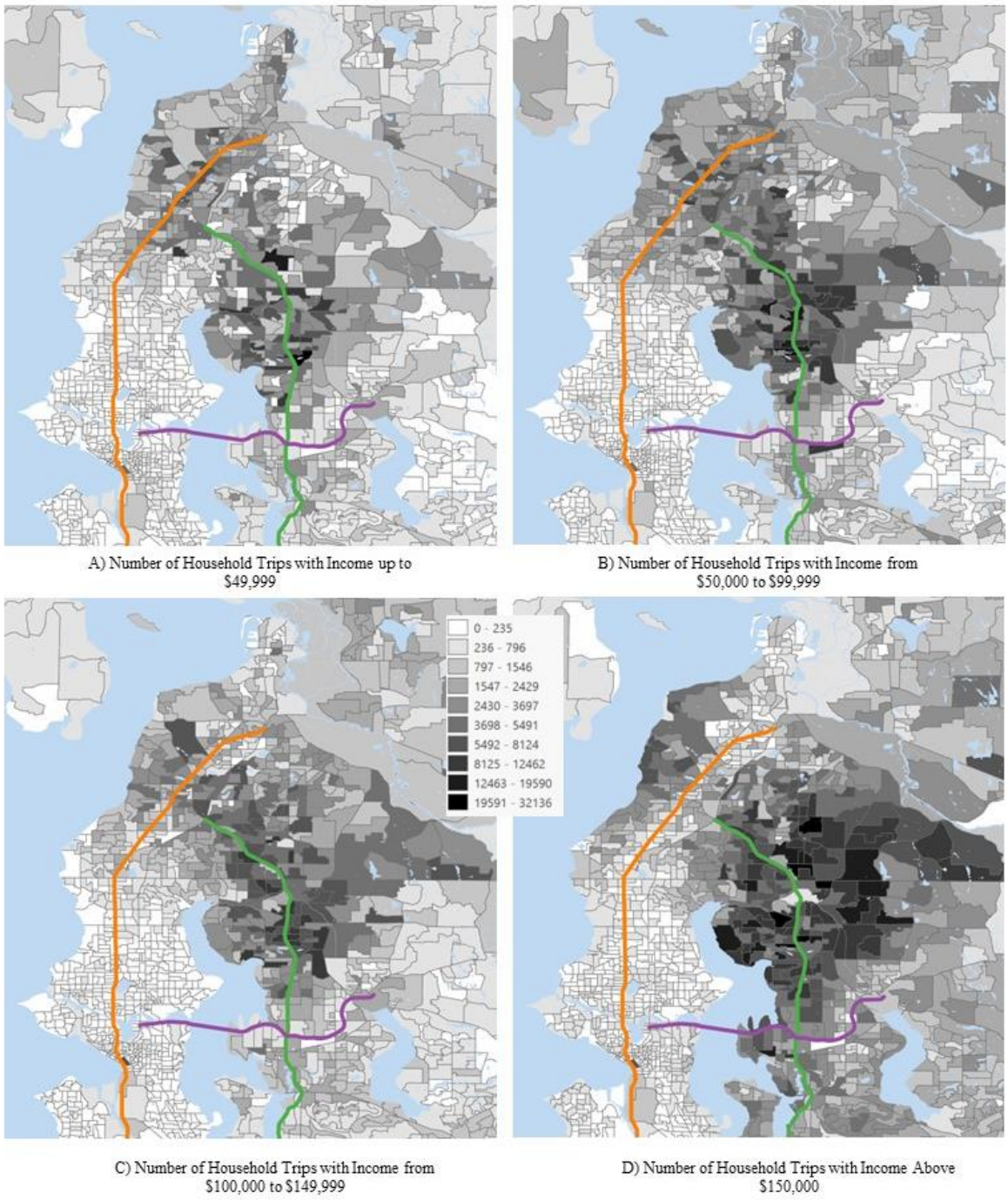

Figure 6.10: Trips Taken on the I-405 ETLs from each CBG Split by Income with Fixed Scale

Figure 6.9 shows the income distribution scaled individually for each income group, showing the distribution of trips within that income group while Figure 6.10 shows the distribution

of income groups at a fixed scale, showing the relative distribution between groups. These figures illustrate that the income levels of I-405 Express Toll Lane users generally follow the greater income trends of the region, with a higher number of lower income users being found in the north geographic area closer to Everett and further from Bellevue and the Eastside communities. In contrast, higher income users tend to skew towards Bellevue and the Eastside communities. This trend is most visible in the highest and lowest income groups, whereas the middle two income groups tend to be more homogenous. This follows our expectation from the ecological regression model results that there would be a relatively even distribution of trips on this facility overall with a slight skew towards higher income individuals.

6.5.2 SR 167 High Occupancy Toll Lanes

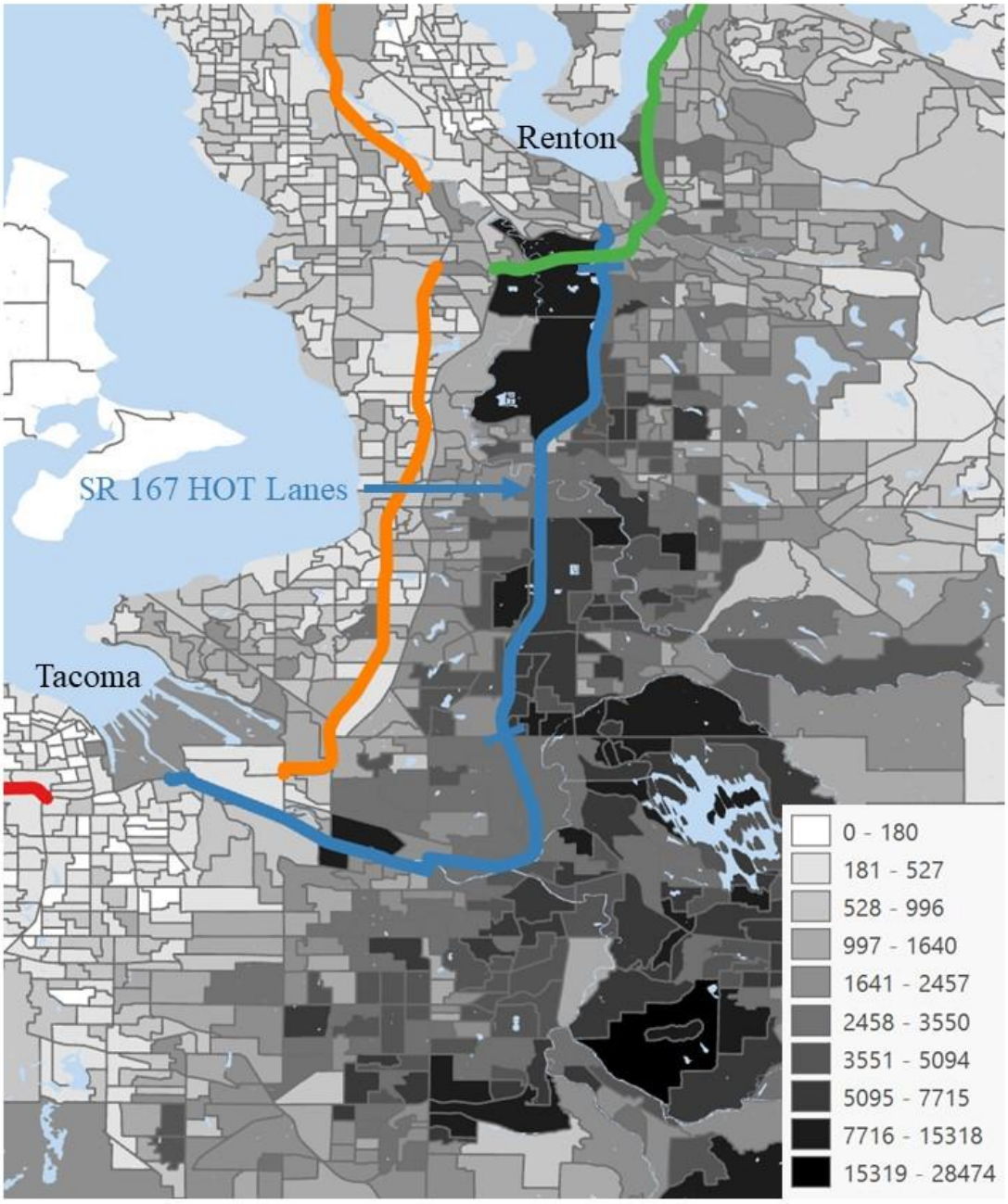

Figure 6.11: Trips Taken on the SR 167 HOT Lanes from each CBG

Figure 6.11 shows the trips made by accounts that used the SR 167 HOT Lane facility. The household location of these trips is concentrated in the central Green River Valley that SR 167

follows through the southern part of King County and the northern part of Pierce County. This generally agrees with our expectations for the use of this facility, as a large fraction of the users are expected to come from households near the facility. A slight note, for the dark sector at the bottom right of the figure though it appears there is a large number of trips originating from this area, this is in reality all one very large CBG. While there are significant trips generating form there, the trips per unit area is in line with the other smaller CBGs neighboring it. For the large area, those trips have been aggregated so there appears to be a higher proportion of trips from this area than there actually is.

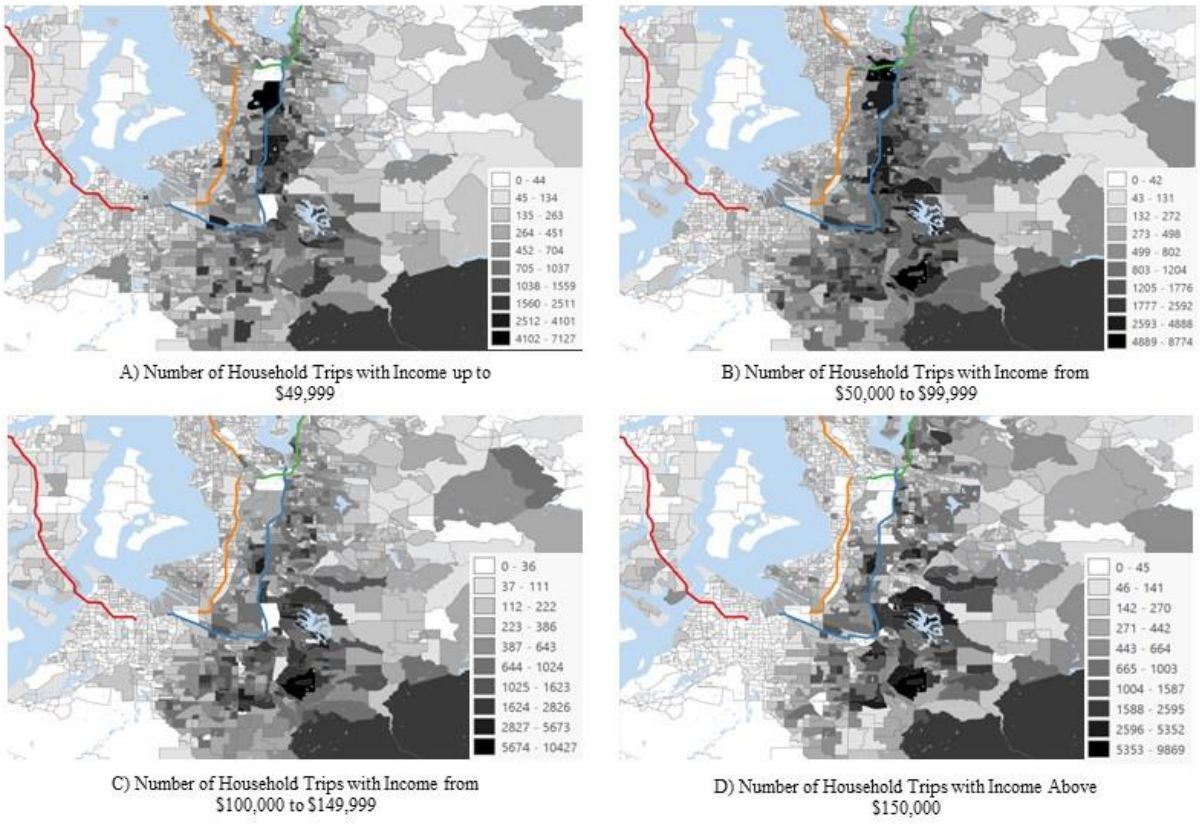

Figure 6.12: Trips Taken on the SR 167 HOT Lanes from each CBG by Income

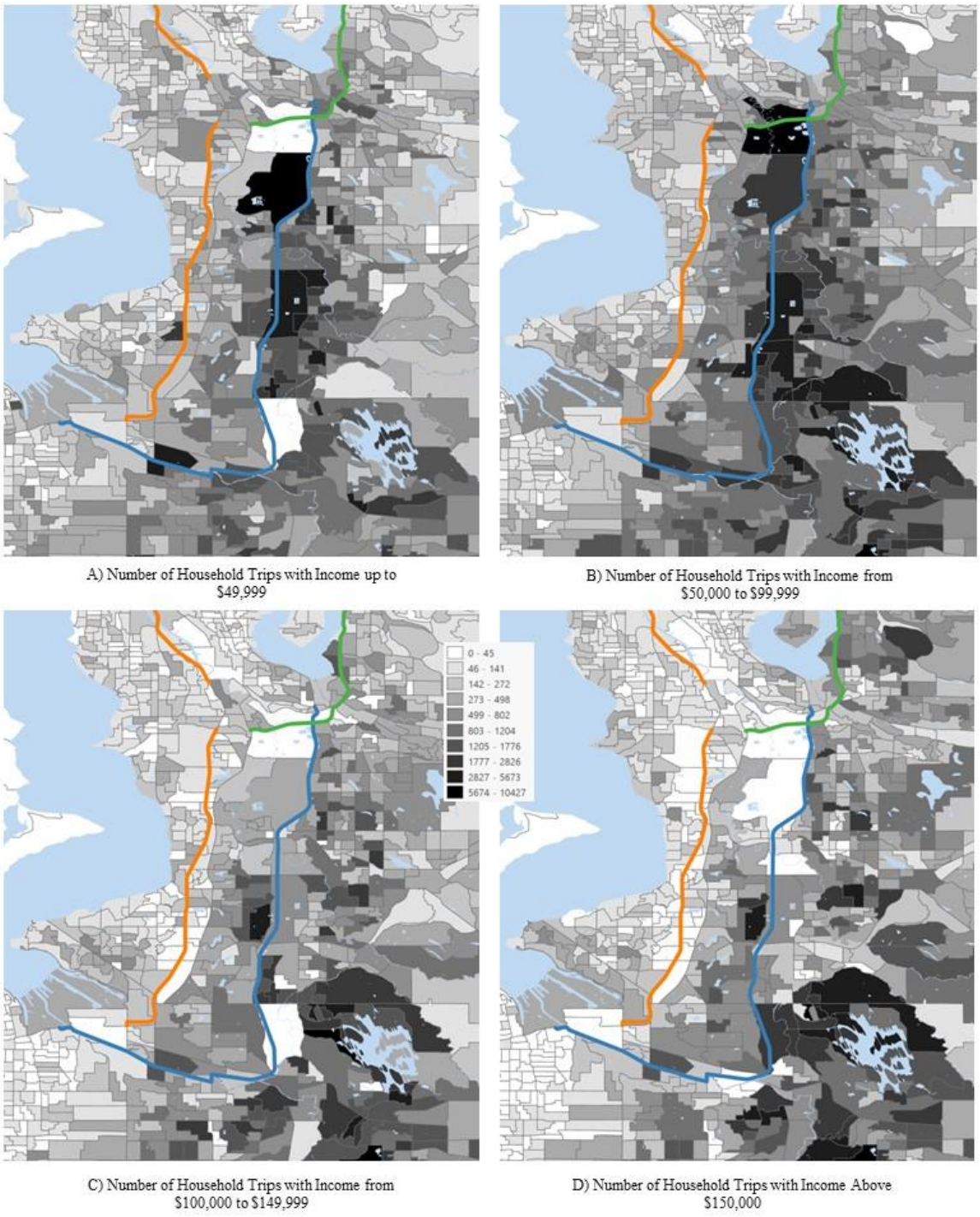

Figure 6.13: Trips Taken on the SR 167 HOT Lanes from each CBG by Income with Fixed Scales

Figures 5.12 and 5.13 show that the distribution of income is relatively even throughout the main catchment area for users, i.e., each income level is relatively evenly represented. This follows the finding of the ecological regression model where the demographic profile of the users of this toll facility is generally uniform compared to the surrounding area. This also is mirrored in Figure 15 showing the income quadrants with a fixed scale. Similar to I-405, we see that each income level is fairly homogenous, though this facility does have slightly higher concentrations of low-income households which is in line with the results of the model showing closer adherence to the overall region demographic profile.

6.5.3 SR 520 Tolled Floating Bridge

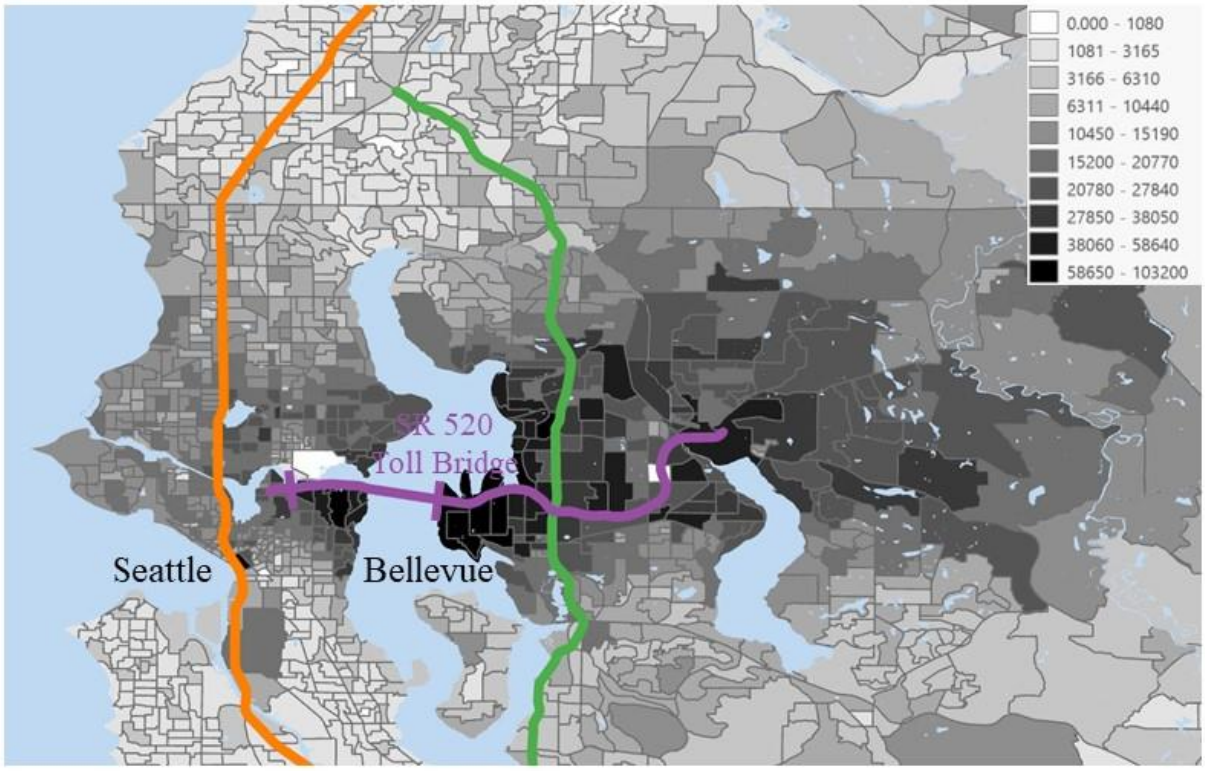

Figure 6.14: Trips Taken on the SR 520 Toll Bridge from each CBG

Figure 6.14 shows the number of trips made across the SR 520 Floating Toll Bridge based on the location of the account's registered address. This figure clearly shows the unsurprising domination of SR 520 users stretching along the SR 520 corridor. This makes sense as people living in these areas are going to gain the most time savings when crossing Lake Washington because of the additional time they must spend detouring to either SR 522 to travel north around the lake or south to I-90. Figures 6.15 and 6.16 show the income distributions for these users of the facility.

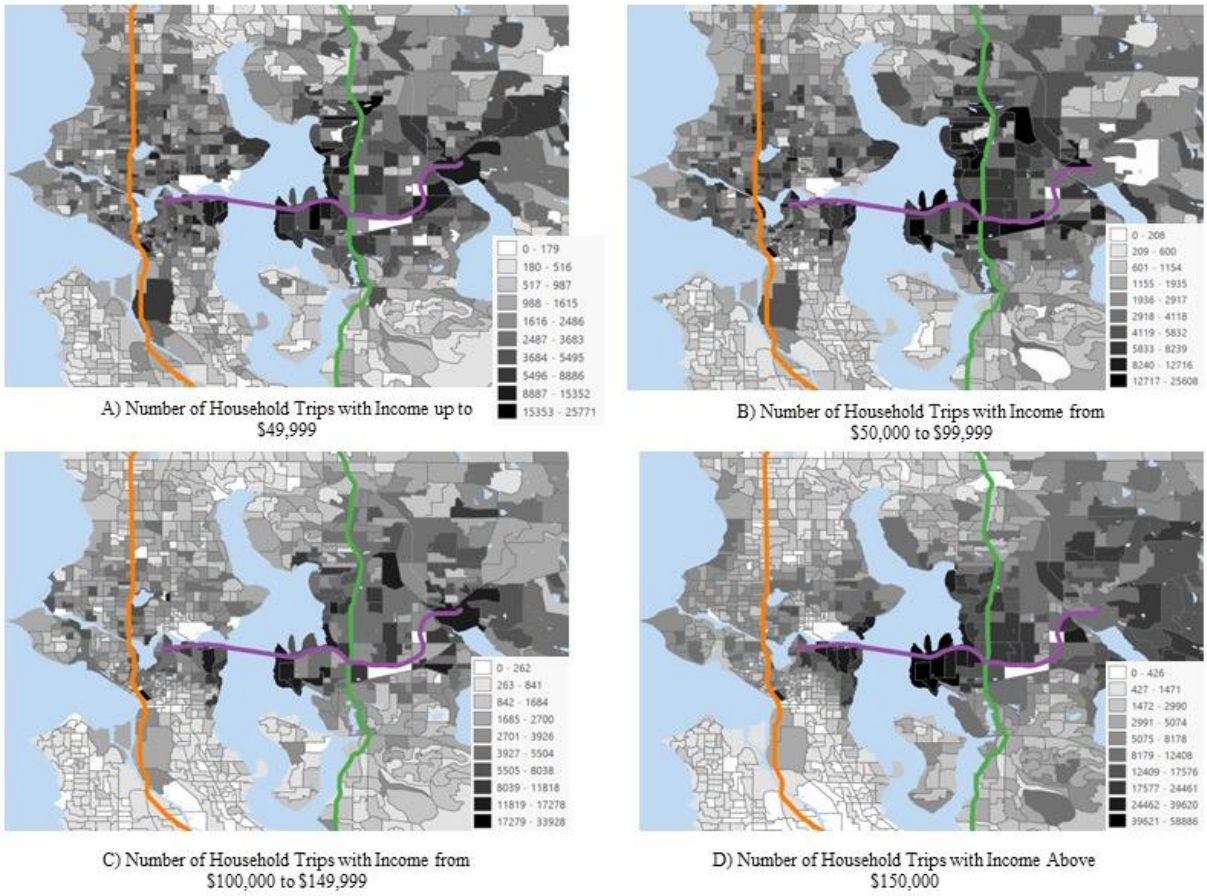

Figure 6.15: Trips Taken on the SR 520 Toll Bridge from each CBG by Income

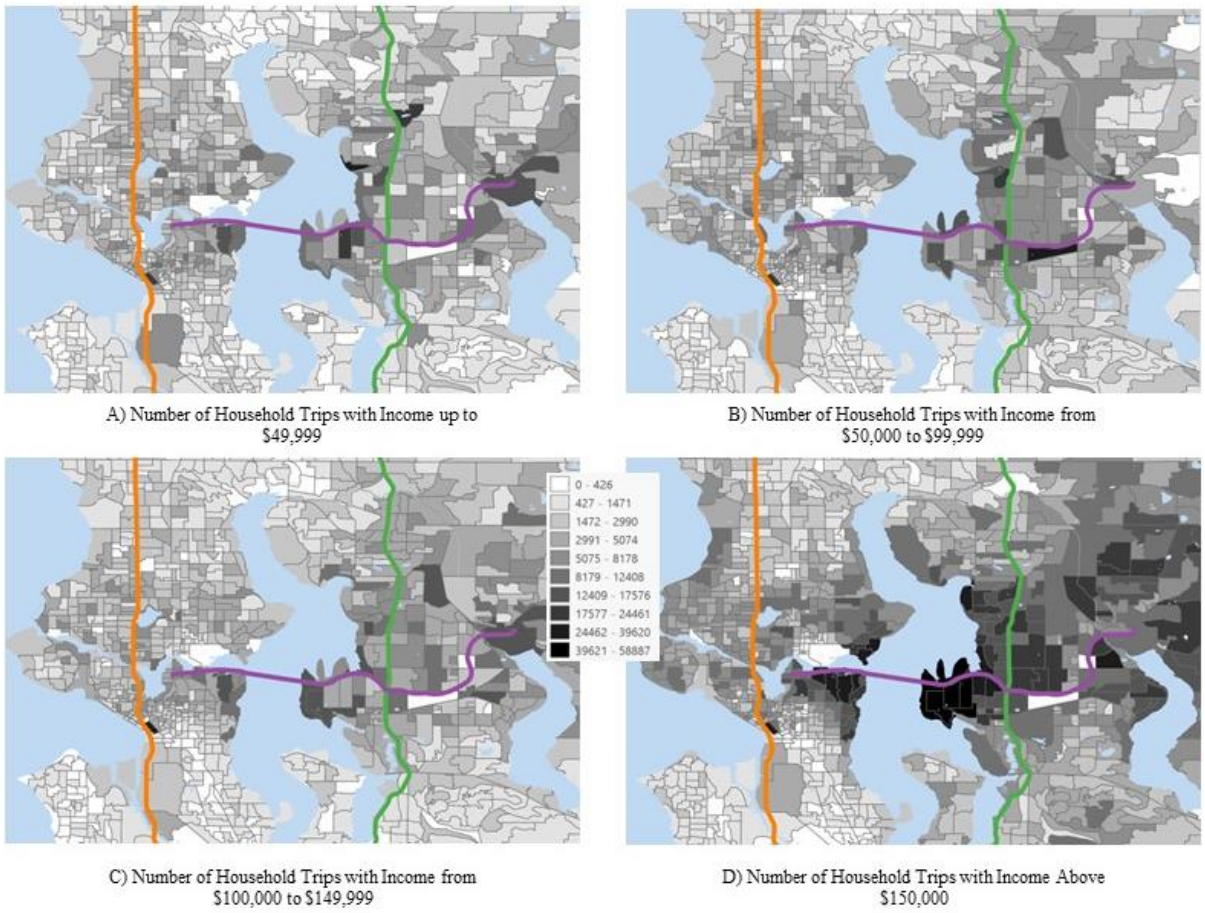

Figure 6.16: Trips Taken on the SR 520 Toll Bridge from each CBG by Income with Fixed Scales

From these figures, we see that the distribution of income is generally what would be expected for this facility. Recall from the income distribution of the trips of this route that this route, more than the other routes, tends to skew towards high income individuals. This can be seen where the distribution of trips for each income level is similar, but the higher income levels generally have 5x more trips per census block group across the SR 520 corridor. This makes sense as this facility bisects and connects some of the wealthiest parts of the region. This is especially evident in the highest income category, where there is a high concentration of trips from the area right next to Lake Washington. We can examine this trend further in Figure 6.16, which explores

the income quadrants on a fixed scale. From this image, we can see the higher rate of use from high-income users. However, we can also see that, even though they are more of a minority on this route, there are still many low-income users along the entirety of the corridor. This matches our expectation from the model results that SR 520 has the most skewed demographic profile of any facility towards high income users.

6.5.4 SR 99 Toll Tunnel

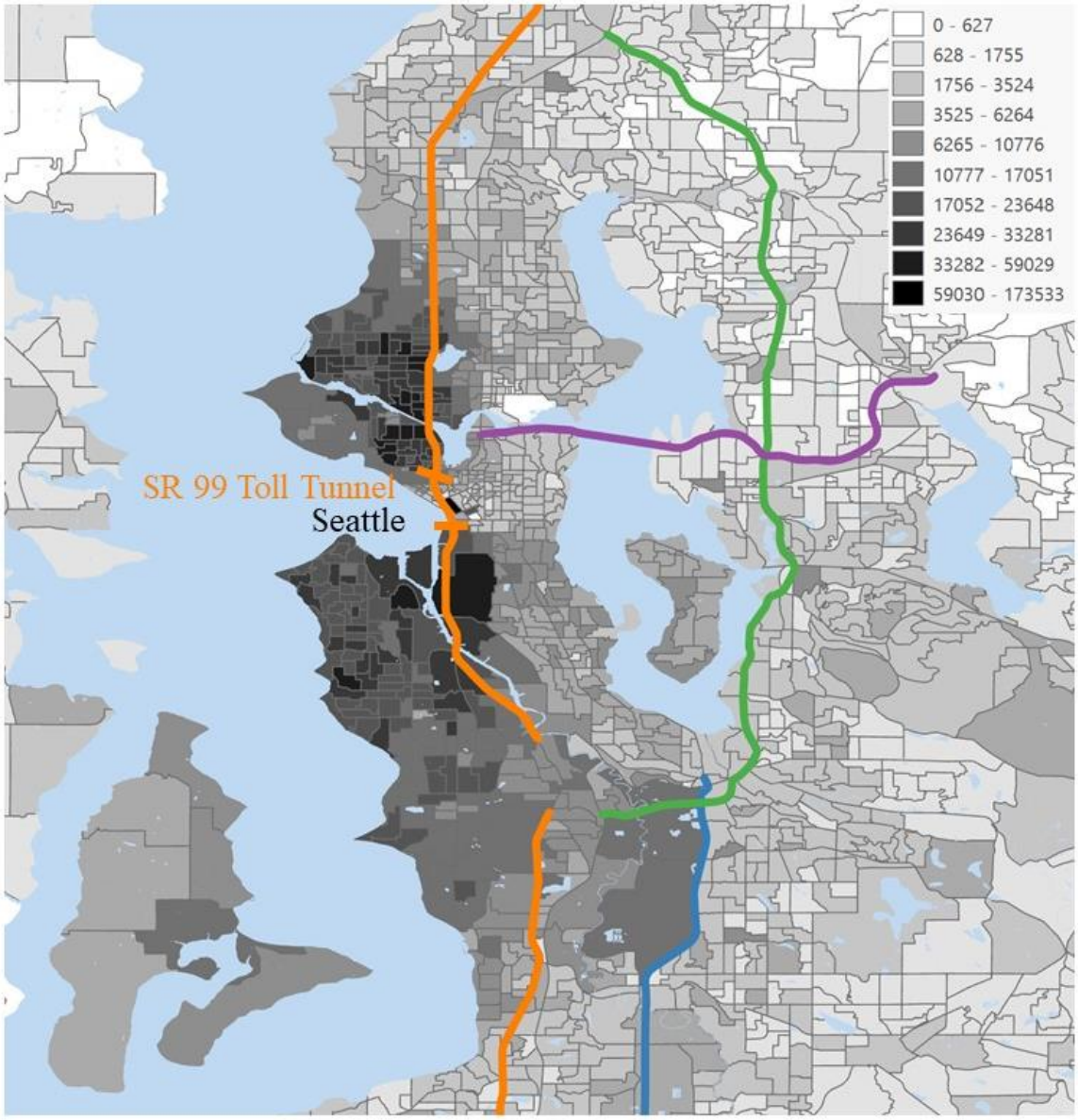

Figure 6.17: Trips Taken on the SR 99 Toll Tunnel from each CBG

Figure 6.17 shows the distribution of trips associated with accounts that use the SR 99 toll tunnel. Similar to all other facilities, the geographic distribution of the households that use the SR 99 tunnel is reflective of the service shed of the facility. The highest concentrations of use come from North Seattle, the Duwamish Valley, and West Seattle. This was exactly the target population

when the tunnel was designed and built (opened to the public in 2019). Other areas of the city are not highly represented because there are reasonable alternatives that do not show relative time added, usually using I-5 or local streets. Similarly, while SR 99 users come from all over the Puget Sound Region, the rest of the region has comparatively low usage.

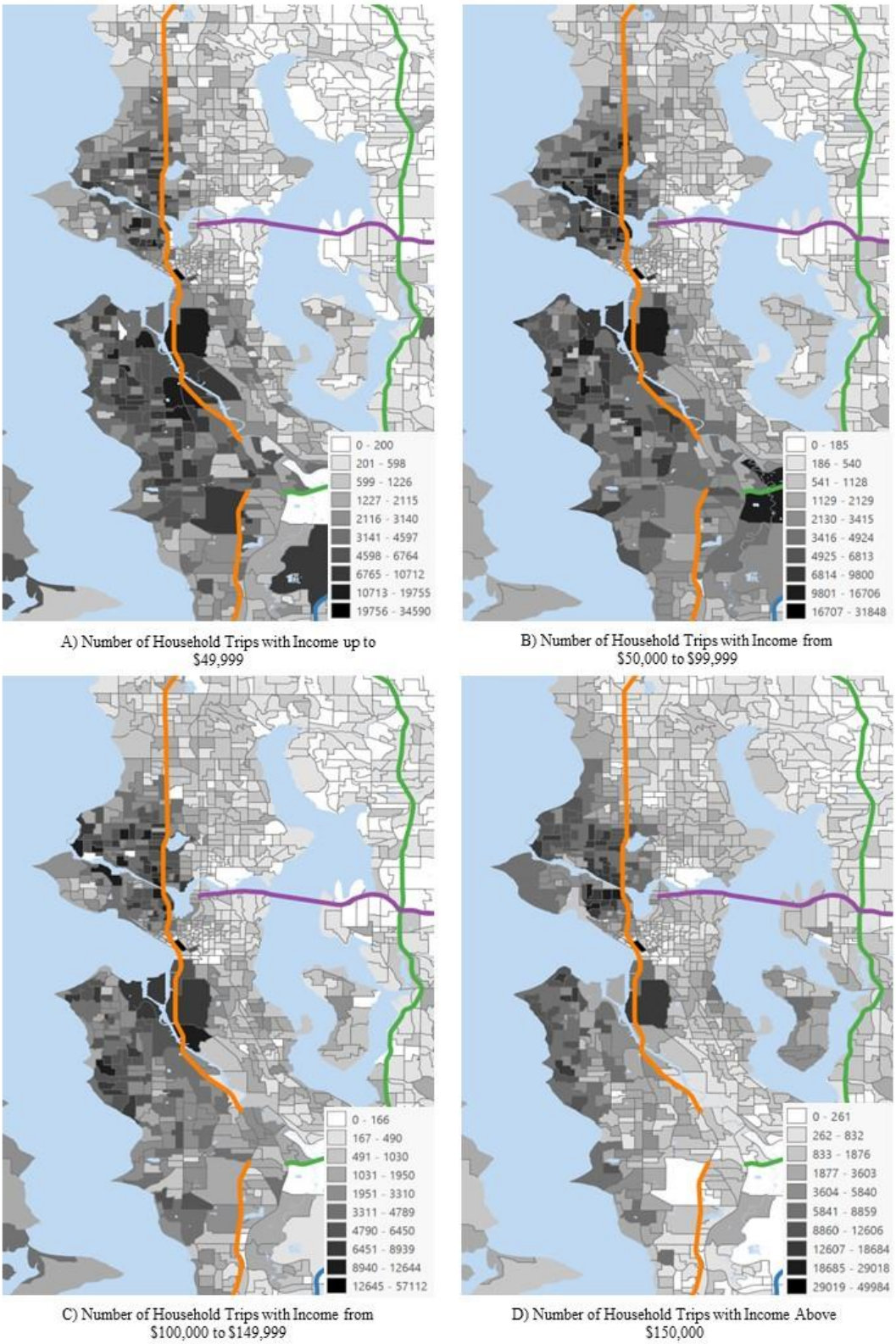

Figure 6.18: Trips Taken on the SR 99 Toll Tunnel from each CBG by Income

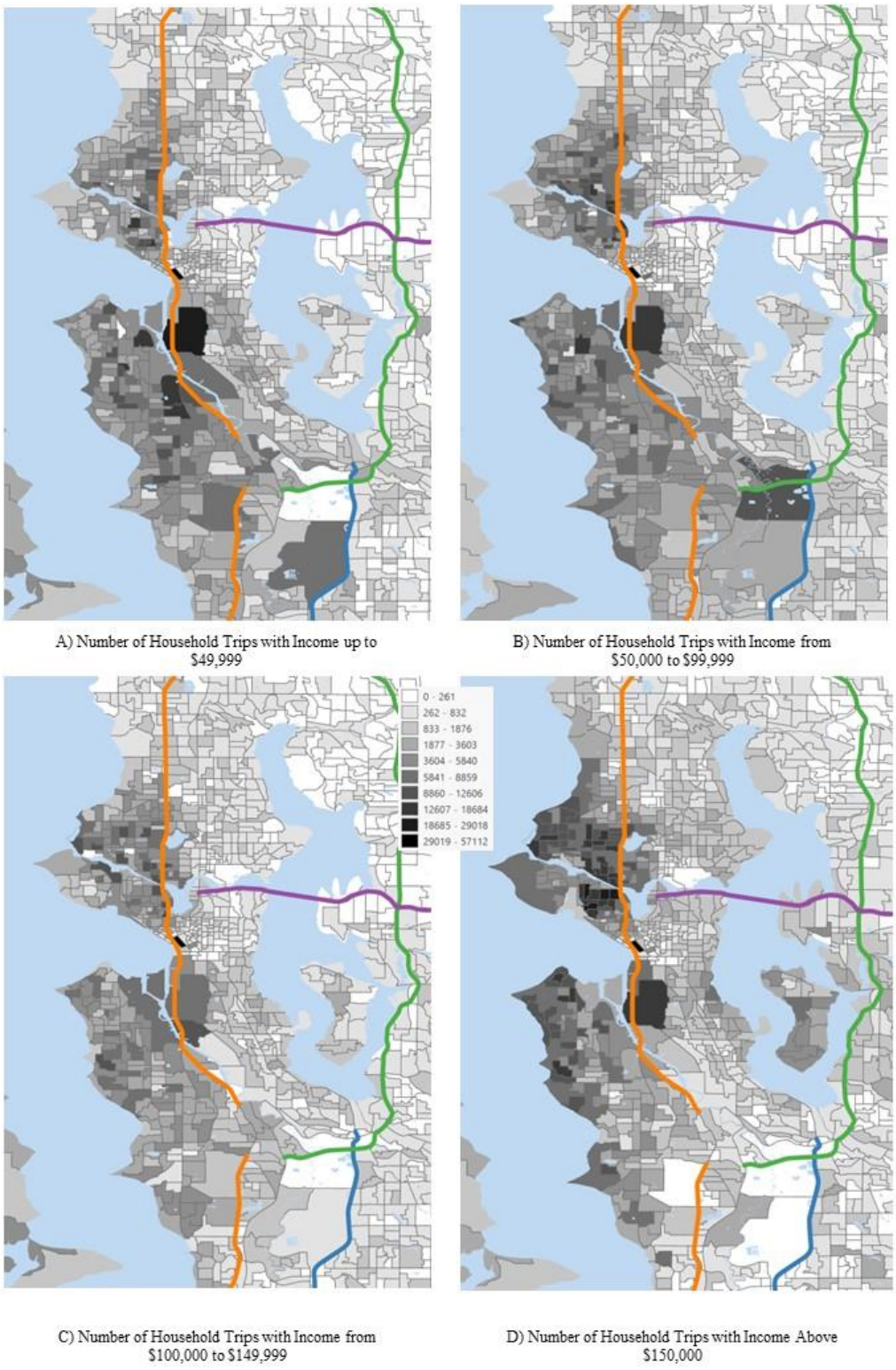

Figure 6.19: Trips Taken on the SR 99 Toll Tunnel from each CBG by Income with Fixed Scales

Figures 6.18 and 6.19 show interesting trends relating to the income distribution of trips that use the SR 99 tunnel. The geographical distribution of trips follows the generally expected results of the ecological regression model that trip distribution would be relatively homogenous for all income levels. Both figures match our expectation for the facility based upon the results of the model. The most notable break from the basic distribution of trips is the high number of low-income trips from the Duwamish Valley area as compared to the other parts of the city. This does follow the income trends in the city, where these areas generally tend to have lower income overall. However, it is also possible that these trips are being made by small businesses located in the industrial area south of downtown Seattle, where the businesses may have only a few company vehicles registered and therefore are not identified as "businesses" by the constraint which requires six vehicles per account to be identified as a business.

6.5.5 SR 16 Tacoma Narrows Tolled Bridge

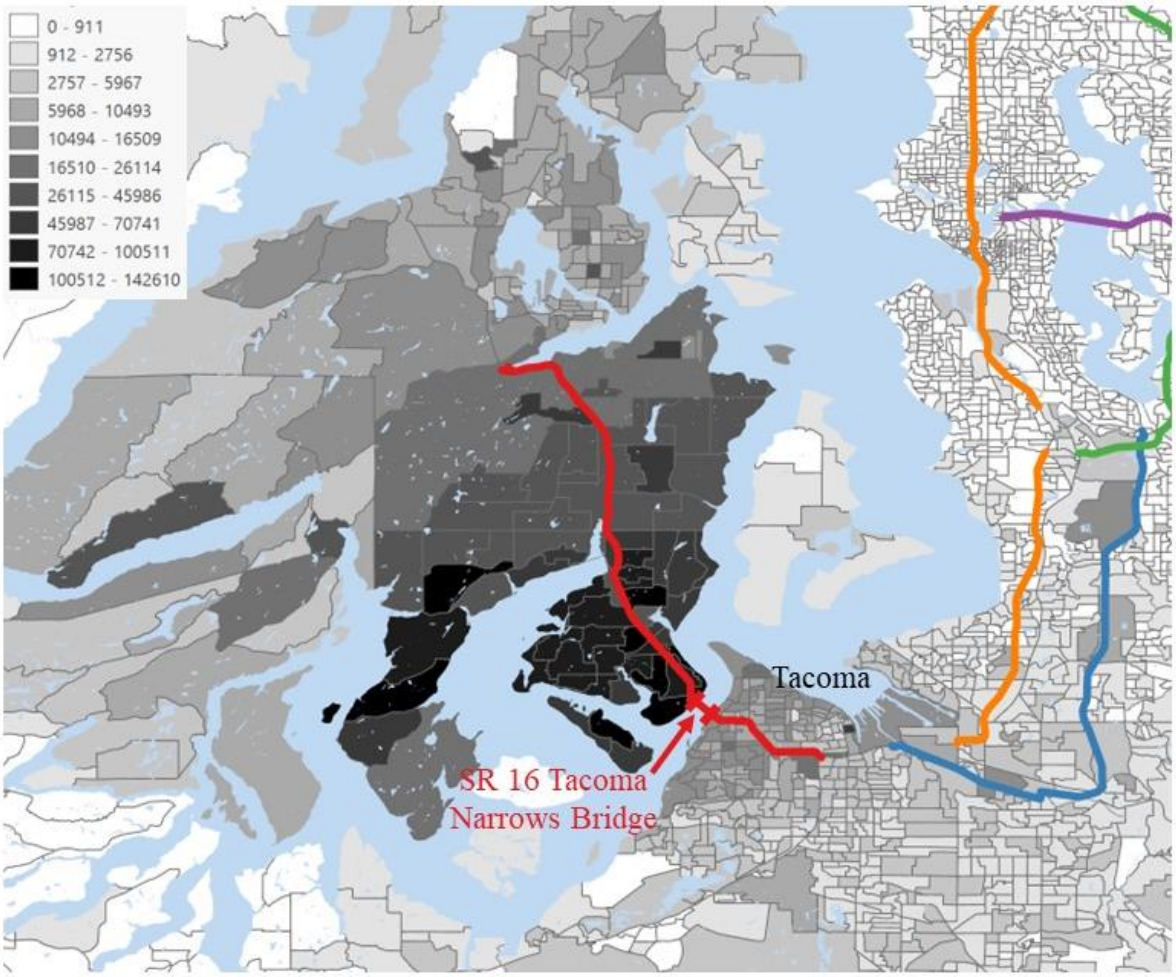

Figure 6.20: Trips Taken on the SR 16 Tacoma Narrows Bridge from each CBG

Figure 6.20 shows the distribution of trips that accounts made across the Tacoma Narrows Toll Bridge on SR 16. Unsurprisingly, the vast majority of these trips come from accounts on the west side of the bridge on the Kitsap peninsula. This follows the widely accepted use pattern for the facility, where it is primarily residents of Kitsap County using the facility to access Tacoma, with a smaller portion of users originating from Pierce County or other eastern counties. This facility is slightly different from the other four facilities, as there is no reasonable alternate route,

so it makes sense that there is a large concentration of people trying to travel into the city of Tacoma from west to east, but fewer making the opposite journey east to west.

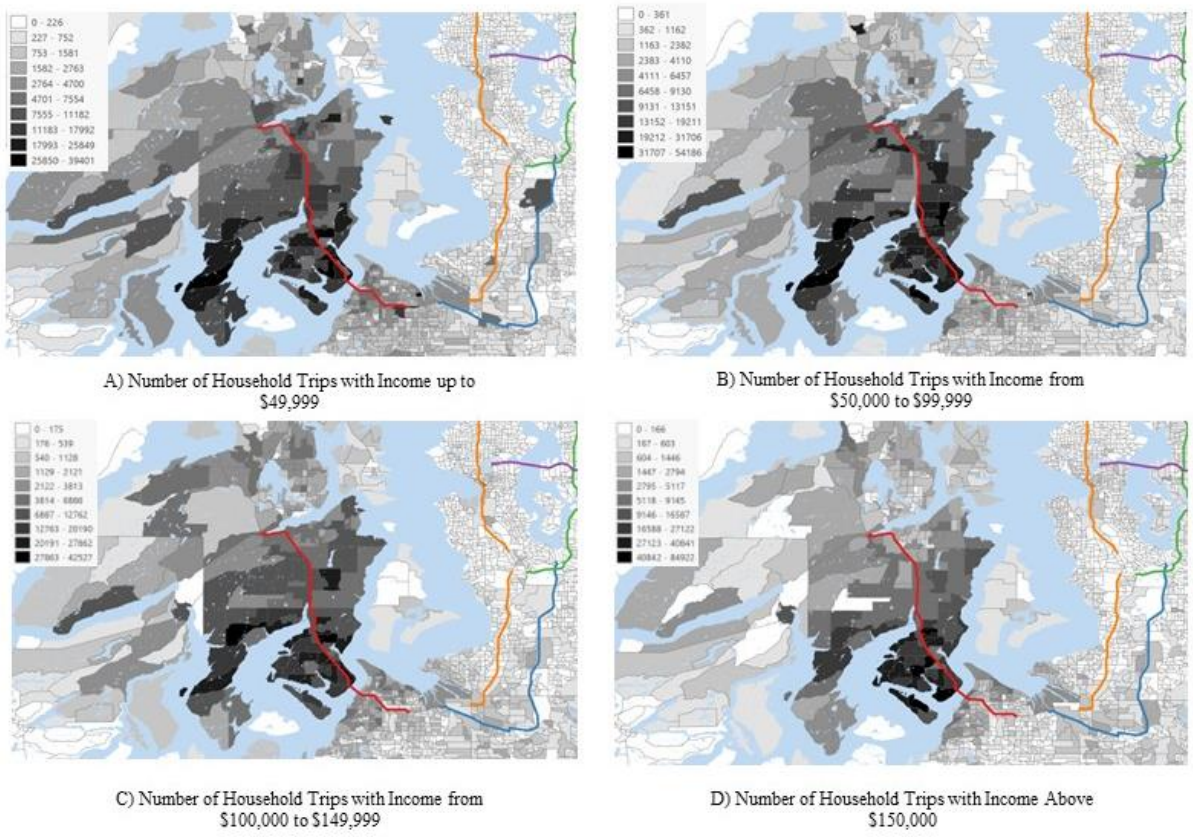

Figure 6.21: Trips Taken on the SR 16 Tacoma Narrows Bridge from each CBG by Income

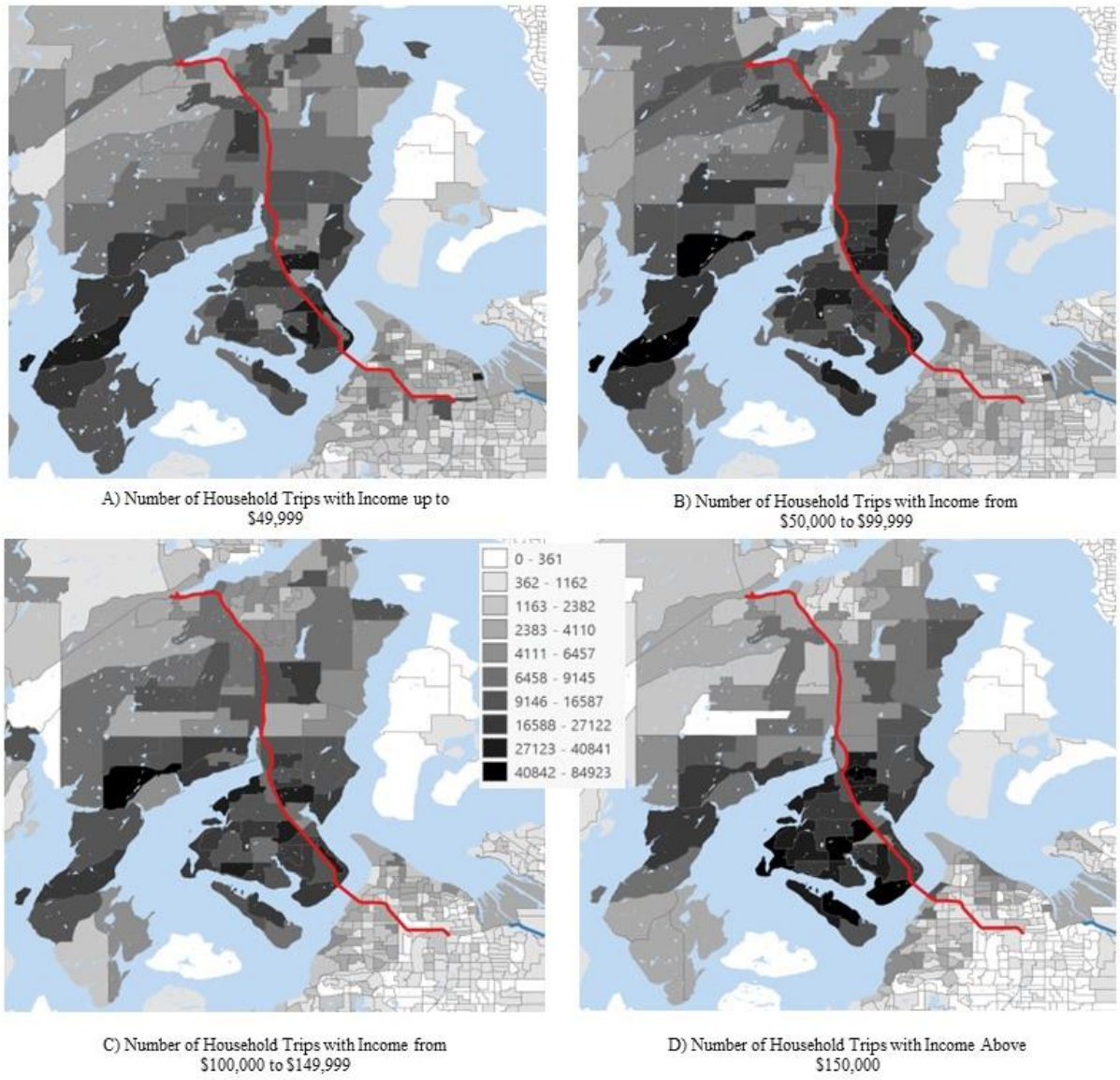

Figure 6.22: Trips Taken on the SR 16 Tacoma Narrows Bridge from each CBG by Income with Fixed Scales

Figures 6.21 and 6.22 show the geographic distribution of household trip making by income level for this route. SR 16 has the greatest homogeneity for the geographic origin of users. That is, a substantial number of users of all household income categories use the facility, and many

of the CBGs have SR 16 users in all income categories. Figure 6.21 shows that the distribution of users is more or less the same for each income level, with the highest concentrations of user households coming from the southern portion of the Kitsap peninsula. This idea is reinforced when looking at Figure 6.22; The four quadrants are very similar, indicating the homogeneity of facility use. This again is indicative of the fact that there is no viable alternative to the Tacoma Narrows Bridge, so if a person wants to make a trip between the peninsula and the mainland, they must pay the toll regardless of income, and they are choosing to pay electronically, rather than choosing to pay using the toll booth option.

6.6 Chapter Summary

In this chapter, we utilize ecological regression models to determine the demographic profiles of income for each tolling facility in the Central Puget Sound Region. From this we demonstrate how this can be used to understand the use of these facilities. Throughout this chapter we 1) provide background on tolling transaction data, 2) model facility usage via ecological regression, and 3) visualize how these trips are distributed geographically throughout the Central Puget Sound Region. The model results show that generally all facilities share a general demographic profile curve to that of the region, however some facilities follow this trend more closely than others. The visualizations corroborate and explain the findings of the ecological regression models. Of note, SR 520 deviates the most with a much higher rate of high-income use. The area surrounding that facility tends to be much higher income, even compared to other parts of the urban metropolitan region. This shows how ecological regression can be applied to understand the travel behavior of populations, especially in regard to tolling facilities.

Chapter 7: Equity Bias Case Studies: WSDOT Tolling

7.1 Overview

Ultimately, all of the data of the Washington State Department of Transportation (WSDOT) tolling facilities described in Chapter 6 attempts to represent the movement of people through these critical transportation facilities. Decisions regarding these facilities are often made through various data sources, and therefore understanding the relative representation between these different data options is critical for decision makers.

This chapter showcases a case study of the methodological framework developed in Chapter 4 to determine the relative representation of various datasets that describe the WSDOT tolling facilities. Two primary data sources are utilized in this case study: WSDOT toll transaction data and Cambridge Systematics (CS) trip data from the LOCUS dataset. This case study provides an opportunity to see how the relative representation of these different datasets can affect real-world decisions and the equity implications of those decisions. This will showcase both the usefulness of the methodological framework as well as the applicability for use in data-driven decision-making process.

7.2 Contributions and Chapter Structure

In this chapter, a case study of the methodological framework developed in Chapter 4 is presented using data describing WSDOT tolling facilities to showcase how this can be used for a real-world decision-making process. We begin with an overview of the real-world decisions that

these data sources might impact. Then we implement the methodological framework on the various datasets that describe the WSDOT toll facilities. Finally, we conclude with a discussion on how the found results impact the real-world decisions surrounding the toll facilities and the equity implications of using these datasets for those decisions. The contributions of the chapter are as follows:

- We showcase the implementation of the methodological framework for real-world datasets.
- We provide insight into the representation of different datasets on WSDOT tolling facilities and the impact this has on real-world decision making.

This chapter is organized as follows: Section 7.3 provides an overview of the real-world decisions impacted by the various tolling datasets. Section 7.4 describes the implementation of the methodological framework for these datasets. Section 7.5 highlights and discusses the results of the analysis. Section 7.6 discusses the implications for the real-world decisions regarding the tolling facilities. Finally, section 7.7 concludes the chapter with a summary and concluding remarks.

7.3 Overview of Real-World Decisions to be Made Regarding Toll Facilities

There are many decisions that must be made regarding the construction and operations of tolling facilities. Most of these decisions can largely be split into two categories: decisions regarding placement of the tolling facilities and decisions regarding the pricing of toll facilities. Decisions on where to place tolling facilities, whether as complete new-build infrastructure or implementation of tolling on a previously non-toll roadway inherently occur only once at the

inception of the tolling facility. Decisions on pricing toll facilities however can be made at any time, and there must necessarily be a decision regarding pricing if a new toll facility is implemented. Therefore, in many ways these decisions are inherently linked, as the type of tolling facility implemented may dictate what kind of pricing schemes can or cannot be implemented for each facility. We will now summarize the different kinds of options for tolling implementation and the different pricing policies that these facility types can accommodate.

The most common type of tolling facility and scheme is flat-rate tolling, where a vehicle is charged a flat rate when using a toll facility regardless of the time of day or operating conditions of the facility (FHWA 2022). These types of tolls can be implemented on a wide variety of roadway types but are most common on freeways. They usually involve the user paying the fixed fee on entrance or exit of the facility either using a booth or a license plate/toll tag reader.

Congestion pricing is another common type of tolling scheme. Congestion pricing schemes vary the price of the toll in various ways to mitigate congestion or other negative externalities related with a roadway by encouraging more efficient use of the roadway (Plotnick 2009). There are many ways to facilitate this kind of pricing scheme. The most common is fixed prices that change depending on the time of day. For example, a tolling facility may have a flat rate throughout the day, except during peak commute hours where the toll price is increased. In this basic example, the idea is that the increased price will shift use away from the toll facility in the most congested times of the day, the commute, leading to more efficient use of the roadway.

Though it is effective for increasing efficiency for general travel patterns through the facility, the major downside of the basic congestion pricing scheme described above is it is not adequate to adapt to real time traffic conditions. Priced managed lanes can address this concern. These toll facilities represent specified lanes alongside an existing general-purpose (GP) facility

where the lanes can change their price in real time based upon the conditions of the facility to ensure the lanes are constantly in a state of free flow (Perez 2012). This type of facility is a subset of congestion pricing as it addresses the same general goals of mitigating facility externalities. It does have several upsides to more general tolling schemes as it allows for real time adaptation to price roads most efficiently, however it does require a parallel route of GP lanes to be effective. Common types of priced managed lanes include express toll lanes (ETLs), high occupancy toll (HOT) lanes, truck-only toll (TOT) lanes, and bus-only toll lanes (Perez 2012).

The final main style of toll collection is through cordon pricing or area charges. Cordon pricing involves charging a fee for vehicles to enter a specified area in order to mitigate congestion and other externalities for that area. Area charges are similar, however instead of charging a one-time entry fee they charge a per mile fee for travel through the designated area (Victoria Transport Policy Institute 2019). Usually, these types of pricing schemes are applied to heavily congested city centers. A summary of all main toll collection schemes is summarized in Table 7.1.

Table 7.1: Summary of Toll Collection Strategies

Pricing Strategy	Description	Objectives
Flat-Rate Tolling	A fixed fee (or fixed fee per mile) for using a facility	Generate revenue
Congestion Pricing	A fee that is higher under congested conditions compared to uncongested	Reduce traffic congestion, generate revenue
Priced Managed Lanes	Rea-time price changes to facilitate free-flow in parallel toll lanes to GP lanes	Reduce congestion, generate revenue

Cordon Pricing	Variable or fixed charges to drive within the designated zone	Reduce congestion, generate revenue, reduce emissions
Area-Wide Charges	Per-mile charges on all roads within a given area	Reduce congestion

All of these facility types/pricing schemes utilize the same basic decision-making process: first, an initial decision must be made to implement the facility with an initial pricing scheme, and second the pricing scheme can be changed through time to adapt to changing use patterns or other issues.

As described in Chapter 6, WSDOT operates a number of toll facilities which utilize many combinations of the above-described tolling strategies. WSDOT operates three facilities that are point flat-rate tolling, the SR 520 Floating Bridge, The SR 99 tunnel, and the SR 16 Tacoma Narrows Bridge. They also operate two congestion pricing facilities, the I-405 Express Toll Lanes (ETLs) and the SR 167 High Occupancy Toll Lanes (HOT Lanes).

WSDOT is also working to implement 3 further tolling facilities throughout its network: the SR 509 and SR 167 facilities as part of the gateway program and an extension of the I-405 ETLs from Bellevue to Renton (Washington State Transportation Commission 2022). These facilities certainly represent major decisions regarding tolling for WSDOT, however since these projects have already been accepted and have already begun construction at the time of writing, these will not be analyzed as part of this case study. Instead, the primary decision that will be assessed herein is the adoption of a low-income toll program by WSDOT. This is an idea for a program that is currently being explored by WSDOT's tolling division. The most basic idea is to implement some type of program to give assistance to low-income toll users to mitigate the

disproportionate burden that may be imposed upon them by toll facilities. The goal of this case study will not to be to give a specific recommendation on what type of low-income toll assistance program to pursue; instead, it will show the implications of how the WSDOT tolling division can use its existing data to make that decision.

7.4 Overview of Data and Methodological Framework Implementation

For this case study, we implemented the methodological framework developed in Chapter 4 on several datasets that describe existing WSDOT toll facilities. These two overarching datasets are the WSDOT toll transaction data and CS LOCUS trip GPS data. For the sake of this analysis, we will treat the data extracted from each of the datasets for each facility as its own dataset.

Both of these datasets cover an entire years' worth of trips, from July of 2021 to June of 2022. Over that time period, there were 51,239,599 trips made on WSDOT toll facilities according to the WSDOT toll transaction data. The CS data included trips made throughout the region, including both those that did and did not use the toll facility. Overall, there were 43,471,412 trips made throughout the region captured in the CS data.

As a review from Chapter 4, the basic process defining the methodological framework is as follows: 1) define the demographic strata that are to be compared, 2) connect these demographic strata to the transportation datasets of interest 3) determine the representation of each stratum compared to the overall community in each dataset 4) calculate the overall representation of each dataset by weighing and averaging the representation of each stratum such that these terms can be compared between datasets.

For step 1 of the framework, we must define the demographic strata that will be used. For each dataset, the key demographic characteristic assessed will be the household income of toll users. Household income was selected because previous studies show that household income correlates with difference sin behavior for use of tolling facilities and that household income is appropriate for conducting equity assessments for tolling facilities (Leung 2019). This can be defined into 16 different income brackets which are as follows: <$10,000, $10,000 to $14,999, $15,000 to $19,999, $20,000 to $24,999, $25,000 to $29,999, $30,000 to $34,999, $35,000 to $39,999, $40,000 to $44,999, $45,000 to $49,999, $50,000 to $59,999, $60,000 to $74,999, $75,000 to $99,999, $100,000 to $124,999, $125,000 to $149,999, $150,000 to $199,999, and >$200,000. For step 2 of the framework, we must connect these demographic strata to the different datasets. To facilitate this, we utilized ecological regression to predict the demographic profiles of each facility based on each dataset as described in Chapter 5. For steps 3 and 4, we utilized the methodologies outlined in section 4.4 to calculate the representation of each stratum and then weigh them.

For each dataset, R_x was calculated utilizing the method of ecological regression following the methods defined in Chapter 5 and comparing the difference of each stratum calculated to the overall demographics of the region. For the WSDOT toll transaction data, O_x was calculated by the number of uncollected transactions from each stratum, where these transactions have an associated CBG but were not collected. This provides a measure of systematic error. For the CS data, O_x was calculated by measuring the difference in representation of the overall CS dataset to the population, which provides a measure of systemic error as theoretically all of the CS trips encompass the entirety of the transportation network. For step 4, we use the weighting techniques defined in Section 4.4, and repeated the weighting calculations several times using different

weighting schemes to showcase how different weighting schemes can affect the results of this framework.

One important distinction to note is that the CS data does have some limitations in terms of its coverage of toll facilities. Because this data is processed routes from GPS points, the data is inadequate to differentiate the difference between the I-405 ETLs, the SR 167 HOT lanes, and the accompanying GP lanes on the same facility. Therefore, this data cannot be used to understand the use of those facilities, so they will not be covered for this dataset.

7.5 Case Study Results and Discussion

Using the above methodologies, the representation of each dataset can be calculated. First, the demographics of the users captured by each dataset can be examined as part of step 2 highlighted above. We can show the difference in demographic profiles for all facilities calculated using ecological regression models. Table 7.2 and 7.3 show the odds ratios for each model as well as the confidence interval values for these ecological regression models showcasing the fitness of all models. Table 7.3 shows the elasticity for each demographic stratum of each model.

Table 7.2: Statistical Significance of Models Using Confidence Intervals for WSDOT Data

	I-405		SR 167		SR 520		SR 99		SR 16	
	Odds Ratio	95% Conf Int	Odds Ratio	95% Conf Int	Odds Ratio	95% Conf Int	Odds Ratio	95% Conf Int	Odds Ratio	95% Conf Int
Intercept	6.64 $\cdot 10^{-4}$	5.9 $\cdot 10^{-6}$	5.24 $\cdot 10^{-4}$	9.3 $\cdot 10^{-6}$	7.13 $\cdot 10^{-4}$	4.7 $\cdot 10^{-5}$	6.51 $\cdot 10^{-4}$	5.5 $\cdot 10^{-5}$	7.23 $\cdot 10^{-5}$	6.5 $\cdot 10^{-5}$
<$10,000	0.998	1.0 $\cdot 10^{-4}$	0.998	1.9 $\cdot 10^{-4}$	1.000	7.1 $\cdot 10^{-5}$	0.999	8.4 $\cdot 10^{-5}$	0.999	1.2 $\cdot 10^{-4}$

$10,000	to	0.995	1.4	0.997	2.4	1.000	8.2	0.998	9.9	0.997	1.3
$14,999			$\cdot 10^{-4}$		$\cdot 10^{-4}$		$\cdot 10^{-5}$		$\cdot 10^{-5}$		$\cdot 10^{-4}$
$15,000	to	1.002	1.0	1.001	2.1	1.001	7.9	1.000	1.0	1.002	1.2
$19,999			$\cdot 10^{-4}$		$\cdot 10^{-4}$		$\cdot 10^{-5}$		$\cdot 10^{-4}$		$\cdot 10^{-4}$
$20,000	to	0.999	1.2	1.001	2.4	1.000	9.4	0.999	1.0	0.999	1.1
$24,999			$\cdot 10^{-4}$		$\cdot 10^{-4}$		$\cdot 10^{-5}$		$\cdot 10^{-4}$		$\cdot 10^{-4}$
$25,000	to	1.000	1.3	0.999	2.4	0.999	9.8	1.002	1.0	1.001	1.2
$29,999			$\cdot 10^{-4}$		$\cdot 10^{-4}$		$\cdot 10^{-5}$		$\cdot 10^{-4}$		$\cdot 10^{-4}$
$30,000	to	0.999	1.3	1.001	1.9	1.000	9.8	1.003	9.6	1.000	9.9
$34,999			$\cdot 10^{-4}$		$\cdot 10^{-4}$		$\cdot 10^{-5}$		$\cdot 10^{-5}$		$\cdot 10^{-5}$
$35,000	to	1.003	1.3	1.004	2.2	1.001	1.1	1.002	1.1	1.002	1.2
$39,999			$\cdot 10^{-4}$		$\cdot 10^{-4}$		$\cdot 10^{-4}$		$\cdot 10^{-4}$		$\cdot 10^{-4}$
$40,000	to	0.999	1.2	1.001	2.0	1.000	8.9	1.001	9.4	1.002	1.1
$44,999			$\cdot 10^{-4}$		$\cdot 10^{-4}$		$\cdot 10^{-5}$		$\cdot 10^{-5}$		$\cdot 10^{-4}$
$45,000	to	1.002	1.2	1.002	2.2	1.002	7.5	1.003	8.3	1.003	1.1
$49,999			$\cdot 10^{-4}$		$\cdot 10^{-4}$		$\cdot 10^{-5}$		$\cdot 10^{-5}$		$\cdot 10^{-4}$
$50,000	to	1.001	7.4	1.002	1.3	1.000	5.8	1.001	6.2	1.001	1.1
$59,999			$\cdot 10^{-5}$		$\cdot 10^{-4}$		$\cdot 10^{-5}$		$\cdot 10^{-5}$		$\cdot 10^{-4}$
$60,000	to	1.001	5.9	1.003	1.1	1.000	4.9	1.001	4.7	1.001	5.2
$74,999			$\cdot 10^{-5}$		$\cdot 10^{-4}$		$\cdot 10^{-5}$		$\cdot 10^{-5}$		$\cdot 10^{-5}$
$75,000	to	1.001	5.0	1.001	9.3	1.000	4.0	0.999	4.7	1.001	5.5
$99,999			$\cdot 10^{-5}$		$\cdot 10^{-5}$		$\cdot 10^{-5}$		$\cdot 10^{-5}$		$\cdot 10^{-5}$
$100,000	to	1.001	5.3	1.002	9.7	1.000	4.0	1.001	4.7	1.001	5.5
$124,999			$\cdot 10^{-5}$		$\cdot 10^{-5}$		$\cdot 10^{-5}$		$\cdot 10^{-5}$		$\cdot 10^{-5}$
$125,000	to	0.999	5.7	1.000	1.1	0.997	3.9	0.998	4.5	0.998	6.2
$149,999			$\cdot 10^{-5}$		$\cdot 10^{-3}$		$\cdot 10^{-5}$		$\cdot 10^{-5}$		$\cdot 10^{-5}$
$150,000	to	1.001	4.8	1.000	1.0	1.000	3.7	1.000	4.9	1.000	5.3
$199,999			$\cdot 10^{-5}$		$\cdot 10^{-4}$		$\cdot 10^{-5}$		$\cdot 10^{-5}$		$\cdot 10^{-5}$
>$200,000		0.998	3.0	0.994	7.5	1.001	1.9	0.999	2.8	0.997	3.8
			$\cdot 10^{-5}$		$\cdot 10^{-5}$		$\cdot 10^{-5}$		$\cdot 10^{-5}$		$\cdot 10^{-5}$
-2 Log Likelihood		34392712		10951446		43957628		34795702		28147695	

Table 7.3: Statistical Significance of Models Using Confidence Intervals for LOCUS Data

	I-405		SR 167		SR 520		SR 99		SR 16		Non-Toll	
	Odds Ratio	95% Conf Int	Odds Ratio	95% Conf Int	Odds Ratio	95% Conf Int	Odds Ratio	95% Conf Int	Odds Ratio	95% Conf Int	Odds Ratio	95% Conf Int
Intercept	2.21 $\cdot 10^{-4}$	9.3 $\cdot 10^{-5}$	1.97 $\cdot 10^{-4}$	9.2 $\cdot 10^{-5}$	1.97 $\cdot 10^{-5}$	2.5 $\cdot 10^{-5}$	1.33 $\cdot 10^{-5}$	2.2 $\cdot 10^{-5}$	2.57 $\cdot 10^{-5}$	2.2 $\cdot 10^{-5}$	248.99	58.04
<$10,000	0.991	4.5 $\cdot 10^{-3}$	0.987	4.9 $\cdot 10^{-3}$	1.001	9.3 $\cdot 10^{-3}$	0.977	2.1 $\cdot 10^{-2}$	1.000	5.7 $\cdot 10^{-3}$	0.990	1.9 $\cdot 10^{-3}$
$10,000 to $14,999	0.993	6.8 $\cdot 10^{-3}$	0.993	6.8 $\cdot 10^{-3}$	1.022	8.9 $\cdot 10^{-3}$	1.016	1.5 $\cdot 10^{-3}$	1.014	5.6 $\cdot 10^{-3}$	1.009	3.9 $\cdot 10^{-3}$
$15,000 to $19,999	0.988	7.1 $\cdot 10^{-3}$	0.990	6.6 $\cdot 10^{-3}$	1.015	1.1 $\cdot 10^{-2}$	1.011	1.6 $\cdot 10^{-2}$	1.018	5.7 $\cdot 10^{-3}$	0.993	2.8 $\cdot 10^{-3}$
$20,000 to $24,999	1.007	4.9 $\cdot 10^{-3}$	1.010	4.6 $\cdot 10^{-3}$	0.981	2.1 $\cdot 10^{-2}$	0.985	2.8 $\cdot 10^{-2}$	0.991	8.7 $\cdot 10^{-3}$	1.010	3.9 $\cdot 10^{-3}$
$25,000 to $29,999	1.019	4.1 $\cdot 10^{-3}$	1.019	4.1 $\cdot 10^{-3}$	0.996	1.6 $\cdot 10^{-2}$	1.019	1.4 $\cdot 10^{-2}$	0.984	1.1 $\cdot 10^{-2}$	0.999	3.4 $\cdot 10^{-3}$
$30,000 to $34,999	1.012	4.7 $\cdot 10^{-3}$	1.010	4.6 $\cdot 10^{-3}$	1.011	1.2 $\cdot 10^{-2}$	1.004	1.9 $\cdot 10^{-2}$	1.007	7.2 $\cdot 10^{-3}$	0.998	2.9 $\cdot 10^{-3}$
$35,000 to $39,999	0.998	5.1 $\cdot 10^{-3}$	1.009	4.7 $\cdot 10^{-3}$	1.003	1.4 $\cdot 10^{-2}$	1.012	1.5 $\cdot 10^{-2}$	1.013	6.9 $\cdot 10^{-3}$	0.995	3.0 $\cdot 10^{-3}$
$40,000 to $44,999	1.001	4.6 $\cdot 10^{-3}$	1.008	3.7 $\cdot 10^{-3}$	1.008	1.4 $\cdot 10^{-2}$	1.008	1.6 $\cdot 10^{-2}$	1.011	6.5 $\cdot 10^{-3}$	1.002	3.0 $\cdot 10^{-3}$
$45,000 to $49,999	1.008	4.2 $\cdot 10^{-3}$	1.005	4.2 $\cdot 10^{-3}$	1.019	1.1 $\cdot 10^{-2}$	1.004	1.7 $\cdot 10^{-2}$	1.016	6.4 $\cdot 10^{-3}$	0.999	3.0 $\cdot 10^{-3}$
$50,000 to $59,999	1.008	2.9 $\cdot 10^{-3}$	1.008	2.9 $\cdot 10^{-3}$	0.985	1.2 $\cdot 10^{-2}$	0.995	1.4 $\cdot 10^{-2}$	1.005	5.3 $\cdot 10^{-3}$	0.996	1.9 $\cdot 10^{-3}$
$60,000 to $74,999	1.007	2.5 $\cdot 10^{-3}$	1.009	2.5 $\cdot 10^{-3}$	1.001	7.8 $\cdot 10^{-3}$	1.011	8.4 $\cdot 10^{-3}$	1.003	4.6 $\cdot 10^{-3}$	0.997	1.6 $\cdot 10^{-3}$
$75,000 to $99,999	1.004	1.7 $\cdot 10^{-3}$	1.005	1.7 $\cdot 10^{-3}$	1.004	5.6 $\cdot 10^{-3}$	1.008	6.3 $\cdot 10^{-3}$	1.009	2.6 $\cdot 10^{-3}$	0.996	1.2 $\cdot 10^{-3}$

$100,000 to $124,999	1.002	$2.2 \cdot 10^{-3}$	1.006	$2.2 \cdot 10^{-3}$	1.005	$6.9 \cdot 10^{-3}$	1.002	$8.1 \cdot 10^{-3}$	1.009	$3.8 \cdot 10^{-3}$	0.993	$1.3 \cdot 10^{-3}$
$125,000 to $149,999	1.010	$2.4 \cdot 10^{-3}$	1.009	$2.7 \cdot 10^{-3}$	1.008	$7.2 \cdot 10^{-3}$	1.002	$1.2 \cdot 10^{-2}$	1.009	$4.3 \cdot 10^{-3}$	0.993	$1.6 \cdot 10^{-3}$
$150,000 to $199,999	1.012	$1.8 \cdot 10^{-3}$	1.009	$2.2 \cdot 10^{-3}$	1.010	$1.6 \cdot 10^{-3}$	1.009	$6.9 \cdot 10^{-3}$	1.002	$3.8 \cdot 10^{-3}$	0.996	$1.3 \cdot 10^{-3}$
>$200,000	1.013	$1.1 \cdot 10^{-3}$	1.005	$1.7 \cdot 10^{-3}$	1.016	$2.9 \cdot 10^{-3}$	1.013	$4.1 \cdot 10^{-3}$	1.012	$2.3 \cdot 10^{-3}$	0.994	$8.4 \cdot 10^{-4}$
-2 Log Likelihood	50491.42		46360.48		9099.627		6112.75		20981.81		54033.22	

Table 7.4: Sensitivity of Ecological Regression Models

	I-405		SR 167		SR 520		SR 99		SR 16		All LOCUS Data
	WSDOT	CS	WSDOT	CS	WSDOT	CS	WSDOT	CS	WSDOT	CS	CS
<$10,000	0.996	0.982	0.996	0.973	1.000	1.001	0.998	0.955	0.998	1.000	0.980
$10,000 to $14,999	0.990	0.986	0.993	0.987	0.999	1.044	0.996	1.032	0.994	1.028	1.018
$15,000 to $19,999	1.004	0.977	1.003	0.980	1.001	1.031	1.001	1.023	1.005	1.036	0.987
$20,000 to $24,999	0.998	1.013	1.003	1.021	1.000	0.962	0.998	0.970	0.998	0.981	1.021
$25,000 to $29,999	1.000	1.039	0.999	1.038	0.998	0.993	1.003	1.039	1.002	0.967	0.998
$30,000 to $34,999	0.999	1.024	1.004	1.020	1.001	1.024	1.006	1.008	1.000	1.014	0.996
$35,000 to $39,999	1.006	0.996	1.008	1.017	1.002	1.007	1.004	1.023	1.004	1.026	0.990
$40,000 to $44,999	0.997	1.002	1.004	1.016	1.001	1.016	1.002	1.017	1.004	1.023	1.004

$45,000 to $49,999	1.004	1.017	1.005	1.010	1.003	1.039	1.005	1.018	1.006	1.032	0.999
$50,000 to $59,999	1.003	1.015	1.004	1.015	1.001	0.971	1.003	0.987	1.002	1.010	0.993
$60,000 to $74,999	1.002	1.014	1.007	1.017	0.998	1.003	1.001	1.023	1.002	1.006	0.993
$75,000 to $99,999	1.002	1.009	1.003	1.009	0.999	1.008	0.998	1.008	1.003	1.019	0.993
$100,000 to $124,999	1.002	1.004	1.004	1.012	1.001	1.011	1.001	1.014	1.001	1.019	0.985
$125,000 to $149,999	0.998	1.020	0.999	1.018	0.995	1.016	0.996	1.003	0.996	1.018	0.986
$150,000 to $199,999	1.002	1.024	0.999	1.019	1.000	1.020	1.000	1.018	1.000	1.004	0.992
>$200,000	0.997	1.026	0.989	1.012	1.002	1.033	0.998	1.025	0.994	1.024	0.988

From these tables we see that all coefficients show reasonably small confidence intervals to claim significance for each income level. Additionally, as expected the elasticity of each income level in each model is directly related to the odds ratio and confidence interval again indicating the fitness of these models. Figures 7.1, 7.2, and 7.3 show the demographic profiles calculated for the facilities that can utilize the CS data while Figures 7.4 and 7.5 show the demographic profiles for the remaining facilities. Note that for I-405 and SR 167 the values for ‘all trips’ are generated from the CS data and represent trips that occur through both the toll and GP lanes.

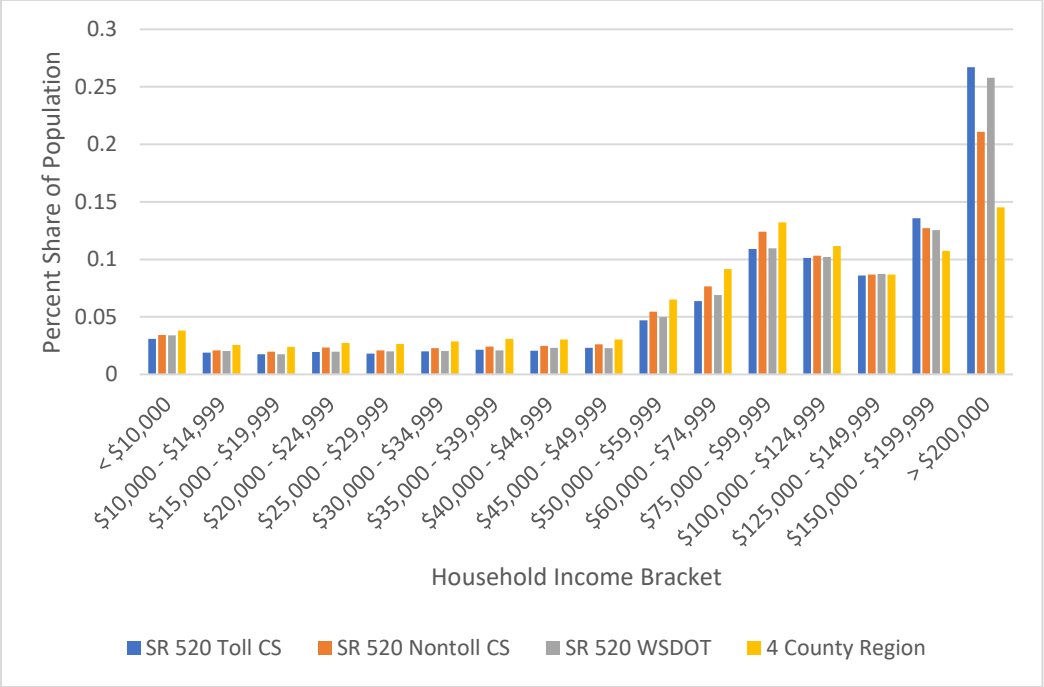

Figure 7.1: Demographic Profiles for SR 520 Floating Bridge

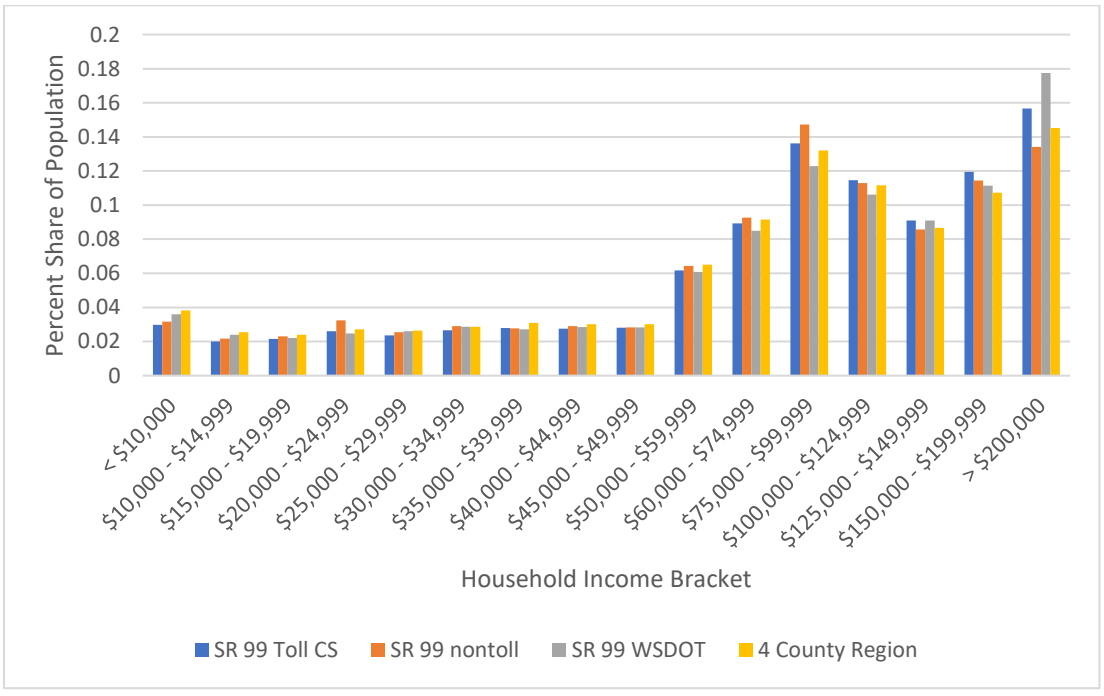

Figure 7.2: Demographic Profiles for SR 99 Tunnel

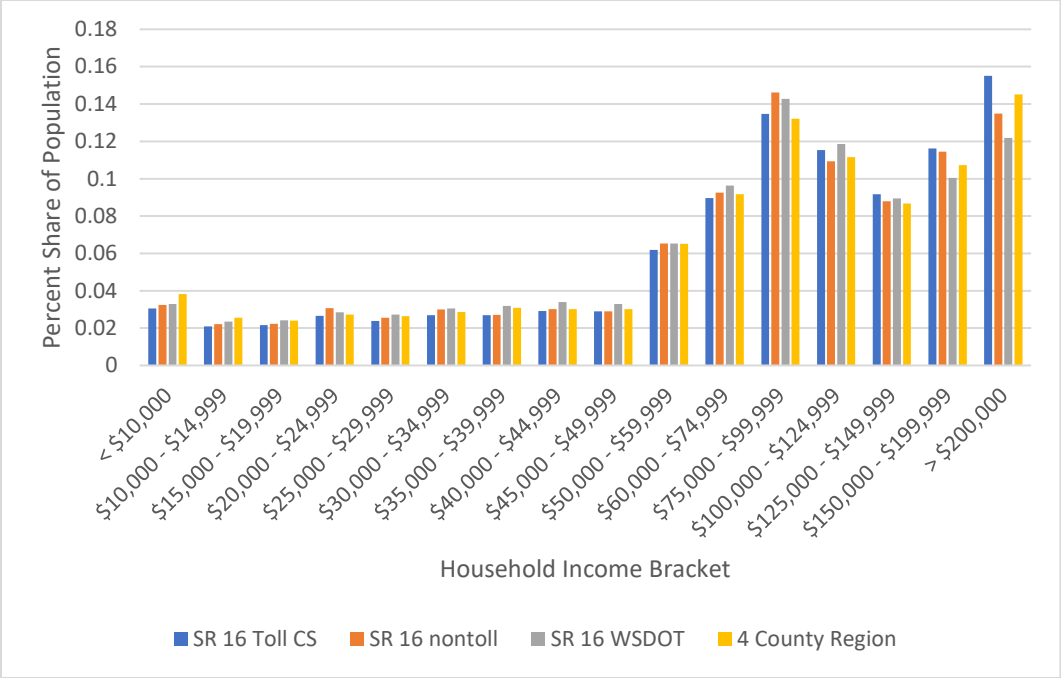

Figure 7.3: Demographic Profiles for SR 16 Tacoma Narrows Bridge

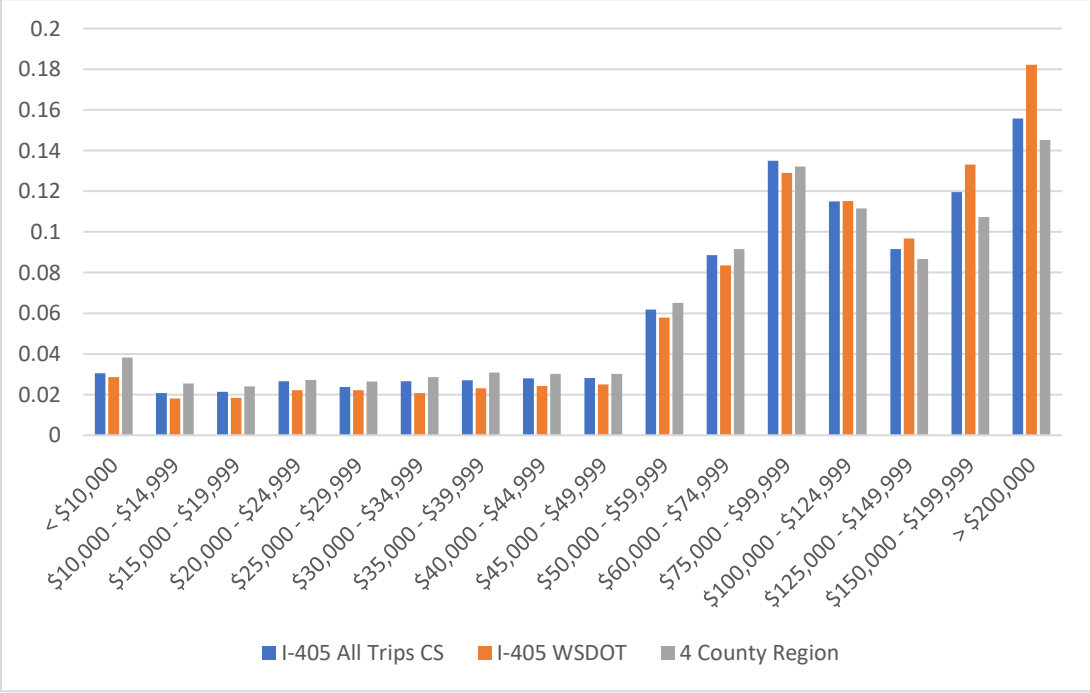

Figure 7.4: Demographic Profiles for I-405 ETLs

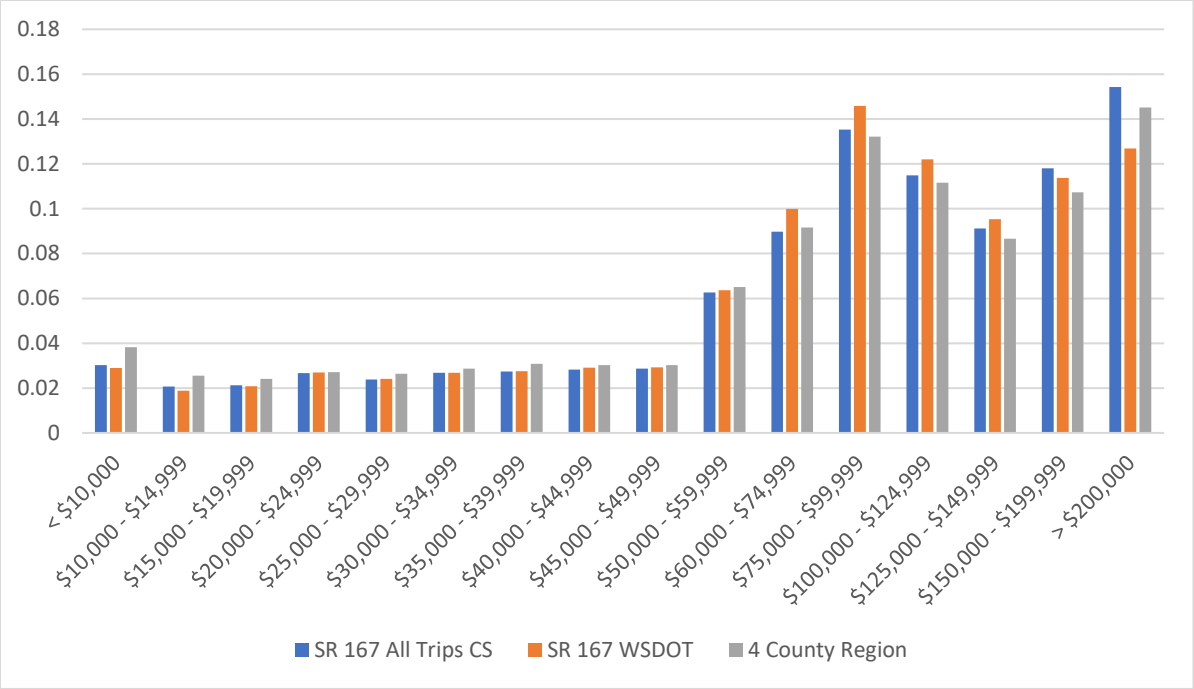

Figure 7.5: Demographic Profiles for SR 167 HOT Lanes

With this information, we can calculate the various terms required to find the representation for each income level for each facility and dataset, with Table 7.5 through Table 7.9 showing these results of the different terms for each facility. We also show the results for the overall CS dataset (both trips that use and do not use a toll facility) in Table 7.10. Note that because the CS data is not sufficient to measure use of the I-405 and SR 167 facilities, there are no columns for CS data in those tables.

Table 7.5 Representation Terms for Income Levels Related to I-405

Income Bracket	R_x	O_x	A	P_x
	WSDOT	WSDOT	WSDOT	WSDOT
<$10,000	0.750	0.998	0.998	0.748

$10,000 to $14,999	0.706	0.999	0.998	0.704
$15,000 to $19,999	0.767	0.999	0.998	0.764
$20,000 to $24,999	0.814	0.999	0.998	0.811
$25,000 to $29,999	0.839	0.999	0.998	0.836
$30,000 to $34,999	0.726	0.999	0.998	0.724
$35,000 to $39,999	0.781	0.999	0.998	0.778
$40,000 to $44,999	0.781	0.999	0.998	0.779
$45,000 to $49,999	0.829	0.999	0.998	0.827
$50,000 to $59,999	0.889	0.997	0.998	0.885
$60,000 to $74,999	0.912	0.995	0.998	0.906
$75,000 to $99,999	0.977	0.993	0.998	0.968
$100,000 to $124,999	0.968	0.994	0.998	0.960
$125,000 to $149,999	0.884	0.995	0.998	0.877
$150,000 to $199,999	0.759	0.993	0.998	0.752
>$200,000	0.746	0.990	0.998	0.737

Table 7.6 Representation Terms for Income Levels Related to SR 167

Income Bracket	R_x	O_x	A	P_x
	WSDOT	WSDOT	WSDOT	WSDOT
<$10,000	0.758	0.998	0.999	0.756
$10,000 to $14,999	0.740	0.998	0.999	0.738
$15,000 to $19,999	0.869	0.998	0.999	0.867
$20,000 to $24,999	0.994	0.998	0.999	0.992

$25,000 to $29,999	0.915	0.998	0.999	0.912
$30,000 to $34,999	0.939	0.998	0.999	0.937
$35,000 to $39,999	0.929	0.998	0.999	0.927
$40,000 to $44,999	0.944	0.998	0.999	0.942
$45,000 to $49,999	0.971	0.998	0.999	0.968
$50,000 to $59,999	0.978	0.995	0.999	0.972
$60,000 to $74,999	0.910	0.991	0.999	0.902
$75,000 to $99,999	0.896	0.988	0.999	0.884
$100,000 to $124,999	0.906	0.990	0.999	0.896
$125,000 to $149,999	0.901	0.992	0.999	0.893
$150,000 to $199,999	0.940	0.990	0.999	0.930
>$200,000	0.874	0.989	0.999	0.864

Table 7.7 Representation Terms for Income Levels Related to SR 520

Income Bracket	R_x		O_x		A		P_x	
	WSDOT	CS	WSDOT	CS	WSDOT	CS	WSDOT	CS
<$10,000	0.888	0.846	0.998	0.954	0.995	0.962	0.881	0.776
$10,000 to $14,999	0.799	0.794	0.999	0.934	0.995	0.962	0.793	0.713
$15,000 to $19,999	0.732	0.788	0.999	0.930	0.995	0.962	0.727	0.705
$20,000 to $24,999	0.728	0.764	0.999	0.941	0.995	0.962	0.723	0.691
$25,000 to $29,999	0.757	0.727	0.999	0.938	0.995	0.962	0.752	0.656
$30,000 to $34,999	0.711	0.748	0.999	0.936	0.995	0.962	0.706	0.673
$35,000 to $39,999	0.709	0.741	0.999	0.981	0.995	0.962	0.704	0.700

$40,000 to $44,999	0.744	0.728	0.999	0.919	0.995	0.962	0.739	0.644
$45,000 to $49,999	0.755	0.810	0.999	0.948	0.995	0.962	0.750	0.738
$50,000 to $59,999	0.767	0.756	0.997	0.959	0.995	0.962	0.760	0.697
$60,000 to $74,999	0.752	0.740	0.996	0.940	0.995	0.962	0.745	0.669
$75,000 to $99,999	0.831	0.874	0.994	0.944	0.995	0.962	0.821	0.793
$100,000 to $124,999	0.916	0.937	0.994	0.970	0.995	0.962	0.906	0.874
$125,000 to $149,999	0.993	0.993	0.995	0.998	0.995	0.962	0.983	0.953
$150,000 to $199,999	0.830	0.754	0.993	0.973	0.995	0.962	0.820	0.706
>$200,000	0.224	0.193	0.986	0.836	0.995	0.962	0.220	0.155

Table 7.8 Representation Terms for Income Levels Related to SR 99

Income Bracket	R_x		O_x		A		P_x	
	WSDOT	CS	WSDOT	CS	WSDOT	CS	WSDOT	CS
<$10,000	0.942	0.816	0.998	0.954	0.994	0.962	0.934	0.749
$10,000 to $14,999	0.940	0.844	0.999	0.934	0.994	0.962	0.933	0.758
$15,000 to $19,999	0.917	0.961	0.999	0.930	0.994	0.962	0.910	0.859
$20,000 to $24,999	0.907	0.981	0.999	0.941	0.994	0.962	0.900	0.922
$25,000 to $29,999	0.986	0.952	0.999	0.938	0.994	0.962	0.978	0.859
$30,000 to $34,999	0.999	0.996	0.998	0.936	0.994	0.962	0.992	0.897
$35,000 to $39,999	0.920	0.961	0.999	0.981	0.994	0.962	0.912	0.907
$40,000 to $44,999	0.921	0.973	0.998	0.919	0.994	0.962	0.914	0.860
$45,000 to $49,999	0.939	0.982	0.998	0.948	0.994	0.962	0.932	0.895
$50,000 to $59,999	0.935	0.987	0.997	0.959	0.994	0.962	0.926	0.912

$60,000 to $74,999	0.928	0.964	0.995	0.940	0.994	0.962	0.917	0.937
$75,000 to $99,999	0.930	0.974	0.993	0.944	0.994	0.962	0.917	0.931
$100,000 to $124,999	0.952	0.997	0.994	0.970	0.994	0.962	0.940	0.936
$125,000 to $149,999	0.951	0.952	0.995	0.998	0.994	0.962	0.940	0.914
$150,000 to $199,999	0.961	0.911	0.994	0.973	0.994	0.962	0.949	0.853
>$200,000	0.778	0.898	0.990	0.836	0.994	0.962	0.765	0.886

Table 7.9 Representation Terms for Income Levels Related to SR 16

Income Bracket	R_x		O_x		A		P_x	
	WSDOT	CS	WSDOT	CS	WSDOT	CS	WSDOT	CS
<$10,000	0.860	0.838	0.998	0.954	0.995	0.962	0.854	0.769
$10,000 to $14,999	0.921	0.879	0.999	0.934	0.995	0.962	0.915	0.789
$15,000 to $19,999	0.991	0.969	0.999	0.930	0.995	0.962	0.985	0.867
$20,000 to $24,999	0.951	0.959	0.999	0.941	0.995	0.962	0.945	0.942
$25,000 to $29,999	0.969	0.963	0.999	0.938	0.995	0.962	0.963	0.869
$30,000 to $34,999	0.935	0.994	0.999	0.936	0.995	0.962	0.929	0.906
$35,000 to $39,999	0.923	0.928	0.998	0.981	0.995	0.962	0.917	0.876
$40,000 to $44,999	0.904	0.974	0.998	0.919	0.995	0.962	0.898	0.907
$45,000 to $49,999	0.911	0.987	0.998	0.948	0.995	0.962	0.905	0.923
$50,000 to $59,999	0.997	0.992	0.997	0.959	0.995	0.962	0.989	0.915
$60,000 to $74,999	0.948	0.960	0.995	0.940	0.995	0.962	0.939	0.940
$75,000 to $99,999	0.919	0.962	0.993	0.944	0.995	0.962	0.908	0.943
$100,000 to $124,999	0.937	0.995	0.994	0.970	0.995	0.962	0.927	0.929

$125,000 to $149,999	0.969	0.945	0.996	0.998	0.995	0.962	0.960	0.907
$150,000 to $199,999	0.936	0.943	0.995	0.973	0.995	0.962	0.927	0.882
>$200,000	0.839	0.885	0.994	0.836	0.995	0.962	0.830	0.896

Table 7.10 Representation Terms for Income Levels Related to All CS Data

Income Bracket	R_x	O_x	A	P_x
	CS	CS	CS	CS
<$10,000	1.00	0.954	0.962	0.917
$10,000 to $14,999	1.00	0.934	0.962	0.898
$15,000 to $19,999	1.00	0.930	0.962	0.895
$20,000 to $24,999	1.00	0.941	0.962	0.905
$25,000 to $29,999	1.00	0.938	0.962	0.902
$30,000 to $34,999	1.00	0.936	0.962	0.900
$35,000 to $39,999	1.00	0.981	0.962	0.944
$40,000 to $44,999	1.00	0.919	0.962	0.884
$45,000 to $49,999	1.00	0.948	0.962	0.912
$50,000 to $59,999	1.00	0.959	0.962	0.922
$60,000 to $74,999	1.00	0.940	0.962	0.904
$75,000 to $99,999	1.00	0.944	0.962	0.908
$100,000 to $124,999	1.00	0.970	0.962	0.933
$125,000 to $149,999	1.00	0.998	0.962	0.960
$150,000 to $199,999	1.00	0.973	0.962	0.936
>$200,000	1.00	0.836	0.962	0.804

From these results we can see some initial trends. The WSDOT toll transaction data has relatively higher O_x compared to the CS data. The toll tag and license plate readers are very mature technologies that tend to have low failure rates indiscriminate of user. The CS data on the other hand relies on GPS data collected from apps which has much more potential for systematic error. Additionally, the CS data appears to have slightly lower R_x values as well, which in combination with the lower O_x leads to generally lower values of overall representation P_x. Another interesting trend is that often the highest income category, household incomes greater than $200,000, regularly has a significantly lower representation than other strata, usually indicating this stratum is generally overrepresented in all datasets. This is especially true on SR 520, where the representation of this stratum is approximately 75% lower than other strata. Now that we have the representation for each stratum, we can weigh each dataset for each route to get an overall score of representation in each case. Table 7.11 shows the final weighted representation scores P for each facility and data source.

Table 7.11: Weighted Representations for Each Facility and Data Type

Facility	No Weights		By 2x for Incomes Below County Median		By Strata Size		By Strata Size and 2x for Incomes Below County Median	
	WSDOT	CS	WSDOT	CS	WSDOT	CS	WSDOT	CS
I-405	0.816	N/A	0.814	N/A	0.840	N/A	0.846	N/A
SR 167	0.899	N/A	0.899	N/A	0.897	N/A	0.898	N/A
SR 520	0.752	0.696	0.755	0.700	0.730	0.673	0.744	0.689

SR 99	0.922	0.880	0.926	0.877	0.908	0.894	0.914	0.894
SR 16	0.924	0.891	0.926	0.889	0.916	0.904	0.920	0.904
All CS Trips	N/A	0.908	N/A	0.907	N/A	0.904	N/A	0.905

From these results we can draw several key conclusions. Firstly, we can see across all weighting types that SR 520 is consistently the least representative. This is because it traverses some of the wealthiest areas of the region and therefore the majority of the trips that are captured on that facility skew towards higher income individuals. Similarly, I-405 has consistently the second lowest level of representation. This also makes sense because I-405 also covers relatively more wealthy areas, but not to the same extent as SR 520. Then we can see that SR 167, SR 99, and SR 16 all tend to have the highest representation at around the same level. This also follows the trend of the prevailing demographics around the facility as these facilities are in areas with greater numbers of low-income households than the other facilities. We can also see that the overall CS data is quite representative, approximately hovering at a value of 0.9, similar to SR 99, SR 167, and SR 16. However, across the board all CS datasets that describe specific facility fall slightly short of the representation of the WSDOT data. Overall, the data describing the SR 167, SR 99, and SR 16 facilities appears to be the greatest representation.

The differences in the weighting schemes can also be examined to shed light on the qualities of the different datasets. Overall, there were not very large differences between the different weighting schemes. There appears to be no noticeable change when strata that are below the median household income for their county are weighted double. However, there is a small but distinct trend where when weighing by strata size, the representation score is slightly reduced. This

did not flip the order of representation between different facilities in any way but did shed some insight into the representation of those datasets. This indicates that generally, the strata that are larger tend to be less representative compared to smaller strata. In this case, the larger strata sizes tend to occur in higher household income brackets, indicating that lower income brackets are closer to true representation than higher income brackets. The likely cause of this is that higher income brackets tend to be overrepresented in these datasets, which leads to a lower P_x. All of the above trends must be kept in mind when using these datasets for decision making.

7.6 Discussion on Implications for Real World Decisions

Understanding the representation of the datasets that describe these different facilities holds many implications for the decisions made by those operating the facilities. The real-world decision that will be analyzed here is the implementation of a low-income toll program by WSDOT. The goal of the program is to support low-income individuals in using the tolling facility in some way. There are many different methods for implementing a low-income toll program. The goal of this section will not be to identify the most appropriate program style for WSDOT; instead, we will assess how WSDOT can use the data available to it to make the decision.

The first key need to understand the equity implications of the data-driven decision-making process is understanding how the data will be used. The WSDOT data will primarily be used to understand the travel behavior of those using the tolling facilities. For this purpose, representation is important, especially for those of lower household incomes, when thinking about a low-income toll program. Therefore, the WSDOT data describing SR 99, SR 167, and SR 16 seem to be the most critical in this regard as they have consistently the highest representation. This trend also

holds true for the CS data. The key feature of the CS data that will make it useful in the decision-making process is that it contains all trips for the region as opposed to just those that used the toll facilities. This allows for the extraction of trips that share an origin/destination pair that do not use a toll facility with trips that do use the toll facility. This allows for the demographics of alternate trips to be analyzed to understand who isn't but could be using the toll facility. Both of these sets of data are critical for designing a low-income toll program; it is important to understand how the changes made will affect both those using the toll facilities and those not using the toll facilities. This framework provides understanding both to the underlying demographics of use of these particular tolling facilities but also the broader understanding of relative representation for each dataset. From the results of the analysis, it seems that the data describing SR 167, SR 99, and SR 16 are most important for this purpose as they are the most representative.

This finding does not, however, diminish the importance of the other datasets. Many low-income individuals use the SR 520 and I-405 facilities to great personal benefit and would also benefit from the low-income toll program. Instead, this finding serves only to inform the decision-makers at WSDOT to consider the fact that the data describing those two facilities is not as representative overall as the other facilities. The same phenomenon is true of the CS data; across the board the CS data is slightly less representative than the WSDOT data. However, this certainly does not diminish its importance as it contains critical information that cannot be found in the WSDOT data. Again, the decision-makers must account for these differences in representation in their analysis and decision making as this program is designed and implemented.

7.7 Chapter Summary

In this chapter, we conducted a case study to calculate the representation of different datasets for WSDOT tolling facilities. Throughout this chapter we 1) showcase the implementation of the methodological framework for real-world datasets and 2) provide insight into the representation of different datasets on WSDOT tolling facilities and the impact this has on real-world decision making. We followed the framework set forward in Chapter 4 to calculate the representation of both WSDOT and CS data for each toll facility in the state of Washington. We found that overall, the data describing SR 520 is the least representative, followed by I-405, with the data for SR 167, SR 99, and SR 16 all being similarly representative. We also found that the CS data was generally slightly less representative for each facility that it could cover. Finally, the impact of these data sources on decision making was assessed with the WSDOT low-income toll program. We conclude that WSDOT must consider the differences in representation of each dataset when designing its program, however that lower representation does not necessarily diminish the importance of those datasets.

Chapter 8: Conclusions, Challenges and Future Directions

8.1 Research Findings and Contributions

As transportation practitioners begin to more frequently use data in the decision-making process, it is critical to ensure that the data used is as representative of stakeholder populations as possible to maintain the highest level of equity. This dissertation addresses this need by introducing a methodological framework to identify and quantify equity biases in the representation of datasets. We began with a thorough literature review to understand the intersection of equity, big data, and decision making, culminating in a definition of equity as it relates to representation. We then define a methodological framework by which the relative representation of different datasets can be quantified such that practitioners can assess some of the equity implications of using different datasets for decision making. Finally, we conducted several case studies utilizing this methodological framework to assess the representation of various datasets that describe Washington tolling facilities and show the impact this has on real world policy decisions to be made about these facilities. The key findings and contributions of this work is summarized:

- A thorough understanding of the interactions between equity, big data, and data bias in the field of transportation.
    - To understand the historical and current practices for transportation equity and big data, Chapter 2 covers a literature review of the current state-of-the-art in these fields. This includes both historic and current studies of equity, as well as a description of how big data is changing the field of transportation.

- To connect these concepts of equity and big data, Chapter 3 introduces the concepts and current practices related to data bias and uses this to form a definition of equity as it relates to representation in data for data-driven decision-making. This includes a review of data biases and the equity implications these biases have on transportation decision making.

- A robust methodological framework to quantify and assess equity biases in transportation dataset.
  - To create the implementable framework to find the representation of a dataset, Chapter 4 introduces a methodological framework which can be used to find the relative representation in a repeatable way. This includes defining the key conceptual step that must be taken to address the definitions of equity found above.
  - To explore the variables that must be included in the above framework, Chapter 4 defines the fundamental terms that must be used to calculate this value. This includes various calculations to ensure continuity through the methodological framework.
- An introduction to the method of ecological regression for various transportation equity applications including the definition of transportation demographics for understanding representation equity.
  - To provide an introduction of the ecological regression method, Chapter 5 showcases the key functionalities of this statistical methodology. This includes how it can be used to calculate the demographic profiles of transportation networks.

    - To showcase how this method can be used to assess equity, Chapter 6 uses ecological regression to find the demographics of a real-world transportation network, the Washington State Department of Transportation (WSDOT) tolling facilities. This includes showing how toll users are distributed geographically and demographically throughout the region.
- A model by which practitioners can understand the equity implications of the datasets being used to make data-informed decisions at all levels.
    - To showcase the implementation of the above methodological framework, Chapter 7 finds the relative representation of several real-world datasets. This includes both the WSDOT tolling data, described above, as well as an alternate data source utilizing cellular trace data which also describes the tolling facilities in Washington.
    - To provide insight into the impacts that the relative representation has on real-world decision making, Chapter 7 assesses the implications of using the above datasets for decision making purposes. This is done through the lens of WSDOT's proposed low-income toll tag program, a real-world policy that is currently in the process of being assessed and developed and showcases how the methodological framework can be used for both understanding the underlying statistics, over- and under-representation, and demographics of the tolling facilities as well as the broader summary statistics.

These primary contributions are of great impact to the transportation engineering community. The understanding of equity and big data, the statistical methodologies, and the methodological framework allow for practitioners to be proactive instead of reactive in addressing

equity in the data-driven decision-making process. In previous examples, such as the Seoul Owl Bus and the cycling network of Helsinki, practitioners were only able recognize the significant bias present in their dataset after initial analyses were complete and then correct accordingly after the fact. With this framework, these practitioners could have recognized these issues earlier and collected more comprehensive data or used the corrective methodologies initially which can increase efficiency and equity for these projects. Since this framework provides output on both a granular and broad level, it can fundamentally change how practitioners understand and approach the data-driven decision-making process.

8.2 Future Directions of Study, Challenges, and Opportunities

There are many future extensions of this work which can further exploration of representation in data. As this is the first implemented framework for assessing representation in data, there is significant room for expansion in this field. The key methodological framework will continue to evolve in both the general formulation and the statistical methodologies used to facilitate the framework. Furthermore, this framework must remain malleable as the representation of more varying datasets are explored.

There are several challenges and further questions these extensions bring forward as the equity implications of representation continue to be studied:

- How do we calculate absolute representation? This methodological framework only calculates relative representation, meaning that the outcomes are very useful for comparison between datasets but on their own do not carry meaningful information.

- What other statistical methodologies can be used to calculate the demographics of transportation networks? This methodological framework relies on statistical methodologies to find the demographics of the transportation network that is being described by the dataset of study. In this dissertation we use ecological regression to calculate this critical term in the framework, however this framework is not tied specifically to this statistical method and other methods may provide better or more replicable results.
- What further data can be gathered to augment this framework? Currently this methodological framework largely relies on census data for all demographic purposes and other transportation data to determine the uses of the transportation, but there are rarely any transportation specific datasets which include accurate demographics.

All of these challenges provide several key areas where research can be conducted to fill critical remaining gaps. To address the question of absolute representation, firstly more work must be conducted in the defining of representation for data. The various definitions presented in this dissertation are sufficient to be used for relative representation, however these are not sufficient to define absolute representation. Creating such a definition could provide a better theoretical understanding of the impacts that representation has on equity which can be applied to this methodological framework. Second, to address the question of different statistical methodologies, this methodological framework must be used with novel statistical methodologies to understand their fitness for this kind of analysis. There are several promising methodologies and technologies that can be used for this purpose. Artificial intelligence and machine learning models show great promise for improving on the performance of more classical methods. These types of models can be applied to this methodological framework to improve its performance. Finally, to address the

question of collecting more data to be used in the methodological framework, there can be new datasets that can augment the results of this framework, specifically datasets that include transportation demographics. In theory, if accurate demographic data were collected for the transportation network of study, this could greatly reduce the need for and importance of the statistical methodologies described above, and potentially even eliminate them. This will be a benefit as we will no longer be reliant on models to predict demographics, instead directly measuring it which can reduce total error. Additionally, this kind of data has many applications for general assessments of transportation equity. However, there are two critical challenges that need to be overcome with regards to this area: collecting direct demographics data by nature is extremely challenging and the data privacy of transportation users must be maintained. These two points are also somewhat self-reinforcing, where stringent data privacy requirements limit the types and amount of identifiable data to be collected. Addressing these challenges while maintaining the privacy of transportation users will support the further exploration of representation and equity in transportation to support real-world data-driven decision-making, creating a better and more equitable transportation network and society overall.

Bibliography

Allen, Jeff, and Steven Farber. 2021. "Suburbanization of Transport Poverty." *Annals of the American Association of Geographers* 111 (6): 1833–50. https://doi.org/10.1080/24694452.2020.1859981.

Antoniou, C., R. Balakrishna, and H. Koutsopoulos. 2008. "Emerging Data Collection Technologies and Their Impact on Traffic Management Applications." 10th International Conference on Application of Advanced Technologies in Transportation. https://doi.org/10.1007/s12544-011-0058-1

Asensio, C., J. M. López, R. Pagán, I. Pavón, and M. Ausejo. 2009. "GPS-Based Speed Collection Method for Road Traffic Noise Mapping." *Transportation Research Part D: Transport and Environment* 14 (5): 360–66. https://doi.org/10.1016/j.trd.2009.03.008.

Bagchi, M., and P. R. White. 2005. "The Potential of Public Transport Smart Card Data." *Transport Policy*, Road User Charging: Theory and Practices, 12 (5): 464–74. https://doi.org/10.1016/j.tranpol.2005.06.008.

Basche, L. 2011. "Says Solving 'Big Data' Challenge Involves More than Just Managing Volumes of Data." *Bus. Wire, San Francisco CA, USA Tech. Rep*, June 2011.

Beiler, Michelle, and Mona Mohammed. 2016. "Exploring Transportation Equity: Development and Application of a Transportation Justice Framework." *Transportation Research Part D: Transport and Environment* 47 (August): 285–98. https://doi.org/10.1016/j.trd.2016.06.007.

Bergman, Cecilia, and Juha Oksanen. 2016. "Estimating the Biasing Effect of Behavioural Patterns on Mobile Fitness App Data by Density-Based Clustering." In *Geospatial Data in a Changing World*, edited by Tapani Sarjakoski, Maribel Yasmina Santos, and L. Tiina Sarjakoski, 199–218.

Lecture Notes in Geoinformation and Cartography. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-33783-8_12.

Bills, Tierra S., and Joan L. Walker. 2017. "Looking beyond the Mean for Equity Analysis: Examining Distributional Impacts of Transportation Improvements." *Transport Policy* 54 (February): 61–69. https://doi.org/10.1016/j.tranpol.2016.08.003.

Bollier, David. 2010. *The Promise and Peril of Big Data*. Washington, DC: Aspen Inst.

Boucher, David, and Paul Joseph Kelly. 1998. *Social Justice: From Hume to Walzer*. Psychology Press.

Bullard, Robert Doyle, Glenn Steve Johnson, and Angel O. Torres. 2004. *Highway Robbery: Transportation Racism & New Routes to Equity*. South End Press.

Burris, Mark, Sunghoon Lee, Tina Geiselbrecht, Trey Baker, and Texas A&M Transportation Institute. 2013. "Equity Evaluation of Sustainable Mileage-Based User Fee Scenarios." SWUTC/14/600451-00007-1. https://rosap.ntl.bts.gov/view/dot/26648.

Carleton, Phillip R., and J. David Porter. 2018. "A Comparative Analysis of the Challenges in Measuring Transit Equity: Definitions, Interpretations, and Limitations." *Journal of Transport Geography* 72 (October): 64–75. https://doi.org/10.1016/j.jtrangeo.2018.08.012.

Chen, Cailian, Tom Hao Luan, Xinping Guan, Ning Lu, and Yunshu Liu. 2017. "Connected Vehicular Transportation: Data Analytics and Traffic-Dependent Networking." *IEEE Vehicular Technology Magazine* 12 (3): 42–54. https://doi.org/10.1109/MVT.2016.2645318.

Chen, Hsinchun, Roger Chiang, and Veda Storey. 2012. "Business Intelligence and Analytics: From Big Data to Big Impact." *MIS Quarterly* 36 (4): 1165–88. https://doi.org/10.2307/41703503.

Chen, Ying, Andreas Frei, and Hani S. Mahmassani. 2014. “From Personal Attitudes to Public Opinion: Information Diffusion in Social Networks Toward Sustainable Transportation.” *Transportation Research Record* 2430 (1): 28–37. https://doi.org/10.3141/2430-04.

Clair, David J. St. 1981. “The Motorization and Decline of Urban Public Transit, 1935–1950.” *The Journal of Economic History* 41 (3): 579–600. https://doi.org/10.1017/S002205070004434X.

Cochran, Stephen. 1994. “Transportation, Social Equity, and City-Suburban Connections.” In *Planning and Community Equity*. Routledge.

Cottrill, Caitlin D., and Sybil Derrible. 2015. “Leveraging Big Data for the Development of Transport Sustainability Indicators.” *Journal of Urban Technology* 22 (1): 45–64. https://doi.org/10.1080/10630732.2014.942094.

Courage, K. G., M. Doctor, S. Maddula, and R. Surapaneni. 1996. “Video Image Detection for Traffic Surveillance and Control,” March. https://trid.trb.org/view/473231.

Crawford, Kate. 2013. “The Hidden Biases in Big Data.” *Harvard Business Review*, 4.

Delgado-Rodríguez, M., and J. Llorca. 2004. “Bias.” *Journal of Epidemiology & Community Health* 58 (8): 635–41. https://doi.org/10.1136/jech.2003.008466.

Di Ciommo, Floridea, and Yoram Shiftan. 2017. “Transport Equity Analysis.” *Transport Reviews* 37 (2): 139–51. https://doi.org/10.1080/01441647.2017.1278647.

Diao, Mi, Yi Zhu, Joseph Ferreira, and Carlo Ratti. 2016. “Inferring Individual Daily Activities from Mobile Phone Traces: A Boston Example.” *Environment and Planning B: Planning and Design* 43 (5): 920–40. https://doi.org/10.1177/0265813515600896.

Dumbaugh, Eric, Jeffrey Tumlin, and Wesley E. Marshall. 2014. “Decisions, Values, and Data: Understanding Bias in Transportation Performance Measures.” *Institute of Transportation Engineers. ITE Journal* 84 (8): 20–25.

Eksler, V., S. Lassarre, and I. Thomas. 2008. “Regional Analysis of Road Mortality in Europe.” *Public Health* 122 (9): 826–37. https://doi.org/10.1016/j.puhe.2007.10.003.

Elman, Colin. 2005. “Explanatory Typologies in Qualitative Studies of International Politics.” *International Organization* 59 (2): 293–326. https://doi.org/10.1017/S0020818305050101.

ESRI. 2022. “Data Classification Methods—ArcGIS Pro | Documentation.” Data Classification Methods. 2022. https://pro.arcgis.com/en/pro-app/latest/help/mapping/layer-properties/data-classification-methods.htm.

Fan, Yingling, Andrew Guthrie, Leoma Van Dort, and Gina Baas. 2019. “Advancing Transportation Equity: Research and Practice,” 81.

FHWA. 2008. “Income-Based Equity Impacts of Congestion Pricing: A Primer.” (FHWA-HOP-08-040). Washington, DC: Federal Highway Administration.

FHWA. 2022. “21st Century Operations Using 21st Century Technology.” Tolling and Pricing Program. Federal Highway Administration.

Frey, William H. 1979. “Central City White Flight: Racial and Nonracial Causes.” *American Sociological Review* 44 (3): 425–48. https://doi.org/10.2307/2094885.

Gal-Tzur, Ayelet, Susan M. Grant-Muller, Tsvi Kuflik, Einat Minkov, Silvio Nocera, and Itay Shoor. 2014. “The Potential of Social Media in Delivering Transport Policy Goals.” *Transport Policy* 32 (March): 115–23. https://doi.org/10.1016/j.tranpol.2014.01.007.

Gamage, Pandula. 2016. “New Development: Leveraging ‘Big Data’ Analytics in the Public Sector.” *Public Money & Management* 36 (5): 385–90. https://doi.org/10.1080/09540962.2016.1194087.

García-Albertos, Pedro, Miguel Picornell, María Henar Salas-Olmedo, and Javier Gutiérrez. 2019. “Exploring the Potential of Mobile Phone Records and Online Route Planners for Dynamic

Accessibility Analysis." *Transportation Research Part A: Policy and Practice* 125 (July): 294–307. https://doi.org/10.1016/j.tra.2018.02.008.

Garrett, Mark, and Brian Taylor. 1999. "Reconsidering Social Equity in Public Transit." *Berkeley Planning Journal* 13 (1). https://doi.org/10.5070/BP313113028.

Gelman, Andrew, David K. Park, Stephen Ansolabehere, Phillip N. Price, and Lorraine C. Minnite. 2001. "Models, Assumptions and Model Checking in Ecological Regressions." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 164 (1): 101–18. https://doi.org/10.1111/1467-985X.00190.

Gong, Hongmian, Cynthia Chen, Evan Bialostozky, and Catherine T. Lawson. 2012. "A GPS/GIS Method for Travel Mode Detection in New York City." *Computers, Environment and Urban Systems*, Special Issue: Geoinformatics 2010, 36 (2): 131–39. https://doi.org/10.1016/j.compenvurbsys.2011.05.003.

Gong, Li, Xi Liu, Lun Wu, and Yu Liu. 2016. "Inferring Trip Purposes and Uncovering Travel Patterns from Taxi Trajectory Data." *Cartography and Geographic Information Science* 43 (2): 103–14. https://doi.org/10.1080/15230406.2015.1014424.

Gonzalez, Maria E., Jack L. Ogus, Gary Shapiro, and Benjamin J. Tepping. 1975. "Standards for Discussion and Presentation of Errors in Survey and Census Data." *Journal of the American Statistical Association* 70 (351b): 5–23. https://doi.org/10.1080/01621459.1975.10481469.

Goodman, Leo A. 1959. "Some Alternatives to Ecological Correlation." *American Journal of Sociology* 64 (6): 610–25. https://doi.org/10.1086/222597.

Greenland, Sander, and Hal Morgenstern. 1989. "Ecological Bias, Confounding, and Effect Modification." *International Journal of Epidemiology* 18 (1): 269–74. https://doi.org/10.1093/ije/18.1.269.

Greenland, Sander, and James Robins. 1994. “Invited Commentary: Ecologic Studies—Biases, Misconceptions, and Counterexamples.” *American Journal of Epidemiology* 139 (8): 747–60. https://doi.org/10.1093/oxfordjournals.aje.a117069.

Griffin, Greg P., Megan Mulhall, Chris Simek, and William W. Riggs. 2020. “Mitigating Bias in Big Data for Transportation.” *Journal of Big Data Analytics in Transportation* 2 (1): 49–59. https://doi.org/10.1007/s42421-020-00013-0.

Griffin, Greg Phillip, and Junfeng Jiao. 2019. “Crowdsourcing Bicycle Volumes: Exploring the Role of Volunteered Geographic Information and Established Monitoring Methods.” SocArXiv Papers. https://doi.org/10.31235/osf.io/e3hbc.

Guenduez, Ali A., Tobias Mettler, and Kuno Schedler. 2020. “Technological Frames in Public Administration: What Do Public Managers Think of Big Data?” *Government Information Quarterly* 37 (1): 101406. https://doi.org/10.1016/j.giq.2019.101406.

Herrera, Juan C., Daniel B. Work, Ryan Herring, Xuegang (Jeff) Ban, Quinn Jacobson, and Alexandre M. Bayen. 2010. “Evaluation of Traffic Data Obtained via GPS-Enabled Mobile Phones: The Mobile Century Field Experiment.” *Transportation Research Part C: Emerging Technologies* 18 (4): 568–83. https://doi.org/10.1016/j.trc.2009.10.006.

Heyer, Johanna, Matthew Palm, and Deb Niemeier. 2020. “Are We Keeping up? Accessibility, Equity and Air Quality in Regional Planning.” *Journal of Transport Geography* 89 (December): 102891. https://doi.org/10.1016/j.jtrangeo.2020.102891.

Höchtl, Johann, Peter Parycek, and Ralph Schöllhammer. 2016. “Big Data in the Policy Cycle: Policy Decision Making in the Digital Era.” *Journal of Organizational Computing and Electronic Commerce* 26 (1–2): 147–69. https://doi.org/10.1080/10919392.2015.1125187.

Hong, Sounman, Sun Hyoung Kim, Youngrok Kim, and Jeongin Park. 2019. “Big Data and Government: Evidence of the Role of Big Data for Smart Cities.” *Big Data & Society* 6 (1): 2053951719842543. https://doi.org/10.1177/2053951719842543.

Huang, Enyang. 2010. “Algorithmic and Implementation Aspects of On-Line Calibration of Dynamic Traffic Assignment.” Thesis, Massachusetts Institute of Technology. https://dspace.mit.edu/handle/1721.1/60808.

Jackson, Christopher. 2006. “Ecoreg Guide.” MRC Biostatistics Unit, Cambridge.

Kammoun, Karim, Aymen Ghédira, Chaker Ben Saad, and Nesrine Bouhamed. 2020. “Analysis of Road Mortality in Digital Age Using Bayesian Ecological Model: The Case of Tunisia.” *World Review of Intermodal Transportation Research* 9 (4): 393–409. https://doi.org/10.1504/WRITR.2020.111063.

Kang, Chaogui, Yu Liu, Xiujun Ma, and Lun Wu. 2012. “Towards Estimating Urban Population Distributions from Mobile Call Data.” *Journal of Urban Technology* 19 (4): 3–21. https://doi.org/10.1080/10630732.2012.715479.

Karner, Alex. 2016. “Planning for Transportation Equity in Small Regions: Towards Meaningful Performance Assessment.” *Transport Policy* 52 (November): 46–54. https://doi.org/10.1016/j.tranpol.2016.07.004.

Karner, Alex, Jonathan London, Dana Rowangould, and Kevin Manaugh. 2020. “From Transportation Equity to Transportation Justice: Within, Through, and Beyond the State.” *Journal of Planning Literature* 35 (4): 440–59. https://doi.org/10.1177/0885412220927691.

King, David. 2009. “Remediating Equity in Transportation Finance.” Special Report 303. Equity of Evolving Transportation Finance Mechanisms. Committee on the Equity Implications of Evolving Transportation Finance Mechanisms: Transportation Research Board.

Krol, Robert. 2017. “How Congestion Pricing Influences Equity.” No. 07756, Mercatus Center: George Mason University.

Kuai, Xuan, and Fahui Wang. 2020. “Global and Localized Neighborhood Effects on Public Transit Ridership in Baton Rouge, Louisiana.” *Applied Geography* 124 (November): 102338. https://doi.org/10.1016/j.apgeog.2020.102338.

Kwayu, Keneth Morgan, Sia Macmillan Lyimo, and Valerian Kwigizile. 2021. “Characteristics of Cyclists Using Fitness Tracker Apps and Its Implications for Planning of Bicycle Transport Systems.” *Case Studies on Transport Policy* 9 (3): 1160–66. https://doi.org/10.1016/j.cstp.2021.06.004.

Lempert, Robert, James Syme, George Mazur, Debra Knopman, Garett Ballard-Rosa, Kacey Lizon, and Ifeanyi Edochie. 2020. “Meeting Climate, Mobility, and Equity Goals in Transportation Planning Under Wide-Ranging Scenarios.” *Journal of the American Planning Association* 86 (3): 311–23. https://doi.org/10.1080/01944363.2020.1727766.

Leung, Shirley, Cory McCartan, CJ Robinson, Kiana Roshan Zamir, Mark Hallenbeck, and Vaughn Iverson. 2019. “I-405 Express Toll Lanes: Usage, Benefits, and Equity.” University of Washington eScience Institute.

Lewis, Elyse O’Callaghan, Don MacKenzie, and Jessica Kaminsky. 2021. “Exploring Equity: How Equity Norms Have Been Applied Implicitly and Explicitly in Transportation Research and Practice.” *Transportation Research Interdisciplinary Perspectives* 9 (March): 100332. https://doi.org/10.1016/j.trip.2021.100332.

Lin, Ying-Tung, Tzu-Wei Hung, and Linus Ta-Lun Huang. 2020. “Engineering Equity: How AI Can Help Reduce the Harm of Implicit Bias.” *Philosophy & Technology*, July. https://doi.org/10.1007/s13347-020-00406-7.

Litman, Todd. 2018. “Evaluating Transportation Equity.” *World Transportation Policy and Practice*, November.

Liu, Sen, Huanhuan Cao, Lei Li, and MengChu Zhou. 2013. “Predicting Stay Time of Mobile Users With Contextual Information.” *IEEE Transactions on Automation Science and Engineering* 10 (4): 1026–36. https://doi.org/10.1109/TASE.2013.2259480.

Lopes, J., J. Bento, E. Huang, C. Antoniou, and M. Ben-Akiva. 2010. “Traffic and Mobility Data Collection for Real-Time Applications.” In *13th International IEEE Conference on Intelligent Transportation Systems*, 216–23. https://doi.org/10.1109/ITSC.2010.5625282.

Martens, Karel, and Aaron Golub. 2021. “A Fair Distribution of Accessibility: Interpreting Civil Rights Regulations for Regional Transportation Plans.” *Journal of Planning Education and Research* 41 (4): 425–44. https://doi.org/10.1177/0739456X18791014.

Martens, Karel, Matan E. Singer, and Aviv Lee Cohen-Zada. 2022. “Equity in Accessibility.” *Journal of the American Planning Association* 88 (4): 479–94. https://doi.org/10.1080/01944363.2021.2016476.

McArdle, Gavin, and Rob Kitchin. 2016. “Improving the Veracity of Open and Real-Time Urban Data.” *Built Environment* 42 (3): 457–73. https://doi.org/10.2148/benv.42.3.457.

Miller, Harvey J. 1999. “Measuring Space-Time Accessibility Benefits within Transportation Networks: Basic Theory and Computational Procedures.” *Geographical Analysis* 31 (1): 187–212. https://doi.org/10.1111/gean.1999.31.1.187.

Mohl, Raymond. 2004. “Stop the Road: Freeway Revolts in American Cities - Raymond A. Mohl, 2004.” *Journal of Urban History* 30 (5): 674–706. https://doi.org/10.1177%2F0096144204265180.

O'Leary, Daniel. 2013. "Exploiting Big Data from Mobile Device Sensor-Based Apps: Challenges and Benefits." *MIS Quarterly Executive* 12:4 (December): 179–87.

Pelletier, Marie-Pier, Martin Trepanier, and Catherine Morency. 2011. "Smart Card Data Use in Public Transit: A Literature Review." *Transportation Research Part C: Emerging Technologies* 19 (4): 557–68. https://doi.org/10.1016/j.trc.2010.12.003.

Pereira, Rafael, and Alex Karner. 2021. "Transportation Equity." *International Encyclopedia of Transportation*, 271–77. https://doi.org/10.1016/B978-0-08-102671-7.10053-3.

Perez, Benjamin G., Charles Fuhs, Colleen Gants, Reno Giordano, David H. Ungemah, and Wayne Berman. 2012. "Priced Managed Lane Guide." FHWA-HOP-13-007. https://rosap.ntl.bts.gov/view/dot/41487.

Pesesky, Lawrence, Deborah Matherly, Leigh Lane, David Aimen, Deva Deka, Michael Smart, Asha Weinstein Agrawal, Bruce Brown, and Anne Morris. 2018. "Assessing the Environmental Justice Effects of Toll Implementation or Rate Changes: Guidebook and Toolbox." *NCHRP Research Report*, no. 860. https://trid.trb.org/view/1498380.

Plotnick, Robert, Jennifer Romich, and Jennifer Thacker. 2009. "The Impact of Tolling on Low-Income Persons in the Puget Sound Region." Olympia, Washington: Washington State Transportation Commission.

Prentice, Ross L., and Lianne Sheppard. 1995. "Aggregate Data Studies of Disease Risk Factors." *Biometrika* 82 (1): 113–25. https://doi.org/10.1093/biomet/82.1.113.

"Puget Sound Regional Council." n.d. Puget Sound Regional Council. Accessed May 9, 2022. https://www.psrc.org/.

Qi, Luo. 2008. “Research on Intelligent Transportation System Technologies and Applications.” In *2008 Workshop on Power Electronics and Intelligent Transportation System*, 529–31. https://doi.org/10.1109/PEITS.2008.124.

Reinsel, David, John Gantz, and John Rydning. 2017. “Total WW Data to Reach 163 ZB by 2025.” Storage Newsletter.

Richardson, Sylvia, and Christine Monfort. 2000. “Ecological Correlation Studies.” *Spatial Epidemiology: Methods and Applications*, 205–20.

Richardson, Sylvia, Isabelle Stücker, and Denis Hémon. 1987. “Comparison of Relative Risks Obtained in Ecological and Individual Studies: Some Methodological Considerations.” *International Journal of Epidemiology* 16 (1): 111–20. https://doi.org/10.1093/ije/16.1.111.

Sanchez, Thomas W, Carrie Makarewicz, Peter M Haas, and Casey J Dawkins. 2006. “Transportation Costs, Inequities, and Tradeoffs.” In 85th Annual meeting of the Transportation Research Board, 16. Washington DC.

Sanchez, Thomas W, Rich Stolz, and Jacinta S. Ma. 2003. “Moving to Equity: Addressing Inequitable Effects Transportation Policies on Minorities.” Cambridge, MA: The Civil Rights Project at Harvard University. https://escholarship.org/content/qt5qc7w8qp/qt5qc7w8qp.pdf.

Sanchez, Thomas W., Rich Stolz, and Jacinta S. Ma. 2004. “Inequitable Effects of Transportation Policies on Minorities.” *Transportation Research Record* 1885 (1): 104–10. https://doi.org/10.3141/1885-15.

Schweitzer, Lisa A., and Nader Afzalan. 2017. “09 F9 11 02 9D 74 E3 5B D8 41 56 C5 63 56 88 C0: Four Reasons Why AICP Needs an Open Data Ethic.” *Journal of the American Planning Association* 83 (2): 161–67. https://doi.org/10.1080/01944363.2017.1290495.

Seely-Gant, Katie, and Lisa M. Frehill. 2015. “Exploring Bias and Error in Big Data Research.” *Journal of the Washington Academy of Sciences* 101 (3): 29–38.

Shearmur, Richard. 2015. “Dazzled by Data: Big Data, the Census and Urban Geography.” *Urban Geography* 36 (7): 965–68. https://doi.org/10.1080/02723638.2015.1050922.

Smith, Stanley K., and Mohammed Shahidullah. 1995. “An Evaluation of Population Projection Errors for Census Tracts.” *Journal of the American Statistical Association* 90 (429): 64–71. https://doi.org/10.1080/01621459.1995.10476489.

“Strategic Plan FY 2022.” 2022. U.S. Department of Transportation (USDOT).

Uhlemann, Elisabeth. 2015. “Autonomous Vehicles Are Connecting... [Connected Vehicles].” *IEEE Vehicular Technology Magazine* 10 (2): 22–25. https://doi.org/10.1109/MVT.2015.2414814.

Victoria Transport Policy Institute. 2019. “Road Pricing.” Congestion Pricing, Value Pricing, Toll Roads, and HOT Lanes.

Wakefield, Jonathan, and Ruth Salway. 2001. “A Statistical Framework for Ecological and Aggregate Studies.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 164 (1): 119–37. https://doi.org/10.1111/1467-985X.00191.

Wang, Xin, Shuxu Zhao, and Liang Dong. 2017. “Research and Application of Traffic Visualization Based on Vehicle GPS Big Data.” In *Proceedings of the Second International Conference on Intelligent Transportation*, edited by Huapu Lu, 293–302. Smart Innovation, Systems and Technologies. Singapore: Springer. https://doi.org/10.1007/978-981-10-2398-9_27.

Washington State Transportation Commission. 2022. “2022 WSTC Tolling Report & Tacoma Narrows Bridge Loan Update.”

Williamson, Andy. 2014. “Big Data and the Implications for Government.” *Legal Information Management* 14 (4): 253–57. https://doi.org/10.1017/S1472669614000553.

Yeganeh, Armin Jeddi, Ralph P. Hall, Annie R. Pearce, and Steve Hankey. 2018. "A Social Equity Analysis of the U.S. Public Transportation System Based on Job Accessibility." *Journal of Transport and Land Use* 11 (1): 1039–56.

Zannat, Khatun E, and Charisma F. Choudhury. 2019. "Emerging Big Data Sources for Public Transport Planning: A Systematic Review on Current State of Art and Future Research Directions." *Journal of the Indian Institute of Science* 99 (4): 601–19. https://doi.org/10.1007/s41745-019-00125-9.

Zeyu, Jiang, Yu Shuiping, Zhou Mingduan, Chen Yongqiang, and Liu Yi. 2017. "Model Study for Intelligent Transportation System with Big Data." *Procedia Computer Science*, Advances in Information and Communication Technology: Proceedings of 7th International Congress of Information and Communication Technology (ICICT2017), 107 (January): 418–26. https://doi.org/10.1016/j.procs.2017.03.132.

Zhou, Xingang, Anthony GO Yeh, Weifeng Li, and Yang Yue. 2018. "A Commuting Spectrum Analysis of the Jobs–Housing Balance and Self-Containment of Employment with Mobile Phone Location Big Data." *Environment and Planning B: Urban Analytics and City Science* 45 (3): 434–51. https://doi.org/10.1177/2399808317707967.

Zhu, Li, Fei Richard Yu, Yige Wang, Bin Ning, and Tao Tang. 2019. "Big Data Analytics in Intelligent Transportation Systems: A Survey." *IEEE Transactions on Intelligent Transportation Systems* 20 (1): 383–98. https://doi.org/10.1109/TITS.2018.2815678.

Zikopoulos, Paul, Dirk Deroos, Krishnan Parasuraman, Thomas Deutsch, James Giles, and David Corrigan. 2012. "Harnass the Power of Big Data The IBM Data Platform." *McGraw Hill Professional*, October.