DIRECT DEMAND ESTIMATION FOR BUS TRANSIT IN SMALL CITIES

by

Nathaniel Julius Shellhamer

A Thesis

Submitted to the Faculty of Purdue University
In Partial Fulfillment of the Requirements for the degree of

Master of Science in Civil Engineering

Lyles School of Civil Engineering
West Lafayette, Indiana
May 2019

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Samuel Labi, Co-Chair

Lyles School of Civil Engineering

Dr. Jon Fricker, Co-Chair

Lyles School of Civil Engineering

Dr. Brigitte Waldorf

Department of Agricultural Economics

Approved by:

Dr. Dulcy Abraham

Head of the Graduate Program

To my family and friends.

It wouldn’t have been possible without you.

ACKNOWLEDGMENTS

I would like to thank Dr. Labi for the countless hours of kind advice and mentorship throughout my academic career. Thank you for all the opportunities you have given me, and for your endless encouragement. I would also like to thank Dr. Fricker for sparking my interest in transit and planning, and for teaching many of my favorite classes at Purdue. You always drive me to make sure my work is the best it can be, and I truly appreciate it.

I would also like to thank Dr. Waldorf for helping me to think outside the box and for showing me that there is more than one way to analyze a problem. Thank you to the rest of the faculty in Civil Engineering for giving me the tools to be successful in all of my endeavors.

To Bryce, Randy, Joshua, and all the staff at CityBus, thank you for providing me the data for this project, and for answering my numerous questions along the way. I really appreciate it.

To my research group, colleagues, and fellow ITE officers, it has been a pleasure working with all of you. Thank you for making grad school enjoyable. I wish you continued success in all your endeavors.

Lastly, thank you to my parents, my brother, and to Monica. Thank you for your endless support and encouragement, and for being there for every step along my journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACS	American Community Survey
BG	Block group
BRT	Bus rapid transit
FTA	Federal Transit Administration
GIS	Geographic information systems
GLPTC	Greater Lafayette Public Transportation Corporation (Operating Name: CityBus)
MLR	Multiple linear regression
MSE	Mean squared error
NIST	National Institute of Standards and Technology
RMSE	Root mean square error
SLR	Simple linear regression
SS	Span of service
VIF	Variance inflation factor
WLCSC	West Lafayette Community School Corporation

ABSTRACT

Author: Shellhamer, Nathaniel, J. MSCE
Institution: Purdue University
Degree Received: May 2019
Title: Direct Demand Estimation for Bus Transit in Small Cities
Major Professors: Samuel Labi and Jon Fricker

Public transportation is vital for many people who do not have the means to use other forms of transportation. In small communities, transit service is often limited, due to funding constraints of the transit agency. In order to maximize the use of available funding resources, agencies strive to provide effective and efficient service that meets the needs of as many people as possible. To do this, effective service planning is critical.

Unlike traditional road-based transportation projects, transit service modifications can be implemented over the span of just a few weeks. In planning for these short-term changes, the traditional four-step transportation planning process is often inadequate. Yet, the characteristics of small communities and the resources available to them limit the applicability of existing transit demand models, which are generally intended for larger cities.

This research proposes a methodology for using population and demographic data from the Census Bureau, combined with stop-level ridership data from the transit agency, to develop models for forecasting transit ridership generated by a given geographic area with known population and socioeconomic characteristics. The product of this research is a methodology that can be applied to develop ridership models for transit agencies in small cities. To demonstrate the methodology, the thesis built ridership models using data from Lafayette, Indiana.

A total of four (4) ridership models are developed, giving a transit agency the choice to select a model, based on available data and desired predictive power. More complex models are expected to provide greater predictive power, but also require more time and data to implement. Simpler models may be adequate where data availability is a challenge. Finally, examples are

provided to aid in applying the models to various situations. Aggregation levels of the American Community Survey (ACS) data provided some challenge in developing accurate models, however, the developed models are still expected to provide useful information, particularly in situations where local knowledge is limited, or where additional information is unavailable.

1. INTRODUCTION

1.1 Introduction

The predominant transportation mode is the private automobile in most small cities. A small city is defined for this thesis as any city having a population no larger than 200,000 persons. Many of these cities are planned and built with this assumption in mind. The effect of such auto-centric planning is that it is difficult for those without access to a vehicle to get around. In addition, other modes of transport, including shared mobility modes such as ridesharing are often less accessible in small cities than they are in larger cities. Additionally, in smaller cities, less expensive pooled rideshare options (such as “Uber Pool” or “Lift Line”) are not generally available, meaning that the only rideshare options available are more expensive single-party point-to-point options. For these reasons, public transportation in these communities is often a lifeline for those who do not have access to a vehicle. In order to be effective, public transportation must adequately serve those who most rely on the service it provides. At the same time, it must operate under the constraints of increasingly limited financial resources, meaning that efficient service is also important. Therefore, service planning, or determining when, where, and how often an agency will provide service, is of utmost importance to many agencies.

Yet, many agencies have limited time and resources to adequately develop a comprehensive service planning program. Some make judgments about where to provide service based on where new developments are opened, such as apartment complexes or shopping centers. This method can work, but requires that the agency have significant prior experience understanding the ridership trends that these types of development can cause, in order to approximate the ridership that they may generate. This can be a challenge for smaller agencies, and those with less experience in service planning. Furthermore, factors such as the lifestyles of the residents of a new apartment complex or the type of shopping available at a shopping complex can have a significant impact on the ridership that can be expected. Finally, some agencies have the resources to purchase expensive service planning software, but not all agencies have the resources to make such investments.

Nearly all agencies maintain extensive archives of data required for federal reporting. This data can include information about ridership, fare payment method, vehicle load factors, and schedule adherence. Additionally, a wealth of data about the communities that these transit agencies serve is readily available from the Census Bureau (American Community Survey (ACS) 2015). It is believed that combining the agency-level service and ridership data with ACS data will prove useful for estimating transit ridership, and this can be an aid to transit agencies in completing service planning tasks.

Ridership prediction is useful to agencies for several reasons. It provides agencies with additional information to be used for current service planning tasks, such as modifying existing transit service, or adding new transit routes. Additionally, it provides agencies with a basis with which to assess the productivity of existing service. For example, if the actual ridership on a transit route is significantly higher than what is expected from ridership prediction, that route could be regarded as "highly performing" and may perhaps be a candidate for future service expansions or enhancements. Ridership prediction also gives agencies a way to forecast what ridership may look like should a significant change in population, land use, or transportation in an area change. For example, an agency could use ridership prediction to estimate the change in ridership resulting from the opening of a new apartment complex near an existing route. These tasks are essential for any agency to appropriately plan and operate service, and ridership prediction methods can aid in more accurately and efficiently completing them.

1.2 Research Significance

Existing work on transit demand estimation tends to be limited in scope and generally focused on larger cities (generally with populations greater than 1,000,000 persons), which makes it difficult to apply to smaller cities. This research is expected to prove useful to transit agencies in small cities for estimating ridership during service planning, an area that is not well represented in existing work. Additionally, the work is expected to demonstrate the usefulness of linking transit agency data with other publicly available data sources, such as American Community Survey (ACS) data. It is hoped that this research thesis will provide motivation for additional future work in this area.

1.3 Study Objectives

This research proposes a methodology for directly estimating demand for transit using Census Bureau data. The research will use spatial analysis to relate stop-level ridership data to block group level ACS data through an aggregation process. Regression techniques will then be used to develop models that estimate yearly block group level ridership using ACS data for the block group. This methodology will be applied to ACS data for the Lafayette/West Lafayette, Indiana metropolitan area, and corresponding ridership data from the Greater Lafayette Public Transportation Corporation (GLPTC), which provides public transportation service to the area. The results of the research will be a set of models that can be used to predict ridership. These models could be used for estimating ridership in areas where transit service is currently not provided. One such area is the Wabash Avenue neighborhood south of downtown Lafayette. Beyond the Lafayette/West Lafayette area, it is intended that the methodology and modeling process be clear enough so that it can be readily applied to other communities, which may have geographic, demographic, or transit service characteristics different than those of the Lafayette/West Lafayette area.

1.4 Overall Framework

Chapter 2 of this thesis presents a literature review which summarizes the existing work in the areas of transit service planning, transit ridership estimation, and transit demand modeling. Chapter 2 also highlights some of the challenges with existing literature that provide a basis for conducting this research.

Chapter 3 provides background information on the Lafayette/West Lafayette area, and the transit system serving the area. It also describes the input data sources, both from GLPTC and from the ACS and provides basic summary statistics and information. It also describes the necessary spatial analysis steps undertaken to prepare the data for modeling.

Chapter 4 provides a detailed methodology for the study. It begins with an introduction to the regression techniques used in analyzing the data. Next, it provides a detailed description of the modeling process, followed by a discussion of the calibration and validation techniques employed to prepare the final models.

Chapter 5 presents the final models, along with all necessary information to apply them for ridership predictions. It also discusses some of the necessary limitations that are associated with the models, which will be useful in determining their applicability beyond the Lafayette/West Lafayette area. Chapter 5 concludes with a discussion of the implications of this study.

Chapter 6 begins with a summary of the study, including work performed and major findings. It also presents several opportunities for expansion of the study through future work. Finally, the thesis closes with a thorough reference list and detailed appendices to aid in future application of the work.

2. LITERATURE REVIEW

2.1 Introduction

Transportation planning traditionally begins with the four-step trip-based process to build a travel demand model, typically following guidelines outlined by the National Cooperative Highway Research Program Report 716 (2012). This process, while thorough, is data-intensive, time-consuming, and has traditionally been focused on private vehicular transportation (as opposed to transit) as noted by Pas (1995).

The time, effort, and cost required to develop a complete origin and destination travel demand model is not insignificant. These factors make the process less practical for predicting and modeling transit demand. Transit planning, unlike the process for planning other types of transportation infrastructure, is a shorter-term process, due to the shorter amount of time it takes to implement a transit project when compared with other infrastructure projects. This is especially true for bus-based local transit systems.

2.2 Need for Demand Estimation

Transit agencies are tasked with providing mobility services that are both cost efficient to operate, and effective for those who use them. To do this, they set principles, or service standards to guide decisions that are to be made regarding service (Mistretta et. al. 2009). These service standards are often driven by actual or predicted ridership, particularly in cases where cost efficiency is a primary goal. In order to evaluate potential new services, or service changes, a predicted ridership is necessary to assess whether or not service is justified. In these cases, a demand estimation process is of utmost importance. In cases where existing service is being evaluated, it is sometimes necessary to predict ridership to compare with the actual ridership to determine the performance of a particular transit route, or transit service in a particular geographic area. Additionally, federal grant programs (such as New Starts grants from the Federal Transit Administration (FTA)) often require agencies to assess the demand in a particular area that may be targeted for improvements (FTA 2018). For these reasons, it is important to have a robust method for estimating transit demand.

2.3 Level of Analysis

An important consideration is the geographic level at which analysis is conducted. Mckee and Miljkovic (2007) noted the important fact that at larger scales of aggregation, resolution and thus predictive power may be lost in the data. On this note, work has been done to develop stop-level ridership prediction models. Pulugurtha and Agurla (2012) developed models to estimate bus boardings at the stop level in the Charlotte, NC area. In a similar work, Dill, Schlossberg, Ma and Meyer (2013) developed a model to predict stop-level ridership based on the urban form and land use characteristics around each stop. This work was completed for several transit agencies of varying size in Oregon. These models were developed using transit data for a variety of system types. Chu (2004) also developed a model to predict ridership at the stop-level, using data from various Census Surveys. While more applicable to smaller service areas, these models are less useful when planning significant service changes, or service to new areas, because precise stop locations generally aren't determined during the planning phase, but rather are determined based on location of major trip generators, street geometry, and other geographic characteristics according to Giannopoulos (1989). Thus, too small of a geographic scale can also present challenges.

Another factor in choosing a geographic scale is data availability. While having finely aggregated data (for example, at the level of one city block) would be a nearly perfect fit for planning transit service, detailed data at this level is not released by the Census Bureau out of concern for privacy for survey respondents (ACS 2015). For these reasons, the block group is chosen to be the level of analysis. It represents a geographic scope larger than the individual stop level, which is more useful for preliminary service planning, where specific stop locations may not be known yet, and also represents the smallest available geographic unit for which data is readily available from the Census Bureau.

2.4 Analytical Techniques

Several analytical techniques are available for analysis of transit demand. Some focus on purely statistical models and techniques, while others take a more econometric modeling approach.

2.4.1 Statistical Techniques

A variety of statistical models have been used in the past to evaluate transit ridership. Li, Yao, and Fu (2016) used a neural network model to evaluate urban rail ridership in Shanghai. The model estimated ridership with great accuracy, however it is of limited use during the planning phase, because it relies on information regarding an existing rail stop. Additionally, due to significant operational differences between urban rail and local bus systems, this model is of limited use to smaller agencies. Dajani and Sullivan (1976) evaluated ridership using census data and a path analysis in which a regression analysis is conducted using variables that are known to be correlated. This technique is potentially useful in certain situations, however the analysis was limited to home-to-work trips in a large urban area. Additionally, the authors had access to a complete origin and destination trip database, something that is not available to many small transit agencies.

Koppelman (1983) uses a multinomial logit model to predict transit ridership in response to changes in transit service. This approach is similar to that employed in the four-step planning process, which also traditionally uses logit models during the mode split phase. However, it is also reliant on the availability of complete origin and destination data, much like the path analysis discussed above.

These approaches presented unique statistical approaches for estimating ridership. However, all of them rely either on data that is typically not available to small transit agencies (a complete origin-destination trip matrix), or model ridership for systems that have significantly different operating characteristics (such as rail transit) than local bus service in small cities.

2.4.2 Econometric Techniques

Schmenner (1796) conducted a route-level econometric analysis in which he evaluated various bus routes as a function of fare, auto operating costs, and various service and demographic characteristics. He sought to determine the extent to which a particular route was reliant on operating subsidies, and what effect (if any) a change in bus company policy with regard to fare might have on ridership. It was found that headway and fare charged are the most significant determinants of transit ridership. While interesting, this particular study is limited in that it is very focused on route revenue, as opposed to ridership. This is of less use during planning phases, particularly in cases where a transit agency is evaluating coverage services, for which revenue is rarely a major concern.

Fricker and Shanteau (1986) used a simple demand model that assessed the change in ridership due to changes in in-vehicle and out-of-vehicle travel time, and fare. This model was used to develop an optimization program that seeks to minimize operating deficits for an entire small city transit network. They emphasize the need for approaches that are not "data-hungry", and for several potential solutions to aid in decision making using the results. Both of these principles will be applied in this thesis.

2.5 Variables Considered

It is also important to consider variable selection when evaluating ridership prediction models. Choosing the wrong set of variables can lead to models that either incorrectly predict ridership, or are not generally applicable beyond the specific dataset and region of analysis. The study by Pulugurtha and Agurla (2012) made use of land use characteristics in addition to demographics and service characteristics. While the inclusion of land use in the models likely increased their predictive power, this data is not easy to obtain, and is generally only available for larger metropolitan areas. This particular study focused on Charlotte, NC. The Dill, Schlossberg, Ma and Meyer (2013) study did include smaller cities as part of the analysis, but also incorporated land use data into the models.

If land use data is available, it can be quite helpful in developing ridership models. However, it is equally important to have ridership models that do not incorporate land use, particularly for cities where such data is not available, or is incomplete.

Another interesting study is by Cervero, Murakami, and Miller (2010), who estimated ridership for bus rapid transit systems in the Los Angeles area. Although this particular study appears to be more useful, it includes service characteristics that are specific to bus rapid transit (BRT), such as the availability of park and ride facilities near stations, the presence of dedicated transit lanes, and the number of feeder routes stopping nearby. If this model is applied more generally to local bus service, ridership is predicted to be very near to (or even less than) zero, due to the inclusion of BRT-specific variables. This makes these models much less useful in predicting ridership on local transit routes.

2.6 Summary

This thesis seeks to expand on many of the works discussed above in order to develop bus transit ridership models for small cities. Estimating demand is important for a variety of reasons, most importantly for service planning tasks. Additionally, it can be used as a tool for evaluating the performance of existing services. In order to estimate demand, an analysis level must be chosen. Choosing an analysis level (such as at the stop level) that is too small can result in models that are not useful for early-stage service planning. Similarly, choosing an analysis level that is too large (such as at the neighborhood level) can result in a model that is less accurate. It is also important to consider the availability of data when choosing an analysis level.

A variety of analytical techniques can be used to estimate ridership, including statistical methods such as regression, or econometric methods such as logit modeling. However, it is often found that the complex data needs for these approaches limit their applicability to larger areas where such data is readily available.

Selecting appropriate types of variables is important for developing models that are both accurate and useful across a wide variety of possible conditions. Including data such as land use characteristics can render models useless in areas for which this data is unavailable. Additionally, including mode-specific parameters (such as those for BRT or rail systems) can limit the applicability of the models to areas that operate similar modes. Choosing the correct number of variables is also important, and model clarity must be balanced with accuracy.

3. DATA COLLECTION AND PROCESSING

3.1 Brief overview of GLTPC Network

The Greater Lafayette Public Transportation Corporation operates public transportation services in the Lafayette/West Lafayette area (GLPTC 2018). The current system map is shown in Figure 1.

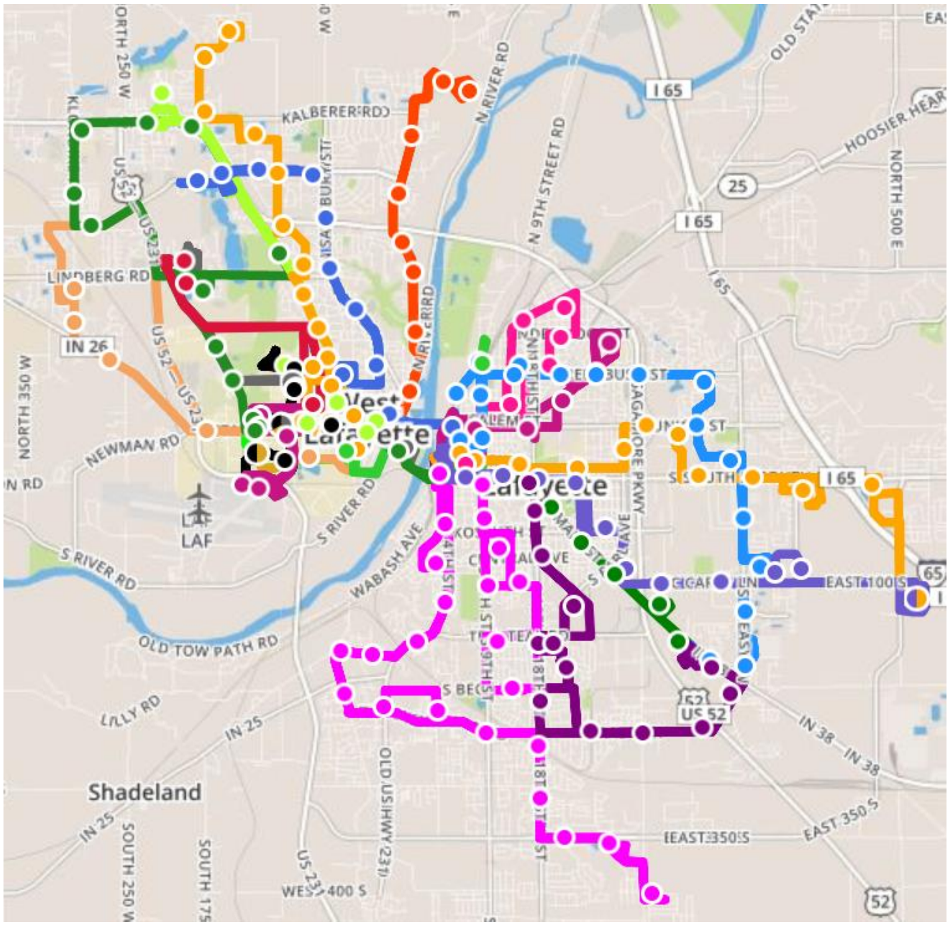

Figure 1. Greater Lafayette Public Transportation Corporation system map (GLPTC 2018).

With the exception of campus loop services, most regular routes begin and end at the CityBus Center, located in downtown Lafayette. Regular routes operate at 30-minute headways during weekday daytime periods. Additional weekday evening and late-night service is provided along select routes, typically at 60-minute headways. Saturday daytime service is provided on some routes at varying headways. Sunday daytime service is provided on some routes at 60-minute headways.

GLPTC also provides campus service to Purdue University, in the form of high-frequency loop routes, as well as shuttle services to nearby student apartment complexes. These services are only offered during times when Purdue classes are in session. Paratransit service provides, as required, curb-to-curb service to those unable to use regular service due to a disability. Finally, tripper service is provided to the West Lafayette Community School Corporation (WLCSC). Paratransit and WLCSC service are not considered as part of this analysis.

3.2 Description of GLPTC data

Ridership data is collected by GLPTC using automated passenger count and automated vehicle location devices installed on each bus in the fleet. These devices automatically count individuals as they pass through the door of the bus and attribute this count to the bus stop location using the vehicle location. Ridership data is also recorded by the farebox installed on each bus, with the operator keying the type of fare each passenger pays. The result of either method is a count of bus boardings at each stop in the GLPTC system. This data is aggregated into monthly totals. For the purposes of this study, the monthly totals are further aggregated into yearly totals. This is done to account for monthly fluctuations in ridership due to weather, road construction, Purdue classes being in session, or other factors outside the scope of this research.

Route and schedule data are needed for each service operated by GLPTC. This data is used both to link bus stops to the routes they are served by, as well as to assess the level of service provided to each stop. According to the Transit Capacity and Quality of Service Manual (TCRP 2013), two important parameters for assessing quality of transit service are headway and span of service. Headway refers to the time between buses, and is the inverse of frequency. For example, a frequency of 2 buses/hour implies a headway of 30 minutes between buses. The term

“Span of service” (SS) refers to the number of hours per day that service is provided to a particular area or stop.

These two parameters vary with the service day. Service is generally provided at longer headways during times of lower demand (evenings and weekends), and shorter spans of service are typically provided during those times as well. To account for this, an average yearly headway and average yearly span of service are calculated for each route. These are shown in Table 1. For example, Route 1A Market Square has a yearly average span of service of 16 hours per day, and a yearly average headway of 36 minutes. It is noted that 2015 was chosen for the analysis year because fewer ridership impacts were experienced in that year due to major construction projects and resulting service changes.

Table 1. Yearly average span and headway for all routes in 2015.

Route Number	Route Name	Yearly Average Span (Hours:Minutes/Day)	Yearly Average Headway (Hours:Minutes)
1A	Market Square	16:00	0:36
1B	Salisbury	16:30	0:38
2A	Schuyler	11:45	0:34
2B	Union	11:45	0:34
3	Lafayette Square	14:04	0:40
4A	Tippecanoe Mall	15:55	0:36
4B	Purdue West	16:04	0:37
5A	Happy Hollow	13:10	0:30
5B	Northwestern	13:05	0:35
6A	South 4th Street	14:00	0:37
6B	South 9th Street	12:15	0:30
7	South Street	16:21	0:38
8	Willowbrook/Klondike Express	3:30	0:30
12	Gold Loop	11:15	0:15
13	Silver Loop	11:10	0:05
14	Black Loop	6:20	0:30
15	Tower Acres	11:00	0:10
16	Bronze Loop	11:05	0:30
17	Ross Ade	11:00	0:10
18	Nightrider	4:00	0:20
19	Inner Loop	10:48	0:15
20	AvTech	10:25	0:20
21	The Avenue	15:09	0:26
23	Connector	12:10	0:17
27	Outer Loop	10:48	0:15

Summary Statistics

Table 2 presents summary statistics regarding the service characteristics and of GLPTC overall. These statistics include information about all routes included in Table 1, which includes both regular/city routes and campus loop service.

Table 2. Summary statistics for GLPTC operations in 2015.

Parameter	**Value**	**Unit**
Annual unlinked passenger trips	4,318,655	Trips
Annual passenger miles	12,027,695	Miles
Average headway	0:27	Hours:Minutes
Average span of service	12:00	Hours/day
Vehicles operated in maximum service	55	Vehicles
Annual vehicle revenue miles	1,773,427	Miles
Number of stops served	802	Stops
Number of block groups served	71	Block groups

Figure 2 presents ridership for each block group visually. High ridership is indicated by areas of magenta and blue, while lower ridership is indicated by areas of yellow and orange. Only block groups included in the study (those with bus stops in them) are shown.

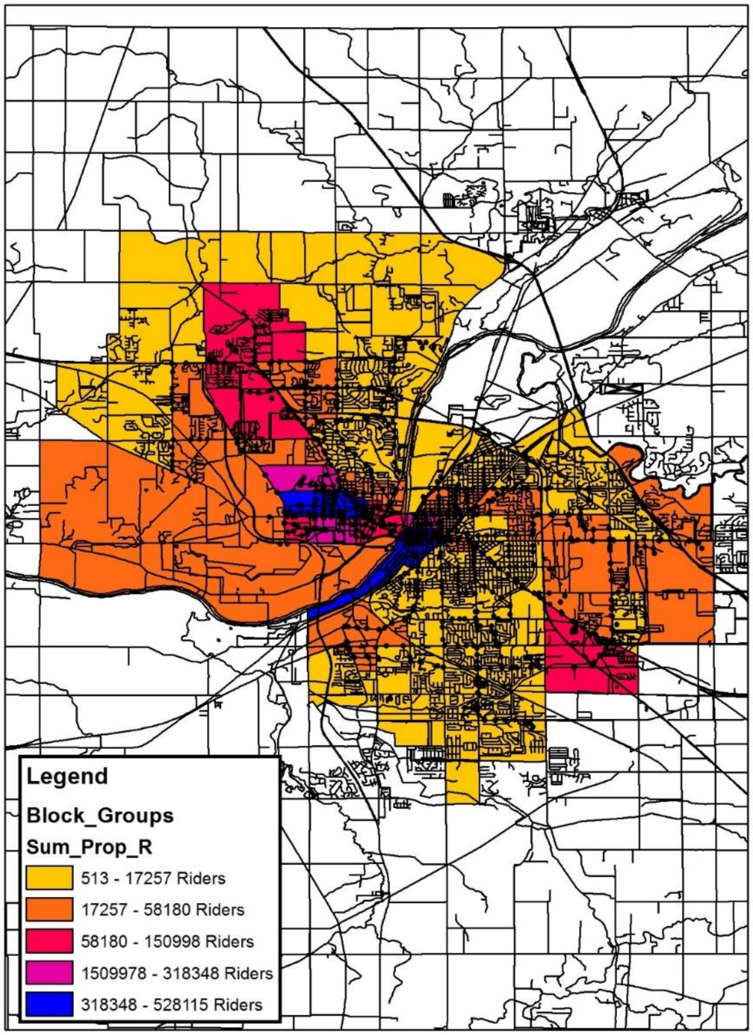

Figure 2. Ridership by block group.

Figure 3 presents population density expressed as persons/square mile for each block group. Together, Figure 2 and Figure 3 show why it often is not adequate to use only population density for ridership modeling. An area with high population density does not necessarily guarantee high transit ridership, and an area with lower population density does not imply low transit ridership.

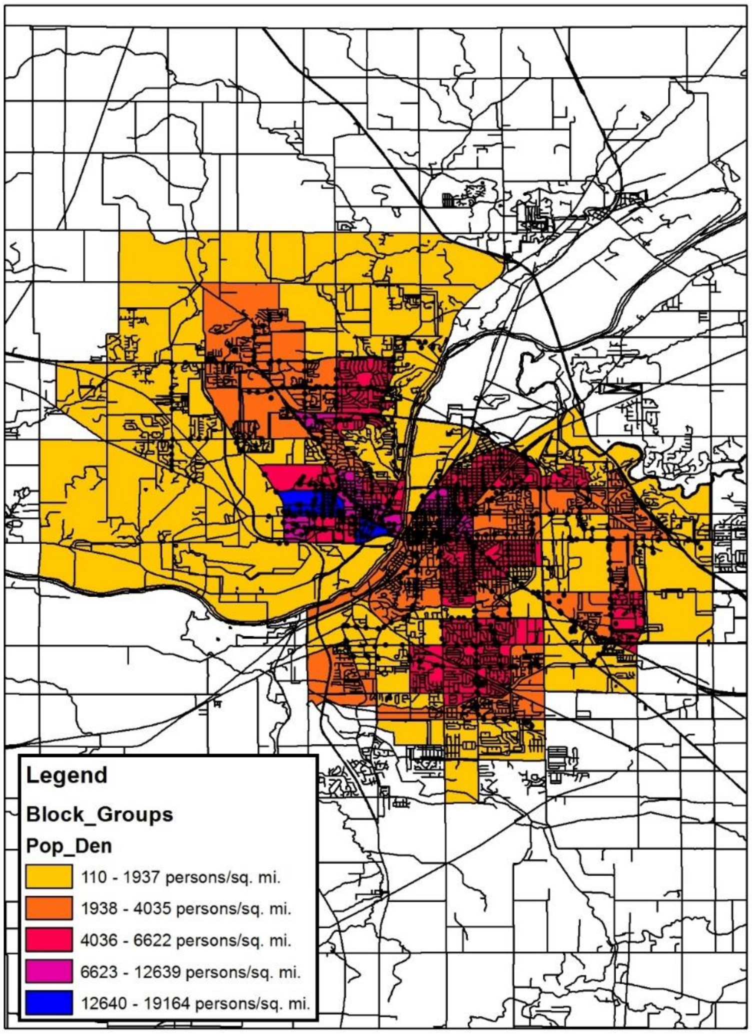
Legend
Block_Groups
Pop_Den
🟨 110 - 1937 persons/sq. mi.
🟧 1938 - 4035 persons/sq. mi.
🟥 4036 - 6622 persons/sq. mi.
🟪 6623 - 12639 persons/sq. mi.
🟦 12640 - 19164 persons/sq. mi.

Figure 3. Population density (persons/square mile) by block group.

Figure 4 presents a plot of BG annual ridership and BG population density. A weak positive trend can be observed, indicating that ridership may be positively correlated with population density. This will be investigated during the modeling process.

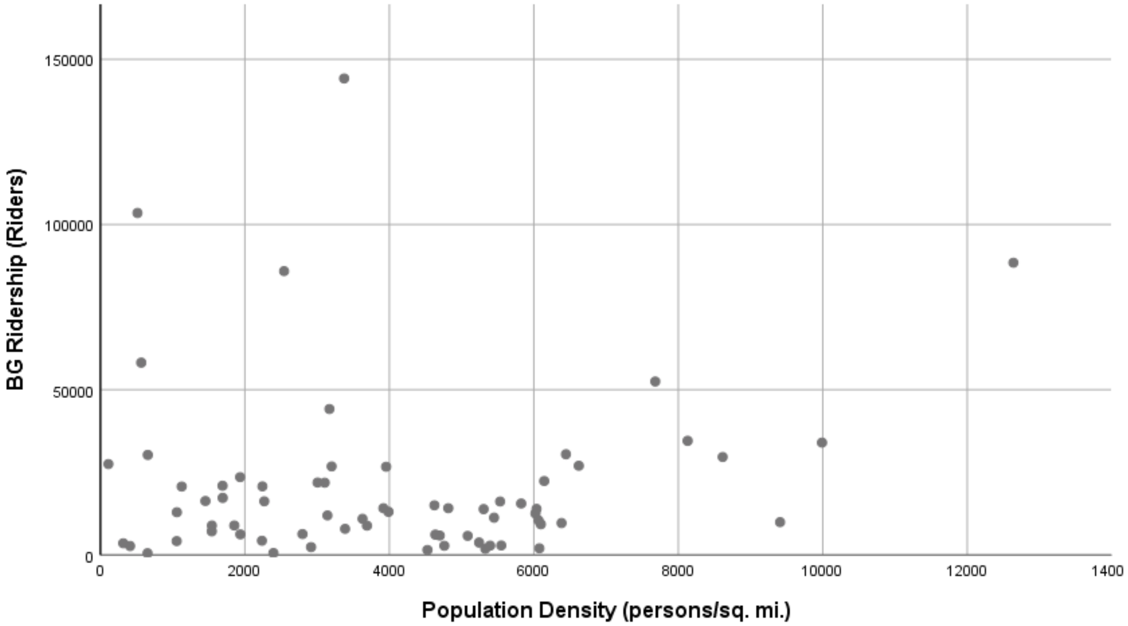

Figure 4. Scatterplot of BG ridership and BG population density.

Figure 5 presents a plot of BG annual ridership versus BG average span of service. It should be noted that BG SS values tend to be clustered in the 12-16 hour region, due to the fact that most routes operate with longer spans of service (>12 hrs). This clustering is a direct result of a policy decision made by GLPTC. It can be seen that a positive correlation exists between ridership and span of service. This supports the claim that higher quality transit service attracts more riders than lower quality transit service.

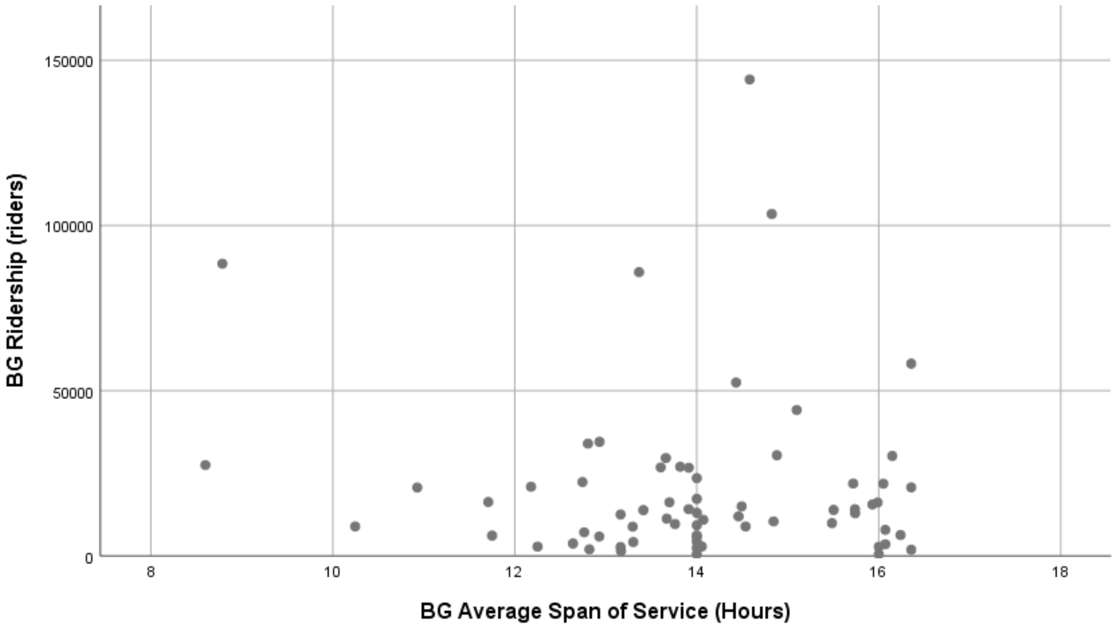

Figure 5. Scatterplot of BG ridership and BG average span of service.

3.3 Description of ACS data

Data from the 2011-2015 American Community Survey (ACS) 5-Year Estimates for Tippecanoe County, Indiana was used to build models (ACS 2015). Five-year estimates were chosen to achieve the greatest reliability compared with 1- or 3-year estimates. The GLPTC service area is entirely contained within Tippecanoe County. Data were obtained at the block-group level, because this is the smallest geographic scale for which data is available. Block groups that are not currently served by transit service were excluded from analysis (for example, rural block groups). There are over 300 datasets available from the Census Bureau. However, a sample of datasets is taken in order to make the analysis manageable. This sample was taken such that information about key demographic indicators believed to influence transit ridership are included in the sample. Table 3 lists these datasets.

Table 3. ACS datasets selected for analysis.

ID	Title
B01003	Total Population
B02001	Race
B03002	Hispanic or Latino Origin by Race
B03003	Hispanic or Latino Origin
B08301	Means of Transportation to Work
B08302	Time Leaving Home to go to Work
B08303	Travel Time to Work
B11001	Household Type
B11016	Household Type by Household Size
B14007	School Enrollment by Detailed Level of School for the Population 3 Years and Over
B15003	Educational Attainment for the Population 25 Years and Over
B16002	Household Language by Household Limited English Speaking Status
B17021	Poverty Status of Individuals in the Past 12 Months by Living Arrangement
B19001	Household Income in the Past 12 Months (In 2015 Inflation-Adjusted Dollars)
B19013	Median Household Income in the Past 12 Months (In 2015 Inflation-Adjusted Dollars)
B19025	Aggregate Household Income in the Past 12 Months (In 2015 Inflation-Adjusted Dollars)
B19055	Social Security Income in the Past 12 Months for Households
B19056	Supplemental Security Income (SSI) in the Past 12 Months for Households
B19057	Public Assistance Income in the Past 12 Months for Households
B19059	Retirement Income in the Past 12 Months for Households
B19101	Family Income in the Past 12 Months (In 2015 Inflation-Adjusted Dollars)
B19301	Per Capita Income in the Past 12 Months (In 2015 Inflation-Adjusted Dollars)
B21002	Period of Military Service for Civilian Veterans 18 Years and Over
B23025	Employment Status for the Population 16 Years and Over
B23027	Full-Time, Year-Round Work Status in the Past 12 Months by Age for the Population 16 Years and Over
B25001	Housing Units
B25002	Occupancy Status
B25003	Tenure
B25004	Vacancy Status
B25006	Race of Householder
B25008	Total Population in Occupied Housing Units by Tenure
B25010	Average Household Size of Occupied Housing Units by Tenure
B25017	Rooms
B25024	Units in Structure
B25041	Bedrooms
B25056	Contract Rent
B25063	Gross Rent
B25070	Gross Rent as a Percentage of Household Income in the Past 12 Months
B25075	Average Home Value (Dollars)
B25077	Median Home Value (Dollars)
B25081	Mortgage Status
B27010	Types of Health Insurance Coverage by Age
C15010	Field of Bachelor’s Degree for First Major for the Population 25 Years and Over
C17002	Ratio of Income to Poverty Level in the Past 12 Months
C24010	Sex by Occupation for the Civilian Employed Population 16 Years and Over

3.4 Data Limitations

Ideally, data for a planning-level study would be available at a geographic scale of adequate granularity (such as block-level) so that more spatially precise ridership estimations could be made. This is particularly an issue for lower density areas located further from the downtown area, because these tend to be represented by larger block groups in the census data. However, due to concerns about privacy, census data is only available at the aggregated block group level. This impedes estimation of ridership for larger block groups, since these block groups likely exhibit greater heterogeneity in land use, socioeconomic, and demographic trends. To accommodate the fact that block groups are of varied geographic size, many of the ACS data points are converted to densities using the area of each block group for use in the modeling process.

Also, ACS data does not adequately represent the student population well. Due to the fact that the student population is much more transient than the non-student population, it is difficult to obtain accurate data about the student population from the ACS. By the time a student completes an ACS survey and that data is included in a published dataset, it is very likely that the student has moved and is no longer a part of the study population. For these reasons, it is difficult to make accurate ridership estimations for Purdue's campus, and the surrounding student housing areas. To overcome this, it can be assumed that the outgoing student who completed the ACS survey would have similar characteristics and background with the incoming student who replaced them. A separate study focusing on estimating ridership for campus areas at a smaller geographic scale would present a good direction of future work in this domain.

3.5 Spatial Analysis

Ridership data is geocoded, meaning that a particular ridership value can be associated with the bus stop (and location) where the boarding occurred. ACS data can also be displayed spatially via the use of shapefiles and geographic information systems (GIS) software, which allow the data to be overlaid on maps of transportation networks, natural features, or other useful maps. This is necessary in order to relate ridership with the characteristics of nearby block groups. Additionally, this can also be useful in conjunction with the ridership predictions developed as a result of this research when used for detailed route planning that may occur later in the service planning process. Information about ridership and demographic characteristics for block groups is obtained from this spatial analysis, and is overlaid on the existing transportation (road) network using GIS software. This can be very useful after it is determined (through ridership estimation) that a particular service change is feasible in planning the precise route and stop locations to serve a particular area.

3.6 Proportional Ridership Allocation

Ridership and schedule data from GLPTC is represented at the bus stop level. While this is useful for analysis of current service, it is less useful for service planning tasks, because precise bus stop locations are typically not determined during the initial planning phase. Additionally, ACS data are represented at the block group level. In order to develop models, the geographic scale of the ridership and schedule data must be adjusted. In this thesis, this was done via a spatial aggregation process.

For transit, a catchment area represents the area surrounding a transit stop from which that stop is expected to pull riders. Typically, this is represented using some radius of distance in all directions from the stop. For local transit, a catchment area with radius of 1/8 mile surrounding the bus stop is suggested (APTA 2009).

Using GIS tools, these catchment areas were overlaid on the block groups. This is shown in Figure 6. It should be noted that the block groups have been color-coded by population in Figure 6 to distinguish each block group from its neighbors.

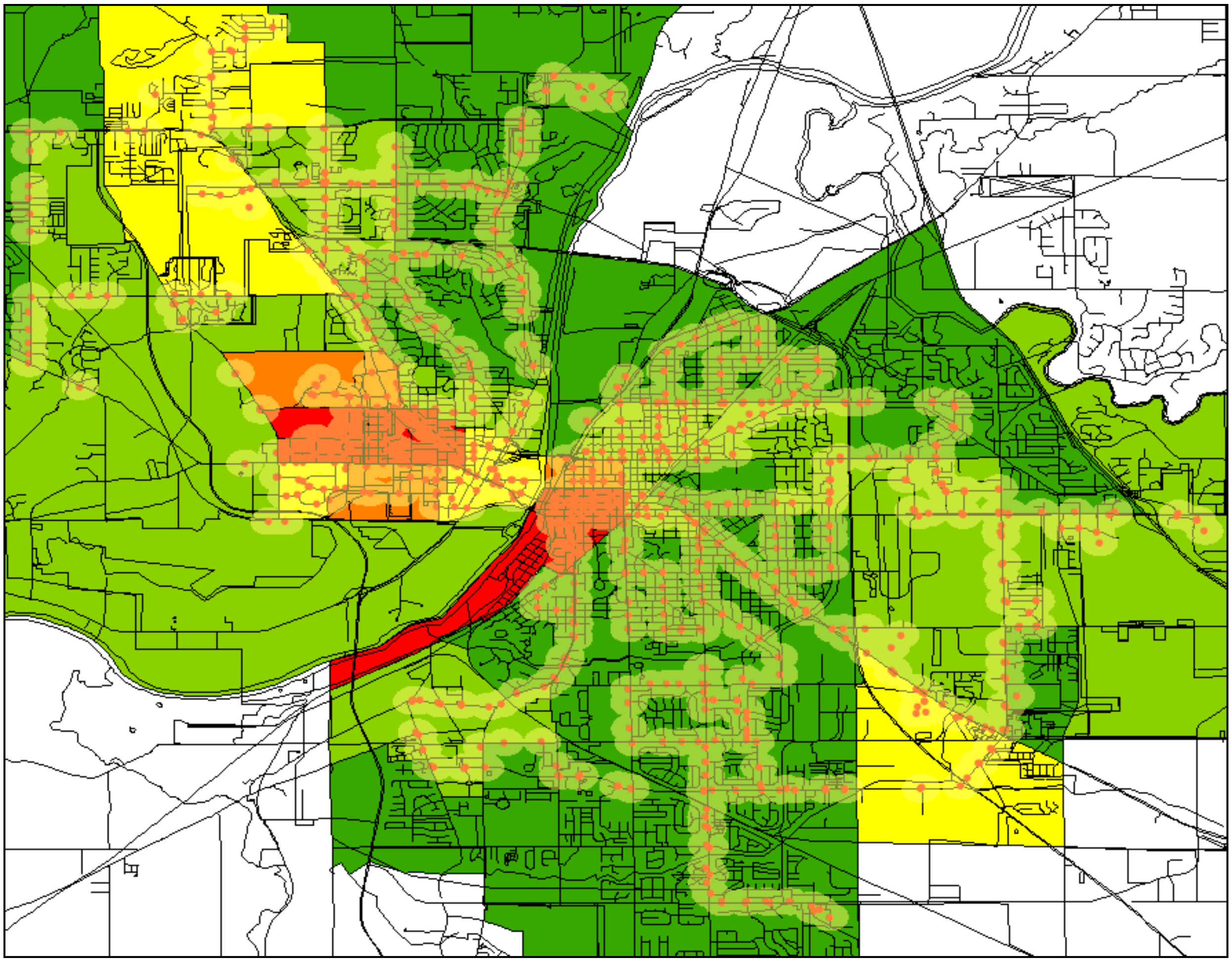

Figure 6. 1/8 mile bus stop catchment areas overlaid on block groups.

Next, the percentage of each catchment area i that falls into each block group is determined (P_i), again using GIS tools. For a catchment area that lies entirely within one block group, this will be 100%. For one that lies on a street bordering two block groups, approximately 50% of the catchment area will be allocated to each block group. Doing this accounts for the fact that riders may cross block group boundaries to board transit.

Next, these percentages are multiplied by the total annual ridership at each stop ($Riders_i$), to determine the percentage of that ridership that originated from each block group (BG Ridership). Lastly, the "fractional ridership" for n stops in a block group are added together. This results in a single ridership value for each block group, representing the total ridership from that block group, adjusted to account for bus stops that may serve more than one block group. The process for calculating total block group ridership is represented by Equation (1).

$$BG\ Ridership = \sum_{i=1}^{n} Riders_i \times P_i \tag{1}$$

A similar process involving weighted averages is used to obtain an average headway and span of service value for each block group. Instead of ridership, the percentage of each catchment area i (P_i) is multiplied by the respective average headway value for each stop, i (HW_i). In this case, the percentage of overlapping area (P_i) is used as a weighting factor. These values are added for all catchment areas overlapping a particular block group. This value is divided by the total of all the percentages for all catchment areas overlapping the block group to complete the weighted average calculation. Equation (2) presents this calculation. The result is an average headway value for each block group. The same is done for span of service (SS). The average span of service value for each stop is represented by SS_i. This calculation is shown in Equation (3).

$$Avg.BG\ HW = \frac{\sum_{i=1}^{n} HW_i \times P_i}{\sum_{i=1}^{n} P_i} \tag{2}$$

$$Avg.BG\ SS = \frac{\sum_{i=1}^{n} SS_i \times P_i}{\sum_{i=1}^{n} P_i} \tag{3}$$

For example, consider the simplified case (shown in Figure 7) of a block group consisting of only two bus stops, A, and B near the BG boundaries. Stop A had a total yearly ridership of 4,260, and 20% of the stop catchment area is located in the block group (80% is located in an adjacent block group). Similarly, stop B had a total yearly ridership of 380, and 95% of the stop catchment area is located in the block group (only 5% is located in an adjacent block group). The total BG ridership for the block group can be calculated as shown in Equation (5).

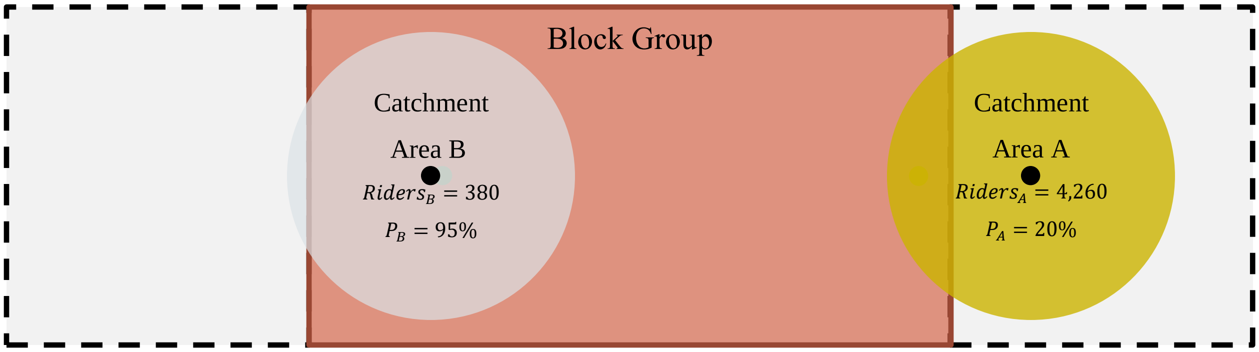

Figure 7. BG ridership calculation example.

$$BG\ Ridership = Riders_A \times P_A + Riders_B \times P_B \quad (4)$$

$$BG\ Ridership = 4{,}260 \times 0.20 + 380 \times 0.95 = 1{,}213\ BG\ riders \quad (5)$$

A similar calculation could be completed for the weighted BG average headway and SS, replacing ridership values with average headway and SS values. The percentage of overlapping area (P_i) for each stop would remain the same. In this case, the result would be divided by the sum of all P_i values in the block group, as here these values are used as weighting factors, not overlapping areas. In this case, the greater the amount of overlapping area, the greater the influence that particular stop has on the overall BG average headway and SS.

The result of these calculations is one ridership, average headway, and average span of service value for each block group. These values are now at the same spatial scale as the ACS data and can be used in the modeling process.

3.7 Summary

Data from GLPTC and the Census Bureau are leveraged to conduct this analysis. Before the modeling process can begin, it is important to understand the data so that logical conclusions can be made later, once the modeling process is complete. It is also important to understand the limitations of the input data, because these likely impact the predictive power of the models that are developed. Lastly, it is important that both the GLPTC data and the ACS data are represented at the same geographic scale before conducting any analysis. In this case, GLPTC data had to be aggregated through a proportional process in order to match the geographic scale of the ACS data, due to the fact that ACS data is not widely available at smaller geographic scales. These steps help to ensure that the modeling process is more successful, and that predictions made as a result of the models are meaningful.

4. STUDY METHODOLOGY

4.1 Introduction

It is desired to develop models that relate ridership at the block group level to a variety of independent variables ranging from service characteristics to demographics to population. Additionally, it is critical that the final models are simple and clear, so that they can be readily applied by someone who may not have knowledge of complex modeling techniques. Still, the models must be accurate, so as to be useful for the planning tasks previously described. For these reasons, regression analysis is used to build models.

Regression seeks to describe the relationship between variables (Kutner 2005). A model can be developed to predict the value of one variable (the dependent variable) using information known about another variable (the independent variable). For example, it might be possible to predict the travel time on a particular road from information about the traffic volume carried on that road. In simple linear regression (SLR), the functional form is Equation (6).

$$Y = \beta_0 + \beta_1 X \tag{6}$$

In this example, Y represents the dependent variable that is being predicted (travel time, in the example above), while X represents the independent variable that is used for the prediction (traffic volume in the example). β_0 and β_1 represent numerical coefficients that are determined during the regression modeling process.

In many cases, a better model (one that offers a better prediction for Y) can be developed through the inclusion of multiple independent variables. In multiple linear regression (MLR), it is possible to include several independent variables. The functional form generally appears as shown in Equation (7).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \tag{7}$$

Here, multiple independent variables $(X_1, X_2, \ldots, X_n)$ are included, each with a regression coefficient. This allows for potentially increased accuracy in comparison to simple linear regression.

In some cases, the relationship between independent and dependent variables is not linear. In these cases, nonlinear regression may provide better models. Unlike SLR and MLR models, nonlinear regression models can take many different forms. In this case, analysis of the type of relationship each independent variable has with the dependent variable can be useful in suggesting the most appropriate model type. This relationship could be linear, exponential, logarithmic, inverse, or other forms.

In this thesis, several SLR, MLR, and nonlinear regression models will be developed. Once these models are developed, it is important to assess how well they predict the desired result, in order to select the most desirable model. There are many ways of doing this, one of which is to evaluate the coefficient of determination, or R^2 value for each model. R^2 is a ratio of the variation in variables explained by the model to the total variation in variables. Values range from zero to one, with a value of one indicating that the model perfectly explains all of the witnessed variation.

R^2 will always increase as additional variables are added to the model, regardless of whether or not they actually increase the model's predictive power. For these reasons, it is often better to evaluate models using the adjusted R^2, which adjusts the original R^2 according to how many variables are included in the model.

4.2 Regression Analysis

Regression techniques are used to develop ridership models using the described data. Block group ridership (Sum_{Prop_R}) is used as the dependent variable, while the ACS data and service characteristics (average headway and average span of service) are used as independent variables. The regression models were developed using SPSS statistical software (IBM 2013).

4.3 Model Building

The stepwise regression technique was selected for model building as a way to accommodate the large number of potential independent variables. In the forward stepwise technique, F-value is used as the decision value. If the addition of an additional predictor improves the statistical significance of the overall model, the predictor is retained. If not, the predictor is omitted. This process is repeated until none of the remaining predictors improve the significance of the model. The resulting models from this procedure are shown in Section 5.3.

4.4 Model Selection

Several methods can be used to select the best of several potential models. A common selection method is using the R^2 criterion, or more properly, the adjusted R^2 for the reasons outlined previously. A model with a larger adjusted R^2 value is generally more desirable than one with a smaller value, because this indicates that the model has better predictive power. However, it is also important to consider the overall model complexity in choosing a model. Adding terms to a model means that additional data is required for the model to be applied, so it should only be done in cases where additional terms provide a significant benefit in the predictive power of the model. In some cases, it may be desirable to select a model with a slightly lower value of R^2 or adjusted R^2 if it requires less input data to use it.

4.5 Model Validation

After the models are developed, it is important to perform steps to ensure that they are accurately predicting ridership. To do this, the root mean square error (RMSE) is used. RMSE is a way to evaluate the difference between predicted and observed values, also known as residuals. In conjunction with RMSE, the F-value is used to assess the level of significance of a particular model. R^2 is also presented, as described previously as a commonly used measure for evaluating the predictive power of a particular model.

Another, more visual method for evaluating the models is to review plots of the observed values plotted against the predicted values. In an ideal scenario, these would fall along the diagonal line

$y = x$, meaning that each observed data point is perfectly predicted by the model. The better the model, the closer the plotted points will fall to this ideal case. These plots are shown in Section 5.5.

5. RESULTS AND DISCUSSION

5.1 Introduction

Chapter 4 provided an overview of the regression process used to develop the ridership models. This chapter presents the models and discusses their accuracy. Several models will be presented, each with different levels of predictive power, and different numbers of variables. This gives the agency several options to choose from when predicting ridership. In some cases, it may be acceptable to sacrifice some accuracy in prediction for a model that is simpler and requires less input data, such as in cases where an estimate needs to be obtained quickly without time to gather lots of input data, or in cases where limited input data is available. However, in other cases, accuracy may be more important, and in those cases it may make more sense to choose a more complex model, which requires more input data, but provides a more accurate prediction, indicated by a larger R^2 or smaller RMSE. For these reasons, several models are presented in this chapter.

5.2 Summary of Variables

A summary of all candidate variable names and a brief explanation of each variable is provided in Table 4.

Table 4. Summary of variable names and descriptions.

Variable Name	Description	Unit
Sum_{Prop_R}	Ridership total for BG (dependent variable)	Persons
BG_Av_Sp	Average span of service for BG	Hours[1]
BG_Av_Hw	Average headway for BG	Hours
$MedHHInc$	Median household income	Dollars ($)
$PAInc_{den}$	Density of persons receiving public assistance income (food stamps, etc.)	Persons/sq. mi.
$RecRetInc_{den}$	Density of persons receiving retirement income (social security, etc.)	Persons/sq. mi.
$PerCapInc$	Per capita income	Dollars ($)
$MedHValue$	Median home value	Dollars ($)
Pop_{den}	Population density	Persons/sq. mi.
Hsg_{den}	Total housing density	Housing units/sq. mi.
$White_{den}$	Density of persons who indicated their race as White	Persons/sq. mi.
$Black_{den}$	Density of persons who indicated their race as Black	Persons/sq. mi.
$Asian_{den}$	Density of persons who indicated their race as Asian	Persons/sq. mi.
$Mult_{den}$	Density of persons who indicated their race as multiracial	Persons/sq. mi.
$Hlat_{den}$	Density of persons who indicated they are of Hispanic/Latino origin	Persons/sq. mi.
$AutoO_{den}$	Density of persons who reported owning at least one automobile	Persons/sq. mi.
HH_{den}	Density of family households[2]	Households/sq. mi.
$Nonfam_{den}$	Density of nonfamily households[3]	Households/sq. mi.
$Enroll_{den}$	Density of persons enrolled in some form of education (K-college)	Persons/sq. mi.
$Engl_{den}$	Density of persons who reported English as their primary language	Persons/sq. mi.
$Span_{den}$	Density of persons who reported Spanish as their primary language	Persons/sq. mi.
Pov_{den}	Density of persons with incomes below the federal poverty line[4]	Persons/sq. mi.
$Empl_{den}$	Density of employed persons	Persons/sq. mi.

[1] For analysis, the Hours:Minutes notation is converted into Hours (using a decimal). For example, 0:15 represents 15 minutes, or 0.25 Hours.

[2] The Census Bureau defines a family as "…a group of two people or more (one of whom is the householder) related by birth, marriage, or adoption and residing together." (Census Bureau 2018).

[3] The Census Bureau defines a nonfamily household as "…a householder living alone (a one-person household) or where the householder shares the home exclusively with people to whom he/she is not related." (Census Bureau 2018).

[4] The Census Bureau defines several poverty thresholds based on family size and annual income (Census Bureau 2018). Those with total incomes (family or individual) below these thresholds are considered to be in poverty, while those with incomes above are not.

Table 4 continued.

$Unempl_{den}$	Density of unemployed persons	Persons/sq. mi.
Own_{den}	Density of homeowners	Persons/sq. mi.
$Rent_{den}$	Density of renters	Persons/sq.mi.
Vet_{den}	Density of veterans	Persons/sq. mi.
Vac_{den}	Density of vacant housing units	Housing units/sq. mi.

5.3 Summary of Models

Several models were developed to predict ridership from a variety of input variables. The first 3 models are summarized in Table 5.

Table 5. Model parameters and significance.

Model	Coefficient		Std. Error	VIF	$Adj. R^2$	F-Value
1	(Constant)	12658.801	3421.079		0.181	153.19***
	Pov_den	8.133	2.059	1.000		
2	(Constant)	26132.393	4942.428		0.305	109.214***
	Pov_den	7.431	1.908	1.011		
	Own_den	-17.300	4.887	1.011		
3	(Constant)	20274.308	26242.672		0.501	68.73***
	PerCapInc	-1.982	0.411	1.024		
	Enroll_den	21.227	4.444	1.024		
	White_den	-14.789	3.250	1.024		
	MedHValue	-0.107	0.042	1.024		
	BG_Av_Sp	4,435.430	1741.708	1.024		
	Vet_den	59.302	23.882	1.024		

***Indicates that the model was found to be statistically significant at the 0.05 significance level.

Of the three models, Model 1 includes the fewest terms, while Model 3 contains the most. Model 3 has a larger adjusted R-square, meaning it has greater predictive power than Model 1. Model 2 had the smallest standard error of the 3 models. Multiple linear regression can exhibit issues with multi-collinearity, so it is important to check models for these issues in MLR models. A common way of doing this is with variance inflation factors (VIF). VIFs are a measure of how much multi-collinearity exists in a particular model (NIST 2003). VIFs have values that range from one to ten, with higher values indicating that significant multicollinearity issues exist. Here, all models have VIFs below 5, which indicates that multi-collinearity is likely not a major issue with the models.

In an effort to obtain a better fit to the data, several alternate forms of the model were tried. Model 4 is a semi-log transformation of Model 3, and is represented by Equation (8). The coefficients for this model are summarized in Table 6.

$$\begin{aligned} Sum_{Prop_R} = A + (B \times \ln(PerCapInc)) + (C \times \ln(Enroll_{den})) + (D \times \ln(White_{den})) \\ + (E \times \ln(MedHValue)) + (F \times \ln(BG\_AV\_Sp)) + (G \times \ln(Vet_{den})) \end{aligned} \tag{8}$$

Table 6. Additional model parameters and significance.

Model	Coefficient		Std. Error	$\boldsymbol{Adj.R^2}$	F-Value
4	A (Constant)	321809.395	99713.817	0.449	17.21***
	B	-25639.591	7162.264		
	C	14078.154	5947.852		
	D	-21794.726	7840.020		
	E	-2819.557	914.237		
	F	24965.210	21611.615		
	G	-1091.584	2369.901		

***Indicates that the model was found to be statistically significant at the 0.05 significance level.

5.4 Model Interpretation

Model 1 (represented by Equation (9) below) is a SLR model incorporating a positive constant term and a positive coefficient on the variable Pov_{den}, which represents the density (persons/square mile) of persons with incomes below the poverty line in a particular block group.

$$Sum_{Prop_R} = 12{,}658.801 + (8.133 \times Pov_{den}) \tag{9}$$

Model 2 (represented by Equation (10)) is a MLR model that incorporates a positive constant term, and a positive coefficient on the variable Pov_{den}. However, Model 2 includes an additional term, Own_{den} which represents the density of homeowners in a particular block group. This term has a negative coefficient.

$$Sum_{Prop_R} = 26{,}132.393 + (7.431 \times Pov_{den}) - (17.300 \times Own_{den}) \tag{10}$$

The constant in these models is positive, which is intuitive because a negative intercept value would indicate that transit ridership could be less than zero under certain conditions. In both models, the coefficient of Pov_{den} is positive, indicating a positive relationship between ridership and the number of low-income persons in a particular block group. When the density of low-income persons in a block group increases, transit ridership is also expected to increase. This is intuitive, because those experiencing poverty often have fewer choices for transportation than those not experiencing poverty.

The coefficient for Own_{den} is negative, meaning that, as the number of homeowners in a block group increase, transit ridership falls. This is also logical, because individuals who own homes are generally more likely to own cars, and therefore less likely to choose public transit for their transportation needs. Since most homeowners are also supporting a family, this trend is also supported by research that validates the idea that transit becomes less appealing to families because they need the flexibility to link trips to care for dependents and accomplish work and personal tasks (TCRP 1998), which can be more difficult using transit.

Model 3 (represented by Equation (11)) is also a MLR model. It incorporates a positive constant term, positive coefficients on enrollment density ($Enroll_{den}$), block group average span of service (BG_{AvSp}), and veteran density (Vet_{den}). It also includes negative coefficients on per capita income ($PerCapInc$), the density of persons who indicated their race as white ($White_{den}$), and median home value ($MedHValue$).

$$\begin{aligned} Sum_{Prop_R} = 20{,}274.308 &- (1.982 \times PerCapInc) + (21.227 \times Enroll_{den}) \\ &- (14.789 \times White_{den}) - (0.107 \times MedHValue) \\ &+ (4{,}435.430 \times BG_{AvSp}) + (59.302 \times Vet_{den}) \end{aligned} \tag{11}$$

As in Models 1 and 2, the coefficients of Model 3 are intuitive. As per capita income increases, transit ridership is expected to decrease. This is because those with higher income are more likely to have additional transportation options (such as the ability to purchase a car), and are therefore less likely to choose public transit for their transportation needs. While the density of enrolled students (defined as any enrollment K-college) increases, transit ridership does as well, because the student population is less likely to have access to cars than the working adult population. When the density of persons who indicated their race as white increases, transit ridership decreases. This is supported by work by McLafferty (1997), which showed that transit ridership tends to be higher among (non-white) racial minority groups. The model also found that, as median home value increases, transit ridership is expected to decrease. This is logical, because home value can be a proxy for income, and for auto ownership, both of which are known to have negative relationships with transit ridership. A positive relationship with span of service is predicted, and this is intuitive, because ridership tends to be higher in areas that are served by better quality transit, such as transit with longer spans of daily service. Finally, the model also predicts a positive relationship between the density of veterans in an area and transit ridership. This is less intuitive, but it is believed that this could be a proxy effect. According to the National Survey of Veterans (2003), over 37 percent of the veteran population is over the age of 65. It is known that older adults are more likely to take transit (Mallett 2018), and it is believed that the density of veterans in an area serves as a proxy for the number of older adults in an area, which, intuitively has a positive relationship with transit ridership.

Model 4 is more complex when compared to Models 1-3, but exhibits a greater level of significance when the F-value is compared with Models 1-3. A decision must be made at the agency level whether the additional time and effort required to use this model is worth the additional significance it provides.

5.5 Model Validation and Selection

Each of the developed models must be validated to assess how well it predicts ridership in comparison with the other models. The root mean square error (RMSE) is used to evaluate each model, along with the F-Value and R^2. The results are shown in Table 7.

Table 7. Model validation results.

Model	RMSE	F-Value	R^2
1	22939.63	153.19	0.181
2	21140.96	109.214	0.305
3	1670.38	68.73	0.501
4	19741.54	17.21	0.449

Model 1, while the simplest of all the models, has the lowest R^2 and also the largest RMSE. For these reasons, this model is not recommended for use in ridership prediction. Model 2 improves on Model 1, with increased R^2, and a smaller value of RMSE. Model 2 is also simpler than Models 3-4, so it could be used in cases where accuracy is less of a concern. Model 3 provides the largest R^2, and also the smallest RMSE, so it is likely the best of all of the models for predicting ridership. Model 4 is a variation on Model 3, and achieves a slightly smaller value of R^2 and larger RMSE when compared to Model 3.

To evaluate the models visually, plots of the observed ridership versus predicted ridership are created. The $y = x$ relationship is also plotted for comparison. Figure 8 represents Model 3, while Figure 9 represents Model 4. Models 1 and 2 are not shown, because they had the poorest performance of all of the models.

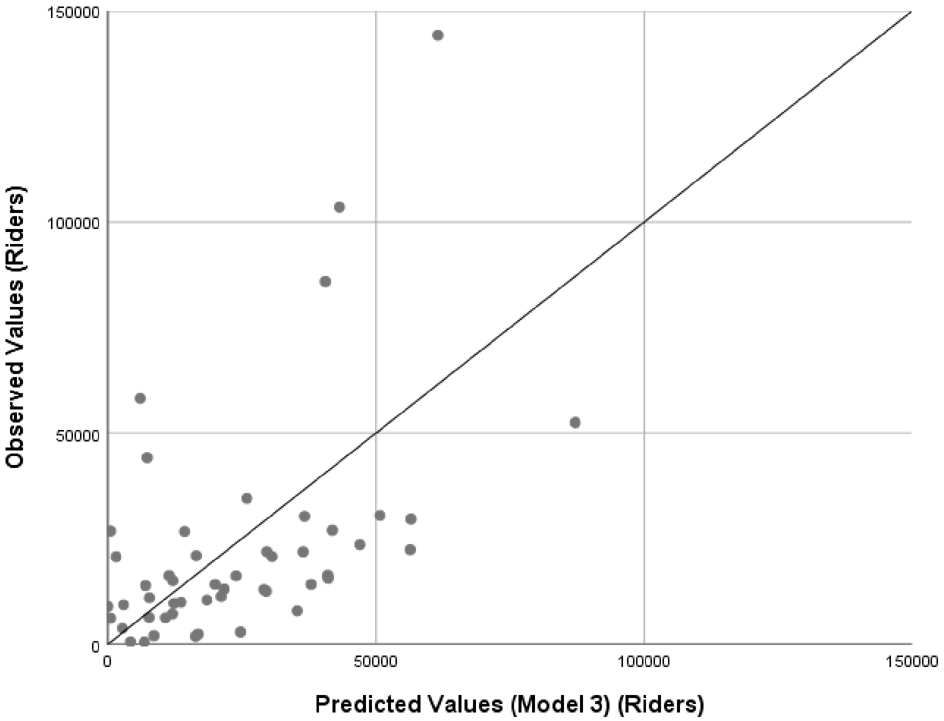

Figure 8. Observed versus predicted values for Model 3.

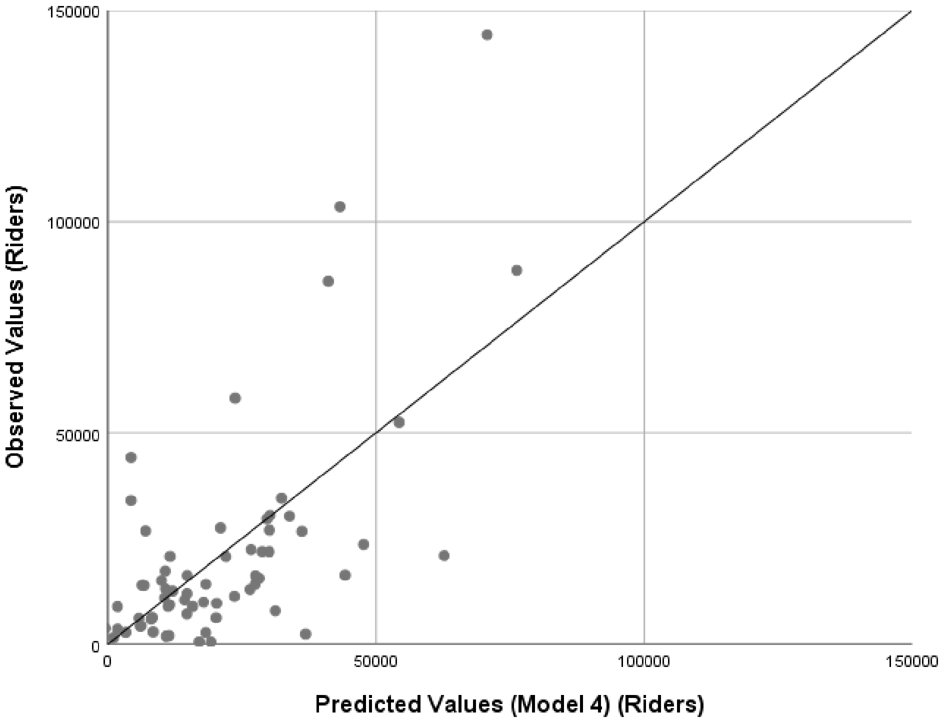

Figure 9. Observed versus predicted values for Model 4.

Model 4 appears to have the best performance when the observed and predicted values are plotted, because the plotted points lie closer to the $y = x$ line, indicating a better match of the observed and predicted points when compared with the other models. It should be noted that none of the models performs particularly well for areas with very high ridership. This is most likely due to the relatively small sample size of block groups with this level of ridership. It is believed that a larger sample size (which would include more block groups with a wider distribution of ridership) would improve the predictive power of the models in this region.

Model 3 has the largest value of R^2 and also the smallest value of RMSE. As discussed above, Model 4 appears to perform best when comparing observed and predicted values. For these reasons, these two models are suggested for use in ridership prediction. However, Model 2 may also be appropriate if limited data is available and lower predictive accuracy is acceptable in certain situations.

An upper bound on ridership when using these models could be set to address the issues discussed for high ridership areas. This bound would likely be set at approximately 50,000 riders per BG. However, this bound places additional constraints on an already limited dataset, reducing the usefulness of the data even further. For these reasons, a bound is not imposed, but caution is suggested in using the models in areas with very high ridership. As previously discussed, using the models in conjunction with prior knowledge will provide the best results as far as service planning is concerned.

5.6 Numerical Examples

The following subsections provide numerical examples for estimating ridership using the models presented in this thesis.

5.6.1 Example 1

Transit service is proposed for an area that has seen significant growth in recent years. Table 8 lists basic information about the area obtained from the Census Bureau. This is representative of typical data used for analysis. An estimate for bus ridership in this area is needed to determine whether or not providing service is justified. For planning purposes, it can be assumed that service will be provided every day of the year for 15 hours each day.

Table 8. Input data for ridership estimation.

Quantity	**Value**	**Unit**
Per capita income	$39,000	Dollars ($)
Total K-college enrollment	570	Persons
Total number of persons who identified their race as White	1157	Persons
Median home value	$112,500	Dollars ($)
Total number of veterans	350	Persons
BG land area	0.78	Square Miles

SOLUTION:

1. All information needed to use Model 3 is provided. Since it provides the best prediction according to R^2, it will be used to calculate ridership.

$$\begin{aligned} Sum_{Prop_R} = {} & 20{,}274.308 - (1.982 \times PerCapInc) + (21.227 \times Enroll_{den}) \\ & - (14.789 \times White_{den}) - (0.107 \times MedHValue) \\ & + (4{,}435.430 \times BG_{AvSp}) + (59.302 \times Vet_{den}) \end{aligned}$$

2. This model requires density values for some inputs. These are calculated next.

$$Enroll_{den} = \frac{570\ persons}{0.78\ sq.mi.} = 730.77\ persons/sq.mi.$$

$$White_{den} = \frac{1157\ persons}{0.78\ sq.mi.} = 1483.33\ persons/sq.mi.$$

$$Vet_{den} = \frac{350\ persons}{0.78\ sq.mi.} = 448.72\ persons/sq.mi.$$

3. The average span of service for this area is needed. For simplicity, it is assumed that this will be the only transit service serving the area, and it is given that it will operate every day of the year for 15 hours each day. Thus, the average span of service is 15 hrs.
4. This information is entered into the model to calculate ridership.

$$Sum_{Prop_R} = 20{,}274.308 - (1.982 \times 39000) + (21.227 \times 730.77) - (14.789 \times 1483.33) - (0.107 \times 112500) + (4{,}435.430 \times 15) + (59.302 \times 448.72)$$

$$Sum_{Prop_R} = 20{,}274.308 - 77298.00 + 15512.05 - 21936.97 - 12037.50 + 66531.45 + 26609.99$$

$$Sum_{Prop_R} = 17{,}655.33$$

The ridership in this area is expected to be approximately 17,655 per year.

5.6.2 Example 2

Transit service is currently provided to an area, but a new student apartment complex is expected to open. The agency seeks to find out how the opening of this apartment complex will affect ridership. Calculate the change in ridership that can be associated with this apartment complex, assuming that nothing else has changed (including service characteristics). Information in Table 9 is provided from a Census Survey with supplemental information collected by the transit agency. It is known that BG ridership before the complex opens is 16,141.

Table 9. Input data for ridership estimation.

Parameter	Value	Unit
$PerCapInc$	$ 35,700	Dollars ($)
$Enroll_{den}$	1489.76	Persons/sq. mi.
$White_{den}$	1131.37	Persons/sq. mi.
$MedHValue$	$91,360	Dollars ($)
BG_AV_Sp	18	Hours
Vet_{den}	33.45	Persons/sq. mi.

SOLUTION:

1. Ridership after the apartment complex is opened can be calculated using Model 4.
   a. $\left(Sum_{Prop_R}\right)_{AFTER} = 321809.395 + (-25639.591 \times \ln(PerCapInc)) + (14078.154 \times \ln(Enroll_{den})) + (-21794.726 \times \ln(White_{den})) + (-2819.557 \times \ln(MedHValue)) + (24965.210 \times \ln(BG\_AV\_Sp)) + (-1091.584 \times \ln(Vet_{den}))$

$$\begin{aligned}(Sum_{Prop_R})_{AFTER} &= 321809.395 + (-25639.591 \times ln(35700)) \\ &+ (14078.154 \times ln(1489.76)) + (-21794.726 \times ln(1131.37)) \\ &+ (-2819.557 \times ln(91360)) + (24965.210 \times ln(18)) \\ &+ (-1091.584 \times ln(33.45))\end{aligned}$$

$$(Sum_{Prop_R})_{AFTER} = 38{,}770.093$$

2. The change in ridership due to the apartment complex opening is calculated.

$$\Delta_{Ridership} = (Sum_{Prop_R})_{AFTER} - (Sum_{Prop_R})_{BEFORE} = 38{,}770 - 16{,}141 = 22{,}629$$

Ridership in the area is expected to increase by approximately 22,629 due to the opening of the student apartments. This represents an increase of approximately 140%.

A reasonableness check should be conducted to verify this result. For the Lafayette/West Lafayette area, a visual survey of existing apartment complexes can be conducted and an approximate number of buildings can be obtained. Using information from the local zoning code (Tippecanoe APC 1998), the approximate number of units per building can be obtained. Using an average household size of 3.2 from the Census Bureau (ACS 2015), and assuming one household per unit gives the total number of residents in an apartment complex of average size for the Lafayette/West Lafayette area. Assuming these residents will each generate 4 trips per day (NHTS 2017), and assuming that nearby transit to the apartment complex causes a diversion of approximately 10% of trips to transit (ITE 2004) gives all the needed information to determine how reasonable this result is.

$$Approx. average\ number\ of\ buildings = 20\ buildings/complex$$

$$Approx. average\ number\ of\ units\ per\ building = 18\ units/building$$

$$Average\ number\ of\ units\ per\ complex = 20\frac{buildings}{complex} \times 15\frac{units}{building} = 300\frac{units}{complex}$$

$$Average\ number\ of\ residents\ per\ household = 3.2$$

$$Total\ number\ of\ residents\ per\ complex = 300\frac{units}{complex} \times 3.2\frac{residents}{unit} = 960\frac{residents}{complex}$$

$$Total\ umber\ of\ daily\ trips\ per\ individual = 4\ trips$$

$$Total\ number\ of\ daily\ trips\ for\ all\ residents = 4\ trips \times 960\ residents$$

$$= 3{,}840\ daily\ trips$$

$$Annual\ trips = 3{,}840\,\frac{trips}{day} \times \frac{365\ days}{1\ year} = 1{,}401{,}600\ annual\ trips$$

$$Transit\ diversion\ rate = 5\%\ of\ total\ trips$$

$$Transit\ trips = 1{,}681{,}920 \times 0.05 = 70{,}080\ transit\ trips\ per\ year$$

This verifies that the result is not unreasonable using approximate values. External factors, such as campus parking policies for university students living in an apartment complex may affect this result.

5.7 Discussion

The developed models can be useful to agencies in predicting transit ridership. For example, ridership can be predicted for areas that do not currently have transit service. Ridership changes as a result of various external factors can also be predicted using the models, such as the opening or closing of an apartment complex, or changes in population characteristics, such as income.

5.7.1 Limitations of Models

While the suggested model provides an estimate for transit ridership using a variety of factors, it is important to consider some limitations. First, the transferability of the model is limited because it was developed using only data for the Lafayette/West Lafayette area. While it is expected to reasonably predict ridership for areas with similar characteristics to those in Lafayette/West Lafayette, further investigation is needed to determine the appropriateness of applying it to areas that are dissimilar to the study area. Next, the model is limited in that it only provides ridership estimates at the block group level. Block groups do vary in geographic scale, and are typically not comprised of one homogeneous land use. It is important to consider how the estimates obtained using this model will vary under heterogeneous land uses, as well as under differently sized block groups to those studied. Lastly, the ACS data used to develop the models is rich and provides a wide variety of information, but there likely exist factors that influence transit ridership that are not captured through the ACS data, and thus are not represented in the models. An example of such a factor is personal preference. Some individuals may choose to take transit for reasons such as level of comfort or convenience, and these factors are difficult to capture in ACS data. These models exist as a tool for estimating transit ridership, but should be used with thorough knowledge of the study areas in question in order to make informed estimates regarding transit ridership. For example, these models may suggest high transit ridership in areas with large proportions of off-campus student housing. However, they would not be able to predict the effects that the university class schedule (summer/holiday breaks) would have on ridership in these areas. Providing service may be very effective in September when classes are in session, but it may be less effective in July, when classes are not in session. This is where local knowledge, in conjunction with the model results, will provide the most useful recommendations.

5.7.2 Impact of Findings

It is expected that the results of this study will be useful to transit agencies as a guide for conducting service planning based on ridership. Service planning tasks could include: providing new service to a previously unserved area, providing additional service to a currently served area (e.g. more frequent service, or service for more hours each day), modifying service to better suit the needs of an area (e.g. using different streets to better serve a high ridership area), or reducing service (e.g. to an area that has experienced a decline in population). While the models have limited transferability due to only incorporating data from Lafayette/West Lafayette, the methodology is also expected to be useful to agencies with different area and service characteristics from those included in the study that seek to develop unique ridership models to suit individual needs. Finally, this work demonstrates the ability to use basic demographic and population data in estimating ridership, and provides a simpler approach to ridership estimation than those currently available.

5.8 Model Summary

The models that were developed provide a predictive tool for estimating transit ridership at the block group level using simple data that reflect the characteristics of the block group. They allow ridership predictions to be made for an area using only basic information about the area. This can be a useful tool, not only for estimating ridership, but also for supporting previous agency predictions and estimations for ridership, such as those made using prior experience.

The models are developed using regression, which allows for simple interpretation of the coefficients included in the models, and also for easy adjustments should an agency find that portions of the model are less suited to their specific area characteristics. Further, the models are validated using common techniques, such as R^2 and RMSE to provide additional evidence regarding the claimed predictive power.

6. CONCLUSION

6.1 Summary and Concluding Remarks

This research combined transit ridership, route, and schedule data for the Lafayette/West Lafayette, Indiana area with ACS data for the same area. Ridership and service data were aggregated to the block group level to match ACS data. Regression techniques were used to develop models to predict transit demand using independent variables that represent the service characteristics, population, and socioeconomic trends of the areas under question. Models were developed and validated to predict transit ridership at the block group level. It is anticipated that this work will be beneficial to GLPTC as well as other agencies in small cities with similar service area characteristics as an aid for transit service planning work.

A major challenge of the study was that the data available through the Census bureau is only provided at the BG level. In many cases, this was found to be too large of a geographic unit for meaningful analysis, as significant amounts of heterogeneity in demographics, land use, and other factors exist within large BGs. Despite the limitations imposed by the data, it is believed that the results of this study will be useful as a preliminary, sketch planning-level ridership estimate, particularly in cases where no other information is available to aid in making service planning decisions.

At the beginning of this study, it was anticipated that this ridership modeling process would be useful for predicting ridership in areas where transit service is not currently provided. One such area is the Wabash Avenue neighborhood. Unfortunately, this neighborhood is combined with much of downtown Lafayette into one Census block group. While geographically this makes sense because the two areas are adjacent, it makes ridership estimation for Wabash Avenue very difficult using existing data, particularly because the demographic and socioeconomic trends of this neighborhood are believed to be significantly different from those of downtown Lafayette.

While ACS data is not useful for estimating ridership in the Wabash Avenue neighborhood, if the required input information to the models could be obtained from another source (e.g. local government) and isolated to include only Wabash Avenue, it would be possible to use the models to predict ridership in this area. This data limitation makes estimation of ridership more difficult in cases where the Census-designated block groups do not align with the areas where ridership estimates are desired.

6.2 Opportunities for Future Work

This study developed planning-phase transit ridership models at the block group level for the Lafayette/West Lafayette, Indiana area. Future work could involve using additional data from other small cities to develop models that would be more applicable to areas with characteristics different than the area under study currently. For example, a community without a major university likely has different demands for transit service than one that does have a major university. Work to develop a similar set of models for medium-sized cities would be beneficial. Work to develop models using data of a different geographic scale than the BG (such as a smaller or larger geographic unit) would also be useful.

Another area of potential future work would be to include additional spatial parameters in the modeling process. For example, including an impedance or distance parameter for each block group to the nearest school, shopping center, hospital, or downtown area could provide additional predictive power to the models. It is possible that using this parameter in conjunction with additional employment or retail floor area parameters may improve the models, particularly in areas that are predominantly non-residential land uses.

Work to develop a similar transit demand model for areas in and around university campuses would be useful for planning transit to serve the student population, many of whom are transit dependent. The BG-level data used in this study are not sufficient to adequately accomplish this, however with a different data source, such as block-level data, data regarding student residences from the university, or information from a university-level travel survey, it is believed that this would be feasible. The scale of ACS data (BG-level) is too large for a meaningful university-level analysis, because large portions of the university campus are included in one BG, meaning that information about smaller-scale travel trends (such as trips occurring within a BG) is lost in the aggregation process. This is important at the university-level, because many trips are of shorter distance (within the university campus), and transit services are often optimized to serve these type of trips on campus (through high-frequency loop or shuttle routes).

REFERENCES

American Community Survey (ACS). (2015). 2015 American Community Survey 2015 5-Year Estimates for Tippecanoe County, Indiana. United States Census Bureau. Washington, DC.

American Community Survey (ACS). (2015). American Community Survey and Puerto Rico Community Survey 2015 Subject Definitions. United States Census Bureau. Washington, DC.

American Public Transportation Association. (2009). Defining Transit Areas of Influence. APTA Technical Report 2009.

Amoroso, S., Catalano, M., Galatioto, F., and Migliore, M. (2010). A demand-based methodology for planning the bus network of a small or medium town. Journal of European Transport 2010: 41-56.

Area Plan Commission of Tippecanoe County, Indiana (1998). Unified Zoning Ordinance, Third Edition. Area Plan Commission of Tippecanoe County, Indiana. Lafayette, Indiana.

Arman, M., Labi, S., and Sinha, K. C. (2013). Perspectives of the Operational Performance of Public Transportation Agencies with Data Envelopment Analysis Technique. Transportation Research Record, 2351(1), 30–37.

Cervero, R., J. Murakami, and M. Miller. (2010). Direct ridership model of bus rapid transit in Los Angeles County, California. Transportation Research Record 2145: 1-7.

Chow, L. F., F. Zhao, H. Chi, and Z. Chen. (2010). Subregional transit ridership models based on Geographically Weighted Regression. Transportation Research Board 89th Annual Meeting Compendium of Papers DVD, Washington, D.C.

Chow, L. F., F. Zhao, X. Liu, M. T. Li, and I. Ubaka. (2006). Transit ridership model based on Geographically Weighted Regression. Transportation Research Record 1972: 105-114.

Chu, X. (2004). Ridership models at the stop level. Report No. BC137-31, Prepared by National Center for Transit Research for Florida Department of Transportation.

Dajani, J. S., and D. A. Sullivan. (1976). A causal model for estimating public transit ridership using census data. High Speed Ground Transportation Journal 10(1): 47-57.

Dill, D., M. Schlossberg, L. Ma, and C. Meyer. (2013). Predicting transit ridership at the stop level: the role of service and urban form. Transportation Research Board Annual Meeting, Washington D.C.

Federal Highway Administration. (2017). National Household Travel Survey. United States Department of Transportation. Washington, DC.

Federal Transit Administration. (2015). National Transit Database. Federal Transit Administration Database. United States Department of Transportation. Washington, DC.

Federal Transit Administration. (2018). Capital Investment Program. Program Overview. United States Department of Transportation. Washington, DC.

Fricker, J. D., and Shanteau, R. M. (1986). Improved service strategies for small-city transit. Transportation Research Record 1051: 30-34.

Giannopoulos, G. (1989). Bus Planning and Operation in Urban Areas: A Practical Guide. Gower Publishing. Brookfield, VT.

Greater Lafayette Public Transportation Corporation. (2018). Route and Schedule Information. Lafayette, IN (https://www.gocitybus.com/).

Greater Lafayette Public Transportation Corporation. (2018). System Map. Lafayette, IN (https://www.gocitybus.com/).

IBM Corp. (2013). IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp.

Institute of Transportation Engineers. (2004). Trip Generation Handbook, Second Edition. Institute of Transportation Engineers. Washington, DC.

Koppelman, F. S. (1983). Predicting transit ridership in response to transit service changes. Journal of Transportation Engineering, 1983, 109(4): 548-564.

Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2005). Applied Linear Statistical Models. McGraw-Hill Irwin Publishers. New York, NY.

Litman, T. (2017). Public transit’s impact on rural and small towns: A vital mobility link. APTA Technical Report 2017.

Mallett, W. J. (2018). Trends in public transportation ridership: Implications for federal policy. Congressional Research Service Technical Document 2017.

Mckee, G. and Miljkovic, D. (2007). Data Aggregation and Information Loss.

McLafferty, S. (1997). Gender, Race, and the Determinants of Commuting: New York in 1990. Urban Geography, 18:3, 192-212, DOI: 10.2747/0272-3638.18.3.192

Mistretta, M., Goodwill, J. A., Gregg, R., and DeAnnuntis, C. (2009). Best Practices in Transit Service Planning. Center for Urban Transportation Research: University of South Florida.

National Cooperative Highway Research Program. (2012). Travel demand and forecasting: parameters and techniques. Transportation Research Board Technical Report.

National Institute of Standards and Technology (NIST). (2003). Variance Inflation Factors. National Institute of Standards and Technology. Gaithersburg, MD.

National Survey of Veterans (NSV9503). (1995). Department of Veterans Affairs, National Center for Veteran Analysis and Statistics.

Pas, E. (1995). The urban transportation planning process. In S. Hanson (Ed.), The geography of urban transportation (pp. 53–77). Guilford Press. New York, NY.

Pulugurtha, S. S., and M. Agurla. (2012). Assessment of models to estimate bus-stop level transit ridership using spatial modeling methods. Journal of Public Transportation, 2012 15(1).

Schmenner, R. W. (1976). The demand for urban bus transit: A route-by-route analysis. Journal of transport economics and policy 10-1, 1976: 68-86.

Taylor, B. D., Miller, D., Iseki, H., and Fink, C. (2008). Nature and/or nurture? analyzing the determinants of transit ridership across US urbanized areas. Transportation Research Part A. doi:10.1016/j.tra.2008.06.007.

Transit Cooperative Research Program. (1998). Transit Markets of the Future. The Challenge of Change. Transportation Research Board. Washington, DC.

Transit Cooperative Research Program. (2013). Transit Capacity and Quality of Service Manual. Transportation Research Board. Washington, DC.

United States Census Bureau. (2018). How the Census Bureau Measures Poverty. United States Census Bureau. Retrieved from https://www.census.gov/topics/income-poverty/poverty/guidance/poverty-measures.html.

Yao, M., Q. Fu, and J. Li. (2017). Forecasting method for urban rail transit ridership at stop level using back propagation neural network. Discrete Dynamics in Nature and Society, 2016.