Ridership Modeling at Stop Level

by

Zhiping She

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Master of Applied Science

in

Civil Engineering

Waterloo, Ontario, Canada, 2015

AUTHOR'S DECLARATION

I hereby declare that I am the sole author of this thesis. This is a true copy of this thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

This thesis presents a ridership modeling method at stop level for public transit using multiple linear regression and Geographic Information System (GIS) analysis tools. This modeling method is applied to one express and one conventional bus route to guide ridership modeling in Waterloo Region. In the developed prediction model, the Dependent Variable (DV) is ridership, while the key Independent Variables (IVs) that affect ridership are: population within stop-based buffer (IV1); number of feeder buses that arrive at each stop (IV2); riders from other origins along the bus route (IV3). In this research, IV1 is extracted from the population data of 2011 Statistics Canada. IV2 is derived from the transit database of Waterloo Region. Finally, IV3 is computed from employment and student data of 2011 Statistics Canada, the Disaggregate Employment Trip Attracting Indices are introduced in IV3 data extraction due to the different trip-attracting strength of the different employment types at each stop. The comprehensive methods are applied in the extraction of those IVs, e.g. the effective service buffer area is decided for each stop by using simple linear regression method. Some close-by stops are combined to a segment level to avoid overlapping counting in buffering. Area-based Fraction Equation is used with combining Spatial Proximity and Weight Methods to improve the accuracy of data extraction. In regression processing, Trip Production (TP) / Trip Attraction (TA) matrices are created, confidence level of ridership is set up to 95%, stop and segment levels along one bus route, and direct and transfer boardings are combined to one prediction model for an accurate estimation. The Least Squares method is used to estimate the relationship between DV and the IVs and to find the coefficients for each bus route.

The developed ridership prediction models are validated through regression results analysis; their accuracies are verified by comparing new observed data to the predicted ridership. The results prove that the prediction models are valid and reliable, and also show that the three regression coefficients for the express model have a significantly larger contribution to the ridership than those of the conventional bus route. Finally, the modeling method is also further validated by applying to different transit service periods such as morning peak, off-peak and afternoon peak hours. The output results are reliable and valid as well.

This research provides a simple and surprisingly precise ridership modeling method. The computational complex and cost of data collection are greatly reduced in comparison to other approaches of ridership prediction. Yet the accuracies of the prediction models are significantly improved. It is expected that the method can be also used to quickly predict other transit routes, thus helping transit agencies plan new routes, evaluate existing transit routes, and manage transit system.

Acknowledgements

I would like to express my sincerest gratitude to my research advisor - Professor Jeffrey Casello for his advice and guidance during my research. He has helped me keep my research on the right track and expanded my knowledge in applying statistical theories and GIS analysis tools to transit planning fields.

I would also like to express my gratitude to Professor Bruce Hellinga and his MASc student -Sahar Tolami for their data support and advice; to all employees of the Geospatial Centre of the University of Waterloo for their data support; and to Geoffrey Keyworth, Garrett Donaher, and Jill Dickinson of the Region of Waterloo for data support and advice.

Furthermore, I would like to thank Professor Liping Fu for his course - Advanced Topic in Urban Travel Demand Modeling and Rasha Kashef for her course - Quantitative Data Analysis and their advice. My thesis benefited a lot from the knowledge I gained from these two courses.

Many thanks are given to Professor Jeffrey West and Xiaodong Lin for reviewing and Jane M. Russwurm from the Writing Center of the University of Waterloo for helping in English grammar.

And at last, I would also like to extend my appreciation to my supervisor from my previous employment in the Region of Waterloo - Henry Lueng for his encouragement and valuable comments; my best friend - Dr. Xing Qian for his help in statistics theories and his encouragement; my fellow students - Wei Ding and Dr. Wenfu Wang for their valuable comments; Pradeep Kumar Bhandari for his help in the study of the course of Urban Travel Demand Modeling; and to my all other friends for their encouragement.

Dedication

This thesis is dedicated to my lovely daughters - Angelica Bao and Yuki Bao! I invested most of my time on my studies. I am not able to pay enough attention to them and feel deeply sorry. By the time I finished this thesis, I noticed that they have already grown-up! They have understood me and given me encouragement throughout this long journey! This study is a testimony of my full love to my daughters.

Table of Contents

AUTHOR'S DECLARATION ii

Abstract iii

Acknowledgements iv

Dedication v

Table of Contents vi

List of Figures ix

List of Tables xi

Chapter 1 Introduction 1

1.1 Background 1

1.2 Motivation 3

1.3 Objectives and Approach 6

1.4 Contributions 8

1.5 Thesis Organization 9

Chapter 2 Literature Review 10

2.1 Factors Affecting Ridership 10

2.2 Ridership Prediction Techniques 11

2.3 This Research Summary 17

Chapter 3 Ridership Modeling Methodology 19

3.1 Case Study in Waterloo Region 19

3.2 The Selection of Study Routes 20

3.2.1 Data Sources 22

3.2.2 Model Structure 23

3.3 Multiple Linear Regression Theory 27

3.4 Ridership Modeling Methodology 28

Chapter 4 Ridership Modeling 30

4.1 Extraction of Dependent Variable from APCS Data 30

4.1.1 APCS Data Analysis and Preprocessing 30

4.1.2 Extraction of Dependent Variable for iXpress 200 31

4.1.3 Extraction of Dependent Variable for Route 12 41

4.2 Extraction of Independent Variables 47

4.2.1 Extraction of the Independent Variable1 (IV1) 47

4.2.2 Extraction of the Independent Variable2 (IV2) 59

4.2.3 Extraction of the Independent Variable3 (IV3) 61

4.3 Correlation Analysis between IVs 70

4.3.1 Correlation Analysis for iXpress 200 Prediction Model 71

4.3.2 Correlation Analysis for Route 12 Prediction Model 72

4.4 Prediction Models 74

Chapter 5 Regression Results Analysis 78

5.1 Regression Results Analysis for iXpress 200 78

5.1.1 Regression Statistics Analysis 79

5.1.2 Regression Properties Analysis 80

5.1.3 Regression Coefficients Analysis 81

5.1.4 Regression Residuals Analysis 83

5.1.5 Verification of the iXpress 200 Prediction Model Accuracy 85

5.2 Regression Results Analysis for Route 12 89

5.2.1 Regression Statistics Analysis 89

5.2.2 Regression Properties Analysis 90

5.2.3 Regression Coefficients Analysis 90

5.2.4 Regression Residuals Analysis 92

5.2.5 Verification of the Route 12 Prediction Model Accuracy 94

5.3 Comparison of iXpress 200 to Route 12 96

5.3.1 The Comparison of the Prediction Models for iXpress 200 and Route 12 97

5.3.2 The Comparison of the Prediction Accuracy for iXpress 200 and Route 12 98

5.4 The Validation of the Modeling Method by Different Transit Service Time 99

5.5 Chapter Summary 103

Chapter 6 Conclusions and Future Work 105

6.1 Conclusions 105

6.2 Future Work 107

Appendix A Ridership Modeling during Morning Peak Period (6:00 am - 10:00 am) 109

Appendix B Ridership Modeling during off-Peak Period (10:00 am - 3:00 pm) 113

Appendix C Ridership Modeling during Afternoon Peak Period (3:00 pm - 7:00 pm) 117

Bibliography 121

List of Figures

Figure 1.1-1 GRT Transit Network in Waterloo Region ... 2

Figure 3.1-1 Proposed Rapid Transit, Express, and Local Routes ... 20

Figure 3.2-1 Case Study Area along iXpress 200 and Route 12 ... 22

Figure 3.4-1 Ridership Modeling Methodology ... 29

Figure 4.1-1 iXpress 200 Stop Name and Number ... 32

Figure 4.1-2 iXpress 200 Weekday Boardings ... 34

Figure 4.1-3 iXpress 200 Average Weekday Ridership ... 37

Figure 4.1-4 iXpress 200 Sample Data Dispersion Degree versus Margin of Error ... 40

Figure 4.1-5 Route 12 Stop Name and Number ... 41

Figure 4.1-6 Route 12 Weekday Boardings ... 44

Figure 4.1-7 Route 12 Average Weekday Ridership ... 45

Figure 4.1-8 Route 12 Sample Data Dispersion Degree versus Margin of Error ... 47

Figure 4.2-1 IV1 - Riders from Stop-based Buffer Area (Stop Level) ... 48

Figure 4.2-2 The Extraction Steps of IV1 ... 49

Figure 4.2-3 iXpress 200 - Boardings versus Residents within Stop-based Buffer Area ... 50

Figure 4.2-4 Route 12 - Boardings versus Residents within Stop-based Buffer Area ... 52

Figure 4.2-5 Buffer Area Selection for iXpress 200 and Route 12 ... 54

Figure 4.2-6 Weight Coefficient Assignment for Stop-based Buffer Area ... 55

Figure 4.2-7 ArcGIS Intersect Analysis ... 57

Figure 4.2-8 IV2 - Number of Feeder Buses That Arrive (Network Level) ... 59

Figure 4.2-9 Transfer Rate at iXpress 200 Stops ... 60

Figure 4.2-10 IV3 - Riders from Other Origins along the Bus Route (Route Level) ... 61

Figure 4.2-11 iXpress 200 - Boardings versus Employment Data from Route Level ... 63

Figure 4.2-12 Route 12 - Boardings versus Employment Data from Route Level 67
Figure 4.3-1 2011 Population Distribution and Ridership 74
Figure 4.4-1 Microsoft Excel's Regression Tool 76
Figure 5.1-1 Hypothesis Testing on iXpress 200 Regression Model Coefficients 83
Figure 5.1-2 iXpress 200 Standard Residuals Distribution 84
Figure 5.1-3 iXpress 200 Normal Probability Plot of the Standard Residuals 85
Figure 5.1-4 iXpress 200 Average 5 Weekday Boardings: Actual versus Predicted 87
Figure 5.1-5 iXpress 200 Average Weekday Boarding Volume at Each Stop 88
Figure 5.2-1 Hypothesis Testing on Route 12 Regression Model Coefficients 92
Figure 5.2-2 Route 12 Standard Residuals Distribution 93
Figure 5.2-3 Route 12 Normal Probability Plot of the Standard Residuals 93
Figure 5.2-4 Route 12 Average 5 Weekdays Boardings: Actual versus Predicted 95
Figure 5.2-5 Route 12 Average Weekday Boarding Volume at Each Stop 96
Figure 5.3-1 The Comparison of the Prediction Accuracy for iXpress 200 and Route 12 98
Figure 5.4-1 Boarding Volume during Service Time 102

List of Tables

Table 2.1-1 Factors Affecting Transit Ridership 10

Table 2.2-1 Prediction Techniques Used by Transit Agencies 11

Table 2.2-2 Examples: Stop Level Prediction Models 15

Table 2.3-1 This Research Summary 18

Table 3.2-1 Dependent Variable and Independent Variables 26

Table 4.1-1 iXpress 200 Weekday Boardings from September 12 to October 12, 2011 33

Table 4.1-2 iXpress 200 Average Weekday Ridership 36

Table 4.1-3 iXpress 200 Margin of Error and Confidence Interval (CI) 39

Table 4.1-4 Route 12 Weekday Boardings in November 2012 43

Table 4.1-5 Route 12 Average Weekday Ridership 45

Table 4.1-6 Route 12 Margin of Error and Confidence Interval (CI) 46

Table 4.2-1 iXpress 200 - The Selection of Effective Service Buffers for 14 Stops 51

Table 4.2-2 Route 12 - The Selection of Effective Service Buffers for 8 Stops and 3 Segments 53

Table 4.2-3 Buffer Width versus Weight Coefficient 56

Table 4.2-4 iXpress 200 - The Data Extraction Method of IV3 64

Table 4.2-5 Route 12 - The Data Extraction Method of IV3 68

Table 4.3-1 iXpress 200 Correlation of Dependent Variable and Independent Variables 71

Table 4.3-2 Route 12 Correlation of Dependent Variable and Independent Variables 72

Table 5.1-1 iXpress 200 Regression Statistics Analysis 80

Table 5.1-2 iXpress 200 Regression Model Properties and Significance F 81

Table 5.1-3 iXpress 200 Regression Coefficients Analysis 82

Table 5.1-4 iXpress 200 Average Error Rate and The Dispersion Degree 86

Table 5.2-1 Route 12 Regression Statistics Analysis 89

Table 5.2-2 Route 12 Regression Model Properties and Significance F 90

Table 5.2-3 Route 12 Regression Coefficients Analysis 91

Table 5.2-4 Route 12 Average Error Rate and the Dispersion Degree 94

Table 5.3-1 The Comparison of the Prediction Models for iXpress 200 and Route 12 97

Table 5.3-2 The Comparison of Average Error Rates for iXpress 200 and Route 12 98

Table 5.4-1 iXpress 200 Regression Results Comparison during Different Service Time.............. 100

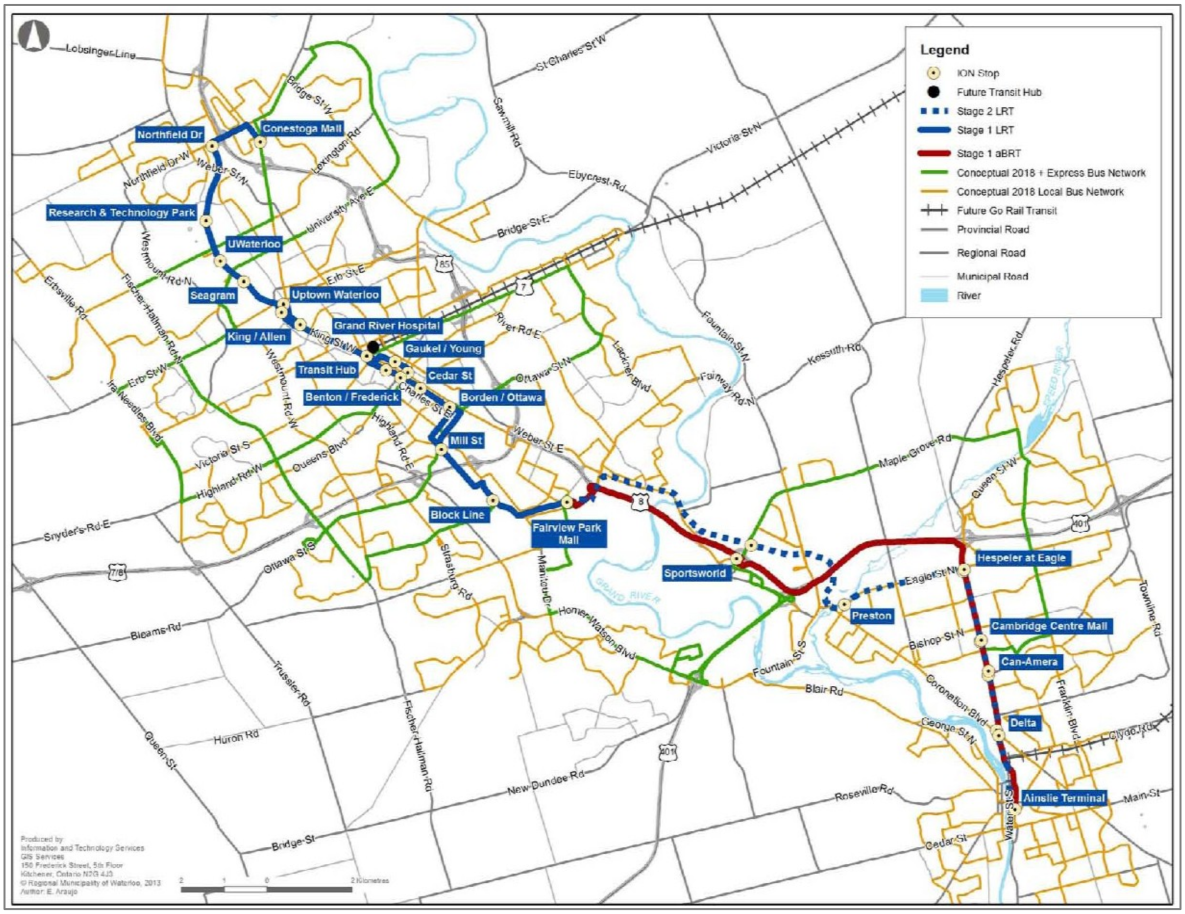
Table 5.4-2 Ridership Distribution Trend at Different Types of Stops 103

Chapter 1 Introduction

1.1 Background

Grand River Transit (GRT), a public transit agency in Waterloo Region, Ontario, Canada, is currently facing a rapidly increasing ridership and a historical change in transportation planning and engineering due to the introduction of adapted Bus Rapid Transit (aBRT), which will be open by 2015 and the Light Rapid Transit (LRT), which will be open by 2017. By then, aBRT and LRT will service the tri-cities of Waterloo, Kitchener, and Cambridge as the trunk line along the central transit corridor. The existing GRT conventional transit network will be reconfigured to fully support the trunk line (Figure 1.1-1). Currently, iXpress - Route 200 (one of GRT's express bus routes) has been serving this corridor since 2005. The express has been used to collect passenger data for the future rapid transit system; from Automatic Passenger Counting Service data analysis, the weekday ridership (boardings) reached 258,798 from September 12 to October 12 in 2011.

Precise ridership prediction has become increasingly important for transit planning and management. It can help decision makers and transit agencies to save budget and improve the quality of services by predicting trend, estimating capital and operating costs, and planning current and future transit services; moreover, adjusting schedule adherence (such as headway, operation periods, route configuration, and service type), adjusting fare changes according to ridership changes, and resizing facilities according to station capacity estimation; in addition, quantifying benefits and helping to achieve public transportation objectives of congestion relief, environmental enhancement, and attracting new passengers (Wilson, 2010) (TCRP, 2006).



Source: The Region of Waterloo

Figure 1.1-1 GRT Transit Network in Waterloo Region

Ridership prediction models are classified by the four levels of transit infrastructure: transit network, route, segment, and stop.

A network-level model, such as the traditional four-step transportation model, is time consuming and expensive to build. It is only suitable for long-term transit planning and prediction, and is not flexible enough to respond quickly to changes in short-term market demand.

Route-level ridership prediction generally has four types of approaches: professional judgment, survey-based methods, cross-sectional data models, and time-series data models. Some route-level models make the assumption of homogeneous service and land use, and thus are not suitable for complicated, long-route ridership prediction such as long routes from downtown to suburbs; some route-level models such as those based on professional judgment estimate suffer from the lack of reliable data and method for making accurate ridership prediction.

The segment-level models group analysis units by time-point stops or fare zones. It still has the same problems of assuming homogeneous socio-demographics along the routes; thus this model cannot reflect stop-level attributes.

Stop-level ridership prediction is considered more precise for monitoring and predicting transit volume as a real-world reflection. Ridership data can be collected at stop level. The stop-level models can reflect ridership changes accurately and quickly from affecting characteristics of each stop such as household, feeder bus service demand, and land use.

1.2 Motivation

This research is motivated by the historical changes in the newly reconfigured transit network system in Waterloo Region and the author's interest in transit planning and management. This thesis will develop a ridership modeling method at stop level by using multiple linear regression and Geographic Information System (GIS) analysis tools. The ridership modeling method will be applied to one express and one conventional bus route in Waterloo Region to develop the prediction models for the two bus routes. In the prediction models, the dependent variable (DV) is defined as the ridership, while the key independent variables (IVs) that affect ridership are identified based on real-world ridership sources from stop, route, and network levels.

Human activities generate a variety of Trip-Production (TP) and Trip-Attraction (TA) trips in urban areas; TP refers to the factors including number of households, income levels, vehicle ownership, and land use characteristics; TA refers to the trips generated by the non-residential ends, such as educational institutions, shopping malls, and work places. This thesis focuses on the number of riders who board at each stop; in other words, only the boardings (ridership) are analyzed. As a result, a total of three IVs are identified based on space accessibility including

IV1 - Population (residents) within the stop-based buffer area (TP)

IV2 - Number of feeder buses that arrive at each stop (TA indirectly)

IV3 - Riders from other origins along the bus route (TA directly)

The prediction models in this thesis calculate the ridership using the above three IVs, plus intercept and errors. IV1 is extracted from the population data of 2011 Statistics Canada Household Survey; IV2 is derived and extracted from the transit database of the Region of Waterloo; IV3 is extracted from employment and student data of 2011 Statistics Canada Household Survey, the Disaggregate Employment Trip Attracting Indices are introduced in IV3 data extraction due to the different trip-attracting strength of the different employment types at each stop (Casello & Smith, 2006).

In this research, the three main aspects will be addressed: First, for the identification and confirmation of the key Independent Variables (IVs), considering the space accessibilities of all boarding sources from transit stop, route, and network levels; avoiding high correlations among IVs. Second, for the extraction of IVs, the comprehensive methods will be considered, e.g. deciding effective service buffer area for each stop by using simple linear regression method; combining some close-by stops to a segment level to avoid overlapping counting in buffering; and moreover, introducing Area-based Fraction Equation with combining Spatial Proximity Method (SPM) and Spatial Weight Method (SWM) to extract the data within the effective service buffer area; in addition, creating Trip Production (TP)/Trip Attraction (TA) matrices with data consistency in time and space. Third, in regression processing, confidence level of ridership is set up to 95%, combining stop and segment levels along one bus route, and direct and transfer boardings to one prediction model for an accurate estimation. The Least Squares method is used to estimate the relationship between DV and IVs in order to find the coefficients for each bus route.

Furthermore, the developed ridership prediction models will need to be validated through regression results analysis, their accuracies will need to be verified through comparing new sample data to the related predicted value. Moreover, the modeling method will be further validated through applying to different transit service periods such as morning peak, off-peak and afternoon peak hours.

This research provides a simple and surprisingly precise ridership modeling method for public transit. The computational complexity and cost of the data collection are greatly reduced in comparison to other approaches of ridership prediction. Yet the accuracies of the prediction models are significantly improved. The purpose is to quickly respond to market demand with proper supply for short term transit planning. It is expected that the modeling method can be also used to quickly predict other transit routes, thus helping transit agencies plan new routes, evaluate transit routes, and manage transit system.

1.3 Objectives and Approach

The research objectives are:

- To develop a ridership modeling method for building ridership prediction models at stop level using multiple linear regression theory and GIS analysis tools
- To provide methods in DV data analysis and preprocessing
- To explore methods in key IVs data identification, confirmation, and extraction
- To establish methods in TP/TA matrices creation
- To check the relationship between DV and IVs and among IVs in correlation analysis
- To establish methods in regress results analysis
- To provide methods in building prediction models
- To provide methods in verifying the accuracies of the prediction models and analyzing the residuals between observation values and prediction values

The research approach used is:

- Review related research work in ridership modeling
- Find the difference between my research work and the related literature
- Give case study corridors - one express route 200 and one conventional bus route 12 of GRT in Waterloo Region as examples
- Collect and analyze DV and IVs data for the two bus routes
- Define DV, analyze and preprocess APCS records, extract DV
- Identify and confirm key IVs through analyzing and repeatedly testing the factors that directly relate to ridership
- Decide effective service area for each stop by using simple linear regression method

- Combine some close-by stops to a segment level to avoid overlapping counting
- Extract residents, employment, and student data from the census tract boundary of 2011 Statistics Canada Household Survey to effective service buffer area based on Area-based Fraction Equation by using GIS analysis tools
- Introduce Spatial Proximity Method (SPM) and Spatial Weight Method (SWM) to improve the accuracy of data extraction for an accurate estimation
- Introduce the Disaggregate Employment Trip Attracting Indices in IV3 data extraction due to the different trip-attracting strength of the different employment types at each stop
- Extract the number of feeder buses that arrive at each stop from the transit database of the Region of Waterloo
- Try to keep TP/TA matrices with data consistency in time and space
- Combine direct and transfer boardings to one model for an accurate estimation
- Test the relationship between DV and IVs and the correlations among IVs
- Create ridership prediction models through regression processing for the iXpress 200 and the conventional bus Route 12 of GRT
- Analyze and validate the regression results
- Verify the accuracy and reliability of the ridership prediction models by comparing new sample observed data to predicted ridership
- Compare and analyze the contributions of the regression coefficients and related IVs to ridership between iXpress 200 and Route 12
- Verify the modeling method by applying to different transit operating periods - morning peak, off-peak, and afternoon peak
- Conclude with recommendations and future work

1.4 Contributions

The key contributions of this thesis include

- A ridership modeling method has been developed at stop level
- The key independent variables (IVs) are identified from transit stop, route, and network levels; redundant and high correlation IVs are avoided in order to reduce the cost of the data collection and to improve the accuracies of the prediction models
- The effective service area are decided by using the simple linear regression method
- Some close-by stops are combined to a segment level to avoid overlapping counting
- TP like residents and TA like employment and student data are collected and extracted by using Area-based Fraction Formula and GIS analysis tools
- Spatial Proximity Method and Spatial Weight Method are introduced for improving the accuracy of the data extraction
- The Disaggregate Employment Trip Attracting Indices (Casello, 2006) are introduced in IV3 data extraction due to the different trip-attracting strength of the different employment types at each stop
- Trip Production/Trip Attraction matrices are created with data consistency in time and space
- The prediction models for the iXpress 200 and the conventional bus route 12 have been developed for different transit service periods - average weekday, morning peak, off-peak, afternoon peak hours

1.5 Thesis Organization

This thesis consists of six chapters. Chapter 1 briefly describes the research background, ridership prediction at different levels, the motivation in developing a ridership modeling method, and the objectives and approach of the model development, and the contribution of the thesis. Chapter 2 is a literature review of the factors affecting ridership and ridership prediction techniques used by transit agencies, also compares the difference between this research and the related literature. Chapter 3 addresses case study area, defines DV, analyzes and identifies key IVs, explains multiple linear regression theory, and presents a ridership modeling method. Chapter 4 covers ridership modeling in the data extraction methods and correlation analysis of DV and IVs, and the ridership prediction regression equations. Chapter 5 delivers the analysis and validation of regression results, the verification of the regression models accuracies, the validation of the modeling method, and the regression coefficients comparisons between iXpress 200 and conventional bus Route 12 of Waterloo Region. Conclusion and visions for future work comprise Chapter 6.

Chapter 2 Literature Review

Chapter 1 introduced the background and motivation of this thesis; listed the research objectives and scope; listed the key contributions; and described the organization of this thesis. Chapter 2 will review the factors affecting ridership and ridership prediction techniques.

2.1 Factors Affecting Ridership

Much research literature analyzes and discusses the many factors that affect transit ridership. These factors can be generalized as two types: Exogenous and Endogenous. Exogenous factors are uncontrollable, such as socio-economic factors, public finance factors, and spatial factors. By contrast, endogenous factors are controllable, such as fare price, service quantity and quality factors. These factors are summarized in Table 2.1-1 (Wilson, 2010).

Table 2.1-1 Factors Affecting Transit Ridership

Exogenous (uncontrollable)	**Endogenous (Controllable)**
Population density	Service supply in fare
Demographics in age and gender	Headway/Schedules
Residential and employment relocation	Waiting time
Employment density	In-vehicle time
Traffic congestion levels	Route design
Parking availability/Policies	Crowding
Auto Ownership/Availability	Comfort
Operating costs	Reliability
Fuel prices	Service quantity and quality
Income	Auto parking costs
Public finance	Query information systems
Public policies and land uses	Marketing

Source: (Wilson, 2010).

Employment and population are the determinants of transit ridership. The employment factor usually affects weekday ridership because of the produced trips between home and work from Monday to Friday. The BRT/LRT modes of transit may attract more employment driven ridership due to their more comfortable and reliable performances compared to conventional bus services. Population mostly drives weekend ridership because of shopping, recreation activities, and so on.

2.2 Ridership Prediction Techniques

The weight magnitude of the above factors is determined by a variety of prediction techniques by the transit agencies. These techniques are summarized in TCRP Synthesis 66 report P19 as shown in a survey in Table 2.2-1 below.

Table 2.2-1 Prediction Techniques Used by Transit Agencies

Prediction Techniques	**No. Agencies Responding**	**Agencies Responding (%)**
Professional Judgment	29	83
Rules of thumb/similar routes	28	80
Service elasticity	22	63
Four-step travel demand model	18	51
Econometric model	7	20
Regression analysis	7	20
Other	7	20
Total Responding	35	100

Source: (Dan Boyle & Associates, 2006)

The four-Step Travel Demand Model (FSTDM) is a traditional transportation planning system technique for network-level prediction. It is a large-scale transportation planning tool for long-term and new services. It concentrates on travel origins and destinations and the equilibrium of the network system (Peng, 1994). FSTDM processes include trip generation, trip distribution, model split, and trip assignment, which require huge data support and are time-consuming. The FSTDM cannot be used to quickly respond to transit demand changes and make ridership prediction at route, segment and stop levels (Stopher & Mulhall, 1992). The rest of the prediction techniques in Table 2.2-1 can be used for route, segment and stop level prediction.

The above techniques can be classified into four types of approaches: professional judgment, survey-based methods, cross-sectional data models, and time-series data models.

Professional judgment can be applied to a variety of market demand changes based on the experience and knowledge of the planners. This approach does not need reliable data and mathematical models (Wilson, 2010).

Survey-based methods use non-committal surveys and/or stated preference surveys. Non-committal surveys are based on (i) surveying potential riders on new services or service changes; (ii) applying survey response data to infer total population at the market segment level; (iii) multiplying by an acceptable adjustment factor to adjust for non-committal bias (the factor can range in practice from 0.05 to 0.50). Stated preference surveys are also called conjoint analysis, which are the emerging feasible statistical tools for assessing probable responses to new transportation system changes. They include detailed, strict survey designs, data analysis, and a series of tradeoffs for planners to rank relative importance of a variety of improvements (Wilson, 2010).

Cross-Sectional Models (C-SMs) are direct demand models based on regression analysis theory; they employ route and demographic data to illustrate route ridership. For example, Agrawal, D. (1978) selected four independent variables (IVs) - average adult fare, vehicle miles of service, jobs, and a miscellaneous-events factor - to explain the dependent variable of ridership in a multiple linear regression model. The regression results show that the four IVs have a strong relationship with transit ridership. C-SMs deal with space variations for different routes.

Time-Series Models (T-SMs) are also direct demand models based on regression analysis theory, they are used to combine the temporal phenomena for prediction purposes to reflect the time variation of the transit ridership. The T-SMs are more accurate for short-time-period ridership predictions. The models can be classified into three types: elasticity analysis models, trend analysis models, and multivariate autoregressive-moving average (ARIMA) models.

Elasticity analysis models have been used extensively for predicting transit ridership changes. Casello, J., & Hellinga, B. (2008) used utility theory and an economic tool to compute the elasticity of transit demand according to the reduction in generalized travel cost and the increase in ridership, and applied the developed model to the iXpress corridor of Waterloo Region as a case study. The findings show that the cost saved by individual Origin-Destination (O-D) pair ranges from zero to as high as 33 percent. Moreover, the model is successfully used to test the sensitivity of the savings of travel cost to the assumed weights of waiting time, transfer time, and value of travel time.

Direct regression models can be used for short-term transit planning at route level. Such models can directly reflect transit service demand through explanatory variables. However, because some regression analyses ignore the inter-route relationship, assuming each route as independent from all others, ridership estimate errors can occur in route-level and network-level systems. Some regression models assume homogeneous social-demographics and land-use along a route segment; and some regression analysis cannot keep data consistency. For example, if a route-level ridership needs to be estimated, the population data should be extracted at a reasonable buffer distance from the census tract boundary instead of using population inside the boundary (Peng, 1994).

Peng et al. (1997) have developed ridership models at segment level by fare zones; Kimpel et al. (2000) have developed ridership prediction models at segment level by defining segments by time-point stops. These research methods improved the prediction accuracy for the segments of a route. However, the segment-level prediction models still have some of the same flaws as route-level prediction models. For example, the social-demographics are assumed homogeneous along a route segment. Moreover, stop-level characteristics and some service variables cannot be reflected from route and segment levels.

Kikuchi & Millkovic (2001) developed ridership prediction models at stop level. The fuzzy inference and directionless stops are applied in their research. This research method is similar to the traditional cross-classification approach. Several stop-level prediction models are summarized in Table 2.2-2.

Table 2.2-2 Examples: Stop Level Prediction Models

Dependent Variable	Independent Variables	R2	Comments	Sources
Average Weekday Boardings at a stop	+ socio-demographics (...) + transit level of service value (…) + street environment for pedestrians (…) + accessibility to population and Employment (…) + interaction with other modes (…) + competition with other TLOS stops (…)	0.54	1. The cost of data collection is increased due to too many IVs (each IV includes many sub-IVs). 2. The redundant IVs may exist and affect the accuracy of the prediction models.	(Chu, 2004) publisher: National center for Transit Research University of South Florida
Ridership at a stop	+ traveler characteristics (White percentage, income percentage) + transit service characteristics (Bus frequency, UT Shuttle Bus, UT Neighborhood) + land use (Multi-family acreage, parking acreage) + other contextual factors (count)	0.28	The prediction model ignored the two important variables from route level due to stop attraction and network level due to feeder bus services, therefore, the accuracy of the prediction model may be poor.	(Park, 2011). Ridership Analysis at the Stop Level: Case Study of Austin, TX
Weekday Ridership at a stop	+ average median household income + major residential + major attractions + supermarkets + retail jobs	0.72	Feeder buses services should be considered.	(Nashua Regional Planning Commission, U.S., 2011) Ridership Prediction Model for the NASHUA TRANSIT SYSTEM

A number of boardings at a stop	+ Resident population in buffer + Number of dwelling units with zero cars in buffer + Number of dwelling units with retirees in buffer + Number of dwelling units with workers in buffer + Total employment in buffer + Service frequency + Route fare in cents + Average parking cost at destination stops + Presence of feeder service	Average Absolute Error (Observed-Predicted): 21-27%	1. IVs covered all riders boarding sources. 2. An algorithm is given to solve overlapping counting caused by close-by stops buffering in modeling.	(Ram M. Pendyala, lke Ubaka, and Nadarajah Sivaneswaran, 2004). GIS-Based Regional Transit Feasibility Analysis and Simulation Tool for the Florida Department of Transportation

To summarize, many studies have explored numerous factors affecting transit ridership; and different factors have been used as independent variables in developed models to predict ridership from route level to stop level. In these studies, the dependent variable, ridership, can be defined as boardings, or boardings+alightings. Ridership can refer to different service times (average weekday, AM/PM peak hours, off-peak hours, and weekends). Independent variables can be selected based on different case study situations, geographic districts, service time, or availability of data.

Although ridership prediction models have already been discoursed in the above studies from network to stop levels, some problems still exist. For example, some regression models have too many IVs with redundant data. These redundant IVs may have high correlation between or among each other and may distort the results; other IVs may have spatial data consistency problems. Furthermore, in some models, transit network level factors are not considered at all in the selection of IVs, which lowers ridership prediction accuracy.

2.3 This Research Summary

This thesis will develop a ridership modeling method at stop-level based on multiple linear regression theory and GIS analysis tools. In the case study in this thesis, the selection of independent variables is limited by the availability of data. Therefore, this thesis mainly focuses on the available data of ridership sources at stop level. The ridership sources are identified from transit stop, route, and network levels in the real world. The selection of stop-based buffer size is repeatedly tested from 300 meters to 1000 meters until the best output result by using simple linear regression method. Data collection is based on different buffer sizes, land use characteristics, and stop attraction types. Data extraction methods are based on Area-based Fraction Equation and GIS analysis tools. The strength coefficients of attractions for different employment types, SPM, and SWM are introduced for helping data extraction. Furthermore, DV and IVs are collected and extracted in keeping data consistency in time and space. This research summary is shown in Table 2.3-1 below.

Table 2.3-1 This Research Summary

Dependent Variable	Independent Variables	R2	Comments
Average Weekday Boardings at a stop	+ IV1 (TP) from stop level - Residents within stop-based buffer + IV2 (TA indirectly) from network level - Number of feeder buses that arrive at each stop +IV3 (TA directly) from Route level - Riders from other origins along the bus route due to stop attraction	0.95-0.98 Average Absolute Error Rate (Observed-Predicted)/Observed: 12.78%-31.85%	1. All boarding sources are covered by the three IVs 2. The high correlations among IVs are avoided 3. The costs of data collection are reduced 4. Effective service buffer area are decided by using simple linear regression method 5. Segment-based buffer is introduced to solve overlapping counting caused by close-by stop in buffering 6. The strength coefficients of attractions in different employment types are introduced for each stop 7. SPM and SWM are introduced for improving the accuracy of the data extraction 8. Data collection are finished at stop Level 9. Stop and segment levels are combined to one model during regression processing 10. Direct boarding and transfer boarding are combined to one model for improving the accuracy of the prediction models

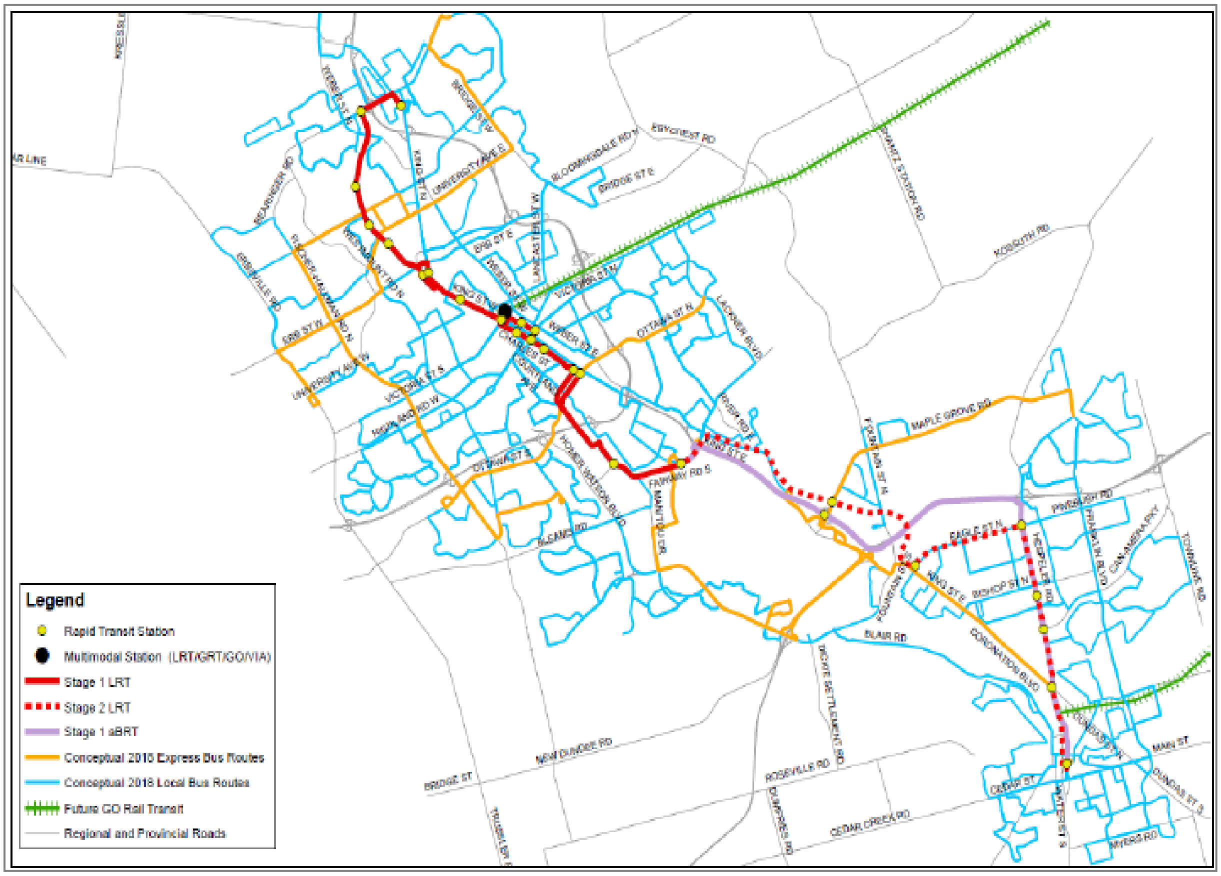
This research considers all boarding sources based on the space accessibility. The redundancy in IVs is avoided to assure the types of IVs are clear, concise, and overall. The cost of the data collection is greatly reduced. The accuracy of the prediction models is significantly improved. The purpose is to use the modeling method to quickly respond to transit demand change and precisely estimate ridership at stop level for transit planning and management.

Chapter 3 Ridership Modeling Methodology

Chapter 2 reviewed the factors impacting ridership and the ridership prediction techniques at different transit service levels; then compared the difference between my research and the related literature. Chapter 3 will first define the dependent variable (DV); then identify the key independent variables (IVs) in the case study area. Next, I will present multiple linear regression theory in order to provide mathematical equations for ridership prediction. Finally, I will develop a ridership prediction modeling method at stop level.

3.1 Case Study in Waterloo Region

Waterloo Region, Ontario, Canada is selected as the case study area. The new GRT transit network configuration is stated in the Region Transportation Master Plan and illustrated in Figure 3.1-1.



Source: Region of Waterloo

Figure 3.1-1 Proposed Rapid Transit, Express, and Local Routes

3.2 The Selection of Study Routes

The iXpress Route 200 (iXpress) and the conventional bus Route 12 (Route 12) of Grand River Transit (GRT) in the Region are selected as case study bus routes. The iXpress is an express bus service with a total length of 37km and 14 stops. It connects Waterloo (from Conestoga Mall), Kitchener (passing through Fairview Mall), and Cambridge (to Ainslie Terminal) along the Region's central transit corridor. The iXpress has been providing express service - fewer stops and less delay since 2005. Therefore, it is attracting more and more riders along its route. It will be upgraded and replaced by adapted Bus Rapid Transit (aBRT) - from Fairview Mall to Ainslie Terminal in 2015 and

Light Rapid Transit (LRT) - from Conestoga Mall to Fairview Mall in 2017. Rapid transit will run an operating speed at 20-25km/h in the downtown area and up to 70km/h within rail corridors away from other pedestrians and traffic.

Route 12 is a conventional bus service servicing a variety of communities of different income levels. It has a total length of 25km and 139 stops. It links Waterloo (from Conestoga Mall) to Kitchener (to Fairview Mall).

These two bus routes have different alignments (roughly parallel), however they pass through the same four stops at Conestoga Mall, University of Waterloo, Wilfrid Laurier University, and Fairview Mall. Sample data have been collected from the two bus routes in order to develop a ridership prediction modeling method at stop level. The case study bus routes of iXpress 200 and Route 12 are shown in Figure 3.2-1. In Figure 3.2-1, 800-meter buffers are created for the two bus routes in order to analyze and compare ridership volume and fluctuation at route and stop levels.

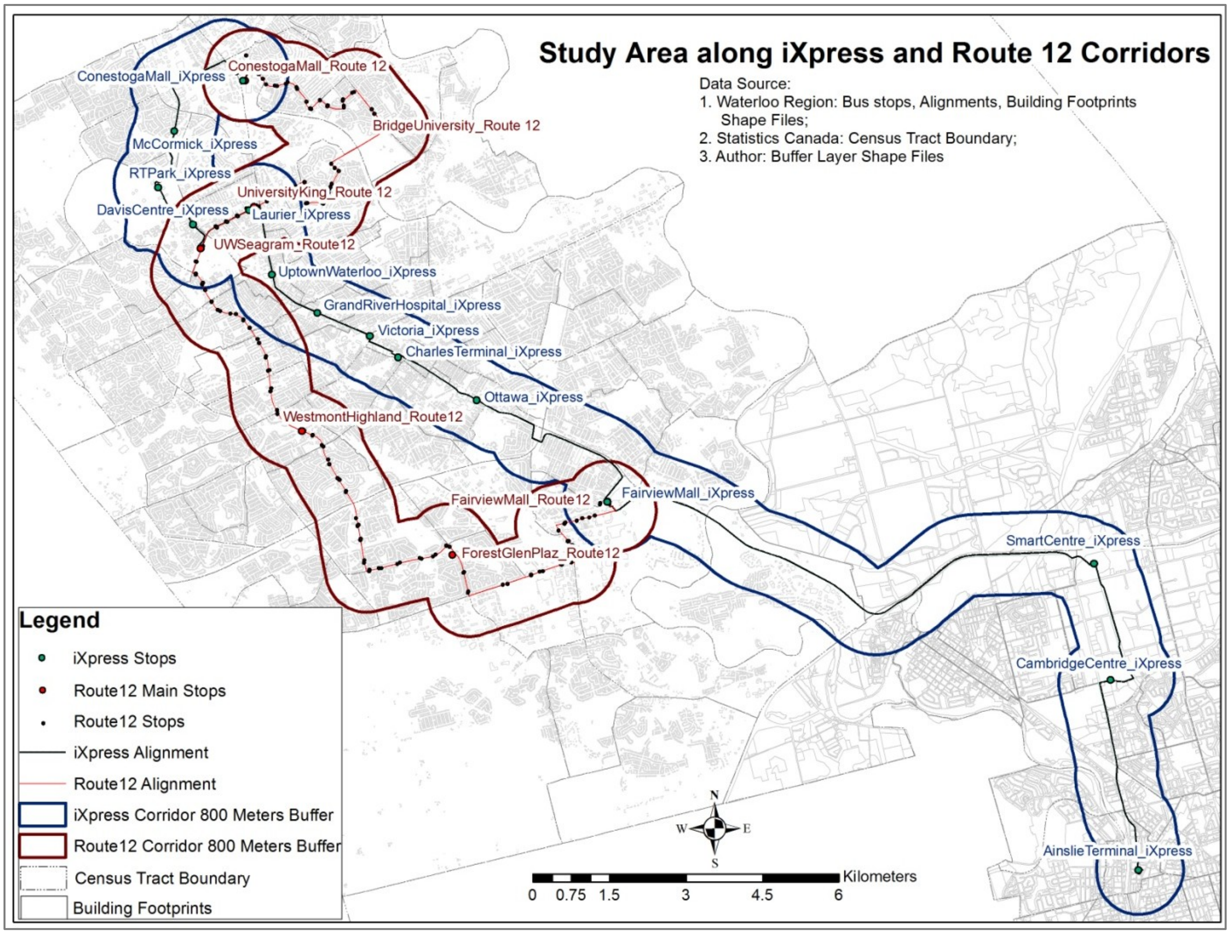


Figure 3.2-1 Case Study Area along iXpress 200 and Route 12

3.2.1 Data Sources

The following sample data are collected from 2011 Statistics Canada; these data are downloaded and processed by the staff of the Geospatial Centre, University of Waterloo:

- Total employed population aged 15 years and over with a usual place of work or no fixed workplace address by mode of transportation (excel file).
- 2011 population data (excel file)
- 2011 number of private dwellings (excel file)

- Total labour force population aged 15 years and over by occupation - National Occupational Classification (NOC) 2011
- Census Tract Boundary (shape file with related attribute database)

The following sample data are provided by the public sector - the Region of Waterloo:

- Grand River Transit iXpress alignment (shape file)
- Grand River Transit iXpress stops (shape file)
- Grand River Transit Route 12 alignment (shape file)
- Grand River Transit Route 12 stops (shape file)
- Building foot prints (shape file)
- Zoning (shape file)
- September 12, 2011 to October 13, 2011 iXpress APCS (Automatic Passenger Counting Service) data by stop (excel file)
- November 2012 Route 12 APCS (Automatic Passenger Counting Service) data by stop (excel file)

3.2.2 Model Structure

This research's purpose is to develop a ridership modeling method at stop level in order to quickly respond to transit demand changes and provide the precise ridership estimation for transit agencies. Therefore, Multiple Linear Regression (MLR) theory is used in this research. Multiple factors or independent variables (IVs) are used to explain the changes of the dependent variable (DV) in the regression model.

Data collection for this research is challenging. For example, a huge dataset is collected from 2006 Statistics Canada (StatCAN), 2006 Transportation Tomorrow Survey (TTS), and the Region of Waterloo. These data are repeatedly tested and analyzed in order to identify and confirm the correct IVs. However, a high correlation is repeatedly observed among IVs.

Data collection is moved to the latest data - 2011 Statistics Canada (2011 TTS data are not available). However, there is no data available for income levels and the number of households with zero, one, two, or more cars. The only available data are population, employment, and total private dwellings from 2011 Statistics Canada. These data are processed to Excel tables by the staff in the Geospatial Centre of the University of Waterloo.

In analyzing and testing population, employment, and total private dwellings data, high correlation coefficients of about 0.9 are found among these three variables. In order to prevent the correlated IVs from distorting the prediction models, the variable of private dwellings is removed from the selection of the IVs. Moreover, population and employment data are classified based on age group and made different components for modeling test, but the preliminary prediction model did not still work well.

As a result, the selection of IVs is limited by the availability of data and the high correlation of the available data. Therefore, this research has to face the fact: only the factors that directly impact on ridership are considered. For example, which kinds of people would board on a bus at stop level? Where are the riders from? How to predict the number of the ridership at each stop? Some observations are taken at the different stop types along the iXpress 200 corridor. In addition, the space accessibilities of all boarding sources are considered from transit stop, route, and network levels. The time accessibilities of all boarding sources are analyzed based on APCS records and land use characteristics at each stop. Moreover, the Quantitative Data Analysis course is taken for helping IVs identification and modeling processing. See text book - Probability and Statistics for Engineers and Scientists (Walpole & Myers, 1978)

Therefore, the final model structure is determined based on the failures of the previous prediction models; the observations at different stop types in the real world; transportation courses study and much research literature study. See text books - Urban Public Transportation Systems and Technology (Vuchic, 2007), Urban Transit: Planning, Operations and Economics (Vuchic, 2005), and Modeling Travel Demand for Urban Transportation Planning (Fu, 2012).

In this research, the dependent variable (DV) is defined as average weekday boardings, while the three key independent variables (IVs) are identified from the transit stop, route, and network levels to explain the changes of the dependent variable (DV) in the regression model. IV1 is identified from stop level and defined as residents within stop-based buffer area; it is referred to as Trip Production (TP). IV2 is identified from transit network level and defined as number of feeder buses that arrive at each stop; it is referred to as Trip Attraction indirectly (TA). IV3 is identified from transit route level and defined as riders from other origins along the bus route due to stop attractions; it is referred to as Trip Attraction directly (TA). The final model structure can be expressed in the multiple linear regression mathematical equation; it is explained by the three independent variables, intercept, and errors. The dependent variable (DV) and the independent variables (IVs) are summarized in Table 3.2-1.

Table 3.2-1 Dependent Variable and Independent Variables

Variables	Description	Data Source
Dependent Variable	Average weekday boardings at stop level for iXpress 200 and Route 12	○ Extracted from the APCS data of the Grand River Transit (GRT) of the Region of Waterloo ○ 22 weekdays of APCS data (Sep. 12-Oct. 13, 2011) for iXpress 200 ○ 22 weekdays of APCS data (Nov. 1 -Nov. 30, 2012) for Route 12.
Independent Variables	IV1 - Trip Production (TP): Residents within stop-based buffer area, the buffer size for each stop is decided by simple linear regression method	○ Data source is from 2011 Statistics Canada. ○ The stop-based buffer data are extracted from Census Tract boundary of 2011 Statistics Canada using Buffer and Intersect analysis tools of ArcGIS and Area based Fraction method.
	IV2 - Trip Attraction (TA) indirectly: Number of feeder buses that arrive at each stop	○ Collected from GRT transit database of the Region of Waterloo.
	IV3 - Trip Attraction (TA) directly: Riders from other origins along the bus route due to stop attraction, most riders are employees and students who go and come back for work and study during weekday	○ Collected from 2011 Statistics Canada and the report (The Region of Waterloo, 2009)

3.3 Multiple Linear Regression Theory

Based on Multiple Linear Regression (MLR) theory, multiple independent variables are used to explain the changes of the dependent variable in the regression model. When the prediction model is linear in coefficients, it is called a multiple linear regression model (Walpole & Myers, 1978) (Kashef, 2014). It is mathematically expressed as:

Real multiple linear regression models:

$$\boldsymbol{Y_i = B_0 + B_1 x_{1i} + B_2 x_{2i} + \cdots + B_k x_{ki}} \qquad \textbf{(Real or underlying)} \qquad \textbf{(3.1)}$$

Predicted multiple linear regression models:

$$\boldsymbol{\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki}} \qquad \textbf{(Estimate or fitted)} \qquad \textbf{(3.2)}$$

Residual:

$$\boldsymbol{\varepsilon_i = Y_i - \hat{y}_i} \qquad \textbf{(3.3)}$$

Where

Y_i: Actual dependent variable

B_k: Determines the contribution of the independent variable x_{ki} to the dependent variable Y_i

x_{ki}: Independent variables

$\hat{y}_i$: Predicted dependent variable

b_k: Estimated coefficients

ε_i: Residual (the difference between the real data and the fitted line)

i: i = 1, 2, ..., n and n > k

The least squares method (fitted equation) is used to calculate $\{b_0, b_1, \dots b_k\}$ for estimates of $\{B_0, B_1, \dots B_k\}$ by finding a fitted line (surface) such that the sum of squares of the errors (SSE) is minimized:

$$\textbf{SSE} = (\boldsymbol{x} + \boldsymbol{a})^{\boldsymbol{n}} = \sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{n}} \boldsymbol{\varepsilon}_{\mathbf{i}}{}^{\mathbf{2}} = \sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{n}} (\boldsymbol{Y}_{\boldsymbol{i}} - \widehat{\boldsymbol{y}}_{\boldsymbol{i}})^{\mathbf{2}} = \sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{n}} (\boldsymbol{Y}_{\boldsymbol{i}} - \boldsymbol{b}_{\mathbf{0}} - \boldsymbol{b}_{\mathbf{1}}\boldsymbol{x}_{\mathbf{1}\boldsymbol{i}} - \cdots - \boldsymbol{b}_{\mathbf{k}}\boldsymbol{x}_{\mathbf{ki}})^{\mathbf{2}} \quad \textbf{(3.4)}$$

The solution is to differentiate SSE with respect to $\boldsymbol{b}_{\mathbf{k}}$.

The matrix representation of the multiple linear regression models is presented as follows:

$$\mathbf{y} = \boldsymbol{X}\mathbf{B} + \boldsymbol{\varepsilon}, \quad \mathbf{y_i} = \boldsymbol{b}_{\mathbf{0}} + \boldsymbol{b}_{\mathbf{1}}\boldsymbol{x}_{\mathbf{1}\boldsymbol{i}} + \boldsymbol{b}_{\mathbf{2}}\boldsymbol{x}_{\mathbf{2i}} + \cdots + \boldsymbol{b}_{\mathbf{k}}\boldsymbol{x}_{\mathbf{ki}} + \boldsymbol{\varepsilon}_{\mathbf{i}}, \quad \mathbf{1} \le \boldsymbol{i} \le \boldsymbol{n}, \;\; \boldsymbol{n} > \boldsymbol{k} \qquad \textbf{(3.5)}$$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & \dots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} \end{bmatrix}, B = \begin{bmatrix} B_0 \\ B_1 \\ \vdots \\ B_k \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \; \hat{y}_i = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} \; \mathrm{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} \qquad \textbf{(3.6)}$$

3.4 Ridership Modeling Methodology

Based on these equations above and the independent variables identified, a multiple linear regression modeling methodology is developed as shown in the flow chart in Figure 3.4-1.

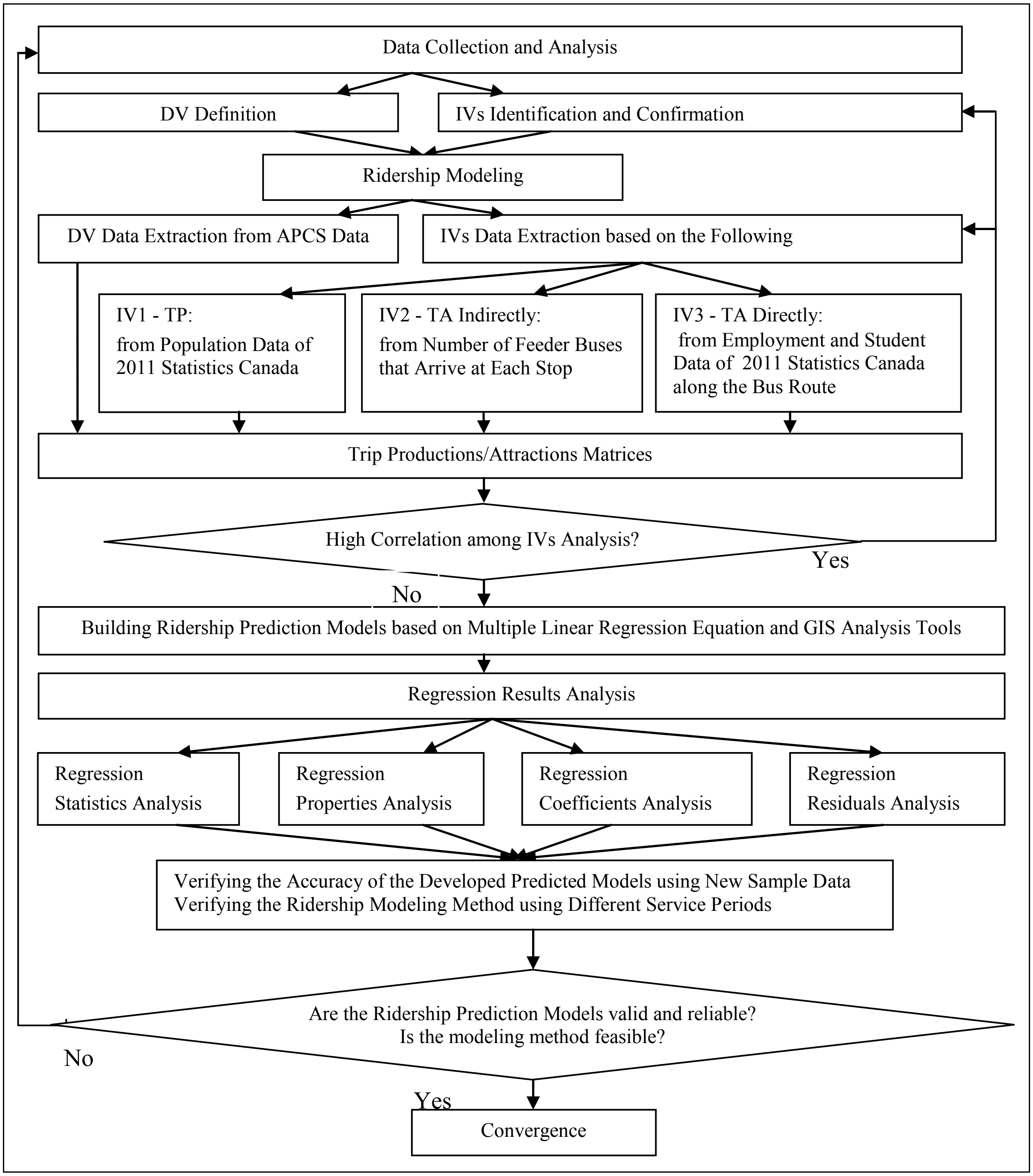


Figure 3.4-1 Ridership Modeling Methodology

Chapter 4 Ridership Modeling

Chapter 3 described case study area and data sources, identified dependent variable and key independent variables to model structure, presented multiple linear regression theory, and developed a ridership modeling methodology. Chapter 4 will process and present the ridership prediction models. First, dependent variable (DV) and independent variables (IVs) will be extracted and the correlation among IVs will be analyzed. Then, the matrices of average weekday trip production/attraction will be generated and the ridership prediction models will be developed based on the method of Least Squares Estimations.

4.1 Extraction of Dependent Variable from APCS Data

The dependent variable (DV) is defined as ridership (or average weekday boardings) in this thesis. The data preprocessing, extraction, and accuracy analysis of the DV will be stated in the following sections.

4.1.1 APCS Data Analysis and Preprocessing

APCS records consist of date, actual arrival time, schedule arrival time, schedule departure time, actual departure time, vehicle number, vehicle name, route name, route number, route direction, stop name, stop number, dwell time, schedule deviation, total in (boardings), total out (alightings), and total load. Each row of these data is recorded when a bus arrives at a stop. These data cannot be used directly and need to be processed properly.

Data analysis and preprocessing are very important because the quality of the data has a great impact on the accuracy of the prediction models. The count numbers of boardings and alightings in APCS records need to be checked to avoid observation errors. The observations with abnormal count numbers in the random sample patterns are defined as outliers in statistical theory. The outliers may skew the regression models; therefore, they should be identified and excluded from the sample data.

For example, based on the APCS data for iXpress, the capacity of one iXpress bus is 37. However, the three records in "In and Out" columns are found unreasonable at the Ainslie Terminal Stop, as shown by the following records:

Ainslie Terminal stop:

Downward:	September 13	14:05,	In: 202,	Out: 172
	October 12	8:04:36,	In: 505,	Out: 505
Upward:	October 12	14:16:00	In: 338,	Out: 338

The count numbers above are far beyond the actual capacity; therefore, these kinds of records will be identified and replaced with the actual capacity in the preprocessing. No outliers are found for Route 12 from the APCS records.

4.1.2 Extraction of Dependent Variable for iXpress 200

Figure 4.1-1 illustrates the stop name and number of iXpress Route 200 (iXpress). From the figure above, a total of 22 weekday APCS records for the 14 stops are available (from 2011-09-12 to 2011-10-12) for iXpress. The values of the DV for each stop will be calculated based on Equation (4.1).

Note that the first 17 weekday APCS records will be used to estimate the coefficients for the prediction model in this chapter. The remaining 5 weekday values will be conducted in Chapter 5 for model validation.

Figure 4.1-1 iXpress 200 Stop Name and Number

$$Boardings\_Weekday_{\_ki} = \sum_{j} IN_{\_jik} \qquad (4.1)$$

Where

IN: The 'IN' (or boardings) value in an APCS record of stop i and a given date

i: The stop number as shown in Figure 4.1-1

j: Arriving bus index

k: A given weekday

The calculation results for weekday boardings (September 12, 2011 - October 12, 2011) are presented in Table 4.1-1 and illustrated in Figure 4.1-2.

Table 4.1-1 iXpress 200 Weekday Boardings from September 12 to October 12, 2011

Stop No. / Date	1 (Conestoga Mall)	2 (McCormick)	3 (R &T Park)	4 (U Waterloo)	5 (Laurier)	6 (Uptown Waterloo)	7 (Grand River Hospital)	8 (Victoria)	9 (Charles Terminal)	10 (Ottawa)	11 (Fairview Mall)	12 (SmartCentre)	13 (Cambridge Centre)	14 (Ainslie Terminal)
Sep-12	1161	960	106	2507	775	579	291	120	1806	175	1393	297	683	909
Sep-13	1416	1186	152	2812	834	698	345	129	1822	189	1465	329	760	986
Sep-14	1108	1060	129	2584	807	635	348	173	1795	160	1412	306	733	987
Sep-15	1042	1132	114	2573	876	726	348	133	1856	180	1472	317	754	952
Sep-16	1305	1036	98	2534	900	744	295	176	1878	140	1445	331	726	910
Sep-19	856	1009	120	2191	713	588	272	129	1548	165	1101	270	711	861
Sep-20	1150	1113	143	2676	853	664	379	154	1837	160	1528	386	786	1056
Sep-21	889	1080	127	2370	757	616	325	124	1589	184	1266	300	743	973
Sep-22	1014	1006	105	2475	790	693	332	159	1807	187	1455	366	742	989
Sep-23	1145	994	127	2436	850	740	254	169	1837	118	1378	321	735	899
Sep-26	821	966	128	2121	696	590	339	154	1710	158	1287	322	660	915
Sep-27	1153	1032	159	2621	788	671	329	171	1686	146	1321	332	723	1006
Sep-28	793	1087	156	2482	789	702	386	181	1804	191	1449	325	789	948
Sep-29	927	1120	128	2471	837	659	378	173	1808	174	1397	312	750	983
Sep-30	1271	1050	116	2919	943	943	346	169	2164	187	1628	407	954	1171
Oct-03	738	983	119	2065	663	585	268	137	1610	173	1267	313	710	844
Oct-04	1039	905	123	2384	693	556	286	155	1524	145	1205	291	644	1060
Oct-05	769	889	97	1991	647	555	339	172	1579	171	1251	291	713	984
Oct-06	815	996	102	2189	766	632	326	163	1658	166	1190	298	663	854
Oct-07	934	905	109	2410	904	654	321	144	1767	176	1451	359	624	815
Oct-11	977	968	120	2345	694	629	308	148	1783	147	1313	296	715	983
Oct-12	752	1025	151	2176	672	537	322	136	1582	158	1209	286	654	916

Note: The calculation results include both directions (upward and downward) of the bus route.

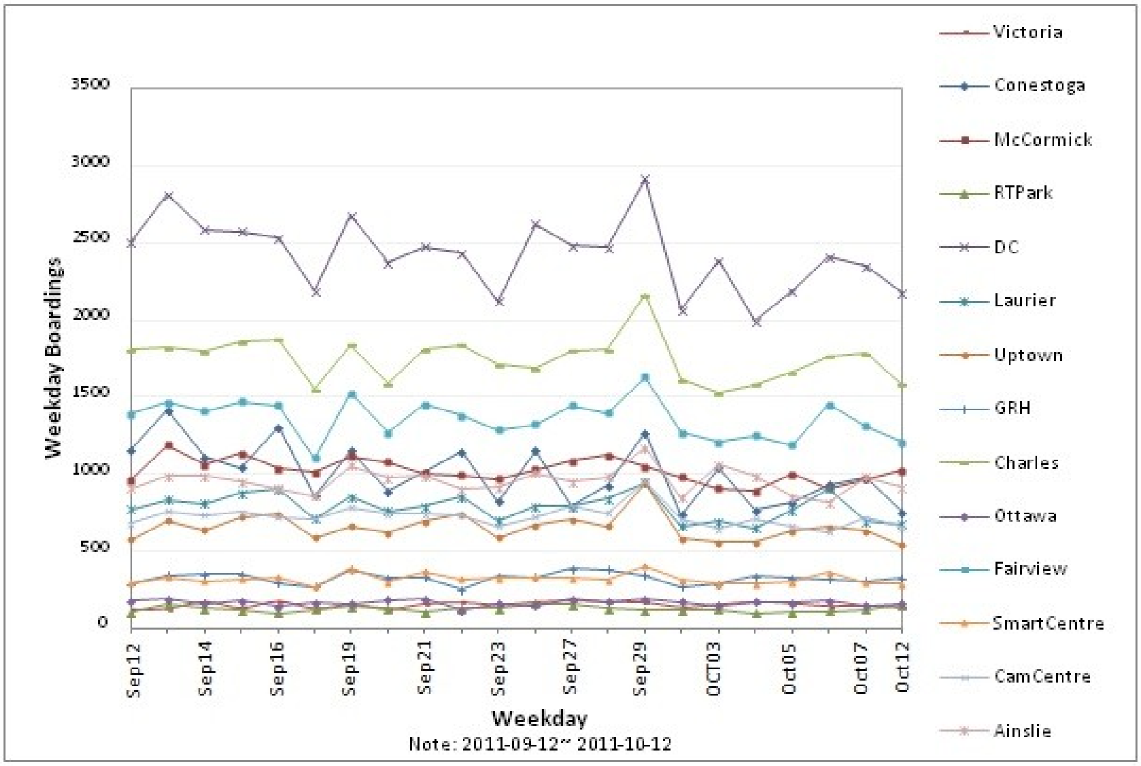


Figure 4.1-2 iXpress 200 Weekday Boardings

Figure 4.1-2 presents the iXpress weekday boardings distribution from September 12, 2011 to October 12, 2011. As shown in the figure above, there are big weekday boardings fluctuations at Davis Centre, Charles terminal, Fairview Mall, Conestoga Mall, and Ainslie terminal stops.

To improve the accuracy of the developed model, the above data need to be processed further so that the average weekday boardings are within the confidence interval (CI). The most common confidence level of 95% will be used in the coefficient estimations in the regression models. In order to calculate CI, average weekday boardings will be calculated as the next step. In order to help ridership analysis in the chapter, average weekday alightings and bus loads will also be calculated at the same time.

4.1.2.1 Average Weekday Boardings, Alightings, and Loads

Average weekday boardings, alightings, and loads for each stop during the first 17 weekdays can be calculated based on the sample mean formula as follows.

$$\boldsymbol{Boardings\_Average\_Weekday_{\_i}} = \frac{1}{17\ weekdays}\sum_{k=Sep.12}^{k=Oct.4} \boldsymbol{Boardings\_Weekday_{\_ki}} \quad \textbf{(4.2)}$$

$$\boldsymbol{Alightings\_Average\_Weekday_{\_i}} = \frac{1}{17\ weekdays}\sum_{k=Sep.12}^{k=Oct.4} \boldsymbol{Alightings\_Weekday_{\_ki}} \quad \textbf{(4.3)}$$

$$\boldsymbol{Load\_Average\_Weekday_{\_i}} = \frac{1}{17\ weekdays}\sum_{k=Sep.12}^{k=Oct.4} \boldsymbol{Load\_Weekday_{\_ki}} \quad \textbf{(4.4)}$$

Where,

i: Stop number

k: A given weekday

$Boardings\_Weekday_{\_ki}$ is calculated by Equation (4.1).

$Alightings\_Weekday_{\_ki}$ is calculated by the same way as $Boardings\_Weekday_{\_ki}$, but using the 'out' field of the APCS records.

$Load\_Weekday_{\_ki}$ is calculated by the same method as $Boardings\_Weekday_{\_ki}$, but using the 'load' field of the APCS records.

Based on Equations (4.2), (4.3), (4.4), the calculation results are shown in Table 4.1-2 and illustrated in Figure 4.1-3.

Table 4.1-2 iXpress 200 Average Weekday Ridership

Stop No.	Boarding	Alighting	Load
1 (Conestoga Mall)	1049	1188	1020
2 (Mccormick)	1042	909	2947
3 (R & T Park)	126	140	3664
4 (U Waterloo)	2484	2279	3841
5 (Laurier)	798	918	3811
6 (Uptown Waterloo)	670	732	3925
7 (Grand River Hospital)	325	304	3874
8 (Victoria)	153	140	3716
9 (Charles Terminal)	1769	1674	3626
10 (Ottawa)	167	158	3537
11 (Fairview Mall)	1381	1487	2677
12 (Smart Centre)	325	360	1848
13 (Cambridge Centre)	741	881	1672
14 (Ainslie Terminal)	968	874	945

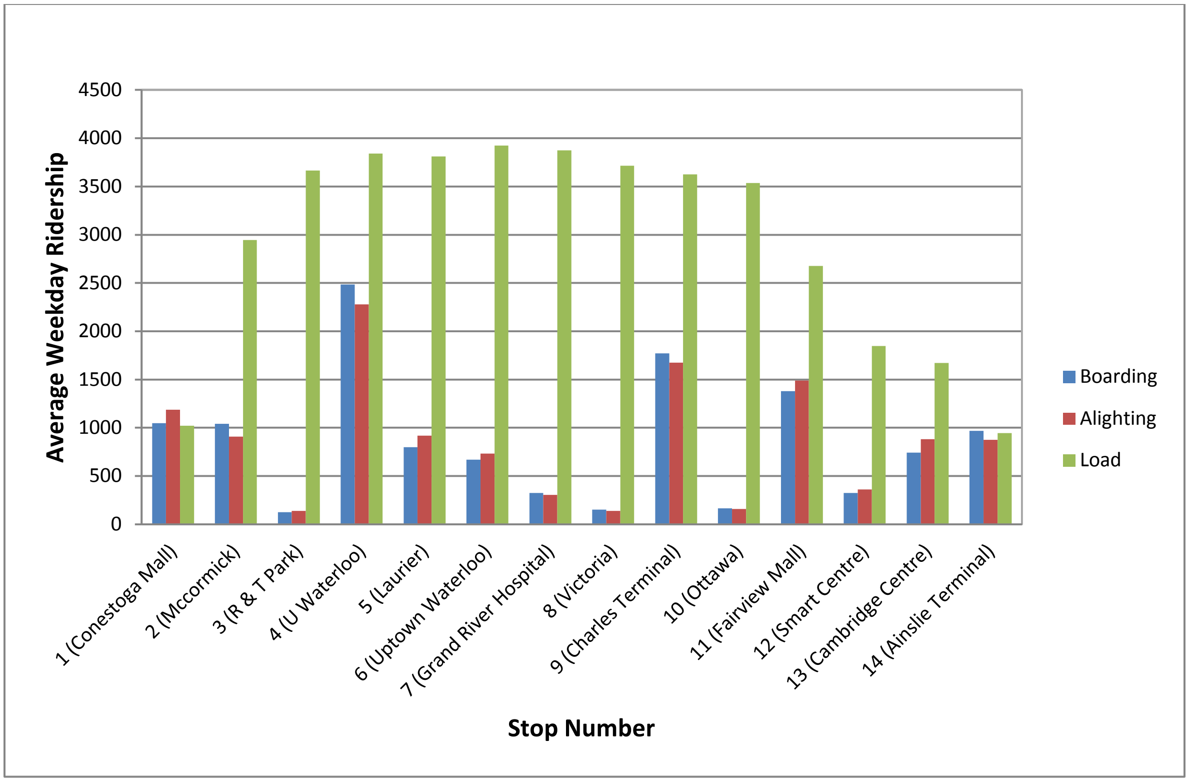


Figure 4.1-3 iXpress 200 Average Weekday Ridership

4.1.2.2 Margin of Error and Confidence Interval

The confidence level is set to 95%, meaning that 95% of the average weekday boardings should be within the Confidence Interval (CI). Mathematically, the CI equals the sample average plus or minus the margin of error.

Since the sample size n is 17 (weekdays), and less than 30, a t distribution is used to find the margin of error, which is calculated based on standard error.

All mathematical equations used in CI calculation are listed as follows:

Degree of freedom: $\boldsymbol{df = n - 1}$ **(4.5)**

Sample standard deviation: $\boldsymbol{s = \sqrt{\frac{\sum_{i=1}^{n}(x-\bar{x})^2}{n-1}}}$ **(4.6)**

Sample standard error: $\mathbf{Se} = \boldsymbol{\frac{s}{\sqrt{n}}}$ **(4.7)**

Margin of error: $\mathbf{Me} = \mathbf{t}_{\frac{\alpha}{2}} * \mathbf{Se}$ **(4.8)**

Confidence interval: $\boldsymbol{(1-\alpha)\% = \bar{x} \pm \mathbf{Me}}$ **(4.9)**

Where

α: Probability that the parameter is NOT within the specified interval

$t_{\frac{\alpha}{2}}$: 2.12 (From the t table, when df = 16 and confidence interval = 95%)

n: Sample size (n = 17)

s: Standard deviation

x: Samples (i.e. Weekday Boardings in Table 4.1-1)

$\boldsymbol{\bar{x}}$**:** Sample mean (i.e. Average weekday boardings in Table 4.1-2)

Se: Sample standard error

Me: Margin of error

Based on Equations (4.2), (4.5), (4.6), (4.7), (4.8), (4.9), the calculation results are shown in Table 4.1-3.

Table 4.1-3 iXpress 200 Margin of Error and Confidence Interval (CI)

Stop no.	$\overline{x}$	S	Se	Me	Me %	CI (Lower95%)	CI (Upper95%)	λ
1 (Conestoga)	1049	192	41	85	8.27%	962	1135	60.70%
2 (McCormick)	1042	72	15	32	3.13%	1010	1075	71.95%
3 (RTPark)	126	18	4	8	6.27%	119	134	75.00%
4 (U Waterloo)	2484	223	47	99	4.05%	2383	2584	42.60%
5 (Laurier)	798	77	16	34	4.38%	763	833	49.70%
6 (Uptown Waterloo)	670	92	20	41	6.23%	628	712	66.65%
7 (Grand River Hospital)	325	41	9	18	5.64%	306	343	54.70%
8 (Victoria)	153	21	4	9	6.05%	144	163	114.40%
9 (Charles Terminal)	1769	152	32	67	3.89%	1701	1838	40.20%
10 (Ottawa)	167	20	4	9	5.55%	157	176	58.95%
11 (Fairview Mall)	1381	128	27	57	4.19%	1323	1438	40.70%
12 (Smart Centre)	325	34	7	15	4.74%	310	340	71.45%
13 (Cambridge Centre)	741	67	14	30	4.09%	711	772	45.70%
14 (Ainslie Terminal)	968	80	17	35	3.73%	931	1004	33.90%

λ : The dispersion degree of the sample data, λ = (maximum of 17 weekdays boardings - minimum of 17 weekdays boardings)/average of 17 weekdays boardings

According to US National Transit Database (NTD) standard, the margin of error is required to 10% precision when using a 95% confidence level (Metropolitan, Council, May 15, 2007). Based on the above calculation, the margin of error rates for all 14 stations are less than 10% and meet NTD's requirement, and also less than the dispersion degree of the sample data at each stop, respectively. For easy comparison, the margins of errors and the related dispersion degrees are illustrated in Figure 4.1-4.

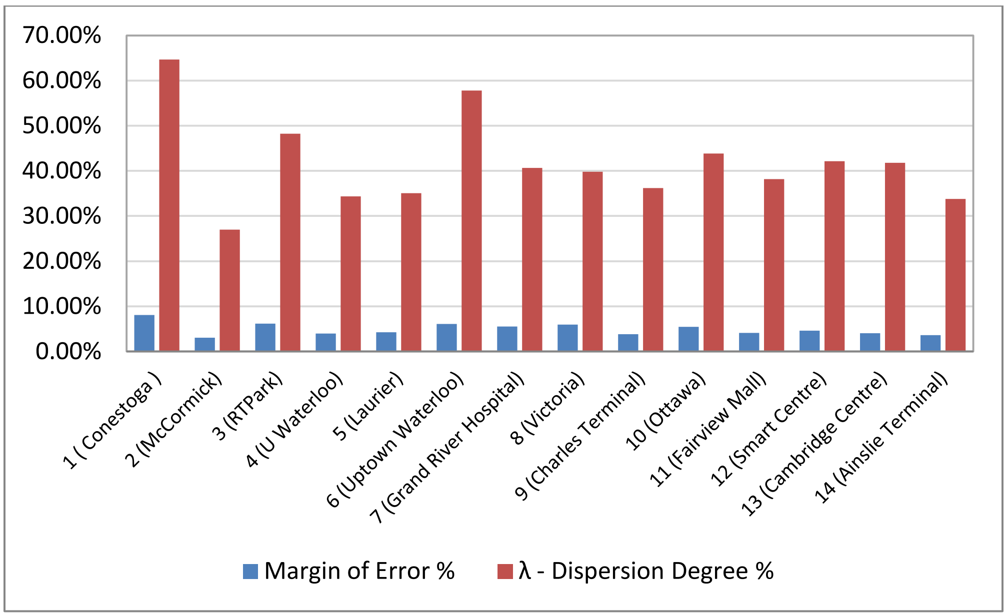


Figure 4.1-4 iXpress 200 Sample Data Dispersion Degree versus Margin of Error

As shown in the above figure, the sample data margins of error are far less than the related dispersion degrees, meaning that the APCS records are accurate with a 95% confidence level.

4.1.3 Extraction of Dependent Variable for Route 12

Route 12 is a conventional bus service with a total length of 25 kilometers and 139 stops. 22 weekdays of APCS data are collected for Route 12 in November 2012. In this thesis, sixteen stops are randomly selected to build a prediction model for Route 12 (see Figure 4.1-5).

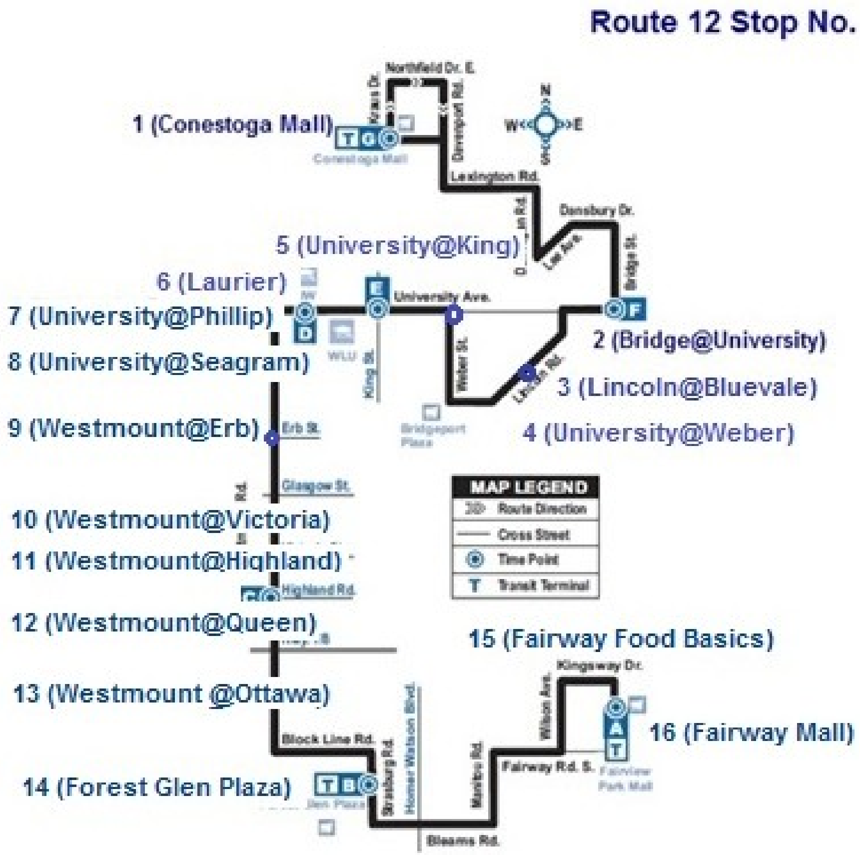


Figure 4.1-5 Route 12 Stop Name and Number

From Figure 4.1-5, the values of the DV for the sixteen stops of Route 12 will be also calculated based on Equation (4.1). Note that the first 17 weekday APCS records will be used to calculate the coefficients for the prediction model of Route 12 in this chapter. The remaining 5 weekday values will be used in Chapter 5 for model validation purposes.

For Route 12, some stop distances are very close, such as 186m between the Laurier and University@King Stops, and 447 meters between University@King and University@Weber stops. Therefore, the segment-based buffers are introduced to solve overlapping area double counting caused by close-by stop buffering in modeling. Univesity@Weber, University@King, and Laurier stops are combined to one segment called S-Laurier; Univesity@Phillip and Univesity@Seagram are combined to one segment called S-U Waterloo; and Westmount@Victoria, Westmount@Highland, and Westmount@Queen are combined to one segment called S-Westmount@Highland. Weekday boardings in November 2012 are calculated for the three segments and the eight stops. The calculation results are presented in Table 4.1.4 and illustrated in Figure 4.1.6, respectively.

Table 4.1-4 Route 12 Weekday Boardings in November 2012

Date \ Stop No.	1 (Conestoga Terminal)	2 (Bridge/ University)	3 (Lincoln/ Bluevale)	4 (S-Laurier)	5(S-U Waterloo)	6 (Westmount/ Erb)	7(S-Westmount/ Highland)	8 (Westmount/ Ottawa)	9 (Forest Glen Terminal)	10 (Fairway/ Food Basics)	11 (Fairview Terminal)
Nov-01	250	68	154	594	762	116	360	141	533	79	565
Nov-02	381	111	191	736	823	157	494	176	703	128	764
Nov-05	234	94	163	580	848	109	442	144	583	108	603
Nov-06	236	72	145	469	657	123	320	124	508	60	486
Nov-07	244	106	225	698	925	130	451	163	664	107	654
Nov-08	356	104	191	799	1007	112	438	191	721	106	730
Nov-09	370	102	165	598	701	160	365	115	592	124	667
Nov-12	212	51	140	446	779	120	316	113	497	90	546
Nov-13	261	85	197	564	799	79	419	142	584	87	546
Nov-14	272	99	204	685	888	47	486	174	596	117	681
Nov-15	270	95	183	639	840	34	429	134	611	120	699
Nov-16	521	106	186	719	799	45	512	183	741	152	847
Nov-19	210	76	151	522	851	21	413	137	599	101	605
Nov-20	286	77	192	590	810	26	397	165	590	93	624
Nov-21	326	111	201	726	973	60	494	158	717	121	741
Nov-22	247	76	141	538	660	80	351	114	501	79	525
Nov-23	601	126	200	743	713	109	443	184	694	157	812
Nov-26	259	104	188	664	931	135	407	165	619	97	638
Nov-27	343	80	170	620	824	78	389	122	520	103	591
Nov-28	263	105	201	625	900	109	458	163	633	111	698
Nov-29	253	87	205	667	830	117	412	142	576	104	583
Nov-30	367	95	166	587	706	133	405	166	639	100	785

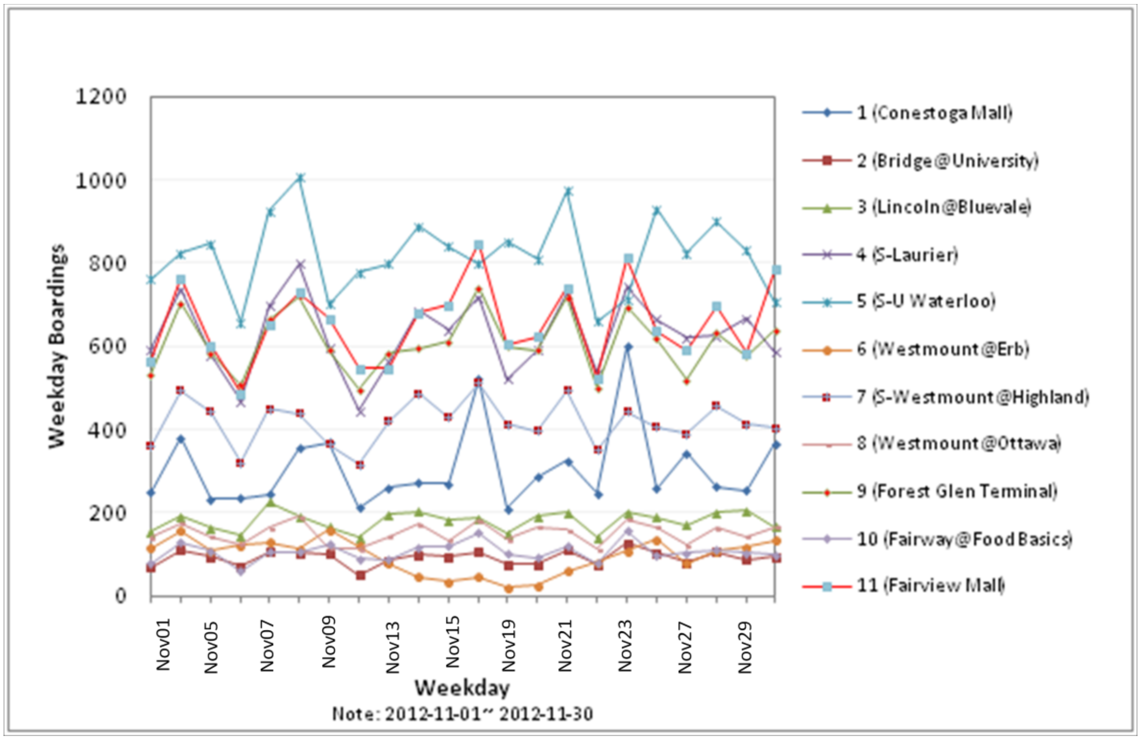


Figure 4.1-6 Route 12 Weekday Boardings

As can been seen from the figure above, the Conestoga terminal has a similar boardings distribution pattern to that of the Fairview Park Terminal. Basically, the two shopping malls have the highest boardings on every Friday and the lowest boardings on every Monday. This distribution pattern reveals riders' behaviors of high shopping activities in late weekdays and low shopping activities in early weekdays. The two universities have the same distribution pattern due to the students' activities; The Forest Glen Terminal has similar fluctuations to the two shopping malls, but less variance range. The rest stops have the smallest fluctuations due to very low boardings.

Similar to the iXpress data analysis, in order to improve the accuracy of the developed model, the data will be processed further so that the average weekday boardings are within the confidence interval (CI). The most common confidence level of 95% will be used for Route 12 as well.

4.1.3.1 Average Weekday Boardings, Alightings, and Load

In order to calculate CI, average weekday boardings will be calculated in this section. To help with ridership analysis in later sections, average weekday alightings and loads will also be calculated at the same time. Based on Equations (4.2), (4.3), (4.4), the calculation results are shown in Table 4.1-5 and Figure 4.1-7, respectively.

Table 4.1-5 Route 12 Average Weekday Ridership

Stop_Name	Average Boardings	Average Alightings	Average Load
1 (Conestoga Terminal)	310	150	242
2 (Bridge/University)	92	90	530
3 (Lincoln/Bluevale)	178	164	721
4 (S-Laurier)	626	581	1245
5 (S-U Waterloo)	814	854	2261
6 (Westmount/Erb)	90	182	1079
7 (S-Westmount/Highland)	419	419	2504
8 (Westmount/Ottawa)	150	146	1082
9 (Forest Glen Terminal)	614	603	760
10 (Fairway/Food Basics)	108	130	860
11 (Fairview Terminal)	653	181	466

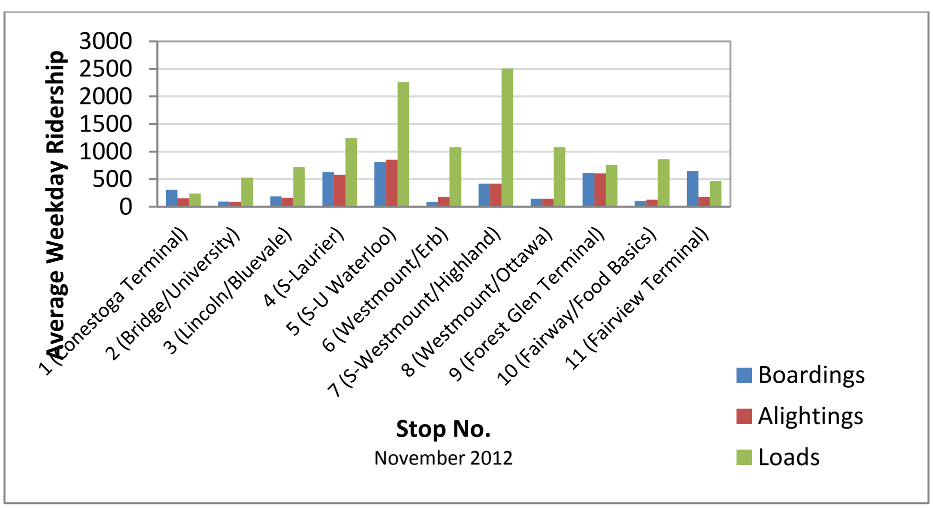


Figure 4.1-7 Route 12 Average Weekday Ridership

4.1.3.2 Margin of Error and Confidence Interval

Route 12 has the same sample size as that of iXpress. Therefore, the same method is used for Route 12 to calculate the Margin of Error and Confidence Interval. Based on Equations (4.2), (4.5), (4.6), (4.7), (4.8), (4.9), the calculation results are shown in Table 4.1-6.

Table 4.1-6 Route 12 Margin of Error and Confidence Interval (CI)

Stop_No	$\bar{x}$	S	Se	Me	Me%	CI_Lower95%	CI_Upper95%	λ
1 (Conestoga Terminal)	310	109	26	56	18.00%	255	366	125.96%
2 (Bridge/University)	92	19	5	10	10.87%	82	102	81.78%
3 (Lincoln/Bluevale)	178	26	6	13	7.39%	165	191	47.71%
4 (S-Laurier)	626	102	25	53	8.41%	574	679	56.37%
5 (S-U Waterloo)	814	100	24	51	6.30%	763	865	43.01%
6 (Westmount/Erb)	90	45	11	23	25.52%	67	113	154.65%
7 (S-Westmount/Highland)	419	61	15	31	7.44%	388	451	46.73%
8 (Westmount/Ottawa)	150	26	6	13	8.88%	137	164	51.84%
9 (Forest Glen Terminal)	614	80	19	41	6.71%	573	655	39.75%
10 (Fairway/Food Basics)	108	26	6	13	12.19%	94	121	90.16%
11 (Fairview Terminal)	653	104	25	53	8.18%	599	706	55.31%

λ : The dispersion degree of the sample data, it is calculated using the formula below:
λ = (maximum of 17 weekdays boardings - minimum of 17 weekdays boardings)/average of 17 weekdays boardings

From the table above, the margin of error of the Conestoga Terminal is 18.00%, which is beyond the NTD's requirement of 10% (Metropolitan, Council, May 15, 2007). A close look at the APCS data reveals that there are boarding peaks on November 16 and November 23, perhaps special events occurred in the two days (Wednesday), which caused the large dispersion of 125.96% in the sample data. The solution is to check the standard residual after the regression processing. If the standard residual of the bus stop is greater than 3, it could be a potential outlier (Fotheringham, et al., 2002), then the boarding data on the two days can be replaced by the average boarding data on the other Wednesdays. Other stops such as Bridge/University, Westmount/Erb, and Fairway/Food Basics are analyzed and processed by the same method. For easy comparison, the margins of errors and the related dispersion degrees are illustrated in Figure 4.1-8.

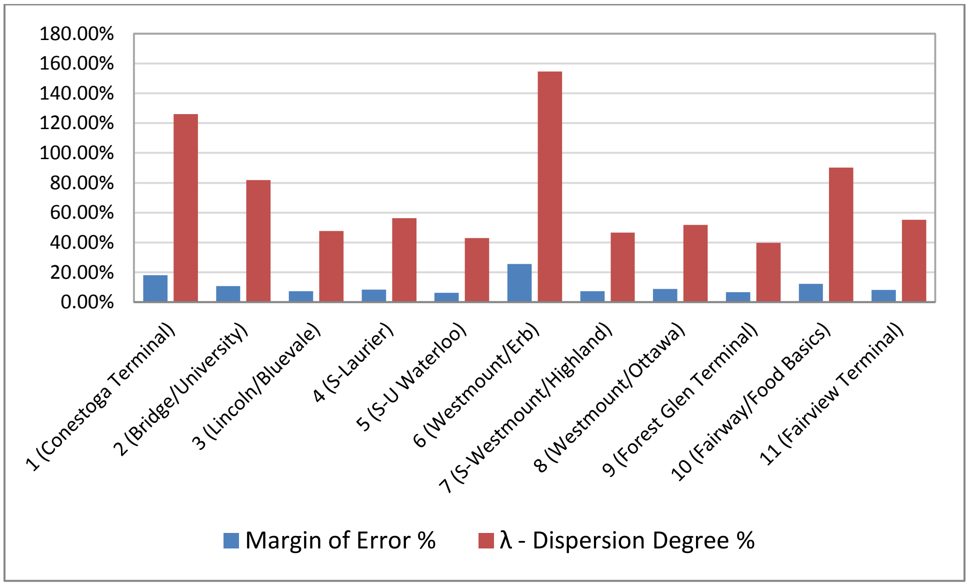


Figure 4.1-8   Route 12 Sample Data Dispersion Degree versus Margin of Error

As shown in Figure 4.1-8, the sample data margins of error are far less than the related dispersion degrees, meaning that the APCS records are accurate with a 95% confidence level for Route 12.

4.2 Extraction of Independent Variables

Section 4.1 analyzed and extracted the dependent variables for the regression models of iXpress and Route 12. The following explains how to obtain the three independent variables (IVs).

4.2.1 Extraction of the Independent Variable1 (IV1)

IV1 is defined as population (residents) within the stop-based buffer area. It can be called as Trip Production (TP). The Davis Centre stop at the University of Waterloo is used as an example for explaining the definition of IV1 (see Figure 4.2-1).

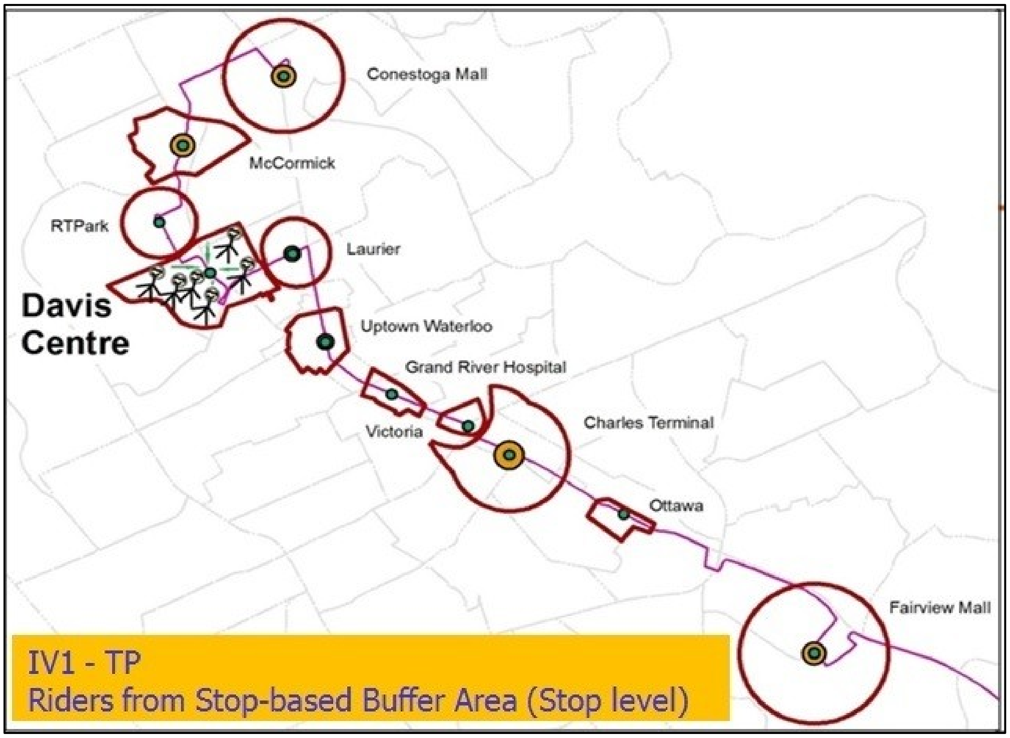


Figure 4.2-1 IV1 - Riders from Stop-based Buffer Area (Stop Level)

Population data are collected from the census tract boundary of 2011 Statistics Canada Household Survey. These data need to be processed to stop-based buffer area using Area-based Fraction Equation and GIS analysis tools. In order to estimate the regression coefficients more efficiently and accurately, different buffer widths are repeatedly tested for different stops so that the extracted data are within the effective service area. Spatial Proximity Method (SPM) and Spatial Weight Method (SWM) are also applied to improve the accuracy of the data extraction. The extraction steps of IV1 are illustrated in Figure 4.2-2.

Step 1 The Selection of Stop-based Effective Service Buffer Area for Population Data

↓

Step 2 The Assignment Method of Weight Coefficients

↓

Step 3 The Extraction Equations of Population Data

↓

Step 4 Overlay Analysis (Conestoga Mall stop as an example)

Buffer areas are overlaid with population data within the Census Tract Boundary.

↓

Step 5 Intersect Analysis (Conestoga Mall stop as an example)

↓

Step 6 Sum Analysis

The small zonal pieces with their attribute data are summed for each stop.

↓

Step 7 Export

The summed results are exported from the GIS system and then are imported to Excel as IV1 for correlation and regression analysis.

Note: The zonal pieces are defined within stop-based buffer area

Figure 4.2-2 The Extraction Steps of IV1

Step 1 The Selection of Stop-based Effective Service Buffer Area for Population Data

EI-Geneidy et al. (2014) determind that "The 85th percentile walking distance to bus transit effective service is around 524 meters from home-based trip origins". However, for different stop types such as shopping mall, bus terminal, and university stops, the effective service buffer area may be different; moreover, for different bus route types such as iXpress 200 and conventional bus route 12, the effective service buffer area may be different as well. Based on simple linear regression method, the relationship between average weekday boardings and residents at stop level can be illustrated as shown in Figure 4.2-3.

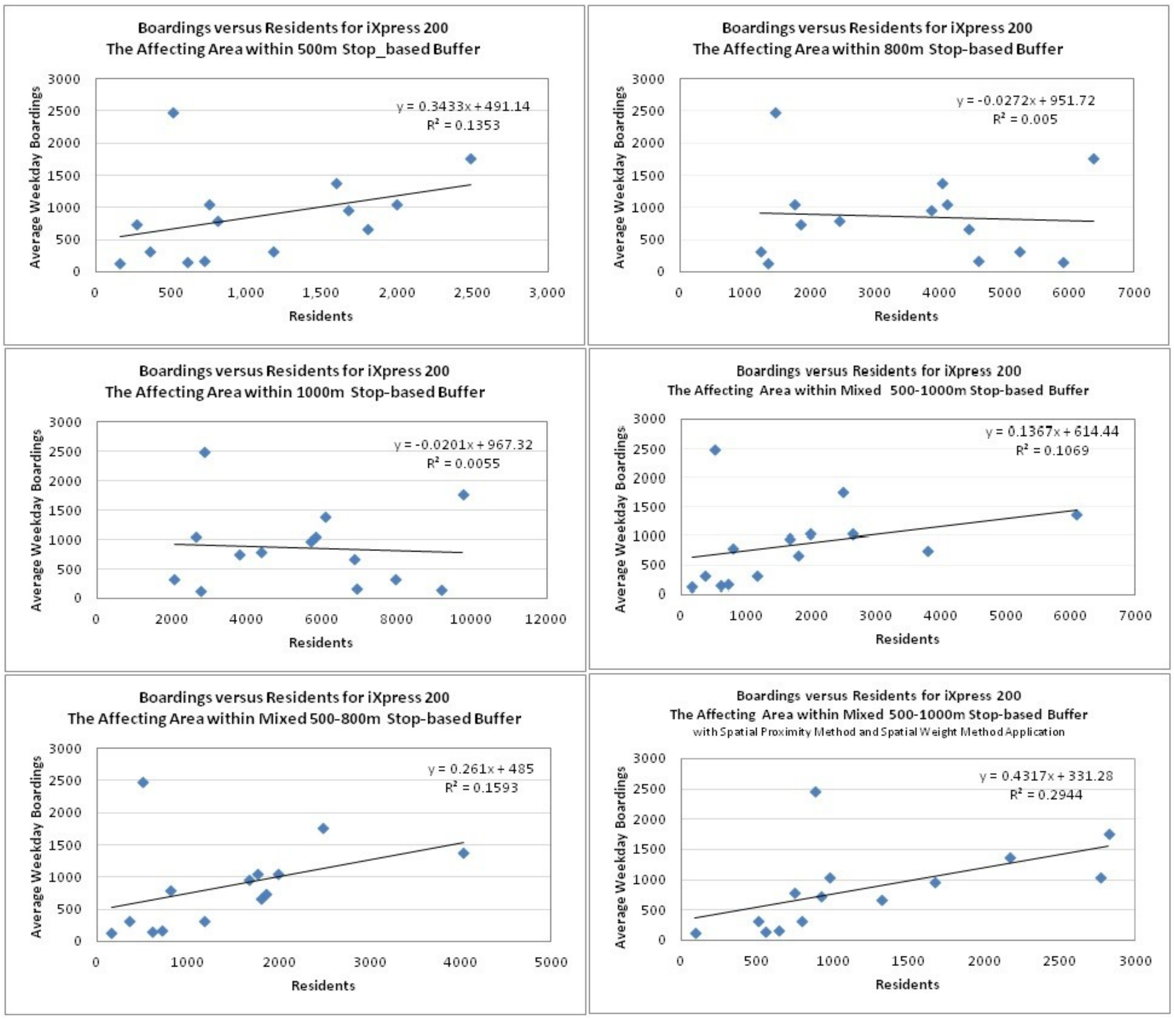


Figure 4.2-3 iXpress 200 - Boardings versus Residents within Stop-based Buffer Area

From Figure 4.2-3, it is found that the highest effective service areas (boarding-affecting buffer area) are within the mixed 500-1000m stop-based buffer with SPM and SWM application. The R square is 0.294. The detailed effective service areas for the 14 stops are summarized in the Table 4.2-1.

Table 4.2-1 iXpress 200 - The Selection of Effective Service Buffers for 14 Stops

Stop_Name	**IV1 - Residents (Trip Production)** within Effective Service Buffer Area	Description
1 (Conestoga Mall)	Stop-based 1000m buffer with SPM and SWM application	Mixed land use in residential, industrial, and commercial area - shopping mall
2 (McCormick)	Stop-based buffer area around residential area boundary	Land use in high density residential area
3 (R & T Park)	Stop-based 500m buffer area	Land use in business park
4 (U Waterloo)	Stop-based buffer area around campus boundary, Census Tract Boundary, and road boundary	Mixed land use in education and residential area
5 (Laurier)	Stop-based 500m buffer area	Mixed land use in education, residential, and retail area
6 (Uptown Waterloo)	Stop-based buffer area around Census Tract Boundary and road boundary	Mixed land use in residential, commercial, and business
7 (Grand River Hospital)	Stop-based buffer area around road boundaries	Mixed land use in hospital and business
8 (Victoria)	Stop-based buffer area are within the northwest of the stop due to bus terminal in the southeast of the stop	Mixed land use in education and commercial area
9 (Charles Terminal)	Stop-based 800m buffer with SPM and SWM application	Mixed land use in residential, government, and commercial area
10 (Ottawa)	Stop-based buffer in northwest-southeast direction around the road boundaries	Mixed land use in residential, old industrial, and commercial area
11 (Fairview Mall)	Stop-based 1000m buffer area with SPM and SWM application due to high density low income residents living at the north of the stop	Mixed land use in residential, commercial, industrial, and highway area
12 (Smart Centre)	Stop-based 800m buffer area with SPM and SWM application	Mixed land use in highway area and commercial area - shopping mall
13 (Cambridge Centre)	Stop-based 1000m buffer area with SPM and SWM application due to big shopping mall attraction	Mixed land use in conservation, residential, and commercial area - shopping mall
14 (Ainslie Terminal)	Stop-based 500m buffer area	Mixed land use in residential and commercial area

Source: Population data (Residents) from 2011 Statistics Canada; Boardings and land use data from the Region of Waterloo
Note: SPM: Spatial Proximity Method; SWM: Spatial Weight Method

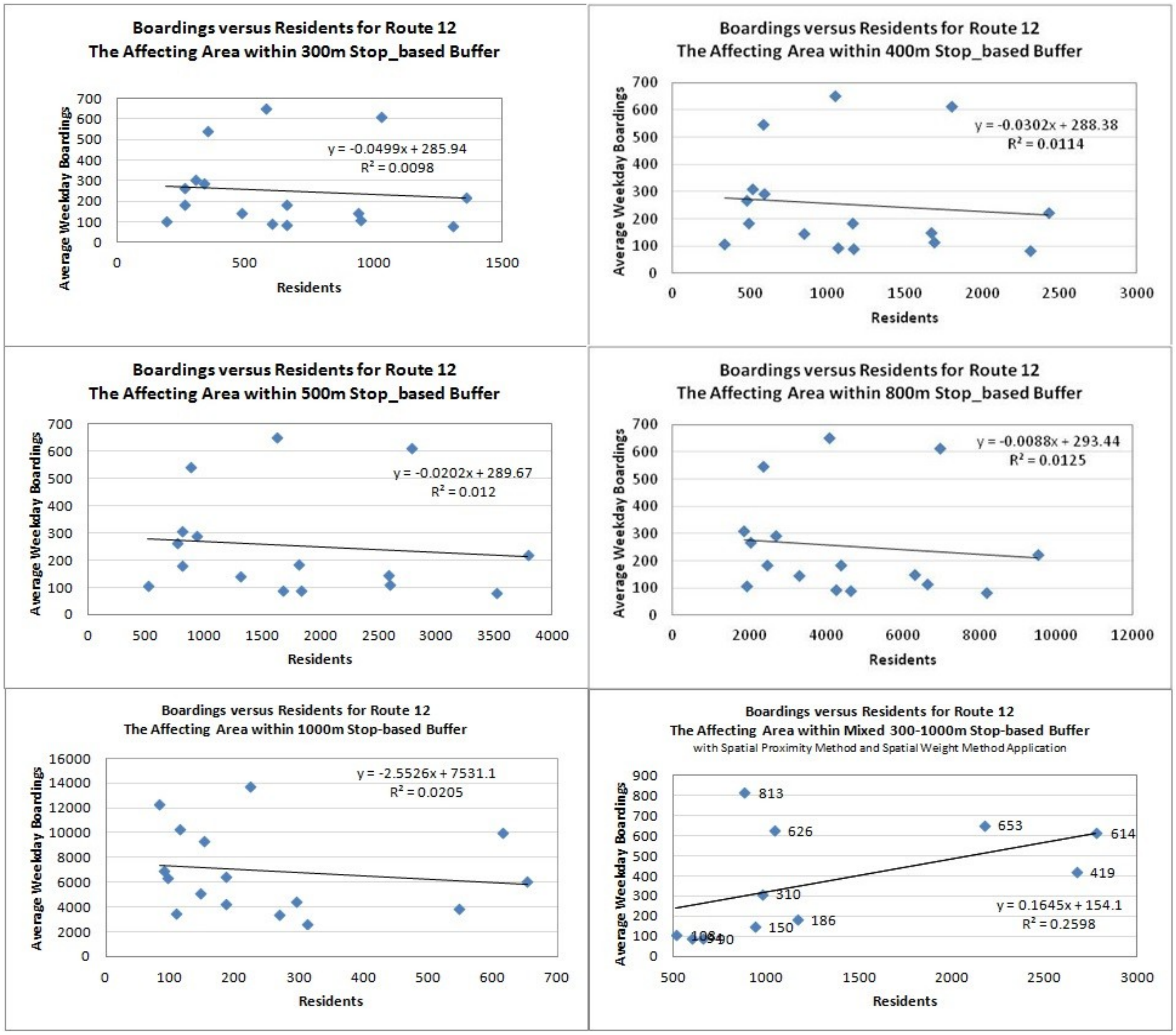


Figure 4.2-4 Route 12 - Boardings versus Residents within Stop-based Buffer Area

From Figure 4.2-4, it is found that the highest effective service buffer areas are within the mixed 300-1000m stop-based buffer with SPM and SWM application. The R square is 0.2598. The detailed effective service buffer areas for the 11 stops are summarized in Table 4.2-2.

Table 4.2-2 Route 12 - The Selection of Effective Service Buffers for 8 Stops and 3 Segments

Stop No.	**IV1 - Residents (Trip Production)** within Effective Service Buffer Area	**Description**
1 (Conestoga Mall)	Stop-based 1000m buffer area with SPM and SWM application	Mixed land use in residential, industrial, and commercial area - shopping mall
2 (Bridge@University)	Stop-based 300m buffer area	Mixed land use in residential, retail, and cemetery
3 (LincolnBluevale)	Stop-based 400m buffer area	Mixed land use in residential, school, retail, and highway
4 (S-Laurier)	Merge University@Weber, University@King, and Laurier stops. 633m Segment-based 300m buffer (from University@Weber to Laurier)	Mixed land use in residential, retails, colleglate, and university
5 (S-U Waterloo)	Merge University@Philip and University@Seagram stops. Campus boundary with road boundary - based polygon buffer	Mixed land use in residential, retails, and university
6 (Westmount@Erb)	Stop-based 300m buffer area	Mixed land use in residential and commercial area
7 (S-Westmount@Highland)	Merge Westmount@Victoria, Westmount@Highland, and Westmount@Queen stops. 1124m segment - based 300m buffer area	Mixed land use in high-density residential, schools, and retails
8 (Westmount@Ottawa)	Stop-based 300m buffer area	Mixed land use in residential, highway, school, and retails
9 (Forest Glen Terminal)	Stop-based 500 m buffer area	Mixed land use in residential, highway, school, park, and retails
10 (Fairway@Food basic)	Stop-based 500 m buffer area	Mixed land use in residential, railway and retails
11 (Fairview Mall)	Stop-based 1000 m buffer area with SPM and SWM application	Mixed land use in high-density low-income residential, highway, school, and big shopping mall

Source: Population data (Residents) from 2011 Statistics Canada; land use data from the Region of Waterloo
Note: SPM: Spatial Proximity Method; SWM: Spatial Weight Method

To summarize, the regression results show that the transit effective service area at stop level for iXpress 200 are greater than that of Route 12, meaning that iXpress 200 attracts more riders. The effective service area selection for iXpress 200 and Route 12 are geographically illustrated in Figure 4.2-5.

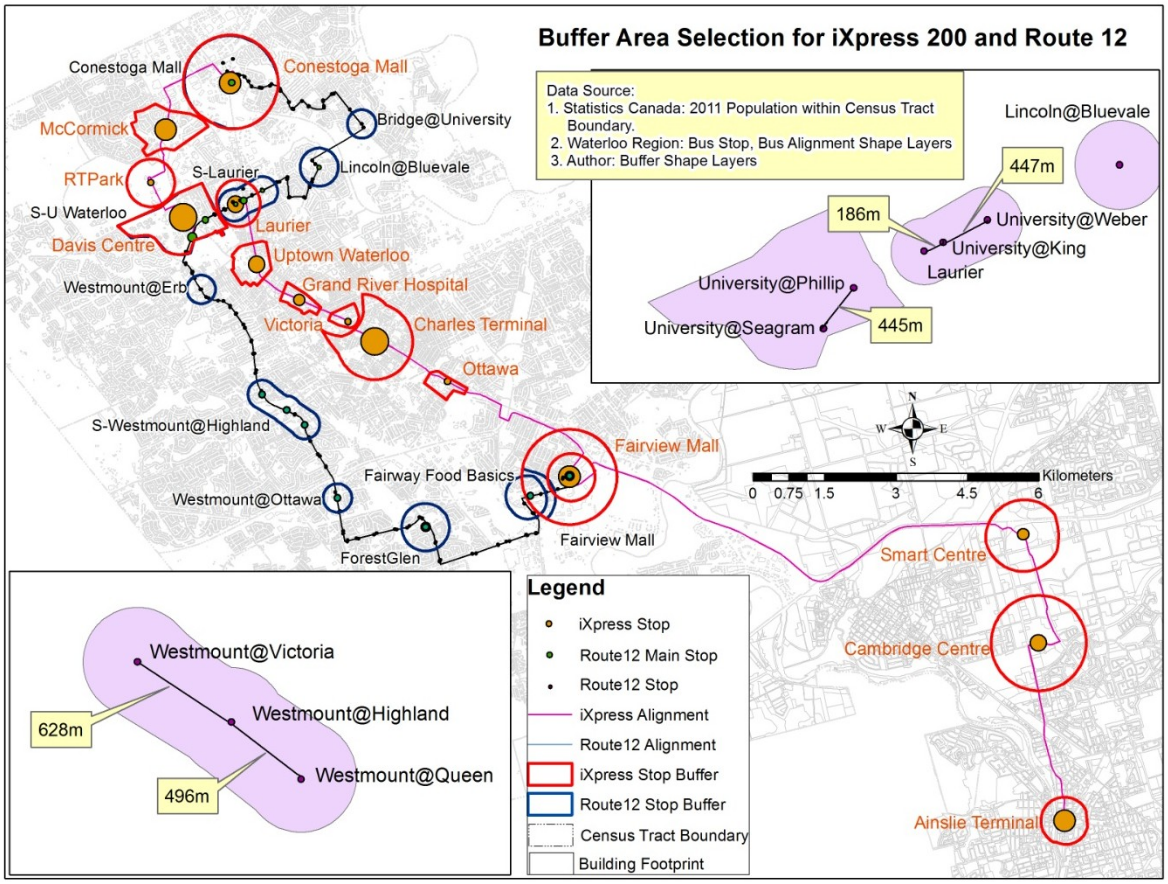


Figure 4.2-5 Buffer Area Selection for iXpress 200 and Route 12

Step 2 The Assignment Method of Weight Coefficients

Weight coefficients are considered in the stop-based effective service area for iXpress 200 and Route 12. They are assigned and illustrated in Figure 4.2-6.

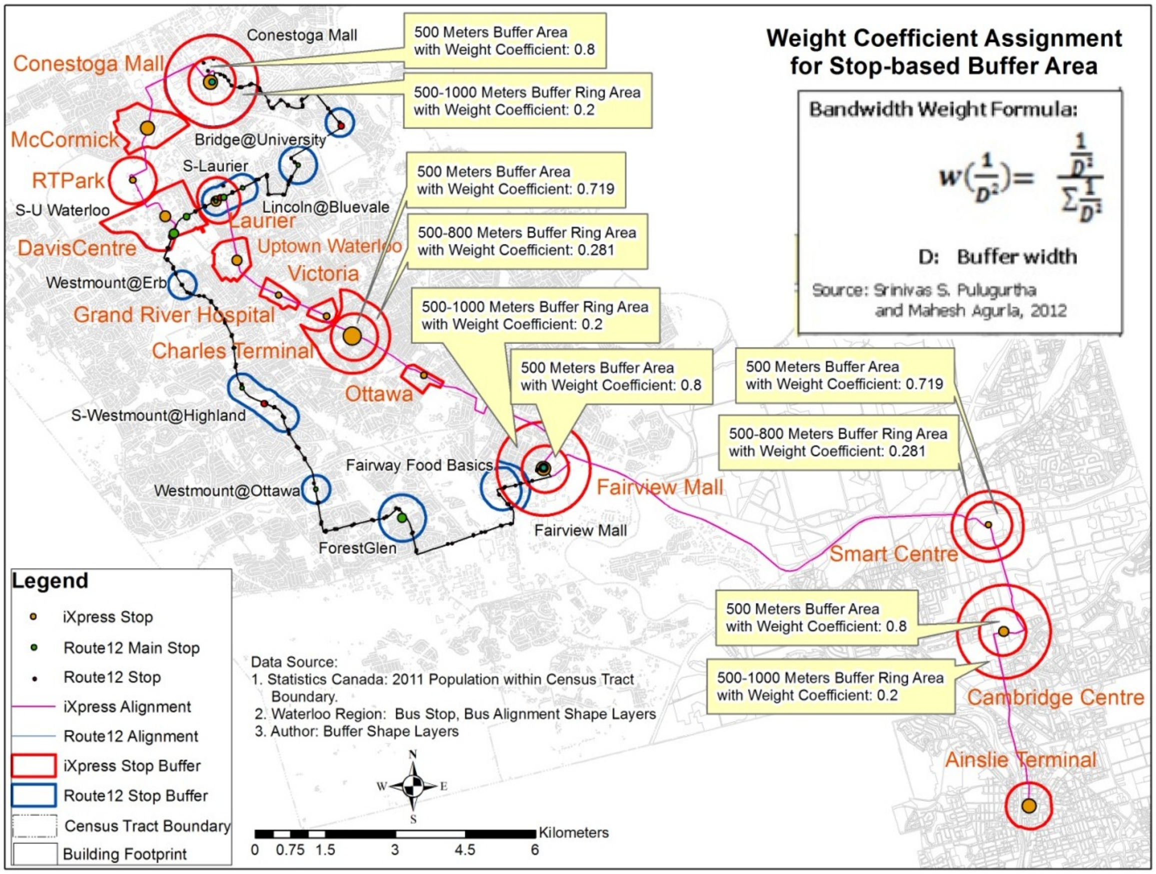


Figure 4.2-6 Weight Coefficient Assignment for Stop-based Buffer Area

In order to extract accurate IV1 data, the Spatial Proximity Method (SPM) is used in the boarding affecting area data extraction (Pulugurtha & Agurla, 2012). Based on a simple linear regression method, the effective service buffer area for iXpress 200 and Route 12 are illustrated in Figure 4.2-5 and stated in Table 4.2-1 and Table 4.2-2.

The Spatial Weight Method (SWM) is also used for the extraction of the population data at the four shopping malls - Conestoga Mall, Fairview Mall, Smart Centre, Cambridge Centre, and GRT bus terminal (Charles Terminal). Based on SPM models for distance decay behavior, 500m buffer areas are assigned a relatively heavy weight coefficient; 500-800m and 500-1000m buffer ring areas are assigned a relatively light weight coefficient. The bandwidth weight formula for the weight coefficient calculation is mathematically expressed as the following equation (Pulugurtha & Agurla, 2012):

Bandwidth Weight: $$\boldsymbol{w}\left(\frac{\boldsymbol{1}}{\boldsymbol{D}^2}\right) = \frac{\frac{\boldsymbol{1}}{\boldsymbol{D}^2}}{\sum \frac{\boldsymbol{1}}{\boldsymbol{D}^2}} \qquad \textbf{(4.10)}$$

Where

D: Buffer width.

The weight coefficients for 500 & 500-800 m buffer ring area and 500 & 500-1000 m buffer ring area are shown in Table 4.2-3.

Table 4.2-3 Buffer Width versus Weight Coefficient

Buffer Width	**Weight Coefficient**
500 m buffer area 500-800 m buffer ring area	0.719 0.281
500 m buffer area 500-1000 m buffer ring area	0.8 0.2

Step 3 The Extraction Formulas of Population Data

Population data are assumed as a uniform distribution within the census tract boundary. For each stop-based buffer area, an area-based fraction method is applied to extract and sum the relevant zonal population data (Pendyala et al., 2004). The population data in small zonal area and stop-based buffer area can be mathematically expressed in Equations (4.11) and (4.12), respectively.

$$Population_{Zonal^k} = \frac{Area_{Zonal\,k}}{Area_{Census\,j}} Population_{Census\,j} \tag{4.11}$$

$$Population_{Stop^i} = \sum_{k=1}^{n} Population_{Zonal^k} \tag{4.12}$$

Where

k: The zonal number , k= 1,2,...,k,...

j: The census tract boundary number, j = 1,2,...,j,...

i: The stop number, i = 1, ...14.

Step 4 Overlay Analysis (Conestoga Mall stop as an example in Figure 4.2-7)

As shown in Figure 4.2-7, a 500 m buffer area and a 500-1000 m buffer ring area are overlaid with population data within the Census Tract Boundary. The heavy weight coefficient 0.8 is assigned to stop-based 500 m buffer area; the light weight coefficient 0.2 is assigned to 500-1000 m buffer ring area.

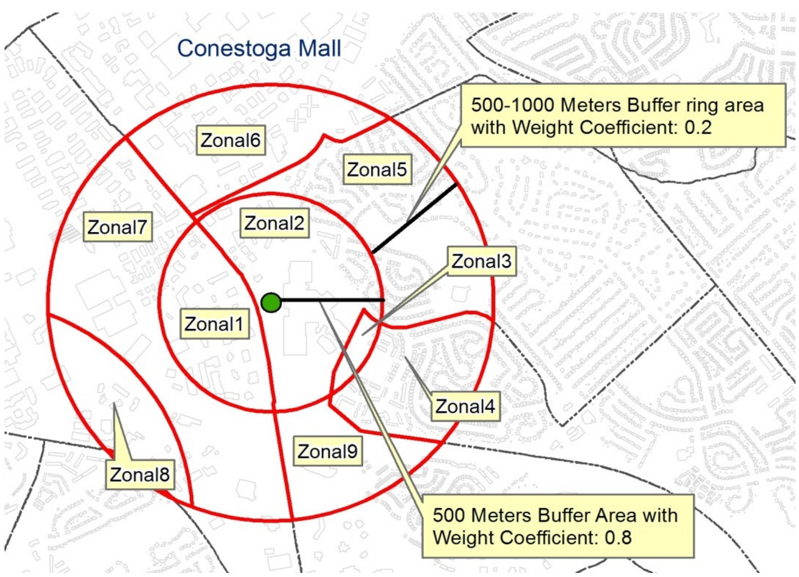


Figure 4.2-7 ArcGIS Intersect Analysis

Step 5 Intersect Analysis

An ArcGIS Intersect Analysis tool is used to process the two overlaid data layers. Population data are broken into small zonal pieces within the stop-based buffer area through the Intersect Analysis. An Area-based Fraction Equation is used for calculating the population data within the small zonal area (assuming population data are uniformly distributed in census tract boundary) based on Equation (4.11).

The buffers, overlay, and intersect functions of the GIS analysis tools are used for the related attribute analysis and the polygon calculations. Moreover, spatial proximity method and spatial weight method are combined in the spatial data processing to obtain precise data extraction for better ridership estimation.

Step 6 Sum Analysis

The population data within stop-based buffer area can be obtained through summing population data of the 9 zonal areas based on Equation (4.12).

Step 7 Export

The summed results are exported from the GIS system and then are imported to Excel as IV1 for correlation and regression analysis.

For Route 12, there are many close-by stops in the conventional bus route, meaning that there are highly overlapped stop-based buffers occurring in the corridor. Therefore, these close-by stops affect the selection of the stop-based buffer distance. Chu (2004) reported the same situation. To avoid the overlapping buffer area between two or more stops, segment-based buffer is introduced to deal with overlapping area double counting caused by close-by stops in buffering (She, 2015). Many tests have been done through steps 1-7. The final result in a high accurate prediction model is reached.

4.2.2 Extraction of the Independent Variable2 (IV2)

IV2 is defined as number of feeder buses that arrive at each stop. It can be referred as Trip Attraction indirectly (TA) because these riders are from transit network level. The Davis Centre stop at the University of Waterloo is used as an example for explaining the definition of IV2 (see Figure 4.2-8).

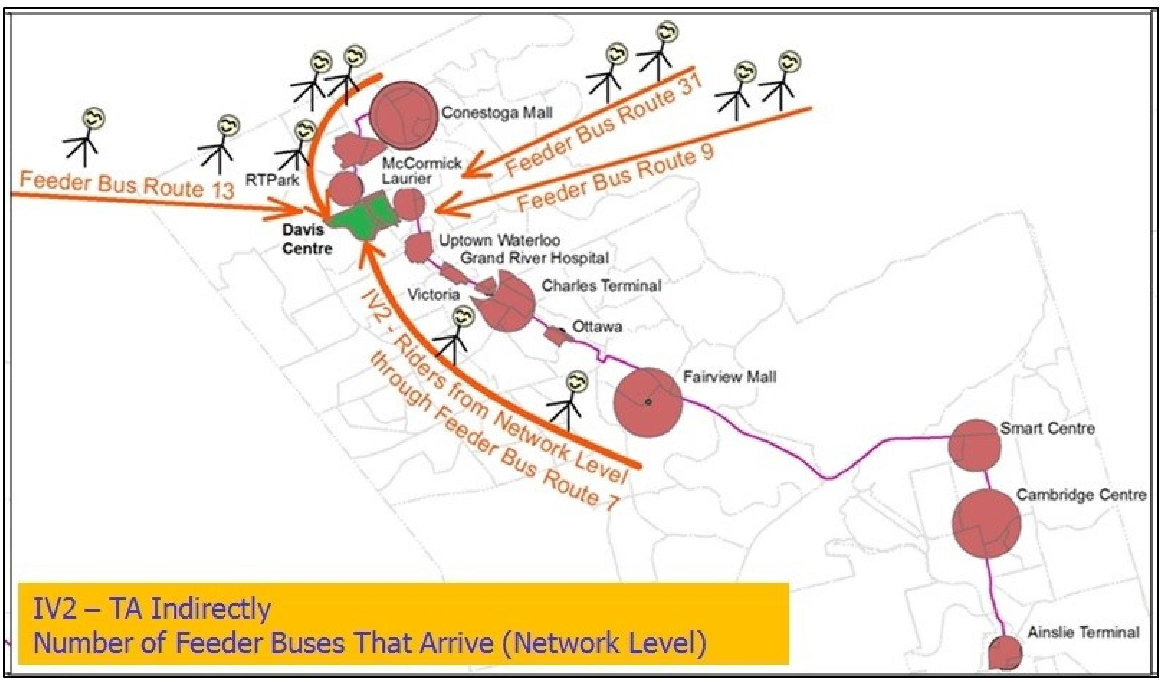
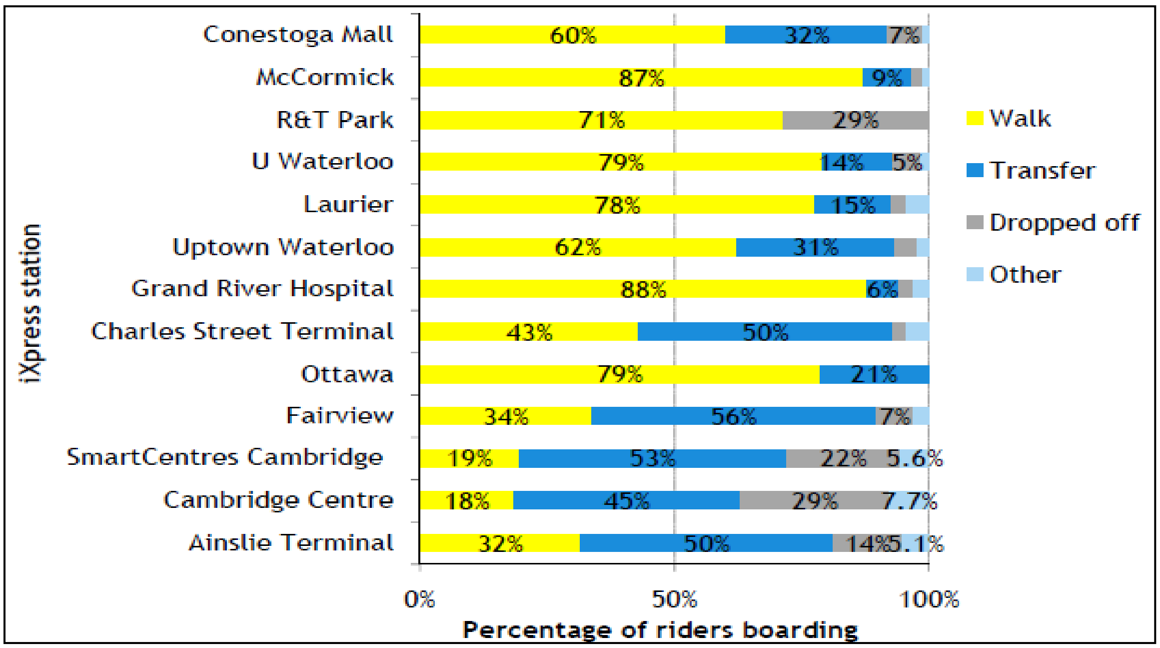


Figure 4.2-8 IV2 - Number of Feeder Buses That Arrive (Network Level)

The data are provided by the Region of Waterloo through routes, stops, stop times, trips, and shapes etc. tables. These tables are imported to Excel software and merged into one table using the VLookup function, and then the filter function is used to count all numbers of feeder buses that arrive at each stop.

Based on the report from the Region of Waterloo (2009), there are high transfer rates of around 50% at the Charles Terminal, Fairview Mall, Smart Centre, Cambridge Centre, and Ainslie Terminal. Therefore, the independent variable - IV2 in feeder buses services has a big contribution to the prediction model and cannot be ignored (see Figure 4.2-9).



Source: (The Region of Waterloo, 2009)

Figure 4.2-9 Transfer Rate at iXpress 200 Stops

IV2 is calculated as

$$(IV2)_i = \textbf{Number of Feeder Buses That Arrive at Stop}_i \qquad \textbf{(4.13)}$$

Where

i: Stop number

Route 7 (from Conestoga Mall to Fairview Mall) and Route 51 (from SmartCentre to Ainslie Terminal) have the same alignment and destination with iXpress 200 (from Conestoga Mall to Ainslie

Terminal). Therefore, there is very low transfer possibility for riders to transfer from Route 7 and Route 51 to iXpress 200. Therefore, the feeder services of Route 7 and Route 51 will not be counted for iXpress 200.

4.2.3 Extraction of the Independent Variable3 (IV3)

In the ridership prediction models, IV3 is used to count for the trip-attraction by non-local residents along the corridor who come and go because of local attractions at each stop. IV3 is defined as riders from other origins along the corridor. It can be referred as Trip Attraction directly (TA) because these riders are from transit route level. The Davis Centre stop at the University of Waterloo is used as an example for explaining the definition of IV3 (Figure 4.2-10).

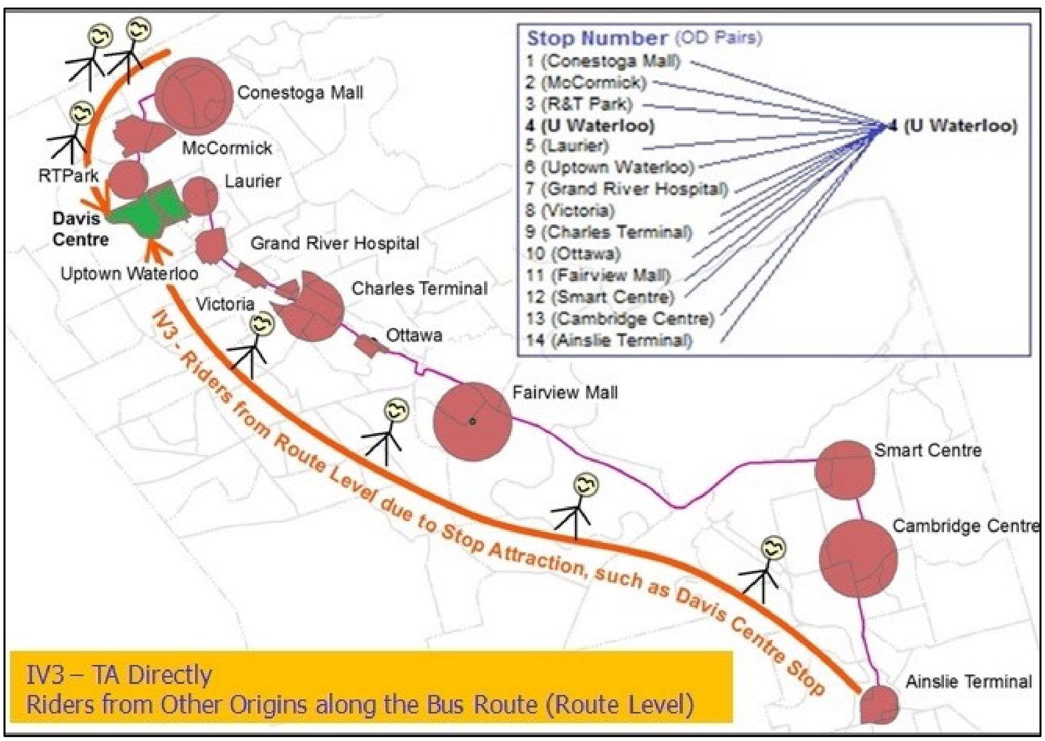


Figure 4.2-10 IV3 - Riders from Other Origins along the Bus Route (Route Level)

Employment and student data along the bus route level are collected from 2011 Statistics Canada Household Survey for IV3. The extraction methods and steps are the same as that of IV1. Based on the IV3 definition and the stop nature along the bus route, the attractive types are different due to the different land use. The strength coefficients of attractions for different employment types are studied by Casello and Smith (2006). For stops at shopping malls, the employment type is considered to be retail jobs, the disaggregate employment trip attracting index is given at 2.56. Therefore, the IV3 is equal to the retail jobs * 2.56. For the stops at the two universities, the strength coefficient of attraction is given at 2 trips/employee/day. The two universities stops also include high volume students approaching from other stop origins along the bus route, therefore, the route level student data (age 18-29) are considered as OD pairs for the two Universities (see Figure 4.2-10). The strength coefficient of attraction for each student is given at 2 trips/student/day. The IV3 for the two universities is taken as (employment data + student data) * 2. For the rest stops, the employment type is considered as service and government employment. The strength coefficient of attraction is given at 1.12, and thus the IV3 for the rest stops is the service type jobs * 1.12.

Step 1 The Selection of Employment and Student Data from Route Level to Stop-based Effective Service Buffer Area (Boarding-affecting Buffer Area)

- **The Selection of Employment Data from Other Origins along iXpress 200 to Stop-based Effective Service Buffer Area**

Based on simple linear regression method, the relationship between average weekday boardings and employment data from iXpress 200 route level can be illustrated in Figure 4.2-11, and the extraction method of IV3 is detailed in Table 4.2-4. From Figure 4.2-11, it is found that the highest Effective Service buffer area for employment type of considering strength of attraction are within the mixed 500-1000m stop-based buffer with SPM and SWM application. The R square is 0.733.

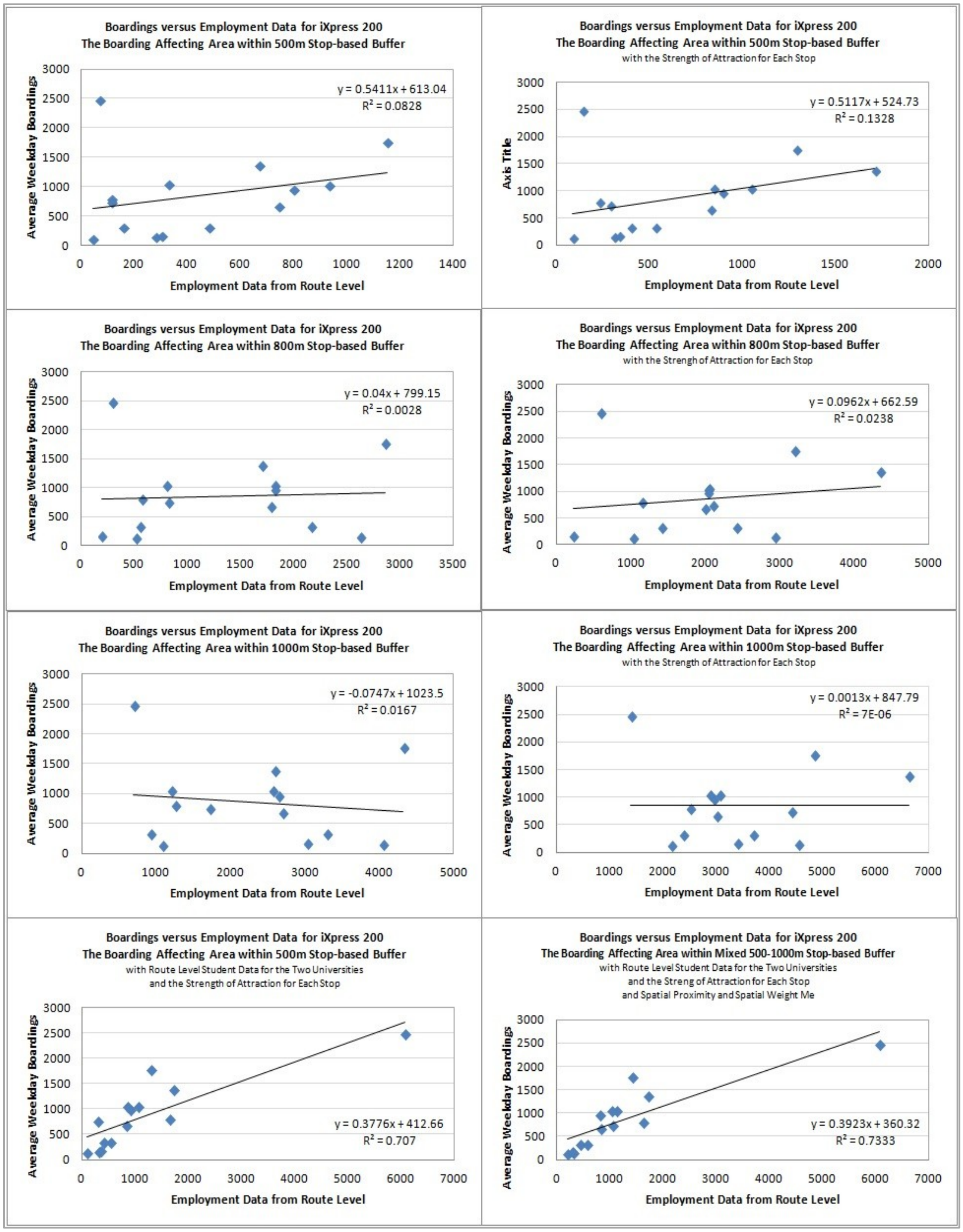


Figure 4.2-11 iXpress 200 - Boardings versus Employment Data from Route Level

Table 4.2-4 iXpress 200 - The Data Extraction Method of IV3

Stop_Name	**IV3 - (Employment + Student) Data** within Effective Service-Buffer Area	Description
1 (Conestoga Mall)	Employment data (as retail jobs) * Strength of attraction (2.56)	1000m buffer with SPM and SWM application. At the stop, most employment data can be considered as retail jobs , therefore the coefficient of the strength of attraction can be taken at 2.56
2 (McCormick)	Employment data (as service jobs) * Strength of attraction (1.12)	500m buffer size is given. Employment types including librarian etc. jobs are considered as service jobs, therefore the coefficient of the strength of attraction can be taken at 1.12
3 (R & T Park)	[Employment data (high-tech) + Student data] * Strength of attraction (2)	500m buffer size is given. Most employment data are high-tech, the strength of attraction is considered as 2 trips/employee/day Co-op students should be from other origins because this is a business park, nearly no residents, the strength of attraction is considered as 2 trips/student/day
4 (U Waterloo)	(Employment data + Students) * the strength of attraction (2)	Road boundary and CTB boundary around polygon area. Most employment data are high-tech and education, the strength of attraction is considered as 2 trips/employee/day 69% of 2011 Statistics Canada (18-29 yrs students from other stop-based buffer area along the iXpress 200 route level based on the report (The Region of Waterloo, 2009). the strength of attraction is considered as 2 trips/student/day
5 (Laurier)	(Employment data + Students) * the strength of attraction (2)	500m buffer area. Most employment data are high-tech and education, the strength of attraction is considered as 2 trips/employee/day 31% of 2011 Statistics Canada (18-29 yrs students from other stop-based buffer area along iXpress 200 route level based on the report (The Region of Waterloo, 2009). The strength of attraction is considered as 2 trips/student/day
6 (Uptown Waterloo)	Employment data (as service jobs) * The Strength of Attraction (1.12)	Employment data in government, community services, sales and service occupations etc. The strength of attraction is given at 1.12

7 (Grand River Hospital)	Employment data (as service jobs) * The Strength of Attraction (1.12)	Road boundary-based buffer. Employment data in health, business, financial, and administration. The strength of attraction is given at 1.12
8 (Victoria)	Employment data (as service jobs) * The Strength of Attraction (1.12)	Road boundary-based buffer. Employment data in education, law and social, community and services. The strength of attraction is given at 1.12
9 (Charles Terminal)	Employment data (as service jobs) * The Strength of Attraction (1.12)	800m stop-based buffer. Employment data in administration, financial, law and social, community and services. The strength of attraction is given at 1.12
10 (Ottawa)	Employment data (as service jobs) * The Strength of Attraction (1.12)	Employment data in sale and services etc. The strength of attraction is given at 1.12
11 (Fairview)	Employment data (as retail jobs) * The Strength of Attraction (2.56)	500m buffer for the shopping mall. At the stop, most employment data can be considered as retail jobs , therefore the coefficient of the strength of attraction can be taken at 2.56
12 (SmartCentre)	Employment data (as retail jobs) * The Strength of Attraction (2.56)	800m buffer with SPM and SWM application. At the stop, most employment data can be considered as retail jobs , therefore the coefficient of the strength of attraction can be taken at 2.56
13 (Cambridge Centre)	Employment data (as retail jobs) * The Strength of Attraction (2.56)	1000m buffer with SPM and SWM application. At the stop, most employment data can be considered as retail jobs , therefore the coefficient of the strength of attraction can be taken at 2.56
14 (Ainslie Terminal)	Employment data (as service jobs) * The Strength of Attraction (1.12)	Employment data in Business, finance and administration; and Health occupations; and education, law and social, community and government services; and art, culture, recreation and sport. The strength of attraction is given at 1.12

Note: All employment data do NOT include the mode of transportation by walk from stop-based buffer area

The Disaggregate Employment Trip Attraction Indices are introduced based on the research paper (Transportation Activity Centres for Urban Transportation Analysis) published by Casello 2006

- **The Selection of Employment Data from Other Origins along Route 12 to Stop-based Effective Service Buffer Area**

Based on simple linear regression method, the relationship between average weekday boardings and employment data from Route 12 route level can be illustrated in Figure 4.2-12. From the figure, it is found that the highest Effective Service buffer area for employment type of considering strength of attraction are within the mixed 300-1000m stop-based and segment-based buffer with SPM and SWM application. The R square is 0.6241.

The extraction method used for IV3 for Route 12 is the same as that of iXpress 200. The details are shown in Table 4.2-5.

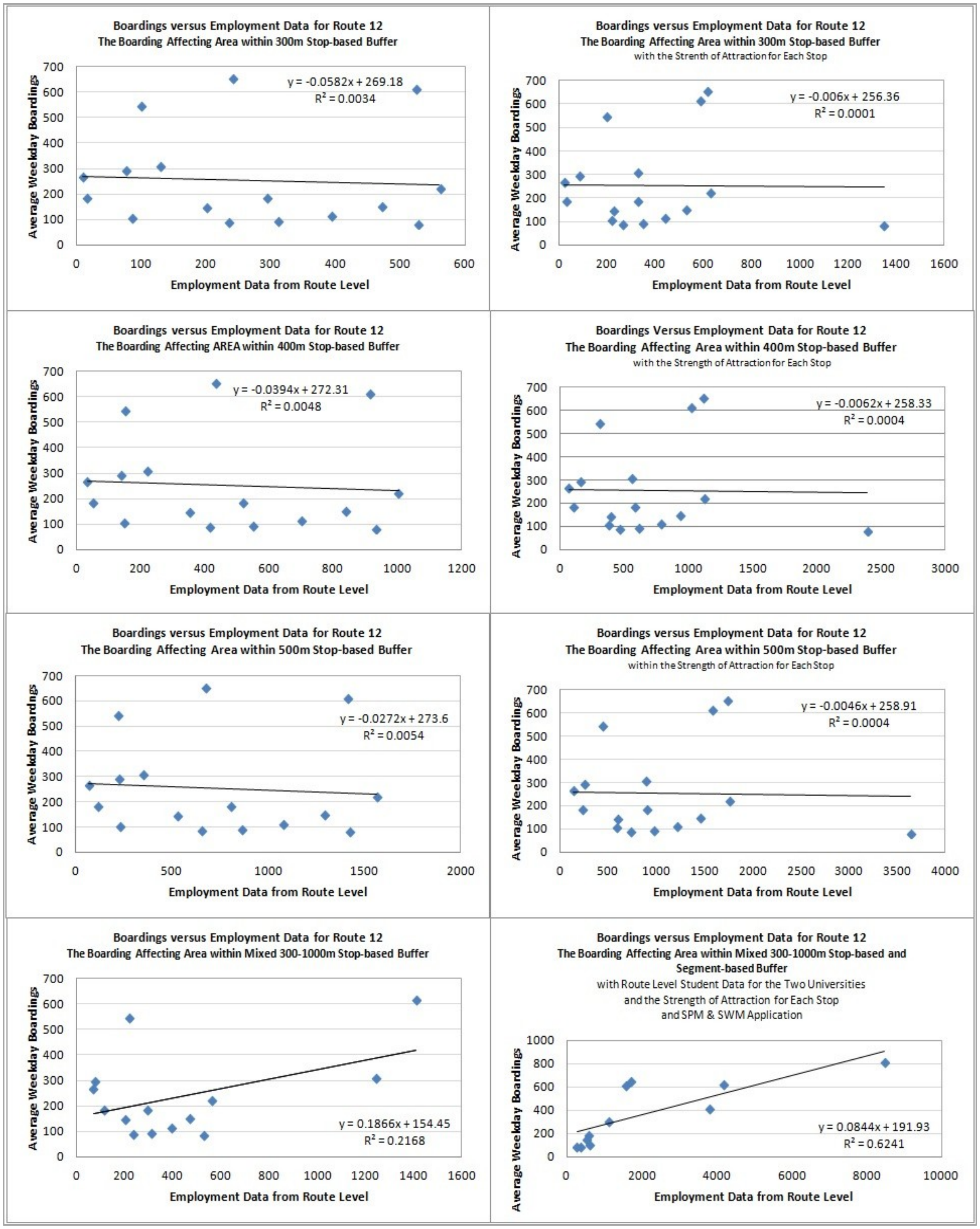


Figure 4.2-12 Route 12 - Boardings versus Employment Data from Route Level

Table 4.2-5 Route 12 - The Data Extraction Method of IV3

Stop No.	**IV3 - (Employment + Student) Data** within Effective Service Buffer Area	**Description**
1 (Conestoga Mall)	Employment data (as retail jobs) * The strength of attraction (2.56)	1000m buffer with SPM and SWM application. At the stop, most employment data can be considered as retail jobs, the coefficient of the strength of attraction can be taken at 2.56
2 (Bridge@University)	Employment data (as service jobs) * The strength of attraction (1.12)	300m buffer is given. Employment types are considered as service jobs, the coefficient of the strength of attraction can be taken at 1.12
3 (LincolnBluevale)	Employment data (as service jobs) * The strength of attraction (1.12)	400m buffer is given. Employment types are considered as service jobs, the coefficient of the strength of attraction can be taken at 1.12
4 (S-Laurier)	(Employment + Student) data * The strength of attraction (2)	633m Segment-based 300m buffer (from University@Weber to Laurier) Most employment data are educational type, the strength of attraction is considered as 2 trips/employee/day. 31% of 2011 Statistics Canada (18-29 yrs) students from other origin along Route 12 based on the report (The Region of Waterloo, 2009). The Strength of attraction is considered as 2 trips/student/day
5 (S-U Waterloo)	(Employment + Student) data * The strength of attraction (2)	Campus boundary with road boundary around segment-based polygon buffer. Most employment data are educational type, the strength of attraction is considered as 2 trips/employee/day. 69% of 2011 Statistics Canada (18-29 yrs) students from other origin along Route 12 based on the report (The Region of Waterloo, 2009). The Strength of attraction is considered as 2 trips/student/day
6 (Westmount@Erb)	Employment data (as service jobs) * The strength of attraction (1.12)	300m buffer is given. Employment types are considered as service jobs, the coefficient of the strength of

		attraction can be taken at 1.12
7 (S-Westmount@Highland)	Employment data (as retail jobs) * The strength of attraction (2.56)	1124m segment - based 300m buffer area. Employment types are considered as retail jobs, therefore, the coefficient of the strength of attraction can be taken at 2.56
8 (Westmount@Ottawa)	Employment data (as service jobs) * The strength of attraction (1.12)	300m buffer is given. Employment types are considered as service jobs, the coefficient of the strength of attraction can be taken at 1.12
9 (Forest Glen Terminal)	Employment data (as service jobs) * The strength of attraction (1.12)	500m buffer is given. Employment types are considered as service jobs, the coefficient of the strength of attraction can be taken at 1.12
10 (Fairway@Food basic)	Employment data (as retail jobs) * The strength of attraction (2.56)	500m buffer is given. Employment types are considered as retail jobs, the coefficient of the strength of attraction can be taken at 2.56
11 (Fairview Mall)	Employment data (as retail jobs) * The strength of attraction (2.56)	1000m buffer is given. Employment types are considered as retail jobs, the coefficient of the strength of attraction can be taken at 2.56

Note: All employment data do NOT include the mode of transportation by walk from stop-based buffer area
The Disaggregate Employment Trip Attraction Indices are introduced based on the research paper (Transportation Activity Centres for Urban Transportation Analysis) published by Casello 2006

Step 2 The Assignment Method of Weight Coefficients for IV3

See Step 2 in the Extraction of IV1 (section 4.2.1).

Step 3 The Extraction Formulas of Employment and Student Data

Employment and student data are taken from the Census Tract Boundary of 2011 Statistics Canada Household Survey. They are assumed to follow a uniform distribution in the boundary, and are processed to stop-based buffer area based on an Area-based Fraction Equation (Pendyala, et al., 2004) using GIS analysis tools. The employment and student data in small zonal area and stop-based buffer area can be mathematically expressed in Equations (4.14), (4.15), (4.16), and (4.17).

$$\boldsymbol{EMPLOYEES_{Zonal^{k}}} = \frac{\boldsymbol{Area_{Zonal^{k}}}}{\boldsymbol{Area_{Census^{j}}}} \boldsymbol{EMPLOYEES_{Census^{j}}} \tag{4.14}$$

$$\boldsymbol{Students_{Zonal^{k}}} = \frac{\boldsymbol{Area_{Zonal^{k}}}}{\boldsymbol{Area_{Census^{j}}}} \boldsymbol{Students_{Census^{j}}} \tag{4.15}$$

$$\boldsymbol{EMPLOYEES_{Stop^{i}}} = \sum_{k=1}^{n} \boldsymbol{EMPLOYEES_{Zonal^{k}}} \tag{4.16}$$

$$\boldsymbol{Students_{Stop^{i}}} = \sum_{k=1}^{n} \boldsymbol{Students_{Zonal^{k}}} \tag{4.17}$$

Where

k: The zonal number , k= 1,2,...,k,...

j: The census tract boundary number, j = 1,2,...,j,...

i: The stop number, i = 1, ...14.

Steps 4-7 for The Extraction of IV3

See the Extraction of IV1 (section 4.2.1).

4.3 Correlation Analysis between IVs

Section 4.1 and 4.2 explain how to extract dependent variable (DV) and independent variables (IVs). Section 4.3 will analyze the relationship between DV and IVs, and check the correlation coefficients among IVs to ensure the reliability and validity of the prediction model.

In multiple linear regression analysis, it is very important to check the IVs' relationships. If IVs have a high correlation to each other, they might distort the prediction model. One possible solution is to remove one or more highly correlated IVs from the regression equation. Another solution is to combine the highly correlated IVs into a best fit IV to improve the accuracy of the prediction model. In the statistical literature, Quantitative Data Analysis, the degree of correlation coefficient (r) can be divided into three categories. For r between [0.8, 1] and [-1, -0.8], it is called strong correlation, between (-0.8, 0.8) is weak correlation, and if r is 0, then no correlation (Walpole & Myers, 1978), (Kashef, 2014).

4.3.1 Correlation Analysis for iXpress 200 Prediction Model

With the extracted data of IV1, IV2, IV3, and DV for iXpress, the correlation coefficients between IVs and DV are analyzed and summarized in Table 4.3-1.

Table 4.3-1 iXpress 200 Correlation of Dependent Variable and Independent Variables

	DV	IV1	IV2	IV3
DV	1			
IV1	0.543	1		
IV2	0.577	0.662	1	
IV3	0.829	0.082	0.167	1

Table 4.3-1 reveals the correlation between the dependent variable, the three independent variables, and the independent variable to each other. These correlation coefficients measure how strong their relationships are. The three IVs have relatively high relationships with the dependent variable (DV). The independent variable3 (IV3) has the highest correlation coefficient of 0.829 with the boardings

dependent variable (DV), meaning that it plays as the most significant independent variable in the prediction model due to stop attractions for riders from other origins along iXpress 200 corridor, followed by IV2 and IV1.

There is no high correlation among independent variables; and the lowest correlation coefficient is between the IV1 and IV3, which is 0.082. This is reasonable in a real world situation - there are very low residents around Davis Centre stop, however, there are very high boardings from other origins along the bus route.

4.3.2 Correlation Analysis for Route 12 Prediction Model

With the extracted data of IV1, IV2, IV3, and DV for Route 12, the correlation coefficients between IVs and DV are analyzed and summarized in Table 4.3-2.

Table 4.3-2 Route 12 Correlation of Dependent Variable and Independent Variables

	DV	IV1	IV2	IV3
DV	1			
IV1	0.510	1		
IV2	0.780	0.228	1	
IV3	0.790	0.130	0.424	1

Table 4.3-2 reveals that the IV2 and IV3 contribute the most to the dependent variable. The IV1 shows a relatively lower correlation coefficient with DV. The APCS records reflect the relationship for Route 12. For example, the Westmonut@Highland stop is located at a high density residential area, but the stop has very low boardings, meaning that most local residents commute by other travel

modes such as driving. This situation is generalized as low traffic congestion, many free parking lots in business parks, and driving habits in Waterloo Region (Hellinga et al., 2007). Another reason is that the effects of the decentralization still exist in the Region of Waterloo. High density residences are still distributed in the suburbs. Therefore, these residents prefer driving to taking public transit. Figure 4.3-1 illustrates the relationship between the 2011 population distribution and ridership. If Route 12 has an alternative schedule to pass through the high density residential area from Westmont@Highland to Forest Glen Plaza, the ridership is expected to improve.

The lowest correlation coefficient is between IV1 and IV3, which is 0.130. For example, in the reality, there are strong attractions for a variety of riders at Conestoga Mall, yet few residents there; there is a high density residential area at Westmount@Highland, yet hardly any riders are attracted there. There are no high correlations between IVs.

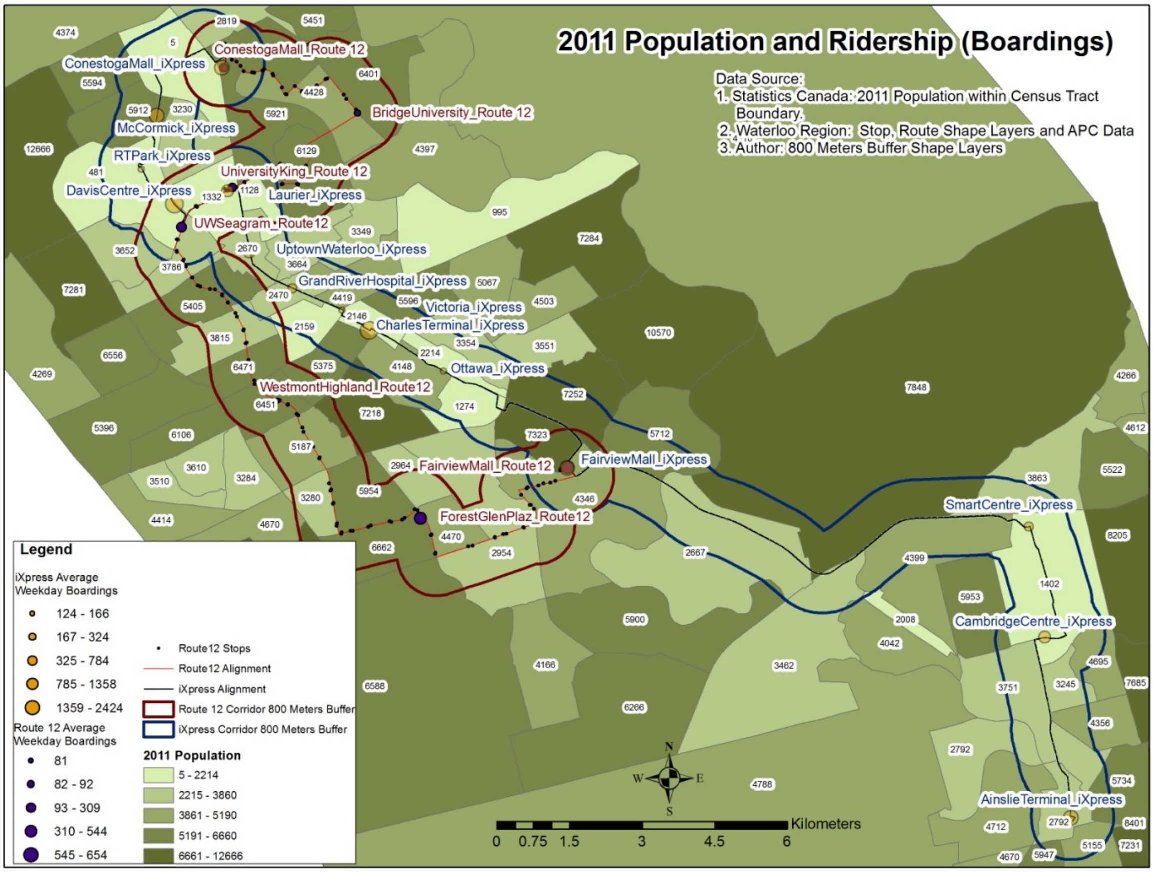


Figure 4.3-1 2011 Population Distribution and Ridership

4.4 Prediction Models

Section 4.1 and 4.2 analyzed and explained how to extract dependent variable (DV) and independent variables IVs; Section 4.3 analyzed the relationship between DV and IVs, and checked the correlations among IVs. This section will develop a multiple linear regression prediction model for iXpress 200 and Route 12 based on the extracted DV and IVs. The regression model includes trip-production (IV1) and trip-attractions (IV2 and IV3). It can be mathematically expressed as:

$$\hat{y}_i = \mathbf{b_0} + \mathbf{b_1} * \boldsymbol{IV1_i} + \mathbf{b_2} * \boldsymbol{IV2_i} + \mathbf{b_3} * \boldsymbol{IV3_i} + \boldsymbol{\varepsilon}_\mathbf{i} \qquad \textbf{(4.18)}$$

Where

$\hat{y}_i$:	Dependent variable - average weekday boardings
b_0:	Intercept (constant value)
b_{1-3}:	Regression model coefficients to be estimated
$IV1_i$:	Population data within stop-based buffer area
$IV2_i$:	Number of feeder buses that arrive
$IV3_i$:	Riders from other origins along bus route
ε_i:	Error
i:	Stop number

The Least Squares method is used to estimate the relationship between average weekday boardings and the three independent variables and to find the coefficients of $\mathrm{b}_0, \mathrm{b}_1, b_2, b_3$ for each route.

The boardings average values of 17 weekdays for 14 stops are used in the regression matrix for the iXpress 200. The matrix representation of the multiple linear regression models is as follows:

$$\boldsymbol{y} = \begin{bmatrix} \mathbf{y_1} \\ \boldsymbol{y_2} \\ \vdots \\ \boldsymbol{y_{14}} \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} \mathbf{1} & \boldsymbol{x_{1,1}} & \boldsymbol{x_{2,1}} & \boldsymbol{x_{3,1}} \\ \mathbf{1} & \boldsymbol{x_{1,2}} & \boldsymbol{x_{2,2}} & \boldsymbol{x_{3,2}} \\ \vdots & \vdots & \vdots & \ddots \\ \mathbf{1} & \boldsymbol{x_{1,14}} & \boldsymbol{x_{2,14}} & \boldsymbol{x_{3,14}} \end{bmatrix}, \boldsymbol{b} = \begin{bmatrix} \mathbf{b_0} \\ \mathbf{b_1} \\ \vdots \\ \mathbf{b_3} \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \boldsymbol{\varepsilon_1} \\ \boldsymbol{\varepsilon_2} \\ \vdots \\ \boldsymbol{\varepsilon_{14}} \end{bmatrix} \qquad \textbf{(4.19)}$$

The boardings average values of 17 weekdays for 11 randomly selected stops are used in the following regression matrix for Route 12. The matrix representation of the multiple linear regression model is as follows:

$$\boldsymbol{y} = \begin{bmatrix} \mathbf{y_1} \\ \boldsymbol{y_2} \\ \vdots \\ \boldsymbol{y_{11}} \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} \mathbf{1} & \boldsymbol{x_{1,1}} & \boldsymbol{x_{2,1}} & \boldsymbol{x_{3,1}} \\ \mathbf{1} & \boldsymbol{x_{1,2}} & \boldsymbol{x_{2,2}} & \boldsymbol{x_{3,2}} \\ \vdots & \vdots & \vdots & \ddots \\ \mathbf{1} & \boldsymbol{x_{1,11}} & \boldsymbol{x_{2,11}} & \boldsymbol{x_{3,11}} \end{bmatrix}, \boldsymbol{b} = \begin{bmatrix} \mathbf{b_0} \\ \mathbf{b_1} \\ \vdots \\ \mathbf{b_3} \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \boldsymbol{\varepsilon_1} \\ \boldsymbol{\varepsilon_2} \\ \vdots \\ \boldsymbol{\varepsilon_{11}} \end{bmatrix} \qquad \textbf{(4.20)}$$

Based on equations (4.19) and (4.20), the prediction models for iXpress and route 12 are developed based on Microsoft Excel's Regression tool (Figure 4.4-1). In the regression processing of the two bus route, confidence level is set at 95% (See Section 4.1 for confidence interval analysis).

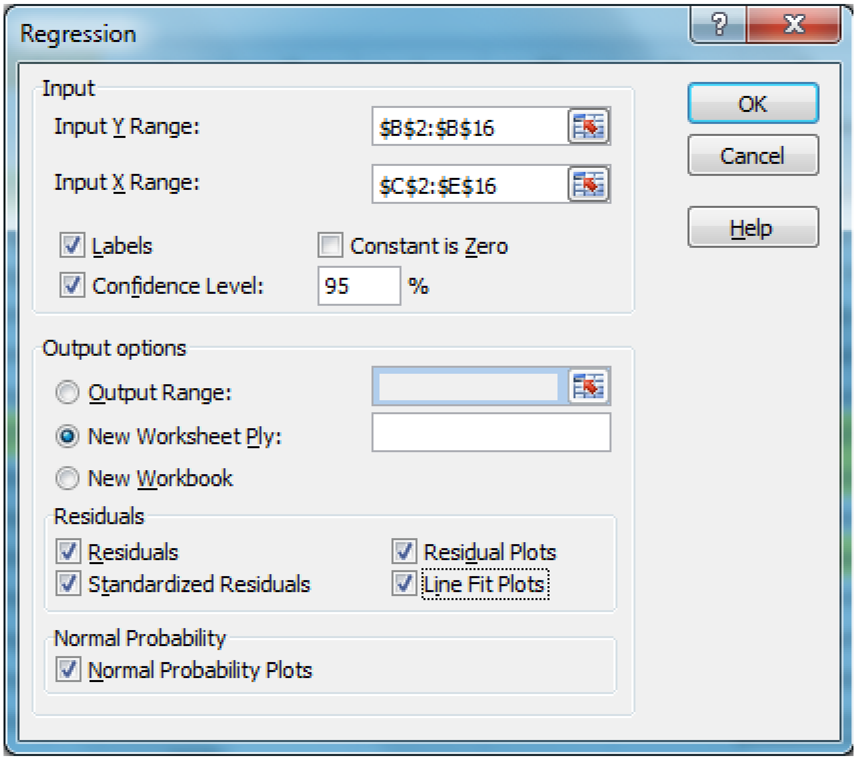


Figure 4.4-1 Microsoft Excel's Regression Tool

The Prediction Model for iXpress 200:

$$DV_{\substack{Average\ Weekday \\ boardings\_iXpress}} = -58.75 + 0.246 * IV1 + 0.22 * IV2 + 0.398 * IV3 \quad \textbf{(4.21)}$$

The Prediction Model for Route 12:

$$DV_{\substack{Average\ Weekday \\ boardings\_Route12}} = -1.79 + 0.107 * IV1 + 0.240 * IV2 + 0.058 * IV3 \quad \textbf{(4.22)}$$

The above prediction models can be explained as the three independent variables IV1, IV2 and IV3 contributing to the dependent variable - average weekday boardings for iXpress 200 and Route 12, respectively.

Chapter 5 Regression Results Analysis

Chapter 4 presented the data extraction methods of dependent variable (DV) and independent variables (IVs); analyzed the correlation among IVs; and developed the ridership prediction models which are applied for iXpress 200 and Route 12, respectively. Chapter 5 will analyze and discuss the reliability and validity of the regression results; verify the accuracy of the developed models by comparing the new sample data to the predicted values; validate the feasibility of the modeling method by applying the method to different service periods - morning peak (6:00 - 10:00) am, off-peak (10:00 am - 3:00 pm), and afternoon peak (3:00 - 7:00) pm; explore the difference between the two bus routes by comparing the regression results; and prove that the ridership modeling method is feasible at stop level estimation.

5.1 Regression Results Analysis for iXpress 200

The multiple linear regression models are developed based on the relationship between a dependent variable (DV) and the three independent variables (IV1, IV2, and IV3).The regression results are obtained using Excel's regression tools. The four results are:

(1) R Square and Adjusted R Square, which reflect overall regression model's accuracy

(2) Significance of F, which explains the probability that the output is not by chance

(3) Regression coefficients with related P-values (which provide the individual regression coefficient's accuracy, respectively)

(4) Residuals, their distribution patterns can be visualized for checking the regression model's validity.

The four results are analyzed and discussed in details as follows.

5.1.1 Regression Statistics Analysis

Table 5.1-1 shows the regression statistics results from Excel's regression processing. R Squared value (coefficient of multiple determinations) is a number from 0 to 1 that reveals how close the estimated values for the trendline correspond to actual sample data. A trendline is more reliable when its R-squared value is at or near 1. The regression result R^2 is equal to 0.9526 which is near 1. It means that the 95.26% of the variance in boardings dependent variable can be explained by the variance of the three independent variables. The higher the R-squared value is, the better the model fit the sample data.

The Adjusted R Squared value at 0.9383 is always less than R Squared value. It expresses a high quality linear fit. When any new input variable is added to the regression model with more accurate results, the Adjusted R Squared value will increase. However, the R Squared value always increases when a new input variable is added, no matter the new input variable enhances the regression model's accuracy or not.

The value of the Standard Error is 167, indicating that the dependent variable (average weekday boardings) has a big variance range. For example, the variance range of the boardings fluctuation distribution at U Waterloo (Davis Centre) stop is from a minimum of 1991 to a maximum of 2919. The 17 weekday boardings show a great dispersion. The solution to decrease the standard error is to increase more observation data and set confidence interval in the sample estimates.

Table 5.1-1 iXpress 200 Regression Statistics Analysis

Regression Statistics	
Multiple R	0.9760
R Square	0.9526
Adjusted R Square	0.9383
Standard Error	167
Observations	14

5.1.2 Regression Properties Analysis

Table 5.1-2 illustrates that the validity of the regression results is confirmed by the small significance F at 0 percent level, meaning that the regression model is significant valid and the result of the model is not by chance.

The regression model properties are defined as follows:

- SSR (Regression Sum of Squares): Amount of variation in average weekday boardings explained by the regression model.
- SSE (Error Sum of Squares): Variation in average weekday boardings due to error that is not explained by the regression model.
- SST (Total corrected Sum of Squares): The variation in average weekday boardings that ideally would be explained by the regression model. The SST equals to the sum of SSR and SSE.
- MSR (Mean Square Regression): MSR = $\frac{SSR}{\mathrm{dfR}}$, which dfR refers to regression of freedom.

- MSE (Mean Square Error): $S^2 = \frac{SSE}{\mathrm{dfE}}$, which estimates S^2 , the variance of the errors; dfE refers to residual of freedom.
- Significance of F: Confirms the validity of the regression result. In this case, the Significance of F equals to 6.36E-07 (near zero). Therefore, the probability of the regression result is not by chance. There is very strong evidence that the regression model is statistically significant (Duever, 2006).

Table 5.1-2 iXpress 200 Regression Model Properties and Significance F

Analysis of Variance					
	df	**SS**	**MS**	**F**	**Significance F**
Regression	3	SSR: 5619172	MSR: 1873057	$\frac{\mathrm{MSR}}{\mathrm{MSE}} = 66.962$	6.36E-07
Residual	10	SSE: 279719	MSE: 27972		
Total	13	SST: 5898890			

5.1.3 Regression Coefficients Analysis

In Table 5.1-3, the coefficients column lists the Least Squares estimates of the three independent variables. The acceptable P-values for the three independent variables are less than 0.03; there is very strong evidence in the three coefficients' validity. The lower the P-value is, the higher the coefficients' validity is; if the p-value is more than 0.1, then the coefficient estimate is not reliable because it has too much dispersion/variance (Duever, 2006). Thus, the coefficients estimate can be predicated as true with a 95% level of confidence.

Table 5.1-3 iXpress 200 Regression Coefficients Analysis

	Coefficients	P-value	Lower 95%	Upper 95%
Intercept	-58.751	0.529	-259.384	141.882
IV1	0.246	0.007	0.083	0.408
IV2	0.220	0.033	0.022	0.419
IV3	0.398	6.6E-07	0.318	0.479

The following is a hypothesis testing on the regression model coefficients: assuming that the three coefficients of the three IVs are B1, B2, and B3.

Then,

Null hypothesis H0: $\boldsymbol{B1 = B2 = B3 = 0}$ **(5.1)**

Alternative hypothesis Ha: At least one coefficient is not 0

Test statistic: $\mathbf{f} = \frac{\mathbf{MSR}}{\mathbf{MSE}} = \frac{\mathbf{1873057}}{\mathbf{27971.88}} = \boldsymbol{66.962}$ **(5.2)**

Where,

MSR: Mean Square Regression

MSE: Mean Square Error

The critical value: $\boldsymbol{F0.05, 3, 10 = 3.7}$ ($\boldsymbol{From\ the\ table\ of\ F - statistics\ P = 0.05}$) **(5.3)**

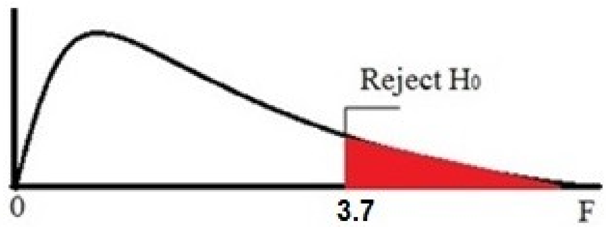


Figure 5.1-1 Hypothesis Testing on iXpress 200 Regression Model Coefficients

The conclusion of the above hypothesis testing: The value of the test statistic equals 66.962 which are far greater than the critical value 3.7. Thus, the null hypothesis falls in the rejection region (see Figure 5.1-1), and it is rejected. Therefore, there is strong evidence in accepting the alternative hypothesis - at least one of the coefficients is not zero, meaning that at least one of the independent variables contributes to the prediction model. The regression coefficients are shown valid.

5.1.4 Regression Residuals Analysis

Based on the regression results and the Equation (3.3) of Chapter 3:

The residual $\boldsymbol{e_i = y_i - \hat{y}_i}$ **i = 1, 2, ..., n** **(5.4)**

Where

i: Stop number

yi: Actual average weekday boardings

$\hat{y}_i$: Predicted average weekday boardings from the regression model

The residual analysis is also an important factor to check the prediction model's validity by checking the residual distributions. Fotheringham et al. (2002) suggested that the standard residual of a data dot's absolute value surpassing 3 could be a potential outlier. Figure 5.1-2 reveals that the residual dots are randomly distributed without pattern along x-axis around zero. Therefore, there are no outliers found from the residual distribution.

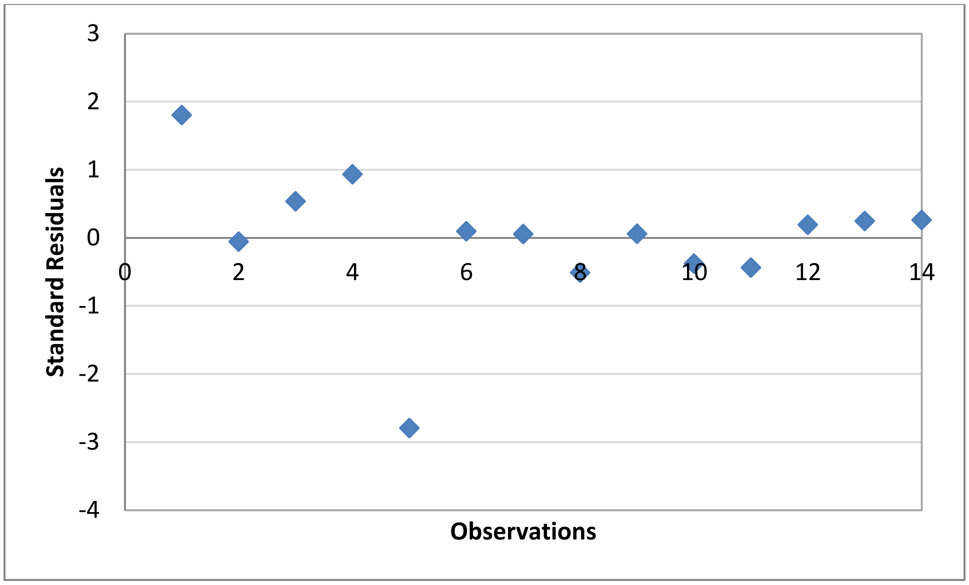


Figure 5.1-2 iXpress 200 Standard Residuals Distribution

Figure 5.1-3 illustrates the sample (14 observations) percentile vs. average weekday boardings. It shows that an 87.71% probability output of the sample data fits a normal distribution. The residuals distribution results show that the regression model is valid. From the regression analysis of the above four output parts, it can be observed that the prediction model for iXpress 200 is statistically significant and reliable.

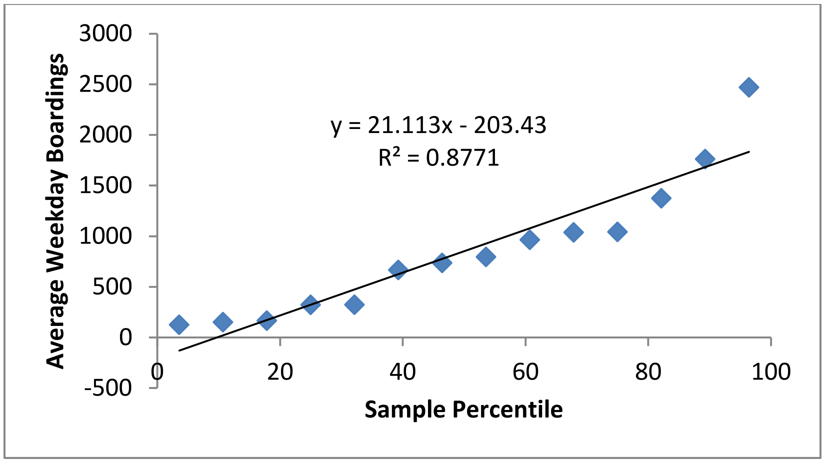


Figure 5.1-3 iXpress 200 Normal Probability Plot of the Standard Residuals

5.1.5 Verification of the iXpress 200 Prediction Model Accuracy

Most transit agencies measure the accuracy and reliability of their prediction models based on the comparison of actual ridership to predicted ridership (Dan Boyle & Associates, 2006). In this thesis, the last 5 weekday boardings (Oct. 5 - Oct. 11, 2011) from 14 stops are used to verify the accuracy and reliability of the prediction model. Based on Equations (4.1), (4.18), (5.4) and the following equations:

$$\boldsymbol{Boardings\_Average\_Weekday_{\_i}} = \frac{\mathbf{1}}{\mathbf{5}\,\boldsymbol{weekdays}} \sum_{\mathbf{Oct.5}}^{\boldsymbol{Oct.11}} \boldsymbol{Boardings\_Weekday_{\_ki}} \qquad (\mathbf{5.7})$$

$$\boldsymbol{AverageError\%_{\_i}} = \frac{\boldsymbol{Average\ Error_i}}{\boldsymbol{Actual\ average\ weekday\ boardings\_i}} \qquad (\mathbf{5.8})$$

$$\lambda_{\_i} = \frac{(\boldsymbol{Actual} \max \boldsymbol{boardings_i} - \boldsymbol{Actual} \min \boldsymbol{boardings_i})}{\boldsymbol{Actual\ average\ weekday\ boardings\_i}} \qquad (\mathbf{5.9})$$

Where,

- i: Stop number
- k: A given weekday
- λ: Refer to the dispersion degree of actual average weekday boardings

The actual average weekday boardings (from October 5 to October 11, 2011) are compared to the predicted average weekday boardings. The average error rate and the dispersion degree of the actual average weekday boardings for each stop are calculated in Table 5.1-4 and illustrated in Figure 5.1-4.

Table 5.1-4 iXpress 200 Average Error Rate and The Dispersion Degree

Stop Number	Actual Average	Predicted Value	Average Errors	Average Error%	Dispersion Degree λ	abs (Error%)
1 (Conestoga Mall)	849	777	72	8.44%	26.50%	8.44%
2 (McCormick)	956	1046	-90	-9.42%	23.35%	9.42%
3 (R & T Park)	115	46	69	59.70%	63.60%	59.70%
4 (U Waterloo)	2220	2330	-110	-4.97%	27.15%	4.97%
5 (Laurier)	736	1204	-468	-63.61%	36.15%	63.61%
6 (Uptown Waterloo)	600	652	-52	-8.69%	21.70%	8.69%
7 (Grand River Hospital)	322	314	8	2.45%	16.05%	2.45%
8 (Victoria)	151	226	-75	-49.77%	61.55%	49.77%
9 (Charles Terminal)	1673	1752	-79	-4.75%	14.10%	4.75%
10 (Ottawa)	163	222	-59	-36.11%	29.25%	36.11%
11 (Fairview Mall)	1282	1438	-156	-12.18%	21.75%	12.18%
12 (Smart Centre)	305	295	10	3.30%	20.70%	3.30%
13 (Cambridge Centre)	673	700	-27	-3.97%	29.60%	3.97%
14 (Ainslie Terminal)	910	926	-16	-1.73%	18.50%	1.73%
Sum:	10955	11930	-975		**Average Error%=**	19.22%

Note:
- Actual Average: Actual average of the rest 5 weekdays' boardings;
- Predicted: Predicted boardings calculated from the model
- Average Error: Actual Average Boardings - Predicted value,
- Aver Error% : Average Error/Actual Average,
- Dispersion Degree λ: (Actual Max Boardings - Actual Min Boardings)/Actual Average Boardings

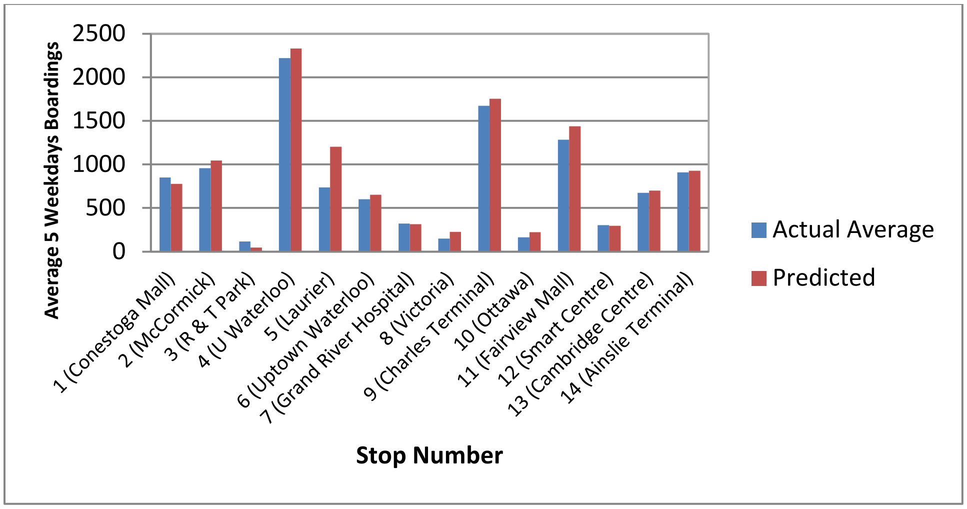


Figure 5.1-4 iXpress 200 Average 5 Weekday Boardings: Actual versus Predicted

Based on the dispersion degrees in the table above, all stops except for Laurier and Ottawa are assumed to be acceptable; the average error rates of Laurier and Ottawa stops are beyond their dispersion degrees, respectively. The reasons can be explained as follows.

The Laurier stop is over-estimated 63.61% higher than its actual average 5 weekday boardings. This is because the numbers of feeder buses that arrive at Laurier stop are higher. These feeder buses provide services around the campus boundary of Laurier University. A site survey needs to be taken so that only the feeder buses that arrive at iXpress 200 stop spot are counted.

The Ottawa stop is over-estimated 36.11% higher than its actual average 5 weekday boardings. Based on the stop ridership distribution, there is no big trip-attraction at the stop. Based on the result of IV1 extraction, there is higher population density at this stop. It is estimated that the most local residents may prefer driving to taking the bus and cause the predicted value to be higher.

The actual average weekday boardings versus predicted values are geographically illustrated for each stop in Figure 5.1-5. From the ridership distribution, the big average weekday boarding volumes are revealed at Davis Centre, Conestoga Mall, McCormick, Charles Terminal, Fairview Mall, and Ainslie Terminal stops. Most trips are generated due to trip-attractions by shopping malls, educational institutions, work places, and bus terminals. The McCormick stop has most trips from stop-based buffer area because there are many students living there.

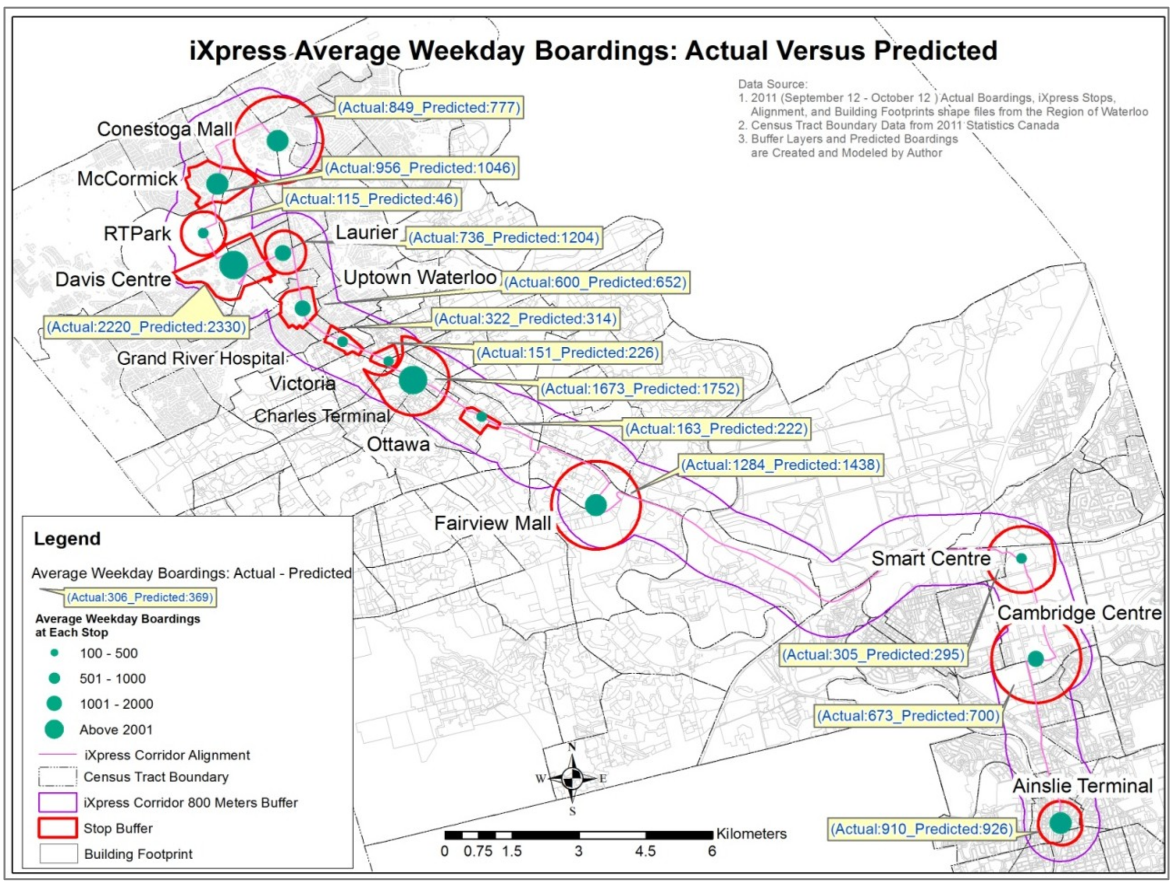


Figure 5.1-5 iXpress 200 Average Weekday Boarding Volume at Each Stop

5.2 Regression Results Analysis for Route 12

The same regression analysis processes are performed for Route 12 with randomly selected 8 stops and 3 segments. Each stop has 22 weekday boardings records in November, 2012. The 17 weekday boardings are used for coefficient estimation; the last 5 weekday boardings are used for verifying the prediction model. The four output parts of the regression results include regression statistics, regression properties, regression coefficients, and regression residuals. They will be analyzed as follows.

5.2.1 Regression Statistics Analysis

Table 5.2-1 presents the regression statistics results from Excel's regression processing. The R-squared value equals 0.9692, which is near 1. This illustrates that 96.92% of the variance in boardings DV can be explained by the variance of the three independent variables (IVs). The Adjusted R Squared value at 0.9559 is always less than R Squared value. When any new input variable is added to the regression model with more accurate results, the Adjusted R Squared value will increase. The Standard Error is 56; it indicates the variance range of the dependent variable. To add more observations can reduce the standard error.

Table 5.2-1 Route 12 Regression Statistics Analysis

Regression Statistics	
Multiple R	0.9845
R Square	0.9692
Adjusted R Square	0.9559
Standard Error	56
Observations	11

5.2.2 Regression Properties Analysis

Table 5.2-2 illustrates the validity of the regression results are confirmed by the small significance F (1.18E-05) at 0 percent level, meaning that there is a strong evidence that the regression model is significantly valid (Duever, 2006; Kashef, 2014). The regression model properties for Route 12 are defined the same as those for iXpress 200.

Table 5.2-2 Route 12 Regression Model Properties and Significance F

Analysis of Variance					
	df	**SS**	**MS**	**F**	**Significance F**
Regression	3	SSR: 690519.9	MSR: 230173.3	$\frac{MSR}{MSE}$=73.358	1.18E-05
Residual	7	SSE: 21963.6	MSE: 3137.657		
Total	10	SST: 712483.5			

5.2.3 Regression Coefficients Analysis

In Table 5.2-3, the coefficients column lists the Least Squares estimates of the three independent variables. The acceptable P-values for the three independent variables are less than 0.002, therefore, there is a strong evidence that the three coefficients are valid (Duever, 2006; Kashef, 2014). These coefficients can be predicted as true with a 95% level of confidence.

Table 5.2-3 Route 12 Regression Coefficients Analysis

	Coefficients	**P-value**	**Lower 95%**	**Upper 95%**
Intercept	-1.790	0.961	-84.497	80.918
IV1	0.107	0.002	0.055	0.159
IV2	0.240	0.0003	0.150	0.329
IV3	0.058	0.0001	0.040	0.077

The following is hypothesis testing on the regression model coefficients, assuming that the three coefficients of the three IVs are B1, B2, and B3.

Then,

Null hypothesis H0: $\boldsymbol{B1 = B2 = B3 = 0}$ **(5.5)**

Alternative hypothesis Ha: At least one coefficient is not 0

Test statistic: f= $\boldsymbol{\frac{\text{MSR}}{\text{MSE}} = \frac{230173}{3137} = 73}$ **(5.6)**

Where

MSR: Mean Square Regression

MSE: Mean Square Error

The critical value: $\boldsymbol{F0.05, 3, 7 = 4.4\ (From\ the\ table\ of\ F - statistics\ P = 0.05)}$ **(5.7)**

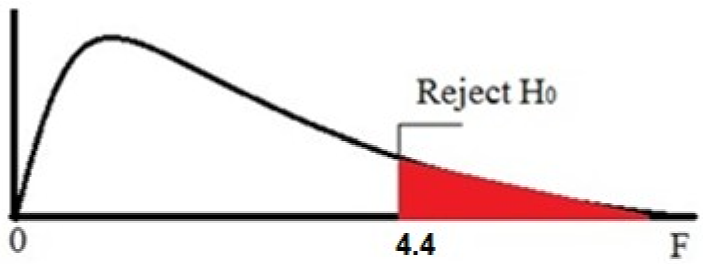


Figure 5.2-1 Hypothesis Testing on Route 12 Regression Model Coefficients

Conclusion: Since the value of the test statistic equals to 73 which is greater than the critical value 4.4, thus it falls in the rejection region, the null hypothesis is rejected. There is strong evidence that at least one of the coefficients is not zero and contributes to ridership - average weekday boardings. Therefore, the regression coefficients are shown valid.

5.2.4 Regression Residuals Analysis

The residual distributions are visualized in Figure 5.2-2. From Figure 5.2-2, the residual dots are randomly distributed without pattern around zero along the x axis. The standard residuals range is greater than negative 1.8 and less than positive 1.6, and thus no outliers are found. The residual distribution results show that the regression model is valid.

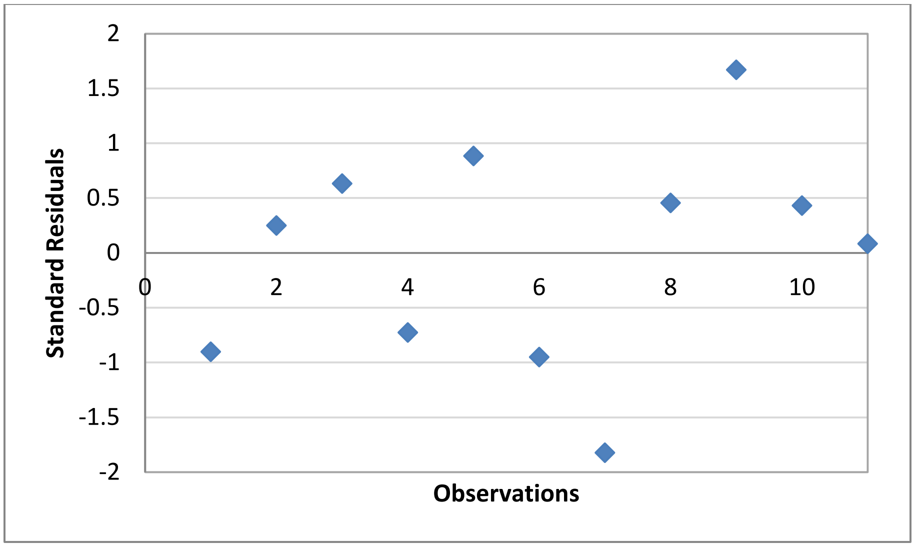


Figure 5.2-2 Route 12 Standard Residuals Distribution

Figure 5.2-3 illustrates that the sample percentile vs. average weekday boardings. This plot determines that a 93.58% probability output of the sample data fits a normal distribution. The plot also shows the regression model for Route 12 is valid.

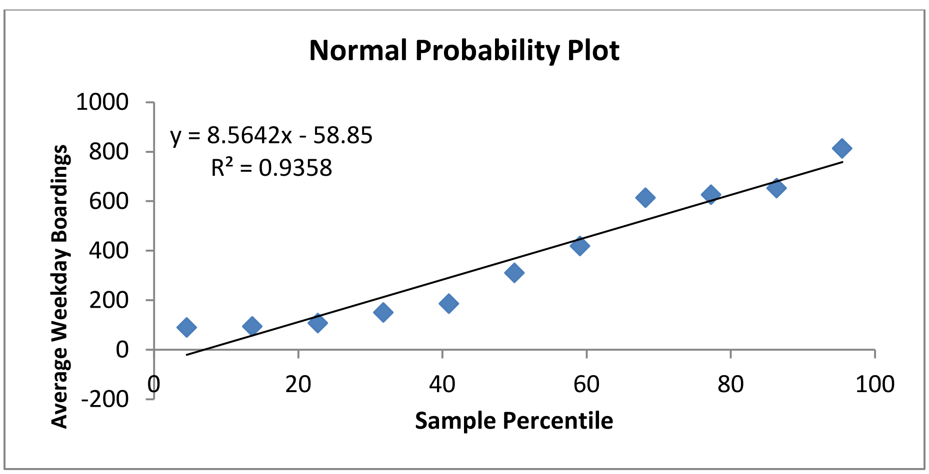


Figure 5.2-3 Route 12 Normal Probability Plot of the Standard Residuals

Therefore, from the above regression analysis of the four output results, the prediction model for Route 12 is statistically significant and reliable.

5.2.5 Verification of the Route 12 Prediction Model Accuracy

The prediction model accuracy is verified through comparing the actual observation values to the predicted boardings. The TCRP report in Table 31 shows that 94% of transit agencies compare actual ridership to predicted value in order to assess the reliability and validity of their prediction methodologies (Dan Boyle & Associates, 2006). Based on Equations (4.1), (4.18), (5.4), (5.7), (5.8), (5.9), the average error rate and the dispersion degree of the average weekday boardings for each stop are calculated in Table 5.2-4 and illustrated in Figure 5.2-4.

Table 5.2-4 Route 12 Average Error Rate and the Dispersion Degree

Stop_No	Actual Average	Predicted Value	Average Errors	Average Error%	Dispersion Degree λ	abs(Error%)
1 (Conestoga Terminal)	297	353	-56	-18.75%	125.96%	18.75%
2 (Bridge/University)	94	83	12	12.37%	81.78%	12.37%
3 (Lincoln/Bluevale)	186	156	30	15.88%	47.71%	15.88%
4 (S-Laurier)	633	659	-26	-4.13%	56.37%	4.13%
5 (S-U Waterloo)	838	769	69	8.22%	43.01%	8.22%
6 (Westmount/Erb)	114	135	-20	-17.61%	154.65%	17.61%
7 (S-Westmount/Highland)	414	504	-90	-21.69%	46.73%	21.69%
8 (Westmount/Ottawa)	152	129	22	14.79%	51.84%	14.79%
9 (Forest Glen Terminal)	597	536	62	10.32%	39.75%	10.32%
10 (Fairway/Food Basics)	103	87	16	15.23%	90.16%	15.23%
11 (Fairview Terminal)	659	649	10	1.53%	55.31%	1.53%
Sum=	4088	4059	28		Average Error%=	12.77%

Note:
- Actual Average: Actual average of the rest 5 weekdays' boardings;
- Predicted: Predicted boardings calculated from the model
- Average Error: Actual Average Boardings - Predicted value,
- Aver Error% : Average Error/Actual Average,
- Dispersion Degree λ: (Actual Max Boardings - Actual Min Boardings)/Actual Average Boardings

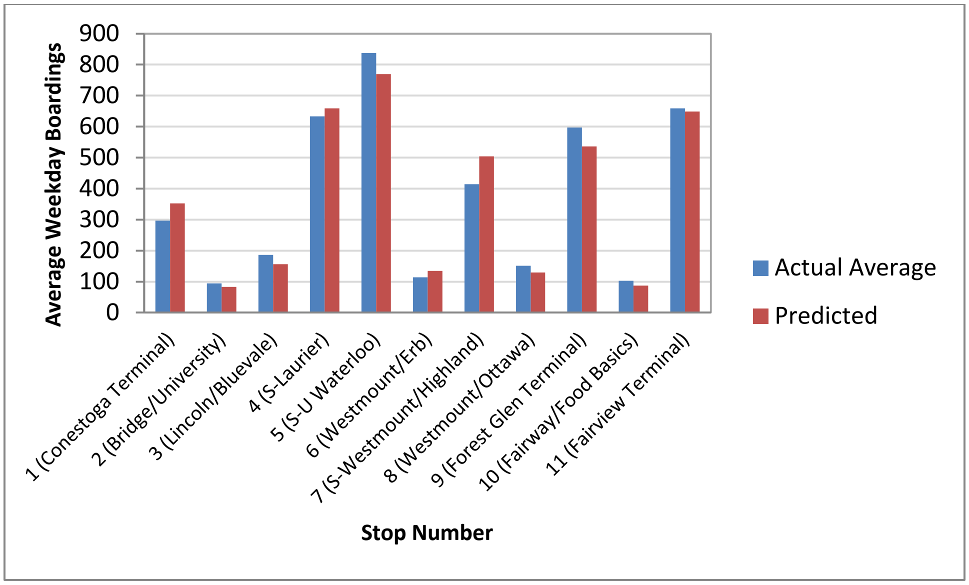


Figure 5.2-4 Route 12 Average 5 Weekdays Boardings: Actual versus Predicted

Based on the dispersion degrees in Table 5.2-4, all stops are assumed to be acceptable; the average error rate is at 12.77%. Route 12 introduced segment-based buffer to solve the overlapping area double counting problem caused by close-by stops in buffering. Therefore, the prediction errors are minimized.

The actual average weekday boardings versus predicted values are geographically illustrated for each stop in Figure 5.2-5. The big average weekday boarding volumes fluctuate at Conestoga Mall, S-U Waterloo, Forest Glen Plaza, and Fairview Mall stops. Most trips are generated due to stop attractions from shopping malls, educational institutions, bus terminal and work places.

From the analysis and validation above, the ridership modeling method is feasible for Route 12 and the accuracy of the model is reliable.

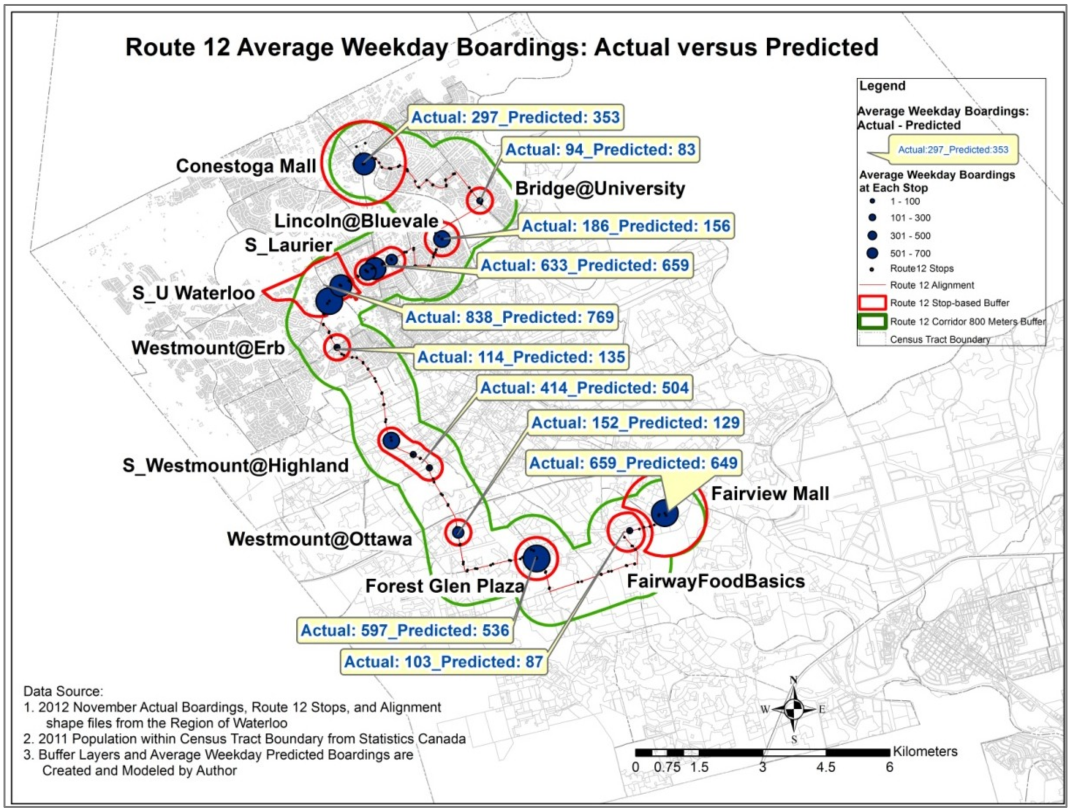


Figure 5.2-5 Route 12 Average Weekday Boarding Volume at Each Stop

5.3 Comparison of iXpress 200 to Route 12

Section 5.1 and 5.2 analyzed and verified the reliability and validity of the regression models of iXpress and Route 12. This section will compare iXpress 200 to Route 12 in regression results and ridership volume at the four stops - Conestoga Mall, U Waterloo, Laurier, and Fairview Mall stops. The purpose is to assess the contribution of each independent variable with its weight coefficient to ridership, and to prove the advance of iXpress 200 over Route 12. In addition, the accuracy of the two prediction models is analyzed for the four stops.

5.3.1 The Comparison of the Prediction Models for iXpress 200 and Route 12

Table 5.3-1 compares and summarizes the regression results for the two prediction models of iXpress 200 and Route 12. From the Table 5.3-1 comparison, the contribution of the three weight coefficients to ridership reveals the preference of riders on iXpress 200. Route 12 has a lower average error rate than that of iXpress 200 because of the introduction of segment-based buffer for solving the overlapping area double counting problem caused by the close-by stops in buffering. However, the modeling method appears to be feasible for the two bus routes.

Table 5.3-1 The Comparison of the Prediction Models for iXpress 200 and Route 12

The Elements of the Two Prediction Models	iXpress 200	Route 12	Analysis and comarison
Observations	14	11	iXpress has 1.3 times observations than Route 12
R^2	95.26%	96.92%	95.26% explained for iXpress; 96.92% explained for Route 12
Intercept	-58.75	-1.79	
The weight coefficient of IV1	0.246	0.107	The weight coefficient of IV1 for iXpress has 2.3 times contribution to boardings than that of Route 12, meaning that the local riders prefer iXpress.
The weight coefficient of IV2	0.220	0.240	The weight coefficient of IV2 for iXpress has 0.9 times contribution to boardings than that of Route 12, meaning that the riders from transit network to iXpress are less than to Route 12 because iXpress has few stops to transfer.
The weight coefficient of IV3	0.398	0.058	The weight coefficient of IV3 for iXpress has 6.9 times contribution to boardings than that of Route 12, meaning that the riders from route level prefer iXpress.
Errors (Observed-Predicted)	-975	28	iXpress focuses on stop-based buffer size, however, Route 12 combines stop and segment-based buffer size to minimize the estimation errors.
Average Error Rates (Average Errors/Actual Average)	19.22%	12.77%	iXpress focuses on stop-based buffer size, however, Route 12 combines stop and segment-based buffer size to minimize the estimation error rate.

5.3.2 The Comparison of the Prediction Accuracy for iXpress 200 and Route 12

iXpress 200 and Route 12 serve the same four stops at Conestoga Mall, U Waterloo, Laurier, and Fairview Mall. The prediction accuracy of the two prediction models at the four stops can be compared in Table 5.3-2 and illustrated in Figure 5.3-1.

Table 5.3-2 The Comparison of Average Error Rates for iXpress 200 and Route 12

Boardings	**iXpress 200**				**Route 12**			
Stop_Name	Actual Average	Predicted Average	Average Error	Average Error%	Actual Average	Predicted Average	Average Error	Average Error%
Conestoga Mall	849	777	72	8.44%	297	353	-56	-18.75%
U Waterloo	2220	2330	-110	-4.97%	838	769	69	8.22%
Laurier	736	1204	-468	-63.61%	633	659	-26	-4.13%
Fairview Mall	1282	1438	-156	-12.18%	659	649	10	1.53%

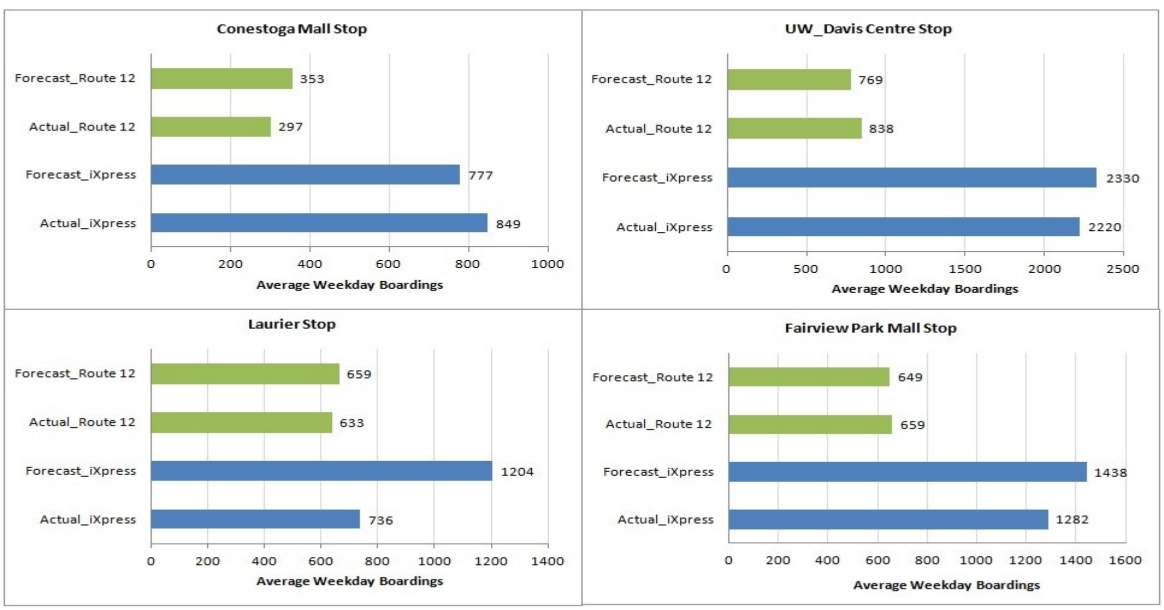


Figure 5.3-1 The Comparison of the Prediction Accuracy for iXpress 200 and Route 12

The comparison of results illustrates the prediction model's accuracy at the four stops.

- Conestoga Mall stop is underestimated 8.44% for iXpress 200 and overestimated 18.75% for Route 12; based on land use types around the stop, there 2/3 industrial and commercial area within 1000 meters stop-based buffer, meaning that 2/3 of riders are from employment data, these employees prefer driving to taking public transit; for Route 12, there are two peak boardings on November 16 and November 23 from APCS records, and the two peaks caused the big average error rate at Conestoga mall stop for Route 12
- U Waterloo (Davis Centre) stop is overestimated 4.97% for iXpress 200 and underestimated 8.22% for Route 12, respectively. iXpress 200 has 3 times boardings than that of Route 12 at the stop
- Laurier stop is overestimated 63.61% for iXpress 200 and 4.13% for Route 12. The number of feeder buses is counted higher at the stop for iXpress 200. A site investigation needs to be taken so that only the number of feeder buses that arrive at iXpress 200 stop-spot can be counted
- Fairview Mall is overestimated 12.18% for iXpress 200 and underestimated 1.53% for Route 12. iXpress 200 has 2 times boardings than that of Route 12 for the stop, and its sample data used for verifying accuracy went down than the sample data used for regression processing at the stop. New sample data can be collected to verify the estimation error at the stop

5.4 The Validation of the Modeling Method by Different Transit Service Time

The same ridership modeling method is applied for iXpress 200 for different service periods - morning peak (6:00 - 10:00) am, off-peak (10:00 am - 3:00 pm), and afternoon peak (3:00 - 7:00) pm.

The regression results are compared in Table 5.4-1 (detail regression results are shown in Appendix A, Appendix B, and Appendix C).

Table 5.4-1   iXpress 200 Regression Results Comparison during Different Service Time

Service Period	**Morning Peak** **(6:00-10:00)am**	**Off-Peak** **10:00am-3:00pm**	**Afternoon Peak** **(3:00-7:00)pm**
R^2	0.9714	0.9875	0.9645
Prediction Model	DV = -12 + 0.136 * IV1 + 0.117 * IV2	DV = -27 + 0.092 * IV1 + 0.274 * IV2 + 0.129 * IV3	DV = 0.714 + 0.441 * IV2 + 0.179 * IV3
Estimation Errors **(Observed - Predicted)**	-52	-251	-412

From Table 5.4-1, morning peak boarding sources are mainly from stop-based buffer and feeder buses services based on normal traveler behaviors in the real world. High boarding volumes can be assumed from stop-based buffer before 10:00 am because riders need to commute for work, study, and other businesses trips etc.. Also non-local riders start boarding can be assumed after 10:00 am, for example, shopping malls open after 10:00 am. Therefore, the IV3 is near zero during morning peak period. Moreover, 97.14% of variance in morning peak boardings can be explained by the variance of the IV1 and IV2 with the estimation errors at 52 riders' difference.

Off-peak boarding sources are mainly from stop-based buffer, feeder buses services, and riders from route level along the bus route based on normal traveler behavior in the real world. For example, students, discretional riders, part time jobs, and irregular routine riders, and so on. Moreover, 98.75% of the variance in off-peak boardings can be explained by the variance of the three independent variables - IV1, IV2, and IV3 with the estimation errors at 251 riders' difference.

Afternoon peak boarding sources are mainly from non-local riders from route level along the bus route and network level, for example, riders from work/school to home, from work/school to shopping malls, and from work/school to bus terminals, and so on. Moreover, 96.45% of the variance in afternoon peak boardings can be explained by the variance of the two variables - IV2 and IV3 with the estimation errors at 412 riders' difference.

From the regression results in the different service periods, the estimation errors can be traced back to the high number of feed buses that arrive at the Laurier stop. The solution is that the site investigation needs to be taken so that only the feeder buses that arrive at iXpress stop-spot at Laurier stop can be counted.

The boarding volume in the different service periods can be illustrated in Figure 5.4-1. From Figure 5.4-1, the boarding volume can be clearly seen for morning peak (6:00-10:00) am with IV1 (residents) and IV2 (number of feeder buses that arrive); off-peak (10:00 am - 3:00 pm) with IV1, IV2, and IV3 (non-local riders from other origins along the bus route); and afternoon peak (3:00-7:00) pm with IV2 and IV3.

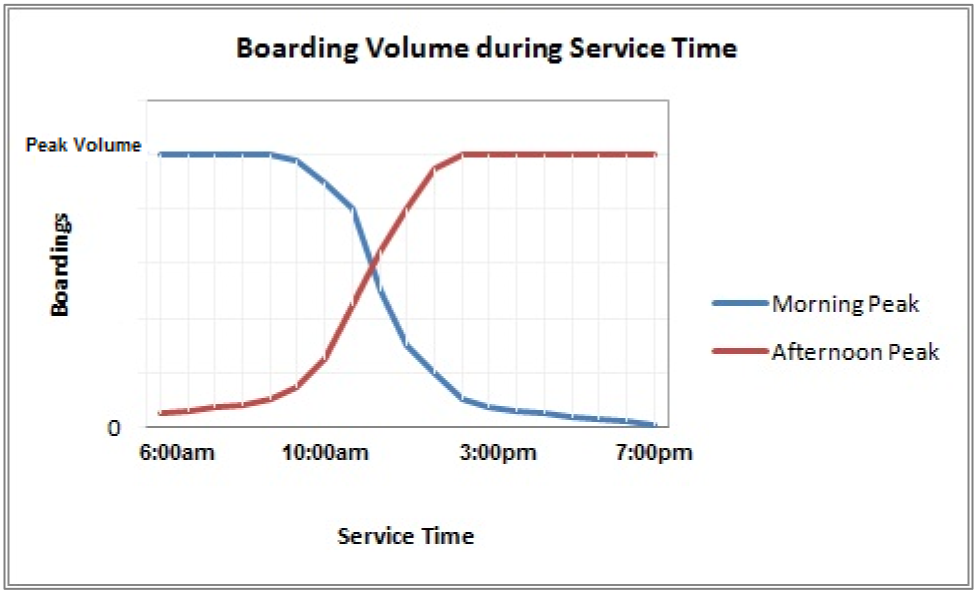


Figure 5.4-1 Boarding Volume during Service Time

Based on the APCS data analysis and land use characteristics for each stop, and the modeling results for the two bus routes during the different service period, the ridership distribution trend can be summarized as 4 types at different stops (see Table 5.4-2).

Table 5.4-2 Ridership Distribution Trend at Different Types of Stops

➢ Stop at Shopping Mall and Bus Terminal ○ Low/middle trip production ○ High transfer rate ○ High trip attraction	➢ Stop at Business Park/Hospital ○ Nearly no/low trip production ○ No/low transfer rate ○ Low trip attraction
➢ Stop at University ○ Low trip production ○ Low transfer rate ○ Very high trip attraction	➢ Stop at Residential Community ○ High trip production ○ Nearly zero transfer rate ○ Very low trip attraction

Table 5.4-2 summarizes ridership distribution trend at different types of stops. The stop categorization can assist transit agencies in data collection and analysis for modeling.

As a result, based on the regression results analysis, the modeling method is feasible for other bus routes with different service periods.

5.5 Chapter Summary

Chapter 5 analyzed the reliability and validity of the regression output results, and verified the accuracy of the prediction models of iXpress 200 and Route 12. By comparing the elements of the prediction model of iXpress 200 to that of Route 12, 70% of residents within stop-based buffer area would prefer iXpress 200; 48% of transferring riders would transfer to iXpress 200; 87% of non-local riders from route level would prefer to choose iXpress 200 due to its excellent performance, such as few stops, few delay, less travel time and short headway.

Based on the validation and verification results of the prediction models of the two bus routes, the ridership modeling method is suitable for stop-based and segment-based ridership prediction. It is expected to apply for other bus routes ridership estimations; and provide assistance for transportation planning and management.

Chapter 6 Conclusions and Future Work

Chapter 5 validated the regression results and the modeling method, verified the accuracy of the prediction models. It has been proved that the modeling method is feasible for stop level ridership estimation. Chapter 6 will summarize the thesis study and vision the future work.

6.1 Conclusions

This thesis developed a ridership modeling method at stop level based on multiple linear regression theory and GIS analysis tools. The modeling method developed is applied to the iXpress 200 and the conventional bus Route 12 in Waterloo Region. The key independent variables (IVs) directly related to ridership are identified first. Then the data extraction methods for dependent variable (DV) and IVs are illustrated step by step. The Trip Production / Trip Attraction matrices are created and the ridership prediction models are developed.

These models have been proven to be valid and reliable by the four part regression outputs analyses: for the two models of iXpress 200 and Route 12 within the average weekday period, the R-Squared of iXpress 200 is at 95.26%, and that of Route 12 is at 96.92%. The regression properties analysis for the two bus routes are at the Significance $F \approx 0$. The related P-values of the three regression coefficients are less than 0.03. Moreover, the regression standard residuals are less than 3 with random distribution without pattern along x axis around zero for the two bus routes. In addition, the accuracy of the prediction models is verified by comparing the new sample values to the predicted ridership for all stops. The results show that the predicted average error rates are at 19.22% for iXpress 200 and 12.77% for Route 12 within the range of their sample variation. The verification results prove that the modeling method is feasible for the two bus routes; and the two prediction models are reliable and valid.

From the regression results, the weight coefficient of IV1 for iXpress 200 has 2.3 times contribution to boardings than that of Route 12, meaning that the local riders prefer iXpress 200; the weight coefficient of IV2 for iXpress 200 has 0.9 times contribution to boardings than that of Route 12, meaning that the riders from feeder buses to iXpress 200 are less than to Route 12 due to fewer stops to transfer; the weight coefficient of IV3 for iXpress 200 has 6.9 times contribution to boardings than that of Route 12, meaning that the riders along the bus route prefer iXpress 200. Based on the analysis and comparison of the two prediction models elements, the results prove the advantage and success of the iXpress 200 over Route 12.

Moreover, the modeling method is also applied to the different service periods of iXpress 200 - morning peak (6:00 - 10:00) am, off-peak (10:00 am - 3:00 pm), and afternoon peak (3:00 - 7:00) pm. The R-Squared is 97.14%, 98.75%, and 96.45% respectively; the regression properties analysis at the Significance F ≈ 0; the three P-values for the related regression coefficients are less than 0.04; and the regression standard residuals are less than 3 with random distribution without pattern along x axis around zero for the three service periods. The regression results are proven to be valid and reliable, and the modeling method appears to be feasible for different service periods as well.

To summarize the key output:

- A Ridership modeling method
- The identification of Key IVs
- The extraction methods of DV and IVs

- Ridership prediction models
- Validation of the prediction models and verification of accuracy of the prediction models
- Guidance on ridership modeling method step by step at stop levels

The models can be directly analyzed to reflect the spatial variation (i.e., land use and socio-economic change from IV1, transit network change from IV2, each stop environment and nature change from IV3). Transit agencies can also use the modeling method to develop other different prediction models at stop level for environmental assessments and transportation impact studies; to predict transit market demands, plan new routes, or evaluate existing transit routes.

6.2 Future Work

- Stop environments can be categorized in order to develop different types of prediction models for other bus routes based on the modeling method. For example, stop-based residential, industrial, educational, and business buffer area prediction models
- More independent variables can be explored to enhance the multiple linear regression prediction models. For example, in residential area, to explore income level, numbers of household with 0, 1, 2 or more cars, retired households; in shopping malls, to explore bus service frequency, parking cost, fare within each stop's buffer zone; in business parks, to survey the possibility of employees taking buses
- Better ways of data collection can be explored. For example, a transit demand real-time monitoring and management system can be developed based on ArcGIS or Intergraph platforms for the Region of Waterloo (or any other interested transit agencies) so that socio-

economic data for each stop-based buffer area, origin-destination pairs, etc. transit-related data are available to use for research study

- The ridership prediction models developed can be optimized using Linear Neural Networks (LNN) to solve nonlinear models by an iterative numerical technique (gradient decent)

In order to improve the accuracy of the prediction model, more data need to be collected and tested as the following steps (when these data are available):

- Collect 2011 TAZ shape file with attribute information
- Collect 2011 TTS data
- Process 2011 TTS data to 2011 TAZ zone
- Process 2011 TAZ zone data to Stop-based buffer
- Re-extract IV1 and IV3 based on service time periods - average weekday, morning peak (6:00 - 10:00) am, off - peak (10:00 am - 3:00 pm), afternoon peak (3:00 - 7:00) pm, and off - peak (7:00 pm -12:00 am); here IV3 can be defined as OD pairs from route level
- Finish regression modeling for the different service periods (two directions)
- Test regression modeling for one direction separately (downward and upward), considering the fact that boardings would decrease when a bus is approaching to the end of its route, therefore, the more factors related to distance, or number of stops, or population etc. need to be considered.

Appendix A Ridership Modeling during Morning Peak Period (6:00 am - 10:00 am)

Table A1 Correlation Analysis

	DV	*IV1*	*IV2*
DV	1		
IV1	0.978678	1	
IV2	0.708029	0.630895	1

IV1 contributes 97.87% to DV, IV2 contributes 70.80% to DV, meaning that most boardings are from local residents and network level during morning peak period (riders from route level only alight for work, study, and business etc. during the time slot, their boardings may start after 10:00 am).

There is no high correlation between IV1 and IV2.

Table A2 Regression Statistics Analysis

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.985617
R Square	0.971441
Adjusted R Square	0.966249
Standard Error	23.88684
Observations	14

The regression result shows that 97.14% of the variance in morning peak boardings can be explained by the variance of the IV1 and IV2.

Table A3 Regression Properties Analysis

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	213495.3	106747.7	187.0858	3.21E-09
Residual	11	6276.393	570.5812		
Total	13	219771.7			

The Significance F is near zero, meaning that the regression model is significant valid and the result of the model is not by chance.

Table A4 Regression Coefficients Analysis

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-11.7606	11.44936	-1.02719	0.326391	-36.9605	13.43926
IV1	0.13573	0.010086	13.45672	3.55E-08	0.11353	0.15793
IV2	0.116981	0.051052	2.291398	0.042673	0.004616	0.229346

The acceptable P-value for the two independent variables are less than 0.04, there is a strong evidence in the two coefficients' validity.

Table A5 Regression Residual Analysis

Observation	*Predicted DV_1-17days*	*Residuals*	*Standard Residuals*
1	140	-18	-0.8047
2	366	14	0.643161
3	0	9	0.388888
4	119	17	0.787489
5	114	21	0.955767
6	173	-26	-1.17879
7	97	-5	-0.21952
8	67	-33	-1.49034
9	441	-18	-0.83058
10	76	-34	-1.54378
11	315	3	0.14635
12	57	20	0.91477
13	129	31	1.410971
14	247	18	0.820312

The standard residual dots are randomly distributed without pattern along x-axis around zero.

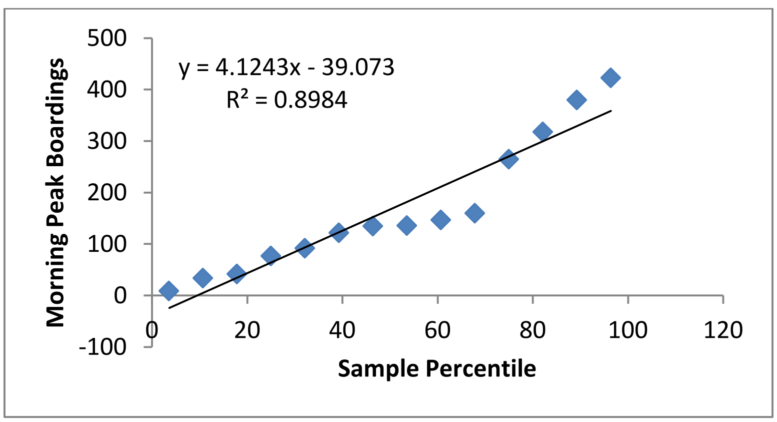


Figure A1 iXpress Normal Probability Plot of the Standard Residuals

Figure A1 illustrates the sample (14 observations) percentile vs. Morning Peak Boardings. It shows that a 89.84% probability output of the sample data fits a normal distribution.

From the regression results, it can be observed that the prediction model for iXpress 200 during morning peak period (6:00 - 10:00) am is statistically significant and reliable.

iXpress 200 Prediction Model during Morning Peak Period (6:00 - 10:00**) am:**

$$\mathbf{DV_{am\text{-}peak} = -12 + 0.136 * IV1 + 0.117 * IV2}$$

Table A6 Average Error Rate and the Dispersion Degree

Stop Number	5weekdaysAver	Predicted	Errors	Error%	λ	abs(Error%)
1 (Conestoga Mall)	110	140	-30	-27.01%	51.80%	27.01%
2 (McCormick)	341	366	-25	-7.44%	46.10%	7.44%
3 (R & T Park)	8	0	8	97.00%	208.30%	97.00%
4 (U Waterloo)	148	119	29	19.80%	21.00%	19.80%
5 (Laurier)	141	114	27	19.17%	60.25%	19.17%
6 (Uptown Waterloo)	126	173	-47	-37.32%	44.75%	37.32%
7 (Grand River Hospital)	93	97	-4	-4.09%	36.40%	4.09%
8 (Victoria)	36	67	-31	-85.16%	85.25%	85.16%
9 (Charles Terminal)	425	442	-17	-3.95%	10.50%	3.95%
10 (Ottawa)	39	76	-37	-94.50%	39.10%	94.50%
11 (Fairview Mall)	324	315	9	2.74%	24.85%	2.74%
12 (Smart Centre)	73	57	16	22.20%	90.25%	22.20%
13 (Cambridge Centre)	157	129	28	17.83%	37.35%	17.83%
14 (Ainslie Terminal)	268	247	21	7.76%	19.70%	7.76%
Sum=	2289	2341	-52		Average=	31.85%

Ottawa stop is overestimated due to high density residents. From the real world observation, the boardings are very low, meaning that most residents go to work by other transportation modes such as by driving instead of taking bus during weekday (Monday - Friday).

Appendix B   Ridership Modeling during off-Peak Period (10:00 am - 3:00 pm)

Table B1   Correlation Analysis

	DV	*IV1*	*IV2*	*IV3*
DV	1			
IV1	0.554831	1		
IV2	0.575688	0.626281	1	
IV3	0.853407	0.099788	0.162655	1

IV3 has the highest correlation coefficient of 0.853 with the DV, meaning that it plays as the most significant independent variable in the prediction model, followed by IV2 and IV1.

There is no high correlation among IVs.

Table B2   Regression Statistics Analysis

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.993732
R Square	0.987504
Adjusted R Square	0.983755
Standard Error	30.90711
Observations	14

The regression result shows that 98.75% of the variance in off-peak boardings period (10:00 am - 3:00 pm) can be explained by the variance of the IV1, IV2, and IV3.

Table B3 Regression Properties Analysis

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	754893.5	251631.2	263.4193	8.2E-10
Residual	10	9552.495	955.2495		
Total	13	764446			

The Significance F is near zero, meaning that the regression model is significant valid and the result of the model is not by chance.

Table B4 Regression Coefficients Analysis

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-26.517	16.07848	-1.64922	0.130118	-62.3421	9.308104
IV1	0.092334	0.012988	7.108975	3.26E-05	0.063394	0.121274
IV2	0.27433	0.050836	5.396412	0.000303	0.161061	0.387599
IV3	0.128814	0.005908	21.8022	9.22E-10	0.115649	0.141978

The acceptable P-value for the three independent variables are less than 0.0003, there is a very strong evidence in the three coefficients' validity.

Table B5 Regression Residual Analysis

Observation	*Predicted DV*	*Residuals*	*Standard Residuals*
1	263.7326	12.26735	0.452547
2	365.0436	0.956446	0.035284
3	8.071083	17.92892	0.661405
4	869.9405	14.05951	0.518661
5	329.8837	-40.8837	-1.50822
6	227.591	14.40898	0.531553
7	104.3375	30.66245	1.13115
8	74.67407	-15.6741	-0.57822
9	625.0804	23.91959	0.882403
10	72.51698	-17.517	-0.64621
11	498.8	-62.8	-2.31672
12	94.30624	-9.30624	-0.34331
13	240.0875	1.91251	0.070553
14	341.9347	30.06525	1.109119

The standard residual dots are randomly distributed without pattern along x-axis around zero.

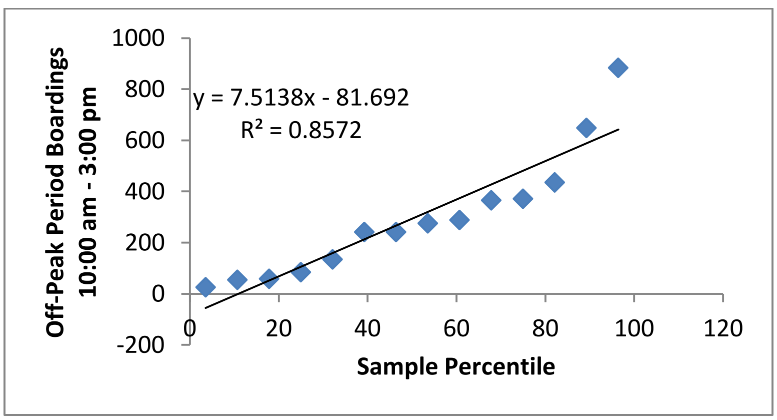


Figure B1 iXpress Normal Probability Plot of the Standard Residuals

Figure B1 illustrates the sample (14 observations) percentile vs. Morning Peak Boardings. It shows that a 85.72% probability output of the sample data fits a normal distribution.

From the regression results, it can be observed that the prediction model for iXpress 200 during off-peak period (10:00 am - 3:00 pm) is statistically significant and reliable.

iXpress 200 Prediction Model during off-Peak Period (10:00 am - 3:00 pm**):**

$$DV_{\text{off-peak}} = -27 + 0.092 * IV1 + 0.274 * IV2 + 0.129 * IV3$$

Table B6 Average Error Rate and the Dispersion Degree

Stop Number	5weekdaysAver	Predicted	Errors	Error%	λ	abs(Error%)
1 (Conestoga Mall)	246	264	-18	-7.13%	34.90%	7.13%
2 (McCormick)	361	364	-3	-0.92%	20.60%	0.92%
3 (R & T Park)	26	8	18	68.93%	101.25%	68.93%
4 (U Waterloo)	804	765	39	4.79%	32.75%	4.79%
5 (Laurier)	263	429	-166	-63.28%	53.35%	63.28%
6 (Uptown Waterloo)	227	227	0	-0.12%	46.95%	0.12%
7 (Grand River Hospital)	131	104	27	20.49%	40.35%	20.49%
8 (Victoria)	52	75	-23	-43.34%	75.30%	43.34%
9 (Charles Terminal)	606	624	-18	-3.00%	21.35%	3.00%
10 (Ottawa)	51	72	-21	-41.88%	59.15%	41.88%
11 (Fairview Mall)	416	498	-82	-19.78%	34.85%	19.78%
12 (Smart Centre)	89	94	-5	-5.89%	76.55%	5.89%
13 (Cambridge Centre)	230	240	-10	-4.31%	25.35%	4.31%
14 (Ainslie Terminal)	354	341	13	3.56%	16.30%	3.56%
	3856	4107	-251		Average=	20.53%

The error rate at Laurier stop is beyond its dispersion degree. The error can be traced back to the data collection at the stop. Numbers of feeder buses that arrive need to be re-counted, only feeder buses that arrive at iXpress 200 stop-spot can be counted. Number of students needs to be confirmed as well.

Appendix C Ridership Modeling during Afternoon Peak Period (3:00 pm - 7:00 pm)

Table C1 Correlation Analysis

	DV	*IV2*	*IV3*
DV	1		
IV2	0.450145	1	
IV3	0.943499	0.196668	1

IV3 has the highest correlation coefficient of 0.943 with the DV, meaning that it plays as the most significant independent variable in the prediction model, followed by IV2 at 0.450.

There is no high correlation among IVs.

Table C2 Regression Statistics Analysis

Regression Statistics	
Multiple R	0.960485
R Square	0.922532
Adjusted R Square	0.908447
Standard Error	82.04316
Observations	14

The regression result shows that 92.25% of the variance in afternoon peak boarding period (3:00 - 7:00) pm can be explained by the variance of the IV2 and IV3 (from the observation in the real world, most of IV1 show alighting during afternoon peak period).

Table C3 Regression Properties Analysis

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	881734.1	440867.1	65.49722	7.77E-07
Residual	11	74041.89	6731.081		
Total	13	955776			

The Significance F is near zero, meaning that the regression model is significant valid and the result of the model is not by chance.

Table C4 Regression Coefficients Analysis

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0.713981	33.7767	0.021138	0.983514	-73.628	75.056
IV2	0.441356	0.137771	3.203542	0.008402	0.138124	0.744589
IV3	0.179226	0.017892	10.01702	7.27E-07	0.139846	0.218607

The acceptable P-value for the two independent variables are less than 0.008, there is a very strong evidence in the two coefficients' validity.

Table C5 Regression Residual Analysis

Observation	*Predicted DV(17wks days)*	*Residuals*	*Standard Residuals*
1	287.5619	111.4381	1.476612
2	194.6186	20.38138	0.270064
3	37.27614	26.72386	0.354105
4	991.2689	72.73111	0.963725
5	494.5297	-232.53	-3.08114
6	182.1743	10.82572	0.143446
7	80.00369	-5.00369	-0.0663
8	69.07659	-25.0766	-0.33228
9	521.6523	1.347734	0.017858
10	55.51421	-3.51421	-0.04657
11	448.8981	10.10192	0.133856
12	103.9452	14.05478	0.186233
13	246.8375	4.162465	0.055155
14	262.6428	-5.64281	-0.07477

The standard residual dots are randomly distributed without pattern along x-axis around zero (except Laurier stop).

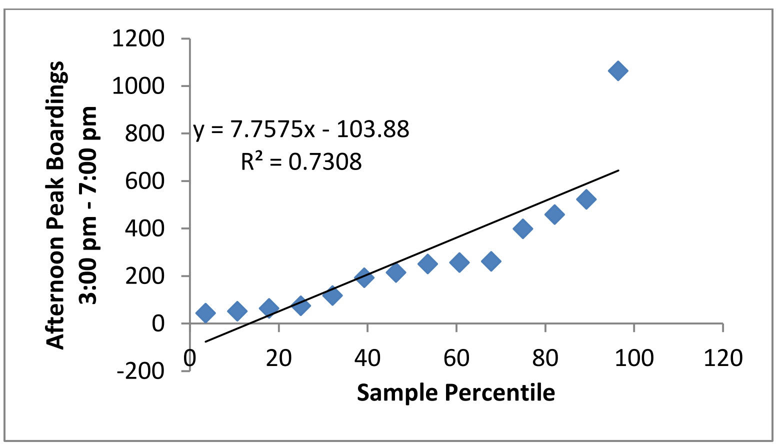


Figure C1 iXpress Normal Probability Plot of the Standard Residuals

Figure C1 illustrates the sample (14 observations) percentile vs. afternoon peak boardings. It shows that a 73.08% probability output of the sample data fits a normal distribution.

From the regression results, it can be observed that the prediction model for iXpress during afternoon peak period (3:00 - 7:00) pm is statistically significant and reliable.

iXpress 200 Prediction Model during Afternoon Peak Period (3:00 - 7:00) pm**:**

$$\mathbf{DV_{afternoon\ peak} = 0.714 + 0.441 * IV2 + 0.179 * IV3}$$

Table C6 Average Error Rate and the Dispersion Degree

Stop_Number	DV(5wks days)	Predicted	Errors	Error%	λ	abs(Error%)
1 (Conestoga Mall)	329	287	42	12.69%	31.30%	12.69%
2 (McCormick)	193	194	-1	-0.71%	50.00%	0.71%
3 (R & T Park)	62	37	25	39.95%	79.65%	39.95%
4 (U Waterloo)	953	990	-37	-3.89%	39.80%	3.89%
5 (Laurier)	233	494	-261	-111.99%	50.15%	111.99%
6 (Uptown Waterloo)	188	182	6	3.21%	49.55%	3.21%
7 (Grand River Hospital)	78	80	-2	-2.44%	37.40%	2.44%
8 (Victoria)	43	69	-26	-60.46%	91.30%	60.46%
9 (Charles Terminal)	485	521	-36	-7.45%	23.85%	7.45%
10 (Ottawa)	56	55	1	0.99%	51.40%	0.99%
11 (Fairview Mall)	397	448	-51	-12.95%	44.00%	12.95%
12 (Smart Centre)	105	104	1	1.13%	70.45%	1.13%
13 (Cambridge Centre)	217	247	-30	-13.62%	73.40%	13.62%
14 (Ainslie Terminal)	220	262	-42	-19.26%	70.40%	19.26%
Sum=	3559	3971	-412		Averasge=	20.77%

The error rate at Laurier stop is beyond its dispersion degree. The error can be traced back to the data collection at the stop. Numbers of feeder buses that arrive need to be re-counted, only feeder buses that arrive at iXpress 200 stop-spot can be counted. Number of students needs to be confirmed as well.

Bibliography

1. Agrawal, D. (1978). The Factors Affecting Mass Transit Ridership: An Analysis. *Transit Journal* , 55-66.
2. Ahmed EI-Geneidy, Mchael Grimsrud, Rania Wasfi, Paul Tetreault & Julien Surprenant-Legault . (2014). New evidence on walking distances to transit stops: Identifying redundancies and gaps using variable service area. Transportation, 41(1), 193-210.
3. Chu, X. (2004). *Ridership Models at the Stop Level.* National Center for Transit Research (NCTR) University of South Florida.
4. Dan Boyle & Associates. (2006). Fixed-Route Transit Ridership Forecasting and Service Planning Methods, A Synthesis of Transit Practice, page31. San Diego,, California, US: Transportation Research Board, Washing, D.C.
5. Duever, T. A. (2006). Applied Statistics in Engineering Course Hand Note. Waterloo, Ontario, Canada, Waterloo Region: University of Waterloo.
6. Fotheringham, A.S., C. Brunsdon, and M. Charlton. (2002). Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. John Wiley & Sons Ltd., West Sussex, England.
7. Fu, L. (2012). *Modeling Travel Demand for Urban Transportation Planning.* Waterloo, Ontario, Canada: The University of Waterloo.
8. Hellinga, B., J. Casello, and S. Farahani. (2007). Impact of the Introduction of an Express Transit Service in Waterloo Region. the Canadian Institute of Transportation Engineers (CITE) Conference, Toronto, Ontario.
9. Jeffrey M. Casello and Bruce Hellinga. (2008). Impacts of Express Bus Service on Passenger Demand. *Journal of Public Transportation, Vol. 11, No. 4, University of Waterloo* .

10. Jeffrey M. Casello and Tony E. Smith. (2006). Transportation Activity Centers for Urban Transportation Analysis. *JOURNAL OF URBAN PLANNING AND DEVELOPMENT © ASCE / DECEMBER* , P253.

11. Kikuchi S. and Miljkovic D. (2001). Use of Fuzzy Inference for Modeling Prediction of Transit Ridership at Individual Stops. *Transportation Research Record 1774: 25-35* .

12. Kimpel TJ, Strathman JG, and Dueker KJ. (2000). Time Point-Level Analysis of Passenger Demand and Transit Service Reliability. *Center for Urban Studies, College of Urban and Public Affairs, Portland State University , Portland, Oregon.*

13. Metropolitan, Council. (May 15, 2007). *Program Evaluation and Audit, HIAWATHA LIGHT RAIL, Ridership Reporting and Fare Compliance,.*

14. Nashua Regional Planning Commission, U.S. (2011). *Ridership Prediction Model for the NASHUA Transit System.* Nashua, United States: Nashua Transit System.

15. Park, H. (2011). Ridership Analysis at the Stop Level: Case Study of Austin, TX. The University of Texas at Austin.

16. Peng Z, Dueker KJ, Strathman J, and Hopper J. (1997). A simultaneous route-level transit patronage model: demand, supply, and inter-route relationship. *Transportation 24: 159-181* .

17. Peng, Z. (1994). A Simultaneous Route-Level Transit Patronage Model: Demand, Supply, and Inter-Route Relationship. *A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Urban Studies* . Porland: Porland State University.

18. Peter Stopher and Shawna Mulhall. (1992). Route Level Patronage Forecasting Methods: A Survey of Transit Operators. *the 71st Annual Meeting, Transportation Research Board.* Washington, D.C.

19. Ram M. Pendyala, lke Ubaka, and Nadarajah Sivaneswaran. (2004). Geographic Information System-Based Regional Transit Feasibility Analysis and Simulation Tool. *Transportation Research Record 1799* , Paper No. 02-2227.

20. Rasha Kashef. (2014). Quantitative Data Analysis. *The University of Waterloo* . Ontario, Waterloo.

21. Ronald E. Walpole and Raymond H. Myers. (1978). *Probability and Statistics for Engineers and Scientists.* New York: Macmillan Publishing Co., Inc.

22. Ronald E. Walpole and Raymond H. Myers. (1978). *Probability and Statistics for Engineers and Scientists.* New York: Macmillan Publishing Co.,Inc.

23. Srinivas S. Pulugurtha and Mahesh Agurla. (2012). Assessment of Models to Estimate Bus-Stop Level Transit Ridership using Spatial Modeling Methods. *Journal of Public Transportation, Vol. 15, No. 1* , The University of North Carolina at Charlotte.

24. Strathman JG, Dueker KJ, and Z Peng. (1997). Issues in the Design of a Stop-Level Transit Patronage Model. *Project Report PR 102. Center for Urban Studies, College of Urban and Public Affairs, Portland State University, Portland, Oregon* .

25. TCRP. (2006). *TCRP Report 118: Bus Rapid Transit Practitioner's Guide.* Kittelson and Associates, Inc. Orlando, FL in association with Herbert S. Levinson Transportation Consultants, New Haven, CT, DMJM+Harris Fairfax, VA.

26. The Region of Waterloo. (2009). *iXpress: Central Transit Corridor Express Project.* Waterloo Region.

27. Wilson, N. (2010). Ridership Forecasting, http://ocw.mit.edu/courses/civil-and-environmental-engineering/1-258j-public-transportation-systems-spring-2010/lecture-notes11/.