Modeling Toll Price Impacts on Traffic Volume in New York City

A Capstone Presented to the Faculty of Architecture, Planning and Preservation

COLUMBIA UNIVERSITY

By Jackson Fordham

May 2022

Abstract

New York City (NYC) is currently developing a congestion pricing policy called the Central Business District Tolling Program (CBDTP). One of the organizations that has advocated for congestion pricing is the Regional Planning Association (RPA) who also serves as the client for this capstone project. This project attempts to answer the following research questions: How can the CBDTP policy decisions be explored for the best possible options for toll prices using traffic volume modeling? How does the toll prices on NYC bridges and tunnels impact traffic volume? This capstone project utilizes mixed methods in an exploratory-sequential approach design. The principal method applied is a quantitative study with a complementary preliminary qualitative study. The qualitative methods applied are semi-structured interviews with professionals in the transportation planning sector who provided information on how planners typically approach understanding policy impacts such as the CBDTP. The quantitative research is conducted in a three part methodology: defining variables, using supervised learning modeling to analyze the correlations between historical traffic volumes and the other features analyzed, and predicting traffic volumes for each crossing based on predicted values of the independent variables using the best performing model. The dependent variable for the analysis will be traffic volume. Independent variables for the analysis include toll price, population, economic factors and temporal events and include 142 features in total that were collected and organized. 42 different Machine Learning models are tested. A Gradient Boosted Regression Trees (GBRT) model is found to predict traffic volumes with high accuracy using only toll prices and employment. This model as well as the limitations and findings of the analyses are incorporated into a report as well as a website for the client, RPA.

Acknowledgements

Thank you foremost to my advisor, Boyeong Hong, whose knowledge and support were invaluable at every step in this capstone project and who helped me grow professionally and personally. I would also like to thank my two readers, Anna Stokes and Josef Szende, for their guidance and willingness to share their expertise.

Contents

Figures and Tables

List of Acronyms

BPM - Best Practices Model
BTA - Balanced Transportation Analyzer
CBD - Central Business District
CBDTP - Central Business District Tolling Program
FHWA - Federal Highway Administration
GBRT - Gradient Boosted Regression Trees
LASSO - Least Absolute Shrinkage Selector Operator
MLP - Multilayer Perceptron
MTA - Metropolitan Transportation Authority
NYC - New York City
NYC DOT - New York City Department of Transportation
NYMTC - New York Metropolitan Transportation Council
OLS - Ordinary Least Squares linear regression
RF - Random Forest
RPA - Regional Plan Association
TBTA - Metropolitan Transportation Authority's Triborough Bridge and Tunnel Authority

1. Introduction

This capstone project hopes to analyze historical traffic volume trends and their impact by toll prices in order to create a model to analyze the New York City (NYC) congestion policy that is currently being developed. This project utilized mixed methods in an exploratory-sequential approach design using primarily quantitative modeling using machine learning algorithms supported with semi-structured interviews with planning professionals. The goal of this capstone project is to create a methodology and tool to analyze the impact on traffic volumes by different toll prices of crossings into the NYC central business district (CBD) for the client of this project, and NYC congestion policy advocate, the Regional Plan Association (RPA).

Downtown Manhattan is infamous for its slow traffic, poor air quality, and traffic noise. One proposed solution to these problems is the idea of attributing a fee to drivers for the externalities related to vehicle traffic flow within downtown Manhattan (MTA 2021). NYC is currently developing a congestion pricing policy called the Central Business District Tolling Program (CBDTP).

Congestion pricing is meant to incentivize car trips to switch to other modes that will not be charged a congestion tax. Some travelers will switch to using some form of transit where available, while others will choose to drive a longer distance to avoid the tax. The success of the congestion tax relies heavily on the availability and attractiveness of other modes, particularly public transit.

The NYC subway, bus, and commuter railroad systems are operated by the Metropolitan Transportation Authority (MTA). The MTA needs a regular source of income to improve and modernize the services and infrastructure (MTA 2021). The MTA reports that transit ridership had increased “nearly 50% in the 20 years” prior to the pandemic, but funding for MTA public transit over that time “has fallen by 8%” (MTA 2021). The CBDTP will decrease traffic congestion and will be used by the MTA to fund much needed investments particularly the MTA “$54.1 billion 2020-2024 Capital Plan” to increase accessibility and improve travel times and transit services for NYC commuters (MTA 2021).

The goal of the CBDTP is to finance the MTA transit investments by charging vehicles to enter the central business district (CBD), which is defined in CBDTP as Manhattan below 60th Street, excluding FDR Drive and West Side Highway (NYS Route 9A). NYC will be the first North American city to implement congestion pricing, joining other world cities such as London, Stockholm, Milan, and Singapore (Fix NYC Advisory Panel 2018).

The final decisions yet to be made for the CBDTP are significant, especially the decisions of how the toll prices across the bridges and tunnels into the CBD might change and impact the fee within the CBDTP. This combination of tolling from the crossings and the CBDTP is being discussed within the policy discussions as *credits*. The policy impacts of credits could decrease the cost of entry into the congestion zone or could unify the toll price to enter Manhattan by making the tolls on the crossings into the CBD the same price. This discussion of credits is particularly significant because it could possibly make the price of entering the NYC CBD the same across different crossings, eliminating any toll shopping behavior that may happen today. This capstone project hopes to generate a model to analyze and predict any toll impacts on traffic volumes on the crossings into Manhattan. This model and its results are part of the final product to the client for this capstone project, the Regional Plan Association (RPA).

1.1. Client

One of the organizations that has advocated for congestion pricing in NYC is the Regional Planning Association (RPA) who also serves as the client for this capstone project. While serving as the client, RPA has been updated and provided feedback as this capstone project has been completed. The meetings discussed the research process and progress as well as the development of the final product.

For 100 years this year (2022), RPA has advocated and worked to create a more "equitable, healthy, prosperous, and sustainable region" of NYC (RPA n.d.). RPA does this work through research in "transportation, economic development, housing, real estate, climate change, open space, and more" (RPA n.d.). RPA continues to advocate for transportation projects such as the collection of projects included in the Gateway Program relating to the updating of the infrastructure of the Northeast corridor rail corridor.

In 1996, RPA published the Third Regional Plan which included recommendations for a rail line called the Triboro Express to connect the Bronx, Queens and Brooklyn using existing rail infrastructure with the possible extension to connect the line to Staten Island. A portion of the proposed Triboro line is currently under environmental review by the MTA under directions from New York Governor Hocul to become the Interborough Express rail service. This line has become a reality partially because of the significant work RPA did in research and advocacy of the new transit line to connect the outer boroughs.

RPA has published several reports about congestion pricing in NYC first in 2003 and most recently in 2019 following the New York State passage of the MTA Reform and Traffic Mobility Act. The act required the MTA Triborough Bridge and Tunnel Authority (TBTA) to create the CBDTP. RPA identified several specifics of the CBDTP that were not specified in the

MTA Reform and Traffic Mobility Act including how toll prices into the CBD could be set equal using credits to decrease any toll shopping behavior. This particular implication of credits for tolls on bridges and tunnels into the CBD are the inspiration for this capstone project's goals.

1.2. Project Objectives

The goal of this project is to create a methodology and tool to analyze the impact on traffic volumes by different toll prices of crossings into the CBD and the best way to predict traffic volume on crossings into the CBDTP. This is done in order to facilitate a greater diversity of methodologies in modeling congestion toll price impacts within NYC with the ability to expand the analysis to other crossings and other cities. This project hopes to answer the following research questions:

- ➔ How can the CBDTP policy options be analyzed for the best possible decisions for toll prices using traffic volume modeling?
- ➔ How does toll price on NYC bridges and Tunnels impact traffic volume ?

By attempting to answer these research questions, the project applys mixed methods in an exploratory-sequential research design. The principal method applied is a quantitative study with a complementary preliminary qualitative study. The qualitative study included semi-structured interviews that supported the quantitative methods. The quantitative methodology includes the analysis of different machine learning models and the creation of a model that can be used to understand the best way to predict traffic volume and other indicators including toll price. The models applied in this capstone project include Linear Regressions (OLS, Ridge, LASSO), Decision Tree, Random Forest, Gradient Boosted Regression, and Multilayer Perceptron (MLP). All of these models are compared in how well they perform.

This project created a unique Machine Learning model to analyze the CBDTP impacts on traffic volume on crossings into the CBDTP. In the creation of this model, various other alternatives were identified, particularly public available models: the Balanced Transportation Analyzer (BTA) by Charles Komanoff and the Best Practices Model (BPM) from the New York Metropolitan Transportation Council (NYMTC). Academic examples using machine learning to model congestion levels were also reviewed prior to the analysis. For the analysis the methodology used mixed methods to create a single robust model to analyze and predict traffic volume on vehicle crossings into Manhattan.

This capstone project is organized as follows. The second section reviews the previous examples of congestion policies and the methodologies of vehicle congestion modeling. The third section discusses the data used in the modeling processes. The research design and methodology is explained in the fourth section, including the the models that were used in this project and comparing the benefits and limitations of the different models. The fifth section discusses the results from the modeling process, testing different models, and developing the final product. The sixth and final section discusses the capstone project conclusions, limitations and future opportunities to continue the research.

2. Literature review

This project investigates a unique way to analyze the traffic volume impacts of the CBDTP. The focus of the project is to create a method or tool that can predict any possible rerouting of traffic, or toll shopping. This will be approached by understanding the CBDTP within NYC, existing methodologies for studying the traffic volume impacts of the policy, and possible machine learning applications for studying the impacts to traffic volume.

2.1. Congestion Pricing Policies Around the World

NYC will not be the first city to implement congestion pricing. Stockholm, London, Singapore and Milan all have existing congestion pricing policies. To fully understand the Central Business District Tolling Program (CBDTP) these international examples of the policy will be discussed.

Singapore's congestion pricing policy was first implemented in 1975, and is known as the Area Licensing System, but was renamed in 1998 as the Electronic Roadway Pricing (ERP). Motorists in Singapore pay a fee to pass through the ERP toll gantries during operation hours of the system (during peak travel times). The ERP fee changes relative to the congestion of road segments and was recently decreased due to the pandemic. After implementing ERP, Singapore experienced a 24% decrease in weekday traffic volume into the CBD (FixNYC Panel, 2018).

In 2003, London implemented their congestion pricing policy. The fee that motorists pay to enter the London CBD is a flat fee unlike the ERP in Singapore. After 15 years of the congestion policy in London, carbon emissions dropped by 20 percent and congestion reduced by 25% (FixNYC Panel, 2018).

In 2008 Milan implemented a pilot congestion zone pricing system called Europass to both reduce congestion and vehicle emissions. After implementing the program Milan banned the worst polluting vehicles from entering the CBD. The toll price is a flat rate. The full policy was redesigned in 2011 and it has been successful in improving Milan's air quality.

In Stockholm the congestion pricing policy was piloted in 2006 and implemented in 2007 after success of the pilot policy implementation. It charges drivers a variable fee to enter the CBD based on time, saving drivers money who avoid peak travel times. After the implementation of the policy, carbon emissions decreased by 10 to 14 percent in the CBD and congestion decreased by 25%. The CBD congestion zone in Stockholm is only accessible via bridges making implementation easier and possibly similar to the CBDTP in NYC (FixNYC Panel, 2018).

2.2. Congestion Pricing in NYC

When the CBDTP is implemented, it will benefit NYC by improving air quality by decreasing congestion, and generate money for the Metropolitan Transportation Authority (MTA) to fund capital projects. The MTA Reform and Traffic Mobility Act was enacted by the NY State Senate in April 2019 (NY Senate n.d.). It requires that the MTA's Triborough Bridge and Tunnel Authority (TBTA) design, develop, build, and operate the CBDTP (NY Senate n.d.). Currently the CBDTP is in the Federal Environment Assessment process (MTA 2021). The MTA Reform and Traffic Mobility Act requires the CDBTP to "charge passenger vehicles only once each day for entering or remaining in the CBD", implement variable tolling based on time of day, provide a tax credit to motorists who make less than $60,000 exempting them from the toll, and make all “authorized emergency vehicles and qualifying vehicles transporting people with disabilities” exempt from the toll (MTA 2021).

There are several aspects of the CBDTP that have yet to be decided on. This includes the price of the fee, credits that might allow for toll payments on bridges or tunnels to count towards the congestion toll charge, and any other possible exemptions for particular motorists. These are particularly important considerations for NYC as the CBD in Manhattan is on an island that must be crossed into via a bridge or tunnel, many of which already have tolls. Currently not all crossings have the same toll price to cross into Manhattan.

2.3. NYMTC Best Practices Model

There are primarily two models that are being used to analyze the impacts of the CBDTP: the Balanced Transportation Analyzer (BTA) by Charles Komanoff and the Best Practices Model from the New York Metropolitan Transportation Council (NYMTC). Both models are publicly available, but the BPM model requires special software and hardware to run compared to the BTA model which can be opened and run in Microsoft Excel on most modern computers. Both models were attempted to be reviewed for this project, but due to the availability of the BPM model, less is known about the BPM methodology. What is known is that the BPM model is a tour-based model for regional demand forecasting where “ use of tour (or paired journeys) are the basic unit of modeling” (NYMTC n.d.).

The BPM uses a “micro-simulation approach to generate forecasts that are discrete choices for individuals” (NYMTC n.d.). A microsimulation approach can be used to model or analyze policy impacts on individual actions. The Office of Operations at the US Department of Transportation Federal Highway Administration (FHWA) has published “Traffic Analysis

Toolbox Volume III: Guidelines for Applying Traffic Microsimulation Modeling Software 2019 Update to the 2004 Version", where the usage of microsimulation for traffic simulations is explained (FHWA 2022). Microsimulation is defined within this document from FHWA as the following:

> "Microsimulation is the modeling of individual vehicle movements on a second or sub-second basis for the purpose of assessing the traffic performance of highway and street systems, transit, and pedestrians." (FHWA 2022)

It is assumed that the BPM model uses a similar methodology described by FHWA (2022). Within the BPM some series of origins and destinations are defined and collected from datasources (NYMTC n.d.). Using a defined road network and networks of other transit modes, travel trips are modeled using a form of 'microsimulation' and projected populations (NYMTC n.d.). The BPM model is a system model for NYC. The projections or outputs of the BPM can be fed into other models to identify local impacts. The BPM is not using machine learning and reports impacts at a different scale than what this capstone project is trying to achieve

2.4. Balanced Transportation Analyzer (BTA)

The Fix NYC Advisory Panel produced a plan which advocated for the creation of the CBDTP. The Fix NYC Advisory Panel analyzed different toll price impacts in the suggested congestion zone in NYC using a model developed by Charles Komanoff's model called the Balanced Transportation Analyzer (BTA). The BTA model has been used to predict various transportation related impacts due to the NYC congestion zone policy. For this capstone project, only the most related parts of the methodologies of the BTA will be discussed: the methodology for analyzing and predicting traffic volumes based on toll price in NYC. This BTA methodology falls under the analysis related to motor vehicles and includes five total parts, but only the first three relate to toll price impacts (BTA model 2021). This methodology is shown in the diagram in Figure 1.

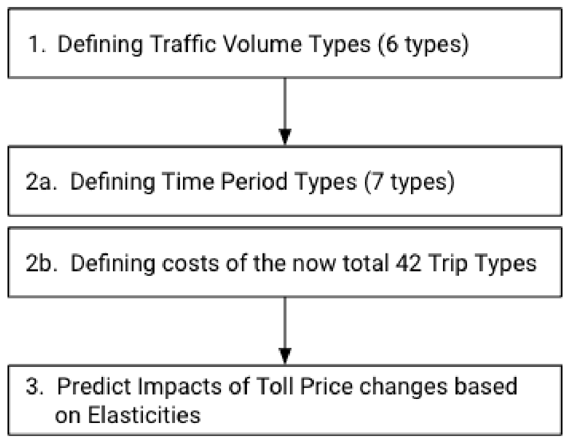

Figure 1: The Balanced Transportation Analyzer (BTA) Methodology for predicting CBDTP price impacts on traffic volumes.

The first part of the methodology is to define the current traffic volumes entering, and exiting of the CBD and to categorize these trips by 1 of 6 trip types: “ (1) auto trips for work purposes, (2) auto trips for non-work purposes, (3) ‘through-trips’ that pass through the CBD, (4) medallion taxi trips, (5) truck and bus trips, and (6) Uber and Lyft trips.” (BTA model 2021). The second part of the methodology is to define the 6 trip types by 7 time periods and to use these 42 different trips to calculate what the BTA methodology calls a ‘baseline’ of all vehicle trips entering the CBD. This part also includes defining all the costs associated with making the trips, such as fuel, parking, insurance, maintenance, trip length and time, and existing toll price costs

(BTA model 2021). The third part of the methodology is to study how these 42 trip types will be impacted due to toll price changes using price elasticities. (BTA model 2021).

The BTA Methodology explains that “price-elasticities play a major role in the BTA” and that they “underlie the algorithms and formulas that predict (or calculate) changes in auto use..in response to implementing congestion pricing”(BTA model 2021). A price-elasticity is the “percentage change in demand or usage associated with a one percent change in the price,” (BTA model 2021). This assumes an impact and the BTA source for this elasticity relationship is informed by a 1977 report by the Tri-State Regional Planning Commission (TSRPC): "Short Term Effects of Transportation Policy Changes on Auto and Transit Ridership" (Interim Technical Report 5303; not Web-available) (BTA model 2021).

The BTA methodology recognizes a possible limitation, that the source for elasticities is informed by a report from 1977 which is more than five decades ago. The BTA methodology responds to this limitation by suggesting that “elasticities are relative measures (percentage changes in parameter Y relative to percentage changes in parameter X), and there is no a priori reason to expect that these relationships should have changed over time,” (BTA Model ‘Elasticities’ sheet, 2021).

The three parts of the BTA methodology shown in Figure 1 are not completely different to the methodology that is deployed for this capstone project. One key difference between the two methodologies is that the BTA model relies on elasticities to predict trip change, while this project's methodology utilizes machine learning algorithms to analyze historical patterns and trends of traffic volume, together with relevant features. This capstone’s methodology provides an alternative model from the BTA model to the client: RPA. When the NYC CBDTP is implemented in the future it will be beneficial to analyze how well any model predicted the changes from the implementation of the congestion zone.

2.5. Accuracy of Congestion Pricing Forecasts

Eliasson et al. analyze the accuracy of congestion pricing forecasts for Stockholm by comparing the forecasted effects of the Stockholm congestion charges with the actual the outcomes. The goal of Eliasson et al. was to find out if the transport models are sufficiently reliable to be used as decision support when designing congestion charging policy and deciding whether to implement such a policy. Eliasson et al. study this by analyzing whether the Stockholm transport model allowed correct conclusions to be drawn regarding the design and preparation of the congestion zone. (Eliasson et al. 2013).

The comparison of outcomes is structured with the same questions from the design and preparation process of the policy, “such as whether the reduction of car traffic would meet the target, whether there would be capacity problems in the public transport system, whether traffic would decrease within the inner city, and whether congestion on circumferential roads would increase.” Eliasson et al. find that most effects were “predicted well enough to allow planners to draw correct conclusions.” (Eliasson et al. 2013).It is not the first model to compare the results of a congestion policy. Eliasson et al. cite the following: de Palma et al., 2006, Santos et al., 2001, Rich and Nielsen, 2007, Eliasson and Mattsson, 2006, Fridstrøm et al., 2000, Kickhöfer et al., 2010. Eliasson et al. (2013) is the first study of “a transport model’s ability to forecast the effects of an urban congestion charging system.” Eliasson et al. (2013) also claim that studying the validity of a model's ability to predict is often difficult due to the long time between the initial forecasts and beginning of operation. (Eliasson et al. 2013).

2.6. Machine Learning for Congestion and Traffic Forecasts

There has been previously research into predicting traffic volume and congestion using machine learning or as Kumar and Raubal discuss, deep learning applications. In their 2021 paper, Kumar and Raubal conduct an extensive literature review process to present the current status of deep learning applications for the detection, prediction and alleviation of traffic congestion (2021). The Kumar Raubal review specifically mentions papers related to the following: deep learning for recurring congestion prediction, deep learning for non-recurring congestion prediction, deep learning for recurring congestion alleviation,, and deep learning for non-recurring congestion alleviation.

Recurring congestion is typically caused by infrastructure constraints and the prediction aspect looks at variations in the time of occurrence and the severity of recurring congestion along a road system. A non-recurring congestion prediction is related to predicting any congestion that is not caused by infrastructure constraints. Kumar and Raubal discuss non-recurring congestion prediction as it relates to predicting congestion caused by accidents. (Kumar and Raubal 2021).

Recurring Congestion alleviation is any method to decrease congestion at the network level. This could mean changing the timing of lights at certain intersections and at the driver level suggesting a change of route to avoid congestion. Kumar and Raubal in their survey explicitly discuss the network aspects for deep learning for recurring congestion alleviation. In their review, Kumar and Raubal begin by an exhaustive explanation of methodological theories and concepts including: deep learning, fully connected layered neural networks, convolutional neural networks and recurrent neural networks, reinforcement learning, commonly used metrics for model evaluation, and relevant concepts and terms in transportation. For this capstone project only papers related to the prediction of congestion levels will be discussed.

Wang et al., 2016 use a model that is a combination of a recurrent neural network (RNN) and a convolutional neural network (CNN) to predict speed and gain insights into congestion sources. Ma et al. 2015 also use a RNN to predict traffic speed at a large transportation network scale. Yu et al., 2017 use a long-short term memory model (LSTM) to do a spatio-temporal analysis of performance study. Sun et al., 2019 use both a LSTM and a CNN to predict speed based on GPS data from taxis. Cheng et al., 2018, also use LSTM and CNN to do traffic congestion forecasting. Ranjan et al., 2020, forecast citywide traffic congestion in Seoul, Korea using CNN and LSTM. Shin et al., 2020 and use a LSTM to forecast congestion level by completing data from the intelligent transportation system that automates congestion mitigation in the streets of Seoul. Fouladgar et al., 2017 also works with data from an intelligent transportation system that monitors traffic conditions and uses RNN and CNN to do short-term traffic forecasting. Liu et al., 2017 uses a variety of LSTM to predict short term vehicle flow, and Li et al., 2017a uses convolutional recurrent neural network (DCRNN) to do traffic vehicle flow forecasting. Rahman and Hasan, 2020 use a LSTM to do real-time predictions for queue length across numerous intersections.

2.7. Machine Learning to Predict Toll Price Using Traffic Volume

The aforementioned papers cover deep learning for recurring congestion prediction, deep learning for non-recurring congestion prediction, Supply side solutions using deep learning for recurring congestion alleviation, deep learning for non-recurring congestion alleviation. These topics show the precedence for using machine learning for congestion modeling, but what these papers do not mention is the relationship between congestion and toll price discussed by Zahedian et al. 2021. Zahedian et al. use three different machine learning algorithms to create a framework to predict dynamic tolls with a time horizon of 30 minutes: random forest, multilayer perceptron, and long short-term memory (Zahedian et al. 2021).

The Zahedian et al. paper involves high-occupancy tolls. High-occupancy lanes are for high-occupancy vehicles (HOV) and public transportation to promote carpooling and using transit. In some areas where the lanes go under-utilized, there has been adoption of a HOT (high-occupancy toll) lane program to allow none HOV to enter the lane but to pay a toll. Zahedian et al. for their study area, include the I-66 inside the Capital Beltway, I-495, located in the Washington DC metropolitan area. Currently the HOT lane toll price is set by the number of vehicles or congestion level of the lane. Zahedian et al. use their machine learning models to estimate toll price based on traffic volumes of the lane and two alternative routes defined in the paper as available as viable alternatives. The methodology of Zahedian et al. includes three parts: defining of three possible routes and existing toll prices, analyzing historical toll prices and vehicle volumes with the models, and finally comparing the model performance to find the best performing model. Their models predict toll price successfully using only historical traffic volume and traffic flow characteristics. The methodology employed by Zahedian et al. differs slightly from the methods that are used for this capstone project as the dependent variable for Zahedian et al. is toll price while for this capstone project it will be traffic volume.

2.8. Literature Discussion and Precedence

This capstone project is unique in that it is studying the impacts of price on traffic volumes on routes using machine learning methods. Zahedian et al. 2021 methodology uses three machine learning algorithms to look at how to predict toll price using traffic volume. The papers listed by Kumar and Raubal 2021 all study topics related to using deep learning to predict congestion but not related to toll price. Eliasson et al. present an analysis of the accuracy of Stockholm's predicted congestion zone impacts. The NYMTC BPM uses microsimulation methods to analyze overall policy impacts and is not specifically created to analyze toll price impacts on traffic volume nor the impacts on specific crossings. The BTA model has a methodology that looks at the impact of toll price on traffic volume, but the analysis uses price elasticities to calculate the CBDTP impact while this capstone project uses machine learning algorithms to analyze historical trends to analyze any impacts.

3. Data Description

3.1. Data Sources

For this capstone project, data is collected from several different sources to use in the quantitative modeling processes. The scope defined the focus of this capstone project as the NYC DOT and MTA TBTA crossings from the time period of 2009-2019. All of the data collected is collected from 2009 to 2019. As much data is collected that is reasonable given the scope and timeline of this capstone project. The data is collected and organized into variable categories that is shown in Table 1.

A majority of the data is collected from the New York Metropolitan Transportation Council (NYMTC) from a publicly available dataset collection NYMTC has named Travel Patterns. The Travel Patterns datasets are organized by NYMTC and collected from various other data sources in order to report average daily vehicle traffic, transit ridership and ferry ridership for the New York City Metropolitan Area. The Travel Patterns datasets are published by quarter and include data reported at a monthly level where appropriate. It may not list ridership for a particular transit service such as a ferry service during a month if that ferry is not operating during that month.

The Travel Patterns quarterly files from NYMTC are collected from 2009-2019 which totalled forty-four different files. These files are organized and combined into a single dataset for this project. This process yields several sets of variables for the methodology. This includes traffic volume crossing data for the units of analysis crossings as well as crossings within the New York Metropolitan Region included in the Travel Patterns datasets. The Travel Patterns dataset also provides ridership data for transit and ferry services within the Metropolitan region of NYC.

Toll price data is collected for the MTA TBTA crossings and PANYNJ crossings. The MTA TBTA crossing toll price data is available from several MTA TBTA reports in which the average price paid for the toll is collected per crossing (Stantec 2020, URS 2009, URS 2002). PANYNJ toll prices are similarly accessed from data from the PANYNJ. The main focus of this capstone project is identifying the role toll price has in predicting traffic volume in the attempt to analyze shopping behavior. PANYNJ toll prices include the peak, non-peak, and cash prices for all PANYNJ crossings. The PANYNJ toll prices are included to provide context as the PANYNJ are in close proximity to the MTA crossings although connecting to different land masses.

Population data is collected from the US Census Bureau from the American Community Survey 5-year Estimates. This population data is collected at the county scale to match the employment data geographic unit which is only available at the county scale. It is important to note that the population data is at an annual temporal unit which is converted to a monthly unit by assuming each month in a year to have the same population.

The employment data is collected from the US Bureau of Labor Statistics (BLS) from the Quarterly Census of Employment and Wages data (QCEW). The employment data is collected as a monthly unit for all industries in total for each of the 31 counties within the NYC Metropolitan Area as defined by both RPA and the NYC Department of City Planning.

ID	Features	Datasource	Spatiality	Temporal Granularity	Calculation
y1	4	NYMTC Travel Patterns (MTA crossings)	crossing	Monthly	None = Monthly
y2	5	NYMTC Travel Patterns (DOT crossings)	crossing	Quarter 4	*Seasonal Calculation % change over year from October*
X1	44	NYMTC Travel Patterns (vehicles, Rail, Bus, Ferry)	-	Monthly, Quarterly	Quarterly => Monthly None = Monthly
X2	4	MTA History and Projection of Traffic, Toll Revenue and Expenses: (2002) (2009) (2020)	crossing	Annual	Same price => Monthly
X3	3	PANYNJ	-	Month of change	Month => Monthly
X4	31	US Census ACS: Table B01003	County	Annual	Annual assumed => Monthly
X5	31	US Bureau of Labor Statistics: QCEW	County	Monthly	None = Monthly
X6	9	Weather Underground	NYC	Monthly	None = Monthly
X7	1	NYS Energy Research and Development Authority	NYC	Weekly	Average Weeks => Monthly

Table 1: Variable Table.

Weather data is collected from Weather Underground which includes key information such as average temperature and precipitation. The weather information is reported for a particular location within NYC that is assumed to be representative of the average weather for NYC. Similarly the average gas price in the NYC metropolitan region is collected from the NYS Energy Research and Development Authority (NYSERDA) Performance Management and Evaluation Services Group. This average gas price during the time period from 2009 to 2019 is assumed to be the most common price paid for gasoline by consumers. All of the data is not available at the same temporal unit which is the discussion of the following section.

3.2. Temporality

While the scope defines the focus of the project between 2009 and 2019, the units of time during this time period are important to define, especially because not all data is available at the same unit of time. The 'Calculation' column of Table 1 shows the change that is made to the collected data in order to convert the data into the chosen unit of time for this capstone project: months. In some cases this calculation or variation is fairly minor. One example of this is the gas price data which is only reported by week. To convert this data into months, the average price of gas in a given month is calculated and attributed to that month to create a new monthly gas price dataset for NYC to be used for the analysis of this capstone project. Some of the temporal availability of the data collected requires some more complex calculations, specifically the Travel Patterns dataset from NYMTC. Most of the data is available at the monthly unit including most of the vehicle crossings and transit ridership data, but the traffic volume data for NYC DOT crossings is not available at monthly rate.

The vehicle volumes reported on the MTA TBTA crossings are available at the monthly unit of time, but the NYC DOT crossing vehicle volume is only available for one month a year. These two differences of temporal data availability are shown in Figure 2. This presented a significant issue to be addressed in this methodology. One option that was considered is to convert all the data into annual data to be at the equivalent unit of time as the NYC DOT crossing traffic volume data. This option is not chosen due to the significant decrease in the number of temporal data points which would decrease the likelihood of good algorithm performance in the analysis process. The solution that is chosen is to estimate seasonality in a given year of the NYC DOT data in order to create what would be ‘expected’ as monthly NYC DOT crossing traffic volumes in the time period of 2009 to 2019.

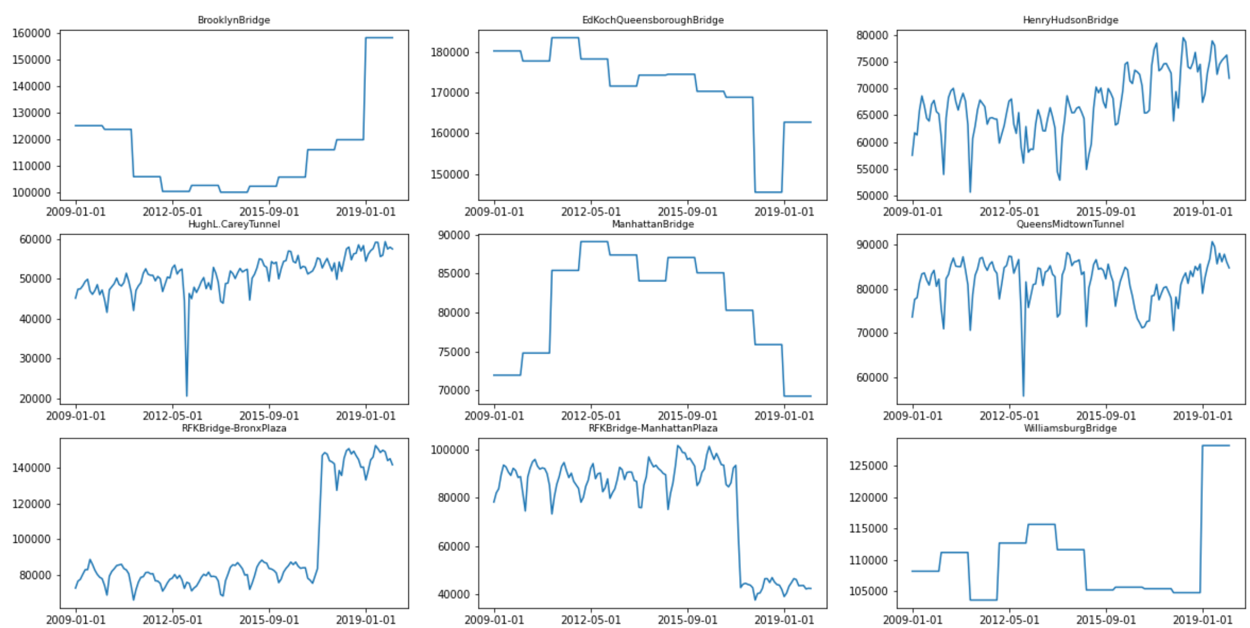

Figure 2: MTA TBTA and NYC DOT Crossing Traffic Volume.

To approximate seasonality, vehicle crossings that did have monthly data are used to calculate the average percent change per month in a given year. The crossings that are used included all of the vehicle crossings with traffic volume data in the NYMTC Travel Patterns data. The crossings include the MTA TBTA units of analysis crossings (Henry Hudson Bridge, Hugh L. Carey Tunnel, Queens Midtown Tunnel, RFK Bridge-Bronx Plaza, and RFK Bridge-Manhattan Plaza) as well as each crossing in the NYC metropolitan region included in the NYMTC Travel Patterns data (Atlantic Beach Bridge, Bear Mountain Bridge, Bronx Whitestone Bridge, Cross Bay Bridge, George Washington Bridge, Goethals Bridge, Holland Tunnel, Lincoln Tunnel, Marine Parkway Bridge, Newburgh-Beacon, Outerbridge Crossing, and Throgs Neck Bridge). All of these crossings are analyzed year by year to calculate the month over month percentage change. The percentage change per month is then averaged across all the crossings to get an average percent change over the months of a given year. This is done for each year from 2009 to 2019. This percentage change is then applied to the DOT data to estimate monthly seasonality change month to month over each year from 2009 to 2019. The result of this calculation is shown in Figure 3 where the previous NYC DOT crossing traffic volume (top) is compared to the new estimated seasonality version of the NYC DOT crossing traffic volume data (bottom).

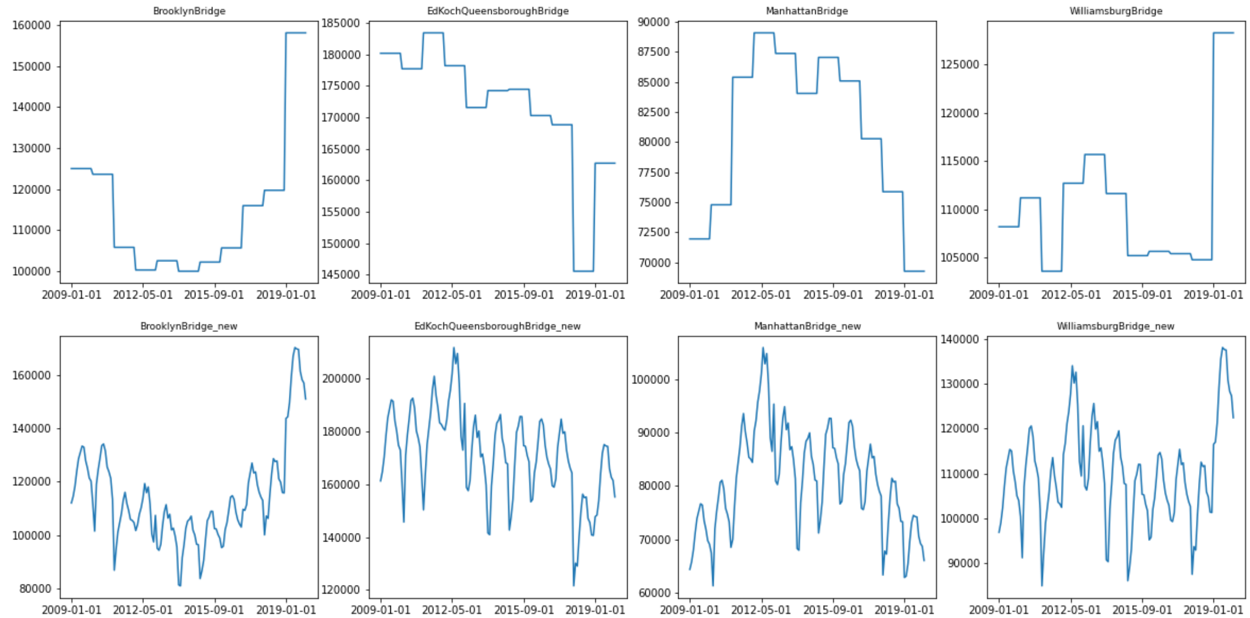

Figure 3: Applying Seasonality to NYC DOT Crossings.

The estimation of seasonality for NYC DOT crossings is the most significant change to any of the data incorporated into the analysis of this capstone project. One other change that is incorporated had to do with ridership data from NJ Transit which is only available at the quarterly level. The NJ Transit services are reported at the average for each quarter in a given year. This value is assumed to be the same across the three months in a quarter. This means if average daily ridership for NJ Transit commuter rail in Q1 of 2009 is reported at 1,000 passengers, it is then assumed that in January, February and March 2009 the average daily ridership is 1,000 passengers. This assumption is applied for all of the NJ Transit data that is unavailable at the monthly unit of time. The final change to the NYMTC Travel Patterns data includes dropping all trip types that are not present in all of the 2009-2019 years. This is done to ignore any new services which included several ferry services with relatively low ridership. This is done because including new services only partially present in the time span from 2009-2019 would cause difficulties in performance for the algorithms that are applied. The goal for this data cleaning process is to provide as much clean monthly data as possible to be used as features in the analysis process of this capstone project.

4. Research Design

The goal of this capstone project is to create a methodology and tool to analyze the impact on traffic volumes by different toll prices of crossings into the CBD. This capstone project utilized mixed methods in an exploratory-sequential approach design. The principal method applied is a quantitative study with a complementary preliminary qualitative study. Figure 4 shows a diagram of methods used in this capstone project. The qualitative methods applied are semi-structured interviews with professionals in the transportation planning sector who provided information on how planners typically approach understanding policy impacts such as the Central Business District Tolling Program (CBDTP). The interviews supported the understanding of the existing models, congestion pricing in NYC, and the data collection and definition. The quantitative methodology was a three step process that included the identification of variables, testing various machine learning algorithms and lastly making predictions. After identifying the best performing model to predict traffic volume the results are organized into a website for RPA.

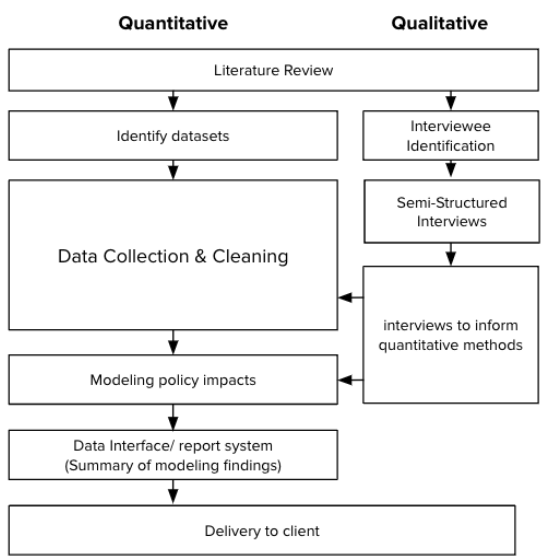

Figure 4: Methodology Chart.

4.1. Scope

For this capstone project, the geographic focus is the New York City metropolitan region, specifically related to the Manhattan CBD and the bridges and tunnels connecting to Manhattan (Figure 5). To explore how toll prices into the CBD could be set equal using credits to decrease any toll shopping behavior the crossings into Manhattan are defined as the units of the analysis. To study historical toll shopping behavior within the study area, the east river crossings are focused on as a priority. On the east river there are two entities that manage crossings.

The NYC Department of Transportation (NYC DOT) operates the Brooklyn Bridge, Queensboro Bridge, Williamsburg Bridge, and Manhattan Bridge. None of the NYC DOT crossings currently have a toll. The East River crossings operated by the MTA TBTA have tolls and include the Hugh Carey Tunnel and the Queens Midtown Tunnel, but the Robert F. Kennedy (RFK) Bridge, and Henry Hudson Bridge are included in the units of analysis in order to include all MTA TBTA crossings and directly compare NYC DOT (Free) crossings and MTA TBTA (tolled) crossings.

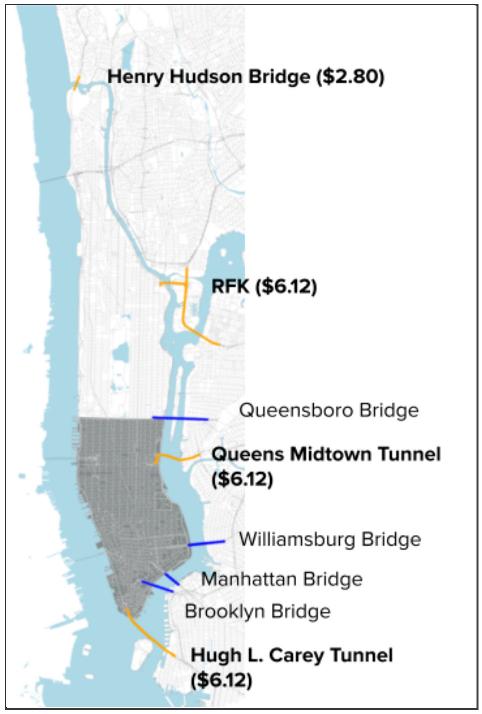

Figure 5: Map of NYC Crossings and the Units of Analysis.

Data Source: NYC DCP 2021, NYC DOT 2021.

The Hudson River crossings include the Holland Tunnel, the Lincoln Tunnel and the George Washington Bridge. All of these crossings have a toll and are operated by a third entity: the Port Authority of New York and New Jersey (PANYNJ). Because there are no free crossings available to travelers into New York City from New Jersey, there is no chance for toll shopping behavior which is the focus of this capstone project. Including the PANYNJ crossings within the units of analysis would not benefit the modeling process because no toll shopping behavior could be included in the PANYNJ crossing data. For this reason, the PANYNJ crossings are not included in the unit of analysis.

The crossings identified as the units of analysis have been selected because of the current differences between MTA TBTA and NYC DOT crossings. To further analyze this difference historical data was chosen to study. The time period from 2009 to 2019 was chosen because of it being from the great recession to before the COVID-19 pandemic. Data from this time period was collected to train the models to identify the best model at analyzing and predicting traffic volume along the identified crossings that are the units analysis.

4.2. Hypothesis

Traffic volume data and toll price data for the MTA TBTA are some of the data collected for this analysis. The analysis hopes to identify any relation between toll price and its impacts on traffic volume on the crossings identified as the units of analysis. The alternative hypothesis is that there is a significant measurable impact on traffic volume by toll price and the null hypothesis is that there is no significant impact of toll price on traffic volume. While toll prices have increased over the time period from 2009 to 2019, traffic volume on the crossings has increased as well. It is expected that the increase in tolls has not significantly affected the number of vehicles on the crossings.

4.3. Qualitative Methods

The qualitative process included several semi-interviews with transportation planners and economists to understand the transportation and toll revenue modeling approaches. Interview questions covered topics such as: the methodologies of demand models and how they typically are used for decision making, the way cities generally study the impacts of congestion pricing and how agencies within NYC are studying the impacts of NYC CBDTP. In several interviews the overall research design of this capstone project is also discussed and feedback was provided by the interviewees. The findings from these interviews assisted in the design and development of the research including the variable definition, modeling, and product development.

There were five interviews with planning professionals who shared their opinions about traffic modeling and making decisions related to transportation policies such as the CBDTP. The individuals are from organizations such as NYC DOT, NYMTC, Manhattan Borough Office and include the creator of the Balanced Transportation Analyzer, Charles Komanoff. The individuals do not represent their organizations and are merely professional individuals with experience analyzing transportation policy impacts in NYC.

4.4. Quantitative Methods

The quantitative process centers around traffic volume modeling of CBDTP policy impacts on the NYC DOT and MTA TBTA crossings. The quantitative process approaches the modeling of congestion pricing in a novel way using supervised learning predictive regression algorithms that provides an alternative to the existing modeling methodologies of predicting congestion pricing impacts. The goal is to provide RPA an alternative method to calculate traffic volume impacts of toll price related to the CBDTP. The resulting model is able to be used to analyze and predict traffic volumes.

In order to complete the quantitative methodology and create the final product three methodological steps are taken shown in Figure 6. The first step is defining variables to include in the modeling process. The second step of the quantitative methodology was a testing of various models to identify the best performing model and incorporate it into the final product.

The final part of the quantitative methodology was to make several scenario predictions of traffic volume on the crossings based on different tolls.

Part 1: Defining variables to include in the modeling process

Part 2: Testing of models to identify the best performing one for final product

Part 3: Predict traffic volumes per route based on predicted toll prices

Figure 6: Quantitative Methodology Sections

4.4.1. Variable Definition

The identified bridges and tunnels will be referred to as crossings. The identified crossings include: Brooklyn Bridge, Ed Koch Queensboro Bridge, Henry Hudson Bridge, Hugh L. Carey Tunnel, Manhattan Bridge, Queens Midtown Tunnel, RFK Bridge-Bronx Plaza, RFK Bridge - Manhattan Plaza, and Williamsburg Bridge. The dependent variable for the analysis is traffic volume of the crossings. The independent variables for the analysis include toll price, population, transportation, employment, gas price, and weather.

The variables are selected based on data availability (shown in Table 1) and literature review. One of the key principles of congestion pricing is that some drivers will change their behavior and switch their trip to either another time or another mode. To capture what the BTA model calls 'mode switching' every transit system incorporated into the NYMTC travel patterns data source is selected to be included in this capstone project. It is expected that some change in the toll price (y2) will be reflected in some transit ridership data (X1). The population in the region (X4) was selected to control for any changes in total population over the time period in relation to changes in traffic volume (y1+y2). It is expected that the vehicles crossing the bridges and tunnels are some segment of the people living within the metropolitan area of NYC. Similarly, employment (X5) in the NYC metropolitan region was selected based on the assumption that a majority of the vehicles at the crossings are people commuting for a job. Weather data (X6) was selected under the assumption that some percent of drivers would change their trip based on significant weather events such as Hurricane Sandy. The average gas prices in NYC (X7) is included assuming that any significant change in gas prices may reflect changes in traffic volume.

To organize these variables into a single model to be used to predict traffic volumes on all the crossings, the variables are organized into a table that can provide the models enough information to predict traffic volume at all the crossings. This table is shown in Figure 7.

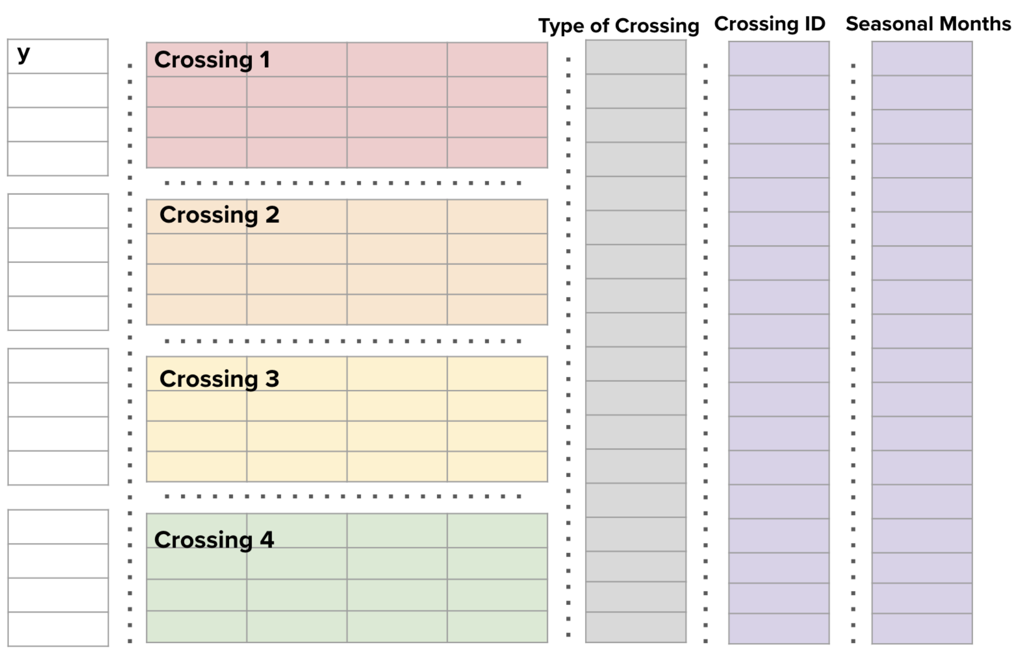

Figure 7: Data Table Concatenation Process.

Figure 7 shows the process of organizing the variables into a single table to be used for analyzing all the crossings. The y column is the result of a concatenation of all the traffic volumes of the crossings previously defined (y1 and y2). The traffic volumes line up with the identified crossing tables which are merely the independent variables as columns and the rows as months from 2009 to 2019. These tables are duplicated and concatenated for each crossing (shown in Figure 7 as Crossing 1, 2, 3, 4). There are also several dummy variables included as columns to indicate if the crossing has a toll or not, which crossing is which following the concatenation process, and finally columns for every month in a year to indicate time or seasonality within the modeling process. This concatenation process of the variables provides substantially more data points for the model to run and allows for the model to be used to predict traffic volumes at all the crossings at the same time. This aligns with the goal of the capstone project: to provide RPA a model that can be used to analyze traffic volume at the crossings and the relationship with toll price.

4.4.2. Models for Prediction

The second part of the methodology is an analysis of the historical traffic volumes for each crossing. This part of the methodology will seek to analyze how the traffic volume can be predicted by different models. This is done using the variable table created in the previous step and using supervised learning modeling. Different algorithms will be applied including: Linear Regressions (OLS, Ridge, LASSO), Decision Tree, Random Forest, Gradient Boosted Regression Trees, and Multilayer Perceptron (MLP). The models are run with the same variables aforementioned to evaluate and compare model performance. Prior to running the models, the variables are first scaled from their previous distributions to be within 1 and 0 using a scaling algorithm incorporated in the Scikit Learn package in python. This was done as not all the data is representing the same as some variables represent vehicles while others represent dollars or inches of rain. By scaling the variables, it allows the models to understand changes over time irrespective of the underlying unit.

Linear Regression (OLS, Ridge, LASSO)

A basic regression model creates predictions of real numbered values relying on the relationship of the independent variable(s). When only one independent variable is used, it is a linear regression, while when multiple independent variables (features) are used it is a multivariate linear regression model. Regression models are often applied for time series applications which makes them applicable for traffic forecasting modeling. Regression models are most suitable for applications when the features are expected to have linear relationships. Aktar et al., 2021 suggest that regression models are now less frequently applied for traffic congestion prediction due to the increasing complexity of datasets. In this methodology three variations of multivariate linear regressions are applied: OLS, Ridge and LASSO.

The linear regression models are applied in this capstone project to compare the results from the linear regression models with the nonparametric models. Ordinary Least Squares (OLS) linear regression minimizes the sum of the squares in the difference between the observed and predicted values. The Ridge and LASSO (Least Absolute Shrinkage Selector Operator) models improve the OLS regression by using L2 regularization which simplifies the regression by adding a penalty parameter to the cost function in the regression. The Ridge regression uses L2 regularization and a hyporamenter of the Ridge linear regression: alpha. Alpha applies a penalty in the regression function, decreasing the magnitude of the coefficients. LASSO (Least Absolute

Shrinkage Selector Operator) also uses alpha to apply a penalty and uses L2 regularization, but also applies a feature selection process using the coefficients to set some features with insignificant coefficients to not be included in the model, further simplifying the regression.

Decision Tree and Random Forest (RF)

The decision tree models used in this capstone project are regression tree models. A decision tree model can be described as a series of if-then statements. All tree models are appropriate for problems with multiple features. Decision trees are based on a series of decision steps (hierarchical). RF models are a randomized series of decision trees where the outputs are averaged. Within a random forest model the features and sampling of the features are randomized. By randomly selecting the sampling and features the correlation between decision trees within a random forest is minimized which results in reduced variance of errors.

Both decision trees and random forests can be used to overfit the data used in the training process. This means the model performs very well with data it has been trained on but performs poorly on data that the model is unfamiliar with. Hyperparameter tuning is used in the modeling processes of decision tree and random forest models to circumvent the overfitting issues. The hyperparameter tuning might include the max depth of the trees, restricting how many layers of splits the trees have, and restricting how well a tree learns the input data. Both decision trees and RF models can handle non-linear data unlike the linear regressions that are explained above. Alajali et al. 2018 used iterations of various decision trees to create projections for traffic volumes within intersections based on multiple features.

Gradient Boosted Regression Trees (GBRT)

A gradient boosted regression trees (GBRT) algorithm is similar to the random forest model described previously, but the trees within the GBRT model are created sequentially using the residuals of the first tree. This process allows the GBRT model to learn the relationships between the features based on errors of previous trees. The GBRT learning process is described as an ensemble learning method. Hyperparameter tuning is also used in the modeling processes to circumvent overfitting issues. Like decision trees and random forests the GBRT model can handle non-linear relationships within the input features. Yang et al. propose GBRT, an ensemble learning method, be used to make short-term traffic prediction based on the traffic volume data.

Multilayer Perceptron (MLP)

The Multilayer Perceptron (MLP) is an artificial neural network (ANN) model that has applications for complex classification and regression problems. ANN models are very simplified versions of the human brain where neurons are connected in different layers, creating a form of neural network. The MLP model consists of an input layer, hidden layers, and the output layer. The layers are made of neurons that, except for the input layer, are "connected with nonlinear functions to a nonlinear model" (Zahedian et al. 2021). MLP models use backpropagation as a supervised learning technique.

MLP models are very sensitive to feature scaling and as previously mentioned in this capstone project all input features are scaled between 0 and 1 to ensure the MLP models perform best. MLP models have strong capabilities for learning non-linear relationships like the other nonparametric models mentioned above (Decision Tree, Random Forest, and GBRT). Zahedian et al. 2021 apply a standard MLP model in predicting toll prices using only historical traffic volume.

Modeling Objective

The models described above will be run using the same scaled input features to compare the model performance of the various algorithms. This research will use supervised learning modeling to analyze correlations between the variables described in Table 1. This research will analyze all of the above models in order to compare the results as part of the methodology, generating a detailed reporting of the performance comparison of the models as part of the findings of this capstone project for the client: RPA.

4.5. Prediction

The final part of the methodology is using the best performing model from the previous step to predict traffic volumes. The goal of this final part is to create a tool to evaluate impacts of predicted toll shopping behavior based on changes to toll price. This final step in the methodology is to apply the selected model in several different scenarios to demonstrate different impacts of toll prices on traffic volumes. The final model will be able to be used by the client, RPA, and the general public to explore how this model predicts toll price changes will impact the traffic volume on different crossings. This will be done with data that is provided from RPA that includes different toll price scenarios and employment projection scenarios.

5. Analysis and Product Development

This capstone project utilized mixed methods in an exploratory-sequential approach design. The principal method applied is a quantitative study with a complementary preliminary qualitative study. From the five interviews conducted, some of the major takeaways came from the interviews with Mr. Komanoff as well an individual from NYMTC about Best Practices Model (BPM). Both interviews discussed the goals and uniqueness about each of the respective models. These two particular interviews helped frame the goals of the quantitative methodology. The BTA model, while very clearly explained and highly accessible in Microsoft Excel, makes predicted impacts using elasticities. In particular, one of the elasticities used to make predicted impacts is cited from a web unavailable report from 1977. This was discussed in the interview and was recognized as a limitation. While making predictions using elasticities has its benefits and limitations this particular elasticity is most likely not truly representative of the decisions drivers are making today.

The interview about the BPM unfortunately did not yield any specifics about the methodology it uses. This highlighted one of the benefits of the BTA model that it is clearly explained and that it can be opened by anyone with Microsoft Excel. Based on the review of BPM and other literature review, it is suspected that the microsimulation within the BPM includes a series of network analysis and can be fed into traffic demand models for the NYC metropolitan region. It is suspected that the BPM model, as a microsimulation, is possibly as unassociated with what drivers are doing on the road today as the BTA model. It appears that both BPM and BTA assume a policy impact when that is not necessarily representative of the real world. In a time when more and more data is being collected in cities and that data is more readily available, this capstone project hopes to utilize public available data to analyze historical traffic volumes for the identified crossings in order to study any possible toll shopping behavior and make predictions using this data.

The goal for the modeling process is to identify which model performs best in predicting traffic volume. In this analysis supervised learning implies that each model is provided a partition of the features for training purposes and another for testing purposes. In all of the models a ratio of 0.75 between training and testing features was used. As mentioned previously the features within the model were all scaled to be within 0 and 1.The different models applied include: OLS, Ridge, LASSO, Decision Tree, Random Forest (RF), Gradient Boosted Regression Trees (GBRT), and Multilayer Perceptron (MLP).

The Ridge and LASSO linear regressions are optimized for the value of alpha or lambda that yields the best results. The tree models (Decision Tree, RF, and GBRT) parameters were tuned using a grid search function. The Decision Tree grid search focused on tree depth. The RF grid search focused on tree depth and number of trees. The GBRT grid search focused on the number of trees, tree depth, learning rate, and the minimum number of samples required to split an internal node of a tree. The MLP model was run without any parameter tuning due to the highest scores obtained with standard model parameters.

The modeling results are shown in Table 2. The Ridge model was run with an alpha of 12 and the LASSO model was run with an alpha of 25. The Decision Tree model was run with a max depth of 8. The RF model was run with a max depth of 10 and the number of trees was 5. The GBRT model was run with 100 trees, a max depth of 5, a minimum split rate of 5, and a learning rate of 0.04.

The adjusted r-squared test scores are reported in the *Test Set* column which is the adjusted r-squared results for how the model performs given data it has not seen previously. The closer the adjusted r-squared number is to 1.00 the higher the percentage of the variance in the target field (y) that is explained by the input features (X variables). The MSE (Mean Square Error) is the standard deviation of the prediction errors (residuals). The lower the MSE the more accurate the model is in predicting the target variable (y). The results from the models in Table 2 are very good.

Model	Evaluation Criteria	Training Set	Test Set
OLS	Adjusted R-Squared	0.86	0.83
	MSE	0.00494	0.00621
Ridge	Adjusted R-Squared	0.87	0.78
	MSE	0.00481	0.00672
LASSO	Adjusted R-Squared	0.74	0.77
	MSE	0.00926	0.00838
Decision Tree	Adjusted R-Squared	1.00	0.93
	MSE	0.00001	0.00231
RF	Adjusted R-Squared	0.99	0.95
	MSE	0.00042	0.00167
GBRT	Adjusted R-Squared	1.00	0.96
	MSE	0.00009	0.00123
MLP	Adjusted R-Squared	0.93	0.88
	MSE	0.00256	0.00363

Table 2: Modeling Results.

The results in Table 2 show that the linear models (OLS, ridge and LASSO) did not perform as well as the nonparametric models which is expected. The models that performed best are the GBRT and RF models. For the testing set results the GBRT and RF have the highest adjusted r-squared number (0.96 and 0.95) and lowest MSE scores (0.00123, and 0.00167). The worse performing models are the OLS, ridge and LASSO models, but they still have acceptable results. Due to this very high performance results from the models it is important to validate that the results are not due to an error or over correlation between the target and one of the input features.

5.1. Performance Verification

In order to validate the results of the models in Table 2, several models are designed to identify any peculiar relationships between the input features and the target. Five different models are designed to compare to the initial results to test any over reliance on the categories of input features. All five models are run with all the algorithms used initially. The five models are described below with model zero the standard model with normal input features.

- **Model 0:** All Input Features - No Change
- **Model 1:** Input Features - No Population Indicators
- **Model 2:** Input Features - No Population and Employment Indicators
- **Model 3:** Input Features - No Population, Employment, and Weather Indicators
- **Model 4:** Input Features - No Population, Employment, Weather, Gas Prices Indicators
- **Model 5:** Input Features - Only Toll Price Indicators and dummy variables

These iterations of the models were designed based on the spatial granularity differences shown in the Variable Table (Table 1). Each model drops a category of variables based on the previous model. Model 1 was designed to address one particular concern that the population indicators were over impacting the performance of the models due to the population features only truly available at an annual unit of time. All models were run with the dummy variables. Model 1 is run with the following input features: transportation, employment, weather, gas prices, and toll prices. Model 2 was designed to address several concerns about the differences in geography for the population and employment features which are at the county level. Model 2 is run with the following input features: transportation, weather, gas prices, and toll prices. Model 3 builds on Model 2 by dropping population and employment features, but also drops weather features. Model 3 is run with the following input features: transportation, gas prices, and toll prices. Model 4 is run with only transportation and toll prices (MTA TBTA) input features. Model 5 is run with only MTA TBTA toll prices and the dummy variables as the input features. The test results of these models are compared with the initial model test results (model 0) in Table 3.

Model	Evaluation Criteria	Model 0	Model 1	Model 2	Model 3	Model 4	Model 5
OLS	Adjusted R-Squared	0.83	0.83	0.84	0.84	0.84	0.85
	MSE	0.00621	0.00612	0.00578	0.00564	0.00562	0.00539
Ridge	Adjusted R-Squared	0.78	0.80	0.81	0.81	0.81	0.79
	MSE	0.00672	0.00724	0.00707	0.00705	0.00706	0.00749
LASSO	Adjusted R-Squared	0.77	0.77	0.77	0.78	0.78	0.78
	MSE	0.00838	0.00838	0.00838	0.00792	0.00792	0.00788
Decision Tree	Adjusted R-Squared	0.93	0.94	0.91	0.92	0.92	0.83
	MSE	0.00231	0.00186	0.00270	0.00237	0.00243	0.00545
RF	Adjusted R-Squared	0.95	0.95	0.92	0.93	0.94	0.83
	MSE	0.00167	0.00171	0.00250	0.00215	0.00204	0.00521
GBRT	Adjusted R-Squared	0.96	0.96	0.96	0.95	0.96	0.83
	MSE	0.00123	0.00126	0.00125	0.00147	0.00119	0.00543
MLP	Adjusted R-Squared	0.88	0.88	0.86	0.89	0.85	0.79
	MSE	0.00363	0.00372	0.00449	0.00348	0.00466	0.00662

Table 3: Performance Testing Results.

The results of the performance testing in Table 3 shows that sets of features that were tested did not significantly change the results of the models. Throughout the five models the adjusted r-squared number is fairly consistent. This shows that not including the tested features did not significantly impact model performance. These tests indicate that the results of the models can be trusted and further analyzed to identify the true best performing model.

5.2. Model Selection

The best performing model in the performance testing was the GBRT model. To further explore how this model performs the residuals were plotted in Figure 8. As expected from the high adjusted r-squared number and low MSE scores, the plotting of the residuals confirmed the accuracy of the GBRT model. The Figure 8 shows that the values of actual where the residual is not 0 are random. This is good as it indicates the model is not better or worse depending on the scale of the value. Figure 8 shows the results from plotting the actual and predicted values for traffic volume at the crossings. The line in Figure 8 shows that many of the points are clustered along it indicating a good match between actual and predicted values. This confirms the GBRT model performance and that it is the best model to incorporate into the final product.

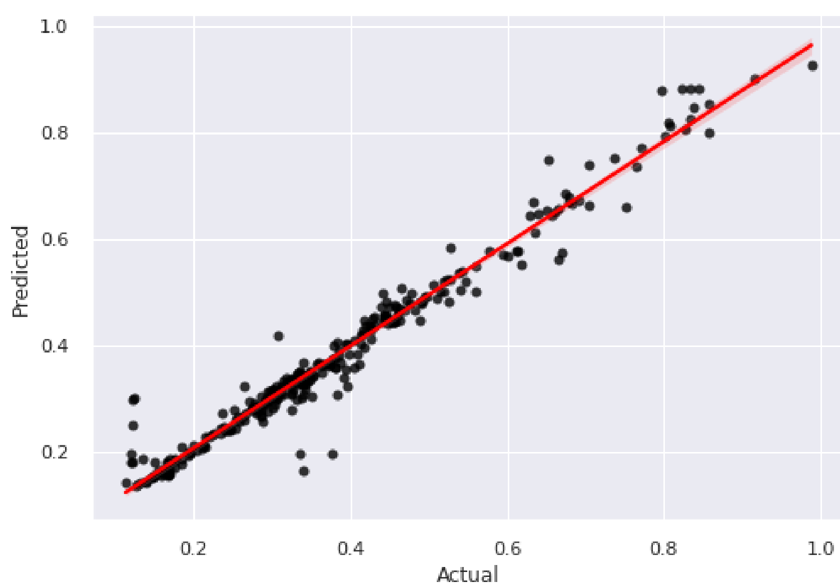

Figure 8: GBRT Scatter Plot of Predicted v. Actual Values.

5.3. Product Development

The final product includes two parts: a report and a website showcasing the model and findings. Both parts allow the client, RPA, to understand how the model is making predictions given different scenarios. The report serves as an accompanying summary of the website and the takeaways from the research processes. The final product enables the client to obtain policy recommendations from the projections of the modeling process.

The website is hosted on Google Colab but is publicly available for RPA's ease of use. The website includes all the data that were collected and organized into variables, the results of the performance verification, and 16 projections of the final GBRT model visualized. The projections included are based on data provided by RPA including 4 different toll price scenarios and 4 employment projections. The website also allows someone to run the model or download it for further exploration. Documentation is included that explains how to use the website as well as explaining this capstone project for anyone not familiar including the objectives, analysis, findings, and limitations.

Provided with the report and interface, RPA will have all the tools necessary to apply this model in various ways. The model can be used to make predictions for any crossing to Manhattan with a high level of accuracy. The model can be used to advocate for specific tolls related to the CBDTP implementation but also the model can continue to be used to advocate for different toll prices for the foreseeable future. Finally the goal is that RPA can adopt these products and overall methodology and support the application of this model to various other contexts and regions to investigate potential congestion zone policies outside of the NYC CBDTP.

6. Discussion and Conclusions

The most significant limitation of this capstone project is due to the use of Machine Learning models that rely on historical data trends. The goals of this project are to predict impacts of the CBDTP in NYC. However this project used machine learning models which rely on historical data trends to make predictions. As NYC has never had a congestion policy in place, there is no historical data to base the model on. Therefore the model in this capstone project can only be used to accurately predict the CBDTP if it is implemented gradually similar to the gradual changes of current toll prices.

Another limitation was the lack of monthly data. This is most present in the NYC DOT crossing traffic volume data which is only available once per year. In this capstone project the attempt was made to circumvent this limitation by approximating the seasonal (month over month) variation of traffic volume on the NYC DOT crossings that was previously mentioned in section 3.2 Seasonality. This issue can not be avoided with the current data management policies in NYC. This severely impacts the ability of any model to accurately predict traffic volume on the NYC DOT crossings. Hopefully, RPA can use the products of this capstone project to advocate for better data management practices in NYC government such as the NYC DOT.

Even with these limitations, the model can predict traffic volumes on the crossings with high accuracy using only toll prices and employment. This model can be used to predict traffic volumes on the NYC DOT and MTA TBTA crossings today with the future potential to include other crossings such as the PANYNJ crossings. This model can be used by RPA and others to apply a Machine Learning methodology to understanding CBDTP impacts. The model can be updated with new data after the implementation of the CBDTP to explore different toll prices for the future of the CDBTP. This model could also be used to just explore MTA TBTA toll prices, PANYNJ toll prices, or one specific crossing etc.

This methodological approach can also be applied to other cities where there is interest in understanding impacts of other congestion zone policies, such as in San Francisco where a congestion pricing policy is being discussed. This methodological approach can also be applied to one of the cities that already has congestion pricing, such as London, to understand how the models are similar or different based on the local contexts.

As congestion pricing continues to operate in several cities around the world, and NYC is on its way to implementing the CBDTP, it is important for the transportation planning field to have a stronger understanding of congestion pricing policies in different cities and how well different models predict the impacts of congestion pricing. This will benefit all cities in having better understanding and interpretations of models for congestion pricing impacts. Hopefully with this capstone project, RPA has the material to start this conversation.

7. Bibliography

Akhtar, Mahmuda, and Sara Moridpour. "A Review of Traffic Congestion Prediction Using Artificial Intelligence." Journal of Advanced Transportation 2021 (2021).

Alajali, Walaa, Wei Zhou, Sheng Wen, and Yu Wang. "Intersection traffic prediction using decision tree models." Symmetry 10, no. 9 (2018): 386

Balta, Musa, and İbrahim Özçelik. "A 3-stage fuzzy-decision tree model for traffic signal optimization in urban city via a SDN based VANET architecture." Future Generation Computer Systems 104 (2020): 142-158.

Cheng, Xingyi, Ruiqing Zhang, Jie Zhou, and Wei Xu. "Deeptransport: Learning spatial-temporal dependency for traffic condition forecasting." In 2018 International Joint Conference on Neural Networks (IJCNN), pp. 1-8. IEEE, 2018.

De Palma, Andre, Robin Lindsey, and Esko Niskanen. "Policy insights from the urban road pricing case studies." Transport Policy 13, no. 2 (2006): 149-161.

Ding, Qing Yan, Xi Fu Wang, Xiu Yuan Zhang, and Zhan Quan Sun. "Forecasting traffic volume with space-time ARIMA model." In Advanced Materials Research, vol. 156, pp. 979-983. Trans Tech Publications Ltd, 2011.

Di, Xiaolei, Yu Xiao, Chao Zhu, Yang Deng, Qinpei Zhao, and Weixiong Rao. "Traffic congestion prediction by spatiotemporal propagation patterns." In 2019 20th IEEE International Conference on Mobile Data Management (MDM), pp. 298-303. IEEE, 2019.

Dusparic, Ivana, Julien Monteil, and Vinny Cahill. "Towards autonomic urban traffic control with collaborative multi-policy reinforcement learning." In 2016 IEEE 19th international conference on intelligent transportation systems (ITSC), pp. 2065-2070. IEEE, 2016.

Eliasson, Jonas, Maria Börjesson, Dirk van Amelsfort, Karin Brundell-Freij, and Leonid Engelson. "Accuracy of congestion pricing forecasts." Transportation Research Part A: Policy and Practice 52 (2013): 34-46. (link)

Eliasson, Jonas, and Lars-Göran Mattsson. "Equity effects of congestion pricing: quantitative methodology and a case study for Stockholm." Transportation Research Part A: Policy and Practice 40, no. 7 (2006): 602-620.

Fan, Zhanyong, Chun Liu, Dongjian Cai, and Shun Yue. "Research on black spot identification of safety in urban traffic accidents based on machine learning method." Safety science 118 (2019): 607-616.

Fix NYC Advisory Panel. "Fix NYC Advisory Panel Report." 2018. <http://www.komanoff.net/cars_II/Fix-NYC-Panel-Report.pdf>

Fouladgar, Mohammadhani, Mostafa Parchami, Ramez Elmasri, and Amir Ghaderi. "Scalable deep traffic flow neural networks for urban traffic congestion prediction." In 2017 International Joint Conference on Neural Networks (IJCNN), pp. 2251-2258. IEEE, 2017.

Fridstrøm, Lasse, Harald Minken, Paavo Moilanen, Simon Shepherd, and Arild Vold. "Economic and equity effects of marginal cost pricing in transport." AFFORD Deliverable A 2 (2000).

Fukuda, Shota, Hideaki Uchida, Hideki Fujii, and Tomonori Yamada. "Short-term prediction of traffic flow under incident conditions using graph convolutional recurrent neural network and traffic simulation." IET Intelligent Transport Systems 14, no. 8 (2020): 936-946.

García de Soto, Borja, Andreas Bumbacher, Markus Deublein, and Bryan T. Adey. "Predicting road traffic accidents using artificial neural network models." Infrastructure Asset Management 5, no. 4 (2018): 132-144.

Genders, Wade, and Saiedeh Razavi. "Asynchronous n-step Q-learning adaptive traffic signal control." Journal of Intelligent Transportation Systems 23, no. 4 (2019): 319-331.

Genders, Wade, and Saiedeh Razavi. "Using a deep reinforcement learning agent for traffic signal control." arXiv preprint arXiv:1611.01142 (2016).

Gupta, Sonu, Deepak Srivatsav, A. Venkata Subramanyam, and Ponnurangam Kumaraguru. "Attentional road safety networks." In 2019 IEEE International Conference on Image Processing (ICIP), pp. 1600-1604. IEEE, 2019.

Gregurić, Martin, Krešimir Kušić, Filip Vrbanić, and Edouard Ivanjko. "Variable Speed Limit Control Based on Deep Reinforcement Learning: A Possible Implementation." In 2020 International Symposium ELMAR, pp. 67-72. IEEE, 2020.

Kickhöfer, Benjamin, Michael Zilske, and Kai Nagel. "Income dependent economic evaluation and public acceptance of road user pricing." In Proceedings of the Kuhmo Nectar conference on transportation economics, vol. 7, pp. 10-03. Technische Universität Berlin, Transport Systems Planning and Transport Telematics, 2010.

Komanoff, Charles "Balanced Transportation Analyzer." (BTA). 2021. <https://nurturenature.org/balanced-transportation-analyzer?_pos=1&_sid=b715849ef&_ss=r>

Kumar, Nishant, and Martin Raubal. "Applications of deep learning in congestion detection, prediction and alleviation: A survey." Transportation Research Part C: Emerging Technologies 133 (2021): 103432.

Li, Linchao, Xi Sheng, Bowen Du, Yonggang Wang, and Bin Ran. "A deep fusion model based on restricted Boltzmann machines for traffic accident duration prediction." Engineering Applications of Artificial Intelligence 93 (2020): 103686.

Li, Pei, Mohamed Abdel-Aty, and Jinghui Yuan. "Real-time crash risk prediction on arterials based on LSTM-CNN." Accident Analysis & Prevention 135 (2020): 105371.

Li, Yaguang, Rose Yu, Cyrus Shahabi, and Yan Liu. "Diffusion convolutional recurrent neural network: Data-driven traffic forecasting." arXiv preprint arXiv:1707.01926 (2017).

Liu, Yipeng, Haifeng Zheng, Xinxin Feng, and Zhonghui Chen. "Short-term traffic flow prediction with Conv-LSTM." In 2017 9th International Conference on Wireless Communications and Signal Processing (WCSP), pp. 1-6. IEEE, 2017.

Liu, Yunxiang, and Hao Wu. "Prediction of road traffic congestion based on random forest." In 2017 10th International Symposium on Computational Intelligence and Design (ISCID), vol. 2, pp. 361-364. IEEE, 2017.

Lin, Yunduan, and Ruimin Li. "Real-time traffic accidents post-impact prediction: Based on crowdsourcing data." Accident Analysis & Prevention 145 (2020): 105696.

Ma, Xiaolei, Haiyang Yu, Yunpeng Wang, and Yinhai Wang. "Large-scale transportation network congestion evolution prediction using deep learning theory." PloS one 10, no. 3 (2015): e0119044.

MTA. “Central Business District Tolling Program.” 2021. <https://new.mta.info/project/CBDTP>

Najjar, Alameen, Shun’ichi Kaneko, and Yoshikazu Miyanaga. "Combining satellite imagery and open data to map road safety." In Thirty-First AAAI Conference on Artificial Intelligence. 2017.

New York City Metropolitan Transportation Council (NYMTC). n.d. New York Best Practice

Model (NYBPM): Current Model. <https://www.nymtc.org/Data-and-Modeling/New-York-Best-Practice-Model-NYBPM/Current-Model>

New York (NY) Senate (n.d.). Legislation: SECTION 553-K - Traffic mobility review board. <https://www.nysenate.gov/legislation/laws/PBA/553-K>

Rahman, Rezaur, and Samiul Hasan. "Real-time signal queue length prediction using long short-term memory neural network." Neural Computing and Applications 33, no. 8 (2021): 3311-3324.

Ranjan, Navin, Sovit Bhandari, Hong Ping Zhao, Hoon Kim, and Pervez Khan. "City-wide traffic congestion prediction based on CNN, LSTM and transpose CNN." IEEE Access 8 (2020): 81606-81620.

Regional Plan Association (RPA). n.d. Regional Plan Association About. < https://rpa.org/about>

Rich, Jeppe, and Otto Anker Nielsen. "A socio-economic assessment of proposed road user charging schemes in Copenhagen." Transport Policy 14, no. 4 (2007): 330-345.

Santos, Georgina, David Newbery, and Laurent Rojey. "Static Versus Demand-Sensitive Models and Estimation of Second–Best Cordon Tolls: An Exercise for Eight English Towns." Transportation Research Record 1747, no. 1 (2001): 44-50.

Shabestray, Soheil Mohamad Alizadeh, and Baher Abdulhai. "Multimodal iNtelligent Deep (MiND) Traffic Signal Controller." In 2019 IEEE Intelligent Transportation Systems Conference (ITSC), pp. 4532-4539. IEEE, 2019.

Shi, Zhongzhi. "Series on intelligence science (Vol. 1): Advanced artificial intelligence." (2011).

Shin, Dong-Hoon, Kyungyong Chung, and Roy C. Park. "Prediction of traffic congestion based on LSTM through correction of missing temporal and spatial data." IEEE Access 8 (2020): 150784-150796.

Stantec Consulting Services, Inc. (Stantec). “Appendix E: History and Calculation of Traffic, Toll Revenues and Expenses and Review of Physical Conditions of the Facilities of Triborough Bridge and Tunnel Authority.” Prepared for MTA Triborough Bridge and Tunnel Authority. 2020.<https://new.mta.info/document/18936>

Sun, Fangzhou, Abhishek Dubey, and Jules White. "DxNAT—Deep neural networks for explaining non-recurring traffic congestion." In 2017 IEEE International Conference on Big Data (Big Data), pp. 2141-2150. IEEE, 2017.

Sun, Shuming, Juan Chen, and Jian Sun. "Traffic congestion prediction based on GPS trajectory data." International Journal of Distributed Sensor Networks 15, no. 5 (2019): 1550147719847440.

Sun, Shiliang, Changshui Zhang, and Guoqiang Yu. "A Bayesian network approach to traffic flow forecasting." IEEE Transactions on intelligent transportation systems 7, no. 1 (2006): 124-132.

Tseng, Fan-Hsun, Jen-Hao Hsueh, Chia-Wei Tseng, Yao-Tsung Yang, Han-Chieh Chao, and Li-Der Chou. "Congestion prediction with big data for real-time highway traffic." IEEE Access 6 (2018): 57311-57323.

United States Federal Department of Transportation: Federal Highway Administration (FHWA). “Traffic Analysis Toolbox Volume III: Guidelines for Applying Traffic Microsimulation Modeling Software 2019 Update to the 2004 Version.” February 4, 2022. <https://ops.fhwa.dot.gov/publications/fhwahop18036/introduction.htm#:~:text=Microsimulation%20is%20the%20modeling%20of,systems%transit%and%pedestrians>.

URS. “Appendix E: History and Calculation of Traffic, Toll Revenues and Expenses and

Review of Physical Conditions of the Facilities of Triborough Bridge and Tunnel Authority." Prepared for MTA Triborough Bridge and Tunnel Authority. 2009. <http://web.mta.info/mta/investor/pdf/2009/URS_TBTA_2009_Report-FINAL.pdf>

URS. "Appendix E: History and Calculation of Traffic, Toll Revenues and Expenses and Review of Physical Conditions of the Facilities of Triborough Bridge and Tunnel Authority." Prepared for MTA Triborough Bridge and Tunnel Authority. 2002. <http://web.mta.info/mta/investor/pdf/tbta2002fina0l90402klind.pdf>

Wang, Jingyuan, Qian Gu, Junjie Wu, Guannan Liu, and Zhang Xiong. "Traffic speed prediction and congestion source exploration: A deep learning method." In 2016 IEEE 16th international conference on data mining (ICDM), pp. 499-508. IEEE, 2016.

Wang, Haizhong, Lu Liu, Shangjia Dong, Zhen Qian, and Heng Wei. "A novel work zone short-term vehicle-type specific traffic speed prediction model through the hybrid EMD–ARIMA framework." Transportmetrica B: Transport Dynamics 4, no. 3 (2016): 159-186.

Wu, Tong, Pan Zhou, Kai Liu, Yali Yuan, Xiumin Wang, Huawei Huang, and Dapeng Oliver Wu. "Multi-agent deep reinforcement learning for urban traffic light control in vehicular networks." IEEE Transactions on Vehicular Technology 69, no. 8 (2020): 8243-8256.

Wu, Yuankai, Huachun Tan, Lingqiao Qin, and Bin Ran. "Differential variable speed limits control for freeway recurrent bottlenecks via deep actor-critic algorithm." Transportation research part C: emerging technologies 117 (2020a): 102649.\

Xiangxue, Wang, Xu Lunhui, and Chen Kaixun. "Data-driven short-term forecasting for urban road network traffic based on data processing and LSTM-RNN." Arabian Journal for Science and Engineering 44, no. 4 (2019): 3043-3060.

Yang, Senyan, Jianping Wu, Yiman Du, Yingqi He, and Xu Chen. "Ensemble learning for short-term traffic prediction based on gradient boosting machine." Journal of Sensors 2017 (2017).

Yu, Rose, Yaguang Li, Cyrus Shahabi, Ugur Demiryurek, and Yan Liu. "Deep learning: A generic approach for extreme condition traffic forecasting." In Proceedings of the 2017 SIAM international Conference on Data Mining, pp. 777-785. Society for Industrial and Applied Mathematics, 2017.

Yen, Chia-Cheng, Dipak Ghosal, Michael Zhang, and Chen-Nee Chuah. "A Deep On-Policy Learning Agent for Traffic Signal Control of Multiple Intersections." In 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), pp. 1-6. IEEE, 2020.

Yuan, Zhuoning, Xun Zhou, and Tianbao Yang. "Hetero-convlstm: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data." In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 984-992. 2018.

Zahedian, Sara, Amir Nohekhan, and Kaveh Farokhi Sadabadi. "Dynamic Toll Prediction Using Historical Data on Toll Roads: Case Study of the I-66 Inner Beltway." arXiv preprint arXiv:2104.10684 (2021).

Zhang, Jie, W. A. N. G. Junhua, and F. A. N. G. Shou'en. "Prediction of urban expressway total traffic accident duration based on multiple linear regression and artificial neural network." In 2019 5th International Conference on Transportation Information and Safety (ICTIS), pp. 503-510. IEEE, 2019.

Zhao, Haitao, Tianqi Mao, Hongsu Yu, Mengkang Zhang, and Hongbo Zhu. "A driving risk prediction algorithm based on pca-bp neural network in vehicular communication." In 2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), vol. 2, pp. 164-169. IEEE, 2018.

Zeng, Jinghong, Jianming Hu, and Yi Zhang. "Adaptive traffic signal control with deep recurrent Q-learning." In 2018 IEEE Intelligent Vehicles Symposium (IV), pp. 1215-1220. IEEE, 2018.

Zhene, Zou, Peng Hao, Liu Lin, Xiong Guixi, Bowen Du, Md Zakirul Alam Bhuiyan, Yuntao Long, and Da Li. "Deep convolutional mesh RNN for urban traffic passenger flows prediction." In 2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), pp. 1305-1310. IEEE, 2018.