INVESTIGATING TEMPORAL EFFECTS ON TRUCK ACCIDENT OCCURRENCES AND SEVERITY LEVELS IN MANHATTAN

By

Robyn Marquis

A Thesis Submitted to the Graduate
Faculty of Rensselaer Polytechnic Institute
in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

Major Subject: TRANSPORTATION ENGINEERING

Approved by the
Examining Committee:

_____
Xiaokun (Cara) Wang
Thesis Advisor

Rensselaer Polytechnic Institute
Troy, NY

April, 2013
(For Graduation May 2013)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to deeply thank my advisor, Professor Xiaokun (Cara) Wang, for her guidance and patience in this process. I would also like to thank Professor Jose Holguin-Veras for getting me involved in the project work that ultimately motivated this research effort.

I am also indebted to Todd Westhuis, Director of NYSDOT's Office of Traffic and Safety Mobility, and his colleague, Andrew Sattinger, for providing me with the accident database for Manhattan that was used in this analysis.

Special shout-outs to all my colleagues in the Department of Civil and Environmental Engineering at Rensselaer, and most importantly to my friends and family whose support has carried me to this finish line.

ABSTRACT

Each year in the United States, traffic accidents are one of the leading causes of fatalities, and also have injury costs in the billions of dollars. Research into these occurrences and the severity levels mostly considers passenger vehicles and highway segments. This research focuses on accidents in which at least one truck as involved, for a three-year period in Manhattan, New York. The author is part of a research team looking at congestion management techniques in this area, namely the adoption of an off-hour delivery program. This approach shifts truck traffic to the overnight hours from 7:00 PM-6:00 AM, and safety impacts are one of the major metrics in assessing this program's effectiveness.

This report is a preliminary analysis into, through separate modeling approaches, the accident severity level and crash occurrences per Census tract, temporally divided into the four time blocks in the Best Practice Model: AM (6:00-10:00 AM), MD (10:00 AM-3:00 PM), PM (3:00-7:00 PM) and NT (7:00 PM-6:00 AM). An ordered probit model was used to predict the severity level of 2,630 unique, geo-located truck accidents, and it was found that a dry roadway surface, dark roads with street lights, and the times of 5:00-6:00 AM and 1:00-3:00 PM increased the odds of the crash being more severe. The hour 5:00-6:00 AM and increases in either the number of vehicles involved or the traffic flow predicted a lower severity.

The crash occurrence models compared standard with zero-inflated negative binomial, and the latter was preferred for all time blocks except PM. An increase in population, vacancy rate, or hourly truck volume would all increase the expected number of crashes for all models, while certain industrial sectors and commuters who walked led to increases in only some of the models. The commuters who used transit was only significant in the MD zero-inflated, and decreased the expected occurrences. The next steps are to update the vehicle flows from an anticipated new data release, and then model occurrences by severity.

1. INTRODUCTION AND BACKGROUND

It is well-documented that motor vehicle crashes are among the leading causes of fatalities annually in United States, topping the list in 2010 with a total of 32,855 [1]. Additionally, over 2.2 million people are injured annually in these accidents, leading to lifetime costs between the injuries and deaths of over $70 billion [2]. In an effort to better understand, predict, and prevent these accidents and related damages, organizations such as the National Highway Traffic Safety Administration (NHTSA) and Centers for Disease Control and Prevention (CDC) devote funding toward identifying high-risk groups, improving in-vehicle mechanisms (seatbelts and airbags, most notably), and child safety. However, the focus of many of their efforts, and in general among researchers, is on passenger vehicles or highway segments [1, 2].

Large metropolitan areas, including New York City (NYC), have drastically different traffic compositions and geometric conditions than general highways. This research examines the borough of Manhattan, which is a grid configuration notorious for its congestion and delays. According to the 2012 Urban Mobility Report, the greater NYC area ranked first with a total travel delay of over 544 million hours, causing almost 256 million gallons of excess fuel consumption. This was a total congestion cost of $11.8 billion, with $2.5 billion coming from trucks, which includes both the value of time ($88 per truck hour) and other operating costs. The city also ranked 4th in terms of yearly delay per auto commuter (59 hours), with each commuter wasting 28 gallons of gas and $1,281 in 2012 [3].

In an effort to relieve some of this congestion, new traffic demand management techniques have emerged recently that strive to either shift travelers from single-occupancy modes to high-occupancy or transit, or to shift trips away from peak hours. In Manhattan, there has been some advancement toward implementing a full-scale off-hour deliveries (OHD) program which deals with the travel time of day. The innovation of this approach is that it focuses on commercial vehicles, primarily trucks, making deliveries during regular business hours, and tries to shift the trips to the off-hours (7:00 PM-6:00 AM). During a pilot test, it was shown that removing these vehicles from the roadways during the daytime decreased travel times for the remaining vehicles by 3-5 minutes apiece. Additionally, the travel speeds of the delivery vehicles increased from an average of 3 mph to 8 mph (under OHD), leading to a travel time savings of 48 minutes per delivery tour [4].

In addition to the benefits mentioned, the OHD pilot showed increases in productivity, decreases in vehicle emissions from reduced idling times, and reductions in parking fines, among others. The combination of all of these has major economic implications, in the range of $100-200 million in savings annually from a full-scale program. A major hurdle, however, is

convincing the receivers, who want to receive goods during business hours, to make the change to either: (1) staffed OHD, which requires paying personnel to stay late; or (2) unassisted OHD, which may require the installation of specialty equipment that would allow the carrier to deposit the goods—both of these options require the receivers to incur a cost, while most benefits are felt by the carriers. This market imbalance necessitates the use of public policies to provide incentives to participating receivers, and a quantification of all benefits in order to assess the effectiveness of the program versus its costs [4].

The research team developing the OHD program, this author included, has been tasked with determining these metrics, which include traffic safety and delays. There are common misconceptions that trucks are the root cause of the traffic delays, and that incidents involving trucks tend to be more severe. While it is true that clearance times for accidents involving trucks are higher than with passenger vehicles, these vehicles only account for about 6% of the daily flows and 2.6% of annual accidents within Manhattan. However, in order to truly capture the full impacts of OHD, it is necessary to look at these truck accidents, both in terms of severity and crash occurrences, prior to and after implementation. This paper includes models for each, with a focus on temporal differences.

The following chapter is literature review on severity and occurrence modeling efforts. Chapter 3 outlines the data sources used in the analyses and highlights basic statistics of key variables. Chapter 4 emphasizes the model specifications for both analyses, while Chapter 5 and Chapter 6 discuss the results of these modeling efforts for the severity levels and crash occurrences, respectively. Future research steps to improve and advance these models are in Chapter 7, and the paper is concluded in Chapter 8.

2. LITERATURE REVIEW

As discussed in the introduction, full quantification of the impacts of shifting truck traffic to the overnight hours requires a study into both the severity levels and the number of occurrences within a spatial unit. The bulk of the literature reviewed below analyzes these separately; however, there has been more recent work into modeling the crash counts by severity level, so a few key studies in this area are also represented, though there are not many available yet.

2.1 Accident Severity Literature

This first set of papers highlights research into the severity level of motor vehicle accidents, and a summary of these models and related data issues can be found in Savolainen, et al. [5]. Among these issues is the underreporting of crashes and omitted variable bias. The former may be the case in the dataset used for this report as accidents below a threshold of property damage were not reported, and the latter stems from the incomplete data reporting, so some factors that are influential are not included. It is also discussed that the severity levels are ordinal by nature, and some model specifications cannot capture this relationship, or may not account for unobserved effects between adjacent category levels, again leading to bias. Spatial and temporal correlations may also be present.

The models used to predict the outcome severity of an accident can, for the most part, be broken into three overarching groups: (1) binary outcome models; (2) unordered multinomial discrete outcome models; and (3) ordered discrete outcome models—the latter is employed in this research analysis, so the bulk of the literature will be on these approaches. The ordered probit model is particularly beneficial because, as noted, severity levels have a natural ordering to them, such as a fatality being most severe, followed by capacitating injury, then non-capacitating, and so on. They may be interdependencies among the outcome levels which can be captured by this model, and the degree to which factors influence the overall severity can be easily determined.

Starting first with the binary methodologies, these are papers in which the severity levels are modeled as injury versus non-injury or fatal versus non-fatal crashes, and the most common approaches are binary logit and binary probit. It should be noted here that one step in the crash occurrence modeling involved the use of binary logit models, as described in Chapter 4, so the methodologies were useful for that process; however, the application to the data is what differs.

Al-Ghamdi [6] employs a standard logistic regression model on accident data that was split into a binary response of either fatal or non-fatal—there was no further distinction made in this secondary category of level of injury or property damage. As is mentioned in Chapter 3, the data for this study had very broad categories and also did not classify levels of personal injury, but the percentage of fatal accidents is so low that this type of binary grouping may not be beneficial. It

was found that the odds of being in a fatal accident are higher at non-intersection locations, when the driver runs a red light, and when the driver was going the wrong way, among other causes, and this cause was found to be the most influential on placement in the fatal category.

One interesting study using the binary logit model is that of Lee and Abdel-Aty [7], where the impact of the presence of passengers is examined. Bivariate probit models are used on different crash-related and passenger-related factors to first predict the probability of a crash occurring given that a passenger is or is not present, and it was shown that someone else in the car does have an impact. The second part of the analysis uses a binary logit model to determine if the presence of a passenger, along with other predictors, more likely leads to a single-vehicle or multi-vehicle accident. It was found that younger drivers with younger passengers, such as a teenager with peers, are more likely to be involved in single-vehicle incidents than any other age group. This shows how these drivers are easily distracted by conditions within the car, and may drive less cautiously in front of friends.

There are also some papers which deal with variations of the binary logit, particularly to deal with the biased parameter estimates when there are within-crash correlations; that is, when there are multiple individuals involved in the same incident, there may need to be additional considerations. Ouyang, et al. [8] examines such a case for car-truck collisions, so multiple vehicles involved, within Washington State. A simultaneous binary logit model was compared to a standard binary logit, which cannot account for any correlations in the same accident, in order to highlight the importance of this inclusion. It was also shown that the parameter estimates for the various roadway design indicators varied greatly between the two models. A similar study was conducted by Huang, et al. [9], instead using a hierarchical model for multi-vehicle crashes with random effects on data from Singapore. This was done as there may be factors common the different drivers in the same crash that are not otherwise observed. It was found that crashes at night and in the right-most lane tended to be more severe, while heavy vehicles such as trucks were more resistant to increased severity. It is important to point out that these modified approaches can only be used when the data separates out the severity level for each driver; in many cases, the overall severity of a single incident is recorded, with no distinction of which vehicle or person was affected.

The next set of papers deals with unordered multinomial discrete outcomes, the most basic of which is the standard multinomial logit (MNL). However, since it has been emphasized that the severity levels are, in fact, ordered by nature and MNL does not account for this fact, these models will not be discussed. There are, however, variations that have proven to be useful, and in some cases even outperform the ordered response models by relaxing some of the constraints.

The nested logit is one that is more popular, as is shown in Shankar, et al. [10]. The goal of this study, as with the others, was to determine the probability that a given accident falls within one of the levels, and these functions have measurable characteristics (such as weather, roadway design, etc.) and a random component. Normally distributed error terms give way to the probit model, which can be difficult to estimate, so a closed form generalized extreme value (GEV) distribution is more commonly used. The GEV approach, however, operates under the assumption that the error terms are not somehow correlated, so there may be specification errors. The nested logit model fixes these issues by determining the probabilities as a difference in the severity level functions, and the nested structure was used to distinguish between property damage only and possibly injury, combined under no evident injury. A motivating factor for their study was the increase in Intelligent Transportation Systems (ITS) as a means to improve vehicular safety, and the authors published an earlier work on crash occurrences related to such improvements, which is discussed in the Chapter 2.2.

The nested logit model is also used in Savolainen and Mannering [11] where motorcycle crashes, either single- or multi-vehicle, are examined more closely due to a drastic increase in motorcycle fatalities per year. Data from Indiana was used, and the levels of no injury and non-incapacitating injury were nested under minor or no injury—the other two levels were fatal injury and incapacitating injury. A variety or roadway, rider, and crash characteristics were examined, with an increase in driver age leading to more severe injuries. Other common factors like alcohol consumption, helmet use, excessive speed, and roadway characteristics were also found to be significant in predicting the severity.

Ordered response models, either probit or logit, are the most commonly used in severity analyses. The literature included here is for ordered probit models as these were used in the analysis, although they are sometimes difficult to estimate; these is little difference between the two, though, just in how the errors are specified. A limitation of these models is that they are highly susceptible to the underreporting of accidents mentioned, and the shifts in threshold from individual variables are constrained to be in the same direction across all levels.

One paper that was particularly interesting for the analysis in this report is Duncan, et al. [12], which uses an ordered probit model to predict the severity of truck-passenger car rear-end collisions. Not much of the available literature focuses on truck accidents, and although this work looks at a very specific subset of incidents, it still provides insights into what factors may be influential in the analysis. It was found in the models that darkness, grades, and high speed differentials increased the severity, while congestion and inclement weather tended to decrease

it—these data are also available in the provided accident files used within this analysis, as outlined in Chapter 3.

The use of ordered probit models appears again in Kockelman and Kweon [13], which examines a wider range of accident types and vehicles involved. There were datasets prepared for all crashes, two-vehicle crashes, and single-vehicle crashes, and different results were determined with and without speed variables. In the results, the full models included variables with low statistical significance, as they were expected to have some impacts and the fact that they did not was addressed. For all crashes, it was found that both older drivers and vehicles were associated with more severe injuries, as was alcohol use. Males and those with prior citations had negative coefficients, which could have been from high involvement in non-severe crashes; these models only focus on accidents that have already occurred, not whether these groups are at a greater risk of more severe crashes in general. As found in Huang, et al. [9], trucks, both light- and heavy-duty, seemed to protect the driver and lead to less severe injuries.

Driver injury severity levels are examined in Abdel-Aty [14] for crashes in Florida, with different ordered probit models estimated for signalized intersections, roadway segments, and toll plazas—the latter is particularly interesting as not many studies look into this. Across all models it was found that the type of vehicle, speed, point of impact on the vehicle, and driver's age, gender (female indicator), and use of a seat belt were significant, and led to more severe injuries. Horizontal curves and lighting conditions were influential for roadway segments, with daylight decreasing the severity. For toll plazas, the presence of electronic tolling led to a higher severity, as the approach speeds for these lanes are greater.

Single-vehicle crash severity in Singapore was the focus of Rifaat and Chin [15], and an ordered probit model was developed for a large variety of factors, including time of day (peak, off-peak, and nighttime), type of vehicle involved, crash type, pre-accident incidents, roadway type, geometry, and surface conditions, and driver attributes such as gender and age—all of these indicators were included in the accident set used in this study and many were examined in the models. There were only three levels of response here: fatality, serious injury, and slight injury; there were no categories examining property damage or no apparent injury. The results were that accidents occurring during the nighttime, classified from 8:00 PM-7:00 AM, had significantly increased severity, and it is discussed that the lower density of traffic during these times allows drivers to speed. The fatality risk for a truck accident during this time was 1.80 times that of a car, due to the higher vehicle momentum in the collision.

In contrast to the probabilistic models above, Edwards [16] uses severity ratios, with the number of fatalities and serious accidents highlight as a proportion of the total. The relationship

between weather conditions and accident severity for road accidents in England, where there are high occurrences of rain, fog, and wind, is determined. These inclement conditions were compared to clear weather, called the nonhazardous state. Since the available data was aggregated at the county level by the local authorities, accident rates were used instead of the actual numbers in order to compare across this spatial unit. This approach, as noted, is difficult to interpret when the numbers are small and there is great variability. It was found that fog led to the most severe injuries for all conditions, but there were large geographical variations across the country. The severity tended to decrease in rain compared to fine weather.

2.2 Accident Occurrence Literature

The set of papers discussed above focus on modeling the injury severity level; however, the bulk of the analysis in this research is on the crash occurrences, so these papers will be discussed next. There are several issues that can be encountered with count data modeling, and these are summarized by Lord and Mannering [17]. The issues that frequently arise with these models are over- or under-dispersion, temporal or spatial correlations, under-reporting of accidents, omitted variables, and endogenous variables, among others—this list is very similar to the issues that were discussed with severity level modeling. There are a variety of model formats that can be used to handle these issues, depending on what is prevalent, yet each has its own limitations, as will be addressed by the specific works below.

The most basic modeling approach is a standard Poisson regression, which can be used when the mean and variance of the data are equal, as there is only one parameter in the model. While there have been a few studies which solely use Poisson to estimate frequencies, these models are very susceptible to low sample means and bias from small sample sizes. Joshua and Garber [18] uses such a model to estimate accidents in which trucks are involved for a highway segment in Virginia, and it was found that the frequency was influenced by, in addition to geometric factors, the total average daily traffic, the percent of this traffic that was classified as trucks, and the speed differential between trucks and non-trucks. Jones, et al. [19] also has a Poisson model for crash counts, in this case for urban freeways around Seattle, and also relates this back to duration of accident as it impacts commuter travel times.

Daniels, et al. [20] considers both a Poisson model and Gamma model for data pertaining to single- and multi-vehicle crashes at roundabouts in Belgium—the latter is used when the crash types are modeled separately as the data are under-dispersed, meaning that the variance is lower than the mean. It is commonly assumed that roundabouts are safer than standard interactions as they have been shown to reduce the number of accidents, and this study further confirms that

notion. There was an increase in counts when cyclists or mopeds were present, and roundabouts with designated cycle lanes were worse.

The most studies, by far, employ the negative binomial (NB) regression model, which is also used in this research; as a result, the literature will focus more on comparing these works. The NB model is used when the assumption behind Poisson of the variance equaling the mean is violated, referred to as over-dispersion when the variance is greater than the mean; however, it cannot handle under-dispersion, and is also impacted by small sample sizes and low sample means.

Miaou [21] is one work which compares Poisson and NB to establish a relationship between road geometries and truck accidents. This study uses hypothetical roadway segments of different truck accident involvement levels to evaluate which models performed better, but there are no conclusions on actual influential factors. It was found that, for maximum likelihood estimations, Poisson and NB were close in performance, so it was recommended to start with a Poisson model to start establishing relationships.

Shankar, et al. [22] is the parallel work to the paper discussed in the severity literature, and again focuses on a highway segment outside of Seattle. The NB model relates roadway geometry, weather, and other seasonal impacts to the frequency—the goal was to better understand the negative effects of inclement weather on segments with difficult geometries, in order to make suggestions for ITS improvements. It was found that segments with multiple curves and higher design speeds led to higher counts, as did snowfall on steeply graded segments, among others.

The effects of ITS are also examined in Carson and Mannering [23], which specifically looks at the use of ice warning signs on accident occurrences under these conditions; it is highlighted that warnings for non-permanent conditions are more difficult to place since they are unpredictable (in contrast to hazards from geometry). Although it was found that the presence of these signs would not significantly reduce counts, those factors that were significant provided insight into placement. A standard NB model was used for state highways, and straight roadways, high daily traffic per lane, longer segments and horizontal curves, and high posted speeds led to increases in ice accidents.

Accidents within the urban setting, specifically at intersections, were the focus of Poch and Mannering [24]. Data across seven years for 63 intersections in Bellevue, Washington, being considered then for operational improvements was evaluated using NB, and it was found again that there are strong interactions between crash counts and geometric designs and traffic characteristics. Different models were estimated for the types of crashes, such as rear-end versus at an angle, as well as for all accidents combined. Intersection approaches with two or more lanes or those with a combined left-through lane showed higher frequencies than those without these

conditions when looking at the total. There were also increases from higher opposing volumes and approach speeds.

Crashes on principal arterials across Washington State were also investigated in Milton and Mannering [25] using a NB model. As was shown in similar studies, the predictors of interest were highway geometries and traffic conditions, and different models were estimated for East and West state routes, as there tend to be vast differences in weather and terrain; however, weather indicators were not directly included. In both models, accident counts were increased by higher volumes, more lanes, and narrow right and left shoulder widths. Frequencies were decreased by having larger horizontal curve radii or sharp curves, higher posted speeds, higher percentage of single-unit trucks in the west and a larger portion of all trucks in the east, and greater peak hour factors. However, the percentage of trucks increases drastically by only adding a few trucks when the daily traffic is low, so this result may be attributed more to the lower chance of conflicts, not the change in composition.

The emergence of panel data and the subsequent issue of heterogeneity are addressed in Karlaftis and Tarko [26], which incorporates socio-economic data in addition to vehicle or roadway factors for Indiana State. Common approaches like the fixed and random effects models cannot be practically applied to count data since marginal effects and accident rates cannot be computed. Instead, the authors used a cluster analysis by disaggregating the dataset into internally homogeneous clusters using a minimum dissimilarities algorithm. Separate NB models were then estimated for each, and it was proven that this approach better described the data than the combined, single model.

The impacts of low sample means and small sample sizes in the NB estimation process are outlined in Lord [27]. Data with low sample means is said to have a low mean problem (LMP), and this is the first study to see how this affects the dispersion parameter—its inverse is important for developing confidence intervals. It was shown that the probability of this parameter being unreliably estimated increases rapidly with a decrease in sample mean and sample size, regardless of the underlying estimation method. These impacts were only compounded further when Bayesian estimates were used.

Malyshkina and Mannering [28] uses a NB model for the frequency portion of the analysis (the severity is discussed previously). There are similar variables to the other studies in this section, such as average annual daily traffic, number of lanes, and roadway curvatures. There are also further considerations for roadway segments on bridges and those with some other design exception, though neither of these provided to be significant. The results did show that, among typical predictors, an increase in the number of freeway ramps on a roadway segment increased

the crashes, which highlights that merging is a dangerous action. Sharper curves were associated with a decrease in incidents, as drivers tend to be cautious in these areas and reduce their speeds.

Looking more closely at count data with many zeros is Ridout, et al. [29], where the distinction is made between inevitable, or structural, zeros and sampling zeros that occur by chance. That is, there may be some observations that will always be zero for the given predictors, while others just happen to have a zero count during the observation period—this is crucial for understanding the zero-inflated models as the inevitable zeros are governed by their own set of variables.

Miaou [21] tests the performance of the zero-inflated Poisson (ZIP) regression against the standard Poisson and NB already discussed. The zero-inflated model is useful for evaluating whether a high occurrence of zeros stem from a different set of predictors than other count outcomes. There were a large number of roadway segments in this study that reported zero accidents for the five-year period, so the ZIP model was used to determine if some of these segments would always be zero, or just happened to be for the time period provided. It was merely proved in this work that the ZIP performed the best of the three, for all estimation approaches, so there is no lengthy discussion on actual influential factors.

Carson and Mannering [23] mentioned above also uses a ZINB, looking at the interstate highways. The South Central and Eastern state regions had higher frequencies, in addition to the variables mentioned in the standard NB model. For the inflated model, crashes within year 1994 were more likely to be in the certain-zero group, and interestingly so were roadway segments without horizontal curves. It is noted that 1994 was an El Nino year, so the temperatures were milder and there may have been less ice. The alpha parameter and Vuong statistic were both quite large, confirming that ZINB was the best model here.

Lord, et al. [30] compares Poisson, NB, ZIP, and ZINB to provide a guide on how to efficiently model the interaction between road safety and traffic exposure. The basis of this analysis is on identifying that a crash is essentially the result of a Bernoulli trial with a success indicating that an accident did occur. This leads nicely into the use of Poisson or variations thereof, as there are a large number of individual trials—every time a vehicle enters an intersection, uses a roadway segment, etc.—and a correspondingly low probability of a success. It was found that, while zero-inflated models may improve the fit measures, there is an assumption of a dual state process that might not be consistent with the data, and the safety of some crash locations may be misinterpreted.

For certain types of crashes, such as angle, the literature have shown inconsistent impacts of left-turn only lanes, as noted in Kim and Washington [31]. Since the introduction of these lanes is

typically driven by the crash occurrences at the intersection, this introduces endogeneity into the model. The authors propose a limited-information maximum likelihood (LIML) estimation to better predict frequencies. The results from this approach are compared to the NB without controlling for the endogeneity. The left-turn lane indicator in the NB has a positive coefficient, which is counter-intuitive as these lanes have been shown to reduce angle crashes; however, this is the point of introducing LIML, and this model has a negative coefficient as originally expected.

In some cases, there is a need to specify the structure of the error term for the dependent variable in the NB model, which can be done using generalized linear models (GLMs). Mountain, et al. [32] uses a GLM in combination with an empirical Bayes (EB) approach to increase the accuracy of the coefficients. It is noted that EB can account for unexplained variations in frequency between otherwise identical—with regard to predictors in the NB regression—spatial units. The locations may have the same expected crash counts using the available regressors, but in reality there are outside influences leading to differences. The study looks at main roads with minor junctions for highways in the United Kingdom, and confirms that frequencies varied in a non-linear fashion based on the traffic flow or link length. This work is expanded in Mountain, et al. [33], which has flexibility for accident risks changing over time. The same approach is used in Cafiso, et al. [34] with similar outcomes regarding exposure variables.

Heydecker and Wu [35] also looks at a Bayesian correction, in this case for a log-linear model that is based on the standard NB, and also for accidents in Britain. The goal was to identify sites that may be appropriate for accident remediation measures based on a threshold value for the number of observed occurrences over the most recent three years. Within this framework, it is noted that there would be selection bias, so instead the expected accident frequency in the future is used to mark trouble locations. The effects of changes to speed limits, geometries, and flows were examined as potential methods to decrease the accident occurrences at these sites.

Hirst, et al. [36] describes that these types of accident prediction models often become outdated so the effect of the improvements is exaggerated. This is due to the fact that it is assumed that the risk of accidents per unit of exposure (such as traffic flow) remains constant over time. In reality, these road safety initiatives are aimed at decreasing such risks, so the ratio is not expected to stay the same. Simulations using typical time periods and accident parameters were generated in order to develop corrections to these models.

Aside from the standard Poisson and NB models, and their respective zero-inflated counterparts, and the GLMs, there are several other models that have been applied toward crash occurrence data. As examples, these range from Gamma (to handle under-dispersion) to random-effects models (when there is temporal or spatial correlation) to Neural Networks (flexibility in

functional form). However, most of these are outside of the scope of this work and will not be further elaborated upon. A few that were found to be interesting and perhaps a basis for further modeling steps with this data are mentioned briefly below.

The models outlined in this review deal mostly with cross-sectional or time series data, and there are some newer models which can better incorporate temporal effects—many accident data sets represent annual trends, so there are repeated measurements over time. Quddus [37] looks at the integer-valued autoregressive (INAR) Poisson model, which is useful in accident frequency analysis because it has the same properties as the Poisson regression with additional capabilities for handling serial autocorrelation. This model performed well for the disaggregated data of small spatial and temporal units where the counts in each were low. It is noted, though, that the model has limitations with seasonality.

Xie, et al. [38] provides an empirical comparison of NB with back-propagation neural network (BPNN) models and Bayesian neural network (BNN) models. Neural network approaches are now being used in statistical applications as it is easy to train the model what the correct output should be for given inputs, and it is not necessary to understand the algorithms in between. BPNN models are useful when there is a large amount of input data but it is unclear how exactly this may relate to the output, and if the solution may change over time; however, it is noted that they tend to over-fit the data, so to deal with this, BNN models are being used. This report showed that the neural network models performed better than the NB, and that the BNN had better generalization than the BPNN.

A similar comparison was made in Li, et al. [39], where Support Vector Machine (SVM) models are used instead of BNN to compare to NB, and these were shown to out-perform the BPNN models, as well. SVM models are a newer class of models for predicting values and are based on the statistical learning theory. This is a new research field that combines statistics, probability, optimization, and computer science to study the performance of computer algorithms in making predictions from training data. These models are universal approximators for any multivariate function with reasonable accuracy, and this is the first case of applying them to crash frequencies. The disadvantages, though, are that they do not have a functional form, similar to NN models, and there are some parameters that need to be found before training the model. The model also takes a long time to run for large data sets.

2.3 Accident Occurrence by Severity Literature

The two major sets of literature above focus individually on injury severity and crash occurrences, and there have been some more recent studies into predicting the occurrences by severity level. Only a handful of papers will be included here as this type of analysis has not yet

been performed, and is part of the next steps following the availability of updated data (see discussion in Chapter 7). As part of that continued effort, a full review into the available literature on this topic will be performed; however, some are included here to gain insight into how to steer that modeling effort.

In Ma and Kockelman [40], it is discussed that the frequencies and severities are typically modeled independently using single equation methods. By not including the information of the other, this could lead to biases or unobserved errors, and reduce the efficiency of the coefficient estimates. A multivariate Poisson (MVP) specification is used to model injury counts by severity level, which is estimated using Bayesian methods—this appears to be the first study to do so. The parameters for the Bayesian statistical inference are determined using a Gibbs sampler and a Metropolis-Hastings (MH) algorithm. Gibbs sampling and the MH algorithm are both Markov chain Monte Carlo (MCMC) methods used to generate random samples for a specified distribution when direct sampling is difficult. It was shown that the MVP outperformed a univariate approach, and that the results make more sense. For instance, fatal injury rates increased with the increase in speed limit, but this was not significant in the univariate model.

An extension of this work is found in Ma, et al. [41], again indicating that individually estimating occurrences separately from the severity levels leads to issues with the results. Although Bayesian methods and MCMC simulations are again used, the overall model here is a multivariate Poisson-lognormal (MVPLN) using rural highway data in Washington State. The MVPLN is more flexible with allowing over-dispersion, and uses parameter estimates from univariate Poisson models as the starting point. It was found that roadway characteristics were especially significant, with sharper curves have more severe injury crashes, while larger roadway shoulders have the opposite impact.

Park and Lord [42] is another paper that focuses on the MVPLN model for considering the crash frequency by severity, in this case looking at a count of accidents, not driver injuries. The discussion of why this model is beneficial is essentially the same as above, and it is also mentioned that these models are more general than the MVP and multivariate NB, as the latter cannot account for negative correlations. An additional MCMC method was adapted (Newton-Raphson) from previous code in order to estimate the parameters within the Bayesian framework. For the crash data at three-legged unsignalized intersections in California, it was found that some variables had counterintuitive outcomes, such as the presence of lighting leading to more crashes in the higher severity levels. It was also found that the sign of some coefficient estimates changed between severity levels.

3. DESCRIPTION OF DATA

The original incident data for these analyses were provided by the New York State Department of Transportation (NYSDOT) for May 2008 to April 2011, and included separate files for: (1) the event information; (2) the vehicles involved; and (3) the contributing factors. All accidents that are classified as reportable, versus non-reportable, are required to have all three of the files. A reportable accident is one in which there is a fatality, a personal injury, or property damage to any individual vehicle of over $1,000.

The first file has a record for every separate accident that was reported and includes information on the date, time, and severity of the crash, in addition to other factors such as lighting, weather, and roadway characteristics. The second file has more data points as each incident could have one or more vehicles involved, so each vehicle has its own record. Within this file there is information on the vehicle itself—such as classification, pre-accident action, and direction of travel—and also on the driver. The third file provides elaboration on what may have led to the accident, e.g., speeding or alcohol. The first two files described were of most importance, and were merged to pair all vehicles with their event record.

Once merged, the data were processed to identify which variables may be of most interest in the severity analysis, and to eliminate incomplete records. The vehicle type indicator was used to filter out those involving trucks specifically, and any rows without geolocation or time of day data were removed, as both spatial and temporal effects are the focus. After further consideration and due to lack of complete data, it was decided that no vehicle-specific variables would be used in the analysis, so duplicates were removed. That is, some incidents may have involved more than one truck, so the data pertaining to the incident itself, as described by the first file above, would have been duplicated in the modeling inputs.

The vehicle flow was also included in the process, and this information was derived from the Best Practices Model (BPM) developed by the New York Metropolitan Transportation Council (NYMTC). The BPM is a roadway network of the greater NYC area that includes, among other features, the flow for each link by time block. These four time blocks are: (1) AM, from 6:00-10:00 AM; (2) MD, from 10:00 AM-3:00 PM; (3) PM, from 3:00-7:00 PM; and (4) NT, from 7:00 PM-6:00 AM. Since the values in the BPM are applicable across the whole block, the flows were divided by the number of hours within each to estimate the average for a single hour.

It should also be noted that the BPM does not include every single link actually present in NYC, so ArcGIS was used to create buffer points of 250 feet, which is roughly the size of a few blocks and the gap between BPM links. This means that each specific point was turned into a circle that would intersect with links within 250 feet, and then the output provided an average

volume in that vicinity based on its closest neighbor links, which was used in the severity level analysis. As described later, the crash occurrences are based on the Census tract level, so the volume data were incorporated into that process by summing the flows on all links intersecting the respective tracts, and then again were divided into hourly averages.

The variables that were considered in the severity analysis are listed in Table 1, and some are highlighted in the descriptive statistics below. Since the majority of the data reported for accidents is in the form of codes used by the police and departments of transportation, this generated a lot of indicator variables as shown. The vehicle flow was then divided by 1,000 to ensure that the coefficient estimates in the modeling process were closer in order of magnitude.

Table 1: Variables used in severity modeling

Variable	Description
num_of_veh	Number of vehicles involved in the incident
flow000	Vehicle flow in 1000s
hour[X]	Time of day, where [X] is a value from 1-24 with hour1 being 12:01-1:00 AM. Some of these were later combined into "hours[X]_[Y]"
daylight	Indicator = 1 for full natural lighting; 0 otherwise
dawn	Indicator = 1 for dawn lighting; 0 otherwise
dusk	Indicator = 1 for dusk lighting; 0 otherwise
dark_lit	Indicator = 1 for dark with streetlights; 0 otherwise
dark_unlit	Indicator = 1 for dark no streetlights; 0 otherwise
clear	Indicator = 1 for clear skies weather; 0 otherwise
cloudy	Indicator = 1 for cloudy skies weather; 0 otherwise
rain	Indicator = 1 for rainy weather; 0 otherwise
snow	Indicator = 1 for snowy weather; 0 otherwise
wintery	Indicator = 1 for hail/sleet/freezing rain; 0 otherwise
strai_lvl	Indicator = 1 for straight, level road; 0 otherwise
strai_grd	Indicator = 1 for straight, grade road; 0 otherwise
strai_cst	Indicator = 1 for straight, at hillcrest; 0 otherwise
curve_lvl	Indicator = 1 for curved, level road; 0 otherwise
curve_grd	Indicator = 1 for curved, grade road; 0 otherwise
curve_cst	Indicator = 1 for curved, at hillcrest; 0 otherwise
surf_dry	Indicator = 1 for dry road surface; 0 otherwise
surf_wet	Indicator = 1 for wet road surface; 0 otherwise
surf_snow	Indicator = 1 for snow on road; 0 otherwise
surf_slush	Indicator = 1 for slush on road; 0 otherwise
surf_flood	Indicator = 1 for flood water on road; 0 otherwise

The dependent variable for the severity modeling was originally provided as five levels, with “injury” and “property damage and injury” as separate values with the latter coded as less severe than injury alone. It was decided to combine these two into one, as there was no distinction in the degree of personal injury. The percentages of observations within each category are shown in Figure 1. Slightly more than half of the incidents (52.17%) fell in the aforementioned combined category, while slightly less than half (46.52%) were property damage only. Less than half of the events led to a fatality.

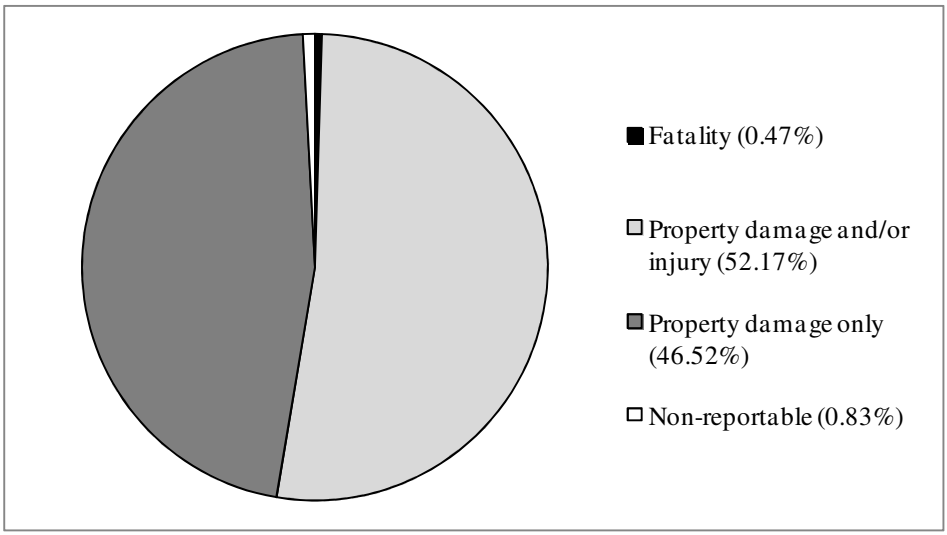

Figure 1: Pie chart of severity responses

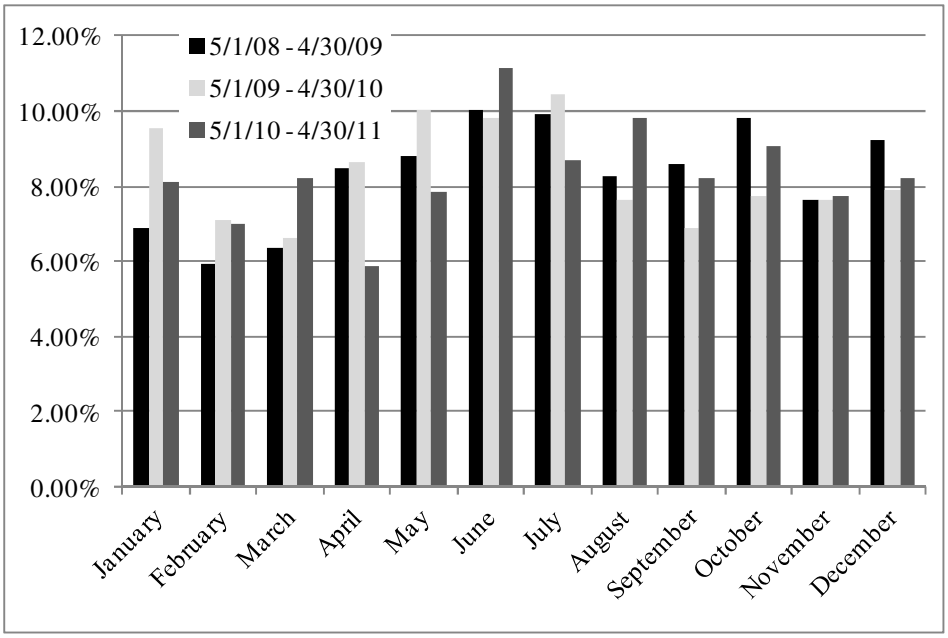

Figure 2: Distribution of annual truck accidents by month

The first set of figures show a breakdown of the accidents by year within each respective variable. For reference, there were a total of 929, 786, and 815 crashes within years one, two, and three, respectively. The date of each event was recorded, so Figure 2 highlights these by month of the year. It can be seen that overall there tended to be more crashes within the warmer months (May through August), with some variation in percentages between individual years. For example, the third year of data shows a much lower percentage in May (7.85%) than in June

(11.17%), while year two has the opposite trend. The values across all years range from 6.64% (February) to 10.32% (June) with most months in the 7-8% level, and an average across all months of 8.33%. It is interesting to note that months that typically experience worse weather actually account for a lower piece of all accidents. This could be due to that fact that in a metropolitan setting such as NYC, traffic moves even slower than usual during inclement weather and some accidents are avoided, or perhaps travelers shift to other modes. The reported weather and road surface conditions for these events are described in Figure 7 and Figure 8, respectively.

Figure 3 and Figure 4 show the distribution of events across the 24 hours in a day, with midnight to 11:59 AM in the first, and noon to 11:59 PM in the second. By examining them both, it is clear to see that the majority of the accidents occurred during the typical business hours when there are more vehicles on the roadways. Across all of the years, the average percentage of events in the AM period ranged from 2.81-7.83%, with an overall average of 5.67% between those four hours in the time block.

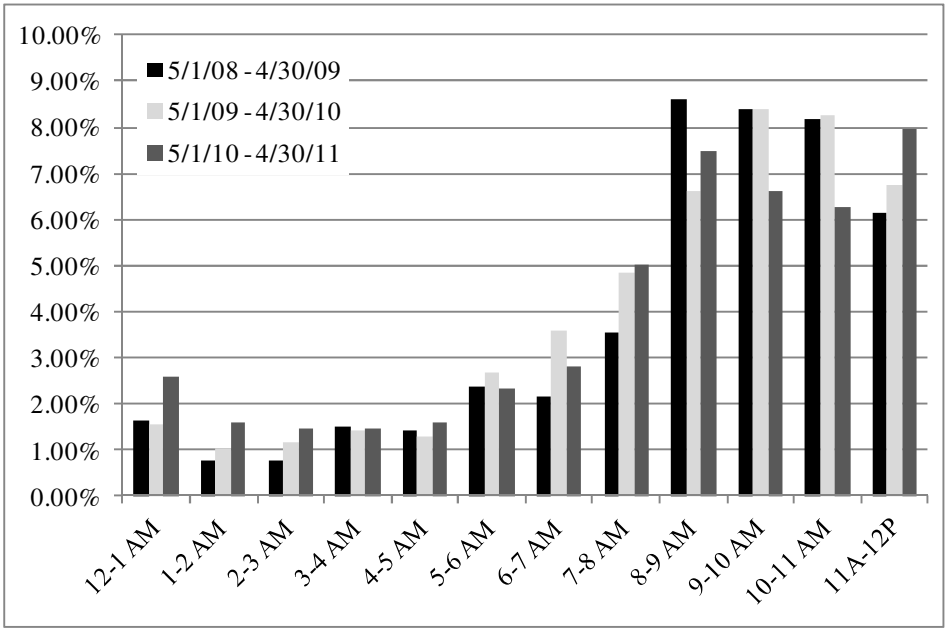

Figure 3: Distribution of annual truck accidents by hour, 12:00-11:59AM

Similarly, the values for the MD, PM, and NT blocks, respectively, were: 6.56-8.62%, 7.30%; 3.52-5.97%, 5.09%; and 1.11-3.28%, 1.86%. The hour from 12-1 PM has the highest average hourly value of 8.62%, and it can be seen that there is a spike in the first year of 10.66% (99 of the 929 crashes). Temporal effects are the primary focus of the crash occurrence analysis, so this breakdown starts to show some interesting trends.

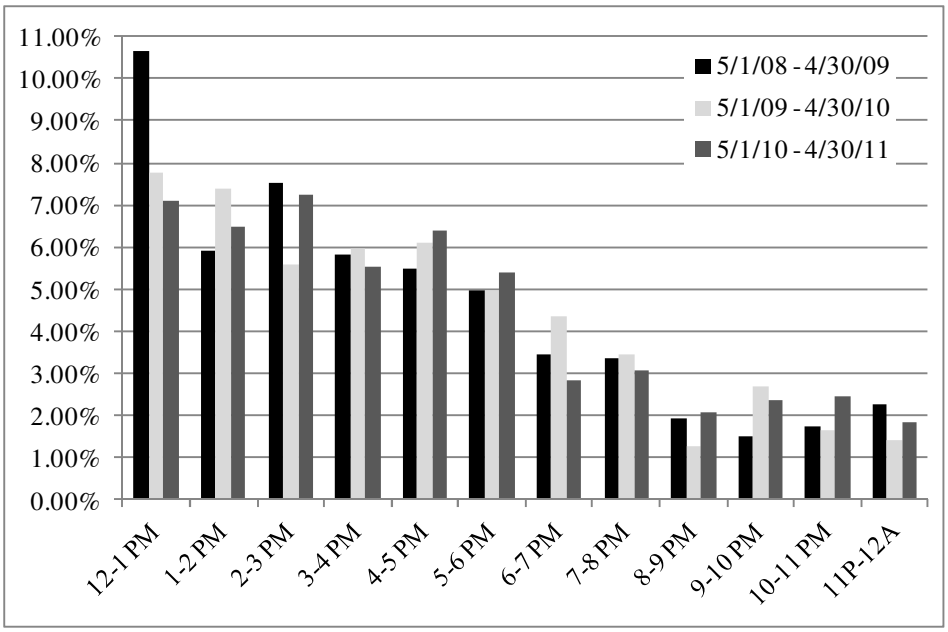

Figure 4: Distribution of annual truck accidents by hour, 12:00-11:59PM

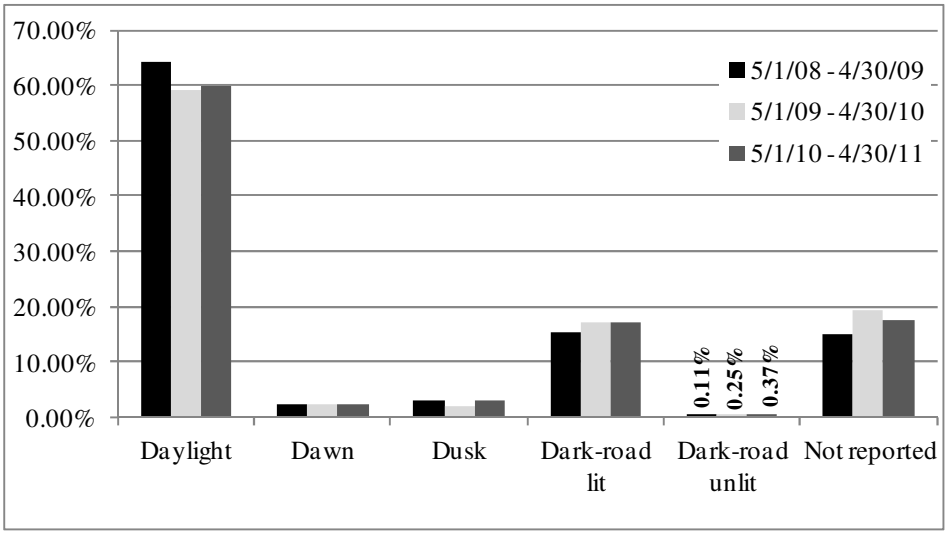

Figure 5: Distribution of annual truck accidents by lighting condition

The remaining set of figures looks at the lighting, weather, and roadway characteristics. Figure 5 has the distribution of natural lighting conditions, which are in some ways related to the time of day. In all years, around 60% of the accidents occurred during daylight, for a total of 1,551 of the 2,530 observations. An average of 16.56% were during darkness on roads lit by streetlights, with barely any (six across the years) under unlighted conditions—given that Manhattan is a major cityscape, almost every roadway has some form of lighting. Roughly 2%

occurred during each the dawn and dusk lighting conditions, and another 15.07-19.21%, depending on the year, did not have lighting reported.

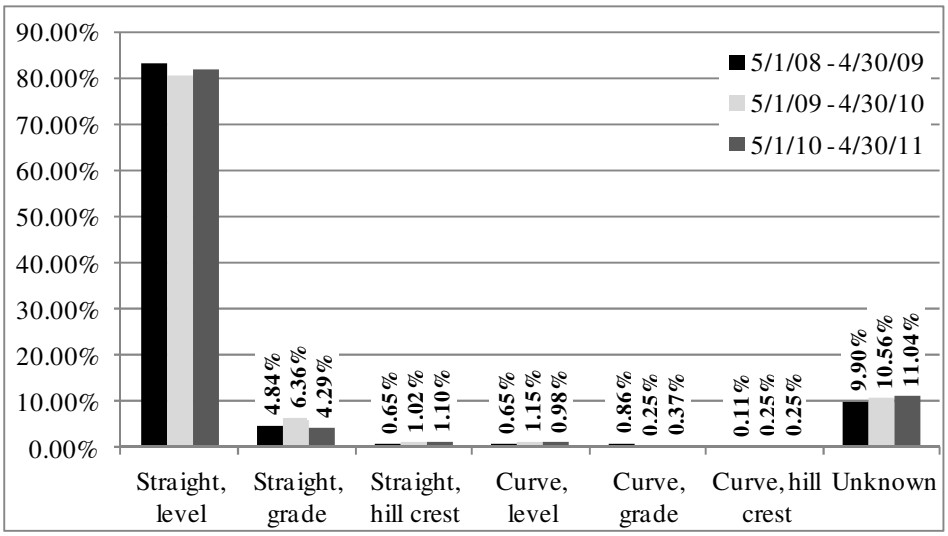

Figure 6: Distribution of annual truck accidents by roadway characteristic

Figure 6 also highlights another characteristic of major metropolitan areas, particularly one like Manhattan where the roadways are a grid, as a large majority (81.86% average) of accidents occurred on straight, level roadway segments. Only straight roads with a grade had notable observations, and still only ranged from 4.39-6.36%. The number of incidents without a road characteristic reported was higher than straight, grade and all of the other non-straight, level possibilities combined, with an overall average of 10.47%, lower than the value observed for lighting conditions. The accidents on straight roads at hillcrest and those on curved roads combined to an average of roughly 2.5% across all three years.

Figure 7 and Figure 8 are related in that the weather conditions will dictate the conditions of the roadway surface. In the first, it is seen that occurrences during clear skies ranged from 60.05-66.63%, while those under cloudy skies accounted for another 12.06-16.54%. The increase in cloudy sky crashes for year two seems to be offset by a corresponding decrease in clear sky crashes. Between these two conditions, there is a total average of 78.30% across all three years, which is in line with the range of dry road surface accidents of 70.10-74.27%. The range of crashes during the rain of 6.50-10.69%, with an average of 8.81%, is also very similar to that of wet surface conditions, which range from 9.94-15.90%. The higher values in the road surface conditions could be as a result of the roadways remaining wet after inclement weather while the skies are still cloudy. The values for wintery weather and surfaces, as well as not reported conditions, are closer in nature.

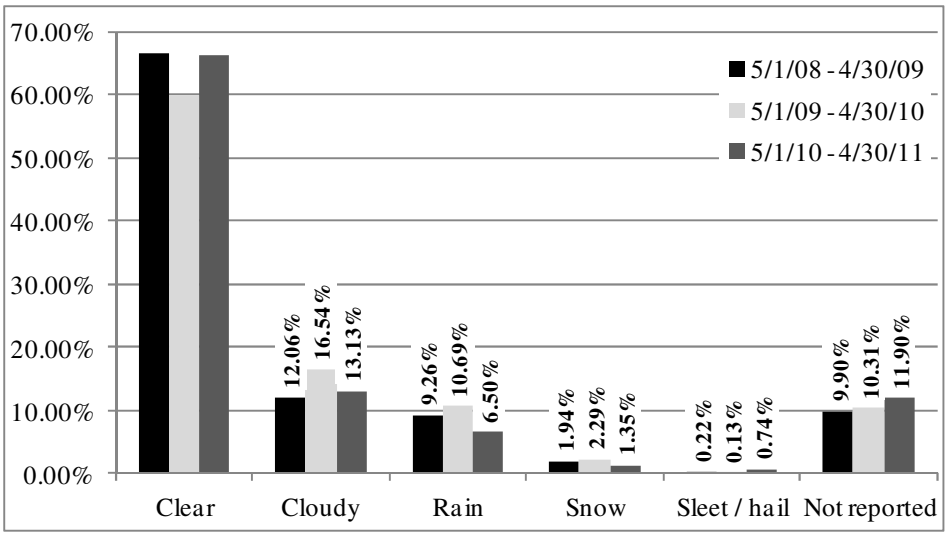

Figure 7: Distribution of annual truck accidents by weather condition

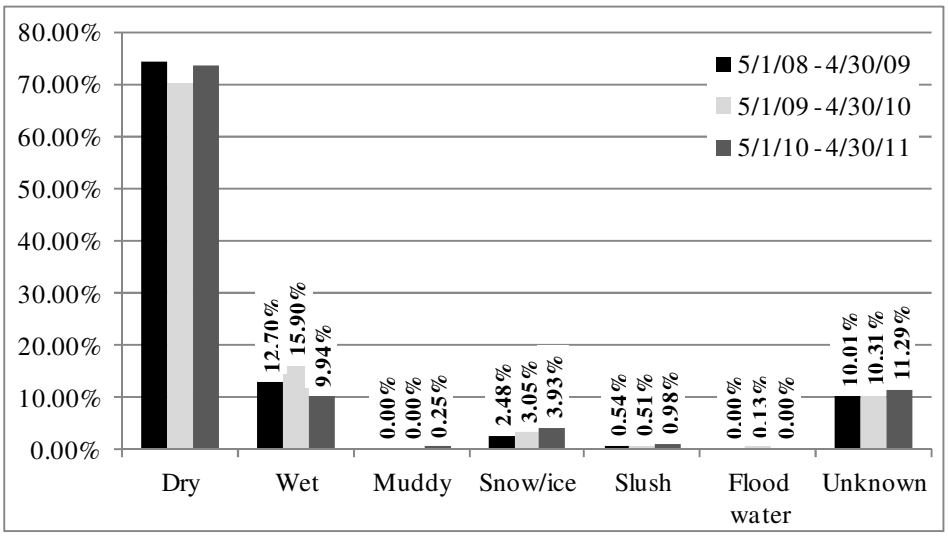

Figure 8: Distribution of annual truck accidents by road surface condition

The accident data were also used in a crash occurrence analysis with the Census tract level as the spatial unit; the GIS file is available for public use by the U.S. Census Bureau, and contains 288 tracts within New York County (Manhattan)—one tract was then removed as it lies entirely in the water and did not have any reported traffic. The map of tracts alone is shown in Figure 9 on the left, with one removal circled, and the BPM network is overlaid on the right in red.

The "Spatial Join" utility in ArcGIS was used to merge the tract data with the geolocated truck accidents, and the output file contained a record for each event with a coding for what tract it occurred within. It must be noted here that six of the original 2,530 records used above fell on highway links outside of the identified tracts and were not able to be merged. The output file was

then collapsed to produce the count of events within each tract both overall and for each BPM block. The summary statistics of these counts are found in Table 2.

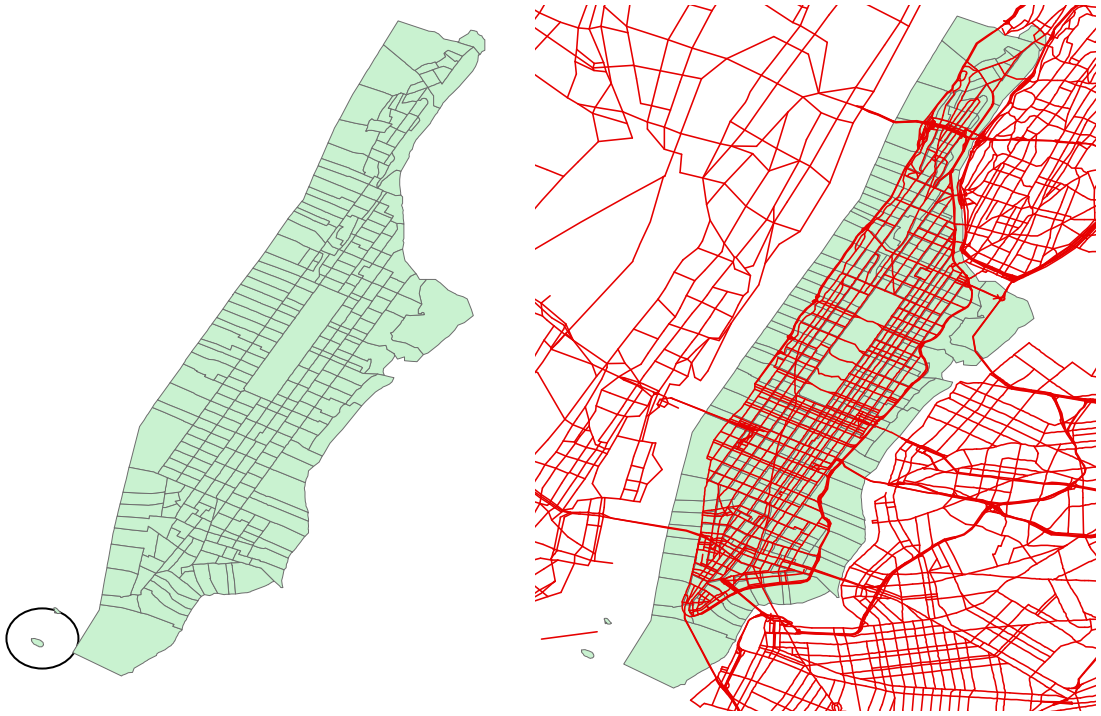

Figure 9: Manhattan Census Tracts and BPM Network

Table 2: Number of truck accidents by BPM time block

Statistic	AM	MD	PM	NT
Mean	1.996516	3.209059	1.787456	1.801394
St Dev	2.253976	3.641955	2.247014	2.302081
Max	14	28	12	19
# of 0s	85	65	94	103
# of 1s	60	45	71	58
# of 2s	52	47	57	49
# of 3s	40	34	23	32
# of 4s	23	26	15	18
# of 5+	27	70	27	27

In Table 2, it is shown that some of the tracts had very high crash occurrences across the three-year time frame; however, full socio-economic Census data was not available for the block level, one geographic unit smaller, so these could not be further broken down. The socio-economic variables used are found in Table 3, and are described in more detail below. From the crash occurrences above, it can be seen that the average number of crashes for the MD block is the highest, even against the two peak periods. The number of zero occurrence tracts is much

lower for this block, while the maximum of 28 is by far the highest, and a substantial amount of the tracts (24.39%) have more than five incidents through the three years. The 2,524 events for this study are broken down into the BPM blocks as 573, 921, 513, and 517, respectively. These values really highlight the similarities seen between AM, PM, and NT, versus the drastic differences with MD—the MD count is 348-408 higher than the other times.

In conjunction with the vehicular flow values from the BPM as described previously, socio-economic variables were used in the occurrence analysis. This information is also publicly available from the U.S. Census Bureau, through the American FactFinder website. For each requested topic, there is either a 2010 Census summary file (SF) or 2010 American Community Survey (ACS) five-year estimate (2006-2010)—the former is a direct observed count, while the latter is an estimate produced from a smaller sampling. Employment values were also provided by the Census through its Longitudinal Employer-Household Dynamics (LEHD) Program. The LEHD Program files are organized by state at the Census block level, and the file used here is the workplace area characteristic data; that is, the information reflects those who actually work in each geographic unit, not those who reside there. These block levels were aggregated into their appropriate tract. The full list of Census variables considered is found in Table 3.

Table 3: Variables used in the occurrence modeling

Variable	Description	Source
pop000	Total population (1000s)	SF
hhsize	Average household size	SF
vac_rate	Vacancy rate of housing units	SF
com_transp	Percentage of commuters who use transit	ACS
com_walkp	Percentage of commuters who walk	ACS
naics_oth000	Total workers in industries listed in Table 4 (1000s)	LEHD
naics44_45_000	Total workers in retail trade (1000s)	LEHD
naics52_000	Total workers in finance and insurance (1000s)	LEHD
naics54_000	Total workers in professional, scientific, technical services (1000s)	LEHD
naics61_000	Total workers in educational services (1000s)	LEHD
naics62_000	Total workers in heath care and social assistance (1000s)	LEHD
naics72_000	Total workers in accommodation and food services (1000s)	LEHD
truck*000	Total hourly truck volume (1000s)	BPM
nontruck*000	Total hourly non-truck volume (1000s) -- taxi, bus, other	BPM

Notes: NAICS means North American Industrial Classification System.
* indicates a time-specific variable, such as "truckam000" for AM block.
SF = Summary File; ACS = American Community Survey
LEHD = Longitudinal Employer-Household Dynamics; BPM = Best Practice Model

The variables in Table 3 were primarily chosen based on *a priori* knowledge of typical factors in crash occurrence models and on their summary statistics. The LEHD Program file includes several additional industries from the North American Industrial Classification System

(NAICS), and the corresponding descriptions for those bundled in "naics_oth000" are in Table 4. The ones isolated were done so based on the highest average percentages and to study different types of activity patterns.

Table 4: Description of NAICS codes

NAICS Code	Description
11	Agriculture, Foresty, Fishing, and Hunting
21	Mining, Quarrying, and Oil and Gas Extraction
22	Utilities
23	Construction
31-33	Manufacturing
42	Wholesale Trade
48-49	Transportation and Warehousing
51	Information
53	Real Estate, and Rental and Leasing
55	Management of Companies and Enterprises
56	Administrative and Support, and Waste Management and Remediation Services
71	Arts, Entertainment, and Recreation
81	Other Services, except Public Administration
92	Public Administration

The summary files and estimates provided by the Census include many more variables that further break down those shown above, such as splitting the population into age groups with a range of five years; however, this fine of an examination was not helpful in the analysis. Some basic statistics of each variable considered were calculated, and these are provided in Table 5. It can be seen there that the tracts have huge variations in values, especially looking at the deviations in commuting percentages, the vacancy rate, and the grouping of NAICS codes from Table 4. In all cases, the true minimum value was observed to be zero, so the non-zero minimum is provided as a reference point.

Table 5: Snapshot of basic statistics for occurrence modeling variables

Variable	Mean	St Dev	Non-zero Min	Max
pop000	5.52568	3.08580	0.002	16.538
hhsize	1.98638	0.50883	1.000	3.430
vac_rate	11.06620	12.14973	0.200	90.000
com_transp	56.03949	17.19759	9.375	91.354
com_walkp	21.39511	14.83754	1.813	100.000
naics_oth000	3.13378	6.84069	0.001	60.830
naics44_45_000	0.45671	0.90500	0.001	9.404
naics52_000	0.96286	2.22549	0.001	17.067
naics54_000	1.02016	2.12013	0.001	13.470
naics61_000	0.35258	0.86691	0.001	12.246
naics62_000	0.69134	1.14153	0.003	7.498
naics72_000	0.54145	0.87496	0.002	4.510

For the percentage of commuters using transit (com_transp), the non-zero minimum is over 9%, and the average is 56.04%. This was expected for Manhattan as public transit is widely available and many residents do not own cars. In can be seen, though, in Figure 10 that even this value of 9.375 is an outlier, as this is less than 3/2 times the lower quartile of roughly 46.5. The least value—where the whisker reaches—is 15.93, and the upper quartile is roughly 68. Despite the maximum value for commuters who walk being 100, greater than transit-users, the average is much lower and there are more outliers, this time on the upper end so they are more than 3/2 times the upper quartile of approximately 28.5. The greatest value shown by the whisker is 55.95, and the lower quartile is around 10. The outlier points are typically in alignment; however, there were two clusters of values that were difficult to read, so the middle point for each was moved to the side for clarity. Although these values are lower than transit, this is still a large portion of the commuting population, only made possible by the density of Manhattan.

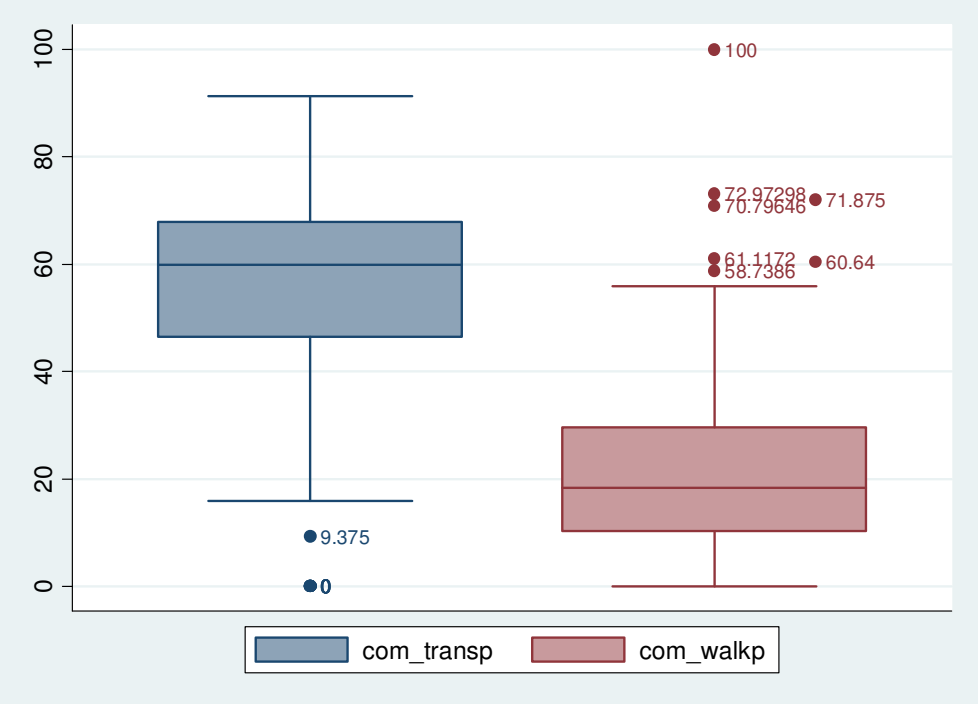

Figure 10: Box plots of transit and walking commuter percentages

Summary statistics were also determined for the vehicle flows in each tract during the respective time blocks, and these were broken down by vehicle classifications provided in the BPM. The designation of "other" includes a summation of single-occupancy vehicles, high-occupancy vehicles, and non-truck commercial vehicles. The values are provided in Table 6 below, and seem to already be showing a correlation between the flow and the crash occurrences;

that is, the total values for AM, MD, and NT are very similar, whereas that of the MD period is noticeably larger. As noted in Table 3, these flows were included in the models as truck and "non-truck" and were divided by 1,000 due to the large values, and the table below reflects this transformation, with the non-truck flows divided into the separate classes.

Table 6: Summary statistics of hourly vehicle flows by BPM time block

		Mean	St Dev	Max			Mean	St Dev	Max
AM	Total	17.2763	16.3372	90.8445	**MD**	Total	21.2924	19.7879	107.7240
	Truck	1.5030	1.4956	8.8481		Truck	1.3228	1.3349	7.8487
	Taxi	1.5489	1.4640	10.7673		Taxi	2.9963	2.8032	17.7887
	Bus	0.3429	0.4084	3.3936		Bus	0.2008	0.2010	1.7909
	Other	13.8816	13.9565	79.1168		Other	16.7726	16.4220	92.0398
PM	Total	18.9758	17.8087	100.2185	**NT**	Total	5.7465	5.8683	37.3684
	Truck	0.5755	0.5984	3.6476		Truck	0.2873	0.3544	2.3424
	Taxi	2.3570	2.1682	13.1823		Taxi	0.5159	0.5036	2.5905
	Bus	0.2823	0.3286	2.4018		Bus	0.0104	0.0115	0.1182
	Other	15.7610	15.4969	92.3808		Other	4.9329	5.2019	33.7402

After the data were cleaned and, in some cases, converted from original codes to the indicators described above, the different datasets were combined into one master for each modeling approach. These files were loaded into Stata to generate the models based on the specifications in the following chapter.

4. MODEL SPECIFICATIONS

The truck accident data were examined more closely to identify factors contributing to the severity of each incident, and separately to understand more about the occurrences per census tract level. As described in the literature review, the approach for the severity was to use an ordered outcome model, while the number of crashes per Census tract was examined using count data models. The datasets described in the chapter above were used as the inputs, and this chapter outlines the specific models.

4.1 Severity Level: Ordered Probit Regression

Each accident was classified according to a discrete numbering system in increasing order of severity, with two of the levels combined as explained in Chapter 3. These values have a natural ordering to them, yet there is no quantitative interpretation; that is, the textual description of each outcome is assigned a value that fits into a ranking system, but the numbers themselves have no actual meaning, and there can be no direct mathematical way to explain the difference, for example, between "1" and "2." Instead, an ordered outcome model—in this case, ordered probit—is used to analyze the outcomes. The specification shown here is based off that used by Kockelman and Kweon [13].

The ordered probit regression is called a "latent variable" model which tries to simplify the complex relationships between the observed variables in the dataset and a latent structure that is unobservable. The variables outlined in Chapter 3 are the manifest variables, which are the x_is, and the modeling generates latent variables. The manifest variables are assumed to be conditionally independent given the latent variables. If there are random variables x_i and x_j, and then a latent variable y_i, then the two xs are independent in their conditional probability distribution given y_i. To put it a different way, given y_i, the probability distribution of x_i is the same for all values of x_j, and vice versa.

The underlying relationship, or latent index, is a y^* that is the exact, unobserved dependent variable, and the function for this is comprised of both observed and unobserved variables. This is in contrast to y itself, which is the assigned outcome by the pre-determined categories. To clarify, in this analysis the focus is on the severity, which can be one of four very broad outcomes; however, the actual severity, particularly when looking at personal injury or property damage, is on a sliding scale. The vehicle in one accident may have been totaled, while in another maybe only the bumper needed to be replaced. These would both be classified as property damage without any further consideration into the actual extent of this damage, which can be more accurately captured by the latent index.

The specification for the latent index is given in Equation 1, and expanded in Equation 2.

$$y_i^* = \boldsymbol{x_i\beta} + \varepsilon_i \quad (1)$$

$$y_i^* = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 \ldots x_{ni}\beta_n + \varepsilon_i \quad (2)$$

where: y_i^* = latent and continuous measure of crash severity for event i,

$\mathbf{x_i}$ = the vector of independent explanatory variables, for index 1 to n,

$\boldsymbol{\beta}$ = the vector of regression coefficients to be estimated, with same index of 1 to n,

ε_i = the random error term, following a standard Normal distribution.

The actual observed severity level, coded as 1-4, is found from the model through Equation 3.

$$y_i = \begin{cases} 1\ if\ -\infty \le y_i^* \le \mu_1 \\ 2\ if\ \mu_1 < y_i^* \le \mu_2 \\ 3\ if\ \mu_2 < y_i^* < \mu_3 \\ 4\ if\ \mu_3 < y_i^* \le \infty \end{cases} \quad (3)$$

where the μ_ks are cutoff thresholds that also need to be estimated. These thresholds are presumably different for each observation, and then the model's output represents the average. There are four severity levels that require three cutoffs, and the value of the latent index will fall within the given ranges.

There are also probabilities for each of the severity levels that show how likely it is that a given observation will have that specific ordered outcome by utilizing the cutoff values described in Equation 3. These are given by Equations 4a-4d, where Φ( ) is the standard normal cumulative distribution function, and each $\Pr(y_i = j)$ is the probability that the occurrence *i* falls within the indicated response *j*.

$$\Pr(y_i = 1) = \Pr(y_i^* \le \mu_1) = \Pr(x_i\beta + \varepsilon_i \le \mu_1) = \Pr(\varepsilon_i \le \mu_1 - x_i\beta) = \Phi[\mu_1 - x_i\beta] = 1 - \Phi[x_i\beta - \mu_1] \quad (4a)$$

$$\Pr(y_i = 2) = \Pr(\mu_1 < y_i^* \le \mu_2) = \Pr(y_i^* \le \mu_2) - \Pr(y_i^* \le \mu_1) = \Pr(x_i\beta + \varepsilon_i \le \mu_2) - \Pr(x_i\beta + \varepsilon_i \le \mu_1) = \Pr(\varepsilon_i \le \mu_2 - x_i\beta) - \Pr(\varepsilon_i \le \mu_1 - x_i\beta) = \Phi[\mu_2 - x_i\beta] - \Phi[\mu_1 - x_i\beta] = 1 - \Phi[x_i\beta - \mu_2] - (1 - \Phi[x_i\beta - \mu_1]) = \Phi[x_i\beta - \mu_1] - \Phi[x_i\beta - \mu_2] \quad (4b)$$

$$\Pr(y_i = 3) = \Pr(\mu_2 < y_i^* \le \mu_3) = \Pr(y_i^* \le \mu_3) - \Pr(y_i^* \le \mu_2) = \Pr(x_i\beta + \varepsilon_i \le \mu_3) - \Pr(x_i\beta + \varepsilon_i \le \mu_2) = \Pr(\varepsilon_i \le \mu_3 - x_i\beta) - \Pr(\varepsilon_i \le \mu_2 - x_i\beta) = \Phi[\mu_3 - x_i\beta] - \Phi[\mu_2 - x_i\beta] = 1 - \Phi[x_i\beta - \mu_3] - (1 - \Phi[x_i\beta - \mu_2]) = \Phi[x_i\beta - \mu_2] - \Phi[x_i\beta - \mu_3] \quad (4c)$$

$$\Pr(y_i = 4) = \Pr(y_i^* > \mu_3) = \Pr(x_i\beta + \varepsilon_i > \mu_3) = \Pr(\varepsilon_i > \mu_3 - x_i\beta) = 1 - \Phi[\mu_3 - x_i\beta] = \Phi[x_i\beta - \mu_3] \quad (4d)$$

4.2 Occurrences: Zero-Inflated Negative Binomial Regression

In contrast to the severity analysis that used an ordered outcome model, the number of truck crashes per Census tract requires the use of a count data model. In this case, the numbers do have actual meaning, and there is a mathematical relationship between them; that is, a value of "1" actually means that one accident was reported, while a value of "2" indicates that two were observed, and the latter is exactly one more incident than the former.

As highlighted in the literature review, there are two basic modeling approaches that are used for count data: (1) Poisson regression, and (2) negative binomial regression. The Poisson regression is very widely used for many data applications including crash occurrences, and the results are relatively straightforward to interpret. The one main assumption behind that specification, however, is that the data are not overdispersed. This means that the mean and variance of the response variable must be roughly the same, and it can be seen in Table 2 that this is not the case for any of the BPM time blocks. The table lists the mean and the standard deviation, which is the square root of the variance and which is always greater than its respective mean. Looking at the AM block as an example, the variance would be 5.076545 and this is not close to the mean value of 1.989583. Using this easy test, it can be seen that the negative binomial regression may be more appropriate for the available data.

The model specification for the negative binomial regression is very similar to a standard Poisson, so it helps to provide them both to highlight the differences. The Poisson model is shown in Equation 5.

$$\Pr(Y = y_i) = \frac{EXP^{-\lambda_i}\lambda_i^{y_i}}{y_i!} \tag{5}$$

where y_i is the outcome count for tract i, and λ_i is the mean number of occurrences. In this format, Poisson is referred to as the log-linear model, and using a basic linear form such as Equation 1, it follows that $E[y_i] = \lambda_i = EXP(\beta X_i)$. Subsequently, $\ln(\lambda_i)$ is then $\beta' X_i$. The exponent constrains the model forecasts to be positive, so there cannot be a negative number of crashes.

The $E[y] = EXP(\beta \boldsymbol{X})$ relationship can be used to find the likelihood function and then the log likelihood, which are shown in Equations 6a and 6b, respectively.

$$L(\beta) = \prod_i \frac{EXP[-EXP(\beta X_i)][EXP(\beta X_i]^{y_i}}{y_i!} \tag{6a}$$

$$LL(\beta) = \sum_{i=1}^{n}[-EXP(\beta X_i) + y_i \beta X_i - LN(y_i!)] \tag{6b}$$

In the case of the negative binomial, $\lambda_i = EXP(\beta' x_i + \varepsilon_i)$ where the $EXP^{\varepsilon i}$ errors are Gamma distributed with a mean of 1 and variance α. This additional parameter yields Equation 7.

$$VAR(y_i) = E[y_i]\{1 + \alpha E(Y_i)\} \tag{7}$$

Starting first with the specification of Poisson allows one to see that the negative binomial model collapses into a standard Poisson when the α parameter is zero.

The other item to look out for in the data is a high number of zero counts, and again looking at Table 2, there are a lot of Census tracts that did not report a single accident throughout the three years. In order of the table, there were 85 (29.6%), 65 (22.6%), 94 (32.8%), and 103

(35.9%) tracts with zero counts. This indicates that a zero-inflated model should be tested, as this is based on a probability distribution that allows for a high number of zero-occurrence observations. The process then allows for estimating a separate model for the certain zero group and another based on the actual count values—by "certain zero," it is meant that these observations will always be zeros based on some set of predictors, whereas other observations may have just happened to have been zero for the given time frame, as dictated by the predictors used on the 1, 2, 3, etc. occurrence tracts, but could have been greater than zero.

The ZINB model, showing both the certain zero group probability and the probability for the observed count group, is shown in Equation 8.

$$\begin{matrix} y_i=0\ with\ probability & p_i+(1-p_i)\left[\frac{\frac{1}{\alpha}}{\left(\frac{1}{\alpha}\right)+\lambda_i}\right]^{\frac{1}{\alpha}} \\ y_i=y\ with\ probability & (1-p_i)\left[\frac{\Gamma\left(\left(\frac{1}{\alpha}\right)+y\right)u_i^{\frac{1}{\alpha}}(1-u_i)^y}{\Gamma\left(\frac{1}{\alpha}\right)y!}\right], y=1,2,3\ldots \end{matrix} \tag{8}$$

The Stata software package generates results for the two separate processes under one combined model, and allows for testing whether or not the zero-inflated negative binomial (ZINB) is a better fit than a standard negative binomial model. This is done through the Vuong test, and if the calculated value is significant, then ZINB is the appropriate specification. The statistic for each observation *i* with a non-nested model, such as the negative binomial, is given by Equation 9a, and then the actual Vuong statistic is in Equation 9b—if this value is positive and greater than the critical value, such as 1.96 for a 95% confidence interval, then model 1 is favored over model 2.

$$m_i = LN\left(\frac{f_1(y_i|X_i)}{f_2(y_i|X_i)}\right) \tag{9a}$$

$$V = \frac{\sqrt{n}\left[(1/n)\sum_{i=1}^{n} m_i\right]}{\sqrt{(1/n)\sum_{i=1}^{n}(m_i-\bar{m})^2}} = \frac{\sqrt{n}(\bar{m})}{S_m} \approx Z_\alpha \tag{9b}$$

where m_bar is the average value of Equation 9a, S_m is the standard deviation, and Z_α is the value to reference against the standard normal tables.

The elasticities from the negative binomial model of the count for individual *i* and continuous variable *k* are calculated in the same manner as the Poisson model, with lambda having the definition used in Equation 7. They are found through Equation 11, and can be collapsed into simply the product of the coefficient estimate with the mean value for each respective variable due to the semi-log nature of the function estimated by Stata; this is Equation 10 [43].

$$\log(y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \tag{10}$$

$$E_{x_{ik}}^{\lambda_i} = \frac{\partial \lambda_i}{\lambda_i} \times \frac{x_{ik}}{\partial x_{ik}} = \beta_k x_{ik} \tag{11}$$

4.2.1 Binomial Logit

In order to determine which variables to include in the inflated model for the zero-inflated binomial regression, a standard binomial logit model can be applied to the variables being considered. This is done by creating a new variable for each time block that has a value of "0" for those that are zero in the original occurrences, and an indicator of "1" for at least one occurrence within that tract. In this particular case, all variables in Table 3 were examined first, and those with a threshold of over 1.00 were further considered in a reduced set to see if their statistics improved without the inclusion of insignificant variables. When deciding which of these to include in the inflated models, a significance of greater than 1.64, or a 90% confidence interval, was required.

In discrete choice modeling, there are special cases when the choice set only contains two alternatives, such as the presence or absence of truck accidents within a Census tract as used here. Other examples could be the choice between driving or taking transit, or choosing between two routes where one has a toll and the other does not. The probabilities associated with each observation choosing one alternative over the other are based on the concept of random utility, and that the outcome observed has a higher utility than the other. This utility function stems from consumer preferences, where each consumer tries to maximize utility; it is considered random because there are unobservable attributes in the analysis, and there may be proxy variables used instead. For the case of choosing between modes, the utility function may include attributes like travel time, travel cost, and availability—these are observable, and are said to be the systematic components, and there are also random disturbances. Putting this explanation into mathematical terms yields Equation 12.

$$\begin{aligned} U_{in} &= V_{in} + \varepsilon_{in} \\ U_{jn} &= V_{jn} + \varepsilon_{jn} \end{aligned} \tag{12}$$

where V_{in} and V_{jn} are the systematic components and ε_{in} and ε_{jn} are the disturbances for the utility of choices i and j. The probability of choose i over j is then given by Equation 13.

$$P_n(i) = Pr(U_{in} \geq U_{in}) = Pr\left(V_{in} + \varepsilon_{in} \geq V_{jn} + \varepsilon_{jn}\right) = Pr(\varepsilon_{jn} - \varepsilon_{in} \leq V_{in} - V_{jn}) \tag{13}$$

The binary logit model has an underlying assumption that the overall error, or disturbance, for an individual observation (ε_n), which is the difference of ε_{jn} and ε_{in}, is logistically distributed.

With μ as its positive scale parameter, the cumulative distribution function and probability density function of ε_n are given in Equation 14a and 14b, respectively.

$$F(\varepsilon_n) = \frac{1}{1+e^{-\mu\varepsilon_n}}, \quad \mu > 0, -\infty < \varepsilon_n < \infty \tag{14a}$$

$$f(\varepsilon_n) = \frac{\mu e^{-\mu\varepsilon_n}}{(1+e^{-\mu\varepsilon_n})^2} \tag{14b}$$

This means that the probability provided in Equation 15 now becomes [44]:

$$P_n(i) = Pr\left(U_{in} \geq U_{jn}\right) = \frac{1}{1+e^{-\mu\left(V_{in}-V_{jn}\right)}} = \frac{e^{\mu V_{in}}}{e^{\mu V_{in}}+e^{\mu V_{jn}}} \tag{15}$$

The analysis of the systematic components in Stata is similar to a linear regression, but the β coefficients are in log-odds units, not a direct one-to-one relationship. The prediction equation can be written as shown in Equation 16, and the coefficients values can be more easily interpreted by converting them to standard odds ratios by exponentiating the coefficient, for all coefficients 1 to k. This process was not necessary for this analysis as only the significance of a variable was taken into consideration.

$$\log\left(\frac{P_n(i)}{1-P_n(i)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \tag{16}$$

Full results from the binary logit step are not provided; however, the variables included in the models in Chapter 6 are those that were shown to be significant from this process.

5. MODELING RESULTS: ACCIDENT SEVERITY

As described in Chapter 4.1, factors contributing to the accident severity were examined using an ordered probit model. The variables listed in Table 1 were considered in the process, and those with significance level of 1.64 or greater were retained. Several variations of the weather, lighting, and road surface condition were changed in and out to determine which were the most influential. Time of day was also a primary interest, and some of the hours proved to be significant. The final model is in Table 7.

Since this ordered model assumes that a higher value corresponds to a more severe accident, the sign of the coefficients can be interpreted in a relatively straightforward manner. If the sign of a β is negative, then this means that the crash severity decreases as the related variable increases, while a positive sign indicates that the severity increases as that variable increases. The positive or negative contribution from these different variables is what leads to the value of the latent index and subsequently places the observation in one of the preset categories. The actual values, however, are more difficult to discuss as the increase in probability attributed to a one-unit increase of a given predictor depends on both its starting value and the given values of the other predictors. In the case of ordered logit, the coefficient values are the ordered log-odds of being in a higher category while the other variables are held constant, and that approach cannot be used here with the ordered probit.

Table 7: Final model for severity analysis

Variable	Coefficient	t Statistic	dydx1	dydx2	dydx3	dydx4
num_of_veh	-0.36086	-8.46	0.00395	0.13971	-0.13725	-0.00641
flow000	-0.08364	-1.68	0.00092	0.03238	-0.03181	-0.00149
dark_lit	0.27760	4.14	-0.00241	-0.10794	0.10407	0.00628
hour6	0.44542	2.83	-0.00287	-0.17151	0.16081	0.01357
hours14_15	0.15248	2.10	-0.00144	-0.05933	0.05765	0.00312
hour18	-0.31599	-2.80	0.00512	0.11778	-0.11891	-0.00400
surf_dry	0.20271	2.79	-0.00183	-0.07890	0.07636	0.00437
cut1	-3.40467		Log Likelihood			-1863.68
cut2	-0.65793		Likelihood Ratio			121.26
cut3	1.77203		McFadden's Pseudo R-squared			0.0315

* marks an indicator variable, where dy/dx will be the discrete change from 0 to 1
dydx1 refers to margins evaluated at outcome 1, and so on.

It can be seen in Table 7 that coefficients are all negative for the variables num_of_veh (-0.36086), flow000 (-0.08364), and hour18 (-0.31599). This means that an increase in the number of vehicles involved in the accident will lead to a lower severity level, as will an increase in the observed vehicle flow—the latter is most likely related to the speeds which can be obtained by

vehicles when the flows are higher. The variable hour18 is an indicator variable, rather than a continuous volume or discrete vehicle count, so this simply implies that accidents that occurred within the hour from 5-6 pm have decreased severity. This does not mean that this time of the day provides a different risk of incidents occurring or having a lower severity overall, but rather that those that have already occurred tended to be less severe.

Conversely, the coefficients for dark_lit (0.27760), hour6 (0.44542), hours14_15 (0.15248), and surf_dry (0.20271) were all positive. Since these are all indicator variables, this means that the presence (a value of "1") of these respective features increased the likelihood of more severe crashes. To elaborate, crashes that occurred when there was only artificial street lighting and not natural lighting, such as during the evening or overnight hours, were associated with a more severe classification. The variables hour6 and hours14_15 correspond to 5-6 am and 1-3 pm, respectively, and accidents occurring during these times were also more severe—it should be noted that 5-6 am, during some times of the year, is before the sun rises, so there may be some crossover with the dark_lit predictor. Finally, incidents when the road surface was dry also increased the likelihood of a more severe outcome, which, similar to the discussion on flow000, could indicate the impacts of speed; in this case, better surface means a higher speed.

The cutpoints, as described in Equation 3, help to differentiate between the response levels, and consequently there are three cuts for the four outcomes. The first, cut1, distinguishes a non-reportable severity from the rest when the predictors are evaluated at zero. This means that observations with a value of -3.404668 or lower for the latent variable would have a non-reportable severity when both the number of vehicles involved and the traffic flow were zero, in addition to all of the indicators being zero—that is, not a dark, lit road, not during 5-6 AM, not during 1-3 PM, not during 5-6 PM, and not when the road surface is dry. Similarly, the second threshold differentiates property damage only from a non-reportable accident and one with injury and property damage. Using the same reasoning of the predictors being evaluated at zero, the observation would fall into the property damage only category for a latent variable value between -3.404668 and -0.6579264; a value between -0.6579264 and 1.772031 would indicate injury and property damage only. When all predictors are evaluated at zero and the latent variable value is above 1.772031, the accident is classified as fatal.

The log likelihood (LL) of the model was found to be -1863.68, and this value is used in the likelihood ratio (LR) Chi-squared test. This test determines if at least one of the predictors' regression coefficients is statistically not equal to zero within the model by coming the likelihood of the null model, with just the response variable, with the fitted model shown. The LR of 121.26 had a p-value of 0.0000, so for an alpha level of 0.1, or even at 0.05, the null hypothesis is

rejected—in this case, that hypothesis is that all of the regression coefficients in the model are zero. While this leads to the conclusion that at least one of the coefficients is not zero, it does not indicate that all of the predictors in the model statistically have non-zero coefficients. The final statistic given is McFadden's pseudo R-squared, which has a value of 0.0315; however, it is difficult to interpret this statistic for the non-linear regression.

The marginal effects of each predictor were also calculated, and are shown in Table 7. These are statistics that are calculated within Stata using predictions of the fit model for fixed values of the variables included, and each individual effect assumes that the others remain unchanged. In this case, there are four different outcomes, so the values provided show a change in probability of being in that category. For example, the discrete change from the accident not occurring from 5:00-6:00 AM to occurring then has positive values on the higher severity levels but negative on the two lower, so these are more likely to be more severe, which is consistent with the discussion.

In all instances, variables with positive coefficient values also have positive margins for the two higher severity levels, and negative for the two lower, while the probability of being property damage and/or personal injury is greater than a fatality. In the case of hour6, the probability of it being a fatality is greater than the other variable margins, yet not greater than its immediate lower severity level. This is due to the fact that very few crashes fell into the two extreme categories—fatality versus non-reportable damages—and the split between the two middle categories was relatively even, as shown in Chapter 3.

6. MODELING RESULTS: CRASH OCCURRENCES

While the previous chapter highlights the modeling efforts on the accident severity level, this chapter outlines the results of the crash occurrences using count data models. The following analysis was the primary focus of this overall research effort, and much more attention was paid to highlighting the temporal differences across the four BPM time blocks, leading to separate models for each block.

Both the standard and non-inflated portion of the zero-inflated negative binomial (ZINB) models are modeling the expected count of truck accidents within a Census tract as a function of the predictor variables. The coefficients in both cases can then be interpreted in the same manner. Given that the other variables in the model are held constant, the difference in the logs of expected accidents is expected to change by the respective coefficient for a one unit change in the predictor variable. Putting this more simply, the number of expected accidents changes by exp(coefficient) for a one unit increase in the corresponding predictor variable, with the others held constant—this is an important distinction to make, and will assumed to be the case during all of the results discussion below, without the need to repeat it for every single coefficient.

For the inflated portion of the ZINB model, the coefficients have a different interpretation than explained above. In that case, a one unit increase in the predictor variable would increase the odds that a Census tract were in the "certain zero" group by a factor of exp(coefficient). As described in Chapter 4, this is a grouping of tracts that will always experience zero accidents, rather than those that may just have had zero occurrences for the time period in the study.

For both models, the LN alpha, alpha, and likelihood ratio test statistic (LR chi2) are provided. LN alpha is merely the estimate of the log of the dispersion parameter, and as discussed in Chapter 4, these models collapse to standard Poisson when the dispersion parameter is zero. The LR chi2 value is a test statistic that indicates whether or not all regression coefficients in the model are simultaneously zero, so a value over the critical one for the respective degrees of freedom means that at least one of the coefficients is not zero, so the fitted model is better than the constant-only model. For regular negative binomial, there is also McFadden's pseudo R-squared value, and as mentioned in Chapter 5, this is not equivalent to the R-squared used in linear regression and is difficult to interpret. The ZINB model results include the Vuong statistic outlined in Chapter 4, which helps choose between this and a standard negative binomial.

6.1 AM Models

The first time block being examined is the AM, which includes the morning peaks and covers 6:00-10:00 AM. The regular and zero-inflated negative binomials models are shown in Table 8.

6.1.1 Negative Binomial Results

It can be seen in the first half of Table 8 that six of the 14 variables had test statistics over the 1.64, 90% confidence interval, threshold value. Those making the cut are pop000, vac_rate, naics_oth000, naics44_45_000, truckam000, and nontruckam000. All had positive coefficients meaning that a unitary increase in each individually would lead to an increase in the number of expected accidents. In the order listed above, the respective increases are: 1.077, 1.013, 1.021, 1.127, 1.206, and 1.008. All of these values are around 1.0, with truckam000 having the highest increase of 1.206 accidents.

Given the data transformations described in Chapter 3, this means that, while holding the other variables constant, an increase of 1,000—the unit here—in population, workers in the combined NAICS group, workers in retail trade, hourly flow of trucks, and hourly flow of non-trucks would all increase the expected crashes by about one. The vacancy rate is the only variable here in a different base unit, so an increase of 1.0 in the percent of vacant housing units would lead to an increase of 1.013 crashes; a positive sign here was unexpected as an increase in residents should bring other changes such as more vehicle flow. The NAICS group of retail trade being significant stands out as this is one of the target industries of the off-hour delivery (OHD) program; it has been shown that this group accounts for 5.03% of the daily truck trips in Manhattan [45]. It is slightly harder to interpret the meaning of the combined group as there are several very different industrial sectors included, so the increase of 1,000 workers could come from any number of combinations of increases within the individual sectors.

The variables that did not have test statistics above the threshold were the household size, two commuting groups, and the remaining individual industrial sectors; however, it is still worthwhile to leave these in the full model as it was expected that they may have an impact, so the fact that they did not still has some indication. Among these, the household size, percentage of commuters using transit, workers in finance and insurance, workers in heath care and social assistance, and workers in accommodation and food services all had negative coefficients, so these would lead to a decrease in expected number of crashes—the latter variable is interesting because it is also a focus of the OHD program, with another 24.2% of the daily truck trips [45]. The use of transit may be represented more clearly in the model through the traffic flow variables, and the other two NAICS groups could be among the transit users. Workers in finance and insurance typically work downtown which is difficult to access via automobile, and workers in health care may report to hospitals where parking is limited for employees.

Looking at the model statistics, it can be seen that the dispersion parameter is 0.281, which is not equal to zero, so this confirms that a NB model is preferred over Poisson; it was also

discussed in Chapter 4 how the summary statistics for the number of crashes in each time block violated the assumption behind Poisson that the mean and variation are the same. The likelihood ratio statistic for the 14 degrees of freedom is 104.26, and it had a p-value of 0.000, meaning that the null hypothesis of all coefficients being simultaneously equal to zero is rejected. As cautioned in Chapter 5, though, this does not necessarily mean that all coefficients are not equal to zero, but rather than at least one is not. The pseudo R-squared value of 0.0956 is quite low—the lowest of all four time blocks—which could be related to the fact that the Vuong statistic was the highest, as discussed in the following section.

Table 8: Crash occurrence models for AM period

	Standard Neg Binomial		*Zero-Inflated Neg. Bin.*	
Variable	**Coefficient**	**t Statistic**	**Coefficient**	**t Statistic**
Constant	-0.2556332	-0.64	-0.2236767	-0.58
pop000	0.0745027	3.60	0.0506702	2.46
hhsize	-0.1897902	-1.34	-0.0658835	-0.47
vac_rate	0.0131426	3.21	0.0125354	3.52
com_transp	-0.0007612	-0.15	0.002946	0.64
com_walkp	0.0057837	1.16	0.0052343	1.16
naics_oth000	0.0203849	2.26	0.0205217	2.78
naics44_45_000	0.1198905	1.76	0.0899947	1.49
naics52_000	-0.0697304	-0.60	-0.0397417	-0.36
naics54_000	0.0662497	0.45	0.0897104	0.69
naics61_000	0.1007127	0.96	0.0955767	1.02
naics62_000	-0.0300711	-0.23	-0.461647	-0.40
naics72_000	-0.0660174	-0.28	-0.2062285	-0.97
truckam000	0.1871808	3.81	0.1777827	3.90
nontruckam000	0.0084227	1.85	0.0038481	0.94
Inflated				
Constant	--	--	-1.107785	-0.80
com_transp	--	--	0.0237998	1.25
naics52_000	--	--	3.688746	2.46
naics72_000	--	--	-13.77024	-2.43
truckam000	--	--	0.2471742	0.51
nontruckam000	--	--	-0.0708537	-1.80
ln alpha	-1.270992	--	-2.014428	-5.01
alpha	0.2805531	--	0.1333967	--
LR chi2	104.26	--	63.92	--
Pseudo R2	0.0956	--	--	--
Vuong	--	--	2.57	--

6.1.2 Zero-Inflated Negative Binomial Results

As mentioned above and shown in Table 8, the Vuong statistic for the AM ZINB model is 2.57, which is the highest of all time blocks, and is greater than 1.96. This seems to indicate that the inflated approach may be best suited for the data in this time block, for the predictors selected. Although the alpha value is smaller at only 0.133, it is still above zero so a Poisson model was

not fit. The LR chi2 value is also much smaller than in the previous model, though still above the critical value for the degrees of freedom, so the null hypothesis is again rejected.

The results in Table 8 show that not all of the variables significant in the regular negative binomial are also significant for the ZINB model, and the constant is again insignificant. The workers in retail trade and the non-truck traffic fell below 1.64 for the non-certain zero observations. The remaining variables and their respective expected increases in the number of crashes are pop000 (1.052), vac_rate (1.013), naics_oth000 (1.021), and truckam000 (1.195). Again, these values are all around 1.0, with the highest being the truck flows. The means that an increase of 1,000 in the population, combined NAICS group, and hourly truck flow for this block would increase crashes, as would a 1.0 percent increase in the vacancy rate of housing units.

Looking now at the variables that were not significant, the same set as the regular negative binomial appears, in addition to those mentioned above—it is particularly interesting that non-truck traffic no longer plays an influential role in predicting the number of truck crashes. Here, the variables with negative coefficients that would decrease the expected number of crashes are household size and the three worker groups mentioned in the previous section, but not the transit commuting group. Both these and those who walk to work are reporting an expected increase in the number of crashes; an increase in commuters using other modes would presumably decrease the non-truck traffic, which is also not significant, so there is not much to be interpreted there.

The binary logit models led to the inclusion of com_transp, naics52_000, naics72_000, truckam000, and nontruckam000 in the inflated model for the certain zero group, and only half of these are shown to be significant here. Workers in finance and insurance had a large positive coefficient, meaning that the odds of being in the certain zero group increase by a factor of 39.995 for a 1,000 increase, holding the other predictors constant. Both workers in accommodation and food services (1.046E-06) and non-truck flows (0.932) had negative coefficients so an increase of 1,000 for either individually would lead to a decrease in odds by the factors shown—the value for food service workers is essentially zero.

6.2 MD Models

The second time block from the BPM has midday off-peak hours of 10:00 AM-3:00 PM, and as mentioned in Chapter 3, these hours had significantly more accidents than the other blocks. The final models for MD are in Table 9 below.

6.2.1 Negative Binomial Results

The midday models were of particular interest leading into the analysis due to the high number of crashes reported. The results for the regular negative binomial, though, are not all that

different from those outlined for the AM period. The variables that were significant are pop000, vac_rate, com_walkp, naics_oth000, and truckmd000. Again, these all had positive coefficients, so the expected number of truck accidents would increase with a 1,000 increase in population, combined NAICS workers, and truck flows, and with a 1.0% increase in the housing unit vacancy rate and commuters who walk to work. The exact increases, in the order shown, are 1.075, 1.016, 1.008, 1.017, and 1.228—the largest value once again is related to the hourly truck traffic. As discussed, a positive correlation for the vacancy rate was not expected, and neither was one for the number of commuters who walk; however, for the latter, it could be that pedestrian flows are an underlying factor that leads to truck accidents.

Of those variables that were not significant, workers in retail trade, workers in professional, scientific, and technical services, workers in educational services, workers in accommodation and food services, and non-truck traffic all had positive coefficients, so these would all also lead to increases in the expected crash counts. Again, food services are a group on which to focus, and although not significant, the increase rather than decrease for this time period could be related to the time at which these establishments are busiest, not only in terms of customer numbers but also when deliveries are made, so the trucks are in these neighborhoods. The portion of commuters using transit is negative here, which was expected, yet interestingly the non-truck flows were not significant as pointed out. The workers in finance and insurance and in health care and social assistance had negative coefficients, consistent with the AM period, so the discussion is applicable here, as well.

The statistics for the model show a dispersion parameter of 0.288, which is not equal to zero, so the negative binomial model is preferred over Poisson. This time period actually had the highest discrepancy between the mean crashes (3.209) and the variance (13.264, the square of the standard deviation of 3.642), almost three times the other blocks. The LR chi2 value of 160.55 is the highest among all time blocks, and is considerably above the critical value for 14 degrees of freedom, so it can be stated that at least one of the coefficients in the model is not equal to zero. The pseudo R-squared value of 0.1213 is still quite low, but again, this is difficult to interpret.

6.2.2 Zero-Inflated Negative Binomial Results

The binary logit models led to the inclusion of the percentage of commuters using transit, workers in educational services, and the hourly truck flows in the inflated portion of the ZINB model. It can be seen in Table 9 that only the transit commuters were significant, and with a negative coefficient. Therefore, the odds of being in the certain zero group would decrease by a factor of 0.771 for every 1.0% increase in such commuters. Workers in educational services also

had a negative coefficients so the odds would be decreased for an increase this predictor, while the truck flows have a positive relationship.

Table 9: Crash occurrence models for MD period

	Standard Neg Binomial		*Zero-Inflated Neg. Bin.*	
Variable	**Coefficient**	**t Statistic**	**Coefficient**	**t Statistic**
Constant	0.1152796	0.30	0.9691993	2.08
pop000	0.0723912	3.94	0.0450148	2.49
hhsize	-0.114515	-0.89	-0.1464449	-1.11
vac_rate	0.0157091	4.19	0.0126044	3.65
com_transp	-0.0048275	-1.07	-0.0089714	-1.87
com_walkp	0.0083077	1.82	0.0032086	0.67
naics_oth000	0.0171397	1.96	0.0173511	2.25
naics44_45_000	0.0923471	1.53	0.0427444	0.78
naics52_000	-0.1465899	-1.41	-0.1436053	-1.52
naics54_000	0.1347374	1.03	0.1322945	1.12
naics61_000	0.0669103	0.63	0.0357787	0.36
naics62_000	-0.1433741	-1.17	-0.1431655	-1.27
naics72_000	0.2134537	0.98	0.1775824	0.91
truckmd000	0.2532498	5.22	0.2114716	4.52
nontruckmd000	0.0042066	1.27	0.003265	1.02
Inflated				
Constant	--	--	1.321769	-1.29
com_transp	--	--	-0.2594797	-1.83
naics61_000	--	--	-16.34198	-1.08
truckmd000	--	--	-0.8814157	1.48
ln alpha	-1.246387	--	-1.606947	-6.05
alpha	0.2875418	--	0.2004988	--
LR chi2	160.55	--	130.80	--
Pseudo R2	0.1213	--	--	--
Vuong	--	--	2.38	--

Within the non-certain zero model, the same variables as the regular negative binomial were significant, with the exception of the percentage of commuters who walk. This was instead replaced by the percentage of commuters using transit, and this was the only negative coefficient in this set, leading to a decrease of 0.991 expected crashes for every 1.0% increase, holding the others constant. Those variables leading to an increase, with their respective increases, are: pop000 (1.046), vac_rate (1.013), naics_oth000 (1.017), and truckmd000 (1.235). These values are all consistent with those in the regular negative binomial of being around 1.0 and truck traffic having the highest influence on its own. As has been stated, the unit increase for population, workers, and flows is 1,000, versus 1.0% for the vacancy rate and commuter proportions. In this case, the model's constant is significant, which gives the predicted number of crashes (2.636) when all of the predictors in the model are evaluated at zero. This would mean, most

significantly, that there is no population and no vehicle flows, which is outside the possibilities here, especially as a lack of vehicles should lead to a lack of vehicular accidents.

Despite there only being one significant variable in the inflated model, the Vuong statistic of 2.38 indicates that the ZINB model is preferred over the standard negative binomial, which was not expected since this time block has the fewest tracts with zero occurrences. The LR Chi-squared value of 130.80 is almost as high as that observed for the standard model, and again means that at least one of the predictor coefficients in the model is not equal to zero, so the constant-only model is rejected. The dispersion parameter alpha is 0.200, so a bit lower than the regular model, but still higher than zero.

6.3 PM Models

The next model covers the evening peak hours, and is for the PM time block from 3:00-7:00 PM. The results for these two models are given in Table 10.

6.3.1 Negative Binomial Results

The results for the regular negative binomial for the PM period are very similar to those in the AM period, which was expected due to both encompassing peak travel hours for a standard work day. In fact, the same six variables are significant, and the difference here is that the model's constant is also significant. The constant value shows that there would be an expected 0.469 accidents given that all model predictors are held at zero, which cannot be the case as was discussed in the previous section, as without vehicle flows, particularly trucks, then no truck incidents would be reported.

A unit increase of 1,000 in the population would increase the expected crashes by 1.060. This same increase in the combined NAICS group workers and in the workers in retail trade would correspond to 1.021 and 1.150 greater expected crashes, respectively and considered independently while the other variables are held constant. This increase in vehicle flows would increase the accidents by 1.470 for more trucks and by 1.015 for more non-truck vehicles. The greatest increase is shown again to be from the unit increase of the hourly truck volume. A 1% increase in the vacancy rate of housing units would produce 1.014 more expected accidents, which is consistent with the two previous time blocks discussed, but this positive relationship is still not one that was expected.

The differences between this period and the AM are shown in the coefficient signs of those variables that were not deemed significant. While the percentage of commuters using transit and the number of workers in finance and insurance both have negative coefficients again, the workers in educational services are also negative, so these would all lead to decreases in the

expected number of crashes. This third group is interesting to consider as the PM period starts at 3:00 PM when most K-12 schools have been dismissed for the day, so these workers would not be contributing to the vehicle flows, or would only be a factor in the beginning of the time block but not as it approaches the standard PM peak. In this block, the workers in health care and social assistance and those in accommodation and food services both had positive coefficients, or would increase the number of incidents.

The model has a dispersion parameter of 0.252, so this shows that the negative binomial does not collapse to a Poisson model as the value is not zero. Although the standard deviation in crash occurrences was the lowest for the PM, it was still 2.247, so squaring this gives a variance of 5.049, and this is greater than the mean of 1.787. The likelihood ratio Chi-squared value is 127.79, again a very high number, and this is greater than the critical value so at least one of the coefficients is not equal to zero. The pseudo R-squared value of 0.1225 is the highest, yet still low, and coincidentally the Vuong statistic is the lowest.

Table 10: Crash occurrence models for PM period

	Standard Neg Binomial		*Zero-Inflated Neg. Bin.*	
Variable	**Coefficient**	**t Statistic**	**Coefficient**	**t Statistic**
Constant	-0.7564528	-1.76	-0.7146355	-1.63
pop000	0.058414	2.78	0.0479318	1.80
hhsize	0.0508402	0.36	0.0448404	0.32
vac_rate	0.0135411	3.22	0.0146137	3.54
com_transp	-0.0040445	-0.80	-0.0009308	-0.17
com_walkp	0.0073288	1.47	0.0086016	1.83
naics_oth000	0.0203093	2.16	0.0168056	2.01
naics44_45_000	0.1393794	2.00	0.1314707	1.98
naics52_000	-0.0629392	-0.55	-0.0254901	-0.24
naics54_000	0.0248538	0.17	-0.0232173	-0.17
naics61_000	-0.1454564	-0.85	-0.1977763	-1.14
naics62_000	0.037261	0.27	0.1037274	0.75
naics72_000	0.0785723	0.34	0.0630009	0.29
truckpm000	0.385228	3.20	0.2771777	2.12
nontruckpm000	0.0145201	3.69	0.0153468	3.47
Inflated				
Constant	--	--	-2.220778	-1.16
pop000	--	--	-0.1035815	-0.60
com_transp	--	--	0.0237191	0.87
truckpm000	--	--	-1.557157	-1.03
ln alpha	-1.380075	--	-1.909602	-3.68
alpha	0.2515596	--	0.1481393	--
LR chi2	127.79	--	104.66	--
Pseudo R2	0.1225	--	--	--
Vuong	--	--	1.01	--

6.3.2 Zero-Inflated Negative Binomial Results

As mentioned above, the Vuong statistic for this model is 1.01, which is the lowest across all four time periods, and is also below the threshold value. This immediately indicates that the inflated model is not preferred over the standard negative binomial discussed in the previous section. It can also be seen in Table 10 that none of the variables from the binary logit process were significant in the inflated model. Although the population, commuters using transit, and truck flows seemed to influence other time periods and/or the non-certain zero groups, this was not the case here. Looking strictly at the signs, though, it can be seen that an increase in population and truck volume would decrease the odds of being in the certain zero group, while the odds would be increased from an increase in transit-using commuters. The signs match expectations as more people and more trucks would tend to lead to truck accidents.

The non-certain zero model includes the same variables as the regular negative binomial, with the inclusion of the percentage of commuters who walk to work. An increase of 1,000 trucks would increase the number of expected truck accidents by 1.319, whereas an increase in non-truck vehicles would lead to 1.015 greater accidents—the same number as the regular negative binomial. The same unit increase also leads to the following higher number of accidents for each variable independently: population (1.049), combined NAICS group workers (1.017), and retail trade workers (1.141). A 1% increase in vacancy rate leads to 1.015 more accidents, and 1.009 more accidents for an increase in walking commuters. These values are consistent with the other time periods, and again show no drastic jumps in the occurrences.

Looking now at the variables that were not consistent, the signs are the same between this ZINB and the standard model, with the exception of workers in professional, scientific, and technical services. The rest of the discussion could be repeated from the previous section. The Vuong statistic has already been highlighted, and the remaining model statistics verify that at least one of the coefficients is not equal and that the negative binomial does not collapse to Poisson—though the alpha parameter is lower here—and to zero. This dispersion parameter may warrant looking further into zero-inflated Poisson models.

6.4 NT Models

The final BPM time block has the overnight off-peak hours from 7:00 PM-6:00 AM, and the models are shown in Table 11. This time period is of particular interest to the off-hour delivery effort, as described in the research background, and will be touched upon further in Chapter 7.

6.4.1 Negative Binomial Results

This model for the overnight time block had some very interesting results, especially comparing back to the other models already described. The population, vacancy rate, workers in educational services, and hourly truck volume all had positive coefficients; the one that stands out from this set is the workers in educational services as it would be expected to have a negative coefficient like the PM period. A 1.0 percent increase in the vacancy rate increases the expected occurrences by 1.009, holding the other variables in the model constant. The unit increase for the others is 1,000, and there would be respective increases of 1.044, 1.245, and 3.381 truck crashes.

It should be noted that this final number is the only one across all time periods greater than roughly one, and is a rather significant increase. This could have some major implications in the analysis of shifting more truck trips to this time period. The variable for workers in health care and social assistance had a negative coefficient, so an increase of 1,000 of these workers would decrease accidents by 0.774. Although services like emergency rooms tend to have higher intakes in the middle of the night, trucks may not be in these areas as the goods they would deliver are often time-sensitive and require managerial approval, such as pharmaceuticals. The constant is also significant here, leading to an expected value of 0.367 occurrences, but given that there are no trucks traveling to get into these accidents.

The remaining variables all have positive coefficients with the exception of workers in finance and insurance, which would decrease the number of expected crashes; this variable has a negative coefficient across all four time periods. One thing to highlight here is that the percentage of commuters using transit has a positive relationship to the number of crashes, and this contradicts the other time periods. This could be due to the fact that people are not usually commuting to and from work during these overnight hours, and for some of them, transit service is not even available.

The likelihood ratio Chi-squared value of 125.63 is much greater than the critical value for 14 degrees of freedom, so at least one of the model coefficients is statistically different from zero. The dispersion parameter of 0.293 seems to indicate that the negative binomial is a better choice than the Poisson model. The pseudo R-squared value of 0.1199 is still low and is not further interpreted.

Table 11: Crash occurrence models for NT period

Variable	*Standard Neg Binomial*		*Zero-Inflated Neg. Bin.*	
	Coefficient	**t Statistic**	**Coefficient**	**t Statistic**
Constant	-1.00184	-2.08	-0.1214715	-0.24
pop000	0.042814	1.94	-0.0041037	-0.18
hhsize	0.0966172	0.63	0.0140259	0.09
vac_rate	0.0085055	1.92	0.0076849	1.97
com_transp	0.0020057	0.37	0.0022343	0.43
com_walkp	0.0063092	1.14	0.0045344	0.89
naics_oth000	0.0132533	1.24	0.0102762	1.15
naics44_45_000	0.1035541	1.48	0.0684596	1.13
naics52_000	-0.1450878	-1.22	-0.1431341	-1.33
naics54_000	0.1994755	1.31	0.156219	1.09
naics61_000	0.2189044	1.92	0.1047143	0.96
naics62_000	-0.2564989	-1.72	-0.1373968	-0.97
naics72_000	0.149535	0.62	0.0912067	0.43
trucknt000	1.218067	6.17	0.8996977	5.27
nontrucknt000	0.0184006	1.49	0.0158008	1.40
Inflated				
Constant	--	--	0.793287	0.71
pop000	--	--	-0.2974746	-1.92
hhsize	--	--	-0.4817853	-0.81
com_transp	--	--	0.022844	1.20
naics54_000	--	--	-0.428514	-0.95
naics61_000	--	--	-9.347501	-1.88
naics62_000	--	--	3.126362	2.20
trucknt000	--	--	-2.350847	-1.85
ln alpha	-1.226268	--	-2.100361	-4.61
alpha	0.2933854	--	0.1224122	--
LR chi2	125.63	--	98.16	--
Pseudo R2	0.1199	--	--	--
Vuong	--	--	2.40	--

6.4.2 Zero-Inflated Negative Binomial Results

The overnight time period had the greatest number of zero occurrence tracts, so the results of this ZINB model were of particular interest. From the binary logit step, the greatest number of variables was found to be significant for consideration toward predicting membership in the certain zero group. Of the seven included, four were found to be significant, with only one having a positive coefficient value; this was the workers in health care and social assistance, a negative relationship in the standard negative binomial, and the odds of being in the certain zero group increase by a staggering factor of 22.791 with an increase of 1,000 of said workers.

The other three variables are population, workers in educational services, and hourly truck volume. For an increase in 1,000 people, the odds of being in the certain-zero group decrease by 0.743, and they decrease by a factor of 0.095 for an increase of 1,000 trucks. This group of workers also decreased the odds based on an increase of 1,000, but the value is 8.718E-05, so it is

essentially zero. This is the only model that has population and truck flows as significant variables in the inflated portion.

The non-certain zero model actually had fewer significant variables than the inflated model, with only the vacancy rate and truck flow statistically influencing the expected number of crashes. For a 1% increase in the vacancy rate, the expected occurrences increase by 1.008. Increasing the number of trucks per hour during this period by 1,000 would increase the expected number of crashes by 2.459, which is lower than the value for the standard negative binomial model, but still greater than the values around 1.0 for all other variables across the other time block models. Again, this is something that will need to be investigated further when developing the OHD program as the core of this approach is to shift trucks from the regular business hours to these times; if done in a full-scale implementation, there could be a significant amount of trucks transplanted. However, the overall reduction in crashes across the other time blocks may outweigh the increases seen here.

The insignificant variables had the same signs as their counterparts in the standard negative binomial, with the exception of population, which has a negative value—this is the only model in which population is not significant. Similar to the MD block, the non-truck volume was not significant in either the ZINB or regular model. The model statistics seems to indicate, though, that the inflated model is preferred, with a Vuong statistic of 2.40. The LR chi2 value was 98.16, lower than most models but still much greater than the critical value, so again at least one of the coefficients is not zero. The dispersion parameter is somewhat low again at only 0.122, so it may be useful to compare these results to a zero-inflated Poisson model when continuing to improve these analyses.

6.5 Summary of Results for all BPM Time Blocks

The previous sections in this chapter discuss the individual models for each BPM time block, and this section will serve as a summary of the results shown above. Table 12 shows the results of all models, and highlights variables that were significant for each model and their corresponding effect on either the expected number of crashes or the odds of being in the certain-zero group for the inflation variables. The test statistics for each model that were emphasized in the discussion are also provided again.

As can be seen in the table, there were some variables that tended to be significant across the board, such as the vacancy rate and the truck volumes. There were two that were included in almost every model, and these were the population (except for NT ZINB, non-certain zero group) and the combined NAICS workers (not included in the NT models)—in further iterations, it may be valuable to separate this group back out and investigate individual industries or select pairings

as this combination as a whole is difficult to interpret. The non-truck flows were only influential during the two peak periods, as was the case with the number of workers in retail trade. It was expected that the commuters either walking or using transit would play a bigger role, as these were only both included in the PM ZINB model; one or the other was significant in the two MD models and in the PM regular negative binomial. The remaining significant factors were a sampling of the individual NAICS groups.

Table 12: Comparison of occurrence models

	AM NB	AM ZINB	MD NB	MD ZINB	PM NB	PM ZINB	NT NB	NT ZINB
Variable	Exp(Coefficient)		Exp(Coefficient)		Exp(Coefficient)		Exp(Coefficient)	
Constant	*0.774*	*0.800*	1.122	**2.636**	***0.469***	*0.489*	***0.367***	*0.886*
pop000	**1.077**	**1.052**	**1.075**	**1.046**	**1.060**	**1.049**	**1.044**	*0.996*
hhsize	*0.827*	*0.936*	*0.892*	*0.864*	1.052	1.046	1.101	1.014
vac_rate	**1.013**	**1.013**	**1.016**	**1.013**	**1.014**	**1.015**	**1.009**	**1.008**
com_transp	*0.999*	1.003	*0.995*	***0.991***	*0.996*	*0.999*	1.002	1.002
com_walkp	1.006	1.005	**1.008**	1.003	1.007	**1.009**	1.006	1.005
naics_oth000	**1.021**	**1.021**	**1.017**	**1.018**	**1.021**	**1.017**	1.013	1.010
naics44_45_000	**1.127**	1.094	1.097	1.044	**1.150**	**1.141**	1.109	1.071
naics52_000	*0.933*	*0.961*	*0.864*	*0.866*	*0.939*	*0.975*	*0.865*	*0.867*
naics54_000	1.068	1.094	1.144	1.141	1.025	*0.977*	1.221	1.169
naics61_000	1.106	1.100	1.069	1.036	*0.865*	*0.821*	**1.245**	1.110
naics62_000	*0.970*	*0.630*	*0.866*	*0.867*	1.038	1.109	***0.774***	*0.872*
naics72_000	*0.936*	*0.814*	1.238	1.194	1.082	1.065	1.161	1.095
truck*000	**1.206**	**1.195**	**1.288**	**1.235**	**1.470**	**1.319**	**3.381**	**2.459**
nontruck*000	**1.008**	1.004	1.004	1.003	**1.015**	**1.015**	**1.019**	1.016
Inflated								
Constant	--	*0.330*	--	3.750	--	*0.109*	--	2.211
pop000	--	--	--	--	--	*0.902*		***0.743***
hhsize	--	--	--	--	--	--	--	*0.618*
com_transp	--	1.024	--	***0.771***	--	1.024	--	1.023
naics52_000	--	**39.995**	--	--	--	--	--	--
naics54_000	--	--	--	--	--	--	--	0.651
naics61_000	--	--	--	7.994E-08	--	--	--	***8.718E-05***
naics62_000	--	--	--	--	--	--	--	**22.791**
naics72_000	--	***1.046E-06***	--	--	--	--	--	--
truck*000	--	1.280	--	*0.414*	--	*0.211*	--	***0.095***
nontruck*000	--	***0.932***	--	--	--	--	--	--
alpha	0.281	0.133	0.288	0.200	0.252	0.148	0.293	0.122
LR chi2	104.26	63.92	160.55	130.80	127.79	104.66	125.63	98.16
Vuong	--	2.57	--	2.38	--	1.01	--	2.40

* indicates a time-specific variable; for example, truck*000 would be truckam000 in the two AM models.
Bold values are the significant variables; *Italic* values are from negative coefficients.

In terms of actual values, almost all of these predictors led to increases in the number of expected occurrences when the other variables in the model were held constant, and all were on the order of roughly one additional crash per unit increase—either 1,000 for the "000" variables or 1% for the vacancy rate and commuter percentages. Disregarding to the two inflation variables that are close to a zero value, the exceptions in terms of sign are: the non-truck volume in the AM

ZINB (-0.932); the percentage of commuters using transit in the MD ZINB (-0.991); the workers in health care and social assistance in the NT NB (-0.774); and the truck volume in the NT ZINB (-0.095). The two notable exceptions in terms of value are 3.381 and 2.459 for the NT regular negative binomial and the NT ZINB model, respectively—for both of these, not many other factors were influential.

The model statistics were somewhat consistent across the time blocks with alpha values above zero indicating that negative binomial was the best choice for the datasets, as was assumed based on the statistics in Table 2; however, the values for the zero-inflated models are not that high, so it may be worthwhile to run Poisson models to check the numbers. The likelihood ratio statistics were all quite high and much higher than the critical value, so no constant-only models were the best fit. In fact, five out of the eight models did not have significant coefficient values, and even for those that did, the expected number of occurrences cannot be validly interpreted due to the assumption that all variables are evaluated at zero. The PM ZINB was the only of these models that did not have a Vuong statistic confirming the fact that a ZINB specification was preferred over the standard model.

Table 13: Predictor elasticities for occurrence models

	AM NB	**AM ZINB**	**MD NB**	**MD ZINB**	**PM NB**	**PM ZINB**	**NT NB**	**NT ZINB**
Variable	Count elasticity		Count elasticity		Count elasticity		Count elasticity	
pop000	0.4117	0.2800	0.4000	0.2487	0.3228	0.2649	0.2366	-0.0227
hhsize	-0.3770	-0.1309	-0.2275	-0.2909	0.1010	0.0891	0.1919	0.0279
vac_rate	0.1454	0.1387	0.1738	0.1395	0.1498	0.1617	0.0941	0.0850
com_transp	-0.0427	0.1651	-0.2705	-0.5028	-0.2267	-0.0522	0.1124	0.1252
com_walkp	0.1237	0.1120	0.1777	0.0686	0.1568	0.1840	0.1350	0.0970
naics_oth000	0.0639	0.0643	0.0537	0.0544	0.0636	0.0527	0.0415	0.0322
naics44_45_000	0.0548	0.0411	0.0422	0.0195	0.0637	0.0600	0.0473	0.0313
naics52_000	-0.0671	-0.0383	-0.1411	-0.1383	-0.0606	-0.0245	-0.1397	-0.1378
naics54_000	0.0676	0.0915	0.1375	0.1350	0.0254	-0.0237	0.2035	0.1594
naics61_000	0.0355	0.0337	0.0236	0.0126	-0.0513	-0.0697	0.0772	0.0369
naics62_000	-0.0208	-0.3192	-0.0991	-0.0990	0.0258	0.0717	-0.1773	-0.0950
naics72_000	-0.0357	-0.1117	0.1156	0.0962	0.0425	0.0341	0.0810	0.0494
truck*000	0.2813	0.2672	0.3350	0.2797	0.2217	0.1595	0.3499	0.2585
nontruck*000	0.1329	0.0607	0.0840	0.0652	0.2672	0.2824	0.1005	0.0863

The elasticities for each predictor were also calculated, and these values are found in Table 13. For the zero-inflated models, only the non-inflated portion of the model was used, as these were easier to estimate and interpret. Since part of the elasticity calculation considers the change in the dependent variable, or number of crashes in this case, it was only appropriate to look at the non-certain zero indicators. Values greater than the absolute value of 1.0 would indicate an elastic

response, meaning that the percent change in accidents is greater than the percent change of the respective variable; conversely, values between zero and magnitude of 1.0 show inelasticity, so that the percent change in accidents of smaller than the percent change in the variable.

It can be seen in the table that all values fall within the inelastic range, which was unexpected, especially given some of the parameter estimates for the overnight period. As noted with the elasticities in Shankar, et al. [22], this could indicate that these variables are nearing levels at which the crash counts have low sensitivity to any change, despite the variables being statistically significant. The values can still be interpreted as the percent change in the crash frequencies resulting from the percent change in each predictor individually. For instance, the number of crashes would change by 0.4117% for a 1.0% increase in the population, measured by thousands, in the AM period for the NB model, which is the largest value for that model. Pop000 also has the greatest positive elasticity in its respective model for the AM ZINB, MD NB, and PM NB. In the other models, the greatest positive elasticity comes from the truck volume (MD ZINB and both NT models) or the non-truck volume (PM ZINB). Looking at negative elasticities, the household size, commuters who use transit, and certain NAICS groups had the largest values, so the number of crashes would decrease by the respective percentages.

7. FUTURE WORK

As discussed in the introduction of this report, a major motivating factor behind this research was to include traffic safety in the list of metrics by which to assess the effectiveness of an off-hour deliveries (OHD) program. Both the accident severity and crash occurrences in Manhattan are important to examine before and after implementation, not only looking at the accidents themselves, but also because there is a tie-in to the estimation of delays and related costs. The models above represent preliminary takes on each of these aspects separately, and it is shown in the literature review that there are newer works that combine the two to estimate occurrences by severity level—this is a logical next step with this data.

Before heading straight into another round of modeling, there are some changes that should be made to the volume data used in this study, which would make the analysis more robust. While the data is sufficient for this first pass of identifying significant factors, there were some simplifications as explained in Chapter 3. The flows from the Best Practice Model (BPM) were provided for time blocks, not for individual hours of the day, so these values were divided by the number of hours in the block to obtain an average hourly volume. Realistically, though, the traffic volumes are different for every hour of the day; for example, the flows for the AM block from 6:00-10:00 AM will differ from 6:00-7:00 AM, from 7:00-8:00 AM, and so on—there will be a peak hour for each block and for the day as a whole.

In order to adjust for this fact, the bridge traffic volumes report produced by the New York City Department of Transportation (NYCDOT) can be used to produce split factors. The numbers of vehicles using the bridge and tunnel crossings into Manhattan would be summed across the time blocks or whole day, and then the percentages that fall within each hour can be determined. Given that there are roughly 20 crossings into Manhattan that support truck traffic, the key to this process would be deciding how to geographically apply the percentage factors; that is, it must be determined which factors to apply to which links based on spatial relationships [46].

It is also worth mentioning that there are inquiries into the accurateness of the BPM, as it has been shown to over- or underestimate certain vehicle-specific flows. It is discussed in Chapter 3 that not all of the links within Manhattan are represented in the model, so for the severity analysis buffer points were created to estimate flows at the actual accident location, and for the occurrences, the volumes were summed within each tract only for the links present. NYCDOT is currently working on a mesoscopic traffic model that will include all links in the city with more accurate traffic counts based on taxi and other data collected. However, this model does not yet reach above Central Park, so it was not available for use in this analysis.

Once the data is updated, the crash occurrences by severity level can be modeled in a combination. These models could then be used in the post-processor being developed by the OHD team to quantify the total monetary impacts of such a program. Following a more comprehensive literature review on these types of models, and after seeing statistics on the updated data, the most appropriate model for this process can then be determined and estimated.

8. SUMMARY AND CONCLUSIONS

Motor vehicle accidents are a leading cause of annual fatalities in the United States, and aside from personal injuries, are known to cause significant property damages, as well. This makes them a viable study topic for many researchers and national safety organizations; however, most of the work done has been on passenger vehicle involvement and highway segment locations. This research focuses on truck accidents in Manhattan, New York, for a three-year period between May 2008 and April 2011.

The motivation behind this study was the author's involvement in the development of an off-hour delivery program, an alternative congestion management technique, which aims to shift freight deliveries, and consequently truck trips, from regular business hours to the period between 7:00 PM and 6:00 AM. Metrics, such as the safety impacts and costs of delays, are being used to evaluate the effectiveness of the program after a full-scale implementation. This research serves as a preliminary analysis into separately predicting the accident severity level and number of crash occurrences per Census tract with current available data sources. Further iterations will introduce newly-anticipated data, primarily with regard to the vehicle flow values, and will examine crash occurrence by severity level.

The results of the ordered probit severity analysis highlighted that an increase in the number of vehicles involved would decrease the likelihood of a more severe accident, as would an increase in the total vehicle volume. Similarly, accidents falling from 5:00-6:00 PM would also be less likely to be as severe, as this is the evening peak period when flows are more restricted. The hours of 5:00-6:00 AM and 1:00-3:00 PM showed crashes that were of higher severity. Roadway conditions of a dry surface and a dark road with street lighting also increased the odds of the crash being more severe. It can be noted that some of these variables are proxies for the speed, which is known to dramatically impact severity.

The crash occurrences analysis was broken into eight models to include a standard negative binomial and a zero-inflated negative binomial for each of the four time blocks identified by the Best Practice Model: AM (6:00-10:00 AM), MD (10:00 AM-3:00 PM), PM (3:00-7:00 PM) and NT (7:00 PM-6:00 AM). The models indicated that the zero-inflated approach was preferred for all but the PM period, meaning that some tracts are always going to have zero occurrences. Variables such as the population, housing unit vacancy rate, and hourly truck volume were significant in all models and, when increased individually, would lead to an increase in expected number of crashes. Other variables leading to increases were certain industrial sectors, such as retail trade during the AM and PM, and the portion of commuters walking. Commuters by transit, when significant, decreased the expected occurrences.

Linking back to the OHD program, it was shown that an increase of 1,000 trucks during the NT period would lead to an increase of 3.381 accidents under the regular negative binomial model, and 2.459 under the zero-inflated model. All of the other positive, significant coefficients led to changes in the order of one, so this increase of 2-3 crashes should be investigated further, especially as a full implementation would send a lot of trucks to this time period; however, as discussed in Chapter 6, the decrease in accidents as a result of removing trucks from the other times may outweigh these increases. It was also interesting to discover that the number of workers in retail trade increased expected occurrences for the AM negative binomial, and for both models during the PM period. This is one of the industrial sectors being targeted for OHD, and these time periods are the busiest. In contrast, it was expected that the works in accommodation and food services would have a similar outcome, and this group was not significant in any models; it was included in the AM zero-inflated, but the value approached zero.

9. REFERENCES

[1] U.S. Dept. of Transportation, "2011 Traffic Safety Facts," Nat. Highway Traffic Safety Admin., Washington, D.C., 2012.

[2] Centers for Disease Control and Prevention, *Injury Prevention and Control: Motor Vehicle Safety*, Centers for Disease Control and Prevention, Oct. 2012. [Online]. Available: http://www.cdc.gov/motorvehiclesafety/ [Accessed: 11 March 2013].

[3] D. Schrank *et al*., "TTI's 2012 Urban Mobility Report," College Station, TX, 2012.

[4] J. Holguín-Veras *et al*., "Overall impacts of off-hour delivery programs in New York City Metropolitan Area," *Transportation Research Rec.: J. Transportation Research Board,* vol. 2238, pp. 68-76, 2011.

[5] P. T. Savolainen *et al*., "The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives," *Accident Anal. & Prevention,* vol. 43, pp. 1666-1676, Sep. 2011.

[6] A. S. Al-Ghamdi, "Using logistic regression to estimate the influence of accident factors on accident severity," *Accident Anal. & Prevention,* vol. 34, pp. 729-741, Nov. 2002.

[7] C. Lee and M. Abdel-Aty, "Presence of passengers: Does it increase or reduce driver's crash potential?," *Accident Anal. & Prevention,* vol. 40, pp. 1703-1712, Sep. 2008.

[8] Y. Ouyang *et al*., "Modeling the simultaneity in injury causation in multivehicle collisions," *Transportation Research Rec.: J. Transportation Research Board,* vol. 1784, pp. 143-152, 2002.

[9] H. Huang *et al*., "Severity of driver injury and vehicle damage in traffic crashes at intersections: A Bayesian hierarchical analysis," *Accident Anal. & Prevention,* vol. 40, pp. 45-54, Jan. 2008.

[10] V. Shankar *et al*., "Statistical analysis of accident severity on rural freeways," *Accident Anal. & Prevention,* vol. 28, pp. 391-401, May 1996.

[11] P. Savolainen and F. Mannering, "Probabilistic models of motorcyclists' injury severities in single-and multi-vehicle crashes," *Accident Anal. & Prevention,* vol. 39, p. 955, Sep. 2007.

[12] C. S. Duncan *et al*., "Applying the ordered probit model to injury severity in truck-passenger car rear-end collisions," *Transportation Research Rec.: J. Transportation Research Board,* vol. 1635, pp. 63-71, 1998.

[13] K. M. Kockelman and Y. J. Kweon, "Driver injury severity: an application of ordered probit models," *Accident Anal. & Prevention,* vol. 34, pp. 313-321, May 2002.

[14] M. Abdel-Aty, "Analysis of driver injury severity levels at multiple locations using ordered probit models," *J. Safety Res.,* vol. 34, pp. 597-603, 2003.

[15] S. M. Rifaat and H. C. Chin, "Accident severity analysis using ordered probit model," *J. Advanced Transportation,* vol. 41, pp. 91-114, 2007.

[16] J. B. Edwards, "The relationship between road accident severity and recorded weather," *J. Safety Res.,* vol. 29, pp. 249-262, 1999.

[17] D. Lord and F. Mannering, "The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives," *Transportation Res. Part A: Policy and Practice,* vol. 44, pp. 291-305, 2010.

[18] S. C. Joshua and N. J. Garber, "Estimating truck accident rate and involvements using linear and Poisson regression models," *Transportation Planning and Technology,* vol. 15, pp. 41-58, 1990.

[19] B. Jones *et al*., "Analysis of the frequency and duration of freeway accidents in Seattle," *Accident Anal. & Prevention,* vol. 23, pp. 239-255, Aug. 1991.

[20] S. Daniels *et al*., "Explaining variation in safety performance of roundabouts," *Accident Anal. & Prevention,* vol. 42, pp. 393-402, Mar. 2010.

[21] S.-P. Miaou, "The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions," *Accident Anal. & Prevention,* vol. 26, pp. 471-482, Aug. 1994.

[22] V. Shankar *et al*., "Effect of roadway geometrics and environmental factors on rural freeway accident frequencies," *Accident Anal. & Prevention,* vol. 27, pp. 371-389, Jun. 1995.

[23] J. Carson and F. Mannering, "The effect of ice warning signs on ice-accident frequencies and severities," *Accident Anal. & Prevention,* vol. 33, pp. 99-109, Jan. 2001.

[24] M. Poch and F. Mannering, "Negative binomial analysis of intersection-accident frequencies," *J. Transportation Eng.,* vol. 122, pp. 105-113, Mar. 1996.

[25] J. Milton and F. Mannering, "The relationship among highway geometrics, traffic-related elements and motor-vehicle accident frequencies," *Transportation,* vol. 25, pp. 395-413, Nov. 1998.

[26] M. G. Karlaftis and A. P. Tarko, "Heterogeneity considerations in accident modeling," *Accident Anal. & Prevention,* vol. 30, pp. 425-433, Jul. 1998.

[27] D. Lord, "Modeling motor vehicle crashes using Poisson-gamma models: Examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter," *Accident Anal. & Prevention,* vol. 38, pp. 751-766, Jul. 2006.

[28] N. V. Malyshkina and F. L. Mannering, "Empirical assessment of the impact of highway design exceptions on the frequency and severity of vehicle accidents," *Accident Anal. & Prevention,* vol. 42, pp. 131-139, Jan. 2010.

[29] M. Ridout, C. G. Demétrio, and J. Hinde, "Models for count data with many zeros," in *Proc. XIXth Int. Biometric Conf.*, Cape Town, South Africa, 1998, pp. 179-192.

[30] D. Lord *et al*., "Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory," *Accident Anal. & Prevention,* vol. 37, pp. 35-46, Jan. 2005.

[31] D. G. Kim and S. Washington, "The significance of endogeneity problems in crash models: An examination of left-turn lanes in intersection crash models," *Accident Anal. & Prevention,* vol. 38, pp. 1094-1100, Nov. 2006.

[32] L. Mountain *et al*., "Accident prediction models for roads with minor junctions," *Accident Anal. & Prevention,* vol. 28, pp. 695-707, Nov. 1996.

[33] L. Mountain *et al*., "The influence of trend on estimates of accidents at junctions," *Accident Anal. & Prevention,* vol. 30, pp. 641-649, Sep. 1998.

[34] S. Cafiso *et al*., "Development of comprehensive accident models for two-lane rural highways using exposure, geometry, consistency and context variables," *Accident Anal. & Prevention,* vol. 42, pp. 1072-1079, Jul. 2010.

[35] B. Heydecker and J. Wu, "Identification of sites for road accident remedial work by Bayesian statistical methods: an example of uncertain inference," *Advances in Eng. Software,* vol. 32, pp. 859-869, Oct. 2001.

[36] W. Hirst *et al*., "Sources of error in road safety scheme evaluation: a method to deal with outdated accident prediction models," *Accident Anal. & Prevention,* vol. 36, pp. 717-727, Sep. 2004.

[37] M. A. Quddus, "Time series count data models: An empirical application to traffic accidents," *Accident Anal. & Prevention,* vol. 40, pp. 1732-1741, Sep. 2008.

[38] Y. Xie *et al*., "Predicting motor vehicle collisions using Bayesian neural network models: An empirical analysis," *Accident Anal. & Prevention,* vol. 39, pp. 922-933, Sep. 2007.

[39] X. Li *et al*., "Predicting motor vehicle crashes using support vector machine models," *Accident Analysis & Prevention,* vol. 40, pp. 1611-1618, Jul. 2008.

[40] J. Ma and K. M. Kockelman, "Bayesian multivariate Poisson regression for models of injury count, by severity," *Transportation Research Rec.: J. Transportation Research Board,* vol. 1950, pp. 24-34, 2006.

[41] J. Ma *et al.* , "A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods," *Accident Anal. & Prevention,* vol. 40, pp. 964-975, May 2008.

[42] E. S. Park and D. Lord, "Multivariate Poisson-lognormal models for jointly modeling crash frequency by severity," *Transportation Research Rec.: J. Transportation Research Board,* vol. 2019, pp. 1-6, 2007.

[43] S. P. Washington *et al.*, *Statistical and Economic Methods for Transportation Data Analysis*. Boca Raton, FL: Chapman & Hall, 2011.

[44] M. E. Ben-Akiva and S. R. Lerman, *Discrete Choice Analysis: Theory and Application to Predict Travel Demand* vol. 9. Cambridge, MA: MIT press, 1985.

[45] J. Holguín-Veras, "Necessary conditions for off-hour deliveries and the effectiveness of urban freight road pricing and alternative financial policies in competitive markets," *Transportation Res. Part A: Policy and Practice,* vol. 42, pp. 392-413, 2008.

[46] New York City Department of Transportation, "New York City Bridge Traffic Volumes 2010," New York, NY, 2012.