TransInfo
Tier I University Transportation Center

FINAL REPORT

Buffalo-Niagara Transportation Data-warehouse Prototype and Real-time Incident Detection

Date of report: November 2017

Andrew Bartlett
Graduate Research Assistant, University at Buffalo
Transportation Engineer, Niagara International Transportation Technology Coalition (NITTEC)
Adel W. Sadek, Ph.D.
Professor, University at Buffalo
Director, Transportation Informatics University Transportation Center
Associate Director, Institute for Sustainable Transportation & Logistics

Prepared by:
Department of Civil, Structural & Environmental Engineering, Univ. at Buffalo

Prepared for:
Transportation Informatics Tier I University Transportation Center
204 Ketter Hall
University at Buffalo
Buffalo, NY 14260

1. Report No.

2. Government Accession No.

3. Recipient's Catalog No.

4. Title and Subtitle

Buffalo-Niagara Transportation Data-warehouse Prototype and Real-time Incident Detection

5. Report Date

November 2017

6. Performing Organization Code

7. Author(s)
Andrew Bartlett & Adel W. Sadek

8. Performing Organization Report No.

9. Performing Organization Name and Address

Department of Civil, Structural and Environmental Engineering
University at Buffalo
204 Ketter Hall
Buffalo, NY 14260

10. Work Unit No. (TRAIS

11. Contract or Grant No.
DTRT13-G-UTC48

12. Sponsoring Agency Name and Address
US Department of Transportation
Office of the
UTC Program, RDT-30
1200 New Jersey Ave., SE
Washington, DC 20590

13. Type of Report and Period Covered

Final – June 2014 – June 2017

14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

In the traffic engineering field, study and analysis often requires the use of multiple datasets. The nature of these data often makes them difficult to work with, especially in conjunction with one another. The overall goal of this study was to not only design a solution to this problem for the Buffalo-Niagara Region of western New York, but to demonstrate its usefulness through a specific application. To meet these objectives a prototype data warehouse was first created. The warehouse was and then tested on two applications. In the first application, the warehouse was used in the creation and validation of three incident detection strategies: a speed threshold detection system, a binary probit model which uses only speed data, and a binary probit model which uses a combination of speed and volume data. In the second application, the data was used to test the accuracy of weather impact models which had been previously developed.

17. Key Words

Data warehouse, automated incident detection algorithms, inclement weather impact, traffic flow.

18. Distribution Statement

No restrictions. This document is available from the National Technical Information Service, Springfield, VA 22161

19. Security Classif. (of this report)

Unclassified

20. Security Classif. (of this page)

Unclassified

21. No. of Pages

37

22. Price

Acknowledgements

Niagara International Transportation Technology Coalition

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the U.S. Department of Transportation's University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

EXECUTIVE SUMMARY

In the traffic engineering field, study and analysis often requires the use of multiple datasets. The nature of these data often makes them difficult to work with, especially in conjunction with one another. The overall goal of this study was to not only design a solution to this problem for the Buffalo-Niagara Region of western New York, but to demonstrate its usefulness through two specific applications. To achieve this, three objectives were designed: (1) outline the structure of a data warehouse for the Buffalo-Niagara region, (2) use the combined data in the prototype warehouse to examine its usefulness in the construction of a real-time incident detection system which not only detects incidents but also tries to predict incident characteristics, and (3) use the data in the warehouse to test the accuracy of weather impact models which had been previously developed by the authors for assessing the impact of inclement weather on average speed and traffic volumes.

To meet these objectives a prototype data warehouse was first created. For the first application involving the development of a real-time incident detection system, three incident detection strategies were created and validated. These were: (1) a speed threshold detection system; (2) a binary probit model which uses only speed data; and (3) a binary probit model which uses a combination of speed and volume data. The prototype data warehouse showed it was possible to construct a fully fleshed-out version for transportation data in the Buffalo-Niagara region with useful results. The speed threshold model which used a 10 minute speed drop of 10 mph to detect incidents had a 62.5% detection rate, as well as favorable false alarm and classification rates. The more complex binary outcome model which used only speed data detected incidents with a success rate of 70.4%, an improvement over the speed threshold model despite worse false alarm and classification rates. It was also able to predict incident type, number of blocked lanes, and incident severity with 75.9%, 70.4%, and 75.9% accuracy, respectively. The binary outcome model which used both speed and volume data had a more impressive detection rate of 75.5% with similar false alarm and classification rates and was slightly better at predicting incident type and severity (both with 77.6% accuracy) but slightly worse at predicting the number of blocked lanes (with 69.4% accuracy). Overall, the combined data model is the best strategy for both detecting incidents and predicting their characteristics, which emphasizes the importance of a transportation data warehouse.

The second application was motivated by the fact that inclement weather could have a significant impact on traffic flow conditions. In previous work by the authors, we have attempted to model average operating speed and hourly traffic volume, respectively, as a function of weather conditions. With the data warehouse constructed, our goal was to validate these models with more recent data (specifically date from the 2013-2014 winter in Buffalo). Using such data, the study showed first that the winter of 2013-2014 appeared to have had significantly lower minimum and average temperatures than other years examined. In terms of the previous models' accuracy, the speed model performed reasonably well, usually achieving results within 5 mph of the observed speed, but accuracy somewhat suffered when inclement weather conditions were harsh or when observed speeds were below 40 mph. The volume model, on the other hand, was not as accurate, and tended to overestimate hourly volumes.

INTRODUCTION

In the traffic engineering field, study and analysis often requires the use of multiple datasets. The nature of these data often makes them difficult to work with, especially in conjunction with one another. This results from the fact that each set of data is collected and stored by different agencies, each with its own formatting and resolution. For example, in the Buffalo-Niagara region of Western New York a variety of datasets, including traffic speeds and volumes on both highways and local roads, traffic accidents, weather, and pavement conditions, are all recorded by different entities and with different formatting and timing schemes.

This study sought to create a solution to this problem through the creation of a data warehouse. According to the Oracle9i Warehousing Guide, "a data warehouse is a relational database that is designed for query and analysis rather than for transaction processing [and] it usually contains historical data derived from transaction data…" (Oracle, 2015). These characteristics of data warehousing make it a fitting solution to the transportation data collection problem discussed here.

In addition to creating a data warehouse, this study also sought to demonstrate potential applications made possible or facilitated by the warehouse. Two applications were chosen for that purpose. The first involved creating a real-time incident detection and characterization system using a combination of speed and volume observations. Since this application would require not only all three of these data sets but also information from them in real time, it is a perfect representation of the utility of a transportation data warehouse.

Furthermore, a real-time incident detection system has the potential to be highly beneficial in and of itself. Faster incident detection times can have a significant impact on incident response and clearance times (PB Farradyne, 2015). Currently, most incidents in the Buffalo-Niagara region are detected via CCTV cameras by the Niagara International Transportation and Technology Coalition (NITTEC). However, there is an inherent limitation in the number of operators watching the camera feeds and the number of cameras they can watch. An incident detection system based on the data collection processes already in place for speed and volume would aid operators in finding and reporting incidents more quickly and thus improving roadway performance and safety.

The second application entailed using the data in the warehouse to test the accuracy of weather impact models which had been previously developed by the authors for assessing the impact of inclement weather on average speed and traffic volumes. This application is motived by the fact that inclement weather conditions, such as fog, rain, snow, and ice, are known to negatively impact the operational efficiency of networks, as well as user safety. According to the Federal Highway Administration (FHWA), for example, weather is responsible for nearly one quarter of all traffic accidents, resulting in over 1.3 million crashes and 6,200 deaths annually in the United States (FHWA, 2014). Therefore, it is important to have a significant understanding of how both inclement weather as a whole and individual weather factors affect the operation conditions of traffic networks.

Recently, the effect of inclement weather on freeway operating conditions near the city of Buffalo, NY has been the topic of multiple studies. Specifically, two studies have attempted to model average operating speed and hourly traffic volume, respectively, as a function of weather conditions (Zhao et al., 2011; Bartlett et al., 2012). Buffalo provided a unique case study for inclement weather research due to the presence of microclimates: small areas which experience a wide variety of weather conditions over a short amount of time.

These two previous studies produced a great amount of meaningful results, but were also subject to some strict limitations, the greatest being data scarcity. While Buffalo has a reputation for harsh weather conditions, some recent winters have been perceived to be exceptionally mild. Due to this, the datasets used by both previous studies had very little data about what was defined as "inclement weather" relative to the period of time over which the data was collected. However, since the release of these studies and the construction of the data ware house, the area has returned to more normal conditions. In particular, the winter of 2013-2014 was perceived as being especially severe with low temperatures and heavy precipitation.

Given the above, the second application we considered in this research, used the data warehouse to: (1) determine if the public perception about weather in Buffalo was true: that recent winters had been milder than normal but the (2013-2014) winter has been harsher; (2) ascertain whether the models developed by the previous two studies, for predicting freeway speed and volumes as function of weather conditions (Zhao et al, 2011 and Bartlett, 2012) would still yield decent results when applied to a winter much harsher than the winters that yielded the data from which the models were developed. To do this, the study examined weather, volume, and speed on seven freeway links near Buffalo (Figure 1) and used data from only winter months, since this data contained the highest frequency of inclement weather.

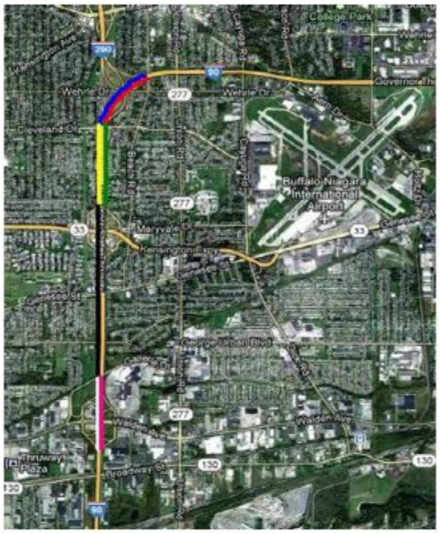

Figure 1: Freeway Links Used in Validation (Bartlett et al., 2012)

There are several aspects of this project which make it both significant and unique. The first is the area of study; neither the data warehouse nor the automated incident detection component has ever been attempted in the Buffalo-Niagara region. Also, while the methodologies explored for incident detection were inspired by those described in the literature review, these specific techniques have not before been applied to this problem, especially the binary outcome models based on speed and volume. Third, the idea of an incident detection system which is also used to predict incident characteristics based on speed and volume data is a concept on which little previous research has been done. Finally, the idea of developing and validating models for assessing the impact of inclement weather on traffic flow and speed is critical for efforts aimed at developing weather-responsive traffic management strategies.

It should be noted that this study had two main objectives. The first was to outline the creation of a transportation data warehouse to bring these different datasets together so that not only are they all in a central location, but also have compatible formatting for simplified cross analysis. This will focus less on the technical aspects of warehouse creation and more on the data processing and formatting required. The second objective was to use the combined data in the prototype warehouse to examine two applications of the data warehouse, as just mentioned above, to show the utility of the data warehouse.

The remainder of this report will be organized as follows. First, the Literature Review section will examine: (1) techniques in data warehouse development; (2) past experiences in creating real-time incident detection systems; and (3) the previous two studies by Zhao et al (2011) and Bartlett et al (2012) which developed models for predicting the impact of inclement weather on traffic flow. Next, the data description and processing, along with the development of the data warehouse prototype will be described; this section will present each dataset used in this project including its source, collection procedure, and formatting. Following this will be the Methodology section which will explain the development of the data warehouse prototype. Two sections will then follow: the first will discuss the incident detection application and the second will describe the inclement weather model validation application. The Conclusions section will summarize the findings of the project and the Future Work section will outline the steps which need to be taken to bring the data warehouse to fruition, and to build upon the two applications considered in tis research.

LITERATURE REVIEW

Data Warehouse Development

Given the scope of this study, this report will not place great focus on the technical design of a data warehouse. However, a brief overview of basic data warehouse development will be discussed here in terms of the attributes which make it ideal for application to transportation data.

The Oracle9i Data Warehousing Guide presents an overview of data warehouses in a business context and describes many traits which make them an ideal technique for storing and related transportation data. The guide explains that data warehouses handle data modification by using bulk data modification techniques which are run on a regular basis. This results in users being unable to directly update the data warehouse. Additionally, data warehouses hold large amounts of historical data and can query thousands or millions of rows. It is also explained that data warehouses often have "staging areas" where data is cleaned in processed before being entered into the warehouse. These are the features of a data warehouse that make it ideal for this project (Oracle, 2015).

Incident Detection

Real-time automatic incident detection has the potential to be beneficial for agencies by decreasing incident response times for either monitoring agencies or emergency response teams. For the lasts several years, several efforts have been made to apply a variety of statistical methods to traffic data as it is collected to see how incidents can be detected most accurately.

One study performed by Wang et. al. (2006) used a partial least squares regression (PLSR) algorithm to detect incidents. A simulated section of an expressway in Singapore was created and 300 incident cases were generated. For each incident, volume, occupancy, and speed were measured upstream and downstream of the incident during the time leading up to and following it. Additionally, that study outlined the criteria used to measure detection effectiveness as detection rate (DR), false alarm rate (FAR), mean time to detection (MTTD), and classification rate (CR). The calculation of these measures is shown below.

$$\mathrm{DR} = \frac{number\ of\ detected\ incidents}{number\ of\ incidents}$$

[Equation 1]

$$\mathrm{MTTD} = \frac{\sum t_i}{m}$$

[Equation 2]

Where t_i represents the time between the incident and the time it was detected for incident *i*, and m represents the number of incidents examined.

$$\mathrm{FAR} = \frac{number\ of\ false\ alarm\ cases}{number\ of\ non - incident\ cases}$$

[Equation 3]

$$\mathrm{CR} = \frac{number\ of\ correctly\ indentified\ incidents}{number\ of\ recorded\ instances}$$

[Equation 4]

It was concluded that this approach was viable, but further work would be needed to enhance its predictive abilities. This was caused by the linear nature of the PLSR used and the use of simulated data (Wang et al, 2006).

Another study was performed by Raiyn and Toledo (2014) which took a different approach. Here, speed was measured both upstream and downstream of the incident and compared to the average speed for that link on the same day of the week and at the same time. If the observed speed was 30 km/hr (18.6 mph) lower than the average, then an incident was predicted to have occurred. While this methodology showed promise, there were several problems encountered when a speed reading of 0 was recorded, since this could indicate either no vehicle or a failure of the recording device to measure the speed correctly (Raiyan and Toledo, 2014).

Another approach to incident detection was undertaken by Motamed and Machemehl (2014), where a dynamic time warping model and support vector machine were used to detect short-term congestion in Dallas, Texas. The algorithms used speed, standard deviation of speed, and occupancy as parameters and examined two thresholds: one to start collecting data, and another to indicate the occurrence of an incident. This was done to decrease the abundance of false alarms found in other detection algorithms. It was found that both algorithms were highly effective but requires large datasets and extensive amounts of training and validation (Motamed and Machemehl, 2014).

A study done by Cheung and Truong (2011) for Auckland Motorways in New Zealand used a time series model to detect incidents in an effort to reduce incident response time. Here, speed and volume (or occupancy) were used to detect accidents and found they were able to differentiate between recurrent congestion and accidents based on the volume/occupancy’s rate of change. In this case, it was concluded that an automated detection system was a critical addition to the CCTV monitoring system already in place in terms of incident detection and response time (Cheung and Truong, 2011).

The final incident detection study examined was performed by Ahmed and Hawas (2012). While all previous studies examined freeway links, this study focused on urban roadways. Vehicle counts and speeds were recorded with stop line detectors and used in a threshold-based linear regression model. DRs and FARs were found to be sufficient in all cases, with the exception of periods with low hourly volumes. The study also concluded that further research would be needed in the areas of sensitivity analysis, collection techniques, and model sophistication (Ahmed and Hawas, 2012).

Inclement Weather models for Western New York

Due to the nature of the second application considered in this study as a validation effort, the literature review pertaining to that application mainly focused on the two previous papers which describe the models being validated.

Average Operating Speed Model

The details of the creation of this model are described in Zhao et al, 2011. One of the contributions of that study was the creation of several indices for weather data which were much more suitable for modeling than the raw data. These included a visibility index, weather type index, temperature index, and precipitation index. Raw wind speed was used, as opposed to an index. In addition to weather data, two other parameters were used in the creation of the operating speed model. First was a day index, which captured the effect of changing traffic patterns across different days of the week. The second was a normal average hourly speed, which was equal to the average speed for an hour of the day averaged over a given time period (e.g. one month). The details of these indices are given in Zhao et al, 2011.

The linear regression model for average operating speed is shown in [Equation 5]. All of the included variables were found to be highly statistically significant with the exception of temperature index. The R^2 value of the model was 56.1%, which compared favorably to previous models reported in the literature attempting to predict weather impact on traffic flow. The sign of each variable was intuitive in terms of its effect on operating speed, further supporting the accuracy of the model.

$$\begin{aligned}
&Average\ Operating\ Speed \\
&\quad = 7.23 + 0.77[Visibility\ Index] + 0.358[Weather\ Type\ Index] \\
&\quad + \ 0.132[Temperature\ Index] - 0.0469[Wind\ Speed] \\
&\quad - 1.92[Cumulative\ Precipitation] + 0.853[Normal\ Hourly\ Speed] \\
&\quad - 0.935[Day\ index]
\end{aligned}$$

[Equation 5]

Hourly Volume Model

This model was both created and validated in Zhao et al, 2012. Specifically, that study created a model to predict hourly traffic volume using many of the indices created in the speed model study, including the weather type index, temperature index, and cumulative precipitation index. However, this model incorporated visibility and wind speed data directly, without the use of indices. In addition, the volume model used a baseline variable, which captured the effect of hour of day.

This study also created a system for defining inclement weather. After determining that there was a significant difference in volume between times with inclement weather and times without, the model was created using only inclement weather data to predict traffic volume.

The linear regression model for average hourly volume is shown in [Equation 6]. The regression model was found to fit the data closely, with an R^2 value of 97.4%. The included parameters of the model were all found to be highly statistically significant with the exception of visibility and temperature index. With the exception of wind speed, the sign of each variable was intuitive as well.

$$\begin{aligned}
Volume = {} & 65.07\ [Weather\ Type\ Index] + \ 24.52[Wind\ Speed] + 28.06[Visibility] \\
& - 419.73[Cumulative\ Precipitation] + 61.06[Temperature\ Index] \\
& + 0.76[Baseline]
\end{aligned}$$

[Equation 6]

Using inclement weather and volume data not used in model creation, the hourly volume model was validated by plotting volumes predicted by the model against observed volumes. The resulting plot was mostly linear, showing the model to have significant predictive ability. However, the study also acknowledged that the lack of data resulted in a small validation sample.

METHODOLOGY

DATA DESCRIPTION AND PROCESSING

In this project there were several datasets which founded the motivation and foundation for the data warehouse. These included the data about freeway volumes, freeway speeds, incidents, and weather. Below, each type of data will be discussed, including the source, raw format, and necessary processing for inclusion in the warehouse and for use in model creation. For demonstration, all data was collected for two months (July 2014 and May 2015) so that one could be used for model creation, while the other for validation. The choice of these two specific months was made due to convenience of data availability.

Thruway Volume

The thruway volume data was obtained from the New York State Thruway Authority (NYSTA) and included count data for all thruway links between exits 49 and 54 of I-90, the section of the thruway which is close to Buffalo. Volumes were recorded in vehicles per lane per 15 minutes. For use in this study, these were converted to hourly volumes for all lanes combined. Links were defined by their thruway ID, a number assigned by the collecting agency.

Thruway Speed

The thruway speed data was obtained from NITTEC for the same links as the volume data. The speeds were given in miles per hour (mph) at 10 minute intervals. Links were defined by their TRANSMIT ID, a number assigned by the collecting agency.

Incident Data

The incident dataset, also obtained from NITTEC, contained a large number of fields containing information about each incident. Of the numerous fields, four were extracted for the purposes of this study: Incident Type, Incident Start Time, Number of Closed Lanes, and Incident Severity. Incident Type categorized the incident as either an accident, congestion, disabled vehicle, or incident (a generic catchall for incidents that do not fall into the first three classifications). Incident Start Time was formulated as a date and time stamp giving the year, month, day, hour, and minute the incident was reported to start. Number of Closed Lanes gave the lane closures associated with the incident, ranging from 0 to the number of lanes of the given link. Finally, Incident Severity ranked the severity of the incident in terms of human injuries and fatalities on a scale from 0 to 3, 0 being property damage only and 3 being most severe.

Weather Data

As with the incident data, the weather data, collected from the National Oceanic and Atmospheric Administration (NOAA), contained numerous fields of information and only a few were selected for use in this study (NOAA, 2015). These weather factors were: Temperature (^{0}F), Precipitation (inches), Wind Speed (mph), and Visibility (miles) . The weather data was recorded approximately hourly, but not regularly. For example, data might be recorded at 12:04 AM, 12:54 AM, 1:54 AM, 2:54 AM, and 3:12 AM. To solve this problem, weather data collection times were “rounded” to the nearest hour and duplicates were removed.

Prototype Data Warehouse

The first organizational step taken in the creation of the prototype data warehouse was to split the structure into two tables. One, the Link ID Table, contained the data about the link, including location, speed limit, number of lanes, road functional class, TRANSMIT ID, thruway ID, length, and direction. The other, the Link Data Table, contained the count, speed, travel time, delay, incident, and weather data for each link. These two tables were joined by a link ID. The link ID was a new number assigned to each link and then matched with the corresponding TRANSMIT and thruway IDs so that no matter what data was being examined there would be a common link identifier.

The processed speed, volume, and weather data was linked by time and date and inserted into the Link Data Table. Since each was recorded at different time intervals, the dataset with the highest resolution (speed) was used. This meant expanding volume and weather data by copying the most recent reading so every 10 minute interval had speed, volume, and weather data. Next the incident data was inserted during the time interval containing the incident start time (e.g. 1:00 PM if the incident occurred at 1:03 PM). Finally the travel time, delay, and density were calculated for each reading using data from both the Link ID and Link Data Tables and the following equations.

$$Travel\ Time = \frac{Distance * 60}{Speed}$$

[Equation 7]

$$Delay = \begin{cases} Travel\ Time - \left(\frac{Distance * 60}{Speed\ Limit}\right), if\ Delay > 0 \\ 0, otherwise \end{cases}$$

[Equation 8]

$$Density = \frac{Volume}{Speed}$$

[Equation 9]

DATA WAREHOUSE APPLICATION 1: REAL-TIME INCIDENT DETECTION

Based on the literature review and the available data, several incident detection strategies were developed and tested to determine how both incidents and their properties could be detected or predicted in real time. Three different methods were analyzed in terms of their ability to detect either incidents only or both incidents and their properties. The methods used were simple speed threshold method, a binary outcome model which uses only speed-related factors as input, and a binary outcome model which uses volume-related factors in addition to the speed data. In addition to detecting incidents, the two types of binary outcome models were used to predict the incident properties, including incident type (congestion vs. accident), severity, and how many lanes were blocked. This section will outline the three methods and how each was applied to the collected data.

Speed Threshold Detection

Speed threshold detection was the simplest incident detection method used and relied solely on the observed speed data for each freeway link. This method attempted to determine whether and incident had occurred based on irregularities in the observed speed. Despite its simplicity, the best way to define the parameters of detection had yet to be determined. First, a decision had to be made between detecting incidents using instances where the speed was significantly below free flow speed or instances where speed was significantly lower than the immediately preceding speed observation. By examining the link speeds which correlated to several incidents, it was seen that regardless of the incident type a large majority were characterized by a sudden drop in speed followed by a return to normal operating speeds some time later, as shown in Figure 2 and Figure 3. In these figures, the vertical red line represents the Incident Start Time. Since the central objective was to detect the incident as soon as it occurs, it was decided that the difference in speed from one observation to the next would be the best measure of speed to detect incidents.

Once it was concluded that speed difference should be used, the threshold of speed change which should be used to determine if an incident had occurred needed to be determined. An obvious trade off exists between using either a high or low value as the threshold. A lower speed change threshold would detect nearly all incidents that occur but would have many false positive readings resulting from normal fluctuations in speed. Alternatively, a high value would avoid many of the false positive readings but could potentially miss some less severe incidents.

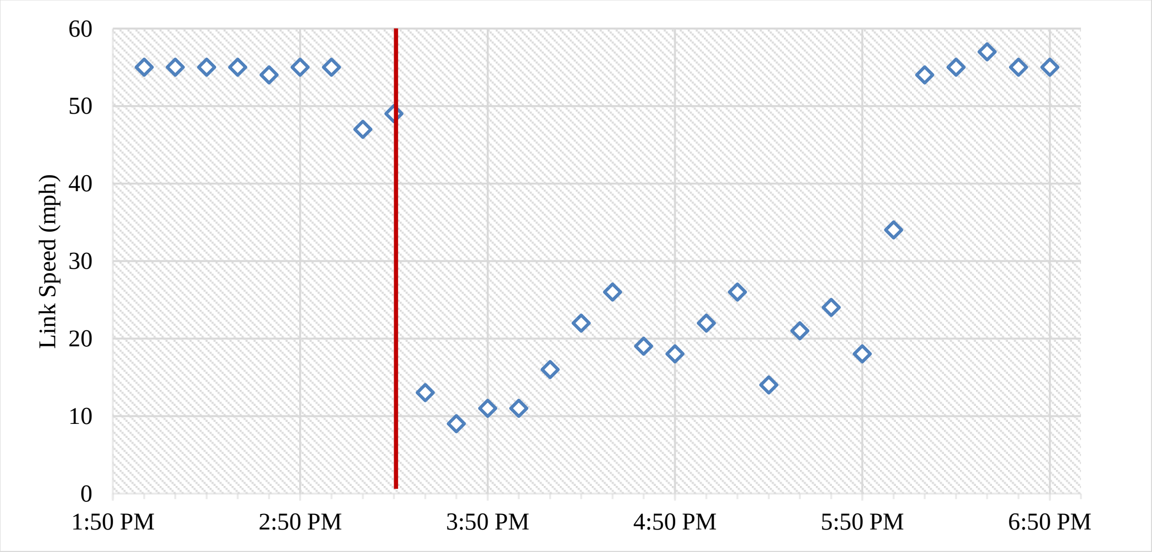

Figure 2: Accident Speed Impact

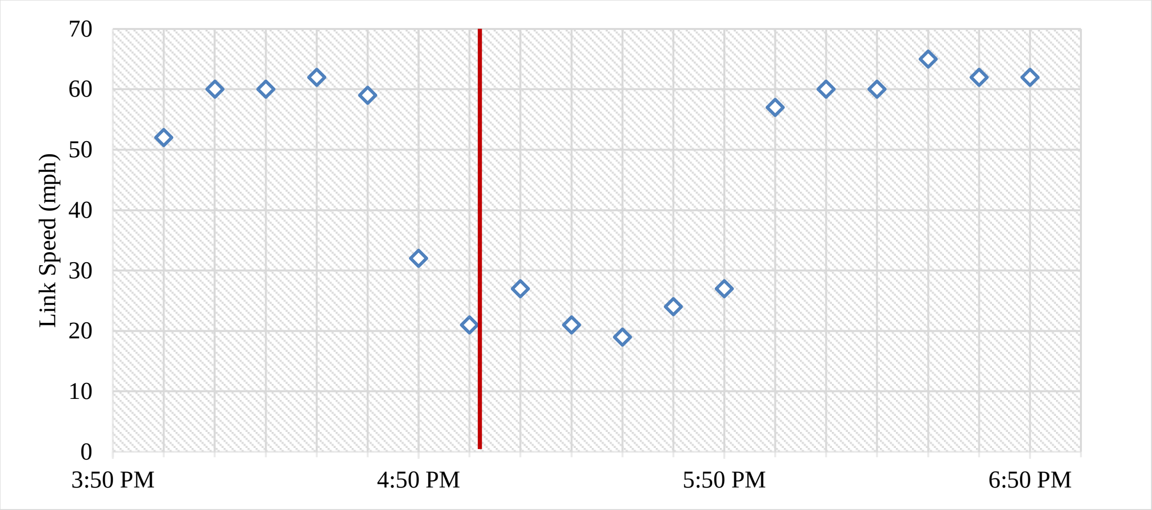

Figure 3: Congestion Speed Impact

Speed-Based Binary Probit Model

Binary probit models are a technique used in discrete choice analysis to predict outcomes when there are only two possibilities. The probability of one alternative outcome (*i*) is given as the probability that the difference between the two alternative systematic utility functions (*V*) is greater than or equal to the difference in the utility errors (ε), where the difference is errors is assumed to follow a normal distribution *(9)*.

$$P_n(i) = \Pr(\varepsilon_{jn} - \varepsilon_{in} \leq V_{in} - V_{jn})$$

[Equation 10]

The systematic utility is composed of independent variables (*x*) and coefficients (*β'*) as shown:

$$V_{in} = \beta'_1 x_1 + \beta'_2 x_2 \ldots$$

[Equation 11]

The statistical modeling software Limdep was used to construct the models in this study based on the available data.

It was decided that binary outcome models for incidents and their properties should first be created using only speed related factors and independent variables and then again with both speed and volume related factors. This was done in order to determine whether the inclusion of volume data allowed by the data warehouse could improve detection accuracy. The speed related factors attempted in the iterations of these models are listed and defined in Table 1.

Table 1: Speed-Related Variables

Variable	Definition	Mean	Min	Max
SPEED	Link speed, measured in miles per hour (mph)	44.414	6	70
TENSPEDI	Difference in speed between the current observation and the pervious observation (10 minutes prior), measured in mph	0.0504	-53	36
LOW	Binary variable, 1 if the observed speed is less than 15 mph below the speed limit, 0 otherwise	0.4872	0	1
LOWFIRST	Binary variable, 1 if the currently observed speed is less than 15 mph below the speed limit but the previously observed speed was not, 0 otherwise	0.0570	0	1
CUMLOW	Binary variable, 1 if the observed speed has been less than 15 mph below the speed limit for at least three consecutive observations, 0 otherwise	0.3857	0	1

When creating iterations of each model, a few criteria were used as general indicators of model quality. First, the p value of each included variable needed to show that the variable was significant in the model. In general practice, this means only using variables with p values less than 0.1, meaning the variable is significant at a 90% level of confidence. Also, no included variables could be highly correlated with other included variables as this leads to problems of multicollinearity. For example, multiple variables derived from raw speed data would not be included in the same model since they are likely highly correlated. Finally, the prediction outcomes were used to gauge one model iterations performance against another. High accuracy was the goal of the incident characteristic models. For the incident detection models, high detection rate was considered more important than low false alarm rate or high classification rate.

Speed-and-Volume-Based Binary Probit Model

After the completion of the speed-based models the volume data was introduced and several new variables were derived. These included both those related only to volume and those which were derived from a combination of speed and volume data. These new volume-related and speed-and-volume-related factors are listed and defined in Table 2. These new factors, along with the original speed-related factors, were used to create new models for the four prediction scenarios examined previously.

Table 2: Volume-Related and Speed-and-Volume-Related Variables

Variable	Definition	Mean	Min	Max
VOLUME	Link volume, measured in vehicles per hour (vph)	3268.479	567	7549
PAVVOL	Ratio of volume to the average volume for the corresponding link and hour of day	1.0781	0.1383	1.9439
PAVVOLO	Binary variable, 1 if the PAVVOL is greater than or equal to 1, 0 otherwise	0.6654	0	1
DENSITY	Link density, the ratio of volume to speed, measured in vehicle per mile (vpm)	89.387	8.723	784.1
TENDENDI	Difference in density between the current observation and the pervious observation (10 minutes prior), measured in mph	-0.5171	-588.1	359.1

RESULTS AND VALIDATION

Speed Threshold Detection

Several speed change thresholds were examined for their accuracy, especially in terms of incidents detected and number of false positive readings. The threshold alternatives used were 15 mph, 12 mph, 10 mph, and 8 mph drops in speed from one observation to the next (over 10 minutes). Following a comparison of all four alternatives the 10 mph threshold was the best option, based on the low percentage of missed incidents as well as relatively low number of false positive readings.

While it was considered that the threshold method could be used to predict incident attributes, the data did not show strong correlation between the magnitude of the speed change and any properties of the incident. Therefore, this method was only examined for its ability to detect incidents, not characterize them. After using the April 2014 speed and incident data to determine that the 10 mph threshold was the best, the May 2015 data was used to validate the speed threshold detection strategy's effectiveness. The validation results are shown in Table 3.

Table 3: Speed Threshold Results

Model	Measure	Speed Threshold
Incident Detection	DR	62.5%
	FAR	4.5%
	CR	92.0%

Speed Models

Using a binary probit model, four models were created using only the speed-related factors: one which detects whether and incident has occurred and three which predict different attributes of incidents. The final model chosen for each detection or characterization goal is described below in terms of its variable composition. The details of each model are shown in Table 4, in which the coefficient (top number) and p value (bottom number) of each variable in given, as well as the predictive ability of each model.

1) **Incident Detection**
   The first model created was used to determine whether an incident had occurred with the binary outcomes being no incident (0) and incident (1). After several trials with different variables and combinations the best model was found to include only TENSPEDI.
2) **Incident Differentiation**
   The next model attempted to use speed related factors to determine incident type. In the data examined, only two incident types were present: congestion (0) accidents (1). Here, the best model used both TENSPEDI and CUMLOW and independent variables.
3) **Blocked Lanes Detection**
   Another model was created to predict lane blockages associated with incidents. There were not enough data points for incidents with many lanes blocked so this model only predicted whether no lanes were blocked (0) or any number of lanes were blocked (1). As with the incident detection model, the best model incorporated only TENSPEDI as an independent variable.
4) **Severity Detection**
   The final model was used to predict incident severity level. NITTEC classifies incident severity in terms of human injuries and fatalities on a scale from 0 to 3, 0 being property damage only and 3 being most severe. However, as with the blocked lanes model, there were not enough instances of high severity incidents, so this model only predicted whether there was no injury (0) or at least minor injury (1). As with the incident detection model, the best model used TENSPEDI and CUMLOW as independent variables.

Speed and Volume Models

As with the speed models four models were created using the binary probit method. These models, however, included speed-related, volume-related, and speed-and-volume-related factors. Each of the final models chosen is described below in terms of its variable composition. The details of each model are shown in Table 4 along with the speed data only models.

1) **Incident Detection**
   In predicting whether an incident had occurred in was found that the best model used TENDENDI and PAVVOLO as independent variables.
2) **Incident Differentiation**
   The best model which was created to determine which type of incident occurred included TENSPEDI, PAVVOLO, and CUMLOW as variables.
3) **Blocked Lanes Detection**
   The Lane blockage detection model which yielded the best results included only TENDENDI as an independent factor.
4) **Severity Detection**
   Finally, the model which best predicted incident severity was found to also use only TENDENDI.

Table 4: Binary Probit Model Results

Speed Data Only Models								
	Incident Detection		Incident Differentiation		Blocked Lanes		Incident Severity	
TENSPEDI	-0.0099		0.0434		0.0381		0.0263	
	(0.0274)		(0.0114)		(0.0372)		(0.0124)	
CUMLOW			-1.4286				-1.7999	
			(0.0002)				(0.0000)	
Actual\Predicted	**0**	**1**	**0**	**1**	**0**	**1**	**0**	**1**
0	62.2%	34.7%	50.9%	14.5%	67.3%	3.6%	81.2%	11.5%
1	0.9%	2.2%	7.3%	27.3%	18.2%	10.9%	5.5%	1.8%
Speed and Volume Data Models								
	Incident Detection		Incident Differentiation		Blocked Lanes		Incident Severity	
TENSPEDI			0.09298171					
			(0.002)					
CUMLOW			-0.92452933					
			(0.0572)					
TENDENDI	0.00199598				-0.01570891		-0.00968812	
	(0.0448)				(0.0234)		(0.1008)	
PAVVOLO			-1.23184803					
			(0.0052)					
Actual\Predicted	**0**	**1**	**0**	**1**	**0**	**1**	**0**	**1**
0	86.1%	9.9%	60.0%	6.0%	64.0%	6.0%	58.0%	4.0%
1	3.1%	1.0%	14.0%	20.0%	18.0%	12.0%	24.0%	14.0%

The speed-only and speed-and-volume models were created using the speed and incident data from April 2014. The May 2015 data was used to validate these models and the results are given in Table 5. The incident detection models were validated in terms of the three common incident detection measures discussed in the literature review: detection rate, false alarm rate, and classification rate. The differentiation, blacked lanes, and severity models are examined only in terms of classification rate as a measure of accuracy.

Table 5: Validation Results

Model	Measure	Speed	Speed and Volume
Incident Detection	DR	70.4%	75.5%
	FAR	35.8%	36.5%
	CR	64.4%	63.9%
Incident Differentiation	CR	75.9%	77.6%
Blocked Lanes	CR	70.4%	69.4%
Incident Severity	CR	75.9%	77.6%

DISCUSSION

The speed threshold detection method showed promising results, especially for such a simple model. While the detection rate was not at high as other models, the false alarm rate and classification rates were very favorable. However, the low detection rate and inability to predict incident characteristics are clear disadvantages of this method.

The binary probit model which used only speed-related factors was an improvement over the speed threshold method in terms of its detection rate, while it suffered from a higher false alarm rate and lower classification rate. This method was also able to predict all three of the examined incident properties with reasonable accuracy.

The binary probit model which included factors related to both speed and volume yielded a significant improvement in detection rate of 5.1%. Smaller improvements in incident differentiation and incident severity detection were found, in addition to slight decreases in detection rate and blocked lane detection accuracy.

These results indicate that, overall, the combination of speed and volume data results in the model which is capable of detecting the greatest percentage of incidents. Without the data warehouse, these two sets of data would not be readily accessible in real time by a single agency, nor would they be in compatible formats. Therefore, the desirable performance of the combined data incident detection model highlights the importance and usefulness of a transportation data warehouse.

DATA WAREHOUSE APPLICATION 2: VALIDATION OF INCLEMENT WEATHER TRAFFIC MODELS IN BUFFALO, NEW YORK

DATA COLLECTION AND PROCESSING

Weather Comparison Data

As mentioned before, the first objective of this study was to discern whether the most recent winter was indeed harsher than previous years, and hence may present a challenge to forecasting the impact on travel conditions using the models previously developed, based on data from those previous years. To achieve that objective, weather data was collected and examined. The weather data used was obtained from the National Oceanic and Atmospheric Administration (NOAA). This data was originally collected at the Buffalo-Niagara International Airport's weather station, which is close to the section of Interstate 90 near Buffalo. Monthly weather summary data used for this comparison was collected from 2000 to 2014 and included number of days with a minimum temperature below freezing, maximum snow depth (mm), total snowfall (mm), minimum monthly temperature (tenths of °C), and average monthly temperature (tenths of °C) (NOAA, 2015).

These five factors were chosen because they each explore a different facet of winter weather. Minimum monthly temperature will capture the extreme temperature reached during the month. Alternatively average monthly temperature will examine a trend in temperature change less effected by outliers. Number of days below freezing will examine the effect of duration of cold temperatures. Total snowfall will capture the impact of all precipitation while maximum snow depth will account for total snow accumulated on the ground.

Before conducting the comparison, the data was modified to improve workability. Maximum snow depth and total snowfall were converted from millimeters to inches and temperature readings were converted from tenths of °C to °F.

Model Validation Data

For the two validations performed in this study, three sets of data were obtained. First, hourly weather data from NOAA was used, including hourly precipitation (inches), temperature (degrees Fahrenheit), wind speed (miles per hour), and weather type (an abbreviation which indicated variety and severity of weather). Weather data was collected from November 2012 to February 2014 to cover the entire period of study.

Link speed data was provided by the Niagara International Transportation Technology Coalition (NITTEC), an international organization comprised of several agencies. NITTEC acts as the region's traffic operations center and collects several types of data, including TRANSMIT speed data. This system uses the detection of E-ZPasses in vehicles to generate speed data for links, recorded at five minute intervals. While the original study used only the three links closest to the airport, this validation expanded the area to include the same seven links as the volume study, to provide more data points. Speed data was examined from December 2012 to March 2013 and from December 2013 to January 2014.

Link volume data was also provided by the New York State Thruway Authority, through NITTEC, and consisted of traffic counts for freeway links in the region. Data from the same seven links as the previous study was used in this validation. Volume data, collected at 15 minute intervals, was examined from November 2012 to March 2013 and from December 2013 to February 2014.

Before the validation could be conducted, several changes had to be made to the datasets to make them usable. First, the weather data was recorded approximately hourly, but irregularly (e.g. first at 12:53 AM, then 1:51 AM). Therefore, the time of each weather reading was rounded to the nearest hour. Next, several pieces of weather data were modified to match that used in the previous studies. The first of these changes was to convert temperature readings to a temperature index based on whether the temperature was below freezing, as shown below.

$$\text{If Temperature} > 32\ ^\circ\text{F, Temperature\_Index} = 0;$$
$$\text{If Temperature} \leq 32\ ^\circ\text{F, Temperature\_Index} = -1.$$

A system created in the previous speed study was used to convert each weather type to a numerical weather type index (Zhao et al, 2011). The system gives each variety of weather a number and also provides modifiers to account for severity. For example, the weather type index of “Heavy Rain” would be -5 (-3 for rain plus -2 for heavy). This is shown in Table 6.

Table 6: Weather Type Index Conversion

Weather Type	Description	Weather_Type_Index
RA	RAIN	-3
DZ	DRIZZLE	-1
SN	SNOW	-3
SG	SNOW GRAINS	-1
GS	SMALL HAIL AND/OR SNOW PELLETS	-3
PL	ICE PELLETS	-5
FG+	HEAVY FOG (FG & LE 0.25 MILES VISIBILITY)	-5
FG	FOG	-3
BR	MIST	-3
FZ	FREEZING	-2
HZ	HAZE	-2
BL (SN)	BLOWING (SNOW)	-3
BCFG	PATCHES FOG	-3
TS	THUNDERSTORM	-2
Modifiers	**Description**	
-	LIGHT	2
+	HEAVY	-2
"NO SIGN"	MODERATE	0

As done in the previous speed study, limited visibility was assumed to only have an effect to a certain extent. Therefore, a visibility index was used as shown below.

If Visibility > 4 miles, Visibility_Index = 5;
If Visibility ≤ 4 miles, Visibility_index = 1.25*Visibility.

The final change made to the weather data was to create a parameter for cumulative precipitation, which represented the sum of all precipitation (in inches) that had occurred during a day up until the given time.

In addition to the weather data, changes were made to the speed and volume datasets as well. From the speed data, data points which did not have a value for speed were removed, as well as any reading that did not occur on the hour. This was done to ensure that the speed data and weather data were recorded as close to one another as possible. In the volume data, volumes from multiple lanes from the same link and direction, as well as from fifteen minute intervals in the same hour, were summed to create single values for hourly volume counts.

To mimic the parameters of the speed study, normal average hourly speed (as described in the Literature Review) and a day index were used. The day index was defined as follows.

If weekday (Monday – Friday), Day_Index = 1,
If weekend (Saturday-Sunday), Day_Index = -1.

Additionally, to recreate the conditions of the original volume study baselines for each hour of the day were taken from the original study, given in Table 7 (Bartlett et al, 2012).

Table 7: Baseline Volumes

Time	Baseline	Time	Baseline
12:00 AM	588.626	12:00 PM	3635.603
1:00 AM	311.978	1:00 PM	3727.665
2:00 AM	239.227	2:00 PM	4232.384
3:00 AM	278.539	3:00 PM	5042.075
4:00 AM	454.486	4:00 PM	5352.192
5:00 AM	1155.670	5:00 PM	5135.184
6:00 AM	3019.478	6:00 PM	3921.348
7:00 AM	4978.068	7:00 PM	2864.226
8:00 AM	4569.668	8:00 PM	2468.197
9:00 AM	3587.800	9:00 PM	2226.611
10:00 AM	3333.133	10:00 PM	1558.301
11:00 AM	3512.792	11:00 PM	1078.142

Only volume data which was recorded at times when inclement weather occurred was used in the creation of the volume model, so only this data was used in the validation. In the previous study, inclement weather was defined as high wind speeds (greater than 16 mph), low visibility (less than 3 miles), and precipitation (greater than 0 inches per hour) (Bartlett et al, 2012).

METHEDOLOGY

To determine if a factor or factors change across several times or locations, a two-way analysis of variance (ANOVA) table without replication is recommended (Helsel and Hirsch, 2002). This method for comparing several independent groups involves examining the variances and errors of group means and determining the significance of a factor's effect on the group measure. This will result in a p value which indicates the factor's significance. If the p value is low (close to 0), then it can be said that the factor varies significantly between groups.

Once a factor has been determined to change across different groups, Tukey's multiple comparison test is a method which can be used to determine which groups differ in terms of that factor. This is done by calculating the least significant range (LSR) for each possible pair of groups and comparing it to a critical value. LSR values are calculated using [Equation 12.

$$LSR = q_{1-\alpha,k,N-k}\sqrt{MSE/n}$$

[Equation 12]

If the LSR for a pair exceeds the critical value, it can be said that the two groups differ with statistical significance in terms of that factor. In addition, if two groups are shown to be different, this test will determine which is greater.

Weather Data Analysis

Prior to the analysis of the recently collected speed, volume, and weather data, it was important to show that the primary motivation of this validation study is true: that the perceived increased severity of weather conditions in the most recent winter is correlated with an actual increase in number of inclement weather days and hence a truly harsher winter.

To show this, winter weather data from 2000 to 2014 was examined with the following objectives:

1. Determine if winter weather conditions varied significantly between specific years
2. If conditions vary, determine which years are significantly different from others

To determine if any of the years are different from the others a two-way ANOVA table was used. This test allowed the effect of the month data was collected in to be ignored and focused only on the year effect. A separate ANOVA table was created for each of the five weather factors and the resulting F value will be compared to a critical value from the F distribution to get a p value. If this p value is significant (below 0.10) then that weather data type can be said to vary across different years.

For any weather factors which were found to vary significantly across years, Tukey’s multiple comparison test was used to determine which years were significantly different. In this test, each difference between years was examined and if it was greater than the critical LSR value, then it was said that those two years are significantly different.

Speed and Volume Data Analysis

To perform the validation of the linear regression model created for speed in the original study, the weather data collected was put into the developed model and the predicted speeds were compared with the actual observed speeds. To assess the predictive ability of the model, the accuracy within several margins of error from the actual observed speed was calculated.

In addition, while the original study used all days, this analysis was repeated for inclement weather days only (as defined in Table 6) to see how extreme conditions affected accuracy.

The analysis of volume data was performed in the same manner as the speed data analysis, comparing the observed traffic volumes with the volumes predicted from the original study’s model. Since volume can vary greatly over time and across different links, volume prediction accuracy was assessed using the percent difference from the observed volume. The volume model was designed for inclement weather, so it was only assessed under those conditions.

RESULTS

Weather Data Comparison

The weather data from each winter between 2000 and 2014 were compared using a two-way ANOVA table to determine if different weather factors varied across years, and Tukey's multiple comparison test was used to determine which years were different.

Table 8 contains the p value for each measure of inclement weather from the two-way ANOVA tables, indicating the significance with which it varies across the years examined (with lower p values indicating greater significance). In addition, the table shows the LSR values which were used in Tukey's multiple comparison test to determine which years were different from one another with respect to each measure of inclement weather.

Table 8: Winter Weather Comparison Results

Measure of Inclement Weather	P Value	LSR
Number of Days Below Freezing	0.011	8.910
Maximum Snow Depth	0.108	11.925
Total Snow Fall	0.451	24.712
Minimum Temperature	0.001	10.701
Average Temperature	0.000	6.917

From the ANOVA tables, it was found that only number of days below freezing, minimum temperature, and average temperature vary significantly across years. Additionally, Tukey's multiple comparison test showed that the winter of 2013-2014 had significantly lower minimum and average temperatures than other years examined, especially those since 2010. These results indicated that the new dataset which included the 2013-2014 winter would have more inclement weather data than the previous studies.

This was later shown by the number of inclement weather data points found in the new dataset. While in the original volume study fewer than 300 data points were collected during inclement weather over a study period of 18 months, this new dataset collected over just 8 months contained over 500 inclement weather data points.

Speed Data Analysis

Each observed speed was matched with its corresponding predicted speed determine from the weather conditions measured at the time. In total there were over 25,000 pairs of matched speed data. The absolute differences between observed and predicted speeds were calculated and analyzed for accuracy. The percentage of differences which fell within different accuracy bounds were found, as presented in Table 9. This was done both for the entire set of data and for inclement weather data only to determine the model's predictive ability under extreme conditions.

Table 9: Accuracy of Average Operating Speed Model

Accuracy Bounds	Accuracy	
	All Data	Inclement Weather Only
± 5 mph	80.75%	39.09%
± 4 mph	51.35%	33.33%
± 3 mph	39.89%	27.31%
± 2 mph	27.93%	17.67%
± 1 mph	13.90%	10.04%

This table shows that at a five-mph accuracy level the model performs reasonably well. It can also be seen that the model is much less accurate under inclement weather conditions than overall.

Figure 4 shows a plot of predicted speed versus observed speed. Plotting all of the data points resulted in an unreadable plot and the trends were not clearly visible, mostly due to the overlap of many points. Therefore, the plot in Figure 4 was created by sorting all of the data by observed speed and averaging every 100 observed and predicted speeds. The corresponding averages were then plotted. As can be seen, with a few exceptions, especially those with quite low observed speeds (specifically less than 40 mph), the predicted values appear to be close to the observed. In addition, the plot shows that the model tends to overestimate speed when the observed speed is low and slightly underestimate it when observed speeds are high. The reason behind the significant difference between the model's predictions and the observed values during low speeds can probably be attributed to the fact that the previous years' data, upon which the model was based, lacked sufficient numbers of days with severe inclement weather. This may point to the need to recalibrate the models using the more recent data.

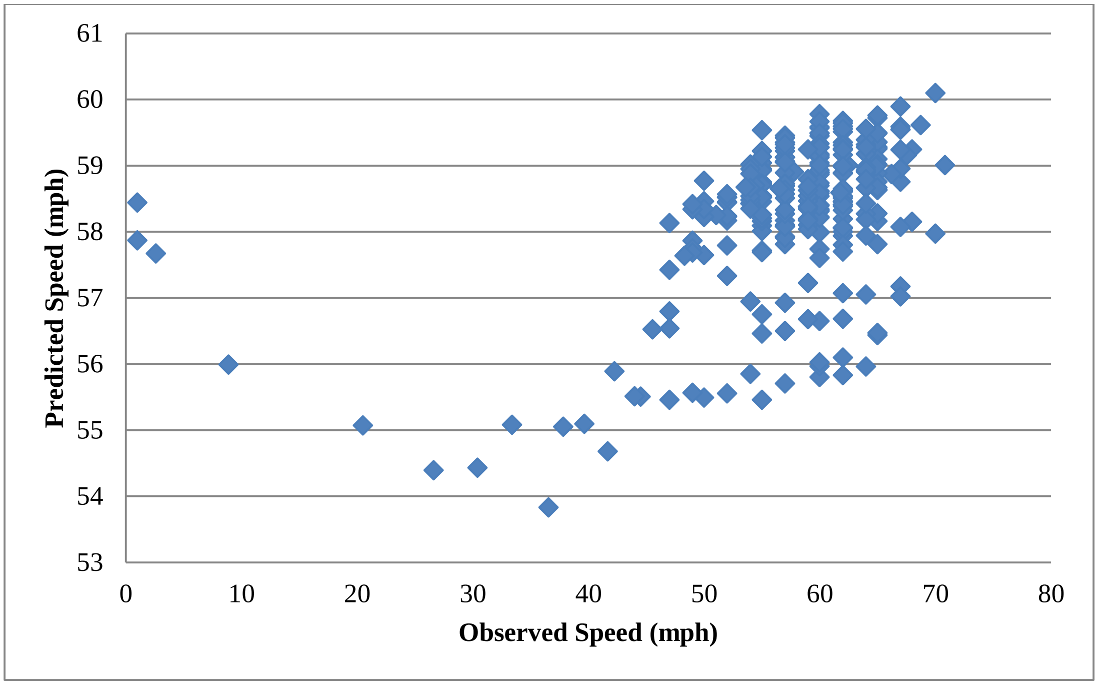

Figure 4: Predicted Speed vs. Observed Speed, Averaged for Every 100 Points

Volume Data Analysis

Accuracy of the hourly volume model was assessed with a similar method to the speed model. Each observed volume under inclement weather was matched with its accompanying predicted volume and the absolute difference between the two values was calculated. However, in this case accuracy as assessed using the percent difference between observed and predicted volumes. In addition to the absolute difference, the difference between predicted and observed volume (non-absolute) was calculated to examine whether the model tended to underestimate or overestimate volume. The results are shown in Table 10. The first three rows give the accuracy of the model within different accuracy bounds. The fourth row shows how often the model does not underestimate volume; in other words, the percentage of observed volumes that were less than the corresponding predicted volumes. Similarly, the last row gives the percentage of observed volumes that were not less than predicted volumes by more than 25%.

Table 10: Accuracy of Hourly Volume Model

Accuracy Bounds	Accuracy
± 150%	84.10%
± 100%	73.18%
± 50%	46.93%
Prediction does not underestimate volume	86.97%
Prediction does not underestimate volume by more than 25%	98.66%

The results in the table show that the volume model was not as accurate as the speed model. As can be seen, for 75% of the model's predictions, the predicted values were within 100% of the observed. The results also clearly demonstrate that model tended generally to overestimate the volume, which means that the model tended to err on the conservative side. Figure 5 is a plot of predicted hourly volume versus observed hourly volume.

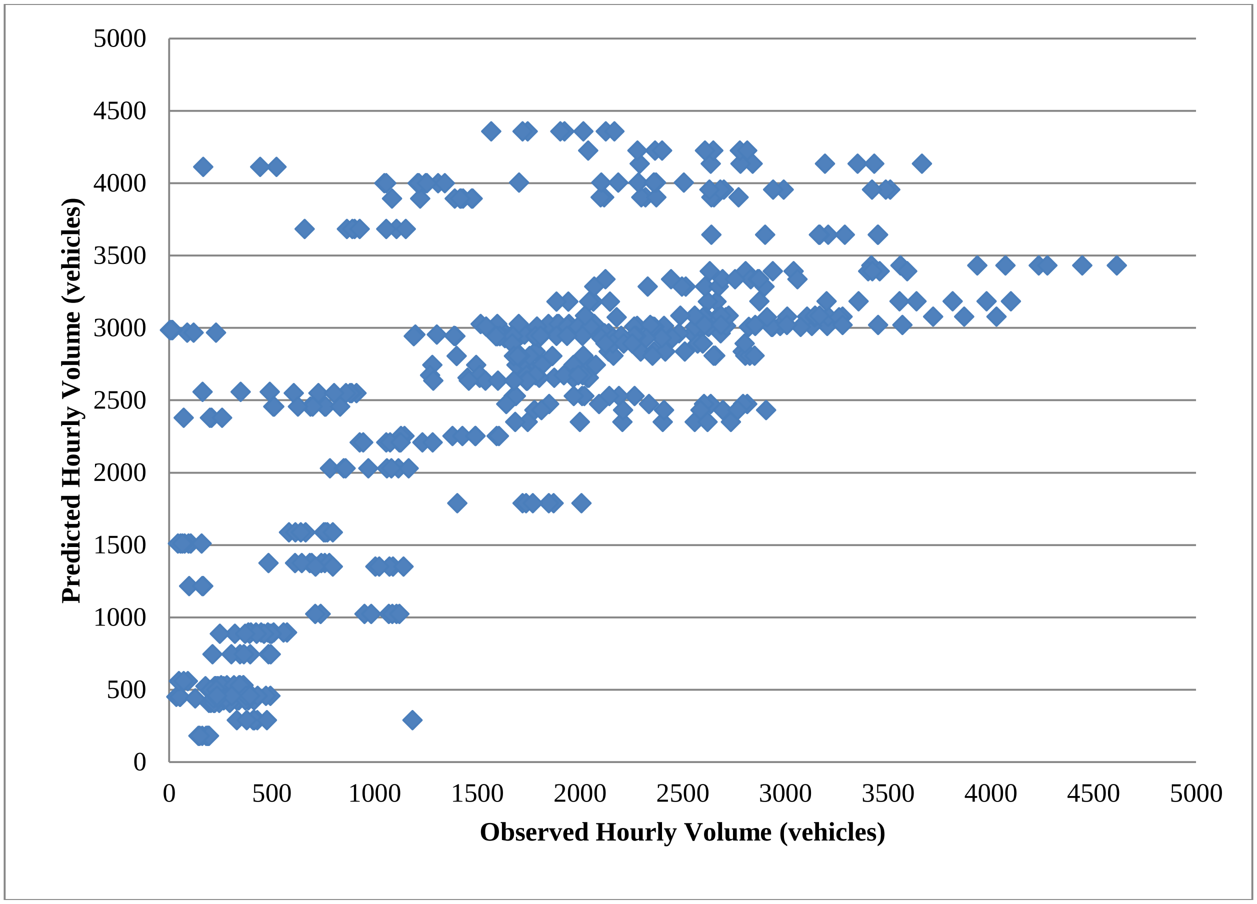

Figure 5: Predicted Hourly Volume vs. Observed Hourly Volume under Inclement Weather Conditions

CONCLUSIONS AND FUTURE RESEARCH

The prototype data warehouse developed in this research shows that a data warehouse can be constructed for transportation data in the Buffalo-Niagara region with useful results. The study then demonstrated the utility of the data warehouse using two specific case studies or application areas; the first involved developing real-time incident detection algorithms, whereas the second used the data warehouse to validate previously developed inclement weather traffic models. The specific conclusions regarding each of the two application areas considered are shown below.

Conclusions Regarding The Real-Time Incident Detection Application

- A simple speed threshold model which used a 10 minutes speed drop of 10 mph to detect incidents had a 62.5% detection rate, as well as favorable false alarm and classification rates
- A more complex binary outcome model which used only speed data detected incidents with a success rate of 70.4%, an improvement over the speed threshold model despite worse false alarm and classification rates
- The speed only model was able to predict incident type, number of blocked lanes, and incident severity with 75.9%, 70.4%, and 75.9% accuracy, respectively
- A binary outcome model which used both speed and volume data had a more impressive detection rate of 75.5% with similar false alarm and classification rates

- The combined data model was slightly better at predicting incident type and severity (both with 77.6% accuracy) but slightly worse at predicting the number of blocked lanes (with 69.4% accuracy)
- Overall, the combined data model is the best strategy for both detecting incidents and predicting their characteristics, which emphasizes the importance of a transportation data warehouse

Conclusions Regarding The Validation Of Inclement Weather Traffic Models Application

- The winter of 2013-2014 had significantly lower minimum and average temperatures than other years examined.
- The speed model performed reasonably well, usually achieving results within 5 mph of the observed speed.
- The speed model did not perform as well during inclement weather or when observed speeds were below 40 mph.
- The volume model did not perform as well as the speed model, usually achieving results within 100% of the observed volume.
- The volume model tended to overestimate volume, providing a safer estimate for volume for use in predictions.

Future Work

One future project that could stem from this work is the creation of a fully functional transportation data warehouse for the Buffalo-Niagara region. This project will allow for all current and historical transportation data in the region to be formatted and linked in a way that makes traffic monitoring and future transportation studies much more efficient.

Another project which can be explored is the deployment of a real-time incident detection system. Now that the best model for incident detection has been determined, an application could be developed which monitors real-time traffic data and detect and characterizes incidents based on the model. This application could be used by agencies like NITTEC to aid in manual detection and facilitate and improve incident detection times and rates.

In terms of the inclement weather traffic models, while the models, especially the speed model, were shown to be valid in this study, there are still improvements which could be made. First, with the new weather data it is possible that more accurate models could be developed for both volume and speed. In addition, a separate model which predicts speed under inclement weather conditions could help to improve accuracy. It could also be possible to develop and calibrate similar models to predict traffic on roadways other than the thruway.

There are several possible applications of the models considered in this study, now that they have been validated. First, the models could be used to provide travelers with more accurate travel information during inclement weather (e.g., estimated arrival times during inclement weather). In addition, they could be used by agencies or municipalities to predict the effects of storms or other weather-related events and prepare accordingly.

REFERENCES

Ahmed, F., Hawas, Y. *A Threshold-Based Real-Time Incident Detection System for Urban Traffic Networks*. Transportation research Arena, 2012.

Bartlett, A. P., Lao, W., Zhao, Y., Sadek, A. *Impact of Inclement Weather on Hourly Traffic Volumes in Buffalo, New York.* Transportation Research Board, Washington, D.C., 2012.

Ben-Akiva, M., Lerman, S. *Discrete Choice Analysis: Theory and Application to Travel Demand*. The MIT Press, 1985.

Cheung, H., Truong, N. *Detection of Incidents using Vehicle Detection Station Data*. IPENZ Transportation Group Conference, 2011.

Federal Highway Administration. *How Do Weather Events Impact Roads?* 15 Jul. 2014 <http://www.ops.fhwa.dot.gov/weather/q1_roadimpact.htm>.

Helsel, D.R. and Hirsch, R. M., 2002. Statistical Methods in Water Resources Techniques of Water Resources Investigations, Book 4, chapter A3. U.S. Geological Survey. <http://water.usgs.gov/pubs/twri/twri4a3/>.

Motamed, M., Machemehl, R. *Real Time Freeway Incident Detection*. Center for Transportation research, 2014.

National Climatic Data Center: National Oceanic and Atmospheric Administration. *Quality Controlled Local Climatological Data*. Available at:
http://cdo.ncdc.noaa.gov/qclcd/qclcddocumentation.pdf, accessed on June 18 2015.

Oracle Corporation. *Oracle 9i Data Warehousing Guide, Release 2*. Available at: http://docs.oracle.com/cd/B10500_01/server.920/a96520/concept.htm, accessed on June 2 2015.

PB Farradyne. *Traffic Incident Management Handbook*. Federal Highway Administration Office of travel Management. Available at: http://ntl.bts.gov/lib/jpodocs/rept_mis/13286.pdf, accessed on June 16 2015.

Raiyn, J., Toledo, T. *Real-Time Road Traffic Anomaly Detection*. Journal of Transportation Technologies, 2014.

U.S. Department of Commerce. NOAA National Weather Service. National Oceanic and Atmospheric Administration. 25 May. 2014 <http://www.weather.gov/>.

Wang, W., Chen, S., Qu, G. *Incident Detection Algorithm Based on Partial Least Squares Regression*. Transportation Research Part C, 2006.

Zhao, Y., Sadek, A., Fuglewicz, D. *Modeling Inclement Weather Impact on Freeway Traffic Speed at the Macroscopic and Microscopic Levels.* Transportation Research Board, Washington, D.C., 2011.