Incorporating Weather Information into Real-Time Speed Estimates: Comparison of Alternative Models

Piyushimita (Vonu) Thakuriah* Nebiyou Tilahun †

Abstract

Weather information is frequently requested by travelers. Prior literature indicates that inclement weather is one of the most important factors contributing to traffic congestion and crashes. In this paper, we propose a methodology to use real-time weather information to predict future speeds. The reason for doing so is to ultimately have the capability to disseminate weather-responsive travel time estimates to those requesting information. Using a stratified sampling technique, we select cases with different weather conditions (precipitation levels) and use a linear regression model (called the base model) and a statistical learning model (using Support Vector Machines for Regression) to predict 30-minute ahead speeds. One of the major inputs

*Associate Professor, Department of Urban Planning and Policy, University of Illinois at Chicago, 412 S. Peoria St, Suite 258, Chicago, IL 60607, USA, vonu-pt@uic.edu

†Assistant Professor, Department of Urban Planning and Policy, University of Illinois at Chicago, 412 S. Peoria St, Suite 254, Chicago, IL 60607, USA, ntilahun@uic.edu

into a weather-responsive short-term speed prediction method is weather forecasts; however, weather forecasts may themselves be inaccurate. We assess the effects of such inaccuracies by means of simulations. The predictive accuracy of the SVR models show that statistical learning methods may be useful in bringing together streaming forecasted weather data and real-time information on downstream traffic conditions to enable travelers to make informed choices.

Introduction

The effect of inclement weather is a much-researched topic in traffic operations and transportation safety. In May 2006, the U. S. Department of Transportation released the National Strategy to Reduce Congestion on Americas Transportation Network, which attributes 15 percent of all secondary causes of transportation system congestion to snow, ice, and fog. The consequent delays in travel, weather-related crashes and secondary crashes, can all accumulate to have significant negative economic and environmental impacts. A voluminous literature briefly reviewed in the next section has noted the effects of inclement weather such as snow, rain, sleet, fog, wet pavement, snowy/slushy pavement, and/or icy pavement, low visibility, wind and temperature on highway capacity and operations. Others have studied how travel demand is affected by inclement weather. Many authors have researched the effects of inclement weather on traffic safety. On average, there are over 6,442,000 vehicle crashes each year, of which more than 24 percent (approximately 1,571,500), are weather-related. Nearly 7,400 people are killed and over 690,000 people are injured in weather-related crashes each year.

Weather-related information continues to be one of the top pieces of information that travelers desire to have in making travel decisions relating to whether or not to make a trip, to change departure times or modes of travel, to take a different route because of congestion, lane closures or debris accumulation, or even to evacuate from an area in the case of flooding or other weather-related hazards. In this paper, we examine the case of using information on real-time weather conditions to predict future speeds along a heavily traversed segment of a Chicago metropolitan area interstate highway. The research approach taken here could support Location-Based Services in a variety of ways. Such information could be disseminated, pre-trip, via handheld devices or through web services; or, it could be streamed into an in-vehicle navigation device, or also to a hand-held device, for en-route decision making. Predicted future speeds along a route can presumably also be broadcast via radio or disseminated to cars by Variable Message Signs. As the quality of weather data that is available improves, for example by incorporating probe vehicle-based measurements to enhance data from other sensors (Drobot *et al.* , 2009), the precision and relevance of such applications are likely to grow.

At the same time, there have been many developments within the fields of knowledge discovery and data mining that offers opportunities for improved extraction of intelligence from vast amounts of traffic data, through advancements in statistical learning, database management, machine learning and artificial intelligence. The major objective of this paper is to explore the performance of two alternative models in making short-term speed forecasts under different weather conditions. Such an approach can ultimately be a part of a weather-responsive traveler information system. We have used an archive of probe vehicle/detector speed data and weather data for the period of a year from a heavily traversed highway segment near the center of the City of Chicago. We consider two different classes of models (a base linear model using Ordinary Least Squares and a statistical learning model, Support Vector Machines for Regression), to predict future speeds. We then compare the performance of the two models under different weather conditions and traffic levels; the experimental conditions under which the models are compared are selected using a stratified sampling strategy. One of the major inputs into a weather-responsive short-term speed prediction method is weather forecasts; however, weather forecasts may themselves be inaccurate. By means of simulations, we "degrade" the quality of observed weather measurements to proxy weather forecast inaccuracies and assess the sensitivity of the speed predictions to inaccuracies in weather forecasts. We then discuss our major findings regarding speed predictions obtained by using the two methods.

The paper is organized as follows: in the next section, we review background research on the effects of inclement weather on speeds and expand on the research questions considered. The study area, data used and primary variables are then described, followed by the stratified sampling design adopted to select the experimental conditions for testing and evaluation. Exploratory results on the relationships between congestion levels, weather conditions and traffic variables, as well as the results from the alternative models are discussed next. The sensitivity of the model predictions to inaccuracies in weather predictions are then presented followed by our conclusions.

Background and Related Literature

Inclement weather can have a range of impacts on the transportation system including increases in frequency of crashes, reductions in throughput, reduced speeds, increases in travel time unreliability, and altered demand. These impacts arise from physical effects that different weather conditions have on the infrastructure and environment (e.g. wetness, slick conditions, snow accumulation, reduced visibility) as well as its impacts on the driving behavior of travelers who may deem conditions too unsafe to follow as closely or to drive at higher speeds. Several authors have looked at the impacts of weather on particular roadways by investigating the changes in speed and volume under various weather conditions. These studies can be categorized into those which examined demand impacts (e.g. Keay & Simmonds (2005); Maze *et al.* (2006); Nookala (2006)), traffic operations (e.g., Agarwal *et al.* (2005); Chung *et al.* (2006); Dailey & Trepanier (2006); Goodwin (2002); Hranac *et al.* (2006); Mahmassani *et al.* (2009); Nookala (2006); Saberi & Bertini (2010); Tu *et al.* (2007)) and safety (e.g., Eisenberg (2004); Golob & Recker (2003)).

Many adverse weather conditions can be linked to decreased traffic performance. For instance, a study by Maze *et al.* (2006) identified the impacts of different weather conditions relative to clear conditions by intensity levels. They found reductions in speed ranging from 2-6% for rain, 4-13% for snow, and 7-12% due to reduced visibility for successively worse conditions. Ibrahim & Hall (1994) estimated a reduction of 1.9-12.9km/h for light rain, and 4.8-16.1km/h for heavy rain. Kyte *et al.* (2001) found a speed reduction of 16 km/h in snow-covered surfaces, and a 9.5km/h drop in wet surfaces. Wind speeds greater than 24km/h were found to have a drop in speeds of about 11 km/hr. The impact of weather conditions vary not only by the weather phenomenon itself, but by the time of day as well. Saberi & Bertini (2010) found significant impacts of rainfall during un-congested times in the range of 3.2-12.9km/h (2-8mph) for light rain and 6.4-16.1km/h (4-10mph) in moderate rain on the I-5 freeway, but did not find pronounced weather effects during congested periods.

A recent workshop report Federal Highway Administration (2011) described emerging analysis, modeling and simulation tools for weather-responsive traffic management. One of the conclusions was that in the future, there would need to be more specificity regarding "what is on the road versus general weather information with location-based systems telling you exactly what is happening on the route you are taking". This paper empirically evaluates the extent to which such weather-responsive traffic management tools can produce information that is accurate enough for meaningful decision-making by travelers.

Data and Study Area

The study site is a 16 km segment of I-290 in the Chicago metropolitan area (locally known as the Eisenhower Expressway) roughly between Des Plaines Ave and Western Avenue for the period from January through December 2006. Three sources of data were combined for this analysis. These were:

- Data from weather sensors at 1-hour time resolution as described below;
- An archive of detector and probe vehicle speeds at 5 minute intervals, available from a private company, NAVTEQ, LLC;
- An archive of loop detector volume and occupancy data, at 5-minute intervals, available from the Lake Michigan Interstate Gateway Alliance (LMIGA) (formerly the Gary-Chicago-Milwaukee Corridor) Information System Data Archive.

As described in an earlier paper Thakuriah *et al.* (2008), the data from six weather sensors in the Chicago metro area were linked to the highway network using link IDs and time-of-day, using a criteria of minimum distance. The final dataset for the Chicago area is over 90 GB with 327 million observations. This data is far too large to allow repeated, exploratory, analysis. The three data sources are stored in three separate SQL tables, each one indexed by any field which might be of interest. There is an additional index based on a random number, which partitions the actual speed data into 10,000 parts. The partition allows the experimental analysis of small parts of the data, selected from the larger data collection. The database system selects the information of interest, in this case typically the actual speed information, finds the matching traffic data and weather information, and presents the merged data for statistical and data mining applications. The LMIGA Data Archive on traffic measurements contains 1.8 TB of data for the period from 2004 through 2010 from different traffic detection systems. Information on incidents and construction are also available from the LMIGA archive. The LMIGA data was used primarily in preliminary analysis, to assess variability in demand on inclement weather days, and not directly used in the analysis described in this paper.

The weather data includes both continuous and categorical variables on a range of weather descriptors. Continuous measurements on precipitation levels, barometric pressure, wind speeds, wind direction, and visibility are available along with categorical descriptions of the sky, precipitation, and temperature (summarized in table 1). These conditions are reported from one of six weather stations around the Chicago area at a one hour time resolution. The precipitation descriptor, for example, includes categories such as drizzle, light rain, rain, heavy rain, as well as different types of snow, thunderstorms and icy conditions. According to the National Climatic Data Center, the weather categories report the highest codes within the hour, meaning if light rain were to be followed by heavy rain within the reporting hour, only the latter would appear in the data.

The volume and occupancy data from the road detectors were merged with the weather data based on timestamps in the two archives. Weather data timestamps increment in strictly 5 minute intervals whereas the LMIGA detector data is often stamped at between five to seven minute intervals. For the data merge, the latter were rounded to the nearest five minutes and merged to the corresponding weather data. Due to the different demand characteristics of weekend days, we limit our analysis to weekday (Monday - Friday) traffic. In addition, all dates designated as federal holidays in 2006 as well as the holiday season at the end of the year (Dec 21st, 2006 to Dec. 31st, 2006) are not included in the analysis. Because the only days in the year where heavy rain was recorded occurred on a weekend and a federal holiday, the heavy rain condition is not included in the analysis below.

Sampling Design

Impacts of weather on traffic can vary by the type of weather condition, the location where the inclement weather conditions prevail, time-of-day and other seasonal factors. The weather data under consideration here mostly includes days with no precipitation. Conditions that can be categorized as inclement weather overwhelmingly occur in the months of January through March, while other months have fewer episodes of inclement weather conditions. Thus, in order to be able to estimate the impact of a range of weather conditions, rather than randomly sampling cases, we proceed by adopting a stratified sampling strategy which ensures that a range of inclement weather conditions are represented in the sample.

The stratified sampling first considers two stratification factors: (i) detector location and (ii) absence/presence of precipitation (i.e., whether drizzle,light rain, rain, heavy rain, light snow, snow, sleet, thundershowers, thunderstorms, or strong thunderstorm conditions were present). The structure is as shown in Table 2. Three roadway sensor locations are randomly selected. Once observations were separated out by location and precipitation conditions, further stratification was done within the "with-precipitation" conditions in order to ensure that all weather conditions in the data are represented. For each condition p signifying precipitation, 70% of the observations for each weather type are randomly sampled for inclusion in the learning sample, with the remaining 30% of the observations left for inclusion in a test sample on which the models will be validated and tested. The 70% "with-precipitation" learning conditions are matched by an equivalent number of no-precipitation cases randomly selected from the "without-precipitation" group by location. This provides the learning data (the data over which the models are estimated). The remaining 30% of observations under the "with-precipitation" conditions is matched by an equivalent number of randomly selected "no-precipitation" cases and used to test the models' prediction accuracy.

The final learning sample contains 25,288 observations and the testing data has 10,867 observations. The prevalence of the different weather conditions in this sample is shown in Table 3. Each sampled record also includes 30 minute lagged observed speeds. If we consider the lagged speeds in each record to be taken at time t, the final dataset then contains speeds and weather information at time $t+\delta$ minutes, where δ is equal to 0.5 hrs. Our interest is in predicting future speeds $S_{t+\delta}$ given weather forecasts for those future time periods. In the analysis presented in the next section, we treat the weather conditions observed at time $t+\delta$ as the forecasted weather condition for that time interval and as an input into predicting $S_{t+\delta}$. Thus, although in reality weather forecasting has its own uncertainties, the modeling, at this stage, treats it as a known quantity without errors. However, the sensitivity of the predicted speeds to inaccuracies in weather forecasts are investigated in a later section, where, as described earlier, we degrade the quality of the observed weather measurement at time $t+\delta$ to reflect inaccuracies in weather forecasts. That section also discusses how predictions could be made if inaccuracies in forecasted weather conditions were present.

Analysis

Preliminary investigation of the data shows that speeds under precipitation conditions are lower than under no-precipitation conditions. Figure 1 shows the hourly average speed observed at each of the sampled locations under no-precipitation and with-precipitation conditions. Though the impact of specific precipitation conditions is masked in this figure, average speeds are consistently lower under the different weather conditions as compared to the no precipitation weather conditions, irrespective of the time of day, i.e., the underlying congestion levels. Figure 2 further shows the speed distributions under specific weather conditions. Substantial speed decreases are observed, for example, under conditions of snow, strong thunderstorms and sleet, while higher speeds are observed under conditions of no precipitation, drizzle and light snow.

Alternative Models

Two classes of models are proposed and estimated using the weather data. The first, a "base" model, uses linear multiple regression to estimate the impact of weather conditions on speeds using ordinary least squares (OLS). The second model is a statistical learning model which uses Support Vector Machines for Regression. Both models are specified the same way and incorporate components for location, day of week, time of day, prevailing speeds and 30

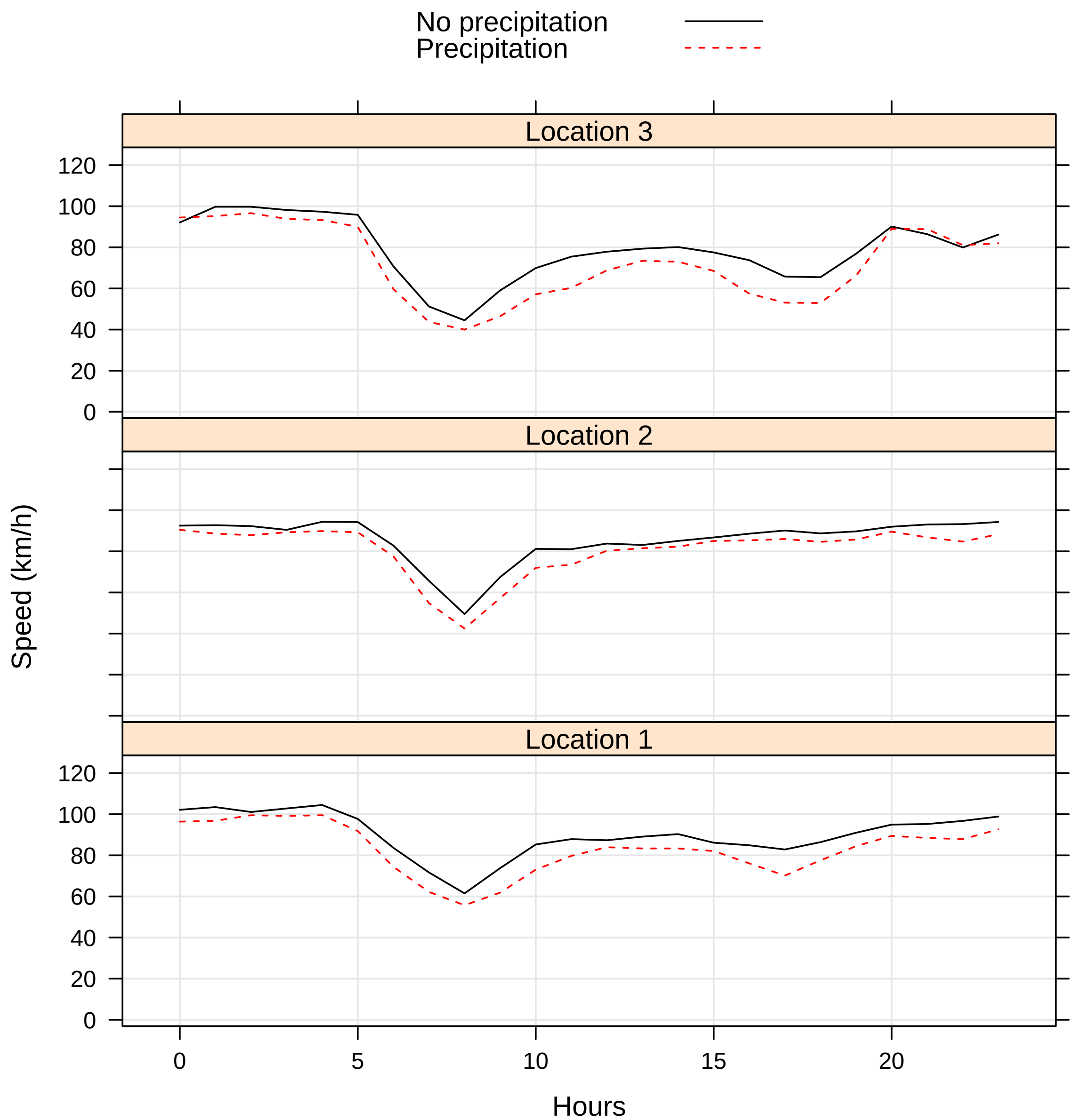

Figure 1: Average speeds by hour-of-day under good and inclement weather conditions for the sampled data

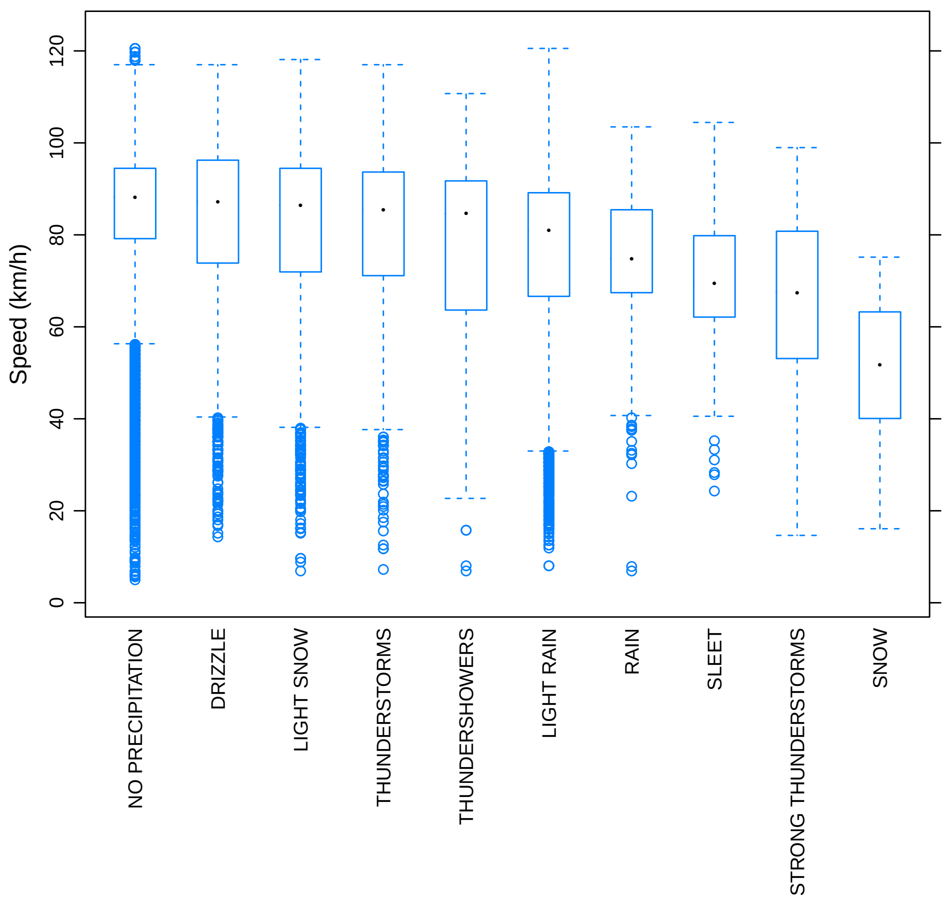

Figure 2: Box plot of speeds under different precipitation conditions for the sampled data (all locations). Top whiskers extend to the minimum of 1.5 times the interquartile range (IQR) or the largest value from the top edge of the box. Bottom whiskers extend to the maximum of the 1.5 IQR or minimum observed speed.

minute ahead weather conditions (which are treated as forecasts), and used to predict the 30 minute ahead speeds. The models estimated here can be used in a situation where roadway detector data and weather forecasts are streamed into vehicle on-board instruments or mobile connected devices that employ such models to predict travel speeds on a traveler's chosen route.

Base Model: OLS Model of Speed as a Function of Weather Factors

The Base (linear regression) model predicts 30-minute forward speeds at location ℓ based on detector/probe data at time t, time-of-day (T), day-of-week (D) and forecasted weather conditions (precipitation and presence of fog dummies) for time $t + \delta$ and is given by:

$$S_{t+\delta,l} = \alpha + \sum_i \gamma_i L_i + \sum_k \phi_k T_k + \sum_d \chi_d D_d + \beta_1 S_{t,l,y} + \sum_j \mu_j P_{j,t+\delta,l} + \zeta F_{t+\delta,l} + \epsilon_{t+\delta,l,y} \quad (1)$$

where:

$S_{t+\delta,l,y}$: Forecasted speed at time $t + \delta$ at location l

L: A location dummy (1 if i=l, 0 otherwise)

T_k: Dummy variables indicating the time-of-day category (1 if the k^{th} time interval includes $t + \delta$, 0 otherwise) - serves as proxy for demand levels; each hour is given its own dummy variable (base = hour 0).

D_d: Dummy variables for each weekday d

$S_{t,l}$: Speed at time t at location l

$P_{j,t+\delta}$: Dummy variables for the j^{th} forecasted precipitation condition at time $t+\delta$ at location l (base = no precipitation)

$F_{t+\delta}$: Dummy variable for presence of fog/ice-fog conditions at location l at time $t+\delta$ (1= presence of fog/ice-fog, 0 otherwise)

$\epsilon_{t+\delta,l}$: error term, assumed to be iid Normal (more discussion on this below)

The model as specified here takes on only main effects. Weather conditions are described by precipitation descriptors and the presence of "fog" or "ice fog". Initial specification of the model included additional weather variables such as sky conditions, temperature conditions, as well as interaction terms (e.g., precipitation $\times$ temperature, precipitation $\times$ fog), which did not substantially improve the model. Further, many interaction terms could not be estimated because they are either not observed or unlikely to happen together.

In the base model, we are making a strong assumption that the error terms are independently distributed. However, the time-dependent nature of the data, as well as the repeated speed measures taken at the same location, time of day, etc. makes the errors correlated. Some of these is mitigated through the sampling structure where consecutive time segments are less likely to be part of the data. We also include location, time of day, and day of week dummy variables in the analysis to control for the mean effects of these factors. We make the simplified assumption about errors primarily because we are not using the model to make inferences but rather as a straightforward prediction tool. For comparison purposes, a model without the weather variables, as well as one without lagged speed are also estimated and are reported. When reporting the models, we omit discussions on standard-errors, but provide the parameter estimates for comparison purposes with the speed reduction values from other studies.

The estimated models are given in Table 4. Model $m1$ includes lagged speed and weather variables, $m2$ includes lagged speeds but omits weather variables, and $m3$ includes weather but omits the lagged speed variable. In terms of level of fit model 1 has the highest R^2, followed closely by $m2$ and finally $m3$. In $m1$, the inclusion of the lagged speed among the independent variables creates some collinearity with the weather variables since buried within the lagged speeds is information about the prevailing weather conditions. Due to this, the mean impact of a given weather condition should be gleaned from estimates of $m3$ where the effect of weather can be interpreted conditional on other factors remaining constant.

The models illustrate that there are location to location, day-of-week, as well as time-of-day differences in observed speeds. These estimates are left out of Table 4 because they are specific to the locations studied. In brief, what they show is that speeds during the morning rush hour (7:00-10:00am) are significantly lower than at any other time during the day. These are followed by the time intervals at the beginning and end of this time interval (6:00-7:00am, 10:00-11:00am) and the afternoon rush hour (3:00pm-7:00pm) the morning rush hour.

The effect of weather conditions are noticeable in the model. We estimate that sleet conditions have the highest impact, on average reducing the expected speed by 19.3km/h (12mph). This is followed by snow and rain, each of which have an impact of a reduction of 12.1km/h (7.5mph) and 10km/h (6.2mph) all other factors staying the same. The Root Mean Square Error (RMSE) of the models are 10.3km/h (6.4mph), 10.5km/h (6.5mph), and 12.9km/h (8mph) respectively for models $m1$, $m2$ and $m3$.

Model Based on Support Vector Machines for Regression

Support vector machines for regression (SVR) allow us to estimate a model using the same data but with a different loss function from what is used in least square model. Models specified in the same way as the linear regression models are estimated using SVR. While the loss function for OLS penalizes all errors and large errors are penalized even more because squared errors are taken, the loss function for SVR penalizes only errors that are greater than a distance ϵ ignoring errors that are smaller. In addition, the loss function is linear for those errors that exceed ϵ. In estimating the SVR plane then, one is choosing a plane which ideally incorporates as many points as possible within the ϵ boundary while also minimizing those errors that are greater than ϵ. These distances greater than ϵ are measured by slack variables (z^+,z^-) defined on either side of this plane.

Suppose the estimated plane is of the form $f(\mathbf{x}) = \mathbf{w}.\mathbf{x} + \mathbf{b}$ and define a margin ϵ within which the model is insensitive to prediction errors. On either side of the plane, slack variable z^+ and z^- are defined for when the observed point lies outside the ϵ margin around the plane. Given a training set $\{(x_1, y1), (x_2, y_2), ..., (x_n, y_n)\}$, the support vector regression finds the plane $\mathbf{w.x} + \mathbf{b}$ that satisfies the following condition:

$$\min_{w,b,z^+,z^-} \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{n}(z_i^+ + z_i^-) \tag{2}$$

$$such\ that$$

$$y_i - f(x) - z_i^+ \leq \epsilon$$

$$f(x) - y_i - z_i^- \leq \epsilon$$

$$z_i^+, z_i^- \geq 0 \quad i = 1, ..., n$$

The norm of the estimated plane ($||\mathbf{w}||$) determines how flat the estimated plane is. The variable C is a cost that is specified by the analyst and trades off the plane's complexity with the extent to which the slack measures are tolerated (see Basak *et al.* (2007) for more discussions). An additional advantage of SVR is that the data can be mapped to a different feature space by employing a kernel function. The idea is to implicitly map the data to a higher dimensional space where the regression is performed and then map that higher dimensional regression back to the original space. Among common kernel functions used for this process are the radial basis function and the linear kernel.

In fitting the model, the independent variables used in the input space for the SVR model are kept the same as what is in the base linear model. Model estimation was performed in R (R Development Core Team, 2009) using the SVM package e1071(Dimitriadou *et al.* , 2009). When using SVR, the kernel type and parameters related to the kernel function, as well as the cost value (C), and the size of the margin (ϵ) need to be specified. Here we use the Radial Basis Function (RBF) which performs a non-linear mapping of the data. The model was first estimated using the default parameters for cost, ϵ and kernel parameter γ ($\epsilon = 0.1$, γ= 1, C =1). The final model has an $\epsilon = 0.5$ and the tuned parameters have values of $\gamma = 0.5$ and $C = 1$. The model achieves a RMSE of 9.5km/h (5.9 mph) which is less than the 10.3km/h (6.4mph) achieved by employing the linear model $m1$. An SVR model specified in the same way as model $m3$ without the lag speeds achieves a RMSE of 12.1km/h (7.5mph) as compared to the 13.2km/h (8.2mph) for $m3$. Further comparisons between the performance of the Base and SVR models, and models with and without a lagged speed variable, are performed using the test data and discussed in the next section.

Comparison of Results

To compare the performance of the base and SVR models, each is employed to predict speeds using the test data that was prepared (discussed earlier in the sampling stage). Base Model $m1$ (with the weather variables) and the similarly specified SVR model (SVR $m1$) are used for these comparisons. The testing data contains the remaining 30% of cases under each precipitation condition and an equivalent number of randomly sampled no-precipitation observations. None of these observations were used in the estimation of the models. We assess performance by:

1. comparing RMSE of SVR $m1$ and Base $m1$ for all observations and under different precipitation conditions;

2. comparing speed prediction error distributions under SVR $m1$ and Base $m1$ under Sleet, Strong Thunderstorms and Rain, which were estimated earlier to lead to greatest reductions in speeds;

3. determining whether model accuracies vary by time-of-day, as a proxy for congestion levels.

Comparisons on RMSE: On the basis of aggregate measures using the test data, the two models perform relatively similarly. The RMSE of the predictions using the test data set is 9.4km/h (5.82mph) for Base $m1$ and 8.7km/h (5.43 mph) for SVR $m1$. Table 5 shows the RMSE of predicted speeds using SVR $m1$ and Base $m1$ (in columns 2 and 4 respectively). The table also shows the percentage improvement in the RMSE by including weather variables in the two $m1$ models over the baseline model of $m2$ where weather variables are not included (percent reduction in RMSE for SVR $m1$ over SVR $m2$ RMSE is in column 3 and the equivalent comparison for the Base Models $m1$ and $m2$ is in column 5), under each of the precipitation conditions. The last column of Table 5 gives the percent improvement in the RMSE of SVR $m1$ over Base $m1$ for each weather condition.

Overall, Table 5 shows that the gain in predictive accuracy is not the same under all precipitation conditions. For the SVR, predictive accuracy increases with the inclusion of weather variables for all precipitation conditions, with the smallest gains under no precipitation conditions. By including weather variables, the largest gains in predictive accuracy of the SVR model accrues under snow, sleet, rain, thunderstorms and strong thunderstorms. On the other hand, for the Base Model, the greatest gains from including weather variables are under sleet, strong thunderstorms and rain. It should be noted that for the Base Model, predictive accuracy using RMSE does not improve by using weather information for some weather conditions including for light snow, snow and thundershowers.

The table also shows that the predictive accuracy of the SVR is higher than that of the Base Model in all precipitation conditions, except when there is no precipitation, although the difference is very small, of 0.05km/h. The highest gains by using the SVR $m1$ model compared to the Base $m1$ model is during rain, followed by snow, thunderstorms, light snow and sleet.

Comparisons of speed prediction error distribution: Table 6 shows the quartiles of the residual speeds $e_{t+\delta,\ell} = S_{t+\delta,\ell} - \widehat{S_{t+\delta,\ell}}$ where $\widehat{S_{t+\delta,\ell}}$ is the predicted speed value for forecast period $t+\delta$ at location ℓ. The performance of SVR $m1$ and Base $m1$ are shown for sleet, strong thunderstorms and rain. Under sleet conditions, the residual speed distribution is narrower under the SVR $m1$, compared to Base $m1$. The largest absolute value of residual speed is 25.52km/h (15.86mph) under SVR, compared to 26.7km/h (16.6mph) under Base. Additionally, the median of the SVR residual distribution is closer to 0 than for the Base distribution median. A similar pattern is seen under rain; however, the median for the Base residual distribution is closer to 0 than for the SVR residual distribution.

Under strong thunderstorm conditions, once again the residual error distribution is narrower under SVR than under Base. The highest absolute value of residual speed is under the Base model, with a deviation of 38.5km/h (23.9mph) difference from the observed. However, the median point is close to 0 for the Base model compared to SVR, although the difference is only about 1.05km/h (0.65 mph).

Comparisons under different demand (time-of-day) conditions: The third comparison looks at how the prediction errors from the different models compare under different demand and precipitation conditions. Time of day is used as a proxy for demand. The RMSE under the different demand and precipitation conditions are summarized in table 7. Columns 3 and 4 present the RMSE for SVR $m1$ and the percentage reduction it achieves over the SVR $m2$ which doesn't include weather variables. Columns 5 and 6 present the same information for the Base $m1$ and $m2$ models. Column 7 compares the RMSE improvements of the SVR $m1$ over the Base $m1$. The SVR $m1$ model which incorporates weather variables consistently predicts with lower RMSE as compared to SVR $m2$ under all time of day/precipitation conditions considered. Larger improvements are especially observed for SVR $m1$ under precipitation conditions than the no-precipitation condition. Similarly the Base $m1$ with weather variables outperforms the Base $m2$ across all demand and precipitation conditions. However, the percentage differences in this case are overall moderate as compared to the comparison between the SVR models.

The last column of table 7 shows the percentage improvement of the SVR $m1$ over the Base $m1$. Here, results are mixed under the no precipitation case. While the Base $m1$ does better for the 9:00-14:59 time range and the evening peak, the SVR $m1$ performed better in the morning peak. Under precipitation conditions, however, the SVR $m1$ does significantly better than the Base $m1$, achieving RMSE reductions of approximately 6.9% for the morning peak and above 10% reductions in the evening peak and overnight hours.

These comparisons show that in the majority of cases considered here, the models which account for weather conditions have lower RMSE as compared to those without weather variables. This is especially true for the SVR model, where the root mean squared errors are lower for all cases whether they are separated out by precipitation conditions alone, or precipitation and demand conditions. In some cases, these improvements are greater than a 20% reduction in RMSE under adverse weather conditions (Sleet and Snow). Secondly, the SVR $m1$ model, except in the case of the No Precipitation condition in Table 5, consistently leads to lower RMSE as compared to those achieved through the Base $m1$ model, with some improvements again exceeding 20% (Rain, Snow) and others in the teens and high teens (Sleet, Light snow, Thunderstorms). In addition, when the effect of demand is considered, the SVM model leads to lower average errors under all the with-precipitation cases. These observations suggest 1) that the SVR $m1$ model overall achieves better outcomes over SVR $m2$ model which doesn't include weather variables, 2) that the SVR $m1$ model, though not always, mostly performs better than the Base $m1$ model especially under adverse weather conditions, and 3) that the inclusion of weather variables in $m1$ leads to better predictions in cases where adverse weather is experienced (Rain, Sleet, Snow, Thunderstorms).

Sensitivity to Weather Forecast Inaccuracies

Up to now, we have treated precipitation conditions at forecast period $t + \delta$ as being known with certainty. However, since weather forecasts have uncertainties, this means that the reliability of the speed predictions will depend on the quality of these forecasts. An evaluation of how such uncertainties affect speed predictions is therefore essential. In addition, a way to handle speed predictions when estimates of weather forecast uncertainty is available is also desirable.

Evaluating sensitivity to weather variables requires us simulate conditions in which the weather variables are degraded while all other inputs to the model remain the same. One way to do this is to sample episodes of different weather conditions from the existing data, systematically change the forecasted precipitation variable to reflect potential forecast errors, and look at what happens to the speed predictions from these models. Predictions from these models can then be compared to the actual experienced speed as well as to speeds predicted using the measured weather conditions at time $t + \delta$.

At least two questions need consideration in using the existing data for such an application. The first is whether it is reasonable to assume that the forecasted weather conditions can be different while all other independent variables remain at the observed levels in the data. The second is what types of weather forecast errors are likely to occur. For the first question, clearly location, time of day, and day of week do not pose problems if kept unchanged while forecasts are changed. However, observed speeds at time t are likely highly correlated with prevailing weather conditions which themselves are likely correlated with the type of forecast that is made. If our presumption is that weather forecasts may be wrong, then prevailing conditions at time t (the time at which prediction is being made) are likely different enough from the circumstances in the observed data. This suggests that taking the speeds as observed may not be reasonable. Were these (or other similar) models to be used in real applications where weather forecasts are provided, the speeds at time t would be observed at the prevailing weather conditions. Our options then are either to estimate what observed speeds would have been and employ the models with lag speeds or employ model $m3$ and its SVR equivalent where lagged speed does not play a role. We have employed this latter method for its simplicity.

The second question relates to how to degrade the weather conditions. We follow two strategies:

- Scenario 1: We assume that each weather type in the next time period has an equal chance of occurrence. This would be the same as saying a given forecast was a random draw from the list of potential weather types. Generating the weather probabilities for the first case is straight forward. Since we have ten precipitation conditions, each is assigned a probability of 0.1 of being reported as a forecast. While this would allow us to evaluate model predictions under any wrong forecast (because every other weather condition is equally likely), such a forecast is however unlikely in reality.

- Scenario 2: A slightly different (and perhaps better) estimate is to use the progression of weather events observed from one time period to the next and assume that forecast errors share some similarities with these progression. For this scenario, we derive probabilities by looking at weather conditions at time t and weather conditions at time $t + 0.5hrs$ and derive probabilities based on the frequency of instances where one weather condition leads into another. For each actual observed condition i at time t, the probability that a forecast of j is mistakenly made is estimated by the proportion of times that j occurs a half hour after i. This is simply calculated by counting the instances in which i occurs at time t and j occurs at time $t+\delta$ (call it c_1), the instances where i occurs at time t without being followed by j at $t+\delta$ (call it c_2). For each (i,j) pair then the probability that j is forecasted when i actually occurs is $p_{ij} = c_1/(c_1+c_2)$. Since each weather condition is most often followed by itself, this has the advantage that the most likely prediction in each case is the correct one (when $i = j$).

Once these probabilities are calculated, to test sensitivity to weather forecast errors, 10 data points observed under each of the precipitation conditions in the data are randomly sampled. Each of these are replicated 1024 times to generate a data set each with 10240 data points. For each observed weather condition, forecasted weather is then simulated based on the probabilities described above to generate two data sets (one based on equal probabilities and another on weather progression). Each of these datasets is then used in model $m3$ and its SVR equivalent to predict speeds.

The resulting speed predictions using the data from scenario 1 are as shown in figure 3. Each box in the figure is labeled with the observed weather condition in the original data. The flat black line shows the observed speeds. The Base and SVR model predictions under each simulated weather condition (labeled on the horizontal axis) is also shown. In this illustrative case, the largest deviations of the predicted speeds occur when erroneous sleet, snow, and rain condition forecasts are made for most weather conditions. In addition, when weather forecasts miss a sleet condition, model predictions considerably overestimate speeds.

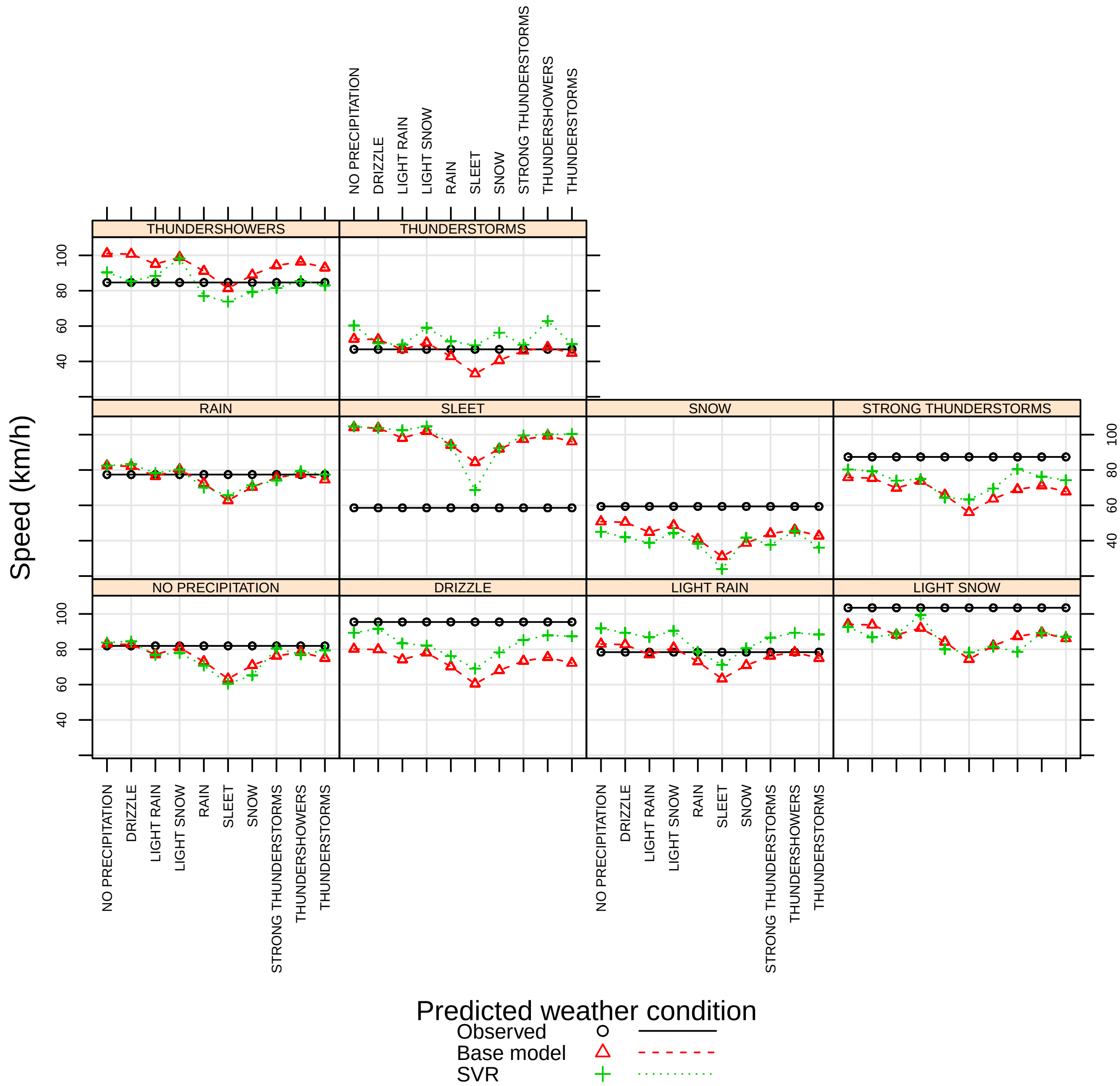

Figure 3: Model sensitivity to weather forecasts - each box is labeled by the actual weather condition that was observed

The data with the degraded weather conditions is also useful in illustrating how the speed prediction models may be used when weather forecasts are provided along with probability of occurrence for each weather type. Such probabilities can be used to generate simulated weather conditions in a similar fashion as we have done in scenario 2. The resulting speed predictions from applying the models to these data provide a range of speed estimates and a measures of variance. Depending on goals (e.g. reducing the probability of delay, reporting most likely outcome etc.) the speed reported can be set to be a certain percentile of the range of speeds forecasted. To illustrate, we use the datasets generated under scenario 1 and 2 above. Each observation has weather conditions that are simulated. The mean speed and standard deviation using each of these simulated conditions under scenario 1 and 2 is as shown in Table 8. As expected, there is more variability in the estimates when forecasts are treated as completely random (scenario 1). Narrower estimates are found when the progression probabilities are used (scenario 2). Within each observed weather condition, all variables in the model have remained the same except the precipitation conditions. Mean speed estimates from the SVR model with weather transition probabilities are on average closer to the observed speeds across all weather conditions (RMSE=12.2km/h (7.6mph) across the ten cases). This is followed by the Base model with weather transition probabilities (RMSE= 15.9km/h (9.9mph)). The speed predictions when using the data from scenario 1 have larger errors since these incorporate weather forecast errors that are less likely to be made (with RMSE on the order of 16-18km/h).

This analysis suggests the following: First, in the presence of significant uncertainty in weather forecast, using only one condition as the forecast variable can lead to large errors in predicted speed. There are however two things that can be done to mitigate such errors. One is that, weather forecast uncertainties, when available, can be used in the manner demonstrated along with these models to provide estimates that are on average close to the eventual prevailing conditions. The approach can also be extended to other variables that may have their own uncertainties. The output distribution of speeds also allows the user of these models to select appropriate estimates of speed based on different goals. Secondly, if lagged speeds were to be used in these models, those can serve a correcting role in the model for wrongly forecasted weather condition by injecting information about prevailing roadway conditions.

Summary and Conclusions

The use of weather information in traveler information systems have proliferated in recent years. Technologies ranging from connected mobile phones and PDAs to vehicle on-board instruments that have the capacity to receive streaming data, perform calculations, and report forecasts of travel conditions are being widely adopted. This paper estimates two types of models that could be used in such instruments to predict travel speeds while incorporating forecasted weather conditions. The models estimated employ a linear regression model (Base model) and a statistical learning model using Support Vector Machines for Regression (SVR). Each type of model was estimated with alternative specifications that included or left out forecasted weather conditions. The different models performance is then compared using a test data set not used in the estimation of the models.

The weather and traffic data for this study are from the Chicago area from 2006. Weekends and holidays are excluded from the analysis. A stratified sampling strategy is adopted to ensure that both the model estimation (learning) data and the testing data incorporated different weather conditions observed throughout the year. The estimated models controlled for location, time of day, day of week factors as well as incorporated lagged speed estimates from a previous time step to predict future speeds.

The performance of the linear and SVR models were compared using the Root Mean Square Error (RMSE) that each achieves using the test data set. Comparisons were based on: 1) overall RMSE, 2) RMSE under specific weather events, 3) RMSE under different demand and precipitation conditions, and 4) the distribution of errors under the most adverse weather conditions. The results of the comparison show that:

- the SVR model which accounts for weather conditions (SVR $m1$) has lower RMSE as compared to SVR model without weather variables (SVR $m2$) under any weather condition;
- the Base model which accounts for weather conditions has lower RMSE as compared to the Base model without weather variables in a majority of cases;
- the SVR $m1$ model consistently leads to lower RMSE as compared to those achieved through the Base $m1$ model in any weather condition except the No-Precipitation

condition;

- when the effect of demand is considered, the SVR model leads to lower average errors under all the with-precipitation cases whereas in the no-precipitation condition, the results are mixed;
- the range of errors under the most adverse weather conditions (Sleet, Strong thunderstorms, Rain) for the SVR model are either smaller or comparable to that from the Base model.

Overall, these observations suggest that the SVR with weather variables (SVR $m1$) achieves the best predictions among the different models considered and under most circumstances.

The estimated models were also used to analyze how uncertainties in weather forecasts may affect prediction quality. This was done by selecting a subset of the data under each observed precipitation condition and creating two new data sets where the weather forecasts were degraded using probabilities based on inclement weather progression and randomly, respectively. The new simulated data is then used to investigate under which conditions large prediction errors occur. The findings suggest that the largest deviations of the predicted speeds occur when erroneous sleet, snow, and rain condition forecasts are made for most weather conditions. In addition, when weather forecasts miss a sleet condition, model predictions can considerably overestimate speeds.

The same simulated data is also used to illustrate that when weather forecast uncertainties are known, they can be used to estimate the possible range of speeds that could occur. These results can be aggregated into expected speed values and reported. Speeds aggregated in this manner mitigate against the possibility that large prediction errors arise as a result of a wrong weather forecast. In addition, they allow the reporting of a range of speeds, where conservative and aggressive estimates may be taken as desired by the user.

Overall, the paper shows that consideration of weather conditions leads to improvements in the RMSE of prediction, that the SVM model outperforms the Base model in making predictions under most conditions, and that when uncertainties in forecast weather conditions are known, that these can be incorporated in the results of the prediction. Similar models could be estimated and used at different time resolutions to serve different purposes. For example, forecasts 5, 10 or 15 minutes into the future could be of great importance for those travelers already en-route, whereas hourly or longer forecasts can help in longer-term trip planning. One of the limitations in using the current data has been the hourly resolution of the weather data. Future efforts will employ weather data at smaller time resolutions. The predictive accuracy of the models on the test dataset used here shows promise that SVR models could be used to bring together streaming forecasted weather data and traffic conditions to inform travelers.

References

Agarwal, M., Maze, T., & Souleyrette, R. 2005. Impacts of weather on urban freeway traffic flow characteristics and facility capacity. *In: Proceedings of the 2005 mid-continent transportation research symposium, ames, iowa.*

Basak, D., Pal, S., & Patranabis, D.C. 2007. Support vector regression. *Neural information processing–letters and reviews*, **11**(10), 230–224.

Chung, E., Ohtani, O., Warita, H., Kuwahara, M., & Morita, H. 2006. Does weather affect highway capacity. *In: 5th international symposium on highway capacity and quality of service.*

Dailey, D.J., & Trepanier, T. 2006. *The Use of Weather Data to Predict Non-recurring Traffic Congestion.* Tech. rept. Technical report to Washington State Transportation Commission, Washington State Department of Transportation, University of Washington TransNow, and Federal Highway Administration.

Dimitriadou, Evgenia, Hornik, Kurt, Leisch, Friedrich, Meyer, David, , & Weingessel, Andreas. 2009. *e1071: Misc functions of the department of statistics (e1071), tu wien.* R package version 1.5-22.

Drobot, S., Mahoney III, W.P., Schuler, E., Wiener, G., Chapman, M., Pisano, P.A., Kennedy, P., McKeever, B.B., Stern, A.D., & Noblis, I. 2009. IntelliDrive road weather research & development-the vehicle data translator. *Proceedings of the 2009 intelligent transportation systems world congress.*

Eisenberg, D. 2004. The mixed effects of precipitation on traffic crashes. *Accident analysis & prevention*, **36**(4), 637–647.

Federal Highway Administration. 2011. *Developments in weather responsive traffic management strategies*. Summary Report: 1st National Workshop and Stakeholder Meeting on Weather Responsive Traffic Management (WRTM), Portland, OR, Oct. 6-7. Federal Highway Administration.

Golob, T, & Recker, W. 2003. Relationships among urban freeway accidents, traffic flow, weather, and lighting conditions. *Journal of transportation engineering*, Jan.

Goodwin, L. 2002. Weather impacts on arterial traffic flow. *Mitretek systems inc.*, Jan.

Hranac, R., Sterzin, E., Krechmer, D., Rakha, H., & Farzaneh, M. 2006. *Empirical studies on traffic flow in inclement weather*. Tech. rept. FHWA-HOP-07-073. US DOT Federal Highway Administration.

Ibrahim, A, & Hall, F. 1994. Effect of adverse weather conditions on speed-flow-occupancy relationships. *Transportation research board*, Jan.

Keay, K., & Simmonds, I. 2005. The association of rainfall and other weather variables with road traffic volume in Melbourne, Australia. *Accident analysis & prevention*, **37**(1), 109–124.

Kyte, M, Khatib, Z, Shannon, P, & Kitchener, F. 2001. Effect of weather on free-flow speed. *Transportation research record: Journal of the transportation research board*, **1776**(-1), 60–68.

Mahmassani, H S, J, Dong, J, Kim, B, Chen R, & B, Park. 2009. *Incorporating weather impacts in traffic estimation and prediction systems*. Tech. rept. FHWA-JPO-09-065. Research and Technology Administration, ITS Joint Program Office.

Maze, TH, Agarwai, M, & Burchett, G. 2006. Whether weather matters to traffic demand, traffic safety, and traffic operations and flow. *Transportation research record: Journal of the transportation research board*, **1948**(-1), 170–176.

Nookala, L.S. 2006. *Weather Impact on Traffic Conditions and Travel Time Prediction*. M.Phil. thesis, University of Minnesota.

R Development Core Team. 2009. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Saberi, M., & Bertini, R.L. 2010. Empirical Analysis of the Effects of Rain on Measured Freeway Traffic Parameters. *In: 89th annual meeting of the transportation research board*.

Thakuriah, P., Yanos, G., Lin, J., Metaxatos, P., Sulo, R., Pu, W., & Mbekeani, L. 2008. Including weather effects in real-time traffic information environment: Exploratory analysis and alternative models. *In: Paper presented at the 15th intelligent transportation systems world congress*.

Tu, H., Van Lint, JWC, & Van Zuylen, HJ. 2007. The impact of adverse weather on travel time variability of freeway corridors. *In: 86th annual meeting of the transportation research board papers cd-rom*.

United States. Dept. of Transportation. 2006. *National strategy to reduce congestion on* *america's transportation network*. United States Department of Transportation.

Table 1: Available Weather Data Summary

Weather descriptor	Units	Number of categories
Sky condition	Categorical	6
Sky descriptor (includes fog/ice-fog conditions)	Categorical	34
Precipitation descriptor	Categorical	77
Temperature descriptor	Categorical	12
Temperature	Celsius	
Wind speed	km/hr	
Wind direction	Degrees	
Humidity	Percents	
Sea level pressure	millibars	
Visibility	km	

Table 2: Sampling Design for location l and precipitation condition j

Use Condition	Precipitation condition j	
	With*	Without
Learning	Randomly sample 70% within condition j	Randomly sample equivalent number of cases
Testing	Keep remaining 30% within condition j	Randomly sample equivalent number of cases

*Precipitation conditions considered are Drizzle, Light Rain, Rain, Heavy Rain, Light Snow, Snow, Sleet, Thundershowers, Thunderstorms, and Strong Thunderstorms

Table 3: Prevalence of different weather conditions in sampled data

Category	Subcategories	Frequency
	No precipitation	50%
	Drizzle	5.7%
	Light rain	29.1%
Precipitation conditions	Light snow	8.7%
	Rain	1.1%
	Sleet	0.9%
	Snow	0.2%
	Strong thunderstorms	0.8%
	Thundershowers	1.5%
	Thunderstorms	2.1%
Foggy conditions	Fog/Ice fog	24.5%
Sample size		25288

Table 4: Base Model: Linear regression model incorporating current traffic information and real-time weather information (speed in km/h)

Variable	Description	$m1$ (All)	$m2$ (without weather)	$m3$ (without speed)
α	Intercept	39.8^{+}	35.08^{+}	105.08^{+}
S_t	Speed (lagged)	0.62^{+}	0.65^{+}	
W	drizzle	0.03		-0.32^{+}
	Light rain	-2.54^{+}		-5.97^{+}
	Light snow	-0.40		-2.14^{+}
	Rain	-4.03^{+}		-9.94^{+}
	Sleet	-8.6^{+}		-19.72^{+}
	Snow	0.44		-12.1^{+}
	Strong thunderstorms	-4.51^{+}		-6.76^{+}
	Thundershowers	$-1.23^{\cdot}$		-4.7^{+}
	Thunderstorms	-2.73^{+}		-8.02^{+}
F	Fog/Ice fog	-1.42^{+}		-3.64^{+}
L	Location factors	significant effects estimated (not presented)		
D	Dow factors	significant effects estimated (not presented)		
T	Time of day factors	significant effects estimated (not presented)		
Residual S.E.		10.32	10.44	13.19
Multiple R^2		0.713	0.706	0.530
F-stat $m1$		1564 on 40 and 25247 DF		
F-stat $m2$		2017 on 11 and 25257 DF		
F-statistic $m3$		730.9 on 39 and 25248 DF		

$^{+}$ pval < 0.01, $^{\cdot}$ pval < 0.05

Table 5: Comparisons on RMSE under different weather conditions (speed in km/h)

Weather condition	SVR $m1$ RMSE (km/h)	% Reduction in SVR RMSE with weather variables	Base $m1$ RMSE (km/h)	% Reduction in Base RMSE with weather variables	% Reduction in SVR $m1$ over Base $m1$
No Precipitation	8.29	-1.65	8.24	-1.61	0.52
Drizzle	10.01	-6.47	10.94	-0.15	-8.53
Light Rain	9.82	-3.44	10.3	-0.75	-4.55
Light Snow	8.42	-5.56	9.96	0.53	-15.46
Rain	7.79	-18.39	10.64	-4.67	-26.64
Sleet	8.1	-22.03	9.38	-27.1	-13.64
Snow	8.96	-23.87	11.8	4.6	-24.00
Strong Thunderstorm	10.72	-7.49	11.62	-8.15	-7.69
Thundershowers	9.00	-4.39	9.78	0.46	-7.95
Thunderstorms	9.27	-10.81	11.44	-0.81	-18.90

Table 6: Distribution of Speed Residuals (km/h) from SVR $m1$ and Base Model $m1$ for Sleet, Strong Thunderstorms and Rain

Residual Quartile (mph)	Sleet		Strong Thunderstorms		Rain	
	SVR	Base	SVR	Base	SVR	Base
Min. - 0%	-20.02	-19.68	-28.15	-38.53	-23.22	-28.42
25%	-7.6	-8.16	-6.71	-8.79	-1.71	-2.67
50%	-1.32	-2.43	1.51	0.47	2.54	0.03
75%	2.7	2.61	7.69	6.44	6.55	4.35
Max. - 100%	25.52	26.72	22.64	18.44	33.23	48.3

Table 7: Comparison of SVR models and Base models by time-of-day and presence/absence of precipitation using RMSE

Precipitation	Time of day	SVR $m1$ RMSE (km/h)	Improvement in SVR RMSE with weather variables (%)	Base $m1$ RMSE (km/h)	Improvement in Base RMSE with weather variables (%)	Improvement in SVR $m1$ over Base $m1$ (%)
No	6:00-8:59	9.19	-1.89	9.72	-3.6	-5.52
	9:00-14:59	8.96	-1.78	8.64	-1.07	3.54
	15:00-18:59	10.56	-2.1	10.38	-1.23	1.77
	19:00-5:59	5.91	-0.53	5.95	-1.21	-0.87
Yes	6:00-8:59	10.43	-2.63	11.2	-0.82	-6.91
	9:00-14:59	10.24	-3.5	10.64	-1.85	-3.86
	15:00-18:59	10.22	-4.27	11.39	-0.61	-10.41
	19:00-5:59	8.45	-8.29	9.41	-1.65	-10.36

Table 8: Predicted speeds with uncertain weather forecasts (km/h)

	Equal probabilities (Scenario 1)				Weather progression (Scenario 2)				
	OLS		SVR		OLS		SVR		
Precipitation	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Observed
No precipitation	87.23	5.63	86.42	6.44	82.88	0.48	83.69	0.64	103.48
Drizzle	76.12	5.63	75.8	7.4	79.66	0.97	91.25	1.45	81.92
Light rain	73.23	5.63	83.04	6.44	77.09	1.13	86.9	1.29	95.43
Light snow	75.96	5.79	85.3	6.28	91.89	1.13	99.14	1.93	78.38
Rain	75.64	5.47	76.44	5.31	72.58	0.64	70.33	1.61	77.41
Sleet	97.04	5.79	96.88	10.46	84.65	1.29	69.52	4.18	58.58
Snow	43.94	5.47	39.59	5.63	38.79	1.13	41.84	0.32	59.38
Strong thunderstorms	68.72	5.79	73.71	5.95	69.04	0.8	79.66	2.74	87.39
Thundershowers	94.15	5.63	84.17	6.6	96.08	0.8	85.46	0.64	84.65
Thunderstorms	45.54	5.63	53.75	4.99	44.9	0.8	50.21	2.74	46.83