
Piyushimita Thakuriah (Vonu)
Urban Planning and Policy and Urban Transportation Center
412 South Peoria Street, Suite 340 CUPPA Hall, Chicago, Illinois 60607 USA
+1-312-355-0447, vonu-pt@uic.edu

George Yanos, gynos@uic.edu, Urban Transportation Center
Jane Lin, janelin@uic.edu, Civil & Materials Engineering
Paul Metaxatos, pavlos@uic.edu, Urban Transportation Center
Rajmonda Sulo, rsulol@uic.edu, Urban Transportation Center
Wenjing Pu, wpu2@uic.edu, Civil & Materials Engineering
University of Illinois at Chicago

Lumumba Mbekeani, Lumumba.Mbekeani@navteq.com
NAVTEQ LLC, 425 West Randolph Street, Chicago, Illinois 60606 USA

Submitted for consideration for presentation in the
Intelligent Transportation Systems World Congress 2008, New York City

Abstract

Inclement weather can seriously affect traffic operations and contributes to increased accident risk. With the proliferation of traveler information systems, the opportunity exists to significantly increase the use of inclement weather information for travel decisions and increase traffic safety. The objective of this paper is to research the expected impacts of weather effects in a real-time traveler information environment and present alternative models to estimate the speed differential between actual and historical speeds. Such reliable information can then be broadcast to ATIS users and affect travel behavior.

The paper reviews the scientific literature on weathers impacts on traffic, discusses issues related to the size of the data involved in the analysis and proposes two alternative modeling frameworks. A parametric linear model with main and interaction effects, proposed as a base model, predicts the impact of weather on the difference between historical and sensor speeds to be in the range of 7 to 10 mph. A supervised learning algorithm, using Random Forests, predicts the previous error to be less than 9 mph but explains a greater percentage of the variability in the variable analyzed, indicating the latter is a robust approach to handling the highly non-linear and massive data sources with strong interactions among different weather variables.

Keywords: weather effects, real-time ATIS, alternative models
INTRODUCTION

Inclement weather can seriously affect traffic operations and contributes to increased accident risk. Weather affects driver behavior, vehicle performance, pavement friction, and roadway infrastructure (Office of Operations, Federal Highway Administration, 2007). An Oak Ridge National Laboratory study estimated the delay experienced by American drivers due to adverse weather conditions in 1999 at 46 million hours (Cambridge Systematics and Virginia Tech Transportation Institute, 2006). Rain, mean temperature and snow were found to significantly affect automobile accidents (Frost, 2002), with snow being the leading variable, as the number of accidents increased sharply with increased snowfalls.

With the proliferation of traveler information systems, the opportunity exists to significantly increase the use of inclement weather information for travel decisions. While weather forecasts have been disseminated via radio for several decades and via the Internet over the last 10 years or so, the proliferation of Variable Message Signs, hand-held mobile devices with real-time communications capability, roadside or wayside, kiosks and in-vehicle navigation systems that is connected via satellite, Dedicated Short-Range Communications (DSRC), HD Radio and other communications methods to real-time traffic information streams provide the basis for instantaneous and ubiquitous decision-making in response to weather. The potential use of real-time weather information in transportation can occur in many different ways including the decisions to not undertake a trip during times of bad weather, change in transportation modes, route or time of departure or simply the expectation of how much delay time is facing the traveler as he or she is stuck in traffic due to bad weather.

The objective of this paper is to research the expected impacts of weather effects in one such real-time traveler information environment. We start by defining the specific problem under consideration. This is followed by a review of prior literature. Subsequent sections discuss the data used in the analysis, alternative models and results. The final section presents conclusions and thoughts for future research.

TRAFFIC INFORMATION ENVIRONMENT AND STATEMENT OF PROBLEM

The problem considered here pertains to a specific configuration of traveler information systems. Let \( \hat{S}_{t,d} \) represent the historical mean speed on link \( \ell \) for the relevant time-period \( t \) and day-type \( d \), actual speed be \( S_{t,d} \) on link \( \ell \) at time \( t \) and day \( d \), \( d \in D \) and \( \Delta_{t,d} = S_{t,d} - \hat{S}_{t,d} \). The quantity of interest here is \( \Delta_{t,d} \), which would be impact due to weather that \( \hat{S}_{t,d} \) is to be factored by, in the event of inclement weather. Our goal is to model \( \Delta_{t,d} = f(\mathbf{X}) \) where \( \mathbf{X} \) is a vector of weather, time-of-day, season and other factors.

As an illustration of the problem, we depict, in Figure 1, a user with a hand-held device or a driver in a vehicle equipped with navigational capabilities on link \( l \) and time \( t \) and day-type \( d \). The hand-held device or vehicle contains \( \hat{S}_{t,d} \) or the historical means. In the event of
inclement weather, the goal would be to broadcast $\Delta_{t,t,d}$, so that the corrected speed estimate becomes $(\hat{S}_{t,t,D} + \Delta_{t,t,d})$.

Figure 1: Potential ATIS applications

In Figure 2, we illustrate two possible broadcasting scenarios for using weather effects in updating static or historical estimates of speeds. In Scenario I, $\Delta_{t,t,d}$ is broadcast centrally and then broadcast to the device or the vehicle whereas in Scenario II, the weather elements of $\mathcal{X}$ are broadcast and $\Delta_{t,t,d}$ is calculated locally, using pre-estimated functions $f$.
Irrespective of the specific dissemination approach used, our goal in this paper is to experiment with ways in which $\Delta_{t,t,d}$ can be calculated.
Figure 2: Possible scenarios for transmitted weather information

**PRIOR LITERATURE**

Although not related specifically to real-time applications, the literature has addressed different adverse weather events (e.g. rain, snow, ice, fog, wind, etc.) and their impacts on transportation systems. Reduced visibility and slippery pavement surface are the main consequences of adverse weather, and they will then reduce vehicle maneuverability and trigger more conservative driving behavior. This will finally result in changes in: (1) traffic operation parameters such as speed, capacity, and travel time (e.g., Goodwin, 2002; Huang and Ran, 2003; Smith et al., 2004; Agarwal et al., 2005; Chung et al., 2006; Dailey, 2006; Hranac et al., 2006; KYTE et al. 2000 and 2001; Nookala, 2006; Tu et al., 2007); (2) traffic safety (e.g., Shankar et al., 1995; Edwards, 1998; Brow and Baass, 1997; Khattak et al., 1998; Eisenberg, 2004); and (3) travel demand (e.g., Hanbali and Kuemmel, 1993; Hanbali, 1994; Knapp, 2000; Keay and Simmonds, 2005; Maze et al., 2006; Nookala, 2006). The focus of this paper is speed changes.

The adverse effect of weather on traffic operations (speed, capacity and travel time variability) can be explained by the fundamental speed-flow diagram in traffic flow theory, as shown in Figure 1 (Huang and Ran, 2003). The solid curve is the speed-flow relationship under normal conditions. Point C and F represent the free flow speed and link capacity, respectively. When a weather event occurs, the link speed-flow relationship shifts to the dashed curve, where link capacity (point E) and free flow speed (point D) are reduced simultaneously. At any traffic state (A) under normal conditions is changed to a new state (B) under inclement weather conditions. The change between A and B is the effect of weather quantified either through direct observations or modeling, to be discussed later.

![Speed-flow relationship and weather impacts](image)

Figure 3: Speed-flow relationship and weather impacts

The effects of severe weather conditions on traffic operations are quantified either through simple descriptive analyses of observations or by modeling between traffic operational parameters (speed, capacity, travel time) and weather variables. Some common modeling techniques include traffic flow theory (Chung et al., 2006), regression analysis (Hranac et al.,
2006), time series analysis (Dailey, 2006), artificial neural network (Huang and Ran, 2003), and statistical range method (Tu et al., 2007).

**DATA USED FOR MODELING**

An analysis of the following sources of data were undertaken for this paper:

- Archive of static or historical estimates of speed ($\hat{S}_{t,\ell,D}$) from road sensors at 5-minute intervals, on link $\ell$ for the relevant time-period $t$ and day-type $D$.
- Archive of observed speed $S_{t,\ell,d}$ on link $\ell$ at time $t$ and day $d$, $d \in D$ at 15 minute intervals.
- Data from weather sensors at 1-hour time resolution on: precipitation (inches), barometric pressure (millibar), wind (miles per hour), visibility (miles), humidity (percent) and cloud base (feet). It needs to be noted that for some of the models, numeric measurements of the weather variables were used whereas in other models, indicator variable representation of the same weather factors were used\(^1\).

The data used here was for the highway network in the Chicago metropolitan area over a period of three years. Weather data was also available, from 6 sensors located throughout the metro area. These datasets were fused together using link ID’s and time-of-day. 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 static estimates and weather information, and presents the merged data for statistical and data mining applications.

**BASE MODEL: EXPERIMENTATIONS WITH MAIN AND INTERACTION EFFECTS**

This model predicts $\Delta_{t,\ell,d}$, which, as defined earlier, is the deviation of actual speed $S_{t,\ell,d}$ on link $\ell$ at time $t$ and day $d$ from the estimates $\hat{S}_{t,\ell,D}$ that represents the historical mean speed

\(^1\) For example, the “wind” variable, which ranges in value from 0 mph to 50 mph over three years on I-290, can be converted into a qualitative variable with $K$ levels ranging from “extremely high winds” to “no wind”. This scheme allows us to define $K-1$ indicator variables (for example, $W_k = \begin{cases} 1 & \text{if } \ell^k \text{ observation denotes "Very Windy" conditions} \\ 0 & \text{otherwise} \end{cases}$)

The rationale for indicator-variable coding is that it allows us to explicitly examine the effects on speed of interactions such as very windy under heavy snow versus very windy under no snow.
on link $\ell$ for the relevant time-period $t$ and day-type $d$, due to the impact of weather. That is,
\[ \Delta_{t,d} = S_{t,d} - \hat{S}_{t,d} \]

The Base Model is:
\[
\Delta_{t,d} = \alpha + \sum_{j=1}^{J} \beta_j W_{j,t,d} + \sum_{k=1}^{K} \gamma_k T_{k,t,d} + \sum_{m=1}^{M} \mu_m M_{m,t,d} + \sum_{r=1}^{R} \sum_{s=1}^{S} \omega_{r,s} (W_{r,t,d} \times W_{s,t,d}) \\
+ \sum_{j=1}^{J} \sum_{k=1}^{K} \varphi_{j,k} (W_{j,t,d} \times T_{k,t,d}) + \varepsilon_{t,d} \tag{I}
\]

where:
- $W$’s are indicator variables related to the $j$th weather condition,
- $T$’s are the time-of-day indicator variables (there are 5 $T$'s in the model: early morning, morning peak, mid-day, evening peak and late evening, with the rest of the day being the base),
- $M$’s are indicators for month of the year (to account for seasonality effects),
- the $(W_r \times W_s)$ terms are interactions between different weather conditions $r$ and $s$, and
- the $W_r \times T$ terms are interactions between weather condition and time-of-day.

The reason for introducing these main and interaction effects is that we believe that the impact that any one weather condition has, on the deviation of the observed speed from the historical estimates, will vary based on other weather conditions. The impact of weather on the speed deviation will also depend on the time of day (proxying congestion levels); for example, heavy rain will affect how differently observed speeds differ from the static estimates during off-peak hours compared to peak hours.

For an illustration of the performance of this model, we present the results from running the model on the linked sensor, static estimate and weather data a 40-link segment [20 in each direction (EB and WB)] on Eisenhower Expressway (I-290) between downtown Chicago and 1st Ave. The data includes three years between 2005, 2006, and 2007. Figure 4 shows the quantiles of the residuals $(\Delta_{t,d} - \hat{\Delta}_{t,d})$ of this model. About 80 percent of the observations are at or very close to 0. Below the 10th and above and 90th percentile points are some large (negative and positive) residuals – however, these large residuals are not for days when there was inclement weather, indicating that introducing the weather effects has explained the variation on the $\Delta_{t,d}$ due to bad weather conditions.

It appears that the large residuals are due to congestion-related factors that the time-of-day proxy variables have not adequately addressed. Most of the large positive residuals are during the evening peak on “good” weather days whereas the large negative residuals are for both morning and evening peaks during such days. Another point to be noted is that the large residuals are not spatially constrained to a few links – these appear to be more or less evenly distributed to over all 40 links, although some links have relatively higher proportion of failures.
Figure 4: Distribution of residuals (difference between actual and predicted values of $\Delta_{t,j,d}$ from applying Base Model on I-290 over 3 years

The models we experiment with have given us Root Mean Square Errors (RMSE) in the range of 7 - 10 miles per hour (that means, the average error in predicting the impact of weather on the static estimates is “off” from the actual impact by 7 to 10 mph). This is not bad, but clearly there is room for improvement. As another indicator, all the terms on the right hand side of (I) above explain only in the range of about 2 to 8 percent of the variation in $\Delta_{t,j,d}$, as measured by the coefficient of determination (the $R^2$). However, $R^2$ might not be the right diagnostic measure since we believe that there are strong non-linearities in the data, which we have not mined as yet in these models. The Random Forest approach given next might be the key to uncovering some of these non-linearities.

**RANDOM FORESTS: A SUPERVISED LEARNING METHOD**

Random Forests (RF) is a supervised learning method given by Breiman (2001a, 2001b). The RF algorithm can be used for both classification and regression. Unlike CART-type methods, RF is based on multiple trees which splits the data into clusters and estimates $\Delta_{t,j,d}$ on the basis of trees. RF runs efficiently on large databases and can handle thousands of input variables without variable deletion. It gives estimates of what variables are important in the classification and splits data on the basis of a random selection of variables at each node. The generated forest can be saved for future use on other data and can be used to estimate $\Delta_{t,j,d}$ given input vector $X$ either centrally or in a mobile device or the vehicle.
RF offers an experimental method for detecting variable interactions; in fact, we selected RF for the weather problem because it has the capability to handle non-linearities and interactions in automated fashion without over-fitting. The method is also able to capture nonlinear, complex relations in traffic historical data. Our goal was to run RF code in regression mode to build several tree models and to analyze the statistical summaries generated from these models and compare them to the Base Models described in the previous section. We have also used RF to cluster similar network links and/or time periods.

In general, the RF algorithm works as follows:

Given $N$ observations and $K$ tree models:
- for $i=0 .. k$
  - Bootstrap sample observations
  - For each sample calculate tree model (unpruned)
  - For each node
    - Sample m predictor variables
    - Choose best partition
- combine models
- calculate summary statistics, variable importance

Two examples from preliminary random forest regressions are now presented. In preliminary analysis of the RF algorithm, the results appeared to be promising but the performance was not too different from the Base Model (using RMSE and percent of variation explained as the main comparison measures).

Table 1. Variable Importance in Illustrative Preliminary Model

<table>
<thead>
<tr>
<th>Variable</th>
<th>% MSE Increase</th>
<th>Increase in Node Purity</th>
</tr>
</thead>
<tbody>
<tr>
<td>link</td>
<td>23.07</td>
<td>6,034,689.7</td>
</tr>
<tr>
<td>year</td>
<td>16.53</td>
<td>427,796.8</td>
</tr>
<tr>
<td>month</td>
<td>19.49</td>
<td>832,265.9</td>
</tr>
<tr>
<td>day</td>
<td>26.43</td>
<td>1,378,135.1</td>
</tr>
<tr>
<td>hour</td>
<td>15.57</td>
<td>1,720,666.5</td>
</tr>
<tr>
<td>minute</td>
<td>4.21</td>
<td>1,602,468.1</td>
</tr>
<tr>
<td>day of week</td>
<td>26.19</td>
<td>1,007,407.0</td>
</tr>
<tr>
<td>temperature</td>
<td>24.64</td>
<td>1,806,195.1</td>
</tr>
<tr>
<td>wind speed</td>
<td>16.72</td>
<td>1,420,225.6</td>
</tr>
<tr>
<td>humidity</td>
<td>8.01</td>
<td>1,704,723.9</td>
</tr>
<tr>
<td>visibility</td>
<td>10.1</td>
<td>573,282.7</td>
</tr>
<tr>
<td>precipitation</td>
<td>8.53</td>
<td>129,786.8</td>
</tr>
<tr>
<td>precip. code</td>
<td>14.34</td>
<td>303,259.8</td>
</tr>
<tr>
<td>snow cover</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>bar. tendency</td>
<td>16.51</td>
<td>1,906,774.7</td>
</tr>
<tr>
<td>sky descr.</td>
<td>19.83</td>
<td>873,035.1</td>
</tr>
</tbody>
</table>
For example, in one preliminary model, we randomly chose 200,000 observations from the 327 million observations in the Chicago linked dataset. We grew 10 trees at each iteration and used 5 variables at each split. The model has a 10.8 RMSE and only 4.7% of the variance explained. Table 1 gives two criteria for variable selection in the model, the % MSE increase, and the increase in node impurity. The % MSE increase means: for each tree, the prediction accuracy on the out-of-bag (training) portion of the data is recorded. Then the same is done after permuting each predictor variable. The difference between the two accuracies are then averaged over all trees, and normalized by their standard deviation. The second measure, increase in node purity, is the total increase in node impurities (the homogeneity of observations in a node to each other) from splitting on the variable, averaged over all trees. For classification, the node impurity is measured by the Gini index. For regression, it is measured by residual sum of squares. In other words, the higher the value of node purity and percent increase in MSE, the more important the variable. For example, the link ID is the most important variable because it scores the highest on both measures.

The model performance is significantly improved by building more trees and with larger sample databases, which provides a greater training set for the RF algorithm. By building 50 instead of 10 trees as in the earlier case, we are able to significantly improve model performance. The percent of variation explained is 27.50 and the RMSE is in the range of 9.3 mph. While the average prediction error is in the same range as the Base Model, the RF Model now accounts for a far greater amount of the variation in $\Delta_{t,d}$, thus providing a more robust approach to modeling speed impacts for weather. As a whole, the predictive power of the model has increased in comparison to the Base Model and the preliminary RF model results.

Table 2: Variable Importance in Models including more trees

<table>
<thead>
<tr>
<th>Variable</th>
<th>% MSE Increase</th>
<th>Increase in Node Purity</th>
</tr>
</thead>
<tbody>
<tr>
<td>link</td>
<td>35.77</td>
<td>9688485.2</td>
</tr>
<tr>
<td>year</td>
<td>22.75</td>
<td>615529.9</td>
</tr>
<tr>
<td>month</td>
<td>13.84</td>
<td>1279389.4</td>
</tr>
<tr>
<td>day</td>
<td>15.74</td>
<td>2025995.5</td>
</tr>
<tr>
<td>hour</td>
<td>18.35</td>
<td>2616395</td>
</tr>
<tr>
<td>minute</td>
<td>3.62</td>
<td>2416053.5</td>
</tr>
<tr>
<td>day of week</td>
<td>28.33</td>
<td>1605130</td>
</tr>
<tr>
<td>temperature</td>
<td>34.89</td>
<td>2604826.3</td>
</tr>
<tr>
<td>wind speed</td>
<td>19.14</td>
<td>2048658.6</td>
</tr>
<tr>
<td>humidity</td>
<td>16.21</td>
<td>2500565.8</td>
</tr>
<tr>
<td>visibility</td>
<td>10.19</td>
<td>753280</td>
</tr>
<tr>
<td>precipitation</td>
<td>5.01</td>
<td>183434</td>
</tr>
<tr>
<td>precip. code</td>
<td>11.96</td>
<td>429566.3</td>
</tr>
<tr>
<td>snow cover</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>bar. tendency</td>
<td>30.21</td>
<td>2790835.6</td>
</tr>
<tr>
<td>sky descr.</td>
<td>14.52</td>
<td>1252401</td>
</tr>
</tbody>
</table>
Bad weather can affect traffic operations and contributes to increased accident risk. The objective of this paper is to research the expected impacts of weather effects in a real-time traveler information environment and present alternative models to estimate the speed differential between actual and historical speeds. Such information can then be broadcast to ATIS users and affect travel behavior.

The paper reviews the scientific literature on weather's impacts on traffic, discusses issues related to the size of the data involved in the analysis and proposes two alternative modeling frameworks. A parametric linear model with main and interaction effects, proposed as a base model, predicts the impact of weather on the difference between historical and sensor speeds to be in the range of 7 to 10 mph.

A supervised learning algorithm, using Random Forests, predicts the previous error to be less than 9 mph but explains a greater percentage of the variability in the variable analyzed, indicating that the latter is a robust approach to handling the highly non-linear and massive data sources in which strong interactions exist among different weather variables. We found that RF offers a robust method for detecting variable interactions in massive datasets and has the ability to handle non-linearities and interactions in automated fashion without over-fitting. We have also used RF to cluster similar network links and/or time periods. The reason for this is that the effect of weather on traffic speed is not the same in all links; the effect on some links can be accounted for by only a few variables whereas others need a large number of variables and interactions among those variables.

REFERENCES

17) Khattak, A.J., Kantor, P., and Council, F.M., 1998. Role of adverse weather in key crash types on limited access roadway-implications for advanced weather system. Transportation research record 1621, 10-19

---

The authors would like to thank NAVTEQ LLC for the data used in this study.