Improving traffic prediction by including rainfall data
Improving traffic prediction by including rainfall data

Leong Wai Leong, Kelvin Lee, Kumar Swapnil, Xiao Li, Ho Yao Tong Victor, Nikola Mitrovic, Muhammad Tayyab Asif, Justin Dauwels, and Patrick Jaillet

Abstract—Weather conditions tend to have measurable impact on traffic conditions of the roads. This relationship is commonly studied at the network level without explicit explanation of the link performances. Furthermore, existing studies typically use high resolution traffic data which may not be available across the entire network and especially during the adverse weather conditions. In this study we explore the impact of rainfall intensity on low-resolution speed band data. We also test whether this additional information about rainfall may improve the prediction accuracy of data-driven models for individual roads. To do so, we incorporate the information about the rainfall intensity into support vector machine (SVM) prediction algorithm. As a benchmark we only consider temporal features to predict near future traffic conditions during rainy weather. Numerical results for 616 road segments in Singapore confirm that rainfall impacts traffic conditions in terms of decreasing the driving speed. This reduction increases with the rain intensity. Furthermore, the results show that additional rainfall data enhances the prediction accuracy for certain number of links; while for the others the rainfall information is not that useful.

Index Terms—Intelligent transportation systems, traffic prediction, SVM, rainfall correlation.

I. INTRODUCTION

Weather conditions tend to have a significant impact on the driving conditions. Various studies have shown that different weather factors such as temperature [1], [2], precipitation [3], [4] and visibility [5], [6], tend to increase the risk of vehicle collisions and reduce the roadway capacity. The decreased throughput results in degradation (up to some extent) in the driving conditions measured in terms of speed, volume and travel time [7].

The rainfall data is frequently used to explore the impact of adverse weather conditions on road traffic [3], [8]–[11]. Previous studies have reported that this impact varies across the different locations, time of the day, road infrastructure, etc. Nonetheless, rainfall is reported to cause speed reduction by 6% to 12% in different locations [9], [12]. Such analysis has been utilized to improve the performance of short-term traffic prediction [11]. These studies typically provide aggregate performance across the network. However, the prediction performance may vary significantly from one road segment to another. Furthermore, they perform analysis by using high resolution traffic data whereas in many cases high resolution data may not be available in real-time.

In this study we explore how the rainfall information can be used to improve the prediction performance of data-driven methods. To this end, we try to predict various levels of congestion that might arise due to rainfall. To do so, we represent the traffic conditions of each road by a speed band. For instance, speed band 1 would mean that the average speed on that road was less than 20 km/h during that time interval. Similarly, speed bands 2, 3, and 4 refer to the speeds between 20 – 40 km/h, 40 – 60 km/h and speeds higher than 60 km/h, respectively. We then combine this speed information with weather data to train support vector machine (SVM) to perform traffic prediction. As a benchmark, we apply SVM to perform short-term prediction (for 5 minute prediction horizon) of the experimental network without utilizing additional weather information. We compare the performance of these two sets of predictors to assess the impact of additional weather information. The results show that for certain links additional rainfall information improves the prediction accuracy. For others, the rainfall data is not that useful.

The rest of the paper is structured as follows. In Section II we explain the data set analyzed in this paper and discuss the pre-processing steps involved in data collection. In Section III we present the meaningful information extracted from raw data. In Section IV we briefly explain the support vector machine (SVM) method in conjunction with our data set. In Section V we provide numerical results for our experiments and discuss further improvements. In Section VI we summarize our work and suggest topics for future research.

II. DATA COLLECTION

In this section, we describe the different steps involved in the collection and processing of rainfall and traffic data sets. The recorded speed represents the average speed of all vehicles which traverse a road segment (link) during the 5-mins sampling interval. The Land Transportation Authority (LTA) of Singapore provided us with speed data for a period of three months (September - November, 2014). The speed
information is provided in the form of a discrete variable with range 1 – 4. We refer to these discrete states as speed bands. We represent the speed data set in the form of a matrix $A$ ($A \in \mathbb{Q}^{m \times n}, Q = \{1, 2, 3, 4\}$). The columns of the matrix $\{a_i\}_{i=1}^n$ contain traffic data from different roads $\{s_i\}_{i=1}^n$. Rows represent time instances $\{t_j\}_{j=1}^m$ at which the traffic data is recorded. Each matrix cell $(a_{ij})$ represents one of the four speed bands at location $s_j$ during the interval of time $(t_i - T, t_i)$ where $T$ is the sampling period (e.g., 5 minutes). In this analysis we select the 616 road segments (hence $A \in \mathbb{Q}^{26208 \times 616}$). Observed links have at least 95% of data. The missing values are not taken in consideration in this study.

We collected rainfall information from National Environment Agency’s (NEA) website [13]. The rainfall data is obtained from the weather radar and published as an image (on average) every 5-10 minutes (see Fig. 1). We collected rainfall data for the same time period as the speed band data and organized it in the form of a matrix $B \in \mathbb{R}^{26208 \times 616}$. Each matrix cell $(b_{ij})$ in the matrix $B$ represents the estimated rainfall intensity at the location $s_j$ during the interval of time $(t_i - T, t_i)$.

We estimate rainfall intensity at the location $\{s_i\}_{i=1}^n$ as follows: first we overlay the rainfall and traffic road maps to estimate the location of the link $s_j$ on the rainfall map (see Fig. 2). To quantify rainfall intensity at a single pixel of the map we develop a custom 0–10 scale where 0 represents no rain and 10 refers to value of heavy rain (see the bottom part of Fig. 2). The rainfall intensity corresponding to a segment $s_j$ is approximated by the average of rainfall intensities in the area of this segment (see the right side of Fig. 2).

Since the rainfall data is irregularly reported (especially during the heavy raining intervals) we use Optical Character Recognition (OCR) to estimate rainfall intensity at time $t_i$ (see the top of Fig. 2). OCR is the process of converting text images into text in ASCII format. In order to increase the accuracy of the OCR, we binarize and scale the text images [14]. To this end, we arrange data in two matrices, speed matrix $A$ and rainfall matrix $B$.

III. DATA ANALYSIS

In this study we aim to explore whether the rainfall impacts the low resolution traffic data and (if yes) how this information can be used to improve the prediction performance of data-driven models. We start by comparing the average speed across the network for rainy and corresponding non-rainy time intervals. Average network speed at time $t_i$ represents the mean speed for the subset of the links $\{c_1, ..., c_k\} \subseteq \{a_1, ..., a_n\}$ where the rainfall intensity $b_{ij}$ is above the certain threshold $\ell$. The baseline speed is the average speed for the identical subset $c$ for the same day (e.g. Wednesday) and period of the day if there is no rain $(b_{ij} = 0)$. Since the light rainfall may not have significant impact on (low-resolution) speed data, we tested various rainfall thresholds $(\{\ell = (0, 1, 2, ..)\})$. Fig. 3 shows that average network speed deviates from its baseline value during the rainy intervals. This deviation increases with the rain intensity (see blue line on Fig. 3). It is somehow natural that heavy rain has stronger impact on traffic conditions than the light rain (see Fig. 3). However, this increase in rainfall threshold leads to the reduction in the size of data set for which $b_{ij} > \ell$ (see Fig. 3). Aforementioned observations confirm that there is a quantitative relationship between rainfall rate and traffic conditions. This relationship can prove useful in improving the prediction accuracy of data-driven models. In the next section, we briefly discuss a commonly applied data-driven regression method known as SVM. In this study, we performed traffic prediction by training SVM using traffic features (speed band data) along side weather information.

IV. SUPPORT VECTOR MACHINE (SVM)

Support vector machine (SVM) is a data-driven prediction algorithm that is often applied for traffic applications [15]–[18]. In this paper we use multiclass SVM since the traffic speed belongs to one of four categories (1 – 4). The multiclass SVM problem is frequently decomposed into multiple binary classification problems [19].

The binary classification aims to find a suitable hyperplane (i.e. decision boundary) to separate the classes. The equation of this hyperplane is:

$$w^T x + b = 0, \text{ where } w \in \mathbb{R}^n, b \in \mathbb{R}. \quad (1)$$

SVM tries to search for such hyperplane which maximizes the margin between the two classes and minimizes the training error [20]. This leads to the following optimization problem:
\[
\begin{align*}
\text{min} & : \frac{1}{2} w^T w + C \sum_i \varepsilon_i, \\
\text{s.t.} & : y_i (w^T x_i + b) \geq 1 - \varepsilon_i \text{ and } \varepsilon_i \leq 0 \quad \forall \ x_i,
\end{align*}
\]

where \( C \) is a cost associated with the error of training data points and \( \varepsilon_i \) are slack variables.

To achieve better performance linear classifiers are frequently replaced with the non-linear. In this case all \( x_i \) are replaced with \( \phi(x_i) \), where \( \phi \) provides the higher-dimensional mapping that leads to standard SVM formulation:

\[
\begin{align*}
\text{min} & : \frac{1}{2} w^T w + C \sum_i \varepsilon_i, \\
\text{s.t.} & : y_i (w^T \phi(x_i) + b) \geq 1 - \varepsilon_i \text{ and } \varepsilon_i \leq 0 \quad \forall \ x_i
\end{align*}
\]

For more detailed information about the SVM, please refer to [21].

In this study, we evaluate the prediction performance of two sets of models for 5 minute prediction horizon. We refer to these models as benchmark and proposed models. In the former, we only use current/past traffic speed values to train and test SVM predictors. In the latter, we use information about the current/past rainfall as additional features. These formulation allows us to analyze the impact of rainfall information in the prediction accuracy of data-driven models. In our study, we evaluate the prediction performance for each link in the network and prediction horizon of 5 min.

Fig. 3: Speed reduction. Comparisons are provided for different threshold values (\( \ell \)) of rainfall intensity.

Fig. 4: Comparison of prediction performance of proposed and benchmark methods. For certain links proposed method outperforms the benchmark approach.

V. RESULTS

First we investigate the prediction performances of multiclass support vector machine (SVM) for the proposed and benchmark methods and prediction horizon of 5 min. We train SVM predictors using two months of data (September and October, 2014). For proposed method, we restrict training...
### TABLE I: Average improvement of prediction performance in respect to the rainfall threshold. Calculated performance refers only to these links where additional rainfall information is useful. Number of these links is presented for each threshold.

<table>
<thead>
<tr>
<th>Threshold ($\ell$)</th>
<th>0</th>
<th>0.5</th>
<th>1</th>
<th>1.5</th>
<th>2</th>
<th>2.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of the links where rainfall is useful</td>
<td>251</td>
<td>226</td>
<td>214</td>
<td>200</td>
<td>181</td>
<td>165</td>
</tr>
<tr>
<td>Accuracy improvement [%]</td>
<td>2.11</td>
<td>2.56</td>
<td>2.66</td>
<td>2.77</td>
<td>3.05</td>
<td>3.35</td>
</tr>
</tbody>
</table>

Fig. 5: Histogram of the links where rainfall data is useful. Results are presented for different threshold values.
data set to these instances when the rainfall intensity is above a certain threshold. Hence, different road segments might have different number of training data points. Similarly, testing set for both methods refers to the November speed band data points \( a_{i,j} \) when the corresponding rainfall intensity \( b_{i,j} \) is above a certain threshold \( \ell \). Fig. 4 shows that, for \( \ell = 0 \), there are certain links (251 out of 616) where information about the rain improves the performance of data-driven models. For other links the performance either remained the same or degraded (see Fig. 4). In further analysis, we only considered these roads where rainfall data proved useful.

Table I shows that prediction accuracy can be improved by choosing higher thresholds for rainfall intensity. For instance, if we consider all the time instances involving rainfall then we get an average improvement of around 2% in prediction performance. On the other hand, if we restrict our analysis to cases involving moderate to heavy rain (\( \ell > 2.5 \)), then we are able to achieve an average improvement of 3.5% in prediction performance. Naturally, heavy showers tend to have significant effect on road traffic and hence make it easier for predictors (SVM) to model this relationship. However, it is not possible to select a very high threshold as it will substantially reduce the size of the training and test data.

Let us now explore the impact of rainfall threshold rate on prediction performance of those links where the rainfall data was useful. We trained the predictors and compared the prediction performance of the proposed and benchmark methods for these time instances when rainfall intensity is above a certain threshold. Fig. 5 shows the frequency of the road segments where proposed method outperformed the benchmark method. Frequencies are presented for different threshold values. From the Fig. 5 it can be seen that for the most of links rainfall data provides minor improvement, regardless to the threshold of rainfall intensity. With the increase in rain intensity threshold, number of these links which have significant improvement increases. Our numerical results show that for \( \sim 16\% \) of the all road segments in the network rainfall data proved to be useful for each threshold.

VI. CONCLUSIONS AND FUTURE WORK

In this paper we explored the impact of rainfall intensity on the prediction performance of discrete traffic data. Our analysis of traffic and rainfall data showed that rainfall reduces the average network speed. Furthermore, we also found that this decrease in speed increases with the rainfall intensity. We then added this information as input feature in support vector machine (SVM) prediction algorithm. Numerical results show that for certain number of links rainfall information enhances the prediction accuracy.

In the future, we will investigate those links where the additional rainfall information is not useful. Moreover, to further support our conclusions we will conduct similar analysis using the high-resolution information such as volume data.

REFERENCES


