Traffic Incident Duration Prediction Based On Partial Least Squares Regression

Xuanqiang Wang\(^a\)*, Shuyan Chen\(^a\), Wenchang Zheng\(^a\)

\(^a\)School of Transportation, Southeast University, Nanjing, 210096, China

Abstract

The prediction of the traffic incident duration is a very important issue to the Advanced Traffic Incident Management (ATIM). An accurate prediction of incident duration makes a lot contributes to making appropriate decisions to deal with incidents for traffic managers. The paper employed the Partial Least Squares Regression (PLSR) to build model between incident duration and its influence factors. Three models were established for three types of incident correspondingly, i.e. stopped vehicle, lost load and accident. Meanwhile, a model without distinguishing the incident type was built as a comparison. The experiments results indicated that the model obtained high prediction accuracy for those incidents which last 20 minutes to 90 minutes. The models got prediction accuracy of 77.24%, 86.59%, 83.33% and 71.30% for stopped vehicle, lost load, accident and all incidents within 20 minutes error, respectively. The results indicated that the PLSR has a promising application to predict traffic incident duration

Keywords: Incident Management; Incident Duration; Prediction; Partial Least Squares Regression.

1. Introduction

Traffic incident is one of the major causes leading to non-recurring congestion. Owning to its unpredictable and warrants immediate response, traffic managers require a full understanding of the nature and effect of the incident to estimate incident duration accurately. Then, they can make reasonable rescue decision and carry out corresponding traffic control measures to reduce the impact of these incidents.

So far, various methodologies have been employed to build models for incident duration forecast. The log-logistic distribution was firstly used to describe freeway incident duration (Jones, Janssen & Mannering, 1991), then, lognormal distribution (Garib, Radwan, & Al-Deek, 1997) and weibull distribution (Nam & Mannering, 1991).
2000) followed. Garib, Radwan and Al-Deek (1997) developed a log-based regression model to predict incident duration, and the adjusted R-square value of his regression model is 0.81. As the incident duration is not easy to predict exactly, many researchers turned to find the probability of the incident duration. Jones, Janssen and Mannering (1991) introduced the important concept of conditional probability (i.e., the probability of an incident ending in the X min given that it has lasted Y min, here Y<X.), and they applied a log-logistic hazard-based duration model to study freeway incident duration in Seattle; Nam and Mannering (2000) followed up on the concept by applying hazard-based models to incident duration. The estimation results by their model showed that a wide variety of factors significantly affect lasting times. In the recent years, with the development of data mining technology, a large number of data mining technologies have been applied to predict incident duration, such as decision trees (Ozbay & Kachroo, 1999), classification tree (Smith, 2001), Fuzzy Logic Theory (Wang, Chen, & Bell, 2002), Bayesian Networks (Ozbay & Noyan, 2006), Artificial Neural Network (Liping, Weiming, Xiangyuan & Luping, 2010) and Support Vector Regression (Wu, Chen, & Zheng, 2011). These models are more complicated to be understood, in fact, some algorithms have a good performance, but some methods are not very promising.

Even though a lot of methods have been proposed to predict incident duration, however, there are deficiencies among the existing methods more or less. Firstly, some algorithms are complicated and take a lot of time to train, e.g. ANN and decision trees. Secondly, some methods are not easy to operate in the practical application. It is a common problem for probability-based model that it only gives a probability of the incident duration. Traffic managers are always faced with the dilemma when the probabilities of the two durations are almost the same. Thirdly, the prediction accuracy of existing methods is not satisfied, and there is a great room for improvement.

The Partial Least Squares Regression (PLSR) is a method for constructing predictive models when the independent variables are a lot and highly collinear while the sample size is not enough. It combines the merits of Principal Component Analysis, Canonical Correlation Analysis and Multiple Linear Regression in the course of modeling (Tobias & Others, 1995). PLSR is an ideal technique to solve the problem mentioned previously. In this paper, we used PLSR to model the duration of urban expressway incidents and obtained expected results.

2. Partial Least Squares Regression

2.1. The basic principle of PLSR

The partial least squares regression was developed in the 1960’s by Herman Wold as an econometric technique, but it became popular first in chemo-metrics (HO Skuldsson, 2005). The general idea of PLS is to try to extract the latent factors, accounting for as much of the manifest factor variation as possible while modeling the responses well (Tobias & Others, 1995).

Assume there are $p$ independent variables $X = \{x_1, \ldots, x_p\}$, $q$ dependent variables $\{y_1, \ldots, y_q\}$ and $n$ samples. Then the data matrices $X_{n \times p}$ and $Y_{n \times q}$ are obtained. Component $t_1$ and $u_1$ are extracted from $X$ and $Y$ respectively under the following conditions (Cai, Wang, Yang, Hua, & Wu, 2008)

- $t_1$ and $u_1$ should represent $X$ and $Y$ to the most extent;
- Relativity between $t_1$ and $u_1$ should be the maximum.

These two demands indicate that $t_1$ and $u_1$ should represent $X$ and $Y$ respectively as well as possible. Meanwhile, independent variable $t_1$ explains dependent variable $u_1$ best. The regressions of $X$ to $t_1$ and $Y$ to $t_1$ are made respectively after extraction of $t_1$ and $u_1$. Algorithm is stopped if the accuracy is enough, which will be talked in the next chapter in detail. Otherwise, second extraction of components would be implemented. Finally if $m$ components $t_1, \ldots, t_m$ are obtained, regression equation of $y_j$ to original variables $x_1, \ldots, x_p$ is established through transition from $t$ to $x$, for $j = 1, 2, \ldots, q$. Finally, we get a model between $X$ and $Y$, as follow

$$y_j = a_{j1}x_1 + a_{j2}x_2 + \ldots + a_{jp}x_p$$  (1)
Where \( \mathbf{a}_j = (\alpha_{j1}, \alpha_{j2}, \ldots, \alpha_{jp}) \) is the coefficient vector of regression equation of dependent variable \( y_j \).

### 2.2. Modeling process

In our research, we treat traffic incident duration as dependent variable \( y \), thus, we will build the single factor partial least squares regression model. The PLSR modeling process (Wang, 1999) mainly includes the following steps,

(1) Normalize matrices \( \mathbf{X} \) and \( \mathbf{Y} \) into \( \mathbf{E}_o \) and \( \mathbf{F}_o \)

\[
\mathbf{E}_{oi} = \frac{(x_i - \bar{x}_i) / S_{xi}}
\]

\[
\mathbf{F}_o = \frac{(y - \bar{y}) / S_y}{(y - \bar{y}) / S_y}
\]

Where: \( \bar{x}_i \) and \( \bar{y} \) are the mean of \( x_i \) and \( y \), \( S_{xi} \) and \( S_y \) are the mean square deviation of \( x_i \) and \( y \).

(2) Extract principle components \( t_1 \) and \( u_1 \)

\[
t_1 = \mathbf{E}_o \mathbf{w}_1 \), \( u_1 = \mathbf{F}_o \mathbf{c}_1 \)

\[
1w_1 = 1, \quad 1c_1 = 1
\]

Where: \( \mathbf{w}_1 \) is the model effect weight, and \( \mathbf{c}_1 \) is the dependent variable weight. According to the two demands described before, we have

\[
\text{var}(t_1) \rightarrow \max; \quad \text{var}(u_1) \rightarrow \max; \quad \text{r}(t_1, u_1) \rightarrow \max
\]

That is:

\[
\text{cov}(t_1, u_1) = \sqrt{\text{var}(t_1) \cdot \text{var}(u_1) \cdot \text{r}(t_1, u_1)} \rightarrow \max
\]

That is to ask for maximum of \( \mathbf{w}_1^T \mathbf{E}_o \mathbf{F}_o \mathbf{c}_1 \) with constraints that \( \mathbf{w}_1 \mathbf{l} = 1 \) and \( \mathbf{c}_1 \mathbf{l} = 1 \). We use Lagrange method to solve it, then

\[
\text{Max } s = \mathbf{w}_1^T \mathbf{E}_o \mathbf{F}_o \mathbf{c}_1 - \lambda_1 (\mathbf{w}_1^T \mathbf{w}_1 - 1) - \lambda_2 (\mathbf{c}_1^T \mathbf{c}_1 - 1)
\]

After calculating partial derivative, we got the latent vector of matrices \( \mathbf{E}_o^T \mathbf{F}_o \mathbf{E}_o \) and \( \mathbf{F}_o^T \mathbf{E}_o \mathbf{F}_o \), namely, \( \mathbf{w}_1 \) and \( \mathbf{c}_1 \).

(3) Get vectors of regression coefficient

After getting principle components \( t_1 \) and \( u_1 \), we write \( \mathbf{E}_o \) and \( \mathbf{F}_o \) as follows

\[
\mathbf{E}_o = t_1 \mathbf{a}_1 + \mathbf{E}_1; \quad \mathbf{F}_o = t_1 \mathbf{b}_1 + \mathbf{F}_1
\]

Where: \( \mathbf{E}_1 \) and \( \mathbf{F}_1 \) are the residual matrix. By using least square method, we get regression coefficient vector \( \mathbf{a}_1 \) and \( \mathbf{b}_1 \) as follows

\[
\mathbf{a}_1 = \frac{\mathbf{E}_o^T t_1}{\|t_1\|^2}; \quad \mathbf{b}_1 = \frac{\mathbf{F}_o^T t_1}{\|t_1\|^2}
\]

(4) Extract principle components \( t_i \) and \( u_i \)

At this step, the algorithm repeats step (2) and step (3) to extract principle components \( t_i \) and \( u_i \) by using the residual matrix \( \mathbf{E}_{ei} \) and \( \mathbf{F}_{ei} \).

(5) Reduction process. After getting \( r \) principle components \( t_1, t_2, \ldots, t_r \), we have equation set as follows

\[
\begin{align*}
\mathbf{E}_o &= t_1 \mathbf{a}_1^T + \cdots + t_r \mathbf{a}_r^T + \mathbf{E}_r \\
\mathbf{E}_o &= t_1 \mathbf{b}_1^T + \cdots + t_r \mathbf{b}_r^T + \mathbf{F}_r
\end{align*}
\]


By substituting \( t_k = w_{k1}x_1 + \cdots + w_{kp}x_p \), \((k = 1, 2, \ldots, r)\) into \( Y = t_1 \beta_1 + \cdots + t_r \beta_r \), then we get the partial least regression equation of traffic incident duration, as follow

\[
y = a_1x_1 + \cdots + a_px_p
\]

(12)

2.3. The cross-validation

In many cases, there is no need to choose all the components \( t_1, \ldots, t_r \), actually, if the subsequently components fail to supply meaningful information for explaining \( F_a \). Using too many components would only damage the understanding of the statistical trend. We can use cross-validation to discriminate a new component contributing to model’s predict ability or not (Wang, 1999).

Firstly, use all the samples except sample \( i \) to build a regression equation based on \( h \) components. Substitute sample \( i \) into the equation to acquire fitting value \( \hat{Y}_{h(-i)} \). Repeat this computation for every sample and define the sum of squared predict error for \( y_i \) as \( \text{PRESS}_{hj} \), that is

\[
\text{PRESS}_{hj} = \sum_{i=1}^{n} (y_{ij} - \hat{Y}_{h(i-j)})^2
\]

(13)

Define the sum of squared predict error for \( Y \) as \( \text{PRESS}_h \), that is

\[
\text{PRESS}_h = \sum_{j=1}^{p} \text{PRESS}_{hj}
\]

(14)

Secondly, use all the samples to build a regression equation based on \( h \) components. Here, note the prediction value of sample \( i \) as \( \hat{Y}_{hij} \) and define the sum of squared error for \( y_i \) as \( \text{SS}_{hj} \), that is

\[
\text{SS}_{hj} = \sum_{i=1}^{n} (y_{ij} - \hat{Y}_{hij})^2
\]

(15)

Define the sum of squared error for \( Y \) as \( \text{SS}_h \), that is

\[
\text{SS}_h = \sum_{j=1}^{p} \text{SS}_{hj}
\]

(16)

The cross-validation of component \( t_h \) is defined as follow

\[
Q^2_h = 1 - \frac{\sum_{k=1}^{h} \text{PRESS}_{hk}}{\sum_{(k-h-1)} \text{SS}_{(h-k)}} = 1 - \frac{\text{PRESS}_h}{\text{SS}_{(h-1)}}
\]

(17)

Generally speaking, \( \text{PRESS}_h \) is always larger than \( \text{SS}_h \), while \( \text{SS}_h \) is smaller than \( \text{SS}_{h+1} \) to certain extent, adding component \( t_h \) is considered to be useful for the improvement of regression accuracy. There are two rules to be used for the cross-validation, that is

- If \( Q^2_h \geq 0.0975 \), the quality of model could be improved by adding a component. Otherwise, the number of component is enough;
- For \( k = 1 \ldots p \), there is at least a \( k \) making \( Q^2_h \geq 0.0975 \) to make sure that more than a component was obtained.

The cross-validity \( Q^2_h \) is the criterion for determining the number of components.
3. Data Description

The data set contains 1853 entries of traffic incident data collected from May 1, 2005 to September 13, 2005 in Utrecht, a central city in the Netherlands (Knibbe, Alkim, Otten, & Aidoo, 2006). Each item contains the parameter information about incident physical trait, response measures and traffic management measures. In this data set, the incident duration is the time between the detection of an incident and the clearance of an incident on the road. There are three basic types of traffic incidents: Stopped vehicle, Lost load and Accident. The parameters characteristic of these incidents have a great difference from each other. Subsequently, a preliminary statistical description for all incident data and each type of incident to present an overview of the data is given in Table 1 and Fig 1, respectively.

According to the Table 1, the mean of the incident duration is about 40 minutes. Besides, the Figure 1 shows that the incident duration between 20 minutes and 60 minutes have a great ratio. Some incidents that last from 90 minutes to 180 minutes have a great effect on the large standard deviation. We will model each type of incident separately.

<table>
<thead>
<tr>
<th>Incident type</th>
<th>Entries number</th>
<th>Minimum (min)</th>
<th>Maximum (min)</th>
<th>Mean (min)</th>
<th>Median (min)</th>
<th>Mode (min)</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stopped</td>
<td>571</td>
<td>2</td>
<td>282</td>
<td>43.73</td>
<td>31</td>
<td>26</td>
<td>37.90</td>
</tr>
<tr>
<td>Lost load</td>
<td>379</td>
<td>1</td>
<td>139</td>
<td>26.64</td>
<td>23</td>
<td>18</td>
<td>19.58</td>
</tr>
<tr>
<td>Accident</td>
<td>903</td>
<td>0</td>
<td>435</td>
<td>42.74</td>
<td>33</td>
<td>25</td>
<td>41.38</td>
</tr>
<tr>
<td>All</td>
<td>1853</td>
<td>0</td>
<td>435</td>
<td>39.75</td>
<td>30</td>
<td>24</td>
<td>37.40</td>
</tr>
</tbody>
</table>

![Fig. 1. Duration distribution frequency histogram of the incident](image)

4. Experiments Study

In our study, we wrote a code for the PLSR on MATLAB, a data analysis software platform, to train and
validate the data. Every time, 25% instances were selected to test the prediction accuracy of PLSR model established by the other 75% instances. The basic steps of the experiment are as follows,

1. Data pre-treatment. We dropped the record whose duration are under 10 minutes or over 180 minutes;
2. Dividing Data into the training set and the testing set;
3. Modelling based on PLSR by using the training set;
4. Predicting and evaluating on the testing set.

After running our programming code in the MATLAB, we got four models for forecast. And the predictive validities of each incident type are shown in Fig. 2. In the Figures, the blue line is for real statistical data, and the red line is for predictive data.

![Fig. 2. Predicted value and actual value of incident duration: (a) Stopped vehicle; (b) Lost load; (c) Accident; (d) All accident](image)

Seeing from Fig.2, we can roughly get the information that the models overestimate the data within 20 minutes and underestimate the data over 90 minutes, generally. But, the models have a good prediction on the data between 20 minutes and 90 minutes, and it has basically met the needs of the traffic management. As summarized in Table 2 below, the percentages of the incident among each time range under a given predict error were listed. It’s obviously that a large error happened when we predict the long incident duration. To sum up, the small predict error points are distributing on the short time range, while the relative big predict error points are among the long-time range. At the same time, the statistical data shows us a minimum accuracy when we predict
all the data together. Thus, it is better to build one prediction model for each different type incident data.

Table 2. Prediction results of PLSR model

<table>
<thead>
<tr>
<th>Incident type</th>
<th>Stopped vehicle (%)</th>
<th>Lost-load (%)</th>
<th>Accident (%)</th>
<th>All (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean absolute error</td>
<td>15.95</td>
<td>12.70</td>
<td>13.51</td>
<td>16.54</td>
</tr>
<tr>
<td>(error less than 10min)</td>
<td>10min-30min</td>
<td>50.94</td>
<td>76.74</td>
<td>42.86</td>
</tr>
<tr>
<td></td>
<td>30min-60min</td>
<td>39.62</td>
<td>23.25</td>
<td>52.38</td>
</tr>
<tr>
<td></td>
<td>60min-90min</td>
<td>9.4</td>
<td>0</td>
<td>2.86</td>
</tr>
<tr>
<td>Total</td>
<td>44.09</td>
<td>52.44</td>
<td>47.30</td>
<td>41.09</td>
</tr>
<tr>
<td>(error less than 20min)</td>
<td>10min-30min</td>
<td>51.58</td>
<td>71.83</td>
<td>46.49</td>
</tr>
<tr>
<td></td>
<td>30min-60min</td>
<td>40</td>
<td>28.17</td>
<td>44.32</td>
</tr>
<tr>
<td></td>
<td>60min-90min</td>
<td>8.42</td>
<td>0</td>
<td>7.57</td>
</tr>
<tr>
<td>Total</td>
<td>77.24</td>
<td>86.59</td>
<td>83.33</td>
<td>71.30</td>
</tr>
<tr>
<td>(error over than 30min)</td>
<td>10min-30min</td>
<td>12.5</td>
<td>33.3</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>30min-60min</td>
<td>6.25</td>
<td>16.67</td>
<td>8.33</td>
</tr>
<tr>
<td></td>
<td>60min-90min</td>
<td>43.75</td>
<td>50</td>
<td>41.67</td>
</tr>
<tr>
<td>Over 90min</td>
<td>37.50</td>
<td>0</td>
<td>50</td>
<td>30.65</td>
</tr>
<tr>
<td>Total</td>
<td>13</td>
<td>7.3</td>
<td>5.4</td>
<td>13.48</td>
</tr>
</tbody>
</table>

Liping, Weiming, Xiangyuan and Luping (2010) applied artificial neural network to predict incident duration. In their study, the prediction accuracy is 33% within 10 minutes error and 63% within 10 minutes error. Compared with their research, ours is better. Jiyang, Zhang and Sun (2008) built a model based on Bayesian Decision Method-Based Tree, and got a prediction accuracy of 79% for lost load incident within 10 minutes error and 65% for the accident within 20 minutes error. Compared with their results, our accuracy that within 20 minutes error is higher, while in the periods of short duration, our model’s performance is inferior to theirs. Wu, Chen and Zheng (2011) employed SVM to predict incident duration and gained a highest accuracy of 76.92% for the lost load incident. Comparing to his prediction results, we made a little improvement and offered a new alternative way to predict duration.

5. Conclusions

In this study, we established four models for the data based on PLSR and validated the models. The prediction accuracy is acceptable compared with the previous research. We found that it is better to analysis the data separately by the incident type. Though the accuracy between 20 minutes and 90 minutes is acceptable, however, it is must point out that some instances that over 90 minutes were underestimated. We summarize the reasons as follow: firstly, sample imbalance that the small and large incident duration instances are less than the middles; secondly, a large number of inconsistent records in the data set, that is, some instances have the same attribute record, but their duration is huge different from each other. Our further study will pay more attention to improving the quality of the data set. On the one hand, we will try to improve the quality of the data set by sampling technical. On the other hand, a large number of inconsistent records in the data set should be manipulated. Meanwhile, we will try more different methods to improve the predict accuracy of duration.
Acknowledgements

This research was supported by the National High Technology Research and Development Program of China (863 Program) (2012AA112304) and the National Natural Science Foundation of China (61074141).

References