Journal of the Eastern Asia Society for Transportation Studies, Vol. 6, pp. 2457 - 2472, 2005

# APPLYING DATA FUSION TECHNIQUES TO TRAVELER INFORMATION SERVICES IN HIGHWAY NETWORK

Chien-Hung WEI
Professor
Department of Transportation and Communication Management Science
National Cheng Kung University
#1, University RD., Tainan city,
701 Taiwan, R. O. C.
Phone: +886-6-2757575-53233
Fax: +886-6-2753882
E-mail: louiswei@mail.ncku.edu.tw

Ying LEE
Ph. D. Candidate
Department of Transportation and Communication Management Science
National Cheng Kung University
#1, University RD., Tainan city,
701 Taiwan, R. O. C.
Phone: +886-6-2757575-53270-5020
Fax: +886-6-2753882
E-mail: r5891101@ccmail.ncku.edu.tw

**Abstract:** Data fusion techniques are applied to traveler information services and used to build an accident duration estimation function. The accident duration is estimated at the initial occasion of an accident. The data fusion procedure clusters the values of each factor into a small number of intervals and effectively smoothes the data noise to the model. In most experiments, the mean absolute percentage errors of the estimated outputs are under 25%, indicating an acceptable forecasting effect. Through the factor sensitivity analysis, time of day, number of vehicles & vehicle type involved in accidents, and geometry have high significance in conducting the accident duration models. The results confirm that the data fusion techniques are practical and reliable for developing traveler information systems. This study is granted by National Science Council, Taiwan, under the project number NSC93-2218-E-006-094. It shows very promising practical applicability of the proposed models in the Intelligent Transportation Systems context.

**Key words:** Data fusion, Traveler information, Intelligent transportation systems, Accident duration.

## 1. INTRODUCTION

Traveler information services are main components of intelligent transportation systems (ITS), which aim at using information and strategies to raise the efficiency of transportation facility. The purpose of Traveler information services is providing the real-time information for the traveler to understand the traffic condition. According to the investigation, the information that most travelers expect to aware of is the accident information such as accident location, accident duration. Therefore, the accident duration is chosen as the first traveler information service to develop in this study.

Relying on the advanced technology, more traffic data can be collected more easily than before. Traffic pattern may be adequately characterized if these data can be processed and analyzed effectively. Therefore, people expect useful information relevant to their travels. This research applies data fusion technique to build an accident duration estimation function between the traffic data and accident duration. The purpose of this study is to provide real-time accident duration information at the initial accident period. Therefore, the model inputs only consider the available data at this period. Through this model, the estimated duration can be provided by plugging in the traffic data and the traveler and traffic manager can roughly fell the accident impact. In this study, the factor sensitive analysis is used to test the factor effect in duration model.

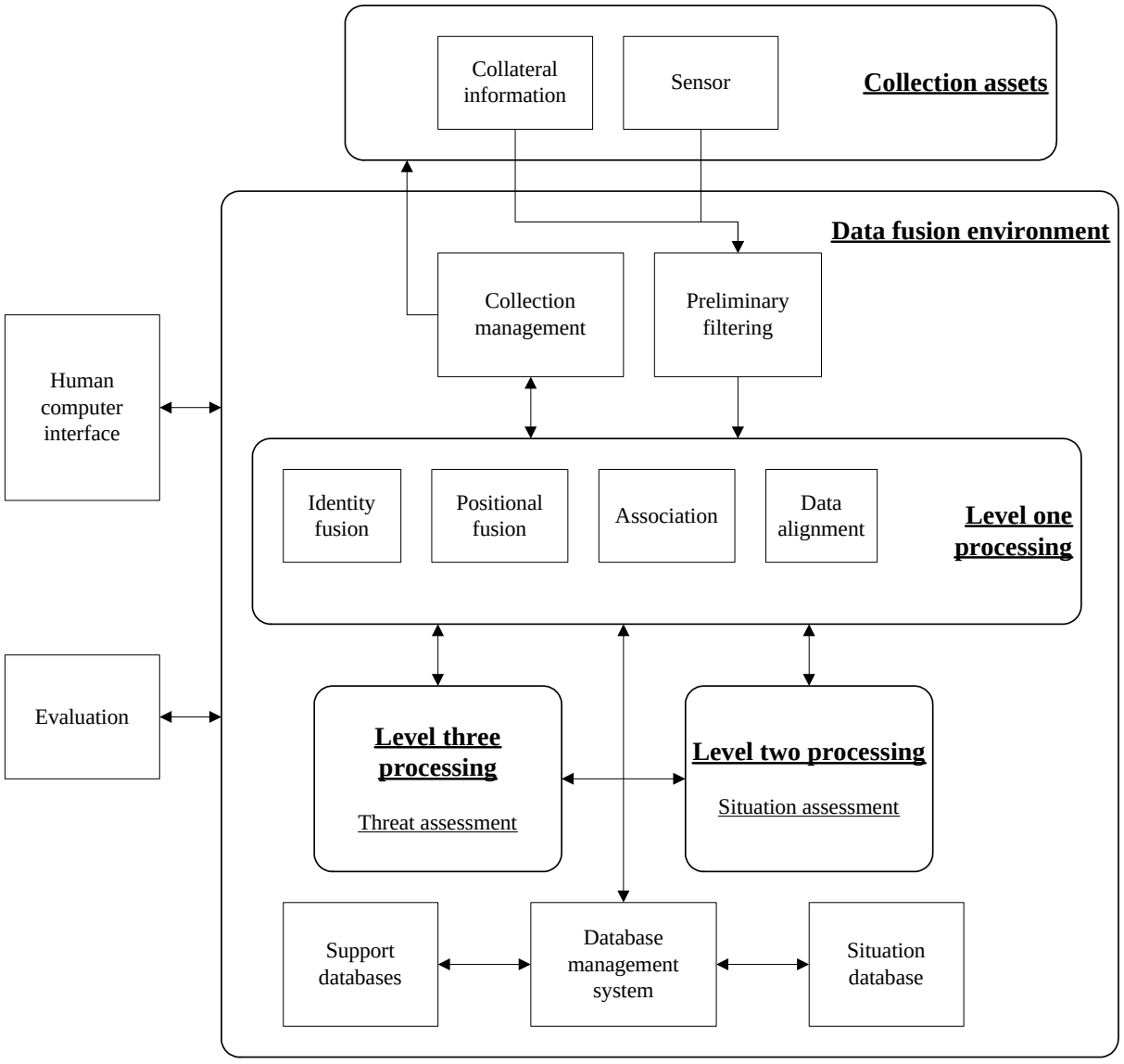
## 2. LITERATURE REVIEW

### 2.1 Data fusion

Data fusion is an evolving technology concerned with the problem of how to combine data from multiple sensors in order to make inferences about a physical event, activity, or situation [Hall, 1992]. The procedure of data fusion includes preliminary data filtering, data collections management, data alignment, association analysis, position fusion, identity fusion and situation assessment, as show in Figure 1.

First, the existing sources of data collection must be evaluated comprehensively. According to the needs of information service, the criteria of the data filtering and data alignment can be performed. Analyzing the association among the collected data usually use the parametric association method or the estimation technique. Through the data analysis, the suitable variables can be chosen for position fusion and identity fusion. The position fusion is to search the data which present the same situation. In the identity fusion, the mission is developing a model to fuse the searched data to the information. Physical models, feature-based inference techniques and cognitive-based models are the three major identity fusion approaches. In order to comprehend the reliability of the fused information, assessing the information accuracy will be included.

Data fusion and data grafting are concerned with combining data coming from different sources [Saporta, 2002]. This technique is not only used to extract information from an individual database, but also to merge the data collected in different databases from administrative sources. Even the data of these databases are composed of different statistical units or levels. This research applies a data fusion technique to develop the accident duration estimating models for the traveler information service in Taiwan highway.




[Hall, 1992]

Figure 1. A Process Model for Data Fusion

## 2.2 Accident duration

The accident duration is usually defined as the time between incident occurrence and roadway clearance. This duration can be divided into three parts, reporting time (the time between accident occurs and accident notification), response time (the time between accident notification and the rescue people arrival) and clearance time (the time between the rescue agency arrival and the accident cleared). This division is illustrated in Figure 2. The accident duration can be forecasted at the one of three moments, the accident occurrence, the accident notification or the rescue agency arrival.

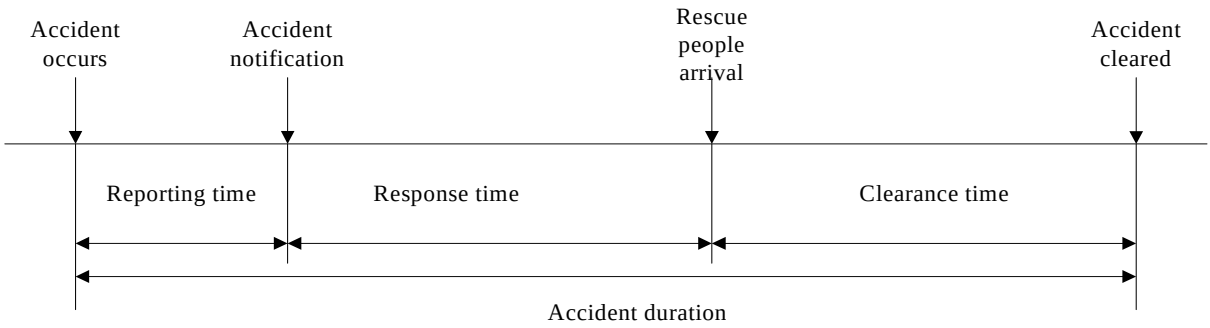



Figure 2. The Components of Accident Duration

Garib et al. [1997] employed the multiple regression analysis to predict the clearance time of the accident. The model show that 81% of variation in accident duration can be predicted by number of lanes affected, number of vehicles involved, truck involvement, time of day, police response time, and weather condition. But the accident duration before rescue people arrival can not be estimated in this study.

Nam and Mannering [2000] apply hazard-based analysis to build three accident duration models from the three moments, the accident occurrence, the accident notification and the rescue people arrival. The model estimation results show that a wide variety factors significantly affect accident duration. At the accident occurrence, the Month, Rain, Geographic information and Accident characteristics (Fatal; Injury) are significant variables in forecasting duration. At the moment of the accident notification, the significant variables are Temporal characteristics (Peak time; Night; Saturday; Sunday) and Geographic information. Usually, the agency who first arrive the accident site is not usually the same in every accident. The third model, forecasting the duration at the rescue agency arrival, consider the lead agency as the model input. The result indicates that the lead agency is significant. The forecast of three models can cover the whole accident period, but the forecast is not continuous from the star of prediction to the accident cleared. In traveler information services, forecasting duration by stage is an unsuitable way for road user.

The previous researches point out several highly significant factors for us in conducting an accident duration model. Differing from the hazard-based model and the regression approaches, this study try to employ the data fusion techniques in modeling. We consider the needs of road user and forecast the accident duration from the accident notification to the accident cleared.

## 3. MODEL CONCEPT

In this study, the data fusion techniques are applied to build the accident duration estimation model. The model intends to provide the real-time information as the moment of accident occurrence. Therefore, the model inputs adopt the available traffic data at the moment. Considering the needs of road user, the accident duration forecasted in this study is the time from accident notification to accident cleared. Artificial neural networks are chosen as the key analytical techniques to found the relation function between the traffic data as the model inputs and the accident duration as the output variable. In order to find out the contribution of each variable to the model effect, the factor sensitivity analysis is one of our study points. Figure 3 shows our modeling.

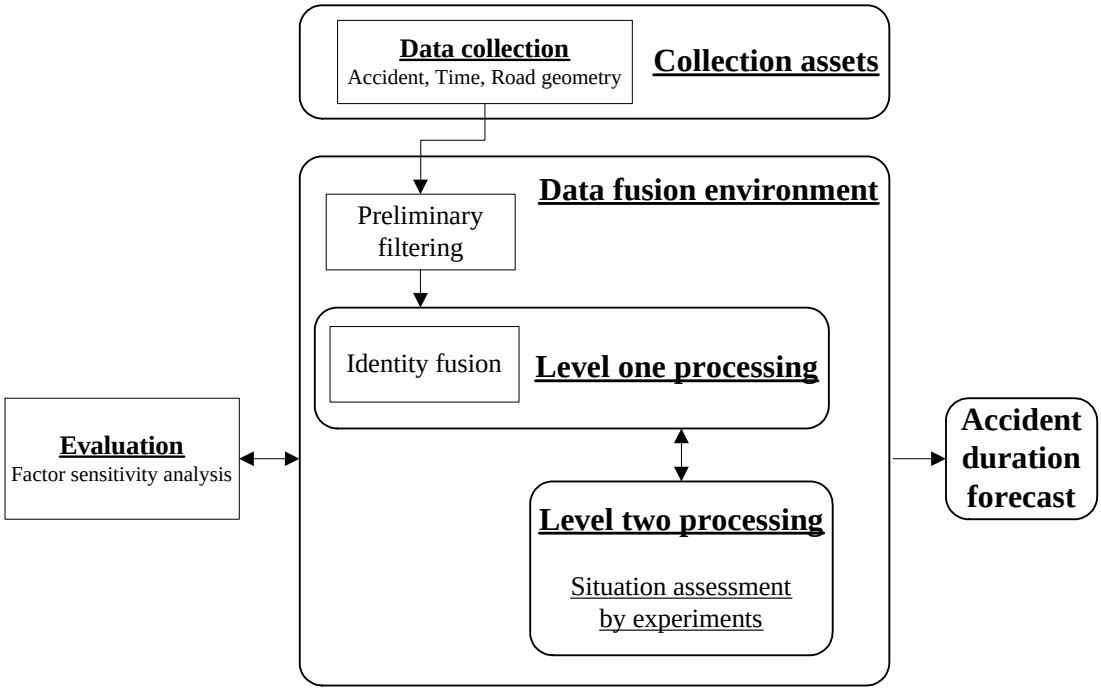



Figure 3. Concept of Accident Duration Model

## 4. DATA PROCESSES

In this research, the model is planed to be implemented at the moment of accident notification. Therefore, this model inputs is dependent on the resources available at this moment. The time factor, the road geometry characteristics and the accident characteristics reported at accident notification are the candidate inputs in this study. The accident raw data were collected 24 hours a day, over a 9-day period in February, 2002. The time period covers weekdays and weekends, peak hours and non-peak hours, which adequately reflects real traffic situations. The chosen geographical site is the southern part of the No.1 national freeway in Taiwan, south bound, from the Tai-chung interchange to the Kang-shan interchange, as shown in Table 1.

Table 1. Geographical Site

| Facility | Mileage(km) | Distance to upstream interchange(km) |
|---|---|---|
| Taic-hung IC. | 178.6 | - |
| Nan-tun IC. | 181.4 | 2.8 |
| Wang-tian IC. | 189.4 | 7.0 |
| Chang-hua IC. | 198.4 | 9.0 |
| Chang-hua SIC. | 208.0 | 9.6 |
| Yuan-lin IC. | 211.0 | 3.0 |
| Yuan-lin T.S. | 218.2 | - |
| Bei-dou IC. | 220.1 | 9.1 |
| Si-luo S.A. | 229.6 | - |
| Si-luo IC. | 230.5 | 10.4 |
| Dou-nan IC. | 240.6 | 10.1 |
| Dou-nan T.S. | 246.7 | - |
| Da-lin IC. | 250.3 | 9.7 |
| Chia-yi IC. | 264.3 | 14.0 |
| Shuei-shang IC. | 270.4 | 6.1 |
| Sin-ying T.S. | 280.7 | 10.3 |
| Sin-ying S.A. | 284.2 | 3.5 |
| Sin-ying IC. | 288.4 | 4.2 |
| Ma-dou IC. | 303.7 | 15.3 |
| An-ding IC. | 311.1 | 7.4 |
| Sin-shih T.S. | 313.6 | - |
| Tai-nan SIC. | 315.5 | 4.4 |
| Yong-kang IC. | 319.6 | 4.1 |
| Tai-nan IC. | 327.4 | 7.8 |
| Ren-de S.A. | 335.1 | - |
| Lu-jhu IC. | 338.3 | 10.9 |
| Gang-shan T.S. | 346.8 | - |
| Gang-shan IC. | 349.4 | 11.1 |

IC.: Interchange, SIC.: System interchange, T.S.: Toll station, S.A.: Service area

### 4.1 Data preprocesses

The incident data are updated every 3 minutes. Each incident is shown repeatedly in the database until it is cleared away from the traffic lane. The time of accident notification, the accident characteristic and accident location are record.

#### 4.1.1 Data calculation

The duration time of each accident can be calculated from each first and last record.

#### 4.1.2 Data filtering

There are two missions in the data filtering. One is to cancel the records that accident duration is less than 9 minutes. The duration of these accident are very short, so they are not necessary to be predicted in the model. The other mission is to make sure that each accident only exist one record in the accident database. The exceeding record will be cancel or merge into one.

## 5. DATA FUSION

Summarizing the accident researches and our domain knowledge, accident duration is affected by the following factors, time of day, day of week, lane occupied by accident, opportunity of vehicle over-turned, number of vehicles and vehicle type involved in accident, the distance from the accident location to the neighboring interchanges. In this section, the identity declaration of these candidate factors and the identity fusion of the duration model are addressed.

### 5.1 Identity declaration

By identity declaration, we can realize the relationship between the duration and the candidate factors. On the other hand, the values of each candidate factor can be clustered into a small number of intervals. This process not only can reduce the data values and the model noise, but also can preserve the character of the original data.

#### 5.1.1 Time of day

Observing the relationship between time of day and the duration in Figure 4, the variance of accident duration is high during 12:00 p.m. to 15:00 p.m. and 17:00 p.m. to 21:30 p.m. The duration spread out between 4000 sec high and 1000 sec low, therefore these periods are declared as the high variance period. The medium variance period is during 9:00 a.m. to 12:00 a.m. Other periods, the duration is about 1500 sec. These periods are declared as the low variance period.

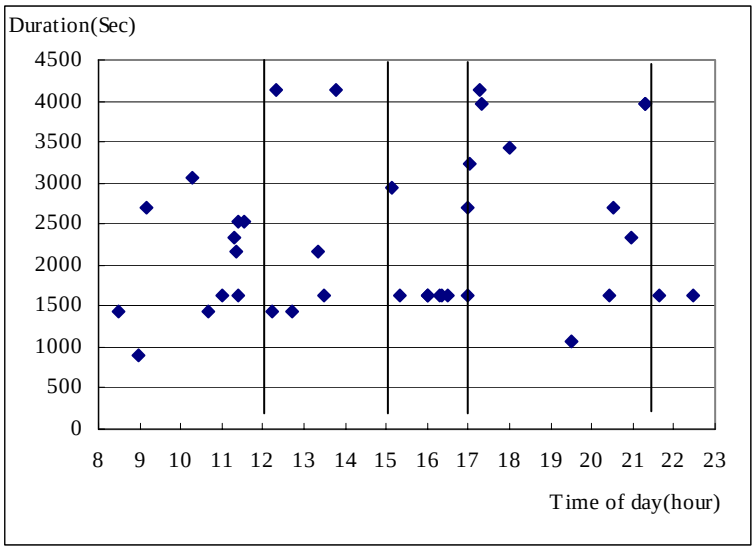



Figure4. The Relationship between the Time of Day and the Duration

#### 5.1.2 Day of week

Figure 5 shows the scatter plots of day of week and duration. From day of week, the trends of duration cannot be generalized. Therefore, the day of week is not considered as the model input.

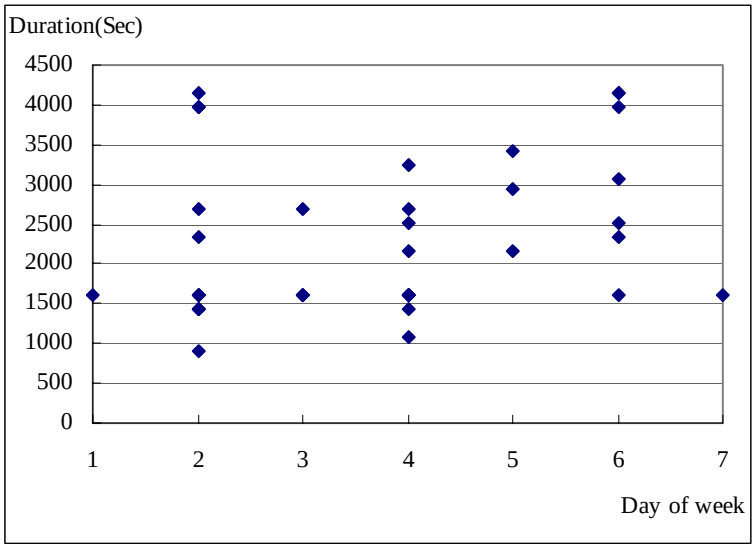



Figure 5. The Relationship between the Day of Week and the Duration

### 5.1.3 Occupied lane

When the accident is reported, lane occupied by accident is one of the important features which are easy to be describe. Table 2 shows five conditions of the occupied lanes and the mean and standard deviation of duration. When only the inside lane or outside lane be occupied in accident, the mean of duration are very close to each other. We merge two conditions into one and consider one lane be occupied. Based on the same reason, the conditions that inside and outside lane be occupied and outside lane and shoulder be occupied are merged and declared two lane be occupied. Figure 6 shows the duration of three conditions after merging.

Table 2. The Duration in Five Conditions of Occupied Lanes

| Occupied lanes | Inside | Inside & Outside | Outside | Outside & Shoulder | Shoulder |
|---|---|---|---|---|---|
| Mean of duration (Sec) | 2047 | 2700 | 1800 | 2430 | 1440 |
| Std. of duration (Sec) | 687 | 1199 | 673 | 718 | 311 |

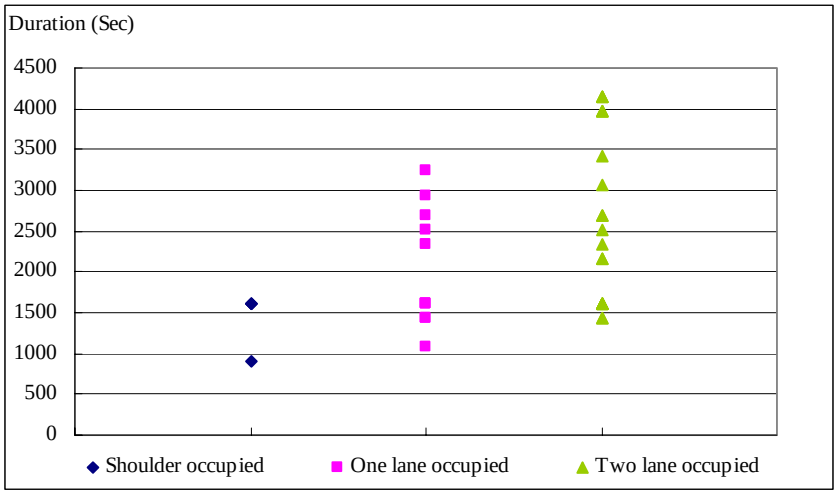



Figure 6. The Relationship between Occupied Lane and the Duration

### 5.1.4 Over-turn vehicles involved

In our study samples, there are nine accidents with duration over 3000sec. Among these nine instances, seven accidents involve the over-turn vehicles.

### 5.1.5 Number of vehicles and vehicle type involved

The relation among the number of vehicles, vehicle type and the duration is displayed in Figure 7. In the figure, vertical is the number of passenger cars in an accident, horizontal is the number of non-passenger cars in an accident and the color represents the accident duration. Each accident is plotted in the figure according to the number of passenger car and non-passenger car involved. As the Figure 7 showed, the more the number of passenger car is involved in each accident, the longer the accident duration is.

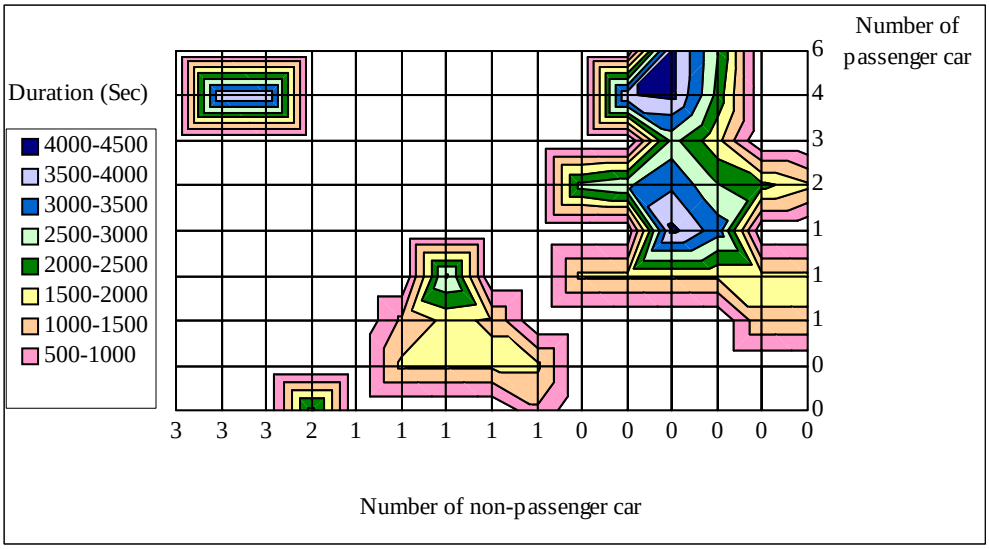



Figure 7. The Relationship among Number of Vehicles, Vehicle Type and the Duration

### 5.1.6 Geometry

According to the geometry, the highway can be classified roughly as interchanger area, service facility area and others area. If the accident locates in the front and rear of the interchange or the service facility within 1.5 km, the accident belongs to the corresponding area. The service facility includes the toll station and the service area. The relationship between three classified areas and the duration is shown in Figure 8. Among these three areas, the distribution of the duration is different from each other.

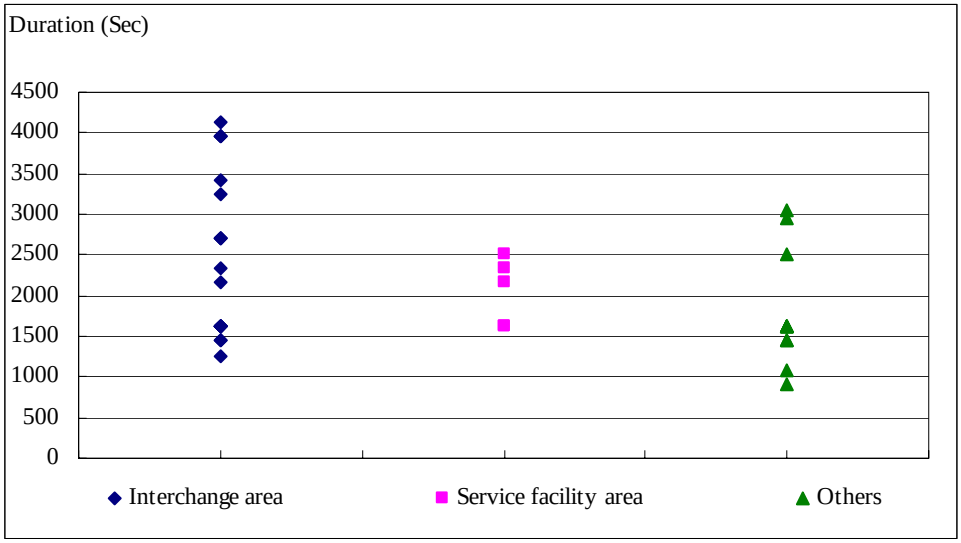



Figure 8. The Relationship between the Highway Geometry and the Duration

### 5.1.7 Distance to the interchange

The distance from the accident location to the interchange is displayed in Figure 9. Vertical is the distance between the accident location and the nearest downstream interchange, horizontal is the distance between the accident location and the nearest upstream interchange. The accident duration are showed by the different color. In the Figure 9, every point represents an accident location and its accident duration. When the distance from the accident location to the interchange is between 4 and 7 km, the accident duration usually is longer. These accidents are marked with the circles in the Figure 9.

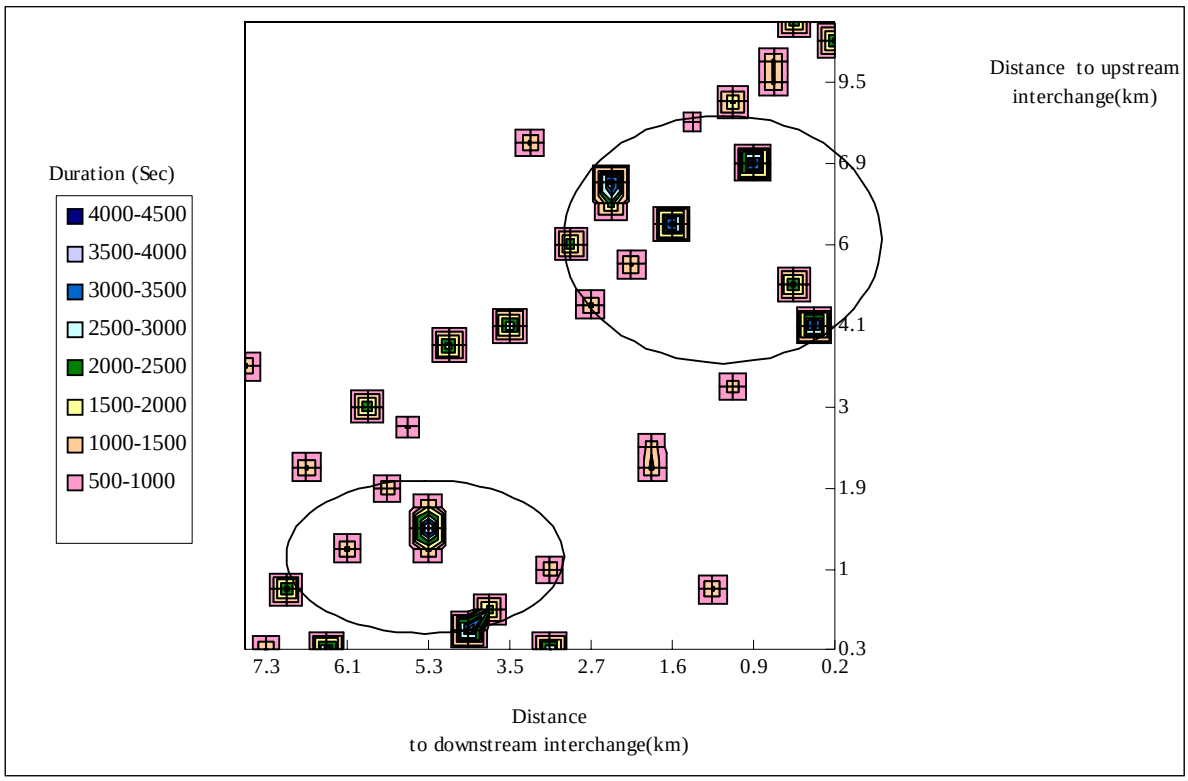



Figure 9. The Relationship between the Distance from the Interchange and the Duration

### 5.2 Identity fusion

After the data analysis and the identity declaration process in Section 5.1, the values of each factor we choose are cluster into a small number of intervals. The suitable variables considered into our model for position fusion are time of day, lane occupied by accident, opportunity of vehicle over-turned, number of vehicles and vehicle type involved, the distance from the accident location to the interchange. In this fusion stage, artificial neural networks (ANNs) have been chosen as the key technique. The ANNs approach is a data-driven, self-adaptive, nonlinear methodology. The model inputs and the structure are shown in Figure 10.

When inputting variables into the network, the weights from input layer to hidden layer are calculated. Through the transfer function in the hidden layer, the input data are rescaled as inputs to the output layer. The output is the estimated accident duration. Since a discrepancy might occur between the estimated output and the actual accident duration, the weights are adjusted repeatedly by a suitable training method until the variation of the resulting error is stabilized. The accident duration function is formed after this training procedure *(Zurada, 1992)*.

| Factor | Input statement | Input value |
|---|---|---|
| Time | P.M.12:00-15:00, 17:00-21:30 | 1:Yes, 0:No |
| | A.M.9:00-12:00 | 1:Yes, 0:No |
| | Other time | 1:Yes, 0:No |
| Occupied lane | Shoulder | 1:Yes, 0:No |
| | One lane | 1:Yes, 0:No |
| | Two lanes | 1:Yes, 0:No |
| Accident type | Over-turn | 1:Yes, 0:No |
| The number of vehicle and the vehicle type be involved in accident | Passenger car | Numbers |
| | Non-passenger car | Numbers |
| Geometry | Interchange area | 1:Yes, 0:No |
| | Service facility area | 1:Yes, 0:No |
| | Other areas | 1:Yes, 0:No |
| Distance to the interchange | The distance between the accident location and the nearest upstream interchange | Kilogram |
| | The distance between the accident location and the nearest downstream interchange | Kilogram |

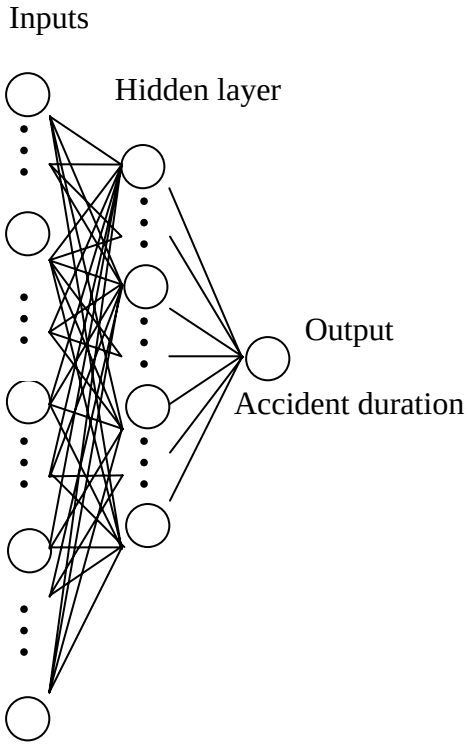



Figure 10. The Model Structure in Identity Fusion

After model training, the mapping between variables and accident duration is formed. Inputting the data from the tested examples into this trained model, the forecasted accident duration is produced. The mean absolute percentage error (MAPE) has been chosen for model evaluation, as shown in equation 1.

$$MAPE = \frac{1}{M}\sum_{k=1}^{M}\left|\frac{\hat{x}(k) - x(k)}{x(k)}\right| \times 100\% \quad \text{..... (1)}$$

$\hat{x}$ : Forecasted accident duration

$x$ : Actual accident duration

$M$ : Total numbers of examples

$k$ : The $k_{th}$ example

Typical MAPE values for data and their assessment are shown in Table 3 (Lewis, 1982). As the MAPE becomes closer to 0, the forecasted value becomes more accurate.

Table 3. Criteria of MAPE for Model Evaluation

| MAPE (%) | Assessment |
|---|---|
| <10 | Highly accurate forecasting |
| 10-20 | Good forecasting |
| 20-50 | Reasonable forecasting |
| >50 | Inaccurate forecasting |

### 5.3 Modeling and results

For our model development, 39 accident duration examples have been collected from a field data. We randomly sample 30 examples for model training, and the rest are used in model testing. The model is trained with the back-propagation algorithm that is the most popular ANN training method. Using the same sampling process, ten experiments are conducted. Table 4 shows the results with ten experiments. The MAPE values are between 20% and 30%. The model effect is acceptable and it indicates that the proposed data fusion procedure is practical and reliable for developing accident duration.

Table 4. Model Results with Ten Experiments

| Experiments | MAPE(%) |
|---|---|
| 1 | 19.57 |
| 2 | 27.42 |
| 3 | 32.83 |
| 4 | 23.16 |
| 5 | 19.17 |
| 6 | 31.36 |
| 7 | 15.84 |
| 8 | 22.12 |
| 9 | 23.25 |
| 10 | 21.21 |
| Mean | 23.59 |
| Std. | 5.14 |

## 6. FACTOR SENSITIVITY ANALYSIS

In Section 5, the duration model is conducted successfully with six factors. In order to compare the effect of each factor to the model, we leave one factor out and use the rest five factors in modeling. If the model which is leaving one factor out has a worse effect than the model with all six factors, it indicates that the leaved factor has a good contribution to estimate the accident duration. Table 5 is the models effect in factor sensitivity analysis with ten experiments. In terms of the means of MAPE for each leave-one-factor-out model, time of day, opportunity of vehicle be over-turn, number of vehicles & vehicle type involved and geometry are the important factors in estimating the duration.

Table 5. Model Effects in Factor Sensitivity Analysis (MAPE%)

| Experiment | No factor leaved out | Time of day | Occupied lane | Accident type | Number of vehicles & vehicle type involved | Geometry | Distance to interchange |
|---|---|---|---|---|---|---|---|
| # 1 | 19.57 | 37.19 | 26.10 | 18.63 | 25.57 | 27.94 | 19.95 |
| # 2 | 27.42 | 46.35 | 24.70 | 25.47 | 30.07 | 27.07 | 24.90 |
| # 3 | 32.83 | 49.20 | 24.56 | 29.33 | 26.13 | 36.15 | 24.90 |
| # 4 | 23.16 | 44.56 | 22.44 | 20.60 | 28.17 | 28.05 | 25.41 |
| # 5 | 19.17 | 37.28 | 22.22 | 23.76 | 22.15 | 21.47 | 18.59 |
| # 6 | 31.36 | 41.65 | 23.15 | 29.33 | 48.05 | 31.46 | 24.53 |
| # 7 | 15.84 | 39.96 | 17.45 | 20.17 | 12.49 | 20.18 | 15.68 |
| # 8 | 22.12 | 24.23 | 24.43 | 25.40 | 28.55 | 25.63 | 24.30 |
| # 9 | 23.25 | 49.49 | 22.13 | 30.15 | 24.82 | 8.58 | 18.84 |
| # 10 | 21.21 | 23.01 | 22.50 | 28.44 | 30.12 | 26.99 | 18.07 |
| Mean | 23.59 | 39.29 | 22.97 | 25.13 | 27.61 | 25.35 | 21.52 |
| Std. | 5.14 | 8.87 | 2.23 | 4.01 | 8.39 | 7.05 | 3.45 |

In addition to analyze the factor importance form the means of MAPE, we individually compare the effect between the models in every experiment. If the MAPE of the leave one factor out model is worse than the six factors model, we use the symbol “+” to represent that the leaved factor is good for our duration model. Observing each leave one factor out model, if the model effects are “+” more than 50% experiments, we define that the influence of the factor to the duration model is high significance. Through the factor sensitivity analysis, the comparison of model effects in every experiment is shown in Table 6. Time of day, number of vehicles & vehicle type involved and geometry are high significant factors to our duration model.

Table 6. Comparison of Model Effects in Every Experiment

| Experiment | Time of day | Occupied lane | Accident type | Number of vehicles & vehicle type involved | Geometry | Distance to interchange |
|---|---|---|---|---|---|---|
| # 1 | + | + | - | + | + | + |
| # 2 | + | - | - | + | - | - |
| # 3 | + | - | - | - | + | - |
| # 4 | + | - | - | + | + | + |
| # 5 | + | + | + | + | + | - |
| # 6 | + | - | - | + | + | - |
| # 7 | + | + | + | - | + | - |
| # 8 | + | + | + | + | + | + |
| # 9 | + | - | + | + | - | - |
| # 10 | + | + | + | + | + | - |
| Factor significance | High | Median | Median | High | High | Low |

According to the comparison in Table 6, we choose the three high significant factors to re-conduct the duration model and compare the model effect between this model and the six-factor model as show in Table 7. Model A is the model considering all six factors. Model B is conducted with three significant factors. In terms of the MAPE, Model B has the better effect. This result indicates that using the factors time of day, number of vehicles & vehicle type involved and geometry can effectively estimate the accident duration.

Table 7. The Effect Comparison by the Two Model (MAPE%)

| Model | A | B | Effect |
|---|---|---|---|
| Experiment 1 | 19.57 | 14.67 | Model B is good |
| Experiment 2 | 27.42 | 23.58 | Model B is good |
| Experiment 3 | 32.83 | 13.41 | Model B is good |
| Experiment 4 | 23.16 | 22.58 | Model B is good |
| Experiment 5 | 19.17 | 17.72 | Model B is good |
| Experiment 6 | 31.36 | 22.59 | Model B is good |
| Experiment 7 | 15.84 | 12.49 | Model B is good |
| Experiment 8 | 22.12 | 32.22 | Model A is good |
| Experiment 9 | 23.25 | 17.04 | Model B is good |
| Experiment 10 | 21.21 | 27.79 | Model A is good |
| Mean | 23.59 | 20.41 | Model B is good |
| Std. | 5.14 | 6.14 | |

Model A: Six factors; Model B: Three significant factors

## 7. CONCLUSION

In this research, we have developed the accident duration model by the data fusion techniques. Empirical study has shown acceptable model performance. The MAPE of forecasted results of the incident duration in the fusion model are mostly under 20%. This shows that the model fits the actual accident duration quite well. In terms of the model effect, data fusion techniques effectively smooth the data noise to the model. With this model, the estimated duration can be provided by plugging in relevant traffic data as soon as the incident is notified. The traveler and traffic manager can generally imagine the accident impact by the forecasted accident duration.

Through the identity fusion procedure, time of day, lane occupied by accident, opportunity of vehicle over-turned, number of vehicles & vehicle type involved and the distance to interchange are chosen as the suitable variables for modeling and clustered into a small number of intervals.

By the factor sensitivity analysis, the significant factors can be identified. The factors, time of day, number of vehicles & vehicle type involved in accident and geometry, have high significance in conducting the duration model.

The infrastructure is already in place to collect the traffic data for various purposes. This study uses the same sources of data for extra applications. The cost of data collection for various factors in this study is not significantly different. The operation and maintenance cost might be considered in future works.

The results of this study confirm that the data fusion technique is practical and reliable for developing traveler information systems. This study shows very promising practical applicability of the proposed models in the Intelligent Transportation Systems (ITS) context.

## ACKNOWLEDGEMENTS

This paper was derived from a research project sponsored by the National Science Council, Taiwan under the contract of NSC93-2218-E-006-094. We are grateful to the Taiwan Area National Freeway Bureau and Polices Radio Station for providing relevant data for this research.

## REFERENCE

1. Garib, A., Radwan, A. E. and Al-Deek, H. (1997) Estimating magnitude and duration of incident delays, **Journal of Transportation Engineering**, Vol.123, No.6, 459-466.
2. Hall, D. L. (1992) **Mathematical techniques in Multisensor data fusion**. Artech House, Boston.
3. Lewis, C. D. (1982) **Industrial and Business Forecasting Method**. Butter worth Scientific, London.
4. Nam, D. and Mannering, F. (2000) An exploratory hazard-based analysis of highway incident duration. **Transportation research part A**, Vol.34, No.2, 85-102.
5. Saporta, G. (2002) Data fusion and data grafting. **Computational statistics & data analysis**, Vol.38, No.4, 465-473.