Journal of Zhejiang University-SCIENCE A (Applied Physics & Engineering)
ISSN 1673-565X (Print); ISSN 1862-1775 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Short-term traffic safety forecasting using Gaussian mixture model and Kalman filter*

Sheng JIN†1, Dian-hai WANG1, Cheng XU2, Dong-fang MA†‡1
(1College of Civil Engineering and Architecture, Zhejiang University, Hangzhou 310058, China)
(2Zhejiang Police College, Hangzhou 310053, China)
†E-mail: jinsheng@zju.edu.cn; mdf2004@163.com

Received Aug. 30, 2012; Revision accepted Jan. 18, 2013; Crosschecked Mar. 26, 2013

Abstract: In this paper, a prediction model is developed that combines a Gaussian mixture model (GMM) and a Kalman filter for online forecasting of traffic safety on expressways. Raw time-to-collision (TTC) samples are divided into two categories: those representing vehicles in risky situations and those in safe situations. Then, the GMM is used to model the bimodal distribution of the TTC samples, and the maximum likelihood (ML) estimation parameters of the TTC distribution are obtained using the expectation-maximization (EM) algorithm. We propose a new traffic safety indicator, named the proportion of exposure to traffic conflicts (PETTC), for assessing the risk and predicting the safety of expressway traffic. A Kalman filter is applied to forecast the short-term safety indicator, PETTC, and solves the online safety prediction problem. A dataset collected from four different expressway locations is used for performance estimation. The test results demonstrate the precision and robustness of the prediction model under different traffic conditions and using different datasets. These results could help decision-makers to improve their online traffic safety forecasting and enable the optimal operation of expressway traffic management systems.

Key words: Forecasting, Traffic safety, Gaussian mixture model, Kalman filter
doi:10.1631/jzus.A1200218 **Document code:** A **CLC number:** U491

1 Introduction

Road traffic accidents are one of the world's largest public health and injury problems. According to the World Health Organization (WHO), more than a million people are killed on the world's roads each year. Thus, the measurement, assessment and forecasting of traffic safety are very important topics that may be applied in transportation planning, operation, and management (Meng *et al*., 2011a; 2011b). Traffic safety is most commonly measured in terms of the number of traffic accidents and the consequences of those accidents in terms of fatalities and injuries of differing severities (Qu *et al*., 2011). This is generally regarded as a reactive rather than a proactive approach in that it is based on the collection and analysis of accident data by the police and other authorities after the accidents have actually happened. Because accidents occur randomly in time and space, traffic safety is a particularly difficult phenomenon to study, thereby making its measurement, assessment and comparison difficult as well (Archer, 2004). For traffic planning purposes, levels of traffic safety for a given traffic site may be predicted based on long-term historical accident data. However, for the purposes of traffic management, short-term traffic safety forecasting is very limited by the coverage, quality and usefulness of the accident data. Thus, the underlying principle for a more effective short-term traffic safety evaluation strategy is to develop proximal safety indicators that represent the temporal and spatial proximity characteristics of unsafe interactions and near-accidents.

‡ Corresponding author

* Project (No. 2011AA110304) supported by the National High-Tech R&D Program of China (863 program)

These “safety indicators” are usually defined as traffic measures that are statistically correlated with the number of road traffic accidents at a particular location. These values are based on the temporal and spatial proximity between road users during safety-critical events. Svensson (1998) stated that for proxy measures or indicators of safety to be useful they must (a) complement accident data and be more frequent than accidents, and (b) have the characteristics of “near-accidents” in a hierarchical continuum that describes all severity levels of road-user interactions, with accidents at the highest level and very safe passages with a minimum of interaction at the lowest level. In the literature, many safety indicators have been applied as traffic safety performance measures (Guido *et al.*, 2011; McAndrews, 2011). The maximum deceleration rate to avoid a crash (DRAC), defined as the difference in speeds between a following vehicle (FV) and its corresponding leading vehicle (LV) divided by their closing time, was first proposed by Almquist *et al.* (1991). The proportion of stopping distance (PSD) was defined by Allen *et al.* (1978) as the ratio of the remaining distance to the point of collision to the driver’s minimum acceptable stopping distance. The deceleration rate (DR) is simply a measure of the highest rate at which a vehicle must decelerate to avoid a collision (FHWA, 2003). The time-to-collision (TTC) was defined by Hayward (1971) as the time interval that separates a given FV from its corresponding LV, where their respective speeds are such that the vehicles are closing in on each other. A further variation around the TTC measure is found in relation to pedestrian (zebra) crossings. The time-to-zebra (TTZ) value was used by Várhelyi (1996) to assess the frequency and severity of critical encounters between vehicles approaching a pedestrian crossing and pedestrians crossing from either the left or right side of the road. Minderhoud and Bovy (2001) considered a more complete and comprehensive analysis of safety performance than the conventional TTC, and proposed two new safety indicators for comparative road traffic safety analyses. The first indicator, time exposed TTC (TET), measures the length of time for which all vehicles involved in a conflict are below a designated TTC minimum threshold. The second indicator, time integrated TTC (TIT), is based on summing the integral of the TTC profile and provides a more qualitative measure of safety for the study period. Another measure similar to the TTC concept is post-encroachment time (PET). This is used to measure the frequency and severity of situations where two road users pass over a common spatial point or area with a temporal difference that is below a stated threshold value, usually between 1–2 s (van der Horst and Kraay, 1986; Hydén, 1996; Topp, 1998). Cunto and Saccomanno (2007) introduced the crash potential index (CPI), expressed as the probability that the DRAC for an individual FV exceeds the vehicle’s braking capability or the maximum available deceleration rate (MADR). However, all of the safety performance indicators above have to be captured by video for individual vehicles, and are thus difficult to obtain directly from traffic management systems. The purposes of short-term traffic safety forecasting are to calculate an integrated indicator for traffic flow in a short time interval, and to predict the indicator in the next time interval. Therefore, to overcome the above problem, we consider the distribution characteristics of TTC samples and propose a new traffic safety indicator for safety performance forecasting.

The rest of the paper is organized into four sections. The second section briefly describes the definition of TTC, the collection of the TTC data, and the basic statistical parameters of TTC. In Section 3, a Gaussian mixture model (GMM) is built to establish the distribution of TTC data. Section 4 describes in detail the improvement and development of a novel online method of forecasting traffic safety, using the GMM and a Kalman filter. The last section concludes the paper with a summary of our findings.

2 TTC data collection and characteristics

2.1 Definition of TTC

TTC is widely accepted as a highly useful and valid safety indicator for traffic conflicts at intersections or on highways (Vogel, 2003). TTC was defined by Hayward (1971) as “the time that remains until a collision between two vehicles will occur if the collision course and speed difference are maintained”. The TTC between two consecutive vehicles is a common traffic parameter applied in safety estimation, obstacle avoidance, the design of collision warning systems, and driving behavior modeling (Jin *et al.*, 2011a; 2012). According to Svensson (1998), TTC is inversely related to accident risk (smaller TTC values indicate higher accident risks, and vice versa), and can be used for safety forecasting.

TTC can be defined as the range between an FV and an LV, divided by the relative velocity between the two consecutive vehicles at a particular time. Thus, it is expressed as follows:

$$\mathrm{TTC}_i(t)=\begin{cases}\dfrac{x_{i-1}(t)-x_i(t)-\mathrm{VL}_{i-1}}{v_i(t)-v_{i-1}(t)}, v_i(t)>v_{i-1}(t), \\ \infty, \qquad\qquad\qquad\qquad \text{otherwise,}\end{cases} \tag{1}$$

where $\mathrm{TTC}_i(t)$ is the TTC of FV i at time t, $x_{i-1}(t)$ and $x_i(t)$ are the positions of the LV $i-1$ and the FV i at time t, respectively, $v_{i-1}(t)$ and $v_i(t)$ are the speeds of the LV $i-1$ and the FV i at time t, respectively, and VL_{i-1} is the length of the LV $i-1$.

It is difficult for traffic management systems to capture the positions and speeds of both the FV and the LV at a particular time, especially when the distance between two consecutive vehicles is large. That is to say, we cannot obtain TTC using Eq. (1). Therefore, TTC has to be calculated through fixed station traffic parameters, which can be measured easily by traffic management systems. Assuming that vehicles have a consistent travel speed through a fixed station in a short time interval, the distance headway $x_{i-1}(t)-x_i(t)$ while the FV is moving through the detector can be estimated by the FV speed multiplied by the time headway (Vogel, 2003). Then, the expression of TTC in the car-following scenario can be rewritten as

$$\mathrm{TTC}_i=\frac{\mathrm{TH}_i\cdot v_i-\mathrm{VL}_{i-1}}{v_i-v_{i-1}}, \quad \forall v_i>v_{i-1}, \tag{2}$$

where $\mathrm{TH}_i=T_i-T_{i-1}$ is the time headway between the two vehicles, where T_i and T_{i-1} are the times when the LV and the FV travel through the station, respectively.

2.2 Data collection

The field TTC data analyzed in this study were collected by traffic management systems between Sitong Bridge and Lianxiang Bridge on the Beijing North Ring III expressway, China. The overhead Ring expressway is a vital infrastructure in Beijing's road system, and has a total length of 48 km. The data were collected from four segment stations over a whole day, Tuesday, June 21, 2011. Station 1 is located on the mainline upstream of an on-ramp (direction W→E); station 2, in a weaving segment downstream of an on-ramp (direction W→E); station 3, on the mainline downstream of an off-ramp (direction W→E); and station 4, in a weaving segment downstream of an on-ramp (direction W←E) (Fig. 1). All segments of the selected expressway are three-lane roads, and the speed limit is 80 km/h. The datasets used to calculate the TTC values included every vehicle's speed, time headway, and length.

2.3 TTC characteristics

To analyze the fundamental characteristics of the TTC samples, we used the raw traffic data from 6:00 am to 9:00 am (including morning peak hour and low volume periods) to calculate TTC values. Some fundamental statistical results for the four segments' TTC values are shown in Table 1. The results show

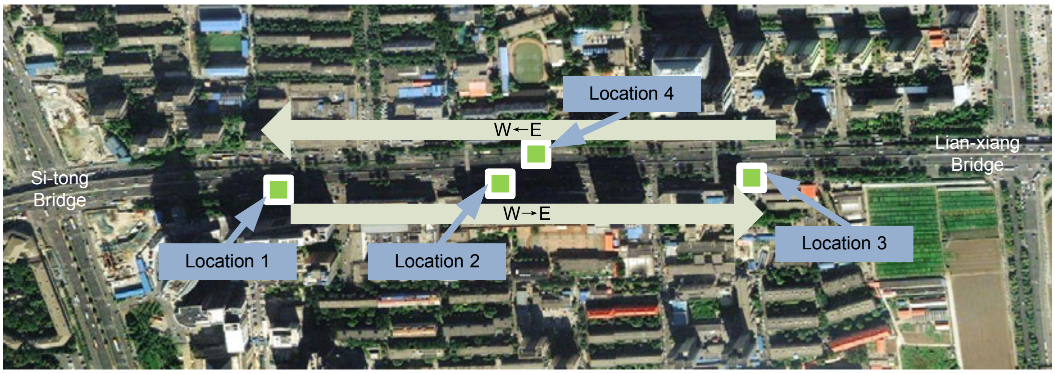

Fig. 1 Data collection sites on the Beijing North Ring III expressway, China

that 9101 TTC samples were collected with respect to the different road environments and traffic conditions. Each segment provided nearly 2275 TTC samples for the analysis. The TTC values in the different segments cover a very wide range, from 0.51 s to 69.82 s, and the standard deviations are large. The data clearly contain both low and high values of TTC, and even the very high values of TTC were included in the calculation of the means and variances. These statistics have more to do with traffic volume than traffic safety. For the purpose of traffic safety forecasting, we need to focus on the low TTC values that are significantly correlated with traffic conflicts.

To determine which TTCs should be included in our safety analysis, the proportion of short TTCs and the shape of the TTC distribution should be determined. Therefore, a common histogram method can be used first, with one predetermined bin-width. The simple histogram can provide brief but significant information about the TTC distribution. Fig. 2 illustrates the TTC distributions for two different segments. These examples indicate that the characteristics and variability of TTC distributions on different segments of a given expressway may have the same trends or patterns. Also, both samples seem to have multimodal and long-tailed distributions. This means that the TTC data may come from different unimodal distributions related to different levels of safety. One part of the data is concentrated around low TTC values (risky situations), while the other part is concentrated around high TTC values (safe situations). For the evaluation and prediction of traffic safety, we are more concerned about the TTC samples representing risky situations.

3 GMM-based distribution of TTC

In previous studies, TTC data have generally been described using a conventional probability density function (PDF), such as a normal, exponential, lognormal, gamma or inverse Gaussian distribution (Lord and Mannering, 2010). However, field TTC data from expressways have been shown to follow bimodal or multimodal distributions (Jin *et al.*, 2011b). From the point of view of safety forecasting, TTC data representing risky and safe situations should follow different distributions, and there is no specific distribution function available to represent mixed TTC data covering both types of situation. Therefore, we need to use a more sophisticated method to model our TTC distribution.

Table 1 Statistical description of the selected expressway segments

Station No.	Number of vehicles	Traffic volume (vehicles/h/lane)	TTC (s)			
			Max (s)	Min (s)	Mean (s)	SD (s)
1	2332	1217	68.91	0.51	16.65	12.52
2	2286	1218	69.82	0.63	14.64	11.17
3	2211	1232	60.36	0.83	16.82	11.85
4	2272	1044	63.42	0.62	16.34	12.21
Average	2275	1178	65.63	0.65	16.11	11.94

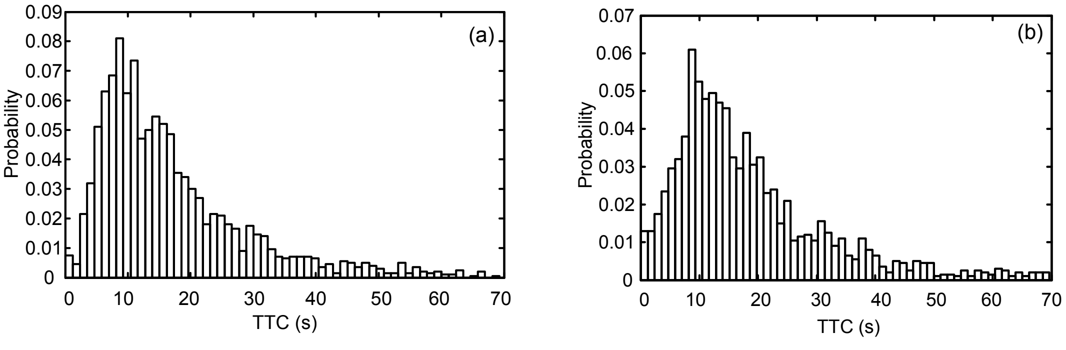

Fig. 2 TTC distributions of two different segments of expressway: (a) station 1 and (b) station 2

To estimate the PDFs (or parameters) of a mixed TTC distribution, the mixture model is a useful tool. The GMM, which is the most popular mixed distribution due to the simplicity of its estimation process, is a parametric PDF represented as a weighted sum of Gaussian component densities. The GMM can be represented by any type of probability distribution, and has been used widely for density estimation in computational, mathematical and optimization contexts. GMMs have been used successfully in a wide variety of fields, such as speaker recognition systems (Hsieh *et al.* 2003), video image processing (Stauffer and Grimson, 1999), pattern classification (Kim and Kang, 2007), and traffic safety parameter estimation (Jin *et al.*, 2011b).

3.1 Gaussian mixture model

A GMM for the TTC distribution can be formed as follows, as a weighted sum of I component Gaussian distributions (Titterington *et al.*, 1985):

$$P\left\{\text{TTC}\left|\left(\omega_i,\mu_i,\sigma_i^2\right)\right.\right\}=\sum_{i=1}^{I}\omega_i g(\text{TTC}|\mu_i,\sigma_i^2), \quad (3)$$

where $P\{\text{TTC}\}$ represents the probability that a specific value, TTC, occurs, ω_i, i=1, 2, …, I, are the mixture weights, and $g(\text{TTC}|\mu_i,\sigma_i^2)$, i=1, 2, …, I, with mean μ_i and variance σ_i^2, are the component Gaussian density functions. Each component density is a univariate Gaussian function of the form:

$$g(\text{TTC}|\mu_i,\sigma_i^2)=\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left\{-\frac{1}{2\sigma_i^2}\left(\text{TTC}-\mu_i\right)^2\right\}. \quad (4)$$

The mixture weights ω_i indicate the percentage of the TTC samples belonging to each category i and satisfy the constraint $\sum_{i=1}^{I}\omega_i=1$. The complete GMM is parameterized by the means, variances and mixture weights from all of the component Gaussian densities. These parameters are collectively represented by the notation $\theta=\left(\omega_i,\mu_i,\sigma_i^2\right)$.

One of the powerful attributes of the GMM is its ability to form smooth approximations of arbitrarily shaped densities. Due to their ability to represent a large class of sample distributions, GMMs can be used to analyze TTC data and capture the component Gaussian distribution patterns. The use of a GMM to represent feature distributions of TTC data may also be motivated by the intuitive notion that the individual component densities may model some underlying set of hidden classes. For example, in the safety context, it is reasonable to assume that the TTC values correspond to different situations. The choice of model configuration (number of components and model parameters) is often determined by the amount of field data available for estimating the GMM parameters, and the specific application.

In this study, TTC data could be classified into two categories (risky and safe) or three categories (such as high risk, medium risk and low risk), and a multimodal model could be developed accordingly based on similar methodology. The main contribution of this study is to propose this bi/multimodal concept to analyze TTC data. Therefore, for simplicity, two separate underlying regimes, leading to a mixture of two different Gaussian density distributions, were considered. The first mixture component, representing the low TTC values, depicts risky situations, while the second, representing the high TTC values, covers safe situations. Thus, in the case of this two-component GMM, the parameter I equals 2, the values of the mixture weights are associated with the safety of the traffic flow, and their sum should be 1.

3.2 Maximum likelihood parameter estimation

For the two-component GMM, five parameters need to be estimated. Given our TTC training samples, we wish to estimate the parameters of the GMM, θ, which in some sense best match the distribution of the training samples. There are several techniques available for estimating the parameters of a GMM (McLachlan, 1988). By far the most popular and well-established method is the maximum likelihood (ML) estimation. The aim of ML estimation is to find the model parameters that maximize the likelihood function of the GMM, given the training data. For a sequence of N TTC$_j$ taken from the training data, the GMM likelihood function, assuming the independence of the TTC$_j$, can be written as

$$L(\theta|\text{TTC}) = \log\prod_{j=1}^{N} P(\text{TTC}_j|\theta) = \sum_{j=1}^{N}\log P(\text{TTC}_j|\theta). \tag{5}$$

The function $L(\theta|\text{TTC})$ is referred to as the likelihood of the parameters given the data, or simply the likelihood function. It is a function of the parameters θ for a fixed TTC. In the ML problem, the goal is to find the θ that maximizes L. That is, we wish to find θ^* where,

$$\theta^* = \arg\max_{\theta} L(\theta|\text{TTC}). \tag{6}$$

Unfortunately, this expression is a non-linear function of the parameters θ and direct maximization is not possible. However, a special case of the expectation-maximization (EM) algorithm can be obtained to estimate the ML parameters. The EM algorithm (Dempster *et al.*, 1977) is a general method for finding the ML estimate of the parameters of an underlying distribution from a given dataset when the data is incomplete or has missing values. The basic idea of the EM algorithm is, beginning with initial parameters θ, to estimate new parameters $\bar{\theta}$, such that $P(\text{TTC}|\bar{\theta}) > P(\text{TTC}|\theta)$. The new parameters then become the initial parameters for the next iteration and the process is repeated until some convergence threshold or iteration number is reached.

The EM algorithm first finds the expected value of the complete data log-likelihood. The evaluation of this expectation is called the E-step of the algorithm, and the second step (the M-step) maximizes the expectation computed in the first step. These two steps are repeated as necessary. Each iteration process is guaranteed to increase the log-likelihood, and the algorithm is guaranteed to converge to a local maximum of the likelihood function. For detailed descriptions of the EM algorithm, please refer to Redner and Walker (1984) and Jordan and Jacobs (1994).

3.3 Results

In this study, the GMM parameters are estimated from the TTC training data using the iterative EM algorithm, with the number of components *I* set to 2 in the empirical observation and analysis. The two categories of TTC data represent risky and safe situations, respectively. Two normal distributions are applied to fit the TTC data and the mixture weight of each distribution reflects the percentages of risky and safe situations in the traffic flow.

Using the field samples from the four stations mentioned earlier, Table 2 shows the number of iterations of the EM algorithm, the estimated Gaussian distribution parameters, and the results of the Kolmogorov-Smirnov (K-S) test for each station. Looking at the values of the weights in each situation, we can see that the risky situations make up between 68% and 75% of the samples. The mean TTC values in risky situations are much lower than those in safe situations. Therefore, the GMM has the ability to distinguish between TTC samples in different situations and describe the characteristics of each Gaussian distribution.

To verify further the fit of the results statistically, the K-S test was adopted as a goodness-of-fit test (Stephens, 1974). In statistics, the K-S test is a nonparametric test of the equality of continuous 1D probability distributions. It can be used to compare a sample with a reference probability distribution (one-sample K-S test), or to compare two samples (two-sample K-S test). The K-S statistic quantifies the distance between the empirical distribution function of the sample and the cumulative distribution function (CDF) of the reference distribution. The null distribution of this statistic is calculated under the null

Table 2 Results from estimation of the GMM distribution parameters and the K-S test

Station No.	Number of iterations	CPU time (s)	Risky situations			Safe situations			K-S test results
			ω_1	μ_1 (s)	σ_1 (s)	ω_2	μ_2 (s)	σ_3 (s)	
1	12	4.21	0.68	10.51	5.29	0.32	29.39	13.52	0
2	6	1.08	0.73	9.82	5.38	0.27	27.41	12.23	0
3	8	2.44	0.75	11.81	5.60	0.25	32.25	12.70	0
4	11	3.54	0.71	10.95	5.62	0.29	29.49	14.02	0

hypothesis that the sample is drawn from the reference distribution. As Table 2 shows, the K-S goodness-of-fit tests all suggested that the GMM performs well. The samples from the four stations are demonstrated to have been drawn from the GMM distribution at a statistical significance level of α=0.05.

Fig. 3 depicts the EM estimation results for the two component Gaussian density function. It shows that the GMM has the ability to fit a two-peak distribution to the TTC data. The GMM fits the empirical data very well, with a small error. Fig. 3a shows the trends in the TTC based on the Gaussian mixture distributions using the TTC sample data from station 1. The mean TTC value is 10.51 s in the risky situations and 29.39 s in the safe situations, a difference of 18.88 s, and the mixing weights are 0.68 for risky situations and 0.32 for safe situations. The TTC pattern appears to follow a bimodal distribution since the mean difference between the two TTC regimes is large. Similarly, for station 2 (Fig. 3b), there is a large difference (17.59 s) between the means (9.82 s with a 0.73 mixing weight for the risky situations and 27.41 s with a 0.27 mixing weight for the safe situations), again indicating a two-peak distribution. Figs. 3c and 3d show similar patterns.

4 Method of forecasting traffic safety

4.1 Proportion of exposure to traffic conflicts

In Section 3, the TTC distribution parameters representing risky situations, which are more significant for safety forecasting, have been extracted from the field data using GMM. However, we do not want to consider all of the TTC data relating to risky situations; instead, we should focus on the sample data that may lead to conflicts. Mainly for this reason, there is a need to develop the concept of the traffic safety indicator further and produce an assessment method that can be used indirectly to measure expressway traffic safety.

For the purpose of safety forecasting, we define a new safety indicator: the proportion of exposure to traffic conflicts (PETTCs) in a given time period (e.g., 5, 10, or 15 min). PETTC represents the proportion of vehicles exposed to dangerous scenarios (or critical encounters). Having obtained the TTC distributions for the four expressway segments (Section 3), the PETTC in a particular interval can be defined as follows:

$$p_{\text{ettc}} = \omega_{\text{rs}} \cdot g_{\text{rs}} \left[\text{TTC} \le \tau \right], \tag{7}$$

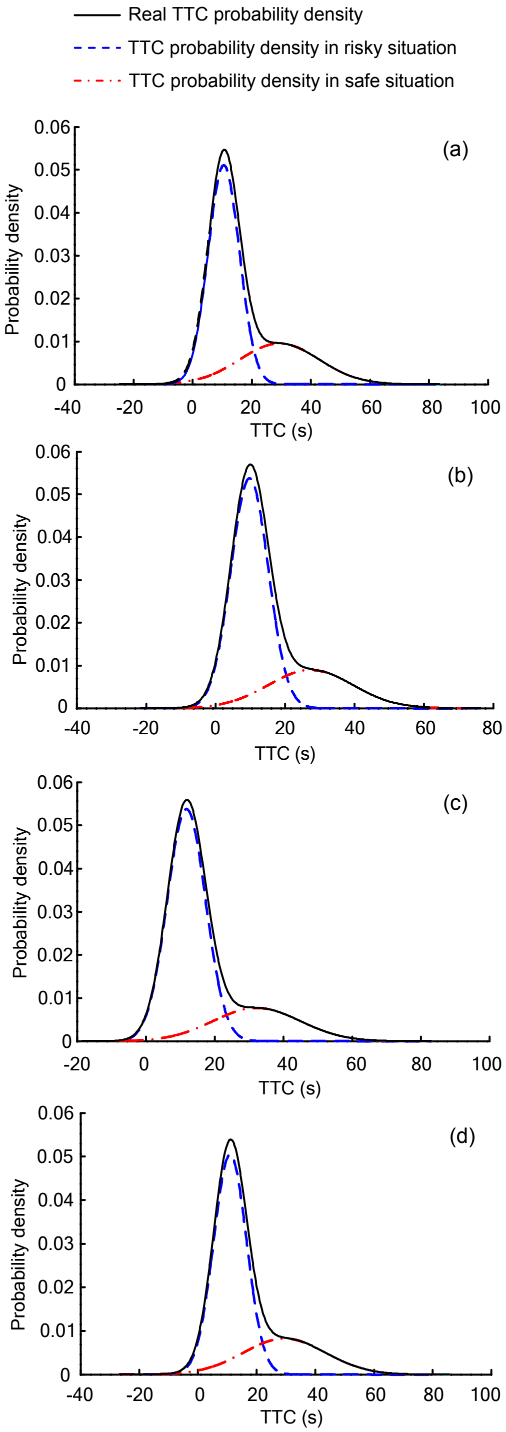

Fig. 3 Gaussian mixture distributions of TTC data from four stations, where (a)–(d) show the trends in the TTC based on Gaussian mixture distributions obtained using sample data from stations 1–4, respectively

where p_{ettc} is the PETTC, ω_{rs} is the weight, $g_{\text{rs}}()$ is the CDF of that part of the TTC sample in risky situations, whose parameters can be estimated using the GMM and EM algorithm, and τ is a predetermined TTC threshold value.

By introducing the probability concept in Eq. (7), the PETTC is easily obtainable once the TTC distributions are determined. Accordingly, PETTC (based on TTC) is a straightforward and useful indicator for forecasting traffic crashes in road sections. PETTC is thus the mixing weight for the risky situations multiplied by the probability that the TTCs in risky situations are lower than a predetermined threshold value. Therefore, the TTC threshold value is usually chosen to distinguish dangerous scenarios (or critical encounters) from relatively safe situations, and is of significance for the proposed new safety indicator. It is widely acknowledged that the TTC threshold should be between 2 and 4 s (Minderhoud and Bovy, 2001; Vogel, 2003). In this study, three conventional values (2, 3 and 4 s) are considered as possible TTC threshold values for safety forecasting.

According to Meng and Qu (2012), there is a strong linear relationship between accident frequencies and conflicts. Therefore, PETTC is a good indicator for the potential risk of accidents and can be used for safety forecasting.

4.2 Kalman filter

Many different algorithms can be found in the literature that have being applied to time series forecasting (Khashei and Bijari, 2012), such as Bayesian networks (Park and Cho, 2012) and support vector regression (Jiang and He, 2012). These algorithms are complex and have little ability to predict online time-series data quickly. Therefore, in this paper, we propose a simple Kalman filter model for safety forecasting. The Kalman filter is an algorithm that operates recursively on streams of noisy input data to produce a statistically optimal estimate of the underlying system state. They have been applied in many areas, including navigation (Yim *et al.*, 2011), water demand prediction (Nasseri *et al.*, 2011), and traffic volume forecasting (Xie *et al.*, 2007). An introduction to Kalman filter theory is given by Haykin (2001).

The Kalman filter model assumes that the true state at step k evolves from the state at $k-1$ according to

$$\boldsymbol{x}_k = \boldsymbol{F}_{k,k-1}\boldsymbol{x}_{k-1} + w_k,\ \ x_k, x_{k-1} \in \mathbb{R}^n, \tag{8}$$

where $\mathbb{R}^n$ represents the n-dimensional real variable domain, $\boldsymbol{x}_k$ and $\boldsymbol{x}_{k-1}$ are the state vectors at steps k and $k-1$, respectively, $\boldsymbol{F}_{k,\,k-1}$ is the state transition model which is applied to the previous state x_{k-1}, and w_k is the process noise, which is assumed to be drawn from a zero mean multivariate normal distribution with covariance $\boldsymbol{Q}_k$, i.e., $w_k \sim N(0, \boldsymbol{Q}_k)$.

At step k, a measurement (or observation) $\boldsymbol{z}_k$ of the true state $\boldsymbol{x}_k$ is made according to

$$\boldsymbol{z}_k = \boldsymbol{H}_k\boldsymbol{x}_k + v_k,\ \ z_k \in \mathbb{R}^m, \tag{9}$$

where $\mathbb{R}^m$ represents the m-dimensional real variable domain. $\boldsymbol{H}_k$ is the measurement model that maps the true state space into the measured space, and v_k is the measurement noise, which is assumed to be zero mean Gaussian white noise with covariance $\boldsymbol{R}_k$, $v_k \sim N(0, \boldsymbol{R}_k)$.

Let $\hat{x}_k^-$ and $\hat{x}_k$ represent the *a priori* and *a posteriori* state estimates at step k, respectively, of which the error covariance matrices are

$$\boldsymbol{P}_k^- = E\left[\left(\boldsymbol{x}_k - \hat{x}_k^-\right)\left(\boldsymbol{x}_k - \hat{x}_k^-\right)^{\mathrm{T}}\right], \tag{10}$$

$$\boldsymbol{P}_k = E\left[\left(\boldsymbol{x}_k - \hat{x}_k\right)\left(\boldsymbol{x}_k - \hat{x}_k\right)^{\mathrm{T}}\right]. \tag{11}$$

The Kalman filter process is shown in Fig. 4. The Kalman filter uses time update and measurement update algorithms to estimate $\boldsymbol{x}_k$. First, a tentative estimate $\hat{x}_k^-$ is calculated based on the value of $\hat{x}_{k-1}^-$; then, the measurement value $\boldsymbol{z}_k$ is used to refine further the value of $\hat{x}_k^-$ in order to obtain $\hat{x}_k$, which is the estimate of $\boldsymbol{x}_k$.

To forecast traffic safety using the Kalman filter, let $p_{\text{ettc}}(k)$ denote the safety indicator, PETTC, for the kth time interval, that is to be estimated. It is assumed that the indicator $p_{\text{ettc}}(k)$ at the time interval k has a linear relationship with the indicators at the last n intervals, as follows:

$$p_{\text{ettc}}(k) = \mathbf{PETTC}_k\boldsymbol{\beta}_k + \varepsilon, \tag{12}$$

where $\mathbf{PETTC}_k$=[$p_{\text{ettc}}(k-1)$, $p_{\text{ettc}}(k-2)$, …, $p_{\text{ettc}}(k-n)$]

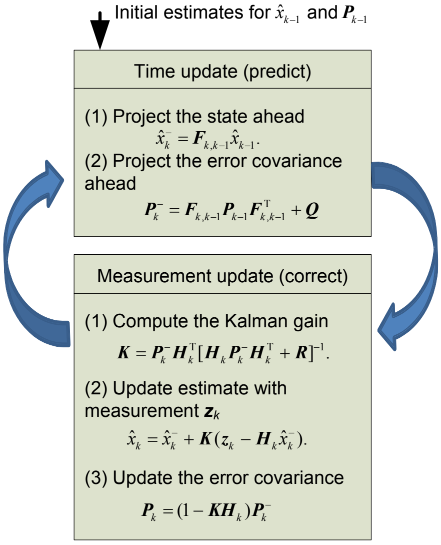
Initial estimates for $\hat{x}_{k-1}$ and $\boldsymbol{P}_{k-1}$
Time update (predict)
(1) Project the state ahead
$\hat{x}_k^- = \boldsymbol{F}_{k,k-1}\hat{x}_{k-1}$.
(2) Project the error covariance ahead
$\boldsymbol{P}_k^- = \boldsymbol{F}_{k,k-1}\boldsymbol{P}_{k-1}\boldsymbol{F}_{k,k-1}^{\mathrm{T}} + \boldsymbol{Q}$
Measurement update (correct)
(1) Compute the Kalman gain
$\boldsymbol{K} = \boldsymbol{P}_k^-\boldsymbol{H}_k^{\mathrm{T}}[\boldsymbol{H}_k\boldsymbol{P}_k^-\boldsymbol{H}_k^{\mathrm{T}} + \boldsymbol{R}]^{-1}$.
(2) Update estimate with measurement $\boldsymbol{z}_k$
$\hat{x}_k = \hat{x}_k^- + \boldsymbol{K}(z_k - \boldsymbol{H}_k\hat{x}_k^-)$.
(3) Update the error covariance
$\boldsymbol{P}_k = (1 - \boldsymbol{K}\boldsymbol{H}_k)\boldsymbol{P}_k^-$

Fig. 4 Kalman filter process

is the measured safety indicator vector, $\boldsymbol{\beta}_k = \left[\beta(k-1), \beta(k-2), \cdots, \beta(k-n)\right]^{\mathrm{T}}$ is the collection of coefficients for each corresponding measured safety indicator in the row vector $\mathbf{PETTC}_k$, and ε is the noise term.

To implement the Kalman filter model, $\boldsymbol{\beta}_k$ is used as the state vector $\boldsymbol{x}_k$ in Eq. (8); $\mathbf{PETTC}_k$ is used as the measurement matrix $\boldsymbol{H}_k$ in Eq. (9); $p_{\text{ettc}}(k)$ corresponds to z_k in Eq. (9); and Eq. (12) is equivalent to the measurement equation shown in Eq. (9). Assume now that there are n+1 observed safety indicators: $p_{\text{ettc}}(k)$, $p_{\text{ettc}}(k-1)$, …, $p_{\text{ettc}}(k-n)$. Based on the Kalman filter prediction algorithm illustrated in Fig. 3, first the *a priori* estimate of $\boldsymbol{\beta}_k$ is calculated ($\hat{\beta}_k^-$), then the measured value of $p_{\text{ettc}}(k)$ is used to update $\hat{\beta}_k^-$ and obtain an *a posteriori* estimate $\hat{\beta}_k$. Since, for short-term forecasting, the transition of the state vector can be regarded as a smooth process, the safety indicator at the next time interval can then be predicted by

$$p_{\text{ettc}}(k+1) = \mathbf{PETTC}_{k+1}\hat{\beta}_k. \tag{13}$$

Several parameters need to be determined before starting the recursive Kalman filter prediction process. The transition matrix $\boldsymbol{F}_{k,k-1}$ is set to be an n×n identity matrix since the transition is generally smooth. The calculated PETTCs can be used as the true safety indicator values so it is assumed that there is no measurement error and the variance of the measurement noise, $\boldsymbol{R}$, is zero. The variance of the process error, $\boldsymbol{Q}$, is obtained by minimizing the following negative log-likelihood function (Digalakis *et al.*, 1993).

$$-\ln\left(L(\boldsymbol{Q})\right) = \sum_{k=1}^{M}\left[\ln(\boldsymbol{X}_k) + \boldsymbol{Z}_k^{\mathrm{T}}\boldsymbol{X}_k^{-1}\boldsymbol{Z}_k\right] + C, \tag{14}$$

where $\boldsymbol{X}_k = \boldsymbol{H}_k\boldsymbol{P}_k^-\boldsymbol{H}_k^{\mathrm{T}} + \boldsymbol{R}$, $\boldsymbol{Z}_k = z_k - \boldsymbol{H}_k\hat{x}_k^-$, $\boldsymbol{P}_k^- = \boldsymbol{F}_{k,k-1}\boldsymbol{P}_{k-1}\boldsymbol{F}_{k,k-1}^{\mathrm{T}} + \boldsymbol{Q}$, C is the constant value and M is the number of sample intervals.

Based on the algorithm outlined in Fig. 4, the traffic safety prediction for the expressway using the Kalman filter is carried out as follows:

1. Let k=n+1 set initial values for $\hat{x}_{k-1}$. $\boldsymbol{P}_{k-1}$ is customarily set to be a matrix with very small values. In this study, $\hat{x}_{k-1}$ is set to be $[1/n, 1/n, \ldots, 1/n]^{\mathrm{T}}$ and $\boldsymbol{P}_{k-1}=10^{-2}\boldsymbol{I}_{n\times n}$.

2. Calculate $\hat{x}_k^-$ and $\boldsymbol{P}_k^-$ using the equations shown in the time update (predict) part of Fig. 4.

3. Let $\boldsymbol{H}_k=[p_{\text{ettc}}(k-1), p_{\text{ettc}}(k-2), \ldots, p_{\text{ettc}}(k-n)]$ and $z_k=p_{\text{ettc}}(k)$. Calculate $\boldsymbol{K}$, $\hat{x}_k$ and $\boldsymbol{P}_k$ using the equations shown in the measurement update (correct) part of Fig. 4.

4. Let $\boldsymbol{H}_{k+1}=[p_{\text{ettc}}(k), p_{\text{ettc}}(k-2), \ldots, p_{\text{ettc}}(k-(n-1))]$, then the predicted value of $\hat{p}_{\text{ettc}}(k+1) = \boldsymbol{H}_{k+1}\hat{x}_k$.

5. Let k=k+1 and go to step 2.

4.3 Forecasting process and performance evaluation

The main idea in this study is to apply a Kalman filter model to estimate the safety indicator PETTC. Fig. 5 shows a flowchart of the traffic safety forecasting procedure. The process includes five main steps: (1) raw traffic data are used to calculate TTC values in each time interval, (2) TTC distribution parameters are obtained based on the GMM, (3) a new traffic safety indicator (PETTC) is calculated using the TTC distribution parameters, (4) a Kalman filter model is applied to predict the PETTC values in the next interval, and (5) it is determined whether the PETTC is larger than a predetermined threshold; if so, a security warning is reported and management and control technologies, such as ramp controls or speed limits, are applied to lower the traffic risk.

As discussed above, this study aims to forecast a safety indicator at short intervals, and a 30-min interval was chosen. Since volumes of traffic early in the morning and late at night are very small, the number of TTC data points is also small, and this period is typically of little concern to researchers. Therefore, only data from 5:00 am to 11:00 pm (a total of 18 h and including 36 time intervals) were used for prediction and performance evaluation.

We employed two commonly used performance indices to evaluate the proposed online forecasting model. The first was the mean absolute percentage error (MAPE), and the second was the root mean square error (RMSE). These indices are given by the following equations:

$$\mathrm{MAPE}=\frac{1}{M}\sum_{k=1}^{M}\left|\frac{\hat{p}_{\mathrm{ettc}}(k)-p_{\mathrm{ettc}}(k)}{p_{\mathrm{ettc}}(k)}\right|\times 100\%, \qquad (15)$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{M}\sum_{k=1}^{M}\left[\hat{p}_{\mathrm{ettc}}(k)-p_{\mathrm{ettc}}(k)\right]^{2}}, \qquad (16)$$

where M is the number of sample intervals.

4.4 Results and analysis

Based on the above flowchart, we calculated the predicted safety indicators for the four stations. Fig. 6 shows scatter plots of the measured and predicted PETTC values, from which it can be seen that the predicted data fit the measured data well, both in low-risk and high-risk situations. The gap between the measured and predicted results throughout the

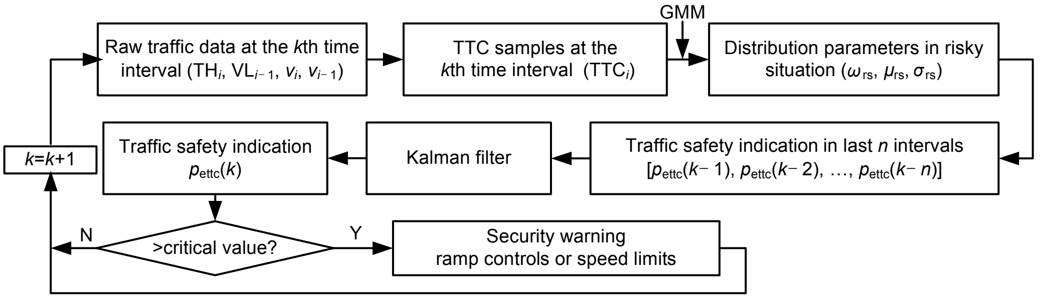
GMM
Raw traffic data at the kth time interval (TH_i, VL_{i-1}, v_i, v_{i-1})
TTC samples at the kth time interval (TTC_i)
Distribution parameters in risky situation (ω_{rs}, μ_{rs}, σ_{rs})
k=k+1
Traffic safety indication $p_{\mathrm{ettc}}(k)$
Kalman filter
Traffic safety indication in last n intervals $[p_{\mathrm{ettc}}(k-1), p_{\mathrm{ettc}}(k-2), \ldots, p_{\mathrm{ettc}}(k-n)]$
N
>critical value?
Y
Security warning
ramp controls or speed limits

Fig. 5 A flowchart of traffic safety forecasting

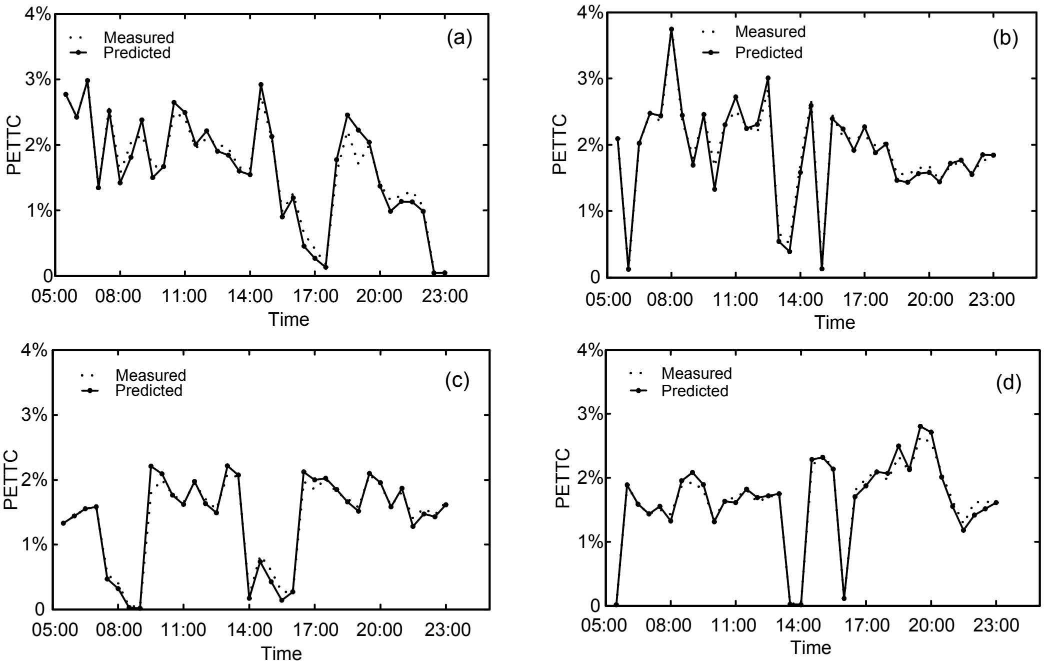

Fig. 6 Scatter plots of measured and predicted PETTCs, where (a)–(d) represent the four expressway stations 1–4, respectively

whole domain is small. Therefore, the proposed model, based on the GMM and the Kalman filter, can be seen to be a very accurate and robust online prediction method for short-term traffic safety forecasting.

Table 3 shows the performance indices of the proposed prediction model. These results indicate that the average MAPE value is about 5.0% and the average RMSE value about 0.00136, and that our prediction method can accurately forecast traffic safety indicators for all four datasets and different TTC threshold values. It is easy to see that the MAPE and RMSE values are both stable and robust, which supports the theory that it is effective to use the safety indicator, PETTC, to improve short-term traffic safety predictions. Therefore, this method can be applied for the online forecasting of relative changes in levels of safety for different road environments or traffic conditions.

5 Summary

Traffic safety prediction is of great significance for the optimal operation of expressway traffic management systems. Accurate and timely online traffic safety prediction can help managers to identify traffic risks quickly, and reduce these risks using ramp controls or speed limits. Therefore, forecasting traffic safety is one of the most important functions of expressway traffic management systems. In this paper, we propose a GMM and Kalman filter-based online traffic safety forecasting method. Firstly, the raw traffic data are used to calculate TTC values at 30-min intervals, and the statistical characteristics of the TTC sample values are presented. Secondly, the Gaussian mixture distribution is used to fit the empirical distributions of the TTC samples, which seem to be bimodal. Therefore, we can use the two-component GMM, and apply the EM algorithm to estimate the TTC distribution parameters. Then, a new traffic safety indicator, the PETTC, is proposed for risk assessment and safety predictions for expressway traffic. This indicator is demonstrated to have a strong correlation with accident frequencies, and can easily be obtained from traffic management systems. A Kalman filter-based prediction model is proposed for online short-term safety forecasting, and a detailed flowchart of the process is presented. Finally, datasets collected from four different locations on the Beijing expressway are used for performance estimation. The results show that the MAPE value is about 5.0% and the RMSE value about 0.00136. The test results demonstrate that the proposed algorithm performs well in online safety forecasting. Further studies could use a more complex prediction model, such as the extended Kalman filter, neural networks, or the discrete wavelet, to improve the performance further.

References

Allen, B.L., Shin, B.T., Cooper, D.J., 1978. Analysis of traffic conflicts and collision. *Transportation Research Record*, **667**:67-74.

Almquist, S., Hyden, C., Risser, R., 1991. Use of speed limiters in cars for increased safety and a better environment. *Transportation Research Record*, **1318**:34-39.

Archer, J., 2004. Methods for the Assessment and Prediction of Traffic Safety at Urban Intersections and Their Application in Micro-Simulation Modeling. PhD Thesis, Royal Institute of Technology, Stockholm, Sweden.

Cunto, F., Saccomanno, F.F., 2007. Micro-Level Traffic Simulation Method for Assessing Crash Potential at Intersections. Proceedings of 86th Annual Meeting, Transportation Research Board, Washington DC.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, **39**(1):1-38.

Digalakis, V., Rohlicek, J.R., Ostendorf, M., 1993. ML estimation of a stochastic linear system with the EM

Table 3 Performance of the proposed safety forecasting model under different datasets and TTC threshold values

Station No.	TTC threshold value=2 s		TTC threshold value=3 s		TTC threshold value=4 s	
	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE
1	5.36%	0.00104	5.91%	0.00137	6.42%	0.00179
2	3.45%	0.00115	3.76%	0.00132	4.07%	0.00150
3	6.22%	0.00113	4.02%	0.00140	5.78%	0.00166
4	5.17%	0.00126	5.51%	0.00139	4.33%	0.00130
Average	5.05%	0.00115	4.80%	0.00137	5.15%	0.00156

algorithm and its application to speech recognition. *IEEE Transactions on Speech and Audio Proceeding*, **1**(4): 431-442. [doi:10.1109/89.242489]

FHWA, 2003. Surrogate Safety Measures from Traffic Simulation Models. Final Report, Publication No. FHWA-RD-03-050, Federal Highway Administration, USA.

Guido, G., Saccomanno, F., Vitale, A., Astarita, V., Festa, D., 2011. Comparing safety performance measures obtained from video capture data. *Journal of Transportation Engineering*, **137**(7):481-491. [doi:10.1061/(ASCE)TE.1943-5436.0000230]

Haykin, S., 2001. Kalman Filtering and Neural Networks. John Wiley and Sons, Inc., New York.

Hayward, J., 1971. Near Misses as a Measure of Safety at Urban Intersections. PhD Thesis, Department of Civil Engineering, The Pennsylvania State University, University Park, PA.

Hsieh, C.T., Lai, E., Wang, Y.C., 2003. Robust speaker identification system based on wavelet transform and Gaussian mixture model. *Journal of Information Science and Engineering*, **19**(2):267-282.

Hydén, C., 1996. Traffic Conflicts Technique: State-of-the-art. *In*: Topp H.H. (Ed.), Traffic Safety Work with Video-Processing, University Kaiserslautern, Transportation Department, Kaiserslautern, Germany.

Jiang, H., He, W., 2012. Grey relational grade in local support vector regression for financial time series prediction. *Expert Systems with Applications*, **39**(3):2256-2262. [doi:10.1016/j.eswa.2011.07.100]

Jin, S., Wang, D.H., Yang, X.R., 2011a. Non-lane-based car following model using visual angle information. *Transportation Research Record Journal of the Transportation Research Board*, **2249**:7-14. [doi:10.3141/2249-02]

Jin, S., Qu, X., Wang, D.H., 2011b. Assessment of expressway traffic safety using Gaussian mixture model based on time to collision. *International Journal of Computational Intelligence Systems*, **4**(6):1122-1130. [doi:10.2991/ijcis.2011.4.6.4]

Jin, S., Wang, D.H., Xu, C., Huang, Z.Y., 2012. Staggered car-following induced by lateral separation effects in traffic flow. *Physics Letters A*, **376**(3):153-157. [doi:10.1016/j.physleta.2011.11.005]

Jordan, M., Jacobs, R., 1994. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, **6**(2): 181-214. [doi:10.1162/neco.1994.6.2.181]

Khashei, M., Bijari, M., 2012. A new class of hybrid models for time series forecasting. *Expert Systems with Applications*, **39**(4):4344-4357. [doi:10.1016/j.eswa.2011.09.157]

Kim, S.C., Kang, T.J., 2007. Texture classification and segmentation using wavelet packet frame and Gaussian mixture model. *Pattern Recognition*, **40**(4):1207-1221. [doi:10.1016/j.patcog.2006.09.012]

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A*, **44**:291-305.

McAndrews, C., 2011. Traffic risks by travel mode in the metropolitan regions of Stockholm and San Francisco: a comparison of safety indicators. *Injury Prevention*, **17**(3): 204-207. [doi:10.1136/ip.2010.029306]

McLachlan, G., 1988. Mixture Models. Marcel Dekker, New York, NY.

Meng, Q., Qu, X., 2012. Estimation of vehicle crash frequencies in road tunnels. *Accident Analysis and Prevention*, **48**(1):254-263. [doi:10.1016/j.aap.2012.01.025]

Meng, Q., Qu, X., Wang, X., Yuanita, V., Wong, S.C., 2011a. Quantitative risk assessment modeling for nonhomogeneous urban road tunnels. *Risk Analysis*, **31**(3):382-403. [doi:10.1111/j.1539-6924.2010.01503.x]

Meng, Q., Qu, X., Yong, K.T., Wong, Y.H., 2011b. QRA model-based risk impact analysis of traffic flow in urban road tunnels. *Risk Analysis*, **31**(12):1872-1882. [doi:10.1111/j.1539-6924.2011.01624.x]

Minderhoud, M.M., Bovy, P.H.L., 2001. Extended time-to-collision measures for road traffic safety assessment. *Accident Analysis and Prevention*, **33**(1):89-97. [doi:10.1016/S0001-4575(00)00019-1]

Nasseri, M., Moeini, A., Tabesh, M., 2011. Forecasting monthly urban water demand using Extended Kalman Filter and Genetic Programming. *Expert Systems with Applications*, **38**(6):7387-7395. [doi:10.1016/j.eswa.2010.12.087]

Park, H.S., Cho, S.B., 2012. Evolutionary attribute ordering in Bayesian networks for predicting the metabolic syndrome. *Expert Systems with Applications*, **39**(4):4240-4249. [doi:10.1016/j.eswa.2011.09.110]

Qu, X., Meng, Q., Yuanita, V., Wong, Y.H., 2011. Design and implementation of a quantitative risk assessment software tool for Singapore's road tunnels. *Expert Systems with Applications*, **38**(11):13827-13834.

Redner, R., Walker, H., 1984. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, **26**(2): 195-237. [doi:10.1137/1026034]

Stauffer, C., Grimson, W.E.L., 1999. Adaptive background mixture models for real-time tracking. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, **2**:246-252.

Stephens, M.A., 1974. EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, **69**(347):730-737. [doi:10.1080/01621459.1974.10480196]

Svensson, A., 1998. A Method for Analyzing the Traffic Process in a Safety Perspective. PhD Thesis, University of Lund, Lund, Sweden.

Titterington, D., Smith, A., Makov, U., 1985. Statistical Analysis of Finite Mixture Distributions. John Wiley & Sons.

Topp, H.H., 1998. Traffic Safety Work with Video-Processing. University Kaiserslautern, Transportation Department, Green Series No. 43, Kaiserslautern, Germany.

van der Horst, R., Kraay, J., 1986. The Dutch Conflict Observation Technique-DOCTOR. Proceedings of

Workshop-Traffic Conflicts and Other Intermediate Measures in Safety Evaluation, Budapest, Hungary.

Várhelyi, A., 1996. Dynamic Speed Adaptation based on Information Technology—a Theoretical Background. PhD Thesis, Bulletin 142, Lund University.

Vogel, K., 2003. A comparison of headway and time to collision as safety indicators. *Accident Analysis and Prevention*, **35**(3):427-433. [doi:10.1016/S0001-4575(02)00022-2]

Xie, Y.C., Zhang, Y.L., Ye, Z.R., 2007. Short-Term Traffic Volume Forecasting Using Kalman Filter with Discrete Wavelet Decomposition. *Computer-Aided Civil and Infrastructure Engineering*, **22**(5):326-334. [doi:10.1111/j.1467-8667.2007.00489.x]

Yim, J., Joo, J., Park, C., 2011. A Kalman filter updating method for the indoor moving object database. *Expert Systems with Applications*, **38**(12):15075-15083. [doi:10.1016/j.eswa.2011.05.037]