Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 852495, 7 pages
http://dx.doi.org/10.1155/2014/852495

*Research Article*

# Unrecorded Accidents Detection on Highways Based on Temporal Data Mining

**Shi An, Tao Zhang, Xinming Zhang, and Jian Wang**

*School of Transportation Science and Engineering, Harbin Institute of Technology, Harbin 150090, China*

Correspondence should be addressed to Xinming Zhang; 12b332001@hit.edu.cn

Received 3 April 2014; Revised 26 May 2014; Accepted 26 May 2014; Published 15 June 2014

Academic Editor: Hamid Reza Karimi



Automatic traffic accident detection, especially not recorded by traffic police, is crucial to accident black spots identification and traffic safety. A new method of detecting traffic accidents is proposed based on temporal data mining, which can identify the unknown and unrecorded accidents by traffic police. Time series model was constructed using ternary numbers to reflect the state of traffic flow based on cell transmission model. In order to deal with the aftereffects of linear drift between time series and to reduce the computational cost, discrete Fourier transform was implemented to turn time series from time domain to frequency domain. The pattern of the time series when an accident happened could be recognized using the historical crash data. Then taking Euclidean distance as the similarity evaluation function, similarity data mining of the transformed time series was carried out. If the result was less than the given threshold, the two time series were similar and an accident happened probably. A numerical example was carried out and the results verified the effectiveness of the proposed method.

## 1. Introduction

Road accidents are regarded as one of the leading causes of death for people between the ages of 5 and 44 according to the World Health Organization [1]. More than that, traffic crashes result in serious economic losses on account of traffic congestion which in turn leads to a wide variety of adverse consequences such as traffic delays, supply chain interruptions, travel time unreliability, and increased noise pollution, as well as deterioration of air quality [2]. Thus, reducing or avoiding traffic collisions is of great significance to traffic safety. High collision concentration location (HCCL) [3] detection is an effective means to find out the accident black spots and take some necessary continuous improvement measures. Historical accident data is the necessary foundation of any research on this subject. One of the main problems of the accident data was considered as the heterogeneity [4] in the previous studies. A great many methods had been proposed to solve this problem, such as latent class clustering [5–8], Bayesian networks (BNs) [9–11], and continuous risk profile (CRP) [12, 13]. One of the commonalities between these methods is that historical accident data, more specifically, recorded historical accident data, were used as the basic data. However, not all traffic crashes are known and recorded by traffic police. It is undeniable that some minor accidents often happened on highways and were settled privately for trivial losses. Usually, traffic accident black spots are identified mainly based on the traffic crash data recorded by traffic police departments [14]. This just helps to find out the collisions we know, while ones we do not know, which did happen, keep unconsidered yet. In part, these observations motivated our study.

Traffic accidents are contingent events and are difficult to detect if there is no alarm. Nonetheless, a traffic accident was bound to create an impact on traffic flow pattern and cause different levels of congestion [15, 16]. The traffic volume, traffic speed, and traffic density were changed by crashes, even minor accidents. Thus, capturing the change of these traffic flow parameters is very helpful for traffic accidents, especially unrecorded traffic accidents detection.

Given this, an automatic traffic accident detection method was proposed in this paper. Traffic flow data was used and simulated by cell transmission model (CTM). According to the different inflow between two cells, time series model was constructed to reflect traffic flow state. Time series pattern, when accidents happened, was established based on historical accident data. To overcome the defect

of Euclidean distance, which does not consider the linear drift in the time domain, and to reduce the computational cost, discrete Fourier transform was implemented to turn the time series from time domain to frequency domain. Leveraging the strengths of temporal data mining at finding time-varying patterns, any time series that were similar to the given pattern could be figured out, namely, the unknown and unrecorded accidents, by similarity search. The premier aim and contribution of this paper are to find out "the hidden accidents," such as compounding in private, using temporal data mining method. A case study using the real highway traffic data in Harbin, China, was conducted for verification.

## 2. Material and Methods

*2.1. Construction of Time Series Reflecting Traffic Flow State.* For highway traffic flow, there was a unique state in each period. One of the ways to describe this was the metaphor of a screen capture for the traffic flow over consecutive periods of time. Every picture reflected the state of the traffic flow at a certain time. These pictures constituted a sequence over time. This was the consideration of building the time series model to describe the evolution of traffic flow in this paper. Traffic conditions estimation was achieved through dynamic traffic assignment (DTA) simulation that utilized temporal aspects of a transportation system. Different values of inflow between cells in CTM were typically expressed as ternary numbers (0, 1, and 2). A series of ternary numbers, generated by CTM, were introduced to illustrate the traffic flow state. Then, time series data were created by converting ternary numbers to decimal numbers. Thus, the traffic flow state could be reflected by time series data, and it was the basic work for unrecorded accidents mining.

*2.1.1. Cell Transmission Model.* To model the propagation of traffic flow and construct time series data in the section below, the spread of highway traffic flow was simulated by CTM in this paper. CTM was proposed by Daganzo [17, 18] and was considered as a proper method. It was believed that the relationship between traffic flow ($q$) and density ($k$) was of the form depicted figurally as follows:

$$q = \min\left\{vk, q_{\max}, \omega\left(k_j - k\right)\right\}, \tag{1}$$

where $v$, $q_{\max}$, $\omega$, and $k_j$ denoted the free-flow speed, the maximum flow (or capacity), the backward wave speed, and the maximum (or jam) density, respectively, as shown in Figure 1.

In Figure 1, if the density is less than $k_1$, the traffic flow $q$ is equal to $vk$; if it is between $k_1$ and $k_2$, then $q$ reaches its maximum, $q_{\max}$; if it is between $k_2$ and $k_j$, $q$ is equal to $\omega(k_j - k)$; and $q$ is 0 when the density reaches $k_j$.

Then, the continuous Lighthill-Whitham-Richards (LWR) equations [19, 20] for a single highway link were discretized through this method and could be approximated by a set of difference equations. The state of the system was updated over time. Thus, the discontinuous changes of traffic flow could be captured.

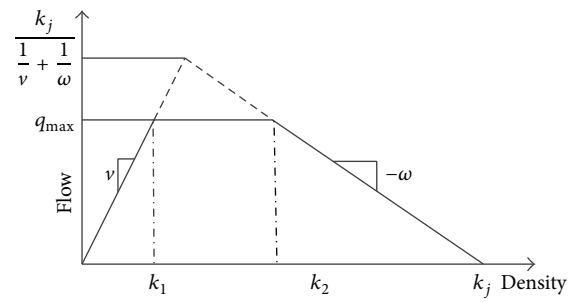


Figure 1: The relationship of traffic flow and density in CTM.

In CTM, a single road was divided into homogeneous sections (cells), $i$, whose lengths equaled the distance traveled by free-flowing traffic speed in one clock interval. The state of the system at instant $t$ was then given by the number of vehicles contained in each cell, $n_i(t)$. The following parameters were defined for each cell.

$N_i(t)$ is the maximum number of vehicles that can be present in cell $i$ at time $t$, and $Q_i(t)$ is the maximum number of vehicles that can flow into cell $i$ when the clock advanced from $t$ to $t + 1$.

These constants could vary with time (e.g., contingent traffic incidents or conscious traffic control measures), but this dependence was able to be ignored for simplicity of notation. The first constant $N_i(t)$ was defined to be the product of the cell's length and its jam density, and the second one was the product of the time interval and the cell's capacity.

If cells were numbered consecutively starting with the upstream end of the road from $i = 1$ to $I$, the recursive relationship of the CTM, as discussed by Daganzo [17, 18], could be expressed as

$$n_i(t+1) = n_i(t) + y_i(t) - y_{i+1}(t), \tag{2a}$$

where $y_i(t)$ was the inflow to cell $i$ in the time interval $(t, t+1)$, given by

$$y_i(t) = \min\left\{n_{i-1}(t), Q_i(t), \delta\left[N_i(t) - n_i(t)\right]\right\}, \tag{2b}$$

where $\delta = \omega/v$.

The formulas (2a) and (2b) constituted the fundamental equations of CTM. Equation (2a) expressed the status updates of cells over time, while the latter gave the variations for updating.

*2.1.2. Constructing Time Series of Traffic Flow.* Time series data was a sequence of data evolving through time. There were two strengths of temporal data mining: data-based and pattern-based. The former was more likely to approach the truth; the latter was more likely to extract the features. Every traffic accident was considered to change the traffic flow state more or less. Features of this change were supposed to be extracted by temporal data mining and these features were able to be used to find out "the hidden accidents."

Gao et al. used the NaSch traffic model to simulate the evolution of traffic flow [21, 22]. The state of traffic flow at

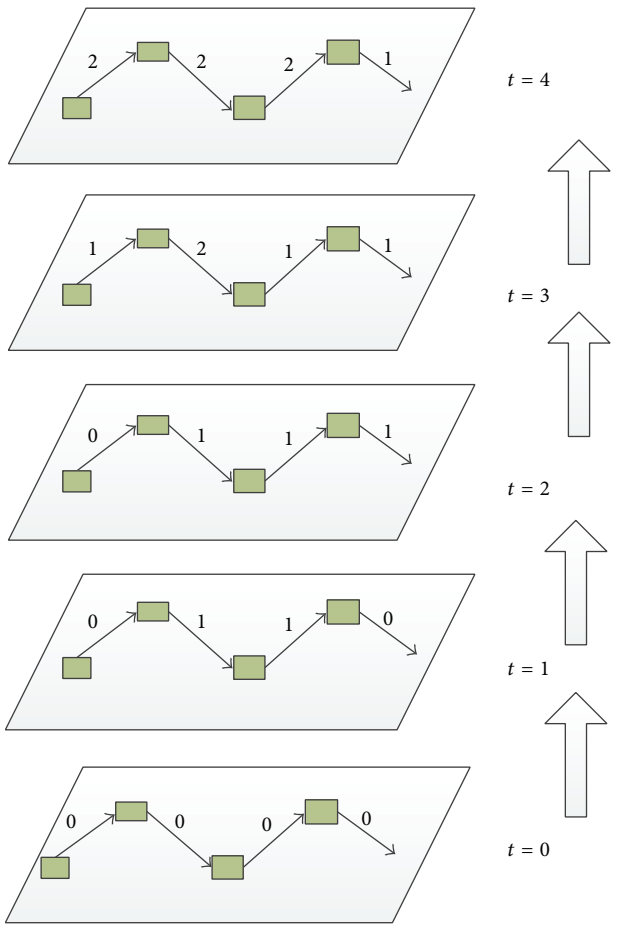


FIGURE 2: The traffic flow trends with the tick of a clock.

each time period was regarded as a node of a network. Then, a complex network was constructed, also known as a multiple-mode system model, which could describe the evolution of traffic flow. Zhao et al. studied state estimation of this kind of systems [23]. However, the temporal relations among these nodes were ignored in their research.

According to CTM, at each time step, the inflow $y_i(t)$ to cell $i$ could be $n_{i-1}(t)$, $Q_i(t)$ or $\delta[N_i(t) - n_i(t)]$. When $y_i(t) = n_{i-1}(t)$, the state of the cell was represented by number 0; when $y_i(t) = Q_i(t)$, it was represented by number 1, otherwise represented by number 2. Then, the state of the system was expressed by a sequence of ternary numbers, for example, $\{0, 0, 1, 2, 1\}$.

The model employed was described as follows. For $N$ cells, we assumed that, at time period $(t, t + 1)$, the state was represented by a set of ternary numbers, namely, $S_t = \{s_1, s_2, \ldots, s_N\}$. Then, $S_t$ is the precursor of $S_{t+1}$. Here, each ternary number $s_i$ can be considered as an element, which can take three different states; that is, $s_i = 0, 1, 2$. Figure 2 depicted the evolution of traffic flow with the tick of a clock.

As shown in Figure 2, when the clock advanced from 0 to 4, the system (a single segment) state change could be discovered clearly using the "screen capturing" method. The state $S_t$ was time-varying and was represented by a sequence of numbers. For convenience, a parameter $M$ was introduced

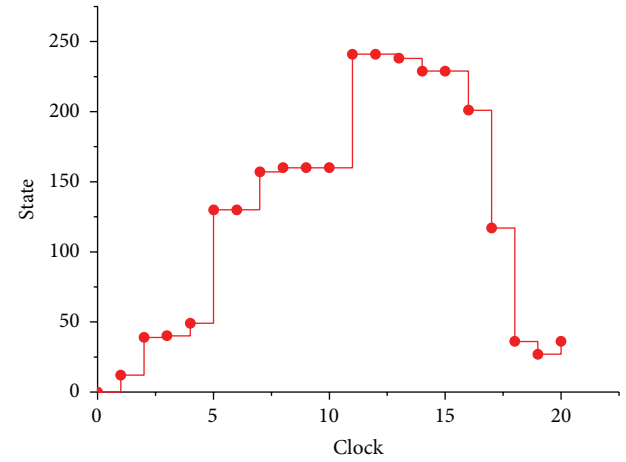


FIGURE 3: Time series data reflecting traffic flow trends of a five-cell road.

in this paper to represent the value of $S_t$ and $M = \sum_{i=1}^{N} 3^i s_i$; thus, ternary numbers were converted to decimal numbers; for example, $\{0, 0, 1, 2, 1\}$ was converted to 16. Thus, time series data was created (see Figure 3).

Each single decimal number represented a system's state. As shown in Figure 3, at time step 9, for instance, the value of the system state was 78; thus, the ternary numbers were $\{0, 2, 2, 2, 0\}$, which was the state of traffic flow.

*2.2. Feature Extraction.* Noise in the raw time series data could reduce accuracy and creditability of data mining. Linear drift was certainly an example. In many clustering analysis methods, $K$-means, for example, Euclidean distance, was frequently used as a similarity measure function. The unrecorded accidents detection method proposed in this paper was mostly based on similarity mining, which would be discussed later. Linear drift was the most important factor to influence the accuracy of the results.

For time series, $X = \{x_1, x_2, \ldots, x_n\}$ and $Y = \{y_1, y_2, \ldots, y_n\}$, if, at time $t$, $x_i - y_i = \varepsilon$, $i = 1, \ldots, n$, where $\varepsilon$ was a constant. That was to say that the relationship between $X$ and $Y$ was linear drift in the time domain, as shown in Figure 4.

In this case, if Euclidean distance was used and $\varepsilon$ was beyond the threshold, these two time series $X$ and $Y$ were not considered to be similar by mining algorithms. While they had similar shape and trend apparently, the judging result was inaccurate obviously. Aiming to prevent such errors and to realize data compression and reduce the computational cost, it was necessary to extract feature from the original time series data, using the image in feature space to replace the original one.

Discrete Fourier transform (DFT), which had unique merits in time series analysis, was an alternative way. For a given time series object, DFT could be used to turn it to frequency domain from time domain. According to Parseval theory, the time-domain energy function was equal to the frequency-domain energy function. And most of the energy in frequency domain concentrated on the first few

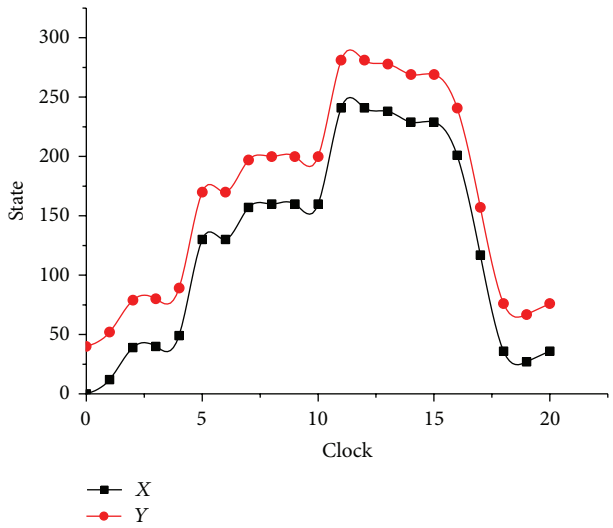


Figure 4: Linear drift between time series $X$ and $Y$.

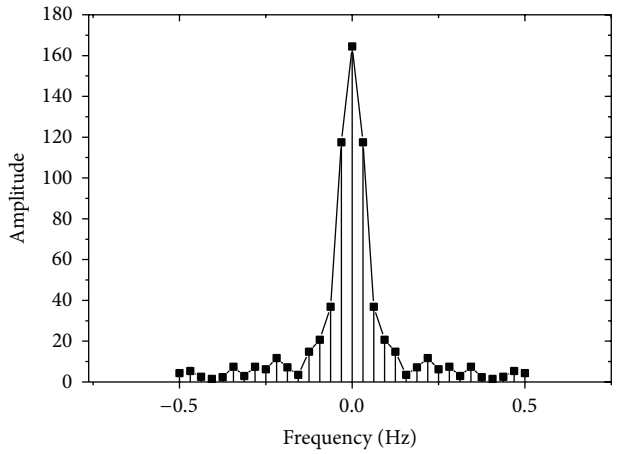


Figure 5: Discrete Fourier transform.

coefficients; hence, other coefficients could be omitted. Thus, the remaining coefficients were able to be seen as the features of the original time series.

For a given time series $X = \{x_t\}$, $t = 0, 1, \ldots, n-1$, translating the series from time domain to frequency domain, the new sequence obtained by DFT was denoted by $\overline{X} = \{x'_f\}$, $f = 0, 1, \ldots, n-1$, where

$$x'_f = \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} x_t \exp\left(-\frac{j2\pi ft}{n}\right), \quad f = 0, 1, \ldots, n-1. \quad (3)$$

Taking the data in Figure 3, for example, the result of DFT was shown in Figure 5.

The area, in which the frequency was greater than 0.0, was the real transformed data of the original finite time series. 17 elements were left out of the initial total of 21 ones. Data compression had been realized and as a result of the transformation from time domain to frequency domain, linear drift no longer existed as well as other noises in the time domain.

*2.3. Unrecorded Accidents Detection Using Similarity Mining.* Similarity search is an important research field in temporal data mining. As mentioned above, each recorded accident could bring a piece of time series data, and all recorded historical accidents data over a period of time could explain the traffic flow trends when accidents happened. After data processing, using the above methods, which could be called data preprocessing, clustering analysis was supposed to be implemented in this paper. The classical $K$-means method would be able to meet the accuracy requirements due to the appropriate data preprocessing. Results of cluster were considered as "normal traffic flow trends" under accidents and the "hidden accidents" could be found out by similarity search.

The method of similarity measurement used in this paper was Euclidean distance. Set the time sequence of "normal traffic flow trends" of a single road segment which was $\{x_i\}$, whose length was $n$. The time series to be measured was denoted by $\{y_i\}$ with its length $N$, $N \geq n$ generally. In similarity search, subsequences of $\{y_i\}$, whose lengths were $n$, were the measuring object. These subsequences were denoted by $\{z_i\}$. It was known that the number of $\{z_i\}$ was $J$, $J = N-n+1$. Then, the similarity metric function could be defined as follows:

$$\min_J \sum_{i=1}^{n} \left(x_i - K_J z_i^J\right)^2, \quad (4)$$

where $K_J$ was the scaling factor. The calculation times were $N - n + 1$ obviously.

So far, for a single road segment, the procedure to detect unrecorded accidents was described in Figure 6.

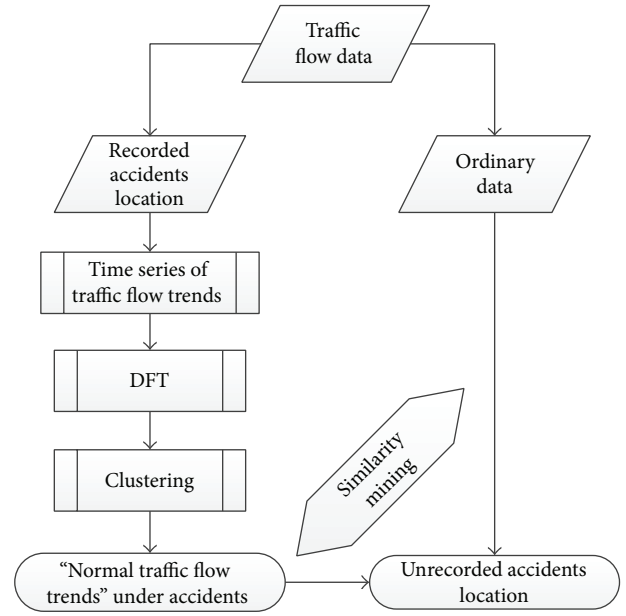


Figure 6: The procedure of unrecorded accidents detection.

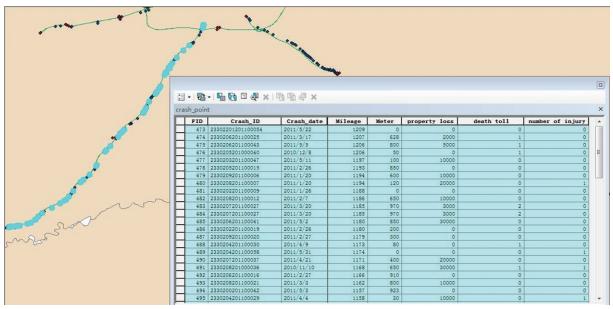

Figure 7: The study site and data.

## 3. Results and Discussions

*3.1. The Study Site and Data Preparation.* A case study was conducted based on the data extracted from records collected on Beijing-Harbin Expressway (G1) between Harbin and Lalinhe from January 2010 to July 2011. The traffic accidents dataset included a total of 73 crashes recorded on the 72-kilometer length road in total. The study site and the traffic accidents data were shown in Figure 7.

In one and a half years, for such a highway with the annual traffic reaching 8–10 million, "only" 73 crashes happened. Experience told us that it did not indicate how safe the highway was, but some minor accidents were not known or recorded. It was meaningful to detect the unrecorded collisions for traffic safety research.

As the limit of detection device, traffic flow data worked out using data collected by highway toll collection system. For detailed calculation process, readers could refer to Weng et al. [24]. The estimation accuracy could meet the requirements technically.

*3.2. Numerical Experiment.* In the numerical experiment, the time horizon was set within half an hour and the time interval was 30 seconds. The accidents happened at the eleventh clock interval, namely, 5 minutes after the start time of simulation. Thus, there were 60 points in one piece of time series data. The free-flow speed was 120 km/h; thus, the length of each cell was the product of the free-flow speed and the unit clock interval that was 120 km/h $*$ 30 s = 1 km. Based on the historical accident data, the accident location was set in the middle cell in CTM and five cells were brought in to represent a single highway segment where a crash happened, as shown in Figure 8.

The virtual cell in Figure 8 was on behalf of the demand generation. There were five inflow states, as shown by the five arrows in the figure above. Using the method shown in Figure 6, first of all, it was significant to identify the "normal traffic flow trends" under accidents. Time series data was conducted using the method mentioned above and DFT was implemented then for each crash record. $K$-means cluster analysis was adopted and the results were indicated in Figure 9.

As is shown in Figure 9, two clusters were generated by this method. In Figure 9(a), there were 34 accidents gathered together. The common point of them was that congestion formed and dissipated gradually in the simulation period. The

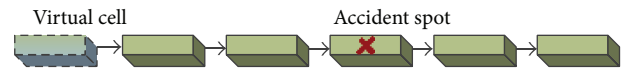


Figure 8: CTM for a single highway segment.

shapes of these time series data were similar to the normal distribution curve. According to the accident records of traffic police, most of these crashes were single-vehicle crashes and both-car accidents. In Figure 9(b), 33 crashes were involved and the congestion was not eliminated in study period. Two-thirds of them were crashes between two heavy trucks and one-third were multivehicle accidents. This kind of crashes could cause severe congestion and influence traffic seriously. Actually, another cluster existed in this case and only five accidents were involved. This cluster had less interference to traffic, but, for the records, they were collisions between vehicles and pedestrians. It was a serious threat to traffic safety but beyond the scope of this paper, so the result was not shown here.

For the unrecorded accidents, usually there were no casualties and little losses. The reason why they were not recorded was that these minor accidents were settled privately and even kept unknown by the traffic police. So, there was every reason to believe that these unrecorded accidents were single-vehicle crashes or both-car accidents. Thus, the cluster shown in Figure 9(a) was our target in the similarity search. For verification, one accident out of the 73 crashes was separated out, pretending to be unknown.

By similarity search, the "prepared" accident was found out and the time series data was shown in Figure 10. The traffic flow trend in Figure 10 was indeed similar to the cluster in Figure 9(a). At the first 10 clock intervals in both figures, the traffic was smooth because of the unsaturated traffic flow and no accidents. Then, congestion formed and dissipated gradually in a certain period. The result of similarity search proved the reliability of the proposed method.

## 4. Conclusions

Unrecorded accidents were significant to identify traffic accident-prone location. Based on the observation that the traffic volume, traffic speed, and traffic density were changed by crashes, even minor accidents, an automatic traffic accident identification method was proposed. As most of the current studies did not pay enough attention to the time factor when studying the relationship between traffic state and crashes on highways, this paper proposed a method to construct time series data using traffic flow data when accidents happened. To avoid the defect of not considering the linear drift in the time domain between two sequences, DFT was carried out to extract features from original time series data. Traffic flow trend could be well understood by clustering analysis. Then, through the method of similarity search, unrecorded accidents, which were believed to be single-vehicle crashes or both-car accidents, were found out. The case study using real data in Harbin showed the feasibleness of the proposed method.

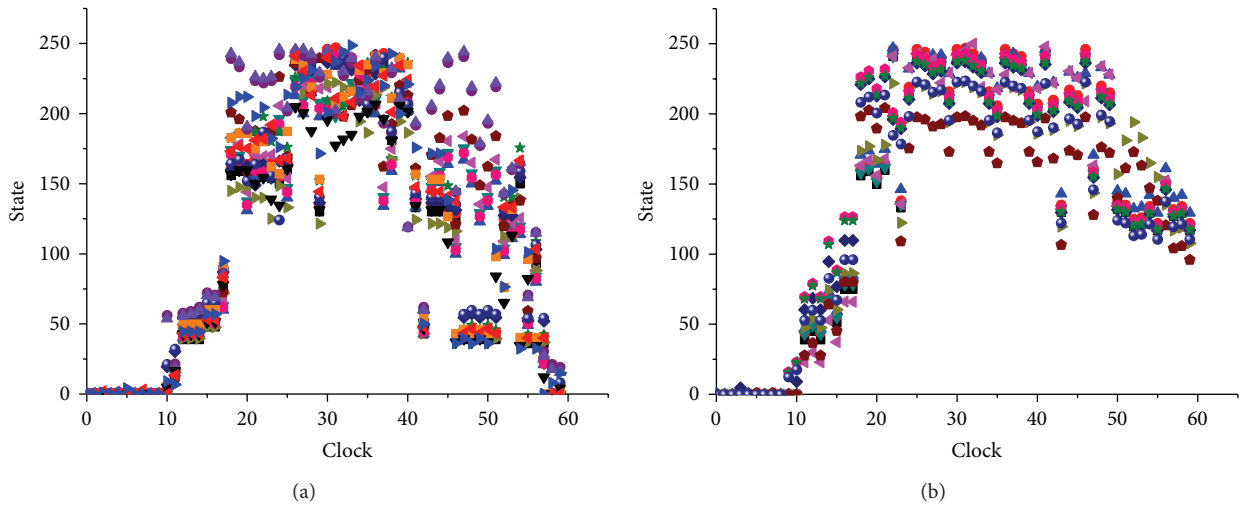


Figure 9: Results of cluster.

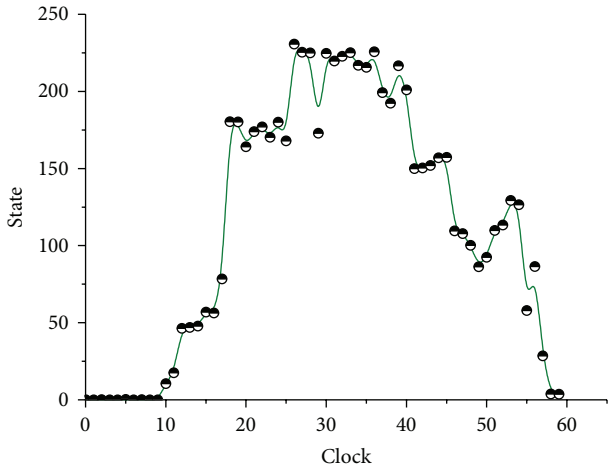


Figure 10: Result of similarity search.

For further research, data of car insurance would be valuable for data mining in this area or for verification of the automatic traffic accidents identification. And further study could focus on the traffic flow trends under accidents, such as the influence diffusion and the elimination of the crash influence.

## Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

## Acknowledgments

This work was supported by the HTRDP (863) under Grant no. 2012AA112310 and Research Fund for the Doctoral Program of Higher Education of Ministry of Education of China (20112302110054).

## References

[1] E. Krug, "Decade of action for road safety 2011–2020," *Injury*, vol. 43, no. 1, pp. 6–7, 2012.

[2] A. Gregoriades and K. C. Mouskos, "Black spots identification through a Bayesian Networks quantification of accident risk index," *Transportation Research C: Emerging Technologies*, vol. 28, pp. 28–43, 2013.

[3] O. H. Kwon, M. J. Park, H. Yeo, and K. Chung, "Evaluating the performance of network screening methods for detecting high collision concentration locations on highways," *Accident Analysis and Prevention*, vol. 51, pp. 141–149, 2013.

[4] P. T. Savolainen, F. L. Mannering, D. Lord, and M. A. Quddus, "The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives," *Accident Analysis and Prevention*, vol. 43, no. 5, pp. 1666–1676, 2011.

[5] B. Depaire, G. Wets, and K. Vanhoof, "Traffic accident segmentation by means of latent class clustering," *Accident Analysis and Prevention*, vol. 40, no. 4, pp. 1257–1266, 2008.

[6] J. M. Pardillo-Mayora, C. A. Domínguez-Lira, and R. Jurado-Piña, "Empirical calibration of a roadside hazardousness index for Spanish two-lane rural roads," *Accident Analysis and Prevention*, vol. 42, no. 6, pp. 2018–2023, 2010.

[7] B.-J. Park and D. Lord, "Application of finite mixture models for vehicle crash data analysis," *Accident Analysis and Prevention*, vol. 41, no. 4, pp. 683–691, 2009.

[8] B.-J. Park, D. Lord, and J. D. Hart, "Bias properties of Bayesian statistics in finite mixture of negative binomial regression models in crash data analysis," *Accident Analysis and Prevention*, vol. 42, no. 2, pp. 741–749, 2010.

[9] M. Simoncic, "A Bayesian network model of two-car accidents," *Journal of Transportation and Statistics*, vol. 7, no. 2-3, pp. 13–25, 2005.

[10] J. De Oña, R. O. Mujalli, and F. J. Calvo, "Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks," *Accident Analysis and Prevention*, vol. 43, no. 1, pp. 402–411, 2011.

[11] R. O. Mujalli and J. De Oña, "A method for simplifying the analysis of traffic accidents injury severity on two-lane highways using Bayesian networks," *Journal of Safety Research*, vol. 42, no. 5, pp. 317–326, 2011.

[12] K. Chung, K. Jang, S. Madanat, and S. Washington, "Proactive detection of high collision concentration locations on highways," *Transportation Research A: Policy and Practice*, vol. 45, no. 9, pp. 927–934, 2011.

[13] K. Chung, D. R. Ragland, S. Madanat, and S. M. Oh, "The continuous risk profile approach for the identification of high collision concentration locations on congested highways," in *Proceedings of the 19th International Symposium on Transportation & Traffic Theory (ISTTT '09)*, pp. 463–480, 2009.

[14] G. Vandenbulcke, I. Thomas, and L. Int Panis, "Predicting cycling accident risk in Brussels: a spatial case-control approach," *Accident Analysis and Prevention*, vol. 62, pp. 341–357, 2014.

[15] Z. Zheng, "Empirical analysis on relationship between traffic conditions and crash occurrences," *Procedia-Social and Behavioral Sciences*, vol. 43, pp. 302–312, 2012.

[16] A. Hamzehei, E. Chung, and M. Miska, "Pre-crash and non-crash traffic flow trends analysis on motorways," in *Proceedings of the Australasian Transport Research Forum*, 2013.

[17] C. F. Daganzo, "The cell transmission model: a dynamic representation of highway traffic consistent with the hydrodynamic theory," *Transportation Research B: Methodological*, vol. 28, no. 4, pp. 269–287, 1994.

[18] C. F. Daganzo, "The cell transmission model, part II: network traffic," *Transportation Research B: Methodological*, vol. 29, no. 2, pp. 79–93, 1995.

[19] M. J. Lighthill and G. B. Whitham, "On kinematic waves I: flow movement in long rivers. II: a theory of traffic flow on long crowded roads," *Proceedings of the Royal Society of London A*, vol. 229, no. 1178, pp. 281–316, 1955.

[20] P. I. Richards, "Shock waves on the highway," *Operations Research*, vol. 4, pp. 42–51, 1956.

[21] G. Zi-You and L. Ke-Ping, "Evolution of traffic flow with scale-free topology," *Chinese Physics Letters*, vol. 22, no. 10, pp. 2711–2714, 2005.

[22] X.-G. Li, Z.-Y. Gao, K.-P. Li, and X.-M. Zhao, "Relationship between microscopic dynamics in traffic flow and complexity in networks," *Physical Review E*, vol. 76, no. 1, Article ID 016110, 2007.

[23] X. Zhao, H. Liu, J. Zhang, and H. Li, "Multiple-mode observer design for a class of switched linear systems," *IEEE Transactions on Automation Science and Engineering*, 2013.

[24] J.-C. Weng, L.-L. Liu, and B. Du, "ETC data based traffic information mining techniques," *Journal of Transportation Systems Engineering and Information Technology*, vol. 10, no. 2, pp. 57–63, 2010.