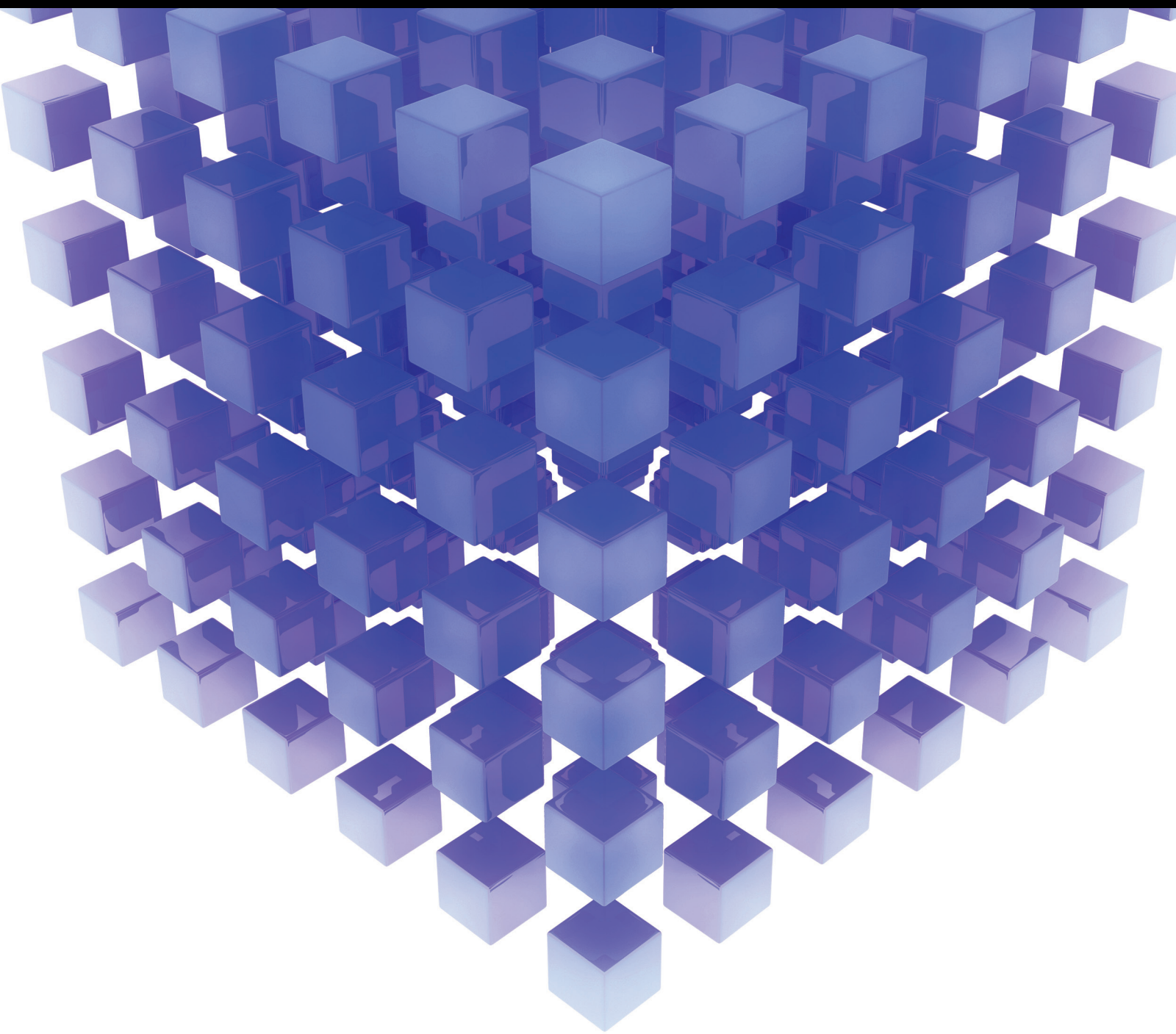
Mathematical Problems in Engineering

Safe, Resilient, and Sustainable Transportation Systems

Guest Editors: Yuanchang Xie, Pan Liu, and Chronis Stamatiadis



Safe, Resilient, and Sustainable Transportation Systems

Mathematical Problems in Engineering

Safe, Resilient, and Sustainable Transportation Systems

Guest Editors: Yuanchang Xie, Pan Liu, and Chronis Stamatiadis

Editorial Board

Contents

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2016, Article ID 5696419, 2 pages
http://dx.doi.org/10.1155/2016/5696419

Editorial

Safe, Resilient, and Sustainable Transportation Systems

Yuanchang Xie,[1] Pan Liu,[2] and Chronis Stamatiadis[1]

[1]*Civil and Environmental Engineering, University of Massachusetts Lowell, 1 University Avenue, Lowell, MA 01854, USA*
[2]*School of Transportation, Southeast University, 2 Sipailou, Nanjing, Jiangsu 210096, China*

Correspondence should be addressed to Yuanchang Xie; yuanchang_xie@uml.edu

Received 2 December 2015; Accepted 2 December 2015

Traffic accidents are among the top 10 causes of death worldwide. They resulted in over 1.3 million fatalities in 2012 alone and cause billions of dollars of loss each year in productivity, property damage, and time and fuel due to congestion. Traffic safety is becoming increasingly important, particularly in developing countries that are facing rapid urbanization and motorization. In addition to traffic safety, there are many recent examples of transportation systems being disrupted by extreme weather events. Transportation activities, on the other hand, also contributed substantially to these extreme weather events either directly or indirectly. Clearly, it is important to build a resilient and sustainable transportation system that is less vulnerable to disasters and has the minimum possible environmental footprint.

In this special issue, several studies investigated highway transportation from the operations, resiliency, and logistics perspectives. X. Zhang et al. developed a cooperative Q-learning algorithm for finding near-optimal path sets for multiple Origin-Destination (OD) pairs simultaneously. The new algorithm was compared with the k-shortest path algorithm and the Dijkstra's algorithm. The comparison result shows that the new algorithm can generate more reliable shortest path solutions that take into consideration failures of road segments and intersections due to congestion or traffic accidents. L. Yang et al. modified the desired minimum gap and structure of the Intelligent-Driver Model (IDM) to better account for the real-time maximum deceleration. They applied the new IDM to Adaptive Cruise Control (ACC) and simulated it under two road conditions. They found that the new model can improve the safety and stability of ACC. It can also increase roadway capacity. Accurate queue length estimation can be very useful in developing effective traffic control strategies to improve mobility and reduce traffic emissions. S Zhao et al. modified the Cell Transmission Model (CTM) for signalized intersection queue length estimation. The new CTM was tested based on the queue length data simulated by a calibrated VISSIM model. It was found to perform well under both undersaturated and saturated traffic conditions and can accurately model both queue forming and dissipating processes. Accurate traffic flow predictions are also important to traffic signal control. A. O. Adewumi et al. proposed a traffic flow prediction model based on Chaos Theory by computing the largest Lyapunov Exponent. To model how intersection and road segment failures propagate in a traffic network, Y. Zhang et al. proposed a Coupled Map Lattice (CML) method, based on which simulations were conducted to evaluate the impacts of various factors on network failures. As an essential component of autonomous/self-driving vehicles, Traffic Sign Detection and Recognition (TSDR) systems are very important for traffic safety. S. B. Wali et al. developed a fast automatic TSDR system based on RGB color segmentation, shape matching, and a support vector machine classifier. The developed system generated promising detection and false positive rates under various lighting and viewing angle conditions. J. A. Marmolejo et al. studied the design of capacity-constrained distribution networks to minimize fixed and transportation costs. They formulated the problem as a mixed-integer program and proposed a Benders decomposition algorithm to solve it. The Benders decomposition algorithm was compared with CPLEX. The results suggest that it can closely approximate the optimal solutions generated by CPLEX with significantly less computation time and is suitable for solving large-scale distribution network design problems. C. Wu et al. developed

a Transportation Supply-Demand Ratio (TSDR) model to quantitatively assess whether an urban area's transportation supply can meet its demand for sustainable development. This TSDR model is based on the system dynamic principle and was evaluated using VENSIM simulation.

Given the rapid development of high-speed rail in China, the aging rail infrastructure in developed countries, and many high-profile rail accidents, railroad safety has been receiving increasing attention globally. P. Xu et al. developed a railroad track condition prediction model and demonstrated its accuracy and robustness using track geometry data. L. Song et al. focused on analyzing the safety and reliability of high-speed Electric Multiple-Unit (EMU) trains. They integrated Fault Tree Analysis (FTA) and Petri nets and demonstrated their effectiveness in analyzing the failure modes of pantograph of high-speed EMU trains. J. Ye et al. proposed an improved Optimal Velocity (OV) car-following model for moving-block train control. Compared to the conventional OV car-following model, the improved OV model takes into consideration uncertainties in distance headway measurements and generates safer and more stable train control performance.

Extreme weather events have caused many airline flight disruptions. B. Zhu et al. developed a two-stage stochastic model to optimize the recovery process of airline flight disruptions. The first stage adjusts flight schedules to minimize delay and flight cancellation cost. The second stage focuses on minimizing aircraft rerouting cost. The authors also developed a simulated annealing algorithm to solve the two-stage model. G. Duan et al. focused on aircraft crash modeling. They improved the Functional Resonance Analysis Method (FRAM) by integrating it with model checking, which allows researchers to use the FRAM for identifying and simulating all possible ways that a hazard factor may lead to accidents. They demonstrated the enhanced FRAM's effectiveness by applying it to analyze a two-aircraft crash in Italy.

Public transit plays a major role in battling traffic emissions and global warming. To make public transit more attractive, it is important to have well-developed transit assignment models for optimal transit network design. S. Luo et al. investigated the independence between the travel times of adjacent transit links, which has been implicitly assumed to be independent in most transit assignment methods. They developed a method to determine the validity of the travel time independence assumption. They also proposed a capacity-restraint assignment method to handle the case where the travel time independence assumption is invalid.

This special issue includes a broad range of topics related to safe, resilient, and sustainable transportation systems. We received 47 manuscripts and 14 of them were accepted for publication, covering highway, rail, air, and transit modes.

Acknowledgments

We would like to sincerely thank all authors who contributed their papers to this special issue. Also, we are extremely grateful to all reviewers for their time and outstanding work, which are critical to the success and quality of this special issue.

Yuanchang Xie
Pan Liu
Chronis Stamatiadis

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2016, Article ID 5656734, 15 pages
http://dx.doi.org/10.1155/2016/5656734

Review Article

Application of Chaos Theory in the Prediction of Motorised Traffic Flows on Urban Networks

Aderemi Adewumi,[1] Jimmy Kagamba,[2] and Alex Alochukwu[1]

[1]*School of Mathematics, Statistics & Computer Science, University of KwaZulu-Natal, Westville Campus, Durban 4000, South Africa*
[2]*African Institute for Mathematical Sciences, Km 2 Route de Joal, Centre for IRD, BP 1418, M'Bour, Senegal*

Correspondence should be addressed to Aderemi Adewumi; laremtj@gmail.com

Received 2 June 2015; Revised 4 October 2015; Accepted 25 October 2015

Academic Editor: Pan Liu

In recent times, urban road networks are faced with severe congestion problems as a result of the accelerating demand for mobility. One of the ways to mitigate the congestion problems on urban traffic road network is by predicting the traffic flow pattern. Accurate prediction of the dynamics of a highly complex system such as traffic flow requires a robust methodology. An approach for predicting Motorised Traffic Flow on Urban Road Networks based on Chaos Theory is presented in this paper. Nonlinear time series modeling techniques were used for the analysis of the traffic flow prediction with emphasis on the technique of computation of the Largest Lyapunov Exponent to aid in the prediction of traffic flow. The study concludes that algorithms based on the computation of the Lyapunov time seem promising as regards facilitating the control of congestion because of the technique's effectiveness in predicting the dynamics of complex systems especially traffic flow.

1. Introduction

In recent times, urban traffic road networks are faced with severe congestion problems as a result of the accelerating demand for mobility. The excessive congestion in the form of immense traffic jams on urban roads has hindered mobility along these roads. This is one of the major challenges encountered in most mega cities around the world with urban road networks and in turn has a serious effect on road users which includes economic, health, and environmental problem such as vehicle emission and air pollution, arising out of increased fuel consumption during the long periods of congestion. U.S. Bureau of Transport Statistics in 2007 recorded that, due to traffic congestion, Americans residing in urban areas were coerced to travel more 4.2 billion hours and spent about $87.2 billion in purchasing extra 2.8 billion gallons of fuel [1, 2].

Urban planning and complex traffic network studies have been explored explicitly to potentially mitigate congestion and its associated problems on urban roads. Several efforts and studies have been made in time past by researchers on two major areas that affect urban traffic, namely, traffic flow modeling and prediction and information communications technology which is meant to give guidance to drivers through updated information about their desired routes [3]. However, without fundamental knowledge of the dynamics of vehicles on road networks, these studies were mainly based on costly and obsolete classical travel surveys on traffic flow and travel times and to some extent failed to provide the necessary information needed by road users in order to cope with the increasing urban demand for mobility [4].

One of the major concerns of traffic managers in traffic management system is traffic volume estimation, a major component of Intelligent Transport System (ITS), as it helps in the decision making and efficient traffic management planning when monitoring the current traffic flows in the road networks. Thus, to reduce the effect of congestion on urban road networks, accurate prediction of the Motorised Traffic Flow as well as traffic estimation is of paramount importance as it provides information on road accidents and level of congestion along the roads [3, 5]. Real time traffic flow data are useful for traffic volume estimation and help in forecasting traffic trends by determining the traffic flow patterns. Traffic data collection, predicting traffic patterns,

and forecasting traffic trends are usually performed for pavement design, fuel-tax revenue projection, and highway planning. However, the monitoring activities necessary for accurate Annual Average Daily Traffic (AADT) estimates are expensive in terms of costs and personnel. Thus, aside providing information for road accidents frequency and congestion, traffic estimation is also an issue for tactical purposes of transportation [6].

Based on the reports on experimental data found in literature, traffic flow patterns are highly predictable and often exhibit irregular and complex behaviours which changes abruptly when entering or leaving a congestion zone [3, 7]. Shang et al. in [8] reported on the irregularity and complexity of traffic flow as one approaches congestion zones in a traffic stream. They stated that the current nature and future dynamics of traffic flows highly depend on continuously interacting properties such as human behaviour and traffic characteristics. [9] noted that some of the main characteristics responsible for the complex behaviour in a traffic flow stream are variations in headways and spacing.

Several methods have been used in time past for short-term traffic flow prediction flows, including ARIMA-type models, Artificial Neural Networks, SARIMA models, Generalised Linear models, Nonparametric Statistical methods, Dynamic Neural Networks, Support Vector Regression models, and STARIMA models just to mention but a few. A brief review of some related work on Traffic Flow Prediction is presented below.

Catriona and Casper in [10] presented a Linear Multi-regression Dynamic Model (LMDM) which uses concept of graph for the traffic flow forecasting where the time series of flows at different sites are represented by the nodes and the structure between the flows at different sites as well as the independence is represented by the edges connecting the nodes. The idea of using graphical dynamic model approach in their work for traffic flow forecasting follows from that of [11, 12], respectively, with focus on forecasting traffic flows in two separate motorway networks using UK as a case study. Based on the distinctive features of their the LMDM, their model can be used for testing real-time instances. They illustrated how the LMDM can be used for forecasting and validated their model on some networks. The performance of the proposed approach was compared with other models in literature.

Dauwels et al. in [13] proposed a unified model by developing different forecasting models that is matrix and tensor based by applying partial least squares (PLS), higher order partial least squares (HO-PLS), and N-way partial least squares (N-PLS) for the time series prediction. Their focus was on collective prediction for multiple road segments and prediction-horizons against the known prediction for individual road segments and prediction horizons. One interesting feature of the developed models was the ability to carry out feature selection efficiently and simultaneously carry out traffic condition forecasting for multiple road segments and prediction-horizon. The computational performance of the proposed models which was validated on generic road networks consisting of expressway and arterial roads, in particular, an urban subnetwork in Singapore by performing a multihorizon speed prediction, showed that the proposed models performed better than the Support Vector Regression (SVR) traditional based model for longer prediction horizons. For the short prediction horizons, lower prediction errors were seen in SVR compared to the PLS based methods with N-PLS achieving higher accuracy when compared with PLS and HO-PLS. In fact their proposed unified models achieved same prediction accuracy as compared to the individual models but can be faster than the traditional based model for moderately sized networks.

ARIMA is one of the most precise methods for traffic flow prediction when compared to other known methods. In particular, Seasonal ARIMA (SARIMA) models have been shown to perform better than the other traditional based models but often times it is faced with some restriction in applicability as a result of using huge historical database for model development. Kumar and Vanajakshi, having this background knowledge in their work in [14], tried to overcome such drawbacks by proposing a prediction scheme approach using the SARIMA model for short-term traffic flow prediction which needs on limited input data for model development. They validated their proposed approach with using both historic and real-time data considering cases where peak period occurred both in morning and evening. The data used for the analysis and model development was from a 3-lane arterial roadway in Chennai, India, with limited flow data from 3 consecutive days. The results of the values of the predicted flows were compared with that of the actual flow values. Thus, the proposed approach will work in most cases where database is a major challenge when using ARIMA for traffic flow prediction model development.

Previous studies have shown that ANN has stable and consistent performance even if there is an increase in the travel time interval for the traffic flow prediction. This was so evident in [15] by Kumara et al. where ANN based model for a neural network was used for short-term prediction of traffic flow with heterogeneous condition for nonurban highway. Their model incorporates speed, density, traffic volume, and time as input variable but considers the speed separately in contrast to most work in literature where average speed of combined traffic flow was considered. For other works in literature that applied Artificial Neural Network or ARIMA-type models for traffic flow prediction, see [16–19]. Moreover, previous and recent research findings have shown that policy makers using existing traffic flow models to predict traffic flows have not been able to mitigate the congestion problem to fairly acceptable levels as expected and hence the need to come up with robust methodology for predicting traffic flows.

Chaos Theory is a novel science paradigm with numerous applications that have not been deeply explored and seems very promising with respect to the analysis and prediction of complex systems like traffic flows, although at the moment little empirical evidence exists to confirm this notion. It can be used to analyze the traffic flow patterns in urban road network by utilizing the intrinsic deterministic nature of the traffic flow in order to reduce congestion on urban road networks. In this paper, we report a systematic review of Chaos Theory and propose an approach to predicting Motorised Traffic Flow on Urban Road Networks based on

Chaos Theory with emphasis on the Largest Lyapunov Exponent method for prediction, the most common, effective, and direct technique of analyzing the presence of Chaos in a given dynamical system. This work contributes to this research field in the sense that the proposed approach is different from other conventional models found in literature and serves as an alternative method for predicting Motorised Traffic Flow on Urban Networks. Also, the effectiveness of the Largest Lyapunov Exponent prediction method seems very promising in terms of prediction accuracy as well as reducing the congestion problems on urban network, although this is yet to be fully validated using computer based algorithm on empirical traffic flow data.

The layout of this paper is as follows. The congestion problem on urban road networks is introduced in Section 1 with brief review of related works on Motorised Traffic Flow Prediction Models. Section 2 gave an insight on Motorised Traffic Flow and Traffic Flow Variability by highlighting some of the main characteristics for the complex behaviour of traffic flow stream. A systematic review of Chaos Theory is presented in Section 3 with emphasis on its application to the analysis and prediction of Motorised Traffic Flow in Urban Road Networks based on the Largest Lyapunov Exponent Prediction Method. The conclusion and directions for future work are drawn in Section 4.

2. Motorised Traffic Flow

2.1. Headway and Spacing. One of the applications of ITS as earlier mentioned is predicting road traffic volumes in order to make efficient traffic management and planning over a network as well as implementing road safety measures. [20] in reporting Shang et al.'s study in their paper noted that the differences in the distribution of various vehicle types, human driving habits (high driver perception-reaction times), space, and time headway, are among the principle causes of chaotic behaviour in traffic flows with the time and space headway been the main factors causing variations in observed traffic distributions and its transformation [9]. A proper knowledge of the above mentioned will be helpful in understanding traffic flow and to some extent provide theoretical foundation for short-term traffic flow forecasting. For the purpose of this study, our focus is on the linking space and time headways and variation of traffic flows on a given road network.

Based on the study carried in [9], suppose we have a traffic stream composed of two consecutive vehicles in a single lane road such that we have a follower-vehicle, i a leader-vehicle, and $i + 1$ as shown in Figure 1.

It can be observed that vehicle, i, is some distance, h_{s_i}, from its pacesetter, $i + 1$, termed as the space headway (usually expressed in metres, m). h_{s_i} comprises of the distance to the leader-vehicle, g_{s_i} (the space gap), and the self-length of the follower-vehicle, l_i. Hence, h_{s_i} is given by

$$h_{s_i} = g_{s_i} + l_i. \tag{1}$$

g_{s_i} is measured from the follower-vehicle's anterior bumper to the leader-vehicle's hind bumper. The hind bumper of the vehicle represents the vehicle's position. Thus, the space headway, h_{s_i}, can be expressed as

$$h_{s_i} = x_{i+1} - x_i. \tag{2}$$

From (1), each of the two vehicles has also a time headway associated with it. Thus, h_{t_i} (measured in seconds, s) comprises a time difference, g_{t_i}, and a time of occupancy, ρ_i, given by

$$h_{t_i} = g_{t_i} + \rho_i. \tag{3}$$

Both space and time headway can be envisaged in a space time diagram as shown in Figure 2. Thus, the positions x_i and x_{i+1} of the two vehicles, i and $i + 1$, can be plotted with respect to time, tracing out two vehicle trajectories, as the vehicles are in motion.

Figure 2 is called a time-space diagram. The respective speeds of the two vehicles can be derived from the diagram by drawing the tangent line. For simplicity, we assume that both vehicles travel at a constant speed resulting into parallel trajectories.

In single-lane traffic (microscopic traffic model), vehicles always keep their relative order. However, for multilane traffic (macroscopic traffic model), this principle can no longer be obeyed due to overtaking manoeuvres, resulting into irregular vehicle trajectories. If the same time-space diagram were to be drawn for several lanes (in multilane traffic), then some vehicles' trajectories would suddenly appear or fade away at the point where there exists a change of lane.

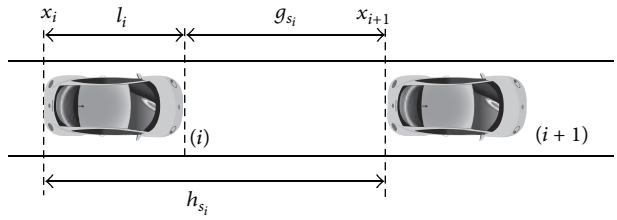


Figure 1: Two consecutive vehicles (a follower-vehicle, i at position, x_i and a leader-vehicle, and $i + 1$ at position, x_{i+1}) in a single lane road (after [9]).

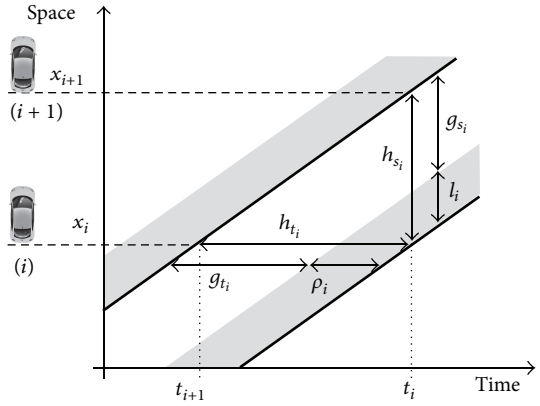


Figure 2: Trajectories of a two-car traffic stream (after [9]).

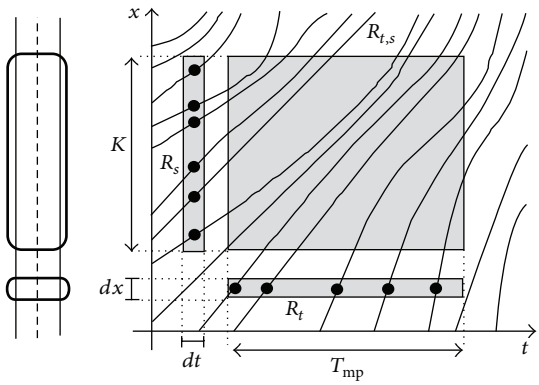


Figure 3: A time-space diagram showing nonlinear trajectories of several vehicles where movements are bounded by three regions of measurement, that is, R_t, R_s, and $R_{t,s}$ (after [9]).

Figure 3 shows the relative trajectories of vehicular traffic stream in a multilane facility.

In Figure 3, the three regions of measurement are always bounded in both time and space (that is to say, a period of measurement, T_{mp}, and a length of road section, K). Black dots were used to represent single measurements made in the diagram. The following describes what the three bounded regions represent:

(i) R_t represents measurements taken at certain locations that are fixed in space, (dx), done in T_{mp} time period. An example of such a measurement is one obtained by an underground automatic inductive loop.

(ii) R_s represents measurements taken at particular instances in time, (dt), along length, K, of a road section. An example is results taken from aerial photographs.

(iii) $R_{t,s}$ represents a region where general measurements are made. This region normally takes other forms of shape other than a rectangular one (as illustrated in Figure 3). An example of such a measurement is results of video cameras.

It becomes more complicated to represent the vehicle trajectories on the space diagram as a result of the disorderliness in the dynamics of the vehicle movements along the traffic stream. This causes variability in the traffic flow.

2.2. Traffic Flow Variability. Traffic flows are subject to variations over numerous time scales, namely, yearly, monthly, weekly, and daily. It also varies directionally as well as from place to place. Aside the fact that roads carry different volumes of traffic, the characteristics of the vehicles using these roads also change depending on the road facility [21]. For example, one road with about 10,000 vehicles per day may have very little truck traffic, while another road with the same volume of vehicles may have 2,000 trucks per day mixed with 8,000 ordinary cars. Similarly, one road section may be traversed by 1,000 heavily loaded trucks per day while a nearby road is used by 1,000 partially loaded trucks

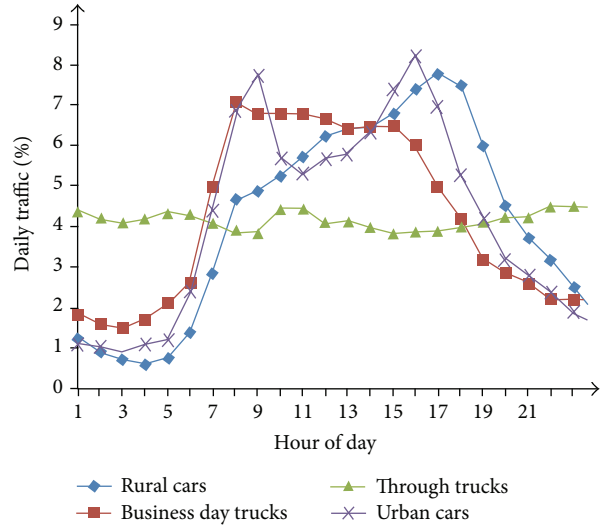


Figure 4: Vehicle volume distributions by classification of vehicles in California in 1996 (source: [22]).

(Traffic Monitoring Guide, 2013). We illustrate with the aid of graph the two major types of traffic volume variations, namely, Time-of-Day and Day-of-Week Variation based on the findings of the Federal Highway Authorities [3].

2.2.1. Time-of-Day Variation. The Federal Highway Administration (FHA) in 1996 reported that most truck travel falls into one of two basic time-of-day patterns, namely, a pattern that is centered on travel during the business hours of a day (working hours) and a pattern that shows almost constant travel all day through (twenty-four-hour day). Figure 4 summarizes the research findings of FHA in 1996.

As can be seen in Figure 4, cars tend to follow either the traditional two-humped urban commute pattern or the single-hump pattern commonly seen in rural areas, where traffic volumes continue to grow throughout the day until they begin to taper off in the evening. However, the truck pattern differs from the rural car pattern; in that it peaks in the early morning (many trucks make deliveries early in the morning to help prepare businesses for the coming workday) and tapers off gradually, until early afternoon, when it declines quickly. The other truck pattern (travel constantly occurring throughout the day) is common with long haul trucking movements. In addition, at any specific location, time-of-day patterns may differ significantly as a result of local trip generation patterns that differ from the norm. For example, Las Vegas, Nevada, generates an abnormal amount of traffic during the night because that city is very active late at night. In heavily congested urban areas, the commute period traffic volume peaks flatten out and can last three or more hours.

A close observation at Figure 4 reveals that cars tend to follow either the traditional two-humped urban commute pattern (double peaked pattern) or the single-hump pattern commonly seen in rural areas, where traffic volumes continue to grow throughout the day until they begin to taper off in

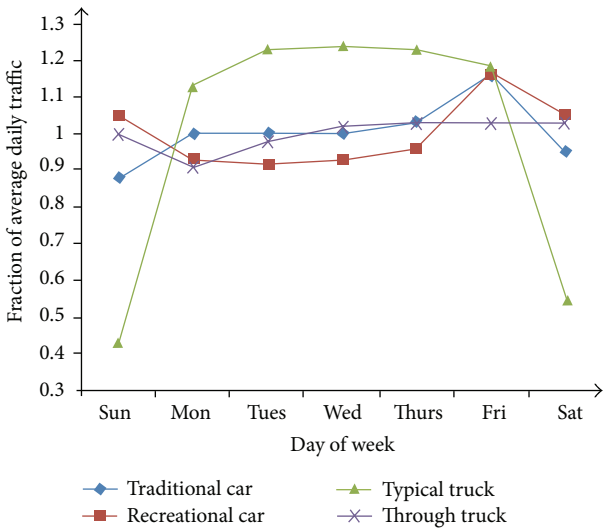


FIGURE 5: Day of the week traffic variations in California in 1996 (source: [22]).

the evening. However, the truck pattern differs from the rural car pattern; in that it peaks early in the morning due to the fact that many trucks make deliveries early in the morning to help prepare businesses for the coming workday and tapers off gradually, until early afternoon, when it declines quickly. The other truck pattern maintains a constant pattern throughout the day which is mostly common with long haul trucking movements.

Moreover, time-of-day patterns usually differ significantly with respect to places at any specific location as a result of local trip generation patterns that differ from the norm. For example, a city with night clubs or recreational facilities will generate an abnormal amount of traffic during the night hours or other hours of operation because that city is very active late at night. Also, in heavily congested urban areas, the commute period traffic volume peaks flatten out and can last three or more hours.

2.2.2. Day-of-Week Variation. The same study also revealed that there exists a large difference in daily patterns of the ordinary vehicle categories and typical trucks since truck travels are mainly business motivated as opposed to ordinary vehicles whose drivers have several travel objectives. Figure 5 illustrates the day-of-the-week variations.

It is evident from the graph that the day-of-week traffic variations are highly responsible for the traffic congestion that comes in form of jams on urban roads. A good example is the stampede observed along Kwame Nkrumah circle in Accra, Ghana, whose immense traffic jams are estimated to have caused annual losses of about $125 million to travelers along this road in 2014 (monetary value of lost time during traffic jam), as pointed out by traffic experts of the Ghana Institute of Engineers [23].

To mitigate this problem of congestion on urban roads, it is very necessary to carry out a substantial traffic estimation which requires a method of high precision to forecast a complex entity such as traffic flow. This is the main reason for proposing an alternative way of addressing complex systems like traffic flows, using effective techniques based on Chaos Theory (which studies dynamic systems) to analyse and predict traffic flow patterns.

3. Chaos Theory Review

3.1. Introduction. Several systems exist in everyday life that evolve with time. Such systems are difficult to predict accurately on long-term scale even with robust statistical prediction models. Examples of such system include weather, turbulent fluids (flowing across planes), population infected by epidemic, and stock market indices and they are generally referred to as dynamical systems [24].

These systems are said to exhibit "Chaos." Chaos in a simple term refers to any state of confusion or disorder that is showing the absence of some kind of particular order. Many work exists in literature that addresses dynamical systems as well as the chaotic behaviour. Kiel and Elliott in [25] described how many disorganised systems can spontaneously acquire organisation. For example, a shapeless liquid mass upon cooling can be transformed into an exquisite shape. Zhang and Jarrett in [26] studied the dynamic behaviour of road traffic flows in an origin-destination network. Their proposed dynamic model is a modification of the static conventional model by Dendrinos which also describes the traffic flow variability of the O-D network flows. They showed that the O-D flow patterns varies depending on whether the dimension is lower or higher. The characterization of the chaotic attractors by positive Lyapunov Exponents and fractal dimensions agrees with the fact that Largest Lyapunov Exponents provide the best measure of Chaos in any dynamical system. See [27] for details on the search for chaos in traffic-flow dynamics.

A Chaotic system can be described as one that is complex, aperiodic (it never exactly repeats), and sensitive to its initial conditions. Chaos Theory is novel Science paradigm in the field of nonlinear analysis which is used to describe the realms of nonrepeating and highly complex dynamic systems. This discipline is accredited to a meteorologist from the Massachusetts Institute of Technology (MIT), Muhmoudabadi [6], who described Chaotic systems to sensitively depend on initial conditions. He termed this behaviour as "*The Butterfly Effect*" (where the flap of butterfly's wings in Brazil sets off a tornado in Mexico). For clarity purposes, Chaotic processes should not be confused with random processes because Chaos does not imply randomness in any sense. Chaotic processes do not have any kind of distribution like random processes such as Brownian motion that exhibit a Gaussian distribution [28]. Furthermore, Chaotic processes are perfectly deterministic while random process are attached to some prior probabilities. Some properties of chaotic systems outlined below will help in understanding the behaviour of Chaotic systems.

3.2. Properties of Chaotic Systems. Chaotic systems have a number of distinctive characteristics which are used to

describe the dynamic evolution of such systems. These characteristics include the following.

Sensitivity to Initial Conditions. As already introduced in Section 3.1, Chaotic systems are highly dependent on initial conditions, a property sometimes regarded as "*The Butterfly Effect*." Two trajectories emerging from two different close-by initial conditions diverge exponentially from one another as the system evolves in phase space (a phase space is a representation of all possible states (configurations) of a dynamic system, and each possible state mapped by unique points [29]) [30]. In order to make accurate prediction of long-term behavior of Chaotic systems, the initial conditions must be known in their entirety and to high levels of precision.

Determinism. Chaotic systems are strictly deterministic. A deterministic system is one where for a given time interval there is only one future state that follows from the current state [31]. These systems can be described by Ordinary Differential Equations (ODE's). At least three variables are needed for Chaos in continuous-time systems as opposed to Chaos in discrete systems that requires only a single variable [29]. The reason is that the space time trajectories have to be aperiodic and finitely bounded in some region. However, it is unlikely to have a single trajectory intersecting itself due to the fact that every point has a unique mapping in space [29].

Nonlinearity. Intuitively, a nonlinear system is a system whose outputs and inputs are not proportional to each other. In other words, a nonlinear system is a system which cannot be decomposed into parts and reassembled into the same thing. This is a situation where the relationship between variables describing a system is not simply static or directly proportional to the output, but instead it is dynamic and varies [32]. Nonlinear dynamic systems exhibit nonlinear time series (discussed later in Section 3.4). In the case of nonlinearity, there is no periodicity (nonrepetitive system) as compared to linearity where the system repeats itself over a time period.

Instability. Chaotic systems have a sustainable irregular manner caused by sensitive dependence on initial conditions and thus predictions for a given system can only be made on short-term scales to high precision [29].

Attractors. These are d-dimensional sets of states, $\mathcal{X} \in \mathbb{R}^d$ (points in phase space) invariant under the system's dynamics where all states in close proximity asymptotically approach each other [33]. Many dynamic systems in nature have attractors and it has been discovered by researchers that all Chaotic systems' dynamics of evolution emerge into a certain type of attractors called strange attractors which are sensitively dependent on their initial conditions [24]. The four known types of attractors are briefly described as follows:

(i) Point attractor: a system is said to have a point attractor if the system evolves to a fixed point, for example, a single singing pendulum bob (see Figure 6(a)).

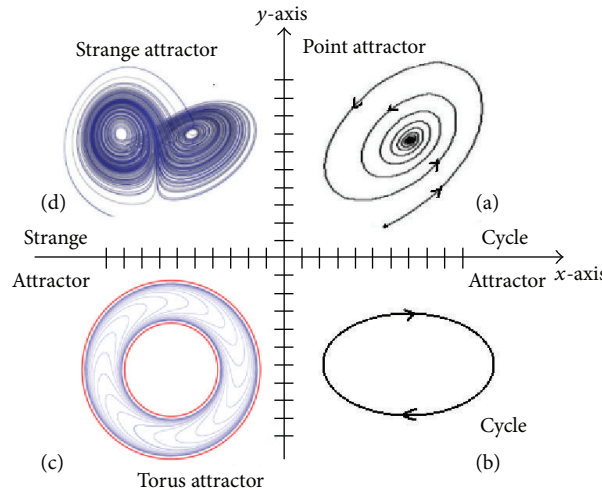


Figure 6: Different types of attractors constructed in 2-dimensional phase space; (a) point attractor, (b) limit cycle, (c) limit torus, and (d) strange attractor (after [29]).

(ii) Limit cycle: if the system is cyclic and its position in the cycle can be predicted, then the system is said to have a limit cycle, for example, planetary motions (see Figure 6(b)).

(iii) Limit torus: a system that has a limit torus is similar to that of a limit except that the system's trajectories are bounded within a region of a ring torus; for example, the "halo" ring of planet Jupiter is a torus composed of mainly dust particles in motion (see Figure 6(c)).

(iv) Strange attractor: if a system takes an aperiodic irregular shape and never repeats itself in time, the system is said to have a strange attractor. Such an attractor can also be described as a limit region (object with fractional (fractal) dimension) within phase space which is ultimately occupied by all trajectories of a dynamical system. Examples of such strange attractors include the famous Lorenz attractor illustrated in Figure 6(d) [30], Hénon attractor, and logistic map attractor.

Fractal Dimensionality. It is an already established fact that that the geometrical dimension of a line, plane, and box is 1, 2, and 3, respectively. However, many examples seen in our everyday life as well as many objects are not geometrically smooth like the ones mentioned above. Complex, noninteger dimensions are called fractal dimensions [35]. This is usually used to measure the complex nature of a given Chaotic system. When a Chaotic system's evolution is represented in phase space, the topological dimension, d, of the space state of the system's trajectories is a noninteger. A famous example of a plot with fractal dimension is Mandelbrot's plot ($z = z^2 + c$), which lies in the category of fractals, which are shapes that infinitely repeat themselves in smaller magnifications (scales) [34]. Figure 7 is an illustration of Mandelbrot's plot in a 2-dimensional complex plane.

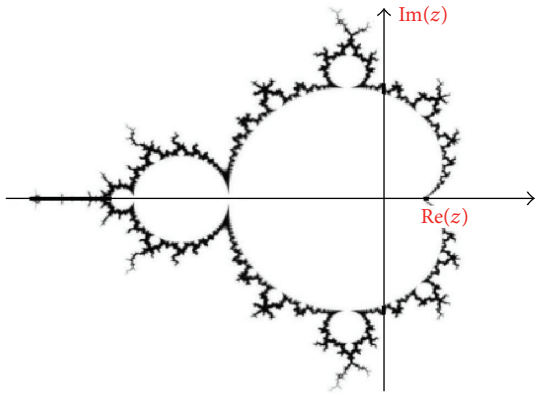


Figure 7: Mandelbrot's plot that is self-replicating according to some predetermined rule such that the boundary of the set has fractal dimensions (drawn in a 2-dimensional complex plane) (after [34]).

Other examples of shapes in nature with fractal dimensions include coastlines and slow flakes [34].

To summarize the properties of Chaotic systems, we note that there are two important characteristics that make chaotic systems very complex and our focus is on these characteristics:

(i) The strange attractor, which contains a large number of unstable system trajectories.

(ii) The ergodicity (ergodicity is a system behaviour that is averaged over time and space for all the system's states) in the dynamics of the system trajectories. In other words, as the system evolves temporarily, a small neighbourhood of every point in one of the unstable orbits within the attractor is visited [29, 36].

We note that in Chaos Theory, there is no need for prior knowledge of probabilities unlike in statistical physics. Under appropriate circumstances, it has been reported that algorithms based on Chaos Theory have shown the capability of attaining high level of performance, far better than those obtained using classical stochastic methods or techniques based on signal processing, and these can be applied in the following areas among others [37].

In meteorology, Chaos Theory is used to predict slight changes in weather, air, and aerosol movements in the atmosphere and so forth as studied by Lorenz in the late 1960s [30]; it is used in most biological processes such as heart beat detection, circadian rhythms, in particular, and electrocardiographic recording of a pregnant woman [29]. In economics and finance, Chaos Theory is used in foreign exchange rates and stock market indices for market crash forecasting. This is based on the Mandelbrot fractal hypothesis which predicts a market crash every two decades starting from 1987 up to date [34]. Moreover, Chaos Theory is also applied in traffic flow predictions, which is still an open and new area for research opportunities. This is the main motivation for this review [38].

3.3. Limitations and Control of Chaotic Systems. Although Chaotic systems have good characteristics that are suitable for analysis of complex behaviours, there are significant factors that hinders one from accurately predicting the behaviour of complex system. They include sensitive dependence on initial conditions which are in most cases unknown as most assumptions made often lead to error, the current stage of this "new" discipline of science (just half a century old) as one is not yet very sure of how much data is required to precisely reconstruct phase space and determine the fractal dimension of a given system (discussed in Section 3.4.1), the nature of the calculations involved in Chaos Theory which are repetitive, high extensive, and tedious which can only be done with the help of computers with high accuracy and precision [29, 30, 39].

Chaos systems can be controlled in order to reduce computational errors due to the adverse effects of the above limitations. Shewalo et al. in [29] stated that Chaos in systems can be controlled in exactly three ways which we have summarised below without full details (see [29] for full details). First, the systems parameters can be changed heuristically so that the range of fluctuations is limited. Secondly, one can apply perturbation to the Chaotic system which causes the system to organise itself using Ott-Grebogi-Yorke method, and finally the relationship between the system and the environment is changed using Pyragas method.

Having established the fact that Chaotic systems exhibit nonlinearity property with time evolution in previous sections, we now briefly describe the Chaotic Time Series and how it can be applied in the prediction of dynamical systems such as traffic flow based on the nonlinearity concept.

3.4. Chaotic Time Series Prediction. Phase space dynamics can be used to analyse and make predictions of dynamical systems. Nonlinear processes resulting in higher dimensional objects (called attractors when drawn in phase space) are characterised by nonlinear time series that intrinsically describe the behaviour of the system under study [40].

One can make prediction for a given time series using phase space techniques which is often referred to as the determinism test of a system. Such techniques are based on the fundamental fact that trajectories in close proximity asymptotically approach each other within the phase space [7]. A dynamical system can be represented in an m-dimensional finite dimensional vector space, $\mathbb{R}^m$, by the following equation:

$$\{X_{n+1}\} = F\left(X_n\right)_{\{n=0,1,2,\ldots\}}, \tag{4}$$

where n is the dimension of the vector space, $\{X_n\}$ represents n-phase space time evolution points, and $F(X_n)$ is an arbitrary function representing the system's behaviour (but is usually unknown). This is as a result of the fact that, in most cases, elements of X_n are very difficult to observe empirically; that is, one may only be able to measure a single variable for a given time series and still have no explicit knowledge on the system's nonlinear dynamics [20].

Taking a look at traffic systems in particular, they highly depend on human and physical factors in a given road facility

and this even becomes more complicated due to the presence of immeasurable quantities such as traffic laws and social codes. Nevertheless traffic flow patterns are deterministic and their time series have been found to be nonlinear [4, 20].

Since Chaotic systems exhibit a nonlinearity property, developing a Chaos prediction model for a given dynamical system is based on nonlinear time series analyses which mainly involves two steps, that is,

(i) reconstruction of the phase space from a given data set,

(ii) developing of a methodology for predicting the phase space dynamics.

These steps can be explored following Takens' Fundamental Embedding Theorem (from 1981) [33]. We note that the reconstruction of the original time series data is done using this theorem known as the foundation of all Chaos based predictions [36].

Theorem 1 (Takens' Theorem). *Let $\mathcal{X} \in \mathbb{R}^d$ be a bounded set. In the Cartesian product space of C^1 mappings on $\mathcal{X}$ and the C^1 function $\mathcal{X} \to \mathbb{R}$, there exists an open and dense subset, U such that if $(T, f) \in U$, then the reconstruction map, $R^{(m)}_{map}$, is embedding whenever $m > 2\dim(\mathcal{X})$. Moreover the embedding is continuously differentiable and also has a differentiable inverse (C^1 diffeomorphism). We have a deterministic system, $T : \mathcal{X} \to \mathcal{X}$, and we also have a read-out function, $f : \mathcal{X} \to \mathbb{R}$. If $m > 2\dim(\mathcal{X})$, there exists a precise deterministic rule, g, for predicting the next state of a time series.*

Interpretation of Takens' Theorem. The proof of Takens' Theorem is omitted in this work but the following definitions and interpretation of the theorem will give an understanding of the theorem.

Definition 2 (a diffeomorphism). A diffeomorphism is a map between manifolds (smooth space system states), which is differentiable and has a differentiable inverse.

$\mathcal{X}$ is called the attractor set corresponding to the following time series:

$$\{x_t\} = \{x_0, x_1, x_2, \ldots\} = \{f(x_0), f(x_1), f(x_2), \ldots\}, \quad (5)$$

and we can rebuild the system's dynamics by the rule, g, which states that

$$\dim(\mathcal{X}) \le m \le 2\dim(\mathcal{X}) + 1, \quad (6)$$

where all the d-dimensional manifolds (space-states) of the system's attractor, $\mathcal{X}$, can be embedded in an $m = (2d + 1)$-dimensional reconstructed space while preserving the geometrical invariants, and $d = \dim(\mathcal{X})$.

This simply means that all the information about the system's complex d-dimensional attractor can still be captured in the discretized reconstructed m-dimensional phase space. Based on the knowledge of the outcomes of Theorem 1, we can now determine the topological parameters of the system's attractor.

3.4.1. Reconstruction of Phase Space. During reconstruction, new space states are created that are (in the sense of diffeomorphisms) equivalent to the original space states so that the relevant geometrical properties of the system are always preserved. The set of reconstructed trajectories, X, corresponds to a matrix in which each row is a vector in phase space; that is,

$$X = \left[\vec{X}_1, \vec{X}_2, \ldots, \vec{X}_i, \ldots, \vec{X}_M\right]^T, \quad (7)$$

where $\vec{X}_i$ is the system state at discrete time, i, and for a real time series with N-points, $\{x_1, x_2, \ldots, x_N\}$, each $\vec{X}_i$ is denoted by

$$\vec{X}_i = \left[x_i, x_{i+\tau}, x_{i+2\tau}, \ldots, x_{i+(m-1)\tau}\right], \quad (8)$$

where τ is the reconstruction delay time (lag) and m is the embedding dimension.

Therefore the matrix X is $M \times m$ matrix given by

$$X = \begin{pmatrix} x_1 & x_2 & \cdots & x_M \\ x_{1+\tau} & x_{2+\tau} & \cdots & x_{M+\tau} \\ x_{1+2\tau} & x_{2+2\tau} & \cdots & x_{M+2\tau} \\ \vdots & \vdots & \vdots & \vdots \\ x_{1+(m-2)\tau} & x_{2+(m-2)\tau} & \cdots & x_{M+(m-2)\tau} \\ x_{1+(m-1)\tau} & x_{2+(m-1)\tau} & \cdots & x_{M+(m-1)\tau} \end{pmatrix}^T, \quad (9)$$

where the constants M, N, m, and τ are related by the equation $M = N - (m-1)\tau$ and by Theorem 1, $m > 2d$ where d is the dimension of the system's attractor.

Now, suppose we have a scalar observed nonlinear times series, say from empirical traffic data,

$$\{x_t\}_{\{t=1,2,3,\ldots,N\}}. \quad (10)$$

The vector for each reconstructed single point time series is given by

$$X_t = \{x_t, x_{t+\tau}, x_{t+2\tau}, \ldots, x_{t+(m-1)\tau}\}, \quad (11)$$

and it follows that

$$X_{t+1} = \{x_{t+1}, x_{t+1+\tau}, x_{t+1+2\tau}, \ldots, x_{t+1+(m-1)\tau}\}, \quad (12)$$

where τ is the time delay and m is the embedding dimension (as before), and that

$$[t + 1 + (m-1)\tau] \le N. \quad (13)$$

Consider Figure 8 illustrating the time series, $\{X_t\}$ and $\{X_{t+1}\}$ in a time-space diagram, and a phase space diagram, respectively.

Figure 8(b) gives a probable representation of the strange attractor of the data set, whose result is set of points of the above two time series plotted in 2-dimensional phase space. However the trajectories of the attractor (as in the diagram) may appear to intersect each other but they actually never cross even in higher dimensions.

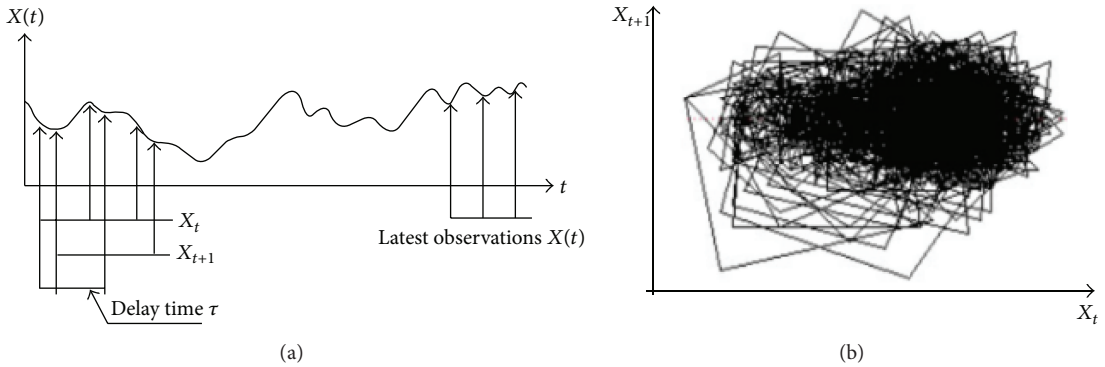


FIGURE 8: Illustration time series, $\{X_t\}$ and $\{X_{t+1}\}$, plotted in a space time (a) and a 2-dimensional phase space plot (b) (after [39]).

Parameters (topological parameters) such as the dimension of d, the attractor, τ, the delay time, and m, the embedding dimension, are necessary for reconstruction of the systems' dynamics in phase space before any predictions can be made. These parameters can be determined by the procedures in the order described below. First, we compute the delay time, τ, as follows.

(i) Determination of Delay Time, τ. There are several approaches for determining the delay time. The first approach as pointed out in [33] is by computing the Auto Correlation Function (ACF) of the data given by the following equation:

$$\mathscr{C}(r) = \frac{1}{N}\sum_{i=0}^{N-1}\left(x_i - \langle x\rangle\right)\left(x_{i+1} - \langle x\rangle\right), \tag{14}$$

where $\langle x\rangle$ is the arithmetic mean of the observations, given by

$$\langle x\rangle = \frac{1}{N}\sum_{i=0}^{N-1} x_i. \tag{15}$$

The choice of τ is determined by the duration after which x_i (or x_t) and x_{i+1} (or $x_{t+\tau}$) become uncorrelated, although [20] claims that it is difficult to obtain this.

Another method of determining τ is to calculate the nonlinear Auto Correlation Function called the Average Mutual Information (AMI), $I(\tau)$. AMI is a standard technique that tells us how much information we can obtain about a measurement taken from one time series, say $\{x_t\}$, that is affected by another measurement taken from another time series, $\{x_{t+\tau}\}$, sampled after a time interval, τ [41]. In other words $I(\tau)$ is a measure of the mutual dependence between two time series, and it is given by

$$I(\tau) = \sum_{ij} p_{ij}\ln p_{ij}(\tau) - 2\sum_{i} p_i \ln p_i(\tau), \tag{16}$$

where p_i is the probability that x_t takes the ith bin of a histogram, p_{ij} is the probability that x_t is in the ith bin, and $x_{t+\tau}$ is in the ith bin.

The concept of bin in a histogram will help in understanding how the information is obtained. We define a bin of a histogram intuitively with the following example. The bar graph of a histogram simply shows how many data points fit within a certain range. That range is called the bin (sometimes called the bin width). See Figure 9.

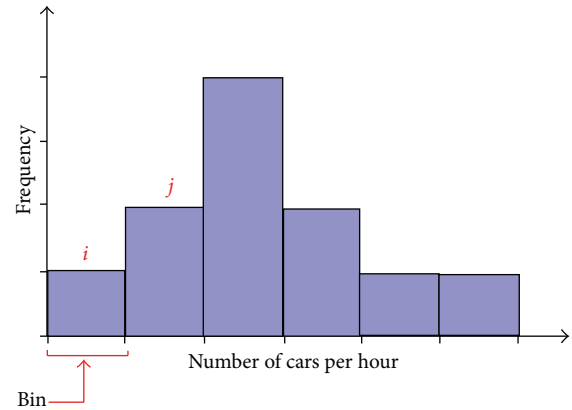


FIGURE 9: An example of a histogram plot illustrating arbitrary bins, i and j (after [42]).

For instance, suppose we want to plot a histogram graph after counting the number of cars passing through a certain area per hour. Using histogram chart in Figure 9, we might decide to plot it using the intervals 1–10, 11–20, 21–30, and so on. In this case, our bin would be 10 and every bar on your histogram represents a range of ten cars. The same data could be plotted on a range of 5 as 1–5, 6–10, 11–15, and so on. Here, our bin would be five. Obviously, the smaller the bin is, the more information we obtain about our data set, and vice versa. The narrower the bin is, the more you miss out on the point of a histogram.

Thus, we can compute the above probabilities (p_i and p_{ij}) and hence $I(\tau)$, by the Fraser and Swinney (1986) algorithm that is fully described by [42]. This algorithm can be directly applied to a given time series.

$I(\tau)$ is plotted against increasing values of τ and this plot is known as the AMI graph. This takes a shape such as the one illustrated in Figure 10.

To obtain the most appropriate value of τ, the first minimum in the AMI graph is chosen. This is because the first

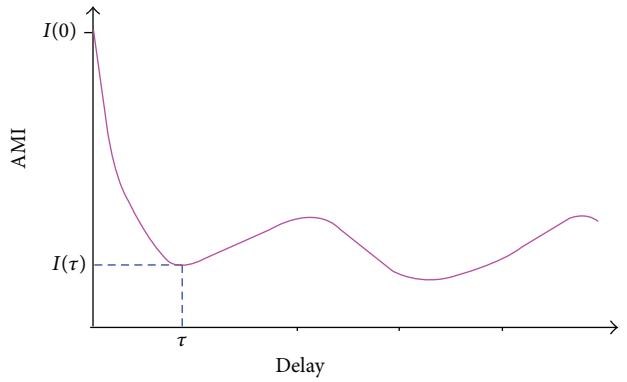


Figure 10: AMI against τ plot (after [41]).

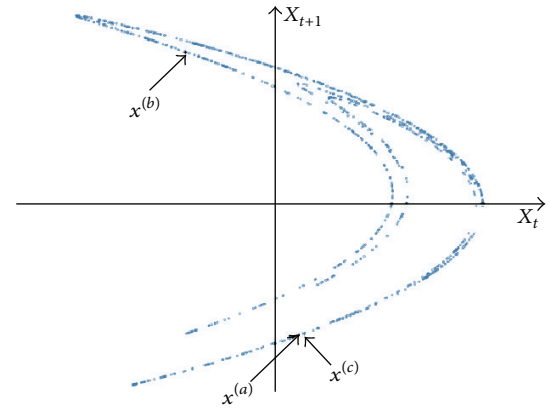


Figure 11: A 2-dimensional plot of the Hénon attractor showing $x^{(b)}$ called the FNN of $x^{(a)}$ and $x^{(c)}$ called the TN of $x^{(a)}$ (after [41]).

minimum preserves both the independence and correlation of the values of the two time series of x_t and $x_{t+\tau}$ and with this we can have a good approximation of the coordinates for the reconstructed vectors [7].

Claim. In [43], the criterion suggests that $I(\tau)/I(0) \approx 1/5$ if τ time series works well for down sampled data. $\tau \leq 5T_s$, where T_s is the sampling time of the data set.

Next, we compute the embedding dimension, m, as follows.

(ii) Determination of Embedding Dimension, m. This is done by computing the False Nearest Neighbours (FNN) Method [43]. This method is based on the assumption that two points that are in close proximity in the appropriate embedding dimension, m, must remain close as we move to higher dimensions [44]. However, if the embedding dimension is too small, then the points that are truly farther apart could seem to be neighbours, and such points are known as FNN.

Now, suppose 2 points, $x^{(a)}$ and $x^{(b)}$, are in close proximity in phase space. We compute the Euclidean distance of the 2 points given by $|x^{(a)} - x^{(b)}|$ in 2 consecutive embedding dimensions, m_0 and $m_0 + 1$ for $(m_0 \geq 2)$. Then, we determine whether a certain ratio (which is a function of the Euclidean distances in dimensions, m_0 and $m_0 + 1$) is greater than some predetermined value. One detects FNN within a given vector when the points close in dimension, m, move a significant distance apart in the following state while doing the computation. In dimension, m_0, the Euclidean distance is obtained as follows:

$$R_{m_0}^2 = \sum_{m=1}^{m_0-1} \left\{x^{(a)}(t + m\tau) - x^{(b)}(t + m\tau)\right\}^2. \quad (17)$$

Moving from dimension, m_0, to dimension, m_0+1, means that position of points in phase space changes by an amount equal to $x(t + d\tau)$ and this has a contribution to each delay vector. It follows that the Euclidean distance in the dimension, $m_0 + 1$, is given by

$$R_{m_0+1}^2 = R_{m_0}^2 + \left|x^{(a)}(t + d\tau) - x^{(b)}(t + d\tau)\right|^2. \quad (18)$$

The relative distance between the 2 dimensions gives the following relationship (a ratio):

$$\frac{\sqrt{R_{m_0+1}^2 - R_{m_0}^2}}{R_{m_0}} = \frac{\left|x^{(a)}(t + d\tau) - x^{(b)}(t + d\tau)\right|}{R_{m_0}}. \quad (19)$$

Based on this criterion, [43] states that if the ratio in (19) above is found to be greater than some predetermined value, R_{tol}, called the tolerance threshold, then the points $x^{(a)}$ and $x^{(b)}$ are characterised as "False Nearest Neighbour" (FNN).

In the same way, $R_{m_0+1} > \sigma/R_{\text{tol}}$, where σ is the statistical standard deviation of the attractor's time series data set around the mean, $\langle x \rangle$.

Claim. Reference [43] showed that, for several dynamical systems, $R_{\text{tol}} \approx 15$.

The authors of [20] stated that the claim presented in [43] was later empirically confirmed by a study on the eruption of Vatnajökull volcano of Iceland that $9 \leq R_{\text{tol}} \leq 17$, and a value of $R_{\text{tol}} = 10$ has proved to give good results. Thus, FNN is calculated for a given observed time series to determine the sufficient delay time necessary for phase space reconstruction.

Consider Figure 11 showing the Hénon attractor to help us intuitively understand the difference between FNN and "True Neighbours" (TN).

The above procedure is repeated for all possible pairs of points in dimensions of ascending order until the fraction of FNN drops to zero (or gets close to zero), a process usually termed as "unfolding" of the attractor. The percentage of FNN should drop to zero when the appropriate embedding dimension, m, is achieved.

For a given dynamical system such as traffic flow, a suitable value of R_{tol} has to be chosen although 10 is usually the best value as stated above. Based on this criterion, we note that a graph of the percentage of FNN against increasing values of embedding dimension, m, is plotted, which takes a shape similar to the one illustrated in Figure 12.

Normally, the value of m corresponding to the first minimum value of FFN% (for curve (a) in Figure 12) above

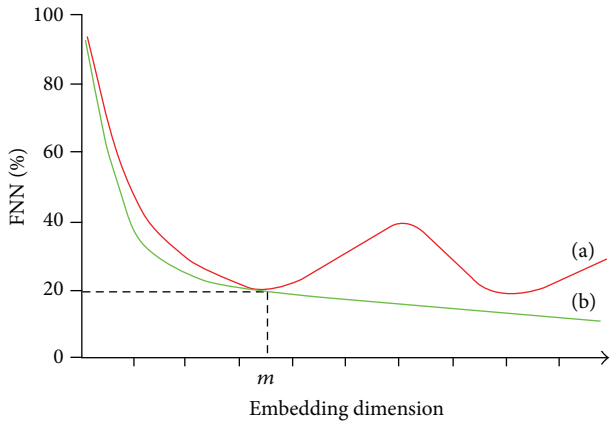


FIGURE 12: Plot of the percentage of FNN against embedding dimension, m (after [41]).

zero is taken as the most appropriate embedding dimension of the reconstructed time series. This is because by then the percentage of FNN has substantially reduced and the attractor is unfolded.

Noise Reduction. In the case of clean Chaotic data (having no random noise), it is expected that the percentage of FNN is reduced to zero when the proper embedding dimension is found. If the time series data is too noisy, however, it is likely that the method fails due to futile attempt of trying to unfold the noise in the data. Apart from determining the optimal embedding dimension, m, the FNN method is a good indicator of a noisy data set. From Figure 12, if FNN% converge in the range of increasing values of m (i.e., $\lim_{m\to\infty}$ FNN% $\to$ 0) as shown in curve (b) of Figure 12, then there is high possibility of random noise, which may be responsible for spreading the data, and therefore, it needs to be filtered [45]. As a stochastic process, noisy data must not unfold at any given dimension in phase space (in this case, we have no clear-cut minimum). Moving average and low-pass filter are commonly used methods for noise reduction in data sets although it is not discussed in this work [6].

We now discuss the different methodologies for prediction of Chaotic system's behaviour having discussed the topological parameters of the attractor.

3.4.2. Methodology for Prediction. Literature suggests that it is very necessary to check for Chaos in a given data set before predictions are made. The reason for the check is that there might be presence of random data, which are often assumed to be chaotic, in the data set.

There are several methods used to test for Chaos in a time series data set of a dynamical system. The following methods covered in this work were briefly discussed. They include computation of the (i) Correlation Dimension, d_c; (ii) Hurst Exponent, $\mathscr{H}$; (iii) Kolmogorov Entropy, K; and (iv) Largest Lyapunov Exponent (LLE), $\lambda_{\max}$.

(i) Correlation Dimension, d_c. This method has been widely used by physicists to test for Chaos in dynamical systems [33]. It provides a measure of which points in a given data set of an attractor affect each other. This parameter provides one of the best measures used in differentiating between stochastic and Chaotic systems.

The Correlation Function, $\mathscr{C}(r)$, is given by

$$\mathscr{C}(r) = \lim_{N\to 0} \frac{2}{N(N-1)} \sum_{i,j=1}^{N} H(u), \tag{20}$$

where $H(u)$ is the Heaviside step function given by

$$H(u) = \begin{cases} 1; & u > 0 \\ 0; & u \le 0, \end{cases} \qquad u = r - \left|X_i - X_j\right|, \tag{21}$$

where r is the radius of the sphere whose center is at X_i or X_j and N is the number of points in the reconstructed attractor's data set.

If the time series is characterized by an attractor, then

$$\mathscr{C}(r) \propto \mu r^{d_c}, \tag{22}$$

where μ is a constant of proportionality and d_c is the Correlation Dimension or the gradient of the $\log \mathscr{C}(r)$ against $\log r$ plot denoted by

$$d_c = \lim_{r\to 0} \frac{\log \mathscr{C}}{\log r}, \tag{23}$$

where d_c can also be estimated by the method of least squares or a smooth line over a certain range of m values referred to as the *scaling region*. This region can be estimated by determining the local slope given by

$$d_c = \frac{d[\log \mathscr{C}]}{d[\log r]}. \tag{24}$$

Reference [33] states that d_c provides the lower bound of the dimension, d, of the attractor and satisfies the inequality

$$d_c \le d. \tag{25}$$

To observe the existence of Chaos in the data, a plot of the Correlation Dimension against increasing embedding dimension values is obtained. The plot takes such a shape as illustrated in Figure 13.

If $d_c \in \mathbb{R} \setminus \{\mathbb{Z}\} < \infty$, then Chaos exists in the data set. The closest integer above the scaling region of the curve gives the least value of phase space variables used in the modeling of the actual dimension, d, of the attractor.

Note. If d_c is unbounded and is observed to increase with increasing embedding dimension, m, that is, $\lim_{m\to\infty} d_c \to \infty$, then the system is considered to be stochastic.

Now we define an upper bound for the d-dimension of the attractor called the *Limit capacity*, d_l, which satisfies the following inequality:

$$d_c \le d \le d_l. \tag{26}$$

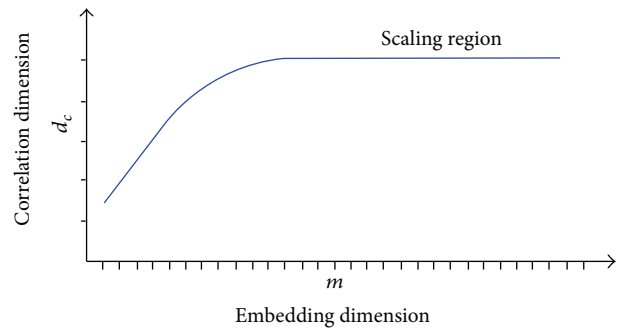


FIGURE 13: Plot of the Correlation Dimension, d_c, against embedding dimension, m (after [20]).

To determine d_l, we let $N(\epsilon)$ be the number of spheres of radius, r, for $0 < r < 1$, such that all the points of the attractor are covered by the spheres. Then it follows that

$$d_l = \lim_{\epsilon \to 0} \frac{\ln N(\epsilon)}{\ln (1/\epsilon)}. \tag{27}$$

In practice, we do not know the prior dimension, d, of the attractor and the most appropriate value of m of the newly reconstructed dynamics. Therefore, the dimensional estimate of d is found by increasing values of m (starting with $m = 2$) until a stable value of d is achieved (as described in Section 3.4.1(ii)).

(ii) Hurst Exponent, $\mathcal{H}$. Similar to the Lyapunov Exponents, a well-established parameter that is commonly used for testing for the Chaos in systems is the Hurst Exponent [38].

The Hurst Exponent, $\mathcal{H}$, is a measure of the degree to which a given time series can be statistically expressed as a random walk (i.e., Brownian motion).

If a time series vector, x_t, on average moves away from its original position by an amount that is directly proportional to $\sqrt{\Delta t}$ (where Δt represents a time interval), it is said that its Hurst Exponent is $1/2$ as stressed by [39] in reporting Kantz and Schreiber's work of 1997.

Therefore, one can determine whether the time series data is randomly distributed or not. This is obtained through the square root relation between increments after a certain time interval as follows:

$$\Delta x^2 \propto \Delta t^{2\mathcal{H}}, \tag{28}$$

where $\mathcal{H}$ is the Hurst Exponent and $0 \leq \mathcal{H} \leq 1$ and Δt is the time interval.

Reference [45] claimed that the relationship between the Hurst Exponent, $\mathcal{H}$, and Correlation Dimension, d_c is

$$d_c = 2 - \mathcal{H}. \tag{29}$$

In a data set where $\mathcal{H} = 1/2$, we conclude that the data is randomly distributed and is not correlated, while for $\mathcal{H} > 1/2$, we say that the data set has a positive correlation, and finally when $\mathcal{H} < 1/2$, the time data set has negative correlation.

(iii) Kolmogorov Entropy, K. A change in volume gives information about the sum of the corresponding Lyapunov Exponents which is equal to the Kolmogorov Entropy, K, given by

$$K = \sum_{\lambda_i > 0} \lambda_i, \tag{30}$$

where λ_i is the spectrum of Lyapunov Exponents (seen later in Section 3.4.2(iv)) [46].

For $N(\epsilon)$-number of spheres (as defined before in part (ii)) and embedding dimension, m, if a time series is completely deterministic (Chaotic), then

$$\lim_{N(\epsilon) \to \infty} \lim_{m \to \infty} K \longrightarrow 0. \tag{31}$$

On the other hand, for a completely random time series, the value K will not converge to single value, that is, $(\lim_{N(\epsilon) \to \infty} \lim_{m \to \infty} K \to \infty)$. Therefore, lower values of K imply higher predictability of the system and vice versa.

(iv) Largest Lyapunov Exponent, $\lambda_{\max}$. As far as we know, computation of Lyapunov exponents provides the best measure of Chaos in any dynamical system [46]. For this reason, we are going to explicitly explain and focus on this method since it is the most direct and most effective technique used for analysing the Chaotic behavior in a given dynamical system which is helpful in making predictions. Lyapunov exponents can clearly explain all the information contained in a time series. Thus, can be used to determine the length of the predicting period for any dynamical system, as argued out by [20].

Having established that the exponential divergence of nearby trajectories is the hallmark of Chaotic behaviour as explained by [30], the Lyapunov spectrum of exponents is given by

$$\lambda_{i\{i=1,2,\ldots,n\}}, \tag{32}$$

where n is the number of points in the reconstructed data set.

If the exponents are arranged in descending order such that

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n, \tag{33}$$

then the following relationships are true:

(i) The length of the principle axis of spectrum is proportional to $e^{\lambda_1 t}$.

(ii) The area determined by 2 principle axes is proportional to $e^{(\lambda_1 + \lambda_2)t}$.

(iii) The volume of the first k-principle axes is proportional to $e^{(\lambda_1 + \lambda_2 + \cdots + \lambda_k)t}$,

where t is time interval for the system to evolve from one state to another in phase space.

To understand the above relationships, we compute the he Euclidean distance between 2 points in phase space. Suppose

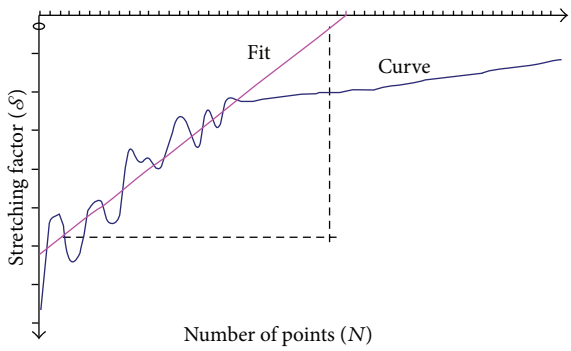


FIGURE 14: A plot of the Stretching Factor, $\mathcal{S}$, against number of points, N, in the data set (after [20]).

that originally we have 2 points in phase space that is $x^{(n_0)}$ and $x^{(n_1)}$ whose Euclidean distance is given by

$$\left|x^{(n_0)} - x^{(n_1)}\right| = \delta_0. \tag{34}$$

After a time interval, t, the system evolves and the new distance is given by $\delta = \delta_0 e^{\lambda_1 t}$, where $\lambda_1 > 0$, called the Lyapunov exponent. Thus, computing the Euclidean distances between points in consecutive higher dimensions will give the area and the volume, respectively.

Our focus is mainly on the Largest Lyapunov Exponent (LLE), $\lambda_{\max} = \lambda_1$, which gives evidence for determinism of a given system. In reporting Rosemstein et al.'s study, Shang et al. in [20] suggest that, after determining the most suitable topological parameters τ and m of the attractor, a point $x^{(n_0)}$ is chosen and all the neighbouring points $x^{(n_i)} = [x^{(n_1)}, x^{(n_2)}, \ldots, x^{(n_j)}]$, called True Neighbours (TN), closer than the distance, r (for chosen arbitrarily between 0 and 1), are found.

A number of N-trajectories are utilized in finding the closest points on the predicted trajectory, $x(t_1 + m\tau)$, which is used as the starting vector during the computation of the LLE. This procedure is repeated for N-number of points along the orbits and an average quantity, $\mathcal{S}$, known as the Stretching Factor given by (35) is calculated. One has

$$\mathcal{S} = \frac{1}{N}\sum_{n_i=1}^{N}\left(\ln\frac{1}{u_{x^{(n_0)}}}\sum\left|x^{(n_0)} - x^{(n_i)}\right|\right), \tag{35}$$

where $u_{x^{(n_0)}}$ is the number of neighbours around $x^{(n_0)}$.

Claim. Xue and Shi in [36] stated that if $20 \leq |u_{x^{(n_0)}}| \leq 30$, then a good approximation of the LLE can be obtained. A plot of $\mathcal{S}$ against the number of points N (or $t = N\Delta t$) yields a curve that has a linear inverse in one region which is followed by a plateau in another region. This plot takes the shape as illustrated in Figure 14.

The least squares approach gives a smooth line (fit) on Figure 14 and its slope gives an estimation of LLE, $\lambda_{\max}$.

Prediction. If $\lambda_{\max} \in]0, 1[$, then the system under analysis is not a Chaotic system but rather a stochastic one, and so we cannot make any predictions based on Chaos Theory.

If $0 < \lambda_{\max} < 1$, then it implies that there is Chaos in the system. For practical purposes, we compute the approximate period limit, $\Delta t_{\max}$ (often called *Lyapunov time*) for accurate prediction since it is a function of the LLE, $\lambda_{\max}$.

The Lyapunov time, $\Delta t_{\max}$, is given by

$$\Delta t_{\max} = \frac{1}{\lambda_{\max}}. \tag{36}$$

If $\lambda_{\max} \to 0$ implies $\Delta t_{\max} \to \infty$, then long-term accurate predictions are possible. Initially, one starts with a vector, $X(t_1)$, followed by selecting k-closest trajectories (not points) on the system's attractor which is then followed by choice of k-closest points to $X(t_1)$ (one on each trajectory). It follows that we precisely know the dynamic evolution of the system after time, $\Delta t_{\max}$.

In the same way, if $\lambda_{\max} \to \infty$ implies $\Delta t_{\max} \to 0$, long-term accurate predictions are not possible, but rather short-term ones can be made. With the Lyapunov time, $\Delta t_{\max}$, we can precisely predict any observed quantity (say traffic flow) for this time [20].

Practically in traffic flow analysis, the one-dimensional traffic flow time series data is replaced with m-dimensional reconstructed data. The reconstructed time series data is then plotted, and this is followed by analysis of the previous observations which are neighbours to the preceding ones, and short-term predictions are finally made.

4. Conclusions

This study have shown how Chaos Theory can be used in the analysis of dynamical systems via a systematic review of the characteristic features of Chaotic system. In particular, it showed how Chaos Theory can be used for Motorised Traffic Flow Time Series Prediction in Urban Transport Network based on the the method of computation of the Largest Lyapunov Exponent, $\lambda_{\max}$, which is the best method so far for analysis and prediction of chaotic behaviours of a given complex system like traffic flow as reported by most researchers in literature. Using the Largest Lyapunov Exponent prediction method, it was shown how the Lyapunov time, $\Delta t_{\max}$, can be obtained which is the time interval for making accurate predictions of traffic flows.

In order to make a complete and robust prediction model for traffic flow, there is need to develop a computer based algorithm that will compute the time delay, embedding dimension, and Lyapunov time of a real time series from empirical traffic flow data. Thus, the validation aspect of the proposed approach and comparison with other known conventional models for traffic flow prediction especially in the area of prediction accuracy is still in progress and left for our future work so as to enable us have access to available traffic flow data sets. Moreover, there is need to come up with a concrete relationship (most preferably a mathematical equation) that links the Lyapunov time with traffic flow so as to aid in proper traffic predictions. The effect of noise on traffic flow data as well as determining the type of noise and magnitude is also an important area to look into in our future work. Thus, by effectively incorporating all these into

the present work, it is believed that the proposed approach will be useful in reducing the congestion problem on urban traffic networks.

Conflict of Interests

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

Acknowledgments

The support of Dr. Abbas Mahmoudabadi of the Road Maintenance and Transport Organization, Tehran, Iran, is appreciated. The financial support of the African Institute for Mathematical Sciences (AIMS), Senegal, is hereby acknowledged.

References

[1] M. Garavello and B. Piccolli, *Traffic Flow on Networks: Conservation Laws Model*, American Institute of Mathematical Sciences, Springfield, Mo, USA, 2006.

[2] U.S. Bureau of Transportation Statistics, National Transportation Statistics Electronic, January 2015, http://www.rita.dot.gov/bts/sites/rita.dot.gov.bts/files/publications/national_transportation_statistics/index.html.

[3] N. H. Gartner and G. Improta, *Urban Traffic Networks: Dynamic Flow Modeling and Control*, vol. 2 of *Transportation Analysis*, Springer, Rome, Italy, 2nd edition, 2011.

[4] A. Y. Abul-Magd, "Modeling highway-traffic head-way distributions using superstatistics," *Physical Review E*, vol. 76, no. 5, Article ID 057101, 2007.

[5] G. Mwesige, H. Farah, U. Bagampadde, and H. N. Koutsopoulos, "A stochastic model for passing rate at passing zones on two-lane rural highways in Uganda," in *Proceedings of the 93rd Annual Meeting of the Transportation Research Board*, Washington, DC, USA, January 2014.

[6] A. Muhmoudabadi, "The assessment of applying chaos theory for daily traffic estimation," in *Proceedings of the International Conference on Industrial Engineering and Operations Management*, vol. 1, Tehran, Iran, January 2014.

[7] A. S. Nair, L. R. Jyh-Cham Liu, and G. Saurabh, "Non-linear analysis of traffic flow," in *Transportation Systems: Theory and Application of Advanced Technology, IFAC Symposium, Tianjin, PRC*, vol. 1, pp. 773–784, 1999.

[8] P. Shang, M. Wan, and S. Kama, "Fractal nature of highway traffic data," *Computers & Mathematics with Applications*, vol. 54, no. 1, pp. 107–116, 2007.

[9] B. De Moor and S. Maerivoet, *Traffic Flow Theory*, vol. 8999.40.-a, 02.50.-r,45.70, 2008.

[10] M. M. Q. Catriona and J. A. Casper, *Forecasting Traffic Ows in Road Networks: A Graphical Dynamic Model Approach*, Department of Mathematics and Statistics, The Open University, Milton Keynes, UK, 2008.

[11] M. E. Whitlock and C. M. Queen, "Modelling a traffic network with missing data," *Journal of Forecasting*, vol. 19, no. 7, pp. 561–574, 2000.

[12] C. M. Queen, B. J. Wright, and C. J. Albers, "Eliciting a directed acyclic graph for a multivariate time series of vehicle counts in a traffic network," *Australian & New Zealand Journal of Statistics*, vol. 49, no. 3, pp. 221–239, 2007.

[13] J. Dauwels, A. Aslam, M. T. Asif et al., "Predicting traffic speed in urban transportation subnetworks for multiple horizons," in *Proceedings of the 13th International Conference on Control Automation Robotics & Vision (ICARCV '14)*, pp. 547–552, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, December 2014.

[14] S. V. Kumar and L. Vanajakshi, "Short-term traffic flow prediction using seasonal," *European Transport Research Review*, vol. 7, no. 3, article 21, 2015.

[15] K. Kumara, M. Paridab, and V. K. Katiyar, "Short term traffic flow prediction for a non urban highway using artificial neural network," *Procedia—Social and Behavioral Sciences*, vol. 104, pp. 755–764, 2013, Proceedings of the 2nd Conference of Transportation Research Group of India (2nd CTRG).

[16] H. Dong, L. Jia, X. Sun, C. Li, and Y. Qin, "Road traffic flow prediction with a time-oriented ARIMA model," in *Proceedings of the 5th International Joint Conference on INC, IMS and IDC*, pp. 1649–1652, Seoul, Republic of Korea, August 2009.

[17] D. K. Parbat and P. B. Nagarnaik, "Artificial neural network modeling of road traffic noise descriptors," in *Proceedings of the 1st International Conference on Emerging Trends in Engineering and Technology (ICETET '08)*, pp. 1017–1021, Civil Engineering Department, G. H. Raisoni College of Engineering, Nagpur, India, July 2008.

[18] G. Fusco and C. Colombaroni, *An Integrated Method for Short-Term Prediction of Road Traffic Conditions for Intelligent Transportation Systems Applications*, Department of Civil, Constructional and Environmental Engineering, Sapienza University of Rome Via Eudossiana, Roma, Italy, 2013.

[19] G. A. Ali and C. F. Bakheit, "Comparative analysis and prediction of traffic accidents in Sudan using artificial neural networks and statistical method," in *Proceedings of the 30th Annual Southern African Transport Conference (SATC '11)*, Sudan University of Science and Technology, Pretoria, South Africa, July 2011.

[20] P. Shang, X. Li, and S. Kamae, "Chaotic analysis of traffic time series," *Chaos, Solitons & Fractals*, vol. 25, no. 1, pp. 121–128, 2005.

[21] Q. C. Diliman, "Traffic impact assessment for sustainable traffic management and transportation planning in Urban areas," in *Proceedings of the Eastern Asia Society for Transportation Studies*, vol. 5, pp. 2342–2351, 2005.

[22] F. L. Hall, *Traffic Stream Characteristics, Traffic Flow Theory*, US Federal Highway Administration, Washington, DC, USA, 1996.

[23] E. Appiah, http://www.myjoyonline.com/news/2015/may-19th/we-lose-over-125m-just-staying-in-traffic-experts-say.php.

[24] P. Cvitanovic, *Chaos, and What to Do About It?*vol. 1 of *Ebook*, 2015.

[25] L. D. Kiel and E. W. Elliott, *Chaos Theory in the Social Sciences: Foundations and Applications*, University of Michigan Press, Ann Arbor, Mich, USA, 1997.

[26] X. Zhang and D. F. Jarrett, "Chaos in a dynamic model of traffic flows in an origin-destination network," *Chaos*, vol. 8, no. 2, pp. 503–513, 1998.

[27] D. S. Dendrinos, "Traffic-flow dynamics: a search for chaos," *Chaos, Solitons & Fractals*, vol. 4, no. 4, pp. 605–617, 1994.

[28] E. Pepke, "Chaos Theory: What is the Difference Between Chaotic Behaviour and Random Behaviour?, Differentiating Chaotic and Random Processes," September 2004.

[29] G. Shewalo, N. Shinde, A. Shirode et al., *Report on Chaos Theory*, Sardar Patel Institute of Technology, Mumbai, India, 2012.

[30] E. N. Lorenz, *The Essence of Chaos*, vol. 1, University of Washington Press, Seattle, Wash, USA, 1st edition, 1993.

[31] S. S. Tiberio, *Stochastic and Deterministic Differential Equation Modeling: The Accuracy of Recovering Dynamic Model Parameters of Change*, University of Notre Dame, 2004.

[32] B. Dictionary, Non-Linearity, Definition of Non-linearity, 2015.

[33] M. Leok and B. Tiong, "Estimating the attractor dimension of the equatorial weather system," *Acta Physica Polonica A*, vol. 85, pp. 27–35, 1994.

[34] B. B. Mandelbrot, *The Fractal Geometry of Nature*, vol. 2 of *Cataloging in Publication Data*, Library of Congress, Washington, DC, USA, 1977.

[35] B. Goertzel, *Chaotic Logic: Language, Thought, and Reality from the Perspective of Complex Systems Science*, vol. 9, Springer, 1994.

[36] J. Xue and Z. Shi, "Short-time traffic flow prediction based on chaos time series theory," *Journal of Transportation Systems Engineering and Information Technology*, vol. 8, no. 5, pp. 68–72, 2008.

[37] S. J. Clement, M. A. P. Taylor, K. V. K. Rao, T. V. Mathew, and P. J. Gundaliya, *From Chaotic Road Traffic to Cooperative Opportunistic Percolation Using Cellular Automata*, NSW Transport and Population Data Centre, 2005.

[38] C. Frazier and K. M. Kockelman, "Chaos theory and transportation systems: an instructive example," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1897, no. 1, 2004.

[39] L. W. Lan, F.-Y. Lin, and A. Y. Kuo, "Testing and prediction of traffic flow dynamics with chaos," *Journal of the Eastern Asia Society for Transportation Studies*, vol. 5, 2003.

[40] J. D. Farmer, E. Ott, and J. A. Yorke, "The dimension of chaotic attractors," in *The Theory of Chaotic Attractors*, pp. 142–149, Springer, 2004.

[41] D. H. Campos, "Non-linear time series analysis of the EEG during sleep," *Revista de la Academia Colombiana de Ciencias Exactas*, vol. 20, no. 78, pp. 491–501, 1996.

[42] C. J. Cellucci, A. M. Albano, and P. E. Rapp, *Statistical Validation of Mutual Information Calculations: Comparisons of Alternative Numerical Algorithms*, Bureau of Medicine and Surgery of the Navy, Washington, DC, USA, 2004.

[43] H. D. I. Abarbanel, *Analysis of Observed Chaotic Data*, Institute for Nonlinear Science, Springer, New York, NY, USA, 1996.

[44] Ixellence, "False nearest neighbors (FNN)," May 2015.

[45] M. S. Khan and T. A. Siddiqui, "Estimation of fractal dimension of a noisy time series," *International Journal of Computer Applications*, vol. 45, no. 10, 2012.

[46] L. W. Lan, F.-Y. Lin, and A. Y. Kuo, "Identification for chaotic phenomena in short-term traffic flows: a parsimony procedure with surrogate data," *Journal of the Eastern Asia Society for Transportation Studies*, vol. 6, pp. 1518–1533, 2005.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 196107, 11 pages
http://dx.doi.org/10.1155/2015/196107

Research Article

Extended FRAM by Integrating with Model Checking to Effectively Explore Hazard Evolution

Guihuan Duan, Jin Tian, and Juyi Wu

School of Reliability and Systems Engineering, Beihang University, Beijing 100191, China

Correspondence should be addressed to Jin Tian; tianjin@buaa.edu.cn

Received 8 June 2015; Accepted 12 October 2015

Academic Editor: Yuanchang Xie

Functional Resonance Analysis Method (FRAM), which defines a systemic framework to model complex systems from the perspective of function and views accidents as emergent phenomenon of function's variability, is playing an increasingly significant role in the development of systemic accident theory. However, as FRAM is typically taken as a theoretic method, there is a lack of specific approaches or supportive tools to bridge the theory and practice. To fill the gap and contribute to the development of FRAM, (1) function's variability was described further, with the rules of interaction among variability of different functions being determined and (2) the technology of model checking (MC) was used for the analysis of function's variability to automatically search the potential paths that could lead to hazards. By means of MC, system's behaviors (normal or abnormal) are simulated and the counter example(s) that violates the safety constraints and requirements can be provided, if there is any, to improve the system design. The extended FRAM approach was applied to a typical air accident analysis, with more details drawn than the conclusions in the accident report issued officially by Agenzia Nazionale per la Sicurezza del Volo (ANSV).

1. Introduction

With the increasing complexity in sociotechnical systems, accident models are playing significant roles in explaining why an accident occurs, and the ways that hazards go to an accident can be identified based on the understanding of accident causes. At present, there are three categories of accident models widely acknowledged: the causal sequence models represented by domino model [1], the epidemiological models like Swiss cheese model [2], and the systematic models such as FRAM (Functional Resonance Analysis Method) [3], and STAMP (Systems-Theoretic Accident Model and Processes) [4]. Different from the former two categories in which accident was considered as a sequence of a series of unexpected incidents or as a result of combinations among factors involving human, working environment and media, the systematic models take accident as emergence due to nonlinear interactions among technical, human, and organizational factors within sociotechnical systems and explain the accident by determining the latent deviations of the system operations from what they should be.

As a typical systemic accident model, FRAM is capable of comprehensively analyzing complex sociotechnical systems from the perspective of function and describing interactions and couplings among the functions. Based on FRAM, an accident would be taken as "resonance" of multiple functions' variability, which is an innovative perspective to look at accident and effectively assist safety analysis and accident investigation. Nevertheless, FRAM is on the way of continuous development for the reasons below.

On one hand, the descriptions of function's variability and the spreading rules among functions need to be further elaborated to ensure the rigor and comprehensiveness of variability analysis. In terms of the classic FRAM, six aspects of a function are described and the variability going between aspects of upstream and downstream functions is explained with a rough categorization of "timing" and "precision," and it lacks details such as the rules about how the variability spreads from a function to another. Hence, we believe that the bias in analysts' mind can hopefully be minimized when they conduct variability analysis, if a set of rigorous and practical instructions are available.

On the other hand, FRAM is more like a conceptual method so far than a mature model [5], and it is significant to call on some corresponding approaches or supporting tools based on FRAM so as to contribute to the development of FRAM as such. Since the variability in a single function and between functions is basically hand-picked, there tends to be low efficiency and poor thoroughness in the analysis. Hence, the efficient supporting tools or approaches, based on the function variability and spreading rules defined by FRAM, are necessary to ensure that all the potential states and behaviors of the given system be checked, with the aim of determining whether the paths that variability spreads among functions may lead to an accident. Both the completeness and efficiency can be guaranteed with the aid of appropriate computer tools such as model checking [6], by which all the potential function states and sequences can be automatically searched.

Therefore, in this paper, FRAM was enhanced to explore the paths of hazard evolution by integrating with model checking. The rest of this paper is structured as follows: the background research and related work are reviewed in Section 2, and the method is described in Section 3. An air accident is taken into case study to illustrate the proposed approach in Section 4. Finally, Section 5 sets out conclusions and future work.

2. Literature Review

As a systemic accident model, FRAM was presented first by Hollnagel [5, 7]. It was pointed out that accidents were the resonance and amplification among functions' variability. In FRAM, structural models were used to describe functions and further to analyze aggregations of function variability. FRAM can be used to identify functional or logic deficiencies in system design, in addition to failures in hardware or software. Some comparisons were made between FRAM and other methods to discuss the pros and cons of both. In 2008, Hollnagel [8] elaborated the shortages of traditional safety analysis technologies and the advantages of FRAM, and concluded that FRAM could be more beneficial to facilitate safety analysis of key information system. Based on the research, Herrera and Woltjer [9] compared Sequentially Timed Events Plotting (STEP) with FRAM from aspects of rationale and application, respectively. The results show that FRAM can identify the accident causes that were not found with STEP and justify the advantage of FRAM to analyze the nonlinear and dynamic systems, such as sociotechnical systems. In addition, facing the challenge of current accidents, Hovden et al. [10] suggested that new theories, models, or methods such as FRAM be developed aiming at the "foresight" for accident prevention.

Moreover, FRAM has been used in different fields and demonstrates its significance and contribution to industrial practice, specifically for accident investigation and accident analysis. Based on FRAM, Woltjer [11] discussed the categories of all the contributing factors in aviation accidents and then explained how the resonance happened among human, technical, and organizational causes. Sybert et al. [12] pointed out that FRAM lacked system assessment on interactions between functions and variability in performance during hazard identification in Air Traffic Control (ATC), by analyzing the elastic characteristics of ATC system and confirming the variability existing in the system behaviors [13]. Besides, FRAM was applied to analyzing air accidents, and its effectiveness was verified by Hollnagel et al. [14] and Sawaragi et al. [15]. FRAM has also been adopted in train control, nuclear power, and electric systems; for example, Belmonte et al. [16] analyzed the safety of Automatic Train Monitoring System (ATS) through FRAM, and Macchi et al. [17] applied FRAM on the maintenance in nuclear power plant to explain the principle of the local maintenance activities and the possible influences on system safety.

The applications above indicate that FRAM can facilitate safety analysis and accident investigation and contribute to identification of more details of hazards than the traditional methods. However, in order to further develop FRAM, one of the key points is how to determine function variability and the rules of variability spreading from a function to another, as well as how to conduct efficient and complete search (based on the spreading rules) through all the combinations of the function variability. To make the rules derived and the search realizable, model checking can be adopted. Model checking is a widely used technology with which system's behaviors are described as transition among system states, and system properties are represented with temporal logic formulae, and thus all the possible behaviors can be automatically searched for any unexpected state sequences [6].

At present, there is significant development in model checking. Many logic expressions and rules have been extended to get adapted better to different systems [18–20]. Furthermore, great efforts were made to solve the key problems from which model checking often suffered such as the "explosion" of state space and the time synchronization [21, 22]. Model checking has a very wide range of applications, covering software, network, chemical industry, and other fields [23–25], in which it is taken as a comparatively mature means for justifying whether the system meets a given specification by modeling and simulating a complex system. Particularly, model checking has been used to develop safety analysis and random probability analysis from the perspective of function in hybrid systems [26]. With enlightenment from the application above, it is assumed that model checking can be used to simulate potential function states and sequences based on FRAM. Gao et al. [27] extended continuous stochastic logic to conditional continuous stochastic logic (CCSL) by introducing a conditional probabilistic operator to describe a richer class of properties for continuous-time Markov chains. In Rushby's research [28], model checking was used to analyze the autopilot accident, and it was concluded that the accident was caused by the fact that the cognitive process of human communication had not yet been covered completely. Zhang et al. [29] applied model checking to building information system to evaluate the risks and develop preventive measures, and Lahtinen et al. [30] discussed the sense that model checking has made in the nuclear industry and proposed a systemic method to justify the model when using it for safety-critical systems.

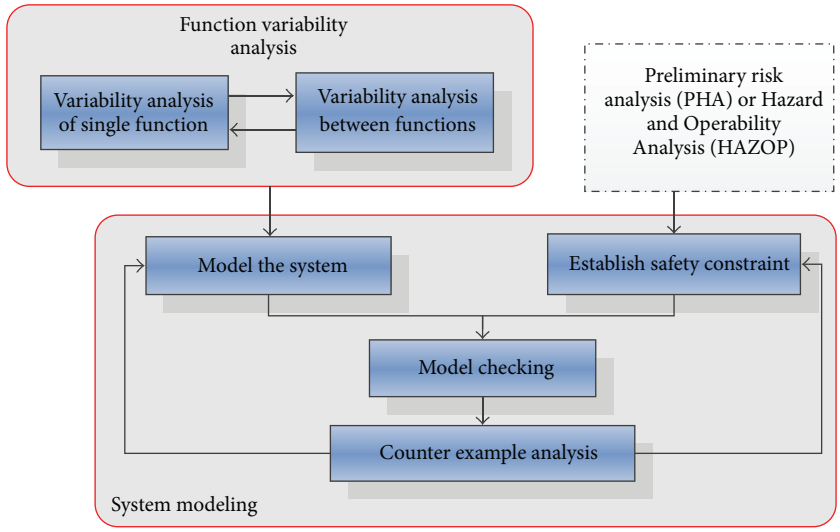


Figure 1: The framework of hazard identification based on the extended FRAM.

Overall, the existing research provides the evidence that it is feasible to analyze accidents by means of model checking, but model checking has not yet been adopted for safety analysis from the perspective of system functions and their variability, although it was recognized to be able to model and simulate system functions. Accidents would be explained in more details by simulating variability of system's behaviors, if combining model checking and FRAM, to demonstrate the scenarios describing why and how accidents occur.

3. Method

Hazards are defined as the states that may lead to any accident or unexpected event [31]. Based on the rationale of FRAM, the hazard evolution can be viewed as a process that variability is propagated among functions with increasing (or decreasing) magnitude that may violate the safety constraints and lead to accidents. The framework of identification of hazards and their evolution is illustrated in Figure 1. Firstly, FRAM was extended by redefining and clarifying some critical terms and deriving the criterion for function variability, both which were taken as supplementary to the original FRAM. Based on the extended FRAM, an approach of system modeling with detailed steps was provided to show how model checking was used to search all the system states for the potential paths which may lead the system to an accident. The two parts are marked with red blocks in Figure 1 and elaborated in the following subsections, respectively.

3.1. Function Variability Analysis. According to the original FRAM, the characteristics of function are described as the hexagon shown in Figure 2, and the six angles are labeled with the aspects of function: Input, Output, Precondition, Resource, Time, and Control, respectively.

Input is used to start or begin a function; Output represents the product or outcome generated from a function; Precondition is the conditions that need to be ready before

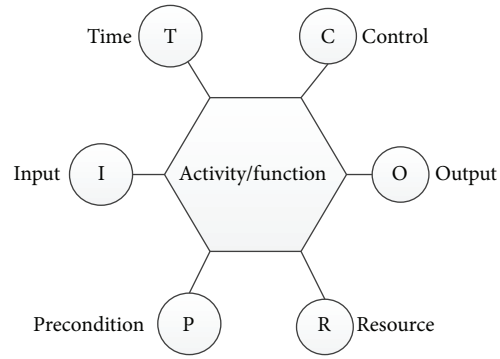


Figure 2: A hexagon representing a function [5].

running a function; Resource refers to the materials that are consumed to run a function or execution condition that is essential for a function; Time is the schedule or the time window that a function needs to follow; and Control is used to supervise and constrain a function, and it can be a set of specific plans, procedures, or guidelines [5]. For convenience, the phrase Five Aspects mentioned in the following text refers specifically to Input, Precondition, Resource, Time, and Control.

The phenomenon that a function does not perform as fully as designed or expected is called function variability [5]. On one hand, due to the instability and variation of the external environment around the systems, the aspects of a function are likely to be affected by the environment, which hence makes the function deviate from its behavior desired by designers. On the other hand, a function may be connected with other function(s) in the same system; for example, the Output of the upstream function can be the Input of the downstream function(s). The variability may possibly spread from the upstream to the downstream function(s) in some forms, once it does exist in the upstream function; therefore, a function's variability may come from

the variability of its upstream function through interactions between the functions. To make it clearer, a pair of definitions is presented as follows.

Definition 1. Variability in a single function that refers to how the Output of a function is influenced by the variability in one or more of the Five Aspects of the same function.

Definition 2. Variability between functions refers to how the aspect(s) of a function (Function A) is influenced by the variability in the Output of its upstream function (Function B), when the Output of Function B is related to one of the Five Aspects of Function A.

Hollnagel defined the term "aggregation" as "functional upstream-downstream coupling" [5], which means that how the upstream Output can have effect on the Five Aspects of the downstream function. It is pointed out that the variability in the upstream Output would correspond to the downstream function, which indicates that the variability would not vary when being passed between the upstream and downstream functions. However, we tend to assume that variability exists not only in a single function but also in the spreading process between functions. For example, in the task "a file is sent from Computer A to Computer B;" the upstream function is "Computer A provides the file," and the downstream function is "Computer B receives the file." The following incident may occur: the file provided by Computer A is normal and correct whereas Computer B receives the file with virus attached due to some unknown attacks during the delivery through the information network. Herein, the Output of upstream function is corresponding to what it is expected to be, but when it becomes to the Input of downstream function, deviation arises which in some sense can be taken as variability. It is shown that the Output variability of upstream function is likely not to correspond to the downstream function, thus in this paper the "aggregation" is classified into two types: variability in a single function and variability between functions, to make safety analysis as complete as possible.

For each function, any of the Five Aspects may deviate from its expected situation, and in most of cases the Output may deviate from its expected states accordingly. However, since the Five Aspects play different roles in a certain function, the variability of them can contribute to the Output variability in different ways. The potential Output variability caused by the Five Aspects of the same function is explained in Table 1. For a specific function, the variability analyzed is likely to be some (not all) of the items shown in Table 1, which actually covers potential variability as complete as possible. With the aid of Table 1, analysts can determine the variability in accordance with the features of each function. It is noted that an aspect (e.g., Input) may have presence which can be accepted by the function and even taken as normal for some reasons even if it is provided improperly, despite the fact that actually it is abnormal from the view of a whole system. In this way, the variability of Output can be relatively predictable once the variability of the other aspects is determined.

Table 1: The Output variability from a single aspect.

Aspect	Variability of the aspect	Output variability
Input	Earlier	Earlier, normal, later, or omitted
	Later	Later or omitted
	Erroneous	Erroneous or omitted
	Imprecise	Imprecise or omitted
	Omitted	No output
Precondition	Earlier	Normal
	Later	Later or omitted
	Imprecise	Normal, erroneous, or omitted
	Omitted	Normal, omitted, or erroneous
	Erroneous	Omitted or erroneous
Resource	Imprecise	Imprecise, erroneous, or normal
	Later	later, imprecise, or erroneous
	Omitted	Omitted
	Erroneous	Erroneous
Time	Shorter duration	Insufficient or erroneous
	Longer duration	Exceeding or overflow
	Omitted	Omitted, erroneous, insufficient exceeding, or overflow
	Imprecise	Imprecise or erroneous
Control	Imprecise	Imprecise, normal, or erroneous
	Earlier	Imprecise or normal
	Later	Imprecise, erroneous, or normal
	Omitted	Erroneous, imprecise, or normal
	Erroneous	Erroneous

In a function, variability may exist more often in two or more aspects simultaneously rather than in a single aspect, since a function may be connected with two or more other functions. For example, the Input and Precondition of Function A are connected, respectively, with the Outputs of Functions B and C. Furthermore, the Five Aspects may have different influences on different types of function; hence, the aspect weighting higher should be paid more attention to if variability exists in more than one aspect. Essentially functions vary with the characteristics of their Outputs. In this paper, functions are classified into four categories: material handling, energy transfer, information/data processing, and state change, according to which the Output variability can be summarized in Table 2.

In case that the variability of two aspects of the same function works simultaneously, more attention should be paid to the variability of the aspect contributing more to the Output. The parameter q_i $(i = 1, 2, \ldots)$ is used to present the relative importance of aspects' variability, and it is determined according to the ratio of impact of the aspect's variability to that of Input's variability. The parameter value is always an integer mostly depending on features of the specific function, but q_i $(i = 1, 2, \ldots)$ is set as 1 in most cases. It is exemplified with the function "calculate aircraft's weight and

Table 2: The Output variability for different types of functions.

Types of function	Potential variability of Output
Material handling	Earlier, later, imprecise (quantitatively), erroneous (with an incorrect target), and no output
Energy transfer	Earlier, later, imprecise (insufficient quantitatively, or unstable), erroneous (quantitatively, or with an incorrect type of energy), and no output
Information/data processing	Earlier, later, imprecise (quantitatively), erroneous, no output
State change	Earlier, later, imprecise (incomplete change from a state into another), erroneous (with an incorrect target), and no output

gravity balance," which is an instance in the type "information/data processing": given the Resource "passenger's weight or calculation formula," the Output would deviate more or less when something wrong happens to the Resource, for example, an incorrect formula provided. Considering that the aspects affect function in different ways, we assume that the parameters can be used to describe variability. Given that the variability of Precondition is represented with Δf_1, the variability of Input with Δf_2 and the contributing weights of the Precondition and the Input are represented with q_1 and q_2, respectively, the variability of the Output is described as in the following equation:

$$\Delta F = q_1 * \Delta f_1 + q_2 * \Delta f_2. \quad (1)$$

Output of the upstream function can be Input, Time, Control, Resource, or Precondition of the downstream function. When analyzing the variability between each pair of functions related to each other, the spreading from upstream to downstream can also be taken as a certain function. Based on the possible variability of upstream Output (for which the possible variability can be determined as Table 2) as well as interactions between the upstream and downstream functions, the possible variability of the downstream aspect can be analyzed accordingly, which is shown in Table 3.

3.2. System Modeling with Model Checking. Given that the rules of variability and its spreading from upstream to downstream function have been established, the issue "whether the system meets the safety requirements and does not violate the safety constraints" could be interpreted into a model describing "whether the state transitions within the system satisfy the temporal logic formulae derived from the rules of variability propagation." Based on the model, the system behaviors can be simulated with model checking [18] by following the three steps below.

Step 1 (describe state transitions within system). First, the definition about state transitions within system is given below.

Definition 3. In terms of FRAM, function variability roughly indicates two states of function: standard and deviate. The state of a function is taken equivalent to that of its Output, as the function variability is basically reflected with its Output.

To depict the changes in system behaviors, State Transition Diagram (STD) is used, where states are represented with circles, and conditions for state transition are represented with links. By means of model checking, the STDs are beneficial for describing the process that the state change in a function influences that in the other functions related to it, so as to model the functional behaviors of the whole system. The model structure of functional state transition is shown in Figure 3.

Table 3: The variability of the downstream aspect in terms of upstream Output.

Aspect of downstream function	Variability of upstream Output	Variability of downstream aspect
Input	Earlier	Earlier, later, or normal
	Later	Later or omitted
	Omitted	Omitted
	Imprecise	Imprecise, normal, or omitted
	Erroneous	Erroneous
Precondition	Earlier	Earlier, later, normal, or omitted
	Later	Later or omitted
	Omitted	Omitted
	Imprecise	Erroneous or omitted
	Erroneous	Erroneous or omitted
	Insufficient	Insufficient, erroneous, or omitted
Resource	Earlier	Normal
	Later	Omitted or insufficient
	Omitted	Omitted
	Insufficient	Omitted or insufficient
	Erroneous	Erroneous
Time	Earlier	Earlier
	Later	Later
	Higher value	Lasting longer
	Lower value	Lasting shorter
	Erroneous	Erroneous
	Omitted	No requirement about time
Control	Earlier	Earlier or normal
	Later	Later or omission
	Imprecise	Erroneous or imprecise
	Erroneous	Erroneous
	Omitted	Omitted

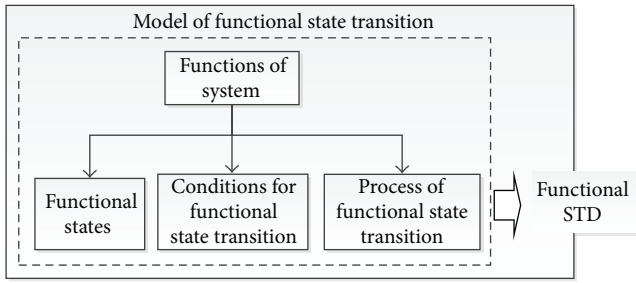


Figure 3: The model structure of a functional state transition.

There is assumed to be six functional states: standard, no output, erroneous, earlier, later, and imprecise, all of which are represented with circles and labeled, respectively. Taking Function 7 "ATC confirmed business jet position" as an example (note: please see Section 4 for all the nine functions identified in the case), "output == 0" means Function 7 is in a standard state, whereas "output == 1" means it is in any of deviate states, that is, any of the latter five states above. Then based on the function variability, the conditions for the transition from deterministic state of a function's Input to the different potential Output are analyzed, and Precondition, Resource, Control, and Time are represented with specific arrays or parameters, as well as the transition from upstream to downstream function with logic formulae or IF statements. To continue the example of Function 7, variability arises in the Output when there is variability in the Resource, where the conditions for state transition can be expressed with "output[j] = res[j]" and the j right means Function 7.

Based on the functional states as well as spreading between functions that have been clarified, the system behaviors can be modeled and simulated by means of model checking. The variability of Input and Output is defined as parameters, respectively, through which variability can be thus characterized. For example, the variability of Output is described with (2) if there is any variability in the sense of timing or precision:

$$v = \begin{cases} i, & \left(\dfrac{\text{Time earlier}}{\text{Value larger}}\right), \\ 0, & (\text{No variation}), \\ -j, & \left(\dfrac{\text{Time later}}{\text{Value smaller}}\right). \end{cases} \tag{2}$$

The numerical level of i and j indicates the degree of variability. The larger the value, the more considerable the variability, so the performance of functions and thus the system states can be described by these parameters. As a preliminary principle of variability comparison proposed in this paper, it is assumed that $i = 1$ when a function varies and $i = 0$ when it does not.

According to the rules of variability spreading between the aspects in a function as well as between functions that are determined in FRAM analysis, the conditions for functional state transition can be defined and then the system model established. For example, assuming that the variability of Input is described as i and the variability of Precondition is described as j, and the weight of Input and of Precondition is represented as q_1 and q_2, respectively, the Output is explained as follows:

$$\text{output} = i * q_1 + j * q_2. \tag{3}$$

The conditions for state transitions are expressed as follows:

$$\{\text{output} = i * q_1 + j * q_2\}, \quad \text{if } (\text{input} == i \,\&\&\, \text{precondition} == j \,\&\&\, \text{input\_weight} == q_1 \,\&\&\, \text{precondition\_weight} == q_2). \tag{4}$$

Step 2 (determine safety constraints). Safety constraints are typically developed in terms of the unexpected states or events and mostly converted into safety requirements during the system design. Safety constraints can be taken as the criteria in position, during model checking simulation, for justifying whether the system is safe or not if going along the given path of functional state transition and for further identifying the path(s) leading to accidents after checking all the potential paths. Herein, safety constraints are categorized into two types: (1) critical functional constraints, which means that some deviation of a certain function causes an accident directly and (2) combining functional constraints, which means an accident is caused by the mixture of deviations of more than one function. The second type is more typical in most cases of sociotechnical systems, since in terms of the rationale of resilience engineering, sociotechnical systems are likely to be self-adjustable and keep working normally even though some of their functions deviate from the desired performance, because the deviations can be dampened or mitigated through the interactions among functions.

In order to develop safety constraints, the potential hazards need to be identified beforehand, which is even treated as the indispensable step within system modeling. The hazards, defined here as those that may cause deaths, injuries, damage to equipment, or environmental pollution, can be identified by means of Preliminary Hazard Analysis (PHA) [32] or Hazard and Operability Analysis (HAZOP) [33]. After being determined, safety constraints are interpreted to the descriptions in the text of Linear Temporal Logic (LTL) which is a widely used text form in model checking. The structure Kripke such as $M \mid / = \neg f$ is applied to the first type of constraints, wherein M describes the system state, f is a formula consisting of logic connectors like &&, ‖, !, and so

forth, and ¬ means a negative logic. For the second type of constraints, the form $p = f_1 \wedge f_2$ is adopted, wherein f_1 describes the first potential deviation, and f_2 describes the second one, and so on. For example, equation "output[4] + ouput[9] == 0" to be involved in case study indicates that both outputs of Functions 4 and 9 are zero, which means accident may occur if either Function 4 or 9 is deviate. Furthermore, safety constraints can be described in mathematical ways and algorithms in model checking, provided their values assigned in terms of corresponding parameters.

Step 3 (simulate and analyze results). Simulation is conducted based on the established model, and the simulation result can be explained in accordance with the following principles.

(a) It is firstly checked whether the model is established correctly, if the simulation result does provide counter example(s) which means the safety constraints are violated under certain conditions. Provided that it is a correct model, the counter example(s) evidently indicates that the combinations of functional states do not satisfy the system safety requirements. After the model has been justified to be established correctly and rationally, the practical significance underlying the counter examples needs to be analyzed.

(b) Considering the practical significance of the counter examples given in the result of simulation, measures should be developed to eliminate hazards or to dampen their evolution in the system. The parameters in the model can be reset according to the measures being developed, and even the model structure can be updated, to check the benefit and effectiveness of the measures.

4. Case Study

Taking an air accident as a case, the approach proposed in this paper is used to analyze why and how the accident may occur, and the comparison is made between the conclusions drawn with this approach and those from the official investigation report, to illustrate the merit of this approach.

(1) The Accident Process. On 8 October 2001, an aircraft crashed at the Linate Airport in Milan, Italy. Scandinavian Airlines Flight 686 carrying 110 people collided with a Cessna Citation CJ2 business jet carrying four people. All 114 people on both aircrafts were killed, as well as four people on the ground. The disaster is the deadliest air disaster in Italian aviation history [34]. On the day of the accident, the visibility at the airport is only 50–100 meters due to heavy fog. Flight 686 was allowed to taxi to the runway R6 on 07:54, while the business jet was allowed to taxi to the runway R5 on 8:05. When having parked on taxiway S4, the crew on the business jet reported to the air traffic controllers. It was unnoticed that business jet accidentally broke into runway R6 along the indicator lights and ground markings. Flight 686 was allowed to take off on 8:09, but it crashed with the business jet stopping on the same runway when it took off.

Table 4: Descriptions for Activity 7 based on FRAM.

Aspect	Details
Input	Order to taxi to the main runway
Output	Enter R6 runway
Precondition	Available runway(s)
	Fulfill the ATC's taxi instructions
	Complete taxi checklists
Resource	Flight crew who is familiar with the airport
	Signal lights and ground markings on the airport
Control	Alarm system for preventing airplanes from breaking into the wrong runway
Time	The whole process

(2) Function Modules. The plane had to take off with the aid of the instructions given by the Air Traffic Control (ATC) due to the poor visibility at the airport. The processes can be broken down specifically into the following 11 activities/events: (1) the air traffic controller guided flight 686 to R6 runway; (2) Flight 686 taxied to R6 runway; (3) ATC guided business jet to R5 runway; (4) business jet reported to ATC in the S4 taxiway position; (5) ATC confirmed business jet position; (6) ATC guided business jet continue to slide to the main runway; (7) Business jet turned left into the R6 runway along signal lights and ground markings; (8) ATC guided the 686 flight takeoff; (9) Flight 686 took off; (10) alarm system prevented the aircraft break into the runway; (11) the ground radar monitored the positions of the aircrafts and vehicles on the airport. The latter two are involved with safety control and monitoring system. Based on FRAM, the eleven activities/events can be treated as function modules of the accident, with specific descriptions for each of them. The details for the instance of Activity 7 are shown in Table 4, and the standard conduction of the activities/events and interactions among the flight 686, the business jet, the ATC, and the monitoring system are shown in Figure 4.

(3) Potential Variability. According to the rules given in Table 1, the potential conduction variability of each activity/event is analyzed based on the features of the activity/event, and the potential variability of Activity 7 is shown in Table 5 as an example.

(4) Modeling and Simulation with Model Checking

(a) Modeling of System Behaviors. The tool Process Analysis Toolkit (PAT), which is a self-contained framework to support composing, simulating, and analyzing dynamic systems [35], was used here to model and verify the double-plane system (including the planes, the crew aboard, the ATC, and the airport situations) involved in this air accident. For simplicity, it is assumed that the upstream Outputs are consistent with the relevant downstream aspects like Inputs or Preconditions, and the effects of only the Input and

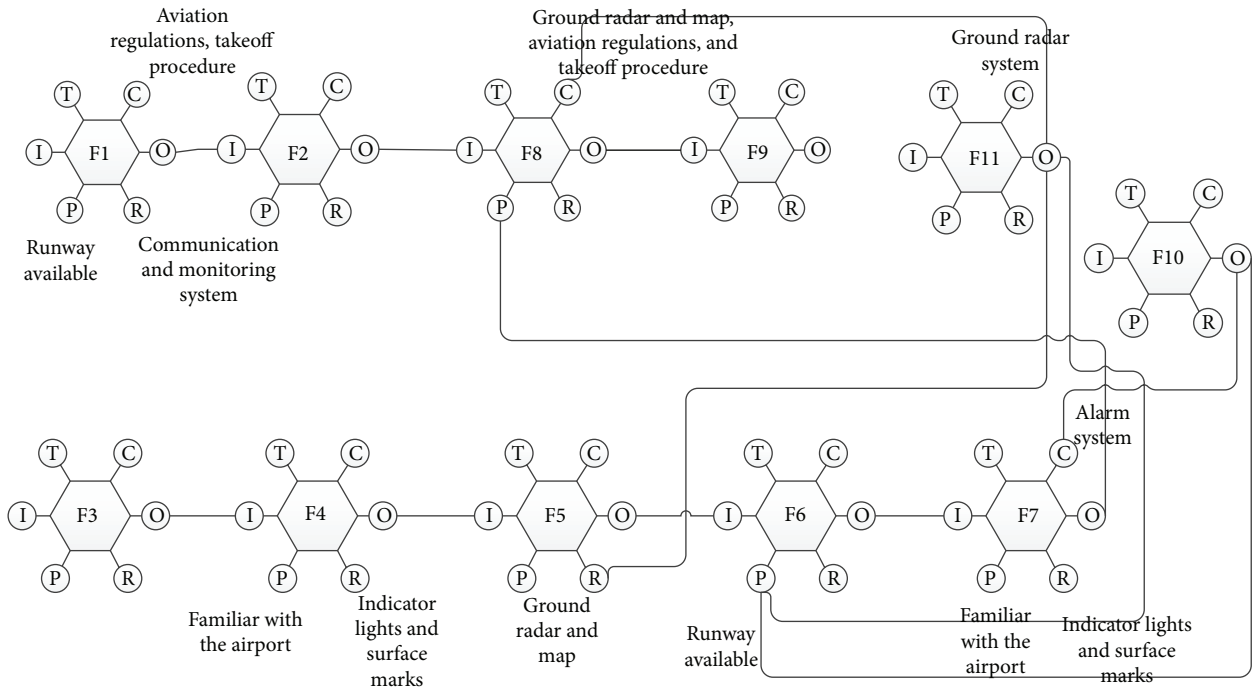


Figure 4: The instantiation of the functional interactions.

Table 5: Potential variability of Activity 7.

Aspects	Variability	Influence on Output
Input	The order by ATC is imprecise	The plane takes the wrong runway
	The order is earlier	The plane takes the wrong runway earlier
	The order is later	The plane takes the wrong runway later
Precondition	The runway is not available	The plane shares the runway with other aircrafts
	The airplane does not fulfill ATC's taxi instructions	The plane takes the wrong runway
Resource	The crew members are not familiar with the airport situations	The plane takes the wrong runway
	The signal lights and the ground markings on airport are misleading	The plane takes the wrong runway
Control	Alarm system does not work properly	The plane is not prevented from breaking into the wrong runway
Time	Too short	The plane takes the wrong runway earlier
	Too long	The plane takes the wrong runway later

Resource on the Output are considered for each function. For the purpose of coding with model checking, the system is described with functions of the two planes, respectively (as shown in Table 6), taking into account the descriptions for the activities identified earlier.

The value for the six aspects of each function is defined as 0 when they are normal and as 1 when there is variability existing. For example, input[1] = 0 indicates that Function 1 has normal Input, and input[1] = 1 means a variant Input to Function 1 from its upstream functions. The similar case applies to res[1] = 0 or output[1] = 0. For the four functions of the flight 686, the upstream Output is equivalent to the downstream Input or some other aspects, which is expressed as follows:

$$\text{input}[i+1] = \text{output}[i]. \quad (5)$$

For the functions performed by ATC, for example, Functions 1, 3, 5, 7, and 9, it is the Resources that have a great influence on the Output. The Output is normal if the Resource and Input are both normal, but there is variability in the Output with the variability in the Resource. The logic correlation is expressed as follows:

$$\text{output}[i] = 1 * \text{res}[i] + 0 * \text{input}[i], \quad \text{if } i\%2 == 1. \quad (6)$$

For Functions 2 and 4, it is the Inputs that have a greater influence on the Output. The Output is incorrect if the Input is wrong; otherwise, the Output is correct if the Input is correct. The logic is expressed as follows:

$$\text{output}[i] = 0 * \text{res}[i] + 1 * \text{input}[i], \quad \text{if } i\%2 == 0. \quad (7)$$

Table 6: The functions of the flight 686 and the business jet.

Airplane	Number	Function
Flight 686	Function 1	The air traffic controller guided flight 686 to R6 runway
	Function 2	Flight 686 taxied to R6 runway
	Function 3	ATC guided the 686 flight take off
	Function 4	Flight 686 took off
Business jet	Function 5	ATC guided business jet to R5 runway
	Function 6	Business jet reported to ATC in the S4 taxiway position
	Function 7	ATC confirmed business jet position
	Function 8	ATC guided business jet continue to slide to the main runway
	Function 9	Business jet turned left into the R6 runway

Similarly, for the six functions of the business jet, the upstream Output is taken equivalent to the downstream Input, which is expressed as follows:

$$\text{input}\left[j+1\right] = \text{output}\left[j\right]. \tag{8}$$

Finally, considering the functions conducted in parallel by the two airplanes, the system state is expressed as follows:

$$\text{System} = \text{flight}() \,|||\, \text{jet}(). \tag{9}$$

(b) Safety Constraints Description. In this case, the accident happened due to the fact that a runway is occupied by the two planes simultaneously, so the safety constraint is described as that no more than one airplane is permitted to be on a runway at a time. According to the parameters' meaning given earlier, output[4] = 0 means the flight 686 is on R6 runway and output[9] = 0 means the business jet is on R5 runway, when there is no function variability, while output[9] = 1 means the business shares R6 runway with the flight 686. Apparently R6 runway cannot be used by two planes at the same time according to air traffic rules; that is, the constraint is interpreted as both the statements output[4] = 0 and output[9] = 0 are true, which is expressed as (10) to indicate that the flight and the business jet take different runways

$$\text{output}\,[4] + \text{output}\,[9] = 0. \tag{10}$$

(c) Simulation and Result Analysis. Given the initial values of parameters preset randomly, how the parameters change is observed for all the potential states by means of model checking. The simulation result shows that there are two scenarios in which the runway may be occupied by the two planes at the same time. The first scenario is exactly the accident process that has truly occurred (and been reported by ANSV) while the second one is a potential accident process which may occur, despite the fact that it has never occurred so far.

Scenario 1. Due to ATC's unclear instructions as well as the misleading ground markings and flight indicator, the business jet entered the wrong taxi way. Moreover, when it even arrived at S4 position, ATC did not correct the direction of the flight because of the map without being updated. Due to the failures of ground radar and alarm device, the business jet was not prevented from breaking into R6 runway. Besides, in spite of the failure to scan the planes' positions with the ground radar, ATC guided flight 686 to take off, which eventually led to the accident. This scenario was explained in ANSV report [34]: the accident was caused basically by the combination of inaccurate order of ATC, the wrong guide lights and ground markings, unfamiliarity with the airport, the map without updated, and the failure of ground radar and alarm system.

Scenario 2. Even if the ATC gives right instructions in some time, the business jet might enter the wrong taxiway due to the error of the markings and flight indicators. When the business jet arrives at S4 position, ATC may not correct the flight direction because of the map without being updated. Furthermore, since the alarm device and ground radar fail to work, the two planes might possibly enter the same runway. This is a potential scenario which could also lead to the accident, and the slight difference from Scenario 1 is whether Function 5 "ATC guided business jet to R5 runway" has normal Output or not. In Scenario 2, all the functions of the planes and ATC are normal before the business jet taxies, but the business jet parks at the wrong location because the crew is unfamiliar with the airport, in addition to the misleading guide lights and ground markings at the airport. When the crew reports to ATC their exact location, ATC does not perceive the deviation of its position due to the fact that neither the report map has been updated nor the ground radar works, so ATC guides the plane to continue taxiing to the main runway R6 instead of the R5 runway that ATC wants. It indicates that the ATC function's Output may deviate from what is desired due to the erroneous Resource, even if the conduction was in accordance with the relevant provisions. Scenario 2 is depicted as in Figure 5 with the abnormal paths being marked as red, to demonstrate how the deviations spread among functions.

5. Conclusions and Future Work

This paper extends FRAM by integrating it with model checking to effectively explore hazard evolution. The extended FRAM refines the understanding of interactions among functions of sociotechnical systems by redefining and categorizing the couplings. It also proposes a process, by means of auto search with a computer tool, for identifying functional deviations as well as their propagation among upstream and downstream functions. The approach is a progress in efforts for the exhaustiveness of heuristic analysis, which in the past depended excessively on the knowledge and experience of analysts, and was hard to traverse all the possible conditions due to limitations of human's recognition. While the exhaustive search across all the scenarios for the one(s) that may lead to accident is conditionally conducted, that is, the search is based on the functions and their variability pre-determined heuristically, more potential couplings among functions can

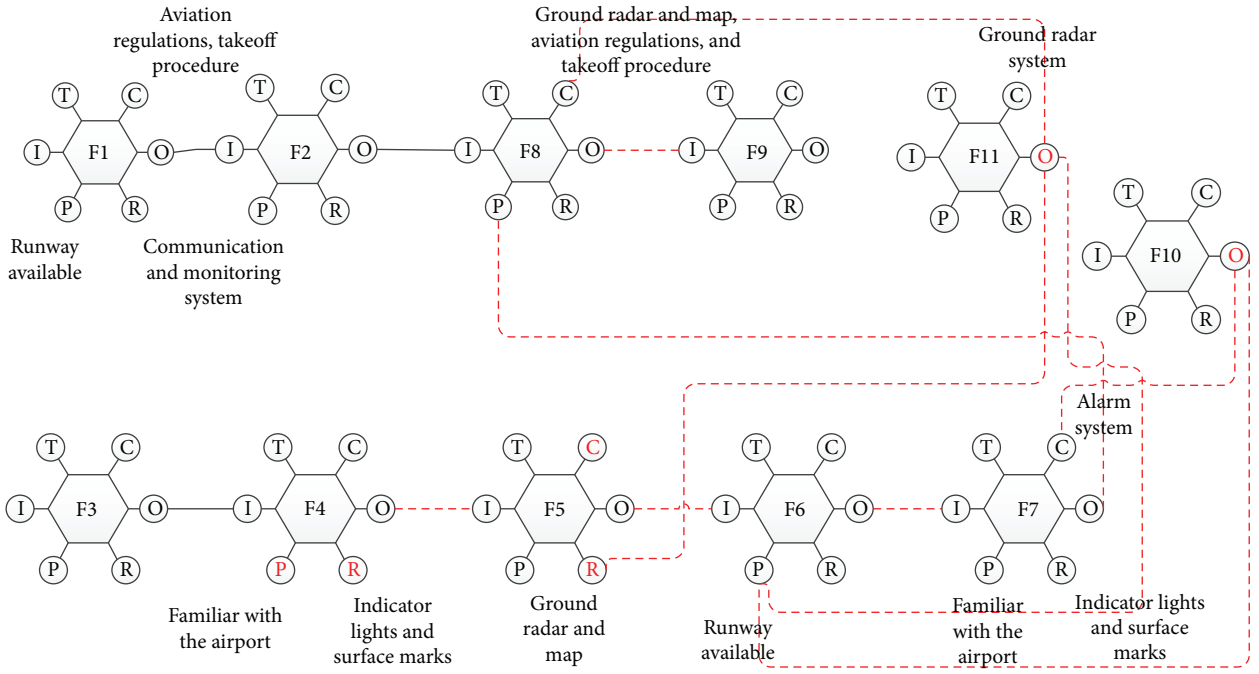


Figure 5: Scenario 2 for the accident.

be identified by relatively rigorous derivation rather than by subjective analysis.

Taking a typical air accident as the case, the plausibility of the approach is illustrated. Additionally, in point of the accident scenarios as such, apparently it is highlighted that an accident was not caused by a simple combination of multiple contributing factors, but by performance variability and its nonlinear propagation among functions. Hence, in order to prevent accidents, it is far more significant to coordinate the functions in the system than put the emphasis on a certain single aspect. For example, to continue the case of the air accident, it is definitely true that the mentoring system should be open, that the map should be updated, and that the indicating lights and ground markings should work regularly. But more efforts need to be concentrated on how to make all of these aspects cooperate and function well simultaneously and continuously, without any aggravation even though an unexpected disturbance happens somewhere and sometime.

It is noted that in this paper we focus more on whether, rather than how, functions vary (in the part of quantitative analysis), as actually the terms of variability in different ways have essential influence on the rationality of a functional model, for the reason that different variability terms of an upstream function may impact its downstream function in different ways. Accordingly, the spreading rules of variability need to be elaborated further based on the specific term of functional variability. In the future work, the approach of functional behavior modeling will be improved and specified based on FRAM, with more deliberation of variability of functions, as well as further development of the rules that describes interactions among functions in the sense of abstraction. Besides, there are some assumptions made to simplify the analysis in the case study; for example, the Time aspect and details of the airplanes' behaviors have not been considered. For the lack of practical illustration of some aspects (e.g., Time and Control), there are not sufficient analysis and discussion with regard to them in functional modeling. Thus, efforts will also be put into case study; that is, the case will be analyzed in detail, with more consideration supplemented involving the aspect of Time, airplanes' states like position and direction, and so on.

Conflict of Interests

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

Acknowledgments

This research was supported by Grants from the National Natural Science Foundation of China (no. NSFC-61403009), from a project of Ministry of Industry and Information Technology of China (no. JSZL2014601B004), and from the Major State Basic Research Development Program of China (973 Program) (no. 2014CB744904). This support is gratefully acknowledged.

References

[1] F. I. Khan and S. A. Abbasi, "Studies on the probabilities and likely impacts of chains of accident (domino effect) in a fertilizer industry," *Process Safety Progress*, vol. 19, no. 1, pp. 40–56, 2000.

[2] C. D. Lorenz and R. M. Ziff, "Precise determination of the critical percolation threshold for the three-dimensional "Swiss cheese" model using a growth algorithm," *The Journal of Chemical Physics*, vol. 114, no. 8, article 3659, 2001.

[3] E. Hollnage, "Barriers and accident prevention," *Ergonomics*, vol. 50, no. 6, pp. 961–962, 2007.

[4] N. G. Leveson, *Engineering a Safer World: Systems Thinking Applied to Safety*, The MIT Press, London, UK, 2011.

[5] E. Hollnagel, *FRAM: The Functional Resonance Analysis Method, Modelling Complex Socio-Technical Systems*, MPG, London, UK, 2012.

[6] M. C. Browne, E. M. Clarke, and O. Grumberg, "Reasoning about networks with many identical finite state processes," *Information and Computation*, vol. 81, no. 1, pp. 13–31, 1989.

[7] E. Hollnagel, *Barriers and Accident Prevention*, Ashgate, Farnham, UK, 2004.

[8] E. Hollnagel, "Critical information infrastructures: should models represent structures or functions?" in *Computer Safety, Reliability, and Security*, vol. 5219 of *Lecture Notes in Computer Science*, pp. 1–4, Springer, Berlin, Germany, 2008.

[9] I. A. Herrera and R. Woltjer, "Comparing a multi-linear (STEP) and systemic (FRAM) method for accident analysis," *Reliability Engineering & System Safety*, vol. 95, no. 12, pp. 1269–1275, 2010.

[10] J. Hovden, E. Albrechtsen, and I. A. Herrera, "Is there a need for new theories, models and approaches to occupational accident prevention?" *Safety Science*, vol. 48, no. 8, pp. 950–956, 2010.

[11] R. Woltjer, *A Systemic Functional Resonance Analysis of the Alaska Airlines Flight 261 Accident*, Human Fators and Economic Aspects on Safety, 2006.

[12] H. Sybert, H. C. Stroeve Mariken, and A. P. Henk, "Studying hazards for resilience modelling," in *Proceedings of the 1st SESAR Innovation Days*, pp. 1–8, Toulouse, France, November-December 2011.

[13] P. V. R. de Carvalho, "The use of Functional Resonance Analysis Method (FRAM) in a mid-air collision to understand some characteristics of the air traffic management system resilience," *Reliability Engineering & System Safety*, vol. 96, no. 11, pp. 1482–1498, 2011.

[14] E. Hollnagel, S. Pruchnicki, R. Woltjer, and S. Etcher, "Analysis of Comair flight 5191 with the functional resonance accident model," in *Proceedings of the 8th International Symposium of the Australian Aviation Psychology Association*, Sydney, Australia, 2008.

[15] T. Sawaragi, Y. Horiguchi, and A. Hina, "Safety analysis of systemic accidents triggered by performance deviation," in *Proceedings of the SICE-ICASE International Joint Conference*, pp. 1778–1781, Busan, South Korea, October 2006.

[16] F. Belmonte, W. Schön, L. Heurley, and R. Capel, "Interdisciplinary safety analysis of complex socio-technological systems based on the functional resonance accident model: an application to railway traffic supervision," *Reliability Engineering & System Safety*, vol. 96, no. 2, pp. 237–249, 2011.

[17] L. Macchi, P. Oedewald, M. H. Rø Eitrheim, and C. Axelsson, "Understanding maintenance activities in a macrocognitive work system," in *Proceedings of the 30th European Conference on Cognitive Ergonomics (ECCE '12)*, pp. 52–57, Edinburgh, Scotland, August 2012.

[18] J. Bentahar, M. El-Menshawy, H. Qu, and R. Dssouli, "Communicative commitments: model checking and complexity analysis," *Knowledge-Based Systems*, vol. 35, pp. 21–34, 2012.

[19] F. Wang, "Efficient model-checking of dense-time systems with time-convexity analysis," *Theoretical Computer Science*, vol. 467, pp. 89–108, 2013.

[20] V. Chapurlat, "UPSL-SE: a model verification framework for systems engineering," *Computers in Industry*, vol. 64, no. 5, pp. 581–597, 2013.

[21] C. Tian, Z. Duan, and N. Zhang, "An efficient approach for abstraction-refinement in model checking," *Theoretical Computer Science*, vol. 461, pp. 76–85, 2012.

[22] R. Gómez, "Model-checking timed automata with deadlines with Uppaal," *Formal Aspects of Computing*, vol. 25, no. 2, pp. 289–318, 2013.

[23] E. Onem, A. B. Gürdağ, and M. U. Çağlayan, "Formal security analysis of ariadne secure routing protocol using model checking," *International Journal of Ad Hoc and Ubiquitous Computing*, vol. 9, no. 1, pp. 12–24, 2012.

[24] M. H. Ter Beek, A. Fantechi, S. Gnesi, and F. Mazzanti, "A state/event-based model-checking approach for the analysis of abstract system properties," *Science of Computer Programming*, vol. 76, no. 2, pp. 119–135, 2011.

[25] S. Konur, C. Dixon, and M. Fisher, "Analysing robot swarm behaviour via probabilistic model checking," *Robotics and Autonomous Systems*, vol. 60, no. 2, pp. 199–213, 2012.

[26] M. Gribaudo, A. Horváth, A. Bobbio, E. Tronci, E. Ciancamerla, and M. Minichino, "Fluid petri nets and hybrid model-checking: a comparative case study," *Reliability Engineering and System Safety*, vol. 81, no. 3, pp. 239–257, 2003.

[27] Y. Gao, M. Xu, N. Zhan, and L. Zhang, "Model checking conditional CSL for continuous-time Markov chains," *Information Processing Letters*, vol. 113, no. 1-2, pp. 44–50, 2013.

[28] J. Rushby, "Using model checking to help discover mode confusions and other automation surprises," *Reliability Engineering & System Safety*, vol. 75, no. 2, pp. 167–177, 2002.

[29] S. Zhang, J. Teizer, J.-K. Lee, C. M. Eastman, and M. Venugopal, "Building Information Modeling (BIM) and Safety: automatic safety checking of construction models and schedules," *Automation in Construction*, vol. 29, pp. 183–195, 2013.

[30] J. Lahtinen, J. Valkonen, K. Björkman, J. Frits, I. Niemelä, and K. Heljanko, "Model checking of safety-critical software in the nuclear engineering domain," *Reliability Engineering & System Safety*, vol. 105, pp. 104–113, 2012.

[31] Department of Defense Standard Practice for System Safety, MIL-STD-882D, 1993.

[32] C. Zhao, M. Bhushan, and V. Venkatasubramanian, "Phasuite: an automated HAZOP analysis tool for chemical processes: part I: knowledge engineering framework," *Process Safety and Environmental Protection*, vol. 83, no. 6, pp. 509–532, 2005.

[33] S. Dowlatshahi, "The role of product safety and liability in concurrent engineering," *Computers & Industrial Engineering*, vol. 41, no. 2, pp. 187–209, 2001.

[34] Agenzia Nazionale per la Sicurezza del Volo (ANSV), *Accident Involved Aircraft Boeing MD-87, Registration SE-DMA and Cessna 525-A, Registration D-IEVX*, Milano Linate Airport, 2001.

[35] J. Zhang, Y. Liu, J. Sun, J. S. Dong, and J. Sun, "Model checking software architecture design," in *Proceedings of the IEEE 14th International Symposium on High-Assurance Systems Engineering (HASE '12)*, pp. 193–200, Omaha, Neb, USA, October 2012.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 250461, 11 pages
http://dx.doi.org/10.1155/2015/250461

Research Article

An Automatic Traffic Sign Detection and Recognition System Based on Colour Segmentation, Shape Matching, and SVM

Safat B. Wali, Mahammad A. Hannan, Aini Hussain, and Salina A. Samad

Department of Electrical, Electronic & Systems Engineering, Universiti Kebangsaan Malaysia, Jalan Reko, 43600 Bangi, Selangor, Malaysia

Correspondence should be addressed to Safat B. Wali; safat.2804@gmail.com

Received 12 June 2015; Revised 2 September 2015; Accepted 25 October 2015

Academic Editor: Pan Liu

The main objective of this study is to develop an efficient TSDR system which contains an enriched dataset of Malaysian traffic signs. The developed technique is invariant in variable lighting, rotation, translation, and viewing angle and has a low computational time with low false positive rate. The development of the system has three working stages: image preprocessing, detection, and recognition. The system demonstration using a RGB colour segmentation and shape matching followed by support vector machine (SVM) classifier led to promising results with respect to the accuracy of 95.71%, false positive rate (0.9%), and processing time (0.43 s). The area under the receiver operating characteristic (ROC) curves was introduced to statistically evaluate the recognition performance. The accuracy of the developed system is relatively high and the computational time is relatively low which will be helpful for classifying traffic signs especially on high ways around Malaysia. The low false positive rate will increase the system stability and reliability on real-time application.

1. Introduction

In order to solve the concerns over road and transportation safety, automatic traffic sign detection and recognition (TSDR) system has been introduced. An automatic TSDR system can detect and recognise traffic signs from and within images captured by cameras or imaging sensors [1]. In adverse traffic conditions, the driver may not notice traffic signs, which may cause accidents. In such scenarios, the TSDR system comes into action. The main objective of the research on TSDR is to improve the robustness and efficiency of the TSDR system. To develop an automatic TSDR system is a tedious job given the continuous changes in the environment and lighting conditions. Among the other issues that also need to be addressed are partial obscuring, multiple traffic signs appearing at a single time, and blurring and fading of traffic signs, which can also create problem for the detection purpose. For applying the TSDR system in real-time environment, a fast algorithm is needed. As well as dealing with these issues, a recognition system should also avoid erroneous recognition of nonsigns.

The aim of this research is to develop an efficient TSDR system which can detect and classify traffic signs into different classes in real-time environment. For detecting the red traffic signs, a combination of colour and shape based algorithm is presented which will up the procedure of the detection stage and for recognition SVMs with bagged kernels are introduced.

This paper is organized as follows: Section 2 presents the related works in the field of development of the TSDR system. In Section 3, the overall methodology is discussed. The experimental results and discussions are summarized in Section 4. In Section 5, the conclusion and some suggestions are made for future improvement on the field of automatic traffic sign detection and recognition.

2. Related Work

According to [2], the first work on automated traffic sign detection was reported in Japan in 1984. This attempt was followed by several methods introduced by different researchers to develop an efficient TSDR system and minimize all the

(a) (b) (c)

Figure 1: Model for sample collections (a) used car with a camera placed on the left side of the dashboard, (b) camera setup including a laptop, and (c) on road camera range and sign detection.

issues stated above. An efficient TSDR system can be divided into several stages: preprocessing, detection, tracking, and recognition. In the preprocessing stage the visual appearance of images has been enhanced. Different colour and shape based approaches are used to minimize the effect of environment on the test images [3–6]. The goal of traffic sign detection is to identify the region of interest (ROI) in which a traffic sign is supposed to be found and verify the sign after a large-scale search for candidates within an image [7]. Different colour and shape based approaches are used by the researchers to detect the ROI. The popular colour based detection methods are HSI/HSV Transformation [8, 9], Region Growing [10], Colour Indexing [11], and YCbCr colour space transform [12]. As the colour information can be unreliable due to illumination and weather change, shape based algorithm is introduced. The popular shape based approaches are Hough Transformation [13–15], Similarity Detection [16], Distance Transform Matching [17], and Edges with Haar-like features [18, 19].

The tracking stage is necessary to ensure real-time recognition. In addition, the information provided by the images of the traffic signs will help verify the correct identification and thus detect and follow the object [20]. The most common tracker adapted is the Kalman filter [18, 21, 22].

Several methods have been used by the researchers for recognizing traffic sign. Ohara et al. [23] and Torresen et al. [24] used the Template Matching technique, which is a fast and straightforward method. Genetic Algorithm is used by Aoyagi and Asakura [25] and de la Eccalera et al. [26] which is said to be unaffected by the illumination problem. The main advantage of the AdaBosst is its simplicity, feature selection for large dataset, and generalization [27]. Li et al. [28] used Adaboost learning containing five classical Haar wavelets and four HoG (Histogram of Oriented Gradient) features. Greenhalgh and Mirmehdi [29, 30] showed a comparison between SVM, MLP, HOG-based classifiers, and Decision Trees and found that a Decision Tree has the highest accuracy rate and the lowest computational time. Its accuracy is approximately 94.2%, whereas the accuracy of the SVM is 87.8% and that of MLP is 89.2%. Neural Network is flexible, adaptive, and robust [31]. Hechri and Mtibaa [12] used a 3-layer MLP network whereas Sheng et al. [32] used a Probabilistic Neural Network for the recognition process. Support Vector Machine (SVM) is another popular method used by the researchers which is robust against illumination and rotation with a very high accuracy. Yang et al. [33] and García-Garrido et al. [34] used SVM with Gaussian Kernels for the recognition whereas Park and Kim [35] used an advanced SVM technique that improved the computational time and the accuracy rate for gray scale images.

For improving the recognition rate of the damaged or partially occluded sign, Soheilian et al. in [36] used template matching followed by a 3D reconstruction algorithm. The distortion-invariant fringe-adjusted joint transform correlation (FJTC) was used by Khan et al. in [37] and Principal Component Analysis (PCA) is used by Sebanja and Megherbi in [38] which have a very high accuracy rate. In [39], Prieto and Allen used a self-organizing map (SOM) for recognition whose main idea was to apply SOM at every level of RSs with a hit rate of 99%.

In our approach, for reducing the processing time RGB segmentation and shape matching based detection and SVM with bagged kernel are used for recognizing the red traffic signs. Grey-scale images are used to make our detection and recognition algorithm more robust to changes in illumination.

3. Methodology

3.1. Image Acquisition. The samples are collected from an inexpensive on board camera (Canon SX170 IS) which is connected to a laptop placed inside of a vehicle (Figure 1). The images were taken in different roads and highways in Malaysia under various weather conditions (Table 1) from 8:00 A.M. to 8:00 P.M. after every two seconds. The camera is placed in the left side of the dashboard so that it can capture the traffic sign of left side. The aim of this section is to create a database of traffic sign images under different variations.

3.2. Image Preprocessing. Image preprocessing is an important part of the TSDR system whose main idea is to remove low-frequency background noise, normalising the intensity of the individual particles images, removing reflections,

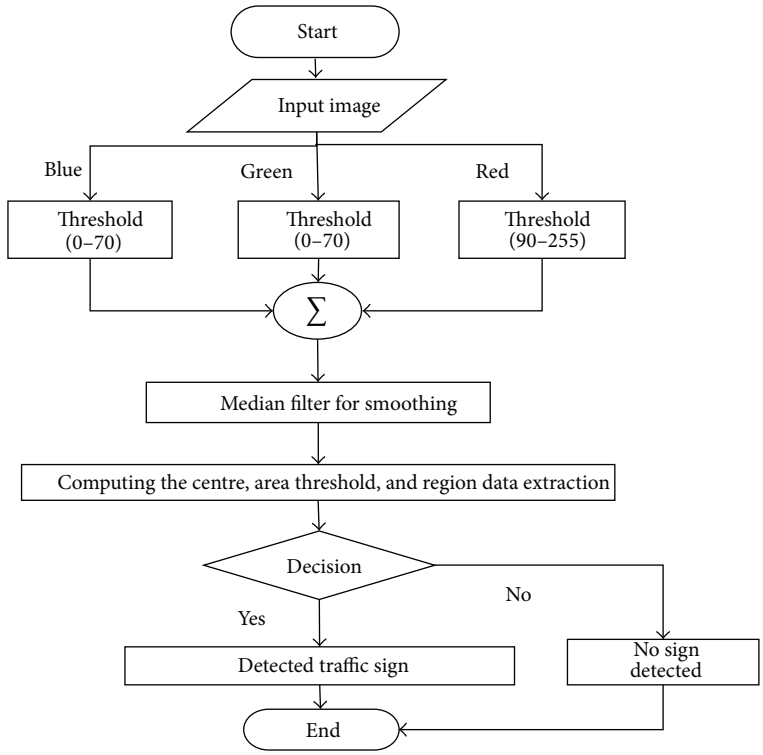


Figure 2: The overall block diagram of the detection system.

Table 1: Environmental condition for image acquisition.

Environment	Real time
Weather	Rainy, sunny, cloudy
Capturing time	8 am–8 pm
Camera specs	24 fps
Image size	3264 × 2448
Image background	Complex; not fixed
Total number of images taken	350
Total number of signs	123
Traffic sign sizes	Different
Traffic sign condition	Faded, blurred, damaged, occluded
Traffic sign type	Stop, do not enter, do not park

and masking portions of images. Below is a description of selected image preprocessing techniques. The input image is divided into channels R, G, and B separately. In the proposed approach, filters are applied on each channel threshold to select those regions of the image where the values of the pixels fall in the range of our target object. For example, for traffic signs with a red background (such as stop signs), the threshold for channel R is pixels with values in the range of 90–255 and for channels G and B the range is 0–70. The region of interest (ROI) is the logical sum of the three filtered channels of R, G, and B.

3.3. Shape Matching Based Detection. The idea is to use colour characteristic of the preferred object to accelerate the procedure without employing model-based classifiers which is a time consuming process [40–42]. After filtering and analysing the features of the detected object, the candidates of the traffic sign are selected based on shape matching. The flow chart of the system is shown in Figure 2.

3.3.1. Objects Features Analysing. One of the important steps is to eliminate noise from the image therefore to better deal with the ROI. Appropriate filters have an enormous effect on accuracy and speed of the procedure without deleting any useful information. In the proposed system, for image smoothing and filling up the smaller region to extract the region of interest, a median filter was used.

3.3.2. Shape Matching and Candidate Selection. As almost all traffic signs containing red colour are round or octagonal, the proposed method drew on these common shapes to detect hypothetical shapes which are close to traffic signs. Those regions with k in the range of 0.7–1.3 are accepted as candidates for traffic signs:

$$k = \frac{a}{\pi \times w^2}, \tag{1}$$

where a is the area of the region and w is the longest width.

3.3.3. Traffic Sign Detection. The area range for road signs determines the distance in which the system can detect the traffic sign. Outside of this range, objects with the same range of pixels value cannot be traffic signs. In this level,

crucial information such as centre, area, and longest width of each region is calculated. This information is used to decide whether or not each region is a traffic sign. The detected traffic sign blob images are then passed to SVM for recognition.

3.4. Support Vector Machine (SVM) Based Recognition. After the detection of traffic sign, the region of interest (ROI) is passed to the SVM for recognition. The SVM is one of the most successful kernel methods with a given labeled training dataset $\{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in R_d$ and $y_i \in \{-1, +1\}$. In the semisupervised SVM, the total image is clustered for building the bagged kernel. Then, the modification of the base kernel is done. A number of SVMs are trained separately using a bootstrap algorithm and are then aggregated via a suitable combination technique. A bagged kernel is a kernel function encoding the similarity between unlabeled samples [43]. For training sample k and given dataset PQ, the bootstrapping is built k replicate training datasets $\{PQ_k^b \mid k = 1, 2, \ldots, k\}$ by random resampling but replacing the values of given dataset PQ repeatedly. For a dataset S, kernel methods calculate the comparison between training samples $S = \{x_i\}_{i=1}^{n}$, using pair-wise inner products between mapped samples. Thus the final kernel matrix is $K_{ij} = K(x_i, y) = (\varphi(x_i), \phi(x_j))$. For the dataset formation, a set of 400 traffic sign images are generated from 100 traffic sign samples and for nonsign images a set of 1000 images are generated from 250 nonsign samples; those are collected randomly by a camera attached with a car in different times of the day and varying weather condition. It also includes partially occulted, slightly damaged, faded, and blurred signs for making the system more successful in real-time environment. All the candidates are scaled down to 25 × 25 pixels and in each step in 1.2 factors to smooth the progress of the features extraction process.

The proposed algorithm is discussed in the following steps.

(1) The computation of the base SVM kernel K_{SVM} is done.

(2) The k-means algorithm with various initializations is performed at t times but with the similar number of clusters k. The result is $p = 1, 2, \ldots, t$ cluster assessments $c_p(x_t)$ for each sample x_t.

(3) A bagged kernel K_{bag} is built based on the time fraction between x_i and x_j and assigned to the same cluster

$$K_{\text{bag}}\left(x_i, x_j\right) = \frac{1}{t}\sum_{p=1}^{t}\left[c_p\left(x_i\right) = c_p\left(x_j\right)\right], \tag{2}$$

where $[c_p(x_i) = c_p(x_j)]$ returns "1" if samples $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same cluster according to the pth realization of the clustering $cp(\cdot)$ and "–1" otherwise.

(4) Consider the sum or the product between the original and bagged kernels,

$$\begin{aligned} K\left(\mathbf{x}_i, \mathbf{x}_j\right) &\longleftarrow K_{\text{bag}}\left(\mathbf{x}_i, \mathbf{x}_j\right) + K_{\text{SVM}}\left(\mathbf{x}_i, \mathbf{x}_j\right), \\ K\left(\mathbf{x}_i, \mathbf{x}_j\right) &\longleftarrow K_{\text{bag}}\left(\mathbf{x}_i, \mathbf{x}_j\right) \cdot K_{\text{SVM}}\left(\mathbf{x}_i, \mathbf{x}_j\right). \end{aligned} \tag{3}$$

(5) With the resultant modified kernel $K(\mathbf{x}_i, \mathbf{x}_j)$, an SVM is trained. The flow chart of the overall SVM with bagged kernel is showed in Figure 3.

Different outcomes are obtained from step (2) because the k-means give various solutions in each process. In the semisupervised setting, a reduced dataset is used to compute the cluster centres. The test pixels can be assigned to the nearest cluster in each of the bagged runs to compute $K_{\text{bag}}(\mathbf{x}^*, \mathbf{x}_i)$. This way, the assignment can be done sequentially or can be parallelized, and only the cluster centres have to be maintained. Intensity correction and histogram equalization are applied to the standard traffic sign images for reducing the effect of variable lighting and illumination and then used to train the SVM.

4. Result and Discussion

4.1. Performance of Image Preprocessing. For saving the storage capacity and reducing the computational complexity, the original images are scaled down into 250 × 250 pixels. In the proposed approach, after the image acquisition process described in Section 2, the image preprocessing is performed by the RGB segmentation approach. In the proposed approach, a filter is applied on each channel threshold field to select just those regions of the image where values of the pixels are in the range of the target object. The region of interest (ROI) is actually the logical sum of the three filtered channels of R, G, and B, as shown in Figure 4. The median filter is applied for image smoothing and filling the smaller regions of the image, which is shown in Figure 4(f).

4.2. Performance of Traffic Sign Detection. The final selected candidates such as range of pixel values, area, and shape are drawn on the image by using extracted data (centre and area) of each of them. In the proposed method, only consider those traffic signs containing red colours. After applying shape-matching technique for the images containing the nontraffic signs, the output is given that "no road sign is detected." The result has been classified into four sections. False positive (FP) is where the sign is not detected correctly. For the false negative (FN), the sign is detected as a nonsign region. True positive (TP) is defined as the sign is correctly detected and in the true negative (TN), a nonsign region is correctly recognised as a nonsign region. The contingency matrix of the detection performance is given in Table 2.

From Table 4, the sensitivity and specificity values are calculated. Sensitivity is defined as the ability of identifying a condition correctly whereas specificity is defined as the ability of excluding a condition correctly:

(i) Sensitivity or recall = TP/(TP + FN) = 83.4%.

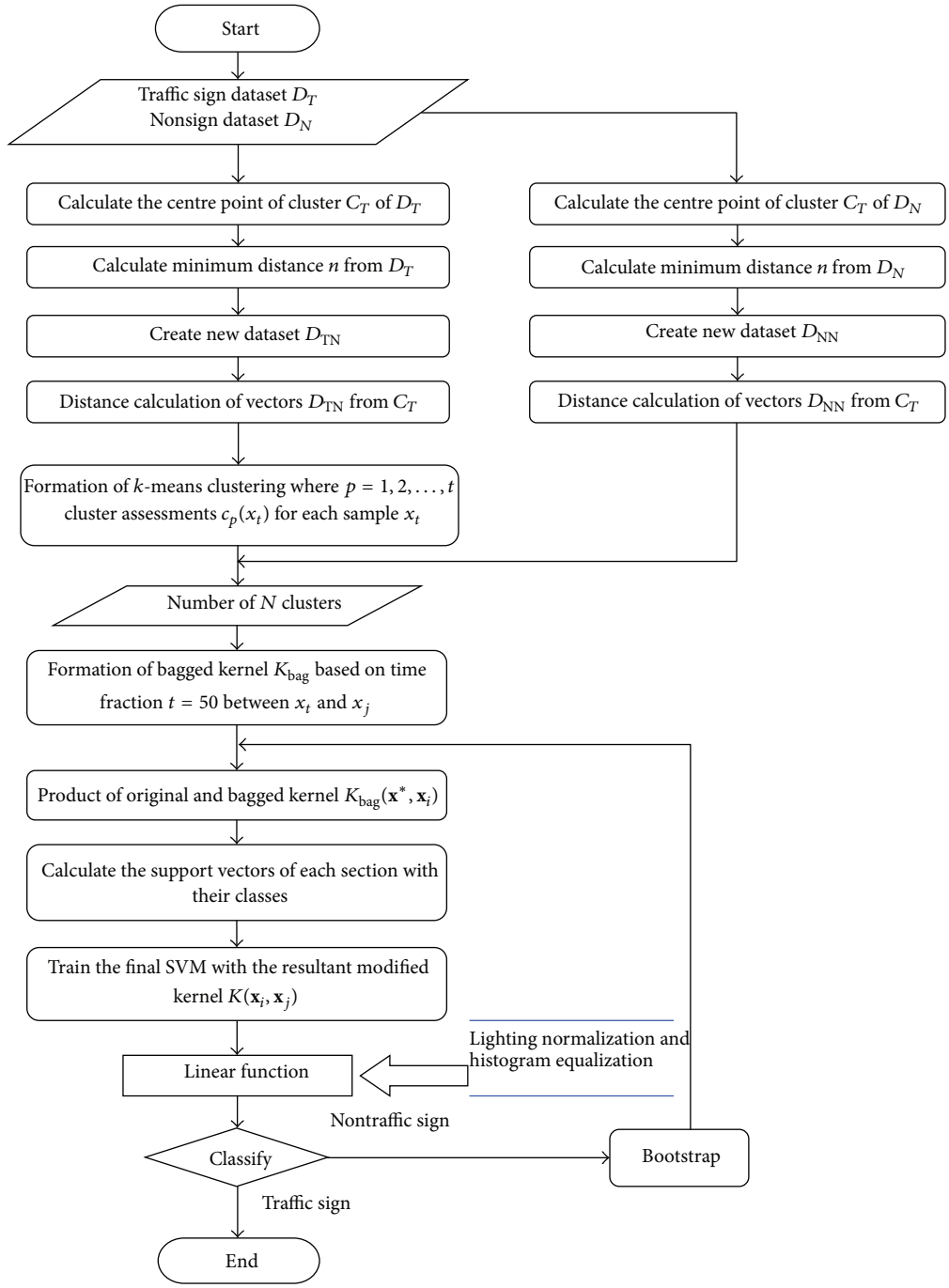


Figure 3: Flow chart of parallel SVM with bagged kernel.

(ii) Specificity = TN/(TN + FP) = %100.

(iii) Accuracy = (TP + TN)/(TP + TN + FP + FN) = 94.85%.

In the tests, it has been concluded that several problems affected the detection performance. Variant lighting conditions, occultation, and illumination of traffic signs are the main reasons of the false detection. The outcome of the proposed detection method shows that the red colour of the traffic sign is segmented and unswervingly illuminated by the sun. This happens because of the property of the colour segmentation using RGB model involved in comparing the RGB values. In the developed system, the computational time is around 0.25 s and the accuracy rate is 94.85%. Figure 4 shows the detection steps of the traffic sign detection system. The result of our detected traffic sign is given in Figures 5 and 6. In Figure 5, first and second columns show the true positives and true negatives, respectively. Third column shows the false negatives.

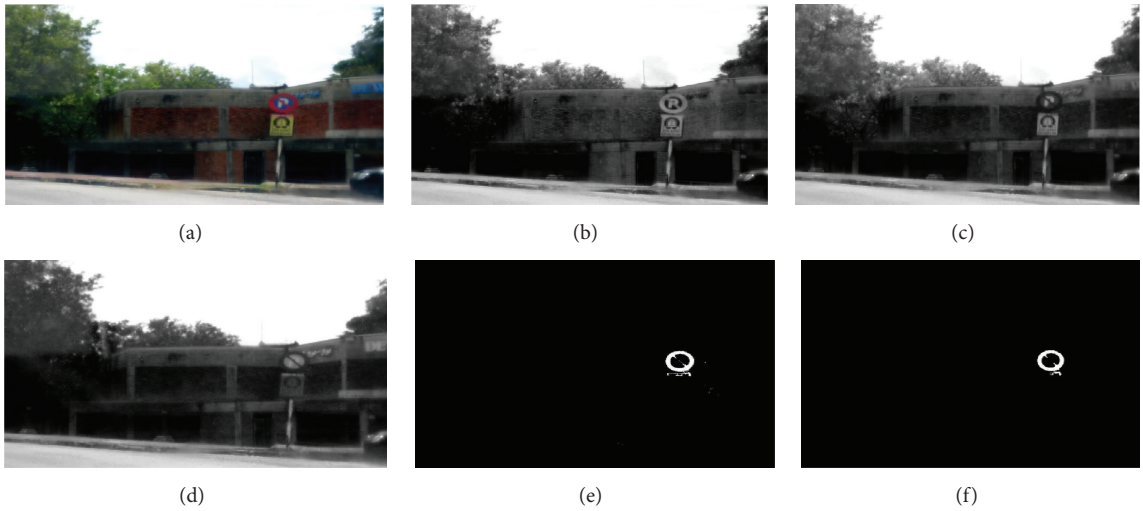


Figure 4: Colour processing for traffic sign detection: (a) original image, (b) R channel after threshold, (c) G channel after threshold, (d) B channel after threshold, (e) logical sum of three channels, and (f) ROI after filtering and smoothing.

Table 2: Contingency matrix of the RGB segmentation and shape matching sign detection method.

				Total
Test outcome	Test outcome positive	TP = 105	FP = 0	105
	Test outcome negative	FN = 18	TN = 227	245
		123	**227**	**350**

4.3. Performance of Recognition. In the proposed system, after the colour segmentation and shape matching, semisupervised SVM is applied and the total image is clustered for building the bagged kernel. After that, the modification of the base kernel is done. A number of SVMs are trained separately using a bootstrap algorithm and then they are aggregated via a suitable combination technique. Intensity correction and histogram equalization are applied to the standard traffic sign images for reducing the effect of variable lighting and illumination and then used to train the SVM. A total number of 350 images consist of two different shapes which are circular and octagonal, respectively. Table 3 shows the database used to train the NN.

Table 4 shows the final recognition results after the colour segmentation and shape matching technique are applied. Among the 123 traffic signs, 79 signs are octagonal considered as G1 and 44 of them are circular traffic signs considered as G2. Among the 44 traffic signs, there are also two different classes such as "no parking" and "do not enter" signs.

According to this data, evaluation parameters are sensitivity or recall, specificity, precision or PPV, FPR, and accuracy rate (AR) based on the number of FP, FN, TP, and TN values as follows:

(i) Sensitivity or recall = TP/(TP + FN) = 89.43%.

(ii) Specificity = TN/(TN + FP) = 99.12%.

(iii) Precision or PPV = TP/(TP + FP) = 98.21%.

(iv) False positive rate (FPR) = FP/(FP + TN) = 0.009.

(v) Accuracy = (TP + TN)/(TP + TN + FP + FN) = 95.71%.

The overall accuracy of the traffic sign recognition is 95.71% whereas the accuracy of the detection phase is 94.85%. According to the data analysis, the TPR is 89.43% whereas the FPR is 0.009%. Octagonal or "BERHENTI" sign has the highest recognition rate of 94.94% and "no parking" sign has the lowest rate of 84.09%. The processing time of the recognition system is 0.18 s. The overall processing time of the TSDR system is 0.43 s. To evaluate the system performance the ROC curve and the area under the curve are shown in Figure 7.

4.4. Performance Comparison of SVM Based Recognition System. A comparison of previous studies in detecting the traffic sign is given in Table 4. From Table 5, it can be observed that SVM used in [44] has the highest recall rate with an overall good accuracy of over 90%. 327 signs out of 340 signs are correctly classified. MSER and HOG based SVM used in [30] had the highest overall accuracy of 97.6% with a false positive rate of 0.85 and 92 signs out of 104 signs are classified correctly. In the proposed system, the lowest false positive rate was 0.009 and accuracy 95.71%. The precision is 98.21% and recall is 89.43%. 112 signs among the 123 detected signs are classified correctly. The proposed method has the highest precision rate (98.21%) and lowest FPR (0.009). The accuracy of the proposed method is 95.71%, which is good compared to other systems.

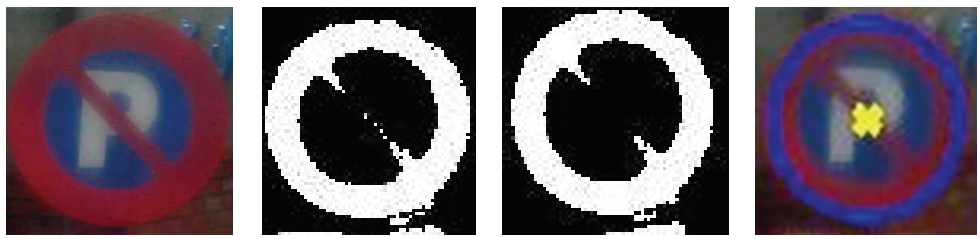
The main limitation of the developed system is that it is only applicable for red traffic signs. The "warning sign" and the "prohibitory sign" contain red which is the most important sign as they are more responsible for traffic accidents. The proposed method has a low detection rate as colour tends to be unreliable due to various factors like

(a) Normal daylight (b) High illumination (c) Shade

(d) Evening (e) Daylight (f) Shade

(g) Daylight (h) Daylight (i) Shade

Figure 5: Examples of TP in variant lighting conditions (a), (b), and (c); example of TN in variant lighting conditions (d), (e), and (f); and examples of false detection (g), (h), and (i).



(a) (b) (c) (d)

Figure 6: Final detection: (a) sample traffic sign, (b) closed curve obtained by colour thresholding, (c) after filtering and smoothing the candidate, and (d) detected ROI after shape matching and candidate selection.

Table 3: Modification of traffic signs used to train NN.

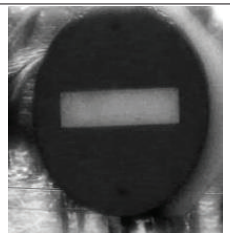
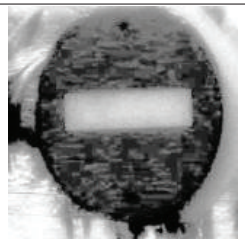
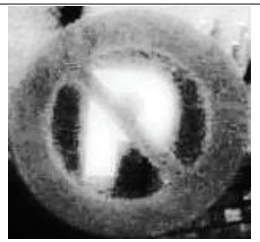
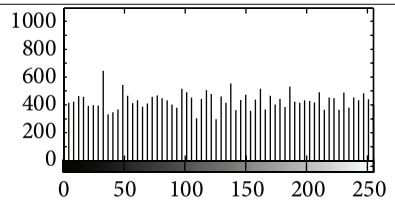
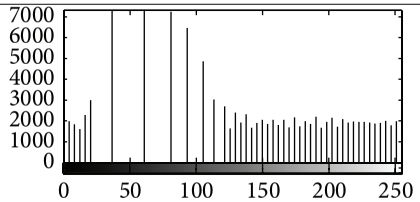
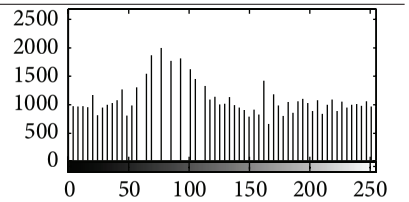
	Octagonal (stop)	Circular (do not enter)	Circular (do not park)
Original			
Greyscale conversion			
Illumination			
Intensity correction			
Applying histogram equalization			

Table 4: Example of traffic signs used to train NN.

				Total
Test outcome	Test outcome positive	TP = 110	FP = 2	112
	Test outcome negative	FN = 13	TN = 225	238
		123	**227**	**350**

Table 5: Comparison between proposed method and several existing methods.

Reference	Total sign	Correctly classified sign	Precision (%)	Recall (%)	False positive rate	Overall accuracy (%)	Processing time (s)
[44]	340	327	96.51	**92.97**	0.13	90.27	0.35
[30]	104	92	88.75	81.35	0.85	**97.6**	—
[8]	104	38	41.03	34.15	0.26	93.6	—
[45]	**650**	—	—	—	1.2	86.7	—
Proposed method	123	112	**98.21**	89.43	**0.009**	95.71	0.43

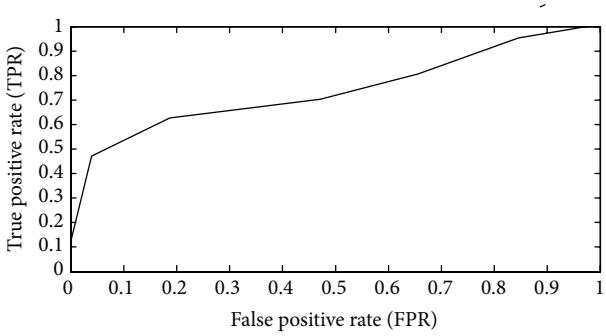


Figure 7: ROC curve; FPR versus TPR.

illumination, variable lighting, blurring, and fading. That is why the recognition process is also affected in terms of overall accuracy rate. Another limitation is the lack of images in the Malaysian traffic sign database. The overall processing time is 0.43 s, which is still in the higher side compared to [44]. To recognize all types of signs in Malaysia, reduce the processing time, and improve the Malaysian traffic sign database can be proposed as a future work.

5. Conclusion

The goal of this research is to develop an efficient TSDR system based on Malaysian traffic sign dataset. In the image acquisition stage, the images were captured by an on board camera under different weather conditions and the image preprocessing was done by using RGB colour segmentation. The recognition process is done by SVM with bagged kernel which is used for the first time for traffic sign classification. The developed system has shown promising results with respect to the accuracy of 95.71%, false positive rate (0.009), and processing time (0.43 s). The recognition performance is evaluated by using ROC curve analysis. The simulation results are compared with the existing methods showing the correctness of the implementation.

Conflict of Interests

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

References

[1] J. P. C. Pascual, *Advanced driver assistance system based on computer vision using detection, recognition and tracking of road signs [Ph.D. thesis]*, Charles III University of Madrid, Getafe, Spain, 2009.

[2] P. Paclík, J. Novovičová, and R. P. W. Duin, "Building road-sign classifiers using a trainable similarity measure," *IEEE Transactions on Intelligent Transportation Systems*, vol. 7, no. 3, pp. 309–321, 2006.

[3] A. Ruta, Y. M. Li, and X. H. Liu, "Detection, tracking and recognition of traffic signs from video input," in *Proceedings of the 11th International IEEE Conference on Intelligent Transportation Systems (ITSC '08)*, pp. 55–60, IEEE, Beijing, China, December 2008.

[4] P. Gil Jiménez, S. M. Bascón, H. G. Moreno, S. L. Arroyo, and F. L. Ferreras, "Traffic sign shape classification and localization based on the normalized FFT of the signature of blobs and 2D homographies," *Signal Processing*, vol. 88, no. 12, pp. 2943–2955, 2008.

[5] S. Lafuente-Arroyo, S. Salcedo-Sanz, S. Maldonado-Bascón, J. A. Portilla-Figueras, and R. J. López-Sastre, "A decision support system for the automatic management of keep-clear signs based on support vector machines and geographic information systems," *Expert Systems with Applications*, vol. 37, no. 1, pp. 767–773, 2010.

[6] G. A. Tagunde and N. J. Uke, "Detection, classification and recognition of road traffic signs using color and shape features," *International Journal of Advanced Technology & Engineering Research*, vol. 2, no. 4, pp. 202–206, 2012.

[7] H. Gündüz, S. Kaplan, S. Günal, and C. Akinlar, "Circular traffic sign recognition empowered by circle detection algorithm," in *Proceedings of the 21st Signal Processing and Communications Applications Conference (SIU '13)*, pp. 1–4, IEEE, New York, NY, USA, April 2013.

[8] S. Maldonado-Bascón, S. Lafuente-Arroyo, P. Gil-Jiménez, H. Gómez-Moreno, and F. López-Ferreras, "Road-sign detection and recognition based on support vector machines," *IEEE Transactions on Intelligent Transportation Systems*, vol. 8, no. 2, pp. 264–278, 2007.

[9] G. A. Tagunde and N. J. Uke, "Detection, recognition and recognition of road traffic signs using colour and shape features," *International Journal of Advance Technology & Engineering Research*, vol. 2, no. 4, pp. 202–206, 2012.

[10] L. Priese and V. Rehrmann, "On hierarchical color segmentation and applications," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '93)*, pp. 633–634, IEEE, New York, NY, USA, June 1993.

[11] M. J. Swain and D. H. Ballard, "Color indexing," *International Journal of Computer Vision*, vol. 7, no. 1, pp. 11–32, 1991.

[12] A. Hechri and A. Mtibaa, "Automatic detection and recognition of road sign for driver assistance system," in *Proceedings of the 16th IEEE Mediterranean Electrotechnical Conference (MELECON '12)*, pp. 888–891, Yasmine Hammamet, Tunisia, March 2012.

[13] G. Overett and L. Petersson, "Large scale sign detection using HOG feature variants," in *Proceedings of the IEEE Intelligent Vehicles Symposium (IV '11)*, pp. 326–331, Baden-Baden, Germany, June 2011.

[14] F. Zaklouta and B. Stanciulescu, "Real-time traffic-sign recognition using tree classifiers," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 4, pp. 1507–1514, 2012.

[15] A. Møgelmose, M. M. Trivedi, and T. B. Moeslund, "Vision-based traffic sign detection and analysis for intelligent driver assistance systems: perspectives and survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 4, pp. 1484–1497, 2012.

[16] S. Vitabile, G. Pollaccia, G. Pilato, and E. Sorbello, "Road signs recognition using a dynamic pixel aggregation technique in the HSV color space," in *Proceedings of the 11th International Conference on Image Analysis and Processing (ICIAP '01)*, pp. 572–577, Palermo, Italy, September 2001.

[17] D. M. Gavrila, "Traffic sign recognition revisited," in *Mustererkennung 1999: 21. DAGM-Symposium Bonn, 15.-17. September 1999*, pp. 86–93, Springer, Berlin, Germany, 1999.

[18] B. Höferlin and K. Zimmermann, "Towards reliable traffic sign recognition," in *Proceedings of the IEEE Intelligent Vehicles Symposium*, pp. 324–329, Xi'an, China, June 2009.

[19] V. A. Prisacariu, R. Timofte, K. Zimmermann, I. Reid, and L. Van Gool, "Integrating object detection with 3D tracking towards a better driver assistance system," in *Proceedings of the 20th International Conference on Pattern Recognition (ICPR '10)*, pp. 3344–3347, IEEE, Istanbul, Turkey, August 2010.

[20] M. A. Hannan, A. Hussain, and S. A. Samad, "Decision fusion via integrated sensing system for a smart airbag deployment scheme," *Sensors and Materials*, vol. 23, no. 3, pp. 179–193, 2011.

[21] C.-Y. Fang, S.-W. Chen, and C.-S. Fuh, "Road-sign detection and tracking," *IEEE Transactions on Vehicular Technology*, vol. 52, no. 5, pp. 1329–1341, 2003.

[22] S. Lafuente-Arroyo, S. Maldonado-Bascon, P. Gil-Jimenez, H. Gomez-Moreno, and F. Lopez-Ferreras, "Road sign tracking with a predictive filter solution," in *Proceedings of the 32nd Annual Conference on IEEE Industrial Electronics (IECON '06)*, vol. 1–11, pp. 3314–3319, IEEE, Paris, France, November 2006.

[23] Y. H. Ohara, I. Nishikawa, S. Miki, and N. Yabuki, "Detection and recognition of road signs using simple layered neural networks," in *Proceedings of the 9th International Conference on Neural Information Processing (ICONIP '02)*, vol. 2, pp. 626–630, IEEE, 2002.

[24] J. Torresen, J. W. Bakke, and L. Sekanina, "Efficient recognition of speed limit signs," in *Proceedings of the 7th International IEEE Conference on Intelligent Transportation Systems (ITSC '04)*, vol. 90, pp. 652–656, Washington, DC, USA, October 2004.

[25] Y. Aoyagi and T. Asakura, "A study on traffic sign recognition in scene image using genetic algorithms and neural networks," in *Proceedings of the IEEE 22nd International Conference on Industrial Electronics, Control, and Instrumentation (IECON '96)*, vol. 3, pp. 1838–1843, IEEE, Taipei, Taiwan, August 1996.

[26] A. de la Eccalera, J. M. Arminogol, and M. A. Salichs, "Traffic sign detection for driver support systems," in *Proceedings of the International Conference on Field and Service Robotics (FSR '01)*, Helsinki, Finland, June 2001.

[27] L. Chen, Q. Li, M. Li, L. Zhang, and Q. Mao, "Design of a multi-sensor cooperation travel environment perception system for autonomous vehicle," *Sensors*, vol. 12, no. 9, pp. 12386–12404, 2012.

[28] Y. Li, P. Sharath, and W. Guan, "Real-time traffic sign detection: an evaluation study," in *Proceedings of the 20th International Conference on Pattern Recognition (ICPR '10)*, pp. 3033–3036, Istanbul, Turkey, August 2010.

[29] J. Greenhalgh and M. Mirmehdi, "Traffic sign recognition using MSER and Random Forests," in *Proceedings of the 20th European Signal Processing Conference (EUSIPCO '12)*, pp. 1935–1939, August 2012.

[30] J. Greenhalgh and M. Mirmehdi, "Real-time detection and recognition of road traffic signs," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 4, pp. 1498–1506, 2012.

[31] M. A. Hannan, S. B. Wali, T. J. Pin, A. Hussain, and S. A. Samad, "Traffic sign recognition based on neural network for advance driver assistance system," *Przegląd Elektrotechniczny*, vol. 1, no. 12, pp. 169–172, 2014.

[32] Y. Sheng, K. Zhang, C. Ye, C. Liang, and J. Li, "Automatic detection and recognition of traffic signs in stereo images based on features and probabilistic neural networks," in *Optical and Digital Image Processing*, vol. 7000 of *Proceedings of SPIE*, p. 12, April 2008.

[33] S. Yang, X. Wu, and Q. Miao, "Road-sign segmentation and recognition in natural scenes," in *Proceedings of the IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC '11)*, pp. 1–4, Xi'an, China, September 2011.

[34] M. A. García-Garrido, M. Ocaña, D. F. Llorca, M. A. Sotelo, E. Arroyo, and A. Llamazares, "Robust traffic signs detection by means of vision and V2I communications," in *Proceedings of the 14th International IEEE Conference on Intelligent Transportation Systems (ITSC '11)*, pp. 1003–1008, IEEE, Washington, DC, USA, October 2011.

[35] J.-G. Park and K.-J. Kim, "Design of a visual perception model with edge-adaptive Gabor filter and support vector machine for traffic sign detection," *Expert Systems with Applications*, vol. 40, no. 9, pp. 3679–3687, 2013.

[36] B. Soheilian, N. Paparoditis, and B. Vallet, "Detection and 3D reconstruction of traffic signs from multiple view color images,"

ISPRS Journal of Photogrammetry and Remote Sensing, vol. 77, pp. 1–20, 2013.

[37] J. F. Khan, S. M. A. Bhuiyan, and R. R. Adhami, "Image segmentation and shape analysis for road-sign detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 1, pp. 83–96, 2011.

[38] I. Sebanja and D. B. Megherbi, "Automatic detection and recognition of traffic road signs for intelligent autonomous unmanned vehicles for urban surveillance and rescue," in *Proceedings of the 10th IEEE International Conference on Technologies for Homeland Security (HST '10)*, pp. 132–138, IEEE, Waltham, Mass, USA, November 2010.

[39] M. S. Prieto and A. R. Allen, "Using self-organising maps in the detection and recognition of road signs," *Image and Vision Computing*, vol. 27, no. 6, pp. 673–683, 2009.

[40] X. Gao, N. Shevtsova, K. Hong et al., "Vision models based identification of traffic signs," in *Proceedings of the 1st European Conference on Colour in Graphics, Imaging and Vision (CGIV '2002)*, pp. 47–51, April 2002.

[41] J. Miura, T. Kanda, and Y. Shirai, "An active vision system for real-time traffic sign recognition," in *Proceedings of the IEEE 2000 Intelligent Transportation Systems Conference*, pp. 52–57, October 2000.

[42] B. Alefs, G. Eschemann, H. Ramoser, and C. Beleznai, "Road sign detection from edge orientation histograms," in *Proceedings of the IEEE Intelligent Vehicles Symposium (IV '07)*, pp. 993–998, Istanbul, Turkey, June 2007.

[43] D. Tuia and G. Camps-Valls, "Semisupervised remote sensing image classification with cluster kernels," *IEEE Geoscience and Remote Sensing Letters*, vol. 6, no. 2, pp. 224–228, 2009.

[44] M. A. García-Garrido, M. Ocaña, D. F. Llorca, M. A. Sotelo, E. Arroyo, and A. Llamazares, "Robust traffic signs detection by means of vision and V2I communications," in *Proceedings of the 14th IEEE International Intelligent Transportation Systems Conference (ITSC '11)*, pp. 1003–1008, IEEE, Washington, DC, USA, October 2011.

[45] T. Bui-Minh, O. Ghita, P. F. Whelan, and T. Hoang, "A robust algorithm for detection and classification of traffic signs in video data," in *Proceedings of the International Conference on Control, Automation and Information Sciences (ICCAIS '12)*, pp. 108–113, Ho Chi Minh City, Vietnam, November 2012.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 821574, 12 pages
http://dx.doi.org/10.1155/2015/821574

Research Article

A Capacity-Restraint Transit Assignment Model When a Predetermination Method Indicates the Invalidity of Time Independence

Haoyang Ding,[1,2] Yu Bao,[1,2] Sida Luo,[1,2] Hanxia Shen,[1,2] Wei Wang,[1,2] and Man Long[1,2]

[1] *Jiangsu Key Laboratory of Urban ITS, Southeast University, Nanjing 210096, China*
[2] *Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies, Nanjing 210096, China*

Correspondence should be addressed to Haoyang Ding; dinghy123@hotmail.com

Received 8 May 2015; Revised 25 July 2015; Accepted 28 July 2015

Academic Editor: Yuanchang Xie

The statistical independence of time of every two adjacent bus links plays a crucial role in deciding the feasibility of using many mathematical models to analyze urban transit networks. Traditional research generally ignores the time independence that acts as the ground of their models. Assumption is usually made that time independence of every two adjacent links is sound. This is, however, actually groundless and probably causes problematic conclusions reached by corresponding models. Many transit assignment models such as multinomial probit-based models lose their effects when the time independence is not valid. In this paper, a simple method to predetermine the time independence is proposed. Based on the predetermination method, a modified capacity-restraint transit assignment method aimed at engineering practice is put forward and tested through a small contrived network and a case study in Nanjing city, China, respectively. It is found that the slope of regression equation between the mean and standard deviation of normal distribution acts as the indicator of time independence at the same time. Besides, our modified assignment method performs better than the traditional one with more reasonable results while keeping the property of simplicity well.

1. Introduction

Urban transit network is becoming a hot issue especially in developing countries like China where transit priority has risen to become a national policy. Developed countries in Europe witness this trend as well. Even in the US where car traffic dominates, as a major focus of transit network, transit accessibility is widely researched (e.g., [1–3]). The essential part of transit network analysis, transit assignment, is paid attention to concerning its framework, algorithm, equilibrium solution, and so on. For example, the solution algorithms for the multicriteria multimodal shortest path problem (M-SPP), which is known as NP-hard, in urban transit network were proposed [4]. This was followed by a label-setting algorithm for finding optimal hyperpaths in large transit networks with realistic headway distributions [5]. Among those models aimed at transit network, many of them have something to do with statistical independence, especially stochastic transit assignment models. A conditional logit model applied to traffic assignment and modal split was established which was able to cope with the independence of irrelevant alternatives (IIA) phenomenon in a very natural way [6]. Nested logit model that is widely believed to overcome the IIA characteristic was put forward as an extension of conditional logit model [7]. Cascetta et al. [8] introduced a commonality factor for overcoming the IIA problem in route choice, which was made good use of to estimate link flow variance and the variance of the path choice proportion in a proposed stochastic assignment model [9, 10]. Independence of route choice probability has been a focus since discrete choice model was brought in the area of traffic assignment.

The issue "independence" acts as one of the prerequisites that make a great difference to the validity of many models targeted at analyzing transit network. Traditional works, however, hardly paid enough attention to time independence of every two adjacent bus links. Substantially, this independence can never be neglected. MNP model, which is short for multinomial probit model, was first used in stochastic traffic assignment [11]. The model has no requirement on independence of irrelevant route choice but requires time independence of every two adjacent bus links. The requirement was applied to various logit models as well [12]. When probit-based stochastic user equilibrium (SUE) models were researched, most of them were implicitly based on the assumption of the correctness of time independence. Those researches varied from algorithms and dynamic pricing to sensitivity analysis [13–15]. Yet none of them concerned the validity of independence assumption, which might lead to inaccurate, even useless mathematical models when it came to their applications to real life. Meanwhile, plenty of research had something to do with independence test method in the statistics field. Entropy Theory is frequently made use of to test independence and the entropy was used as a measure of dependence, like serial dependence and spatial dependence, in these studies [16–18]. In addition, many other methods were proposed to conduct independence test. Broock et al. [19] presented an independence test method on the basis of an in-depth analysis on correlation dimension. Sugiyama and Suzuki [20] introduced least squares theory to nonparametric independence test. Those methods of test are subtle and theoretic. It is, however, rather hard to put them into practice for their complexity, especially to introduce them to time independence test in large urban transit networks.

This work is targeted at engineering practice of transit systems planning and management while two simple methods are proposed. To start with, we put forward a method to predetermine time independence of every two adjacent bus links. A characteristic quantity, α, the indicator of time independence, is derived to conduct independence predetermination, laying a foundation of the usage of many models. It is found that time independence is invalid when $\alpha > 0.5$, and data collected in metropolitan area of Nanjing city, China, is used to make a preliminary test on the practicability of our method. In the case $\alpha > 0.5$, transit network models on the basis of MNP lose their effects, especially those probit-based stochastic transit assignment models, so the number of alternative assignment models decreases. Meanwhile, capacity-restraint assignment method lacks the multipath characteristic [21] though widely used in engineering practice out of its simplicity and not bad precision. Therefore, a modified capacity-restraint transit assignment method is proposed following the predetermination method while taking good advantage of some predetermination results. The method maintains its effectiveness under the circumstance of invalidity of time independence. After introducing 95% quantile of normal distribution to the traditional capacity-restraint method to realize multipath transit assignment, it turns out that the results of the modified method are more sensible than the traditional one through a small contrived transit network.

2. Methodology

2.1. Time Independence Predetermination

2.1.1. Basic Assumptions

(1) Most bus drivers are technical in the network.

(2) The V/C (the ratio of road traffic volume to road capacity) can be somewhat large but no congestions occur when there exists a transit exclusive lane. Otherwise, V/C is medium or low.

(3) There is only a little interference to transit operations caused by pedestrians and bicycles.

(4) Transit priority control is advanced, so intersections have limited effect on transit operations.

These four basic assumptions are not impractical. Assumptions (1), (2), and (3) are valid in most cities of developed countries, while assumption (3) may be a bit questionable in some developing countries like China. These countries are known for mixed traffic flow that does make a big difference to transit operations in their big cities like Shanghai. Anyway, bicycles and pedestrians have little effect on public transit on the road with separation infrastructures. The effects are also quite limited in most of their medium and small cities. The most demanding assumption is (4) probably, which is easy to be satisfied in many European countries. It is, however, very difficult to guarantee an advanced transit priority control system. Anyhow, the priority control does exist in most cities of developed countries and many big cities of developing countries. In a word, it is likely that assumption (1) is the most accessible while assumption (4) is to the contrary.

As for time independence predetermination, the "time" here is to be defined strictly. The two sides of a bus line have a bus stop. When a bus begins to pull over at the upstream stop, the time is t_1. Similarly, t_2 is defined. We call $|t_2 - t_1|$ "stop-stop time" that is simply denoted by SST, which is composed of the parking time at the upstream bus stop and travel time between two adjacent stops. If all basic assumptions are satisfied, we could infer that SST at the bus link conforms to normal distribution according to the data in Nanjing. This can be seen from the section of distribution test below. To analyze the whole transit network basically satisfying all the assumptions above, we assume that all the links in the network conform to normal distribution, which is the foundation of the proposed methodology. More insight concerning the normal distribution can be seen in Discussion.

2.1.2. Mean and Standard Deviation Fitting. We fit the mean and std, which is short for standard deviation in this paper, of the normal distribution mentioned above. Bus links that satisfy all the basic assumptions to the greatest degree can be selected in a transit network. Accordingly, we could get several normal distributions: $N(\mu_1, {\sigma_1}^2), N(\mu_2, {\sigma_2}^2), \ldots, N(\mu_n, {\sigma_n}^2)$. μ and σ are estimated using the samples collected in transit survey. Note that μ and σ have the same dimension, so regression analysis is decided to be aimed at the std rather than variance. Furthermore, $\mu > 0$ and $\sigma > 0$, so instead of doing linear regression directly, we fit the μ and

σ in the Double Logarithmic Coordinate System and get the regression equation as follows:

$$\ln \sigma = \alpha \ln \mu + \theta. \tag{1}$$

The regression model is the fundamental of time independence predetermination and the modified capacity-restraint model. According to the data in Nanjing, when all chosen links meet the assumptions well, the goodness of fitting is perfect with high R-square which can be seen in Section 4. To analyze the whole transit network, we assume the regression equation has high goodness of fitting. If the regression model is not sufficiently significant, we may appropriately select other or more bus links that satisfy all basic assumptions well to avoid poor goodness of fitting. More analysis about the regression model can be seen in Section 5.

2.1.3. Indicator of Time Independence. After the regression analysis, we could predetermine time independence of every two adjacent bus links in a network. Time independence predetermination is to predetermine the independence of SST at every two adjacent links. Note that all bus links in the network conform to normal distribution and the goodness of fitting of regression equation is high on the basis of our analysis above. In this part, we put forward an indicator of independence whose value can simply predetermine time independence with practicability.

Considering normal distribution $N(x, y^2)$ with the mean x and std y, according to the previous section of fitting,

$$\ln y = \alpha \ln x + \theta \quad (x, y > 0). \tag{2}$$

α and θ are regression coefficients of the regression equation. The slope α is supposed to be greater than zero ($\alpha > 0$) because the std will increase when the mean increases. Another form of equation derived by deforming the regression equation above is

$$y = \mu x^{\alpha}, \quad \text{where } \ln \mu = \theta,\ \mu > 0. \tag{3}$$

Then the variance is given by the following expression:

$$\mathrm{VAR} = y^2 = \mu^2 x^{2\alpha} = \lambda x^{\beta} \quad \text{where } \lambda = \mu^2 > 0,\ \beta = 2\alpha. \tag{4}$$

A transit route is given in Figure 1. Here A represents the origin and B represents the destination. The mean of a random variable D_{ab}, the difference between the moment when a bus stops at station A and that at station B, is set at fixed value T. S_k ($0 \le k \le n$) represents the intermediate bus stop, and t_r ($0 \le r \le n+1$) is the mean of SST at bus link. Obviously, T can be expressed as follows:

$$T = \sum_{k=1}^{n+1} t_k. \tag{5}$$

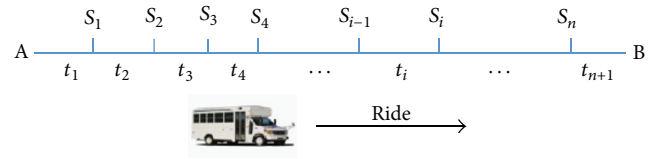


Figure 1: A transit route with its stops and SST.

Supposing that SST at every two adjacent links are mutually independent, then D_{ab} is normally distributed with variance ${\sigma_n}^2$ according to the linear superimposition principle of normal distribution. Considering there are n stops between station A and station B, we obtain

$${\sigma_n}^2 = \sum_{k=1}^{n+1} \mathrm{VAR}_k = \sum_{k=1}^{i-1} \lambda t_k^{\beta} + \lambda t_i^{\beta} + \sum_{k=i+1}^{n+1} \lambda t_k^{\beta}. \tag{6}$$

If there are $n+1$ stops between station A and station B, now we assume naturally the stop S_{n+1} is set up between the S_{i-1} and S_n. Thus, t_i is divided into two parts as t_j and t_{n+2} ($t_i = t_j + t_{n+2}$). Similarly, the variance is given as follows under this circumstance:

$${\sigma_{n+1}}^2 = \sum_{k=1}^{n+2} \mathrm{VAR}_k = \sum_{k=1}^{i-1} \lambda t_k^{\beta} + \lambda t_j^{\beta} + \lambda t_{n+2}^{\beta} + \sum_{k=i+1}^{n+1} \lambda t_k^{\beta}. \tag{7}$$

By subtracting ${\sigma_n}^2$ from ${\sigma_{n+1}}^2$, the result is

$$\begin{aligned} {\sigma_{n+1}}^2 - {\sigma_n}^2 &= \lambda \left(t_j^{\beta} + t_{n+2}^{\beta} - t_i^{\beta} \right) \\ &= \lambda \left(t_j^{\beta} + t_{n+2}^{\beta} - \left(t_j + t_{n+2} \right)^{\beta} \right), \quad \lambda > 0. \end{aligned} \tag{8}$$

Function $f(t_j, t_{n+2})$ is constructed where t_j, $t_{n+2} > 0$. Its partial deviation is obtained:

$$\begin{aligned} f\left(t_j, t_{n+2}\right) &= t_j^{\beta} + t_{n+2}^{\beta} - \left(t_j + t_{n+2} \right)^{\beta} \\ \frac{\partial f}{\partial t_j} &= \beta \left(t_j^{\beta-1} - \left(t_j + t_{n+2} \right)^{\beta-1} \right) \\ \frac{\partial f}{\partial t_{n+2}} &= \beta \left(t_{n+2}^{\beta-1} - \left(t_j + t_{n+2} \right)^{\beta-1} \right). \end{aligned} \tag{9}$$

The analytical results could be obtained. When $\beta > 1$ or $\beta < 0$ and $f(t_j, t_{n+2})$ is a decreasing function, then $f(t_j, t_{n+2}) < f(0,0) = 0$, which is equivalent to ${\sigma_{n+1}}^2 < {\sigma_n}^2$; when $\beta = 1$, $f(t_j, t_{n+2}) = 0$, and ${\sigma_{n+1}}^2 = {\sigma_n}^2$; when $\beta = 0$, $f(t_j, t_{n+2}) = 1$, and ${\sigma_{n+1}}^2 > {\sigma_n}^2$; and when $0 < \beta < 1$ and $f(t_j, t_{n+2})$ is an increasing function, then $f(t_j, t_{n+2}) > f(0,0) = 0$, which is equivalent to ${\sigma_{n+1}}^2 > {\sigma_n}^2$.

As is analyzed before, $\beta = 2\alpha$, and D_{ab} conforms to normal distribution. Note that $\alpha > 0$; the results are given as follows:

(1) When $\alpha > 0.5$, the variance of D_{ab} decreases with the increasing of the number of intermediate bus stops.

(2) When $\alpha = 0.5$, the variance of D_{ab} is constant with the changing of the number of intermediate bus stops.

(3) When $0 < \alpha < 0.5$, the variance of D_{ab} increases with the increasing of the number of intermediate bus stops.

Under the premise of a fixed mean of D_{ab}, the variance should increase with the number of intermediate stops increasing. This is because the increasing of the number of bus stops results in starting and stopping more frequently, causing the uncertainty of travel time to become greater. The variance is a strong indicator of uncertainty. To sum up, all the analysis leads to a conclusion of predetermining time independence.

When $\alpha \geq 0.5$, if the independence of SST holds, a clear contradiction exists. The SST at every two adjacent bus links, therefore, is not mutually independent. Time independence is not valid in this case.

When $0 < \alpha < 0.5$, the SST at every two adjacent links is probably mutually independent, where a further demonstration is needed. In this case, we could not decide whether or not time independence is valid. This is what "predetermination" derives from.

Therefore, the parameter α is not only a regression coefficient of the equation expressed as $\ln y = \alpha \ln x + \theta$, but also an indicator of time independence. On the one hand, the larger α is, the higher time uncertainty is in urban transit network. Considering uncertainty avoidance theory, we hope α is not too large. On the other hand, the relative magnitude of the indicator and 0.5 predetermines whether SST in transit network is independent or not. This decides whether some stochastic models are valid to a large extent. Taking MNP-based assignment model as an example, it should not be applied if $\alpha \geq 0.5$. Anyway, we cannot determine the validity of this model under the condition $0 < \alpha < 0.5$.

2.2. A Modified Capacity-Restraint Transit Assignment Method. If time independence is predetermined to be not valid, many transit assignment methods such as MNP-based models should never be used. Under our basic assumptions, methods based on logit models are untenable due to the normal distribution for SST. As a result, many assignment methods prove ineffective. Traditional capacity-restraint transit assignment method is not affected, though it could not take time uncertainty into consideration. So this part proposes a modified capacity-restraint transit assignment method on the basis of normal distribution for SST and regression equation between the distribution mean and std. When the invalidity of time independence has been predetermined, unlike many assignment methods, the proposed one keeps its effectiveness. It remains useful even though time independence is proved to be solid.

Compared with traditional capacity-restraint method, the modified one considers time uncertainty by using 95% quantile of normal distribution while keeping the advantages of traditional method like the simplicity of calculation and so on. It is believed that traditional heuristic assignment methods are faced with the "common lines" problem, and many modifications have been made to overcome the obstacle [22]. These modifications can be used in our method as well if the "common lines" problem is supposed to be considered in a specific case. It is not difficult to generalize the proposed method using the modifications but requires a little future work. So, in this paper, the "common lines" problem is not much concerned including the contrived network we use below. Flow chart of the modified capacity-restraint transit assignment method is shown in Figure 2. When given a traditional transit network, we could regard the time at bus links as mean value of SST. Then, the previous regression equation is made use of to work out the std at all links. In this way, the normal distribution of every bus link is determined. As Figure 3 shows where A, B, C, D, and E represent bus stops and $N(\mu, \sigma^2)$ stands for normal distribution, the modified method is based on transit network with SST treated as random variable rather than deterministic variable, which is closer to reality.

The rationale of the modified method is similar to the traditional one, though the coefficient of assignment γ ($0 < \gamma < 1$) is introduced to the model. We assign 100γ percent of the OD, which is short for origin-destination, to the shortest path on the basis of the mean value of SST to obtain the passenger volume X_a. And $100(1 - \gamma)$ percent is assigned to the shortest path after each SST values their 95% quantile on the shortest route to obtain Y_a. The 95% quantile is chosen because it has a good sense in probability studies. With this method, the total travel time of shortest path with mean SST could be maximized by using 95% quantile of the SST of each bus link composing the route, which is aimed at achieving multishortest paths. It is suggested that γ should be specified as 0.75 [23]. So the final passenger volume Z_a could be calculated using the formula

$$Z_a = 0.75X_a + 0.25Y_a. \tag{10}$$

Note that the generalized cost of bus travel changes after loading one part of OD into transit network, but the std of SST is to remain the same unless some changes about the mean of SST take place. Then the remaining parts of the modified method are the same as that of the traditional one. To make this revised method more useful in engineering practice, efficiency should be paid special attention to. As a result, we recommend dividing the whole OD into three shares by 50%, 30%, and 20%, respectively.

3. Data Preparation

3.1. Bus Line Selection. Data has been collected through transit survey in metropolitan area of Nanjing, China. Considering all the basic assumptions of our methods and the traffic

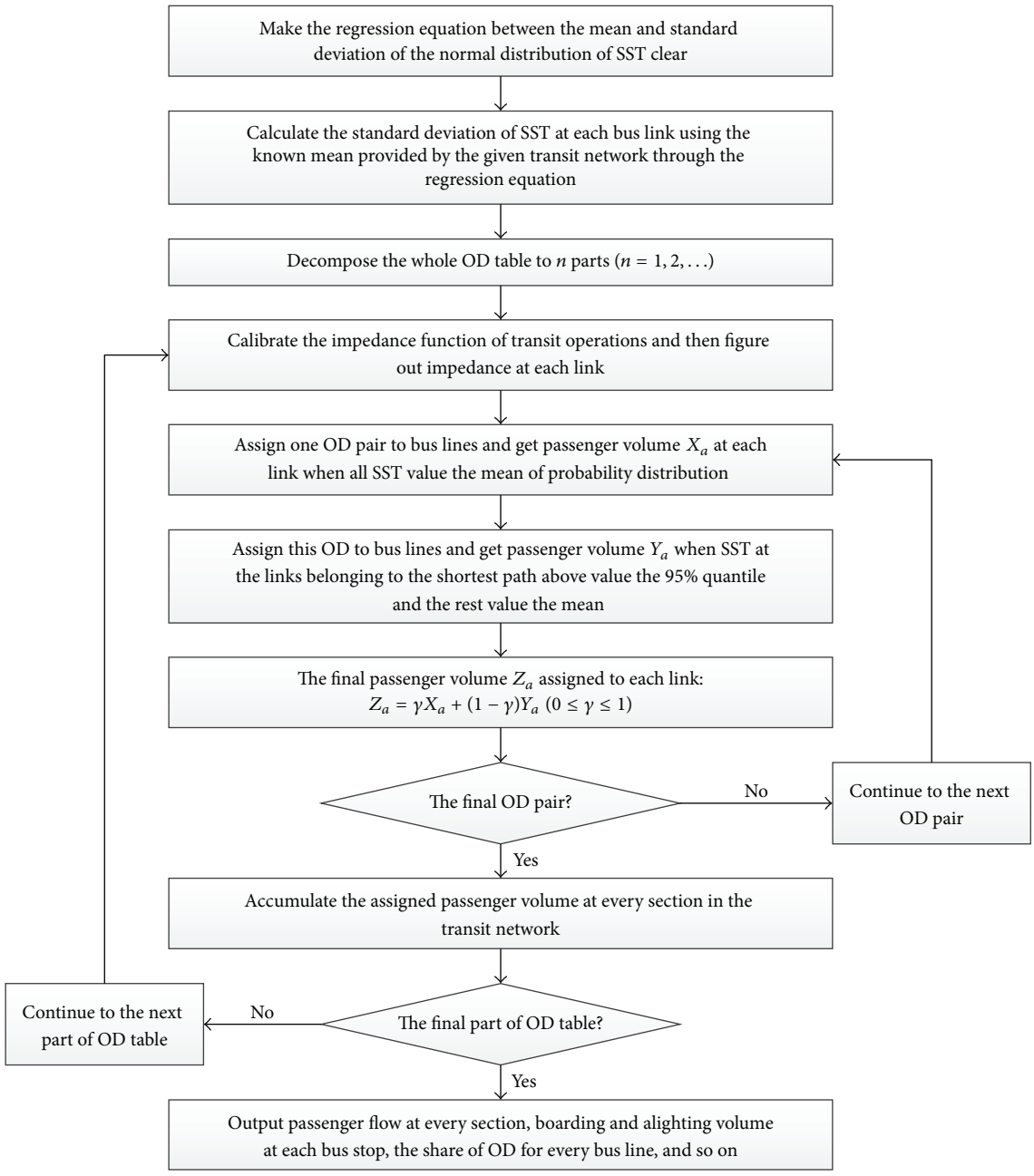


Figure 2: Flow chart of the modified capacity-restraint assignment method.

environment in real world, the bus lines are selected according to the following points:

(a) All the bus lines pass the metropolitan area of Nanjing, which are widely distributed in the city center.

(b) Bicycles and pedestrians make a little difference to bus operations, especially where the side median exists.

(c) Buses have priority on signalized intersections.

(d) Survey had better be conducted on a sunny day to avoid small probability events.

Taking all these factors into consideration, 7 groups of adjacent bus stops were randomly picked out, and corresponding bus lines were selected in the metropolitan area of Nanjing. Note that there may exist several choices of bus lines when a pair of bus stops is determined. So it is rationally assumed that there exists no difference between these lines concerning their operations between the pair of bus stops. To avoid waiting for a long time at bus stops, generally bus lines with too low frequency were never selected in this survey.

3.2. The Survey. The transit survey was conducted from 4:00 p.m. to 7:00 p.m. on weekdays to collect data at peak hours.

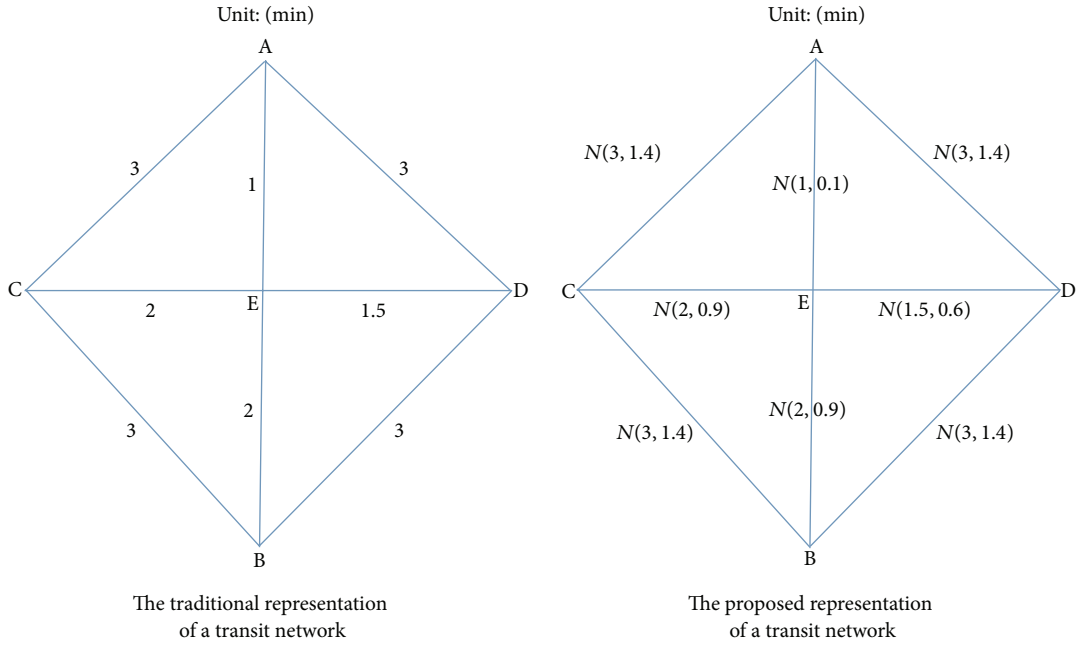


Figure 3: Comparison of the proposed and traditional abstraction of networks.

The investigators must record the time when their targeted buses stop for the first time at the bus stop. Then they should go on their buses and record the time when the buses pull over at the next stop. In theory, investigators should repeat the process for 16 times in single direction. In this way, however, it is hard for the survey to get finished in 3 hours. So investigators were permitted to repeat the process for 8 times in dual directions if the bus frequency was rather low. The error is typically rather limited if the bus stop on the other side of road is very close to the stop on this side in terms of the distance along the road.

3.3. Preliminary Data Analysis. From the recorded time when buses pull over at stops, SST samples are figured out and displayed in Table 1. The quality of our data is evaluated by the seven frequency histograms in Figure 4. According to these figures, it could be found qualitatively that the SST at bus links might conform to normal distribution, which is consistent with our conjecture in the former section. In addition, the correlation coefficient between the logarithm of sample mean and std is 0.996. This means the logarithm of the two basic statistics is highly correlated indicating a significant linear relationship. Therefore, the data acquired from this survey is preliminarily considered reliable. It can be made use of in the following analysis of the proposed methods.

Table 1: Result of the SST survey in metropolitan area of Nanjing.

Number	201	100	792	405	126	160	16
1	159	32	42	61	55	35	49
2	127	37	49	52	56	36	46
3	166	32	42	59	54	37	51
4	158	28	46	56	65	43	50
5	122	31	41	67	67	41	58
6	134	33	51	60	56	35	60
7	138	35	41	62	59	35	47
8	148	33	45	53	51	36	52
9	166	29	42	58	52	39	51
10	140	31	47	70	52	37	53
11	155	35	42	71	62	35	51
12	138	35	40	67	55	40	57
13	151	30	45	58	48	36	52
14	144	33	49	64	55	37	62
15	138	36	45	57	54	35	48
16	132	33	48	57	52	38	51
Mean	144.8	32.7	44.7	60.8	55.8	37.2	52.4
std	13.4	2.5	3.4	5.7	5.1	2.4	4.6

4. Results

4.1. Transit Network in Metropolitan Area of Nanjing

4.1.1. Distribution Test of SST. Although it has been assumed that SST at bus links conforms to normal distribution, distribution test could be made to give a preliminary proof. We conduct Jarque-Bera test on the basis of the data in Nanjing using MATLAB. According to the results displayed in Table 2, h values 0 under the level of significance of 0.05. Besides, it is seen that the scatters are basically uniformly distributed around the line that indicates a normal distribution from one of the test graphs (Figure 5). In summary, this assumption of normal distribution is tenable.

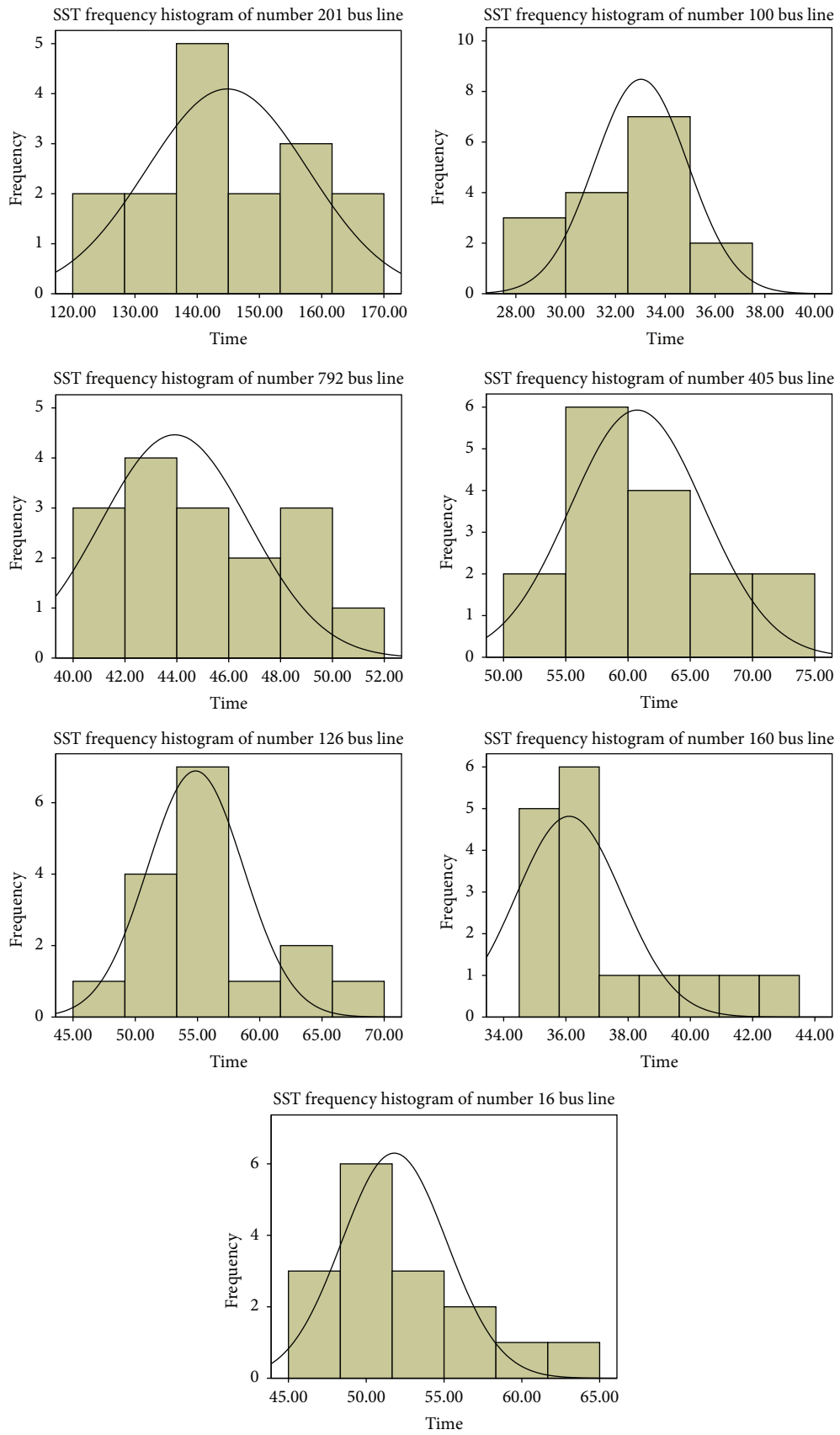


Figure 4: Frequency histogram of SST of each bus line.

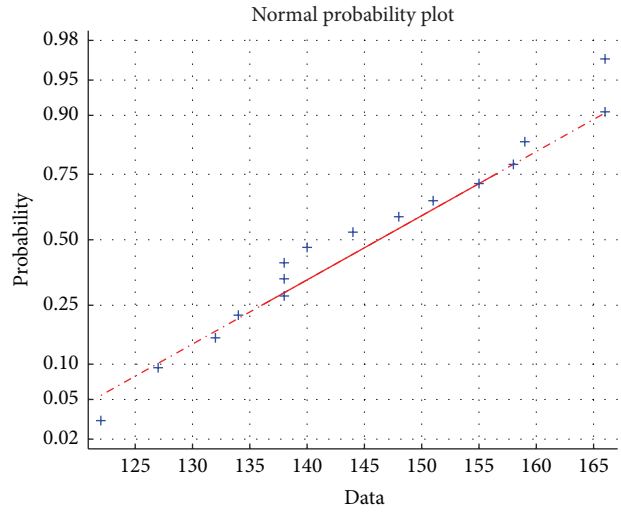


Figure 5: Fitting curve of the SST of number 201 bus line.

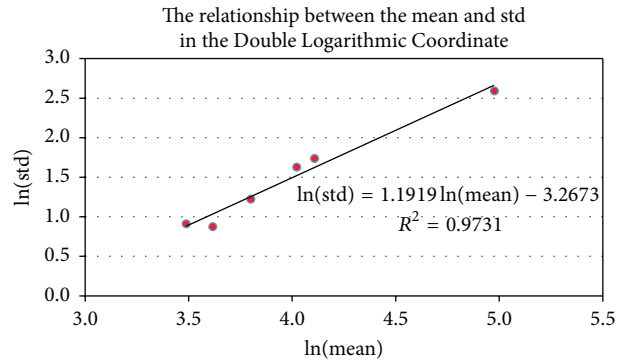


Figure 6: Regression analysis between the mean and std of normal distributions.

4.1.2. Regression Analysis between the Mean and Standard Deviation. Data from 6 lines is used for fitting the relationship between the mean and std, and the rest is for examining the fitting result simply. According to Section 2.1.2, the regression analysis is conducted after taking logarithm with the help of MATLAB followed by an evaluation of its results. The coefficients of regression and some statistics are listed in Table 3 while the regression equation is shown in Figure 6.

Judging from the results, the coefficient of determination is close to 1 and the probability of significance values 0.0003 that is smaller than 0.05 under the level of significance of 0.05. Furthermore, the R-square is 0.9731 indicating that the regression relationship appears reliable. To determine the validity of the equation, the rest of the bus lines are used for checking through comparing the calculation from the regression equation with the observed data. For the std of SST of line number 16, the result by the regression equation is 4.2665 while data shows the value 4.6025. The relative error is 7.3%, which is small enough to prove the validity of the regression analysis. To sum up, the regression equation is significant. In addition, it has been assumed that the equation

Table 2: Result of the distribution test of SST.

Bus lines	h	CV	Mean	Variance
201	0	3.4140	144.75	180.47
100	0	3.4140	32.69	6.36
792	0	3.4140	44.69	11.56
405	0	3.4140	60.75	32.47
126	0	3.4140	55.81	26.30
160	0	3.4140	37.19	5.90
16	0	3.4140	52.38	21.18

Note: h = result of the test; if $h = 0$, X conforms to normal distribution; if $h = 1$, X does not conform to normal distribution; CV = the critical value whether to reject the null hypothesis or not.

Table 3: Result of the regression analysis.

b	1.1919		−3.2673	
bint	[0.9169, 1.4669]		[−4.3756, −2.1590]	
stats	0.9731	144.8071	0.0003	0.0139

Note: b = the least-square estimated value of the two regression coefficients; bint = the 95% confidence interval of the regression coefficients; stats = the coefficient of determination, F statistics of variance, the probability of significance of variance, and the estimate of variance, respectively.

between the mean and std of normal distribution of SST has high goodness of fitting. The results of the regression analysis can well support this assumption.

4.1.3. Time Independence of Every Two Adjacent Bus Links. On the basis of our regression equation (11), the indicator of independence α is 1.1919 and another parameter θ values −3.2673 where the unit of time is the second:

$$\ln(\text{std}) = 1.1919 \ln(\text{mean}) - 3.2673. \quad (11)$$

α is larger than 0, which suits our common sense. More importantly, α is larger than 0.5. First, it means that the independence of SST at every two adjacent bus links is invalid in metropolitan area of Nanjing. Consequently, models like MNP-based assignment can hardly be used to estimate passenger flow of transit system in the city center. It is worth mentioning that most logit-based models are groundless to be adopted in this city area because SST conforms to normal distribution, let alone the invalidity of time independence. It is these widely used models that lose their effects. Second, the time uncertainty of transit operations is never low because α is significantly larger than zero. These may account for the reasons why many models adopted by traffic engineers appeared ineffective on the transportation network in Nanjing, which indicates the significance of the proposed indicator to a great extent. More analysis could be made through comparison with that of other cities.

4.2. A Contrived Transit Network. A contrived transit network is established to compare our proposed assignment method with the traditional one. Impact from other traffic modes on transit operations is not considered in this example. There are three transit lines in Figure 7. The mean of SST

Table 4: Cost update of transit routes.

OD pair	Transit route	Our cost during loading 50%/30%/20% of OD (95% quantile SST if necessary)
Origin 1 to destination 3	1 → 2 → 3	4.9 (5.52)/8.1 (9.216)/10.02 (11.4336)
	1 → 5 → 6 → 3	28.5/32.7/34.5
	1 → 9 → 10 → 6 → 3	11.7/13.1/14.66
Origin 1 to destination 6	1 → 3 → 6	12.9/17.7/20.58
	1 → 5 → 6	8.7 (9.63)/10.8/ 11.7 (13.095)
	1 → 9 → 10 → 6	8.9/10.3 (11.757)/11.86
Origin 1 to destination 12	1 → 2 → 3 → 4 → 12	10.9 (11.55)/17.1 (19.611)/20.82
	1 → 9 → 10 → 6 → 3 → 4 → 12	27.9/34.5/39.54
	1 → 5 → 6 → 7 → 12	17.7/19.8/20.7 (23.49)
	1 → 9 → 10 → 6 → 7 → 12	23.7/25.8/28.14

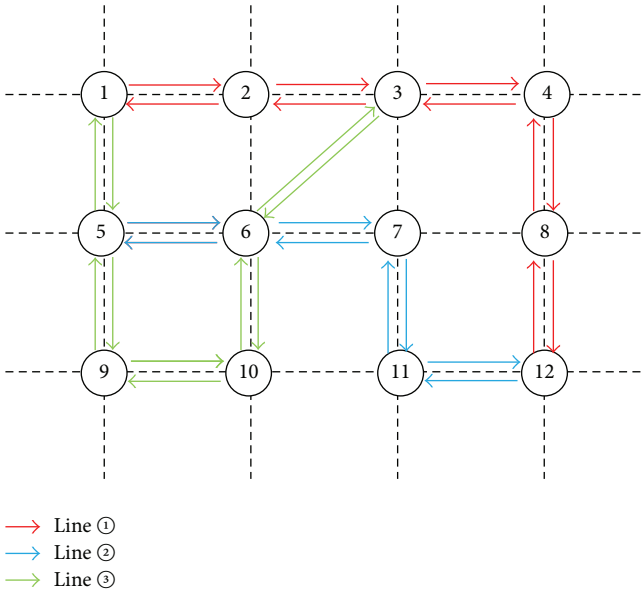


Figure 7: A contrived transit network.

between two adjacent stops is 2 min while it is 2.8 min between stop 3 and stop 6. Their fees are all 0.9 units. The frequencies and vehicle capacities of these bus lines are deterministic, and it is worked out that the rated capacity and maximum capacity of all the three lines are 500 passengers per hour and 800 passengers per hour, respectively [24]. OD from stops 1 to 3, 1 to 6, and 1 to 12 is 300, 400, and 500 passengers per hour, respectively.

Generalized cost function is commonly used in transportation studies [25]. Referring to the document mentioned above, the shortest path defined in this example is on the basis of generalized cost while passenger congestion constitutes one part of the cost. Meanwhile, we consult the form of generalized cost function used in [26]. Combining what is referred altogether, the cost of a route is determined as follows:

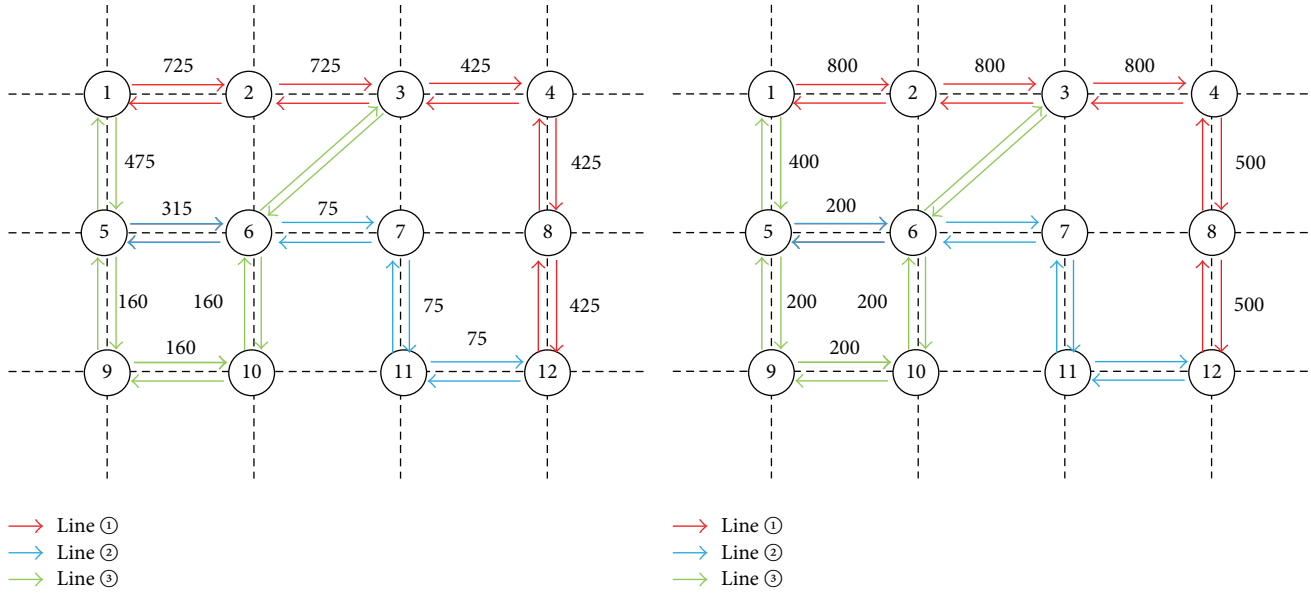
$$\begin{aligned} \mathrm{GC} &= \left(\eta \sum T \cdot k + F\right) \cdot q \\ V &\le \mathrm{MC}, \end{aligned} \tag{12}$$

where GC is generalized cost; T is mean SST; F indicates fees; η is value of travel time; k indicates discomfort coefficient; q is transfer penalty; V is passenger volume; MC is maximum capacity.

As is recommended in [26], it is decided that η is 1 unit/min and k is defined to be $V/C+1$, where V/C is the ratio of passenger volume to the rated capacity. And if i denotes transfer times, q equals $1.5i$, while it values 1 unless a transfer occurs. Traditional capacity-restraint method in this case is under the condition that OD is divided into 3 parts: 50%, 30%, and 20%, which is the same as that of our modified method. And the coefficient of assignment γ values 75% as recommended. If some routes have the same cost, passenger flow will be assigned to the route with the fewest transfers while they share the equivalent flow if their transfer times are the same. All routes that are feasible to be assigned are listed in Table 4 which provides the cost update during the assignment. These exclude the routes with too much cost (e.g., route 1 → 3 → 6 → 7 → 12 with the cost of 46.5 units) and reverse routes (e.g., route 1 → 5 → 9 → 5 → 6). Table 5 presents how our assignment method performs in each OD share, and results of the assigned passenger flow are shown in Figure 8.

5. Discussions

As for the independence predetermination, it is hoped that the indicator of time independence α is small when an urban transit network has low time uncertainty. Meanwhile, chances are that time independence is valid when α is small; to be specific, $0 < \alpha < 0.5$. Therefore, the most ideal situation is $0 < \alpha < 0.5$ when the transit network has low time



(a) Passenger volume obtained by the modified method

(b) Passenger volume obtained by traditional method

Figure 8: Results of transit assignment.

Table 5: Process of passenger flow assignment.

Transit route	50% OD share	30% OD share	20% OD share
1 → 2 → 3	400	240	85
1 → 5 → 6	150	30	60
1 → 9 → 10 → 6	50	90	20
1 → 2 → 3 → 4 → 12	250	150	25
1 → 5 → 6 → 7 → 12	0	0	75

Note: routes in Table 4 that are assigned zero flow during the process are not listed in this table.

uncertainty and network models requiring time independence, like MNP-based assignment models, can probably be used. With regard to the modified capacity-restraint method, it has proved to be more accurate than the traditional one. Heuristic assignment algorithms like capacity-restraint method are more common in engineering practice, so the modified method with the same time complexity as that of the traditional method is highly prospective. Comparing our assignment methods with equilibrium ones, it is valueless to take many SUE models into account when time independence is not valid. But the assertion is hard to be made that the modified heuristic model will be more accurate than SUE models when $0 < \alpha < 0.5$. Anyhow, the two methods we put forward have an advantage over others on simplicity. They can be adopted in engineering practice in any urban transit network with basic assumptions satisfied by and large, which is not a tough requirement.

The results of metropolitan area of Nanjing indicate the key point before modeling urban transit network. It is the basic assumptions that researchers should lay great emphasis on. Instead of asserting that all model assumptions could be satisfied with qualitative analysis, some quantitative analysis is also indispensable. The proposed indicator of time independence along with some statistics of probabilistic tests exactly plays the role. Some uncomplicated transit survey mentioned in this paper could be conducted to figure out these key indicators before selecting reasonable models.

Some basic assumptions could be strengthened with the development of urban transportation. For assumption (2), big cities in developing countries are promoting the construction of transit special lane network. This strategy is aimed at ensuring "Transit Priority" by providing successive transit right of way. For assumption (3), separation of different types of traffic flow is being advocated all over the world considering efficiency and safety issue. Transit operations are becoming increasingly steady. For assumption (4), with the devotion to transit priority at intersections, actuated signal control systems have a brilliant future. Anyway, the basic assumptions are aimed at guaranteeing a normal distribution of SST to the maximum extent. As an adequate alternative of assumption (4), bus delays at intersections must be very small.

Although it is assumed that SST conforms to normal distribution when bus links satisfy all the four basic assumptions, we could conduct distribution test to confirm the normal distribution assumption just like the case study in Nanjing, China. Only in a few cases may such SST at bus links not conform to normal distribution though there is no exception in Nanjing. When the assumption of normal distribution does not pass the test, the modified assignment method will lose its effect, and the method to predetermine time independence cannot be used in the whole urban transit network. However, we can focus on the subnetworks that SST

at all bus links are tested to conform to normal distribution, so the predetermination method is able to be adopted in these subnetworks. In this way, the results of predetermination could also act as part of evidence to decide the validity of some models.

Another assumption is made in which the regression equation between the mean and std of normal distribution in the Double Logarithmic Coordinate is significant. Nevertheless, this cannot be guaranteed even though all bus links satisfy those basic assumptions. To obtain high goodness of fitting, samples could probably be optimized by selecting locations. This measure is quite flexible, which can make successive adjustments in the whole network, and there is a high chance that a significant regression equation exists characterizing a transit network that satisfies the basic assumptions. Even if poor goodness of fitting remains unsettled, which is not likely to happen, our proposed assignment method will work all the same if all std of SST at bus links could be worked out through transit survey. Then, it is evident the transit network cannot be too large. But in this case, anyhow, our method to predetermine time independence will become invalid.

6. Conclusions

A simple method has been proposed to predetermine time independence of every two adjacent bus links, which can prejudge the validity of many stochastic network models for public transit. Then a modified capacity-restraint transit assignment method is put forward and aimed at engineering practice when such independence is predetermined to be invalid.

Some basic assumptions have been made to elicit two direct assumptions to the proposed methods: First, SST at bus links conform to normal distribution. Second, the regression equation between the mean and standard deviation has high goodness of fitting. These two assumptions are tenable according to the results in Nanjing. Then the indicator of time independence α is proposed to show the predetermination result. When $\alpha \geq 0.5$, SST at every two adjacent bus links interact with each other. When $0 < \alpha < 0.5$, time independence has potential validity that still needs proving. Meanwhile, α acts as the slope of the regression equation, which can reflect the time uncertainty in urban transit network. The modified assignment method is put forward, which appears effective when predetermination indicates the validity of time independence. We introduce the coefficient of assignment γ and 95% quantile of probabilistic distribution on the basis of the traditional capacity-restraint method. The traditional transit network with fixed SST at bus links is transformed into the network with normal distributed SST, which supports the modified method.

We take good advantage of the data acquired in Nanjing to conduct a case study. From the result of distribution test, all the seven bus links conform to normal distribution with a good significance level of 0.05. Then regression analysis is made in the Double Logarithmic Coordinate with a large R-square of 0.9731 reflecting its significance. The indicator of time independence or the regression coefficient is 1.1919 that is larger than 0.5. As a consequence, time independence is not valid in Nanjing, so many network models could not be used in the city. Researchers have to turn to other methods. Finally, the proposed capacity-restraint assignment method is preliminarily tested by a simple network. Compared with traditional capacity-restraint method, the modified method inherits the property of simplicity and has a better performance with more sensible results.

Although the findings in this paper are some of engineering practices, this work is limited by some issues. Firstly, it is very difficult to completely satisfy basic assumption (4) nowadays. Even if we turn to searching for signalized intersections with much small bus delays, it may have some influence on the normal distribution assumption of SST at bus links. Besides, the sample size in this study is somewhat limited for the huge workload of data acquisition though the proposed methodology is of theoretical basis. There are great numbers of two adjacent links in an urban transit network, which is not that suitable to acquire SST data manually.

Further research could focus on how to loosen the proposed assumptions. Even if normal distribution is invalid, others like Negative Binomial Distribution and others may replace it. Different distributions can be introduced to a network simultaneously, and accordingly the methods of predetermination and assignment might exist as well, which is probably complicated to be put into practice yet highly prospective. In addition, existing approaches could be introduced to have our assignment model well consider the "common lines" problem as with the development of previous heuristic capacity-restraint models. Related details are worth exploring. Moreover, the validity of the modified assignment method could be verified using real data of transit passenger flow before getting adopted by traffic engineers. Last but not least, work could be continued in the case $0 < \alpha < 0.5$, and it is expected that further conclusions concerning the time independence will be drawn in future.

Conflict of Interests

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

Acknowledgments

This research is funded by the National Key Basic Research Program of China (no. 2012CB725402), the Fundamental Research Funds for the Central Universities (no. KYLX_0171), and the Scientific Research Foundation of the Graduate School of the Southeast University.

References

[1] S. Chandra, M. E. Bari, P. C. Devarasetty, and S. Vadali, "Accessibility evaluations of feeder transit services," *Transportation Research Part A: Policy and Practice*, vol. 52, pp. 47–63, 2013.

[2] K. Sivakumaran, Y. Li, M. Cassidy, and S. Madanat, "Access and the choice of transit technology," *Transportation Research Part A: Policy and Practice*, vol. 59, pp. 204–221, 2014.

[3] S. Farber, M. Z. Morang, and M. J. Widener, "Temporal variability in transit-based accessibility to supermarkets," *Applied Geography*, vol. 53, pp. 149–159, 2014.

[4] L. Z. Liu, J. H. Yang, H. B. Mu, X. J. Li, and F. Wu, "Exact algorithms for multi-criteria multi-modal shortest path with transfer delaying and arriving time-window in urban transit network," *Applied Mathematical Modelling*, vol. 38, no. 9-10, pp. 2613–2629, 2014.

[5] Q. F. Li, P. Chen, and Y. Nie, "Finding optimal hyperpaths in large transit networks with realistic headway distributions," *European Journal of Operational Research*, vol. 240, no. 1, pp. 98–108, 2015.

[6] D. Heidemann, "Conditional logit model and the independence of irrelevant alternatives phenomenon," in *Proceedings of the Conference on Traffic and Transportation Studies (ICTTS '98)*, pp. 255–264, July 1998.

[7] H. C. Williams, "On the formation of travel demand models and economic evaluation measures of user benefit," *Environment and Planning A*, vol. 9, no. 3, pp. 285–344, 1977.

[8] E. Cascetta, A. Nuzzolo, F. Russo, and A. Vitetta, "A modified logit route choice model overcoming path overlapping problems: specification and some calibration results for interurban networks," in *Proceedings of the 13th International Symposium on Transportation and Traffic Theory*, pp. 697–711, Lyon, France, July 1996.

[9] K. S. Chan and W. H. K. Lam, "A stochastic traffic assignment model for estimating the variance of the path choice proportion," in *Proceedings of the 1st Conference of Hong Kong Society for Transportation Studies*, pp. 31–44, 1996.

[10] K. S. Chan and W. H. K. Lam, "Estimation of link flow variance by stochastic traffic assignment," in *Proceedings of the Transportation Systems Conference*, pp. 1241–1246, Chania, Greece, June 1997.

[11] C. F. Daganzo and Y. Sheffi, "On stochastic models of traffic assignment," *Transportation Science*, vol. 11, no. 3, pp. 253–274, 1977.

[12] E. T. Kenneth, *Discrete Choice Methods with Simulation*, University of California, Berkeley and National Economic Research Associates, 2003.

[13] Q. Meng and Z. Liu, "Mathematical models and computational algorithms for probit-based asymmetric stochastic user equilibrium problem with elastic demand," *Transportmetrica*, vol. 8, no. 4, pp. 261–290, 2012.

[14] K. Zhang, H. S. Mahmassani, and C.-C. Lu, "Dynamic pricing, heterogeneous users and perception error: probit-based bicriterion dynamic stochastic user equilibrium assignment," *Transportation Research Part C: Emerging Technologies*, vol. 27, pp. 189–204, 2013.

[15] S. D. Clark and D. P. Watling, "Sensitivity analysis of the probit-based stochastic user equilibrium assignment model," *Transportation Research Part B: Methodological*, vol. 36, no. 7, pp. 617–635, 2002.

[16] A. Dionísio, R. Menezes, and D. A. Mendes, "Entropy-based independence test," *Nonlinear Dynamics*, vol. 44, no. 1–4, pp. 351–357, 2006.

[17] M. Matilla-García and M. R. Marín, "A non-parametric independence test using permutation entropy," *Journal of Econometrics*, vol. 144, no. 1, pp. 139–155, 2008.

[18] F. López, M. Matilla-García, J. Mur, and M. R. Marín, "A non-parametric spatial independence test using symbolic entropy," *Regional Science and Urban Economics*, vol. 40, no. 2-3, pp. 106–115, 2010.

[19] W. A. Broock, J. A. Scheinkman, W. D. Dechert, and B. LeBaron, "A test for independence based on the correlation dimension," *Econometric Reviews*, vol. 15, no. 3, pp. 197–235, 1996.

[20] M. Sugiyama and T. Suzuki, "Least-squares independence test," *IEICE Transactions on Information and Systems*, vol. E94-D, no. 6, pp. 1333–1336, 2011.

[21] W. Wang and X. W. Chen, *Traffic Planning*, Southeast University, Nanjing, China, 2007.

[22] Y. Liu, J. Bunker, and L. Ferreira, "Transit userś route-choice modelling in transit assignment: a review," *Transport Reviews*, vol. 30, no. 6, pp. 753–769, 2010.

[23] H. J. Huang, *Urban Transportation Network Equilibrium Analysis: Theory and Practice*, Beihang University, Beijing, China, 1994.

[24] W. Wang and X. W. Chen, *Urban Transit Systems Planning and Management Techniques*, Southeast University, Nanjing, China, 2002.

[25] L. Zhang and J. F. Jia, "Study on generalized cost of passenger trip and product split of passenger transport," *Journal of Beijing Jiaotong University*, vol. 35, no. 3, pp. 68–71, 2011.

[26] L. Q. Yang, W. Huang, and N. Zhang, "Transfer preferential benefit of public transport based on generalized cost," *Journal of Transport Information and Safety*, vol. 27, no. 3, pp. 20–23, 2009.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 230616, 7 pages
http://dx.doi.org/10.1155/2015/230616

Research Article

Stability Analysis of Train Movement with Uncertain Factors

JingJing Ye,[1] KePing Li,[2] and XueDong Jiang[1]

[1]*School of Electrical Engineering, Beijing Jiaotong University, Beijing 100044, China*
[2]*State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing 100044, China*

Correspondence should be addressed to KePing Li; kpli@bjtu.edu.cn

Received 27 May 2015; Revised 22 July 2015; Accepted 11 August 2015

Academic Editor: Yuanchang Xie

We propose a new traffic model which is based on the traditional OV (optimal velocity) car-following model. Here, some realistic factors are regarded as uncertain quantity, such as the headway distance. Our aim is to analyze and discuss the stability of car-following model under the constraint of uncertain factors. Then, according to the principle of expected value in fuzzy theory, an improved OV traffic model is constructed. Simulation results show that our proposed model can avoid collisions effectively under uncertain environment, and its stability can also be improved. Moreover, we discuss its stability as some parameters change, such as the relaxation time.

1. Introduction

In the past 60 years, the theory of traffic flow has been gradually developed and improved. Now, it becomes the foundation which is used to explain complex traffic phenomena, solve some difficult traffic problems, and increase the efficiency of traffic system. A number of theoretical calculations, numerical simulations, and empirical observations are reported. Even so, the theory of traffic flow still faces many great challenges of some realistic traffic problems.

In the theory of traffic flow, car-following model is an important microscopic simulation model which has been widely extended and applied. Since the car-following model uses realistic driver behavior and detailed vehicle characteristics, it can be used to simulate various traffic phenomena which are observed in realistic traffic [1–3], such as soliton wave, shock wave, and kink wave. A famous car-following model is optimal velocity model which was proposed by Bando et al. in 1995 [4]. Afterwards, many improved and extended car-following models were reported which were based on Bando's model. For example, Helbing and Tilch found that collision would occur if accelerating rate is too large; then they derived general force model (GF) [5]. Jiang et al. studied start process of static motorcade by GF and proposed full velocity difference model [6]. Li et al. studied stabilization strategies of relaxation time of driver based on general nonlinear car-following model [7].

Usually, in railway traffic, there are two kinds of block modes for controlling the train movement: the fixed block and moving block modes. With the fixed block mode, the train movement is controlled by signal lights. However, with the moving block mode, one train runs following its leading train with the help of wireless communication. Moving block mode which is based on communication is the most advanced signal control system in the world. At present, it is extensively used in the urban railway traffic and high-speed railway traffic. With the moving block mode, train runs at higher speed and smaller time gap. Figure 1 displays the principle of train movement with moving block mode.

The key step with moving block mode is about the safety of train movement. In Figure 1, the safety of train movement is determined by headway distance between leading train 2 and following train 1:

$$\begin{aligned} L_A &= L_B + L_Z, \\ L_B &= \tau v + \frac{v^2}{(2b_{br})}. \end{aligned} \tag{1}$$

Here, L_A is the safe headway distance, τ is relaxation time, v is the velocity of train movement, and b_{br} is the maximum

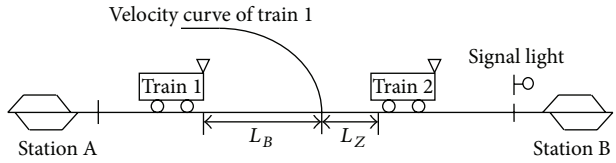


Figure 1: The Moving Block Principle.

deceleration. L_Z is a margin which cannot be less than the length of whole train.

In realistic traffic, vehicle's movement is under the constraint of many uncertain factors, such as weather, driver's judgment, and passenger demand. Recently, such a problem attracts more and more attention. Zheng et al. studied the urban pickup and delivery problem in which the velocities of vehicles were regarded as fuzzy and uncertain [8]. In 2010, Chen and Zhou proposed a new model called the reliable mean-excess traffic equilibrium model, in which it reflected driver's risk preferences under uncertain environment [9]. Zheng and Van Zuylen proposed a probabilistic delay distribution model with stochastic arrivals and departures [10]. In railway traffic, the passenger demand and waiting time usually were considered as uncertain factors, which were treated as fuzzy variables [11]. Wei et al. proposed location-scheduling model based on genetic algorithm, where transportation risk was considered as time-dependent fuzzy random variables [12]; Yang et al. studied credibility-based rescheduling model in a double-track railway network, where low-probability incident was taken as uncertain factor [13].

In present work, the headway distance of vehicle is regarded as fuzzy variable. Based on traditional OV car-following model, we propose a new OV car-following model. Since we have accounted for some uncertain factors, our proposed model is more realistic than traditional model. This paper is organized as follows. In Section 2, traditional OV car-following model is introduced, our proposed model is proposed in Section 3, and simulation results are shown in Section 4. Finally, a conclusion is drawn in Section 5.

2. Optimal Velocity Car-Following Model

In principle, car-following model is a kind of response-stimulus model, in which a follower tries to maintain a space gap with its leader. Early car-following models were proposed by Reuschel and Pipes [14, 15]. In order to account for the time lag, in 1958, Chandler et al. suggested an improved car-following model [16]:

$$\ddot{x}_n(t+T) = \lambda\left[\dot{x}_{n+1}(t) - \dot{x}_n(t)\right], \tag{2}$$

where T is a response time lag and λ is the sensitive coefficient. For a more realistic description, Newell presented the optimal velocity car-following model [17]:

$$\ddot{x}_n(t+T) = V^{\text{opt}}\left(\Delta x_n(t)\right), \tag{3}$$

where the function V^{opt} is $V^{\text{opt}}(\Delta x_n(t)) = v_0(1-\exp[-(\Delta x_n - x_c)/(v_0 T_f)])$, Δx_n, x_c, and v_0 are the headway distance, safe distance, and the desired velocity, respectively, and T_f is a safe time interval characterizing the car-following behavior.

In 1995, Bando et al. proposed a famous OV car-following model which is written as follows [4]:

$$\ddot{x}_n(t) = \frac{1}{\tau}\left(V^{\text{opt}}\left(\Delta x_n(t)\right) - v_n(t)\right). \tag{4}$$

Here, V^{opt} is the desired optimal velocity function. τ is the relaxation time. The desired optimal velocity of the nth vehicle V_n^{opt} is derived as follows:

$$V_n^{\text{opt}} = \frac{v_{\max}}{2}\left\{\tanh\left[\Delta x_n(t) - S_m\right] + \tanh\left(S_m\right)\right\}. \tag{5}$$

Here, $v_{\max}$ is the maximum velocity of vehicles, and $\Delta x_i(t)$ is the distance between the ith vehicle and the $(i+1)$th vehicle. S_m is called the minimum safety distance.

In fact, the OV car-following model belongs to the response-stimulation equation. The stimulation is the difference between the velocity of vehicle and optimal velocity. In the past decades, OV car-following model attracts more and more concerns because not only it is simple in numerical calculation, but also it can be analyzed by using iteration equation. For instance, the kink-antikink waves can be obtained.

3. Our Proposed Model

Usually, from road to railway and civil aviation, vehicles encounter many disturbances of uncertain factors, for example, the disturbance of weather, and the error of driver's judgment. To a larger extent, these uncertain factors directly affect vehicles' movement. In existing research reports, the influences of these factors are always ignored, especially in the railway traffic. This directly leads to the fact that the constructed simulation model or mathematical computations are not consistent with some realistic observations. Furthermore, some theoretical results using existing models or methods have unrealistic defects.

In our proposed model, the headway distance Δx_n is considered as a fuzzy variable. In railway traffic, the headway distance Δx_n is actually the following gap, and it can be obtained by calculating the distance between two successive trains. In general, the position of train must be detected and transmitted to control center uninterrupted. Because of the errors caused by the detection devices, environments, or communication system, the position of train cannot be obtained accurately. This makes the headway distance (i.e., the following gap) Δx_n be an uncertain value. In this paper, Δx_n is assumed to be a random value. It is divided into two parts: one part is actual value Δx_{na} and the other is error x_{er}.

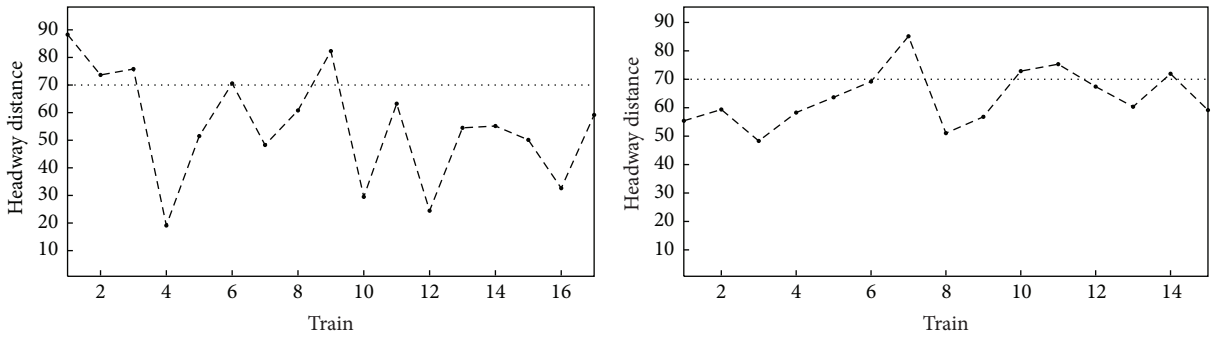


FIGURE 2: The distribution of headway distance for (a) $d_m = 60$ and (b) $d_m = 20$.

Because it is difficult to predict error value, we take x_{er} as a random value, and it is calculated by $x_{\text{er}} = d_m n_r$; here, n_r is a random number. So, Δx_n can be taken as $\Delta x_n = \Delta x_{na} + x_{\text{er}}$. Here, Δx_{na} represents the accurate headway distance between the nth vehicle and the $(n+1)$th vehicle. $x_{\text{er}} = d_m n_r$, d_m represents the maximum error of headway distance, and n_r is a random number in the region $[-1\ 1]$.

In this paper, we apply expected value principle in fuzzy theory to deal with the fuzzy variable Δx_n. Firstly, let x_{er} change randomly for N_m times from the formula of (5), so that N_m values of V^{opt} can be obtained. Averaging these N_m values of V^{opt}, a mean value V_{ra} is given. Repeating the above step for N_{nm} times, and then averaging N_{nm} values of V_{ra}, a desired optimal velocity V_{rp} is obtained.

At time t, the executed process of the proposed model can be described as follows.

Step 1. Initial parameters are $N_{rm} = 0$ and $V_{rm} = 0$.

Step 2. Initial parameters are $N_r = 0$ and $V_r = 0$.

Step 3. Let x_{er} change randomly: $N_r = N_r + 1$.

Step 4. According to (5), calculating the optimal velocity V^{opt}, $V_r = V_r + V^{\text{opt}}$.

Step 5. If $N_r < N_m$, go to Step 3; else $V_{ra} = V_r/N_m$, $N_{rm} = N_{rm} + 1$, and $V_{rm} = V_{rm} + V_{ra}$.

Step 6. If $N_{rm} < N_{nm}$, go to Step 2; else $V_{rp} = V_{rm}/N_{nm}$.

In our method, the boundary condition is open. From the site $x = 1$ to the site $x = S_m$, if there is no vehicle with the region $[0, S_m]$, a vehicle with the velocity $v_n = 0$ is created at the site $x = 1$. The newborn vehicle immediately moves according to the proposed model. At the site L, vehicles simply move out of the simulated system. L is the length of the considered simulated system.

4. Numerical Computation

We use the proposed OV car-following model to simulate the evolution of traffic flow in single-lane railway traffic. Here, all vehicles must move and only the minimum safety distance S_m is kept among them. In railway traffic, the minimum safe distance is calculated by $S_m = \tau V_{\max} + {V_{\max}}^2/2b_{\max} + \Delta x_{\text{mag}}$. Here, τ is relaxation time, $V_{\max}$ is maximum speed, $b_{\max}$ is the maximum deceleration, and Δx_{mag} is margin which is fixed. According to this formula, when $V_{\max}$ and $b_{\max}$ are constant, S_m is constant too. The length of the simulated system is set to be $L = 2000$, and the maximum iteration time step is $T = 1000$. A station is designed at the middle site of the simulated system. At the station, all vehicles should stop for the same time T_d and then leave. T_d is called the dwell time (red light time). The parameters $v_{\max}$, S_m, τ, and T_d are taken as 10, 70, 2.5, and 5, respectively.

In reality traffic, the safety factor is very important in which collisions between two vehicles must be avoided. This means that the distance between two successive vehicles must be larger than the minimum safety distance. In order to investigate the evolution of traffic flow under uncertain environment, we use the proposed OV car-following model to simulate the vehicle's movements. The uncertain environment is completed by $\Delta x_n = \Delta x_{na} + x_{\text{er}}$. After sufficient transient, we begin to record the traffic series of the headway distance at a given time t, $\{D_i\}$, where D_i is the distance from the ith vehicle to the $(i+1)$th vehicle. Figure 2 shows the distribution of the headway distance D_i at $t = 600$. In Figure 2, the minimum safety distance S_m is indicated by the dotted line, and the simulation results are drawn by the dash-dot line. Figure 2(a) displays the fact that the distribution of headway distance is less than S_m, and most of headway distances are within the region $[20, 60]$. When d_m is equal to 20, the distribution of headway distance shown in Figure 2(b) is larger than 50. Comparing Figure 2(a) with Figure 2(b), we can conclude that as d_m is larger, the probability of collisions among trains is higher. In this case, the safety of train becomes worse.

Applying excepted value principle, the train's velocity can be optimized. So we use such a method to deal with the vehicle's velocity under uncertain condition. The simulation environment is the same as that adopted in Figure 2(a). Figure 3 shows the distributions of optimized headway distance under $d_m = 60$. In Figure 3, we can see that although most of the simulation results are still smaller than S_m, the

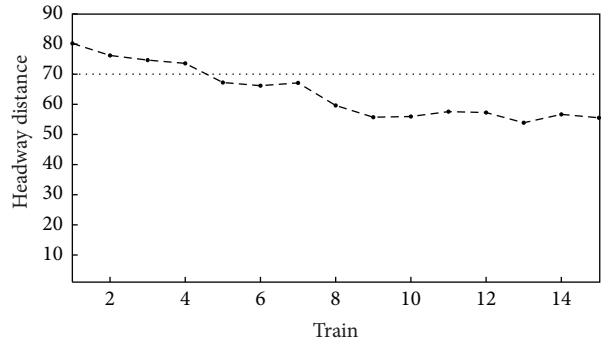


Figure 3: The distribution of headway distance under the excepted value principle for $d_m = 60$.

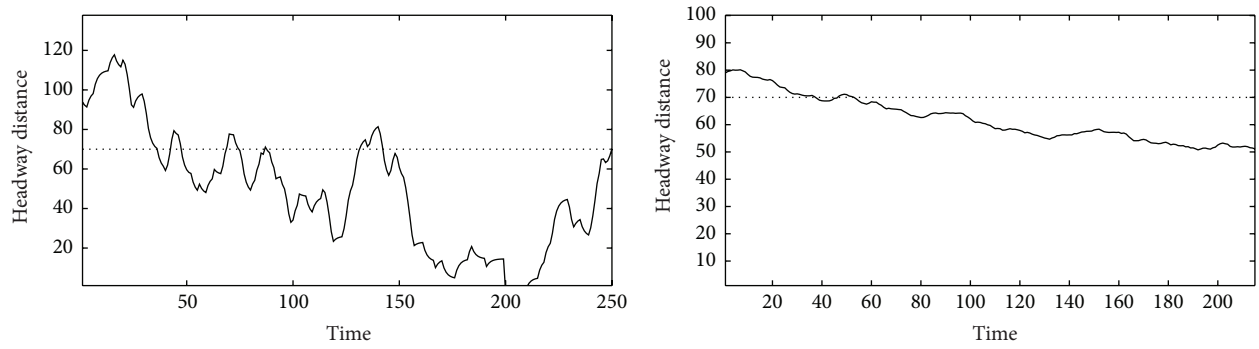


Figure 4: The distribution of headway distance of a single tracked train for $d_m = 60$ (a) without the excepted value principle and (b) with the excepted value principle.

values of headway distance vary slowly within the region [50, 80]. Compared to Figure 2(a), it is obvious that our proposed model which uses the expected value principle can effectively improve the safety of train.

In order to further study the safety of train, the headway distance $D(t)$ of one tracked train is measured. Here, $D(t)$ is the distance from the tracked train to its leading train at time t. Figure 4 shows the simulation results of the tracked train which departs at the time $t = 600$. From Figure 4(a), it is clear that most of distance $D(t)$ is smaller than the minimum safety distance. Moreover, within some time intervals, the headway distance of tracked vehicle is zero. In this case, collision occurs. The result demonstrates that train travelling is very dangerous under the uncertain condition. Figure 4(b) shows the headway distance $D(t)$ of the tracked train. Here, although $D(t)$ is still smaller than the minimum safety distance, they are larger than 50. Moreover, it also shows that the values of $D(t)$ change slowly. In this case, the travelling comfort is also improved.

In the theory of traffic flow, the stability of traffic model is a major character. In general, the stability of traffic model is described by the variation amplitude of the headway distance which is got by calculating the difference between the maximum value and average value of the headway distance. Figure 5(a) shows the variation amplitude of a tracked train at all time under the uncertain environment. Here the variation amplitude of the headway distance is within the region [0, 50]. However, from Figure 5(b), the amplitude of the headway distance varies within the region [0, 20], and most of the values are less than 10. Numerical results indicate that not only the collision can be avoided, but also the fluctuation of amplitude is small compared with Figure 5(a). This means that the stability of the tracked train simulated using our proposed car-following model is improved.

In order to further study the stability character of train travelling, the variation amplitude of headway distance of all trains at a given time is measured. Figure 6 shows the results in which the variation amplitude of headway distance is obtained at $t = 600$. The variation amplitude of the headway distance using the excepted value principle shown in Figure 6(b) is small and slow, and it is less than that shown in Figure 6(a). In terms of the simulation results which are shown in Figure 6, we can conclude that train travelling is not stable under the uncertain factor, but the situation can be improved using our proposed model where the expected value principle is used.

We study the stability of one tracked train. The variation acceleration of the tracked train which departs at $t = 600$ is measured. Figure 7(a) shows that the acceleration of the tracked train varies sharply and frequently. It demonstrates that the tracked train cannot obtain stable travelling. From Figure 7(b), the amplitude of the acceleration is within a small

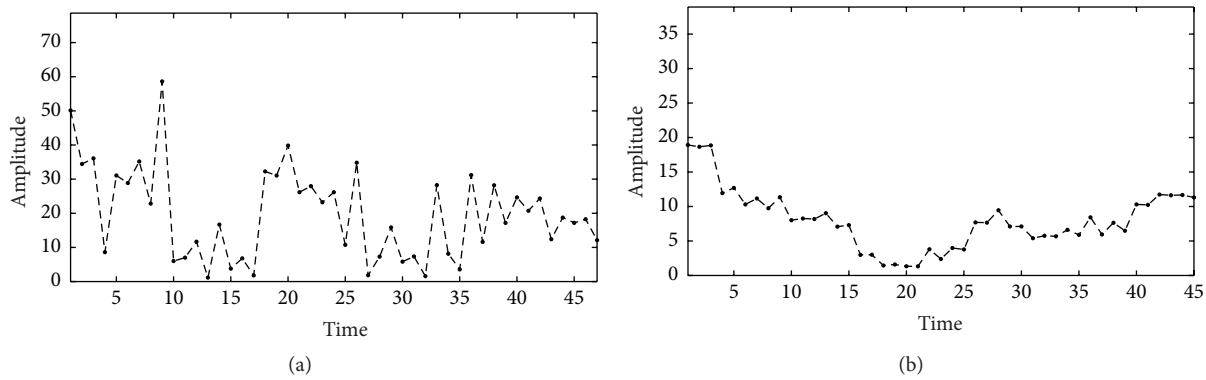


Figure 5: The variation amplitude of headway distance of a tracked train for $d_m = 60$ (a) without the excepted value principle and (b) with the excepted value principle.

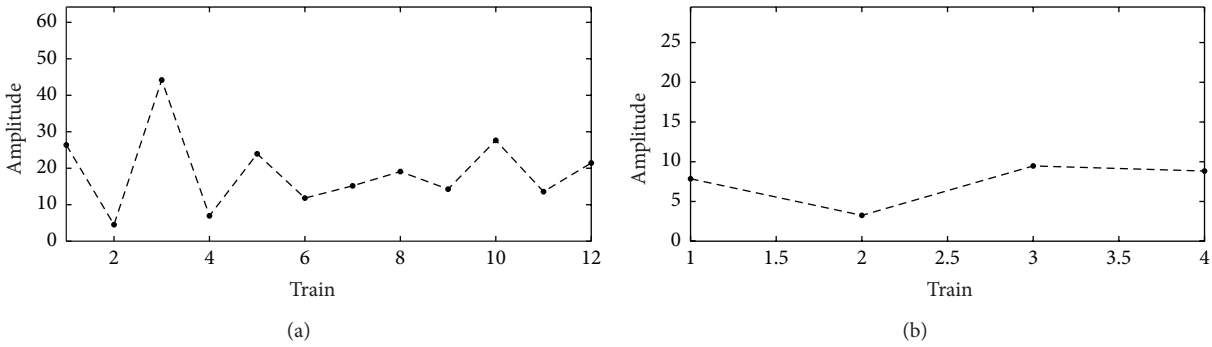


Figure 6: The variation amplitude of headway distance of all trains for $d_m = 60$ (a) without the excepted value principle and (b) under the excepted value principle.

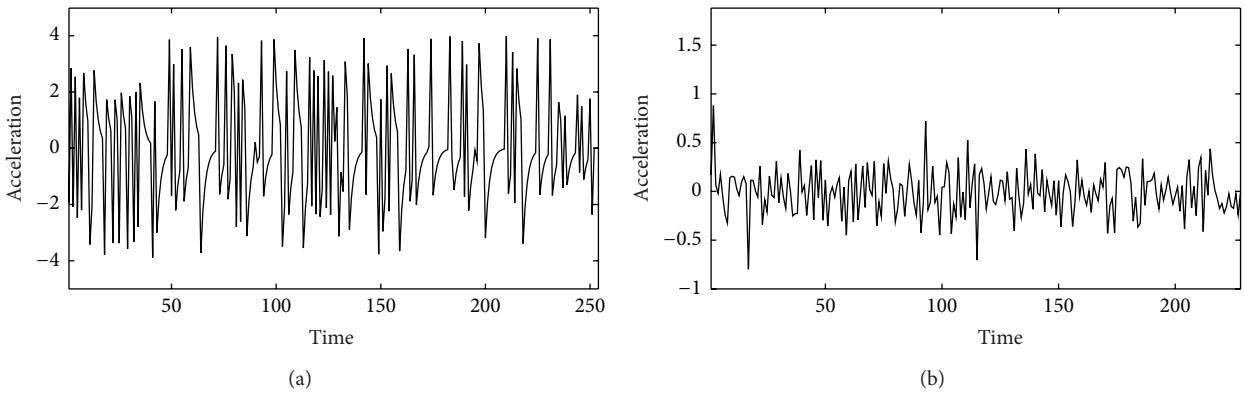


Figure 7: How the acceleration varies with the time for $d_m = 60$ (a) without the excepted value principle and (b) with the excepted value principle.

region $[-1, 1]$. Most of the measured values are far smaller than 0.5. This means that the driver only needs to adjust the velocity of the tracked train simulated using our proposed model, and the stability of train travelling can be enhanced.

Relaxation time τ is one of the major parameters of car-following model which describes the response extent of the driver. In general, the driver's response greatly affects the train travelling. In theory, as the relaxation time is larger, the response of driver is slower. Figure 8 shows the variation amplitude of headway distance of one tracked train for different relaxation time. Here, the headway distance is uncertain. Comparing Figure 8(a) with Figure 8(b), it is clear that the values of amplitude shown in Figure 8(a) are larger than that shown in Figure 8(b). It indicates that the stability of train traveling at $\tau = 1$ is better than $\tau = 4$. The simulation results are consistent with some observation results.

When the excepted value principle is applied to our proposed car-following model, how the variation amplitude changes with the relaxation time varies. Figure 9 shows the variation amplitude of one single tracked train for different

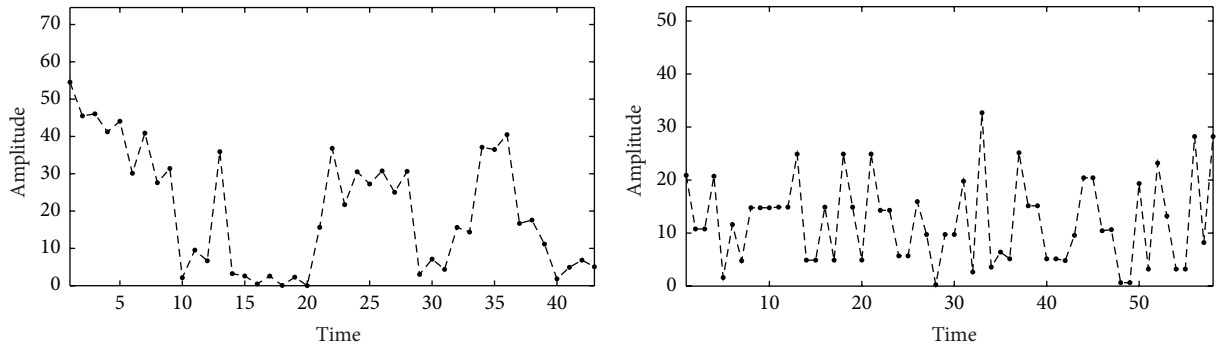


Figure 8: The variation amplitude of headway distance of one tracked train for $d_m = 60$. (a) $\tau = 4$ and (b) $\tau = 1$.

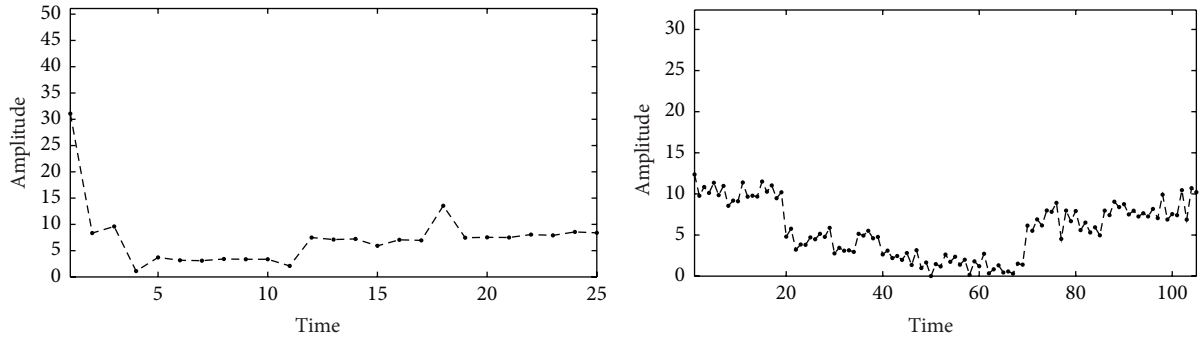


Figure 9: The variation amplitude of headway distance under the excepted value principle for (a) $\tau = 4$ and (b) $\tau = 1$.

time τ, where the excepted value principle is used. Here we can see that there is no obvious oscillation which is similar to that shown in Figure 8. Moreover, we observe that the overall values of Figure 9(b) are smaller than that shown in Figure 9(a). The results further demonstrate that the quick response of driver contributes to the stability of train travelling.

In Figures 8 and 9, we find that even if one train has a greater acceleration, the stability of system consisting of many trains is good. The reason is that, in our paper, the stability of train movement is described by the difference among the headway distances of tracked trains. So the stability is used to investigate the systematic character of many trains.

5. Conclusion

In summary, we use an extended OV car-following model to simulate the evolution of traffic flow under uncertain condition. Here, the headway distance of train is regarded as fuzzy variable. The simulation results demonstrate that not only the disturbance of uncertain factors can be overcome effectively, but also the stability character of our proposed model can be improved.

We also discuss that the relaxation time affects the stability based on the proposed model. The stability is improved with the shortening of the relaxation time. Because optimizing of the excepted value principle weakens the influence of error headway distance, the improvement of stability under the excepted principle is not very obvious.

Here, we only take a simple uncertain factor into account; however, some complex factors, such as the resistance force of vehicle movement and the disturbance of wind and sand, are not considered. Even so, we provide a good way to handle the uncertain factor in the theory of traffic flow. We will do them in our future works. Moreover, our results are useful for optimizing the design and plan of urban traffic system.

At present, there are some better extensions of the OV model, such as the FVD [18] model and MMVD model [19]. Although they are mainly applied for the road traffic, they may be good for solving some railway traffic problems.

Conflict of Interests

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

Acknowledgment

This work was supported by the Fundamental Research Funds for the Central Universities (no. 2014JBM109).

References

[1] T. Komatasu and S. Sasa, "Kink solution charactering traffic congestion," *Physical Review E*, vol. 52, pp. 5574–5581, 1995.

[2] M. Muramatsu and T. Nagatani, "Soliton and kink jams in traffic flow with open boundaries," *Physical Review E*, vol. 60, no. 1, pp. 180–187, 1999.

[3] H. X. Ge, S. Q. Dai, and L. Y. Dong, "An extended car-following model based on intelligent transportation system application," *Physica A: Statistical Mechanics and Its Applications*, vol. 365, no. 2, pp. 543–548, 2006.

[4] M. Bando, K. Hasebe, K. Nakanishi, A. Nakayama, A. Shibata, and Y. Sugiyama, "Phenomenological study of dynamical model of traffic flow," *Journal de Physique I*, vol. 5, no. 11, pp. 1389–1399, 1995.

[5] D. Helbing and B. Tilch, "General force model of traffic dynamics," *Physical Review E*, vol. 58, no. 1, pp. 133–138, 1998.

[6] R. Jiang, Q. S. Wu, and Z. J. Zhu, "Full velocity difference model for a car-following theory," *Physical Review E*, vol. 64, no. 1, Article ID 017101, 2001.

[7] S. K. Li, L. X. Yang, Z. Y. Gao, and K. P. Li, "Stabilization strategies of a general nonlinear car-following model with varying reaction-time delay of the drivers," *ISA Transactions*, vol. 53, no. 6, pp. 1739–1745, 2014.

[8] S. F. Zheng, J. D. Cao, X. M. Lian, and K. P. Li, "Urban pickup and delivery problem considering time-dependent fuzzy velocity," *Computers & Industrial Engineering*, vol. 60, no. 4, pp. 821–829, 2011.

[9] A. Chen and Z. Zhou, "The α-reliable mean-excess traffic equilibrium model with stochastic travel times," *Transportation Research Part B: Methodological*, vol. 44, no. 4, pp. 493–513, 2010.

[10] F. Zheng and H. Van Zuylen, "Uncertainty and predictability of urban link travel time: delay distribution-based analysis," *Transportation Research Record*, vol. 2192, pp. 136–146, 2010.

[11] L. X. Yang, K. P. Li, and Z. Y. Gao, "Train timetable problem on a single-line railway with fuzzy passenger demand," *IEEE Transactions on Fuzzy Systems*, vol. 17, no. 3, pp. 617–629, 2009.

[12] M. Y. Wei, X. Li, and L. Yu, "Time-dependent fuzzy random location-scheduling programming for hazardous materials transportation," *Transportation Research Part C: Emerging Technologies*, vol. 57, pp. 146–165, 2015.

[13] L. X. Yang, X. S. Zhou, and Z. Y. Gao, "Credibility-based rescheduling model in a double-track railway network: a fuzzy reliable optimization approach," *Omega*, vol. 48, pp. 75–93, 2014.

[14] A. Reuschel, "Fahrzengbewegungenin der kolonne oesterr," in *Der Österreichische Ingenieur- und Architekten-Verein*, vol. 4, pp. 193–215, 1950.

[15] L. A. Pipes, "An operational analysis of traffic dynamics," *Journal of Applied Physics*, vol. 24, no. 3, pp. 274–281, 1953.

[16] R. E. Chandler, R. Herman, and E. W. Montroll, "Traffic dynamics: studies in car following," *Operations Research*, vol. 6, no. 2, pp. 165–184, 1958.

[17] G. F. Newell, "Nonlinear effects in the dynamics of car following," *Operational Research*, vol. 9, no. 2, pp. 209–229, 1961.

[18] R. Jiang, Q. S. Wu, and Z. J. Zhu, "Full velocity difference model for a car-following theory," *Physical Review E*, vol. 64, no. 1, Article ID 017101, 4 pages, 2001.

[19] T. Wang, Z. Y. Gao, and X. M. Zhao, "Multiple velocity difference model and its stability analysis," *Acta Physica Sinica*, vol. 55, no. 2, pp. 634–640, 2006.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 328712, 12 pages
http://dx.doi.org/10.1155/2015/328712

Research Article

CTM Based Real-Time Queue Length Estimation at Signalized Intersection

Shuzhi Zhao, Shidong Liang, Huasheng Liu, and Minghui Ma

College of Transportation, Jilin University, Changchun 130022, China

Correspondence should be addressed to Huasheng Liu; liuhuasheng521@163.com

Received 28 May 2015; Revised 29 July 2015; Accepted 6 August 2015

Academic Editor: Yuanchang Xie

Queue length is an important index of the efficiency of urban transport system. The traditional approaches seem insufficient for the estimation of the queue length when the traffic state fluctuates greatly. In this paper, the problem is solved by introducing the Cell Transmission Model, a macroscopic traffic flow, to describe the vehicles aggregation and discharging process at a signalized intersection. To apply the model to urban traffic appropriately, some of its rules were improved accordingly. Besides, we can estimate the density of each cell of the road in a short time interval. We, first, identify the cell, where the tail of the queue is located. Then, we calculate the exact location of the rear of the queue. The models are evaluated by comparing the estimated maximum queue length and average queue length with the results of simulation calibrated by field data and testing of queue tail trajectories. The results show that the proposed model can estimate the maximum and average queue length, as well as the real-time queue length with satisfactory accuracy.

1. Introduction

It has been generally recognized that the vehicular queue length is a crucial quantitative measure used in the evaluation of the performance of signalized intersections [1–3] and signal optimization [4–9]. Due to the importance of this traffic parameter, many scholars studied this topic, and their proposed models can be classified into three types: input-output accumulation curve methods, statistical methods, and the shockwave theory.

The first type models are based on the construction of accumulation curves by means of analysis of vehicles input-output to a signal link, where vehicles between two curves are accounted for as being in a queue. The method was first demonstrated by Webster [4] and later improved by many scholars [5, 10–12]. Although such models are widely used to calculate the queue length, it is hard to obtain the real-time queue length. In addition, the accumulation input-output curve is difficult to structure under complex traffic conditions. Therefore, such methods are limited in their ability to describe complex queuing processes.

According to statistics methods, the vehicle arrival and release processes are regarded as some certain probability distributions, where the queue length is the mathematical expectation of the difference between the numbers of arriving and released vehicles [13]. Usually, such methods are used for steady-state traffic flows, because the distribution should change in different traffic states. Commonly, it is assumed that the arrival distribution follows the Poisson distribution [14] under free flow and negative binomial distribution under saturated flow [15]. In other words, the robustness of such models is not strong enough, as they cannot adapt to the situation when traffic state changes irregularly, such as in the real world.

The shockwave theory is another powerful tool used in queue length estimation, which was proposed by Lighthill and Whitham [16, 17] and Richards [18] to describe the queuing process of uninterrupted traffic flows on freeways. Later, Stephanopoulos et al. [19] expanded the theory for continuum traffic flows at signalized intersections. The validity of this method was proved by many scholars [20, 21]. These elegant theoretical models only perform well in the situation when the traffic flow can be regarded as a steady flow macroscopically. For example, such models assume that vehicle flows transform from one state to another instantaneously.

Given these observations, the common deficiencies of current methods can be described as their inability to obtain the real-time spatial distribution of queue lengths when traffic flows constantly change. To solve this problem and describe traffic flows at signalized intersections, a discrete macroscopic traffic flow model, the Cell Transmission Model (CTM), is introduced in this paper. The model was proposed by Daganzo [22]. This method can be used to describe traffic changes over time and space, including transient phenomena such as building, propagation, and dissipation of queues. The CTM can overcome computational difficulties caused by the substitution of difference equations for partial differential equations. The method automatically generates appropriate changes in the density at locations where the hydrodynamic theory produces shockwaves and sudden changes in the density, those typically seen at the end of a queue. Therefore, complex calculations required by classical methods are eliminated.

The CTM of motorway traffic is a discrete version of the simple continuum model of traffic flow that is convenient for computer implementations. The first order hydrodynamic model of traffic flow can be programmed using cell transmission scheme sending and receiving functions. Muñoz et al. [23] proposed a semiautomated method to calibrate the parameters of the model. According to the authors, a least-square fitting approach is applied to the loop detector data to determine free-flow and congestion speeds, as well as jam densities for specific subsections of a freeway segment. Bottleneck capacities are estimated by measuring mainline and onramp flows. Sun et al. [24] applied the mixture Kalman filtering to the CTM. Their method is able to estimate vehicle densities and congestion statuses at locations, which are not directly observed. The program runs efficiently, making it possible to carry out estimations in real time. Tampère and Immers [25] proposed another traffic estimation and prediction model based on the CTM. In their model, they provided parameters for automatic estimation of traffic conditions. Hadiuzzaman and Qiu [26] utilized the CTM in variable speed limit on freeways. They developed an analytical model based on the CTM capable of describing active bottlenecks, where the capacity drops once the incoming flow exceeds the capacity, and variable free-flow speeds in cells operated using VSL control.

Kim and Keller [27] and Pohlmann and Friedrich [28] introduced the CTM to the problem of traffic control in urban networks. They use the CTM to model the effects of different signal settings. The model is capable of adapting every 15 minutes to the currently estimated traffic and optimizing signal plans and coordination according to the demand in the network. Xie et al. [29] formalized the description of the urban CTM and evaluated its capability to model traffic volumes and jams using a microscopic traffic simulator. The CTM was further extended by Chiou and Huang [30] to describe hybrid traffic flows on urban roads. In their work, the mixed traffic CTM is proposed to replicate traffic behavior on Asian urban streets, where mixed traffic of cars and motorcycles is prevailing. The mixed CTM uses the ratio of cars to motorcycles in the last upstream cell to determine the amount of roadway resources allocated to cars and motorcycles.

Therefore, the advantages of the CTM can be summarized as follows:

(1) Satisfactory performance for description of traffic phenomena such as the queuing process and traffic wave propagation.

(2) Relative simplicity for modeling, compared to traditional traffic flow models, which is convenient for applications using computer programs.

(3) Flexibility for applications, since the traffic parameters can be estimated for both online and offline, which ensures that the queue length can be estimated in time.

The objective of this paper is to provide a strong approach to real-time queue length estimation. We focus on obtaining the maximum and average queue length within a signal cycle and tracking the queue tails trajectories. The performance of the proposed methodology partly relies on the traffic flow theory used for describing queuing processes at signalized intersections. Therefore, the CTM was tested to check its adaptation to traffic flows at signalized intersections. Accordingly, some rules were modified to adapt the original model to urban roads. Besides, the density plays an important role in queue length estimation within a discrete time interval. The density in each cell is analyzed to identify the queue tail cell and eventually calculate the exact position in this cell. It is worth mentioning that the fixed detectors of traditional signal control systems (such as SCOOT) can provide the data required by the proposed model.

The rest of the paper is organized as follows. In Section 2, we introduce the Cell Transmission Model, test it using numerical simulations, and modify it accordingly. In Section 3, we present an estimation mode of queue length and the general formula for queue length calculation. We also discuss three special cases. In Section 4, we show the testing results, including the comparison of the average and maximum queue lengths obtained in estimations and their corresponding real values, and the results on the trajectories of queue tails. Section 5 discusses some limitations of our approach and presents future work.

2. Proposed Model for Traffic Flow Description

The traditional Cell Transmission Model partitions a one-way road into several, say I, homogenous cells. The length of each cell is equal to the distance traveled by free-flowing traffic in a time interval T. So, in the case of free flow, all the vehicles travelling at the maximum speed can move from one cell to the next one in the unit time interval. In each cell i during time period k, the density is approximated by $\rho_i(k)$, and the exit flow $q_i(k)$ is expressed as the minimum between the upstream demand and the downstream supply.

To describe the traffic flow at a signalized intersection, two modifications in the model are proposed. The first modification concerns the rules for the end downstream cell

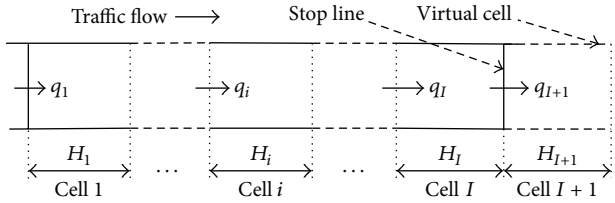


Figure 1: Transmission mechanism.

during the red and green phases separately. The second one modifies the rules for the discharging process of the queue based on numerical simulation analysis.

2.1. Density Dynamics. To apply the CTM in the case of traffic flows at signalized intersections, we assume that the road, from the loop detector upstream to the stop line downstream, is divided into homogeneous cells, and the length of each cell is defined as H_i. Figure 1 illustrates the vehicle transmission mechanism.

In period $k+1$, the number of vehicles in cell i, $n_i(k+1)$, decreases by the number of exiting vehicles and increases by the number of entering vehicles, which can be written as

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

where $y_i(k)$ represents the number of vehicles travelling into cell i from k to $k+1$. The number of vehicles in cell i and travelling between two cells can be represented using basic traffic parameters, the density and traffic flow, respectively:

$$n_i(k) = \lambda \cdot \rho_i(k) \cdot H_i, \tag{2}$$

$$y_i(k) = q_i(k) \cdot T, \tag{3}$$

where $\rho_i(k)$ represents the traffic density in cell i for each lane, from $k-1$ to k; and $q_i(k)$ represents the traffic volume from cell $i-1$ to cell i in the last time interval for each lane. Here, λ means the number of lanes; H_i refers to the length of the cell $v_f \cdot T$.

Using (1), (2), and (3), we can derive the relationship between $\rho_i(k)$ and $q_i(k)$, with the density dynamics expressed by

$$\rho_i(k+1) = \rho_i(k) + \frac{T}{\lambda \cdot L_i}\left(q_i(k) - q_{i+1}(k)\right). \tag{4}$$

The traffic flow between two adjacent cells depends on the space in the downstream cell and demand of the upstream cell, as well as the capacity of each cell $Q_{\max}$. In the traditional CTM, traffic running in each cell is characterized by the FD shown in Figure 2. In the linear fundamental diagram, from the origin to point A', the traffic flow is free, and vehicles travel at maximum velocity. Between points A' and A, traffic becomes congested, and the traffic volume stays at the level of capacity, instead of increasing along with the growing density. From point A, the traffic flow switches to a jam, and the shockwave starts to propagate to the upstream.

The interpretation of the demand and supply functions can be illustrated as follows. If the cell density $\rho_i(k)$ is

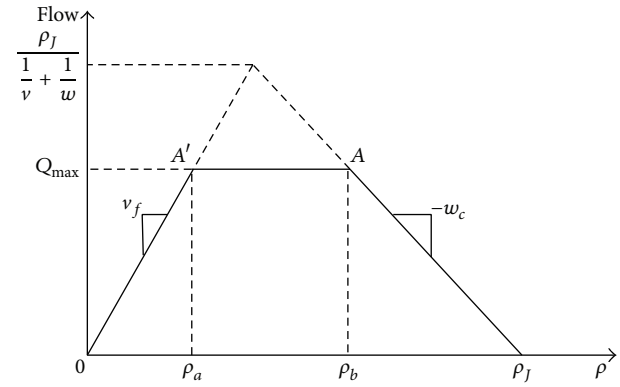


Figure 2: Fundamental diagram of the Cell Transmission Model.

lower than the density of ρ_a, cell i demands space from its downstream cell $(i+1)$ for a flow of $v_f \cdot \rho_i(k)$, where v_f means the velocity of the free flow. Otherwise, not all vehicles can travel to the downstream cell even at the saturation flow rate, because of the limitation of capacity. The demand for space from the upstream cell $(i+1)$ for a flow of $\phi_i(k)$ is expressed as

$$\phi_i(k) = \begin{cases} \lambda \cdot v_f \cdot \rho_i(k), & \text{if } \rho_i(k) < \rho_a, \\ Q_{\max}, & \text{if } \rho_i(k) \geq \rho_a. \end{cases} \tag{5}$$

If the spare space in cell i is enough for the coming vehicles from cell $i-1$, it provides space to the upstream for saturated traffic flow. Alternatively, if the shockwave occurs, the supplied space will be less than the traffic capacity. The supply for the upstream cell for a flow of $\varphi_i(k)$ can be written as follows:

$$\varphi_i(k) = \min \begin{cases} Q_{\max}, \\ \lambda \cdot w_c \cdot \left(\rho_J - \rho_i(k)\right). \end{cases} \tag{6}$$

Having the supply and demand function, we can calculate the traffic volume in those cells using the following function:

$$\begin{aligned} q_i(k) &= \min\left\{\phi_{i-1}(k), \varphi_i(k)\right\} \\ &= \min\left\{v_f \cdot \rho_{i-1}(k), Q_{\max}, w_c \cdot \left(\rho_J - \rho_i(k)\right)\right\}. \end{aligned} \tag{7}$$

Now, knowing the density and flux values in the given time interval, using (7), the density evolution in all cells can be estimated using (4). The proposed traffic flow model can be implemented for any number of cells in the same manner as explained above.

2.2. Output for Boundary Cells at Signalized Intersection. For finite roads, boundary conditions can be specified by means of input and output cells. The input flow can be obtained from the loop detector directly, which is discussed further in the following section on implementation. Either the data from a fixed detector or signal timing at the intersection can model the output flow. Because the data is assumed to be collected

by the SCOOT system in this paper, there is no detector at the stop line. Therefore, another estimation method for the output cell, using signal timing parameters, is proposed.

Since in the traditional CTM signal control is not considered, we provide further analysis for the output cell during the red and green phases. When the signal light is red, none of the vehicles can pass the stop line. Therefore, the traffic flow is zero, although there can be significant supply in each cell $i+1$ due to its upstream cell i. The supply and demand space can be written as follows:

$$\phi_I^r(k) = \begin{cases} \lambda \cdot v_f \cdot \rho_I(k), & \text{if } \rho_I(k) < \rho_a,\ k \in \text{redphase}, \\ Q_{\max}, & \text{if } \rho_I(k) \geq \rho_a,\ k \in \text{redphase}, \end{cases} \quad (8)$$

$$\varphi_{I+1}^r(k) = 0, \quad \text{if } k \in \text{redphase},$$

where $\phi_I^r(k)$ means the demand space from cell $I+1$ and $\varphi_{I+1}^r(k)$ represents the supply space for cell I when t represents the red phase.

When the signal switches to green, at the beginning, the vehicles enter the intersection at the saturation flow rate. During the end of the green phase, the output flow depends on the density of the output cell. If the density is larger than its critical density, the output flow is still at the saturation rate. Otherwise, the vehicles leave the cell at the rate of $v_f \cdot \rho_I(k)$. Regarding the supply space in cell I, as there is no block downstream, it can be regarded as plus infinity, as shown in Figure 3. This can be expressed as follows:

$$\phi_I^g(k) = \begin{cases} \lambda \cdot v_f \cdot \rho_I(k), & \text{if } \rho_I(k) < \rho_a,\ k \in \text{greenphase}, \\ Q_{\max}, & \text{if } \rho_I(k) \geq \rho_a,\ k \in \text{greenphase}, \end{cases} \quad (9)$$

$$\varphi_{I+1}^g(k) = +\infty, \quad \text{if } k \in \text{greenphase},$$

where $\phi_I^g(k)$ means the demand space from cell $I+1$ and $\varphi_{I+1}^g(k)$ represents the supply space in cell I when k is in the green phase.

Thus, the traffic flow functions for the output cell can be written as follows:

$$\begin{aligned} q_{I+1}(k) &= \begin{cases} \min\{\phi_I^r(k), \varphi_{I+1}^r(k)\} & \text{if } k \in \text{redphase} \\ \min\{\phi_I^g(k), \varphi_{I+1}^g(k)\} & \text{if } k \in \text{greenphase} \end{cases} \quad (10) \\ &= \begin{cases} 0 & \text{if } k \in \text{redphase} \\ \min\{\lambda \cdot v_f \cdot \rho_I(k), Q_{\max}\} & \text{if } k \in \text{greenphase}. \end{cases} \end{aligned}$$

2.3. Modification of the CTM. Since the original model was developed for freeways, a numerical simulation to check its correctness in describing of vehicle aggregation at signalized intersections is necessary. To be closer to real situations, the following parameters are assumed in the simulation: a one-way road with three lanes, the maximum traffic density on road being 0.4332 veh/m, the critical traffic density being 0.1 veh/m, the length of each cell being 60 m, and the number of these cells being 8. At each time interval of 4 seconds, the density values of all cells were estimated using a Matlab program implementing the discussion above.

As shown in Table 1, the increase in density values means the aggregation of vehicles, and the decrease in density values refers to the release of vehicles. Therefore, to a certain extent, the basic CTM can describe the queue process qualitatively. However, it is noticeable that the density evolutions in cells are different from each other during the process of queue discharging. It becomes clear in Table 1, which shows the density values in each cell from 419 s to 520 s. In cell 8, the density values changed to the critical density within 40 s and in cells 7, 6, and 5 within 60 s, 76 s, and 84 s, respectively. However, according to the assumption of the CTM, the cells of the road are homogenous, so that the density evolutions should be similar, and the time discussed above has to be approximately equal for each cell. So, the basic model should be modified to describe the queue discharging process.

The analysis of the real queue release process shows that the vehicles in the output cell leave first, and the vehicles in the immediate upstream cell cannot flow into the output cell until the density gets close to the critical density. This rule is applied to every cell up to the queue tail cell. During the queue discharging process, the traffic density in the downstream is larger than both the upstream and critical densities. The function of flow can be written as follows:

$$q_i(k) = \begin{cases} 0, & \text{if } \rho_{i-1}(k) > \rho_i(k),\ H_i \cdot \rho_i(k) - Q_{\max} \cdot T > \rho_b, \\ \dfrac{(\lambda \cdot (2 \cdot \rho_b - \rho_i(k)) \cdot H_i)}{T}, & \text{if } \rho_{i-1}(k) > \rho_i(k),\ H_i \cdot \rho_i(k) - Q_{\max} \cdot T < \rho_b, \\ \min\{\lambda \cdot \rho_i(k) v_i(k), Q_{\max}, \lambda \cdot (\rho_j - \rho_i(k)) \cdot H_i\}, & \text{else.} \end{cases} \quad (11)$$

The density values were generated once again according to the modified model, by using a Matlab program. The results are shown in Figure 4.

As can be clearly seen in Figure 4, the slopes of contour line are similar during the queue discharging process, which means the evolutions of density in the cells are similar to each

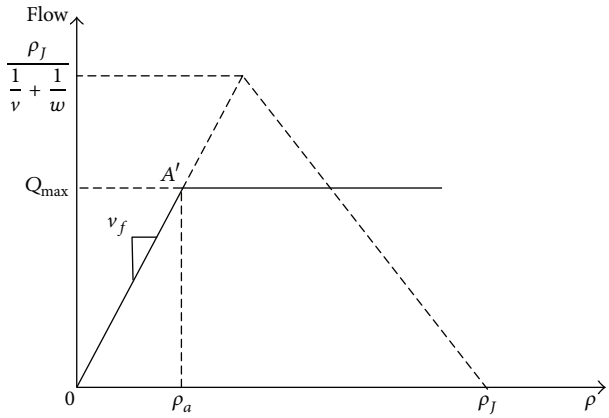


Figure 3: FD for cell $I+1$ during the green phase.

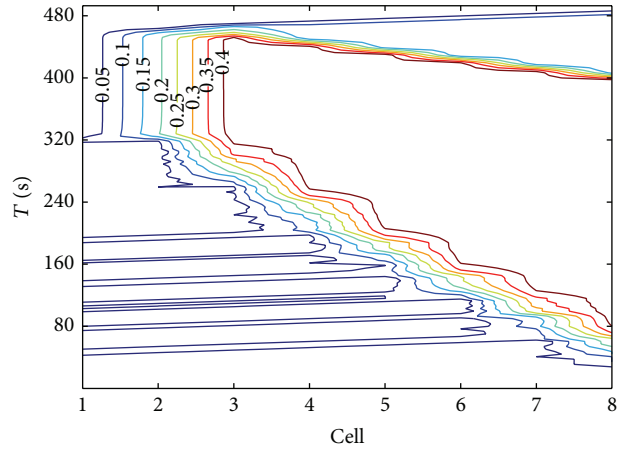


Figure 4: Density evolution based on the modified model.

Table 1: Density values evolution based on the basic model.

t (s)	Cell						
	2	3	4	5	6	7	8
412	0.1884	0.4332	0.4332	0.4332	0.4332	0.4332	0.4332
416	0.1884	0.4332	0.4332	0.4332	0.4332	***0.4332***	0.3332
420	0.1884	0.4332	0.4332	0.4332	*0.4332*	***0.4032***	0.2632
424	0.1884	0.4332	0.4332	**0.4332**	*0.4242*	***0.3612***	0.2143
428	0.1884	0.4332	0.4332	**0.4305**	*0.4053*	***0.3171***	0.18
432	0.1884	0.4332	0.4324	**0.4229**	*0.3789*	***0.276***	0.156
436	0.1884	0.433	0.4296	**0.4097**	*0.348*	***0.24***	0.1392
440	0.1883	0.4319	0.4236	**0.3912**	*0.3156*	***0.2098***	0.1274
444	0.1879	0.4294	0.4139	**0.3685**	*0.2838*	***0.1851***	0.1192
448	0.1868	0.4248	0.4003	**0.3431**	*0.2542*	***0.1653***	0.1134
452	0.1842	0.4174	0.3831	**0.3164**	*0.2275*	***0.1497***	0.1094
456	0.1795	0.4071	0.3631	**0.2898**	*0.2042*	***0.1376***	0.1066
460	0.1717	0.3939	0.3411	**0.2641**	*0.1842*	***0.1283***	0.1046
464	0.1599	0.3781	0.318	**0.2401**	*0.1675*	***0.1212***	0.1032
468	0.1434	0.3601	0.2946	**0.2183**	*0.1536*	***0.1158***	0.1023
472	0.1214	0.3404	0.2717	**0.1989**	*0.1423*	***0.1118***	0.1016
476	0.0936	0.3198	0.2499	**0.1819**	*0.1331*	0.1087	0.1011
480	0.0596	0.2989	0.2295	**0.1673**	*0.1258*	0.1064	0.1008
484	0.0193	0.278	0.2108	**0.1548**	*0.12*	0.1047	0.1005
488	0	0.2306	0.194	**0.1444**	*0.1154*	0.1035	0.1004
492	0	0.1589	0.1791	**0.1357**	*0.1118*	0.1025	0.1003
496	0	0.0827	0.1661	**0.1285**	0.109	0.1019	0.1002
500	0	0.0025	0.1548	**0.1227**	0.1069	0.1014	0.1001
504	0	0	0.0642	**0.1179**	0.1052	0.101	0.1001
508	0	0	0	0.0837	0.104	0.1007	0.1001
512	0	0	0	0	0.0879	0.1005	0.1
516	0	0	0	0	0	0.0885	0.1
520	0	0	0	0	0	0	0.0886

other. This result is similar to some traditional traffic flow theories, which indicates that the modified model can be used to depict the discontinuous flow at intersections with signal control.

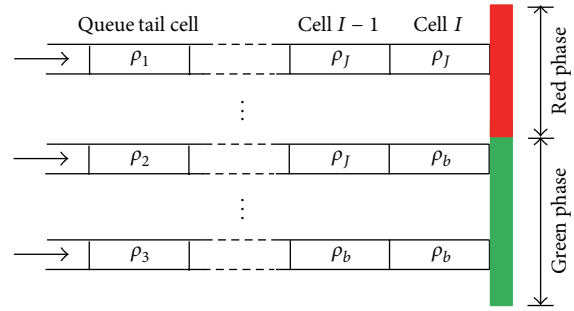


Figure 5: Density values evolution in a signal cycle.

3. Queue Estimation Model

This research proposes a real-time queue estimation model that aims to identify the point where the tail of the queue is located at every moment. Further, the maximum and average queue length can be calculated based on the real-time queue lengths. There are two steps according to the proposed method. The first step is to identify the cell where the tail of the queue is located, based on the density values in every cell. In the second step, the exact queue tail point is identified based on the density distribution in the cell selected in the first step. To overcome traffic density abrupt changes, it is assumed that in the queue tail cell the traffic density transforms into another state linearly. The formulation and analysis of this linear change are presented in Section 3.2.

3.1. Queue Tail Cell Identification. To understand the queue discharging process, a typical cycle of the evolution of cell density is shown in Figure 5. Similar processes are repeated in all cycles.

The queue starts to discharge at the beginning of the green phase, and the densities of cells between the stop line and queue tail are equal to the jam density. While discharging, some densities in those downstream cells become critical, while other upstream cells are still under the jam condition. Sometime after the green light started, the maximum queue length is achieved, and at this moment, all the densities of

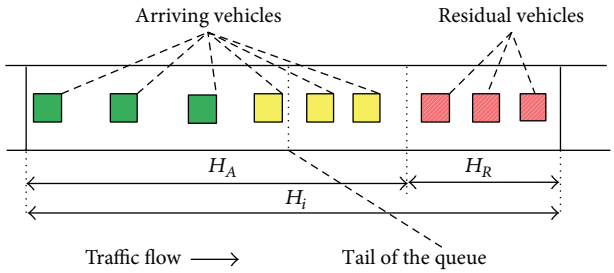


Figure 6: Vehicles spatial distribution in queue tail cell.

the downstream of the queue tail cell are equal to the critical density. Therefore, it can be concluded that if the cell i and all its downstream cell densities are larger than the critical density and if the upstream cell densities are lower than the critical density, the queue tail locates in cell i or cell $i-1$.

Further, to analyze whether the queue tail still stays in cell i or moves to cell $i-1$, the demanded space from cell $i-1$ and the supplied space from cell i should be compared. If all the last time interval vehicles of cell $i-1$, $n_{i-1}(k-1)$, flow into cell i in this time interval, it suggests that the traffic flow is free and the queue tail is still in cell i. Otherwise, the tail of the queue is located in cell $i-1$. Thus, it can be summarized that the queue tail is located in cell.

3.2. Method of Queue Length Calculation. To get the exact point where the tail of the queue is located, it is necessary to analyze the distribution of the density in the queue tail cell identified in Section 3.1. The queue tail cell is divided into a number of homogenous sections. In this paper, the queue tail is defined as to where the shockwave propagates. So, the queue length can be estimated as long as the exact space is identified, whose density is ρ_b. The point A in Figure 2 represents the initial formation of the queue with the increasing density.

The vehicles in the queue tail cell can be divided into three categories, according to the difference in densities. The rules for classification are as follows:

(1) The first category refers to the vehicles that did not move to the downstream cell in the last time interval, since the supply space was not enough. They are shown in Figure 6 as red ones. The density of this section approximately equals its immediate downstream cell ρ_{i+1}.

(2) The vehicles of the second category are those that came in the last time interval from the upstream cell; they meet the first category vehicles tail. Due to different braking time of the vehicles in this category, the density ρ_{trans} is not uniform, and it gradually changes from the density of the first category to the one of the third category vehicles. The vehicles are indicated as yellow ones in Figure 6.

(3) All other vehicles arrive to the queue tail cell in the last time interval and do not reach the tail of the first category vehicles, so that these vehicles do not have to brake, and their density ρ_{in} is approximately that of the upstream cell in the last interval. The green ones in Figure 6 represent these vehicles.

According to the analysis above, the density of cell $i+1$ is greater than the critical density ρ_b and the density of the third category vehicles is less than ρ_b. Therefore, there must be a section in the transform area (the second category) with the density equal to ρ_b, and the location of the queue tail is shown in Figure 6 as the red line. To calculate the queue tail location, the numbers of vehicles in the three categories should be obtained.

The spatial length H_R, which corresponds to the vehicles that remained after the last time interval, equals the number of such vehicles divided by the density of cell $i+1$, which can be written as

$$H_R = \frac{\rho_i(k-1)\cdot H_i - q_{i+1}(k-1)\cdot T}{\lambda\cdot\rho_{i+1}(k)}. \tag{12}$$

The supplied space H_A for the vehicles arriving in the current interval is the rest of the length of the queue tail cell. As all the vehicles in cell $i-1$ move into the queue tail cell, the density of the third category vehicles can be written as

$$\rho_{\text{in},i}(k) = \frac{n_{i-1}(k)}{H_{i-1}}. \tag{13}$$

Although the densities of sections in the transform area are not uniform, they can be regarded as the average value of $\rho_{\text{in},i}(k)$ and $\rho_{i+1}(k)$. First, the supplied space H_A is filled with vehicles of categories 1 and 2. Second, the vehicles that have arrived to the queue tail cell in the last time interval consist of two parts:

$$H_S = \frac{\chi_1\lambda\left(\rho_{\text{in}} + \rho_{i+1}(k)\right) + 2\chi_2}{\lambda^2\left(\rho_{\text{in}}^2 + \rho_{\text{in}}\rho_{i+1}(k)\right)}. \tag{14}$$

The numbers of vehicles in the two parts can be calculated. The functions can be written as follows:

$$\chi_1 = \lambda\frac{\rho_{\text{in},i}(k)^2 H_A + \lambda\rho_{\text{in},i}(k)\rho_{i+1}(k)H_A - 2\rho_{\text{in},i}(k)\rho_i(k)\cdot T}{\rho_{i+1}(k) - \rho_{\text{in},i}(k)},$$
$$\chi_2 = \frac{q_i(k-1)\cdot T\cdot\rho_{i+1}(k) - \lambda\cdot\rho_{\text{in},i}(k)\cdot\left(\rho_{\text{in},i}(k)\cdot H_A + \rho_{i+1}(k)\cdot H_A - \rho_i(k)\cdot T\right)}{\rho_{i+1}(k) - \rho_{\text{in},i}(k)}. \tag{15}$$

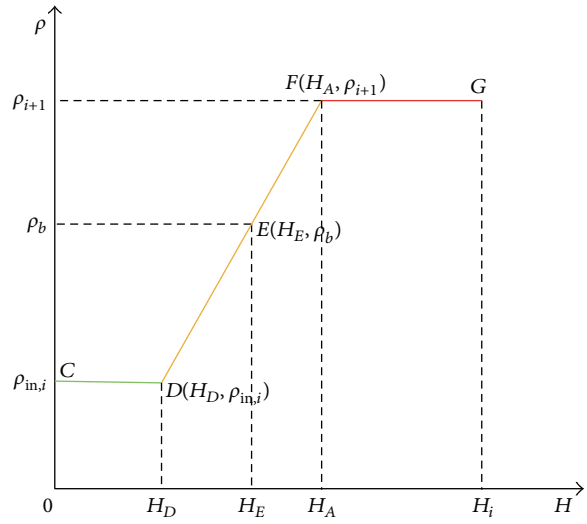


FIGURE 7: General density distribution in the queue tail cell.

We assume that the densities change from $\rho_{i+1}(k)$ to $\rho_{\text{in},i}(k)$ linearly in the transform area. For further calculations, the coordinate system is used for the density values and their positions, as shown in Figure 7.

Point E indicates the queue tail position, and its ordinate value is known as ρ_b. The ordinate value of the line segment FG means the density of the first category vehicles, and point F represents the tail of these vehicles with coordinates $(\chi_1/\lambda\rho_{\text{in},i}(k) + 2\cdot\chi_2/(\lambda\rho_{i+1}(k) + \lambda\rho_{\text{in},i}(k)), \rho_{i+1}(k))$. The line segment CD represents the third category vehicles, and the coordinates of its starting point D are $(\chi_1/\lambda\rho_{\text{in},i}(k), \rho_{\text{in},i}(k))$. To obtain the coordinates of point E, its relationship with the known points F and G should be established.

As illustrated in Figure 7, Points D, E, and F are located on the same line, with the slope that can be written as

$$\frac{\chi_1/\lambda\rho_{\text{in},i}(k) + 2\cdot\chi_2/\left(\lambda\rho_{i+1}(k) + \lambda\rho_{\text{in},i}(k)\right) - \chi_1/\lambda\rho_{\text{in},i}(k)}{\rho_{i+1}(k) - \rho_{\text{in},i}(k)} = \frac{H_E - \chi_1/\lambda\rho_{\text{in},i}(k)}{\rho_b - \rho_{\text{in},i}(k)}. \tag{16}$$

The abscissa of point E can be obtained using (15) and (16), as shown below:

$$H_E = \frac{\left(\rho_b - \rho_{\text{in},i}(k)\right) 2\cdot\chi_2}{\lambda^2\rho_{\text{in},i}(k)\left(\rho_{i+1}(k) - \rho_{\text{in},i}(k)\right)\left(\rho_{i+1}(k) + \rho_{\text{in},i}(k)\right)} + \frac{\chi_1}{\lambda\rho_{\text{in},i}(k)}. \tag{17}$$

To avoid zero denominators, three cases should be analyzed according to their physical meanings.

As shown in Figure 8(a), if the density in cell $i+1$ is zero, $\rho_{i+1}(k) = 0$, then the signal just switched to red and vehicles begin to accumulate. There are no vehicles in the downstream cell of the output cell. The traffic density of output cell I becomes jam ρ_J during the red phase, so it can be assumed that $\rho_{I+1}(k)$ equals ρ_J until the green phase begins. In addition, because no vehicles flow into cell $I+1$, the traffic flow from cell I is zero. So, the function to calculate H_E can be written as follows:

$$H_E = \frac{\left(\rho_b - \rho_{\text{in},i}(k)\right) 2\cdot\chi_2}{\lambda^2\rho_{\text{in},i}(k)\left(\rho_J - \rho_{\text{in},i}(k)\right)\left(\rho_J + \rho_{\text{in},i}(k)\right)} + \frac{\chi_1}{\lambda\rho_{\text{in},i}(k)}. \tag{18}$$

When the input density is zero, $\rho_{\text{in},i}(k) = 0$, it means that there were no vehicles flowing into cell i from cell $i-1$ in the last time interval, $q_i(k-1) = 0$. According to Figure 8(b), it can be seen that point E does not exist, and the queue tail is represented by point F. Therefore, the queue tail is located at the end of the remaining vehicles from last time interval:

$$H_E = H_R - \frac{n_i(k-1) - q_{i+1}(k)\cdot T}{\lambda\cdot\rho_{i+1}(k)}. \tag{19}$$

Because the density value cannot be negative, $\rho_{i+1}(k)$ is equivalent to $\rho_{\text{in},i}(k)$. In addition, as analyzed in Section 3.1, the density of the third category vehicles is equal to or less than the critical density, and the vehicle density in cell $i+1$ is equal to or is greater than ρ_b. Therefore, if $\rho_{i+1}(k) = \rho_{\text{in},i}(k)$, then both are equal to the critical density. This indicates that the tail of the queue is located at the boundary of cell i and cell $i-1$, as shown in Figure 8(c).

The queue length $H_q(k)$ in a certain short time interval T can be calculated based on the queue tail location and can be written as

$$H_q(k) = \sum_{j=I-i-1}^{I}\left(H_i - H_E + H_j\right). \tag{20}$$

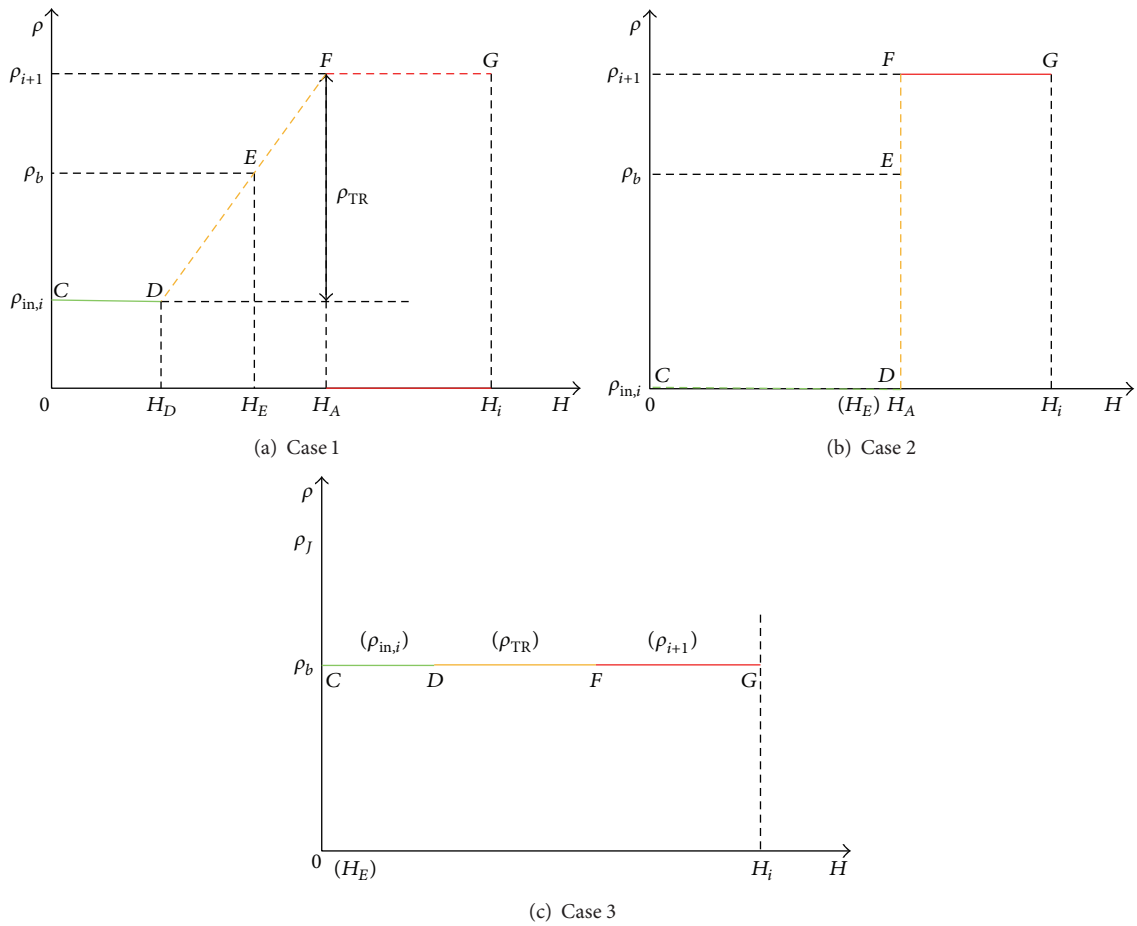
In signal cycle m, the average queue length H_m^{ca} is the arithmetic mean of every queue length in this cycle:

$$H_m^{ca} = \sum_{p=0}^{C/T} H_q(k)\left(m\cdot C + p\right). \tag{21}$$

The maximum queue length in signal cycle is the longest queue in every short time interval.

4. Implementation

4.1. Implementation Procedure. Figure 9 emphasizes the idea that the procedure of queue length estimation consists of two major steps. First, traffic flow parameters in all cells are calculated based on the modified CTM. If the density in cell I is larger than the critical density and all the upstream densities are larger than the critical density down to cell $i-1$, it can be judged that the queue tail is located in cell i or cell $i-1$. The exact queue tail cell is identified by verifying whether all the vehicles from cell $i-1$ moved to cell i in the last time interval. Second, the queue tail is located within the cell by transforming the problem into searching of



(a) Case 1

(b) Case 2

(c) Case 3

Figure 8: Density distributions in queue tail cell for three special cases.

the point whose ordinate value is the critical density ρ_b. Then, the general formula for the queue length in the queue tail cell is deduced. Three special cases mentioned in this paper should be considered in addition to the general estimation of the queue tail. Finally, the queue length in the current time interval can be easily calculated by adding the queue length in the queue tail cell and the lengths of all the cells downstream.

4.2. Simulation Evaluation Based on Field Data. The intersection of Yatai Street and Ziyou Road in Changchun, China, is selected as the investigation site, because they are two arterial roads without much interruption due to secondary roads, and traffic flows include both free and saturated conditions. The actual data of the test road was used for calibrating the related parameters of the simulated traffic environment such as the free-flow speed 54 km/h, critical density 33.3 veh/km/lane, jam density 144.4 veh/km/lane, the distribution of the traffic flow in the test road, and the capacity. Simulation model was verified by comparing the output parameters with the field data collected from the video detector temporally located on the street. The input traffic parameters for simulation software are investigated, including the signal cycle 200 s, green ratio 0.36, and traffic volume measured every 5 minutes during 35 signal cycles. Figure 10 shows the site of test. The length of the road between the fixed detector and stop line is 480 m. The time interval was chosen to be two seconds. Accordingly, the selected street is divided into 16 sections. VISSIM output includes the second-by-second data collected by the fixed detector, maximum queue length, and average queue length in one signal cycle.

The queue length is calculated by a Matlab program, and the proposed model based on the modified CTM is tested. We also selected two other traditional approaches for comparison. One method is based on the shockwave theory. According to Liu et al. [21], the queue back is the maximum queue length in a signal cycle. We compared the maximum queue length estimated by the proposed model with the shockwave theory. The key of shockwave based method is to identify points where queuing shockwave meets discharging shockwave. The location of the identified point represents the rear of the maximum queue. Therefore, the maximum queue length is obtained. The performance of the two methods is shown in Figure 11. For the evaluation of the average queue length, we selected a statistical model based on the vacation queuing theory [31, 32]. According to the model, the queuing and discharging processes are regarded as the fixed vacation time following the M/G/1 distribution. The input process follows the passion distribution, service time

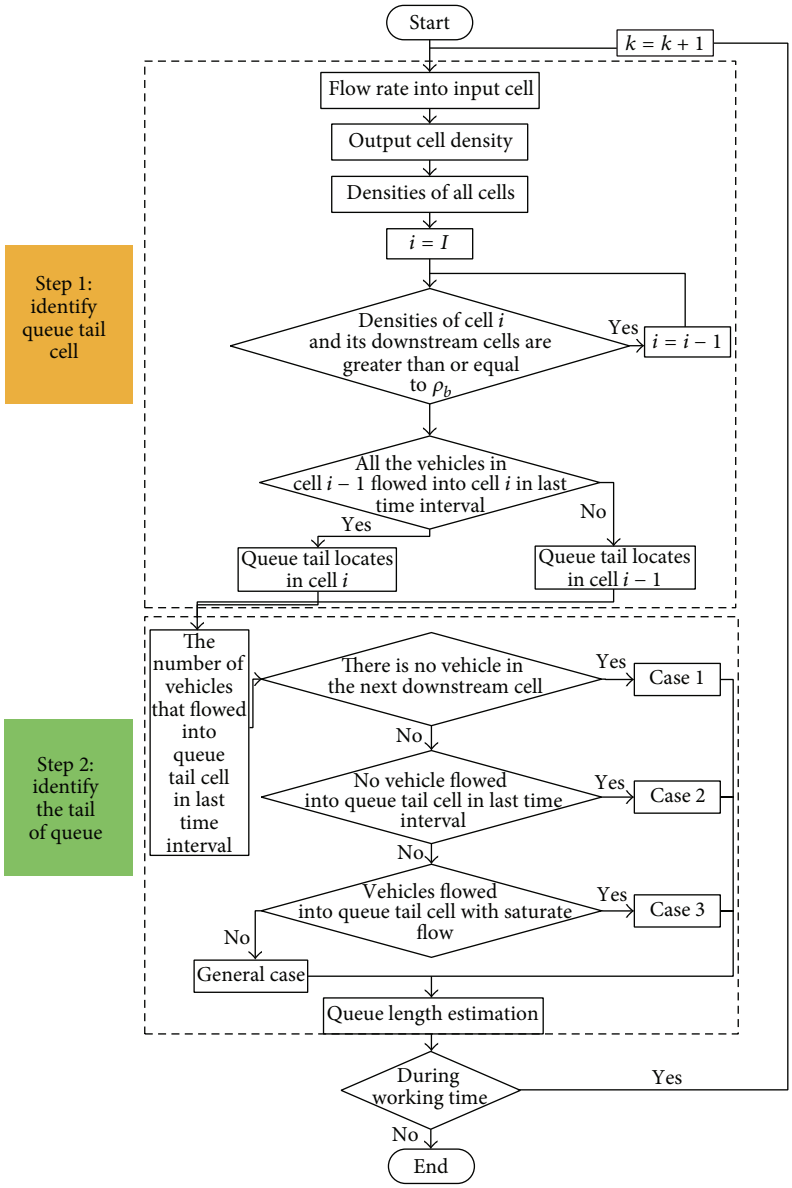


Figure 9: Flow chart of the implementation of queue length estimation.

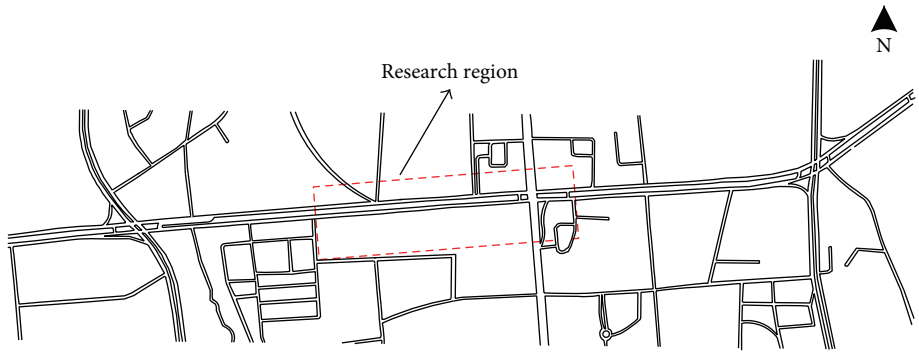


Figure 10: Intersection for testing.

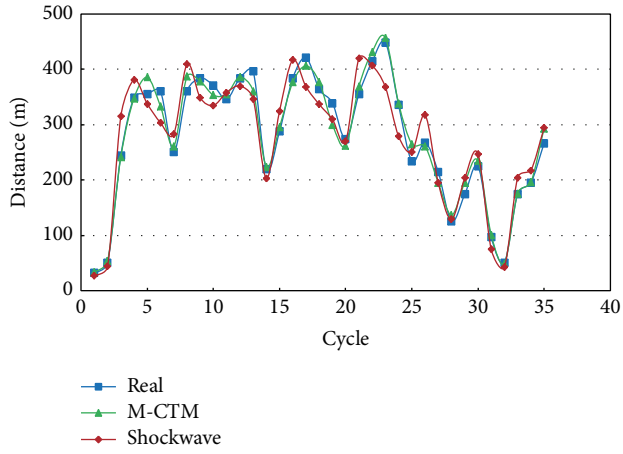


Figure 11: Maximum queue length estimation comparison.

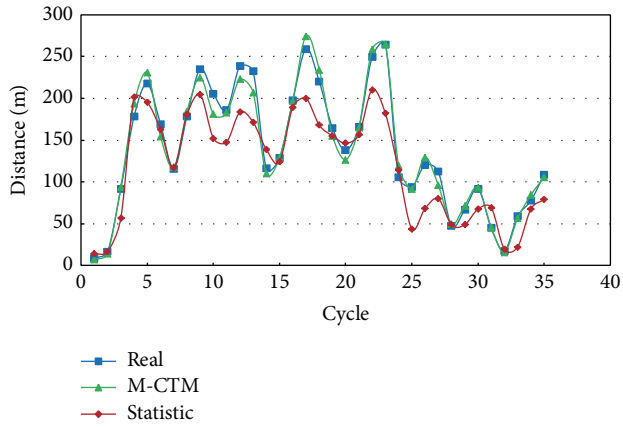


Figure 12: Average queue length estimation comparison.

follows the general distribution, and vacation time is the red phase time in a signal cycle. Then, the average queue length is the expected value of the queue length. The performance of our method and the average queue length method is shown in Figure 12. Tables 2 and 3 show the Mean Absolute Errors (MAE) and Mean Relative Errors (MRE) calculated using the following formulas:

$$\text{MAE} = \frac{1}{n}\sum_{n}\left|\text{output} - \text{estimation}\right|,$$
$$\text{MRE} = \frac{1}{n}\sum_{n}\left|\frac{\text{output} - \text{estimation}}{\text{output}}\right| \times 100\%, \quad (22)$$

where n is the sample size.

As presented in Table 2, the proposed model based on the modified CTM can estimate the maximum queue length in one signal cycle with the MRE of 4.79%, where the absolute error is about 12.79 meters. As shown in Figure 11, almost all the relative errors are less than 10%. The traditional shockwave based model can estimate the maximum queue with the MRE of 11.59% and absolute error of about 31 meters. Figure 11 also presented that almost all the relative errors are less than 16%. The comparison results indicate that the proposed model can exactly describe the traffic flow at a signal intersection of an urban road and estimate the maximum queue length with satisfactory accuracy.

Table 2: Errors of maximum queue length estimation.

	Shockwave based model	M-CTM based model
MAE	30.66 m	12.79 m
MRE	11.59%	4.79%

Table 3: Errors for average queue length estimation.

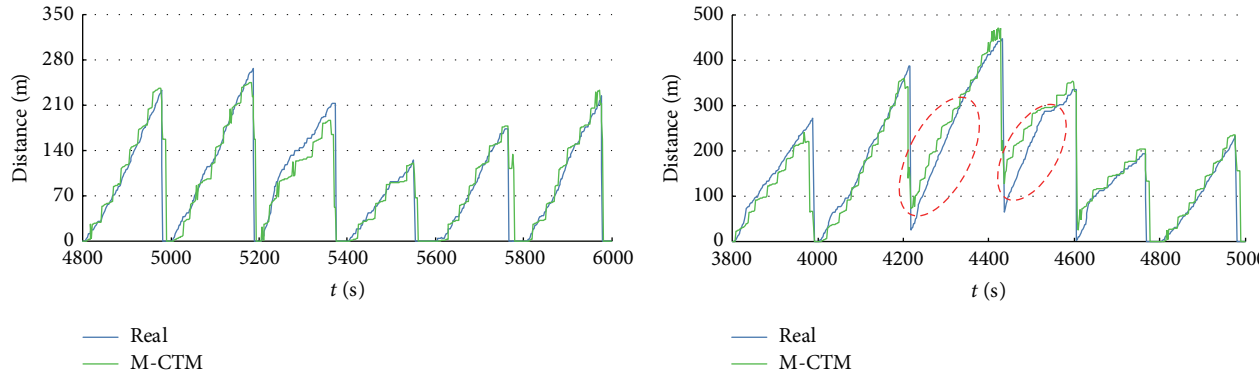
	Statistic based model	M-CTM based model
MAE	26.25 m	7.67 m
MRE	20.79%	6.11%

Because the average queue length is calculated using the mean queue values of every time interval during one signal cycle, corresponding results can partly represent real-time estimation performance. As indicated in Figure 12, the proposed model also performs well in estimating the average queue length in every signal cycle, with relatively low errors compared to the maximum queue length estimations. The MRE of the proposed estimation model based on the M-CTM is about 6% lower than the traditional statistical model having the MRE of 20.79%. The proposed model performed especially well compared to the traditional method when the traffic flow fluctuations were high.

As shown in Figure 13, the proposed model performs well in describing the queue formation and discharging process for both long and short queues. Here, a short queue is the one when all vehicles can pass the stop line in one signal cycle, and a long queue is the one where some vehicles in the queue need to wait for the next cycle. Overall, the proposed model can successfully estimate the maximum and average queue lengths and succeeds in tracking real-time queue trajectories with satisfactory accuracy.

5. Discussion and Future Research

In this paper, we propose the M-CTM based model to estimate the real-time queue length at a signalized intersection, instead of just calculating the maximum or average queue length as traditional methods often do. The queue length in one time interval is estimated in two steps. At the first step, we estimate the densities in all cells based on the M-CTM. The CTM model was modified to describe the traffic flow at a signalized intersection of an urban road. Figure 4 shows the accuracy of the modified CTM in describing the queue forming and discharging processes. At the second step, the density estimated in the first step is analyzed to determine the queue tail position. First, the cell where the queue tail is located is identified. Then, through the construction of the spatial distribution density in the queue tail cell, the exact position of the queue tail is located. There are three special cases, presented in Figure 8, which cannot be calculated using a general formula. Every queue length in a short time interval is calculated; and the queue trajectory is estimated as well.



(a) Short queue dynamics estimated

(b) Long queue dynamics estimated

Figure 13: Queue trajectory.

The proposed model was tested on VISSIM simulation platform, which was calibrated using field data. The results show the following:

(1) As indicated in Table 1, the traditional CTM can describe the queue forming process for signalized intersections on urban roads; however, the description of the releasing process is not quite accurate. As a basis of the queue length estimation, the exact estimation of the density is necessary. In comparison, Figure 4 shows that the evaluation of the queue discharging process obtained by the proposed modified model is close to the real conditions.

(2) The test results of Section 4.2 suggest that the method proposed in this paper can estimate the maximum and average queue lengths in every signal cycle regardless of whether there is congestion or not. It should be mentioned that, as shown in Figures 11 and 12, the M-CTM based model performs better than the traditional CTM, especially during a saturated period. It means that the proposed approach is more robust.

(3) Because each time interval is very short, we assume that the queue length estimated in a time interval represents the average length in the current time interval. Every queue length is calculated in short time intervals, second-by-second, to obtain real-time queue length estimations. As can be seen in Figure 13, the proposed model performs well in tracking the queue tail trajectory with satisfactory accuracy, under both free-flow and saturated conditions.

Although the proposed model performs well according to the test results, there are several limitations, which can be addressed in future work:

(1) The proposed model needs further testing and validation and evaluation using field data collected by fixed detectors directly. We will conduct more experiments to test the performance of the proposed model under different traffic conditions. There are some other problems, such as errors caused by detectors. Therefore, preprocessing of the raw data collected by fixed detectors is generally needed.

(2) In this paper, the vehicles are not allowed to change lanes after passing detectors. However, the distance between detectors and the stop line is long enough, so, in real world situations, drivers have enough time and space to change the lane. Therefore, some methods for taking lane changing in relative publications into account should be considered into the proposed model.

(3) As is indicated in Figure 13(b), highlighted with circles, during 4200–4600 s the proposed model overestimated the queue length. It means that the error accumulation exists in the proposed model, and the problem is especially critical during congestion periods, because the system has no chance to clear the vehicles and reset the density values.

(4) Although the time interval adopted in testing is very short, the span still exists. The queue trajectory estimated is not as smooth as the real one. A filtering method, such as Kalman filter, should be used to smooth the estimated results, and improve the accuracy of the proposed model.

Conflict of Interests

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

Acknowledgment

The work was supported by the National Natural Science Foundation of China (Grant no. 51378237).

References

[1] F. V. Webster and B. M. Cobbe, "Traffic signals," Road Research Technical Paper 56, Road Research Laboratory, Her Majesty's Stationery Office, London, UK, 1966.

[2] W. B. Cronje, "Analysis of existing formulas for delay, overflow, and stops," *Transportation Research Record*, vol. 905, pp. 89–93, 1983.

[3] K. Balke, H. Charara, and R. Parker, "Development of a traffic signal performance measurement system (TSPMS)," Report 0-4422-2, Texas Transportation Institute, 2005.

[4] F. Webster, "Traffic signal settings," Road Research Technical Paper 39, Road Research Laboratory, Her Majesty's Stationery Office, London, UK, 1958.

[5] G. F. Newell, "Approximation methods for queues with application to the fixed-cycle traffic light," *SIAM Review*, vol. 7, no. 2, pp. 223–240, 1965.

[6] D. H. Green, "Control of oversaturated intersections," *Journal of the Operational Research Society*, vol. 18, no. 2, pp. 161–173, 1967.

[7] P. G. Michalopoulos and G. Stephanopoulos, "Oversaturated signal systems with queue length constraints—I: single intersection," *Transportation Research*, vol. 11, no. 6, pp. 413–421, 1977.

[8] T.-H. Chang and J.-T. Lin, "Optimal signal timing for an oversaturated intersection," *Transportation Research Part B: Methodological*, vol. 34, no. 6, pp. 471–491, 2000.

[9] P. B. Mirchandani and Z. Ning, "Queuing models for analysis of traffic adaptive signal control," *IEEE Transactions on Intelligent Transportation Systems*, vol. 8, no. 1, pp. 50–59, 2007.

[10] A. D. May, "Traffic flow theory-the traffic engineers challenge," in *Proceedings of the Institute of Traffic Engineering*, pp. 290–303, 1975.

[11] R. Akcelik, *A Queue Model for HCM 2000*, ARRB Transportation Research, Vermont South, Australia, 1999.

[12] G. Vigos, M. Papageorgiou, and Y. Wang, "Real-time estimation of vehicle-count within signalized links," *Transportation Research Part C: Emerging Technologies*, vol. 16, no. 1, pp. 18–35, 2008.

[13] N. Geroliminis and A. Skabardonis, "Prediction of arrival profiles and queue lengths along signalized arterials by using a markov decision process," *Transportation Research Record*, vol. 1934, no. 1, pp. 116–124, 2005.

[14] H. Heffes and D. M. Lucantoni, "Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance," *IEEE Journal on Selected Areas in Communications*, vol. 4, no. 6, pp. 856–868, 1986.

[15] M. Poch and F. Mannering, "Negative binomial analysis of intersection-accident frequencies," *Journal of Transportation Engineering*, vol. 122, no. 2, pp. 105–113, 1996.

[16] M. J. Lighthill and G. B. Whitham, "On kinematic waves. I. Flood movement in long rivers," *Proceedings of the Royal Society of London A*, vol. 229, no. 1178, pp. 281–316, 1955.

[17] M. J. Lighthill and G. B. Whitham, "On kinematic waves. II. A theory of traffic flow on long crowded roads," *Proceedings of the Royal Society of London A*, vol. 229, no. 1178, pp. 317–345, 1955.

[18] P. I. Richards, "Shock waves on the highway," *Operations Research*, vol. 4, no. 1, pp. 42–51, 1956.

[19] G. Stephanopoulos, P. G. Michalopoulos, and G. Stephanopoulos, "Modelling and analysis of traffic queue dynamics at signalized intersections," *Transportation Research Part A: General*, vol. 13, no. 5, pp. 295–307, 1979.

[20] A. Skabardonis and N. Geroliminis, "Real-time estimation of travel times on signalized arterials," in *Proceedings of the 16th International Symposium on Transportation and Traffic Theory*, pp. 387–406, College Park, Md, USA, July 2005.

[21] H. X. Liu, X. Wu, W. Ma, and H. Hu, "Real-time queue length estimation for congested signalized intersections," *Transportation Research Part C: Emerging Technologies*, vol. 17, no. 4, pp. 412–427, 2009.

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

[23] L. Muñoz, X. Sun, D. Sun, G. Gomes, and R. Horowitz, "Methodological calibration of the cell transmission model," in *Proceedings of the American Control Conference*, vol. 1, pp. 798–803, IEEE, Boston, Mass, USA, June-July 2004.

[24] X. Sun, L. Muñoz, and R. Horowitz, "Mixture kalman filter based highway congestion mode and vehicle density estimator and its application," in *Proceedings of the 2004 American Control Conference (AAC '04)*, pp. 2098–2103, July 2004.

[25] C. M. J. Tampère and L. H. Immers, "An extended Kalman filter application for traffic state estimation using CTM with implicit mode switching and dynamic parameters," in *Proceedings of the IEEE Intelligent Transportation Systems Conferenc (ITSC '07)*, pp. 209–216, IEEE, Seattle, Wash, USA, October 2007.

[26] M. Hadiuzzaman and T. Z. Qiu, "Cell transmission model based variable speed limit control for freeways," *Canadian Journal of Civil Engineering*, vol. 40, no. 1, pp. 46–56, 2013.

[27] Y. Kim and H. Keller, "On-line traffic flow model applying the dynamic flow-density relation," in *Proceedings of the 11th International Conference on Road Transport Information and Control*, pp. 141–145, London, UK, March 2002.

[28] T. Pohlmann and B. Friedrich, "Online control of signalized networks using the cell transmission model," in *Proceedings of the 13th International IEEE Conference on Intelligent Transportation Systems (ITSC '10)*, pp. 1–6, IEEE, Funchal, Portugal, September 2010.

[29] B. Xie, M. Xu, J. Härri, and Y. Chen, "A traffic light extension to Cell Transmission Model for estimating urban traffic jam," in *Proceedings of the IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC '13)*, pp. 2566–2570, IEEE, September 2013.

[30] Y.-C. Chiou and Y.-F. Huang, "Stepwise genetic fuzzy logic signal control under mixed traffic conditions," *Journal of Advanced Transportation*, vol. 47, no. 1, pp. 43–60, 2013.

[31] Y. Levy and U. Yechiali, "Utilization of idle time in an M/G/1 queueing system," *Management Science*, vol. 22, no. 2, pp. 202–211, 1975.

[32] G. Choudhury, "On a batch arrival poisson queue with a random setup time and vacation period," *Computers & Operations Research*, vol. 25, no. 12, pp. 1013–1026, 1998.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 891204, 7 pages
http://dx.doi.org/10.1155/2015/891204

Research Article

A Decomposition Strategy for Optimal Design of a Soda Company Distribution System

J. A. Marmolejo,[1] I. Soria,[2] and H. A. Perez[2]

[1]*Faculty of Engineering, Anahuac University, 52786 Mexico City, MEX, Mexico*
[2]*CADIT, Anahuac University, 52786 Mexico City, MEX, Mexico*

Correspondence should be addressed to J. A. Marmolejo; jose.marmolejo@anahuac.mx

Received 2 May 2015; Revised 9 July 2015; Accepted 28 July 2015

Academic Editor: Pan Liu

This work presents a distribution problem of products of a soda bottling company. Commodities are produced at several plants with limited capacity and the demand of distribution centers is satisfied by shipping via cross-docking warehouses. The decomposition strategy is proposed to determine which warehouse needs to be opened to consolidate the demand and by which warehouse each distribution center is served exclusively. The objective is minimizing fixed costs and total transportation costs. The model presented is a mixed-integer programming model with binary variables for which we propose a decomposition strategy based on Benders algorithm. Numerical results show that the proposed strategy can provide the optimal solution of several instances. A large-scale case study based on a realistic company situation is analyzed. Solutions obtained by the proposed method are compared with the solution of full scale problem in order to determine the quality bound and computational time.

1. Introduction

According to international organisms, Mexico is the second biggest consumer of bottled soda with an average consumption of 160 liters per person a year after the United States, where 94% of the population consumes these types of drinks. The main point of sale of soda in Mexico is in small stores where 75% of the sales are carried out, 24% is in restaurants, and the remaining are in self-service stores [1]. For the company, it is very important to have an effective distribution network that provides the possibility to keep a high level of service to the client at the smallest possible cost; that is, the clients must have products in the opportune moment at minimum cost.

In 2012, the bottling company offered a level of service above 99.6% (delivery requested/customer request), but in 2013 the first supply problems were presented due to the productive capacity of the plants; this situation originated the specialization of the production lines of the company plants, holding the product readiness in the market at minimum cost; see Figure 1.

For this reason, we proposed a strategy for optimal design of a soda bottling company distribution system based on [1]. The proposed distribution network is constituted by plants, cross-dock warehouses, and distribution centers. Commodities are produced at several plants with limited capacity and the demand of distribution centers is satisfied by shipping via cross-docking warehouses. The problem is to determine which warehouse needs to be opened to consolidate the demand and by which warehouse each distribution center is served exclusively. The objective is minimizing fixed costs and total transportation costs; see Figure 2.

The idea is to establish a network of arcs which enables the flow of products in order to satisfy demand of distribution centers. A proper design can yield better operation levels and cost reductions. In this problem, these reductions can be significant. To the best of our knowledge, the specific model we pose in this study is not addressed in the literature; however, the area of fixed charge network design is closely related.

This paper is organized as follows. In Section 2, we present a brief review of the network design problem. Section 3 contains the specific problem formulation and then Section 4 describes the solution methodology. Section 5 presents computational results of the application of Benders

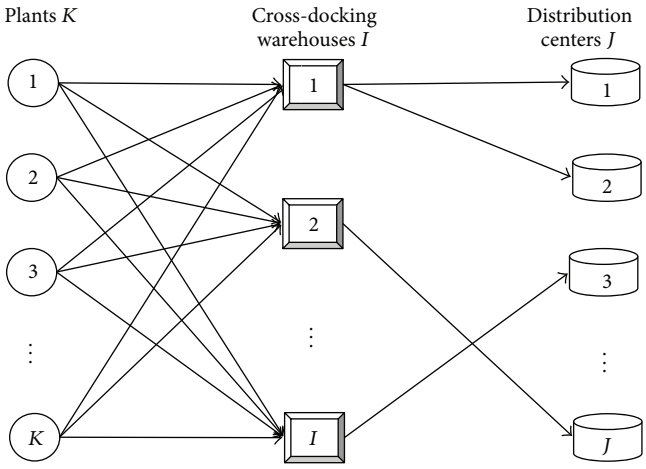


Figure 1: Proposed distribution system.

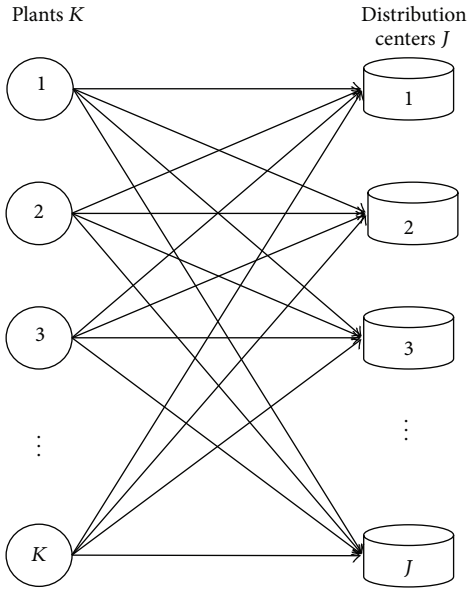


Figure 2: Current distribution system.

Decomposition in several instances. Finally, conclusions are in Section 6.

2. Literature Review

Nowadays, companies still pay attention to efficiency because of tactical and operational impact at the executive level; these aspects remain relevant when a new framework is proposed, because all decisions will depend on it to continue the pursuit of all the company objectives. This issue is referred to as supply chain network design problems by most researchers and frequently begins from real-life applications. In this sense, a traditional paper is [2]. They present a new method for the solution of the problem addressing the optimal location of distribution centers between plants and customers. A particular interest involves a study in facility location that is reviewed in [3], where the authors explain some important characteristics, methods, and application in the context of supply chain management. Furthermore, they mentioned that the discrete facility location problems could be categorized as supply chain network design problems. Despite depending on situations, many techniques used in the design of a distribution network are from the research operations area, especially those presented in [4]; some contributions are made to the state of the art that follows this subject for facility locations models. However, several industries are still facing problems related to design of the distribution network because it is a decision linked to improving the level of service, and many other important metrics. Along the same line, in [5], they use a case study related with a model for the distribution network of a region of consumer goods company taking into account several tactical decisions like lead time, credit performance, power, and distribution's reputation. Among various types of problems discussed in supply chain management, facility location has been studied for a long time by some researchers [6]. In [7], the authors follow an objective to propose a novel scheme of distribution. In these investigations, they include a classical facility location problem, and they presented cases of studies solved by appropriate techniques.

Typically, these problems are presented as a mixed-integer programming (MIP) formulation [8, 9]. In [2], the authors develop an algorithm based on Benders Decomposition for solving multicommodity distribution network design problem. Another classical model is presented in [10]. They consider a model to solve a minimization function which includes fixed cost in warehouses and distribution centers and transportation cost for multicommodities from plants to warehouses and finally to customers. Moreover, they consider a tri-echelon, multicommodity system concerning production, distribution, and transportation planning. The authors use a Lagrangian relaxation-based heuristic to provide an effective feasible solution for the problem. Also, they reconsider other different characteristics for solving the main problem of integrated logistics model [8, 11, 12]. In [13], they consider an integrated distribution network design and site selection problem arising in the context of transportation planning faced by the freight-forwarding industry; in this sense, they consider a strategic level multicommodity network design where each commodity is defined by a unique pair of origin and destination points and known required flow amount and other considerations proper to the real problem but using Benders Decomposition for its solution. Similarly, they illustrate the efficiency and the effectiveness of this approach. As in [14, 15] also a Benders Decomposition approach is used, first in combination with an intelligent algorithm to improve the time solution for the master problem and then in modified version to exploit the mathematical formulation of the problem in deterministic, multicommodity, single-period contexts, respectively. In [16], the authors exposed that while these production-distribution problems focused to pursue exact solutions efficiently by using optimization software (only for small instances and small dimensions), the main reason to avoid it is because they contain a large number of constraints and variables. Besides, they

propose heuristics for this bilevel mathematical problem using Stackelberg's equilibrium.

3. Mathematical Formulation

Let K be the set of manufacturing plants. An element $k \in K$ identifies a specific plant of the company. Let I be the set of the potential cross-dock warehouses. An element $i \in I$ is a specific cross-dock warehouse. Finally, let J be the set of distribution centers; a specific distribution center is any $j \in J$. Let $\mathbb{Z}$ denote the set of integers $\{0, 1\}$.

Parameters are as follows:

Q_k: capacity of plant k.

K_i: capacity of cross-dock warehouse i.

F_i: fixed costs of opening cross-dock warehouse in location i.

G_{ki}: transportation cost per unit of the product from the factories k to cross-dock warehouse i.

C_{ij}: cost of shipping the product from the cross-dock i to distribution center j.

d_j: demand of the distribution center j.

Decision Variables. We have the following sets of binary variables to make the decisions about the opening of the cross-dock warehouse, and the distribution for the cross-dock warehouse to the distribution center:

$$Y_i = \begin{cases} 1 & \text{if location } i \text{ is used as a cross-dock warehouse,} \\ 0 & \text{otherwise,} \end{cases}$$
$$X_{ij} = \begin{cases} 1 & \text{if cross-dock } i \text{ supplies the demand of distribution center } j, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

W_{ki} is the amount of product sent from factory k to cross-dock i and is represented by continuous variables.

We can now state the mathematical model as a (P) problem

$$\min_{W_{ki}, Y_i, X_{ij}} Z = \sum_{k \in K} \sum_{i \in I} G_{ki} W_{ki} + \sum_{i \in I} F_i Y_i + \sum_{i \in I} \sum_{j \in J} C_{ij} d_j X_{ij} \tag{2}$$

subject to the following constraints:

Capacity of the plant:

$$\sum_{i \in I} W_{ki} \leq Q_k, \quad \forall k \in K. \tag{3}$$

Balance of product:

$$\sum_{j \in J} d_j X_{ij} = \sum_{k \in K} W_{ki}, \quad \forall i \in I. \tag{4}$$

Single cross-dock warehouse to distribution center:

$$\sum_{i \in I} X_{ij} = 1, \quad \forall j \in J. \tag{5}$$

Cross-dock warehouse capacity:

$$\sum_{j \in J} d_j X_{ij} \leq K_i Y_i, \quad \forall i \in I. \tag{6}$$

Demand of items:

$$p Y_i \leq W_{ki}, \quad \forall i \in I, \ \forall k \in K, \tag{7}$$

$$p = \min \left\{ d_j \right\}, \tag{8}$$

$$W_{ki} \geq 0, \quad \forall i \in I, \ \forall k \in K, \tag{9}$$

$$Y_i \in \mathbb{Z}, \quad \forall i \in I, \tag{10}$$

$$X_{ij} \in \mathbb{Z}, \quad \forall i \in I, \ \forall j \in J. \tag{11}$$

Objective function (2) considers in the first term the cost of shipping the product from the plants k to cross-dock warehouse i. The second term contains the fixed cost to open and operate the cross-dock warehouse i. The last term incorporates the cost to supply the demand of the distribution center j. Constraints (3) imply that all that is produced in plant k does not violate the capacity of plant k. Balance constraints (4) ensure that the amount of products that arrive at a distribution center j is the same as that sent to the plants k. The demand of each distribution center j will be satisfied by a single cross-dock warehouse i, which is achieved by constraints (5). Constraints (6) bound the amount of products that can be sent to a distribution center j from a cross-dock warehouse i that has been open. Finally, constraints (7) guarantee that any opened cross-dock warehouse i receives at least the minimum amount of demand for distribution centers j. The demand is satisfied by shipping via cross-dock warehouse with each distribution center being assigned exclusively to a single cross-dock warehouse. The possible locations for the cross-dock warehouses are given, but the particular facilities to be used will be selected as a result of the minimum total distribution cost. As a result, efficient load consolidation arises as an important opportunity for profitability in this work. To the best of our knowledge, the specific model we pose in this study is not addressed in the literature; however, the area of fixed charge network design is closely related. Our formulation facilitates the use of a Benders Decomposition framework for its solution.

4. Solution Methodology

Because of the economic importance of the network design problems and their combinatorial nature, several solution

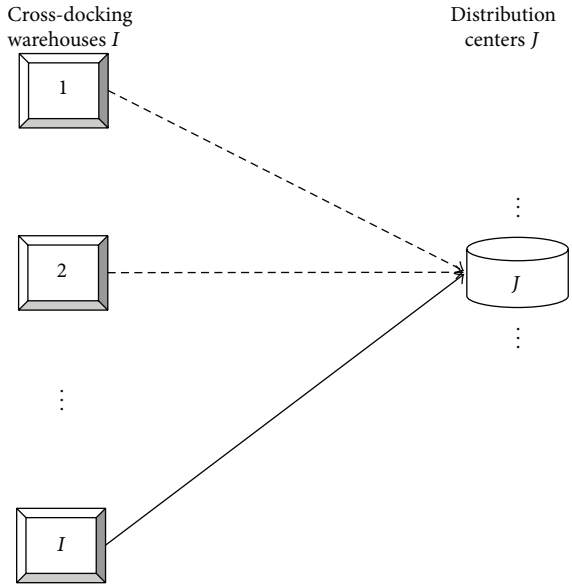


Figure 3: One cross-dock warehouse to one distribution center.

methodologies have been developed. In this paper, we describe a decomposition approach for solving the soda bottling company problem. The approach is based on Benders Decomposition, which is one of the most successful solution approaches. Model (2)–(11) presents a structure well-suited for a primal decomposition approach (Benders Decomposition).

This solution method is based on the situation that we can decompose the original problem obtaining several smaller and thus easier to solve subproblems. In this case, Benders Decomposition is used to meet the optimal location of intermediate warehouses between plants and distribution centers. As in [2], this model proposes that no customer zone is allowed to deal with more than one cross-dock warehouse, since X_{ij} must be 0 or 1 and not fractional. Additionally, we can mention that the real case presented has a large number of binary variables compared to the number of constraints. Thus, the Benders Decomposition is the better way to solve the optimal design of the distribution system, because it exploits the special structure of the original problem. In Figure 1 a graphical representation of the proposed distribution system is shown. This graphical illustration is different from the one presented in Figure 2. The difference between the current system and the proposed is because the demand is satisfied by shipping via cross-dock warehouse with each distribution center, which is assigned exclusively to a single cross-dock warehouse; see Figure 3.

4.1. Benders Decomposition. The optimal design of a soda bottling company distribution system is a problem that has a large number of complicating binary variables and thus involves a large CPU time to find an optimal integer solution; for this reason, we applied Benders Decomposition framework [17]. This method projects problem (2)–(11) onto the space defined by the binary variables X_{ij} and Y_i. The original problem P is decomposed into two different problems: a restricted master problem and Benders subproblem. The subproblem (SP) is a dual lineal problem (LP) of P and is obtained by fixing the variables to either 0 or 1. Benders subproblem provides an upper bound of the original problem. Let (SP) be the dual problem of P and the dual variables associated with constraints (3), (4), and (7), respectively. We have the following:

Subproblem (SP):

$$\begin{aligned}
\max_{\omega_k,\alpha_i,\beta_{ik}} \Phi = {} & \sum_{k\in K} Q_k\omega_k + \sum_{i\in I}\sum_{j\in J}\left(d_j\overline{X}_{ij}\right)\alpha_i + \sum_{k\in K}\sum_{i\in I}\overline{Y}_i\beta_{ik}, \\
& \omega_k + \alpha_i + \beta_{ik} \le G_{ki}, \quad \forall i \in I,\ \forall k \in K, \\
& \omega_k \le 0, \\
& \alpha_i \text{ unrestricted}, \\
& \beta_{ik} \ge 0.
\end{aligned} \tag{12}$$

Relaxed Master Problem (RMP):

$$\min_{Y_i,Xij} \Omega, \tag{13}$$

$$\Omega \ge \sum_{i\in I} F_iY_i + \sum_{i\in I}\sum_{j\in J}\left(C_{ij}d_jX_{ij}\right)\overline{\omega}_k + \sum_{i\in I}\sum_{k\in K} Y_i\overline{\beta}_{ik} + \sum_{i\in I}\sum_{j\in J}\left(d_jX_{ij}\right)\overline{\alpha}_i, \tag{14}$$

$$\sum_{i\in I} F_iY_i + \sum_{i\in I}\sum_{j\in J}\left(C_{ij}d_jX_{ij}\right)\overline{\omega}_k + \sum_{i\in I}\sum_{k\in K} Y_i\overline{\beta}_{ik} + \sum_{i\in I}\sum_{j\in J}\left(d_jX_{ij}\right)\overline{\alpha}_i \le 0, \tag{15}$$

$$\sum_{i\in I} X_{ij} = 1, \quad \forall j \in J, \tag{16}$$

$$\sum_{i\in I} X_{ij} \le 0, \quad \forall j \in J, \tag{17}$$

$$\sum_{j\in J} d_jX_{ij} \le K_iY_i, \quad \forall i \in I, \tag{18}$$

$$\sum_{j\in J} d_jX_{ij} \le 0, \quad \forall i \in I, \tag{19}$$

$$Y_i, X_{ij} \in \{0, 1\}. \tag{20}$$

The master problem is an integer problem and its variables are considered as complicating variables. To determine the values of these variables, the Benders master problem must be solved. Since the number of primal cuts is too large, Benders [17] proposed to solve a Relaxed Master Problem (RMP), by taking only a subset or Benders cuts, and to generate these cuts, one by one, in each iteration of the algorithm. Relaxed Master Problem provides feasible solutions through Benders optimality cuts (14) because they are based on optimality conditions of the subproblem and through valid constraints called feasibility cuts in every iteration of the algorithm. Feasibility cuts (15), (16), (17), (18), and (19) enforce necessary conditions for feasibility of the

Table 1: Comparison of the instances.

Instance	K	I	J	Continuous variables	Binary variables	Constraints
INST 0	2	2	2	4	6	10
INST 1	4	5	17	20	90	36
INST 2	4	10	17	40	180	51
INST 3	6	25	40	150	1025	121
INST-R CASE	44	56	254	2464	14280	466

{Initialization}
Y_i, X_{ij} := Fix integer variables to given feasible integer values
LB := $-\infty$
UB := $+\infty$
while UB − LB > ε **do**
  {solve subproblem SP}
  if Unbounded **then**
    Get unbounded ray $\overline{\omega}_k, \overline{\beta}_{ik}, \overline{\alpha}_i$
    Add feasibility cut to master problem (MP)
  else
    Get extreme point $\overline{\omega}_k, \overline{\beta}_{ik}, \overline{\alpha}_i$
    Add optimality cut to master problem (MP)
    UB := min{UB, SP}
  endif
  {solve master problem (MP)}
  LB := $\overline{\Omega}$
endwhile

Procedure 1: Benders Decomposition $(Y_i, X_{ij}, \omega_k, \beta_{ik}, \alpha_i)$.

primal subproblem. Master problem provides a lower bound of the original problem (P).

Benders Algorithm. A brief summary of the algorithm is given in Procedure 1 for completeness.

5. Computational Experience

The full scale model and the decomposition strategy proposed were implemented in GAMS [18] using the solver CPLEX [19] for MIP and LP problems (master problem and Benders subproblem). All mathematical models were carried out on an AMD Phenom II N970 Quad-Core with a 2.2 GHz processor and 4 GB RAM. Because the major difficulty of Benders method is the solution of master problem, and because [2] suggests that the Benders master problem should not be solved to optimality, we set GAMS parameter OPTCR at 0.0015; that is, the relative termination tolerance is within 0.15% of the best possible solution. In the first iteration of algorithm, we fixed all integer variables to 1. Additionally, the size of all MIP models was reduced through presolver phase of CPLEX. Benders algorithm stops when the values of lower bound and upper bound are equal, except for a small tolerance $\varepsilon = 0.15\%$:

$$\varepsilon = \left[\frac{(\mathrm{UB} - \mathrm{LB})}{\mathrm{UB}}\right] \cdot 100\%. \quad (21)$$

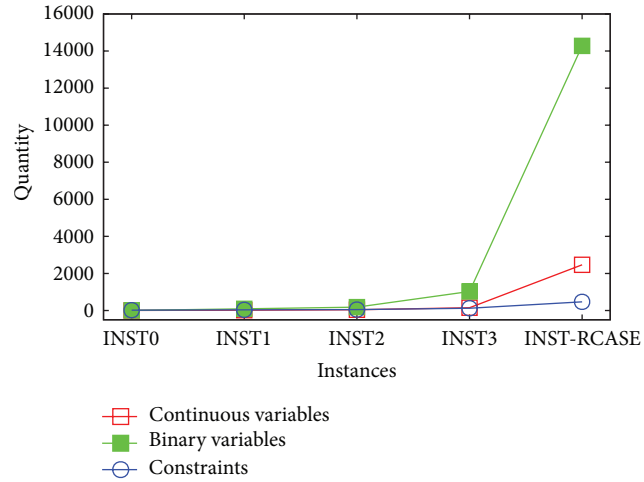


Figure 4: Model complexity.

To test the efficiency in terms of CPU time and quality of our solution method, we solved instances with data having demand distribution characteristics that reflect real applications. We generated randomly four instances with different number of plants, cross-dock warehouses, and distribution centers. Additionally, we solved a realistic case of a soda bottling company. Table 1 indicates the specific size of each instance.

The model complexity of each instance can be seen in Figure 4. This figure shows the increase in problem size in terms of the number of constraints, continuous variables, and binary variables.

5.1. Results. Benders Decomposition described in the previous section was used to design the optimal distribution system for the soda bottling company. Because the amount of memory and the computational effort needed to solve the instances grow significantly with the number of variables and constraints, we propose a decomposition strategy based on Benders Decomposition. The comparison results demonstrate that the ε-optimal solution is very close to the optimal solution (GAMS-CPLEX Solution).

Table 2 illustrates that the comparison results demonstrate that the CPU time of proposed decomposition strategy is less than direct solution with GAMS-CPLEX. The GAP is less than 0.2%:

$$\mathrm{GAP} = \left[\frac{(\mathrm{GAMS\ Solution} - \mathrm{Benders\ Decomposition\ Solution})}{\mathrm{GAMS\ Solution}}\right] \cdot 100\%. \quad (22)$$

Table 2: Performance of Benders Decomposition and direct solution of full scale problem.

Instance	Opt. value GAMS-CPLEX	CPU time (sec.)	LB	UB	GAP (%)	CPU time (sec.)
INST 0	1270	1.5	1270	1270	<0.2	0.5
INST 1	179170	2.5	179168	179168	<0.2	1.5
INST 2	485140	3.5	485142	485142	<0.2	2.5
INST 3	$1.00E+07$	930	$1.00E+07$	$1.00E+07$	<0.2	630
INST-R CASE	$8.45E+10$	4160	$8.43E+10$	$8.43E+10$	0.2	2160

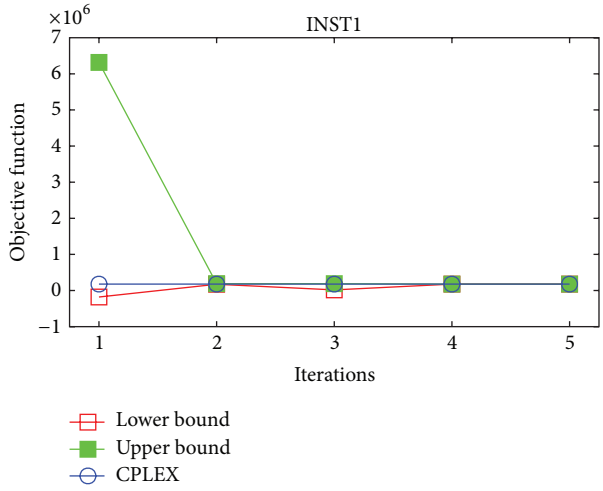


Figure 5: Benders versus full scale solution.

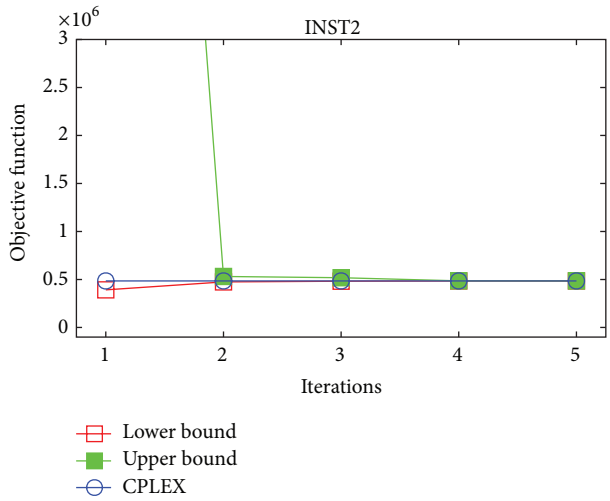


Figure 6: Benders versus full scale solution.

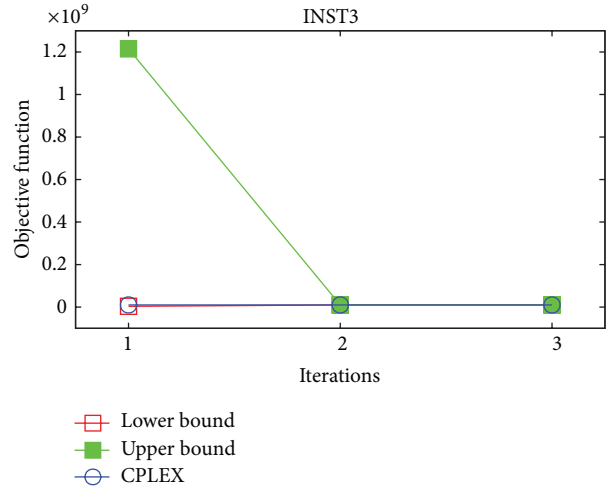


Figure 7: Benders versus full scale solution.

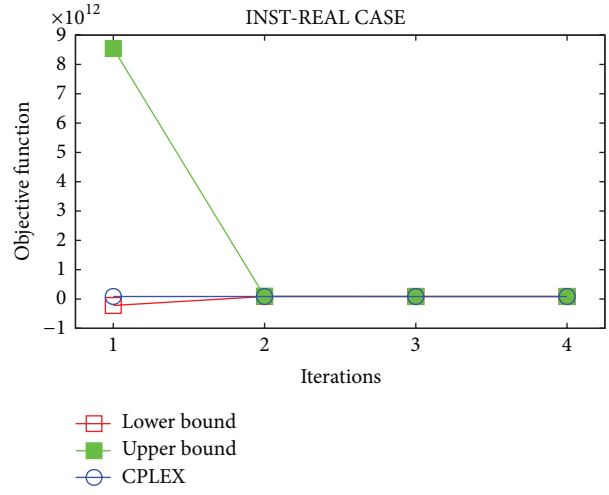


Figure 8: Benders versus full scale solution.

The progression of upper bound and lower bound values over the iterations around the optimum in all instances is plotted in Figures 5, 6, 7, and 8. Clearly, the use of feasibility cuts is very effective in speeding up convergence and they affect the quality of both the upper and the lower bounds. In our numerical studies, we observe similar high performance in convergence with feasibility cuts in other instances. For all the instances solved, tight lower and upper bounds were obtained with a small GAP in just a few iterations of the proposed algorithm.

These measures reported show that the performance of the Benders Decomposition seems to be indifferent to its size when feasibility cuts are implemented. The results of the experiments reported in Table 2 show that the proposed Benders procedure produces very good feasible solutions compared to the optimal/best available ones generated by CPLEX in significantly less CPU time.

6. Conclusions

In this paper, we present an optimal design of a soda bottling company distribution system problem. We proposed a decomposition strategy that shows acceptable convergence properties. In all instances, the number of iterations required to get convergence is less than five. The feasibility cuts allow reaching an ε-optimal solution. The total CPU time required to solve the large-scale case study based on the realistic company situation was less than 2200 seconds. This instance has 14280 binary variables, 2464 continuous variables, and 466 constraints. The proposed decomposition strategy was shown to be very efficient for the case study. For this class of problems, global optimality is not guaranteed in a reasonable time, but the solution through Benders Decomposition provides good feasible solutions and good bounds to the optimum.

Moreover, it is important to mention that the main objective was to compare the CPU effort and the quality bounds of the full scale solution through a commercial solver and the proposed decomposition strategy. We can finally conclude that we could propose a strategy for designing the distribution system of a soda bottling company in competitive CPU times. The solutions obtained are close to the optimal values reported by commercial optimizers. This ensures having a flexible and scalable solution tool when the company decides to increase the number of its facilities.

Conflict of Interests

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

References

[1] I. Soria, *Rediseño de la cadena de abastecimiento de un grupo embotellador de bebidas [M.S. thesis]*, ITESM Campus Toluca, Toluca, Mexico, 2008.

[2] A. M. Geoffrion and G. W. Graves, "Multicommodity distribution system design by benders decomposition," *Management Science*, vol. 20, no. 5, pp. 822–844, 1974.

[3] M. T. Melo, S. Nickel, and F. S. da-Gama, "Facility location and supply chain management—a review," *European Journal of Operational Research*, vol. 196, no. 2, pp. 401–412, 2009.

[4] A. Klose and A. Drexl, "Facility location models for distribution system design," *European Journal of Operational Research*, vol. 162, no. 1, pp. 4–29, 2005.

[5] A. Cintron, A. R. Ravindran, and J. A. Ventura, "Multi-criteria mathematical model for designing the distribution network of a consumer goods company," *Computers & Industrial Engineering*, vol. 58, no. 4, pp. 584–593, 2010.

[6] M. S. Jabalameli, B. Bankian Tabrizi, and M. Javadi, "Capacitated facility location problem with variable coverage radius in distribution system," *International Journal of Industiral Engineering & Production Research*, vol. 21, no. 4, pp. 231–237, 2010.

[7] P. N. Thanh, O. Péton, and N. Bostel, "A linear relaxation-based heuristic approach for logistics network design," *Computers & Industrial Engineering*, vol. 59, no. 4, pp. 964–975, 2010.

[8] V. Jayaraman and H. Pirkul, "Planning and coordination of production and distribution facilities for multiple commodities," *European Journal of Operational Research*, vol. 133, no. 2, pp. 394–408, 2001.

[9] H. Sadjady and H. Davoudpour, "Two-echelon, multi-commodity supply chain network design with mode selection, lead-times and inventory costs," *Computers and Operations Research*, vol. 39, no. 7, pp. 1345–1354, 2012.

[10] H. Pirkul and V. Jayaraman, "Production, transportation, and distribution planning in a multi-commodity tri-echelon system," *Transportation Science*, vol. 30, no. 4, pp. 291–302, 1996.

[11] V. Jayaraman, "An efficient heuristic procedure for practical-sized capacitated warehouse design and management," *Decision Sciences*, vol. 29, no. 3, pp. 729–745, 1998.

[12] H. Pirkul and V. Jayaraman, "A multi-commodity, multi-plant, capacitated facility location problem: formulation and efficient heuristic solution," *Computers & Operations Research*, vol. 25, no. 10, pp. 869–878, 1998.

[13] H. Üster and H. Agrahari, "A Benders decomposition approach for a distribution network design problem with consolidation and capacity considerations," *Operations Research Letters*, vol. 39, no. 2, pp. 138–143, 2011.

[14] W. Jiang, L. Tang, and S. Xue, "A hybrid algorithm of tabu search and benders decomposition for multi-product production distribution network design," in *Proceedings of the IEEE International Conference on Automation and Logistics (ICAL '09)*, pp. 79–84, August 2009.

[15] H. M. Bidhandi, R. M. Yusuff, M. M. H. M. Ahmad, and M. R. Abu Bakar, "Development of a new approach for deterministic supply chain network design," *European Journal of Operational Research*, vol. 198, no. 1, pp. 121–128, 2009.

[16] J.-F. Camacho-Vallejo, R. Muñoz-Sánchez, and J. L. González-Velarde, "A heuristic algorithm for a supply chain's production-distribution planning," *Computers & Operations Research*, vol. 61, pp. 110–121, 2015.

[17] J. F. Benders, "Partitioning procedures for solving mixed-variables programming problems," *Computational Management Science*, vol. 2, no. 1, pp. 3–19, 2005.

[18] A. Brooke, D. Kendrick, and A. Meeraus, *GAMS: A User's Guide*, Boyd & Fraser Publishing Company, 1998.

[19] GAMS Development Corporation, *GAMS: The Solver Manuals; CONOPT; CPLEX; DICOPT; LAMPS; MILES; MINOS; OSL; PATH; XA; ZOOM*, GAMS Development Corporation, 1994.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 642021, 9 pages
http://dx.doi.org/10.1155/2015/642021

Research Article

The Research of Car-Following Model Based on Real-Time Maximum Deceleration

Longhai Yang, Xiqiao Zhang, Jiekun Gong, and Juntao Liu

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

Correspondence should be addressed to Jiekun Gong; gongjiekun@126.com

Received 3 April 2015; Accepted 18 June 2015

Academic Editor: Yuanchang Xie

This paper is concerned with the effect of real-time maximum deceleration in car-following. The real-time maximum acceleration is estimated with vehicle dynamics. It is known that an intelligent driver model (IDM) can control adaptive cruise control (ACC) well. The disadvantages of IDM at high and constant speed are analyzed. A new car-following model which is applied to ACC is established accordingly to modify the desired minimum gap and structure of the IDM. We simulated the new car-following model and IDM under two different kinds of road conditions. In the first, the vehicles drive on a single road, taking dry asphalt road as the example in this paper. In the second, vehicles drive onto a different road, and this paper analyzed the situation in which vehicles drive from a dry asphalt road onto an icy road. From the simulation, we found that the new car-following model can not only ensure driving security and comfort but also control the steady driving of the vehicle with a smaller time headway than IDM.

1. Introduction

Driving assistance systems and vehicle-sensor designs are becoming mature along with the development of intelligent transportation. Adaptive cruise control (ACC), which is a component of driving assistance systems, is gradually replacing part of the driving control of the human driver. ACC not only alleviates driving fatigue but also improves driving safety. The operation of ACC depends on the information on the vehicle motion and the external environment, which is detected by vehicle sensors. Therefore, reasonable and effective use of sensor data can improve the car-following safety and stability of ACC and can improve the road capacity effectively.

The tire-road friction coefficient, which determines the real-time maximum deceleration, is a significant parameter for ensuring the vehicle driving safely and avoiding rear-end collisions. The estimation methods can be divided into two types: direct detection with sensors and vehicle dynamic model estimation. Direct detection methods, such as detecting the coefficient with optical or acoustic sensors, need expensive sensors which limit the use of the method [1]. Its reliability and robustness is low, because the detection precision varies with the environment. The estimation of the coefficient with the vehicle dynamic model has higher robustness, but the accuracy still needs to be improved [2]. In order to improve the estimation accuracy of the latter method, the combination of a vehicle dynamic model with data fusion of multiple sensors has become a hot topic of research [3]. Recently, estimation methods with vehicle dynamic models, which include slip-slope methods, Kalman filter-based friction coefficient estimation, and methods based on lateral dynamics, have concentrated on estimating the mean coefficient value of four wheels [4]. However, the latest researches indicate that friction coefficient estimation of individual wheels has greater value [5].

The friction coefficient is used in vehicle collision-avoidance systems, antilock braking systems (ABS), and electronic stability systems (ESC). The real-time friction coefficient has important significance for influencing the car-following behavior and guaranteeing traffic safety. But there are few researches which have included the real-time friction coefficient in car-following models until now. Most of the traditional car-following models regard the maximum deceleration as a constant value and lack consideration of the real-time road conditions [6]. In the following controls of ACC and CACC, the driver can choose different modes according to the real-time driving environment that the driver can

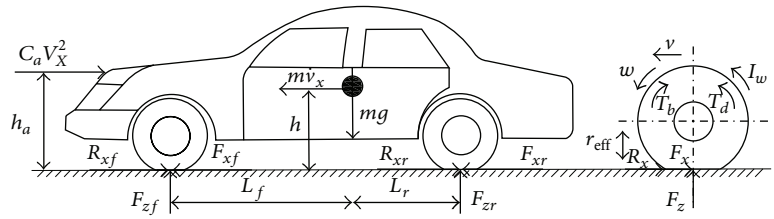


Figure 1: Longitudinal vehicle dynamics schematic.

perceive. Different modes have corresponding gaps. For the ACC system, the available gap settings are 1.1, 1.6, and 2.2 s. For the CACC system, the available gap settings are 0.6, 0.9, and 1.1 s [7]. The gap controls of ACC and CACC ignore differentiation between different vehicles on the same road. Therefore, the gap should be optimized based on the detected road condition.

In this paper, we propose a new car-following model based on the maximum deceleration, which is estimated with the vehicle dynamic model. This paper is organized as follows. The study of the car-following model in real-time conditions is presented in Section 2. In Section 3, the approach of real-time friction coefficient estimation is introduced. In Section 4, the maximum deceleration is calculated according to the longitudinal vehicle model. The new car-following model is established in Section 5 and simulated in Section 6.

2. Real-Time Conditions in Car-Following Model

Generally, the road condition is defined as a congested road or an uncongested road. The car-following models have different parameter values, which are calibrated by experimental data, in the different road conditions [8]. Kesting et al. summarized traffic situations as free traffic, upstream front, congested traffic, bottleneck, and downstream front [9]. The car-following model detected the real-time traffic situation and used corresponding parameters, which differed under different situations, to avoid congestion actively.

Soria et al. [10] assessed four car-following models with field data which were collected under different traffic (congested versus uncongested) and weather conditions (rain versus clear sky). The four car-following models were Gipps, Pitt, MITSIM, and modified Pitt. The parameters of the four car-following models were calibrated by field data, and the corresponding parameters were applied to evaluate the simulation precision of four car-following models under different traffic and weather conditions.

Tang et al. [11, 12] developed a car-following model to study the influence of real-time road conditions on driving behavior. The real-time road conditions, which were divided into good, moderate, and bad, were regarded as a separate parameter modifying the full velocity difference (FVD) model. A good road was defined as having no influential factors along the road (e.g., expressway and freeway). A bad road was defined as having many influential factors along the road (e.g., bus station, sidewalk).

The above works analyzed the effect of real-time traffic, weather, and road conditions on car-following by adjusting the parameters of the models. But the real-time road surface conditions were omitted. The real-time road surface condition can be expressed as the tire-road friction coefficient or the theoretical maximum deceleration. In ACC and collision-avoidance systems, the tire-road friction coefficient is used to adjust the braking stance [13]. In the research on car-following, only the collision-avoidance model takes into account the real-time road surface condition [14, 15]. The theoretical maximum deceleration is used to calculate the safe following distance.

3. Friction Coefficient Estimation

3.1. Longitudinal Vehicle Model. A longitudinal vehicle model provides theoretical support for individual wheel friction coefficient estimation. In this paper, we assume that the vehicle performs a linear motion on a level road and other degrees of freedom of motion are ignored. The static force models of the vehicle are shown in Figure 1.

The equations of the longitudinal vehicle model are given by

$$m\dot{v}_x = F_x - R_x - C_a v_x^2, \tag{1}$$

$$I_\omega \dot{\omega}_i = T_{di} - T_{bi} - r_{\text{eff}} F_{xi}, \tag{2}$$

$$F_{zf} = \frac{mgL_r - m\dot{v}_x h - C_a v_x^2 h_a}{L}, \tag{3}$$

$$F_{zr} = \frac{mgL_f + m\dot{v}_x h + C_a v_x^2 h_a}{L}, \tag{4}$$

$$s_i = \frac{|\omega_i r_{\text{eff}} - v_x|}{\max(\omega_i r_{\text{eff}}, v_x)}, \tag{5}$$

$$\gamma = \frac{F_{xi}}{F_{zi}}, \tag{6}$$

where m is the mass of the vehicle, $i = fl, fr, rl, rr$ are used to separately represent the four wheels of the vehicle, v_x is the longitudinal speed, F_x is the total longitudinal tire force, which is the summation of the tire forces generated at four tires, F_{zi} is the normal forces of each wheel, $F_{zfl} = F_{zfr} = F_{zf}/2$, $F_{zrl} = F_{zrr} = F_{zr}/2$, R_x is the rolling resistance, C_a is an aerodynamic drag parameter, I_ω is the rotational dynamics of each wheel, ω is the wheel speed, T_d and T_b represent the drive and brake torques, r_{eff} represents the effective radius of the tire, $L = L_f + L_r$ is the wheelbase, L_f represents the horizontal distance between the center of gravity and the front wheel, L_r represents the horizontal distance between the center of

gravity and the rear wheel, h and h_a are the heights of the center of gravity and wind loading above the ground, and γ is the normalized traction force of the tire.

The slip ratio s_i is calculated by (5). During braking, v_x, which is greater than $\omega_i r_{\text{eff}}$, is used as the denominator of (5). During acceleration, $\omega_i r_{\text{eff}}$, which is greater than v_x, is used as the denominator of (5) [5].

The longitudinal tire force varies with the slip ratio, friction coefficient, and normal force. In some researches on ABS, in order to improve the efficiency of control and to shorten the braking distance, the slip ratio is adjusted by ABS to maximize the normalized traction force of the tire [16].

3.2. Friction Coefficient Calculation. The approach which is used in this paper to estimate the individual wheel friction coefficient consists of the following three steps:

(1) estimating the longitudinal tire force;

(2) measuring the longitudinal slip ratio at the wheel;

(3) using a recursive least-squares parameter identification algorithm to calculate the tire-road friction coefficient.

The approaches of longitudinal tire force and longitudinal slip ratio estimation mainly perform friction estimation using GPS and torque measurements, torque measurements and an accelerometer, and GPS and an accelerometer. The first approach has greater estimation accuracy than the other two approaches [1]. Therefore, the first approach is utilized to estimate the longitudinal tire force and longitudinal slip ratio in this paper.

In a small slip ratio interval, where the slip ratio is smaller than 0.15 and when the slip ratio is greater than 0.005, the normalized traction force has a linear relationship with the slip ratio. For any given road, this linear relationship in the range of a small slip ratio is established [1]

$$\gamma = \frac{F_x}{F_z} = Ks. \tag{7}$$

K presents the slip-slope, whose value changes with the road surface condition and the type of tire. According to previous experimental data, it is found that the tire-road friction coefficient has a linear relationship with the slip-slope [5]. This linear relationship is shown as

$$u = AK + C, \tag{8}$$

where $A = 0.026$ is the proportionality constant and $C = 0.047$ is a bias constant.

According to (8), we know that the slip-slope is the key to estimate the tire-road friction coefficient and its estimation accuracy determines the estimation accuracy of the tire-road friction coefficient. Consequently, a recursive least-squares parameter identification algorithm is utilized to improve the estimation accuracy.

4. Maximum Deceleration Estimation

ESC can make each wheel reach the peak of the normalized traction force by achieving the best slip ratio. The tire-road friction coefficients of the front and rear wheels correspond to u_f and u_r. We assume that the vehicle brakes with the maximum deceleration that the vehicle can attain during emergency braking on a given road. Combining (1), (3), (4), and (6), the following equation can be obtained:

$$\begin{aligned}&\frac{u_r}{L}\left(mgL_f + ma_{\max}h + C_a v_x^2 h_a\right)\\&\quad + \frac{u_f}{L}\left(mgL_r - ma_{\max}h - C_a v_x^2 h_a\right) - R_x - C_a v_x^2\\&\quad = ma_{\max}.\end{aligned} \tag{9}$$

Simplifying (9), the maximum deceleration can be presented as

$$\begin{aligned}a_{\max} &= \frac{L_r u_f + L_f u_r}{L + h\left(u_f - u_r\right)}g\\&\quad + \frac{C_a v_x^2 h_a\left(u_r - u_f\right) - R_x L - C_a v_x^2 L}{m\left[L + h\left(u_f - u_r\right)\right]}.\end{aligned} \tag{10}$$

The air resistance and the rolling resistance account for only a small proportion of the longitudinal force. Ignoring the air resistance and the rolling resistance, the maximum deceleration can be presented as

$$a_{\max} = \frac{L_r u_f + L_f u_r}{L + h\left(u_f - u_r\right)}g. \tag{11}$$

5. Car-Following Model

The change of real-time maximum deceleration is not significant when a vehicle is steering in a similar section. It is difficult for the driver to judge a subtle change in maximum acceleration through his or her own senses and to adjust the following gap through a subtle change. However, a computer can detect subtle changes of maximum acceleration and then control the throttle valve and brake valve. Consequently, the car-following model considered the real-time maximum deceleration should be applied in ACC or automatic driving. Automatic driving technology is still in its infancy. ACC technology has gradually become mature. In the current study of ACC, the vehicle controls its speed according to the distance and the relative speed between itself and leading vehicle [17]. The ACC system designs different following gaps which the driver can manually select, but its presupposed gap value is constant [7]. The ACC system fails to automatically calculate a more reasonable desired gap according to the real-time road surface conditions.

In order to guarantee computational efficiency, the model which is utilized in ACC should have only a few parameters and should be simple on the basis that the model can control the vehicle effectively and securely. The IDM is an eligible model which meets these criteria [9].

5.1. IDM. The IDM is expressed by

$$\dot{v}(t) = a_0\left[1 - \left(\frac{v(t)}{v_0}\right)^4 - \left(\frac{s^*\left(v(t), \Delta v(t)\right)}{s}\right)^2\right]. \tag{12}$$

This expression can be divided into two parts. The first two terms on the right, $\dot{v}_{\text{free}}(t) = a_0(1 - (v(t)/v_0)^4)$, mean the acceleration when the vehicle is driving on a road without congestion. The last part, $\dot{v}_{\text{brake}}(t) = -a_0(s^*/s)^2$, means the acceleration which is dominant when the following vehicle approaches the leading vehicle and the following vehicle must decelerate to avoid a rear-end collision

$$s^*(v(t), \Delta v(t)) = s_0 + v(t)T + \frac{v\Delta v}{2\sqrt{a_0 b}}, \tag{13}$$

where $s^*(v(t), \Delta v(t))$ expresses the desired minimum distance of car-following and is rewritten as s^* in the following text. a_0 is the desired maximum acceleration, b is the desired deceleration, $v(t)$ and v_0 are the actual speed and desired speed, Δv is the speed difference between the following vehicle and the leading vehicle, s is the actual gap, s_0 indicates that the minimum distance in congested traffic is 2 m, and the desired safe time headway T is equal to 1.5 s for a car on an uncongested road.

5.2. Dynamic Properties of IDM. IDM was established to simulate the car-following behavior on a freeway [18]. v_0 takes the value of 120 km/h for a car. IDM was analyzed for the four following situations:

(1) equilibrium traffic: in equilibrium traffic of arbitrary density, $\dot{v}(t) = 0$ and $\Delta v = 0$;

(2) low density: in this situation, s is very large, the term $(s^*/s)^2$ of (12) is negligible, and the vehicle speeds up to the desired speed;

(3) braking as a reaction to high approach rates: when a vehicle approaches slower or standing vehicles at a sufficiently high approach rate, the term $s_0 + vT$ of the desired minimum distance can be neglected;

(4) braking in response to small gaps: the gap s is much smaller than s^*, but there are no large velocity differences.

In order to directly analyze the IDM, (12) is transformed into

$$\dot{v}(t) = a_0\left[1 - \lambda\left(\frac{v(t)}{v_0}\right)^4 - \beta\left(\frac{s^*}{s}\right)^2\right]. \tag{14}$$

In the first mode, with $\lambda = 0$ and $\beta = 1$, the IDM is equivalent to the proportional derivative (PD) control law [19]. In the second mode, with $\lambda = 1$ and $\beta = 0$, the IDM is similar to the cruise control law, and the motion has nothing to do with the desired time headway T [7]. In the third and fourth modes, the IDM controls deceleration under the range of desired deceleration. In an emergency braking situation, the vehicle also decelerates more strongly than the desired deceleration to avoid a rear-end collision.

In the applicability analysis of IDM, the assumptions are strictly limited [18, 20]. Nevertheless, the limited assumptions cannot fully represent the actual operation. When the fleet moves at an even and high speed, this situation in which $\dot{v}(t) = 0$, $\Delta v = 0$, and $v(t) \to v_0$ is an equilibrium traffic situation. In this situation, (12) should be expressed as $\dot{v}(t) = -a_0(s^*/s)^2$, and if and only if $s \gg s^*$, we have $\dot{v}(t) \to 0$. However, when the fleet moves with the desired minimum following distance in this situation, $\dot{v}(t) = -a_0 < 0$, which contradicts the precondition $\dot{v}(t) = 0$.

In the judgment of car-following behavior, traffic flow theory describes car-following as driving with a gap that is less than 125 m [21]. Weidemann and Reiter consider that the vehicle continues car-following when the gap is less than 150 m [22]. We can clearly see that the IDM neglects judgment of car-following behavior. When a vehicle is driving in low density traffic, the vehicle is independent of the vehicle in front and the car-following model becomes invalid. For the ACC system, the computer changes the gap regulation controller into a gap-closing controller when the gap is greater than the gap threshold. Gap control and cruise control are separated to control the vehicle steadily.

5.3. Desired Minimum Distance Modification. The ACC system tends to design a comfortable system, and relatively large headways are applied [23]. CACC achieves smaller headways and more moderate natural driving due to the wireless communication between vehicles [24]. A small headway can improve the road capacity. Therefore, we should try to reduce the headway on the basis of ensuring passenger comfort and safety.

The term $v\Delta v/2\sqrt{a_0 b}$ of the desired minimum distance does not take into account the effect of real-time road surface conditions on the braking distance, and the calculated result is greater than the minimum braking distance. The comfort of passengers is ensured on the basis of avoiding rear-end collisions because the greater desired following distance leads to moderate braking. A desired following distance that is smaller than the result of (13) might trigger extreme deceleration during emergency braking. However, the vehicle performs emergency braking only if the leading vehicle performs emergency braking in the actual operation of the vehicles. The probability of emergency braking is relatively small. Consequently, reducing the following gap on the basis of ensuring driving safety to improve the road capacity is a desirable method.

Based on the above analysis, the minimum desired following distance is modified to

$$s_l^* = s_0 + v(t)T + \frac{v^2(t)}{2a_{\max}} - \frac{v_l^2(t)}{2a_{l\max}}, \tag{15}$$

where $a_{\max}$ and $a_{l\max}$ present the maximum deceleration of the following vehicle and the leading vehicle. The last two terms of (15) express the minimum distance necessary to avoid a rear-end collision. Taking into account that wireless communication is not mature, we assume that $a_{\max}$ is equal to $a_{l\max}$; that is, $a_{\max} = a_{l\max}$.

In a cut-in scenario, if the minimum desired following distance is calculated by (15), excess adjustment might occur. But the IDM has better performance. These two different results are illustrated in Figure 5. Consequently, in order to guarantee the comfort of passengers, the minimum desired

following distance model of the IDM is retained by the modified model in the cut-in scenario.

There are two different minimum desired following distance models in the cut-in scenario and the other scenario. Before turning to the minimum desired following distance model, the cut-in scenario should be identified. In this paper, the cut-in scenario is defined as a vehicle changing lane and driving into the middle of a two-vehicle fleet which is moving faster than 6 m/s. When a vehicle cuts into a two-vehicle fleet, the following gap difference between the subsequent following gap and the original following gap is greater than the possible maximum braking distance in the detection period. Therefore, the control law turning into the cut-in scenario should be given as

$$s(t-\Delta t)-s(t)>a_{\max}\cdot\Delta t, \tag{16}$$

where Δt presents the detection period. The cut-in scenario stops and the other scenario begins once the actual gap is close to the desired minimum following distance.

5.4. New Car-Following Model. According to the above analysis, (12) is modified as follows:

$$\dot{v}(t)=\begin{cases} a_0\left[1-\left(\dfrac{v(t)}{v_0}\right)^4\right], & s>s_T,\\ a_0\left[1-\left(\dfrac{s_l^*}{s}\right)^2\right], & s\le s_T, \end{cases} \tag{17}$$

where s_T represents the threshold of judging the car-following behavior and is equal to 125 m. When s is smaller than s_T, the gap controller is triggered, and when the reverse is true, the cruise control is triggered. Equation (17) resolves the contradiction which occurs when a vehicle is moving at an even and high speed. The third and fourth following modes also can be controlled well.

Combining (11), (15), and (17), the new car-following model is expressed as

$$\begin{aligned} \dot{v}(t)&=a_0\left[1-\left(\frac{s_l^*}{s}\right)^2\right],\\ s_l^*&=s_0+v(t)T+k1\left(\frac{v^2(t)}{2a_{\max}}-\frac{v_l^2(t)}{2a_{l\max}}\right)\\ &\quad+k2\frac{v\Delta v}{2\sqrt{a_0 b}},\\ a_{\max}&=\left[\frac{l_r u_f+l_f u_r}{L+h\left(u_f-u_r\right)}\right]*g, \end{aligned} \tag{18}$$

where

$$k1=\begin{cases}1, & \text{others},\\ 0, & \text{cut-in},\end{cases}\qquad k2=\begin{cases}0, & \text{others},\\ 1, & \text{cut-in}.\end{cases} \tag{19}$$

In the new car-following model, if the actual speed has reached the desired speed, the vehicle will continue to speed up even if s^* is smaller than s under the control of (18). This result is not in accordance with reality. To avoid this problem, we add to (18) the constraint that the vehicle maintains a uniform speed even when the result of (18) is a positive number.

5.5. Equilibrium Flow-Density Relation. In equilibrium traffic, $\dot{v}(t)=0$, $v(t)=v_l(t)$, and $a_{\max}=a_{l\max}$. The fleet moves stably with the minimum desired following distance; namely, $s=s_l^*=s_0+v(t)T$. The stable velocity can be presented as $v_e=\min(v_o,(s-s_0)/T)$. From a macroscopic point of view, the equilibrium traffic can be characterized by the stable traffic flow $Q=\rho v$ as a function of the traffic density ρ. According to the relation between gap and density, $1000/\rho-l=s$ (l represents the length of the vehicle, which is 5 m in this paper); the relation between flow and density of the new model is shown in Figure 2.

Figure 2(a) shows the relation between flow and density under the premise that the desired speed is 120 km/h. We can clearly see that the theoretical road capacity of the new model is greater than that of the IDM at the same desired speed and time headway. The theoretical road capacity is equal to the maximum flow. The theoretical capacity of the new model increases with decreasing the time headway. Figure 2(b) shows the relation between gap and density under the premise that the time headway is 1.5 s. We find that the theoretical capacity of the new model increases with increasing the desired speed, but the growth rate is negligible. Therefore, we can conclude that reducing the time headway contributes to improving the road capacity more than improving the desired speed.

6. Simulation

In this section, we firstly simulate the new model and the IDM in normal road surface conditions to test the applicability of the new model. Simulations of the new model and the IDM in changing road surface conditions are also carried out to compare the performance of the new model and IDM in real-time road surface conditions. All of the simulations are conducted with MATLAB/Simulink.

6.1. Car-Following Properties in Normal Road Surface Conditions. We carried out two simulations in normal road surface conditions. The first experiment aims to test and contrast the stabilities of the new model and IDM in normal traffic. The second experiment aims to test and contrast the stabilities of the new model and IDM in the cut-in scenario. In these two simulations, the vehicles are driving on a dry asphalt road and the difference between tires is ignored. We assume that the maximum deceleration is 7 m/s^2.

(1) Normal Traffic. In normal traffic, to observe the car-following behavior, a four-vehicle fleet is used and moves at high speed. In the first cycle, the leading vehicle drives at 25 m/s for 20 s and then accelerates at a constant acceleration

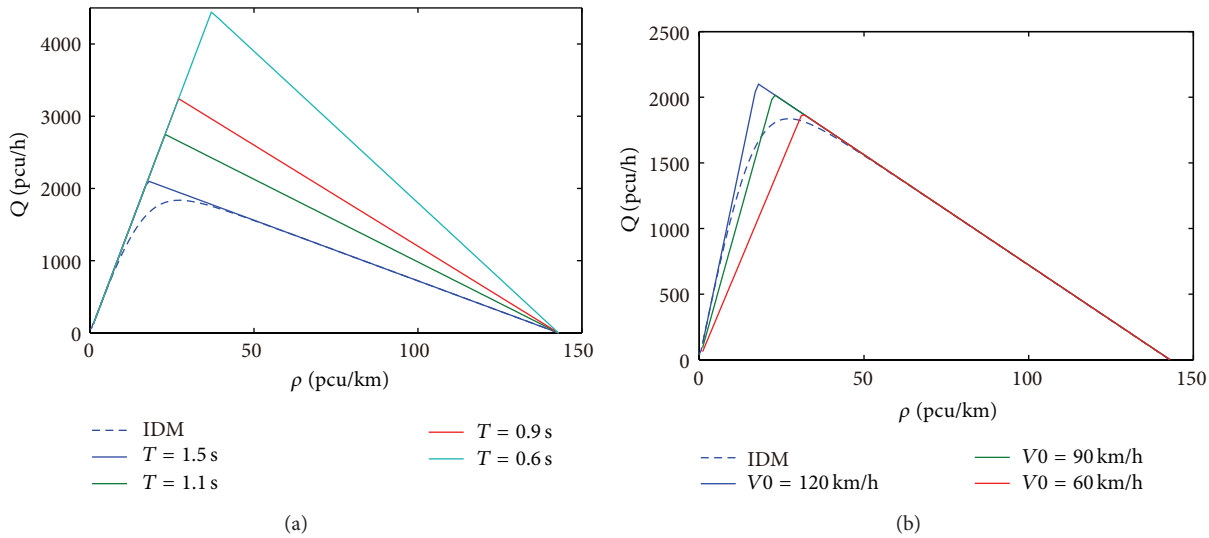


Figure 2: Equilibrium flow-density relation.

of 0.25 m/s^2 to reach 30 m/s. The vehicle decelerates at a constant deceleration of 0.25 m/s^2 to return to 25 m/s and remains at a constant speed for 20 s. In the second cycle, it accelerates at a constant acceleration of 0.33 m/s^2 to reach 30 m/s and maintains a constant speed for 20 s. Then, the vehicle decelerates at a constant deceleration of 0.5 m/s^2 to return to 25 m/s and maintains a constant speed for 15 s. In the third cycle, it accelerates at a constant acceleration of 0.5 m/s^2 to reach 30 m/s and maintains a constant speed for 10 s. Then, the vehicle decelerates at a constant deceleration of 1 m/s^2 to return to 25 m/s and maintains a constant speed for 10 s. In the last cycle, it accelerates at a constant acceleration of 1 m/s^2 to reach 30 m/s and maintains a constant speed for 20 s. The three other following vehicles run with the new model control and IDM control. For the comparison of the new model and IDM, the same desired speed (120 km/h) and the same desired safe time headway (1.5 s) are used. At the beginning of the fleet movement, the four vehicles have the same initial speed of 25 m/s. And the three following vehicles have the same following gap of 39.5 m.

Figures 3 and 4 show the car-following with the IDM control and new model control. It can be found from Figures 3(a) and 4(a) that the response time of the IDM is longer than that of the new model. The longer response time means the following vehicle must maintain a sufficient gap to avoid a rear-end collision. This result is consistent with Figures 3(c) and 4(c). The IDM keeps the actual time headway of the fleet around 2.1 s compared with 1.6 s in the new model. The longer response time of the IDM controller determines that the vehicle needs a larger car-following gap and time headway, but its corresponding acceleration and deceleration are milder than those of the new model controller. It can be found from Figures 3(b) and 4(b). Table 1 shows the parameters selected from the new model and the IDM controller. We can find that the scope of acceleration and deceleration of the new model controller is greater than that of the IDM controller, but it is in the acceptable range and the comfort of passengers is still good. Consequently, the new model is suitable and more effective than the IDM from the capacity point of view.

Table 1: Selected parameters of the new model and the IDM controller.

		Headway (s)	Gap (m)	Deceleration (m/s^2)
AVG	New model	1.60	43.67	−0.21
	IDM	2.11	57.12	−0.17
MAX	New model	1.84	56.60	−1.05
	IDM	2.49	74.05	−0.70

(2) Cut-In Scenario. When two vehicles are driving in the same lane and other vehicles want to enter the lane between them, sudden and unexpected cut-ins will occur in a short period of time. The sudden decrease of the following gap should lead to extreme deceleration for safety. Consequently, the braking efficiency should be considered to test the performance of the new model and the IDM in the cut-in scenario. For the sake of clarity, a two-vehicle fleet response in the cut-in scenario is shown in Figure 5.

The leading vehicle and the following vehicle have the same initial speed of 25 m/s and the initial gap is 40 m. The leading vehicle is driving at a constant speed (25 m/s) and the following vehicle is tracking the leader's speed. Around the 60th second, a cut-in vehicle merges between them, and the following vehicle and the new leading vehicle make up a new two-vehicle fleet. To demonstrate the necessity of cut-in judgment, the minimum desired following distance model of the other scenario, which is also shown in Figure 5 ($k1 = 1$, $k2 = 0$), is simulated in the cut-in scenario.

It is obvious that the IDM has the best performance. The peak of the following deceleration of the IDM is near 1.9 m/s^2. In the new model ($k1 = 0$, $k2 = 1$), when the following

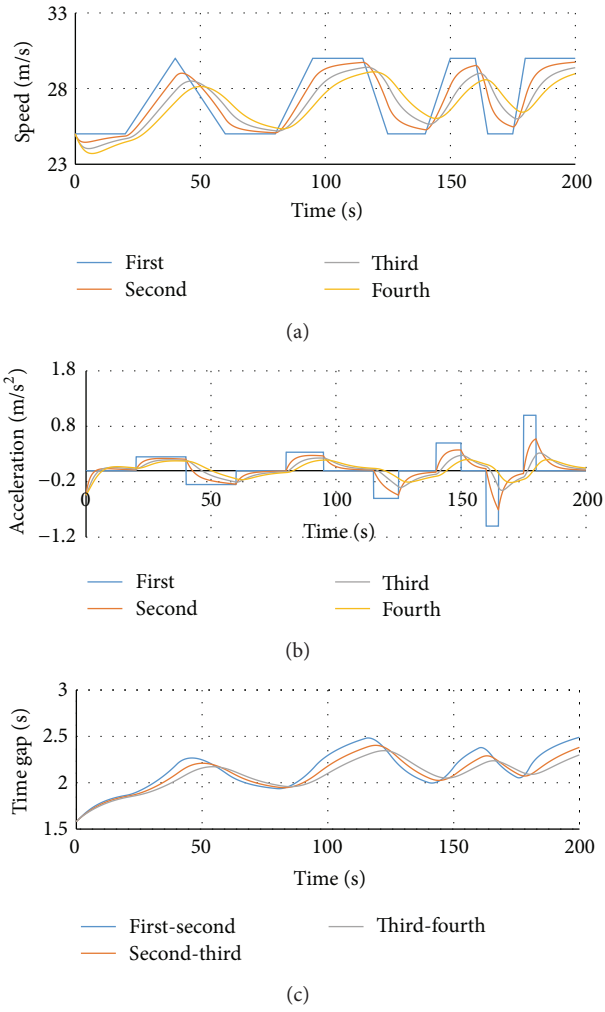


FIGURE 3: Four-vehicle car-following properties with IDM control in normal traffic.

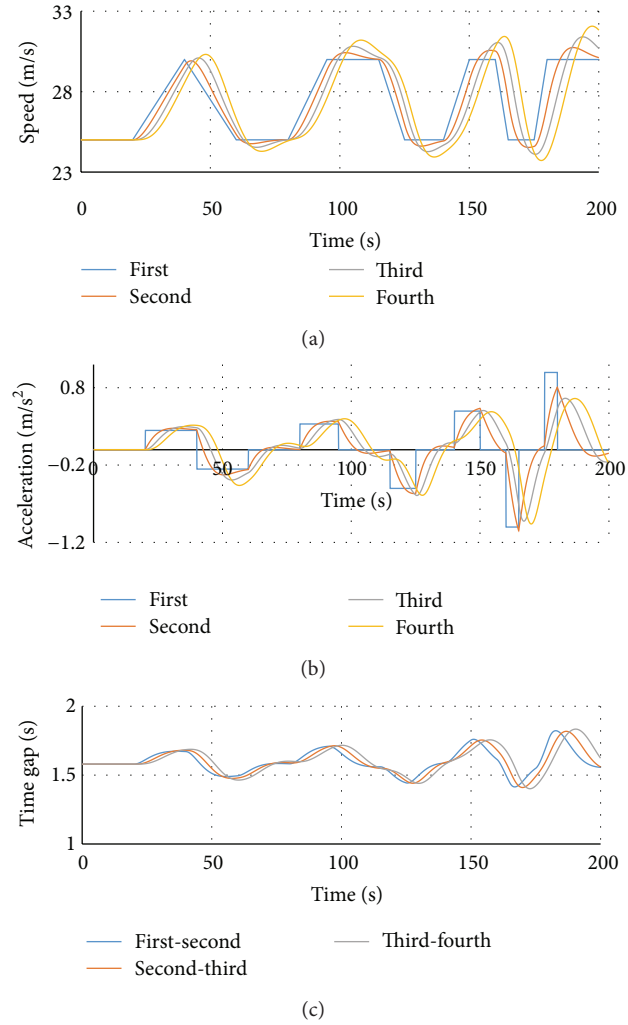


FIGURE 4: Four-vehicle car-following properties with new model control in normal traffic.

vehicle detects the cut-in vehicle, the following vehicle brakes to increase the gap between itself and the new leading vehicle. The peak deceleration of the following vehicle in the new model is near 4.3 m/s^2 and is greater than that in the IDM, but it is in the range of acceptable deceleration. The new model has smaller following gap than IDM; therefore, the vehicle controlled by new model brakes stronger to avoid a rear-end collision. However, the probability of cut-in scenario is small, because the small following gap of the new model does not provide chance for other vehicles to cut in.

By comparing the gray solid line ($k1 = 0$, $k2 = 1$) with the yellow solid line ($k1 = 1$, $k2 = 0$), we find that the peak decelerations are equal. However, the model whose minimum desired following distance model in the cut-in scenario is replaced by the others has the disadvantage of excess adjustment. So, we can conclude that applying a different minimum desired following distance model for new model is necessary and suitable.

6.2. Car-Following Properties in Changing Road Surface Conditions. The real-time maximum deceleration varies with the real-time road surface condition. Consequently, the parameter a_{max} is different on diverse road surfaces. The emergency braking distance is dependent on a_{max}. On the good road surface (dry asphalt road), the emergency braking distance is relatively short because of the large a_{max}. However, the emergency braking distance on the bad road surface (icy road) is longer than the good road surface because of the smaller a_{max}. Therefore, the driver can maintain a greater following gap on the bad road surface.

The IDM considers the real-time road conditions by adjusting the desired time headway. The minimum desired following distance in the new model is a function of the real-time maximum deceleration. To test the performance of the IDM and the new model in real-time road surface conditions, a simulation is carried out. In the simulation, two vehicles move from a dry asphalt road to an icy road at 190 m. The total length of the road section is 530 m. The maximum decelerations on the dry asphalt road and the icy road are 7 and 1.764 m/s^2. The leading vehicle is driven by a human, and the driver slows down before moving onto the icy road because the human driver can see the condition of the road in

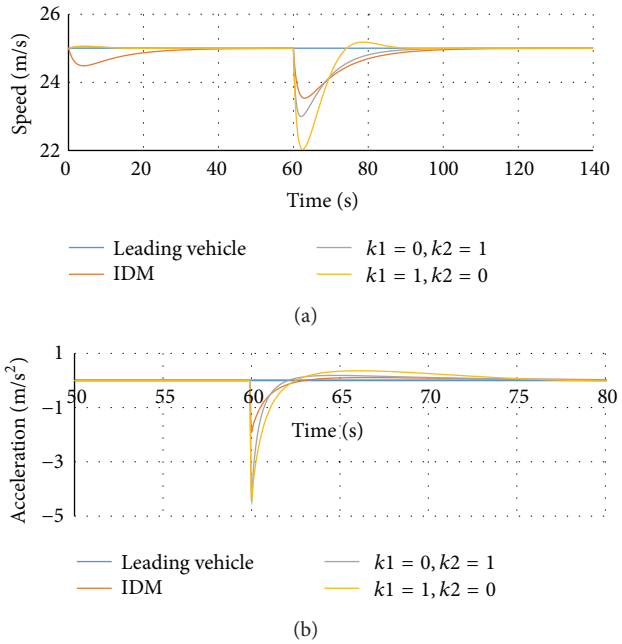


FIGURE 5: Cut-in response.

front and the front vehicle of the leading vehicle brakes before the leading vehicle drives onto the icy road.

The leading vehicle moves onto the icy road at 13.6 s. In the IDM control, the driver manually changes the desired time headway from 1.5 to 2.2 s after observing the change in the road condition. The sudden change in the minimum desired following distance of the IDM which is shown in Figure 6(c) (IDM-s^*) is attributed to the change in the desired time headway. However, the desired time headway is a constant value in the new model. The driver does not need to manually change the desired time headway. The change in the real-time maximum deceleration represents the change in the real-time road condition.

In the new model control, the maximum deceleration is updated after the change in the road condition is detected by the sensors. Because the minimum desired following distance of the new model is a function of the real-time maximum deceleration, it also changes suddenly with changes in the real-time maximum deceleration. From Figure 6(c), we can find the sudden change (new model-s^*). The sudden change of the new model is smaller than that of the IDM. Consequently, the new model controller is more stable and decelerates more smoothly than the IDM controller. The result is in accordance with Figure 6(b). The maximum deceleration of the IDM is $-2.95\,\text{m/s}^2$ and the maximum deceleration of the new model is $-2.51\,\text{m/s}^2$.

Comparing the response time of the following vehicle of the new model with that of the IDM after the vehicles drive onto an icy road, we find that the new model controller has greater delay than the IDM controller according to Figure 6(a). The reason is that the desired time headway is adjusted manually by the human driver before the vehicle drives onto the icy road. However, the parameters of the new model are adjusted automatically when the vehicle drives onto the icy road.

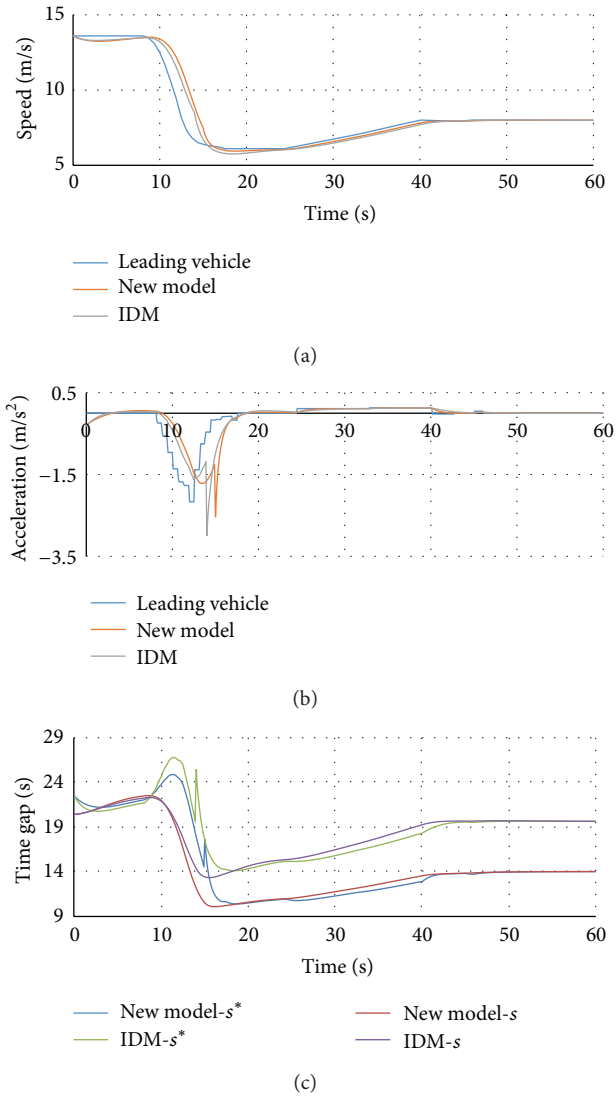


FIGURE 6: The car-following properties on a changing road.

7. Conclusion

In this paper, the real-time maximum deceleration estimation approach is introduced. We analyzed the IDM and found a disadvantage which could occur when the vehicle moves at high and even speed. According to the disadvantage and the real-time maximum deceleration, a new car-following model is established on the basis of the IDM. The minimum desired following distance of the new model contains minimum collision-avoidance distance and ensures the safety of the following vehicles.

The equilibrium flow-density relations of the new model and the IDM are compared and the result indicates that the new model controller improves road capacity more than the IDM controller and that reducing the time headway is more efficient for improving the road capacity than increasing the desired speed.

The new model can keep the vehicle moving with smaller time headway than the IDM on the basis of ensuring safety

and comfort according to the simulation of equilibrium traffic on a normal road. The smaller time headway means greater road capacity which can alleviate the traffic jam. Moreover, the peak of actual deceleration of the new model may be roughly greater than that of IDM in the cut-in scenario, but it is within the acceptable range. In changing road conditions, the new model can have smaller deceleration than the IDM and the controlled vehicle decelerates smoothly. Consequently, considering the effect of real-time maximum deceleration on car-following not only can improve the driving comfort and safety but also can alleviate the traffic jam which has important significance for sustainable transportation.

Conflict of Interests

The authors declare that they have no competing financial interests.

Acknowledgment

The authors would like to thank the reviewers for their careful reading and for providing some pertinent suggestions.

References

[1] B. Li, H. Du, and W. Li, "Comparative study of vehicle tyre–road friction coefficient estimation with a novel cost-effective method," *Vehicle System Dynamics*, vol. 52, no. 8, pp. 1066–1098, 2014.

[2] Z. P. Yu, J. L. Zuo, and L. J. Zhang, "A summary on the development status quo of tire-road friction coefficient estimation techniques," *Automotive Engineering*, vol. 28, no. 6, pp. 546–549, 2006.

[3] X. B. Fan and P. Deng, "Study on the tire/road friction coefficient estimation," *Auto Engineer*, vol. 12, pp. 47–50, 2013.

[4] R. Rajamani, N. Piyabongkarn, J. Lew et al., "Tire-road friction-coefficient estimation," *IEEE Control Systems Magazine*, vol. 30, no. 4, pp. 54–69, 2010.

[5] R. Rajamani, G. Phanomchoeng, D. Piyabongkarn, and J. Y. Lew, "Algorithms for real-time estimation of individual wheel tire-road friction coefficients," *IEEE/ASME Transactions on Mechatronics*, vol. 17, no. 6, pp. 1183–1195, 2012.

[6] D.-H. Wang and S. Jin, "Review and outlook of modeling of car following behavior," *China Journal of Highway and Transport*, vol. 25, no. 1, pp. 115–127, 2012.

[7] V. Milanes, S. E. Shladover, J. Spring, C. Nowakowski, H. Kawazoe, and M. Nakamura, "Cooperative adaptive cruise control in real traffic situations," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 1, pp. 296–305, 2014.

[8] T. Dijker, P. H. L. Bovy, and R. G. M. M. Vermijs, "Car-following under congested conditions: empirical findings," *Transportation Research Record*, no. 1644, pp. 20–28, 1998.

[9] A. Kesting, M. Treiber, M. Schönhof, and D. Helbing, "Adaptive cruise control design for active congestion avoidance," *Transportation Research Part C: Emerging Technologies*, vol. 16, no. 6, pp. 668–683, 2008.

[10] I. Soria, L. Elefteriadou, and A. Kondyli, "Assessment of car-following models by driver type and under different traffic, weather conditions using data from an instrumented vehicle," *Simulation Modelling Practice and Theory*, vol. 40, pp. 208–220, 2014.

[11] T. Q. Tang, Y. P. Wang, X. B. Yang, and Y. H. Wu, "A new car-following model accounting for varying road condition," *Nonlinear Dynamics*, vol. 70, no. 2, pp. 1397–1405, 2012.

[12] T. Q. Tang, J. G. Li, H. J. Huang, and X. B. Yang, "A car-following model with real-time road conditions and numerical tests," *Measurement*, vol. 48, no. 1, pp. 63–76, 2014.

[13] K. Yi and J. Taeyoung, "Observer based estimation of tire-road friction for collision warning algorithm adaptation," *JSME International Journal, Series C: Dynamics, Control, Robotics, Design and Manufacturing*, vol. 41, no. 1, pp. 116–124, 1998.

[14] D. Y. Qu, X. F. Chen, W. S. Yang, and X. H. Bian, "Modeling of car-following required safe distance based on molecular dynamics," *Mathematical Problems in Engineering*, vol. 2014, Article ID 604023, 7 pages, 2014.

[15] L. H. Xu, Q. Luo, J. W. Wu, and Y. G. Huang, "Study of car-following model based on minimum safety distance," *Journal of Highway and Transportation Research and Development*, vol. 10, no. 10, pp. 95–100, 2010.

[16] R. Hoseinnezhad and A. Bab-Hadiashar, "Efficient antilock braking by direct maximization of tire–road frictions," *IEEE Transactions on Industrial Electronics*, vol. 58, no. 8, pp. 3593–3600, 2011.

[17] Z. H. Wang and S. M. Shao, "Simulation of adaptive cruise control algorithm of vehicle," *Computer Engineering and Design*, vol. 2, pp. 604–608, 2014.

[18] M. Treiber, A. Hennecke, and D. Helbing, "Congested traffic states in empirical observations and microscopic simulations," *Physical Review E: Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, vol. 62, no. 2, pp. 1805–1824, 2000.

[19] P. Zheng and M. McDonald, "Manual vs. adaptive cruise control—can driver's expectation be matched?" *Transportation Research Part C: Emerging Technologies*, vol. 13, no. 5-6, pp. 421–431, 2005.

[20] D. Helbing, A. Hennecke, V. Shvetsov, and M. Treiber, "Micro- and macro-simulation of freeway traffic," *Mathematical and Computer Modelling*, vol. 35, no. 5-6, pp. 517–547, 2002.

[21] N. H. Gartner, C. J. Messer, and A. K. Rathi, "Traffic flow theory (update of TRB special report 1165)," Tech. Rep., TRB, Washington, DC, USA, 1997.

[22] R. Weidemann and U. Reiter, "Microscopic traffic simulation. The simulation system-mission," Project ICARUS (V1052) Final Report, University Karlsruhe, 1992.

[23] J. Ploeg, B. T. M. Scheepers, E. van Nunen, N. van de Wouw, and H. Nijmeijer, "Design and experimental evaluation of cooperative adaptive cruise control," in *Proceedings of the 14th International IEEE Conference on Intelligent Transportation Systems (ITSC '11)*, pp. 260–265, IEEE, Washington, DC, USA, October 2011.

[24] F. Bu, H. Tan, and J. Huang, "Design and field testing of a cooperative adaptive cruise control system," in *Proceedings of the American Control Conference (ACC '10)*, pp. 4616–4621, IEEE, Baltimore, Md, USA, June 2010.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 502739, 12 pages
http://dx.doi.org/10.1155/2015/502739

Research Article

Model for Estimation Urban Transportation Supply-Demand Ratio

Chaoqun Wu,[1,2] Yulong Pei,[3] and Jingpeng Gao[4]

[1]*School of Transportation Science and Engineering, Harbin Institute of Technology, Harbin 150090, China*
[2]*School of Automobile and Traffic Engineering, Heilongjiang Institute of Technology, Harbin 150050, China*
[3]*Traffic College, Northeast Forestry University, Harbin 150040, China*
[4]*College of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China*

Correspondence should be addressed to Chaoqun Wu; wuchaoqunwcq@126.com

Received 24 January 2015; Revised 30 June 2015; Accepted 24 August 2015

Academic Editor: Chronis Stamatiadis

The paper establishes an estimation model of urban transportation supply-demand ratio (TSDR) to quantitatively describe the conditions of an urban transport system and to support a theoretical basis for transport policy-making. This TSDR estimation model is supported by the system dynamic principle and the VENSIM (an application that simulates the real system). It was accomplished by long-term observation of eight cities' transport conditions and by analyzing the estimated results of TSDR from fifteen sets of refined data. The estimated results indicate that an urban TSDR can be classified into four grades representing four transport conditions: "scarce supply," "short supply," "supply-demand balance," and "excess supply." These results imply that transport policies or measures can be quantified to facilitate the process of ordering and screening them.

1. Introduction

This paper describes a methodology for estimating the value of urban transportation supply-demand ratio, TSDR for short, that is the result of interaction between transport system support capacity and inhabitants' travel requests and can determine whether transport conditions are balanced.

In a multitude of papers, travel demand and transportation supply are shown to act as a means of traffic phenomenon analysis to shed light on the theories and techniques underlying high-effect transportation.

There are several popular theories deriving from supply and demand consideration. For example, kinematic wave theories of lane-changing traffic flow [1] and merging traffic flow [2] have been described by Jin. Microsimulation models have emphasized both day-to-day and within-day variability in both demand and supply dynamic condition for real-time transport strategies in automated highway and responsive traffic signal control systems [3]. Considering that transport supply and demand varied incessantly under the influence of travel time-variations [4], a theoretical framework for designing a reliable transportation network [5] was developed and formulated as a bilevel program, with the upper level specifying the objective with respect to the optimized subject to demand growth and economic constraints and the lower comprised of time-dependent use of equilibrium models. After calibrating all demand and supply parameters, a dynamic traffic assignment model for highly congested urban networks required a modified treatment of acceptance capacity [6]. Under stochastic demand and supply, a robust traffic assignment process of the expected residual minimization model places emphasis on the planner's perspective and stochastic cell transmission model [7] and can capture the mean and standard deviation (SD) of density of the traffic flow and the propagation of SD over time and space.

A variety of transport problems originating from the contradiction between travel demand and transportation supply have been discussed in the relational literatures. To tackle taxi service refusal [8], pricing policies and regulations should take into consideration its impact on demand-supply equilibrium in both monopolistic and competitive market. Empirical examination, based on data from a survey of the "parking marketplace" site around the UK's 25 busiest passenger airports, indicated that some car parking requirements in

Table 1: Characteristic of complex system and transportation system.

Complex system [20] characteristic	Transportation system characteristic
Large numbers of elements are manifold	Elements: People (drivers, passengers, and pedestrian) Vehicle (motor and nonmotor vehicle, train) Facilities (road, tunnel, bridge, signal system, etc.) Management (laws and regulations, management technique) Environment (economic and culture)
Interaction among the elements is more important than the element itself	The crux of transportation system's maximum efficiency is the coordination of the elements
Multiple causality among the elements	Transportation system consists of subsystems such as economy, number of vehicles, environment, travel demand, transport supply, and a traffic congestion subsystem. Every subsystem has causality and there is special causality among the subsystems [21]
Dynamic and nonlinear	Transportation system's elements are in a stochastic condition; that is, they vary with time and space; their linear relationship, because of complex causality, cannot satisfy the requirements of modeling to simulate real transportation, so the modeling method has undergone several processes: statistics, differential equations, system dynamic, the models of complex network, and modeling method based on Agent [22]
Self-organization and self-adaptiveness	A transportation system, because of randomness and complexity, can only operate in orderly fashion by using a self-feedback function. This is, in order for traffic flow tend to be in ordered under certain conditions, the transportation system should self-adjust, based on real-time traffic status, by control and management technologies [23], so that it has self-organization and self-adaptiveness

the airport would be experiencing rapid growth with a relatively small supply of overall airport parking, so airport operators and local authorities should be cognizant of the necessity of alternative parking provision [9]. A model analyzed the vicious cycle of a bus line [10], in which high demand will induce the operator to increase supply, in turn resulting in a higher level-of-service requirement and a subsequent increase in passenger numbers, triggering another round of service improvements.

The theories and techniques mentioned above can be aimed at any part of the transportation system that could make greater contribution. The question of whether transport supply and travel demand influence the whole transportation system has also been discussed.

Travel demand and transportation supply modeling methodology was presented through an Upper-Silesian Conurbation in Poland [11] example. To agglomerate them, the Interval Fractional Transportation Problem has adopted the expression of intervals with left and right limits [12]. Supply-demand equilibrium [13] has been discussed in terms of a hypernetwork (an abstract network on which a route was chosen) in the disaggregate demand models on a mathematically consistent basis for congested transportation systems. The new method of estimating the effect of travel demand variation and link capacity degradation was applied in the expected reliability of a roadway network: travel time reliability and capacity reliability [14].

The demand of transportation can be generally defined in terms of inhabitant trips, but the supply aspect had different assumptions according to the object or the aim. While route choice was regarded as a supply aspect of the urban network, the supply curves [15] were sensitive to the temporal and spatial distribution of demand, and its shape also differed from Origin-Destination movements within a given network; activity-based modeling and dynamic traffic assignment were combined [16] and the benefit of responsive pricing and travel information was quantified [17]. By improving bus and metro capacities contributing to the transportation supply, a framework for evaluating the dynamic impacts of a congestion pricing policy [18] can show how supply dynamics affect the travel demand of individuals and their choice of different transportation modes, and the method [19] of design and implementation of efficient transit networks can be applied to designing a high-performance bus network in Barcelona (Spain).

The above literatures are aimed at developing solution or a corresponding theory for a transport problem. However, because the various parts of transportation systems are interactive, the solution of a transport problem is bound to bring up new problems, so this paper proposes a macroscopic analysis method for estimating the TSDR.

2. Previous Work

2.1. Methods and Tools. A transportation system is an open complex system, and Table 1 describes the characteristic of complex system and transportation system. Peer experts have applied the theories and methods of system engineering in their approaches for modeling in the transportation area.

System dynamics is an approach to understanding the behavior of complex systems over time, and it is able to deal with internal feedback loops and time delays that affect the behavior of an entire system. This approach was well-suited to

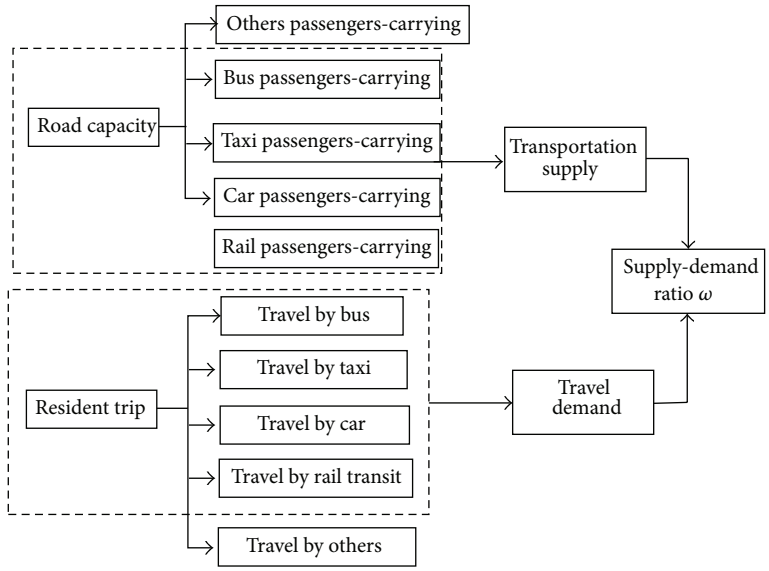


Figure 1: Idealized system of urban transport system.

strategic issues and could provide a useful tool for supporting policy analysis and decision-making in the transportation field [24]. The areas of application include the take-up of alternate fuel vehicles, supply chain management affecting transport, highway maintenance, strategic policy, and a set of emerging application areas.

This paper analyzes the transportation system using the methods of system dynamics and will estimate the TSDR by VENSIM, an industrial-strength simulation software package for improving the performance of real systems; it has a rich feature set emphasizing model quality, connections to data, flexible distribution, and advanced algorithms.

2.2. The Idealize System. As stated in *Transportation System Analysis* [25], "a transportation system is a collection of elements and the interactions between them that produce both the demand for travel in a given area and the provision of transportation services to satisfy this demand." In other words, the transportation system consists of two main components: travel demand and transportation supply. Figure 1 depicts the idealized system studied in the paper.

(1) Transportation Supply. The transportation supply component as described in *Transportation System Analysis* [25] is made up of facilities (roads, parking spaces, railway lines, etc.), services (transit lines and timetables), regulations (road circulation and parking regulations), and prices (transit fares, parking prices, road tolls, etc.) that produce travel opportunities.

This is to say that the purpose of measures such as constructing transport facilities, improving services, strengthening management, and pricing reasonably is to offer additional travel service. It is then necessary that transportation supply is quantized by the maximum passenger-carrying capacity of the transport system per unit time. This leads to Definition 1.

Definition 1. *Transportation supply* is the maximum amount of passenger-carrying capacity contributed by the metropolitan's transport system; it has three characteristics: resources restriction, multiformity of transport modes, and multilevels of service targets. Transportation supply and demand have settled into dynamic equilibrium.

The facilities for mass transit in China include urban road, rail transit, and ferry; rail transit mainly includes tram, light rail, rapid rail (metro), monorail, and funicular. Ferry and funicular play an auxiliary role in some cities with special geography conditions such as a river passing through or mountainous composition of the city's landform; tram generally cannot exist in Chinese cities because of ever more crowded transport. Almost all of the rail transit entities operate independently in underground tunnels or on viaducts so as not to interfere with surface transport, so Hypothesis 2 can be given.

Hypothesis 2. There are two transport facilities for metropolitan inhabitant: urban road and rail transit, and both coexist and are independent of one another.

Most often, types of motorized and nonmotorized vehicles operating on urban roads include bus, car, truck, motorcycle, and bicycle; their purpose and travel times vary so traffic composition varies with city limits and times. Because trucks were forbidden in the daytime in most urban districts, cargo traffic's influence on the transportation supply and travel demand balance can be overlooked in the daytime. Buses, disaggregated by purpose, are composed of public buses, commuter buses, and intercity buses; the latter two are a minority on urban roads and have flexibility for choosing congestion-free routes, so "bus" will refer specifically to public buses to simplify the model. Motorcycles and bicycles are suitable for traveling short distances but rarely run on

Table 2: Information for chosen cities.

Metropolitan area	Year	Population (million)	Gross Domestic Product (billion, RMB)	Urban area (Km^2)	Significant transport event
Beijing	2011	20.186	1,625.19	8579	
Tianjin	2011	13.546	1,130.73	1103	
Shanghai	2011	23.475	1,919.57	9589	
Guangzhou	2012	8.223	1,355.12	2910	
Hangzhou	2013	7.253	780.20	2060	
Shenzhen	2012	1.047	1,150.55	5256	The 181 km new road and the 133 km reformed road in 2012
Shenzhen	2013	1.055	1,295.00	5282	
Shenyang	2012	8.228	660.68	1504	Metro line 1 and line 2 opened, respectively, on September 27, 2011, and February 9, 2012
Shenyang	2011	7.227	591.49	1495	
Shenyang	2010	7.196	501.71	1485	
Nanjing	2012	8.161	720.16	1215	

※These data from *China City Statistical Yearbook*.

the expressway and major arterial roads because of the great distances between workplace and home in metropolitan areas. "Car" comes in three forms: private car, official vehicle, and taxi; the first two have similar traits [26] and the characteristics of taxi trips [27] are, dynamically using road resources, different from the former two. So, Hypothesis 3 is given.

Hypothesis 3. Traffic composition on urban roads consists of car, public bus, and taxi; the term "car" includes both private cars and official vehicles.

(2) Travel Demand. Travel demand, also described in *Transportation System Analysis* [25], derives from the need to access urban functions and services in different places and is determined by the distribution of households and activities within the area.

Travel demand is trip need or expectation of a city inhabitant for business or entertainment and is the natural outgrowth of economic development and urban population increase. It is always reflected in the trip structure of the metropolitan area, combined into several trip modes such as car, bus, taxi, rail transit, bicycle, and walking. The impact factors that affect travel mode choices of urban residents derive from the spatial and temporal nonuniformity of transport facilities' use, the purpose of people trips, and the convenience of the transport system, so Definition 4 applies. In the meantime, Hypothesis 5 can be deduced from Hypotheses 2 and 3.

Definition 4. Travel demand is the movement requirement using public transport facilities from one functional area to another; it varies with the spatial and temporal nonuniformity of transport facilities' use, the purpose-sets of people trips, and the convenience-sets of the transport system. Transportation demand can be measured by the total number of trips.

Hypothesis 5. The trip structure within the metropolitan area can be characterized by four trip modes: car, public bus, taxi, and rail transit. The "car" designation includes both private cars and official vehicles.

2.3. Data. This study examined certain cities that exhibited one of the two characteristics listed below. Table 2 gives specific information about these cities.

(a) The cities are densely populated and relatively well-developed economically so they use superior transportation systems and are able to provide sets of data representing different traffic states for quantitative analysis.

(b) Significant transport events occurred in recent years; examples would include the Shenyang metro being in operation and Shenzhen's new and reformed long urban road. The before-and-after data comparison can reveal an event's impact on the balance of the transportation system.

The sources of the micro- and macrodata involved in this paper are mainly focused on the following ways.

(a) The Official Data. They include *China City Statistical Yearbook*, *Yearbook of China Integrated Transport*, and the existing data from the Internet published by municipal transportation commissions.

The digital information from *China City Statistical Yearbook* had macroeconomic messages (including population and Gross Domestic Product), investment amounts in transportation from government, divergence of residents' earnings, urban road grades and their lengths, and distribution of the vehicles' owners. Information from the *Yearbook of China Integrated Transport* provided the total number of resident trips, the distribution of travel modes, the vehicle speed in road net, the lengths of urban roads classified, the operational information about public bus and railway routes (kilometrage, number of operating vehicles, and passengers

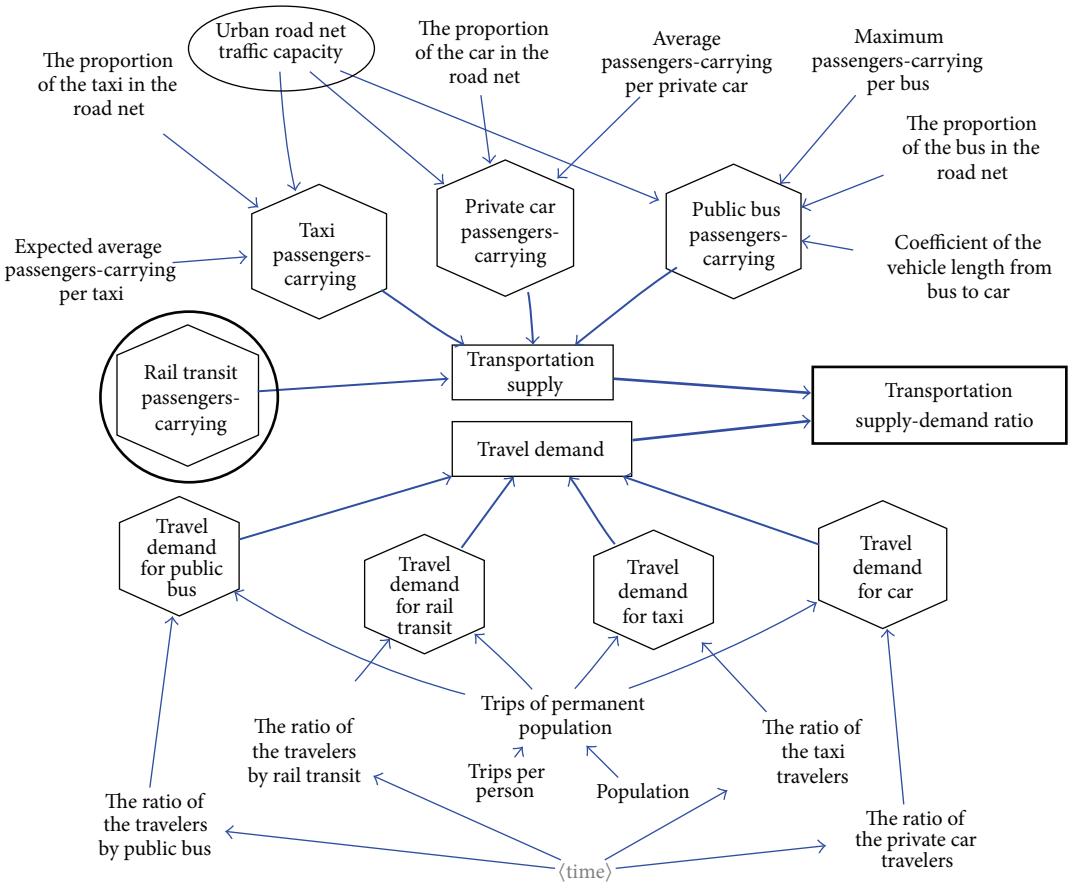


Figure 2: Urban TSDR estimating framework.

carried), and so on. The municipal transportation commission published operating information about railway, public bus, and taxi in real time.

(b) The Data from the Traffic Administrative Department of the Public Security Organ and the Public-Transport Operation Corporation. Traffic monitoring systems and transportation information collection systems set up in Chinese cities provide traffic microdata such as traffic flow, headway, traffic density, and speed, which can be used for estimating Lanes Comprehensive Utility Coefficient and Intersections Effective Utility Coefficient of classified roads.

The operational data from the public-transport operation corporation (bus, metro, or taxi operation company) includes not only passengers carried but also Passenger Load Factor, OD (Origin and Destination), and kilometrage of the passengers, used both in directly estimating the supply-demand ratio and reciprocally verifying with official statistics.

(c) Surveying on the Spot. The purpose of surveying on spot in a city was to gain the necessary data that cannot be obtained by the above two modes; this includes vehicles average travel distance, the respective proportion of the public buses, taxis, and cars in the urban road, and their average number of passengers carried.

Manual records and interview surveys were adopted. The proportion of public buses, taxis, cars and their Average Carried Passengers were manually recorded in forty sections of expressway, major arterial road, minor arterial road, and collector streets during peak hours and off-peak hours during both working days and Sunday. In a roadside interview survey, at least 200 drivers in a city were asked, "Normally, how many kilometers do you drive a day"; the vehicles average travel distance was the average value of these answers.

3. Model

The transportation supply-demand ratio used to quantitatively describe the transport system state is the ratio between transportation supply and travel demand, that is, between the maximum amount of passengers carried and the total trips according to Definitions 1 and 4. As the value of this ratio increases, urban transport would become more and more unobstructed.

3.1. Model for Estimating Urban TSDR. Figure 2 shows the supply-demand ratio estimating framework. Letting ω denote the TSDR and letting S and D separately denote transport supply and travel demand, the estimating model can be written as

$$\omega = \frac{S}{D}. \tag{1}$$

(1) Transportation Passengers-Carrying Supply. In Figure 2, the transportation passengers-carrying supply S is composed of travelers carried by public bus S_{bus}, rail transit S_{rail}, taxi S_{taxi}, and car S_{car}, resulting in

$$S = S_{\text{bus}} + S_{\text{rail}} + S_{\text{taxi}} + S_{\text{car}}. \tag{2}$$

By Hypothesis 2, the bus passengers-carrying supply (S_{bus}), taxi passengers-carrying supply (S_{taxi}), and car passengers-carrying supply (S_{car}) are limited by net urban road traffic capacity C_{roadnet}. By Hypothesis 3, the traffic capacity of urban road net C_{roadnet} and the rail passengers-carrying supply S_{rail}, as shown by the circled section in Figure 2, are independent of one another, and they will be separately introduced in Sections 3.2 and 3.3.

Considering that the traffic capacity unit of measurement is the PCU (Passenger Car Unit), the mathematical formulation of bus passengers-carrying supply S_{bus} after vehicle equivalent conversion can be stated as

$$S_{\text{bus}} = \frac{C_{\text{roadnet}} \times \alpha}{\phi} \times D_{\text{bus}}^{\max}, \tag{3}$$

where α = the proportion of the bus in the road net; ϕ = conversion coefficient of the vehicle length form bus to car; $D_{\text{bus}}^{\max}$ = maximum passengers-carrying per bus, restricted by the transport policies grounding in service-level oriented or transportation-capacity oriented.

Similarly, taxi passengers-carrying supply S_{taxi} can be expressed as the product of the average expected passengers-carrying per taxi and the number of the taxies, so

$$S_{\text{taxi}} = C_{\text{roadnet}} \times \beta \times D_{\text{taxi}}^{a}, \tag{4}$$

where β = the proportion of taxis in the road net and D_{taxi}^{a} = average expected ridership per taxi, dependent on whether carpooling is permitted.

The car passenger-carrying supply S_{car} is also related to the number of cars and average passenger-carrying per car, so

$$S_{\text{car}} = C_{\text{roadnet}} \times \gamma \times D_{\text{car}}^{a}, \tag{5}$$

where γ = the proportion of private cars in the road net; D_{car}^{a} = average passenger-carrying per car.

(2) Travel Demand. Trip demand D is the sum of trip demand for bus, rail transit, taxi, and car and similarly to the transportation passenger-carrying supply estimating process it can be expressed by the mathematical formula

$$D = D_{\text{bus}} + D_{\text{rail}} + D_{\text{taxi}} + D_{\text{car}}, \tag{6}$$

where D_{bus}, D_{taxi}, D_{rail}, and D_{car} are, respectively, trip demand for bus, taxi, rail transit, and car. They are expressed by the ratios of their travelers to the total, so these equations are

$$\begin{aligned} D_{\text{bus}} &= PAR_{\text{bus}}, \\ D_{\text{taxi}} &= PAR_{\text{taxi}}, \\ D_{\text{rail}} &= PAR_{\text{rail}}, \\ D_{\text{car}} &= PAR_{\text{car}}, \end{aligned} \tag{7}$$

where P is the total population in the city; A is average trip; and R_{bus}, R_{taxi}, R_{rail}, and R_{car} are the proportion of bus, taxi, rail transit, and car travelers in the total trips.

(3) Passengers-Carrying Supply-Demand Ratio (PCSDR). To embody the urban transport system supply-demand relation, if the values of the four modes' passengers-carrying supply-demand ratio, PCSDR for short, are estimated at the same time, then

$$\begin{aligned} \omega_{\text{bus}} &= \frac{S_{\text{bus}}}{D_{\text{bus}}}, \\ \omega_{\text{rail}} &= \frac{S_{\text{rail}}}{D_{\text{rail}}}, \\ \omega_{\text{taxi}} &= \frac{S_{\text{taxi}}}{D_{\text{taxi}}}, \\ \omega_{\text{car}} &= \frac{S_{\text{car}}}{D_{\text{car}}}. \end{aligned} \tag{8}$$

3.2. Urban Road Net Traffic Capacity. The traffic capacity of the urban road net is the maximum number of vehicles running on the urban road net at a certain time; it is limited by the characteristics of the net and the traffic conditions. Figure 3 shows its estimation module.

All the transportation modes except rail transit are restricted to the urban road net traffic capacity. And urban road net traffic capacity [28] C_{roadnet} is expressed by the ratio of the spatial and temporal resources of road net RE_d to the spatial and temporal consumption of traffic unit RE_t based on the Specific Vehicle Saturation ρ; it takes the form

$$\begin{aligned} C_{\text{roadnet}} &= \frac{\text{RE}_d}{\text{RE}_t} \times \rho \\ &= \left(\sum_{i=1}^{4} L_i \cdot d_i \cdot \eta_{1i} \cdot \eta_{2i} \right) \times T \\ &\quad \times \frac{1000}{\left(l_p/V\right)\left((t/3.6)\,V + l_v + l_s\right)} \times \rho, \end{aligned} \tag{9}$$

where $i = 1, 2, 3, 4$, respectively, replace the expressway, major arterial road, minor arterial road, and collector street and L_i = length of the roads classified. Its value is calculated by the investment or use existing data. η_{1i} = Intersections Effective Utility Coefficient (IEUC) for roads of each grade; η_{2i} = Lanes Comprehensive Utility Coefficient (LCUC) for roads of each grade; T = service time, h; L_p = vehicles average travel distance, km; V = vehicles average travel speed, km/h; t = driver's reaction time, s; l_v = car length, m; l_s = minimum safe distance between two cars when they are static, m.

In the above formulas, many parameters may change in implementation of transport policies or measures, so the estimated result of the model can reflect the effect of such changes. For example, investment in transportation may cause length of the roads classified and length of rail transit line increase. By using and generalizing advanced technology like Intelligent Transportation Systems, driver Information

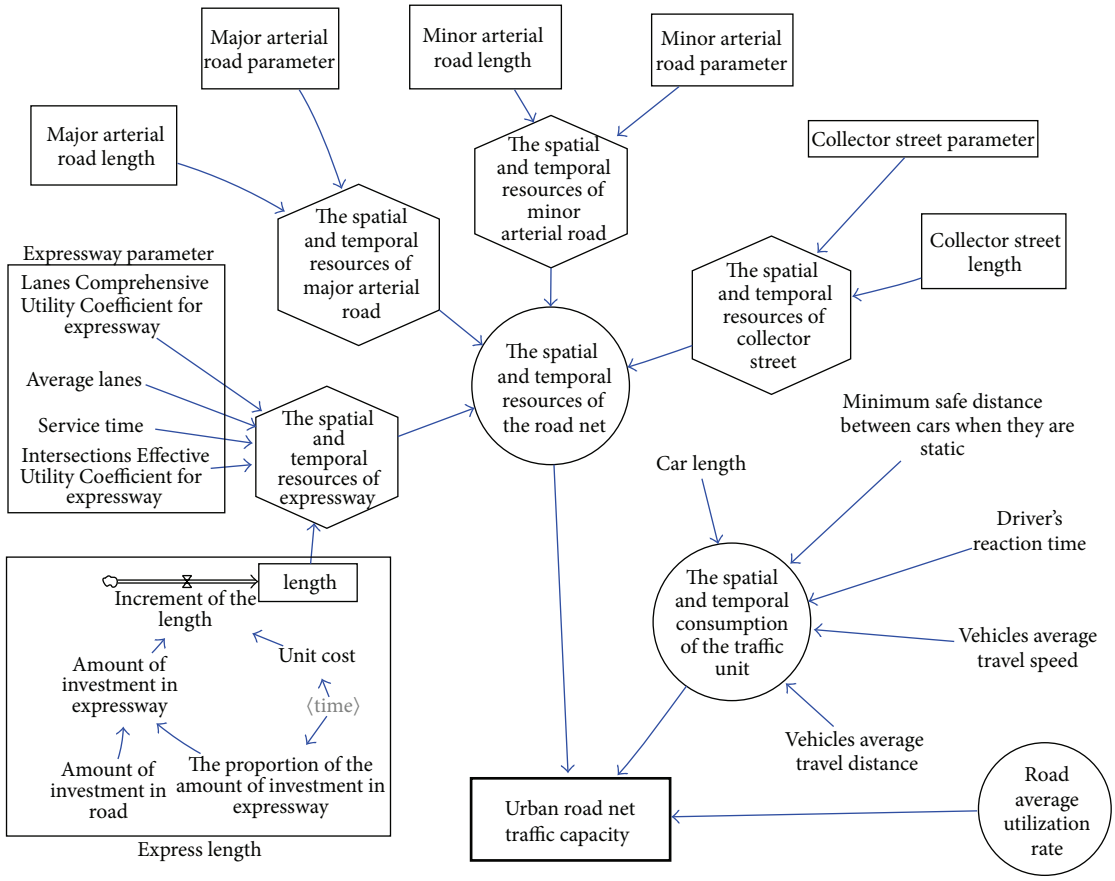


Figure 3: Estimation module for traffic capacity of urban road net.

System, urban Traffic Area-wide Cooperation Control Systems, and Urban Pedestrian Systems the values of Intersections Effective Utility Coefficient and Lanes Comprehensive Utility Coefficient will increase.

The values of IEUC and LCUC for roads classified are displayed by Time Occupancy [29]; that is, they are the percentage of all the time that vehicle occupied in road; then

$$\eta_{1i}\left(\eta_{2i}\right) = \sum_{i=1}^{n} \frac{t_i}{T_o}, \tag{10}$$

where T_o = observed duration, t_i = time that number i vehicle moves through the observed cross-section, and n = number of the vehicles observed.

There are some things to be aware of while observing the IEUC and LCUC:

(i) One should try to choose the cross-sections less affected by intersections and the period during which the traffic becomes saturated.

(ii) The LCUC should be dissected for every lane.

(iii) Entries to the intersection while measuring the IEUC should be observed.

(iv) Select all types of the intersections and cross-sections, but not all of them.

(v) IEUC and LCUC average aimed at the roads of each grade should be calculated.

Table 3 [29] lists the suggested values of IEUC and LCUC.

3.3. Rail Transit Passengers-Carrying Supply. The rail transit passengers-carrying supply S_{rail} is expressed as a sum of all of the city's rail lines ridership capacity (C_i), and it is strongly influenced by the Departing Interval (I_{rail}) and Passengers-carrying per train (N_i). Figure 4 describes the rail transit passengers-carrying supply estimating module.

Generally, the following formulas state the algorithmic method:

$$\begin{aligned} S_{\text{rail}} &= \sum_{i=1}^{n} \left(C_i \times N_i\right), \\ C_i &= \frac{T}{I_{\text{rail}}} \times 3600. \end{aligned} \tag{11}$$

(1) Departing Interval (I_{rail}). In reality, the factors affecting the Departing Interval fall into two categories:

(i) those restricted by telecommunication and signal control technology, letting $I_{\text{rail}}^{\min}$ denote the minimum tracking interval dependent on the signaling system, whose value is provided by the manufacture of the rail line used,

Table 3: Suggested values of IEUC and LCUC.

Urban road classification	Expressway	Major arterial road	Minor arterial road	Collector street
Intersections Effective Utility Coefficient	0.75	0.55~0.65	0.45~0.55	0.40~0.50
Lanes Comprehensive Utility Coefficient	0.9	0.85~0.95	0.80~0.90	0.85~0.95

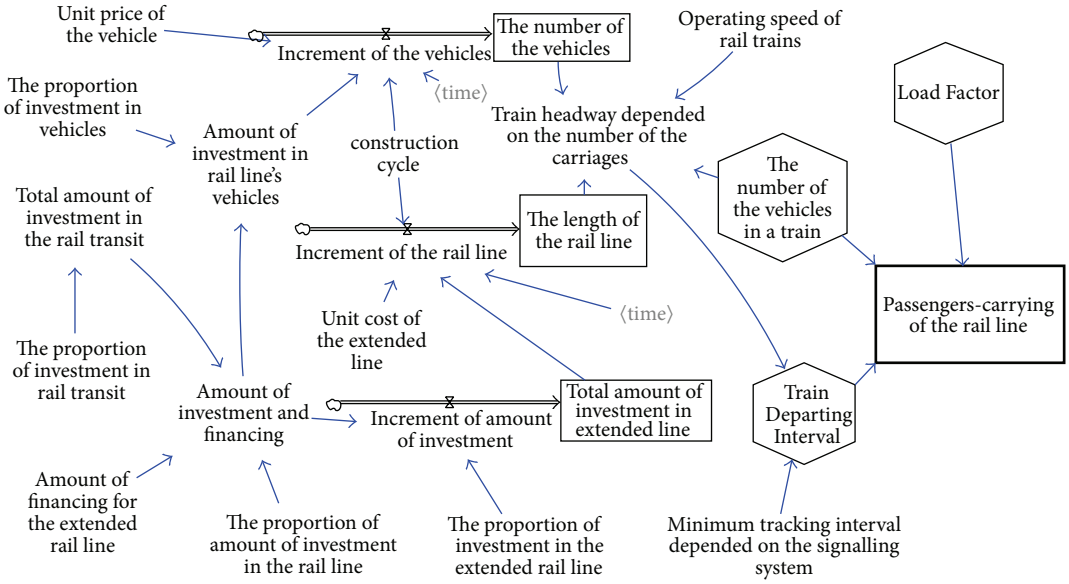


Figure 4: Estimated module for rail transit passengers-carrying supply.

(ii) those restricted by the horizontal and vertical curves of rail facilities and the number of the trains. The horizontal and vertical curves determine the train running speeds and the number of the trains supports the Departing Interval. Let I^d_{rail} denote the Train Departing Interval dependent on the facilities, and the result is the turnaround time (T_t) divided by the number of the trains (N_t). The derivation is as follows:

$$I^d_{\text{rail}} = \frac{T_t}{N_t}, \tag{12}$$

$$T_t = \frac{2L}{V}, \tag{13}$$

$$N_t = \frac{M}{m}, \tag{14}$$

where L = length of rail line; V = average travel speed, km/h; m = number of the vehicles in a train; M = number of the vehicles.

In the model shown in Figure 4, the rail-line length and the vehicle number are calculated by investment; that is, their values after a construction cycle are determined based on the actual investment and the construction cost. The meaning of the method is a quantitative prediction of the implementation effect of policies like the investment focus and reducing the engineering cost; the construction cycle is generally 5 years or so. Certainly, they can either use the existing data for the existing line or define the measures.

After considering the ratio of the spare and maintaining trains to operating trains (τ), converting hours into seconds, substituting (13) and (14) into (12) produces

$$I^d_{\text{rail}} = \frac{2L}{V} \times \frac{m}{M} \times (1 + \tau) \times 3600. \tag{15}$$

For the Departing Interval choose the max value between $I^{\min}_{\text{rail}}$ and I^d_{rail}, producing

$$I_{\text{rail}} = \max\left\{I^d_{\text{rail}}, I^{\min}_{\text{rail}}\right\}. \tag{16}$$

(2) Passengers-Carrying per Train (N_i). Since passengers-carrying per train depends on the prescribed passengers-carrying per vehicle (D_{rail}) and Load Factor (σ), then

$$N_i = m \times D_{\text{rail}} \times \sigma. \tag{17}$$

(3) Example Verification. To verify the rail transit passengers-carrying supply estimating framework, Table 4 lists the error rate of the estimated value compared with the actual value obtained in 2012 from the 15 Beijing subway lines.

In Table 4, most absolute values of the error rate are less than 3%, and only values of Changing Line and Fangshan Line are greater than 3%. The max value is 5.98%, which was

Table 4: Beijing subway passengers-carrying estimated verification in 2012.

Rail line	Passengers-carrying estimated (person/hour)	Passengers-carrying in reality (person/hour)	Error rate (%)
Line 1	42336	42840	−1.18
Line 2	37800	38556	−1.96
Line 4	26779	26677	0.38
Line 5	31129	31328	−0.64
Line 8	17640	17520	0.68
Line 9	11760	11680	0.68
Line 10	32072	32296	−0.69
Line 13	33075	32130	2.94
Line 15	12027	11680	2.97
Changping Line	13569	13140	3.26
Fangshan Line	9284	8760	5.98
Yizhuang Line	11386	11680	−2.52
Batong Line	29223	29988	−2.55
Airport Express	2680	2688	−0.30
Daxing Line	22800	22603	0.87
Average error rate (absolute value)			1.84

※The data of passengers-carrying in reality from *2012 Yearbook of China Integrated Transport*.

caused by reserved actual passengers-carrying of Fangshan Line that was early in the running. Thus the way of the rail transit passengers-carrying supply estimation is effective and feasible.

3.4. Passengers-Carrying Supply Estimation on Bus Lane. The bus lane is one of two types depending on whether it is independent of the urban road net:

(i) Open Up in the Road. This type of bus lane essentially occupies road resources, so the intersections and cross-sections with bus lanes are separated into dedicated lanes, permitting independent observations of the IEUC and LCUC.

(ii) Individually Built, That Is, Bus Rapid Transit (BRT). It is similar to rail transit and is determined to the rail transit passengers-carrying supply estimating framework.

4. Estimation and Results

In this section, fifteen groups' data were chosen from eight cities to analyze the TSDR. Table 5 separately lists the estimated values of the PCSDR of bus, taxi, rail, and car and the TSDR.

Theoretically, the value of supply-demand ratio is near 1: a value greater than 1 means that the supply exceeds the demand, and a value less than 1 means that the supply is short. However, in view of the complexity of the metropolitan transport system and the limitation of the model, the PCSDR of bus, taxi, rail, and car is, respectively, determined, and its underlying causes are explored to lay a foundation for describing the transport conditions represented by the value range of the TSDR.

(1) Estimated Result of the PCSDR of Bus, Taxi, Rail, and Car during Peak Hours. During peak hours, the values of sample cities' TSDR, except for Hangzhou's, are less than 0.95, and four of these are lower than 0.8; the supply and demand for the bus, taxi, rail transit, and car, however, vary considerably.

It is common for the values of bus PCSDR to remain steady during the peak hour interval [30]. In spite of a support of the public bus priority policy, such as bus lanes and bus traffic signal priority, bus operation efficiency is still of concern in the worst road traffic situations. Buses, frequently running under acceleration, deceleration, and idling conditions, seldom arrive at predictable intervals and may be so crowded that passengers who must ride the bus must either contend with a crowded and throttling atmosphere or miss one connection after another following an unsuccessful struggle to aboard.

The values of taxi PCSDR are the largest, and they exceed 0.8 in the overwhelming number of major cities. Because the expectation of taxi is the lowest from perspectives of both supply and demand, on the supply side, taxi drivers try to avoid operation to reduce costs, especially under crowded traffic conditions, and on the demand side, passengers often do not choose taxis travel because of the higher trip charge (compared to bus) and the longer trip time (compared to rail transit).

Most of the values of car PCSDR are less than 0.8, especially for Beijing, just 0.22 during peak hours. From the Chinese standpoint, it may be essential to own private cars because families with children, the elderly, and the infirm must use cars as travel tools to avoid dealing with the congestion of urban public traffic. Other reasons for owning a car, a symbol of identity, might include winning more social respect and even bringing about more economic benefits. Therefore, when Beijing and Shanghai use lottery system to limit new vehicle registrations, vehicle possession still had respective increases of 225,000 and 59,000 in 2013 and reached totals of 5.2 and 2.8 million. Relative to urban road mileage, 28,608 and 17,316 kilometers, private car demand greatly exceeds the supply of urban road net, so excessively larger vehicle possession is the root cause of the urban road net congestion. On the other hand, the low average passengers-carrying, only 1.17 from survey data, is another important contributor to congestion.

The estimated values of rail PCSDR range wildly, from 0.37 to 1.01, and the more developed the rail transit system becomes, the lower its value will be. In vast metropolitan subway networks (like those in Beijing, Shanghai, and Guangzhou), during peak hours a horde of people fight their way off the train while another such horde barely waits before fighting their way on, and the train can hardly get moving because of all the people crammed in and blocking doors; subway attendants help by shoving the last people onto the train. In

Table 5: Estimated result on TSDR.

City	Year	Time period	Bus PCSDR	Rail PCSDR	Taxi PCSDR	Car PCSDR	TSDR
Beijing	2011	Peak hour in working day	0.82	0.50	0.91	0.22	0.47
Guangzhou	2010	Peak hour in working day	0.92	0.37	0.95	0.56	0.68
Shanghai	2011	Peak hour in working day	0.81	0.57	0.93	0.60	0.69
Nanjing	2012	Peak hour in working day	0.86	0.84	0.90	0.66	0.76
Tianjin	2011	Peak hour in working day	0.83	0.96	0.86	0.76	0.81
Shenzhen	2012	Peak hour in working day	0.94	0.70	1.03	0.77	0.84
Shenzhen	2013	Peak hour in working day	0.97	0.71	0.98	0.90	0.92
Shenyang	2010	Peak hour in working day	0.87	—	0.75	0.73	0.82
Shenyang	2011	Peak hour in working day	0.88	0.90	0.75	0.75	0.83
Shenyang	2012	Peak hour in working day	0.99	0.83	0.81	0.81	0.89
Hangzhou	2013	Peak hour in working day	0.90	1.01	0.86	0.84	0.93
Beijing	2011	Off-peak hour in working day	1.13	0.96	0.85	0.93	0.97
Beijing	2011	Weekend daytime	1.09	0.96	1.00	0.90	1.02
Shanghai	2011	Weekend daytime	1.07	1.12	0.96	1.05	1.06
Nanjing	2012	Weekend daytime	1.10	1.18	1.17	1.13	1.13

Table 6: Value range of the TSDR under different transport conditions.

Value range	Significance	Transport conditions
$\omega \leq 0.8$	Scarce supply	Urban transport system cannot meet the challenge of residents travel: excessively crowded rail transit and bus, severe congestion on the road net, and taxi shortage.
$0.8 < \omega \leq 0.95$	Short supply	In the general case, urban transport system may meet the challenge of residents travel. However, when a sudden event (even a small perturbation) or bad weather is encountered, rail transit will be crowded, many roads will become jammed, and buses will be delayed. In other words, the system has weak ability to withstand disturbance.
$0.95 < \omega \leq 1.10$	Supply-demand balance	In most situations, urban transport systems can meet the challenge of residents travel and have self-adjustment ability. When a sudden event or bad weather is encountered, some roads will become jammed, and the number of rail and bus passengers will increase. These disturbances will often be quelled in short times without the interposition of managers.
$1.10 \leq \omega$	Excess supply	Under any circumstances the urban transport system can meet the challenge of residents travel.

spite of such conditions, more and more people are willing to choose subway travel because this is the only way to arrive at their destination on schedule. It is a kind of inevitable phenomenon that the demand for rail transit in Chinese cities will exceed the supply both now and in the future.

(2) Transport Conditions Represented by the Value Range of the TSDR. The TSDR's estimated values, listed in Table 5, range from 0.47 to 1.13 and include all transportation conditions of the Chinese cities studied. After long-time observation of these conditions and using the analysis above, TSDR has been classified into four grades. Table 6 shows the relationship of the supply-demand ratio's value range and the transport conditions.

(3) Political Direction. The TSDR's four grades reflect only conditions during a particular period. In most Chinese cities, prevailing transport conditions are typically in a "scarce supply" or "short supply" condition, so pushing up transport supply has become the main attention focus of policymakers.

In Table 5, the Shenzhen TSDR increased 0.08 from 2012 to 2013 because the new urban road was 181 km long, the reformed road was 133 km long, and the new bus lanes were at least 100 km long. At the same time, the PCSDR values of bus, rail, and car are rising; road net expansion is therefore helpful in improving the service levels of transport system.

The value of Shenyang's TSDR also grew from 0.82 in 2010 to 0.89 in 2013, because the Shenyang metro line 1 (operational with 27.8 km and 22 stations) and line 2 (with 27.36 km and 21 stations), respectively, opened on September 27, 2011, and January 9, 2012. The formation of the Shenyang metro network improves the supply capacity of the transport system, but, at the same time, it changed the structure of transportation supply and demand. With more and more

people choosing subway travel, the value of TSDR rose, while the values of bus, taxi, and car PCSDR dropped.

In most Chinese cities, it is feasible to think that pushing up transport supply can be adopted to increase the transport system's efficiency in the near future. However, in metropolitan areas, this approach has lost its foundation because of limited urban space, and reducing travel demand is palliative.

The worst metropolitan transport situation is described in Table 6. It is difficult to improve because it is rooted in various factors, such as greater and greater urban population, growing city areas, and uneven distribution of public facilities. To fundamentally optimize urban transport systems, traffic policies such as public transport priority, limited new vehicle registrations, vehicle bans, and rail transit network construction are not at work, and it is necessary to do more effective and feasible overall urban planning.

5. Conclusion

This paper has developed a method for estimating the TSDR and completed the following tasks: (1) the TSDR estimation model was constructed using VENSIM, after idealization based on system dynamic principles, and (2) the estimated TSDR results were analyzed by comparison with fifteen data sets about the eight cities' transport conditions refined through long-time observation.

The model can provide a basis for transport policy-making because it shows and quantifies the interaction between transport system supply and demand. The TSDR values symbolize the specific transport conditions and a synthetic result of economic, policy, and traffic development. At the same time, the contribution from traffic policies or measures to the TSDR can be evaluated, so investment projections and transport policies or measures can be ordered and screened.

The results of the model will be different for the various selected regions. The TSDR values in the paper reflect the collective transport condition of the cities, but the unequal population density of each region in a city leads to imbalance of their TSDR values. For example, in Guangzhou in 2012 the urban population density was 2060 persons per square kilometer, while values for Yuexiu and Nansha were 15112 and 795, and their TSDR's values were 0.52 and 0.91 during working-day peak hours, so the geographical scope for the model should be selected according to the particular regional goals for transport policies or measures.

Taken together, this paper sheds light on the nature of likely interaction between transportation supply and demand. However, much work remains to be done because the idealized transport system considered here has a certain distance from reality. Other aspects that clearly deserve further research involve bicycles and motorcycles on the urban road net and changes in the traveling intensity and modes.

Conflict of Interests

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

Acknowledgment

This work was partially supported by National Natural Science Foundation of China (Project no. 51278158).

References

[1] W.-L. Jin, "A kinematic wave theory of lane-changing traffic flow," *Transportation Research Part B: Methodological*, vol. 44, no. 8-9, pp. 1001–1021, 2010.

[2] W.-L. Jin, "Continuous kinematic wave models of merging traffic flow," *Transportation Research Part B: Methodological*, vol. 44, no. 8-9, pp. 1084–1103, 2010.

[3] R. Liu, D. Van Vliet, and D. Watling, "Microsimulation models incorporating both demand and supply dynamics," *Transportation Research A: Policy and Practice*, vol. 40, no. 2, pp. 125–150, 2006.

[4] G. D. Wang, H. Shao, and D. X. Cao, "A mixed equilibrium traffic assignment model for transportation networks with ATIS under demand and supply uncertainties," in *Proceedings of the International Joint Conference on Computational Sciences and Optimization (CSO '09)*, vol. 2, pp. 132–136, IEEE Computer Society, Sanya, China, April 2009.

[5] L. Xu and Z. Y. Gao, "Bi-objective urban road transportation discrete network design problem under demand and supply uncertainty," in *Proceedings of the IEEE International Conference on Automation and Logistics (ICAL '08)*, pp. 1951–1955, IEEE, Qingdao, China, September 2008.

[6] M. E. Ben-Akiva, S. Gao, Z. Wei, and Y. Wen, "A dynamic traffic assignment model for highly congested urban networks," *Transportation Research Part C: Emerging Technologies*, vol. 24, pp. 62–82, 2012.

[7] C. Zhang, X. Chen, and A. Sumalee, "Robust Wardrop's user equilibrium assignment under stochastic demand and supply: expected residual minimization approach," *Transportation Research B: Methodological*, vol. 45, no. 3, pp. 534–552, 2011.

[8] C. Yuan, D. Wei, and H. Liu, "The impact of service refusal to the supply-demand equilibrium in taxicab market," in *Proceedings of the 93rd Annual Meeting of the Transportation Research Board*, Washington, DC, USA, January 2014.

[9] L. Budd, S. Ison, and T. Budd, "An empirical examination of the growing phenomenon of off-site residential car parking provision: the situation at UK airports," *Transportation Research Part A: Policy and Practice*, vol. 54, pp. 26–34, 2013.

[10] A. Bar-Yosef, K. Martens, and I. Benenson, "A model of the vicious cycle of a bus line," *Transportation Research Part B: Methodological*, vol. 54, pp. 37–50, 2013.

[11] G. Karon, "Travel demand and transportation supply modelling for agglomeration without transportation model," in *Activities of Transport Telematics*, J. Mikulski, Ed., pp. 284–293, Springer, New York, NY, USA, 2013.

[12] H. G. Kocken, I. Emiroglu, C. Guler et al., "The fractional transportation problem with interval demand, supply and costs," in *Proceedings of the International Conference on Mathematical Sciences and Statistics (ICMSS '13)*, Z. K. Eshkuvatov, A. Kilicman, and L. W. June, Eds., pp. 339–344, American Institute of Physics, Melville, Australia, 2013.

[13] Y. Sheffi and C. Daganzo, "Hypernetworks and supply-demand equilibrium obtained with disaggregate demand models," *Transportation Research Record*, no. 673, pp. 113–121, 1978.

[14] H. Al-Deek and E. B. Emam, "New methodology for estimating reliability in transportation networks with degraded link capacities," *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, vol. 10, no. 3, pp. 117–129, 2006.

[15] R. H. Liu, T. May, and S. Shepherd, "On the fundamental diagram and supply curves for congested urban networks," *Transportation Research Part A: Policy and Practice*, vol. 45, no. 9, pp. 951–965, 2011.

[16] D.-Y. Lin, N. Eluru, S. T. Waller, and C. R. Bhat, "Integration of activity-based modeling and dynamic traffic assignment," *Transportation Research Record*, vol. 2076, pp. 52–61, 2008.

[17] L. M. Gardner, S. D. Boyles, and S. T. Waller, "Quantifying the benefit of responsive pricing and travel information in the stochastic congestion pricing problem," *Transportation Research Part A: Policy and Practice*, vol. 45, no. 3, pp. 202–218, 2011.

[18] S. Y. Liu, K. P. Triantis, and S. Sarangi, "A framework for evaluating the dynamic impacts of a congestion pricing policy for a transportation socioeconomic system," *Transportation Research Part A: Policy and Practice*, vol. 44, no. 8, pp. 596–608, 2010.

[19] M. Estrada, M. Roca-Riu, H. Badia, F. Robusté, and C. F. Daganzo, "Design and implementation of efficient transit networks: procedure, case study and validity test," *Transportation Research Part A: Policy and Practice*, vol. 45, no. 9, pp. 935–950, 2011.

[20] S. Y. Auyang, *Foundations of Complex-system Theories*, Cambridge University Press, Cambridge, UK, 1999.

[21] J.-F. Wang, H.-P. Lu, and H. Peng, "System dynamics model of urban transportation system and its application," *Journal of Transportation Systems Engineering and Information Technology*, vol. 8, no. 3, pp. 83–89, 2008.

[22] Z.-L. Cheng and X.-G. Qiu, "Overview of traffic system modeling research," in *Proceedings of the 31st Chinese Control Conference (CCC '12)*, pp. 7279–7285, July 2012.

[23] D. Levinson and B. Yerra, "Self-organization of surface transportation networks," *Transportation Science*, vol. 40, no. 2, pp. 179–188, 2006.

[24] S. Shepherd, "A review of system dynamics models applied in transportation," *Transportmetrica B: Transport Dynamics*, vol. 2, no. 2, pp. 83–105, 2014.

[25] E. Cascetta, *Transportation Systems Analysis: Models and Applications*, Springer, 2009.

[26] Z.-A. Zhang and S.-W. Feng, "Game analysis of private cars and government-owned vehicles under road pricing regulation," *Journal of Transportation Systems Engineering and Information Technology*, vol. 2, article 013, 2008.

[27] Y.-H. Li, Z.-Z. Yuan, X.-H. Xie et al., "Analysis on trips characteristics of taxi in Suzhou based on OD Data," *Journal of Transportation Systems Engineering and Information Technology*, vol. 7, no. 5, pp. 85–89, 2007.

[28] Z. Xizhao, L. Canqi, and Y. Peikun, "Time—space sources of the network of urban road and space—capacity of traffic," *Journal of Tongji University*, vol. 24, no. 2, pp. 392–397, 1996.

[29] Z. X. L. Chaoyang, "Supply-demand models for time and space resources of urban road and their application," *Journal of Shanghai Maritime University*, vol. 3, p. 2, 1999.

[30] J. Odeck and S. Bråthen, "Travel demand elasticities and users attitudes: a case study of Norwegian toll projects," *Transportation Research Part A: Policy and Practice*, vol. 42, no. 1, pp. 77–94, 2008.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 680609, 9 pages
http://dx.doi.org/10.1155/2015/680609

Research Article

A Stochastic Programming Approach on Aircraft Recovery Problem

Bo Zhu, Jin-fu Zhu, and Qiang Gao

College of Civil Aviation, Nanjing University of Aeronautics and Astronautics, Nanjing 211100, China

Correspondence should be addressed to Bo Zhu; acoustic.zhu@gmail.com

Received 12 May 2015; Revised 16 August 2015; Accepted 23 August 2015

Academic Editor: Yuanchang Xie

The unexpected aircraft failure is one of the main disruption factors that cause flight irregularity. The aircraft schedule recovery is a challenging problem in both industrial and academic fields, especially when aircraft restoration time is uncertain, which is often ignored in previous research. This paper established a two-stage stochastic recovery model to deal with the problem. The first stage model was a resource assignment model on aircraft schedule recovery, with the objective function of minimizing delay and cancellation cost. The second stage model used simple retiming strategy to adjust the aircraft routings obtained in the first stage, with the objective function of minimizing the expected cost on recourse decision. Based on different scenarios of restoration time, the second stage model can be degenerated as several linear models. A stochastic Greedy Simulated Annealing algorithm was designed to solve the model. The computational results indicate that the proposed stochastic model and algorithm can effectively improve the feasibility of the recovery solutions, and the analysis of value of stochastic solution shows that the stochastic model is worthy of implementation in real life.

1. Introduction

In dynamic operation circumstances, airlines flight schedules will face different kinds of inevitable stochastic disruptions and will deviate from regular operations. As the development of air transportation, the flight schedule is planned pretty tight and the disruption often propagates in the flight network. Flight irregularity is a serious and widespread problem all over the world, which imposes significant cost to airlines, passengers, and the society. In 2013, the average on-time ratio was 78.4% in the U.S. according to the 16 main carriers' data from BTS. Each irregular flight will bring around $16,600 loss on average, including expenses for fuel, maintenance, crew, the passenger time loss, and estimate of welfare loss. In China, the average on-time ratio was only 72.34% in 2013. The average delay time increased to almost 60 minutes, and around 2,100 irregular flights were handled per day. 40% of irregular flights are caused by airlines themselves, which is the most compared to other disruption factors in China. Aircraft breakdown, schedule temporary change, passenger issues, and so forth will hinder the flights operation regularly, numerous flights will be disrupted, and thousands of passenger itineraries will be destroyed. Aircraft are the most treasured resources for airlines; it is significant for dispatchers to retime the flight schedule and reassign aircraft and crews to recover the flight schedule as soon as possible.

The research on flight recovery problem has more than 60 years history, among which the aircraft recovery problem (ARP) is one of the most concerned. Teodorović and Guberinić studied how to recover the flight schedule to minimize the total passengers delay when unexpected aircraft failure happened. They used the branch and bound algorithm to solve some small scale examples [1]. Argüello et al. discussed the flight schedule recovery problem with temporary shortage of aircraft and applied GRASP algorithmic framework to rearrange aircraft routings [2]. Rosenberger et al. studied the aircraft schedule recovery problem under shortage of aircraft or change in airport capacity. They designed heuristic algorithm framework to solve the model [3]. Bratu and Barnhart studied the flight delay and cancellation decision considering the passenger arrival delay cost [4]. Tang et al. revised the GRASP method and designed Greedy Simulated

Annealing (GSA) method to solve the recovery model [5]. Eggenberg et al. developed a column generation scheme to solve ARP [6]. Petersen et al. are known as the first scholars that studied the full integrated recovery formulation and approach with computational results presented [7]. Le and Wu presented iterative tree growing with node combination method to solve aircraft and crew recovery simultaneously [8]. Chan et al. established a model that integrates aircraft and passenger recovery, but no solution was offered [9]. Sinclair et al. designed a large neighborhood search heuristic algorithm to solve the integrated recovery of aircraft and passenger [10]. Hu et al. solved the integrated recovery problem of aircraft and passenger based on reduced time-band network and passenger transiting relationship [11]. Although some theoretical researches show good results in computational tests, they can barely be implemented well in real world because of the following reasons. Firstly, the disruptions are simply assumed as deterministic. For example, the restoration time of aircraft is assumed to be known as constant before decision making, which is usually hard to predict precisely even for the sophisticated maintenance staff. Secondly, as in dynamic circumstances, the recovery solution from deterministic model may be lack of robustness in operation. When the random variables become realized as time passes, the previous recovery plan may be infeasible or not satisfactory. Thus, it is necessary to study the stochastic model and algorithm on the problem. There are some researches on uncertain theory in air transportation field, such as design and optimization on flight network [12] and the flight scheduling problem [13–15]. In airline operation area, Rosenberger et al. worked on the simulation software that controls the uncertain delay time [16]. Mou and Zhao built an uncertain programming model with chance constraint and solved it based on classic Hungarian algorithm to deal with the recovery problem under stochastic flight time [17]. Arias et al. proposed a combined methodology using simulation and optimization techniques to cope with the stochastic aircraft recovery problem [18].

In this paper, we developed a two-stage stochastic model to formulate the stochastic ARP and designed a stochastic algorithm based on GSA to solve the model. As far as we know, this paper is the first to bring the uncertain aircraft restoration time into the recovery problem.

2. Problem Statement and Model

When aircraft failure happens, there are several strategies to recover the flight schedule back to the regular status. The basic strategies to recover the flight timetable are delay and cancellation. For aircraft rerouting problem, strategies such as aircraft swap, type substitution, reserved aircraft, and ferry can be used. In Figure 1, a small example of aircraft routings is illustrated. The grey area means aircraft A2 found failure at 07:30, and the anticipated recovery time will be 13:15. The Airline Operation Control Center (AOCC) can choose strategy to cancel flights 4 and 5; or they can just delay flights 4–7 in a row; or aircraft A1 and A2 can switch routings at 08:00 and so forth. The figure shows a classic deterministic aircraft schedule recovery problem, and all the rescheduled plans are generated on the premise that the recovery time of A2 is known in advance.

However, the recovery time above is an expected value which is usually given by airline maintenance staff. The value barely equals the actual one, which may make the current recovery plan not satisfactory or even infeasible. For example, if at 07:30 AOCC chooses to delay flight 4 until 13:15, but when it comes to the time 13:15, aircraft A2 is not available to use yet, more delays or cancellations will be incurred. Anther situation is that A2 is ready for use earlier than 13:15; then, a more cost-saving plan might be optional. Since the new disruption information will be updated frequently, it will be time consuming to redo the whole optimization iteratively. An intuitive thought is to generate a robust recovery plan and when the random restoration time of aircraft is determined, it is still feasible and satisfactory with simple recourse decision.

In this paper, the concept of stochastic aircraft recovery time is introduced, and a two-stage aircraft schedule recovery model is established. The classic two-stage stochastic fixed recourse linear model is proposed by Dantzig [19] and Beale [20]. The model is designed to choose one decision, which makes the cost of current decision and the expectation of future recourse cost minimized [21]. For flight recovery problem, the two-stage model can evaluate the influences of different rescheduled plans and the uncertainty of the disruption factors, thereby making robust decisions. In our model, the first stage model is the deterministic resource assignment model of ARP. Based on different stochastic scenarios of aircraft recovery time, the recourse model will adjust the recovery plan obtained in the first stage. The strategy of recourse model is retiming the flights but maintaining aircraft routings generated in the first stage. It ensures the feasibility of recourse model and the simple linear formulation can guarantee the computational speed. Cancellation and aircraft swap can also be implemented as strategies in recourse model, but they will not change the essence of the model.

2.1. Stochastic Model. In the research of deterministic aircraft recovery problem, resource assignment model is one of the most prevalent ones because it can describe the problem in a complete and concise way. Our first stage model is referred to Argüello et al.'s model [2]. Flights are implicitly generated as routings which will be assigned to aircraft. The notions are defined as follows:

(1) Sets are as follows:

F: flight set, indexed by i.
K: available aircraft set, indexed by k.
A: airport set, indexed by a.
P: feasible aircraft routing set, indexed by j.

(2) Parameters are as follows:

$a_{i,j}$: equal to 1 if flight i is in aircraft routing j, otherwise, equal to 0.
$b_{j,a}$: equal to 1 if aircraft routing j will end at airport a, otherwise, equal to 0.

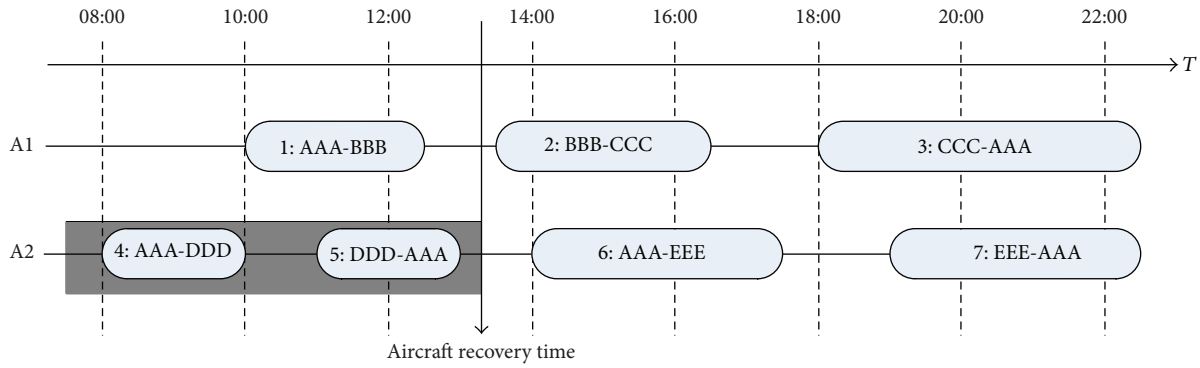


Figure 1: A small case of aircraft routings.

c_i: the cancellation cost of flight i.

h_a:, the required amount of aircraft at airport a, at the end of recovery process.

d_j^k: the delay cost of assigning aircraft k to routing j.

(3) Decision variables are as follows:

x_j^k: equal to 1 if aircraft k is assigned to routing j, otherwise, equal to 0.

y_i: equal to 1 if flight i is cancelled, otherwise, equal to 0.

Using the above notations, the first stage resource assignment model for aircraft recovery problem is

$$\min \quad Z = \sum_{k\in K}\sum_{j\in P} d_j^k x_j^k + \sum_{i\in F} c_i y_i \tag{1}$$

$$\text{s.t.} \quad \sum_{k\in K}\sum_{j\in P} a_{i,j} x_j^k + y_i = 1, \quad \forall i \in F \tag{2}$$

$$\sum_{k\in K}\sum_{j\in P} b_{ja} x_j^k \geq h_a, \quad \forall a \in A \tag{3}$$

$$\sum_{j\in P} x_j^k = 1, \quad \forall k \in K \tag{4}$$

$$x_j^k = 0, 1, \quad \forall (j,k) \in P \times K \tag{5}$$

$$y_i = 0, 1, \quad \forall i \in F. \tag{6}$$

The objective function (1) minimizes the cost of flight delay and cancellation. Constraints (2) are flight coverage constraints. For any flight i, it either be cancelled or assigned to a routing. Constraints (3) are aircraft balance constraints, which require certain amount of aircraft in different airports at the end of recovery process to preserve the future regular operation. Constraints (4) confine that each aircraft can only be assigned to one routing. Constraints (5) and (6) are nonnegative constraints for decision variables.

The deterministic model has an underlying work: the aircraft routings are already generated on the premise that aircraft recovery time is fixed. However, as we mentioned above, it is hard to determine the time in real operation. The research on aircraft reliability and maintainability [22] also supports this point of view. Therefore, an expected cost that is incurred by stochasticity is added to the optimization model; it reflects the possible changes of the rescheduled plans in the first stage. The general stochastic model formulation is as follows:

$$\text{RP} = \min W = \min \quad \left(Z + \mathbb{Q}(x, y)\right) \tag{7}$$

$$\text{s.t.} \quad \mathbf{A}[x, y] = \mathbf{b} \tag{8}$$

$$x, y = 0, 1. \tag{9}$$

The objective function (7) of the stochastic model consists of two parts. The first one is the objective function (1); the second one $\mathbb{Q}(x, y)$ is the expected cost of the future recourse decision on the rescheduled plans obtained at the first stage. Here, and in the following text, x and y are the simplified symbols which denote x_j^k and y_i in the deterministic model, respectively. Formula (8) is the general form of constraints (2)–(4). Constraints (9) are the nonnegative constraints. It is a standard two-stage recourse stochastic integer programming model.

2.2. Recourse Model. Since operations of flight schedule weave so many resources together, frequent severe changes on recovery plan are not preferred. Thus, it is meaningful to get a flexible, robust but also cost-saving recovery plan when disruption happens. Particularly, as the time passes, when the uncertain variables are determined, the selected rescheduled plan in the first stage can be implemented smoothly with or without minor adjustments. To obtain such rescheduled plan quickly is more acceptable than simply the pursuit of optimal solution in deterministic model of NP-Hard problem.

Figure 2 illustrates one rescheduled plan obtained in the first stage model from the same example in Figure 1. It swaps aircraft routings of aircraft A1 and A2 and delayed the flights 1–3. Obviously, the plan is drawn on the given aircraft recovery time, which is the end of the grey interval. In reality, the A2 recovery time may be a random variable with

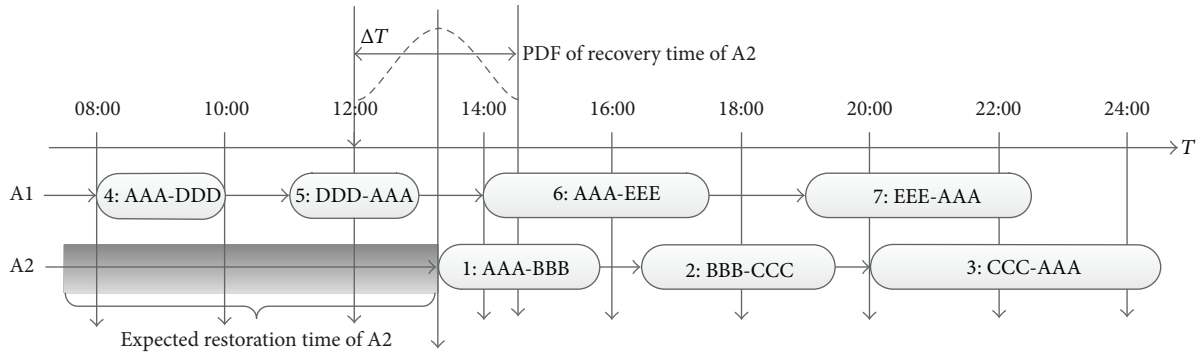


Figure 2: Illustration on recovery time of aircraft and recovery plan.

probability density function (PDF) curve in the figure, and its range is ∇T. If A2 turns out to be ready at 14:00, then flights 1–3 will be redelayed in a row; if the new arrival time of flight 3 is beyond the curfew time of airport AAA, it will be cancelled, which will break the aircraft balance also; or it will be delayed until the curfew time is over, which will impose severe delay to the flight. This situation will be reflected in terms of risk cost in the recourse model.

The deterministic aircraft schedule recovery problem is an NP-hard problem; so no algorithm can be proved to be capable of obtaining optimality in polynomial time. A quick recovery solution is preferred and sometimes required. For two-stage stochastic model, there are a bunch of recourse models to be solved on each feasible solution obtained in the first stage. It requires the recourse model to be simple to solve. Let $\mathbf{T}$ denote the aircraft recovery time vector. It consists of every disrupted aircraft recovery time, which is considered as a continuous variable, and every available time for undisrupted aircraft which is a constant variable. Then, the objective function of the recourse model can be modeled as $\mathbb{Q}(x, y) = \int \mathcal{Q}(x, y, \mathbf{T})\mathrm{d}\mathbf{T}$. Since the objective function is nonlinear and the PDF of the random variable is usually hard to obtain as well, we can discretize the aircraft recovery time without losing precision. The combinations of discretized points from every aircraft construct the finite scenario set Ω. Let $\omega \in \Omega$ denote one scenario (combination), and $\Pr(\omega)$ is the probability of ω. The recourse cost of the rescheduled plan can be expressed in the following:

$$\mathbb{Q}(x, y) = \sum_{\omega} \mathcal{Q}(x, y, \omega) \Pr(\omega). \tag{10}$$

Besides notions in the first stage model, some other notions used in recourse model are listed as follows:

(1) Parameters are as follows:

t_i: fly time of flight i.

ρ_i: unit delay cost of flight i (per minute).

d_i^{s}: original scheduled time of departure of flight i.

$d(i)$: departure airport of flight i.

$r(i)$: arrival airport of flight i.

t_a: starting time of curfew on airport a.

φ_i: cost of breaking curfew regulation of flight i.

g_k: minimum turnaround time of aircraft k.

t_ω^k: recovery/ready time of aircraft k under scenario ω; so the random vector $\xi(\omega) = (t_\omega^k, k = 1, \ldots, |K|)$.

Δ_i: delay time of flight i obtained from optimization on the first stage.

$p(i)$: predecessor flight of flight i in the same aircraft routing after optimization on the first stage.

(2) Decision variables are as follows:

$d_{i,\omega}$: new estimated time of departure of flight i under scenario ω.

$r_{i,\omega}$: new estimated time of arrival of flight i under scenario ω.

$\Delta_{i,\omega}$: estimated delay time of flight i under scenario ω.

$v_{i,\omega}$: equal to 1 if flight i violates the curfew requirement under scenario ω, otherwise, equal to 0.

The recourse model can be established as follows:

$$\begin{aligned} \min \quad & \mathbb{Q} = \sum_{\omega} \mathcal{Q}(x, y, \omega) \Pr(\omega) \\ & = \sum_{\omega} \left(\sum_{i} \rho_i (\Delta_{i\omega} - \Delta_i) + \varphi_i v_{i,\omega} \right) \Pr(\omega) \end{aligned} \tag{11}$$

$$\text{s.t.} \quad r_{i,\omega} = d_{i,\omega} + t_i, \quad \forall i \in F, \ \forall \omega \in \Omega \tag{12}$$

$$d_{i,\omega} - d_i^{\mathrm{s}} = \Delta_{i,\omega}, \quad \forall i \in F, \ \forall \omega \in \Omega$$

$$d_{i,\omega} \geq t_\omega^k \sum_{j \in P} a_{i,j} x_j^k, \quad \forall (i,k) \in F \times K, \ \forall \omega \in \Omega \tag{13}$$

$$d_{i,\omega} - r_{p(i),\omega} \geq g_k \sum_{j \in P} a_{i,j} x_j^k, \quad \forall (i,k) \in F \times K, \ \forall \omega \in \Omega \tag{14}$$

$$\begin{aligned} &v_{i,\omega:d_{i,\omega}\geq t_{d(i)}\|r_{i,\omega}\geq t_{r(i)}} \geq 1, \quad \forall i \in F,\ \forall \omega \in \Omega \\ &v_{i,\omega:d_{i,\omega}<t_{d(i)}\&r_{i,\omega}<t_{r(i)}} \leq 0, \quad \forall i \in F,\ \forall \omega \in \Omega \end{aligned} \tag{15}$$

$$\begin{aligned} &d_{i,\omega}, r_{i,\omega}, \Delta_{i,\omega} \geq 0, \quad \forall i \in F,\ \forall \omega \in \Omega \\ &v_{i,\omega} = 0, 1, \quad \forall i \in F,\ \forall \omega \in \Omega. \end{aligned} \tag{16}$$

The objective function of the recourse model (11) minimizes the expected cost on the recourse strategy of the first stage plan due to uncertainty of aircraft recovery time. For each scenario ω, the recourse cost contains two parts: one is cost of flight retiming; if the aircraft can be ready before the expected time, $\Delta_{i\omega} - \Delta_i$ will be negative value; the other one is the risk cost of curfew breaking after retiming in the recourse stage. Constraints (12) are flight consistency constraints, which define the relationship between flight departure, arrival, and delay time. Constraints (13) require that aircraft cannot fly flights until it is ready. Constraints (14) require the minimum turnaround time (MTT) of adjacent flights in one aircraft routing. Notice that variables t_ω^k and $p(i)$ depend on the decision variables x and y in the first stage; once they are determined and passed to the recourse model, constraints (13) and (14) are degenerated to be linear constraints. Constraints (15) determine the value of $v_{i,\omega}$ for each flight i under scenario ω; if new departure/arrival time of i violates the curfew time of its departure/arrival airport, $v_{i,\omega}$ is forced to be 1; otherwise, it should be 0. Constraints (16) are the nonnegative constraints of the model.

3. Algorithm

Although solving the deterministic model of ARP is already very complicated; there are some successful results on the research of the algorithm. Precise algorithm such as column generation [6] and heuristic methods such as GRASP [2] and GSA [5] can obtain satisfactory solutions in tractable time. To extend the deterministic model to stochastic, one will make the scale of the problem larger and increase the computation complexity. If the recourse model is linear and the scenarios are limited, the stochastic model can always be transferred to equivalent deterministic model, which makes the problem scale even larger. Different from long-term scheduling problem, a quick solution is required for recovery problem. To solve the two-stage stochastic model efficiently, we design stochastic algorithm framework combining GSA [5] and simple retiming strategy, which concerns the stochastic structure of the problem.

For the first stage model, decision variables x and y can be obtained through GSA algorithm. As paper [5] states, the algorithm has 3 steps generally as follows:

(1) Construct initial feasible solution by delaying the disrupted aircraft routings.

(2) Generate neighboring solutions through 5 operations on disrupted pairs of aircraft routings. Disrupted pair of aircraft routings refers to two aircraft routings which include at least one disrupted aircraft in order to conserve the undisrupted aircraft routings. The 5 operations on each disrupted pair of aircraft routings are flight cycle insertion, flight string insertion to the routing tail, flight string swap, tail string swap, and flight cycle cancellation.

(3) Choose neighboring solutions from a Restricted Candidate List (RCL, which consists of cost-saving neighbors) or a Back Restricted Candidate List (BRCL, which consists of cost-increase neighbors) to substitute the original routings to get new solution.

Once the decision variables in the first stage model are passed to the second stage model, the model can be degenerated to several easy-solving linear optimization recourse models. As the aircraft routing's flights are fixed in the recourse model, the only adjustment of the rescheduled plan is to retime the flights as tight as possible (only consider aircraft recovery time and MTT) to minimize the objective function as long as it is feasible. If the retiming solution is not feasible, which can only be curfew violation under such condition, there are two ways to deal with this: one is to cancel the violation flight; the other one is to delay the flight until curfew is over. The cost of curfew breaking will be represented as φ_i. Therefore, $\mathcal{Q}(x, y, \omega)$ will be obtained for each scenario $\omega \in \Omega$, and $\mathbb{Q}(x, y)$ can be computed according to (10) since scenarios are independent of each other.

The end criterion can be set as upper bound of the computational time since the recovery problem requires quick response in real operation. It can also be set as enough number of solutions or maximum iteration number, which will give the decision maker a lot of flexibility in operation. The detail algorithm steps are described as follows.

Step 1 (initialization). Let l denote iteration number. Z^0 denotes the initial objective function value and Z^l denotes the objective function value of the first stage model in the lth iteration. $\mathcal{Q}^{l*}(\omega)$ denotes the optimal value of the recourse model under scenario ω and $\mathbb{Q}^{l*}$ denotes the optimal value of the second stage model in lth iteration. W^* denotes the best objective function value of the stochastic model so far. W^l denotes the current objective function value in lth iteration of the stochastic model. Let $l = 0$, $Z^0 = Z^l = 0$, $\mathcal{Q}^{l*}(\omega) = \mathbb{Q}^{l*} = 0$, and $W^* = W^l = \infty$.

Step 2. Construct initial feasible solution in the first stage model. Take the expectation of aircraft recovery time as constant variable; then, delay the aircraft routings in a row to get Z^0. Set $l = 1$ and $Z^l = Z^0$.

Step 3. Construct neighboring solutions through 5 different operations; choose the optimal neighboring solution for every pair of aircraft routings.

Step 4. Evaluate the neighboring solution. If the neighboring solution can decrease the objective function value of the first stage model, then add it to RCL; otherwise, add it to BRCL.

Step 5. If RCL is not empty, randomly select some neighboring solutions. If RCL is empty, randomly select neighboring solutions from BRCL, and determine whether to accept them

according to Metropolis acceptance criterion from Simulated Annealing algorithm.

Step 6. Substitute the original aircraft routings using chosen neighboring solutions; obtain Z^l.

Step 7. Pass the values of decision variables of the current solution in first stage model to the recourse model, and degenerate the recourse model.

Step 8. For each disruption scenario ω, solve the recourse model by retiming the stochastic aircraft routings, and obtain the optimal objective function value $\mathcal{Q}^{l*}(\omega)$.

Step 9. Compute $\mathbb{Q}^{l*} = \sum_\omega \mathcal{Q}^{l*}(\omega)P(\omega)$, and the total cost of the two-stage model $W^l = Z^l + \mathbb{Q}^{l*}$. If $W^l < W^*$, update $W^* = W^l$ and preserve the current recovery plan.

Step 10. Set $l = l + 1$; if end criterion is met, the current plan is the best one so far; quit the algorithm, if not, go back to *Step 3*.

The flow chart of the algorithm is illustrated in Figure 3.

3.1. Algorithm Complexity. In the lth iteration, for the first stage, an aircraft routing that consists of p flights will need $O(p)$ time to delay the flights to obtain the initial feasible solution. For a pair of aircraft routings that consist of p and q flights, respectively, the time to construct neighboring solutions by 5 different operations is $O(p^2q^2)$. Suppose the total number of aircraft routings is n; m of them are disrupted due to aircraft breakdown; there will be $m(n-1)$ combinations of routing pairs, and the time of construct neighboring solutions in l iterations is $O(lm(n-1)p^2q^2)$. For one disruption scenario, the recourse model has m disrupted aircraft routings; if they contain r flights in each routing, then the computational time on the recourse model will be $O(mr)$. Suppose every disrupted aircraft has ω discrete recovery times; the scenarios of the whole problem will be ω^m, and the objective function of recourse model needs $O(mr\omega^m)$ to compute. In real world operation, an aircraft cannot execute too many flights in one day; usually $p, q, r \leq 10$. To sum up, the algorithm time complexity will be $O(lnm^2\omega^m)$. It is almost impossible that many aircraft have unexpected maintenance at the same time; so the value of m cannot be very large; meanwhile, since there is no need to discretize the aircraft restoration time interval to get too many points as we mentioned before, the value of ω for each aircraft cannot be very large. Thus, in real operations, the computational time can be controlled and the algorithm can be regarded as quasilinear.

4. Computational Test

A case from a Chinese airline is studied in this section. Table 1 shows the original flight schedule snapshotted from the daily flight schedule. Std and Sta in the first row mean original scheduled time of departure and arrival, respectively.

Table 1: Original flight schedule.

Aircraft	Flight	Dep.	Arr.	Std.	Sta.
A1	F11	ICN	PVG	12:50	14:35
	F12	PVG	MFM	15:40	18:20
	F13	MFM	PVG	19:10	21:40
	F14	PVG	HAN	22:40	26:00
A2	F21	PVG	CSX	12:10	14:00
	F22	CSX	PVG	14:55	16:30
	F23	PVG	HAK	17:30	20:25
	F24	HAK	PVG	21:20	23:40
A3	F31	HRB	PVG	13:30	16:10
	F32	PVG	HAK	17:05	19:55
	F33	HAK	PVG	20:50	23:20
A4	F41	CGQ	PVG	11:40	14:00
	F42	PVG	KWL	14:45	17:15
	F43	KWL	CAN	18:10	19:10
	F44	CAN	KWL	20:10	21:00
	F45	KWL	PVG	21:55	23:55
A5	F51	PVG	CGQ	13:05	15:25
	F52	CGQ	PVG	16:20	18:45
	F53	PVG	DYG	19:15	21:30
	F54	DYG	PVG	22:25	24:05
A6	F61	TAO	PVG	11:00	11:55
	F62	PVG	CTU	14:55	18:20
	F63	CTU	PVG	19:15	21:35

Table 2: Probability distribution of restoration time for A4.

Restoration time (min)	200	260	300	330	360	390	450
Probability	0.14	0.15	0.2	0.15	0.13	0.12	0.11

23 flights that operated by 6 aircraft are studied in this case. The unit delay cost ρ_i is set to be 20 per minute, and one cancellation cost is regarded to be equivalent as 8 hours delay, which is 9,600 minutes. The cost of breaking curfew regulation φ_i is 10,000 considering possible cancellation and aircraft balance. All the domestic airports have the same curfew time window [02:00, 06:00] every day; some foreign airports such as HAN do not have curfew time. A4 undergoes unexpected failure at 11:40, according to maintenance staff; its expected restoration time is 320 minutes; that is, it will be ready at 17:00. The probability distribution of the restoration time is presented in Table 2, which is obtained by statistical data from airline. The range of the restoration time varies from 200 to 450 minutes. The tests are performed on a laptop with 4 GB installed RAM and i5-3317U CPU 1.70 GHz.

Based on the definition in paper [23], the first stage model uses the expected values of random variables; it has model formulation $\mathrm{EV} = \min_{x_1} \Phi(x_1, E(\varepsilon))$, where x_1 represents the decision variables in the first stage model and ε represents the random variables in the recourse model. Our first stage deterministic resource assignment model is an instance of EV. The expected value of EV can be represented as

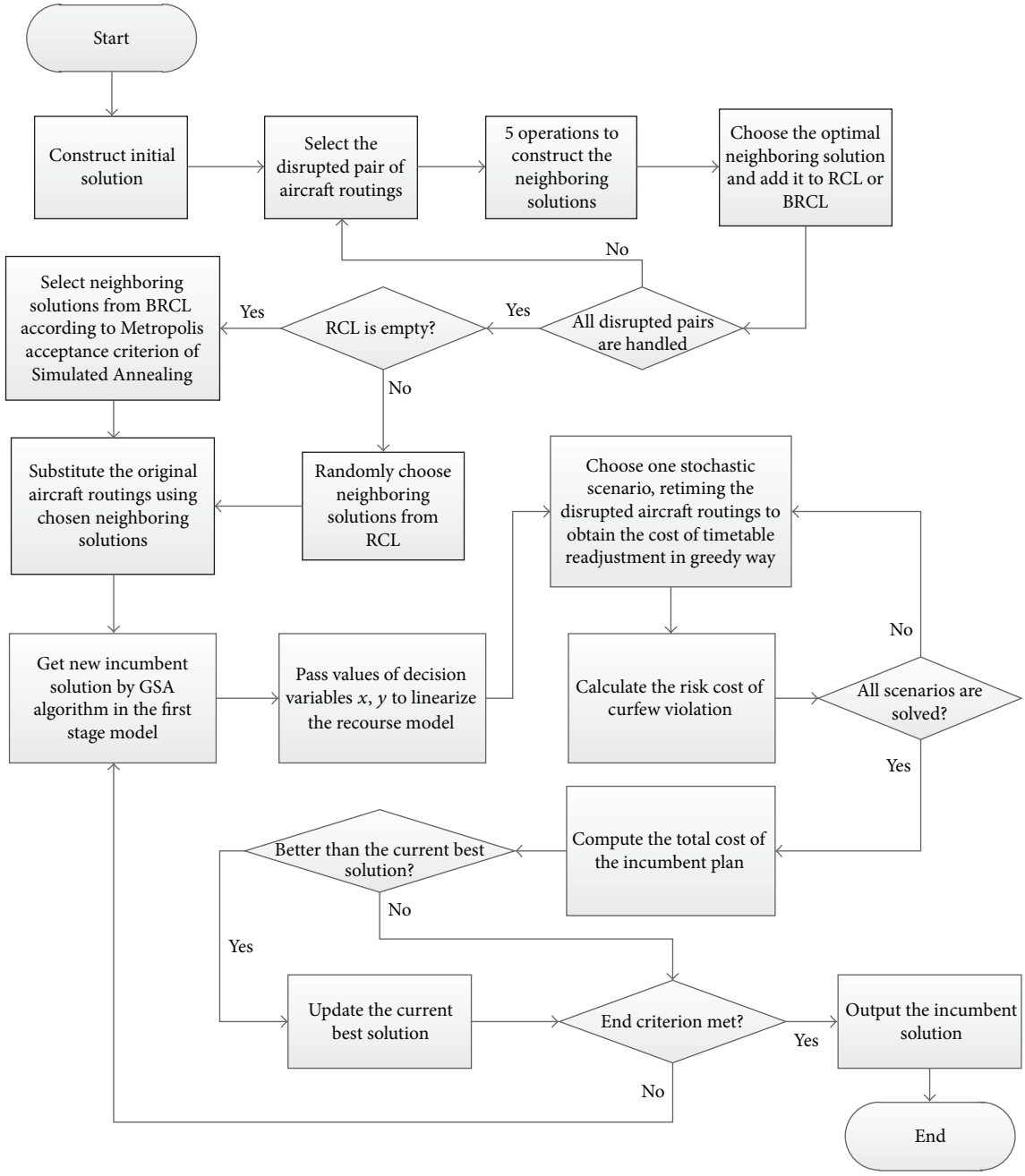


Figure 3: Flow chart of algorithm.

EEV $= E_{\varepsilon}[\Phi(\bar{x}_1(\bar{\varepsilon}), \varepsilon)]$. A perfect information solution would choose optimal first period decisions for each realization of scenario. The expected value of this solution is known as "wait-and-see" (WS) solution, where WS $= E_{\varepsilon}[\min_{x_1} \Phi(x_1, \varepsilon)]$. To study the ARP in a thorough way, we solve and compare the example using models EEV, WS, and the RP proposed in Section 2, respectively.

In solving EV, the restoration time of aircraft A4 is fixed to be 320 minutes; that is, it will be ready for use until 17:00. We run the model by GSM algorithm from Steps 1 to 6; the end criterion is set as 5 minutes of computational time or 300 incumbent solutions. Table 3 lists the final solution of the deterministic model and the total delay cost EV $= Z = 15{,}900$. The deterministic solution will have expected recourse cost $\mathbb{Q} = 5{,}928$; it represents the cost of adjustment due to stochasticity if deterministic solution is implemented. The stochastic expected cost of the deterministic solution will be EEV $= W = Z + \mathbb{Q} = 21{,}828$. Similar method is used to run the WS model, where each value of probabilistic aircraft restoration time is treated as deterministic in the first stage model. After optimization and the computation on the expected value according to the probability table, we obtain

Table 3: Deterministic recovery solution.

Aircraft	Routing	Rescheduled stop over time
A1	F11/F62/F63/F14	26:00
A2	F21/F22/F12/F13	23:00
A3	F31/F32/F33	23:20
A4	F52/F23/F24	26:00
A5	F51/F41/F53/F54	24:05
A6	F61/F42/F43/F44/F45	23:55

Table 4: Stochastic recovery solution.

Aircraft	Routing	Rescheduled stop over time
A1	F11/F12/F13/F14	26:00
A2	F21/F22/F32/F33	23:20
A3	F31/F62/F63	23:15
A4	F52/F53/F54	24:40
A5	F51/F41/F23/F24	25:00
A6	F61/F42/F43/F44/F45	23:55

WS = 16,007. It represents the expected objective value if the decision maker can get the perfect deterministic values of random variables before optimization.

Two-stage stochastic model RP is solved by proposed stochastic algorithm; the end criterion is set as 10 minutes of computational time or 300 incumbent solutions. Table 4 shows the stochastic recovery plan result. The delay cost Z = 16,300, and the cost of second stage recourse model is $\mathbb{Q}$ = 4,257; so the objective function value of the stochastic model will be RP = $W = Z + \mathbb{Q}$ = 20,557.

As we can observe, recovery plan in Table 3 has shorter delay cost (15,900) compared with stochastic plan (16,300) in Table 4 without considering the uncertainty of aircraft recovery time. However, in Table 3, aircraft A4 will operate flight F24, and it will arrive at the PVG at 02:00 in the morning; there is 50% probability that F24 is cancelled or delayed a long time due to curfew breaking and will incur more cost and loss of passenger willing if that happens. That is why the stochastic recovery plan in Table 4 has lower cost (20,557) compared with deterministic solution (21,828) in Table 3 considering the recourse actions. In the former solution, most aircraft will land in the airports before some buffer time to the curfew time window. The changes brought by uncertainty can be absorbed; so the plan is more robust and flexible. More specifically, the expected value of perfect information (EVPI) can be defined as EVPI = RP−WS, which represents the effect of uncertainty in stochastic programs. The value of the stochastic solution VSS = EEV − RP, which is the difference between the result of using an expected value solution and the recourse problem solution. In this example, EVPI = RP − WS = 20,557 − 16,007 = 4,550. The uncertainty has brought about a lot of cost to the rescheduled solution. Compare the expected scenario analysis of rescheduled plans obtained from deterministic model and the objective function value of the two-stage stochastic model; we have VSS = EEV−RP = 21,828−20,557 = 1,271. The stochastic solution can decrease 5.8% of cost of deterministic solution, and that will be a great amount of operational cost to airlines.

5. Conclusion

Aircraft recovery problem due to shortage of aircraft is one of the most challenging problems in the airline operations. In this paper, the uncertainty of aircraft restoration time is introduced. The stochastic aircraft recovery problem is modeled as a two-stage stochastic recovery model. The first stage model is a deterministic resource assignment model and the second model evaluates the retiming adjustment on the solution obtained from the first stage model. Since the stochastic problem is an NP-hard problem and needs quick solution in real operation, it is impossible to use regular method to traverse the whole solution space. We designed stochastic Greedy Simulated Annealing algorithm, which combined conventional heuristic framework and simple greedy recourse method, to solve the problem. It shows the ability of obtaining satisfying solution in tractable time. A real life example is computed to analyze the proposed model and algorithm. The computational results of the stochastic model indicate the significance of considering stochastic disruption factors in the recovery problem. The study of the EVPI and VSS shows the importance of precise information and the cost-saving performance of the proposed stochastic model and algorithm.

Some interesting problems are raised for future work during the research. Besides aircraft failure, some other stochastic disruptions such as airport capacity decrease, airport temporary close, and en-route capacity change are also worthy of research. In order to increase the feasibility and accuracy of the stochastic model and algorithm in real operations, the data collection and data mining in flight irregularity should be paid more attention. Moreover, a full recovery plan for airlines consists of not only aircraft routings but also crew pairings and passenger new itineraries. Thus, research on integrated stochastic flight recovery problem is one of the interests in future work.

Conflict of Interests

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

Acknowledgments

This work was supported by National Nature Science Foundation of China (nos. 61079014 and 71171111), funding of Outstanding Doctoral Dissertation in NUAA (no. BCXJ13-14), and funding of Jiangsu Innovation Program for Graduate Education (no. CXZZ13_0174).

References

[1] D. Teodorović and S. Guberinić, "Optimal dispatching strategy on an airline network after a schedule perturbation," *European Journal of Operational Research*, vol. 15, no. 2, pp. 178–182, 1984.

[2] M. F. Argüello, J. F. Bard, and G. Yu, "A GRASP for aircraft routing in response to groundings and delays," *Journal of Combinatorial Optimization*, vol. 1, no. 3, pp. 211–228, 1997.

[3] J. M. Rosenberger, E. L. Johnson, and G. L. Nemhauser, "Rerouting aircraft for airline recovery," *Transportation Science*, vol. 37, no. 4, pp. 408–421, 2003.

[4] S. Bratu and C. Barnhart, "Flight operations recovery: new approaches considering passenger recovery," *Journal of Scheduling*, vol. 9, no. 3, pp. 279–298, 2006.

[5] X. W. Tang, Q. Gao, and J. F. Zhu, "Research on greedy simulated annealing algorithm of irregular flight schedule recovery model," *Forecasting*, vol. 29, no. 1, pp. 66–70, 2010.

[6] N. Eggenberg, M. Salani, and M. Bierlaire, "Constraint-specific recovery network for solving airline recovery problems," *Computers & Operations Research*, vol. 37, no. 6, pp. 1014–1026, 2010.

[7] J. D. Petersen, G. Sölveling, J.-P. Clarke, E. L. Johnson, and S. Shebalov, "An optimization approach to airline integrated recovery," *Transportation Science*, vol. 46, no. 4, pp. 482–500, 2012.

[8] M.-L. Le and C.-C. Wu, "Solving airlines disruption by considering aircraft and crew recovery simultaneously," *Journal of Shanghai Jiaotong University (Science)*, vol. 18, no. 2, pp. 243–252, 2013.

[9] F. T. S. Chan, S. H. Chung, J. C. L. Chow, and C. S. Wong, "An optimization approach to integrated aircraft and passenger recovery," in *Proceedings of the Institute of Industrial Engineers Asian Conference*, pp. 729–737, Taipei, Taiwan, July 2013.

[10] K. Sinclair, J.-F. Cordeau, and G. Laporte, "Improvements to a large neighborhood search heuristic for an integrated aircraft and passenger recovery problem," *European Journal of Operational Research*, vol. 233, no. 1, pp. 234–245, 2014.

[11] Y. Hu, B. Xu, J. F. Bard, H. Chi, and M. Gao, "Optimization of multi-fleet aircraft routing considering passenger transiting under airline disruption," *Computers & Industrial Engineering*, vol. 80, pp. 132–144, 2015.

[12] W. Ge, J.-F. Zhu, W.-W. Wu, and X.-H. Wu, "Stochastic optimization for uncapacitated *p*-hub median problems," *System Engineering—Theory & Practice*, vol. 33, no. 10, pp. 2674–2678, 2013.

[13] A. J. Schaefer, E. L. Johnson, A. J. Kleywegt, and G. L. Nemhauser, "Airline crew scheduling under uncertainty," *Transportation Science*, vol. 39, no. 3, pp. 340–348, 2005.

[14] J. W. Yen and J. R. Birge, "A stochastic programming approach to the airline crew scheduling problem," *Transportation Science*, vol. 40, no. 1, pp. 3–14, 2006.

[15] M. Dunbar, G. Froyland, and C.-L. Wu, "Robust airline schedule planning: minimizing propagated delay in an integrated routing and crewing framework," *Transportation Science*, vol. 46, no. 2, pp. 204–216, 2012.

[16] J. M. Rosenberger, A. J. Schaefer, D. Goldsman, E. L. Johnson, A. J. Kleywegt, and G. L. Nemhauser, "A stochastic model of airline operations," *Transportation Science*, vol. 36, no. 4, pp. 357–377, 2002.

[17] D. Mou and W. Zhao, "An irregular flight scheduling model and algorithm under the uncertainty theory," *Journal of Applied Mathematics*, vol. 2013, Article ID 361926, 8 pages, 2013.

[18] P. Arias, D. Guimarans, M. M. Mota, and G. Boosten, "A methodology combining optimization and simulation for real applications of the stochastic aircraft recovery problem," in *Proceedings of the 8th EUROSIM Congress on Modelling and Simulation (EUROSIM '13)*, pp. 265–270, IEEE, Cardiff, Wales, September 2013.

[19] G. B. Dantzig, "Linear programming under uncertainty," *Management Science*, vol. 1, pp. 197–206, 1955.

[20] E. M. L. Beale, "On minimizing a convex function subject to linear inequalities," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 17, no. 2, pp. 173–184, 1955.

[21] J. R. Birge and F. V. Louveaux, *Introduction to Stochastic Programming*, Springer, Berlin, Germany, 2011.

[22] S. Y. Sohn, K. B. Yoon, and I. S. Chang, "Random effects model for the reliability management of modules of a fighter aircraft," *Reliability Engineering & System Safety*, vol. 91, no. 4, pp. 433–437, 2006.

[23] J. R. Birge, "The value of the stochastic solution in stochastic linear programs with fixed recourse," *Mathematical Programming*, vol. 24, no. 1, pp. 314–325, 1982.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 169731, 12 pages
http://dx.doi.org/10.1155/2015/169731

Research Article

Research and Application of FTA and Petri Nets in Fault Diagnosis in the Pantograph-Type Current Collector on CRH EMU Trains

Long-long Song,[1] Tai-yong Wang,[1,2] Xiao-wen Song,[3] Lei Xu,[3] and De-gang Song[3]

[1]*School of Mechanical, Electronic and Control Engineering, Beijing Jiaotong University, Beijing 100044, China*
[2]*School of Mechanical Engineering, Tianjin University, Tianjin 300192, China*
[3]*R&D Center, CSR Qingdao Sifang Co., Ltd., Qingdao 266111, China*

Correspondence should be addressed to Long-long Song; song_bjtu@163.com

Received 28 May 2015; Revised 8 September 2015; Accepted 14 September 2015

Academic Editor: Yuanchang Xie

A fault tree is established based on structural analysis, working principle analysis, and failure mode and effects analysis (FMEA) of the pantograph-type current collector on the Chinese Rail High-Speed Electric Multiple Unit (CRH EMU) train. To avoid the deficiencies of fault tree analysis (FTA), Petri nets modelling is used to address the problem of data explosion and carry out dynamic diagnosis. Relational matrix analysis is used to solve the minimal cut set equation of the fault tree. Based on the established state equation of the Petri nets, initial tokens and enable-transfer algorithms are used to express the fault transfer process mathematically and improve the efficiency of fault diagnosis inferences. Finally, using a practical fault diagnosis example for the pantographs on CRH EMU trains, the proposed method is proved to be reasonable and effective.

1. Introduction

In recent years, high-speed electric multiple unit (EMU) trains have become an important mode of transportation in China. The Chinese Rail High-Speed (CRH) EMU train is a large-scale intelligent system with complex structures. The reliability of an EMU train is directly related to the safety and efficiency of the high-speed, heavy-load rail transportation system [1]. The high-voltage traction drive system is one of the most critical aspects of an EMU train system. As a critical subsystem of the high voltage traction system, the pantograph-type current collector plays an important role in providing power for an EMU train by coupling directly with the electric catenary wire [2, 3]. Efficient and accurate fault diagnosis for the pantograph is vital for the safe and stable operation of CRH EMU trains.

Complexity is defined as a state in which many different parts (hardware, software, organizational, and human elements) are related to each other in an interconnected manner [4]. It is defined here in two forms: structural complexity and fault complexity. The CRH EMU train system is complex both in its structure and in its failure modes. Though the structure of the pantograph seems relatively simple, the working principles and failure modes of pantographs are surprisingly complex. For example, there are many complex potential failure modes for a pantograph, which are related to mechanical structures, electrical facilities, pneumatic transmission, network control, and many other aspects. The operation of the pantograph on a CRH EMU train is automatically controlled by microprocessors. The pantograph system involves mechanical structures such as guide bars, electrical equipment such as the traction converter, and rubber products such as the base insulator. The lifting process is controlled by air pressure. The complex failure modes of the pantograph can involve cracks, breaks, fatigue, pitting, wear, and discharge breakdown. The relationship between different failures is complex and uncertain. Consequently, the pantograph system of a CRH EMU train is treated as a complex system. A useful approach combining fault tree analysis and Petri nets theory is applied to deal with the complexity of fault diagnosis in the pantograph system.

Fault tree analysis (FTA) is a useful analytical tool for identifying and classifying hazards and calculating system reliability for both simple and complex engineering systems [5]. It is a systematic way to assess the reliability of complex systems both qualitatively and quantitatively [6]. Much research has been conducted and outcomes include extended fuzzy FTA (FFTA) methodology in the petrochemical process industry in fuzzy environments [7]; the use of a fuzzy-logic-based reliability approach to evaluate basic events in fault tree analysis for nuclear power plant probabilistic safety assessment [8]; FFTA applied to reducing uncertainty in expert judgment in the safety barriers analysis of an offshore drilling system [9]; and the use of fault tree analysis and analytic hierarchy processes to analyze the risks associated with the use of shield tunnel boring machines (TBMs) [10].

However, the modelling power of FTA is limited to the static evaluation of a single criterion at a time and FTA is not capable of describing dynamic system behavior with redundant components, degraded system states, and repair or test activities [11]. Leveson and Stolzy proposed the safety analysis of dynamic systems using time Petri nets [12]. Bobbio et al. provided an algorithm to convert a parametric fault tree (PFT) into a class of high-level Petri nets to exploit the modelling power and flexibility of the stochastic well-formed nets (SWN) formalism [13]. Robidoux et al. used a framework named dynamic reliability block diagrams (DRBD) for modelling the dynamic reliability behavior of computer-based systems and presented an algorithm that automatically converted a DRBD model into a colored Petri net [14]. With increasing complexity in the relationship between system structures and faults, fault trees become more difficult to establish. Problems relating to the Nondeterministic Polynomial (NP), which can lead to combinatorial explosion as the numbers of calculations grow exponentially, appear when solving the minimal cut sets. Makajic-Nikolic et al. proposed an approach to cutting the number of minimal cut sets in a fault tree using reverse Petri nets [15]. There have also been some studies on the application of FTA and Petri nets in transportation systems. For example, Wang et al. proposed reliability modelling and evaluation issues of electric vehicle motors by using FT FTA-based extended stochastic Petri nets [16]; Nguyen et al. combined a Petri nets (PN) method with extensions of FTA to adapt to dynamic systems and applied the method to a satellite-based railway system [17]; and Song et al. applied T-S Fuzzy theory and FTA to diagnose pantograph faults in a multisource heterogeneous knowledge environment [18].

Some other mathematical/statistical methods such as fuzzy reasoning and Bayesian Network (BN) also have been successfully used in the railways risk and safety domain. For example, An et al. proposed a railway risk management system for railway risk analysis using fuzzy reasoning approach and fuzzy analytical hierarchy decision making process [19, 20]; Noori and Jenab developed a fuzzy Bayesian traction control system for rail vehicles with speed sensors in intelligent transportation systems [21]; Muttram applied a Safety Risk Model combining fault tree analysis and cause/consequence techniques to predict residual levels of railway safety risk [22]; Bouillant et al. developed a decision support tool based on Bayesian Networks to evaluate, compare, and optimize various operating and maintenance strategies of Paris metro underground rails [23]; Xie et al. introduced the Bayesian inference and investigated the application of a Bayesian ordered probit (BOP) model in driver's injury severity analysis and verified that BOP model could produce more reasonable parameter estimations and better prediction performance than the ordered probit (OP) model [24]; Bernardi et al. generated Repairable Fault Tree and Bayesian Network models for railway modelling by a model-driven approach called DAM-Rail approach [25]; Mahboob and Straub compared fault tree and Bayesian Networks for modeling safety critical components in railway systems [26]; and Washington and Oh incorporated Bayesian methodology with expert judgment for countermeasure effectiveness under uncertainty and applied the approach in improving the safety of railroad crossings [27].

From the above, it appears that the conventional analysis approaches, such as FTA, Petri nets (PN), fuzzy reasoning, and BN, have been widely used in the railways risk and safety analysis field. However, each approach has its own advantages and limitations. Fault tree, Petri nets, and BN have a strong similarity in many aspects. Fault tree can be established more easily only when the cause-and-effect relationships between events were clear. But it suffers severe limitations of statics structure and uncertainty handling. Fuzzy reasoning and BN are effective tools for quantitative analysis because they are based on probabilistic and uncertain knowledge. BN approach allows dealing with issues such as prediction or diagnosis, optimization, and data analysis of feedback experience [28]. It is more often used in prediction because of good quantitative analysis ability, but the training process of BN is complex. And the dependent probabilities among events are required in BN approach which may be difficult to obtain in some cases. Petri net (PN) is also a powerful and widely used graphical and mathematical modelling tool for the description of sequence dependent behaviors of dynamic systems. When the failure status of a system evolves from one subsystem/component to other subsystems/components, Petri net modelling is more intuitive and more effective to describe the process by carrying mathematical matrix computations, which is easier to handle by computer. At the same time, it is more suitable for large complex systems because the modelling process is simplified by eliminating repeat basic events and reducing the building elements. Through the above analysis, FTA's translation into Petri nets is more suitable in this paper to analyze the dynamic behaviors of failure statuses of Pantograph-Type Current Collector on CRH EMU trains mathematically as far as the extension, fast modelling and dynamic behavior analysis of the system are mainly concerned.

Studies on reliability and failure mode analysis of CRH EMU trains and their subsystems have recently begun. The focus of the research has turned from design and manufacture to maintenance management. FTA and Petri nets approaches are applicable and useful analysis tools in the risk management of complex engineering systems, but there are some deficiencies in their application. Considerable research has been conducted on these problems. The main aim of

this paper is to extend FTA and Petri nets methodology to maintenance and fault diagnosis in CHR EMU train systems. This section introduces some existing applications of FTA and Petri nets in a range of industries. Structural analysis, working principles, failure mode and effects analysis (FMEA), and FTA modelling of pantographs are outlined in Section 2. Petri nets modelling of the fault tree and the use of relational matrix analysis in solving minimal cut sets are provided in Section 3. In Section 4, the place mark and enable-transfer algorithm of the Petri nets and an actual case study on dynamic transition and diagnosis of pantograph diagnostics are provided. The last section emphasises the highlights of this research.

2. Structural Analysis and Fault Tree Modelling of Pantograph

2.1. Structural Analysis and Working Principle of Pantograph. The pantograph of a CRH EMU train is fitted on the roof of the train and is essential to allow the train to get power from the main overhead wires. A CRH EMU train is an eight-car multiple unit configured with four motorised cars and four trailer cars. There are two power units in all and each power unit consists of two power cars and two trailer cars, arranged in T-M-M-T mode. DSA250 pantographs are fitted on number 4 and number 6 cars. Their maximum lifting height is 3000 mm and the width of the head is 1990 mm. In normal operation, there will be only one pantograph collecting current, with another in a folded state. However, when two CRH EMUs are attached together, two pantographs will work simultaneously. The working principle and structural components of pantograph systems for CRH EMUs are shown in Figure 1. The structural parameters of the pantograph are shown in Table 1.

The pantograph is raised to access high voltage power by allowing the carbon skateboard to contact the catenary wire and descends when its compressed air supply is exhausted. The 25 kv single-phase power alternating current (AC) from the catenary is transferred to the traction transformer from the high-voltage electrical equipment by the pantograph, which outputs 1500 V single-phase AC power to the traction converter. A pulse rectifier converts the single-phase AC into direct current (DC) and outputs 2500–3000 V DC to the traction inverter through the DC circuit. Then the traction inverter outputs three-phase AC power, where the voltage and frequency are all adjustable, to drive the traction motor.

When the pantograph rises, air is compressed into the drive cylinder through the cushion valve and the cylinder piston moves left, overcoming the pressure of the reset spring. Then the lower arm rises and rotates clockwise under the action of the guide bar and the spring. At the same time, the upper arm also rises with the drive of the top guide bar.

When the pantograph descends, the compressed air is removed from the drive cylinder through the cushion valve. Then the piston is pushed to the right with the reset spring releasing pressure. The guide bar also moves right to force the lower arm to rotate anticlockwise, thus forcing the upper arm to descend.

The collector head contacts the catenary wire when the pantograph rises. Current is led to the bottom frame through the collector head, the upper arm, and the pushrod. The power cable installed on the bottom frame then leads the current to the vehicle. Since current will flow through the entire pantograph frame in the power supply state, all the pantograph hinges are equipped with bridge connections to prevent the current damaging the bearings.

The performance of the pantograph largely depends on contact pressure. If there is too little pressure, contact resistance will vary easily, resulting in poor contact and arcing. However, too much pressure will increase the friction, aggravating wear on the carbon skateboard and wires and reducing the life of the carbon skateboard.

2.2. Failure Mode and Effects Analysis (FMEA) of Pantograph System. Pantographs on CRH EMU trains are installed on the roof, exposed to the environment. The working environment is complex, volatile, and sometimes very harsh. There are many and complex ways in which the pantograph can fail, which are related to mechanical structures, electrical facilities, pneumatic transmission, network control, and many other factors. The normal operation of the pantograph on a CRH EMU train is automatically controlled by microprocessors. The pantograph system involves mechanical structures such as guide bars, electrical equipment such as the traction converter, and rubber products such as the base insulator. The lifting process is controlled by air pressure. Because pantographs connect directly with the extra high voltage (EHV) catenary wire, they can be easily damaged by partial high voltage discharges. By classifying and analyzing fault tracking records, the most common pantograph failure modes are summarized as follows.

(i) Failure Mode 1. Pantograph rises to an abnormal position.

Probable cause:

When two EMUs attach together, the space between the two pantographs is less than 190 m.

(ii) Failure Mode 2. Pantograph rises normally, while the monitor (MON) does not show it correctly.

Probable causes:

(a) The pressure sensors are not operating correctly.
(b) The pressure switches have failed.

(iii) Failure Mode 3. Pantograph cannot rise.

Probable causes:

(a) Electricity generation system (EGS) is closed.
(b) Vacuum circuit breaker (VCB) is closed.
(c) Pressure in the auxiliary air cylinder is too low.
(d) [Pantograph·VCB] NFB is in the OFF position.
(e) [Pantograph Rising] NFB is in the OFF position.
(f) Pantograph pressurised air ducts leak.
(g) White air ducts connected to the pantograph leak.
(h) Air ducts in the pantograph control valve board leak.

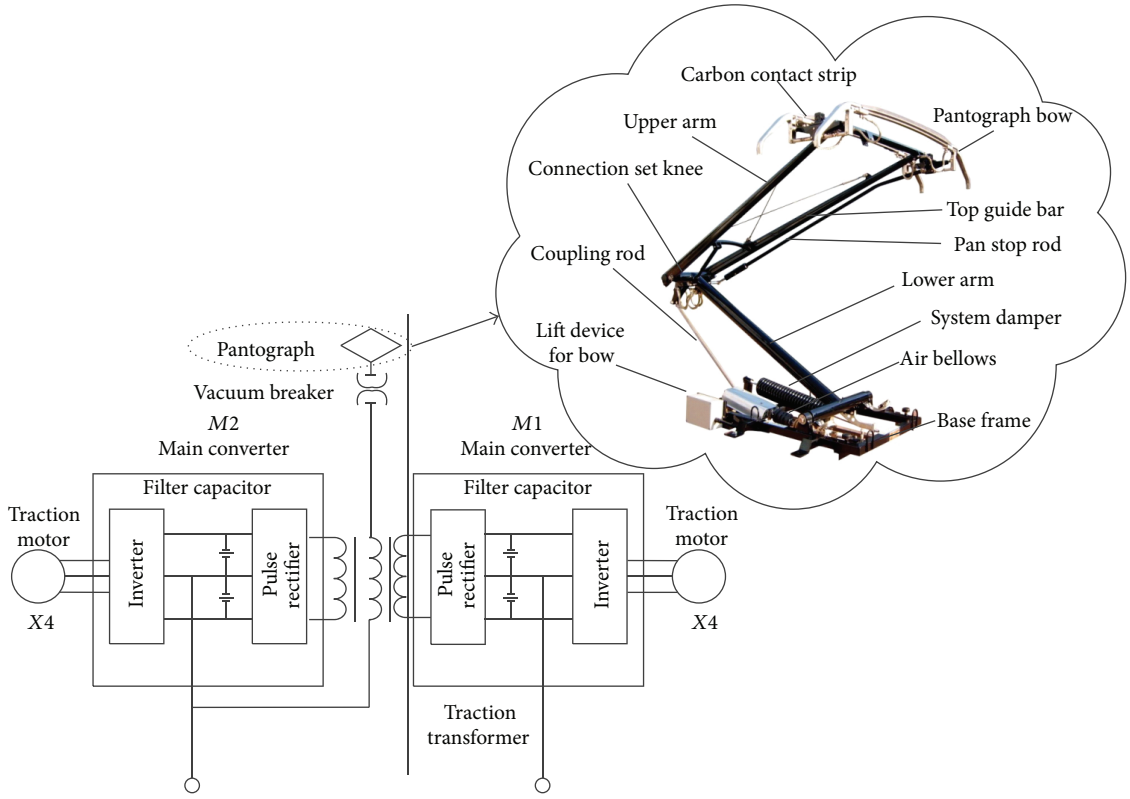


Figure 1: Working principle of traction drive systems and pantograph structure in a CRH EMU train.

Table 1: Structural parameters of the pantograph on a CRH EMU train.

Structural parameters	Values	Structural parameters	Values
Model	DSA250	Vent lifting height	3000 mm (including insulator)
Structure	Single arm pantograph	Maximal action height	2800 mm (including insulator)
Rated voltage	25 kV	Minimal action height	888 mm (including insulator)
Rated current	1000 A	Folded height	593 mm (including insulator)
Weight	115 kg	Contact pressure	70 N ± 5 N
Ambient temperature	−40°C~+60°C	Design life	30 years

(i) Pantograph bellows are damaged.
(j) The pantograph's carbon skateboard leaks.

(iv) Failure Mode 4. Pantograph cannot be lowered.

Probable cause:

Contacts in the pantograph's rising relay have adhered.

(v) Failure Mode 5. Pantograph cannot pass neutral section automatically.

Probable cause:

Terminal RXCB board has failed.

(vi) Failure Mode 6. Supporting insulator is damaged.

Probable causes:

(a) Surface defects created during manufacture.
(b) Insulator being burned by electrical arcing.
(c) Fog flashovers discharged because of high humidity or dust.
(d) Fatigue cracks.
(e) Insulator being struck by foreign objects.

(vii) Failure Mode 7. Pantograph descends automatically.

Probable causes:

(a) Pressure air ducts in the pantograph leak.
(b) White air ducts connected with the pantograph leak.
(c) Air ducts in the pantograph control valve board leak.
(d) Pantograph's bellows are damaged.

(e) The pantograph's carbon skateboard leaks.
(f) Incorrect manipulation.
(g) Pantograph descending loop is instantly energized by electromagnetic interference.
(h) Pressure from the catenary is too low.
(i) There has been discharge between the pantograph bracket and the roof.

(viii) Failure Mode 8. The catenary voltage transformer is damaged.

Probable causes:

(a) Struck by foreign objects.
(b) Burned by electrical arcing.
(c) Quality control issues in components.

2.3. Fault Tree Modelling of the Pantograph. Fault tree analysis (FTA) is one type of FMEA, which is integrated with mechanics, graph theory, optimization theory, and artificial intelligence techniques [29]. It has been widely applied in many fields including aerospace systems, atomic reactors, large-scale equipment, and electronic computer systems. FTA can indicate causal relationships between complex faults and can be useful for logical analysis and diagnosis of complex system faults. It is a systematic way to assess the reliability of complex systems both qualitatively and quantitatively.

FTA is a systematic risk analysis method that deals with the occurrence of an undesired event. The fault event that analysts do not expect to happen is usually the focus of the FTA method. Analysts apply top-down logic to find all direct and indirect fault events relating to the incident. They can then establish logical relationships between the events, form a fault tree, and undertake quantitative or qualitative analysis. In the process above, the focus fault event is the top event. Selection of the top event is crucial to fault tree modelling. If the top event is too general, it is difficult to analyse the fault tree. On the other hand, if the event is too specific, the fault tree will fail to show the causal relationships of the system fully. In general, the failure of the system analyzed will be selected as the top event.

Combined with analysis of the structures and working principles of pantographs on CRH-trains, a pantograph fault tree has been established as shown in Figure 2. "T" represents the top event, "M" represents the intermediate event, and "X" represents the basic event.

Meanings of events in Figure 2 are shown as follows.

T: pantograph fails to work,

$M1$: pantograph rises to an abnormal position,

$M2$: pantograph rises normally, but the MON does not show it correctly,

$M3$: pantograph cannot rise,

$M4$: pantograph cannot be lowered,

$M5$: pantograph cannot pass neutral section automatically,

$M6$: supporting insulators are damaged,

$M7$: pantograph descends automatically,

$M8$: the catenary voltage transformer is damaged,

$M9$: pressure of the auxiliary air cylinder is too low,

$X1$: two EMUs are attached together,

$X2$: spacing of two pantographs is less than 190 m,

$X3$: the pressure sensors do not work properly,

$X4$: the pressure switches have failed,

$X5$: EGS is closed,

$X6$: VCB is closed,

$X7$: in the cab switchboards at both ends, [Pantograph·VCB] NFB is in the OFF position,

$X8$: in the running switchboards of the number 4 and number 6 cars (CRH-200, CRH-300)/number 4 and number 13 cars (CRH long-distance seat car, CRH long-distance sleeper car), [Pantograph Rising] NFB is in the OFF position,

$X9$: pressure air ducts of the pantograph leak,

$X10$: air ducts in the pantograph control valve board leak,

$X11$: pantograph bellows are damaged,

$X12$: pantograph's carbon skateboard leaks,

$X13$: white air ducts connected with the pantograph leak,

$X14$: contacts in the pantograph's rising relay have adhered,

$X15$: terminal RXCB board has failed,

$X16$: surface defects due to manufacturing errors,

$X17$: burned by electrical arcing,

$X18$: fog flashovers discharge,

$X19$: fatigue cracks,

$X20$: struck by foreign objects,

$X21$: driver error,

$X22$: pantograph descending loop is instantly energized by electromagnetic interference,

$X23$: discharge between the pantograph bracket and the roof,

$X24$: pressure of the catenary is too low,

$X25$: burned by electrical arcing,

$X26$: air ducts of the air compressor are damaged,

$X27$: the air compressor fails to work,

$X28$: MR ducts are damaged,

$X29$: main air cylinder leaks.

The fault tree in Figure 2 shows the causal relationships between the pantograph fault events, improving the accuracy and efficiency of fault diagnosis. However, the modelling power of FTA is limited to the static evaluation of a single criterion at a time and cannot describe dynamic system

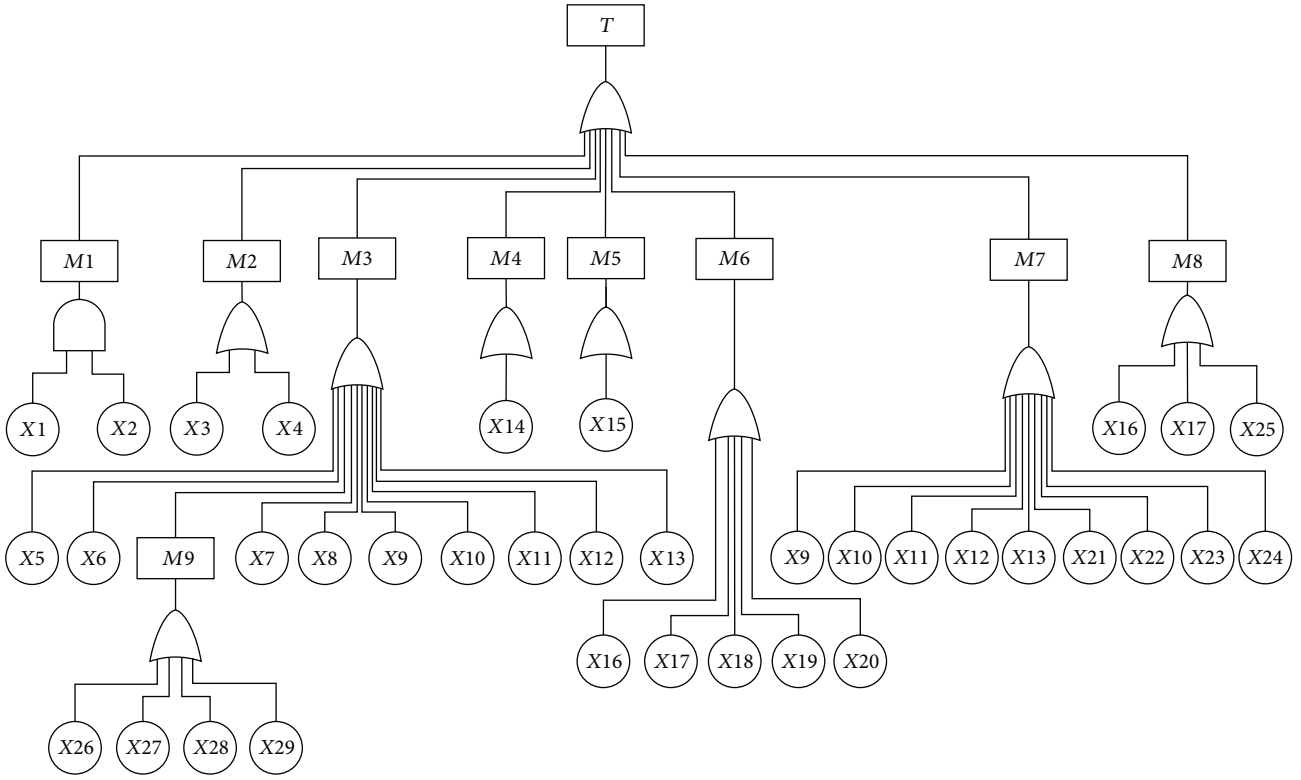


Figure 2: FTA modelling of the pantograph-type current collector.

behavior. It cannot show the transmission and evolution of faults in the system. It also needs too much computation to solve the minimal cut sets of the large-scale fault tree, which may lead to combinatorial explosion because of Nondeterministic Polynomial (NP) problems. For example, if ($i = 1, 2, \ldots, n$) is set to the ith minimal cut set of the fault tree, then the top event can be expressed as

$$T = \bigcup_{i=1}^{n} C_i. \tag{1}$$

The occurrence probability of the top event can be expressed as

$$P(T) = P\left(\bigcup_{i=1}^{n} C_i\right). \tag{2}$$

The logical relationship between the basic events and the minimal cut set is *and*. If the probabilities of the basic events are known, the occurrence probability of the minimal cut set can be expressed as

$$P(C_i) = P\left(\bigcup_{j=1}^{k} x_i\right). \tag{3}$$

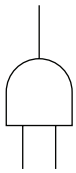
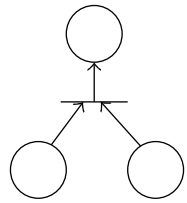
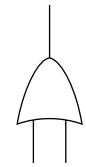
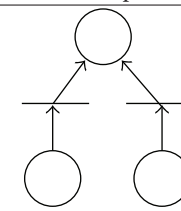
Based on the occurrence probabilities of the minimal cut sets, the occurrence probability of the top event can be expressed as

$$P(T) = \sum_{i=1}^{n} P(C_i) - \sum_{i<j=2}^{n} P\left(C_i C_j\right) + \cdots + (-1)^{m-1} P\left(\bigcup_{i=1}^{n} C_i\right). \tag{4}$$

According to the equation above, the process of determining the probability of the top event consists of 2^{n-1} parts. There are 28 minimal cut sets for the pantograph fault tree in Figure 2, so it needs 2^{27} parts to calculate the probability of the top event, which means 134,217,728 items in all. The combinatorial explosion of the calculation will make quantitative analysis difficult. In addition, the FTA method is generally for static analysis and cannot reflect dynamic changes in multiple states in the system.

The dynamic and structural properties of Petri nets are used to simplify and analytically calculate the pantograph fault tree. Relational matrix analysis is used to solve the minimal cut set equation for the fault tree. Based on the established state equation of Petri nets, initial token and enable-transfer algorithms are used to express the transfer process of faults mathematically.

Table 2: Petri net expression of the FTA logic gates.

AND logic gate		OR logic gate	
Fault tree expression	Petri net expression	Fault tree expression	Petri net expression
			

3. Petri Nets Modelling and Relational Matrix Analysis

3.1. Petri Nets Modelling of the Fault Tree. The Petri net was proposed in 1962 by Petri to express an information flow model for reticular structures [30]. It is a mathematical and graphical analysis tool to express the static structures and dynamic changes in a system [31–33].

In a Petri net, the system state is indicated by "○." A change in state is indicated by "|." An ordered pair is indicated by a directed arc " → " and a token is indicated by "●." The Petri net expressions of the FTA logic gates are shown in Table 2. The transform between Petri net modelling and the fault tree is shown in Figure 3.

One Petri net can be defined in a sextuple $N = (P, T, I, O, M, M_0)$. The six elements should meet the following conditions.

(1) $P = \{P_1, \ldots, P_n\}$ is a finite set of places. $n > 0$ is the number of places.

(2) $T = \{t_1, \ldots, t_m\}$ is a finite set of transactions. $m > 0$ is the number of transactions. $P \cap T = \varnothing$.

(3) $I : P \times T \rightarrow N$ is the input function, defining the set of repetitions or the weight of the directed arcs from set P to set T. $N = \{0, 1, \ldots\}$ is a set of nonnegative integers.

(4) $O : T \times P \rightarrow N$ is the output function, defining the set of repetitions or the weight of the directed arcs from set T to set P. $N = \{0, 1, \ldots\}$ is a set of nonnegative integers.

(5) $M : P \rightarrow N$ is the set of the identification distribution of every place.

(6) $M_0 : P \rightarrow N$ is the set of the initial identification distribution of every place.

As shown in Figures 2 and 3, the top event is replaced by the top place in the Petri net. All the probable events that may lead to the top place are represented as middle places or basic places. The logical gates of the fault tree are denoted by *Transaction* and *Directed arc*. Repeating events in the fault tree no longer exist in Petri net modelling.

Petri net modelling avoids repeating basic events and can achieve a 20% decrease in the number of places. The simplifying effect is more obvious in large-scale systems.

3.2. Relational Matrix Analysis in Solving the Minimal Cut Sets. Besides simplifying the fault tree, a Petri net also effectively overcomes the shortcomings of the traditional FTA in solving the minimal cut sets. The process of Petri net modelling can be undertaken more easily on a computer.

The structure of the Petri net can be translated into a matrix representation. The value of the input function from place P to transaction t is a nonnegative integer ω, recorded as $I(P, t) = \omega$, represented by a directed arc from P to t with the side note. The value of the output function from transaction t to place P is a nonnegative integer ω, recorded as $O(P, t) = \omega$, represented by a directed arc from t to P with the side note ω. The side note is omitted when $\omega = 1$. The directed arc is also omitted when $I(P, t) = 0$ or $O(P, t) = 0$. I and O can both be represented as $n \times m$ nonnegative integer matrixes. The difference between O and I is called the relational matrix, recorded as $A = O - I$.

The steps for solving the minimal cut sets using the relational matrix method are as follows.

Step 1. Find out the row consisting only of "1" and "0" in the relational matrix A. This will be assigned the top place, with only inputs and no outputs.

Step 2. Search for "−1" by column from 1 in the top place. Any row corresponding to "−1" will be regarded as an input to the top place. If there are multiple "−1" values in the column, then there will be multiple inputs for the same transaction. The logical relationship between the inputs is *AND*.

Step 3. Search for "1" by row from "−1" determined in Step 2. Rows including "1" will be regarded as occupying a middle place. Then search repeatedly, following the second step, until a row without "1" is located, which will become a basic place. If there are multiple occurrences of "1" in the row, the logical relationship between the places corresponding to "1" will be *OR*.

Step 4. Continue to search following the second step and the third step until all the bottom basic places are located.

Step 5. Expand the bottom places according to the logical relationships *AND* and *OR*. Then obtain all the cut sets for the system.

Step 6. Find the minimal cut sets according to the Boolean absorption rate or the prime number method.

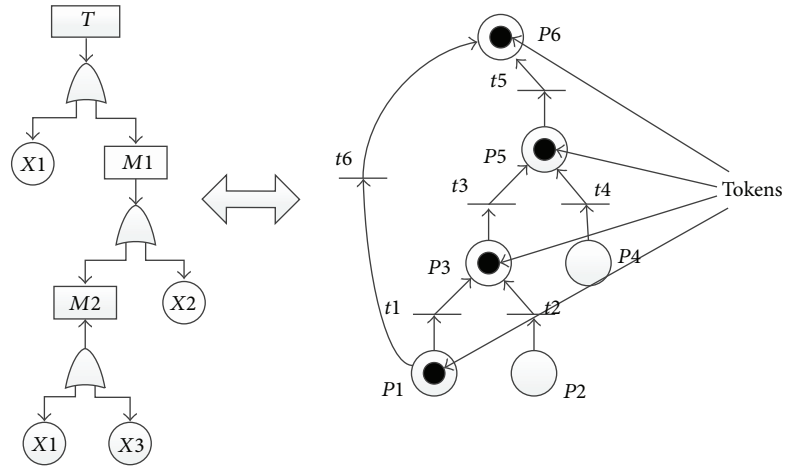


Figure 3: Transform between Petri net modelling and the fault tree.

The fault tree in Figure 3 is used as an example to illustrate solving the minimal cut sets with a relational matrix. Respectively solve the input matrix I, the output matrix O, and the relational matrix A according to the steps above:

$$I = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$O = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix},$$

$$A = O - I = \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & -1 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ 1 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}. \quad (5)$$

3.3. Comparison between Petri Net Modelling and Fault Tree Modelling. The complexity of quantitative analysis depends on the number of nodes and the logical gates in a fault tree. Comparing Figures 3 and 4 to Figure 2, we can see that there are two repeating basic events and seven repeating events in two traditional fault trees. The repetition rates can, respectively, reach 17% and 24%. These repeating events do not exist in the corresponding Petri net modelling. The building of the fault tree needs 6 types of elements which involve basic events, intermediate events, top events, AND logical gate, OR logical gate, and a relation line, while the building of the Petri net model needs only three types of elements involving places, transactions, and directed arcs. Though the fault tree in Figure 2 seems to provide better visualization than the Petri nets in Figure 4, there are more elements and repeating events in the fault tree. The relationships among fault events in the fault tree are more complex. The reduction of building elements and repeating events make the Petri net modelling more convenient and efficient than fault tree modelling.

In addition, Petri nets are graphical and mathematical tools for modelling systems and their dynamics, while fault tree modelling cannot be updated in a timely manner when the system changes. The relationship between fault events is displayed statically in a fault tree. When the failure status of a system evolves from one subsystem/component to other subsystems/components, Petri net modelling is more intuitive and more effective than fault tree modelling because of the use of token transferring and enable-transferring processes. This process in Petri nets can be described by mathematical matrix computations, which is easier to handle by computer (as shown in Section 4).

4. Dynamic Transition and Diagnosis of Pantograph Faults

4.1. Place Mark and Enable-Transfer Algorithm of the Petri Net. A Petri net consists of places "○," transactions "|," directed arcs "→," and tokens of places "●." The places represent logical descriptions of the system states, and the transactions represent the arising of system events. Relational matrixes and state equations are major tools for Petri net analysis:

$$M_{k+1} = M_k + A^T X_k, \quad k = 1, 2, \ldots. \quad (6)$$

In (6), M_k is the initial identification set of the system faults before ignition. M_{k+1} is the result identification set of the system faults after ignition. A^T is the relational matrix and X_K is the transfer sequence for ignition. The

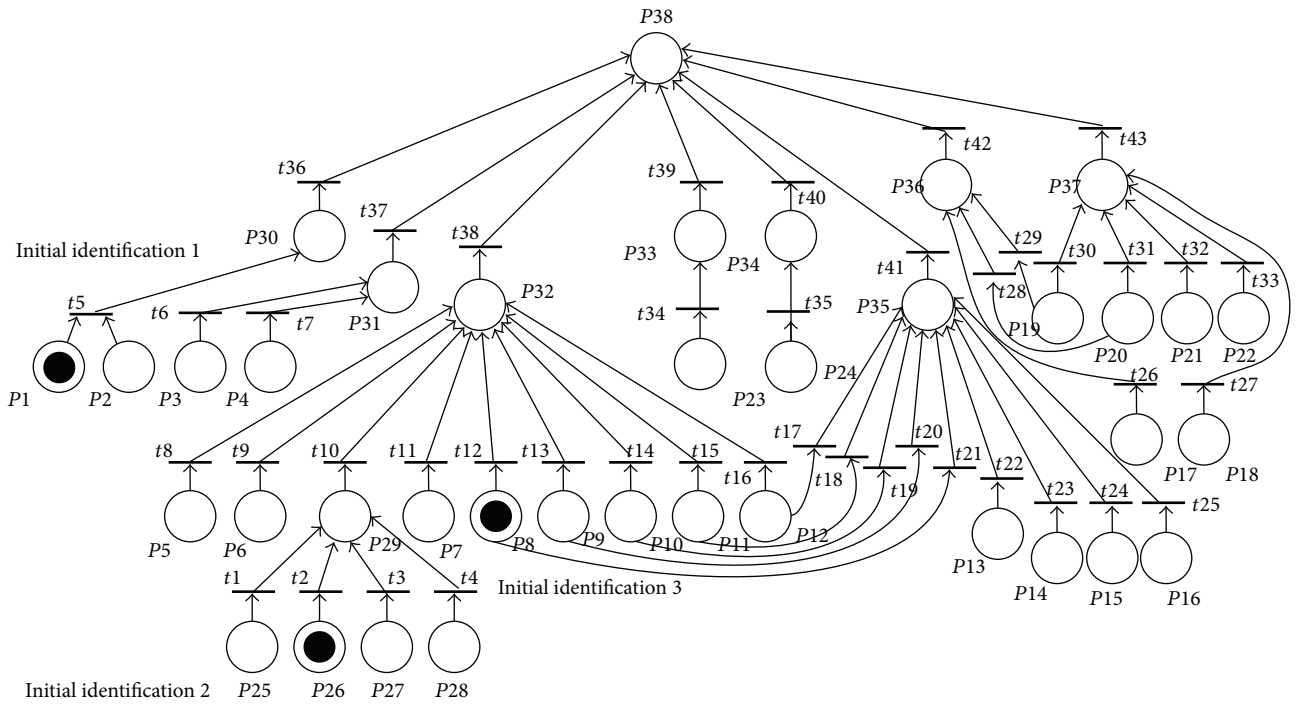


Figure 4: Transfer of tokens in the Petri net of the CRH EMU pantograph.

initial identification M_0 will convert to M_d after the ignition sequence $X_0, \ldots, X_d$. The process can be expressed as

$$M_d = M_0 + A^T \sum_{k=0}^{d-1} X_k. \tag{7}$$

If the element representing the top place is not less than 1 in M_d after a series of transfers, the fault event represented by the top place will occur.

4.2. Dynamic Transition and Diagnosis of Pantograph Faults. In the process of fault diagnosis based on the use of a Petri net, the initial identification of the input place is regarded as the initial symptom of the event occurrence. A token is assigned to the input place if the symptom appears; otherwise the place is assigned a null value. The future states and final identification of the system can be circularly solved through the state equations and transfer sequences, until the token value of the target place is found. The fault occurs when the number of tokens in the target place is not zero.

For the *AND* gate in the Petri net, the tokens cannot transfer downward to the next level if there are empty input places. That is to say, for an *AND* gate, the fault will transfer downward only when all the events of the input places occur. For the *OR* gate in a Petri net, the tokens will transfer downward to the next level place if at least one input place contains tokens. According to the rules of the network theory, the next level place will get two tokens if the two input places both contain tokens. Regardless of the number of tokens, the occurrence of the fault that is represented by the target place is only related to whether there are tokens in the target place. Transfer of the token in the Petri net of the CRH EMU pantograph is shown in Figure 4.

As shown in Figure 4, $P1$ represents "two EMUs attached together," $P2$ represents "spacing of two pantographs of less than 190 m," $P8$ represents "pantograph bellows that are damaged," $P26$ represents "MR ducts that are damaged," $P29$ represents "pressure in the auxiliary air cylinder that is too low," $P30$ represents "pantograph rise to an abnormal position," $P32$ represents "pantograph that cannot rise," $P35$ represents "pantograph that descends automatically," and $P38$ represents "pantograph that fails to work." Three representative fault transfer paths were selected to study the transfer expression of pantograph faults in the Petri net. The initial identification sets of the three paths were, respectively, recorded as M_0^1, M_0^2, and M_0^3:

$$\begin{aligned} M_0^1 &= \left[1, 0, \ldots, 0\right]^T, \\ M_0^2 &= \left[0, \ldots, 0, 1, 0, \ldots, 0\right]^T, \\ M_0^3 &= \left[0, \ldots, 0, 1, 0, \ldots, 0\right]^T. \end{aligned} \tag{8}$$

The ignition sequences of the three paths were, respectively, recorded as X_0^1, X_0^2, and X_0^3:

$$\begin{aligned} X_0^1 &= \left[0, \ldots, 0, 1, 0, \ldots, 0\right]^T, \\ X_0^2 &= \left[0, 1, 0, \ldots, 0\right]^T, \\ X_0^3 &= \left[0, \ldots, 0, 1, 0, \ldots, 0, 1, 0, \ldots, 0\right]^T. \end{aligned} \tag{9}$$

The input matrix I, output matrix O, and the relational matrix A^T of the Petri net are all 38 × 43 matrixes which can be solved using the relational-matrix method:

$$A^T = O - I = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}, \tag{10}$$

$$a_{ij} = O_{ij} - I_{ij} \quad 1 \le i \le 38, \ 1 \le j \le 43. \tag{11}$$

In (10), $m = 38$, $n = 43$:

$$\begin{aligned} M_1^1 &= 0, \\ M_2^1 &= 0. \end{aligned} \tag{12}$$

As (12) shows, tokens cannot transfer downward to the next level in the first path. Thus there is no token in the top place at the end of the procedure, meaning that the fault event represented by $P38$ does not occur.

In the second path, M_1^2, M_2^2, and M_3^2 are as follows:

$$\begin{aligned} M_1^2 &= \left[\underbrace{0, \ldots, 0}_{1 \le i \le 28}, 1, \underbrace{0, \ldots, 0}_{30 \le i \le 38} \right], \\ M_2^2 &= \left[\underbrace{0, \ldots, 0}_{1 \le i \le 31}, 1, \underbrace{0, \ldots, 0}_{33 \le i \le 38} \right], \\ M_3^2 &= \left[\underbrace{0, \ldots, 0}_{1 \le i \le 37}, 1 \right]. \end{aligned} \tag{13}$$

As (13) shows, the tokens in place $P26$ are transferred into place $P29$ after the first ignition and then transferred into $P32$ after the second ignition. Finally, the tokens are transferred into the top place $P38$ after the third ignition, meaning that the top event occurs in this situation.

In the third situation, M_1^3 can be calculated through a single transfer calculation when the default value of the tokens in place $P8$ is 1:

$$M_1^3 = \left[\underbrace{0, \ldots, 0}_{1 \le i \le 7}, -1, \underbrace{0, \ldots, 0}_{9 \le i \le 31}, 1, 0, 0, 1, \underbrace{0, \ldots, 0}_{36 \le i \le 38} \right]. \tag{14}$$

Thus, when there are two transfer paths for one token in the bottom basic place, there will be negative values in the next level state matrix after ignition if the default value of the tokens in the bottom basic place is still 1. There is no practical significance to this phenomenon. To avoid negative values, change the default value of the tokens in place $P8$ from 1 to 2 to produce $M_1^{3\prime}$:

$$M_1^{3\prime} = \left[\underbrace{0, \ldots, 0}_{1 \le i \le 31}, 1, 0, 0, 1, \underbrace{0, \ldots, 0}_{36 \le i \le 38} \right]. \tag{15}$$

Then M_2^3 can be solved according to $M_1^{3\prime}$:

$$M_2^3 = \left[\underbrace{0, \ldots, 0}_{1 \le i \le 37}, 2 \right]. \tag{16}$$

According to M_2^3, there will be two tokens in the top place $P28$ and the top event will inevitably happen when tokens indicate the fault transfer. Thus the fault event in the top place will happen if there is at least one token in the top place. In the Petri net, if an input place is a bottom basic place containing tokens with a value of n (n is a positive integer greater than 1), directed arcs out, then the default value of the tokens in the input place is n. The fault event presented indicated by the top place would will inevitably happen when there are tokens in the top place.

Failure Mode 1. When two EMUs are attached together and at the same time the spacing of the two pantographs is less than 190 m, the pantograph will rise to an abnormal position. However, if only one condition of the two is satisfied, the pantograph will work normally.

Failure Mode 2. Damage to MR ducts causes low pressure in the auxiliary air cylinder, which means the pantograph cannot rise successfully.

Failure Mode 3. Damage to the pantograph bellows may cause the pantograph to descend automatically or possibly fail to rise. Both of the two conditions will lead to a fault in the pantograph.

The token transfer in the Petri net is consistent with the logic analysis of the faults. It can clearly and effectively describe the dynamic processes of the system faults transfer, achieving fast and efficient fault diagnosis.

5. Conclusions

In this study, the working principles and failure modes of pantographs on CRH EMU trains were analyzed systematically. Petri net modelling was used as a graphical modelling tool to simplify the logical relationships and events of the fault tree into a network with places and transactions as nodes. The events, logical gates, and logical relation lines of FT were transformed to places, transitions, and directed arcs, respectively. The complexity of modelling was reduced by 20% by avoiding repeating events. Thus, the process of solving the minimal cut sets was simplified, which effectively saved calculation time for the minimal cut set solutions for complex large-scale fault trees.

The changes in system states and the evolution of failures are described well by the dynamic properties of Petri net modelling. A mathematical model for the Petri net of the pantograph fault tree was established. The equivalence and correctness of the token-transfer description for fault diagnosis inference in a Petri net were verified. Three different fault paths were used to explain and verify the algorithm. The initial identification sets of the three paths were marked.

The three corresponding ignition sequences were calculated using matrix transformations. Finally, the system status was assessed correctly using a mathematical method which can be handled easily in a computerised system.

This work analyzes the evolution of failure events of pantograph system. The process of other critical systems needs to be further examined. Future work will collect and analyse test data from other systems in CRH EMU trains to extend the methodology to cover the whole maintenance process for CRH EMU train systems.

Conflict of Interests

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

Acknowledgments

This research was supported by the National Natural Science Foundation of China (no. 51475324). The authors would like to thank the reviewers for their constructive comments, which enhanced the quality of this paper.

References

[1] F. C. Jiao and B. J. Wang, *The Management and Maintenance of Locomotive Rolling Stock*, Beijing Jiaotong University Press, Beijing, China, 2013.

[2] L. L. Song, T. Y. Wang, X. W. Song, L. Xu, and D. Song, "Fault diagnosis of pantograph type current collector of CRH electric multiple units based on Petri net modelling and fault tree analysis," *Chinese Journal of Scientific Instrument*, vol. 35, no. 9, pp. 1990–1997, 2014.

[3] I. Aydin, M. Karakose, and E. Akin, "Anomaly detection using a modified kernel-based tracking in the pantograph—catenary system," *Expert Systems with Applications*, vol. 42, no. 2, pp. 938–948, 2015.

[4] E. Zio, "Reliability engineering: old problems and new challenges," *Reliability Engineering and System Safety*, vol. 94, no. 2, pp. 125–141, 2008.

[5] D. M. Shalev and J. Tiran, "Condition-based fault tree analysis (CBFTA): a new method for improved fault tree analysis (FTA), reliability and safety calculations," *Reliability Engineering and System Safety*, vol. 92, no. 9, pp. 1231–1241, 2007.

[6] Y. E. Senol, Y. V. Aydogdu, B. Sahin, and I. Kilic, "Fault tree analysis of chemical cargo contamination by using fuzzy approach," *Expert Systems with Applications*, vol. 42, no. 12, pp. 5232–5244, 2015.

[7] S. M. Lavasani, A. Zendegani, and M. Celik, "An extension to fuzzy fault tree analysis (FFTA) application in petrochemical process industry," *Process Safety and Environmental Protection*, vol. 93, pp. 75–88, 2014.

[8] J. H. Purba, "A fuzzy-based reliability approach to evaluate basic events of fault tree analysis for nuclear power plant probabilistic safety assessment," *Annals of Nuclear Energy*, vol. 70, pp. 21–29, 2014.

[9] N. Ramzali, M. R. M. Lavasani, and J. Ghodousi, "Safety barriers analysis of offshore drilling system by employing Fuzzy Event Tree Analysis," *Safety Science*, vol. 78, pp. 49–59, 2015.

[10] K.-C. Hyun, S. Min, H. Choi, J. Park, and I. Lee, "Risk analysis using fault-tree analysis (FTA) and analytic hierarchy process (AHP) applicable to shield TBM tunnels," *Tunnelling and Underground Space Technology*, vol. 49, pp. 121–129, 2015.

[11] S. Verlinden, G. Deconinck, and B. Coupé, "Hybrid reliability model for nuclear reactor safety system," *Reliability Engineering and System Safety*, vol. 101, pp. 35–47, 2012.

[12] N. G. Leveson and J. L. Stolzy, "Safety analysis using Petri nets," *IEEE Transactions on Software Engineering*, vol. 13, no. 3, pp. 386–397, 1987.

[13] A. Bobbio, G. Franceschinis, R. Gaeta, and L. Portinale, "Parametric fault tree for the dependability analysis of redundant systems and its high-level Petri net semantics," *IEEE Transactions on Software Engineering*, vol. 29, no. 3, pp. 270–287, 2003.

[14] R. Robidoux, H. Xu, L. Xing, and M. Zhou, "Automated modeling of dynamic reliability block diagrams using colored Petri nets," *IEEE Transactions on Systems, Man, and Cybernetics. Part A:Systems and Humans*, vol. 40, no. 2, pp. 337–351, 2010.

[15] D. Makajic-Nikolic, M. Vujosevic, and N. Nikolic, "Minimal cut sets of a coherent fault tree generation using reverse Petri nets," *Optimization*, vol. 62, no. 8, pp. 1069–1087, 2013.

[16] B. Wang, G. Tian, Y. Liang, and T. Qiang, "Reliability modeling and evaluation of electric vehicle motor by using fault tree and extended stochastic petri nets," *Journal of Applied Mathematics*, vol. 2014, Article ID 638013, 9 pages, 2014.

[17] T. P. K. Nguyen, J. Beugin, and J. Marais, "Method for evaluating an extended Fault Tree to analyse the dependability of complex systems: application to a satellite-based railway system," *Reliability Engineering and System Safety*, vol. 133, pp. 300–313, 2015.

[18] L. L. Song, T. Y. Wang, X. W. Song et al., "Fuzzy intelligent fault diagnosis of pantograph type current collector under the multi-source heterogeneous knowledge environment," *Chinese Journal of Scientific Instrument*, vol. 36, pp. 1283–1290, 2015.

[19] M. An, Y. Chen, and C. J. Baker, "A fuzzy reasoning and fuzzy-analytical hierarchy process based approach to the process of railway risk information: a railway risk management system," *Information Sciences*, vol. 181, no. 18, pp. 3946–3966, 2011.

[20] M. An, S. Huang, and C. J. Baker, "Railway risk assessment—the fuzzy reasoning approach and fuzzy analytic hierarchy process approaches: a case study of shunting at Waterloo depot," *Proceedings of the Institution of Mechanical Engineers F: Journal of Rail and Rapid Transit*, vol. 221, no. 3, pp. 365–383, 2007.

[21] K. Noori and K. Jenab, "Fuzzy reliability-based traction control model for intelligent transportation systems," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 43, no. 1, pp. 229–234, 2013.

[22] R. I. Muttram, "Railway safety's safety risk model," *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, vol. 216, no. 2, pp. 71–79, 2002.

[23] L. Bouillaut, O. Francois, and S. Dubois, "A Bayesian network to evaluate underground rails maintenance strategies in an automation context," *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, vol. 227, no. 4, pp. 411–424, 2013.

[24] Y. Xie, Y. Zhang, and F. Liang, "Crash injury severity analysis using Bayesian ordered probit models," *Journal of Transportation Engineering*, vol. 135, no. 1, pp. 18–25, 2009.

[25] S. Bernardi, F. Flammini, S. Marrone et al., "Enabling the usage of UML in the verification of railway systems: the DAM-rail approach," *Reliability Engineering and System Safety*, vol. 120, pp. 112–126, 2013.

[26] Q. Mahboob and D. Straub, "Comparison of fault tree and Bayesian Networks for modeling safety critical components in railway systems," in *Advances in Safety, Reliability and Risk Management: Proceedings of the European Safety and Reliability Conference (ESREL 2011)*, chapter 12, pp. 89–95, CRC Press, 2011.

[27] S. Washington and J. Oh, "Bayesian methodology incorporating expert judgment for ranking countermeasure effectiveness under uncertainty: example applied to at grade railroad crossings in Korea," *Accident Analysis and Prevention*, vol. 38, no. 2, pp. 234–247, 2006.

[28] P. Weber, G. Medina-Oliva, C. Simon, and B. Iung, "Overview on Bayesian networks applications for dependability, risk analysis and maintenance areas," *Engineering Applications of Artificial Intelligence*, vol. 25, no. 4, pp. 671–682, 2012.

[29] N. Khakzad, F. Khan, and P. Amyotte, "Safety analysis in process facilities: comparison of fault tree and Bayesian network approaches," *Reliability Engineering and System Safety*, vol. 96, no. 8, pp. 925–932, 2011.

[30] M. Gao, M. Zhou, X. Huang, and Z. Wu, "Fuzzy reasoning petri nets," *IEEE Transactions on Systems, Man, and Cybernetics A: Systems and Humans*, vol. 33, no. 3, pp. 314–324, 2003.

[31] G. J. Tsinarakis, N. C. Tsourveloudis, and K. P. Valavanis, "Modeling, analysis, synthesis, and performance evaluation of multioperational production systems with hybrid timed Petri Nets," *IEEE Transactions on Automation Science and Engineering*, vol. 3, no. 1, pp. 29–46, 2006.

[32] M. P. Cabasino, A. Giua, S. Lafortune, and C. Seatzu, "A new approach for diagnosability analysis of Petri nets using verifier nets," *IEEE Transactions on Automatic Control*, vol. 57, no. 12, pp. 3104–3117, 2012.

[33] Y.-S. Huang, Y.-S. Weng, and M.-C. Zhou, "Design of traffic safety control systems for emergency vehicle preemption using timed petri nets," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 4, pp. 2113–2120, 2015.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 796171, 12 pages
http://dx.doi.org/10.1155/2015/796171

Research Article

Developing an Enhanced Short-Range Railroad Track Condition Prediction Model for Optimal Maintenance Scheduling

Peng Xu,[1] Chuanjun Jia,[1] Ye Li,[2] Quanxin Sun,[1] and Rengkui Liu[1]

[1] *MOE Key Laboratory for Urban Transportation Complex Systems Theory and Technology, Beijing Jiaotong University, No. 3 Shang Yuan Cun, Haidian District, Beijing 100044, China*
[2] *Railroad Maintenance-of-Way Department of Nanchang, 125 Er Shi Qi South Road, Nanchang, Jiangxi 330002, China*

Correspondence should be addressed to Peng Xu; xu.peng.bjtu@gmail.com

Received 5 April 2015; Revised 8 September 2015; Accepted 9 September 2015

Academic Editor: Yuanchang Xie

As railroad infrastructure becomes older and older and rail transportation is developing towards higher speed and heavier axle, the risk to safe rail transport and the expenses for railroad maintenance are increasing. The railroad infrastructure deterioration (prediction) model is vital to reducing the risk and the expenses. A short-range track condition prediction method was developed in our previous research on railroad track deterioration analysis. It is intended to provide track maintenance managers with two or three months of track condition in advance to schedule track maintenance activities more smartly. Recent comparison analyses on track geometrical exceptions calculated from track condition measured with track geometry cars and those predicted by the method showed that the method fails to provide reliable condition for some analysis sections. This paper presented the enhancement to the method. One year of track geometry data for the Jiulong-Beijing railroad from track geometry cars was used to conduct error analyses and comparison analyses. Analysis results imply that the enhanced model is robust to make reliable predictions. Our in-process work on applying those predicted conditions for optimal track maintenance scheduling is discussed in brief as well.

1. Introduction

Transportation systems play a critical role in development of society and economy. Railway system accounted for the largest part of national freight ton-miles, for example, 38.2% in 2005 in USA [1] and 49.70% in 2005 in China [2]. Railroad infrastructure as a base element of railway system has great and direct influences on safety and cost efficiency of rail transport. It is believed that as railroad infrastructure becomes older and older, the risk to safe transport and the expenses for preserving the infrastructure will increase. Specifically, when the infrastructure grows up over a certain age, the risk and the expenses will increase exponentially [3]. The last ten years (as a small portion of the entire infrastructure evolution process) has seen linearly increasing expenses per mile for class I railroad infrastructure of USA, as illustrated by Figure 1. Furthermore, the recent development of rail transportation towards higher speed and heavier axle load is also believed to increase the risk and expenses.

Practices in transportation infrastructure management try to balance the cost associated with potential damage resulting from unfavorable infrastructures, as well as the cost for Maintenance and Renewal (M&R) activities in order to minimize the total cost. Management practices of highway pavement and some other infrastructures have been implemented into some tools [4]. But such tools for railroad infrastructure management are rare. Among issues in achieving the balance between the two categories of costs, railroad infrastructure deterioration modeling is vital [5]. The infrastructure deterioration models fall into two categories: long-range and short-range deterioration (or prediction) models. The long-range models assist infrastructure management departments in making budget plan to minimize the planning horizon cost under constraints. The short-range models are

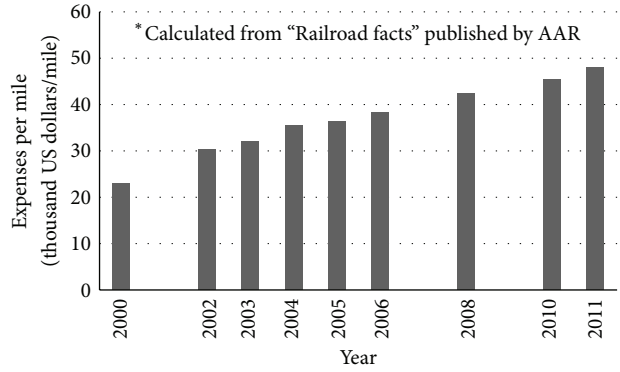


Figure 1: Railroad infrastructure expenses per mile for the last ten years.

necessary to optimally schedule R&M activities, constrained by limited budgets and other maintenance resources allocated through long-range models, and acceptable infrastructure, so as to minimize effects of the activities on rail traffic. According to the R&M scheduling, resources are allocated to each R&M activity within the planning horizon, and, accordingly, the balance between the cost associated with potential damage and the cost for R&M activities is achieved. The categorization of deterioration models applies to infrastructure of all transport modes. If infrastructure condition predictions by deterioration models are not accurate enough, plans will be in question, and sometimes damage might be caused by some of the unpredicted infrastructure failures. Therefore, both long-range and short-range prediction models should be characterized by high accuracy and extensive suitability for infrastructure in various conditions.

A large number of models have been developed for highway pavement deterioration. For instance, Markov decision process was employed to formulate pavement deterioration [6–10]. Kobayashi et al. modeled the pavement deterioration process through a hidden Markov model [11]. Using a time series method, Durango-Cohen formed a model for the pavement deterioration process [12]. Several investigations have been performed on mathematical modeling for railroad infrastructure deterioration. Kawaguchi et al. applied a double exponential smoothing method to the track geometry evolution process and formed a track deterioration model [13]. Based on a track degradation database, Alfelor et al. established a one-to-one linear relationship between track deterioration and one specified contributing factor through least square method [14]. Based on a common denominator regarding track deterioration, Veit and Marschnig used an exponential model to describe track deterioration process [15]. Meier-Hirmer et al. fitted gamma stochastic process to the evolution rate of track surface over a 1000-meter-long section of track [16]. He et al. considered that track deterioration rate increases linearly with the current track condition and exponentially with the sum of individual effects of five contributing factors and formulated their track deterioration model [17]. Liu et al. and Xu et al. thought track deterioration processes between two adjacent maintenance activities may be any forms of smoothly nonlinearly nondecreasing functions and therefore employed piecewise linear regression models to describe track deterioration [18, 19].

From the above literature review, it is seen that some investigations consider infrastructure deterioration stochastic, while others assume that relationships between deterioration and impact factors are characterized by deterministic formulations, for example, linear, polynomial, exponential, and piecewise linear. From the perspective of approximating a curve with straight line segments, the piecewise linear relationship can be used to approximate any smoothly nonlinear relationships. In other words, if properly modeled, the piecewise linear relationship can be more applicable than linear, polynomial, and exponential ones.

In previous research on track (geometry) condition prediction technologies, the authors proposed a method employing the piecewise linear regression to make short-range predictions for condition indices [18, 19]. The proposed method uses track inspection data within time periods of equal length to estimate deterioration rates for all unit sections of a track. Specifically, the lengths of time periods are considered equal in both temporal and spatial dimensions. The estimation of deterioration rates is triggered by availability of new track inspection data. Detail about the estimation will be described in Section 3. The deterioration rate of a unit section for a time period is then used to predict track condition of the unit section for future two or three months. From the brief introduction, it is easily seen that the length of time periods is an extremely key parameter for the method. Hereafter, the time period is referred to as the time span. In the previous research, the length of the time span is determined mainly according to railroad field engineers' knowledge. Recent data analyses discovered that in some cases errors in track condition predictions are not normally distributed around 0 mm. Further rigorous analyses on these cases revealed that the undesirable error distributions may arise from the inappropriate length of the time span.

This paper attempts to formulate an optimization model to estimate the time span length for each unit section. The estimated time spans are varying along a track. The previous model takes a constant time span length for an entire track. This is the difference between the enhanced model and the previous one. Then the estimated time span and a normally distributed random variable are incorporated into the enhancement of the previously proposed prediction method. The enhanced prediction model allows maintenance-of-way departments to acquire accurate track condition two or three months in advance, depending on railroad's transportation focuses, that is, freight and passenger, million gross tons, and traffic speeds.

The reminder of the content is organized as follows. The effects of impact factors on track condition deterioration are descriptively analyzed in Section 2 in order to form a basis for the track condition prediction model. Section 3 presents the enhanced track condition prediction model based on characteristics of track condition deterioration under the impact factors. Section 4 presents error analysis results for track condition predictions. Using the track condition predictions to optimally schedule track maintenance is briefly

discussed in Section 5 for future research. Finally, conclusions regarding the research in this paper are drawn in Section 6.

2. Descriptive Analysis on the Effects of Influential Factors

Railroad track geometry deviations are usually termed track irregularities. Generally speaking, track irregularities are the result of cumulative comprehensive effects of seven categories of impact factors [3, 4, 20–23]: (1) wheel loads on rails, (2) track configuration, (3) materials and manufacture of track components, (4) track design and construction, (5) track maintenance, (6) environmental factors, and (7) terrain. The cumulative wheel load is by far the principal cause for track deterioration and is characterized mainly by axle loads, traffic speed, traffic density, track condition, and train condition. The track configuration plays a critical role in resisting track deterioration. Track deterioration usually begins with small imperfections in the materials and errors in the manufacture of rails and other track components. Good performances of the materials and efficacy of the manufactured components are crucial for preserving tracks in satisfactory condition. Influences of errors during the design and construction of the tracks add to the influences of the materials and the manufacture, as the initial track irregularities with which track deterioration begins. The maintenance is intended to restore tracks to good (or their original) condition. But, during a maintenance activity, survey errors, measurement errors, and maintenance machine tolerances are unable to achieve the desire. Moreover, different kinds of maintenance machines usually have different effectiveness. During track deterioration, in addition to the wheel loads, environmental factors directly deteriorate the tracks as well, for instance, track buckles as a result of extremely high temperature. As the base of railroad track, terrain has obviously direct and considerable influences on track geometry. Any variations in terrain will be reflected immediately by sudden changes in track geometry, for example, Taiwan high-speed rail subsidence [24]. What is more, the vertical stiffness of terrain along a track is varying, resulting in longitudinally varying track deterioration processes.

Under the effects of these seven categories of impact factors, track deterioration processes fall into three groups, gradual deterioration, sudden deterioration (more precisely, damage), and improvement in track condition, as shown in Figure 2. As discussed above, the gradual deterioration process is the result of the effects of moderate environmental factors and the first four categories of impact factors, as demonstrated by solid dark curves in Figure 2, and the improvement process is caused by track maintenance, as demonstrated by the dashed gray lines in Figure 2. Happenings of the other factors including variations in terrain and extreme environmental factors make track deteriorate suddenly, which is not presented in the figure.

In the process of track deterioration, some categories of the impact factors influence each other. For instance, wheel loads deteriorate tracks in terms of track geometry, condition of track components, and performances of materials of the track components. Simultaneously, the deteriorated tracks increase the wheel loads and decease track resistance to deterioration. Such interactive influences among these three categories of the impact factors continue as trains run over the tracks. Sadeghi and Askarinejad have quantitatively investigated influences of some impact factors on track deterioration [23]. Therefore, the track deterioration rate usually shows an accelerated trend. It is the basis for researchers to formulate track deterioration models.

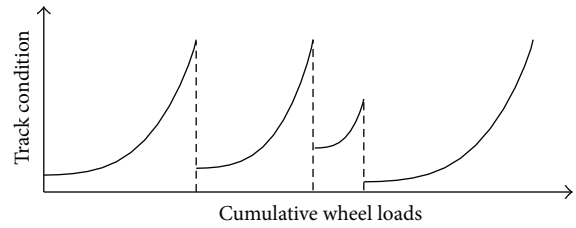


Figure 2: Two groups of track deterioration process.

Track deterioration proceeds under the influences of all these seven interactive categories of impact factors. Actually, the influential degree on track deterioration varies along tracks. In other words, each track location has its own distinctive track deterioration process. The distinctive characteristic of track deterioration has been experienced by railroad field engineers during the past several decades of track management practices. Values of geometrical parameters for track positions vary considerably along a track section, even when the whole track section is maintained during one maintenance activity with the same maintenance machine. By now, only few of the impact factors can be measured, and interactive effects among the impact factors are unable to be measured. What is more, even data for those most possibly accessible influential parameters, that is, million gross tons and train travelling speeds, is often unavailable to investigators because of confidential regulations of operation divisions. Moreover, measurement data of these impact factors that can be measured is often contaminated with slight noises [16, 25–27]. Because of the unavailability of data for most impact factors and uncertainty in the measurement data, track deterioration is usually considered stochastic and is formulated with the independent variable of time.

3. Enhanced Track Condition Prediction Model

Because of the distinctive characteristic of track deterioration, the proposed track condition prediction model is formulated for each analysis object. The spatial and temporal dimensions of a practical analysis object are determined first in light of facts of track geometry measurement data by track geometry cars. Hereafter, the analysis object is often named analysis section for compatibility with railroad industry. For a specified analysis section, track condition prediction model is proposed to predict its condition two or three months in advance. Last, the optimal time span estimation model is formulated for the model.

3.1. Determination of the Analysis Section. Track geometry measurement data from track geometry cars are usually used to analyze track deterioration [5]. Track geometry cars at the speed of train travelling measure various track geometry parameters, positioning parameters (i.e., milepoint), and comfort related parameters at a constant sampling distance. Two sampling distances, 0.25 m and 0.3048 m (1 foot), are used worldwide by current measurement systems of track geometry cars. It is, thus, impossible to acquire historical measurement data for most track points in concerned tracks. Secondly, there are errors in milepoint measurements [27]. These errors make it impossible to acquire historical geometry measurements for even those sampled track positions. Lastly, geometry measurements are usually contaminated with slight noises. This fact makes it difficult to mine slightly contaminated historical geometry measurements for accurate deterioration processes of track points. Therefore, the analysis object cannot be track points but should be a track section. Because of the distinctive track deterioration processes, the analysis track section should not be too long. To determine the length of the analysis track section, the above three factors have to be taken into consideration.

Track geometry cars throughout China Railways today have only one kind of sampling distance, that is, 0.25 m. Due to many reasons [26–28], milepoint measurements from track geometry cars are inconsistent with actual milepoints in field. To deal with errors in milepoint measurements, two mathematical models have been developed by the authors, Key Equipment Identification (KEI) [28] and Dynamic Sampling Point Matching (DSPM) [26, 27]. KEI is intended to automatically identify sampled points associated with key locations of some of track equipment (horizontal curves and diverging tracks of turnouts) in a specified inspection data file and then to revise milepoints of the data file according to actual milepoints of those identified points. After being processed by KEI, milepoint errors are reduced considerably, but two processed inspection data files over same track still have small differences in milepoints of almost same sampling points. Such kind of milepoint differences is referred to as milepoint shift in the references. To reduce milepoint shifts, DSPM was developed. DSPM is formulated to automatically match sampling points on one inspection with the closest sampling points on another inspection. After being processed by both KEI and DSPM, milepoint shifts, in most cases, are reduced to the level below one standard sampling distance, that is, 0.25 m. Readers for details about the above-mentioned data processing models are referred to [26–28].

As for slight noises in geometry measurements, based on inspection data processed by KEI and DSPM, an optimization model was formulated to minimize the effects of the slightly noisy geometry measurements. The noisy effects are quantified through the deviation of the geometrical parameter values predicted by the proposed model from those measured ones. For each inspection, the deviations associated with each analysis section length were calculated. The optimal model uses the sum of the squared deviations as the object function to choose the optimal analysis length that produces the best fit (i.e., minimizing the objective function). The minimization of the object function was accomplished through comparing the sums linked to different lengths. Two years of inspection data from the Jiulong-Beijing Railroad administered by Jinan bureau of China Railways was used to accomplish the comparison analysis. The length of 0.5 m is attained as the best analysis track section.

3.2. Formulation of the Prediction Model

3.2.1. Method for Predicting Track Condition. As noted in Section 2, track condition is the result of the cumulative comprehensive influences of all impact factors. Track condition evolution for an analysis section is mathematically expressed as the equation $c(t) = c(t_i) + \int_{t_i}^{t} f(x)dx$, where $c(t)$ and $c(t_i)$ represent track condition at the time points t and t_i (wherein the subscript i denotes the ith inspection of track geometry cars after a maintenance work covering the analysis section), respectively, and $f(t)$ denotes the deterioration rate of track condition at the time point t. The deterioration rate function $f(t)$ quantifies the comprehensive effects of all impact factors at a point in time. The cumulative influences over the time range from t_i to t are modeled through the integral of the deterioration rate function $f(t)$ over the time period, that is, $\int_{t_i}^{t} f(x)dx$. The equation shows that accurate estimation for the deterioration rate function is crucial to accurately predict track condition.

3.2.2. Track Condition Prediction Model. Occurrences of sudden deteriorations are unpredictable in advance mainly because of unavailability of data for those impact factors causing such category of deteriorations. As shown in Figure 2, after sudden deteriorations, deteriorated (or damaged) tracks are restored to satisfactory condition, which provides safe running surface for trains. This indicates that the restored tracks need not to be worried about very much from the perspective of train safety for research on track condition prediction. But for research on long-range track maintenance optimal scheduling, the restored tracks have to be taken into account in order to maximize benefit functions or to minimize cost functions over a planning horizon. Considering the focus of the current paper, it is assumed that deterioration rate estimation is made for a gradual deterioration process between occurrences of two adjacent sudden deteriorations.

If the value of $f(t)$ over the time range from t_i to t is assumed to be equal to $f(t_i)$, the equation $c(t) = c(t_i) + \int_{t_i}^{t} f(x)dx$ is rewritten as $c(t) = c(t_i) + f(t_i)(t - t_i)$. Actually, the approximation is basically true when the time range is short, like less than half a year [5, 19]. The reason for this is that within a span of such time range there is a small probability for track components to experience large performance degradations, which will result in rapid changes in deterioration rate. Therefore, the assumption for the currently concerned problem (i.e., short-range prediction) is acceptable. But the approximation of the deterioration rate function $f(t)$ by $f(t_i)$ assuredly introduces errors into the predicted track condition $c(t)$. Considering such consequence, a normal random variable, e, is incorporated into the approximate

equation. Accordingly, the short-range prediction mode for an analysis section is formulated as

$$\widehat{C}(t) = C(t_i) + \left(\overline{f}(t_i)(t - t_i) + e\right)V, \quad (1)$$

where $C(t_i)$ is a condition column vector of a specified geometrical parameter measured by a track geometry car at the time point t_i over sampling points in the analysis section, $\widehat{C}(t)$ is a column vector of predicted condition values at the time point t for the parameter on the sampling points, $\overline{f}(t_i)$ is the average of deterioration rates of the parameter at t_i on the sampling points and will be estimated in Section 3.3, and V is a column vector with the identical dimension to $C(t_i)$ and all elements equal to 1. In (1), the average deterioration rate $\overline{f}(t_i)$ rather than individual deterioration rate is used to make predictions. Such treatment of the deterioration rate is feasible because the analysis section is only 0.5 m in length and influence degrees of all impact factors are almost identical along the analysis section. Furthermore, the use of the average deterioration rate may also reduce the effects of slight noises in track geometry measurements.

According to the above discussion, the proposed model is built for the deterioration process of a geometrical parameter on a sampling point in an analysis section, and the deterioration process within a short time range is approximated by a linear model which has a normally distributed random component to quantify errors introduced by the approximation. Parameters of (1), $C(t_i)$ and $\overline{f}(t_i)$, have to be updated continuously as track condition evolves. In the current research, updating the parameters is triggered by availability of new inspection data from track geometry cars.

3.3. Optimal Time Span Estimation. As concluded in Section 3.1, each geometrical parameter of each analysis section has its own distinctive deterioration process. This indicates that the length of a time range within which historical measurement values are used to estimate the average deterioration rate $\overline{f}(t_i)$ through least squares method varies among seven geometrical parameters and varies along track. As for a geometrical parameter of sampling points on a given analysis section, differences between measurement values and prediction values by (1) on all the sampling points are calculated. The minimum sum of the squared differences is used as the objective function, formulated as (2) to determine the time range length at the time point t_i for the geometrical parameter on the given analysis section:

$$\begin{aligned} t_i^* &= t_i - t_{i-j^*} \\ &:= \arg\min_{1 \le j \le i-1} \left(\sum_{1 \le k \le n} \left(C(t_{i+k}) - \widehat{C}(t_{i+k}, j)\right)^T \right. \\ &\quad \left. \cdot \left(C(t_{i+k}) - \widehat{C}(t_{i+k}, j)\right) \right) \end{aligned} \quad (2)$$

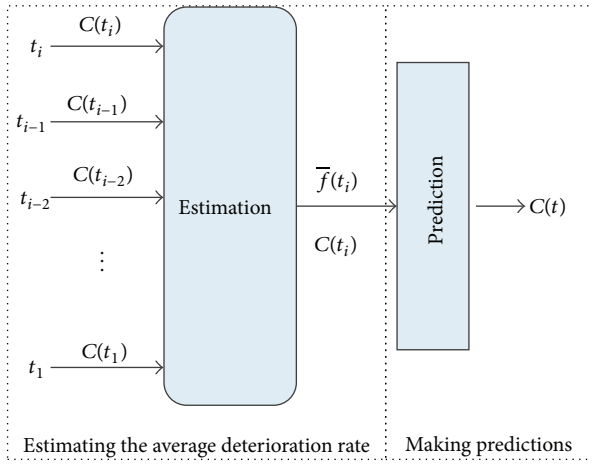


Figure 3: The formulated prediction model at the time point t_i.

where

$$\widehat{C}(t_{i+k}, j) = C(t_i) + \left(\overline{f}(t_i, j)(t_{i+k} - t_i) + e\right)V, \quad (3)$$

$$\overline{f}(t_i, j) = \frac{(j+1)\sum_{l=0}^{j}(t_{i-l} - t_i)\overline{C}(t_{i-l}) - \left(\sum_{l=0}^{j}(t_{i-l} - t_i)\right)\left(\sum_{l=0}^{j}\overline{C}(t_{i-l})\right)}{(j+1)\sum_{l=0}^{j}(t_{i-l} - t_i)^2 - \left(\sum_{l=0}^{j}(t_{i-l} - t_i)\right)^2}. \quad (4)$$

In the objective function, $\widehat{C}(t_{i+k}, j)$ denotes the prediction of the column vector $C(t_{i+k})$ when the last $(j+1)$ measurement column vectors $\{C(t_{i-k}),\ 0 \le k \le j\}$ are used to calculate the average deterioration rate $\overline{f}(t_i, j)$, and the constant n equals the number of track geometry cars' inspections in the time period which the proposed model covers. Equation (3) is the prediction model developed in Section 3.2. In (4), $\overline{C}(t_{i-l})$ is the mean of elements of the measurement column vector $C(t_{i-l})$, and $\overline{f}(t_i, j)$ is calculated through the least squares method from the point set $\{(t_{i-k} - t_i, \overline{C}(t_{i-k})),\ 0 \le k \le j\}$.

When the optimal solution to the object function, that is, j^*, is attained, the value of $\overline{f}(t_i, j)$ in (4) is the average deterioration rate which the proposed model uses to make predictions, as shown in Figure 3. As track inspection cars continue to inspect track condition, the process illustrated in Figure 3 will keep repeating itself.

4. Performance Analysis

This section analyzes the performance of the enhanced model in terms of statistical analyses of errors in geometry parameters values predicted by the enhanced model and comparison analyses of errors between the enhanced and original models. Since the middle of 2007, the authors' team has been collaborating with a couple of bureaus of China Railways on optimal track maintenance scheduling. The Jinan bureau is one of these collaborative bureaus. The collaboration with these bureaus entitles us to access their track

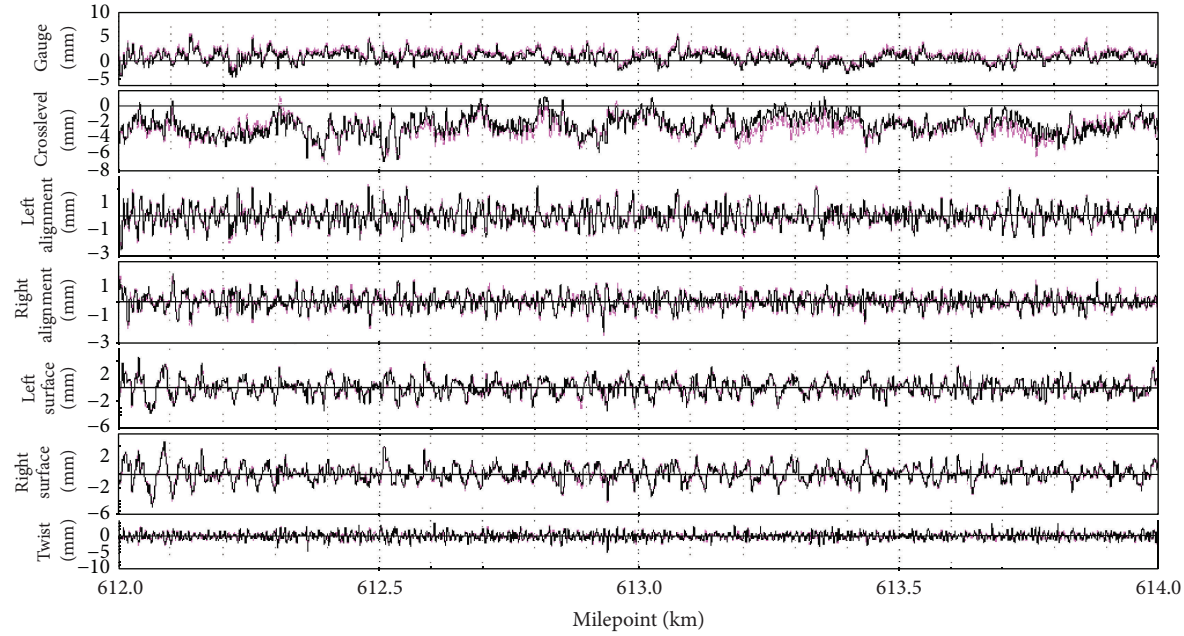


Figure 4: Superposed waves of track geometry data on February 20 and March 6.

configuration data and their various kinds of inspection data for track, such as track geometry data from track geometry cars and track geometry trolleys. The demonstration in this section uses track geometry data from track geometry cars for a 2-kilometer long track section in Jiulong-Beijing railroad, which is administrated by the Jinan bureau.

4.1. Data Preparation. The Jiulong-Beijing railroad is one of class I railroads connecting Beijing, the capital of China, to south part of the country. It transported approximately 68 million passengers and 42 MGT freights in the year of 2008. The inspection frequency of track geometry cars for the Jiulong-Beijing railroad is basically 3 times per 2 months. The track section whose track geometry data are used to perform the performance analyses starts at the milepoint of K612+000 and ends at K614+000.

Rails in this track section were manufactured in early 2003 and were placed on March 1 of 2003. They are all standard length rails, and their weight per meter is 60 kilograms. Because the Jiulong-Beijing railroad is a skeleton one, its rails are all continuously welded. Ties were manufactured in 1994 and were placed in 1995. All ties are of type II concrete ties and the number of ties per kilometer is 1760. The rails are fastened onto the ties with fastening systems of type I. Under the ties, 30 cm thick granite ballast and 20 cm thick granite subballast layers were laid in 1994. The section from K582+84512 to 650.33300, apparently spanning the track section to be analyzed, was rehabilitated in 2003.

Track geometry data that are used to do the performance analysis were acquired through a track geometry car of GJ-4, the most extensively used model of track geometry car in China Railways. Due to errors in milepoint measurements mentioned in Section 3.1, track geometry data were processed by both KEI and DSPM. Figure 4 as an example shows the superposed waves of track geometry data on February 20 and March 6, 2008, for the analyzed track section. In Figure 4, the waves of February 20 are plotted with the black color and the waves of March 6 with the gray color. From this figure, it is hard to differentiate the superposed waves of track geometry parameters. The reasons are that milepoint measurements are almost corrected by the milepoint correction models, namely, KEI and DSPM, and most track positions normally deteriorate slightly within a short period of time, that is, 16 days in the demonstrated example.

4.2. Statistical Analyses of Prediction Errors. After each inspection run of the track geometry car, track conditions within following two months, namely, values of each geometrical parameter on sampling points in the analyzed track section, were predicted by the original and enhanced models, respectively. The parameter of the length of the time period, within which track geometry data are processed to calculate the deterioration rate, was determined for the original model according to experiences of field engineers and takes on the value of four months, whereas the value for the enhanced model varies. After the inspection on October 30, track condition prediction is made for each of days from October 31 to December 31, 2008. Within this date range, there are three inspections by the track geometry car on November 13, December 12, and December 25.

Figure 5 plots predicted values versus actual values on December 12 for each geometry parameter. The predicted values are plotted along the vertical axis, whereas the actual ones are along the horizontal axis. Table 1 tabulates three statistical indices of errors in predictions by the original and enhanced models on these three days for each of geometrical

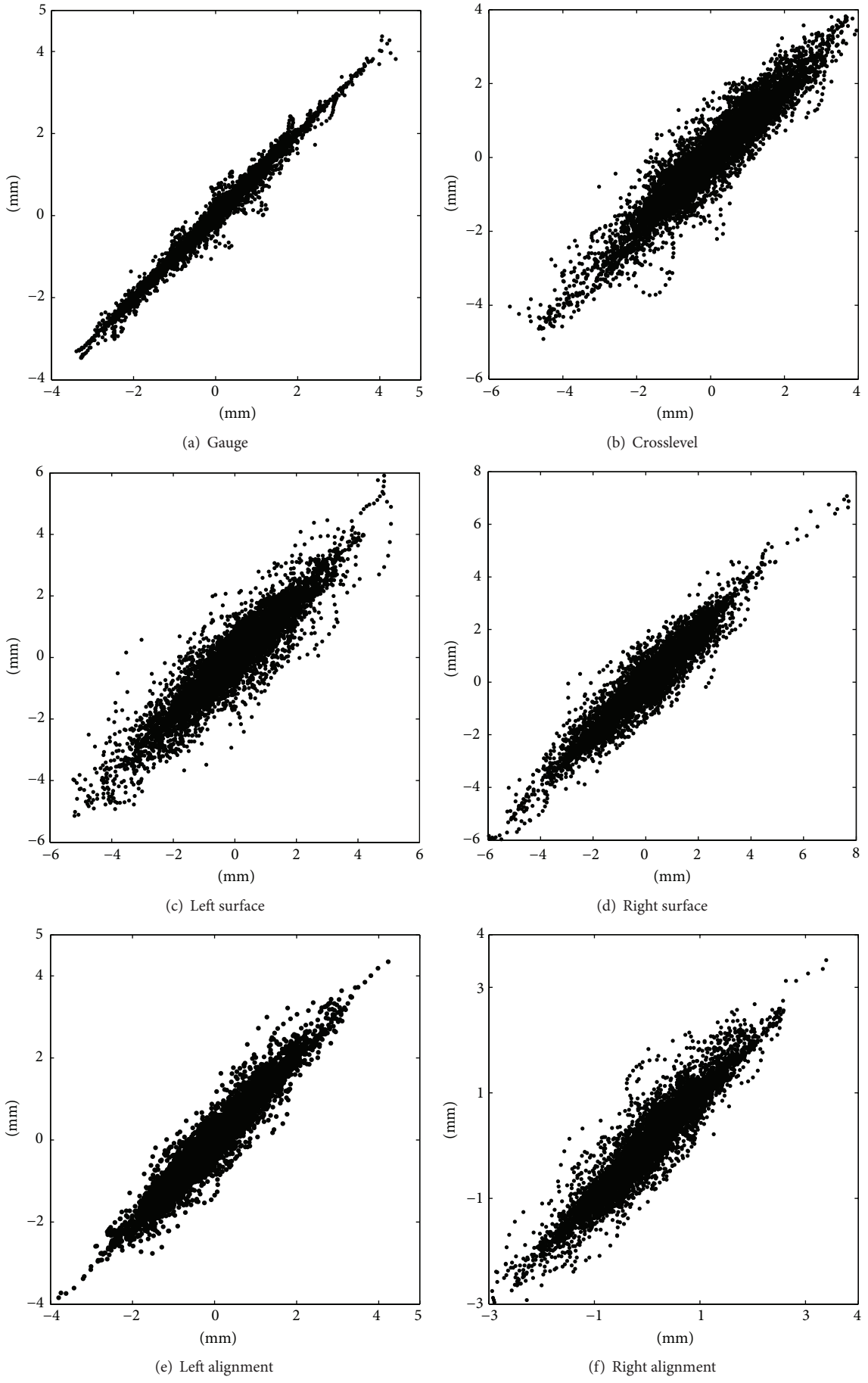
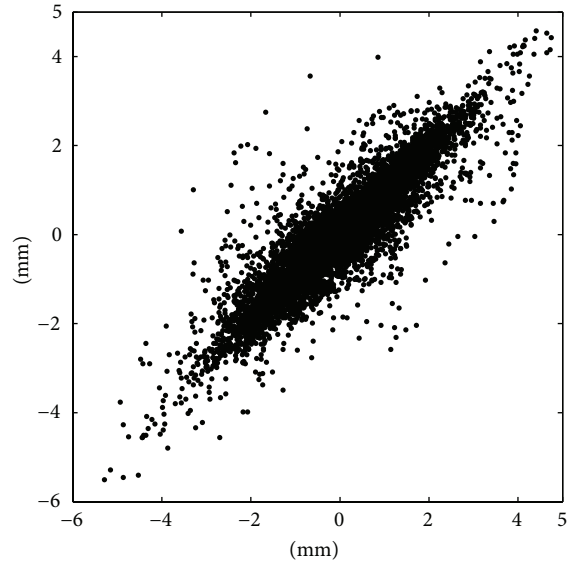


Figure 5: Continued.



(g) Twist

Figure 5: Comparison between actual values and predictions by the enhanced model on December 12.

Table 1: Statistics about errors in predictions by the original and enhanced models.

Geometrical parameters	Error statistics	November 13		December 12		December 25	
		Enhanced	Original	Enhanced	Original	Enhanced	Original
Gauge	μ (mm)	−0.0005	0.0000	−0.0016	0.0000	−0.0021	0.0000
	σ (mm)	0.1191	0.1554	0.1299	0.3478	0.0991	0.4155
	ρ	0.9947	0.9909	0.9935	0.9547	0.9962	0.9371
Crosslevel	μ (mm)	−0.0041	0.0000	−0.0125	0.0000	−0.0163	0.0000
	σ (mm)	0.5673	0.6313	0.4478	0.7209	0.4559	0.9123
	ρ	0.9119	0.8920	0.9469	0.8702	0.9517	0.8125
Left surface	μ (mm)	0.0008	0.0000	0.0024	0.0000	0.0031	0.0000
	σ (mm)	0.4752	0.4959	0.5304	0.6909	0.4135	0.6114
	ρ	0.9452	0.9410	0.9350	0.8935	0.9609	0.9192
Right surface	μ (mm)	−0.0008	0.0000	−0.0023	0.0000	−0.0031	0.0000
	σ (mm)	0.3722	0.4044	0.4450	0.5997	0.4272	0.6210
	ρ	0.9657	0.9603	0.9539	0.9200	0.9584	0.9175
Left alignment	μ (mm)	−0.0012	0.0000	−0.0038	0.0000	−0.0050	0.0000
	σ (mm)	0.3340	0.3597	0.2761	0.4291	0.2556	0.4713
	ρ	0.9511	0.9438	0.9662	0.9225	0.9711	0.9097
Right alignment	μ (mm)	−0.0012	0.0000	−0.0035	0.0000	−0.0046	0.0000
	σ (mm)	0.3267	0.3564	0.2906	0.4291	0.2682	0.4478
	ρ	0.9284	0.9159	0.9432	0.8833	0.9525	0.8786
Twist	μ (mm)	0.0000	0.0000	−0.0001	0.0000	−0.0001	−0.0001
	σ (mm)	0.5184	0.5474	0.5696	0.7441	0.4085	0.6026
	ρ	0.9037	0.8949	0.8996	0.8347	0.9547	0.9006

parameters. The tabulated three statistical indices are mean (μ), standard deviation (σ), and Pearson's correlation coefficient (ρ).

The plots in Figure 5 clearly show that the predicted values for each geometry parameter are pretty close to the actual ones. For the parameter of gauge, 95 percent of the predicted values have errors less than 0.2717 mm; errors of 95% predictions for crosslevel, left surface, right surface, left alignment, right alignment, and twist are less than 1.1409 mm, 1.4894 mm, 1.2346 mm, 0.7765 mm, 0.8202 mm, and 1.6558 mm, respectively.

Table 1 shows that the mean and standard deviation for each geometry parameter are far below its corresponding theoretical measurement accuracy of the track geometry

Table 2: TQI management thresholds specified by MOR.

Railroad class (defined by allowable speed, v_{max})	$v_{max} \leq 160$	$160 < v_{max} \leq 200$	$200 < v_{max} \leq 250$	$300 < v_{max} \leq 350$
TQI management threshold (mm)	15	10	8	5

car. The measurement accuracies for all geometry parameters, gauge, crosslevel, left/right surface, left/right alignment, and twist are ±0.8 mm, ±1.0 mm, ±1.0 mm, ±1.5 mm, and ±1.0 mm, respectively. The correlation coefficients in Table 1 quantitatively confirm the closeness of the predicted values to the actual values again. From these figures, numbers, and already obtained facts, it is concluded that the enhanced model can make fair accurate prediction for track condition two months in advance.

4.3. Comparison Analysis between the Original and Enhanced Models. From Table 1, it is clear that the mean of errors in predictions by the enhanced model for each parameter on each day is almost identical to the one by the original model and approximately equals 0. When it comes to the other statistical index (standard deviation), values for the enhanced model are less than the ones for the original model, whereas the correlation coefficient associated with the enhanced model is greater than the one with the original model. Those inferences imply that in comparison with the original model the enhanced model makes more accurate prediction.

Figure 6 shows the values of standard deviation in Table 1 as vertical bars for each geometry parameter. From these figures, it is apparent that standard deviations associated with the enhanced model are less than the ones with the original model. More importantly, standard deviations linked with the original model basically increase as the time span between the date of the available inspection, for example, October 30, and the date on which predictions were made, for example, November 13, lengthens; however, the standard deviation connected with the enhance model stays robust in the whole time span. This inference implies that compared with the original model the enhanced model possesses a characteristic of robustness.

5. Discussion about Predictions Usage for Optimal Track Maintenance Scheduling

After each inspection run of track geometry cars, the inspection data is used to calculate two categories of track condition indices: track geometrical exceptions and Track Quality Index (TQI) [29–31]. Track geometrical exceptions are characterized as the geometry parameter, the maximum value, the exception class, and the length. Track geometrical exceptions fall into four classes according to their maximum values: I, II, III, and IV. The lower and upper limits of each exception class for each geometry parameter with a speed range are specified by MOR in the Railway Line Maintenance Regulations. Figure 7 depicts an exception of class III. For geometry exceptions of each class, MOR recommends certain measures to be taken. TQI is calculated for a track unit section and is the sum of standard deviations of seven geometry parameters over the unit section [19]. MOR also specifies the management thresholds of TQI for each class of railroads, as listed in Table 2. It is recommended by MOR that track unit sections with TQI greater than the corresponding TQI threshold should be considered when scheduling track tamping maintenance.

A program having the enhanced model has been coded by the authors. The two categories of track condition indices are therefore available two or three months in advance. According to the Railway Line Maintenance Regulations, future two or three months of maintenance works are available in advance. Given maintenance works of two or three months in a planning horizon, they can be optimally scheduled so as to minimize over the planning horizon the sum of the cost for travels between maintenance sites and travels between maintenance sites and depots, the cost for maintenance works themselves, and the influential cost of completing the works on rail transportation. The object function of the optimal scheduling problem is subject to the constraints of available maintenance resources including maintenance machines, required materials, crew members, and track windows left in train timetable.

6. Conclusions and Following Research Areas

As railroad infrastructure becomes older and older and rail transportation is developing towards higher speed and heavier axle, the risk to safe rail transport and the expenses for railroad infrastructure maintenance are increasing. The railroad infrastructure deterioration (prediction) model is vital to reducing the risk and the expenses. This paper enhanced our previous railroad track prediction model. The previous model considers that the length of the historical period (within which track geometry data from track geometry cars are used to estimate track deterioration rates) is constant for all analysis sections in a railroad track. The value of the length for the previous model is determined according to field engineers' knowledge. Comparison analyses between track geometry exceptions calculated from track condition measurement data from track geometry cars and those from predictions by the model imply that the method cannot provide reliable track condition predictions for all analysis sections. The enhanced model, according to track deterioration process revealed by lining up historical track geometry data, employs the minimum sum of squared differences between prediction values and measurement values to estimate the optimal length of the historical period for each analysis section. One year of track geometry data for the Jiulong-Beijing railroad was used in the section of performance analysis to perform error analyses and comparison analysis between the original and enhanced models. The analysis results show

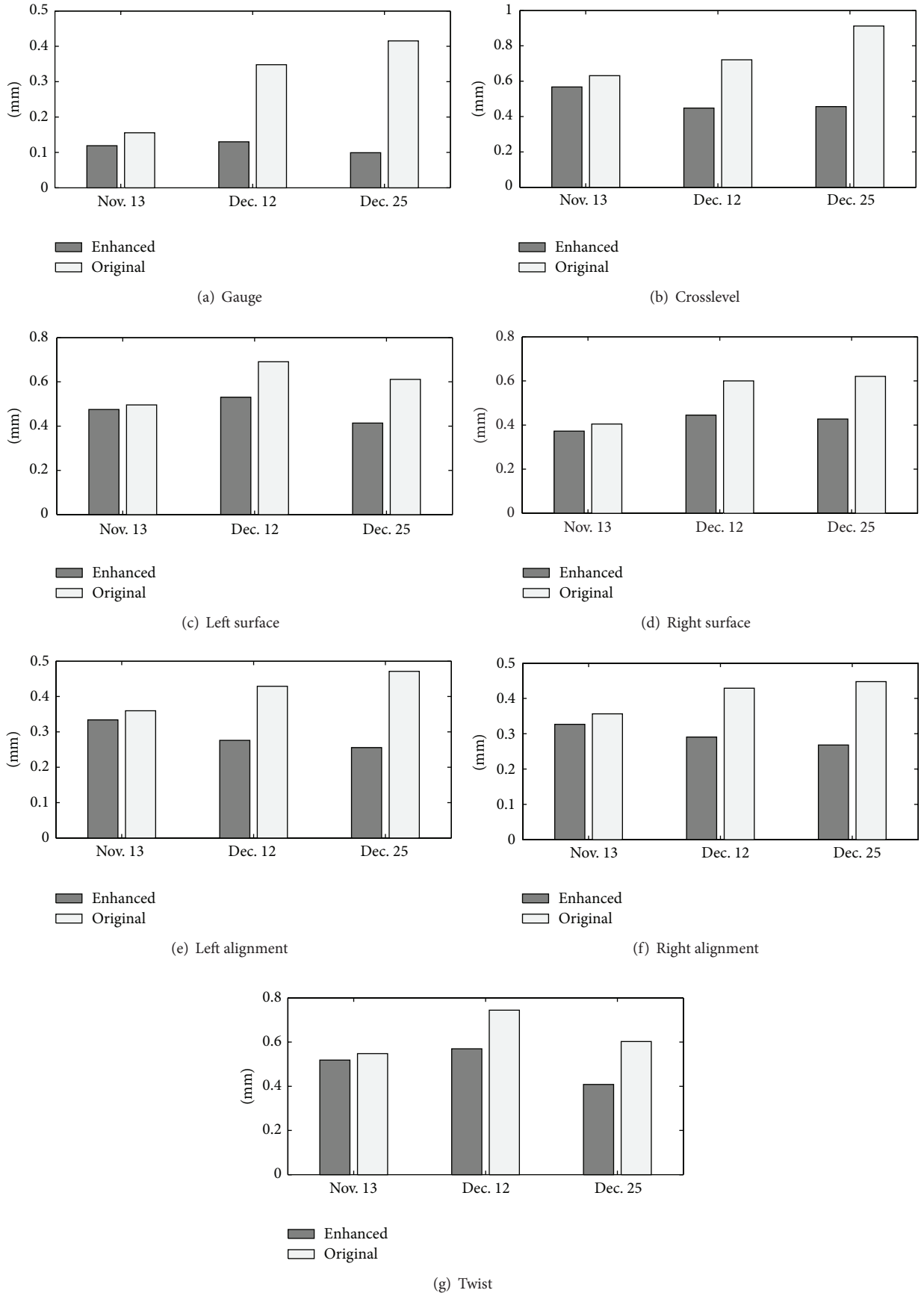


Figure 6: Comparison of standard deviations of prediction errors between the original and enhanced models.

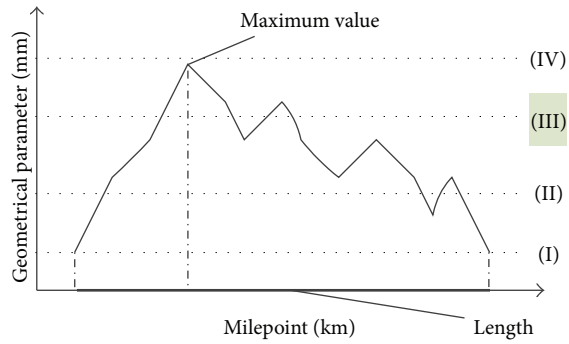


Figure 7: Illustrative example of a track geometrical exception of class III.

that the enhanced model not only makes more accurate predictions for track geometry condition but also possesses the characteristic of robustness. What is more, because the enhanced model uses actual deterioration processes revealed by historical track geometry data to formulate prediction model for each analysis section, it has extensive suitability for analysis sections in various conditions.

This paper also gave a brief discussion about how to use the predictions by the enhanced model to optimally schedule track maintenance works. The research on mathematical formulation of the optimal track maintenance scheduling is in process and will come out in a near future.

Conflict of Interests

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

Authors' Contribution

Peng Xu and Chuanjun Jia contributed equally to this paper.

Acknowledgments

This research was funded by Beijing Jiaotong University under Grant 2014RC003 and by the Fundamental Research Funds for the Central Universities under Grant 2015JBM051. Jinan bureau of China Railways provided track condition measurement data from the track geometry car to this research. The current chief engineer of the maintenance-of-way inspection center of Jinan bureau, Mr. Hong Sun, and the former chief engineer, Mr. Zhigang Li, shared their knowledge regarding effects of influential factors on track deterioration.

References

[1] RITA, "Table 1: U.S. Domestic Freight Ton-Miles by Mode," 2007, http://apps.bts.gov/publications/special_reports_and_issue_briefs/special_report/2007_07_27/html/table_01.html.

[2] National Bureau of Statistics of China, *China Statistical Yearbook 2012*, National Bureau of Statistics of China, Beijing, China, 2012, http://www.stats.gov.cn/tjsj/ndsj/2012/indexeh.htm.

[3] E. T. Selig and J. M. Waters, *Track Geotechnology and Substructure Management*, Thomas Telford Ltd, London, UK, 1994.

[4] L. Ferreira, "Planning Australian freight rail operations: an overview," *Transportation Research Part A: Policy and Practice*, vol. 31, no. 4, pp. 335–348, 1997.

[5] P. Xu, *Mileage correction model for track geometry data from track geometry car & track irregularity prediction model [Ph.D. thesis]*, Beijing Jiaotong University, Beijing, China, 2012.

[6] J. V. Carnahan, W. J. Davis, M. Y. Shahin, P. L. Keane, and M. I. Wu, "Optimal maintenance decisions for pavement management," *Journal of Transportation Engineering*, vol. 113, no. 5, pp. 554–572, 1987.

[7] P. L. Durango-Cohen and S. M. Madanat, "Optimization of inspection and maintenance decisions for infrastructure facilities under performance model uncertainty: a quasi-Bayes approach," *Transportation Research Part A: Policy and Practice*, vol. 42, no. 8, pp. 1074–1085, 2008.

[8] P. L. Durango and S. M. Madanat, "Optimal maintenance and repair policies in infrastructure management under uncertain facility deterioration rates: an adaptive control approach," *Transportation Research Part A: Policy and Practice*, vol. 36, no. 9, pp. 763–778, 2002.

[9] K. Golabi, R. B. Kulkarni, and G. B. Way, "A statewide pavement management-system," *Interfaces*, vol. 12, no. 6, pp. 5–21, 1982.

[10] S. R. Seyedshohadaie, I. Damnjanovic, and S. Butenko, "Risk-based maintenance and rehabilitation decisions for transportation infrastructure networks," *Transportation Research—Part A: Policy and Practice*, vol. 44, no. 4, pp. 236–248, 2010.

[11] K. Kobayashi, K. Kaito, and N. Lethanh, "A statistical deterioration forecasting method using hidden Markov model for infrastructure management," *Transportation Research Part B: Methodological*, vol. 46, no. 4, pp. 544–561, 2012.

[12] P. L. Durango-Cohen, "A time series analysis framework for transportation infrastructure management," *Transportation Research Part B: Methodological*, vol. 41, no. 5, pp. 493–505, 2007.

[13] A. Kawaguchi, M. Miwa, and K. Terada, "Actual data analysis of alignment irregularity growth and its prediction model," *Quarterly Report of RTRI*, vol. 46, no. 4, pp. 262–268, 2005.

[14] R. M. Alfelor, G. A. Carr, and M. Fateh, "Track degradation assessment using gage restraint measurements," *Transportation Research Record*, vol. 1742, pp. 68–77, 2001.

[15] P. Veit and S. Marschnig, "Sustainability in track—a precondition for high speed traffic," in *Proceedings of the Joint Rail Conference*, Urbana, Ill, USA, 2010.

[16] C. Meier-Hirmer, G. Riboulet, F. Sourget, and M. Roussignol, "Maintenance optimization for a system with a gamma deterioration process and intervention delay: application to track maintenance," *Proceedings of the Institution of Mechanical Engineers Part O: Journal of Risk and Reliability*, vol. 223, no. 3, pp. 189–198, 2009.

[17] Q. He, H. Li, D. Bhattacharjya, D. P. Parikh, and A. Hampapur, "Railway track geometry defect modeling: deterioration, derailment risk and optimal repair," in *Proceedings of the Transportation Research Board Annual Meeting*, The Academy of Transportation Research Board, Washington, DC, USA, 2013.

[18] R. K. Liu, P. Xu, and Q. X. Sun, "A novel algorithm for predicting track irregularities of unit track sections," in *Proceedings of the Joint Rail Conference (JRC '10)*, Urbana, Ill, USA, April 2010.

[19] P. Xu, Q. Sun, R. Liu, and F. Wang, "A short-range prediction model for track quality index," *Proceedings of the Institution of Mechanical Engineers Part F: Journal of Rail and Rapid Transit*, vol. 225, no. 3, pp. 277–285, 2011.

[20] B. Indraratna, W. Salim, and C. Rujikiatkamjorn, *Advanced Rail Geotechnology—Ballasted Track*, CRC Press, Taylor and Francis Group, London, UK, 2011.

[21] A. Hamid, K. Rasmussen, M. Baluja, and T. L. Yang, *Analytical Descriptions of Track Geometry Variations: Main Text*, vol. 1, ENSCO, Springfield, Va, USA, 1981.

[22] T. B. Maertens Jr., *The relationships of climate and terrain to maintenance of way on the Norfolk Southern Railroad between Norfolk, Virginia, and Portsmouth, Ohio [Ph.D. thesis]*, The University of Tennessee, Knoxville, Tenn, USA, 1990.

[23] J. Sadeghi and H. Askarinejad, "Influences of track structure, geometry and traffic parameters on track deterioration," *IJE Transactions B: Applications*, vol. 20, no. 3, pp. 292–300, 2007.

[24] H. C. Shih, "Government to act on high-speed rail subsidence problem," 2011, http://www.taipeitimes.com/News/taiwan/archives/2011/07/26/2003509173.

[25] L. M. Quiroga and E. Schnieder, "A heuristic approach to railway track maintenance scheduling," in *Proceedings of the 12th International Conference on Computer System Design and Operation in the Railways and other Transit Systems —(COMPRAIL '10)*, B. Ning and C. A. Brebbia, Eds., pp. 687–699, Beijing, China, September 2010.

[26] P. Xu, R. Liu, Q. Sun, and L. Jiang, "Dynamic-time-warping-based measurement data alignment model for condition-based railroad track maintenance," *IEEE Intelligent Transportation Systems Magazine*, vol. 16, no. 2, pp. 799–812, 2014.

[27] P. Xu, Q. Sun, R. Liu, R. R. Souleyrette, and F. Wang, "Optimizing the alignment of inspection data from track geometry cars," *Computer-Aided Civil and Infrastructure Engineering*, vol. 30, no. 1, pp. 19–35, 2015.

[28] P. Xu, Q.-X. Sun, R.-K. Liu, and F.-T. Wang, "Key equipment identification model for correcting milepost errors of track geometry data from track inspection cars," *Transportation Research—Part C: Emerging Technologies*, vol. 35, pp. 85–103, 2013.

[29] Ministry of Railways of the People's Republic of China, *Regulations for Existing Speed-Raised Railway Lines and Bridge Facilities with the Allowable Speed between 200 and 250 km/h*, China Railway Press, Beijing, China, 2007.

[30] Ministry of Railways of the People's Republic of China, *Beijing-Tianjin Passenger Dedicated Railway Operation Maintenance*, China Railway Press, Beijing, China, 2009.

[31] The Ministry of Railways of the People's Republic of China, *Railway Line Maintenance Regulations*, China Railway Press, Beijing, China, 2010.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 101059, 8 pages
http://dx.doi.org/10.1155/2015/101059

Research Article

Analysis of Road Traffic Network Cascade Failures with Coupled Map Lattice Method

Yanan Zhang,[1,2] Yingrong Lu,[1,2] Guangquan Lu,[1,2] Peng Chen,[1,2] and Chuan Ding[1,2]

[1] *Beijing Key Laboratory for Cooperative Vehicle Infrastructure Systems and Safety Control, School of Transportation Science and Engineering, Beihang University, Beijing 100191, China*
[2] *Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies, SiPaiLou No. 2, Nanjing 210096, China*

Correspondence should be addressed to Guangquan Lu; lugq@buaa.edu.cn

Received 26 March 2015; Revised 18 June 2015; Accepted 21 June 2015

Academic Editor: Yuanchang Xie

In recent years, there is growing literature concerning the cascading failure of network characteristics. The object of this paper is to investigate the cascade failures on road traffic network, considering the aeolotropism of road traffic network topology and road congestion dissipation in traffic flow. An improved coupled map lattice (CML) model is proposed. Furthermore, in order to match the congestion dissipation, a recovery mechanism is put forward in this paper. With a real urban road traffic network in Beijing, the cascading failures are tested using different attack strategies, coupling strengths, external perturbations, and attacked road segment numbers. The impacts of different aspects on road traffic network are evaluated based on the simulation results. The findings confirmed the important roles that these characteristics played in the cascading failure propagation and dissipation on road traffic network. We hope these findings are helpful to find out the optimal road network topology and avoid cascading failure on road network.

1. Introduction

In many large-scale networks, the failure of a node or edge would make the other nodes fail and lead to a chain reaction due to the coupling relationships among nodes. This phenomenon is known as network cascading failure. Cascading failure problems may take place on many natural or artificial networks, such as the Internet [1, 2], power grids [3–6], and traffic networks [7–11]. The effects from large destruction may be caused by cascading failures on the entire networks. Many cities have suffered from serious traffic paralysis that brought great inconvenience to people's normal life (e.g., Beijing urban traffic was shut down completely due to the rainstorm on July 21, 2012). Therefore, it is essential to understand the cascading failure on traffic network to prevent or reduce the influences of large-scale failure.

Many scholars have studied the impacts of network topology [7, 12], network connectivity [13], different attack strategies [4, 14, 15], and network robustness [16–18] on cascading failure. To describe the cascading failure, coupled map lattice (CML) model has been widely applied in previous literatures. For example, using the basic CML method, Xu and Wang [12] studied the cascading failures in different network topologies. Based on the proposed edge-based CML method, Di et al. [19] investigated the cascading failure on random networks and scale-free networks. Though most studies paid attention to the artificial network on cascading failure, the research that applies CML model to investigate the natural road traffic network on cascading failure is limited.

Because the properties of natural road traffic network are different from artificial networks, the particular road traffic network properties need to be concerned when we use CML model. One of the particular properties is aeolotropism. Due to the fact that there are one-way and two-way streets in the city, the road traffic network is supposed to be described as directed graphs. Another particular property is restorability, which means that road congestions can dissipate over a certain range. The road traffic network consists of intersections and road segments. Vehicles travel on the network and form the distributed traffic flow. If traffic congestions occur in one or some road segments, congestions can be gradually dissipated after a period of time due to the redistribution

of traffic flow. These two particular properties may lead to unique cascading failures rules in road traffic network.

Considering the above particular properties, the original CML model will be improved for analyzing cascading failures of road traffic network. The improved CML model is expected to express the aeolotropism of road traffic network topology, which will be proposed in the following section. Besides, in order to match the pattern of road congestion dissipation, a recovery mechanism has been put forward in the next part. For the purpose of deliberating cascading failures roundly, an empirical network in Beijing is tested to investigate the impacts of different attack strategies, coupling strengths, external perturbations, and attacked road segment numbers on road traffic network.

The remainder of this paper is organized as follows. The next section will introduce the improved CML model and the recovery mechanism. Then, the simulations based on the empirical network are conducted. Finally, the highlights of this paper are concluded.

2. Road Traffic Network Cascading Failures Model Based on CML

The original CML model is formulated as follows [12]:

$$x_i(t+1) = \left| (1-\varepsilon) f(x_i(t)) + \varepsilon \sum_{j=1, i\neq j}^{N} \frac{a_{i,j} f(x_j(t))}{k(i)} \right|, \quad (1)$$

where $x_i(t)$ is the state of the ith node at the tth time step. $\varepsilon \in (0, 1)$ is defined as the coupled strength. N is the sum of all nodes. $k(i)$ represents the degree of the ith node. Adjacency matrix $A = (a_{ij})_{N\times N}$ is used to represent the topology of the network. If there is an edge between node i and node j, then $a_{ij} = a_{ji} = 1$; otherwise, $a_{ij} = a_{ji} = 0$. Chaotic Logistic map $f(x) = \mu x(1-x)$ with $\mu \in (0, 4]$ is used to denote dynamic behaviors of nodes. μ is closer to 4; the value of $f(x)$ is more evenly distributed throughout the region of 0 to 1. Therefore, it is always set to be $\mu = 4$. The absolute value notation is used in (1) for ensuring nonnegative saturation state of each node.

To describe the aeolotropism of road traffic network, an improved CML model is proposed to investigate the cascading failures on road traffic network. In the original CML model, $x_i(t)$ means the state of ith node at the tth time step, while, for road traffic network, it is expressed by road saturation:

$$x_i(t+1) = \left| (1-\varepsilon_1-\varepsilon_2) f(x_i(t)) + \varepsilon_2 \sum_{i=1, i\neq j}^{N_1} \frac{b_{ij} f(x_j(t))}{k^-(i)} + \varepsilon_1 \sum_{j=1, i\neq j}^{N_2} \frac{b_{ji} f(x_i(t))}{k^+(i)} \right| \quad (2)$$
$$i, j = 1, 2, \ldots, n.$$

In (2), $x_i(t)$ means the road saturation of the ith road segment at the tth time step. b_{ij} is the value of adjacency matrix $B = (b_{ij})_{N\times N}$ of the road traffic network. If there is an edge from node i to node j, then $b_{ij} = 1$; otherwise, $b_{ij} = 0$. $\varepsilon_1 \in (0, 1)$ and $\varepsilon_2 \in (0, 1)$ delegate the coupled strengths of the start point and endpoints, respectively. N_1 is the sum of all nodes' out-degree, and N_2 is that of in-degree. $k^+(i)$ and $k^-(i)$, respectively, represent the in-degree and out-degree of the ith node which means the number of downstream segments and upstream segments for the road traffic network.

Cascading failure on road traffic network may be triggered by some internal and external factors (e.g., traffic congestion or crash) that lead to the failure of one or more roads. To describe this situation, an external perturbation $R \geq 1$ is added to the node k at the $(m+1)$th time as follows:

$$x_k(m+1) > 1 = \left| (1-\varepsilon_1-\varepsilon_2) f(x_k(m)) + \varepsilon_2 \sum_{k=1, k\neq j}^{N_1} \frac{b_{kj} f(x_j(m))}{k^-(k)} + \varepsilon_1 \sum_{j=1, k\neq j}^{N_2} \frac{b_{jk} f(x_k(m))}{k^+(k)} \right| + R \quad (3)$$
$$i, j = 1, 2, \ldots, n.$$

If $0 < x_k(t) < 1$ when $t \leq m$, the node k is in a normal state; if $x_k(m+1) \geq 1$, the node k is defined to be failed at the $(m+1)$th time step. For the situation when the node k fails at the $(m+1)$th time step, $x_k(t) \equiv 0$ with $t > m+1$ is defined in previous studies [12, 19]. However, with regard to the road traffic network, the failure state could not continue all the time due to the fact that traffic congestion will gradually dissipate with the redistribution of traffic flow. A recovery mechanism to fit with road traffic characteristics is proposed in this study as follows.

If an external perturbation R is added to the node k at the $(m+1)$th time step, $x_k(m+1) > 1$ means that road segment k has been in a blocked state at $(m+1)$th time step. If the upstream vehicles cannot enter, coupled strength of upstream segments and road segment k is set to be 0 (i.e., $\varepsilon_2 = 0$). The saturation state of the kth node from $(m+2)$th step to $(m+n+1)$th step can be represented by (4). If $x_k(m+n+1) < 1$ after n steps, the saturation state of the kth node returns to normal, represented by (2). The recovery mechanism of the road traffic network on cascading failure is shown in Figure 1.

Consider

$$x_i(t+1) = \left| (1-\varepsilon_1) f(x_i(t)) + \varepsilon_1 \sum_{j=1, i\neq j}^{N_2} \frac{b_{ji} f(x_i(t))}{k^+(i)} \right| \quad (4)$$
$$i, j = 1, 2, \ldots, n.$$

The proportion of failed nodes at each time step is used to characterize road cascading failure process as shown in (5) and I is applied to represent the final $P(t)$ to describe the size of cascading failure in the end:

$$P(t) = \frac{N'(t)}{N}. \quad (5)$$

3. Simulation Test and Analysis

To detail the computational experiment of cascading failure on road traffic network, the real road network of

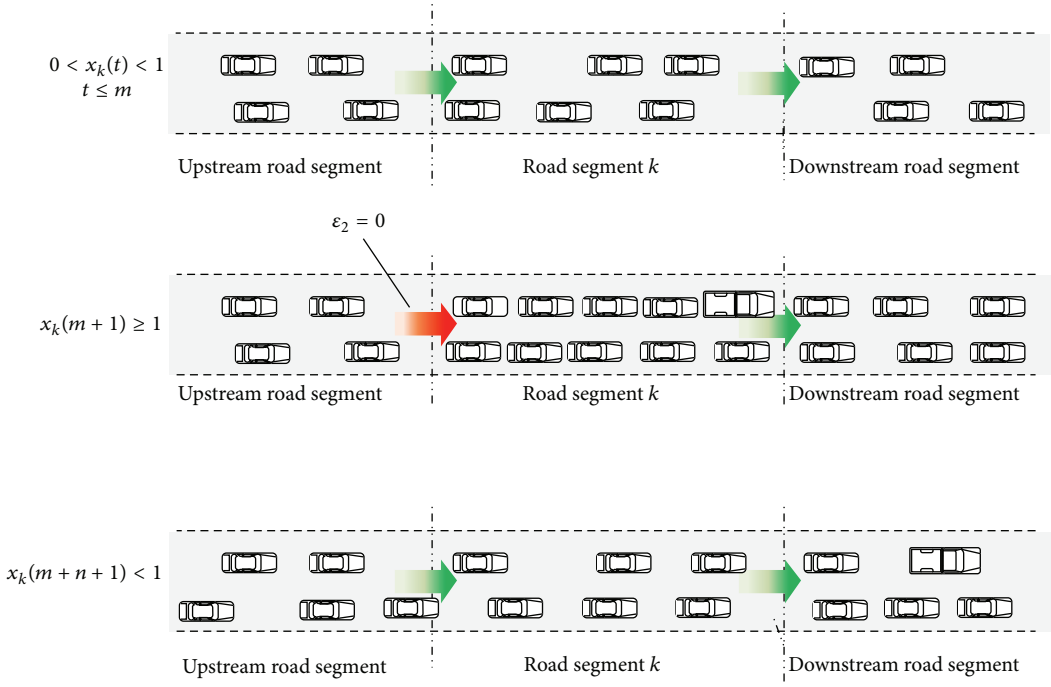
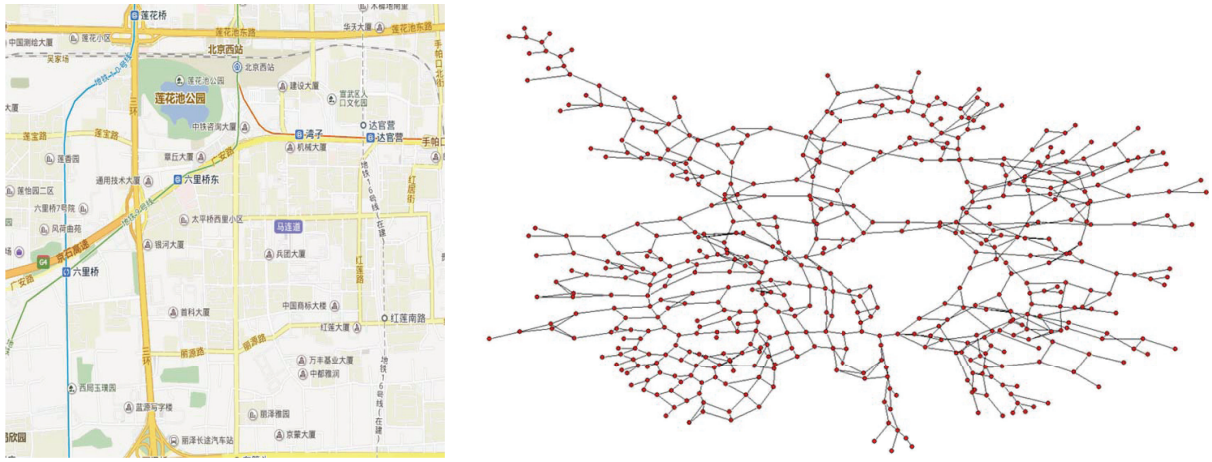


Figure 1: Recovery mechanism of road traffic network on cascading failure.



(a) (b)

Figure 2: (a) Map of Liuliqiao area in Beijing. (b) Network topology of Liuliqiao area in Beijing.

Liuliqiao area in Beijing (total road segments number $N = 1004$) is selected as the empirical network in this study, as shown in Figure 2(a). This area of more than 22 km^2 contains the largest train station in Beijing and is considered a typical region showing transition between free flow and congestions. For the road network, nodes represent the intersections and edges represent the road segments between two intersections as shown in Figure 2(b). Using the proposed model, cascading failure on road traffic network is tested based on different attack strategies, coupling strengths, external perturbations, and attacked road segment numbers.

3.1. Different Attack Strategies. Different attack strategies would lead to different cascading failure on the network. In this study, to obtain the influences of different cascading failure, four kinds of attack strategies are tested: the deliberate attack based on betweenness (BA), the deliberate attack based on saturation (SA), the deliberate attack based on combination of betweenness and saturation (BSA), and the random attack (RA).

Parameter λ is used to characterize the three attack strategies (BA, SA, and BSA) as follows:

$$f_i(t) = \lambda x_i(t) + (1-\lambda) b_i(t), \quad (6)$$

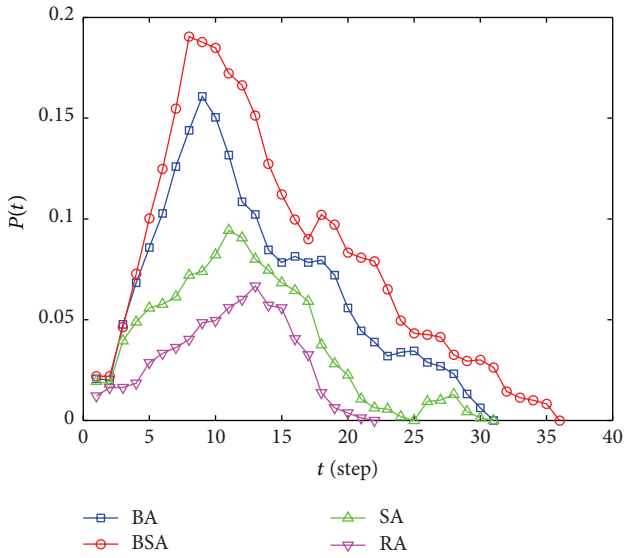


FIGURE 3: $P(t)$ based on different attack strategies.

where λ ($0 \leq \lambda \leq 1$) is a weight coefficient. $x_i(t)$ is the saturation state and $b_i(t)$ is the betweenness of the ith node at the mth time step. $f_i(t)$ is the combination of betweenness and saturation of the ith node with a fixed value λ. $\lambda = 0$ represents BA and the initial failure nodes are deleted in turn according to betweenness. If multiple maximum betweenness nodes exist, we choose the node being attacked in a random manner and $\lambda = 1$ refers to SA which means that the nodes are attacked gradually with saturation. Similarly, the node is randomly selected to be attacked when maximum saturation nodes exist. For BSA, each attack selects the node with maximum combination of degree and betweenness, while the value of λ equals 0.5. RA means that the node to be attacked randomly.

For BA, SA, and BSA, external perturbation $R = 1.5$ is added to a node with the largest value of f_i (corresponding to different values of λ). For RA, $R = 1.5$ is added to a randomly chosen node. Figure 3 shows the results of four kinds of attacking strategies with $\varepsilon_1 = \varepsilon_2 = 0.6$.

Figure 3 shows the occurrence of failure and the process of recovery. This phenomenon is in conformity with the actual road traffic flow. Figure 3 also shows that BA triggers cascading failures more easily than SA. The scale of failure recovery time under BA is longer than SA. This phenomenon implies that betweenness has more destructive impacts on cascading failures than saturation. As shown in Figure 3, RA is least likely to trigger cascading failures and the failures recovery time is also the shortest. It is reasonable for that the node being randomly attacked is usually not that node which has deteriorated impact on network cascading failure. Interestingly, comparing to other three attacks, BSA has the most serious impacts on network, including the largest number of failed nodes, the fastest propagation rate of failure, and the longest recovery time. This implies that those road segments, which have the largest value of combination of betweenness and saturation, are the key nodes causing large-scale cascading failures once attacked. These findings give the guidance on daily traffic control that the potential cascading failures could be avoided by supervising the key road segments and their adjacent segments.

3.2. Different Coupling Strength. The deficiency of giving the coupled strength a fixed value subjectively [19] is overcome in this study. In the case of different values of ε_1 and ε_2, cascading failures are triggered by adding the same external perturbation $R = 1.5$ on one node with RA. The simulation results are shown in Figure 4. Figure 4(a) plots the proportion of failed nodes $P(t)$ versus time step t with fixed value of coupling strength ε_2 ($\varepsilon_2 = 0.6$) and varying values of ε_1 ($\varepsilon_1 = 0.1, 0.2, \ldots, 0.9$). Oppositely, Figure 4(b) presents the time series of $P(t)$ with fixed value of coupling strength ε_1 ($\varepsilon_1 = 0.6$) and varying values of ε_2 ($\varepsilon_2 = 0.1, 0.2, \ldots, 0.9$).

According to Figures 4(a) and 4(b), when the value of ε_1 or ε_2 is below 0.5, road traffic network cascading failures hardly occur. As the value of the coupling strength increases, especially larger than 0.6, the number of failed nodes and the failure recovery time increase sharply.

Figure 5 shows coupled strength against ratio of total failed nodes number I for the different attack strategies. Road traffic network cascading failures are triggered by external perturbation $R = 1.5$. We can see that the size of cascading failures increases as the value of the coupling strength increases. This is consistent with the findings from Figure 4. Figure 5 also shows that there is a threshold for each attack strategy. Only when the value of coupling strength is larger than the threshold, the cascading failures occur. For example, the BA curve shows that the cascading failures occur when ε_1 is larger than 0.4 in Figure 5(b). Comparing four attacks, the threshold of BSA is the smallest. This illustrates that the deliberate attack based on combination of betweenness and saturation is likely to cause cascading failures even under the low coupling strength.

Comparing Figure 5(a) with Figure 5(b), the curves of RA and SA present a different trend. In Figure 5(a), if the value of coupling strength ε_2 is less than 0.6, the size of cascading failure with RA is smaller than SA. When the value of ε_2 is larger than 0.7, the size of cascading failure with SA is smaller than RA. In Figure 5(b), SA is always larger than RA. This phenomenon may be due to the fact that random attack might select the noncritical nodes or the critical nodes, leading to different results.

3.3. Different External Perturbation R. The impact of different external perturbation R on the cascading failure on road traffic network is analyzed by adding varying external perturbation R on a fixed node. Figure 6 shows that the proportion of failed nodes $P(t)$ varies with the time step t in the case of different value of R with $\varepsilon_1 = \varepsilon_2 = 0.6$. The inset of Figure 6 shows the time series of $P(t)$ for the value of R between 1 and 2 ($R = 1, 1.2, 1.4, 1.6, 1.8, 2$).

According to Figure 6, with the increase of R value, the number of failed nodes and failure recovery time increase. There is a threshold R_c for R. Only when the value of R is larger than the R_c, the large-scale cascading failures occur seriously. The inset of Figure 6 shows that R_c value is between

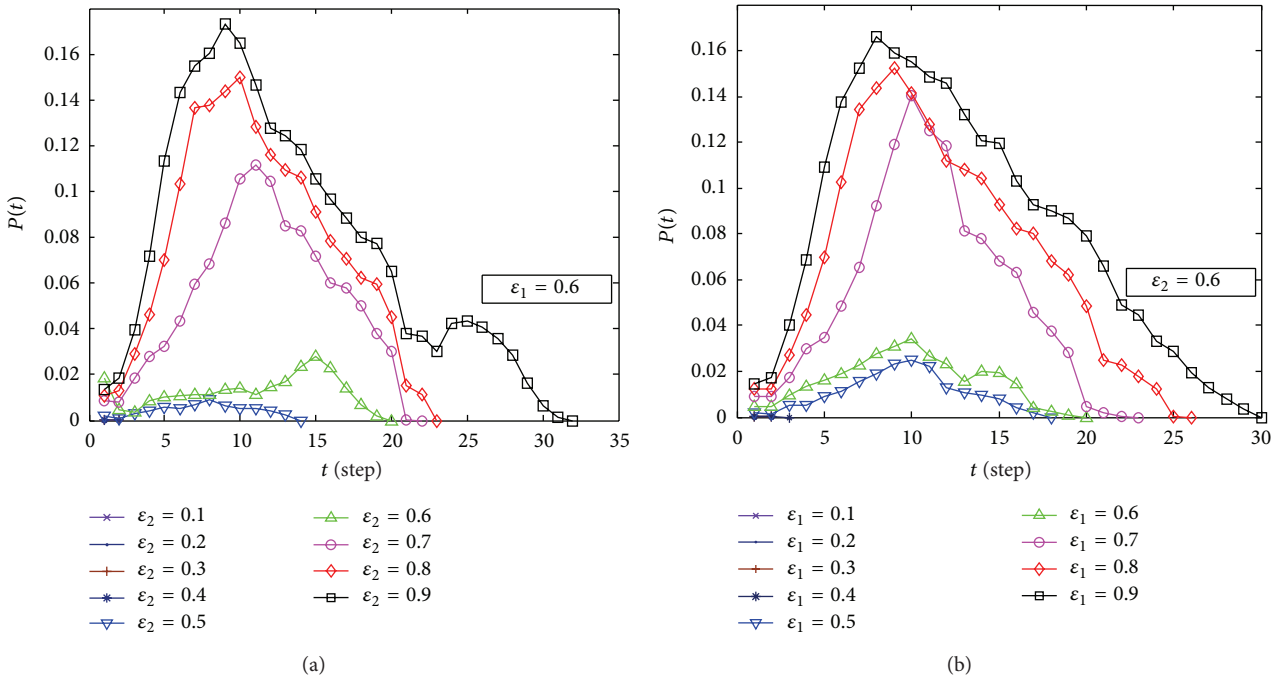


FIGURE 4: $P(t)$ based on different coupled strength.

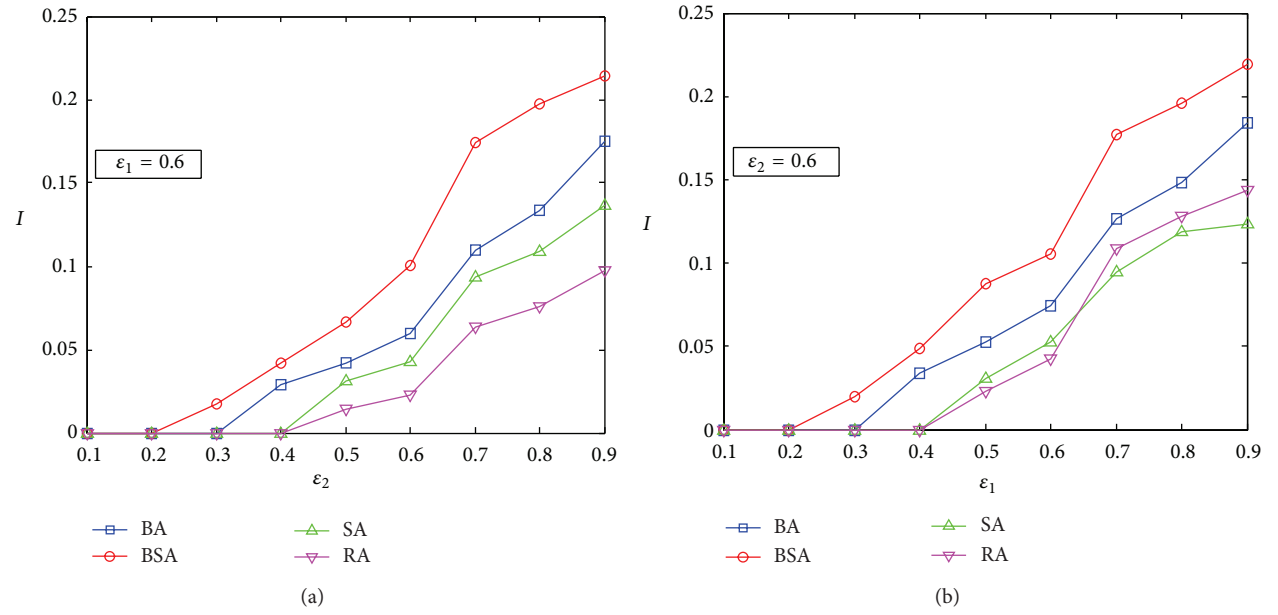


FIGURE 5: I based on different coupled strength and attack strategies.

1 and 2. There are a few failed nodes (less than 1% of N) in the road traffic network when the value of R is smaller than 1.2. The finding is useful that we could prevent the occurrence of large-scale failures by controlling the value of R less than the threshold.

Figure 7 shows that the ratio of total failed nodes I is positively correlated to the external perturbation R and the results are highly dependent on the attack strategies. BSA is the most likely to trigger cascading failures even under small value of external perturbation R.

3.4. Different Number of Road Segments Being Attacked. To confirm the number of nodes being attacked causing large-scale cascading failure, the simulation based on different number of road segments is conducted. Figure 8 shows $P(t)$ based on different value of n (percentage of segments being

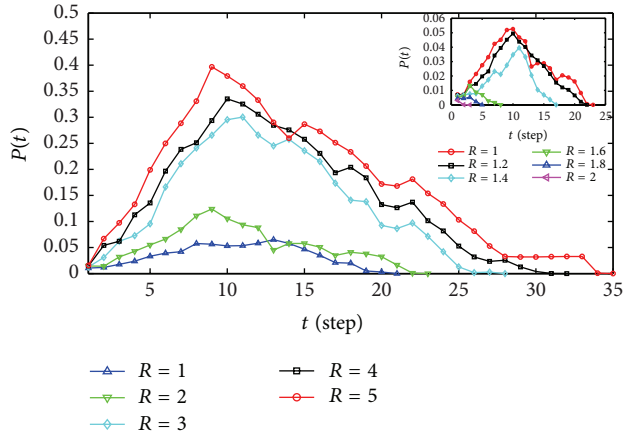


Figure 6: $P(t)$ based on different value of R.

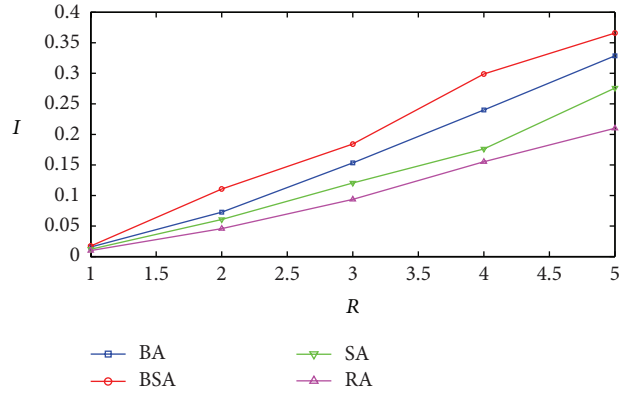


Figure 7: I based on different external perturbation R and attack strategies.

attack) and different attack strategies. Figure 9 indicates n against the proportion of total failed nodes I with different attack strategies. The simulation parameters are set to be $\varepsilon_1 = \varepsilon_2 = 0.6$ and $R = 1.5$.

In Figure 8, for the same n value, the destruction impacts under four kinds of attack strategies are different. This is consistent with the previous analysis. However, in this section's numerical simulation, the whole network might totally become failed, which means that the size of cascades is equal to the size N of the network ($I = 1$). Figure 8 clearly shows that the network can be restored when the number of nodes being attacked is small, while as n increases to a certain value, all of the nodes in the network will be ineffective and difficult to recover. The certain value is the critical value n_c. We should pay more attention to the critical value n_c to avoid the devastating failure. In our simulations, for the RA strategy (i.e., Figure 8(d)), we find that $n_c = 4.5\%$. In Figure 8(b), the n_c for BSA is 3%. The values of n_c under different attack strategies are various and the descending order is as follows: RA, SA, BA, and BSA. This implies that the large-scale failure most likely happens under BSA.

According to Figure 9, the more the nodes being attacked are, the larger the size of cascading failure will be. The simulations also express that, for the same n, the scale of the network cascading failures under BSA is the largest. n_c for each attack strategy can be seen more obviously in Figure 9. For the road traffic network of Liuliqiao area, all values of n_c under four attacks are smaller than 5%, which illustrates that out-road traffic network's invulnerability is low. Therefore, we should take measures to ensure the reliability of the network and control the number of nodes being attacked to be less than the critical value n_c.

4. Conclusion

This paper investigated the cascading failures based on the improved CML model. The improvements of CML model depended on the study of particular road traffic network properties, which are the aeolotropism of road traffic network topology and road congestion dissipation in traffic flow. With a real urban road traffic network in Beijing, the cascading failures are tested using different attack strategies, coupling

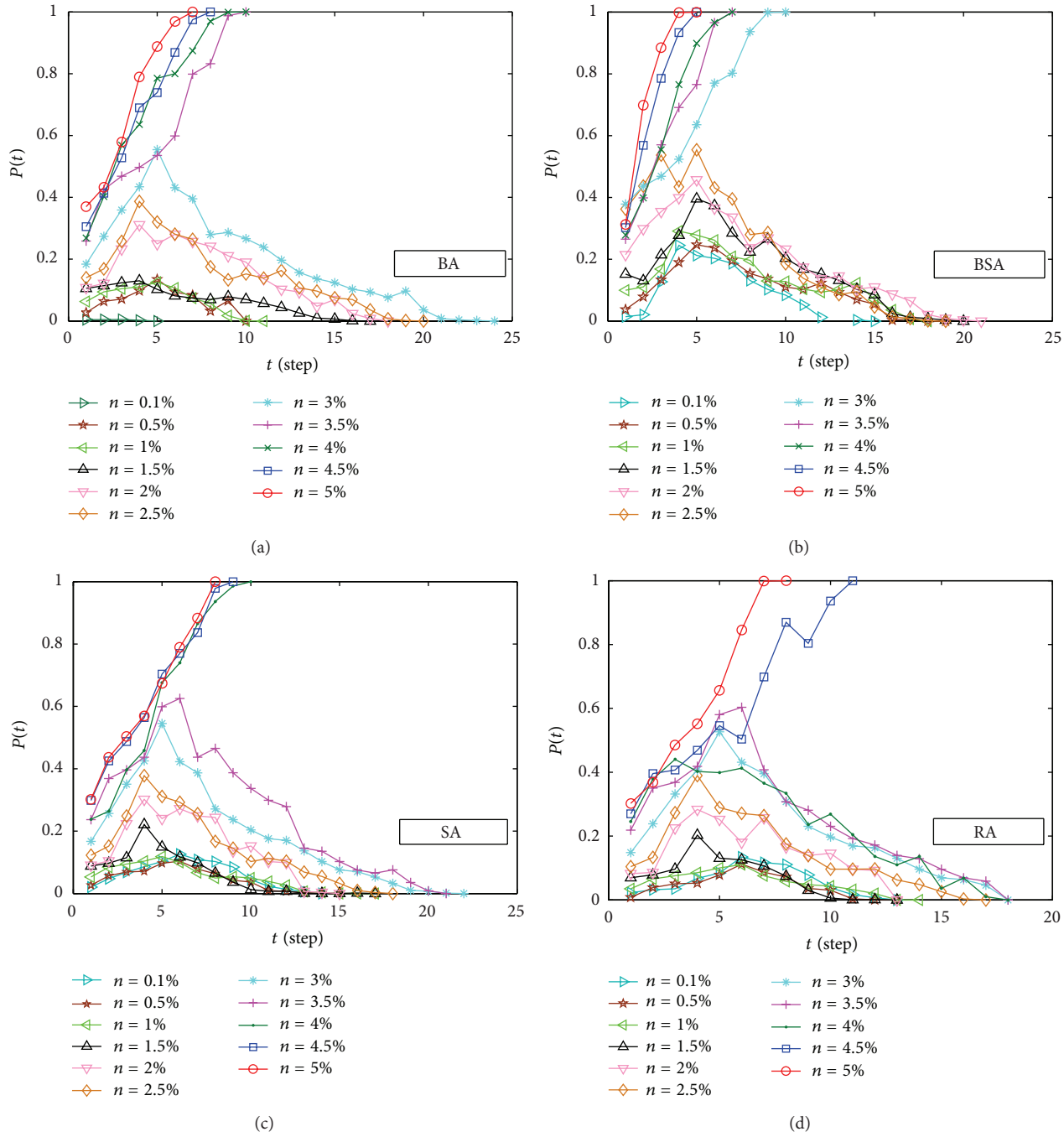


FIGURE 8: $P(t)$ based on different value of n.

strengths, external perturbations, and attacked road segment numbers; we found the following: (1) the aeolotropism and congestion dissipation for the road traffic network topology should be considered; (2) BSA leads to the largest number of failed nodes, of which the propagation rate of failure is the fastest and the failure recovery time is also the longest; (3) as the value of the coupling strength increases, the scale of the network cascading failure increases, and the scale of cascading failures is highly dependent on different attacks; (4) only when the value of external perturbation R is larger than the corresponding threshold of R_c, the large-scale cascading failures would occur, and the number of failed nodes and failure recovery time increase with the increase of R value; (5) the more the nodes being attacked are, the larger the size of cascading failure will be. If the number of nodes being attacked is larger than threshold of n_c, the entire network failure would happen. The road traffic network of Liuliqiao area's invulnerability is very low because the values of n_c for

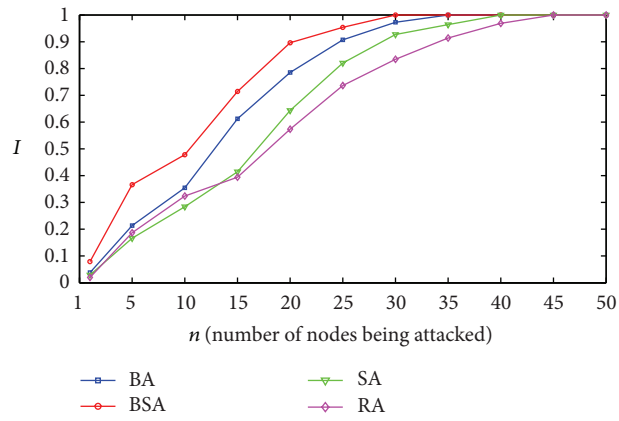


FIGURE 9: I based on different attack strategies.

different attacks are very small. The above findings might be useful in avoiding or alleviating large-scale failures of road traffic network.

Conflict of Interests

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

Acknowledgment

This study is supported by the National Basic Research Program of China (2012CB725404).

References

[1] R. Guimerà, A. Arenas, A. Díaz-Guilera, and F. Giralt, "Dynamical properties of model communication networks," *Physical Review E*, vol. 66, no. 2, Article ID 026704, 2002.

[2] P. Crucitti, V. Latora, and M. Marchiori, "Model for cascading failures in complex networks," *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, vol. 69, no. 4, Article ID 045104, 2004.

[3] R. Kinney, P. Crucitti, R. Albert, and V. Latora, "Modeling cascading failures in the North American power grid," *The European Physical Journal B*, vol. 46, no. 1, pp. 101–107, 2005.

[4] X. Fang, Q. Yang, and W. Yan, "Modeling and analysis of cascading failure in directed complex networks," *Safety Science*, vol. 65, pp. 1–9, 2014.

[5] J. Chen, J. S. Thorp, and I. Dobson, "Cascading dynamics and mitigation assessment in power system disturbances via a hidden failure model," *International Journal of Electrical Power & Energy Systems*, vol. 27, no. 4, pp. 318–326, 2005.

[6] Y. Koç, M. Warnier, R. E. Kooij, and F. M. T. Brazier, "An entropy-based metric to quantify the robustness of power grids against cascading failures," *Safety Science*, vol. 59, pp. 126–134, 2013.

[7] J. J. Wu, H. J. Sun, and Z. Y. Gao, "Cascading failures on weighted urban traffic equilibrium networks," *Physica A: Statistical Mechanics and its Applications*, vol. 386, no. 1, pp. 407–413, 2007.

[8] Z. Su, L. Li, H. Peng, J. Kurths, J. Xiao, and Y. Yang, "Robustness of interrelated traffic networks to cascading failures," *Scientific Reports*, vol. 4, article 5413, 2014.

[9] L. Daqing, J. Yinan, K. Rui, and S. Havlin, "Spatial correlation analysis of cascading failures: congestions and Blackouts," *Scientific Reports*, vol. 4, article 5381, 2014.

[10] J.-F. Zheng, Z.-Y. Gao, and X.-M. Zhao, "Modeling cascading failures in congested complex networks," *Physica A: Statistical Mechanics and its Applications*, vol. 385, no. 2, pp. 700–706, 2007.

[11] Z. Liu, M.-B. Hu, R. Jiang, W.-X. Wang, and Q.-S. Wu, "Method to enhance traffic capacity for scale-free networks," *Physical Review E*, vol. 76, no. 3, Article ID 037101, 2007.

[12] J. Xu and X. F. Wang, "Cascading failures in scale-free coupled map lattices," *Physica A: Statistical Mechanics and its Applications*, vol. 349, no. 3-4, pp. 685–692, 2005.

[13] D. J. Watts, "A simple model of global cascades on random networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 9, pp. 5766–5771, 2002.

[14] A. E. Motter and Y.-C. Lai, "Cascade-based attacks on complex networks," *Physical Review E*, vol. 66, no. 6, Article ID 065102, 2002.

[15] J. Wang, C. Jiang, and J. Qian, "Robustness of Internet under targeted attack: a cascading failure perspective," *Journal of Network and Computer Applications*, vol. 40, no. 1, pp. 97–104, 2014.

[16] Y. Duan and F. Lu, "Robustness of city road networks at different granularities," *Physica A: Statistical Mechanics and Its Applications*, vol. 411, pp. 21–34, 2014.

[17] A. G. Smart, L. A. N. Amaral, and J. M. Ottino, "Cascading failure and robustness in metabolic networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 36, pp. 13223–13228, 2008.

[18] P. Li, B.-H. Wang, H. Sun, P. Gao, and T. Zhou, "A limited resource model of fault-tolerant capability against cascading failure of complex network," *The European Physical Journal B: Condensed Matter and Complex Systems*, vol. 62, no. 1, pp. 101–104, 2008.

[19] C. Di, G. Zi-You, and Z. Jian-Feng, "Tolerance of edge cascades with coupled map lattices methods," *Chinese Physics B*, vol. 18, no. 3, pp. 992–996, 2009.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 146070, 10 pages
http://dx.doi.org/10.1155/2015/146070

Research Article

A Cooperative Q-Learning Path Planning Algorithm for Origin-Destination Pairs in Urban Road Networks

Xiaoyong Zhang, Heng Li, Jun Peng, and Weirong Liu

School of Information Science and Engineering, Central South University, 22 South Shaoshan Road, Changsha 410075, China

Correspondence should be addressed to Jun Peng; pengj@csu.edu.cn

Received 25 May 2015; Accepted 21 September 2015

Academic Editor: Chronis Stamatiadis

As an important part of intelligent transportation systems, path planning algorithms have been extensively studied in the literature. Most of existing studies are focused on the global optimization of paths to find the optimal path between Origin-Destination (OD) pairs. However, in urban road networks, the optimal path may not be always available when some unknown emergent events occur on the path. Thus a more practical method is to calculate several suboptimal paths instead of finding only one optimal path. In this paper, a cooperative Q-learning path planning algorithm is proposed to seek a suboptimal multipath set for OD pairs in urban road networks. The road model is abstracted to the form that Q-learning can be applied firstly. Then the gray prediction algorithm is combined into Q-learning to find the suboptimal paths with reliable constraints. Simulation results are provided to show the effectiveness of the proposed algorithm.

1. Introduction

Recent years have seen a growing interest in the study of route-guidance system in intelligent transportation systems, due to its advantages in reducing traffic congestion and CO_2 emissions, minimizing travel time, and conserving energy [1]. More and more vehicle manufacturers have installed the route-guidance system into their products to assist the drivers' travel.

As an essential part of the route-guidance system, path planning is usually modeled as the shortest-path problem in graph theory [2–8]. When a vehicle departs from the origin and travels to its destination, the map it is involved in can be abstracted as a graph by treating streets as edges and intersections as nodes. The weight of an edge represents the average travel time over the street, which may dynamically change when traffic flows fluctuate. For the graph of a static network, the most efficient one-to-one node shortest path algorithm is Dijkstra's algorithm [2]. When the dynamic graph is considered, A^* algorithm might be a better choice to solve the Origin-Destination shortest path problem [3]. A^* algorithm estimates the minimum distance between the destination and a node to determine whether the node is on the optimal route.

However, even if the generated route is the shortest one, it may not be always available because of traffic emergencies such as sudden accidents. So it may be more practical to provide a number of candidate paths rather than just one optimal path. Lee revealed that finding multiple paths instead of one is a good way to avoid the path overload phenomenon [4]. This optimal path will even accelerate the deterioration of the road network when the overload phenomenon occurs.

Traditionally, the alternative paths could be calculated by two categories of algorithms in graph theory, namely, the k-shortest path algorithm proposed by Eppstein [5] and Jiménez and Marzal [6] and the totally disjoint path algorithms proposed by Dinic [7] and Torrieri [8]. These so-called alternative path planning methods typically find the optimal path using Dijkstra's algorithm first. Then the candidate path set can be generated by applying link weight increment methods. These algorithms seek for the next suboptimal path iteratively until the generated alternative path satisfies some given constraints.

However, the generated way of alternative paths of these algorithms unavoidably lengthens the response time, especially when the network is huge and the traffic load is crowded and time varying. These algorithms need to adjust the link weight of the generated optimal path and then recalculate

the suboptimal paths using Dijkstra's algorithm repeatedly, thus leading to heavy computation burden. In addition, these algorithms generally concern path planning of just one vehicle, while it is essential to simultaneously consider all vehicles' path in practical city road networks.

With the development of intelligent science, some researchers have focused on path planning using reinforcement learning in guidance systems. Reinforcement learning is a category of machine learning algorithms, in which a group of agents can decide how to behave according to their interaction with environment and achieve an optimal objective [9]. Recently, multiagent reinforcement learning has been proposed to find the best and shortest path between the origin and the destination. Some studies treat each intersection as one agent, which needs a large amount of information interaction between traffic intersections to find the optimal path [10] while more studies cast each intersection as the state and take each link as the action in the model, which could deal with the road networks on the whole [11, 12]. Thus our proposed Q-learning adopts the latter method, treating the intersections as states in the model.

With Q-learning, the computational complexity of path planning algorithm could be reduced significantly and the efficiency would be improved. While most existing Q-learning algorithms are designed to solve the optimal path planning for just one OD pair in the literature, the proposed Q-learning algorithm in this paper aims to seek multiple paths for different OD pairs simultaneously. By choosing the suboptimum Q-value of every intersection, it is convenient to provide some alternative paths rather than seeking every alternative route incrementally. This paper makes the following contributions in particular.

First, the multipath set is found for different OD pairs simultaneously using Q-learning. Compared with other multipath algorithms, the proposed algorithm significantly reduces the computational complexity.

Second, some reliability constraints are introduced to choose suboptimal paths in Q-learning. It would not be appropriate to increase the dimension of the multipath set without considering the overall reliability, which ensures that at least one alternative path is available at all times [13].

Third, the FNN prediction is combined with Q-learning. In order to improve the real-time capability, short-term traffic prediction is essential [14, 15]. This paper adopts the FNN prediction mechanism in the Q-learning scheme to predict the traffic condition, with which the reward of the action can be computed in advance.

Fourth, the multiagent cooperative mechanism is applied to path planning. The cooperative mechanism introduced in Q-learning coordinates the actions and strategies among agents with different OD pairs for long-time benefits.

In this paper, we propose a new multiagent reinforcement learning (MARL) algorithm using Q-learning with prediction for multipath planning for OD Pairs in the road navigation system. Compared with traditional multipath algorithms, it reduces computational complexity and improves the efficiency of vehicles' guidance with traffic prediction. The scheme could improve the overall performance of urban traffic networks and balance the traffic flow.

Figure 1: Eastern Town of Changsha.

The rest of the paper is organized as follows. Section 2 describes the model of road networks. The Q-learning based cooperative multiagent multipath planning algorithm for OD pairs is proposed in Section 3. The simulation results are shown and analyzed in Section 4. The conclusion is drawn in Section 5.

2. Model of Road Networks

2.1. Graph Abstraction of Road Networks. For urban areas, two important elements of traffic guidance are intersections and roads. During the process of modeling, the intersection can be seen as the node and the road can be seen as the edge connecting two nodes. The weight on the line stands for the traffic condition of the road, and the arrows mean the allowable direction of forward motion for vehicles. By this abstracting, a graph $G = (S, E)$ with a nonempty finite set of intersections (nodes) $S = \{s_1, s_1, \ldots, s_N\}$ and a set of roads $E \subseteq S \times S$ can be used to describe the road map. Once we have the model and the route algorithm, we can find the needed optimal route.

For instance, Eastern Town of Changsha in China could be taken as an example, whose map is shown in Figure 1. The abstract graph model of Figure 1 is showed in Figure 2. S_i stands for each intersection that is taken as one state in reinforcement learning. S_i has three or four directions to neighbor intersections, including the loop direction that returns to S_i. For example, if one vehicle at intersection S_1 drives west, it will return to S_1. The setting is convenient to model the complex road networks.

The weight of each direction will contain two elements: traffic condition (w) and the distance from the destination S_3 (r). These two elements will be illustrated in the next section.

2.2. Model Using Reinforcement Learning. To address the model more clearly, it is necessary to provide the background on reinforcement learning (RL). Reinforcement learning is a kind of multiagent intelligent algorithms, in which agents select the best actions to maximize the cumulative reward by interacting with the environment. The RL agent interacts with its environment over a sequence of discrete time steps to pick out the optimal actions. The agent in this paper is a processing

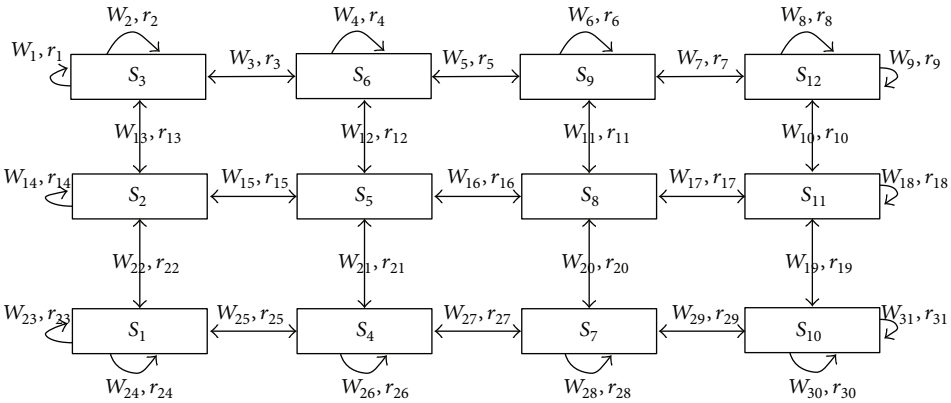


Figure 2: Road model of Figure 1.

center that deals with the path planning from one origin to one destination.

The underlying concept of RL is the finite Markov Decision Process (MDP), which is defined by the tuple $\langle S, A, \phi, \rho \rangle$, where S is a finite set of environment states, A is a finite set of agent actions, ϕ: $S \times A \times S \rightarrow [0, 1]$ is the state transition probability function, and ρ: $S \times A \times S \rightarrow R$ is the reward function. The MDP models an agent's action in an environment where it learns (through prior experiences and short-term rewards) the best control policy (a mapping of states to actions) that maximizes the expected discounted long-term reward. This mapping can be stochastic π: $S \times A \rightarrow [0, 1]$ or deterministic π: $S \times A \rightarrow 0 \parallel 1$.

For deterministic state transition models, the transition probability function ϕ reduces to ϕ: $S \times A \times S \rightarrow 0 \parallel 1$ and, as a result, the reward is completely determined by the current state and the action; that is, ρ: $S \times A \rightarrow R$. The state-action pair's value is called the Q-value and the function that determines the Q-value is called the Q-function. An agent can find the optimal control policy by approximating its Q-values using prior estimates iteratively, the short-term reward $r = \rho(s, a) \in R$, and discounted future reward. This model-free successive approximation technique is called Q-learning. One way to satisfy this criterion is adopting ε greedy approach where a random action is performed with probability ε and the current knowledge is exploited with probability $1 - \varepsilon$.

This paper denotes the link from intersection i to intersection j as a paired index of ij. And, accordingly, the reward of link ij is defined as r_{ij}; the mean travel time of link ij is defined as t_{ij}; the distance between the intersection i and the intersection j is defined as d_{ij}. To cast the path planning problem of the road network as a RL problem, we identify an individual agent i's states (S_i), available actions (A_{S_i}), and reward (r_{ij}).

First, the two weights of each direction should be illustrated. Traffic condition can be described by the mean travel time t_{ij} [13, 16–19]. The distance from the destination that points out the close degree of destination is needed by the idea of reinforcement learning.

Mean travel time t_{ij} could be obtained by probe vehicles such as taxis. Equipped with GPS sensors, probe vehicles can collect data on position, speed, and direction, store them, and send reports at regular intervals of time. By analyzing these data, the mean travel time of each link could be calculated. For instance, Hellinga and Fu advanced the method of probe based arterial link travel time estimation [16]. Tomio et al. used probe vehicle data to identify routes and predict travel times [17]. The distance d_{ij} stands for the Euclidean distance between the intersection i and the intersection j.

States. States of each interaction are form the basis for making choices. We model one intersection as one state S_i, which can easily indicate the location of vehicles.

Actions. Actions are the choices made by the agent. In one state, the vehicle will have four actions A_{S_i}: turning left, turning right, going straight, and turning round. For example, $A_{S_i} = \{\text{up}, \text{down}, \text{right}, \text{left}\}$.

Rewards. Rewards are the basis for evaluating choices. We model the link weight of each link as the rewards, which will contain two elements: the reward of mean travel time (r_{ij}^t) and the reward of the distance from the destination (r_{ij}^d).

For reinforcement learning, everything inside the agent should be completely known and controllable by the agent; everything outside is incompletely controllable but may be or may not be completely known. A policy is a stochastic rule by which the agent selects actions as a function of states. The agent's objective is to maximize the amount of rewards it receives over time. The return R_{ij} is the function of future rewards that the agent seeks to maximize:

$$R_{ij} = r_{i^1 j^1} + \beta r_{i^2 j^2} + \beta^2 r_{i^3 j^3} + \cdots = \sum_{k=0}^{N} \beta^k r_{i^{k+1} j^{k+1}}, \quad (1)$$

where β ($0 < \beta < 1$) is called the discount rate. A whole path has N links.

3. Cooperative Multipath Planning for OD Pairs

In this section, we propose a cooperative multipath planning method for OD pairs. In the proposed method, the agent is a processing center that deals with the path planning from

one origin to one destination. In practical applications, there are typically many paths that are planned synchronously. By introducing the concept of multiagent systems, the path planning problem can be modeled to a form such that reinforcement learning is applicable. Then we introduce a multiagent reinforcement learning mechanism to optimize the Q-value of different paths.

3.1. Reward of Traffic Flow Using FNN Prediction. The travel time of each link ij is defined as t_{ij}. For given OD pairs (O_i, D_j), a set of binary variables are given to represent the selection of links on a path (i.e., a path solution). Thus, for a given path P, its travel time T could be calculated by [20]. Consider

$$T = \sum_{ij \in P} t_{ij}. \tag{2}$$

After getting the history data of travel time t_{ij} of each link ij, we can use some prediction algorithms to compute the future data for improving the real-time characteristics of the guidance system. In this paper, T-S FNN (fuzzy neural network) is introduced to predict the future travel time. T-S FNN is a highly adaptive fuzzy system that can automatically update the membership function of fuzzy subset [13, 21]. This FNN is defined by the following if-then rules. In the case of R^i, fuzzy reasoning is as follows:

$$\begin{aligned} R^i&\text{: If } t_1 \text{ is } A_1^i,\ t_2 \text{ is } A_2^i, \ldots,\ t_k \text{ is } A_k^i, \\ &\text{then } \widehat{t}^i = p_0^i + p_1^i t_1 + \cdots + p_k^i t_k, \end{aligned} \tag{3}$$

where A_j^i is the fuzzy set of the fuzzy system and p_j^i is the parameters of the fuzzy system, $(j = 1, 2, \ldots, k)$. $\widehat{t}^i$ is the predictive output based on the ith fuzzy rule. The input part is fuzzy, while the output part is deterministic, which is the linear combination of the input.

Suppose that the input of t_{ij} is $[t_1, t_2, \ldots, t_k]$, and then the membership degree of each input variable t_j can be computed by fuzzy rules:

$$\mu_{A_j^i} = \exp\left(-\frac{\left(t_j - c_j^i\right)^2}{b_j^i}\right) \qquad j = 1, 2, \ldots, k;\ i = 1, 2, \ldots, n, \tag{4}$$

where c_j^i and b_j^i are the center and width of the membership function, respectively, k is the number of the inputs, and n is the number of fuzzy subsets.

All membership degrees are computed by the fuzzy operators:

$$w^i = \mu_{A_1^i}(t_1) * \mu_{A_2^i}(t_2) * \cdots * \mu_{A_k^i}(t_k) \qquad i = 1, 2, \ldots, n. \tag{5}$$

The output of this fuzzy model is computed by the above results:

$$\widehat{t} = \frac{\sum_{i=1}^n w^i \left(p_0^i + p_1^i t_1 + \cdots + p_k^i t_k\right)}{\sum_{i=1}^n w^i}. \tag{6}$$

FNN is divided into four layers: input layer, fuzzy layer, fuzzy rules calculating layer, and output layer. The input layer is connected with the input vector t_k, so the number of nodes is equal to the dimensions of input vectors. Fuzzy Layer obtains fuzzy membership values μ using membership functions and fuzzy input values (4). Fuzzy rules' calculating layer gets w by the fuzzy multiply equation (5). The output layer uses (6) to calculate the fuzzy neural network outputs.

The parameters of FNN are updated by the following equations. The error e between the desired output and actual output is defined as follows:

$$e = \frac{1}{2}\left(\widehat{t} - t\right)^2, \tag{7}$$

where $\widehat{t}$ is the desired output, t is the actual output, and $\widehat{t}$ is the predictive mean travel time.

The parameters p_j^i of FNN are updated by

$$\begin{aligned} p_j^i(k) &= p_j^i(k-1) - \alpha \frac{\partial e}{\partial p_j^i} \\ \frac{\partial e}{\partial p_j^i} &= \frac{\left(\widehat{t} - t\right) w^i}{\sum_{i=1}^n w^i \cdot t_j}, \end{aligned} \tag{8}$$

where α is the learning rate, t_j is the input, and w^i is the weight computed by (5).

The center c_j^i and the width b_j^i of membership function are updated by (9) and (10), respectively. Consider

$$c_j^i(k) = c_j^i(k-1) - \beta \frac{\partial e}{\partial c_j^i}, \tag{9}$$

$$b_j^i(k) = b_j^i(k-1) - \beta \frac{\partial e}{\partial b_j^i}. \tag{10}$$

When one finds multiple paths using Q-learning, it is important to determine how every action is assessed. A link can be simply described as unblocked, normal, and busy. To simplify the learning process, the precise specific flow density is neglected because finding multipath is the final goal. Thus some links having a very small difference can be regarded as the same optimal choice.

This paper gives discrete weights in terms of the traffic condition and the distance from the destination to simplify the Q-learning reward's iterative calculations. For discretion of the mean travel time t_{ij}, we first gather the maximum travel time of current intersection as "−1"; then we get the difference value between the current maximum travel time and the minimum travel time. So the travel time can be graded into "0," "1," "2," and "3" as the reward r_{ij}^t. The distance from the destination (r_{ij}^d) can be graded into two levels: "1" and "0"; the nearest neighbor intersections are "1," while the others are "0."

So r_{ij} is deduced in the following equation, where τ $(0 \le \tau \le 1)$ is the scaling factor between r_{ij}^d and r_{ij}^t:

$$r_{ij} = \tau r_{ij}^t + (1 - \tau) r_{ij}^d. \tag{11}$$

Initialize $Q(s, a)$ arbitrarily
Repeat (for each episode):
  Initialize s
  Repeat (for each step of episode):
    Choose a direction a from s using policy derived from Q (e.g., ε-greedy)
    Take action a, observe r, s'

$$Q_{n+1}^{i}(s,a) = (1-\alpha_n) Q_n^i(s,a) + \alpha_n^i \left[r + \gamma^i \max_{a' \in A} Q_n^i(s', a') \right]$$

Until that s is the terminal state.

Algorithm 1: The policy iteration algorithm.

3.2. Cooperative Multiagent Multipath Planning Algorithm. If each agent acts independently without cooperation, the Q-learning procedure at node i can be written as

$$Q_{n+1}^{i}(s,a) = \left(1-\alpha_n^i\right) Q_n^i(s,a) + \alpha_n^i \left[r + \gamma^i \max_{a' \in A} Q_n^i\left(s', a'\right) \right], \tag{12}$$

where $\alpha_n \in (0, 1]$ is the learning factor and $\gamma \in [0, 1)$ is the discount factor.

RL have been well developed for discrete-time systems to solve the optimal problem online by using adaptive learning techniques to determine the optimal value function.

An iterative solution technique is given by Algorithm 1.

An agent is a processing center that deals with the path planning from one origin to one destination. The above algorithm focuses on one single agent, while path planning agents with different destinations will observably impact on each other. Thus we propose a cooperative multiagent reinforcement learning (MARL) multipath path planning method, in which all Q-values of different path plans for every intersection are considered, and the maximum value is chosen to ensure that all path planning is optimized under the consideration of each other. It is worth mentioning that, in the proposed algorithm, the decision-making process is assumed ideal, and then the waiting time at the intersections is thus ignored.

In this approach, the Q-value estimated at each autonomous agent is updated based on the individual rewards as well as on information obtained from other agents in the neighborhood. "The neighborhood" here refers to a group of agents that own different destinations. Every agent exchanges the largest Q-value that is associated with its current state with every other agent in its neighborhood. The value iteration procedure at agent i for the state-action pair (s_i, a_i) can be summarized as

$$Q_{n+1}^{i}\left(s^i, a^i\right) = \left(1-\alpha_n^i\right) Q_n^i\left(s^i, a^i\right) + \alpha_n^i \left[r^i\left(s^i, a^i\right) + \gamma^i \sum_{j \in N^i} w(i,j) \cdot \max_{a^j \in A^j} Q_n^i\left(s^j, a^j\right) \right], \tag{13}$$

where $w(i, j)$ is the weight to reflect the effect of agent j to agent i and N_i refers to the set of neighboring agents of i. The simplest strategy for computing the weights $w(i, j)$ is to just consider the total number of agents in the neighborhood; that is, $w(i, j) = 1/|N_i|$, in which case $\sum_j w(i, j) = 1$. It is possible to adopt more complex strategies to take into account the different effects of the different neighbors. When the additional information obtained from agents was incorporated into the value iteration procedure, each agent can ensure that the agent's strategies are decided based on all its neighbors' actions.

3.3. Constraint Conditions of Multipath Set. By the policy iteration algorithm in Section 3.2, we can derive the Q-value table for multipath planning. By comparing the four Q-values of one intersection, we can easily find which action is the best. Then the optimal path is easily obtained.

Although the obtained optimal path is the fastest one, it may encounter traffic emergencies such as a sudden accident, resulting in unavailability of the optimal planning path for the running vehicle. So it is essential to provide several candidate paths rather than just one path, avoiding the deterioration of the road network environment when one guided vehicle is well popular.

In most cases, there may be several actions forming a best action set. Then we can find the multipath by choosing the best actions. However, when the road network is huge, it is difficult to find multipath because there exists only one optimal action for all intersections in most cases. So we should find the suboptimum action that satisfies the following constraints.

First, we introduce B_i as the average Q-value of one intersection. Then we compute the average difference between each $Q(s, a)$ and the average Q-value of one intersection. Furthermore, we can get τ value, the average difference for all states:

$$B_i = \frac{\sum_{a_j \in A_{S_i}} Q\left(S_i, a_j\right)}{4}, \qquad \tau = \partial * \frac{\sum_{S_i \in S} \sum_{a \in A_{S_i}} \left| B_i - Q\left(S_i, a\right) \right|}{4N}. \tag{14}$$

When vehicle arrives at one intersection, it has to make the choice about which way to go by computing the difference

Table 1: Q value table of $4 * 4$ road network with destination Intersection 11.

$Q(s, a)$	Value
$Q(1, 2)$	−25.2838
$Q(1, 5)$	−25.2838
$Q(2, 6)$	−21.8556
$Q(2, 3)$	−22.7127
$Q(10, 11)$	0
$Q(6, 2)$	−27.2838
$Q(6, 7)$	−16.4274
$Q(6, 10)$	−16.4274
$Q(7, 11)$	0
$Q(7, 3)$	−22.7127

between this $Q(s, a)$ and the average Q-value of this intersection:

$$\left|Q\left(S_i, a\right) - B_i\right| \leq \tau. \tag{15}$$

Once the Q-value table has been calculated, we can select ∂ ($\partial \in (0, 1)$) value and compute the corresponding τ value. So we must solve the problem of how to ensure the ∂ value. When ∂ is closer to 1, τ is larger. And there are more paths that could be taken as candidates. When ∂ is closer to 0, τ is smaller, resulting in fewer candidate paths. While more candidate paths sometimes are not stable, the reliability of a path set S should be taken into account.

The reliability of a path can be defined as the probability of not encountering an abnormal delay during a trip along the path, which can be estimated by the reliability of a series of links of the path. Under dynamic conditions, some or all candidate paths may fail together, resulting in a joint failure. In this situation, the reliability of the path set shown in (10) will be weakened. Thus, in the calculation of the candidate paths, it is important to reduce the chance of joint failure of candidate paths [18, 19]:

$$\Phi s = \sum_{i=1}^{M} \left[\prod_{j_{1i}=1}^{J_{1i}} r_{j_{1i}} \prod_{j_{2i}=1}^{J_{2i}} \left(1 - r_{j_{2i}}\right) \right], \tag{16}$$

where Φs stands for the reliability of the path set S, M is the number of disjoint subpaths in the subpath set S, j_{1i} is the jth link in a normal state on the ith disjoint subpath, $r_{j_{1i}}$ is the reliability of the jth link on the ith disjoint subpath, J_{1i} is the number of links in a normal state on the ith disjoint subpath, j_{2i} is the jth failed link on the ith disjoint subpath, $r_{j_{2i}}$ is the reliability of the jth link in a failed state on the ith disjoint subpath, and, finally, J_{2i} is the number of links in a failed state on the ith disjoint subpath.

Generally, the higher the reliability of the candidate path set, the less the chance that all candidate routes will be unacceptable during one trip. Given Φs value (such as 0.9), we can choose ∂ value. Then we find the stable multipath paths.

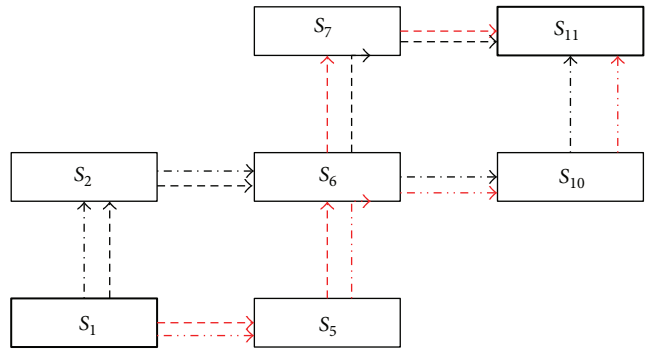


Figure 3: Multipath planning from Intersection 1 to Intersection 11.

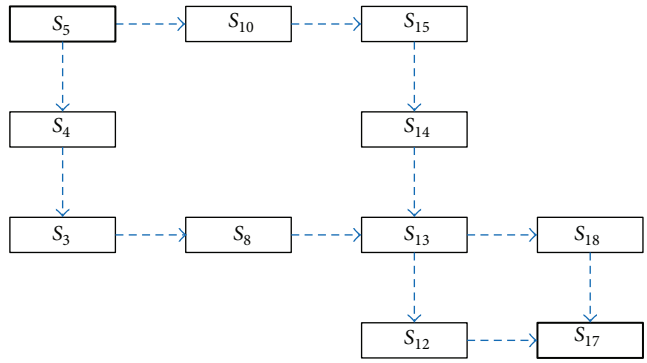


Figure 4: Multipath planning from Intersection 5 to Intersection 17.

4. Simulation Results and Analysis

To test the proposed method in different traffic environments, simulations have been conducted in randomly generated grid road networks. In the simulation scenarios, there are 6 to 25 nodes with 10 to 50 edges. The general network discussed in this paper is shown in Figure 1. In the graph, a node is represented by a box with the node's number shown in it. The start and destination nodes are represented by the rectangles with thicker outlines. The mean travel time and the distance from the destination of all links are set by a random matrix.

4.1. Path Planning for Single Agent. The scenario is shown in Figure 1. The single agent only computes the path towards one single destination. The main objective of Figure 3 and Table 1 is to calculate the shortest path from Intersection 1 to Intersection 11. This is drawn (shortest path) by the use of MATLAB software. Therefore four candidate paths exist. When the car arrives at Intersection 1, it can choose either Intersection 2 or Intersection 5. When the car arrives at Intersection 6, it can choose either Intersection 7 or Intersection 10. In this case, ∂ is equal to 0.1, and Φs is equal to 0.91 by computing.

Given the matrix of Q-value, we can quickly get the optimal multipath. Compared with the other multipath algorithm, our proposed algorithm only uses the information of the whole road networks once.

Table 2: Q value table of $5 * 5$ road network with destination Intersection 17.

$Q(s, a)$	Value
$Q(5, 10)$	−112.7293
$Q(5, 4)$	−112.9003
$Q(4, 3)$	−106.3563
$Q(4, 9)$	−107.0302
$Q(3, 8)$	−97.5428
$Q(3, 2)$	−100.6263
$Q(8, 13)$	−79.9226
$Q(10, 15)$	−105.3436
$Q(15, 14)$	−96.4487
$Q(15, 20)$	−98.3531
$Q(14, 13)$	−79.9226
$Q(13, 12)$	−59.5865
$Q(13, 18)$	−59.1124
$Q(12, 17)$	0

Table 3: Q value table of $5 * 5$ road network with destination Intersection 14.

$Q(s, a)$	Value
$Q(1, 2)$	−87.6625
$Q(1, 6)$	−93.7241
$Q(2, 3)$	−75.1119
$Q(2, 6)$	−84.9330
$Q(3, 4)$	−62.8019
$Q(3, 8)$	−67.8715
$Q(8, 13)$	−57.0315
$Q(8, 9)$	−46.9100
$Q(13, 14)$	0
$Q(4, 9)$	−46.9100
$Q(4, 5)$	−60.3841
$Q(9, 10)$	−50.4666
$Q(9, 8)$	−67.8715
$Q(9, 14)$	0

Table 4: Cooperative Q value table of $5 * 5$ road network with destination Intersection 17.

$Q(s, a)$	Value
$Q(5, 10)$	−112.729
$Q(5, 4)$	−112.9003
$Q(4, 3)$	−106.3563
$Q(4, 9)$	−107.0302
$Q(3, 8)$	−97.5428
$Q(3, 2)$	−100.6263
$Q(8, 13)$	−79.9226
$Q(10, 15)$	−105.3436
$Q(15, 14)$	−96.4487
$Q(15, 20)$	−98.3531
$Q(14, 13)$	−79.9226
$Q(13, 12)$	−82.4741
$Q(13, 18)$	−69.2806
$Q(18, 17)$	0

Table 5: Cooperative Q value table of $5 * 5$ road network with destination Intersection 14.

$Q(s, a)$	Value
$Q(1, 2)$	−100.6263
$Q(1, 6)$	−93.7241
$Q(6, 7)$	−87.8625
$Q(6, 11)$	−87.7966
$Q(7, 8)$	−87.5428
$Q(7, 12)$	−92.4741
$Q(8, 13)$	−79.9226
$Q(8, 9)$	−79.0302
$Q(9, 14)$	0
$Q(11, 12)$	−82.4741
$Q(11, 16)$	−94.6919
$Q(12, 13)$	−79.9226
$Q(12, 17)$	−86.1356
$Q(13, 14)$	0

4.2. Cooperative Path Planning for Multiple Agents. This simulation contains two agents. One agent generates the path from Intersection 5 to Intersection 17. The other gives the path from Intersection 1 to Intersection 14. The proposed algorithm introduces the idea of multiagent to lessen the influence that is displayed in Tables 2–5 and Figures 6–9. From the results we can see advantages of MARL.

There is a significant difference between Figures 6 and 7 because Intersection 3 is still busy in Figures 4 and 5 (after cooperation). So after cooperation in Figure 7, all subsequent paths avoid Intersection 3 and 2. Instead, all paths choose Intersection 6 as the next intersection. There is a smaller difference between Figures 4 and 5 as there are only two overlapping intersections between the path from Intersection 5 to 17 and the path from Intersection 1 to 14. Intersection 13 is considered to be the essential intersection, so the algorithm gives up Intersection 12.

4.3. The Comparison of Several Multipath Algorithms. This simulation is made up of 30 nodes with 49 edges in Figure 8. Each circle stands for one intersection S_i. Each edge has three parameters: mean travel time (3 or 7), levels of mean travel time (−1 or −3), and reliability of this edge (0.98 or 0.87).

Under the same simulation environment, we have the following results in Table 6 by Dijkstra's algorithm, k-shortest planning, Q-learning multipath path planning. From the results, we can draw the follow conclusions: Dijkstra's algorithm can give the shortest path with the least time, while the reliability of this path is lower because reliability of each edge is less than 1. The last two algorithms are almost the same. The mean cost of the four paths of k-shortest planning is less than Q-learning multipath path planning, while the reliability of Q-learning multipath path planning is superior to k-shortest planning.

Next, the performance comparison is given for the last two algorithms on the planning time elapsed and the path

Table 6: The comparison results of different algorithms.

Algorithm	Multipath	Costs	Reliability	Time
Dijkstra's algorithm	1-6-7-12-13-14-19-24-25	25	0.57	0.010142
k-shortest path planning	1-6-7-12-13-14-19-24-25	25	0.77	0.104754
	1-6-11-16-21-22-23-24-25	26		
	1-2-7-12-13-14-19-24-15	27		
	1-6-7-12-13-14-19-20-25	27		
Q learning multipath path planning	1-6-11-16-21-22-23-24-25	26	0.95	0.110398
	1-6-11-16-17-18-19-20-25	31		
	1-2-7-12-13-18-19-20-25	30		
	1-2-3-13-18-23-19-20-25	29		

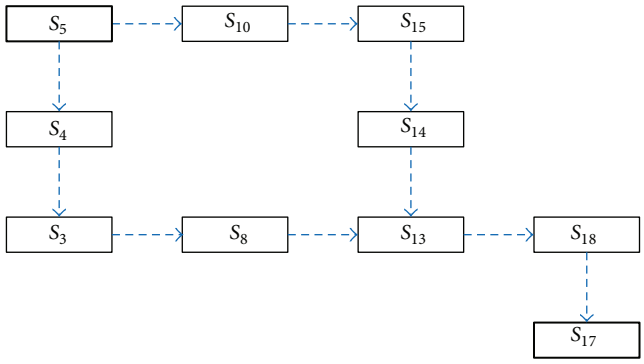


Figure 5: Cooperative multipath planning from Intersection 5 to Intersection 17.

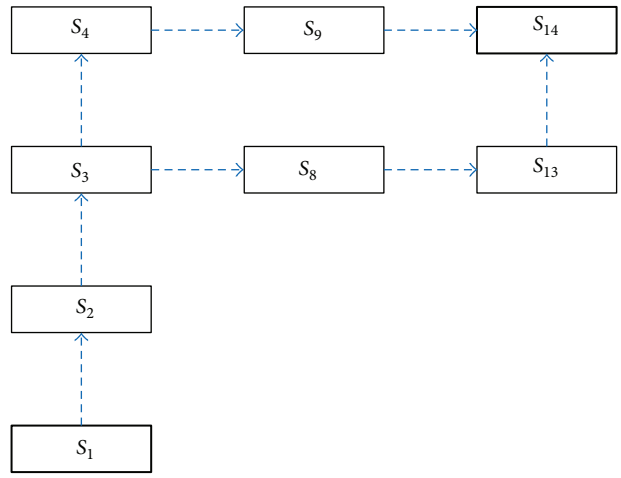


Figure 6: Multipath planning from Intersection 1 to Intersection 14.

reliability. Figure 9 shows that k-shortest planning's time increases linearly over time, while Q-learning multipath planning increases logarithmically. In addition, we can ensure that the reliability of Q-learning multipath planning is more than 0.9. But the reliability of k-shortest planning is less than 0.8.

5. Conclusion

This paper proposes a new Q-learning algorithm to solve the multipath planning problem for OD pairs in a city urban road network. Different from traditional multipath algorithms, this paper focuses on multipath planning via FNN-based Q-learning, an algorithm that makes it easier to choose alternative paths by recurring to the suboptimum Q-value. Furthermore, the paper uses logic to increase the response speed and imposes constraint conditions on path generating to ensure the reliability of the set of candidate paths, which has rarely been taken into account in existing works. Simulation results validate the efficiency and adaptability of the proposed algorithm. In the future work, we will further consider the waiting time at intersections in the algorithm.

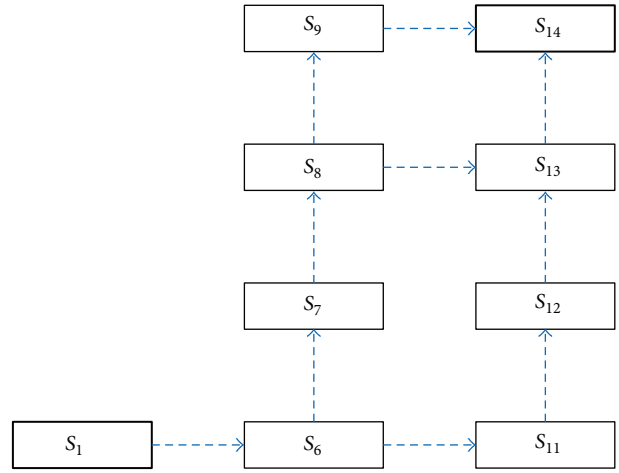


Figure 7: Cooperative multipath planning from Intersection 1 to Intersection 14.

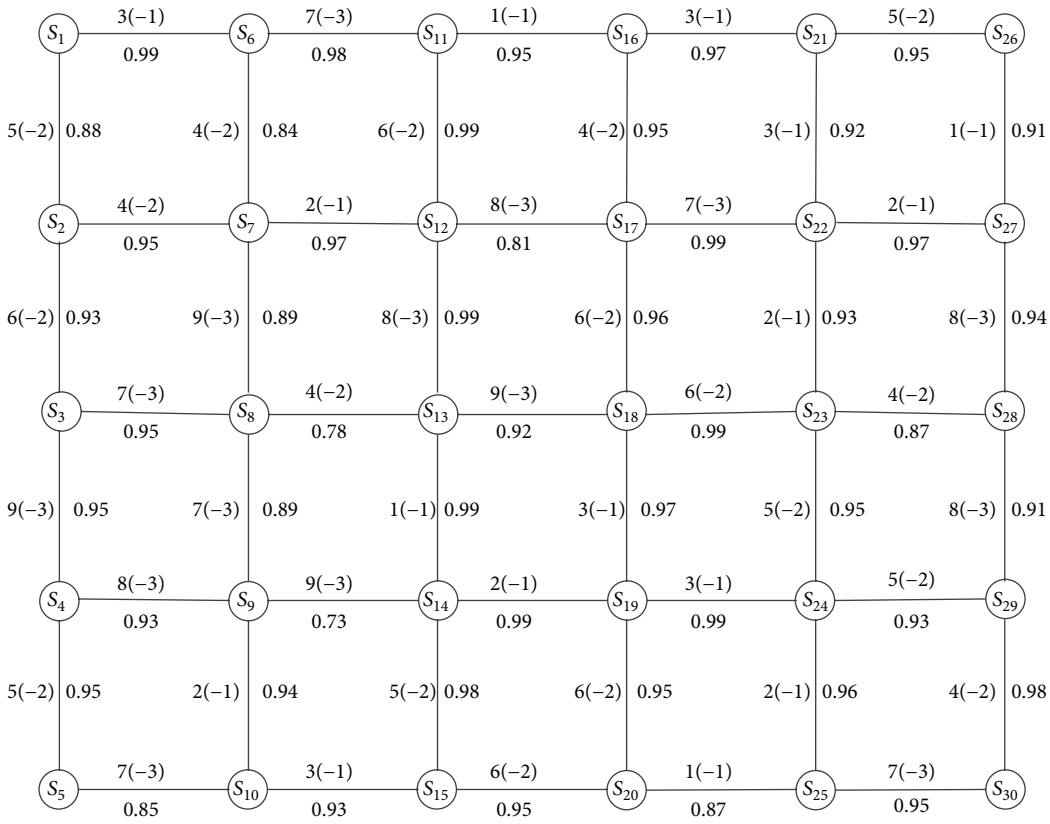


Figure 8: Road networks with 30 nodes.

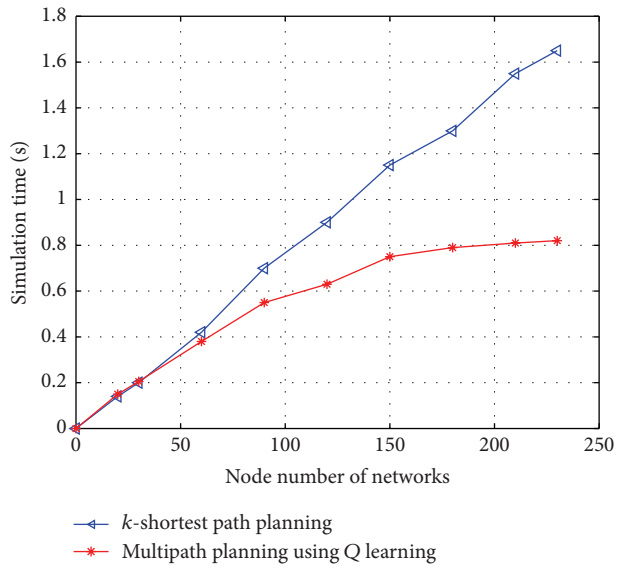


Figure 9: Planning time over the increase of node number of networks with different methods.

Conflict of Interests

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

References

[1] H. Shimoura and K. Tenmoku, "Development of elemental algorithms for future dynamic route guidance system," in *Proceedings of the Vehicle Navigation and Information Systems Conference (VNIS '94)*, pp. 321–326, Yokohama, Japan, 1994.

[2] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische Mathematik*, vol. 1, no. 1, pp. 269–271, 1959.

[3] P. Hart, N. Nilsson, and B. Raphael, "A formal basis for the heuristic determination of minimum cost paths," *IEEE Transactions on Systems Science and Cybernetics*, vol. 4, no. 2, pp. 100–107, 1968.

[4] C.-K. Lee, "A multiple-path routing strategy for vehicle route guidance systems," *Transportation Research C: Emerging Technologies*, vol. 2, no. 3, pp. 185–195, 1994.

[5] D. Eppstein, "Finding the k shortest paths," *SIAM Journal on Computing*, vol. 28, no. 2, pp. 652–673, 1999.

[6] V. M. Jiménez and A. Marzal, "Computing the k shortest paths: a new algorithm and an experimental comparison," in *Algorithm Engineering*, vol. 1668 of *Lecture Notes in Computer Science*, pp. 15–29, 1999.

[7] E. A. Dinic, "Algorithm for solution of a problem of maximum flow in a network with power estimation," *Soviet Math*, vol. 11, no. 5, pp. 1277–1280, 1970.

[8] D. Torrieri, "Algorithms for finding an optimal set of short disjoint paths in a communication network," *IEEE Transactions on Communications*, vol. 40, no. 11, pp. 1698–1702, 1992.

[9] R. S. Sutton and A. G. Barto, "Reinforcement learning: an introduction," *IEEE Transactions on Neural Networks*, vol. 9, no. 5, p. 1054, 1998.

[10] M. Z. Arokhlo, "Route guidance system using multi-agent reinforcement learning," in *Proceedings of the 7th International Conference on Information Technology in Asia (CITA '11)*, pp. 1–5, Kuching, Malaysia, July 2011.

[11] M. Zolfpour-Arokhlo, A. Selamat, and S. Z. M. Hashim, "Self-adaptive and multi-agent reinforcement learning in route guidance system," in *Proceedings of the 5th Malaysian Conference in Software Engineering (MySEC '11)*, pp. 383–387, IEEE, Johor Bahru, Malaysia, December 2011.

[12] Z. Zhang and J.-M. Xu, "A dynamic route guidance arithmetic based on reinforcement learning," in *Proceeding of the 4th International Conference on Machine Learning and Cybernetics*, pp. 3607–3611, August 2005.

[13] C. Wu, O. Satoshi, S. Ohzahata, and T. Kato, "Flexible, portable, and practicable solution for routing in VANETs: a fuzzy constraint Q-learning approach," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 9, pp. 4251–4263, 2013.

[14] K. Y. Chan and T. S. Dillon, "On-road sensor configuration design for traffic flow prediction using fuzzy neural networks and taguchi method," *IEEE Transactions on Instrumentation and Measurement*, vol. 62, no. 1, pp. 50–59, 2013.

[15] C. Li, S. G. Anavatti, and T. Ray, "Short-term traffic flow prediction using different techniques," in *Proceedings of the 37th Annual Conference of the IEEE Industrial Electronics Society (IECON '11)*, pp. 2423–2428, Melbourne, Australia, November 2011.

[16] B. R. Hellinga and L. Fu, "Reducing bias in probe-based arterial link travel time estimates," *Transportation Research, Part C: Emerging Technologies*, vol. 10, no. 4, pp. 257–273, 2002.

[17] M. Tomio, S. Takaaki, and M. Taka, "Route identification and travel time prediction using probe-car data," *International Journal of ITS Research*, vol. 2, no. 1, pp. 21–28, 2004.

[18] Y. Y. Chen, M. G. H. Bell, and K. Bogenberger, "Reliable pretrip multipath planning and dynamic adaptation for a centralized road navigation system," *IEEE Transactions on Intelligent Transportation Systems*, vol. 8, no. 1, pp. 14–19, 2007.

[19] P. Jindahra and K. Choocharukul, "Short-run route diversion: an empirical investigation into variable message sign design and policy experiments," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 1, pp. 388–397, 2013.

[20] T. Xing and X. Zhou, "Reformulation and solution algorithms for absolute and percentile robust shortest path problems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 2, pp. 943–954, 2013.

[21] K. Y. Chan, S. Khadem, T. S. Dillon, V. Palade, J. Singh, and E. Chang, "Selection of significant on-road sensor data for short-term traffic flow forecasting using the Taguchi method," *IEEE Transactions on Industrial Informatics*, vol. 8, no. 2, pp. 255–266, 2012.