DEVELOPMENT OF AN ACCIDENT RISK PREDICTION APPROACH FOR DYNAMIC ROUTE GUIDANCE

BY

GUOBIN MAO

A thesis submitted in comformity with the requirements
for the degree of Master of Applied Science
Graduate Department of Civil Engineering
University of Toronto

Dedicated to

my beloved wife Linda Su

and

my lovely daughter Baihui Mao

Thesis title: Development of an Accident Risk Prediction Approach for Dynamic Route Guidance

Degree: Master of Applied Science, 2003

Name: Guobin Mao

Department: Civil Engineering

University: University of Toronto

ABSTRACT

Dynamic route guidance systems (DRG) aid drivers in choosing the best routes based on real-time conditions in networks. Whether for the evaluation of DRG's impact on a whole-network traffic safety or for determination of the safest routes by DRG, suitable accident prediction models are required. However, such models are not available for links of all kinds of roads. The objective of this research is to develop an accident prediction approach for links on freeways and urban streets suitable for DRG. Firstly, datasets were established by collecting accident data from police reports, link geometric and traffic data from a simulation network. Then, based on a concept we call "link neighbors", a model form was set up. The model performance was measured using the standard deviation of the model's outputs, and through model optimization, model parameters were determined.

ACKNOWLEDGEMENT

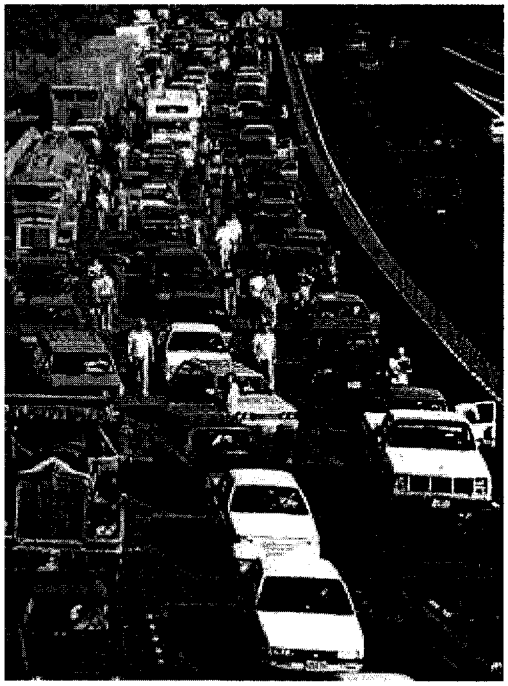
I would like to express my sincere thanks to my supervisor Professor Baher Abdulhai for providing invaluable guidance and funding for my research.

I would also like to thank Asmus, Celine, Kenny and Saikat for sharing their knowledge with me. Particularly, as issues related to Paramics and Microsimulation.

My most sincere gratitude goes to my wife Linda, who has been so supportive and patient throughout my research.

TABLE OF CONTENTS

1 INTRODUCTION



When an auto driver travels on a road from one place to another in a city, there may be many routes to choose. Certainly an ideal one should be fast and safe; auto drivers would try to avoid congestion and accidents. In fact, to be safe and to be fast are not only the goals of drivers, but also the goals of traffic engineering in general.

To realize these goals, one approach is to help drivers to choose optimal routes. In Intelligent Transportation Systems(ITS)(1), dynamic route guidance systems(DRG) are designed to provide drivers with routing information which will assist them to choose routes that are fast and safe . ITS Architecture for Canada(2) describes DRG as:

> *This market package offers the user advanced route planning and guidance which is responsive to current conditions. The package combines the autonomous route guidance user equipment with a digital receiver capable of receiving real-time traffic, transit, and road condition information which is considered by the user equipment in provision of route guidance.*

There are two main characteristics in DRG. One is dynamic, which means that DRG gives route guidance based on real-time information on roads, such as current traffic volumes or speeds. DRG should not be based on static traffic information such as annual average daily traffic volume(AADT), because current traffic on roads changes all the time. The other is route guidance, which means determine a route to meet drivers' goals. Some researchers work to minimize average trip time (3-5), which definitely is very important. However, safety is also very important. If DRG recommends drivers with fast but less safe routes, or if with application of DRG, the travel time is lessened but number of accident is increased for the whole network in which DRG is applied, then the benefit of DRG will be discounted or doubtable. Therefore, safety issues have to be concerned (6-7). In fact, some related researches were done.

Lord, D.(7), for example, developed several traffic accident prediction models, and then applied them to determine the safest paths on digital networks and to evaluate effects of DRG on safety. Look(8), et al, by developing a micro traffic simulation model and integrating it with accident prediction models, established relationships between the number of network-wide accidents and DRG market penetration to quantify the detailed effects of DRG on traffic safety. Their work is meaningful. However, the reliability of their results largely depended on the accident prediction models used. Proper accident prediction models are very essential to accurate evaluation of traffic safety and determination of safest routes.

There are two kinds of accident prediction models developed by Lord. D. One is for nodes, which are at-grade intersections; the other is for links, the sections between adjacent nodes. In the models for nodes, the expected number of accidents is a function of hourly flow, which is a real-time variable. However, in the model on links, the independent traffic variable is AADT. Apparently, this model is static one, so it dose not meet the dynamic requirements of DRG. Look and Abdulhai used this model on DRG due to lack of a model for links based on real-time traffic such as hourly traffic flow. It was concluded that the development of a real-time accident prediction model for links is very necessary.

The objective of this research is to develop an accident prediction approach for links of freeways and urban roads based on key geometric features and hourly traffic characteristics, so that the model can be used in DRG for safety evaluation or the choice of the safest routes in a network.

The thesis firstly reviews some accident prediction models and their limitations. Then it introduces a method of traffic data collection, which uses traffic simulation on computer instead of unavailable field data. Next, an accident prediction model, which is based on a new concept of "link neighbors", is introduced. To assess model performance, an evaluation criterion was established, which was also used in the parameter calibration stage. Finally, the research results are obtained and analyzed.

2 LITERATURE REVIEW

2.1 FUNCTION OF ACCIDENT PREDICTION MODELS

Many accident prediction models were developed for various purposes in the literature. Some models use one variable to identify its effect on traffic safety on only those study roads. Some models use two or more variables to find out their combined effect on accident occurrence on some specific class of roads. Some use macroscopic traffic data (annual average) while others use microscopic traffic data (hourly or shorter).

The reason for developing one-variable models is mainly because the impact of that specific variable on accident rates is of interest. Liu et al(9), for example, developed models for casualty rate with average travel speed or speed differentials in order to develop countermeasures on speed regulations to reduce the number of collisions on highways in Saskatchewan, Canada. The authors claimed that both average travel speed and speed differentials were correlated with casualty rates. Those models show that the higher the average speeds or speed differential on those highways, the higher the casualty rates. Garber et al(10) studied relationships between number of crashes and occupancy for specific sites in order to incorporate crash risk while selecting congestion-mitigation strategies in several specific sites. Persaud et al(11) developed both microscopic and macroscopic accident prediction models for freeways, only based on volume.

Only considering one factor that affects accident occurrence, these one-variable models may work well on their study roads if all other neglected factors are homogeneous. However, if other unexplained factors vary largely, the estimates from these models will become unreliable. Therefore, multi-variable models are required to predict accident risk.

Some multi-variable models are developed for specific kinds of road. Garber et al(12) established models with variables of mean speed, standard deviation of speed, flow per lane, lane width and shoulder width to predict crash rates on roadways in the state of Virginia with speed limits of 89 or 105 km/h. Their models show that crash rate is not solely decided by one variable, but by an interaction of those variables. Lee et al (13), using variables including variability of speed, density, road geometry, weather and time of day, also developed a multi-variables model based on data collected from an expressway in Toronto, Canada. For urban arterial roadways in two cities in British Columbia, Canada, Sawalha et al(14) established accident prediction models and claimed that the variables with a significant effect on accident occurrence were "traffic volume, section length, unsignalized intersection density, driveway density, pedestrian crosswalk density, number of traffic lanes, type of median, and nature of land use".

These three kinds of models consider several factors instead of one. However, they are limited to only certain types of roads, freeway or urban arterials, respectively. All of these models cannot be applied to a road network, which includes lower class of roads. Some researchers did accident prediction studies on both urban and rural highways (15, 16), yet in their models, the measure for traffic volume is AADT, which is a static measure, so the models are not suitable for prediction in dynamic route guidance systems.

An ideal accident prediction model for DRG should include multiple important variables, using real-time traffic data and that is applicable to all kinds of road in a network. Up to now, no model meets all these requirements. Establishment of such a model is the goal of this research.

2.2 ACCIDENT RISK ESTIMATION TECHNIQUES

The usual method to develop accident prediction models [12, 13] includes the following 3 steps:

1. Data collection, including accident data on roads, and road attributes;
2. Attributes are categorized and their observed values are averaged. Then, the data are grouped according to the combination of attribute category, so that every group is relatively homogenous. Accident risk is then computed in one group, while attributes in that group form a set of independent variable values are averaged.
3. Based on data generated in step 2, some regression technique is used to develop the mathematical function or the accident prediction model.

In every step above, there are some problems that researchers face. The first is data collection. It is not easy to collect large sets of traffic and accident data for links, especially on low-class roads. If the dataset is not large enough, statistical precision can not be guaranteed.

The second problem lies in step 2, where the produced group should be homogenous. If the important attributes are not properly categorized, their variations may be so large that within group homogeneity is jeopardized. Consider for instance the case of two variables in a dataset and one is very important, while the other is not important. To enhance within-group homogeneity, the former should be divided into more categories that the latter. If both of these variables have equal number of categories, then the categories of the important one are less homogeneous. When many variables are considered, this problem will become very complex because it is hard to know how many categories are suitable for each variable, and how to categorize those variables.

In step 3, there are some problems as well. First, due to attribute categorization, sample size in each category would be small.. Second, measurement errors could be affect model accuracy, especially the dependent variable, which is not even measured but statistically estimated. Third, the assumptions on which the model is developed may be wrong. Finally, the model function may miss some important variables, or include unimportant variables.

All these problems should be eliminated or alleviated to produce a good model. We address those issues as explained later in the subsequent sections.

3 METHODOLOGY

This research attempts to overcome all the limitations mentioned in Chapter 2. The main parts of this thesis include data collection, model form (or structure) establishment, model performance evaluation and model parameter optimization.

To develop an accident risk prediction model, three kinds of data should be collected. They are traffic data, geometric data and accident data. Because it will be very time consuming and expensive to collect a large number of traffic data from every link in a network, traffic data would be collected from a simulated network on computer instead of a real network. Geometric data of links were also gathered from the simulated network that has link dimensions and speed limits for the real network. Accident data were obtained from police reports.

These three kinds of data formed a large dataset, which was the base for developing the accident risk prediction model. The accident risk rate of any link can be estimated by data (number of crashed vehicles and volume, etc.) of those links which have similar traffic and geometric features. The main methods include:

1. To measure the similarity between any two links, a concept of "link distance" was introduced.

2. A concept of “link neighbors” was introduced. In the dataset the links with the shortest “link distance” from the link whose accident risk is to be predicted are chosen as its “link neighbors”. The accident risk rate of that link is estimated as the total number of crashed vehicles in its “link neighbors” divided by the total amount of exposure in its “link neighbors”. Exposure is defined as the total traveled length of vehicles within study area during study period.
3. A threshold for “link neighbors” was set to control the size of “link neighbors”.
4. To obtain a model of the highest performance, a model performance measure was set up.
5. By optimizing model performance, all parameters in the model were determined.

The following chapters elaborate on the above steps.

4 DATA COLLECTION

The data for developing the model were collected from two sources, a simulated network which simulates Waterfront Network in Toronto(Abdulhai et al 2002), Canada, and police accident reports (from 1996 to 1999) for Toronto, Canada.

The simulated network includes expressways, ramps, main arterials, collector roads and local roads. There are 1,966 links with total length of about 242km. The length distribution with different speed limits is shown in Table 4.1.

Table 4.1 Lengths of Study Links with different speed limits

Speed Limit(km/h)	20	30	40	50	60	70	90	Total
Link Length(M)	574	77,227	39,185	66,973	3,356	26,138	29,124	242,577

For every link, attribute values collected are as shown in Table 4.2. The reason that these attributes were used for model developing is based on following conditions:

1. All attributes of a link should be obtainable from the simulated network and police reports, and technically also be obtainable from real networks;
2. From the literature, these geometric and traffic attributes were taken as important factors for accident risk prediction;
3. In the accident data set, number of crashed vehicles was used instead of number of accidents, because this research concerns the possibility of a vehicle getting involved in a reportable accident on a link when the vehicle runs on that link.

Table 4.2 Attributes Considered for every link

Geometric Data	Traffic Data*	Accident Data*
1. Speed limit 2. Number of lanes 3. Link length	1.Average Moving speed 2.Variance of average moving speed 3.Average volume per lane	Number of vehicles involved in reportable accidents during year 1996 to 1999.

*For the period from 8:00 a.m. to 9:00 a.m. in weekdays.

4.1 SIMULATED NETWORK

In the summer of 2002, using a software tool called Paramics (See http://www.paramics-online.com/), other researchers at University of Toronto finished the simulated Waterfront Network(Abdulhai et al 2002).

Paramics is "used to model the movement and behavior of individual vehicles on urban and highway road networks". It is based on two inputs, road network data and travel demand data. The traffic in the simulated network using Paramics is similar as traffic in the real network, but it is much easier to obtain traffic data from the simulated network.

A Paramics network is made up of Paramics links and nodes. A node in Paramics is a junction, which may be an intersection or not. For example, within a roadway section, a junction between a tangent and curve, or a junction where number of lanes or width of the link changes, is also a node. A Paramics link is the connection from one node to another.

The definition of a link in this thesis is different from that of Paramics link. In this thesis, a link is defined as the traffic lanes in one direction between two adjacent intersections, while an intersection should be a node connected with three or more nodes. If a Paramics node is connected with two nodes, then it is only a point on a link.

In Figure 4-1, AD and DA are links. Node A and D are intersections. AB, BC, CD,DC, CB and BA are Paramics links. All A, B, C and D are Paramics nodes.

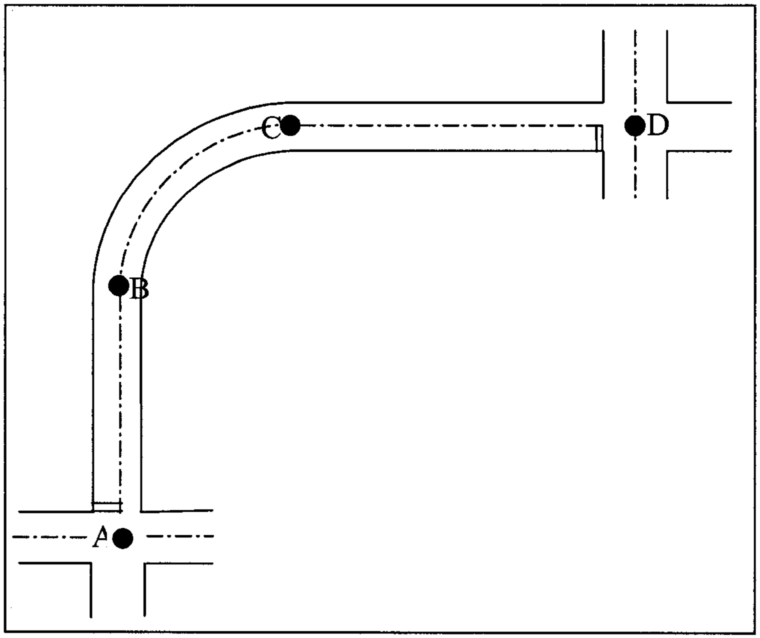


Figure 4-1 Illustration of Different Link Definitions in Paramics and in This Thesis

In the simulated network, nodes and links have some important attributes. The location of a node is determined by the X and Y coordinates, which uses a plane coordinate system called Universal Transverse Mercator(UTM). Properties of a Paramics link include the connected nodes, road classification, speed limit, number and width of lanes, and length.

For all links in the Waterfront network, geometric data were gathered from the inputs of the simulated network, while traffic data were gathered from multiple simulation runs of the network model.

4.2 DATA COLLECTION

Geometric Data Collection

Three kinds of geometric data were collected: speed limit, number of lanes and link length. As mentioned before, these data are collected or calculated from the Waterfront network simulation model. It is not difficult to collect or calculate those data for every Paramics links.

Speed limit and number of lanes of a Paramics link can be collected based on the link's category.

The length of any Paramics link is the distance between its two nodes. It can be calculated from nodes' X/Y values, and from the arc radius and X/Y values of the arc center if the link is an arc.

Because Paramics link is different from link concept needed in this thesis, every Paramics link which is only a part of real link(In Figure 4-1, Paramics link AB is a part of link AD) should join its neighboring Paramics links to form a complete link, which is the connection between two intersections. The number of lanes of the joined link is the average of numbers of lanes weighted by lengths, as calculated by formula 4.1, while the length of the joined link is the calculated by formula 4.2,where the width of any intersection is supposed as 25 meters.

The number of lanes $=(\sum N_i L_i)/ \sum L_i$ for all i, (4.1)

The length of a link$=\sum L_i - 25$ for all i, (4.2)

where N_i=the number of lanes of Paramics link I;

L_i=the length(m) Paramics link I;

Traffic Data Collection

A Paramics Simulation model for the Waterfront network in Toronto was used for the morning peak period in typical weekdays from 1996 to 2001.

To obtain traffic characteristics on every link from the simulation model, a loop detector is added to the middle part of every Paramics link. While the simulation model is running, it records the speed of any vehicle passing the detector.

The simulation period is from 7am to 9 am. Simulation before 8am is to warm up traffic in the network and that between 8am and 9am is to collect volumes, and speeds from which mean moving speed and standard variance of moving speed can be calculated.

For each link made of several Paramics links, mean volume and mean moving speed are calculated from the volume and mean speed of its Paramics links. The standard variance of moving speed of a link is calculated from all the moving speeds during that hour in all its Paramics links, rather than from standard variances of moving speed of its Paramics links.

The simulation has been run for 39 times. Because every simulation run is stochastic, the simulation outcome is always a little different each time. This is similar as the fact that traffic in a network every day is a little different. Since we never know which time of simulation will occur more frequently in real world, and we do not know either which will tend to cause more traffic accidents than the rest, it is supposed that each situation in 39 simulations has equal possibility to occur in the real word, and has same chance to cause traffic accidents.

The volume divided by number of lanes gets volume per lane.

Accident Data Collection

From police reports, a four-years database of accidents happened from 1996 to 1999 in Toronto is available. About 180 thousands vehicles every year were involved in accident in this city and every accident was recorded by policemen in detail. However, only those data meeting all following conditions are useful for the research.

1) Happened in morning peak hour(from 8:00 to 9:00)
2) Happened on weekdays(From Monday to Friday)
3) Classified by police as reportable accidents
4) Happened on links in Waterfront network.

There are no difficulties in the first 3 filtrations. For the fourth condition, it is a problem of judgment on whether the location of an accident is within the range of any link.

The first step for 4) is coordinate transform of accident locations. All 4-year accident locations in the database from police were recorded in terms of spherical coordinates Latitude and Longitude, but those in 1996 were also recorded in terms of UTM, the planar coordinates X and Y. Since locations in the simulated network are described in UTM, all accident locations in Latitude/Longitude should be converted to UTM. The operation of such a conversion is called projection.

After projections, locations of accidents and locations of nodes in waterfront network are in same coordinates. It can be determined whether an accident happened on a specific Paramics link, based on the following link information:

1, X/Y values of the two nodes of the Paramics link;

2, Whether the two nodes of the link are intersections or not;

3, The link direction and width;

4, Arc radius and location of center, if applicable.

After the filtration and mapping, there remained only about 300 crashed vehicles in one year.

What may need to be mentioned here is that the damage caused to every vehicle may vary greatly. In Ontario, drivers need to contact police if a vehicle collision causes more than $1,000(before January, 1998, the reporting threshold was $700) in damages, or any injury is caused, no matter how small is the damage estimate. The reporting threshold refers to the total damages for all vehicles involved in the accident. Therefore, the above accident data only indicate how many vehicles involved in reportable accidents, no matter seriously or not.

4.3 DATASETS

It is found that some links have very low simulated volume, for example, 1 veh/h/lane, but some accidents happened. It seems abnormal. Therefore these links should be excluded. In this research, the links whose volume per lane is less than 60 veh/h/lane would not be considered in model development.

Too short links will also be excluded. The threshold is: the distance between two nodes of a link is less than 25 meters.

In the final database, there are 1966 links, 909 crashed vehicles, 249 million*km*veh of exposure. The average accident risk is 3.639 veh/(million*km*veh).

Two data sets were produced. One dataset was used to estimate accident risk. It recorded all traffic data in 39 times of simulation, so for each link there are 39 records. Apart from 3 traffic attributes and 3 geometric attributes in this dataset, it also included exposure and number of crashed vehicles. Exposure of a link was defined as over the study period the number of vehicles passing the link times the link length. Since one link has 39 simulation results, the exposure in one record of a link is estimated as:

EXP=(volume per lane)*(number of lanes)*(link length)*(5/7*365*4)*1/39

In the second dataset, each link has only one record, in which 3 traffic attributes are the average of 39 records from the first dataset. Apart from 3 traffic attributes, 3 geometric attributes are also included, but exposure and number of crashed vehicles are not included.

Based on traffic and geometric attributes, accident risk of links in the second dataset can be estimated by the first dataset. To easily distinguish the two data sets, the first one is called Predicting Dataset, the second one Predicted Dataset.

5 FORM OF ACCIDENT RISK PREDICTION MODEL

5.1 ACCIDENT PREDICTION MODEL FORM

The Predicting Dataset, which was obtained in the last chapter, has thousands of link samples. Each link sample has geometric attributes, traffic attributes, exposure and number of crashed vehicles. Given this dataset, the problem is how to predict accident risk of a link whose geometric and traffic attributes are given. The main idea of accident prediction in this research is based a concept called "link neighbors".

The "link neighbors" are chosen from the Predicting Dataset. These links have very similar geometric features and traffic characteristics as the link whose accident risk is to be predicted. The similarity of link attributes between two links is measured by "link distance", which is a function of weighted attribute differences. All "link distances" between the link to be predicted and all links in Predicting Dataset are calculated; those links with "link distance" smaller than a certain value are chosen as "link neighbors", and then the accident risk of that link to be predicted is estimated as

$$\frac{\text{the total number of crashed vehicles in "link neighbors"}}{\text{the total exposure in "link neighbors"}}$$

The description of the form of the accident prediction model in this research is as shown in Table 5.1.

Table 5.1 Form of Accident Involvement Rate Prediction Model

For a link, the accident involvement rate(AIR), which is defined as the expected number of vehicles involved in reportable accidents every one million-kilometer-vehicle exposure, is estimated as following:

$$\text{AIR} = \frac{N}{EXP} \qquad (5.1)$$

where, N=The total number of crashed vehicles happened on the "link neighbors" in Predicting Dataset.

EXP=The total exposure of the "link neighbors" in Predicting Dataset;

"Link neighbors" are the links with shortest "link distance" from the link whose AIR is to be estimated.

"Link distance" is calculated as in (5.2).

$$d(x1,x2)=(\sum_{r=0}^{n}\left(b_r\frac{a_r(x_1)-a_r(x_2)}{\bar{a}_r}\right)^2)^{0.5} \qquad (5.2)$$

where, d(x1,x2)=the "link distance" between link x1 and x2;

n=Total number of link attributes considered;

b_r= A weight that determine the contribution of attribute difference between the two links to the distance; $b_r \geq 0$;

a_r= The value of attribute r (traffic characteristics or geometric features) of a link;

$\bar{a}_r$= Average of attribute r, weighted by exposure for all links in Predicting Dataset.

The following is the procedure of choosing "link neighbors" for a link:

1. Calculate all distances between links in Predicting Dataset and the link whose AIR is to be estimated;
2. Sort ascendingly the links according to the distances;
3. From low distance to high distance, include links in Predicting Dataset as neighbors, until the total number of crashed vehicles happened on those neighbors is over a certain value, which is called here as "Threshold for Link Neighbor".

Nine variables (including composite variables) are considered for this model, their descriptions and $\bar{a}_r$ are listed in table 5.2.

Table 5.2 Attributes Considered for the Model and Their Average Values

r	Attributes(Variables)	$\bar{a}_r$
0	Speed limit (km/h)	71.90075
1	Number of lanes	2.799721
2	Link length (m)	1559.189
3	Moving speed (km/h)	70.86688
4	Standard Variance of moving speed (km/h)	12.98597
5	Volume per lane(veh/h/lane)	1084.349
6	Randomly produced value(from 0 to 1)*	0.5
7	(5)/(3)(veh/km/lane)	14.49135
8	(3) – (0) (km/h)	6.752294

*A randomly produced value apparently has no use to prediction, but it was included to test the following modeling method.

Two tasks need completion for the model to be useful. They are:

1. What is the threshold for the neighbors?
2. What are the values of b_r.

5.2 THRESHOLD FOR LINK NEIGHBORS

In function (5.1), if we consider only one link neighbor, which has the smallest "link distance" in function (5.2), then this neighbor's N and EXP may be very small. In the most cases, the N is 0, and the N/EXP will be 0. It is apparently not correct that the accident involvement rate of a link is 0. If we consider two "link neighbors", the situation will be similar, but the N and EXP will all increase. The more the neighbors, the higher the possibility that the value of N is large. In this research, I set a threshold for N so that the neighbors included in formula (5.1) have total number of crashed vehicles just over that threshold.

If the threshold is too small, then the links in "link neighbors" are very similar to the link to be predicted, but both N and EXP will be so small that the statistical precision of estimated value for accident risk in (5.1) cannot be guaranteed. In such a case, a small change of the threshold may change the estimate greatly, so the prediction is unstable. Therefore, the threshold should not be too small.

However, if the threshold is too large, the "link distance" will be so great that links in "link neighbors" are not relatively homogeneous. Those links, which have quite different traffic and geometric features, may be considered as the "link neighbors", so the estimation will be inaccurate. In this case, N and EXP will be so large that the accident risk estimate in (5.1) will be insensitive to link's traffic and geometric features, and then the estimates for different kinds of links will always be similar. One extreme is to set the threshold as 909, which is the total number of crashed vehicles in Predicting Dataset, then for any link, its estimated AIR is always 3.639 veh/(M*veh*km). It is useless if we want to compare the traffic safety among different links. Therefore, the threshold should not be too large.

The threshold for N should be set by balancing prediction stability and link homogeneity in "link neighbors". The optimal threshold was finally found to be 90 crashed-vehicles; the reason is discussed in Chapter 7. Therefore, in formula (5.1) all the nearest neighbors will be taken into account so that the N is just a little bigger than 90.

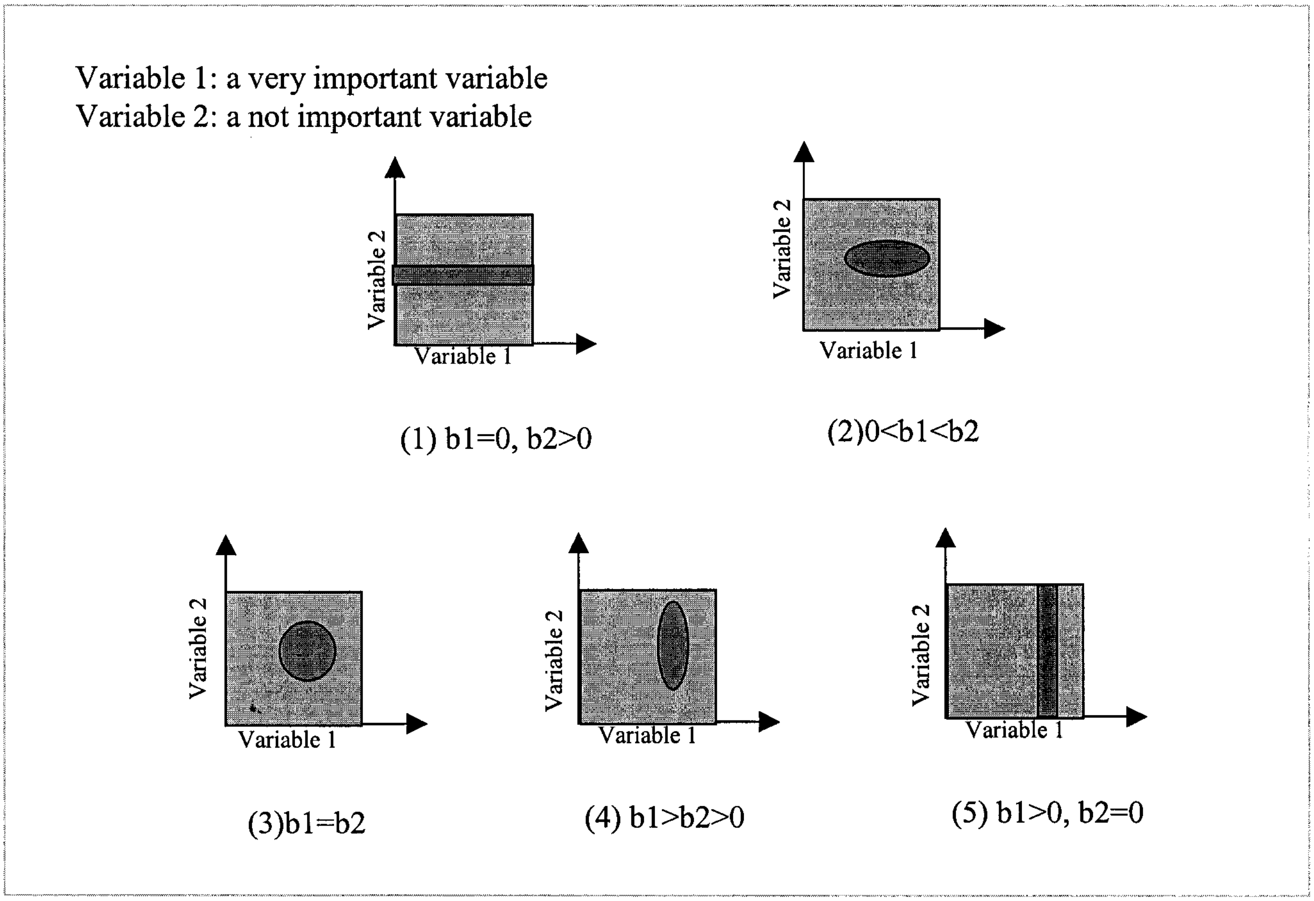
$$N > 90 \tag{5.3}$$

5.3 ATTRIBUTE WEIGHTS

In formula (5.2), weights $b_r(r=0,1,\ldots 8)$ are unknown, but they are important. These attribute weights and attribute differences decide the "link distance", and then have an effect on the selection of "link neighbors". Whether the model is good or not depends on the selection of "link neighbors", so parameters b_r have a great effect on model performance. Usually important attributes should be assigned high weights, while useless attributes should have a weight of zero. Figure 5.1 illustrates influence of attribute weights on selection of "link neighbors" and performance of a two-variable model.

In Figure 5.1, we suppose to develop a model having one very important variable and one not so important variable. Figure 5.1 shows 5 situations with different weights b_r. The dark shaded area is the range of selected "link neighbors" according to the b_r. We can see in situation (1) the model is the worst, because the variation of the important variable in "link neighbors" is the biggest. From (2) to (4) it is better and better. If the unimportant variable is like a randomly produced variable which is uncorrelated with the dependent variable, then in situation (5) the model is the best, because the "link neighbors" are homogenous in the highest degree in term of the important variable.

Chapter 7 will discuss determination of weights b_r to produce a model with the highest performance.



(1) b1=0, b2>0

(2)0<b1<b2

(3)b1=b2

(4) b1>b2>0

(5) b1>0, b2=0

Figure 5.1 Illustrations of "link neighbors" Affected by Weights br

5.4 SIMPLEST MODELS

In formula (5.2), if one of weights is set to non-zero while all others zero, then only one attribute will be considered in the model, which should be the simplest. As shown in Table 5.2, there are 9 variables, so nine one-variable models can be established.

With these models, the AIRs of all links in Predicted Dataset were estimated. Figure 5.1~5.9 show the results.

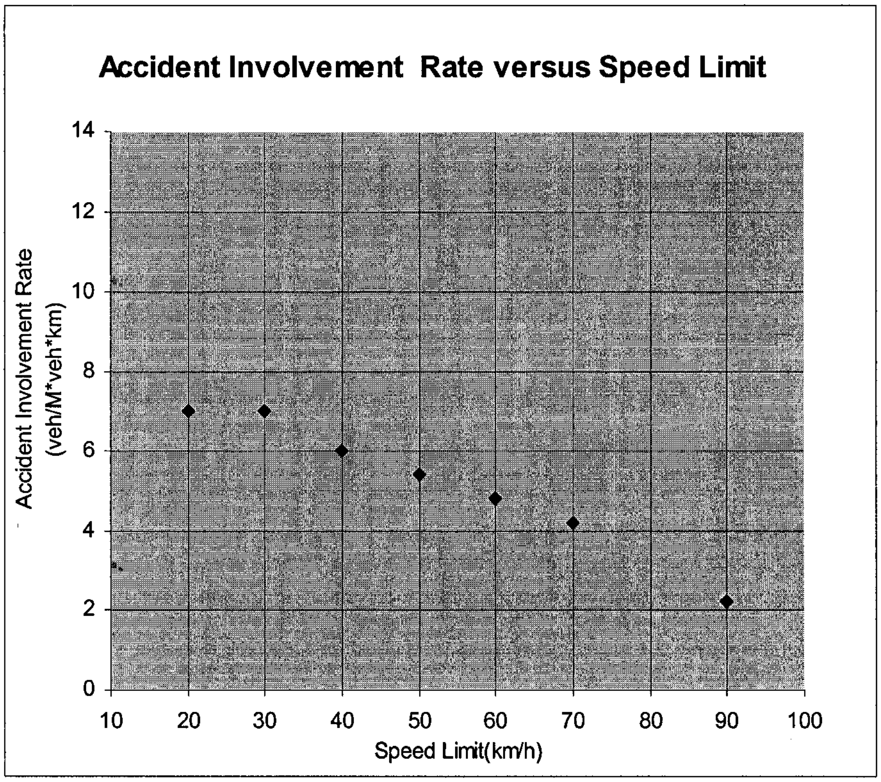


Figure 5.1 Accident Involvement Rate Estimated Only by Speed Limit

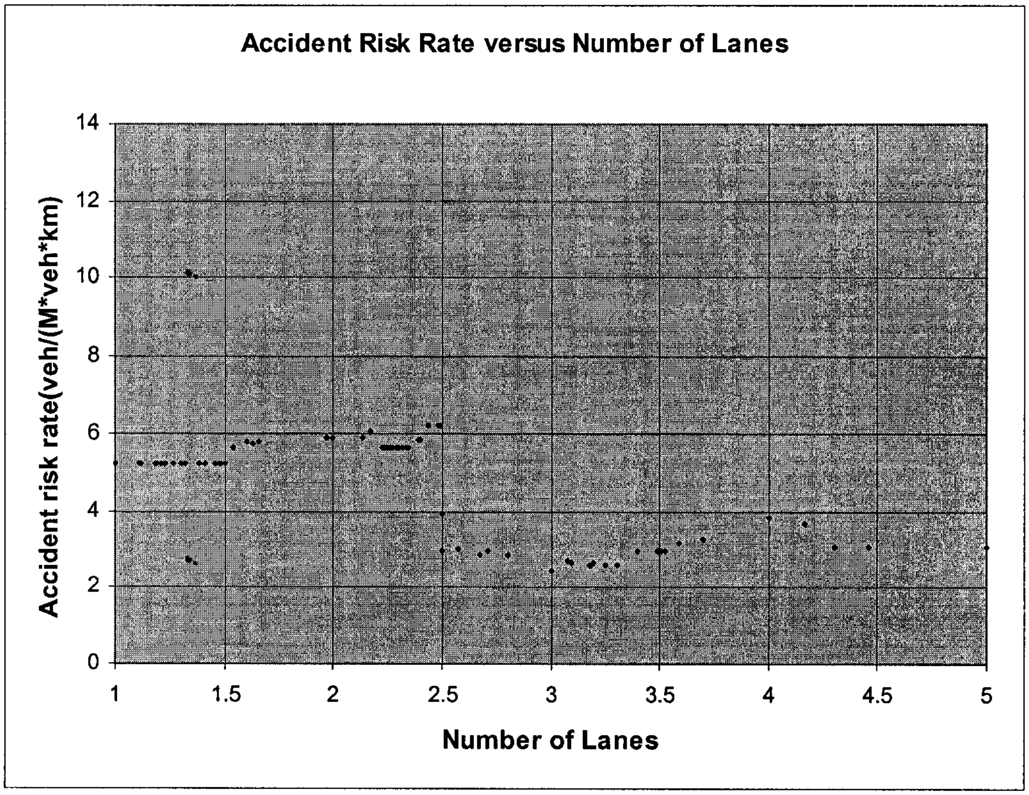


Figure 5.2 Accident Involvement Rate Estimated Only by Number of lanes

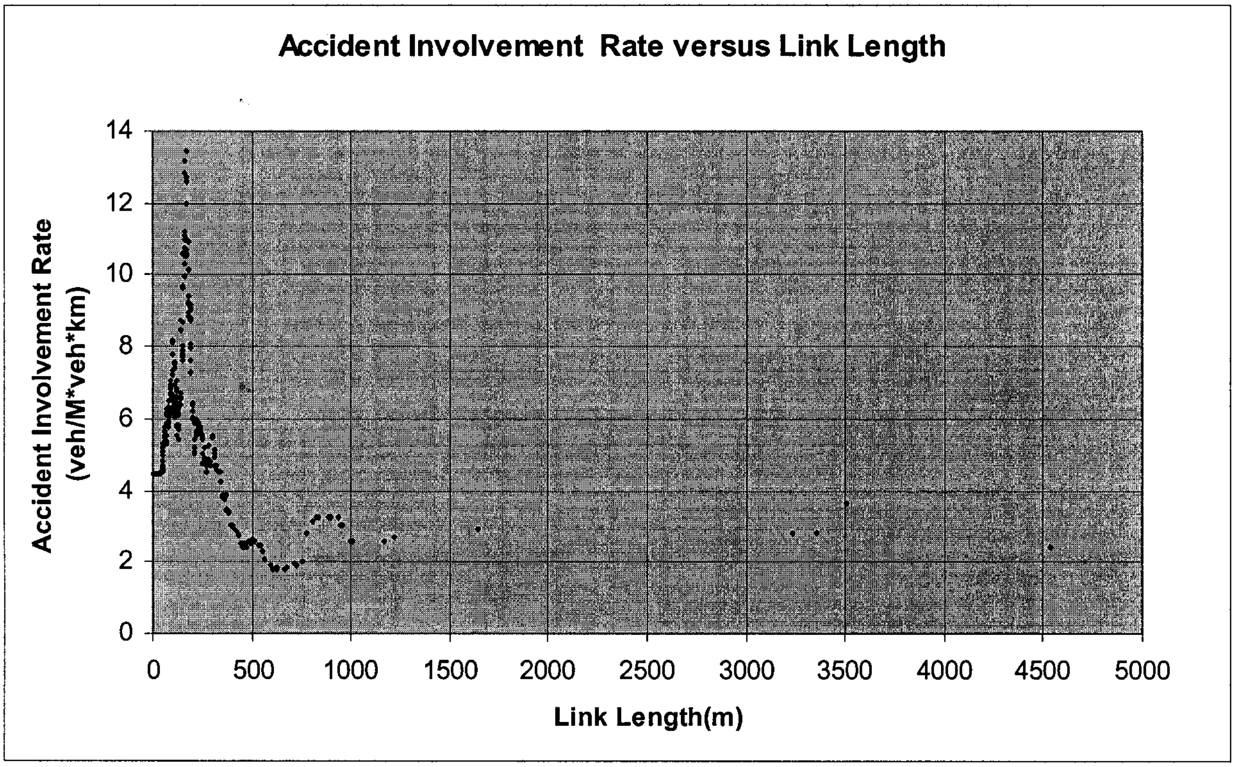


Figure 5.3 Accident Involvement Rate Estimated Only by Link Length

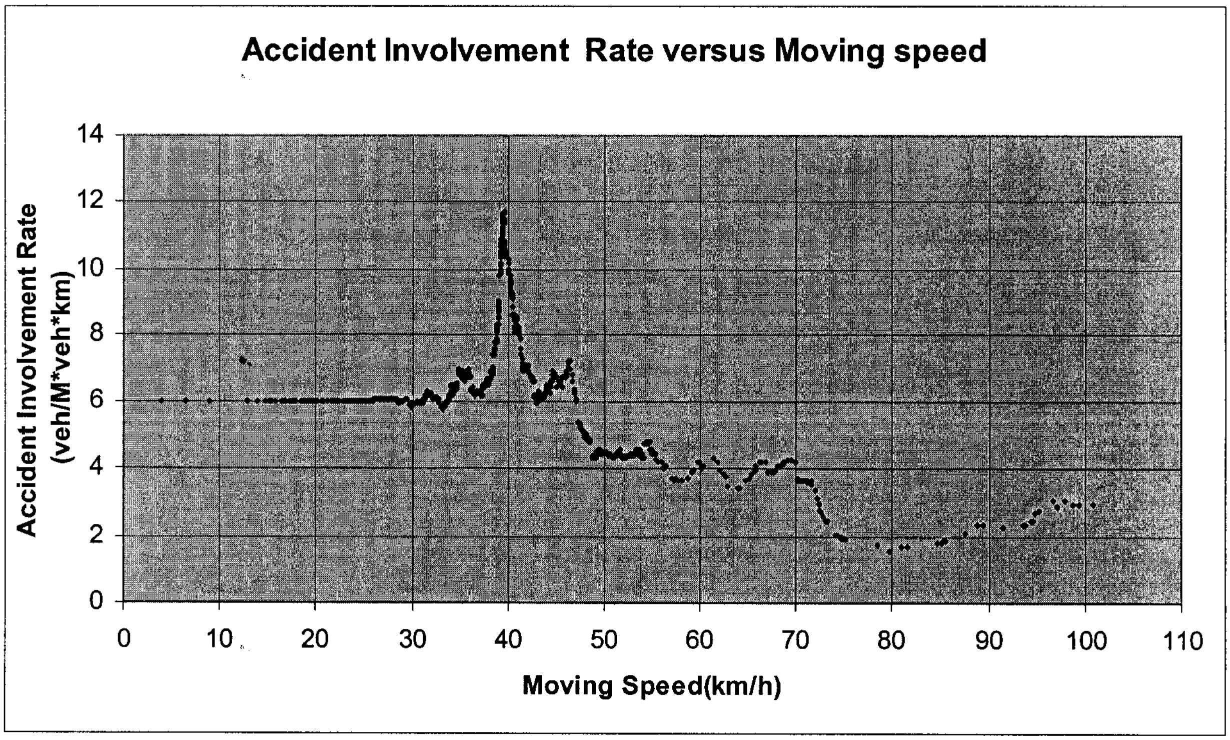


Figure 5.4 Accident Involvement Rate Estimated Only by Moving Speed

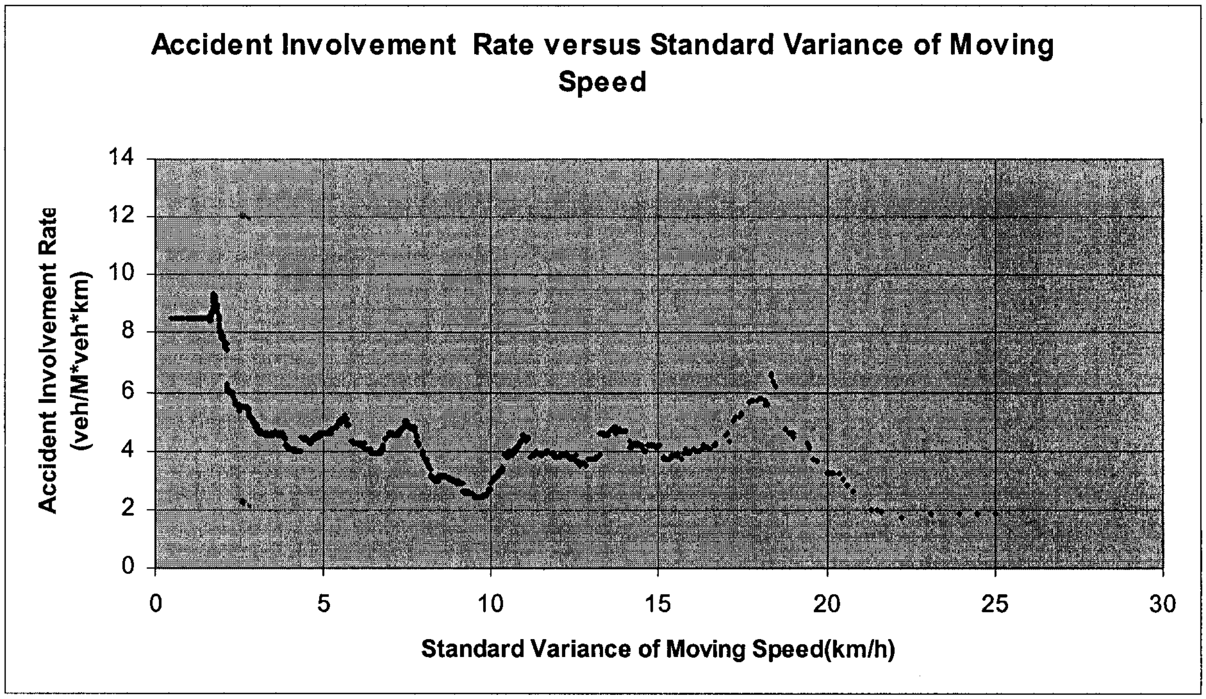


Figure 5.5 Accident Involvement Rate versus Standard Variance of Moving Speed

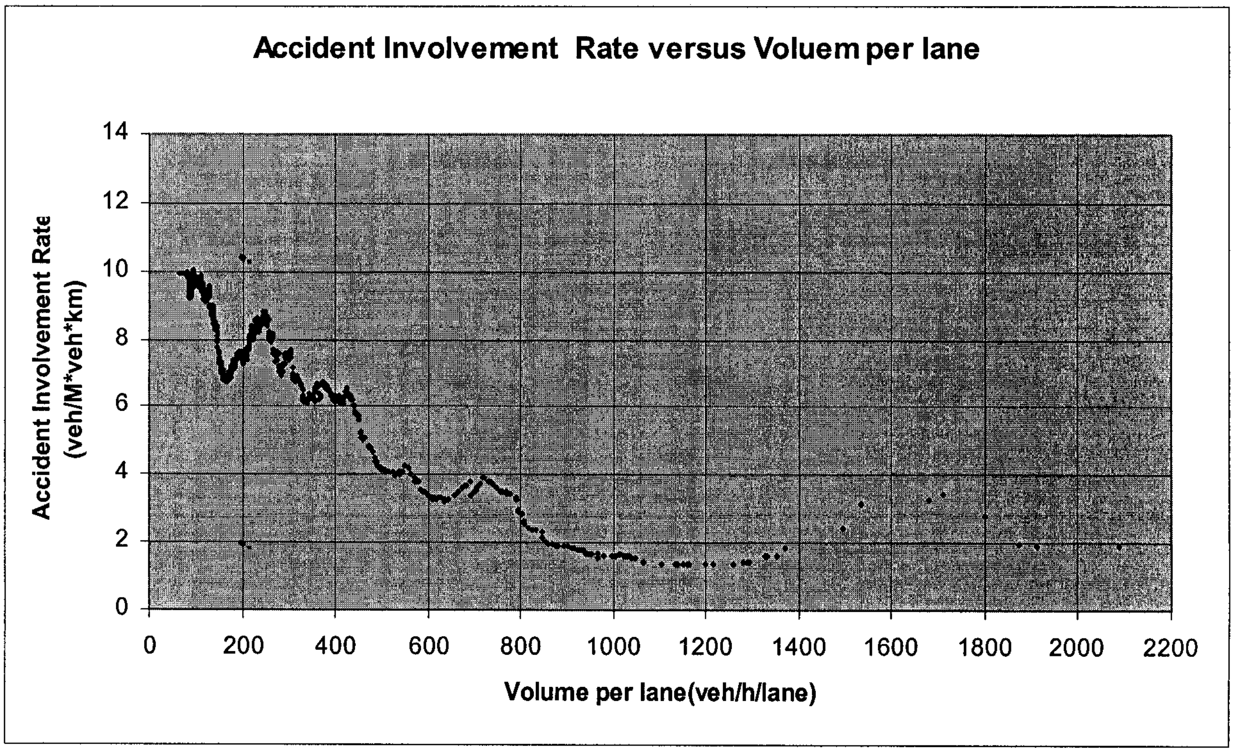


Figure 5.6 Accident Involvement Rate Estimated Only by Volume per lane

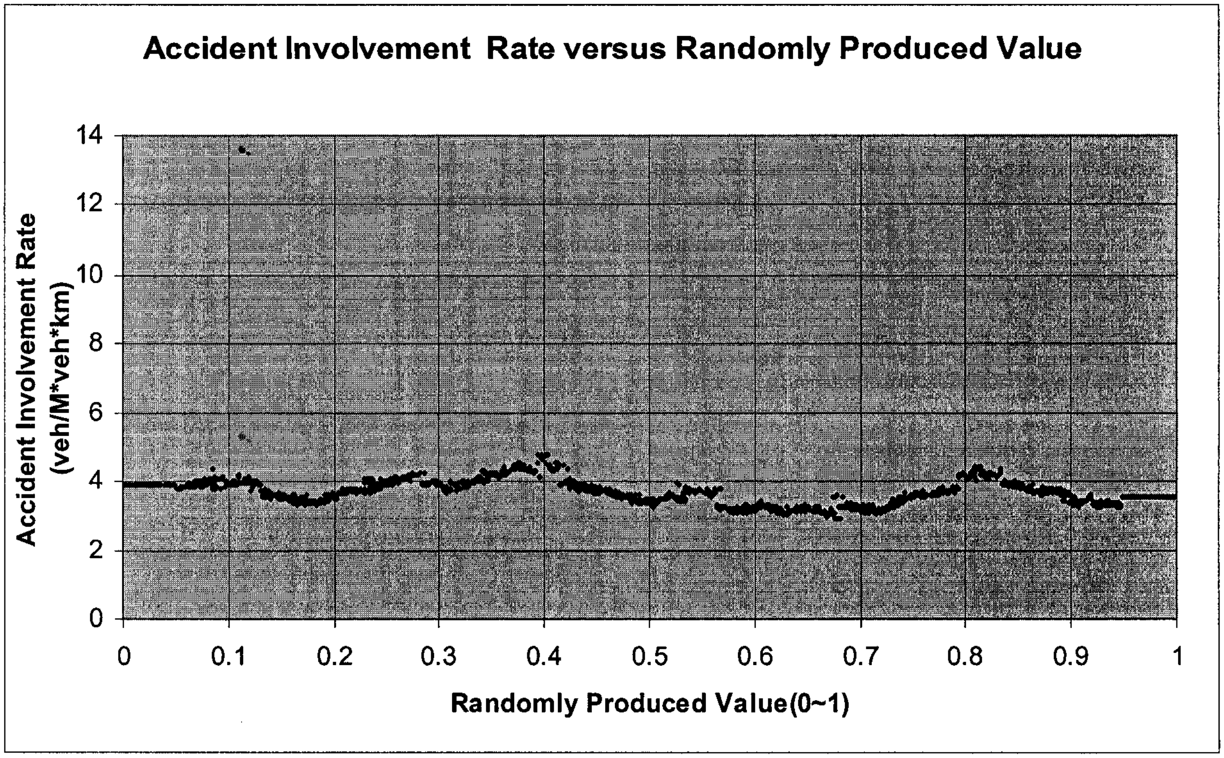


Figure 5.7 Accident Involvement Rate Estimated Only by Randomly Produced Value

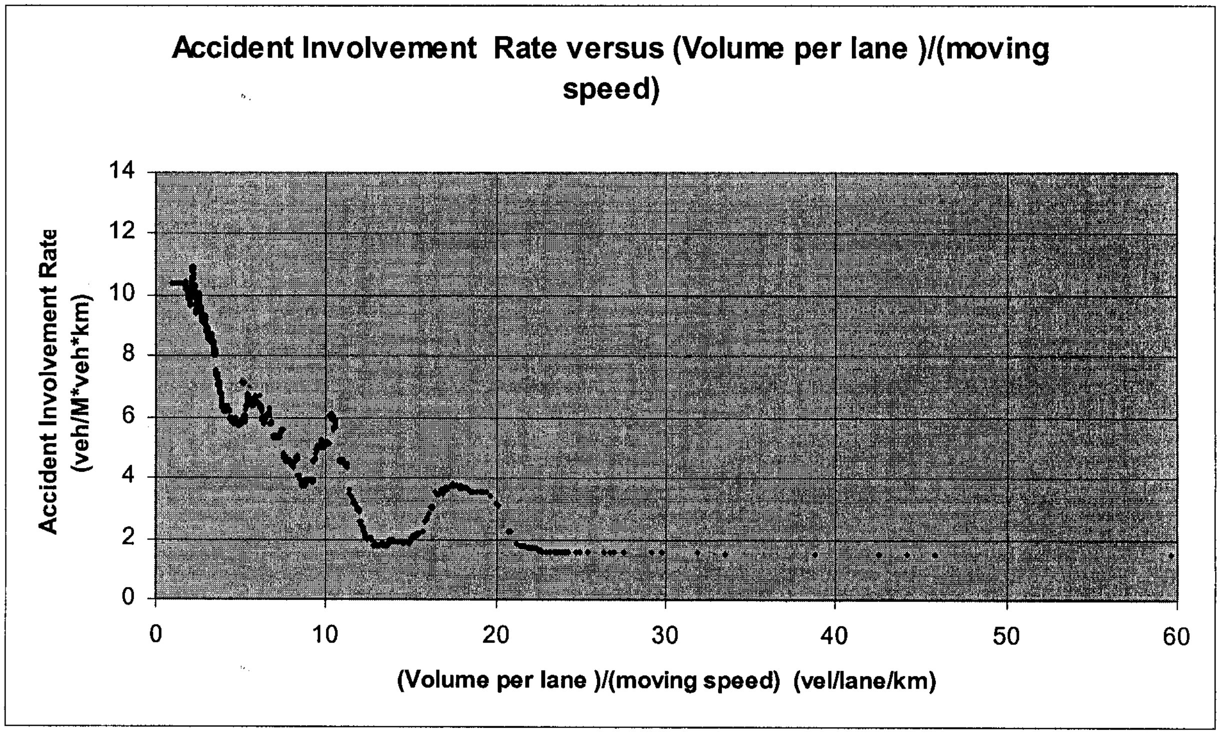


Figure 5.8 Accident Involvement Rate Estimated Only by (Volume per lane)/(moving speed)

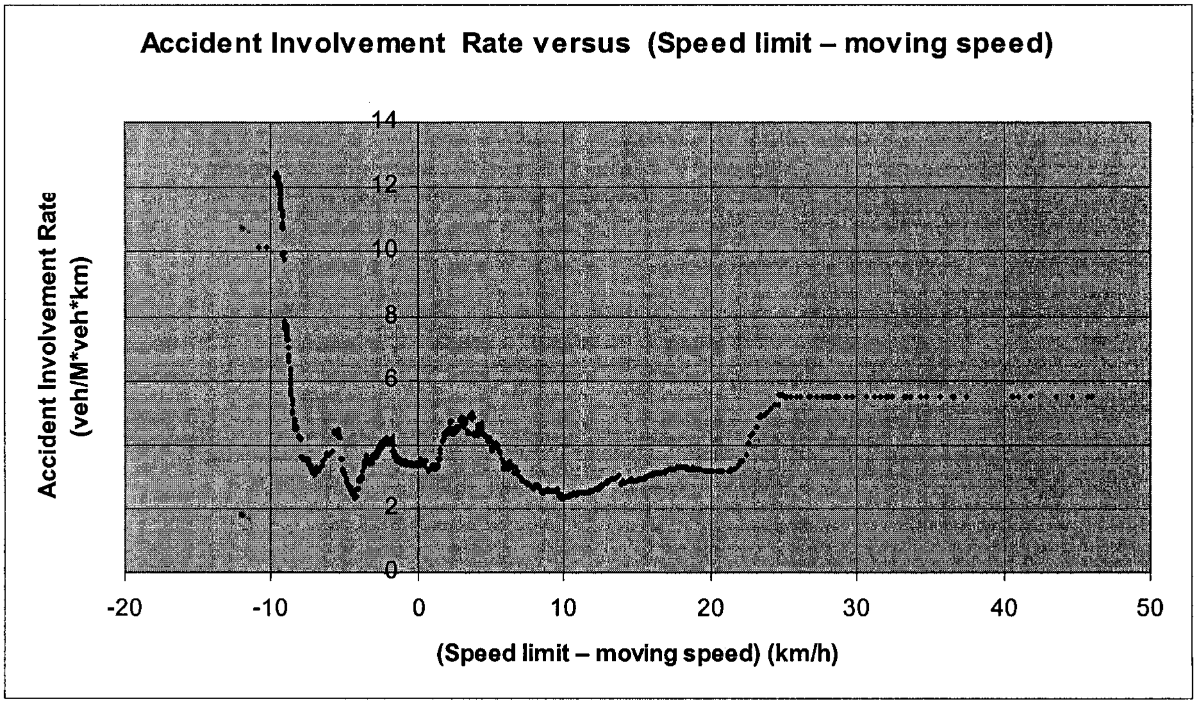


Figure 5.9 Accident Involvement Rate Estimated Only by (Speed limit – moving speed)

From above figures, following results are likely to be drawn from above for the Waterfront Network.

1. Generally, high-speed-limit links are safer than low-speed-limit links;
2. Links with 3 or more lanes are safer than other links;
3. Links longer than 400 meters are safer;
4. Links with moving speed of about 40km/ h are less safe than others; links with moving speed of about 80km/h are the safest;
5. Links with high standard variance of moving speed are usually safer;
6. When the volume per lane is about 900 to 1300 veh/h/lane, the links are the safest;

7. Randomly produced variable is not correlated with accident involvement rate;
8. Links with high (volume per lane)/(moving speed) are safer;
9. Links with moving speeds much higher than their speed limits are less safe.

However, the above results may be incorrect, because the nine models are too simple to give precise prediction.

5.5 COMPARING SIMPLEST MODELS

20 records were randomly chosen from Predicted Dataset. Their accident involvement rates were estimated based on those one-variable models. The samples are shown in Figure 5.10.

From Figure 5.10, it is found that for a link with 9 features, the accident risk estimations based on each single feature will be quite different.

Since all variables except randomly produced variable seem to have explanatory power, but estimation results based on only one variable differ significantly form another, more than one variable is required to estimate accident risk.

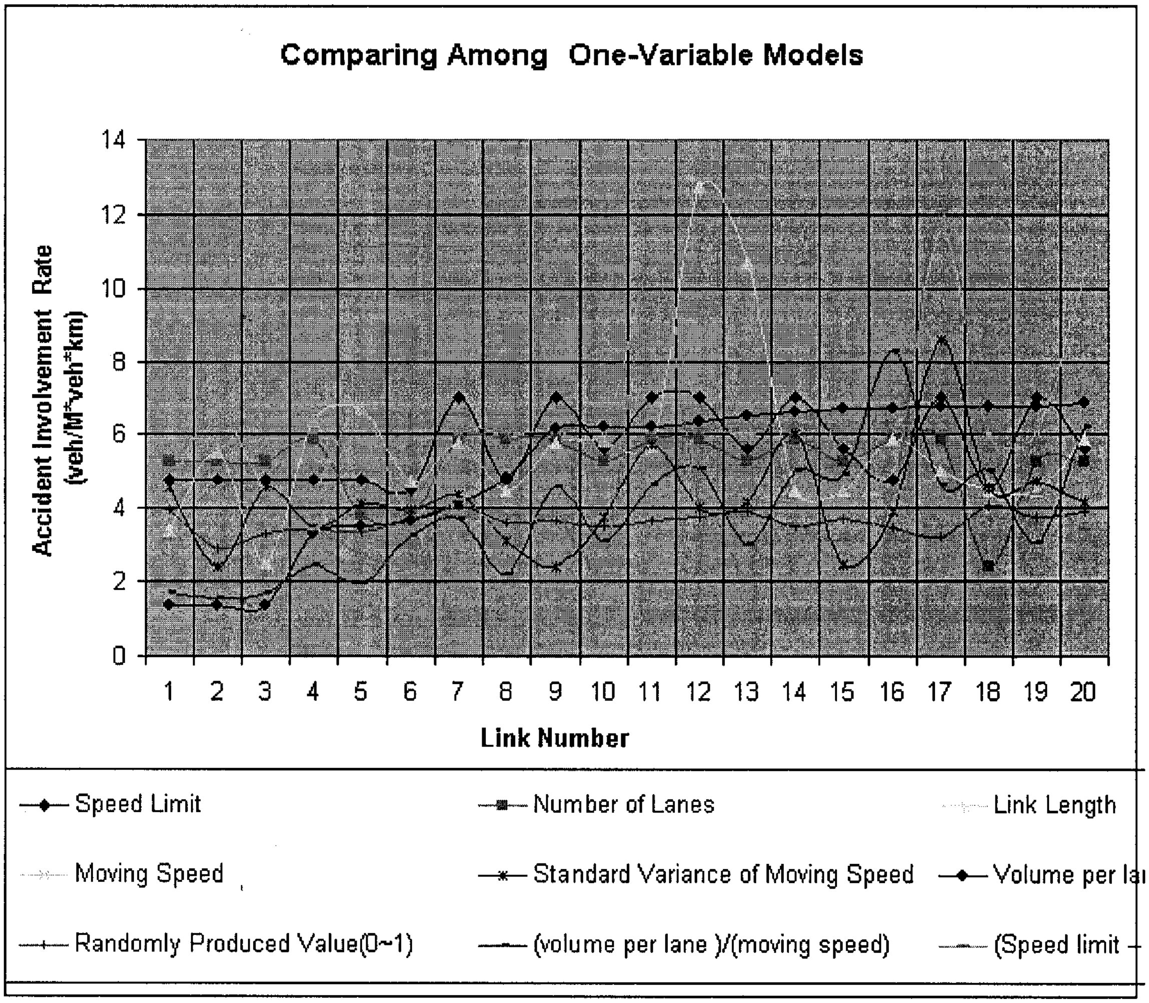


Figure 5.10 Accident Involvement Rates Estimated by Different One-Variable Models

6 EVALUATION OF MODEL PERFORMANCE

According to the model form in Table 5.1, when the parameters (b_r) vary, an infinite number of models can be produced, each with different performance. Therefore, to obtain the best model, a performance measure is required to help with the parameter optimization process.

6.1 A PROBLEM IN MODEL PERFORMANCE EVALUATION

Usually model performance is evaluated by comparing the model's outputs with the true values of the dependent variable. These true values are usually observed so they are considered correct.

However, accident risk can not be observed; it is estimated. Rarely there is a link on which many accidents happened under the same traffic characteristics. Researchers usually categorize each attribute so that in each category that attribute is homogeneous. From any combination of categorizations, a sample, or a set of independent variables and a dependent variable is produced. Such a sample is used for model development or model performance evaluation. However, the accuracy of such a sample depends on the degree of homogeneity which is also dependent on the data, the choice of attributes and the categorization. Often, the data is insufficient; the choice of attributes looks easy but is difficult; categorization is improper. All these cause inaccurate samples, so there is a big problem when these samples are taken as correct in model performance evaluation.

6.2 A MEASURE FOR ACCIDENT PREDICTION MODEL PERFORMANCE

A link has many attributes, some of which are highly correlated with traffic accident risk, while others are less or not. Here let me say the attributes that are highly correlated with accident risk are important attributes, and the links are homogenous if they are homogenous in aspect of their important variables.

After the threshold for "link neighbors" is determined, whether a model is good or not depends on homogeneity of the links in "link neighbors". Under a certain threshold for "link neighbors", the prediction will be more accurate if the links in "link neighbors" are more homogeneous.

A good model should give dispersive estimates for different links. This can be explained as follows. For any two different links whose accident risk rates are predicted by a good model, two sets of "link neighbors" will be produced. Although the links in each "link neighbors" are homogeneous, the links in one set of "link neighbors" and links in another set of "link neighbors" will be heterogeneous, so the estimated accident risk rates from these two "link neighbors" should be different. On the contrary, a poor-performance model is not so able to tell the big difference among different links, because in any set of "link neighbors" the links are more heterogeneous. The estimate results from a group of heterogeneous links tend to be close to an average estimated from all links in the dataset. Therefore, a poor-performance model can not give highly dispersive estimates. The more dispersive the estimates, the better the model. This situation can also be seen in results in Chapter 5.

There are nine one-variable models in last chapter, as shown in Figures 5.1~5.9. Apparently, the model in Figure 5.7, whose independent variable is a randomly produced variable, is the worst. Because the variable is uncorrelated with accident risk, the model can not tell the difference between different links. On the contrary, a much better model using variable of volume per lane in Figure 5.6 can make quite different estimates about accident risk for different links. Volume is considered as one of the most important variables in accident prediction models in literature. From these, it shows that a good model uses important variables, and makes the differences between different links; a poor-performance model gives estimates close to an average so that it can not distinguish amongst different links. This result is based on a condition that the threshold for "link neighbors" is the same.

In this research, dispersion is measured by standard deviation. Therefore, when N, the threshold for "link neighbors", is the same, the best model is the one whose outputs for all links in Predicted Dataset have the highest standard deviation. Standard deviation is calculated as formula (6.1).

$$\sigma = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \mu)^2} \qquad (6.1)$$

where, n=total number of links to be estimated in the Predicted Dataset;

y_j=accident risk rate of link j estimated by a model;

μ = mean of y_j.

Formula (6.1) should be improved because links have different lengths and volumes. Therefore link length, number of lanes and volume per lane are added as weights. The standard deviation of estimated accident involvement rate (denoted by SD) in this research is calculated as in formula (6.2).

$$\text{SD}=\sqrt{\frac{l_j * NL_j * VPL_j * \sum_{j=1}^{n}(y_j-\mu)^2}{\sum_{j=1}^{n}(l_j * NL_j * VPL_j)}}$$

(6.2)

where, n=total number of links to be estimated in Predicted Dataset;

l_j=length of link j;

NL_j=number of lanes of link j;

VPL=volume per lane of link j;

y_j= estimated accident involvement rate of link j;

μ = weighted mean of y_j, $\mu=\frac{\sum_{j=1}^{n}(l_j * NL_j * VPL_j * y_j)}{\sum_{j=1}^{n}(l_j * NL_j * VPL_j)}$.

The evaluation criterion for models is described as:

> For a model developed from Table 5.1, after the threshold for link neighbors is properly set, the performance of the model can be measured by the standard deviation of accident involvement rates estimated by the model for all links in Predicted Dataset, as calculated in (6.2). The greater the SD, the more dispersive the estimated accident involvement rates, then the better the model.

Table 6.2 lists the SD values of those 9 simplest models as shown in Figure 5.1~5.9.

Table 6.2 SD of One-Variable Models
(Threshold for Link Neighbors=90 Crashed-Veh)

Model	**Variables**	**b0**	**b1**	**b2**	**b3**	**b4**	**b5**	**b6**	**b7**	**b8**	**SD**
1	(0) Speed limit (km/h)	1	0	0	0	0	0	0	0	0	1.733
2	(1) Number of lanes	0	1	0	0	0	0	0	0	0	1.451
3	(2) Link length	0	0	1	0	0	0	0	0	0	2.094
4	(3) Moving speed	0	0	0	1	0	0	0	0	0	2.188
5	(4) Standard Variance of moving speed	0	0	0	0	1	0	0	0	0	1.491
6	(5) Volume per lane	0	0	0	0	0	1	0	0	0	2.49
7	(6) Randomly produced value	0	0	0	0	0	0	1	0	0	0.309*
8	(7)=(5)/(3)	0	0	0	0	0	0	0	1	0	2.256
9	(8)=(5)-(3)	0	0	0	0	0	0	0	0	1	1.728

According to evaluation criterion, among these 9 one-variable models, the model using the variable of volume per lane is the best one, while the model based on randomly produced value has the poorest performance for safety evaluation.

7 MODEL PERFORMANCE OPTIMIZATION

A model can be optimized by setting its parameters with proper values so that the model performance is maximized. Based on results of Chapter 5 and Chapter 6, the chapter is to determine the suitable threshold for “link neighbors” and attribute weights for the accident prediction model.

7.1 PARAMETERS TO BE SET IN THE MODEL

For a model as described in Table 5.1, two kinds of parameters should be set before the application of the model. They are threshold for “link neighbors” and weights (b_r).

Threshold for “link neighbors” should be set by balancing two different requirements:

1. It should be small enough so that the model’s output, the accident involvement rate (AIR), is sensitive to the changes of link attributes. In this case, SD will be high.
2. It should be large enough so that the model’s output is insensitive to the changes of threshold for “link neighbors”. That means a small change of the threshold would not affect estimated AIR much. In this case, the model prediction is stable.

Weights should be set so that the model can effectively compare the differences of AIR among different links; the standard deviation of the accident involvement rates (SD) estimated for links in Predicted Dataset should be as high as possible. The process of adjusting weights for the highest SD is model optimization, which requires an optimization algorithm.

7.2 ALGORITHMS FOR THE BEST MODEL PERFORMANCE

Table 7.1 lists algorithms of searching for suitable weights in (5.2) to obtain the highest performance measured in SD. Under such algorithms, the optimized weights are determined given` threshold for "link neighbors", and the initialized weights.

Table 7.1 Model Optimizing Algorithms

```
• Set threshold for "link neighbors" to a certain value
• Initialize each weight $b_i$. *
• Load Predicted Dataset.
• Load Predicting Dataset, calculate exposure weighted average of every attribute, $\bar{a}_i$.
• $SD_{max} := f(b_0, b_1, \ldots, b_8)$;
• SearchFailCount :=0;
• m :=1
• While m $\leq$ 1000
        • $\bar{b}$ := average of $b_i$
        • For each attribute i:
                $SD_i = f(b_0, b_1, \ldots, b_i + \frac{\bar{b}}{m}, b_{i+1 \ldots}, b_8)$
        • If $Max(SD_i) > SD_{max}$
            • Then $SD_{max} := Max(SD_i)$,
                if $SD_i = SD_{max}$,   $b_i := b_i + \frac{\bar{b}}{m}$
        • $\bar{b}$ := average of $b_i$
        • For each attribute i:
                $SD_i = f(b_0, b_1, \ldots, b_i - \frac{\bar{b}}{m}, b_{i+1 \ldots}, b_8)$
        • If $Max(SD_i) > SD_{max}$
            • Then $SD_{max} := Max(SD_i)$,
                if $SD_i = SD_{max}$,   $b_i := b_i - \frac{\bar{b}}{m}$
• Return SDmax and $b_i/\sum b_i$, i from 0 to 8;
```

* Let σ_i denote SD value in Table 6.2. To reduce optimization time, b_i was initialized as a function of σ_i, because the variable with higher σ_i is more important and is expected to have a higher value of b_i in the optimal model. To know whether different initial b_i will affect optimization result or not, 10 different sets(n=0,1,...9) of b_i were tried. In each set, b_i was initialized as $(\sigma_i)^{0.3*n}$.

Table 7.2 shows the optimal (b_r), in which weights were initialized into different values($\sigma_i^{0.3*n}$, see table 7.1) and thresholds of neighbors were set as 90 and 60 crashed vehicles, respectively. The main calculation structure of obtaining results in Table 7.2 is as shown in Figure 7.1.

Table 7.2 Final Weights Searched from Different Start Points*

Threshold for Link Neighbors	n*	Weights									SD	AIRmax**	AIRmin**
		b_0	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8			
90	0	.0000	.1171	.1739	.1846	.0004	.3794	.0000	.1447	.0000	3.045	18.06	1.32
	1	.0004	.1107	.1723	.1722	.0000	.4105	.0000	.1336	.0002	3.043	17.33	1.32
	2	.0000	.1303	.1591	.2083	.0000	.3303	.0001	.1720	.0000	3.038	18.55	1.35
	3	.0000	.1162	.1714	.1707	.0000	.4012	.0000	.1405	.0000	3.043	17.85	1.32
	4	.0000	.1105	.1716	.1998	.0000	.3901	.0002	.1277	.0000	3.042	18.39	1.33
	5	.0000	.1134	.1658	.1892	.0000	.3889	.0017	.1401	.0009	3.043	17.99	1.32
	6	.0000	.1172	.1733	.1703	.0000	.3896	.0017	.1479	.0000	3.043	17.71	1.33
	7	.0004	.1314	.1503	.1987	.0000	.3394	.0000	.1796	.0000	3.037	17.99	1.34
	8	.0002	.1153	.1775	.1690	.0000	.3800	.0004	.1576	.0000	3.044	17.52	1.32
	9	.0000	.1167	.1732	.1845	.0000	.3800	.0009	.1447	.0000	3.044	17.91	1.32
60	0	0.003	0.083	0.182	0.134	0.000	0.595	0.000	0.004	0.000	3.245	18.58	1.04
	1	0.000	0.083	0.200	0.131	0.000	0.585	0.001	0.000	0.000	3.242	19.28	1.04
	2	0.002	0.086	0.182	0.137	0.001	0.585	0.001	0.006	0.000	3.247	19.18	1.04
	3	0.000	0.084	0.167	0.144	0.000	0.602	0.004	0.000	0.000	3.242	18.39	1.04
	4	0.000	0.085	0.201	0.130	0.000	0.581	0.001	0.002	0.000	3.241	19.26	1.04
	5	0.000	0.086	0.184	0.138	0.000	0.590	0.002	0.000	0.000	3.246	19.27	1.04
	6	0.000	0.083	0.178	0.134	0.000	0.595	0.000	0.009	0.000	3.244	18.45	1.04
	7	0.004	0.082	0.185	0.131	0.000	0.594	0.003	0.000	0.000	3.244	18.40	1.04
	8	0.000	0.086	0.184	0.138	0.000	0.590	0.002	0.000	0.000	3.248	18.58	1.04
	9	0.000	0.085	0.184	0.136	0.000	0.592	0.004	0.000	0.000	3.248	19.27	1.04

* Initial $b_i = (\sigma_i)^{0..3*n}$.

** Highest and lowest accident involvement rates estimated from Predicted Dataset.

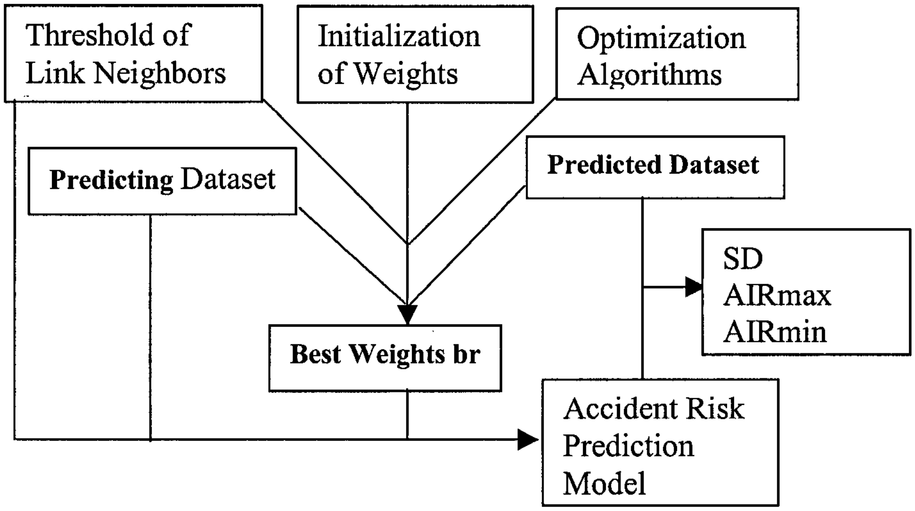


Figure 7.1 Structure Flow for Optimization of Weights and Calculation of SD, AIR_{max} and AIR_{min}

It was found from Table 7.4 that if the threshold for "link neighbors" changed, br, SD, AIRmax and AIRmin would all change much, no matter the initialization of weights is same or not. However, when threshold for "link neighbors" was same but initialization of weights changed, br, SD, AIRmax and AIRmin did not change much. Therefore, we can conclude that SD, AIRmax and AIRmin are determined only by the threshold for "link neighbors".

7.3 DETERMINATION OF THRESHOLD FOR LINK NEIGHBORS

Since values of SD, AIR_{max} and AIR_{min} are mainly determined by threshold for "link neighbors", a suitable threshold for "link neighbors" can be selected from its relationships with SD, AIR$_{max}$ and AIR$_{min}$, so that the model under that threshold for "link neighbors" performs well.

Figure 7.2~7.4 show the relationships between threshold for "link neighbors" and the above three measures. The relationships were obtained from algorithms in Figure 7.1 and Table 7.1, where the threshold was set as from 10 to 145, with step=5.

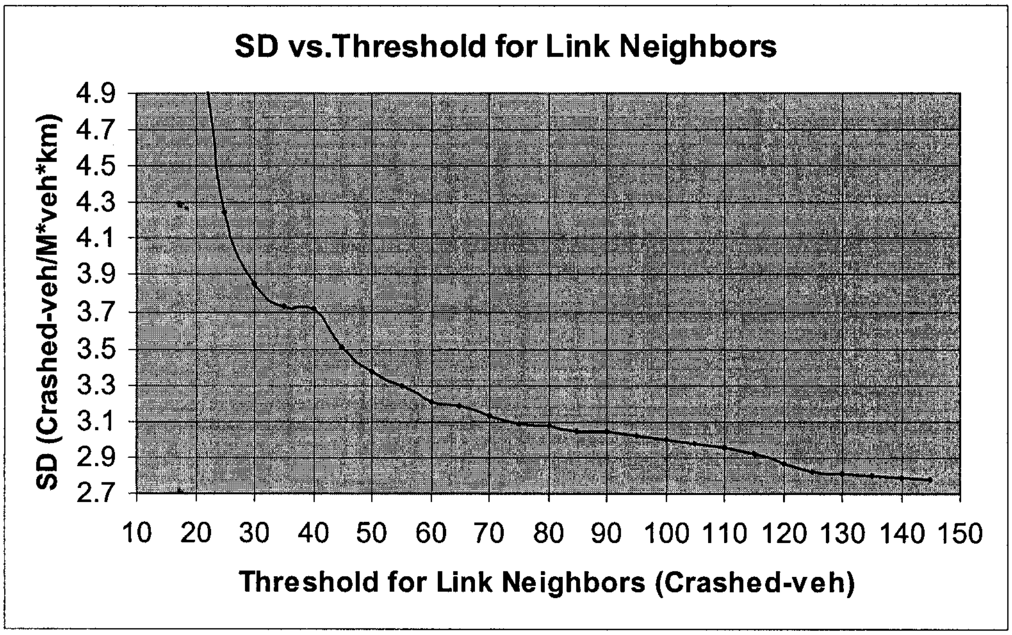


Figure 7.2 SD Versus Threshold for Link Neighbors

If a model can estimate accident risk steadily, then to increase or decrease the threshold for "link neighbors" a little should not greatly affect estimates for most links, so the SD should not change much. Otherwise, the model's estimates should be taken as unstable. From Figure 7.2, we can see that if the threshold for "link neighbors" is less than 80, the model's estimate is unstable. When the threshold is around 90, the model seems much better. However, if the threshold is over 90, the model outputs are insensitive to attributes. Therefore, to let the model be sensitive to changes of attributes, but insensitive to changes of threshold, the threshold of 90 is a good choice.

It should be also true for the highest or lowest estimated accident involvement rates. If a model's estimate is stable, the estimates for the safest or least safe link should not fluctuate much after threshold for "link neighbors" changes a little. From Figure 7.3, it shows that the model is not reliable if the threshold is less than 45 and not very reliable if the threshold is between 45 and 70. When the threshold is between 85 and 100, the model prediction seems stable; It is same in Figure 7.4,

Based on above results and analyses, the threshold was set as 90 crashed vehicles.

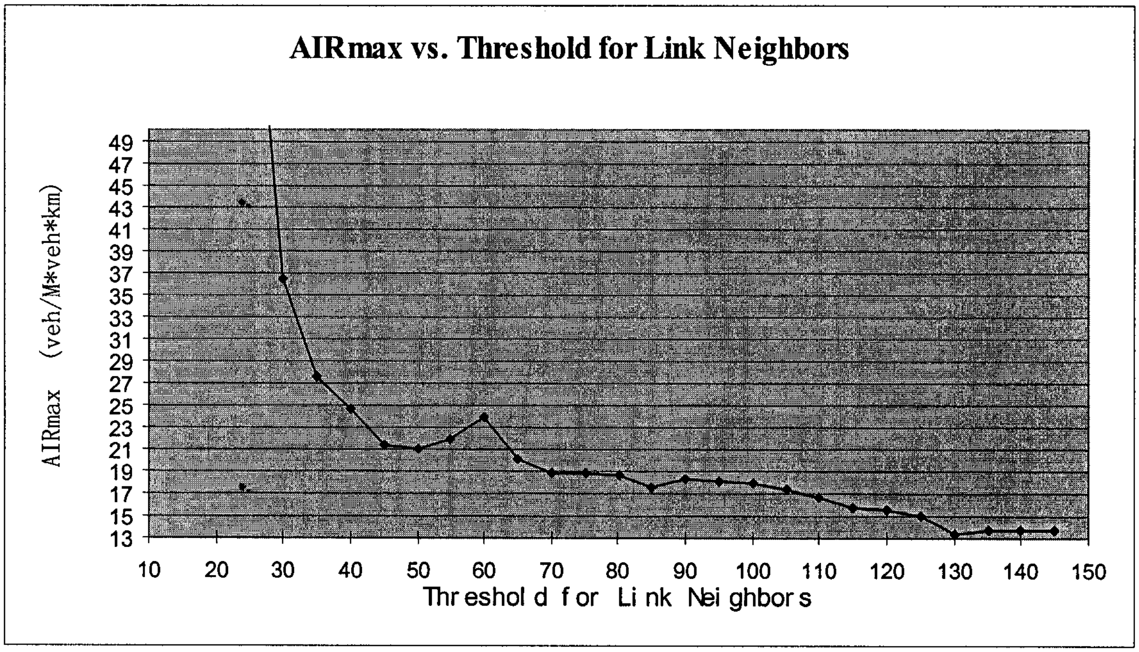


Figure 7.3 AIRmax versus Threshold for Link Neighbors

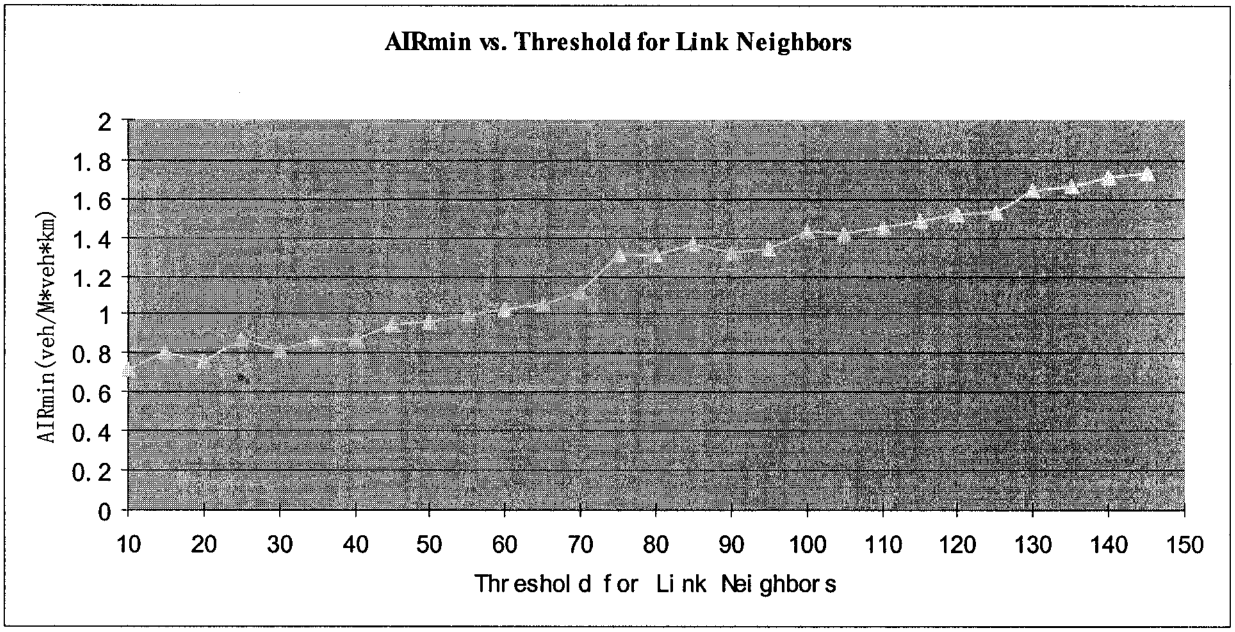


Figure 7.4 AIRmin versus Threshold for Link Neighbors

7.4 DETERMINATION OF WEIGHTS

When the threshold for "link neighbors" is 90, Table 7.4 shows that 4 weights are always 0 or near 0 in all 10 models. To simplify modeling, all these 4 weights were set to 0.

With algorithms in Table 7.1 and the threshold of 90, all other 5 non-zero weights were optimized again. The final results are shown in Table 7.3.

Table 7.3 shows that although weights were initialized to different values, the optimized weights in all 10 models were very similar. Because Model 1 has the highest SD, it is chosen as the model we seek. Definitely, if more optimization work is done, higher SD may be obtained, but gain of SD may be very limited, and the weights will be similar as those in model 1. Actually, the difference between Model 1 and Model 2 is so small that we cannot say that one is apparently better than the other. The case should be the same for the real best model and Model 1. Therefore, it is reasonable that Model 1 is taken as the best model.

Table 7.3 Optimization Results Using Five Variables (Sorted by SD)

Model	n*	SD	br[1]**	br[2]**	br[3]**	br[5]**	br[7]**
1	**0**	**3.04497**	**0.11712**	**0.17396**	**0.18466**	**0.37955**	**0.14471**
2	2.7	3.04433	0.11687	0.17331	0.18466	0.38032	0.14484
3	2.4	3.04348	0.11535	0.17765	0.16915	0.38029	0.15755
4	0.3	3.04306	0.11077	0.17246	0.17230	0.41080	0.13367
5	0.9	3.04249	0.11622	0.17142	0.17066	0.40123	0.14047
6	1.5	3.04245	0.11368	0.16623	0.18965	0.38994	0.14050
7	1.8	3.04231	0.11740	0.17361	0.17058	0.39028	0.14812
8	1.2	3.04209	0.11054	0.17168	0.19984	0.39021	0.12773
9	0.6	3.03829	0.13032	0.15910	0.20829	0.33030	0.17199
10	2.1	3.03652	0.13149	0.15041	0.19884	0.33957	0.17969

* See Table 7.1

** See Table 5.1 and 5.2

8 RESULTS AND DISCUSSION

8.1 VARIABLES USED IN THE MODEL

8.1.1 Results

Based on the accident, traffic and geographic datasets, and on the optimization algorithms in Table 7.1 and concept of SD in (6.2), the best model, which is the most efficient but also is stable, is established in the form described in Table 5.1. The model includes four independent variables. They are: number of lanes, link length, average moving speed and average volume per lane. One composite variable is also included in the model. It is (average volume per lane) / (average moving speed).

The parameters (see Table 5.1) of the model are listed in Table 8.1.

Table 8.1 Parameters of Accident Risk Prediction Model in Table 5.1

Variable	br	$\overline{a_r}$
Number of lanes	0.11712	2.7997
Link length	0.17396	1559.2m
Moving speed (peak hour)	0.18466	70.867 km/h
Volume per lane	0.37955	1084.3 veh/h/lane
(Volume per lane) / (Moving speed)	0.14471	14.491 veh/km/lane
Threshold for "link neighbors": 90 crashed vehicles in reportable accidents		

8.1.2 Discussion

Among nine variables, four variables and one composite variable remain in the model while other four do not. That means, for a given dataset and a given model form, the model may not perform the best with all variables in. It also indicates that the best model is multi-variable, not one-variable.

The final weights of those variables are the balancing result among homogeneity degrees of important link attributes in "link neighbors". In one-variable model, for example, the links in "link neighbors" are relatively highly homogeneous in one attribute, while highly different in other attributes. If another variable is added to the model, to let the links be relatively homogeneous in the second attribute, some links in the "link neighbors" will be eliminated, while other links originally outside will join in. As a result, compared with the "link neighbors" in one-variable model, the "link neighbors" in two-variable model have higher degree of homogeneity in second variable, but lower in the first variable. This is the same for multi-variable model. When a new variable is added, the variance of any other variable among the links in the "link neighbors" will increase. The task of optimization algorithm in this research is to select variables and assign weights so that the links in "link neighbors" reach the highest degree of homogeneity in consideration of all available important link attributes.

From the weight b_r, the importance of a variable can be roughly measured. Volume per lane seems the most important variable, while all other excluded variables, whose weights are 0, are useless. Therefore among the "link neighbors" of a link, the relative range of an important variable will be narrower than the relative range of a less important variable. This result increases homogeneity of important variables in the "link neighbors" from where a link's accident risk is estimated.

8.2 THE SAFEST AND LEAST SAFE LINKS

8.2.1 Results

The Accident Involvement rates (AIRs) of 1,966 links in the dataset were evaluated with this model. Table 8.2 and 8.3 list the 20 safest and 20 least safe links, respectively.

Table 8.2 The Safest 20 Links Evaluated by the Model

Road Category	speed limit (km/h)	Variables					Accident Involvement rate (Crashed-veh per M*veh*km)
		A	B	C	D	E	
		Number of lanes	Link Length (m)	Moving Speed (km/h)	Volume/lane (veh/h/lane)	D/C	
Expressways	90	2.4	726	85	1257	15	1.32
Major Arterials	50	2.4	567	50	1160	23	1.33
Expressways	90	4.0	220	85	1372	16	1.34
Expressways	90	3.0	667	89	1292	15	1.34
Ramps	50	1.0	379	55	1215	22	1.34
Expressways	90	4.0	776	95	1198	13	1.36
Expressways	90	4.5	1221	95	1332	14	1.36
Expressways	90	3.0	939	97	1331	14	1.37
Expressways	90	3.5	241	80	1534	19	1.38
Expressways	90	3.0	302	91	1355	15	1.41
Major Arterials	70	3.0	73	42	1153	28	1.43
Ramps	50	1.0	433	48	1290	27	1.44
Expressways	90	4.0	590	88	1021	12	1.44
Expressways	90	2.0	755	72	1711	24	1.44
Major Arterials	70	3.0	29	52	1281	25	1.45
Major Arterials	50	1.0	458	51	1136	22	1.46
Major Arterials	70	3.0	29	57	1282	22	1.46
Major Arterials	50	3.0	131	52	1328	25	1.47
Expressways	90	4.0	534	89	1800	20	1.47
Expressways	90	2.0	491	94	1067	11	1.47

Table 8.3 The Least Safe 20 Links Evaluated by the Model

Road Category	speed limit (km/h)	Variables					Accident Involvement rate (Crashed-veh per M*veh*km)
		A	B	C	D	E	
		Number of lanes	Link Length (m)	Moving Speed (km/h)	Volume/lane (veh/h/lane)	D/C	
Minor Arterials	50	2.2	172	41	110	3	18.06
Collector Roads	40	2.0	231	43	115	3	17.73
Transit Arterials	30	2.0	314	40	102	3	17.55
Local Roads	40	2.0	160	43	115	3	17.29
Collector Roads	50	2.0	172	41	120	3	17.14
Collector Roads	40	2.0	231	42	103	2	17.08
Minor Arterials	40	2.0	223	43	126	3	17.05
Transit Arterials	30	2.0	317	39	118	3	16.92
Transit Arterials	30	2.0	274	40	127	3	16.81
Minor Arterials	50	2.4	166	43	110	3	16.63
Collector Roads	40	2.0	126	41	107	3	16.32
Collector Roads	40	2.0	120	42	103	2	16.31
Minor Arterials	50	2.0	89	43	126	3	16.26
Transit Arterials	30	2.0	359	39	118	3	16.24
Minor Arterials	50	2.0	97	41	111	3	16.22
Transit Arterials	30	2.0	170	39	125	3	16.08
Transit Arterials	30	2.0	185	39	109	3	16.08
Minor Arterials	50	2.0	128	44	104	2	15.95
Minor Arterials	40	2.0	223	43	137	3	15.87
Major Arterials	50	2.3	127	37	109	3	15.83

8.2.2 Discussion

The safest 20 links belong to high-class roads, with over 1000veh/h/lane of volume per lane. On the contrary, the least safe 20 links belong mostly to low-class roads, with volume of a little over 100 veh/h/lane and moving speed of about 40km/h; these links usually have 2 lanes. The phenomenon of high accident risk on 2-lane low-class roads with light traffic can be explained as follows.

1. Low-class roads provide poor road and traffic conditions for driving, while those conditions on high-class roads are much better.

2. Although the 2-lane low-class roads provide poor driving conditions, the low traffic volume and low speed make drivers overestimate safety of the traffic conditions. On high-class roads, because the volume is relatively high, drivers will be more alert.

3. Traffic speed on low-class roads is much lower than on high-class roads, so drivers have to spend more time to travel same distance on low-class roads than on high-class roads.

4. Volume and traffic safety will affect each other. Apparently safer roads with good geometric features will attract more drivers than less safe roads.

Therefore, it is easy to understand that more time on overestimated poor driving-condition roads is much less safe than less time on correctly-estimated good roads.

8.3 ACCIDENT PREDICTION BY ATTRIBUTES

8.3.1 Results

Figure 8.1 to 8.5 shows the distribution of AIR estimated for all links in the Predicted Dataset versus variables.

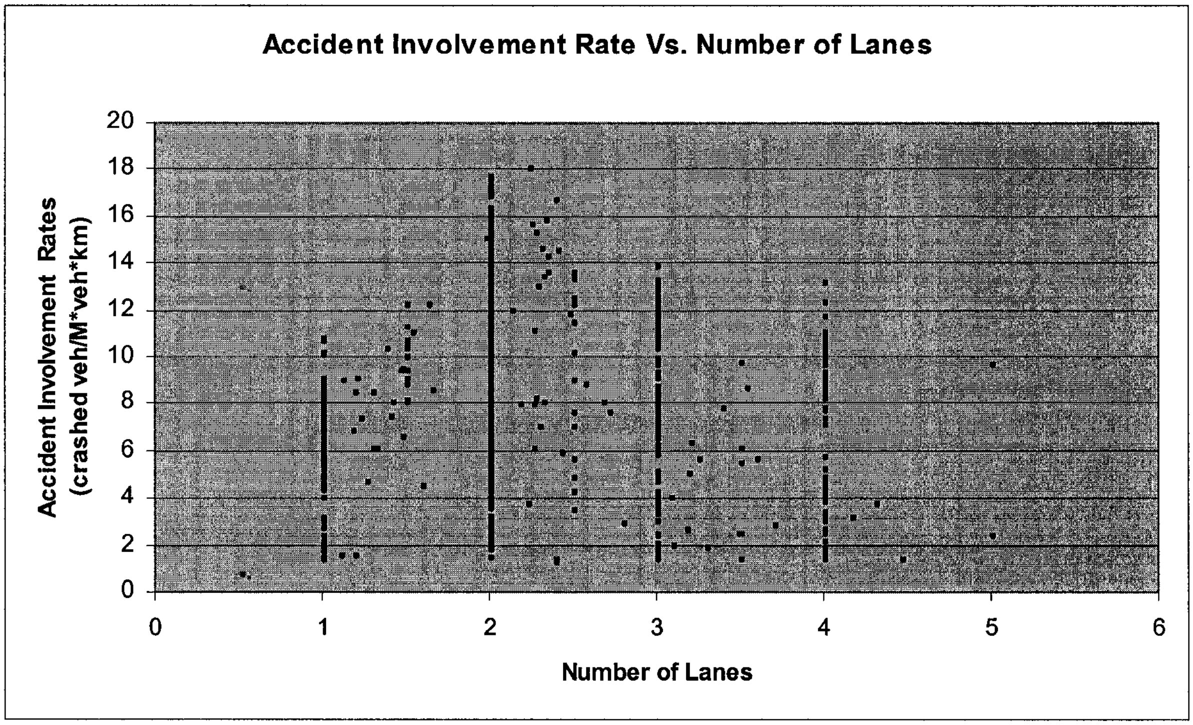


Figure 8.1 Estimates of Accident Involvement Rate versus Number of Lanes

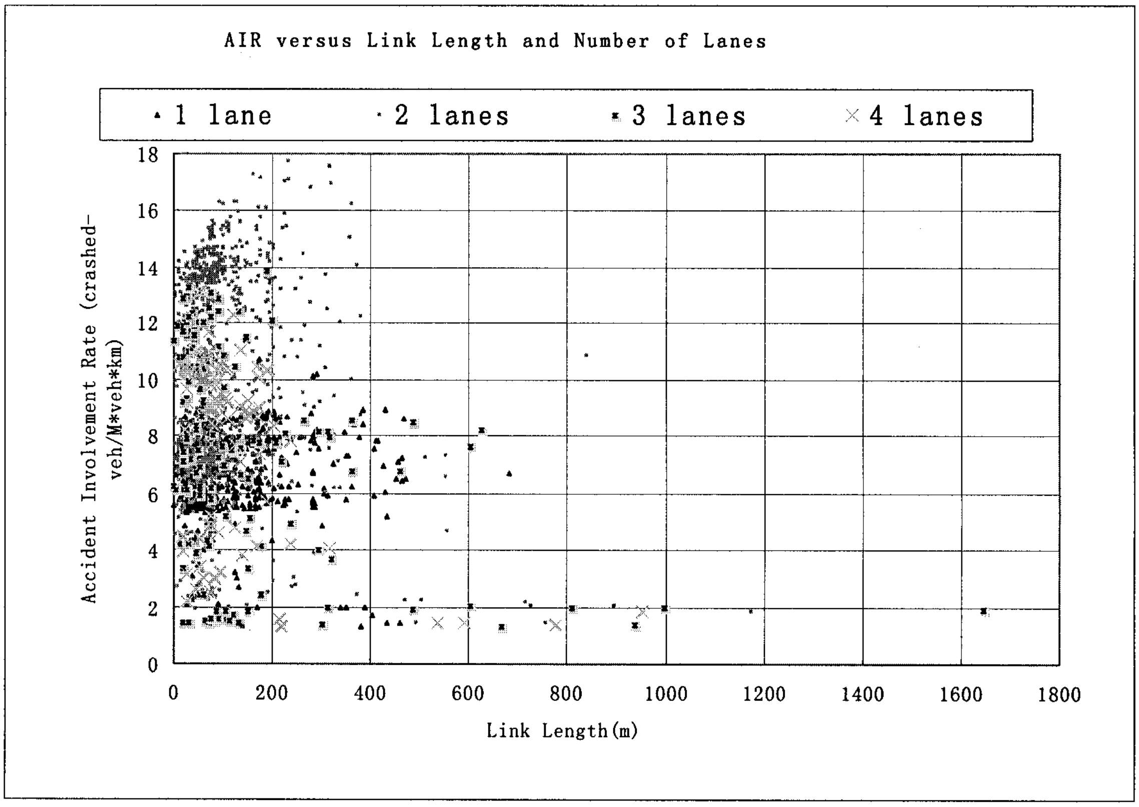


Figure 8.2 Estimates of Accident Involvement Rate versus Link Length and Number of lanes

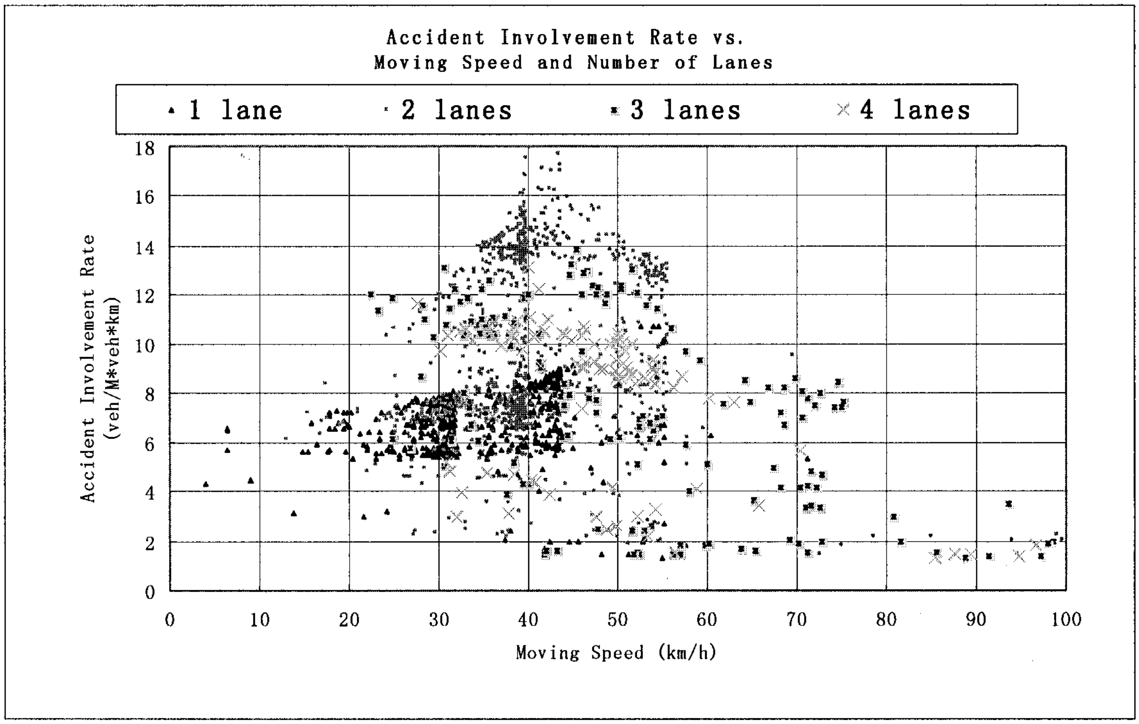


Figure 8.3 Estimates of Accident Involvement rate Versus Average Moving Speed

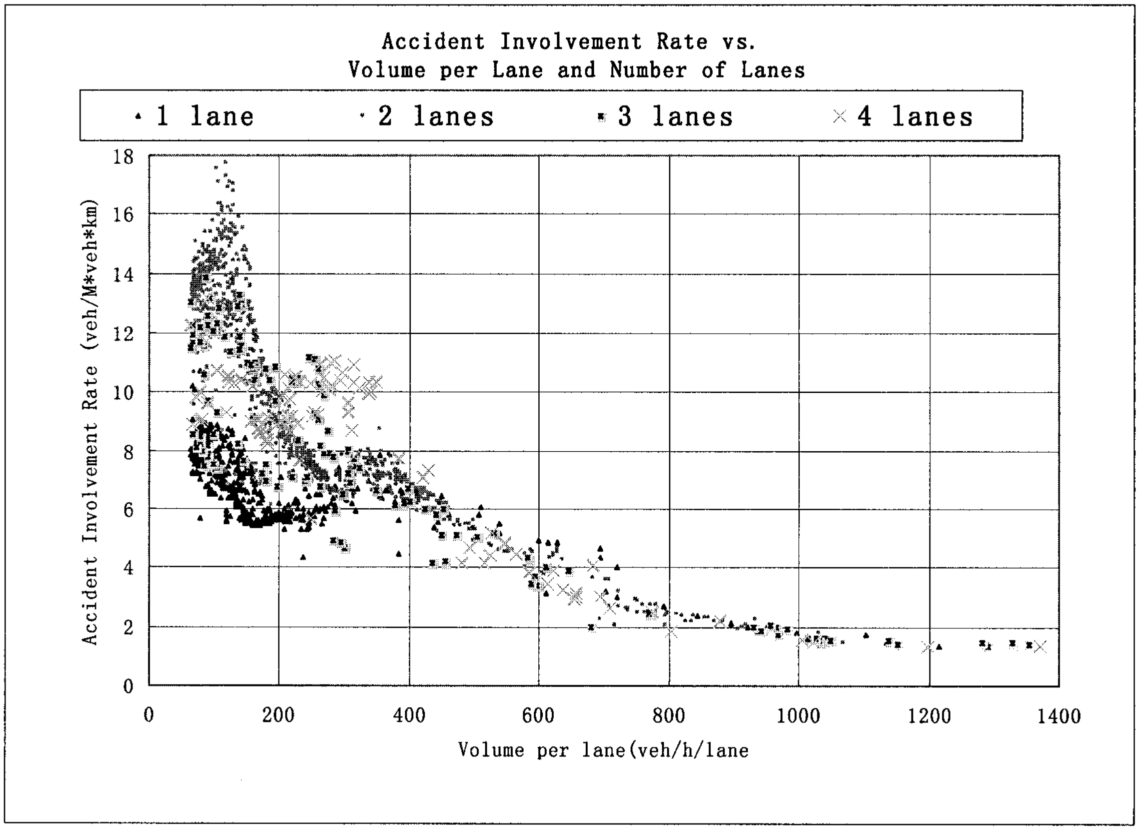


Figure 8.4 Estimates of Accident Involvement Rate Versus Volume Per Lane and Number of Lanes

From above four figures, we do not know the detailed trends of accident risk with links of different attributes, so in the following 4 figures(Figure 8.5~8.8), three attributes are categorized. In each combination of categorization, we can obtain the accident risk trends with links of the different fourth attribute.

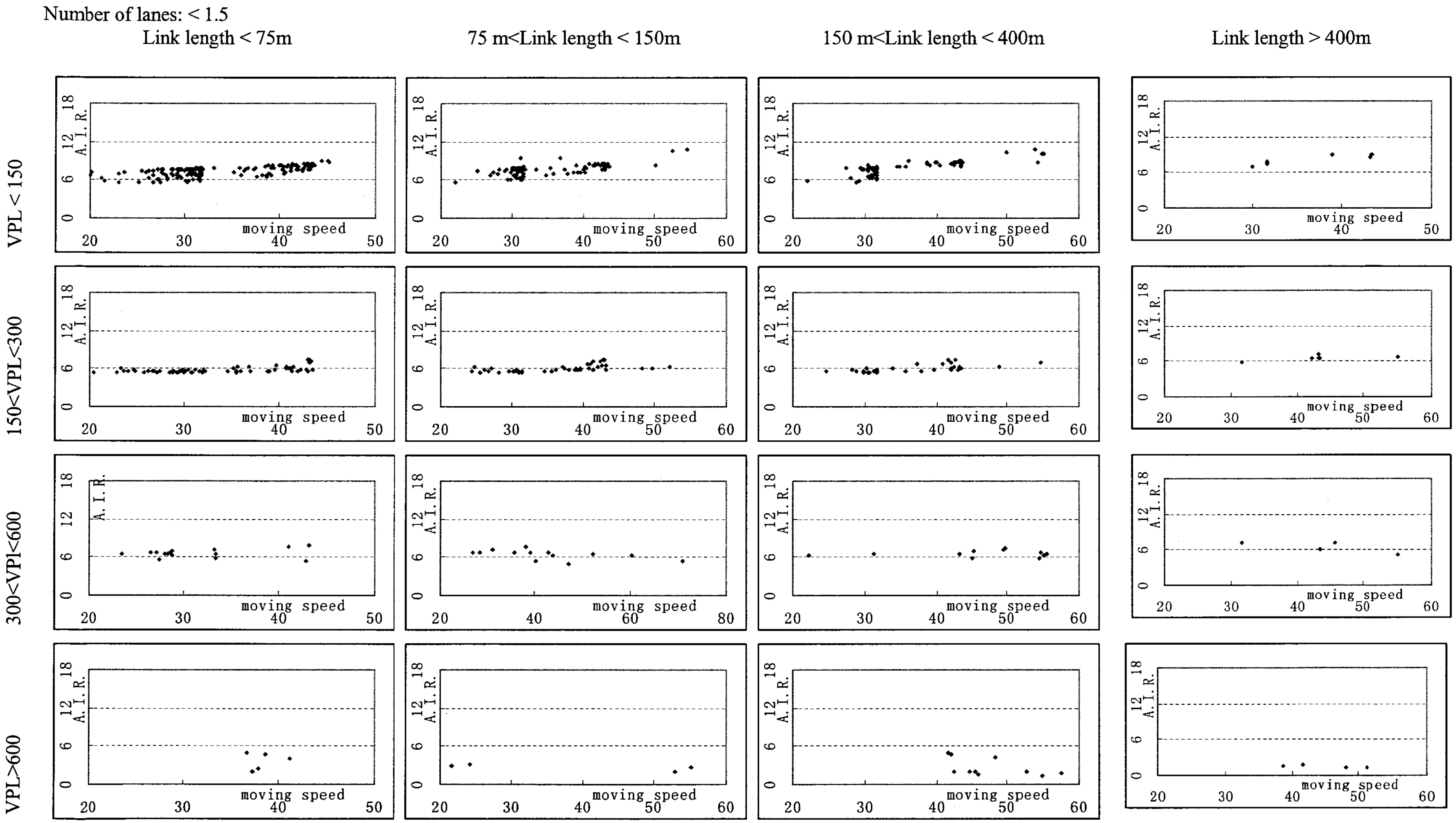


Figure 8.5 Accident Risk versus Moving Speed, Volume per Lane and Link Length for One-Lane Links

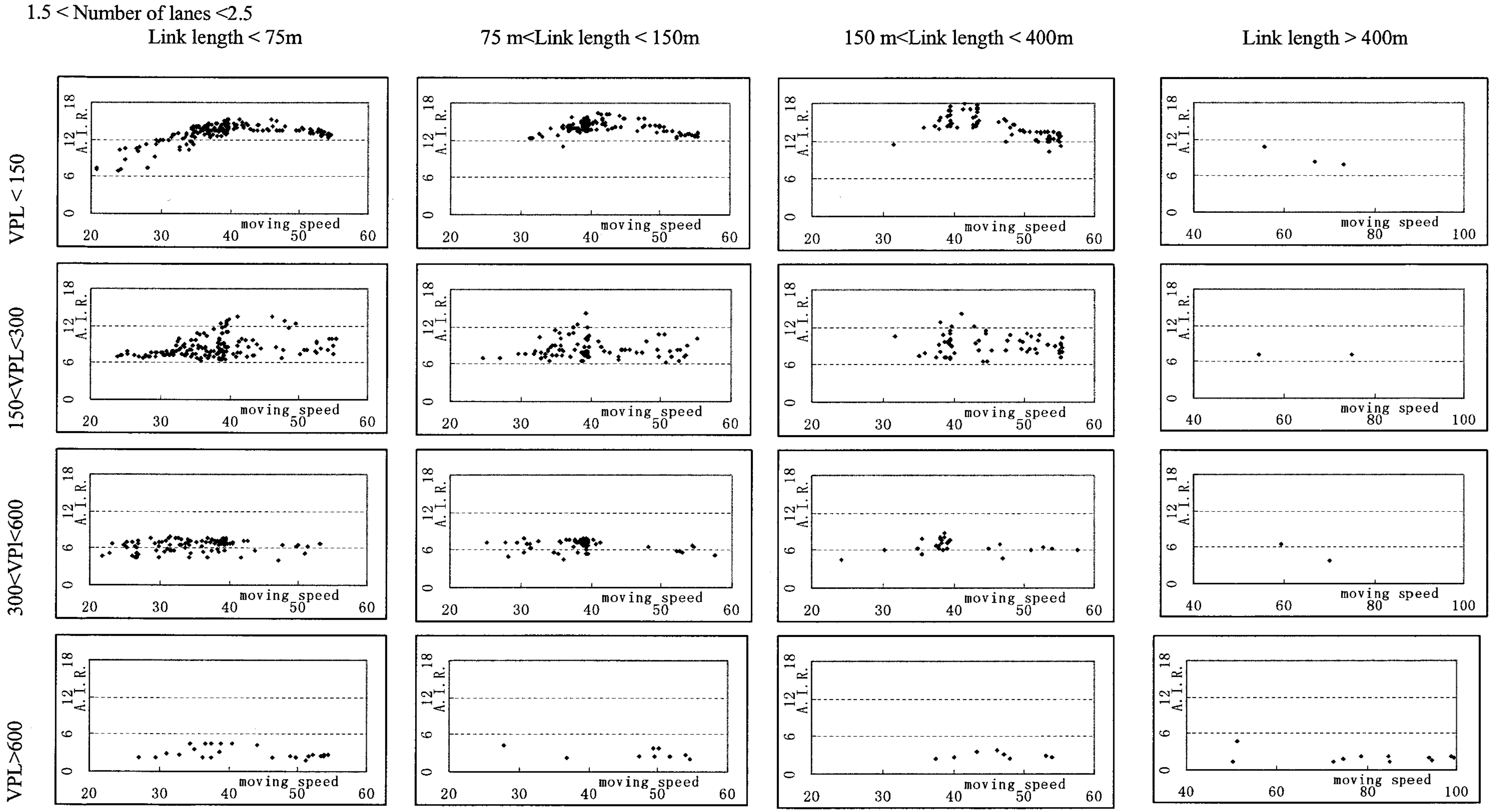


Figure 8.6 Accident Risk versus Moving Speed, Volume per Lane and Link Length for Two-Lane Links

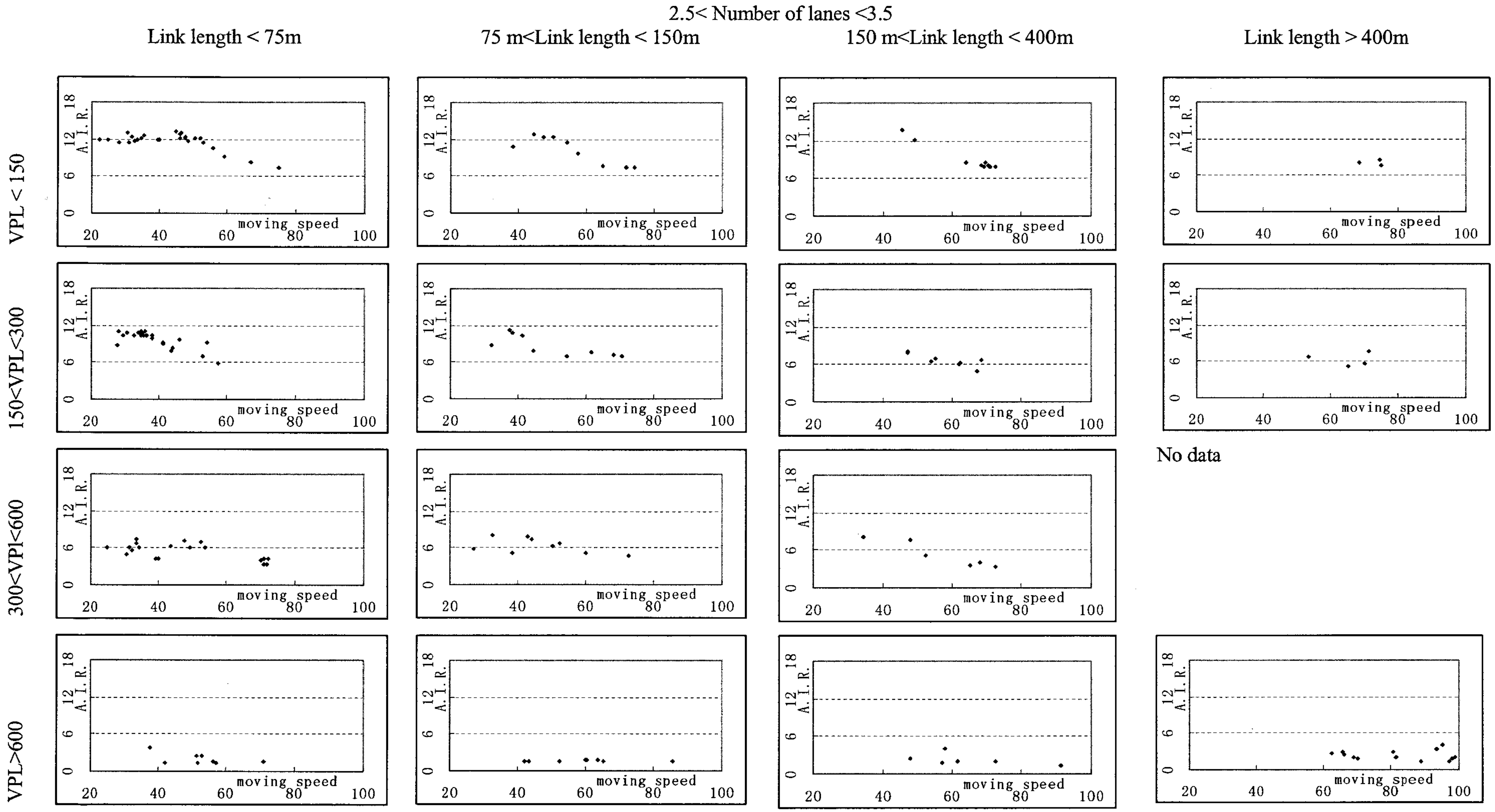


Figure 8.7 Accident Risk versus Moving Speed, Volume per Lane and Link Length for Three-Lane Links

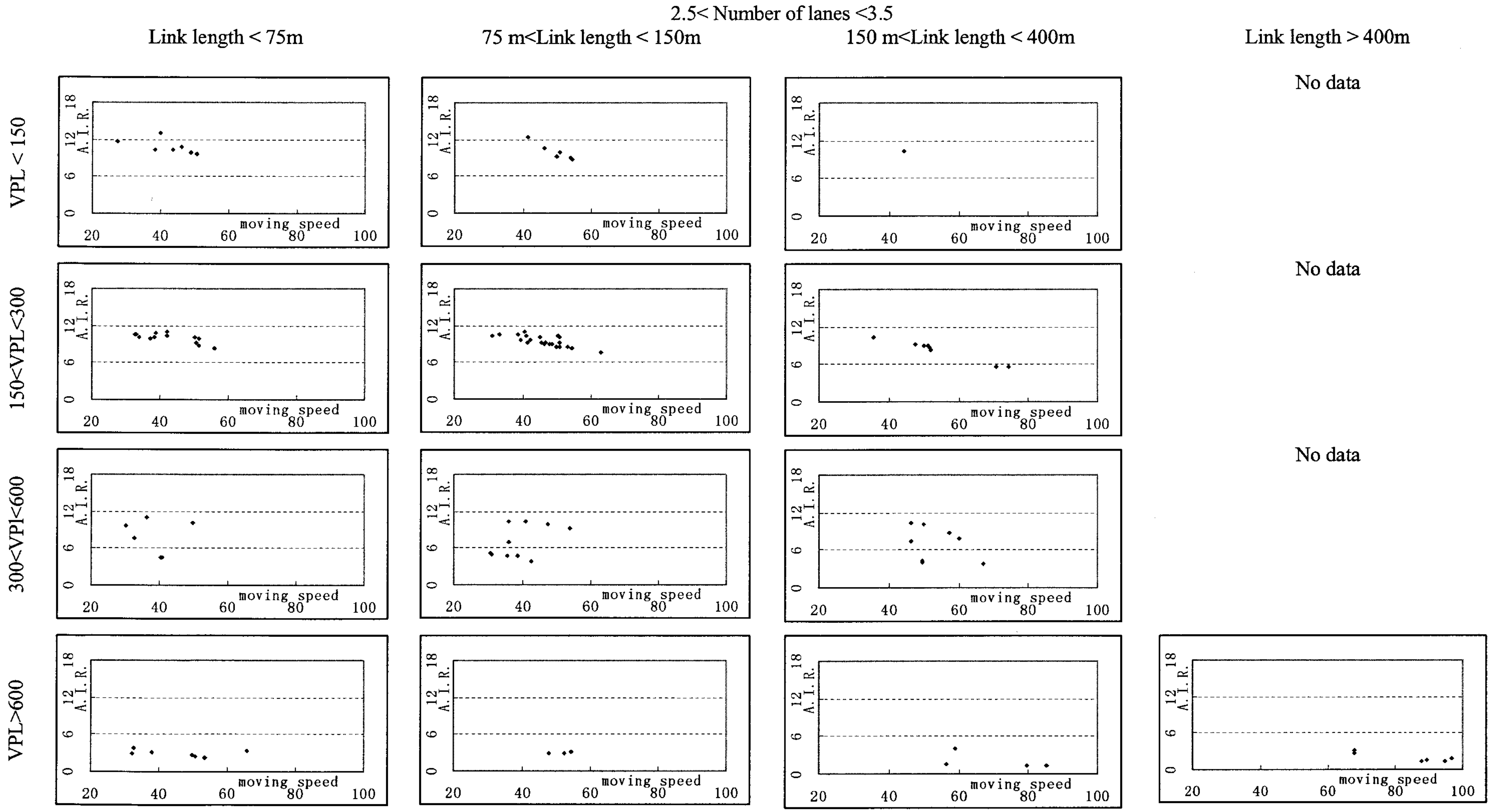


Figure 8.8 Accident Risk versus Moving Speed, Volume per Lane and Link Length for Four-Lane Links

From above 8 figures, we can predict accident risk according to link attributes as following:

1. Among the links with same number of lanes, link length and moving speed, the links with high volume per lane are safer(See Figure 8.4~8.8).

2. Links whose volume per lane is over 1000 veh/h/lane have similar accident risk rates, and they are the safest, no matter what values of other three attributes(See Figure 8.4).

3. Among links with volume per lanes less than 200 veh/h/lane, one-lane links are the safest, while two-lane links are the least safe(See Figure 8.4).

4. Among links with volume per lane between 200 and 400 veh/h/lane, four-lane links are the least safe(See Figure 8.4).

5. Among one-lane links, link length has little effect on traffic safety(See Figure 8.5).

6. Among multi-lane links, if the link length is less than 400m, link length has little effect on safety. If all other attributes are the same, links over 400m long are safer than those with length less than 400m, (See Figure 8.6~8.8).

7. Among one-lane links whose volumes are less than 150 veh/h/lane, links with higher moving speed are less safe, if all other attributes are the same(See Figure 8.5).

8. Among two-lane links whose volume is less than 300 veh/h/lane, the links are the least safe if their moving speeds are around 40km/h(See Figure 8.6).

9. Among 3-lane and 4-lane links, the links with high moving speed are usually safer, if all other attributes are the same(See Figure 8.7~8.8).

8.3.2 Discussion

Based on one attribute, in most cases it is difficult to predict traffic safety. Usually three or 4 attributes are needed to give a precise accident risk estimate. However, from Figure 8.4, it seems that if a link's volume and number of lanes are given, we can estimate the accident risk roughly.

The correlation between an important attribute and traffic safety is not same for all kinds of links. For some links, the correlation is positive, while for other links, it may be zero or negative. For some links, the correlation is strong, while for other links, it may become weak. Therefore, it is difficult to use a mathematic equation to express accident risk as a function of all four attributes.

8.4 COMPARING MODELS WITH DIFFERENT NUMBER OF VARIABLES

8.4.1 Results

If the threshold for links is still set as 90 crashed-veh, but one or more variables are excluded before optimization, models with different number of variables can be established. Table 8.4 lists the models with the highest SD in various numbers of variables.

Table 8.4 Best Models with Different Numbers of Variables

		Models with Highest SD				
		1 variable	**2 variables**	**3 variables**	**4 variables**	**>4 variable**
Weight	(1)Volume per lane	1.000	0.871	0.871	0.514	0.380
	(2)Number of lanes	/	0.129	0.129	0.116	0.117
	(3)Moving speed	/	/	/	0.200	0.185
	(4)Link length	/	/	/	0.170	0.174
	(5)=(1)/(3)	/	/	0.000	/	0.145
SD		2.490	3.016	3.016	3.024	3.045

To compare differences of accident risk estimates between the more-than-4-variable model(i.e., the best model obtained) and other models, the cumulative percentage of links versus estimate difference was calculated as shown in Figure 8.6.

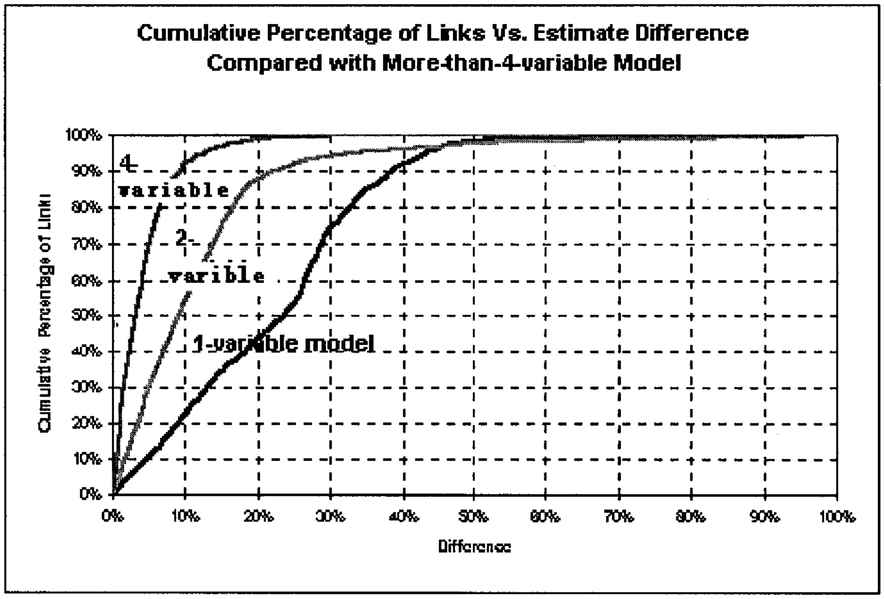


Figure 8.6 Comparison between More-than-4-variable Model and Other Models

8.4.2 Discussion

From Table 8.4, apparently the most important variable among 6 variables is volume per lane. The two most important variables are volume per lane and number of lanes. Just by this single variable, maybe its explanatory power for accident risk estimation is not as strong as other variables, but it is the best one to be combined with volume per lane to form a two-variable model. The reason may be that the number of lanes has the least correlation with the variable of volume per lane.

Figure 8.6 shows that the estimate difference between 1-variable model and more-than-4-variable model is great. As the number of variables in a model increases, the model performance is closer to that of the more-than-4-variable model.

9 CONCLUSIONS

Traffic simulation on computer is a good way to collect traffic data. It saves time and money. Traffic simulation not only is efficient but also makes it possible to obtain large sample of traffic data.

The concept of "link neighbors" formed in this research has many advantages.

1. Based on "link neighbors", the accident risk of a link can be directly estimated from a dataset with traffic, geometric and accident data.
2. The estimation accuracy of accident risk is improved by increasing the homogeneity of important variables among links in "link neighbors". This is realized by adjusting attribute weights in "link distance".
3. The proper threshold for "link neighbors" makes the size of "link neighbors" be at a reasonable value which is not too small and not too large.
4. It is not necessary to establish a mathematical formula which is usually error-prone.
5. Unlike most of other accident prediction models, the model based on "link neighbors" is applicable for all kinds of roads in Waterfront Network of Toronto.

SD, or the standard deviation of the estimated accident involvement rates, is a good performance measure for accident prediction models. The higher the SD, the better the model. Parameters in a model can be optimized so that the model has the highest SD. Therefore, unlike other modeling approaches, the one in this research does not require variable analyses. All kinds of variables, whether considered as important or not, correlated with each other or not, can be added into the optimization model.

Accident risk of a link depends on many geometric and traffic attributes. However, not all attributes will be included in the best model. Based on the dataset, the model includes volume per lane, number of lanes, moving speed and link length.

The correlation between an attribute and traffic safety may change as other link attributes change. It is difficult to use a mathematic equation to correctly express accident risk as a function of link attributes.

The model makes accident risk estimates based on accident data and exposure of the "link neighbors". The links in "link neighbors" should be homogeneous in attributes which are correlated with traffic safety. The more homogeneous those attributes, the better the estimates. If a link whose traffic and geometric features is far different from most of the links in "link neighbors", the estimate may be incorrect. Therefore, the model may not be reliable for other networks, or off-peak period of the study network. To alleviate this problem, more data are required.

The model only predicts reportable accident involvement rates. The severity difference among accidents has not been considered.

REFERENCES

1. Transport Canada. Intelligent Transportation Systems. http://www.its-sti.gc.ca/

2. Transport Canada. ITS Architecture for Canada, Release 1.1. Last updated: 2002-11-13 http://www.its-sti.gc.ca/Architecture/english/web/mpATIS4.htm

3. Kaufman, D.E., J. Nonis, and R.L. Smith. "A Mixed Integer Linear Programming Model for Dynamic Route Guidance". Transportation Research: Part B, Methodology, Vol. 32B, No. 6, August 1998, pp. 431–440.

4. Joseph L. Schofer, Frank S. Koppelman and William A. Charlton, "Perspectives on Driver Preferences for Dynamic Route Guidance Systems," presented at the 76th Annual meeting of the Transportation Research Board, 1996, Transportation Research Record 1588, pp. 26-31.

5. Wunderlich, Kaufman, Smith, "Improved Link Travel Time Prediction for Decentralized Route Guidance Architectures", IEEE Transactions on ITS, March 2000.

6. Chatterjee, K. and McDonald, M. "The network safety effects of dynamic route guidance". ITS Journal, 4(3), 1998.

7. Lord, D. "The prediction of Accidents on Digital Networks: Characteristics and Issues Related to the Prediction of Accident Prediction Models". Ph.D. Thesis, Department of Civil Engineering, University of Toronto, Toronto, Canada.

8. Look, H.; and Abdulhai, B. "Accident Risk Assessment Using Microsimulation for Dynamic Route Guidance". 80th Annual Meeting of the Transportation Research Board, 2001.

9. Liu, G. X. and A. Popoff. Provincial-Wide Travel Speed and Traffic Safety Study in Sasketchewan. InTransportation Research Record 1595, TRB, National Research Council, Washington, D.C., 1997, pp8-13. (studying the relationship between vehicle speed and collisions)

10. Garber N.J. and Subramanyan S. Incorporating Crash Risk in Selecting Congestion-Mitigation Strategies. Transportation Research Record 1746, PP. 1-5 (number of crashes versus occupancy, freeway)

11. Persaud, B. and Dzbik, L. Accident Prediction Models for Freeways. Transportation Research Record, 1401, Transportation Research Board, National Research Council, Washington D.C., 1993, p55-60. (use both macroscopic(average daily traffic) and microscopic(hourly volume) data)

12. Garber, N. J., and A. Ehrhart. Effect of Speed, Flow, and Geometric Characteristics on Crash Rates for Different Types of Virginia Highways. Report 00-R15. Virginia Transportation Research Council, Charlottesville, Virginia, 2000

13. Lee C., Saccomanno F. and Hellinga B. Analysis of Crash Precursors on Instrumented Freeways Analysis of Crash Precursors on Instrumented Freeways. 81st Annual Meeting of the Transportation Research Board, 2001.

14. Sawalha, Z., Sayed, T. (2001) "Evaluating the Safety of Urban Arterial Roadways", Journal of Transportation Engineering, ASCE, Vol. 127(2), pp. 151-158.

15. Persaud, B. N. Estimating Accident Potential of Ontario Road Sections. Transportation Research Record, 1937, Transportation Research Board, National Research Council, Washington D.C., 1991, pp47-53.

16. Resende, P. T. V., Benekohal, R.F. "Development of Volume-to-Capacity Based Accident Prediction Models," Proceedings of Traffic Congestion and Traffic Safety in the 21st Century conference (Chicago), ASCE, P. 215-221, June 1997.