Diagnosis and Prediction of Traffic Congestion on Urban Road Networks Using Bayesian Networks

Jiwon Kim and Guangxing Wang

This paper proposes a Bayesian network (BN) analysis approach to modeling the probabilistic dependency structure of causes of congestion on a particular road segment and analyzing the probability of traffic congestion given various roadway condition scenarios. A BN approach was used to encode the joint probability distribution over a set of random variables that described scenario variables, which represented factors affecting the congestion level of a target segment such as time of day, incident, weather, and traffic states on adjacent links, as well as output variables, which represented traffic performance measures of the target segment such as flow, density, and speed. The study developed a method to build a BN model according to historical traffic and event data and demonstrated the BN-based traffic analysis with a study network in Brisbane, Queensland, Australia. The paper discusses applications of the proposed BN model in urban traffic congestion management, by focusing on identifying leading causes for congestion diagnosis and identifying critical scenarios for congestion prediction.

In dealing with traffic congestion problems, three main questions asked by traffic managers: Why is this congestion occurring? Can we anticipate when it will occur? and How can we prevent or mitigate it? The first question leads to addressing a congestion diagnosis problem, which aims to understand the causes of a particular congestion situation and identify the most critical causes; the second question leads to addressing a congestion prediction problem, which aims to predict the occurrence of traffic congestion on the basis of the observations of the causes. These two problems are closely connected and together provide important grounds for answering the third question; that is, the ability to accurately diagnose and predict traffic congestion allows traffic managers to identify and closely monitor the root causes of congestion and therefore proactively detect and manage congestion hot spots.

The general objective of this study was to apply a probabilistic graphical modeling approach and machine learning techniques to solve congestion diagnosis and prediction problems in urban road networks. For that purpose, this paper proposes a Bayesian network (BN) model that is capable of capturing the probabilistic dependency structure of causes of congestion on a particular road segment and assessing the probability of traffic congestion given various roadway condition scenarios.

Faculty of Engineering, Architecture, and Information Technology, University of Queensland, Room 555, Level 5, Advanced Engineering Building 49, Staff House Road, Brisbane, Saint Lucia, Queensland 4072, Australia. Corresponding author: J. Kim, jiwon.kim@uq.edu.au.

Transportation Research Record: Journal of the Transportation Research Board, No. 2595, Transportation Research Board, Washington, D.C., 2016, pp. 108–118.
DOI: 10.3141/2595-12

BAYESIAN NETWORKS

A BN is a probabilistic graphical model that represents probabilistic relationships between a set of variables via a directed acyclic graph (*1–3*). A BN consists of a set of nodes and a set of arcs, in which nodes represent random variables and arcs connecting pairs of nodes represent direct dependencies between variables. In general, constructing a BN model requires the following three steps: (*a*) defining variables (nodes), (*b*) specifying structure (arcs), and (*c*) specifying parameters (conditional probability distribution for each node). The second step is to determine a qualitative property of the BN approach, which is causality or dependence relationships between variables, and the third step is to determine the quantitative part, which consists of probability distributions that quantify these relationships. Once a set of nodes of a BN are defined, specifying its structure and parameters can be done in two ways: manual specification based on domain expert knowledge or automatic specification using machine learning techniques. In this study, the approach taken is to manually specify the network structure while learning parameters from data. Once built, a BN provides a compact representation of the full joint probability distribution over its variables, which allows one to compute the probability of each state of a node conditioned on any subset of other variables. This process is called probabilistic inference—computing the posterior distribution of variables X given evidence e, $P(X|e)$—and there are a number of efficient exact and approximate inference algorithms for performing complex probabilistic reasoning tasks in a BN approach. Reasoning can be performed in two different directions, that is, from known causes to unknown effects (predictive reasoning) and from known effects to unknown causes (diagnostic reasoning). These features make the BN a powerful tool for diagnosing and predicting traffic congestion under uncertainty. Examples of questions that can be answered by a BN model include what is the probability of having severe congestion at a particular link when there is rain (congestion prediction)? and what is the probability that there is rain if severe congestion is observed (congestion diagnosis)?

While the use of the BN model in transportation research areas is relatively new, there are a number of areas in which the BN approach has been applied. Two main areas of application are traffic estimation and forecasting and congestion prediction (*4–9*) and accident detection and crash prediction (*10–13*).

METHOD

Building a BN Model for Traffic Congestion Diagnosis and Prediction

This paper considers a BN model that represents the dependency structure of link-level measures. For a given link, a BN model is designed that describes relationships between link performance measures (e.g., flow, density, and speed) and external factors that affect the target link (e.g., time of day, weather, and incident). The goal of this model is to assess the effects of external factors on traffic conditions on a target link. Thus, the effects of upstream or downstream traffic conditions are not considered in this model. The BN model in this study is considered to be static in that the model represents a time-independent knowledge of dependency relationships between variables (that is, long-term average patterns). To take into account the temporal dimension, one may use dynamic BNs. This paper uses a static BN; for more information about dynamic BNs, see Murphy (*14*).

Variables

Variables used in the proposed BN model are presented in Table 1. A total of 12 variables were selected. In this study, only discrete variables will be considered, that is, nodes that take discrete values. The variables are categorized into three groups: network environment, external event, and traffic condition variables, as follows:

- Network environment variables represent networkwide environmental factors, such as link direction, day of week, time of day, and weather conditions. Direction (DR) specifies the direction of a link, which in this paper takes two values {southbound and northbound} as will be described in the case study below. Day of week (*D*) takes two discrete states {weekday and weekend}; time of day (*H*) takes five states {morning, a.m. peak, off peak, p.m. peak, and night}; and weather (*W*) takes three states based on rain intensity {clear, light rain, moderate rain, and heavy rain}. The detailed descriptions for the state definitions are presented in Table 1.
- External event variables represent events or activities that are external to the traffic stream itself and cause interruptions to traffic flow, such as incidents, work zones, and traffic control signals. In this study, only incident factors are included, and they are defined as three different variables: incident on a target link (*I_O*), incident on upstream links (*I_U*), and incident on downstream links (*I_D*). The incident variables take two states: {no incident and incident}, for which the "incident" state indicates that there is at least one incident occurrence detected during a measurement interval of link traffic parameters.
- Traffic condition variables represent link performance measures describing traffic states on a target link. This study includes five variables consisting of three basic traffic stream parameters, flow (*F*), occupancy (*O*), and speed (*S*), and two indicators, level-of-service (*L*) and congestion indicator (*C*). Flow (*F*), occupancy (*O*), and speed (*S*) take four discrete states {very low, low, high, and very high}. These states are defined according to the respective value range of each variable as depicted in Figure 1. For each link, the diagrams of flow–occupancy–speed relationships are first plotted, and the maximum value for each parameter is identified. By dividing the parameter values by the associated maximum value, one obtains normalized flow, occupancy, and speed values, which range between 0 and 1, as shown in Figure 1. Then, the range [0, 1] is divided into four equal parts, and the states {very low, low, high, and very high} are assigned accordingly. Level of service (*L*) takes five states {A, B, C, D, and E–F}, where Level E and Level F are merged into one state because the frequency of Level F is too low to be categorized as a separate state. The congestion indicator (*C*) is a binary variable that indicates whether the given link is congested; it takes two states {uncongested and congested}. Occupancy and flow values are used to determine the value of *C*: "uncongested" if occupancy < Occ_{crit} and "congested" if occupancy ≥ Occ_{crit}, where Occ_{crit} represents the critical occupancy at which the link flow becomes maximum, as shown in Figure 1*a*. This binary congestion indicator *C* will be used as a main target variable in the BN in performing the congestion diagnosis and prediction analysis.

TABLE 1 Variables and State Definitions for Proposed BN Model

Variable by Node Group	Description	States and State Definitions
Group 1. Network Environment		
DR	Direction	Southbound Northbound
D	Day of week	Weekend: Saturday, Sunday Weekday: Monday–Friday
H	Time of day	Morning: 1 a.m.–6 a.m. (5 h) a.m. peak: 6 a.m.–10 a.m. (4 h) Off-peak: 10 a.m.–4 p.m. (6 h) p.m. peak: 4 p.m.–8 p.m. (4 h) Night: 8 p.m.–1 a.m. (5 h)
W	Weather	Clear: 0 mm/h Light rain: <2.5 mm/h Moderate rain: 2.5–7.6 mm/h Heavy rain: ≥7.6 mm/h
Group 2. External Event		
I_U	Incident on upstream links	No incident Incident
I_O	Incident on a target link	No incident Incident
I_D	Incident on downstream links	No incident Incident
Group 3. Traffic Condition		
F	Flow (vphpl)	Very low: <0.25[a] Low: 0.25–0.5 High: 0.5–0.75 Very high: ≥0.75
O	Occupancy (%)	Very low: <0.25[a] Low: 0.25–0.5 High: 0.5–0.75 Very high: ≥0.75
S	Speed (km/h)	Very low: <0.25[a] Low: 0.25–0.5 High: 0.5–0.75 Very high: ≥0.75
L	Level of service (LOS)	A B C D E–F
C	Congestion indicator	Uncongested: occupancy < Occ_{crit} Congested: occupancy ≥ Occ_{crit}

NOTE: vphpl = vehicles per hour per lane.
[a]Based on normalized parameter values, which range between 0 and 1.

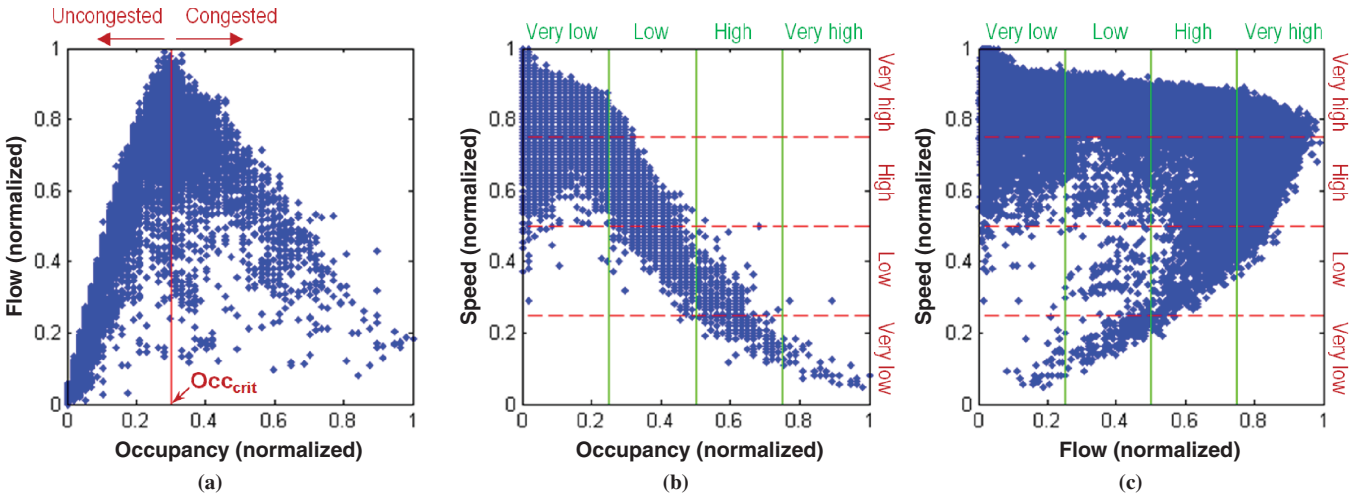

FIGURE 1 Example of data discretization for variables *F*, *O*, *S*, and *C*: (*a*) occupancy by flow, (*b*) occupancy by speed, and (*c*) flow by speed.

Structure

Once the nodes of the BN are specified, the next step is to specify qualitative relationships between variables. While the structure of the BN can be learned from data by searching for the best model with the use of various learning algorithms, this paper adopts the approach of manually specifying a set of alternative model structures and selecting the best model from them (*1, 15–18*). This approach allows one to incorporate knowledge about the variables into the model building process and compare different configuration assumptions more clearly. To develop candidate models systematically, first relationships between node groups are determined and then the group-level relationships are applied to the individual nodes. For instance, if one believes that Group 1 affects (or causes) Group 2, then one adds arcs from all nodes in Group 1 to all nodes in Group 2. Given the three node groups defined above, the following two observations can be made:

Group 1 (network environment) affects Group 2 (external event) or Group 3 (traffic condition) but not the other way around (e.g., weather can cause an incident but an incident cannot cause a weather event).

Group 2 (external event) and Group 3 (traffic condition) are dependent on each other, but the causal relationship can be in either direction, that is, either Group 2 affects Group 3 or Group 3 affects Group 2.

On the basis of these observations, a total of seven possible network configurations can be identified, as shown in Figure 2.

Of these seven configurations, type (*g*) in Figure 2 is considered as the main model structure as it seems to be reasonable to assume that the environment variables directly affect incident occurrence and traffic conditions while there will also be a direct influence of an incident occurrence on link traffic. To keep the model structure simple, the variables in each node group are assumed to be conditionally independent given their common parent or common descendant, that is, no direct arcs between nodes in the group. This assumption is apparent in Group 1 as variables DR, *D*, *H*, and *W* do not affect one another. For Group 2, the indication is that the incident occurrence on a link is influenced by the incident occurrence on its upstream or downstream link only through a third variable, such as network environment and traffic condition, and this assumption seems to be reasonable. The conditional independence assumption is also viewed as acceptable for the variables in Group 3, as those are all different measures of the same link traffic condition and it is not assumed that one variable causes another in this group. The final graph representation for the proposed BN model is thus as shown in Figure 3. The model validation results are presented in a later section.

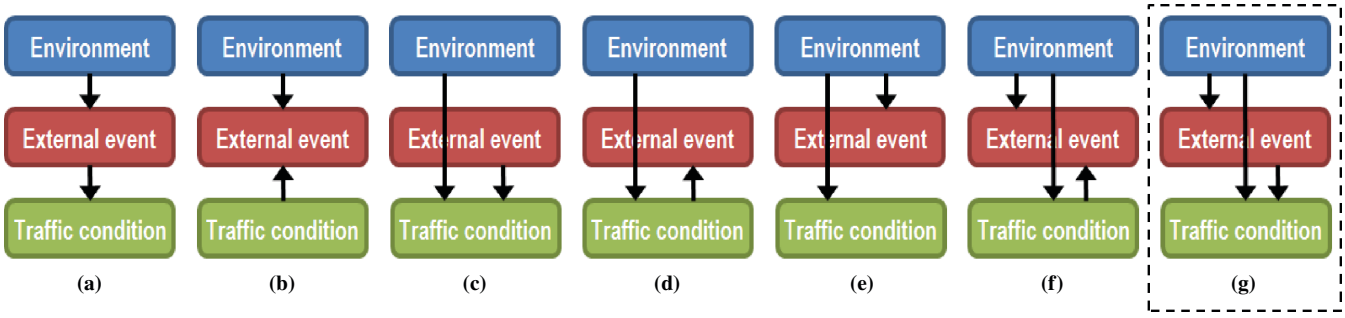

FIGURE 2 Seven possible configurations for proposed BN model.

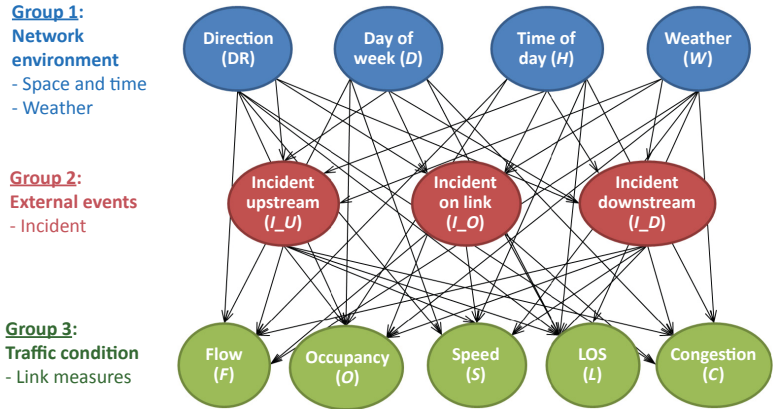

FIGURE 3 Selected graph representation for proposed BN model (configuration type *g*).

Parameter Learning

Once the structure of the BN is determined, the next step is to quantify the relationships between connected nodes. This step is done by computing a conditional probability distribution for each node, after considering all possible combinations of values of its parent nodes. For discrete variables, conditional probability distribution is expressed in the form of a conditional probability table (CPT), in which each element of the table represents the probability that a given variable takes a particular value given a particular combination of its parent node values. For instance, the CPT of Node F represents the probability values of all possible configurations of $P(F = f|D = d, H = h, W = w, I\_U = iu, I\_O = io,$ and $I\_D = id)$. For a node that does not have parents, the CPT becomes the marginal distribution of the node itself. The size of the CPT can become very large if a node has many parents or if the parents can take many states. Thus, in many real-world applications it is not feasible to calculate CPTs manually and instead, machine learning techniques are used to learn the CPTs from the data automatically.

EXPERIMENTAL SETUP AND RESULTS

Study Site

This section presents a case study implementing the proposed BN approach. The study area is Brisbane, Queensland, Australia, and the study sites are 19 highway links selected from Pacific Motorway in Brisbane, as shown in Figure 4. Table 2 presents the basic information about these 19 links.

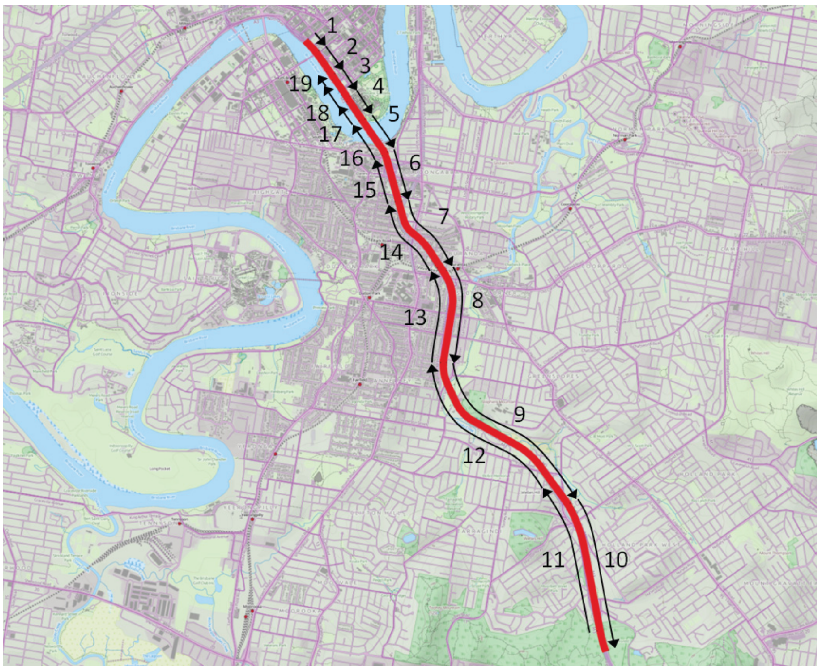

FIGURE 4 Selected links for case study.

TABLE 2 Description for Selected Links

Link	Length (m)	Number of Lanes	Design Speed (km/h)
Pacific Motorway Southbound			
1	143	2	100
2	361	3	70
3	334	3	70
4	346	4	100
5	520	4	100
6	642	3	90
7	1,116	3	100
8	1,177	3	100
9	2,500	3	100
10	1,925	3	100
Pacific Motorway Northbound			
11	1,988	3	100
12	2,047	3	100
13	1,333	3	100
14	1,049	4	90
15	612	3	90
16	580	4	100
17	283	4	100
18	329	3	70
19	116	3	70

Data Collection and Discretization

Traffic and incident data were obtained from the Queensland Department of Transport and Main Roads through the public traffic data system, which provided link measures, including flow, occupancy, speed, and level of service from available loop detectors around the South East Queensland network and the list of incidents reported to the Queensland Department of Transport and Main Roads' incident management system. Link measures in the public traffic data system's traffic data were recorded every 3 min. Weather data were received from the Bureau of Meteorology station located in the study area. Weather data reported weather parameters, including precipitation, visibility, temperature, and humidity, which were recorded every 30 min. For the case study, traffic, incident, and weather data were collected from 608 days between January 1, 2011, and September 30, 2013. These data were used to learn the parameters of the BN model in Figure 3. To match the data with the model, the data should be organized in a matrix form, in which columns represent the 12 variables defined in Table 1 and rows represent observations for these variables for the entire study period covering all study links. To accomplish that task, weather and incident data were mapped to traffic data such that each 3-min traffic observation (link measures) had its associated weather and incident information attached. After the data were fused in this way, a 4,787,932-by-12 matrix for all 19 links was obtained; each row of the matrix represented a 3-min traffic, incident, and weather observation. Since the data were originally numerical and continuous, the data were discretized on the categorical states defined in Table 1. After the discretization, the matrices of discrete (categorical) data converted from the original numeric matrices were obtained. An example of the processed data matrix is provided in Figure 5.

Model Validation

Once the model was built, it was important to assess how good the selected BN model was (that is, how well the model described the underlying data), compared with the other alternatives. Two different tests were performed to assess the goodness of fit of a BN: (*a*) measuring a network-level scoring function from the fitted BN model and (*b*) measuring a variable-specific classification error by performing cross validation. A number of scoring functions were available for assessing a BN structure (*19*). Four well-known scores were used: log likelihood, Akaike's information criterion (AIC), Bayesian information criterion (BIC), and K2 (*20*). For measuring classification error, *k*-fold cross validation was performed; for the cross validation, the original data set was randomly partitioned into *k*-equal size subsamples and the model was trained with $k - 1$ subsamples (called the training set) and then validated on the one remaining sample (called the testing set). This procedure was repeated *k* times. The classification error for a particular variable was measured as the percentage that the BN model predicted a wrong state for this variable during testing, averaged over the *k*-folds. In this study $k = 5$ was used. Figure 6 shows the network scores and classification errors obtained from the seven BN structures in Figure 2, including the one selected for this study (type *g*). Figure 6*a* presents the network scores; the higher the score, the better the model fits the data. Overall, the scores were low in *a* and *b* and high in *c*, *d*, *e*, *f*, and *g*. By closely looking at the numbers, one finds that the score of *g* is the highest of the seven in AIC and K2, the second highest in log likelihood, and the third highest in the BIC. Figure 6*b* presents the classification errors tested for five variables, *F*, *S*, *L*, *O*, and *C*; the lower the error, the better a learned model predicts the state of a

DR	D	H	W	I_U	I_O	I_D	F	O	S	L	C
Southbound	Weekday	p.m._peak	Clear	No_incident	No_incident	No_incident	High	Low	High	C	Uncongested
Southbound	Weekday	p.m._peak	Clear	No_incident	No_incident	No_incident	High	Low	High	D	Uncongested
Southbound	Weekday	p.m._peak	Clear	No_incident	No_incident	No_incident	High	Low	High	C	Uncongested
Southbound	Weekday	p.m._peak	Clear	No_incident	No_incident	No_incident	High	High	Low	E_F	Uncongested
Southbound	Weekday	p.m._peak	Clear	No_incident	No_incident	No_incident	High	High	Low	E_F	Congested
	:	:	:	:	:	:		:	:	:	:

FIGURE 5 Example of processed data.

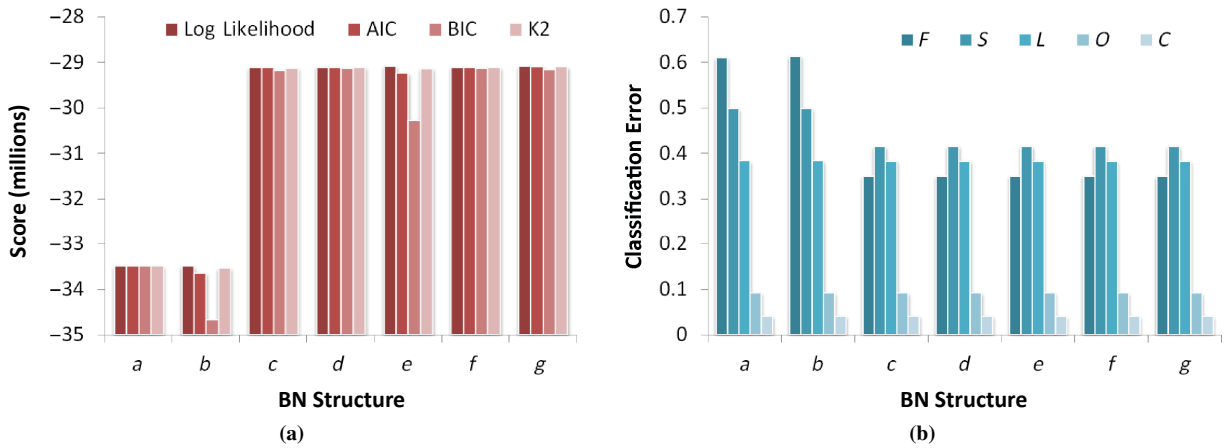

FIGURE 6 Comparing goodness of fit of seven BN structures on the basis of (*a*) network scores (the higher, the better) and (*b*) classification errors for target variables *F*, *O*, *S*, *L*, and *C* (the lower, the better).

variable. The classification error for *F* shows the highest variability across models; the error is the highest in *a* and *b* and the lowest in *f* and *g*. By comparing the numbers, one finds that the error rate produced by *g* was the lowest of the seven models when predicting *S*, *L*, and *C* and the second lowest when predicting *F* and *O*. From these results, it is concluded that of the seven tested structures, the model structure selected for this study, *g*, described the underlying data best.

Parameter Estimation Results

The parameters of the BN, that is, the CPT for each node, were learned from the data that contained a total of 4,787,932 observations from all 19 links. In this paper, the R software package bnlearn was used for parameter learning and probabilistic inference (*21*). Figure 7 presents the estimation result for the marginal probability distribution of each node. The distributions of direction (DR), day of week (*D*), and time of day (*H*) are consistent with what can be expected from the state definitions. The probability that the weather (*W*) was clear is 93.3%, and rain events occurred with the probabilities of 5.2% for light rain, 1.0% for moderate rain, and 0.4% for heavy rain. For any given link, the probability of incident occurrence (*I_O*) was 0.44%. The probabilities of having an incident on its upstream link (*I_U*) and downstream link (*I_D*) were 0.43% and 0.44%, respectively. Since all target links were upstream or downstream links of one another, the marginal distributions of these three incident variables should be very similar. From the distributions of flow (*F*), occupancy (*O*), and speed (*S*), it was observed that *O* is heavily skewed compared with *F* and *S*, showing that 90.9% of the time the occupancy was very low and the percentage of high or very high was 1.7%. According to the congestion indicator (*C*), the probability that any given link in the study corridor was congested was 4% and the probability of its being uncongested was 96%.

ANALYSIS

In this section, a number of analysis methods are presented to identify factors that affect traffic congestion with the BN model constructed above. The main variable of interest is the congestion indicator *C*; its parent nodes DR, *D*, *W*, *H*, *I_U*, *I_O*, and *I_D* are potential causes of congestion or congestion factors, which will be called scenario variables. The focus of the analysis is on understanding the relationships between the target variable and the scenario variables by performing various diagnostic and predictive reasoning tasks.

Identifying Leading Causes for Congestion Diagnosis

With the BN model, diagnostic reasoning can be performed, that is, reasoning from effect (congestion occurrence) to cause (scenario variables). Figure 8 shows the posterior probability distribution of each scenario variable *S* given that congestion has been observed, that is, $P(S|C = \text{congested})$. The probability that C = congested is 100% in Figure 8, meaning that the belief about the congestion state has been updated; for example, traffic congestion has been observed, and there is no uncertainty in variable *C*. Compared with the prior distributions $P(S)$ in Figure 7, the following changes have been made to the distributions of scenario variables:

- [DR]. The probability that a link is northbound when congestion has been observed is 56.3%, increased from 47.5% when there is no information about the congestion state.
- [*D*]. The probability that it is a weekday has increased from 71.1% to 97.1%.
- [*H*]. The probability of being p.m. peak has increased from 17% to 49% and that of being a.m. peak has increased from 17% to 38.1%.
- [*W*]. The probability that it is raining (regardless of the rain intensity) has increased by 3.1%.
- [*I_U*, *I_O*, *I_D*]. The probabilities that there exists an incident on the downstream link, the current link, and the upstream link have increased from the level of 0.44% to 1.85%, 1.41%, and 1.00%, respectively.
- [*F*]. The probability that the flow rate is very high has increased from 5.6% to 26.8%.
- [*O*]. The probability that the occupancy is high or very high has increased from 1.7% to 7.9%.
- [*S*]. The probability that the speed is very low or low has increased from 4.1% to 82.6%.
- [*L*]. The probability that the level of service is E–F has increased from 3.6% to 15.6%.

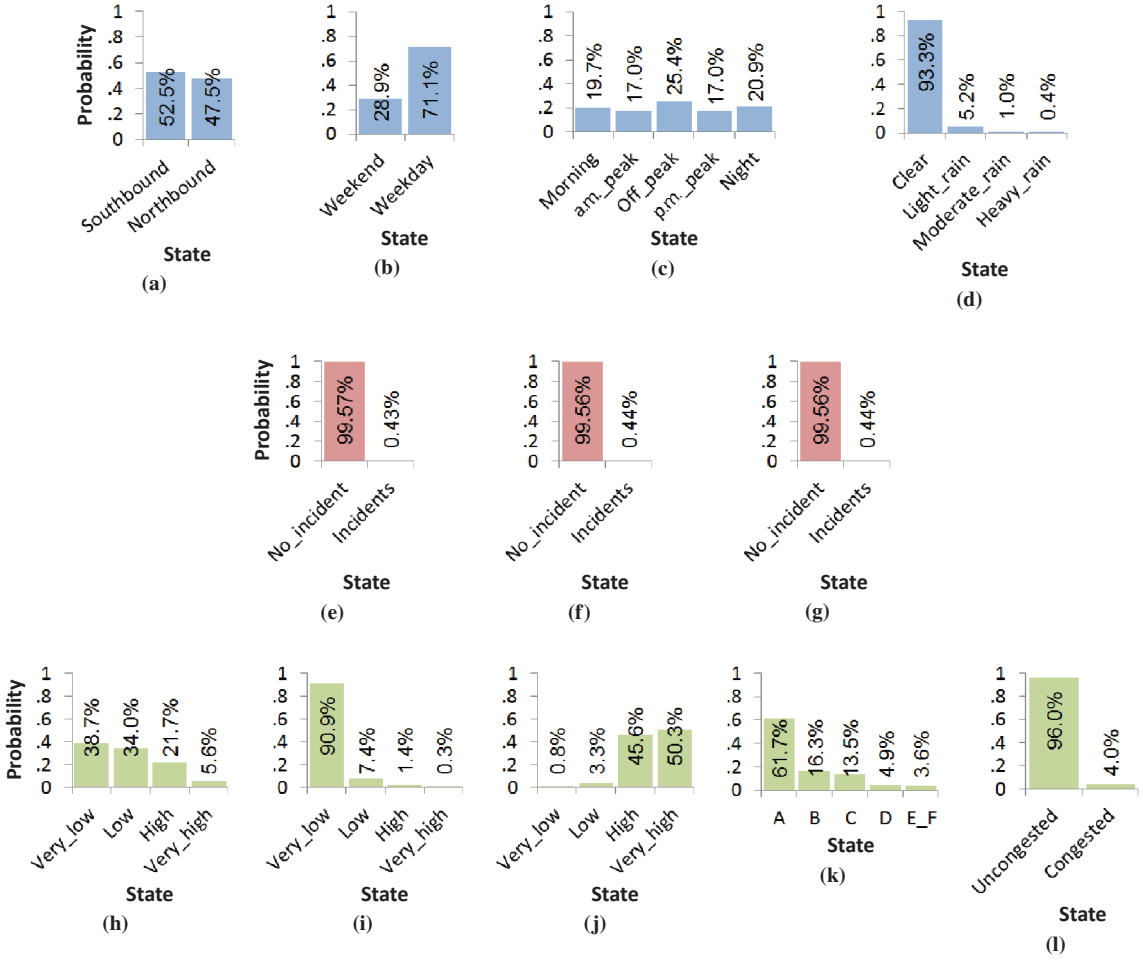

FIGURE 7 Marginal probability distributions for 12 variables in proposed BN model: (*a*) DR, (*b*) *D*, (*c*) *H*, (*d*) *W*, (*e*) *I_U*, (*f*) *I_O*, (*g*) *I_D*, (*h*) *F*, (*i*) *O*, (*j*) *S*, (*k*) *L*, and (*l*) *C*.

How strongly a particular scenario state is associated with the congestion occurrence can be further quantified by measuring the odds ratio. The odds of an event occurring is the ratio of the probability that the event will happen to the probability that the event will not happen. The odds ratio (OR) is the ratio of the odds of an event occurring in one group to the odds of it occurring in another group, providing a way to measure how strongly the event is associated with the first group compared with the second group. For instance, if one considers the event of "congestion" and compares the odds of congestion for two groups, "rain" and "no rain," then

$$\text{OR} = \frac{\dfrac{P(\text{congestion}|\text{rain})}{P(\text{no congestion}|\text{rain})}}{\dfrac{P(\text{congestion}|\text{no rain})}{P(\text{no congestion}|\text{no rain})}}$$

An odds ratio greater than 1 (less than 1) indicates that the event is more likely (less likely) to occur in the first group. In the above example, OR < 1 indicates that the congestion is more likely to occur when there is "rain" compared with when there is no rain. The fact that the distribution of the OR is highly skewed (the odds ratio is limited to zero at the lower end but unlimited at the upper end), however, makes it difficult to compare the strength of association across two regimes 0 < OR < 1 and OR > 1. To overcome this difficulty, the logarithm of the OR (logOR) is used to convert the scales of less likely and more likely regimes from (0, 1) and (1, ∞) to (−∞, 0) and (0, ∞), respectively.

By denoting the event of having congestion and the event of having a particular scenario state simply by *C* and *S,* respectively, the logOR for the problem is defined as follows:

$$\log \text{OR} = \log\left(\frac{\dfrac{P(C|S)}{P(\sim C|S)}}{\dfrac{P(C|\sim S)}{P(\sim C|\sim S)}}\right) = \log\left(\frac{P(C,S)\cdot P(\sim C,\sim S)}{P(C,\sim S)\cdot P(\sim C,S)}\right) \quad (1)$$

where

$P(S)$ = probability of scenario event S occurring [e.g., $P(W =$ clear)];
$P(\sim S)$ = probability that scenario event does not occur [e.g., $P(W \neq$ clear)];
$P(C)$ = probability that congestion occurs, that is, $P(C =$ congested); and
$P(\sim C)$ = probability that congestion does not occur, that is, $P(C \neq$ congested).

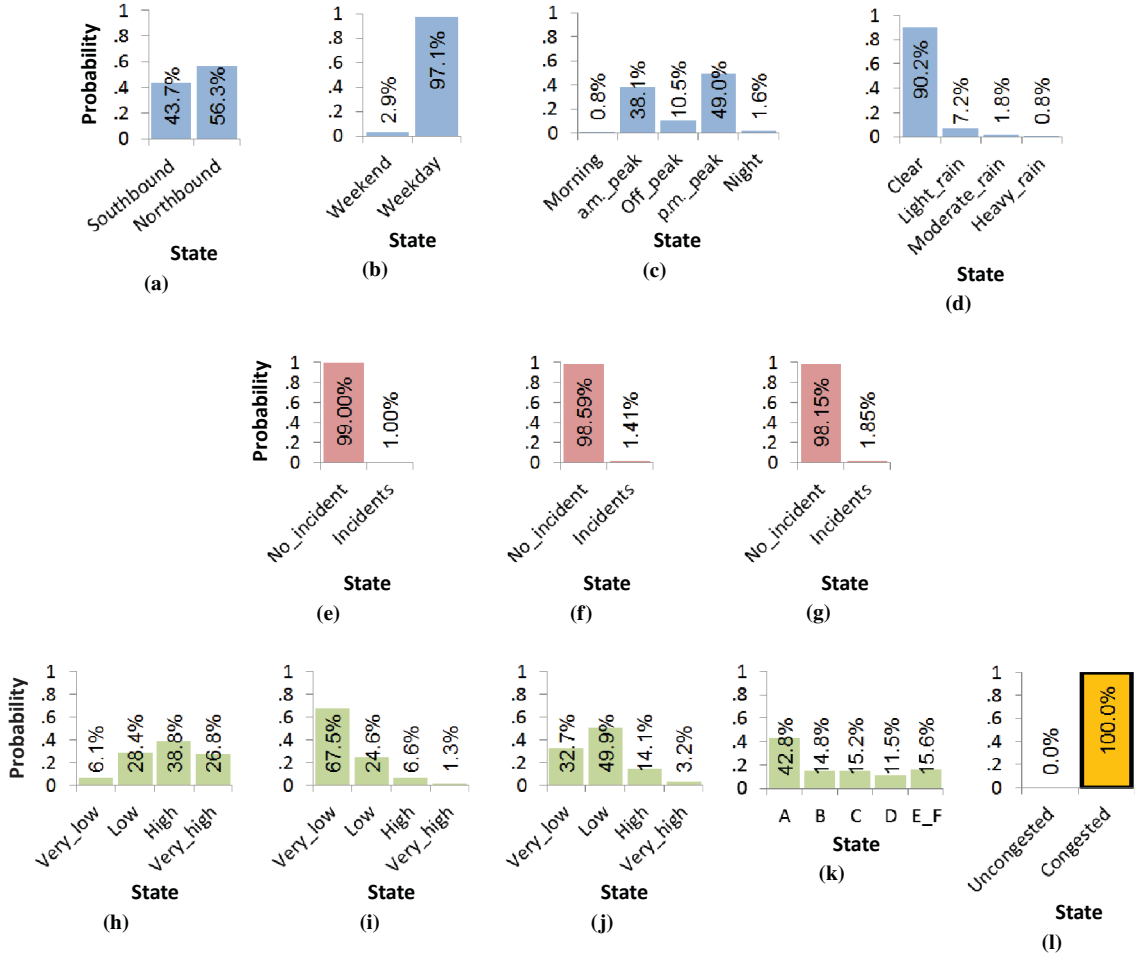

FIGURE 8 Posterior distributions of scenario variables S given that congestion has been observed, that is, $P(S|C$ = congested): (*a*) DR, (*b*) *D*, (*c*) *H*, (*d*) *W*, (*e*) *I_U*, (*f*) *I_O*, (*g*) *I_D*, (*h*) *F*, (*i*) *O*, (*j*) *S*, (*k*) *L*, and (*l*) *C*.

The OR can also be defined relative to the joint probabilities, as shown in the last term in Equation 1, where the expression becomes the product of the probability that both C and S occur and the probability that both C and S do not occur divided by the product of the probabilities that only one of them occurs. If the logOR is greater than 0, then having scenario event S is considered to be associated with having congestion C and events S and C are more likely to occur together. If the logOR is less than 0, Events S and C are less likely to occur together. All of the probability values required to compute logOR can be obtained from the BN model; the results are presented in Figure 9. The highest logOR is found in scenario event D = weekday; it is 1.15, indicating that weekday and congestion occurrence are very strongly associated (highly likely to occur together). The events H = p.m. peak and I_D = incident are also strongly associated with congestion occurrence, with the logORs of 0.71 and 0.70, respectively. The next highest logORs are found in I_O = incident and H = a.m. peak, with the values of 0.56 and 0.51, respectively. Scenario events DR = southbound, D = weekend, H = morning, H = off-peak, H = night, W = clear, I_O = no incident, I_D = no incident, and I_U = no incident are all showing logORs of less than 0, indicating that these events are less likely to occur together with the congestion event.

Identifying Critical Scenarios for Congestion Prediction

So far, the relationships between scenario variables and congestion variables have been investigated by focusing on individual scenario variables separately. It is, however, possible to consider all seven scenario variables simultaneously to quantify their effects on congestion occurrence. Now assume that scenario event S is the combination of all seven scenario variables: for example, S = {DR = southbound, D = weekday, H = a.m. peak, W = clear, I_U = no incident, I_O = incident, I_D = no incident}. The focus is on identifying the most important scenarios that are highly associated with congestion event C, namely, scenarios that have high joint probabilities with the congestion event, $P(S, C)$. While all possible combinations of $P(S, C)$ can be computed and the scenarios with the highest joint probabilities identified, the scenarios identified in this manner might include very rare scenarios because $P(C, S) = P(S) \times P(C|S)$ and the combination of very low $P(S)$ and very high $P(C|S)$ can still produce high $P(C, S)$. A better strategy would be to identify scenarios that produce high $P(S)$ and high $P(C|S)$. This approach is similar to the concept of identifying risk scenarios in risk analysis. In quantitative risk analysis, risk is often expressed as

$$\text{risk} = \text{probability} \times \text{impact}$$

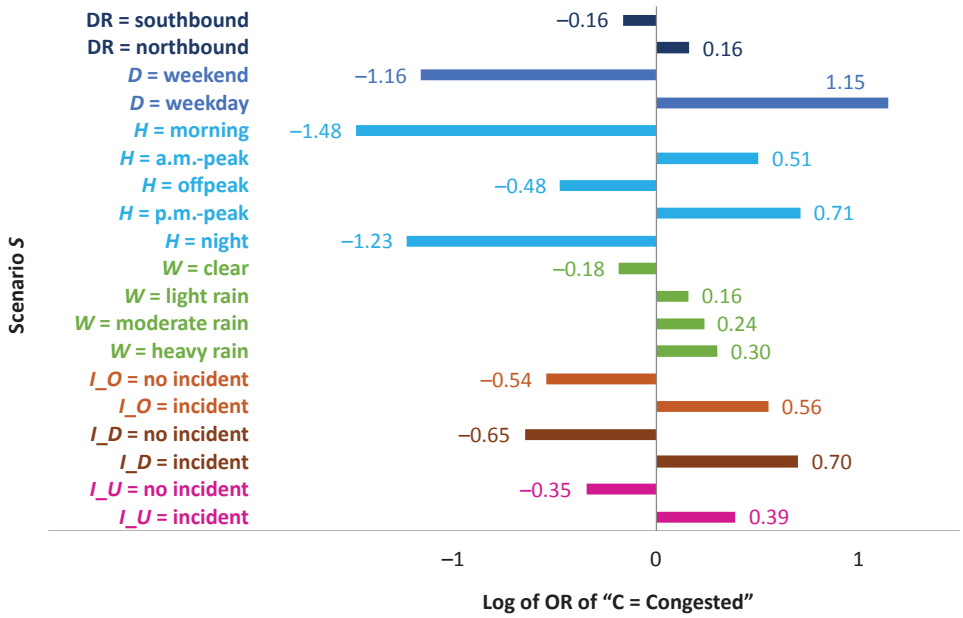

FIGURE 9 LogOR between each scenario event and congestion occurrence.

The relationship $P(C, S) = P(S) \times P(C|S)$ can be interpreted similarly as follows:

$$\underbrace{\text{scenario risk}}_{P(C,S)} = \underbrace{\text{probability of scenario}}_{P(S)} \times \underbrace{\text{impact of scenario}}_{P(C|S)}$$

where

$P(C, S)$ = overall importance or risk of scenario S,
$P(S)$ = likelihood of scenario, and
$P(C|S)$ = effect of scenario expressed in regard to probability that congestion occurs given scenario occurring.

On the basis of this framework, the impact-probability chart is created to identify high probability–high impact scenarios (that is, high $P(S)$ and high $P(C|S)$) as shown in Figure 10. This chart allows one to rate potential risks or the importance of a scenario on two dimensions and select appropriate cutoff points for $P(S)$ and $P(C|S)$ according to the number of scenarios that one wishes to include in the final scenarios set. In this study, the top 40 important scenarios were selected from the high-probability and high-impact region in the chart, and the boundaries for this region were selected as $P(S) \geq 0.0003$ and $P(C|S) \geq 0.0085$, as depicted in Figure 10*b*. The selected 40 scenarios are presented in Figure 11 in the form of a scenario tree. The column

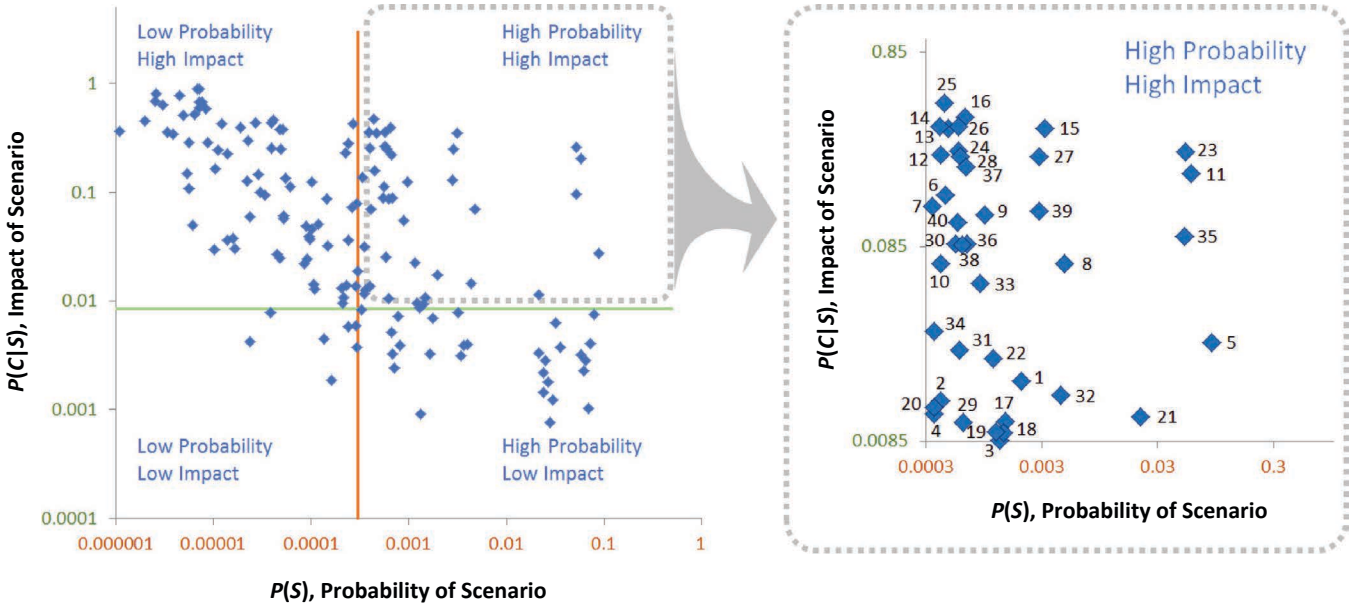

FIGURE 10 Probability–impact chart for identifying high probability–high impact scenarios in regard to scenario probability $P(S)$ and scenario impact $P(C|S)$.

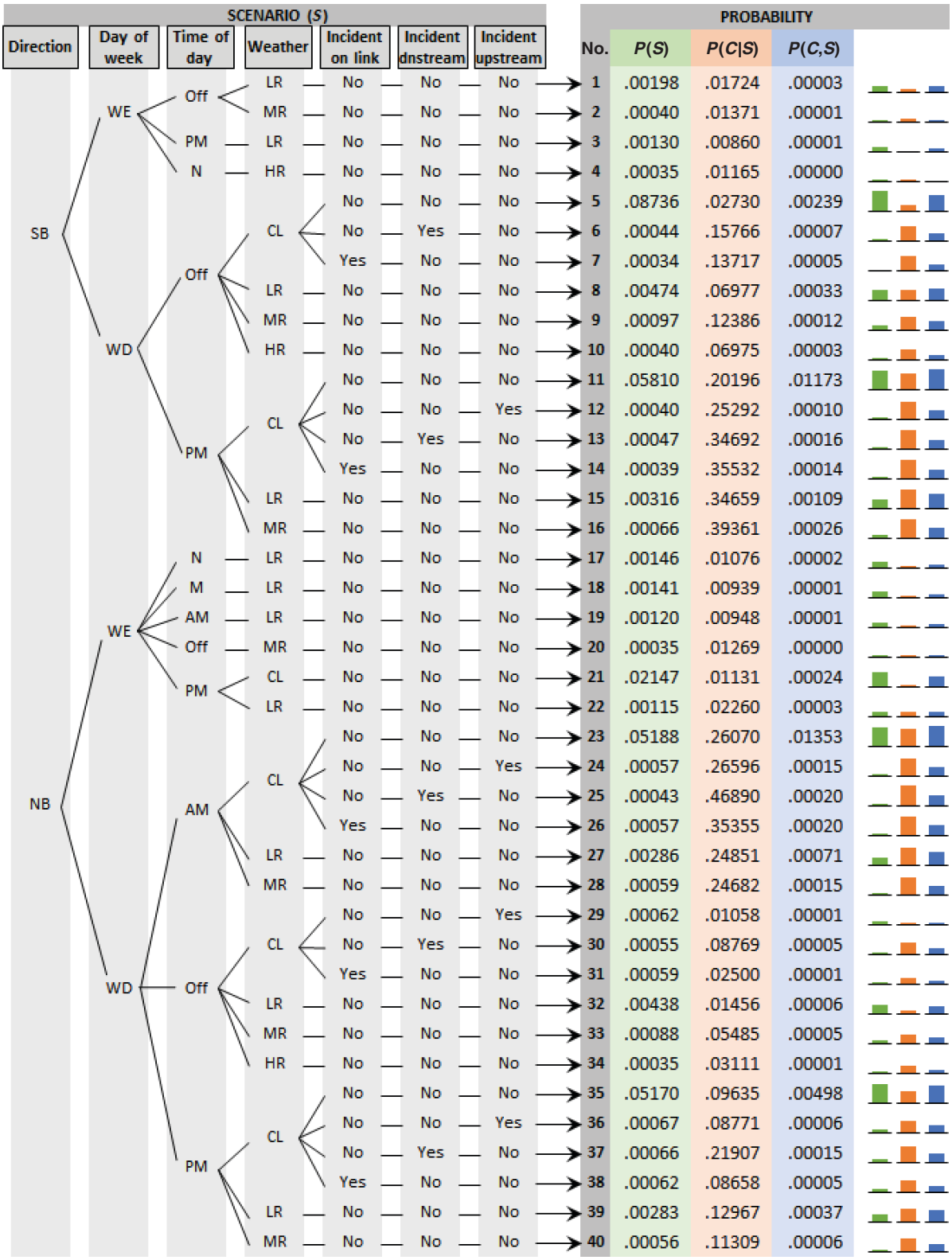

FIGURE 11 Scenario tree of top 40 important scenarios that have high association with congestion occurrence (SB = southbound direction; NB = northbound direction; WE = weekend; WD = weekday; Off = off-peak period; PM = p.m. peak period; N = night; M = morning; AM = a.m. peak period; LR = light rain; MR = moderate rain; HR = heavy rain; CL = clear; dnstream = downstream).

chart in the rightmost column of the figure visualizes the relative magnitude of each of $P(S)$, $P(C|S)$, and $P(C, S)$ across 40 scenarios, making it easy to find the ranking of a scenario in the selected scenario group. For instance, Scenario 5 has a relatively high $P(S)$ but a low $P(C|S)$, and Scenario 15 has a relatively low $P(S)$ but high $P(C|S)$ within those 40 scenarios. The resulting $P(C, S)$ values are relatively high in both scenarios.

Identifying critical scenarios that affect traffic performance is an important task in many decision-making situations for transportation planning and operations. The proposed approach provides a systematic framework to rank and prioritize important scenarios and can be used in various scenario analysis applications, such as scenario-based travel time reliability analysis and simulation modeling (*22, 23*).

CONCLUSIONS

Managing traffic congestion requires accurate knowledge of the causes of congestion on a given road network as well as their relative significance and respective solutions. Previous research focused on estimating the relative effect of different causes of congestion, which included incidents, weather, work zones, and special events, with statistical analysis such as linear regression methods. Traditional statistical methods, however, have limitations in capturing complex dependencies and uncertainties in external events and traffic states in urban networks. This study proposed a BN analysis approach to modeling the probabilistic dependency structure between causes of congestion on a particular road segment and analyzing the probability of traffic congestion given various roadway condition scenarios. A BN approach was used to encode the joint probability distribution over a set of random variables that described scenario variables, which represented factors affecting the congestion level of a target segment, such as time of day, incident, weather, and traffic states on adjacent links, as well as output variables, which represented traffic performance measures of the target segment, such as flow, density, and speed. A properly configured BN model can be used to (*a*) quantify the contribution of each cause or the combination of multiple causes to traffic congestion, thereby allowing the identification of leading causes for the purpose of congestion diagnosis; (*b*) predict future congestion levels on the basis of the current network situations; and (*c*) analyze the likely scenarios (combinations of causes) that produce the worst traffic congestion in a study network and their occurrence probabilities. This study developed a method to build a BN model on the basis of historical traffic and event data and demonstrated the BN-based traffic analysis with a study network in Queensland, Australia. The study discussed applications of the proposed BN model in urban traffic congestion management, focusing on its capability to provide a comprehensive data-driven and probabilistic analysis platform for congestion diagnosis and prediction.

ACKNOWLEDGMENT

This work was funded by the University of Queensland through the New Staff Research Start-Up Fund.

REFERENCES

1. Heckerman, D. *A Tutorial on Learning with Bayesian Networks.* Publication MSR-TR-95-06. Microsoft Research, 1995.
2. Jensen, F. V., and T. D. Nielsen. *Bayesian Networks and Decision Graphs.* Springer, New York, 2007.
3. Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann Publishers, San Francisco, Calif., 1988.
4. Sun, S., C. Zhang, and G. Yu. A Bayesian Network Approach to Traffic Flow Forecasting. *IEEE Transactions on Intelligent Transportation Systems,* Vol. 7, No. 1, 2006, pp. 124–132.
5. Pascale, A., and M. Nicoli. Adaptive Bayesian Network for Traffic Flow Prediction. Presented at 2011 IEEE Statistical Signal Processing Workshop (SSP), Nice, France, 2011.
6. Samaranayake, S., S. Blandin, and A. Bayen. Learning the Dependency Structure of Highway Networks for Traffic Forecast. Presented at 2011 50th IEEE Conference on Decision and Control and European Control Conference, Orlando, Fla., 2011.
7. Yu, Y. J., and M.-G. Cho. A Short-Term Prediction Model for Forecasting Traffic Information Using Bayesian Network. *Third International Conference on Convergence and Hybrid Information Technology,* No. 1, 2008, pp. 242–247.
8. Hofleitner, A., R. Herring, P. Abbeel, and A. Bayen. Learning the Dynamics of Arterial Traffic from Probe Data Using a Dynamic Bayesian Network. *IEEE Transactions on Intelligent Transportation Systems,* Vol. 13, No. 4, 2012, pp. 1679–1693.
9. Castillo, E., J. M. Menéndez, and S. Sánchez-Cambronero. Predicting Traffic Flow Using Bayesian Networks. *Transportation Research Part B: Methodological,* Vol. 42, No. 5, 2008, pp. 482–509.
10. Zhang, K., and M. A. P. Taylor. Effective Arterial Road Incident Detection: A Bayesian Network Based Algorithm. *Transportation Research Part C: Emerging Technologies,* Vol. 14, No. 6, 2006, pp. 403–417.
11. Gregoriades, A., and K. C. Mouskos. Black Spots Identification Through a Bayesian Networks Quantification of Accident Risk Index. *Transportation Research Part C: Emerging Technologies,* Vol. 28, 2013, pp. 28–43.
12. Sun, J., and J. Sun. A Dynamic Bayesian Network Model for Real-Time Crash Prediction Using Traffic Speed Conditions Data. *Transportation Research Part C: Emerging Technologies,* Vol. 54, 2015, pp. 176–186.
13. Hossain, M., and Y. Muromachi. A Bayesian Network Based Framework for Real-Time Crash Prediction on the Basic Freeway Segments of Urban Expressways. *Accident Analysis and Prevention,* Vol. 45, 2012, pp. 373–381.
14. Murphy, K. P. *Dynamic Bayesian Networks: Representation, Inference and Learning.* University of California, Berkeley, 2002.
15. Friedman, N., and D. Koller. Being Bayesian About Network Structure. A Bayesian Approach to Structure Discovery in Bayesian Networks. *Machine Learning,* Vol. 50, No. 1/2, 2003, pp. 95–125.
16. Chickering, D. M., D. Heckerman, and C. Meek. A Bayesian Approach to Learning Bayesian Networks with Local Structure. In *Proceedings of 13th Conference on Uncertainty in Artificial Intelligence,* San Francisco, Calif., 1997, pp. 80–89.
17. Maghrebi, M., and S. T. Waller. Exploring Experts Decisions in Concrete Delivery Dispatching Systems Using Bayesian Network Learning Techniques. Presented at 2014 2nd International Conference on Artificial Intelligence, Modelling and Simulation (AIMS), Madrid, Spain, 2014.
18. Buntine, W. L. A Guide to the Literature on Learning Probabilistic Networks from Data. *IEEE Transactions on Knowledge and Data Engineering,* Vol. 8, No. 2, 1996, pp. 195–210.
19. de Campos, L. M. A Scoring Function for Learning Bayesian Networks Based on Mutual Information and Conditional Independence Tests. *Journal of Machine Learning Research,* Vol. 7, 2006, pp. 2149–2187.
20. Cooper, G. F., and E. Herskovits. A Bayesian Method for the Induction of Probabilistic Networks from Data. *Machine Learning,* Vol. 9, No. 4, 1992, pp. 309–347.
21. Scutari, M. Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software,* Vol. 35, No. 3, 2010, pp. 1–22.
22. Kim, J., H. S. Mahmassani, P. Vovsha, Y. Stogios, and J. Dong. Scenario-Based Approach to Analysis of Travel Time Reliability with Traffic Simulation Models. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2391,* Transportation Research Board of the National Academies, Washington, D.C., 2013, pp. 56–68.
23. Mahmassani, H. S., J. Kim, Y. Stogios, K. Currie, and P. Vovsha. Incorporating Reliability Performance Measures in Operations and Planning Modeling Tools. Presented at 92nd Annual Meeting of the Transportation Research Board, Washington, D.C., 2013.

The Standing Committee on Artificial Intelligence and Advanced Computing Applications peer-reviewed this paper.