Study work

Prediction and Diagnosis of Road congestion using Bayesian networks

Shaocheng Liu

1. Introduction

1.1 Introduction about traffic congestion

In today's era of continuous urbanization and population growth, traffic congestion has been on the rise all over the world. Due to the high density of people and vehicles, especially in large cities, traffic congestion takes up a lot of people's time and energy and becomes a major traffic and social problem. Traffic congestion is also a serious problem in the development of autonomous driving, because one of the things that autonomous driving is trying to achieve is to reduce congestion and make it easier for people to get around.

Congestion is a traffic condition characterized by slower speeds, longer travel times and increased vehicle queues. And in dealing with traffic congestion problems, there are several points that people need to think about: the first thing is about the diagnosis of traffic congestion, which means how to define a traffic congestion when it happens; the second thing is about the causes of traffic congestion, which aims to understand the reasons that cause a road congestion; the third thing is about the prognosis of traffic congestion, which means after the causes of traffic congestion are found and accurately defined, how can we predict the traffic congestion and proactively report it before it happens.

Bayesian network is a type of probabilistic graphical model that can be used to represent the joint probability distribution over a set of random variables under uncertainty, and it can provide an efficient way to capture the complex relationships between different factors and their effects on traffic congestion. Using Bayesian networks, traffic managers can diagnose congestion and predict the causes of congestion.

1.2 Related researches of traffic congestion

There are some existing studies and articles which use Bayesian network to solve the traffic problems, after searching and reading these studies, I classify them into studying two aspects of traffic problems:

Incident Prediction: a study in 2017 explores an application of Bayesian network theory based on probability risk analysis to causation analysis of road accidents, taking Adelaide Central Business District (CBD) in South Australia as a case, which includes driver, road, environment, vehicle and road crash as variable class, and divides them in different variables. The results provide theoretical support for urban road management authorities when analyzing induction factors and improving safety performance within their respective systems. The relevant Bayesian network is showed in this picture

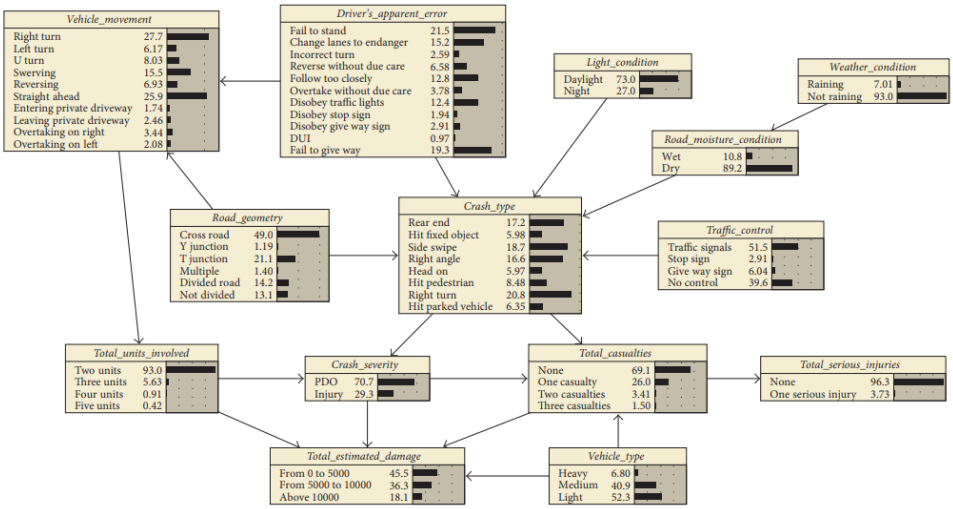

FIGURE 4: The Bayesian network model after parameter learning in Netica 6.02.

Figure 1: BN Model for road incident

Another study in 2011 used references to expert knowledge and data fusion method to explore a topological structure of Bayesian network, which apply a joint tree engine to infer the probability distribution of traffic accident types under the influence of factors such as vehicle type, accident location and traffic participants.

Congestion prediction and diagnosis: There are two different methods when these studies deal with congestion prediction. One method is to use a normal Bayesian network. A study in 2021 proposes a Bayesian network based probabilistic congestion estimation approach for monitoring traffic conditions. The proposed BN-based approach considers both speed (which is called Speed Performance Index) and volume (which is called Level of Services) related measures to provide an estimate of the probable congestion state in terms of probability. The study builds two different Bayesian networks for recurring and non-recurring congestion, and the dataset for non-recurring congestion was selected because a hurricane, a natural disaster, occurred during this time. The comparison of these two Bayesian networks is shown below:

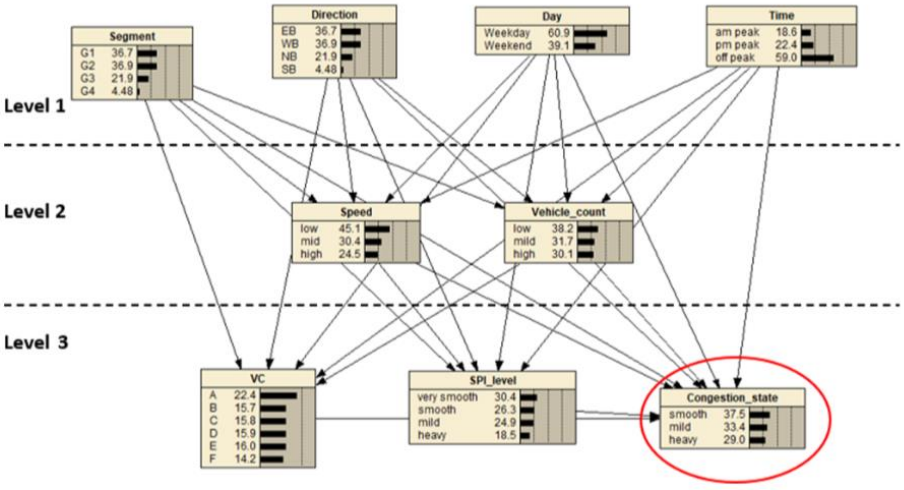

Figure 2: BN Model for recurring situation

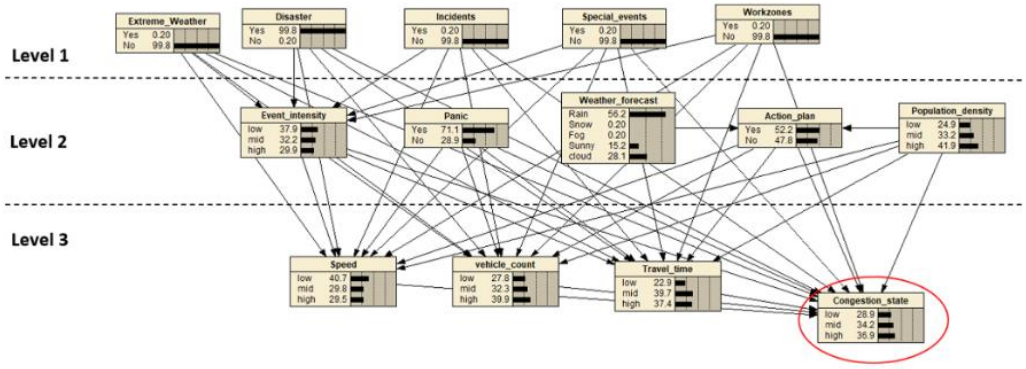

Figure 3: BN Model for nonrecurring situation

Another study in 2020 discusses a method for multi-cause automatic real-time identification of urban road traffic congestion based on the Bayesian network. The proposed model has high flexibility and strong interpretability, which can help better express the correlation between nodes and achieve real-time automation. The results of a study conducted in Quanxiu Street, Quanzhou City, showed that five causes of traffic congestion had higher detection accuracy rates than contrast methods, such as pedestrian influence, peak traffic, parking occupied roads and unreasonable signal timing. A 2016 study proposes a Bayesian Network (BN) analysis approach to model the probabilistic dependency structure of causes leading to traffic congestion on a given road segment. It also analyses the probability of traffic congestion under different road condition scenarios, such as time of day, incident, weather, and conditions on adjacent links. This article serves as my main reference for my research work, and the Bayesian network constructed by this study is shown below:

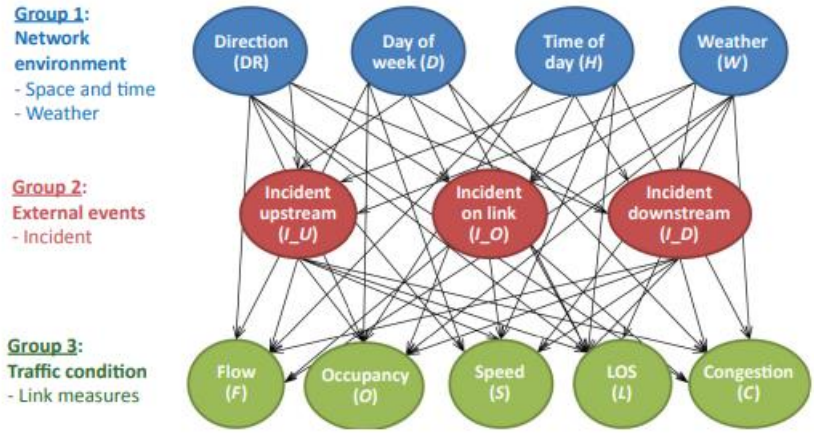

Figure 4: BN Model of the study in 2016

The other method is to use a dynamic Bayesian network. In a study in 2018, a dynamic Bayesian network model is proposed to describe the change and dissipation of road congestion. The prediction results show that this method is feasible in predicting the flow state and dissipation time of vehicles, providing drivers with the shortest routes to less congested roads. Another study in 2021 discusses a new dynamic Bayesian graph convolutional network (DBGCN) that can be used to

characterize congestion propagation in road networks, and it is able to simulate congestion propagation processes for customized scenarios by learning latent rules from observed data, and reveal variations in congestion patterns according to road network structure. The overall framework of DBGCN is shown below:

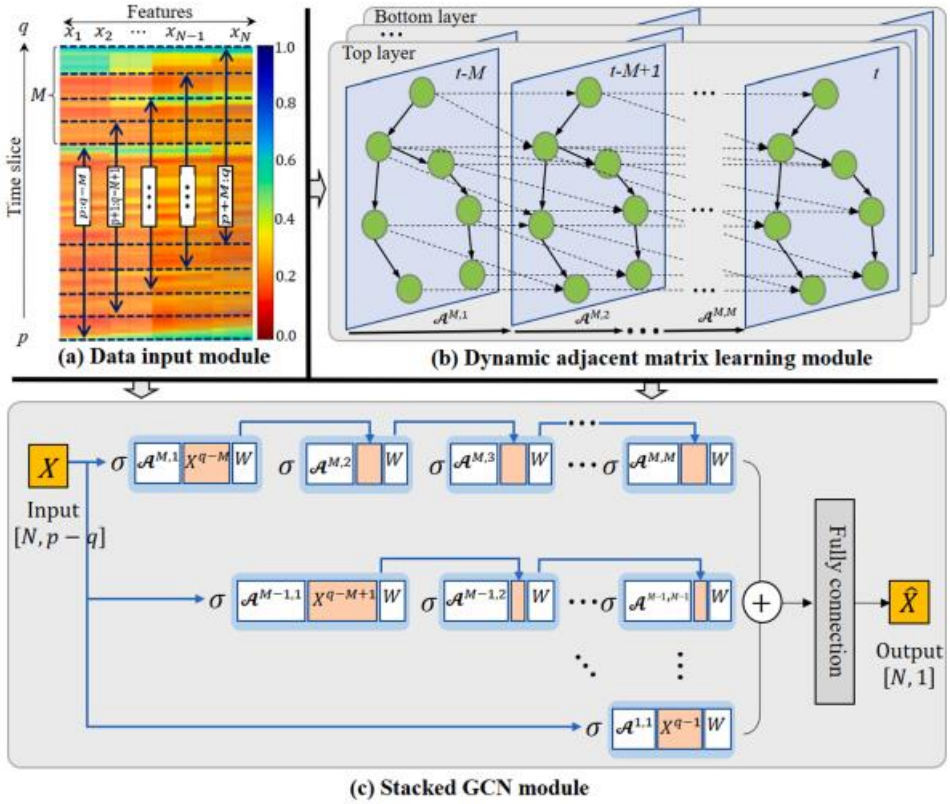

Figure 5: DBGCN Model of study in 2021

Also, a study in 2006 presents a new arterial road incident detection algorithm called TSC_ar. This algorithm uses Bayesian networks to model the causal dependencies between traffic events and parameters, allowing for robust and dynamic knowledge base in order to detect incidents with high accuracy while keeping false alarm rate low. Additionally, incorporating intersection traffic signals into data processing further improves accuracy of results.

2. Data analysis

2.1 Introduction about open datasets

Since most of the papers do not explicitly state the source of their data, the reason may be that they do not use public datasets, but rather datasets obtained in cooperation with relevant authorities. I found several relevant datasets online, a small number of which were obtained from the papers, and I briefly describe them below.

1) Florida Department of Transportation's Traffic Information (FDOT)
   Traffic Information (fdot.gov)
   This dataset provides statistical traffic information for Florida's State Highway System, which provides not only historical information about the traffic situation, but also a website to watch the traffic data in real-time.

2) South Australian Government Data Directory
   Road Crash Data - Dataset - data.sa.gov.au
   This department for Infrastructure and Transport provides a dataset for road crash data, which includes time, location, type of crash, weather when crash happened etc.
3) Open Data Portal of Queensland Government
   Transport and Main Roads - Organisations - Open Data Portal | Queensland Government
   This dataset provides also many datasets for transport, which include traffic data and also crash data of different roads in Queensland.
4) Chicago traffic tracker on data.gov
   Chicago Traffic Tracker - Historical Congestion Estimates by Segment - 2018-Current - Catalog (data.gov)
   This dataset is from a official dataset of the United States government, which contains the historical data of traffic such as speed, street name etc., which can be used to estimate congestion.
5) Caltrans Performance Measurement System (PeMS)
   Caltrans PeMS
   This dataset provides over ten years of data for historical analysis in the California-area from different detectors and sensors, which contains almost all kinds of traffic data, and there is also a real-time website which shows the real-time traffic situation in California.
6) National Center for Environmental Information
   National Centers for Environmental Information (NCEI) (noaa.gov)
   This dataset manages one of the largest archives of atmospheric, coastal, geophysical and oceanic research in the world, and the weather condition part is useful for our study, because it provides data which is collected by different sensors all over the world hourly, such as hourly precipitation, visibility and windspeed etc.

2.2 Introduce about my dataset and variables

After searching and learning about these relative datasets, I chose PeMS and NCEI as the sources of my dataset. A vehicle detection sensor (VDS) on the I-105-E freeway in Los Angeles was chosen as the source of my data because this part of the freeway is one of the busiest areas in the United States, connecting two freeways I-105 and I-710, so it's also one of the most congested areas. The design of this part of the freeway and the location see below:

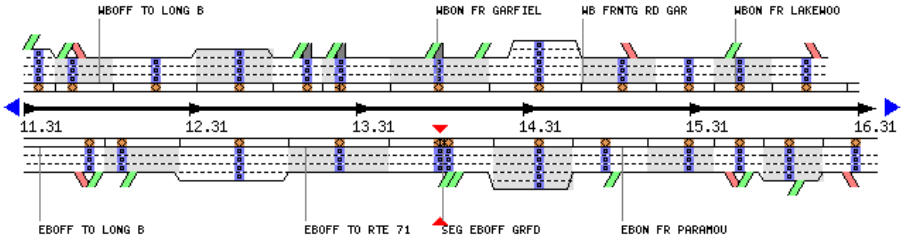

Figure 6: draft of selected area

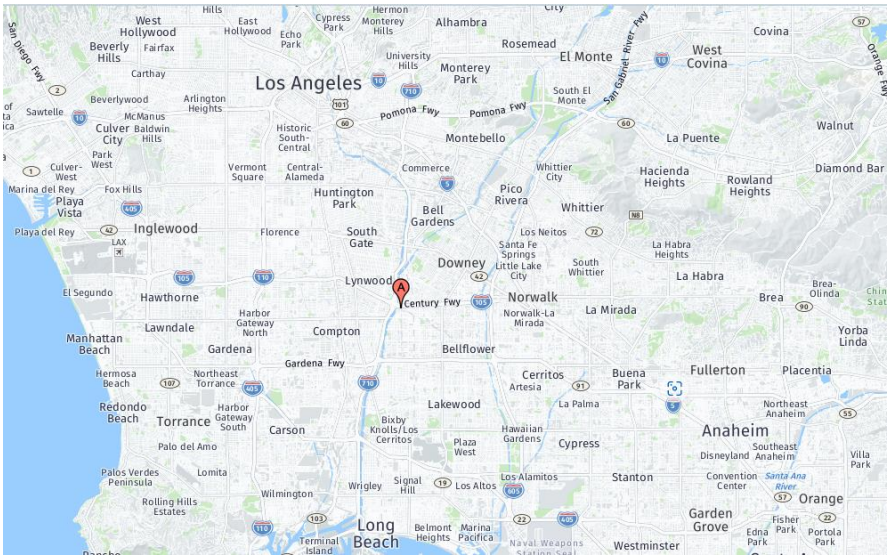

Figure 7: location of selected area

And then I choose a Station which records weather data near this vehicle detection sensor, which named Los Angeles downtown USC, and the location shows below:

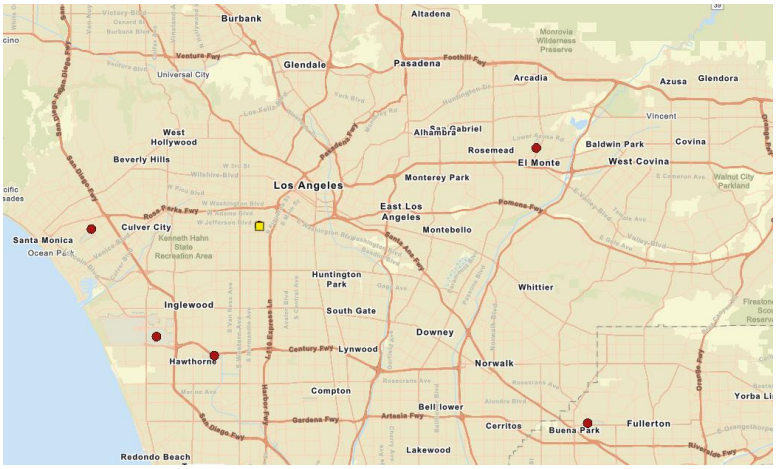

Figure 8: location of selected weather station

The traffic and weather data are collected from 2022.12.1-2023.2.28 and is recorded hourly, which obtained as a 2161-by-8 matrix. the traffic data is collected from the Vehicle Detection Sensor, which contains Speed (mph), Flow (Vehicle/hour) and Occupancy (%), and the weather data is collected from the weather station, which contains Hourly precipitation (inch) and Hourly visibility (mile). And the variables used in the Bayesian network model are presented in the table. A total of 8 variables were selected, and only discrete variables will be considered, that is, nodes that take discrete values. And the variables are categorized into 2 groups: Network Environment and Traffic condition as follows:

Network environment variables represent the environmental factors, Time of day, Day of week, Hourly precipitation and Hourly visibility are considered as environmental factors which have the possibility to influence the traffic congestion. Time of day takes 3 states {AM peak; PM peak; Off peak}, Day of week takes 2

states {Weekday; Weekend}, Hourly precipitation takes 4 states {No rain; Light rain; Moderate rain; Heavy rain}, and Hourly visibility takes 2 states {Clear; Haze}. The detailed descriptions for these state definitions are presents in Table 1, and these 4 variables can be recognized as background variables, which has a causal influence on problem variable.

Traffic condition variables represent link performance measures describing traffic states on the target area. This study includes 4 variables consisting 3 basic traffic steam parameters, Speed, Flow and Occupancy. These three variables take 4 same discrete states {very low, low, high, very high}, and these states are defined according to the value range of each variable, which is also called as normalization. In this study, each variable will be normalized, and corresponding to the range between 0 and 1. And then it will be divided into 4 states as above. The Congestion is a binary variable that indicates whether this area is congested. As the study in 2016, occupancy and flow values are used to determine the value of congestion. It takes two states {congest; uncongest}, “uncongest” if occupancy<Occcrit and “congest” if occupancy>Occcrit, where Occcrit represents the critical occupancy at which flow becomes maximum, as shown in the figure below.

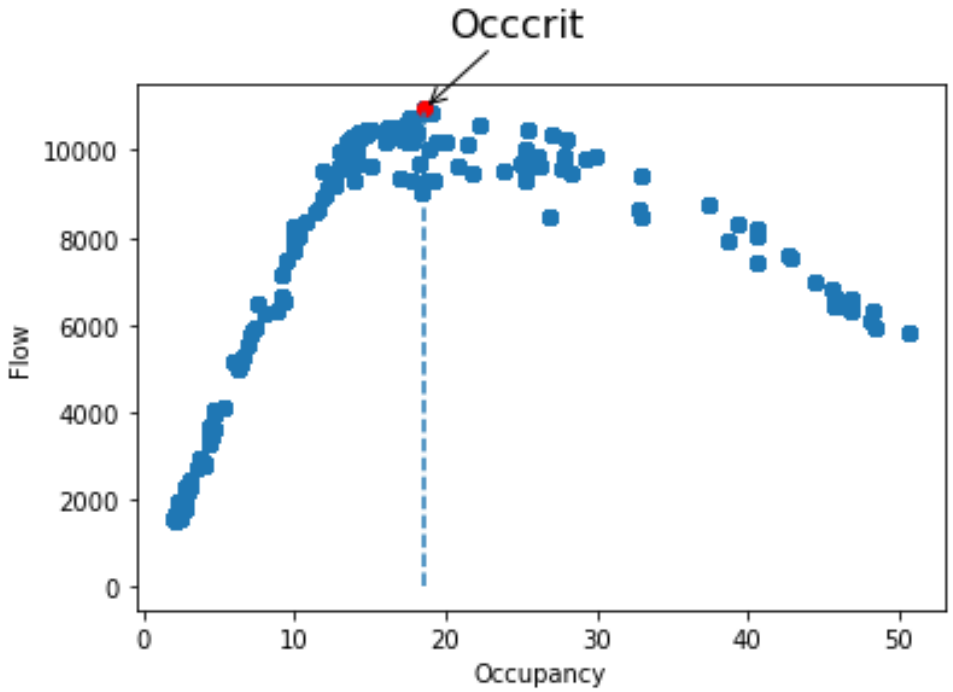

Figure 9: Occupancy-by-speed scatter

In this dataset, Occcrit = 18.6%, maximum flow = 10956. This binary congestion indicator will be used as the problem variable, which the Bayesian network model want to compute the posterior probability given observations of values for information variables (which contain symptom variables and background variables), and this variable will be used in performing the congestion diagnosis and prediction. And the data discretization for occupancy by speed and flow by speed show below:

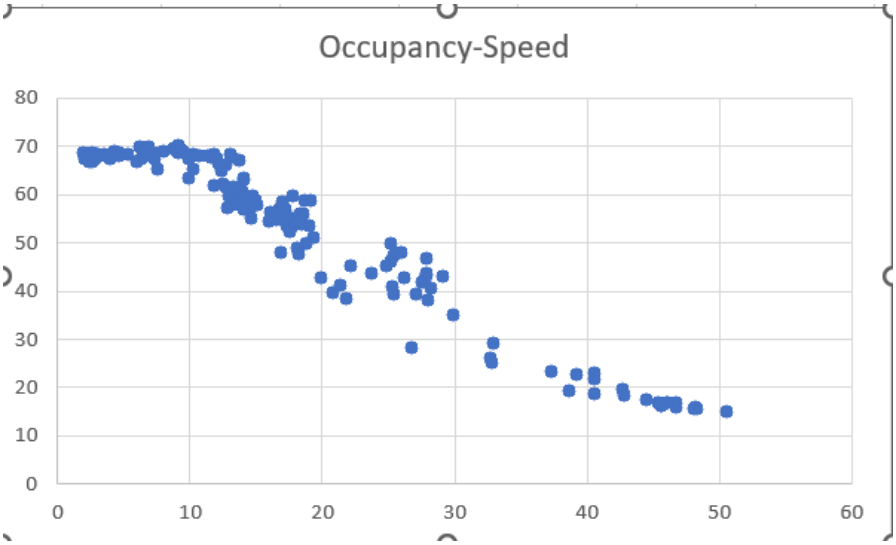

Figure 10: Occupancy-by-speed scatter

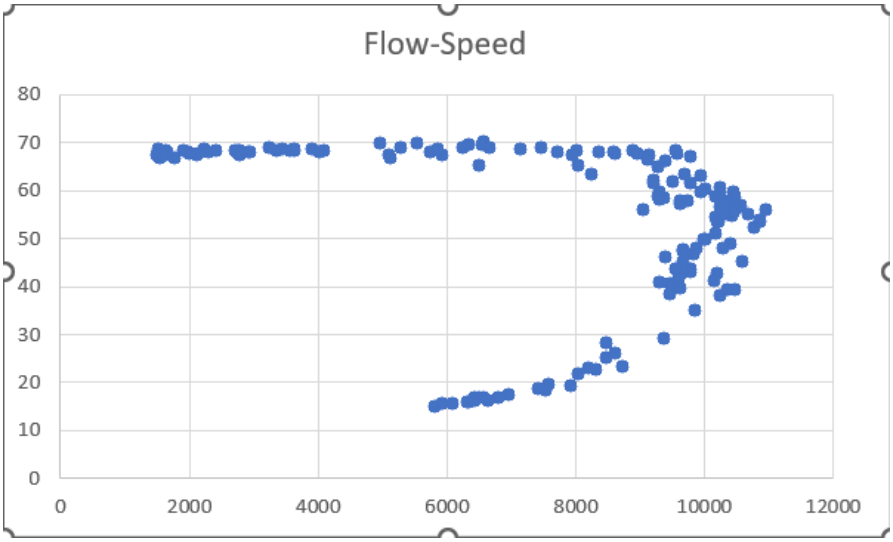

Figure 11: Flow-by-Speed scatter

An example of the dataset is shown in the picture below:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Hour	Speed	Flow	Occupanc	Q (VMT/V	HourlyPre	HourlyVisi	Classification	Day	Time of da	Rain state	Speed sta	Flow state	Occupanc	Day of we	Visibility state	
2	2022/12/1 0:00	68.3	3574	4.7	68.3	0	10	uncongest	Thursday	Off Peak	No rain	very high	low	very low	Weekday	Clear	
3	2022/12/1 1:00	67.9	2299	3.1	67.6	0	10	uncongest	Thursday	Off Peak	No rain	very high	very low	very low	Weekday	Clear	
4	2022/12/1 2:00	67.6	1657	2.4	67.2	0	10	uncongest	Thursday	Off Peak	No rain	very high	very low	very low	Weekday	Clear	
5	2022/12/1 3:00	67.2	1650	2.5	66.9	0	10	uncongest	Thursday	Off Peak	No rain	very high	very low	very low	Weekday	Clear	
6	2022/12/1 4:00	68	2804	4.1	67.8	0	10	uncongest	Thursday	Off Peak	No rain	very high	low	very low	Weekday	Clear	
7	2022/12/1 5:00	69.7	6587	9.2	69.6	0	10	uncongest	Thursday	Off Peak	No rain	very high	high	very low	Weekday	Clear	
8	2022/12/1 6:00	59.6	9300	17.8	59.6	0	10	uncongest	Thursday	AM Peak	No rain	very high	very high	low	Weekday	Clear	
9	2022/12/1 7:00	49.6	10014	25.2	49.6	0	10	congest	Thursday	AM Peak	No rain	high	very high	low	Weekday	Clear	
10	2022/12/1 8:00	43.4	9743	27.9	43.4	0	10	congest	Thursday	AM Peak	No rain	high	very high	high	Weekday	Clear	
11	2022/12/1 9:00	42.6	9662	26.2	42.6	0	10	congest	Thursday	AM Peak	No rain	high	very high	high	Weekday	Clear	
12	2022/12/1 10:00	53.5	10202	17.8	53.5	0	10	uncongest	Thursday	AM Peak	No rain	very high	very high	low	Weekday	Clear	
13	2022/12/1 11:00	56.4	10304	16.7	56.4	0	10	uncongest	Thursday	Off Peak	No rain	very high	very high	low	Weekday	Clear	
14	2022/12/1 12:00	56.6	10527	17.1	56.6	0	10	uncongest	Thursday	Off Peak	No rain	very high	very high	low	Weekday	Clear	
15	2022/12/1 13:00	53.5	10871	19.1	53.5	0	10	congest	Thursday	Off Peak	No rain	very high	very high	low	Weekday	Clear	
16	2022/12/1 14:00	34.9	9867	29.9	34.9	0	10	congest	Thursday	PM Peak	No rain	low	very high	high	Weekday	Clear	
17	2022/12/1 15:00	21.7	8045	40.6	21.7	0	10	congest	Thursday	PM Peak	No rain	low	high	very high	Weekday	Clear	
18	2022/12/1 16:00	16.9	6506	46.1	16.9	0	10	congest	Thursday	PM Peak	No rain	very low	high	very high	Weekday	Clear	
19	2022/12/1 17:00	15.4	5945	48.4	15.4	0	10	congest	Thursday	PM Peak	No rain	very low	high	very high	Weekday	Clear	
20	2022/12/1 18:00	15.9	6319	46.8	15.9	0	10	congest	Thursday	PM Peak	No rain	very low	high	very high	Weekday	Clear	
21	2022/12/1 19:00	25.2	8487	32.9	25.2	0	10	congest	Thursday	PM Peak	No rain	low	very high	high	Weekday	Clear	
22	2022/12/1 20:00	57.7	9629	15.1	57.8	0	10	uncongest	Thursday	Off Peak	No rain	very high	very high	low	Weekday	Clear	
23	2022/12/1 21:00	66.9	9795	13.8	66.8	0	10	uncongest	Thursday	Off Peak	No rain	very high	very high	low	Weekday	Clear	
24	2022/12/1 22:00	67.6	8628	11.6	67.6	0	10	uncongest	Thursday	Off Peak	No rain	very high	very high	very low	Weekday	Clear	
25	2022/12/1 23:00	68.8	6264	8	68.8	0	10	uncongest	Thursday	Off Peak	No rain	very high	high	very low	Weekday	Clear	
26	2022/12/2 0:00	68.1	4105	5.3	68	0	10	uncongest	Friday	Off Peak	No rain	very high	low	very low	Weekday	Clear	
27	2022/12/2 1:00	68.3	2731	3.5	68.3	0	10	uncongest	Friday	Off Peak	No rain	very high	very low	very low	Weekday	Clear	
28	2022/12/2 2:00	67.7	2014	2.8	67.9	0	4	uncongest	Friday	Off Peak	No rain	very high	very low	very low	Weekday	Clear	
29	2022/12/2 3:00	66.6	1782	2.7	66.8	0.01	10	uncongest	Friday	Off Peak	Light rain	very high	very low	very low	Weekday	Clear	
30	2022/12/2 4:00	67.4	2791	4	67.5	0	5	uncongest	Friday	Off Peak	No rain	very high	low	very low	Weekday	Clear	

Figure 12: example of dataset

And accurate definition of variables and state for Bayesian network model is shown below:

Variable	States and definitions
Environment variables (background variables)	
Time of day	AM peak: 6 am – 10 am in weekdays (5 hours) PM peak: 2 pm – 7 pm in weekdays (6 hours) Off peak: the other time in weekdays and all weekends
Day of week	Weekday: Monday – Friday Weekend: Saturday and Sunday
Hourly precipitation	No rain: 0 inches/hour Light rain: 0 < Hourly precipitation <= 0.1 inches/hour Moderate rain: 0.1 < Hourly precipitation <=0.3 inches/hour Heavy rain: >0.3 inches/hour
Hourly visibility	Clear: >= 3 miles Haze: < 3 miles
Traffic condition variables	
Speed (mph) Occupancy (%) Flow (vehicles/hour)	very low: < 0.25*max_value low: 0.25*max_value <= value < 0.5*max_value high: 0.5*max_value <= value < 0.75*max_value very high: 0.75*max_value <= value < max_value+1
Congestion variables	
Congestion	1: Congest: Occupancy >= Occcrit 0: Uncongest: Occupancy < Occcrit

3. Road congestion

3.1 Bayesian network model structure

1) The Bayesian network model in study

Since the variables of the Bayesian network model are specified, the next step is to specify the qualitative relationships between the variables. First, I

conducted the following Bayesian network modeling with reference to the 2016 study, in which the environment variables directly affect the traffic condition and also congestion. And the environment variables day of week, time of day, hourly precipitation, and hourly visibility do not affect each other, and the conditional independence assumption can also apply to the traffic condition variables.

However, if I choose a different way to define the time of day, where AM and PM Peak exist only in the weekday, the variables Day of Week and Time of Day are not independent. After considering all the conditions, the Bayesian network model is shown below:

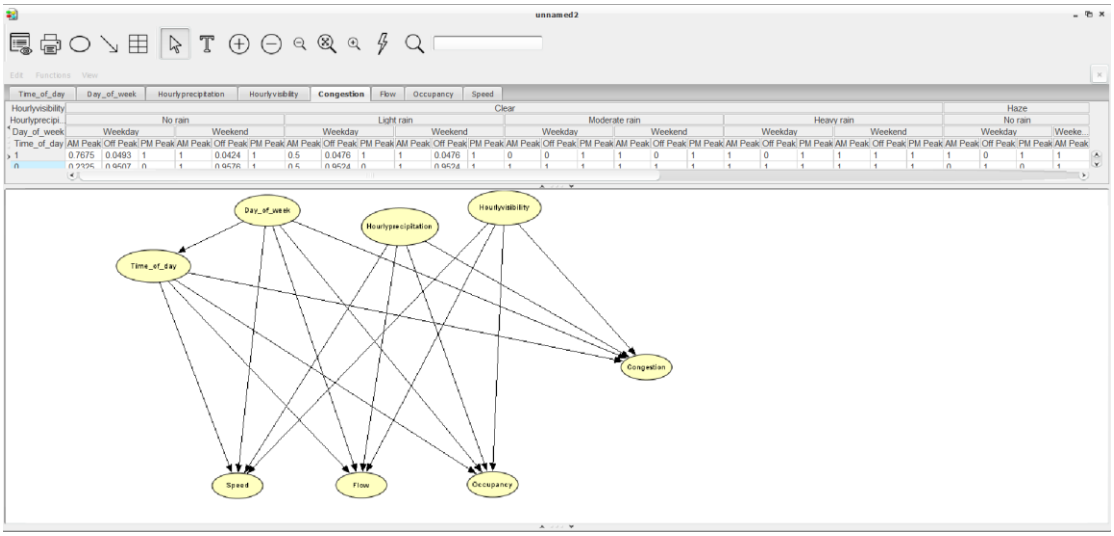

Figure 13: the first BN Model of this study based on the study in 2016

3.2 Realization of Bayesian network model

I choose nearly half of my dataset as a training model, which means the data of training model is from 2022.12.1 – 2023.1.18, so that the data contains 7 whole weeks. And especially the relationship between Day of week and Time of day is shown below, which means AM and PM Peak only exist in Weekdays:

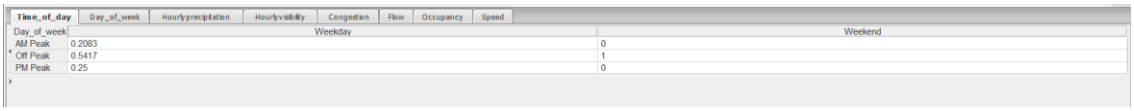

Figure 14: relationship between nodes Time of day and Day of week

After calculating the probability of these variables, I specify the states of these variables and the conditional probability table of each node, and then run the network, the result is shown below:

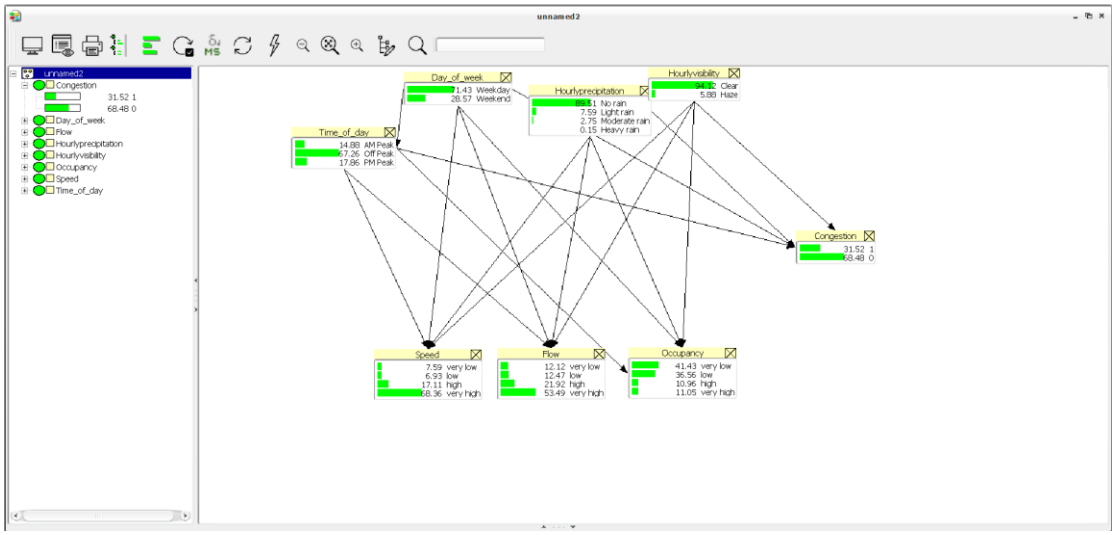

Figure 15: Probability distribution of the first BN Model

The next step is to diagnostic the reasons of congestion when it occurs with this Bayesian network model. The Figure below shows the posterior probability distribution of each variables when congestion has been observed, which means P (Congestion = congest) = 1:

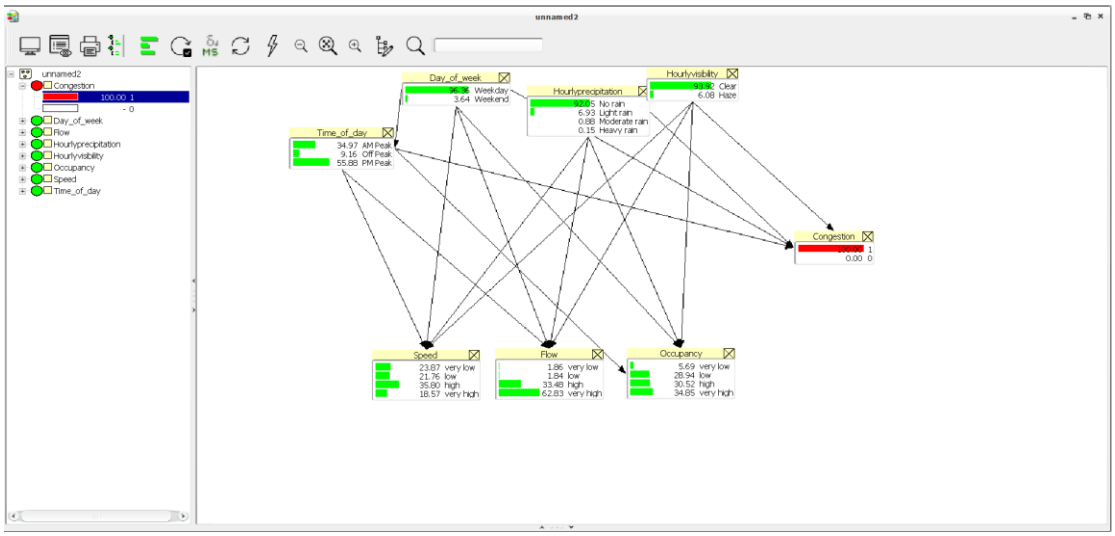

Figure 16: Probability distribution of the first BN Model when congestion occurs

From Bayesian network model we can also prognostic the congestion. The example shows as followed Figure, which includes the situation that when Time of Day is PM Peak, Day of week is Weekday, Hourly precipitation is No rain and Hourly visibility is Clear, the probability of congestion is 76.75%.

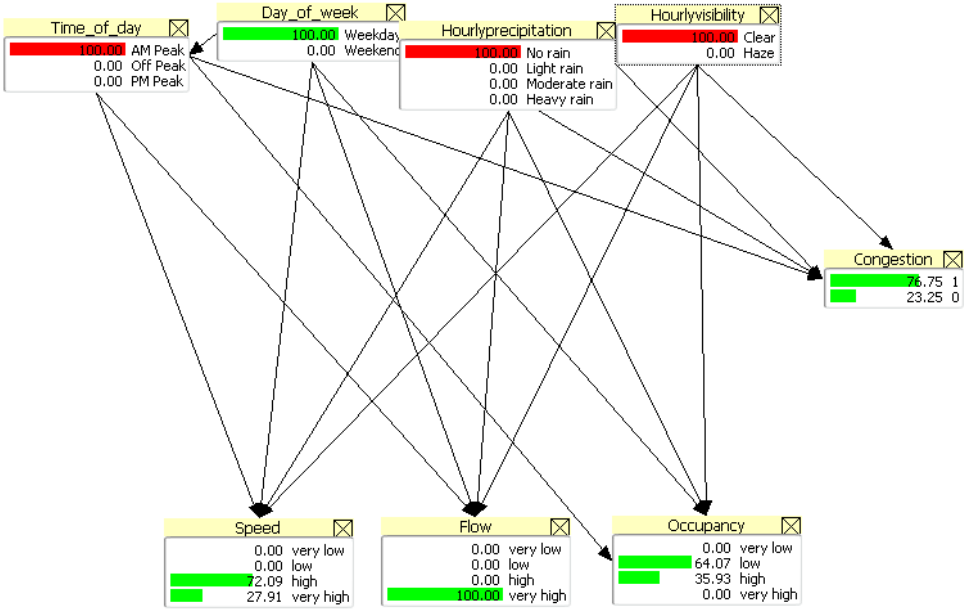

Figure 17: Probability distribution of the first BN Model in a selected situation

3.3 discussion about Bayesian network model above

1) change of probability distribution

when a congestion has been observed, the posterior probability distribution is shown above, which can be compared with the prior probability distribution in Figure:

- Time of day: the probability of being AM Peak has increased from 14.88% to 34.97%, and the probability of being PM Peak has also increased from 17.86% to 55.88%, meanwhile the probability of being Off Peak has decreased from 67.26% to 9.16%.
- Day of week: the probability of being Weekday has increased from 71.43% to 96.36%.
- Hourly precipitation: the probability of being No rain has lightly increased from 89.5% to 92.05%.
- Hourly Visibility: the probability of being Clear has lightly decreased from 94.12% to 93.92%.

The probability distribution of traffic condition variables has changed obviously, which can be discussed in the way below:

- Speed: the probability that shows the Speed is very high has decreased from 68.36% to 18.57%, meanwhile the probability that shows the speed is very low increased from 7.59% to 23.87%.
- Flow: the probability that shows the Flow is very high has increased from 53.49% to 62.83%, meanwhile the probability that shows the flow is very low decreased from 12.12% to 1.86%.
- Occupancy: the probability that shows the Occupancy is very high has increased from 11.05% to 34.85%, meanwhile the probability that shows the Occupancy is very low has decreased from 41.43% to 5.69%

2) Analyze of this Bayesian network model

After discussing the difference of the probability distribution between the prior and posterior probability, we also need to analyze the Bayesian network model with the help of Analysis Wizard in HUGIN, to access how good the Bayesian network model is. I choose the data from 2023.1.19-2023.2.28 as the data to test the model. The results are shown below:

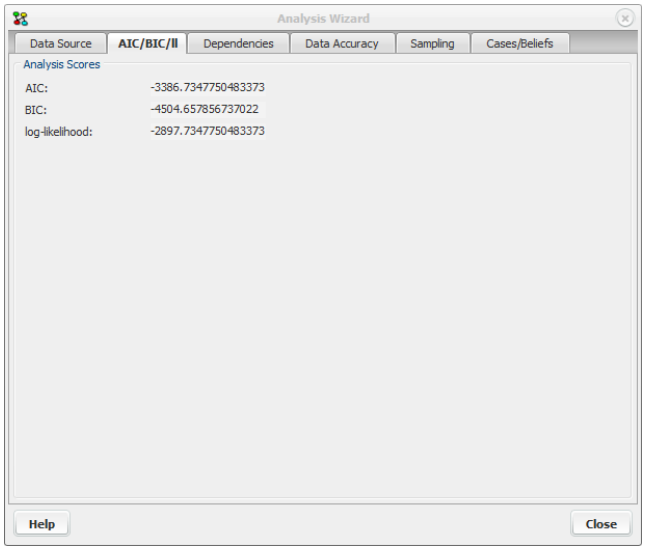

Figure 18: Scoring function of the first BN Model

These scoring function numbers are available to evaluate a Bayesian network model; the smaller the numbers, the better the Bayesian network model performs.

The next step is to evaluate the dependence between the variables. First, from the difference in the probability distribution, we can roughly speculate that time of day and day of week have relevant significant effects on congestion, but hourly precipitation hardly affects the congestion states, which can also be seen in the Analysis Wizard:

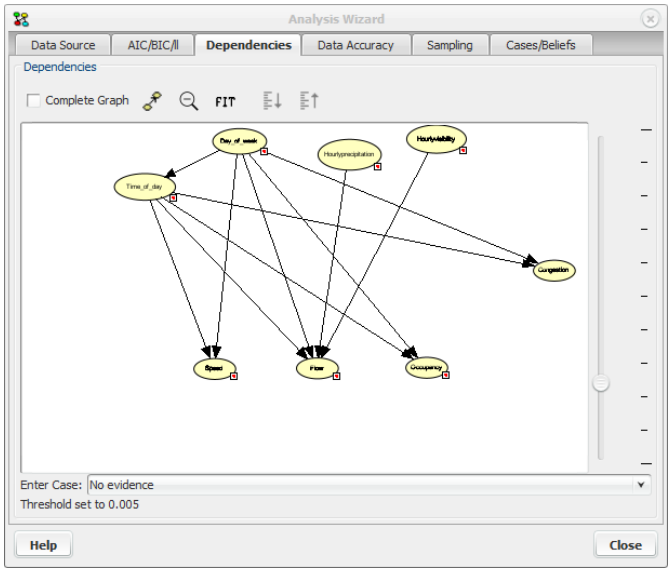

Figure 19: Dependencies of the first BN Model

When the threshold is set to 0.005, the dependence between hourly precipitation and hourly visibility has disappeared. Although at first I thought that precipitation and visibility must affect the congestion state, but from the dataset we can learn that even if there is heavy rain and poor visibility, when the flow and occupancy is very low, for example at 1 am, there is unlikely to be congestion.

The final step is data accuracy, which evaluates the Bayesian network model through the confusion matrix. The result of the congestion data accuracy is shown below:

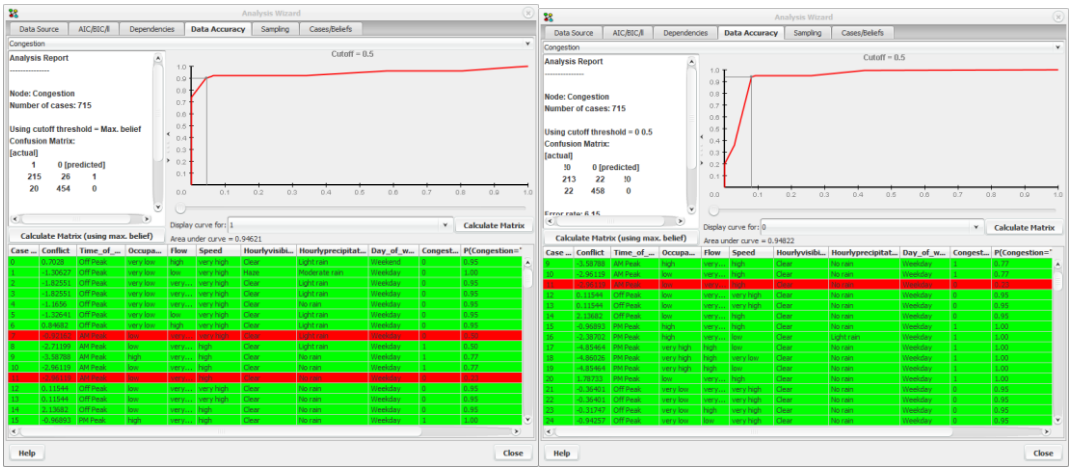

Figure 20: Data accuracy of congestion node of the first BN Model

4. Discussion

4.1 Discussion about sensor uncertainties

Sensor uncertainties are important to the Bayesian network model, because sometimes the data from sensors are not accurate, but the sensors have probability to measure a wrong data, which means the real data is distributed in a range which based on the measured data, for example, a measured data add a Gaussian distribution which represents the uncertainty of the sensor.

In our study, it is also possible that the traffic sensor and the weather sensor have uncertainties, which means the data are not so accurate. But the parameters of sensor uncertainties are normally not given by the open datasets, therefore based on the others studies, I defined the variables in discrete values instead of continuous values, to reduce the influence of sensor uncertainties.

4.2 Discussion about data labeling

Data labeling is also a important part of the Bayesian network model, because each variable has different states, which represent different situations. And with the labels we can easily recognize the prior and posterior probability with the name of labels, for example, P (Congestion=1 | Day of week=Weekday, Time of day=AM Peak, Hourly precipitation=light rain, Hourly visibility=haze) =0.8889 means when Day of week is Weekday, Time of day is AM peak, Hourly

precipitation is light rain and hourly visibility is haze, the probability of congestion is 0.8889.

4.3 Discussion about another type of Bayesian network structure for congestion diagnostic

The Bayesian network model is based on the study in year 2016, but since the variables can be defined as background variable, symptom variable (which can be observed as a consequence of the presence of the problem and hence will be available after the occurrence of the problem), mediating variable and problem variable, there is also another structure of probabilistic network which is different from the Bayesian network above, which means congestion also affect the traffic variables such as Speed, Flow and Occupancy:

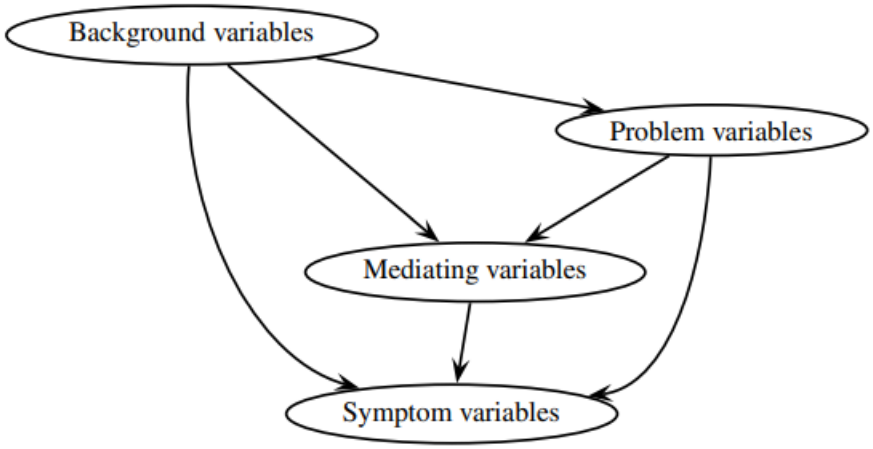

Figure 21: Typical casual structure of a BN Network

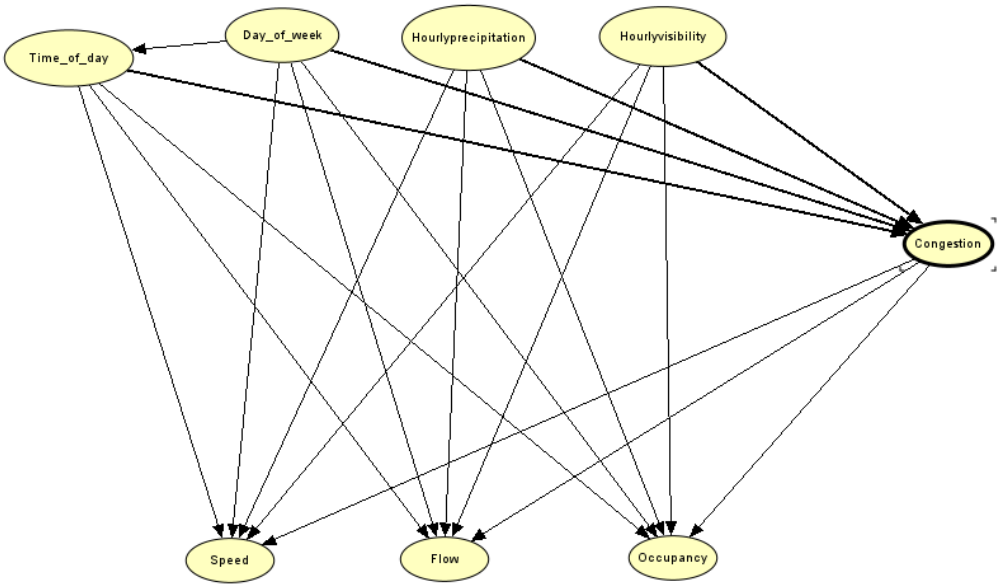

Figure 22: The Second BN Model for congestion diagnosis

In order to identify which structure is better, I also calculate the prior probabilities of this Bayesian network model and test it with the testing data, and

from the Analysis Wizard we can assess these two structures, and the results are shown below:

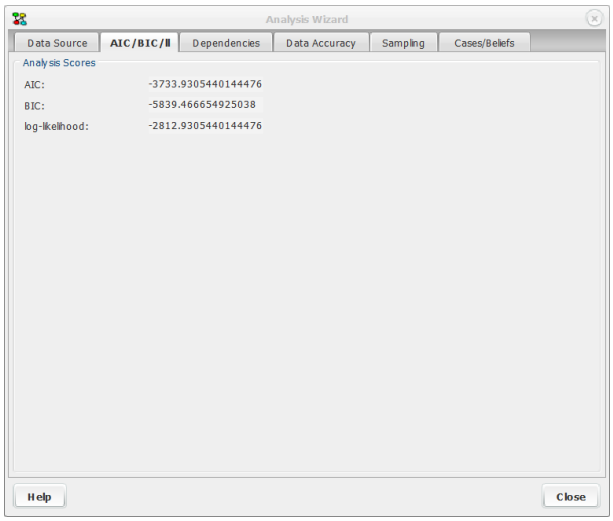

Figure 23: Scoring function of the second BN Model

The parameters of AIC and BIC is bigger than the parameters of the first network, but the log-likehood is lower than the parameter of the first network.

And the results of confusion matrix and the relative curves are shown below, which on the left side belong to the first network, and on the right side belong to the second network, Speed is used as the variable:

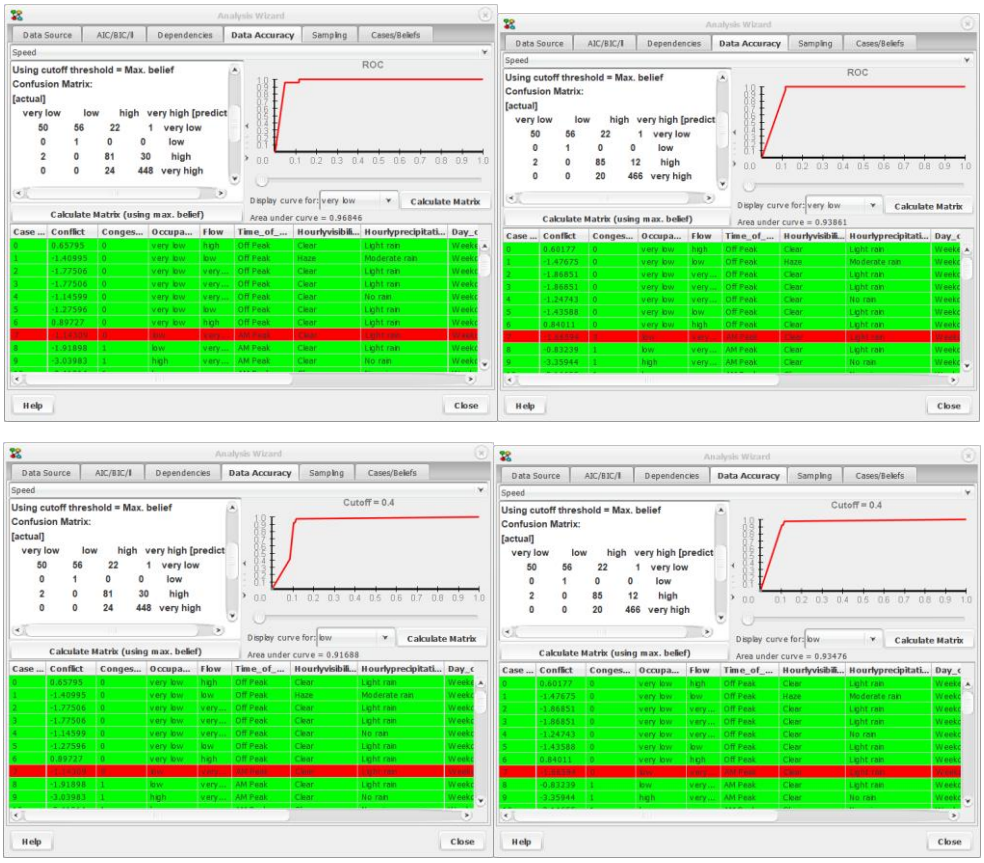

Figure 24: Data accuracy of Speed between the first and second BN Model

The results show that the second Bayesian network structure has a higher accuracy on traffic parameters, which can be also proved with the data accuracy of the confusion matrix of congestion:

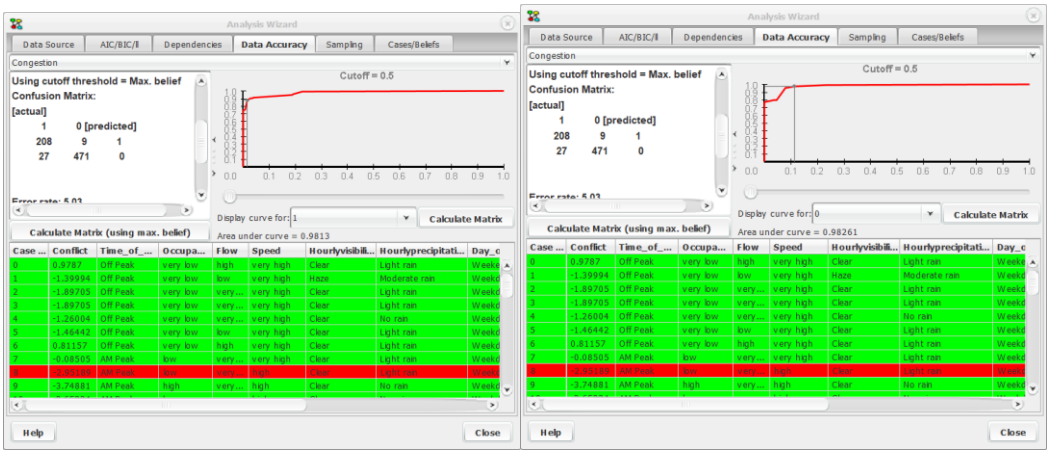

Figure 25: Data accuracy of congestion of the second BN Model

Which has also less error than the first Bayesian network model, in other words, it can diagnostic the congestion more accurately than the first model.

4.4 Discussion about another Bayesian network structure for congestion Prediction

After the discussion with Prof. Weidl, I know that we can also take the

congestion state as a predicted variable, which represents a causal consequence from the influence variables, such as Speed, Flow and Occupancy. And I also build a Bayesian network model in order to predict the congestion, the model is shown below:

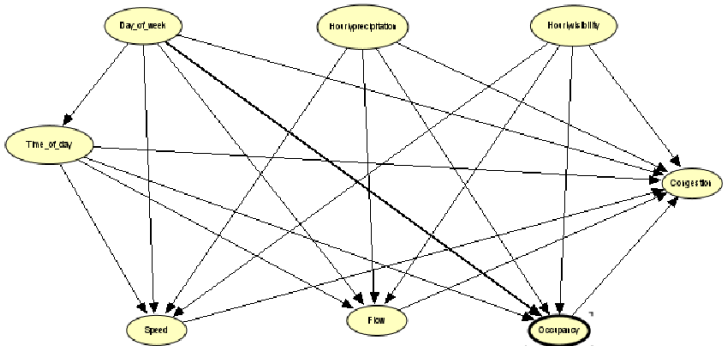

Figure 26: the third BN Model for congestion prediction

When I calculate the probability distribution, I also find the difference between this model and the model to diagnostic congestion: At first, we define a situation of traffic variables and background variables as S, and congestion situation as C. the probability distribution of the model for congestion diagnosis is about when a congestion occurs or not, what is the probability distribution of traffic variables (which can be also called symptom variables, writes as P(S|C)). Which means the training data of traffic variables will be divided into different groups based on the congestion states (writes as P(C)), so that when we test this Bayesian network model, the result of congestion diagnosis will be based on the observed traffic variables states with the formula P(C|S)=P(S|C)*P(C)/P(S).

The probability distribution of the model for congestion prediction is about when the states of traffic variables are observed, what is the probability distribution of congestion states (write as P(C|S)), which means the training data of congestion states will be divided into different groups based on the states of traffic variables, so that we can use this Bayesian network model to predict the congestion states with the prior probability distribution in this model.

The results of the model for congestion prediction are shown below:

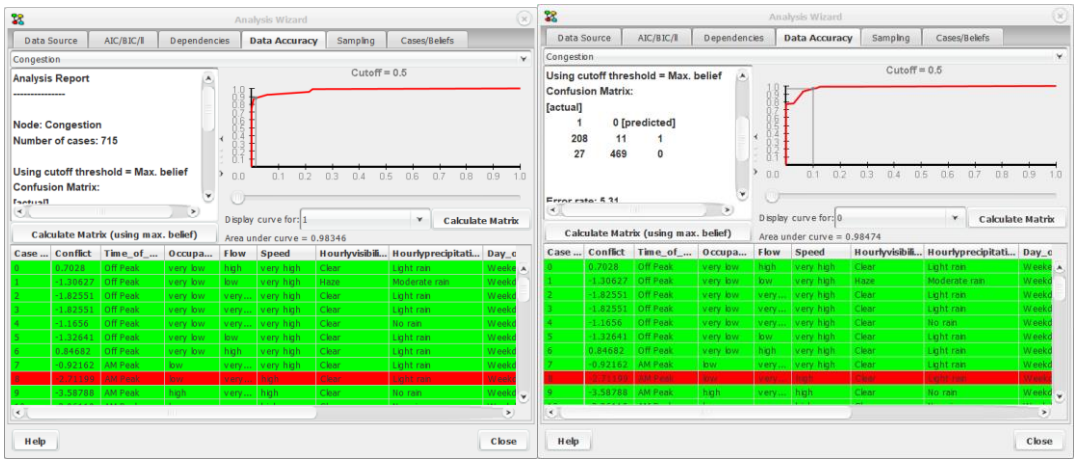

Figure 27: Data accuracy of congestion of the third BN Model

The result shows that the data accuracy of this network is also better than the first Bayesian network model, and the result is closed to the result of the Bayesian network model for congestion diagnosis.

5. Conclusion

From this study, I learned a lot of knowledge about Bayesian network and probability distribution, and also knowledge about how to build and run a Bayesian network model with Hugin. And from the historical studies I learned the Road congestion with Bayesian network, for example, how to build a Bayesian network model to predict and diagnose the road congestion, which is better than the traditional statistical methods, because they have limitations in capturing complex dependencies and uncertainties in environmental variables and traffic variables in urban networks.

And from the 3 different BN Models we built for the different situations of road congestion, which all contain the probabilistic dependency structure between environmental variables and traffic variables, but are different from the relationship between the traffic variables and congestion states, we can not only use them to quantify the contribution of each cause or the combination of multiple causes to traffic congestion, in order to diagnose the traffic congestion, but also predict the traffic congestion based on the previous data and situations.

From calculating the probability distributions, I also had a deeper understanding about the difference between diagnosis and prediction in Bayesian network, and also the relationship between prior and posterior probability distributions, which is useful to the further learning of data analyze etc.