

Diagnostic and Predictive Models for Traffic Congestion

Using Bayesian networks to study the relationship between time and congestion

Xingyu Xing

TH Aschaffenburg

MT_27: Anwendungen der Mechatronik (Student Research Project)

Prof. Weidl

13/3/2023

Diagnostic and Predictive Models for Traffic Congestion

This paper presents a Bayesian network (BN) analysis method for modelling Diagnostic and Predictive Models of the causes of congestion and analysing the probability of traffic congestion under various time-condition, in order to achieve congestion Diagnosis of causes or congestion prediction based on existing situations.

Overview

The paper begins with background information detailing the causes of today's traffic congestion problems and the importance to solve them, followed by a discussion of how this paper will solve the problem of congestion diagnosis and prediction through BN. As the paper goes, the selection of variables, database selection, data selection, data analysis, etc., encountered in the process of building the model are explained in detail. The paper will conclude with a display of the completed model and its results.

Background Information

With the development of modern transport and the automobile industry, cars, as a very important means of transport, the number and per capita possession of them, is constantly increasing. This increase in numbers and usage has made life easier for people, but has also led to serious traffic congestion today. At the same time, traffic congestion is limiting the development of urban transport and cities as a whole. Therefore, the traffic congestion is a problem that must be solved on the way to the development of a modern transport industry.

Topic Discussion

In order to solve traffic congestion, one of the questions we have to answer is what causes traffic congestion: assuming that traffic congestion is currently occurring, there are various reasons behind such a phenomenon, such as weather, visibility, time of day, traffic

accidents, etc. These reasons are the key to solving the traffic congestion. At the same time, these causes can also be used as a reference for predicting upcoming traffic congestion, so that we can anticipate congestion in advance and intervene as early as possible based on the current information.

Therefore, analyzing the relationship between these causes and traffic congestion is very important. The effects of these causes are complex, and not only do they interact with each other, e.g. weather affects visibility, but their effects on traffic congestion are not absolute, only a possibility. Therefore, based on such facts, we use Bayesian networks in this study to investigate such a network of relationships, with the aim of constructing a Bayesian network consisting of factors affecting congestion and congestion situations, in order to achieve congestion Diagnosis of causes or congestion prediction based on existing situations.

Bayesian network

A BN is a probabilistic graphical model that represents probabilistic relationships between a set of variables via a directed acyclic graph. 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. 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.

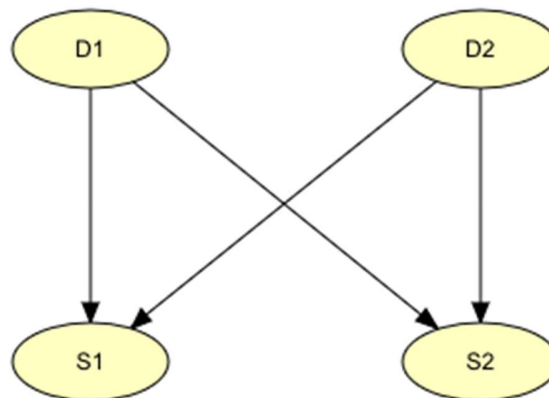


Figure 1 An example of Bayesian Network

Thus, the Bayesian network fits our experiment perfectly: The objective of this experiment is to investigate the probability of traffic congestion due to a combination of complex factors (weather, time of day, traffic accidents, etc.), i.e. prediction, and to analyse what causes traffic congestion and how likely it is to occur, i.e. diagnosis. Therefore, we can use a Bayesian network to model and eventually implement the function by following the steps of (a) defining variables, (b) specifying structure, and (c) specifying parameters

Factors Selection and the Networks

Based on life experience and reading of references, we have collected many factors that influence the level of congestion: for example, at the aspect of the environment: weather conditions, road slipperiness, visibility (Figure 2); at the aspect of time conditions: time of day, day of the week, month (Figure 3); and at the aspect of accidental conditions, such as traffic accidents (Figure 4), etc.

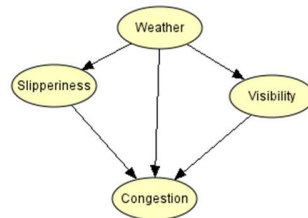


Figure 2 BN of Environment

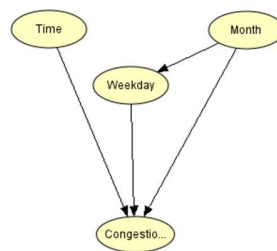


Figure 3 BN of time

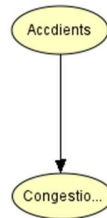


Figure 4 BN of Accidental Conditions

Selection of the Variables

Their respective Bayesian networks are shown in the figure, and at the same time they can interact with each other. This results in the following complex Bayesian network (Figure 5). It is worth noticing that although it may seem complex, it is actually made up of three relatively independent Bayesian systems and their interactions, which can be seen as three separate parts: environment, time, and accident. Although we are free to choose the factors we wish to study, our study needs to be based on the available and reliable database we have so far. So, we have chosen to study one of them. It is worth saying that even though we have studied one part of the network in this study, this part can be later fused with another part of the network that is also relevant to achieve a more comprehensive model construction.

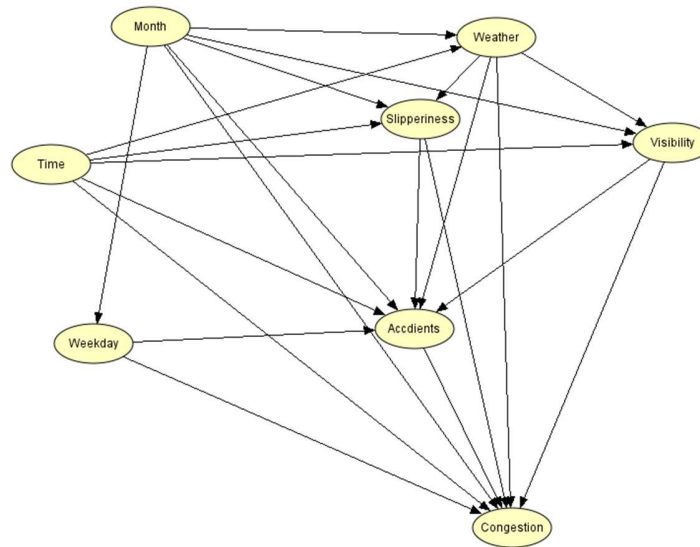


Figure 5 BN that is made up from three parts

In the end we chose the time part to build our Bayesian network, mainly because we lacked a database that could support the other part, and we found the appropriate database that could support us in this part of the experiment.

Build up of the Bayesian network

This Bayesian network should be explained. First, there is an influence of time of a day on congestion. In a 24-hour day, different time of day is related to different activities of people's life, and the relationship is usually fixed. At the same time, the activities of people's life are also more closely related to congestion, for example, at 5pm most people leave work, and they need a car to leave work, so congestion will occur. Therefore, the time of day has an impact on congestion.

Similarly, the influence of days of the week and months on congestion is also based on their impact on people's living activities, for example, people prefer to stay at home on

weekends, or in winter when snow has an impact on visibility and slippery ground, its coldness also leads to fewer or more cars and therefore has an impact on congestion.

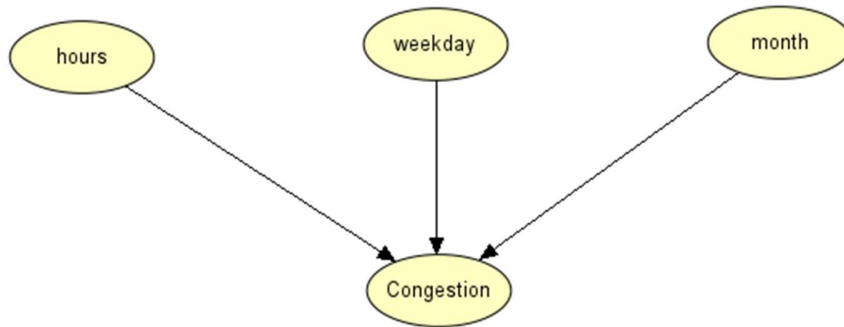


Figure 6 Corrected BN of Time/ Our Planned Model

It is also worth clarifying the effect of the month on the number of days of the week (in Figure 3), which requires some explanation and correction. It is true that the number of days of the week varies from month to month. For example, it is possible that there are five Mondays in January and only four Sundays. It depends on the different number of days in each month and the year in which the month falls. Therefore, if a Bayesian network is constructed for a specific year, this line is indispensable. But from a statistical point of view, the different weekdays should be independent of the month, since their circulation is not affected by changes in the month, and at the same time, the different weekdays are statistically equal in front of the month. In other words, the number of Mondays to Sundays occurring in a month should theoretically be the same in a sufficiently large amount of data. Since the model we wish to construct is applicable to a wider range, instead of a specific year, we have removed this line from the statistical point of view later on and made some corrections to the data. And finally, we get the corrected BN of time, and it's also our planned model.

Data Selection

Once we have set the model, we need to select data based on the model. The requirements the database need to meet are:

1. The data can determine the traffic congestion.
2. It contains the relevant factors we need to analyse.
3. the data should be realistic and specific.

Here we found the official Chicago database of congestion on all roads in the city from 2018 to the present “*Chicago Traffic Tracker Historical Congestion Estimates by Segment 2018-Current*”. The characteristics of the database fully meet our requirements above, but it also contains some drawbacks, such as the large volume of data and the large amount of useless data. Therefore, data selection has to be carried out before data processing.



Figure 7 Brief Introduction of the Database

Firstly, the data was selected through time. It is worth noting that the outbreak of the covid in 2019 and 2020 has had a significant impact on people's travel, with many people

choosing to work from home and so on, which may also have a great influence on the traffic congestion. Therefore, we have chosen the years 2021 and 2022 after the outbreak to ensure that people's lifestyles are back to normal and that we have chosen two years of data to obtain more valid statistics.

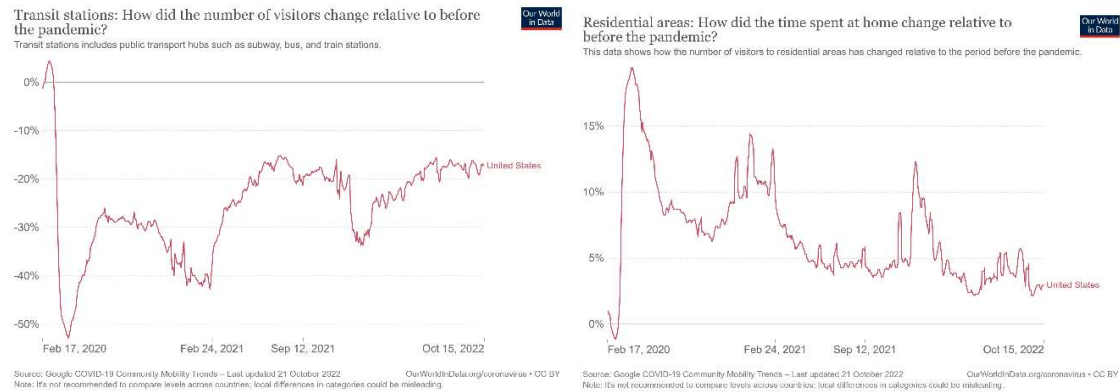


Figure 8 How the Covid-19 influence the Traffic Congestion

Second, the data is selected through location. The database contains congestion on all roads in Chicago, recorded using average passing speeds as a measure. Firstly, we could not select all roads in the city for the study. Because congestion in one part of the city would reduce congestion in another part of the city at the same time, and the interaction between the two would lead to more changes in terms of location rather than time, which would lead to errors in our judgement of congestion. Similarly, we cannot select multiple streets that run parallel or cross each other as the subject of our study. We therefore need to select just one street as the subject of the study.

Also, for the direction on that street, we could not choose to average the two-way lanes, because for most lanes, congestion to and from a place will show different levels of congestion at the same point in time, so using the speed to measure will wipe out the congestion feature.

Therefore, we need to choose the same direction on the same street as the study to calculate the feature

The location of the street should also be chosen. Firstly, it should be in the heart of the city, i.e. the amount of data needed should be high and the probability of congestion should be high, but it should also avoid continuous congestion. In other words, the congestion on that street needs to be obvious over time so that we can also make it easier for us to produce reasonable statistics.

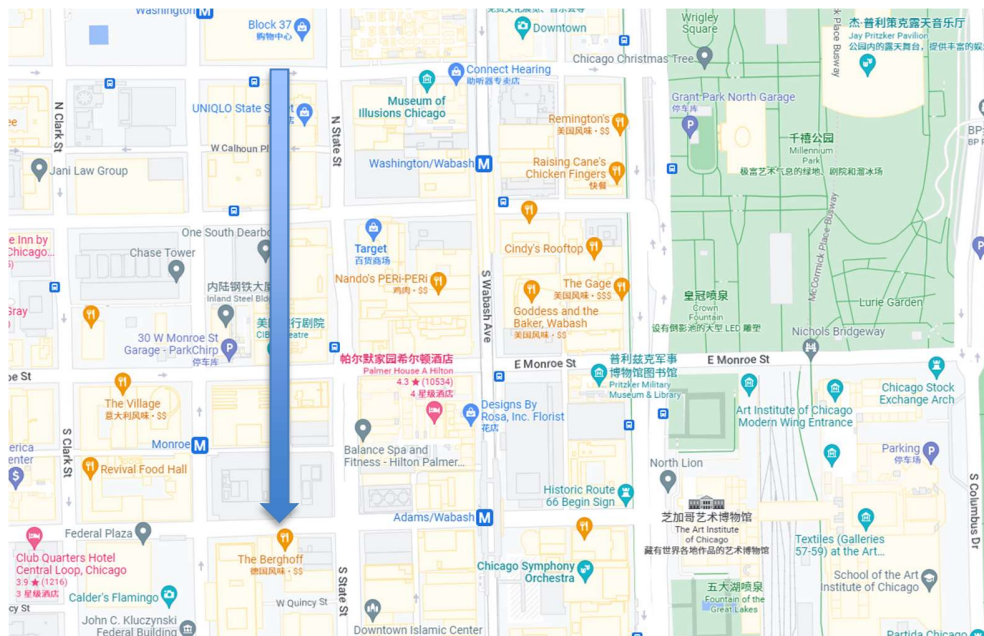


Figure 9 the Location of the Selected Street

We therefore chose State Street from Congress to Roosevelt University as the subject of our study, as it is a busy street in the heart of the city that perfectly meets all our requirements.

Finally, we also need to select the various data provided by the database, and we end up keeping the following columns: the time when the data was generated, the speed, the street name

and street direction for later checking of the data, as well as the time of day, the day of the week and the month. Among those, we also remove some of the invalid data.

Here are the columns of the selected database:

Table 1 the Columns of the Selected Database

TIME	the time when the data was generated	MM/DD/YYYY HH:MM: SS AM/PM
SPEED	Estimated traffic speed in miles per hour. Used to determine if there is congestion	NUMBER
STREET	Street name of the traffic segment. Used to check if data is eligible	“State”
DIRECTION	Traffic flow direction for the segment. Used to check if data is eligible	“SB”
FROM_STREET	Start street for the segment in the direction of traffic flow Used to check if data is eligible	“Congress”
TO_STREET	End street for the segment in the direction of traffic flow. Used to check if data is eligible	“Roosevelt”
HOUR	Hour of the day.	NUMBER from 0 to 23
DAY_OF_WEEK	Day of the week. Sunday = 1	NUMBER from 1 to 7
MONTH	Month of the year.	NUMBER from 1 to 12

TIME	SPEED	STREET	DIRECTION	FROM_STR	TO_STREET	HOUR	DAY_OF_W	MONTH
03/21/201	16	State	SB	Congress	Roosevelt	17	4	3
03/21/201	16	State	SB	Congress	Roosevelt	17	4	3
03/21/201	18	State	SB	Congress	Roosevelt	17	4	3
03/21/201	17	State	SB	Congress	Roosevelt	17	4	3
03/21/201	18	State	SB	Congress	Roosevelt	16	4	3
03/21/201	15	State	SB	Congress	Roosevelt	16	4	3
03/21/201	15	State	SB	Congress	Roosevelt	16	4	3
03/21/201	17	State	SB	Congress	Roosevelt	16	4	3
03/21/201	17	State	SB	Congress	Roosevelt	16	4	3
03/21/201	20	State	SB	Congress	Roosevelt	16	4	3
03/21/201	25	State	SB	Congress	Roosevelt	15	4	3

Figure 10 the selected data

Data processing and parameterisation

Before the number of individual data with different characteristics is counted from the selected data, we still need some processing.

Definition of the Variables and the Events

The time, days of the week and month have corresponding data. Congestion is defined by the official Chicago speed limit, which is 30mph, so it is officially considered to be above 20mph as no congestion and below 20mph as congestion. So we use this as a filter to calculate the number of data for each condition.

The screenshot shows the Chicago Traffic Tracker website. The header includes the logo "CHICAGO TRAFFIC TRACKER" and the text "Realtime Traffic Information for Chicago". Below the header, there is a section titled "About Chicago Traffic Tracker(beta)" which explains that the website is presented by the Chicago Department of Transportation (CDOT) and provides realtime traffic conditions on arterial streets, ADT volumes, traffic signal locations, pedestrian counts, and automated red-light enforcement program intersections. A section titled "Realtime Arterial Traffic" explains that arterial traffic condition is estimated utilizing realtime GPS probes received from transit buses operated by the Chicago Transit Authority (CTA). It also includes a "Speed Color Codes" legend: green for over 20 mph, yellow/orange for 10 to 20 mph, and red for below 10 mph. The text states that the segment speed color code on the map is based on the speed limit of the street segment, and that on most city streets, the speed is limited to 30 mph. Consequently, the congestion map shows green for speed above 20 mph, yellow/orange for 10 to 20 mph and red for speed below 10 mph.

Figure 11 the Definition of Congestion in Chicago

We then define some events:

We use letters to represent the corresponding events: Congestion(C), Time(T), Weekday(W), Month(M)

C_0 : the situation, that there is no congestion now (speed ≥ 20 mph).

C_1 : the situation, that there is a congestion now (speed < 20 mph).

T_i : it's i ($i = 0, 1, 2 \dots, 23$) o'clock now.

W_i : it's the i ($i = 1, 2 \dots, 7$) weekday of a week. (e.g. $i=1$ main it's Monday).

M_i : it's the i ($i = 1, 2 \dots, 12$) month of a year. (e.g. $i=1$ main it's January).

the number of individual data with different characteristics can then counted from the selected data.

	A	B	C	D	E	F	G
1	month	1	1	1	1	1	1
2	week	1	1	2	2	3	3
3	hour	≥ 20	< 20	≥ 20	< 20	≥ 20	< 20
4	0	22	7	22	1	14	3
5	1	15	4	17	7	15	2
6	2	8	2	8	2	6	1
7	3	7	4	6	0	8	1
8	4	8	1	11	1	8	0
9	5	14	1	17	3	16	2
10	6	20	4	25	3	23	0
11	7	17	9	14	15	20	2
12	8	22	4	10	17	15	6

Figure 12 Sections of statistical results for different categories of data

Probability of the Variables

We then calculate the probability of occurrence of the three variables according to the data:

$$P(T_i) = \frac{\text{Number of data that meet the condition}}{\text{Number of all data}}$$

$$P(W_j) = \frac{\text{Number of data that meet the condition}}{\text{Number of all data}}$$

$$P(M_k) = \frac{\text{Number of data that meet the condition}}{\text{Number of all data}}$$

However, in practice our data are not collected evenly and there is a large amount of missing data (firstly, the test interval does not guarantee that the same number of data are collected at the same feature point, in other words, for example, the number of data points at different times does not reflect the probability of occurrence at that point in time, and we also have missing data from speed measurements, which we have removed in the) The data we have obtained do not correlate well with time of day, day of week and month, i.e. the number of data points does not represent their probability of occurrence. Therefore, we make a correction here, i.e. we modify the proportions according to objective facts, as follows:

$$P(M_k) = \frac{\text{The number of hours that match the time feature during the measurement time period in practice}}{\text{The number of hours of the time period in practice}}$$

Also, as we have previously made a correction to the model based on statistical significance, the actual meaning here is:

$$P(M_k) = \frac{\text{The number of hours that match the time feature during the measurement time period that long enough}}{\text{The number of hours of the time period that long enough}}$$

The same is true for the other two variables, then there are:

$$P(T_i) = \frac{\text{The number of hours that match the time feature during the measurement time period that long enough}}{\text{The number of hours of the time period that long enough}} = \frac{1}{24}$$

$$P(W_j) = \frac{\text{The number of hours that match the time feature during the measurement time period that long enough}}{\text{The number of hours of the time period that long enough}} = \frac{1}{7}$$

$$\begin{aligned} P(M_k) &= \frac{\text{The number of hours that match the time feature during the measurement time period that long enough}}{\text{The number of hours of the time period that long enough}} \\ &= \frac{\text{Number of days in the month}}{\text{Number of days in a year}} \end{aligned}$$

It is worth noting that before further changes are proposed (in Figure 3), the probability formula to be modified therein should be:

$$P(W_i|M_j) = \frac{P(W_i, M_j)}{P(M_j)}$$

After further modification (in Figure 6), with W and M independent of each other, so it should be:

$$P(W_i|M_j) = P(W_j)$$

Because:

$$P(W_i, M_j) = P(W_i) \times P(M_j)$$

Probability of the congestion

Next, we calculate the probability of congestion for the corresponding congestion at that time characteristic, according to the Bayesian formula:

$$P(C_0 | T_i, W_j, M_k) = \frac{P(C_0, T_i, W_j, M_k)}{P(T_i, W_j, M_k)}$$

And for the two part in it:

$$P(C_0, T_i, W_j, M_k) = \frac{\text{Number of data that meet the condition}}{\text{Number of all data}}$$

$$P(T_i, W_j, M_k) = \frac{\text{Number of data that meet the condition}}{\text{Number of all data}}$$

The probabilities calculated here are independent of the probability of the occurrence of this time feature itself, so the correction we make ahead does not have an impact on the results here.

0	0.758621	0.241379	0.956522	0.043478	0.823529	0.176471
1	0.789474	0.210526	0.708333	0.291667	0.882353	0.117647
2	0.8	0.2	0.8	0.2	0.857143	0.142857
3	0.636364	0.363636	1	0	0.888889	0.111111
4	0.888889	0.111111	0.916667	0.083333	1	0
5	0.933333	0.066667	0.85	0.15	0.888889	0.111111
6	0.833333	0.166667	0.892857	0.107143	1	0
7	0.653846	0.346154	0.482759	0.517241	0.909091	0.090909
8	0.846154	0.153846	0.37037	0.62963	0.714286	0.285714

Figure 13 Sections of statistical results for different categories of data (after Bayesian formula)

With this, we have calculated all the data needed to build our model. Our model has also been constructed (Figure 13).

Model display and Conclusion

Our final model is shown in the diagram, with its parameters set alongside.

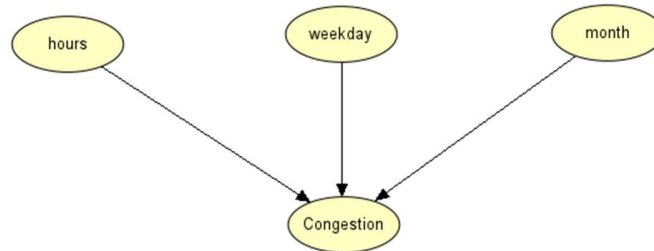


Figure 14 The final model

hours	weekday	month	Congestion
0	0.416666		
1	0.416666		
2	0.416666		
3	0.416666		
4	0.416666		
5	0.416666		
6	0.416666		
7	0.416666		

Figure 15 Parameters of "hours"

hours	weekday	month	Congestion
Monday	0.142857		
Tuesday	0.142857		
Wednesday	0.142857		
Thursday	0.142857		
Friday	0.142857		
Saturday	0.142857		
Sunday	0.142857		

Figure 16 Parameters of "weekday"

hours	weekday	month	Congestion
January		0.084931	
February		0.076712	
March		0.084931	
April		0.082192	
May		0.084931	
June		0.082192	
July		0.084931	
August		0.084931	

Figure 17 Parameters of "month"

hours	weekday	month	Congestion
month			
weekday			
hours			
no congest...	0.758...	0.789...	0.8
congestion...	0.241...	0.210...	0.2

Figure 18 part of the Parameters of "Congestion"

Explanation of the Model

In summary, we can conclude that:

1. The probability of traffic congestion on the road is 30.1%

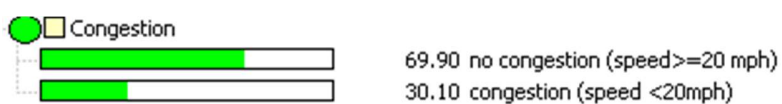


Figure 19 Probability of congestion and no congestion

2. The model can be used as a diagnostic.

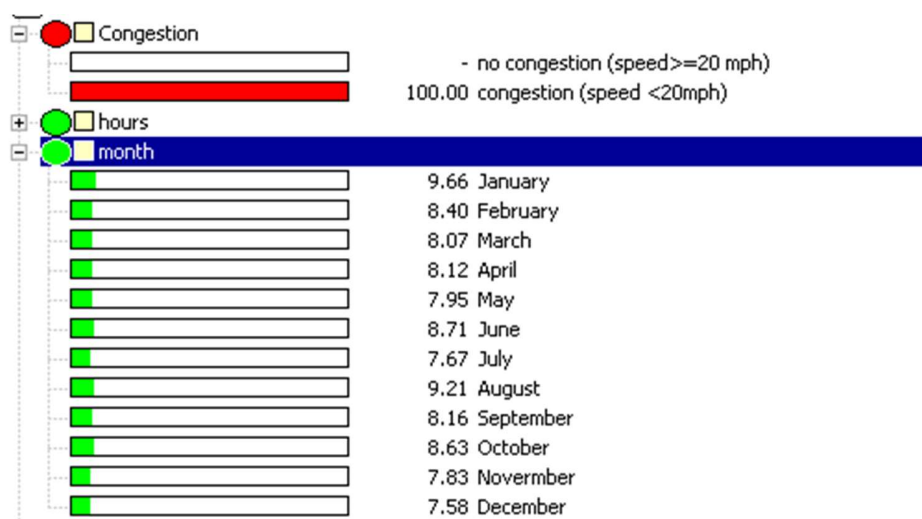


Figure 20 if congestion, the probability of month

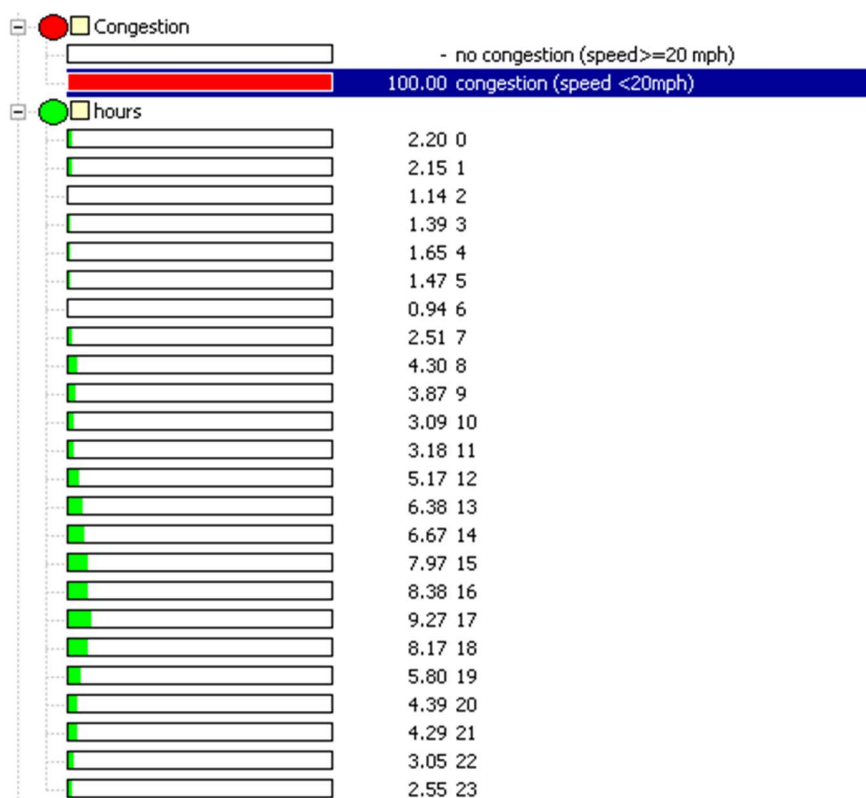


Figure 21 if congestion, the probability of hours

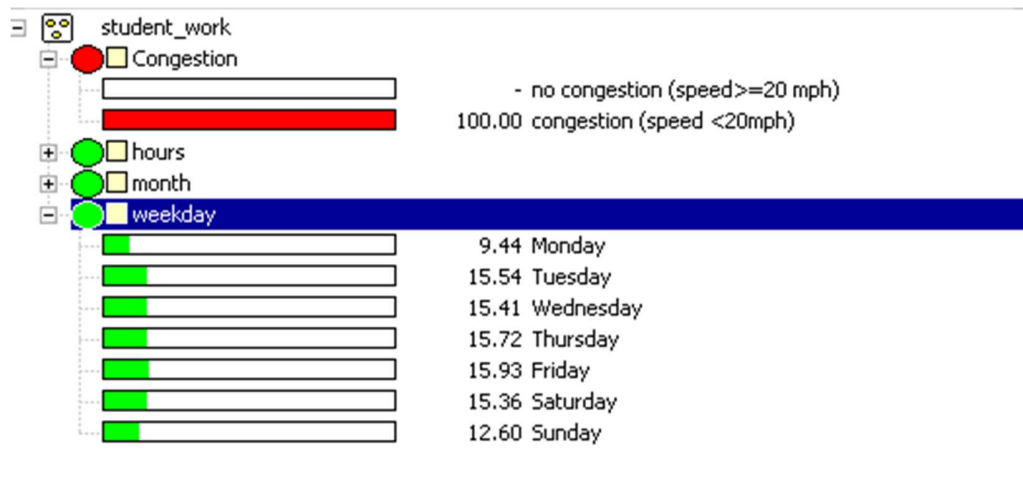


Figure 22 if congestion, the probability of weekday

If there is current congestion, the most likely time is 17:00 (Figure 21), January (Figure 20), Friday (Figure 22). At the same time, if more conditions are set, it is possible to determine how likely it is that this is the cause, i.e. the cause analysis.

For example, if the traffic is currently congested and it is a Wednesday in August, at 5pm, then the probability of the traffic congestion being caused by "5pm" in this case is 11.17% (Figure 23).

For example, if it is 5pm on a Wednesday in August, the probability of congestion is 82.69%, i.e. there is a high probability of traffic congestion, and early intervention can be made (Figure 24).

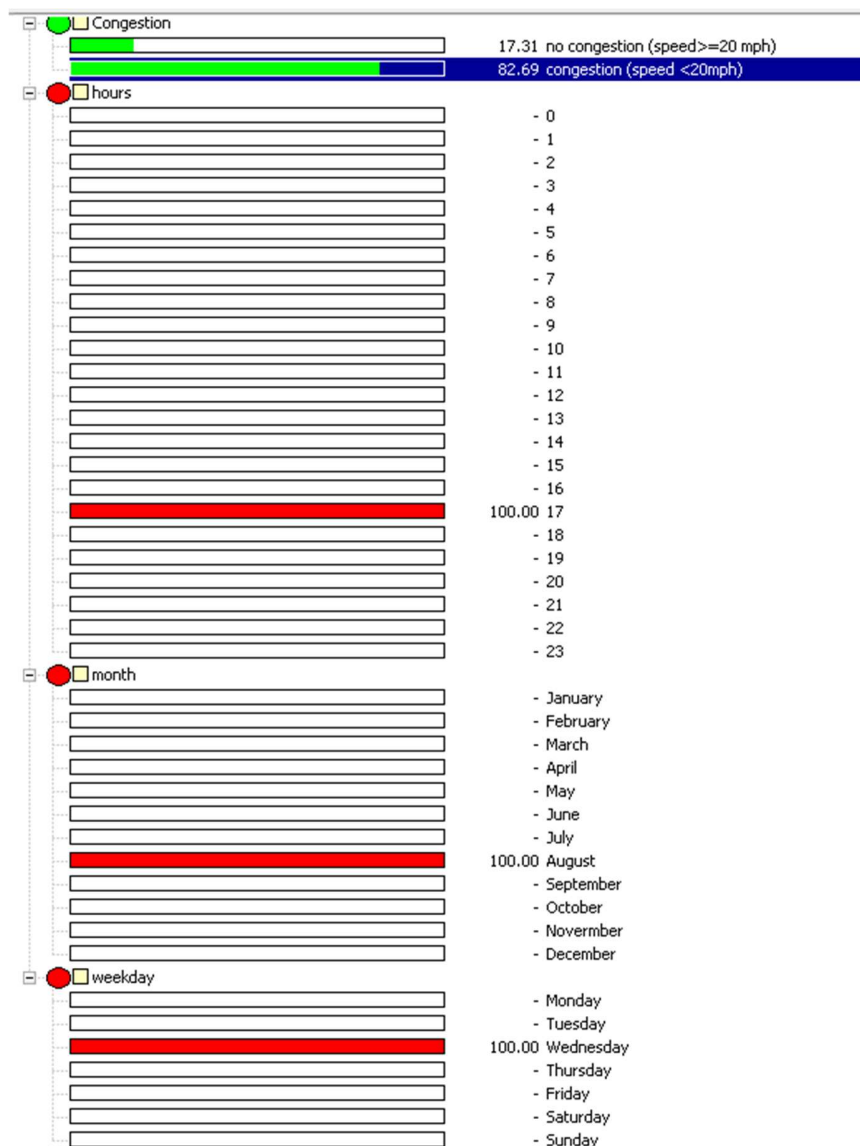


Figure 24 an example of congestion prediction

If it is currently 7am on a Tuesday in April, then the probability of congestion is only 15.56%, i.e. there is a high probability of no congestion (Figure 25).

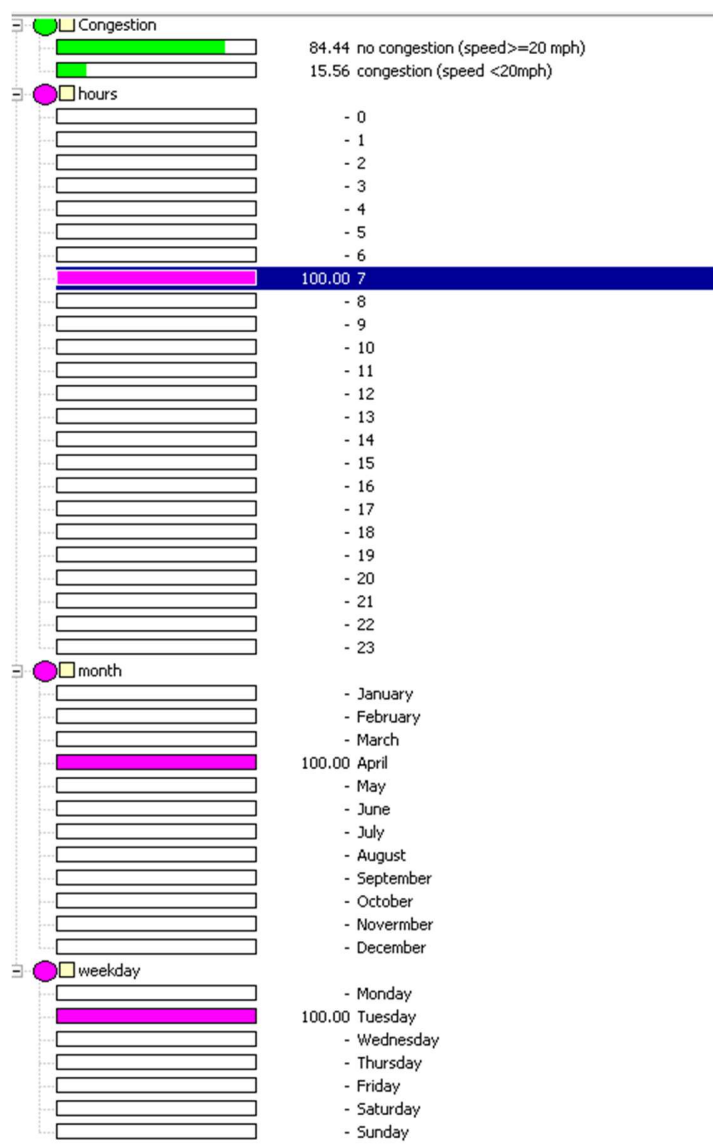


Figure 25 an example of no congestion prediction

As above, an early forecast of congestion can be made

Drawbacks of the Model

The current model also has a number of drawbacks and problems.

1. The amount of data we have is still not sufficient, which is determined by our database, for which we have only taken two years of data, 2021 and 2022. The database has the potential to

be expanded, but since the 2019 and 2020 epidemics have a more significant impact on the data model, it is not convenient to use it, and the solution is to wait for the later 2023 data to be updated before carrying out the analysis.

2. Our correction of the relationship between month and day of the week is in fact a simplification of the model. If there is sufficient data in the latter, we can refine the model by refining the effect of month on day of the week according to the year to obtain a more accurate model. The solution would still be to introduce a larger database.

3. For the data itself, the statistical period is so long that there will be some equipment damage or equipment replacement during the statistical process, but also because the data period is so long that we can only complete two statistical cycles, so it will result in too little data at the moment of the occurrence of an unexpected situation. The solution to this is to increase the number of measurement cycles, thus offsetting the impact of unexpected situations on the data.

4. Again due to the data problem, there is a problem of too small a sample, which leads to predictions such as "100% certainty of no congestion", which is not in line with the laws of statistics. The solution is still to increase the amount of data and expand the database so that from a statistical point of view there should not be a 100% probability.

Conclusion

In this student work, the importance of diagnosing and predicting traffic congestion is explained from the background analysis; then based on the reality and related literature, several variables affecting traffic congestion are collated and analyzed, and "time of day", "day of weeks" and "month" are selected to construct a Bayesian network model. The Bayesian network model was constructed by selecting "time", "day of the week" and "month".

After thorough data screening and analysis of the official traffic congestion data obtained from Chicago on the internet, all the corresponding parameter data in the Bayesian network were calculated and the final network was constructed.

The final Bayesian network model not only has the function of diagnosing the causes of traffic congestion, but also has the function of predicting traffic congestion according to the current state. Although there were some drawbacks due to the lack of data, the initial objectives were met and the task set for this student work was well accomplished.

References

- Heckerman, D. A Tutorial on Learning with Bayesian Networks. Publication MSR-TR-95-06. Microsoft Research, 1995.
- 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.
- 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.
- 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
- Chicago Traffic Tracker - Historical Congestion Estimates by Segment - 2018-Current.
<https://data.cityofchicago.org/Transportation/Chicago-Traffic-Tracker-Historical-Congestion-Esti/sxs8-h27x>
- Matt Burdett. Traffic congestion. 2020. <https://www.geographycasestudy.com/traffic-congestion/>
- Automated Speed Enforcement Frequently Asked Questions.
https://www.chicago.gov/city/en/depts/cdot/supp_info/children_s_safetyzoneporgramautomaticspeedenforcement/automated_speed_enforcementfrequentlyaskedquestions.html
- About Chicago Traffic Tracker(beta). <https://webapps1.chicago.gov/traffic/about.jsp>
- Google COVID-19 Community Mobility Trends. <https://OurWorldData.org/coronavirus>

Google map. <https://www.google.com/maps/>

Congestion Explorer. Congestion Management Program · SFCTA Prospector.

<https://congestion.sfcta.org/#close>