Early Warning Evaluation of Tuberculosis and Meteorological Factors in Shanxi Province Based on Dynamic Bayesian Network

Meichen Li
Shanxi Medical University School of Public Health
Zhuang Zhang
Shanxi Medical University School of Public Health
Hao Ren
Shanxi Medical University School of Public Health
Yueling Fan
Shanxi Province Disease Prevention and Control Center: Shanxi Center for Disease Control and Prevention
Weimei Song
Shanxi Medical University School of Public Health
Liyu Fan
Shanxi Province Disease Prevention and Control Center: Shanxi Center for Disease Control and Prevention
Dichen Quan
Shanxi Medical University School of Public Health
Jianwei Gao
Shanxi Province Disease Prevention and Control Center: Shanxi Center for Disease Control and Prevention
Limin Chen
Shanxi Provincial Peoples Hospital
Lixia Qiu (qlx_1126@163.com)
School of Public Health, Shanxi Medical University

Research Article

Keywords: Tuberculosis, Multiple indicator panel data, Moving percentile method, Dynamic Bayesian Network

Posted Date: March 1st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-259121/v1

Abstract

Background: Tuberculosis is a major global public health problem. However, it is still in the exploratory stage that the study of the meteorological factors related to the incidence of tuberculosis in Shanxi Province. Therefore, it is very urgent to establish an early warning system that easily operate of tuberculosis.

Method: The epidemiological characteristics of tuberculosis in Shanxi Province were described, and the Dynamic Bayesian Network early warning model was established by time series cross-correlation analysis and Bayesian Network.

Results: 1. The reported incidence of tuberculosis in Shanxi Province showed an overall downward trend from 2008 to 2017, showing a phenomenon of high in the middle and low at both ends each year, with certain seasonal characteristics. 2. Based on the results of cross-correlation analysis, it is reasonable to use dynamic Bayesian model fitting with meteorological factors lagging for 2 months; the monthly average temperature and monthly precipitation are positively correlated with the incidence of tuberculosis, but the monthly mean air pressure is negatively. 3. Comparison of classification and recognition performance of the three models shows that DBN has the highest classification accuracy in the two regions, which indicates that DBN is better than the other two models in reflecting the performance of minority classes, and better for the comprehensive classification of minority classes and majority classes.

Conclusion: 1. Shanxi Province has tuberculosis clustering in time, space and time and space. Incidence peak is in spring and early summer. March is the highest month in the year. Seven meteorological factors such as monthly precipitation are the main factors affecting the incidence of tuberculosis in Shanxi Province. 2. The classification and recognition performance of the Dynamic Bayesian Network early warning model of tuberculosis-meteorological factors established in this study is significantly better than that of static Bayesian Network and support vector machine model, and can better predict the future.

Background

Tuberculosis (TB), one of the deadliest infectious diseases in the world, is a major global public health problem[1–3]. Although China has made great achievements in tuberculosis control[4], it continues to be the third-largest TB burden country and not optimistic in the world[5], with obvious regional differences and a serious situation in the central, western and poor regions[6, 7]. But in recent years, with the increase of population and spread of multidrug-resistant tuberculosis, tuberculosis has been paid more and more attention all over the world[3, 8]. The aspiring strategy of WHO to end TB aims to reduce TB incidence and mortality in 2035 by 90%, and 95%, respectively compared to the 2015 cases[3, 9].

Situated in the center of China, Shanxi Province is an economically underdeveloped area, which TB incidence is not balanced in each city. At present, it is still in the exploratory stage that the study of the

meteorological factors related to the incidence of tuberculosis in Shanxi Province. Therefore, it is very urgent to establish an early warning system that easily operate of tuberculosis.

This study is aimed to explore the influence of meteorological factors on tuberculosis in Shanxi Province, to understand the incidence and epidemic characteristics of tuberculosis in Shanxi Province in the past ten years. So as to lay a foundation for the establishment of a dynamic Bayesian early warning model of tuberculosis-meteorological factors and to provide the decision-making basis for the prevention and control strategy of tuberculosis. Some studies have pointed out that the incidence of tuberculosis has certain seasonal distribution characteristics[10, 11], and this study found that tuberculosis in Shanxi Province also has similar characteristics, which speculates that the occurrence of tuberculosis may be related to specific climatic conditions[12].

The Bayesian Network (BN) proposed by Judea Pearl in the 1980s, but it is difficult for the traditional BN to reflect the influence of the time factor on the result of the whole event[13], while the Dynamic Bayesian Network (DBN) is an extension of the static BN in the time dimension[14–16], it continues the advantages of static Bayes and makes the reasoning process more continuous and cumulative. The traditional epidemiological analysis mainly studies the distribution characteristics of the disease in time, space and population[6, 17], while time series analysis can determine whether the occurrence of the disease has periodicity and peak period, and can predict the incidence[18]. DBN has great potential for data mining applications[19], but it has few in medicine, which are mainly in gene regulatory Networks. In 2010, Lingling Ge studied the construction method of gene regulatory Network based on dynamic Bayesian model. And In 2014, Yi Zhou improved it.

If the trend of tuberculosis incidence can be predicted by scientific methods, the purpose of effective early warning can be achieved and unnecessary losses can be reduced. Therefore, the DBN early warning model was established based on the tuberculosis incidence and meteorological data from January 2008 to December 2017 in Shanxi Province to make a reasonable prediction of tuberculosis incidence.

The data used in this study is a group of repeated measurements data of monitoring objects at different time points. The time series analysis commonly used in this data, which mainly constructed by time series analysis method at present. The DBN model is one of the effective models, it is accepted by more and more people because of its ability in generality and data mining[19].

Materials And Methods

Data:

1. 198,968 cases of tuberculosis from January 2008 to December 2017 were collected in 11 cities of Shanxi Province;
2. Demographic data of each city in Shanxi Province, from Shanxi Statistical Yearbook;
3. This study selected monthly data of 18 meteorological stations in Shanxi Province.

Method:

In this study, the clustering method of panel data was used to analyze the aggregation characteristics of pulmonary tuberculosis in spatial distribution[7]. The risk grade of tuberculosis incidence was divided by moving percentile method, so as to explore the regularity of time aggregation[20, 21]. The meteorological data are processed by IDW interpolation to represent the real situation of cities. The level of tuberculosis incidence is related to many factors, and there is a certain lag in time[6, 12]. This paper uses time series cross-correlation analysis to determine the lag period of the influence of meteorological factors on the incidence of tuberculosis in Shanxi Province from 2008 to 2017. The main meteorological influencing factors of tuberculosis were judged by principal component analysis. After Bayesian structure and parameter learning, the DBN early warning model is established and compared with BN and Support Vector Machine (SVM) [22].

Construction of DBN

It is the extension of BNs that the DBN has expanded the BN to the time dimension, so that forming a model that can deal with time series data[16]. DBN is used to describe the dynamic process of random variables, and X^t represents the state of node variables at time T. We define Dynamic Bayesian network as DBN$(B_0, B_{\rightarrow})$[22]. The Bayes at the initial time obtained by taking X^0 as the node is represented by B_0, and $B_{\rightarrow}$ is the Bayes fragment at the transfer network, and the node includes $x^t \cup x^{t+1}$, x^t represents the current state, with no parent node[23]; x^{t+1} represents the state at the next time with conditional probability $P(x^{t+1}/parent(x^{t+1}))$. The transition probability distribution of transfer network $B_{\rightarrow}$ is defined as follows:

$$P(x^{t+1}/x^t) = \prod_{i=1}^{n} P\left(x_i^{t+1}/\mathrm{parent}(x_i^{t+1})\right)$$

For any t, the X^t joint probability distribution of DBN is similar to that of BN, as follows:

$$P(x^0, \ldots, x^t) = P(x^0) \prod_{i=1}^{n} P(x^t/x^{t-1})$$

That's mean a DBN model defines the probability distribution of an infinite trajectory in a dynamic stochastic process. (Fig.1 The example of DBN.) In the process of building DBN, firstly, the related variables are extracted and the state values are determined, then the topological structure graph is established according to the dependence among variables, and finally, the conditional probability table is established according to the dependence among variables[22, 24]. According to the combination of expert knowledge and data learning, this study takes into account both efficiency and practical application, and constructs the DBN model.

The purpose of DBN reasoning is to calculate the influence of observed variable nodes on the probability distribution of other variables, the complexity is that both the evidence and the posterior probability distribution are indexed by time. The reasoning in DBN is that given an observation sequence, we can extend the network to the whole time series by copying the Bayesian segments of each time point, and then we can apply the reasoning algorithm in BN.

The measurement criteria of warning model classification performance

To evaluate and compare the DBN model established in this study, that is, to evaluate its classification effect. The evaluation indexes of classification effect included accuracy (ACC), sensitivity (TPR), specificity (TNR), precision, as well as comprehensive evaluation indexes F-measure, G-measure. The area AUC under ROC curve was compared and the Z test was used. The difference was statistically significant at $P<0.05$.

Software

Data collation and statistical description were performed using Excel and IBM SPSS Version 24.0 IDW and plotting were implemented using R3.6.1. Static and Dynamic Bayesian Network model are constructed by weka3.8 and GeNie2.4, and SVM classification model is built in R3.6.1.

Results

Clustering of Panel datah

Panel data is three-dimensional dynamic data, which needs to be reduced before clustering. When dividing the risk areas of the disease, all indicators should be taken into account to cluster the results as follows Fig. 2 YangQuan was left out because of data problems. (Fig. 2 The results of Panel data clustering in each city of Shanxi Province)

Finally, the 10 cities according to the result of clustering can be divided into two types of risk areas, as follows:

The first region: TaiYuan, Changzhi, Jincheng, Jinzhong, Lvliang;

The second region: Datong, Shuozhou, Yuncheng, Xinzhou, Linfen

Incidence and number

According to the distribution of tuberculosis incidence in Shanxi Province from 2008 to 2017, the overall incidence of tuberculosis showed a downward trend, showing a phenomenon of high in the middle and low at both ends every year (Fig. 3 Tuberculosis incidence in Shanxi Province from 2008 to 2017). The number of reported cases was the lowest from January to February, and there were multiple cases of tuberculosis from March to June, which indicates a significant seasonal increase (Fig. 4 The total numbers of tuberculosis in Shanxi Province from 2008 to 2017).

Risk grade

The moving percentile method was used to classify the risk grade of tuberculosis incidence, so as to explore the regularity of time aggregation. Finally, the results of TB risk classification in two regions of Shanxi Province from January 2011 to December 2017 are shown in Fig. 5–6. (Fig. 5. Results of TB risk grade classification in the first region. Figure 6. Results of TB risk grade classification in the second region.)

Lag periods

Whether the lag effect of meteorological factors on the incidence of disease can be correctly inferred is the key to the success of the study. This study can be seen that the climatic conditions vary greatly from region to region. On the whole, some meteorological factor indicators and monthly incidence changes regularly in different regions; the results of time series cross-correlation analysis between meteorological indicators and monthly incidence are shown in additional file 1. According to the results, related literature and the characteristics of tuberculosis, it is more reasonable to use the meteorological factors with a lag of 2 months to fit the dynamic Bayesian model.

Principal component regression analysis

To lay the establishment foundation of the dynamic Bayesian model, the principal components included in the model were selected by regression to determine the main influencing factors. It can be seen from Table 1 that the eigenvalues of the first four principal components are 2.970, 1.972, 0.952 and 0.542 respectively. Until the fourth principal component, the contribution rate reaches 91.932%. According to the principle of principal component extraction, the first four principal components are extracted.

Table 1
Principal component extracted information

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.970	42.422	42.422	2.970	42.422	42.422
2	1.972	28.174	70.596	1.972	28.174	70.596
3	0.952	13.600	84.196	0.952	13.600	84.196
4	0.542	7.736	91.932	0.542	7.736	91.932
5	0.277	3.953	95.885			
6	0.176	2.510	98.395			
7	0.112	1.605	100.000			

The four principal component scores were further analyzed by multiple linear regression using stepwise (α_{in} = 0.05 and α_{out} = 0.15) with the data of monthly incidence of tuberculosis.

The model fitting results (Table 2) show that the estimation coefficients of the first, third and fourth principal components are statistically significant (P < 0.05), while the second principal component could not be included in the model. And the related meteorological factors of each principal component show in Table 3. Therefore, all the seven preselected meteorological factors were retained for the next step of DBN modeling.

Table 2
Model fitting table of meteorological factors related to tuberculosis incidence

Model	Unstandardized Coefficients		Standardized Coefficients	*t*	*p*
	b	*SE*	Beta		
constant	4.593	0.052	-	89.123	0.000
z3	0.237	0.053	0.127	4.489	0.000
z1	-0.117	0.030	-0.111	-3.907	0.000
z4	0.226	0.070	0.092	3.227	0.001

Table 3
Related meteorological factors of each principal component

component	the closely related meteorological factors
first component	Monthly precipitation, mean temperature, average relative humidity, daily precipitation ≥ 0.1mm days.
second component	Mean pressure, hours of sunshine, average wind velocity.
third component	Mean pressure, mean temperature, average wind velocity, hours of sunshine.
forth component	Mean pressure, hours of sunshine, average wind velocity.

Discretization of meteorological factors

Because the DBN model can only deal with classified data, and the influence factors of some continuous variables can only be discretized. In this study, the equidistance method is used to discretize the meteorological factors into 5 grades, which are carried out in GeNie.

Structural learning

Principal Component Regression analysis showed that all seven meteorological factors were important risk factors of tuberculosis, so all of them were used to establish the DBN model.

Firstly, the model network structure is constructed by the structure learning algorithm based on Tabu Search algorithm. After that, the initial network is adjusted according to the actual situation and expert learning opinions, until the whole network structure meets the actual requirements and can have an accurate prediction of the results. The process is implemented by Weka and GeNie.

As can be seen from Fig. 7 (Fig. 7. The structure learning of Bayes network in the first region), the risk level of TB outbreaks in the first region is the child node of city, mean air pressure, mean temperature, monthly precipitation, mean relative humidity, mean wind velocity and sunshine hours, indicating that the risk level of TB outbreak is directly related to these influencing factors. And the daily precipitation ≥ 0.1 mm was the parent node of sunshine hours and average relative humidity, and was also the child node of monthly precipitation, indicating that the daily precipitation ≥ 0.1mm is indirectly related to the risk level of TB outbreak through sunshine hours, average relative humidity and monthly precipitation.

It is worth noting that the risk level node has an arrow pointing to itself, which means that the risk level of the previous month has an impact on the node probability of the next time. The mean temperature is the parent node of the mean air pressure, the average relative humidity is the parent node of the average wind speed and sunshine hours, and the sunshine hours are the parent node of the average temperature, which shows that these factors are not independent of each other. They also indirectly affect the risk level of TB outbreak through this complex relationship.

According to Fig. 8 (Fig. 8. The structure learning of Bayes network in the second region), the risk level of TB outbreak in the second region is the same as that in the first region. The difference is that the number of days with daily precipitation ≥ 0.1mm is the parent node of mean air pressure and mean relative humidity, and is also the child node of monthly precipitation, indicating that the number of days with daily precipitation ≥ 0.1mm is indirectly associated with the risk level of tuberculosis outbreak through mean air pressure, mean relative humidity and monthly precipitation. The remaining nodes also have complex relationships and are not listed.

Parameter learning

The parameters learning that used expectation maximization (EM) algorithm are studied according to the Bayesian structure learning network.

After the outbreak risk classification of the incidence data, the data for the period from January 2011 to December 2017 were remained. A total of 300 sets of data from January 2011 to December 2015 were used as the training set for five cities in each of the two regions for the parameters learning. Finally, 120 sets of data from January 2016 to December 2017 were used as the validation set to evaluate the effect of classification recognition. Through parameter learning, it is found that the distribution of

meteorological factors in the two regions is different in different grades, and their prior probabilities are shown in Table 4.

Table 4
The prior probability of each node.

region	meteorological factor	prior probability				
		1	2	3	4	5
the first region	Monthly precipitation	0.610	0.193	0.105	0.050	0.043
	Mean pressure	0.250	0.348	0.140	0.169	0.093
	Mean temperature	0.107	0.207	0.150	0.238	0.298
	Average relative humidity	0.055	0.279	0.269	0.286	0.112
	Hours of sunshine	0.076	0.236	0.417	0.226	0.045
	Average wind velocity	0.090	0.533	0.300	0.060	0.017
	Daily precipitation ≥ 0.1mm days.	0.250	0.329	0.250	0.136	0.036
the second region	Monthly precipitation	0.548	0.205	0.110	0.079	0.060
	Mean pressure	0.264	0.207	0.129	0.126	0.274
	Mean temperature	0.088	0.205	0.198	0.279	0.231
	Average relative humidity	0.131	0.226	0.310	0.264	0.069
	Hours of sunshine	0.019	0.138	0.288	0.376	0.179
	Average wind velocity	0.048	0.355	0.412	0.164	0.021
	Daily precipitation ≥ 0.1mm days.	0.310	0.314	0.174	0.152	0.050

The risk levels of tuberculosis outbreaks in both regions have 6 parent nodes in this study. Taking TaiYuan as an example, assuming that the average air pressure, average temperature, monthly precipitation, average relative humidity and average wind speed at t time are all grade 1, the transfer probability of risk level nodes is shown in Table 5.

Table 5
Transition probability of TaiYuan risk level node

level of risk(t)	**level of risk (t + 1)**	
	Warning	**No warning**
Warning	0.9431818261828593	0.8823357944000647
No warning	0.05681817381714072	0.1176642055999353

Verification and comparison of the warning effect of the model

Among the 420 sets of data collected in this study, 34 groups were classified as "outbreak" early warning, the ratio of "early warning" to "non-early warning" was about 11:1 in the first region. And 26 groups were classified as "outbreak" early warning, the ratio of "early warning" to "non-early warning" was about 15:1 in the second region. In the first and second regions, the proportion of risk level of training and validation set with "outbreak" warning was 1.6:1, and 1.4:1, respectively.

It can be seen from Table 6 that the classification accuracy of the three models in the two regions is all over 75% and 80%, respectively. And DBN is the highest at 95.00% and 97.5%, respectively. Other indexes such as precision, TPR, TNR, F-Measure and G-measure are also the highest in DBN. Among them, the F-measure value of DBN is the largest, the two regions are 0.77 and 0.84 respectively, indicating that DBN reflects the performance of minority classes better than the other two models; G-measure values are the largest, the two regions are 0.86 and 0.85 respectively, indicating that its comprehensive classification performance for minority and majority categories is better.

Table 6
Comparison of classification accuracy of the three models in identifying "outbreak" early warning

Region	index	DBN	BN	SVM
the first region	ACC(%)	95.00	86.67	79.17
	precision	0.77	0.36	0.20
	TPR	0.77	0.31	0.31
	TNR	0.97	0.93	0.58
	F-Measure	0.77	0.33	0.56
	G-measure	0.86	0.54	0.42
the second region	ACC(%)	97.50	95.00	83.33
	precision	1.00	1.00	0.36
	TPR	0.73	0.45	0.45
	TNR	1.00	1.00	0.66
	F-Measure	0.84	0.62	0.62
	G-measure	0.85	0.67	0.54

ROC analysis

DBN, BN and SVM models respectively fit the two-regional data, the AUC of the area under the ROC curve was calculated and the statistical difference was tested, as shown in Table 7 and Fig. 9. (Fig. 9. ROC curves of three models in the first region and second region)

Except for the SVM model fitted in the first region, the ROC curve area AUC of the other models is more than 0.6. The AUC of the two regions is DBN > BN > SVM from the largest to the smallest. Pairwise comparison was carried out on the AUC of the three models in the two regions, the results show that in the first region, the AUC of BN and SVM is statistically different from DBN, but there is no statistical difference between BN and SVM. In the second region, the AUC of the three models were statistically different from each other.

Table 7
AUC values and multiple comparisons of the three models

Region	Model	AUC	SE	95%CI
The first region	DBN	0.871^{a}	0.070	(0.733,1.000)
	BN	0.621^{b}	0.092	(0.440,0.802)
	SVM	0.593^{b}	0.91	(0.416,0.771)
The second region	DBN	0.864^{a}	0.082	(0.704,1.000)
	BN	0.727^{b}	0.100	(0.531,0.924)
	SVM	0.668^{c}	0.097	(0.478.0.858)
*Multiple comparisons between models with different letters are different at the level of α = 0.05.				

In general, for the panel data with spatial and temporal dimensions, the classification and recognition performance of the DBN warning model of TB-meteorological factors established in this study is significantly better than that of the static BN and SVM models, whether it is to distinguish the minority categories or the majority categories.

Discussion

The SARS, influenza and the COVID-19 all showed the necessity and urgency of establishing the infectious disease early warning system[25, 26]. In 2008, China officially launched an automatic early warning system for infectious diseases based on the whole country[27]. At present, although there are many methods to establish an effective early warning system for infectious diseases, most of the traditional methods are based on historical incidence data to predict the fluctuation of future

incidence[28], that is, to predict the incidence at a certain time in the future. It only studies the temporal and spatial characteristics of infectious diseases at the incidence level, and the relationship between them and other factors, such as socio-economic factors and meteorological factors, is not explored in depth[6]. If there is a period of missing historical data or large errors, it is difficult to predict. Therefore, the DBN is more suitable for the conditional probability method of uncertain inference for the lack of deterministic prediction.

At present, there have been many studies on the epidemiological characteristics of tuberculosis and the influence of meteorological factors[29, 30]. Meteorological factors are one of the important factors affecting the occurrence and spread of infectious diseases[31]. Meteorological factors can not only affect immunity and health, but also affect pathogenic microorganisms in the environment, thus affecting the prevalence of infectious diseases[32–34]. Shanxi Province, located in central China, is a typical mountainous plateau with the level of economic development at the end of the country. In addition to the serious environmental pollution, dense population, and bad weather such as smog, which will accelerate the spread of tuberculosis. Shanxi Province had a total of 198968 registered cases of tuberculosis from 2008 to 2017, and the average reported incidence rate was 55.67 / 100000, which is the middle level of the country. At present, the study on the meteorological factors related to the incidence of tuberculosis in Shanxi Province is still at the exploratory stage. Therefore, this study focuses on this and describes the epidemiological characteristics of tuberculosis in Shanxi Province, to explore the relationship between the incidence of tuberculosis and meteorological factor, and to establish an early warning model of tuberculosis based on DBN.

This study found that the reported incidence of tuberculosis in Shanxi Province showed a downward trend during the decade, with an average incidence of 55.67/100,000, but the number was still large, therefore prevention and control measures should not be relaxed. It shows a phenomenon of high in the middle and low at both ends every year, with certain seasonal characteristics. March is the peak of the whole year, and spring and early summer have an obvious seasonal increase[2, 6, 17], which is roughly consistent with previous studies. The results of this study show that meteorological factors have a lagging effect on the incidence of tuberculosis. Monthly average temperature, monthly precipitation and sunshine hours are positively correlated with the incidence of tuberculosis, while monthly mean air pressure is negatively correlated. Through principal component analysis, the results show that seven meteorological factors such as monthly precipitation are the main factors affecting the incidence of tuberculosis in Shanxi Province.

BN is one of the most effective theoretical models in the uncertain reasoning[19, 35]. The DBN early warning model of tuberculosis-meteorological factors established in this study has a significantly better classification and recognition performance than other models for panel data with spatial and temporal dimensions. It can accumulate the law and experience of variables changing with time to better predict the future moment.

Innovations: (1) the DBN model has many applications in continuous speech recognition, finance and fault diagnosis, but it is not widely used in medicine. In particular, its use in the prediction of TB incidence has not been reported. (2) No research has been reported on the prediction of DBN model based on meteorological factor data and tuberculosis incidence data in Shanxi Province. This study can provide a new idea for the prevention and control of tuberculosis in Shanxi Province, it lays a foundation for the intelligent identification of the risk of tuberculosis.

Shortcomings: (1) tuberculosis is a chronic infectious disease affected by various factors, and it is difficult to include other factors since monthly data are used in this study;(2) we discretized the original continuous data before modeling, which would lose part of the information and maybe affect the probability estimation in the results. In future studies, we should further improve the BN algorithms and methods, or find more advantageous methods for better analysis of continuous data. And collect more TB related risk factors, to construct and improve the early warning model of tuberculosis and provide more comprehensive recommendations for prevention and control of tuberculosis in Shanxi Province.

Conclusion

1. In this study, the monitoring data of tuberculosis in Shanxi Province from 2008 to 2017 were analyzed from the ecological perspective. The results show that the incidence of 10 years to overall a downward trend, showing a phenomenon of high in the middle and low at both ends every year, which had certain seasonal characteristics. The cluster analysis result of multi-index panel data divides 10 cities into two regions in order to model and analyze the different regions.

2. The time series cross-correlation results showed that meteorological factors had a lag effect on the incidence of tuberculosis, and monthly mean temperature, monthly precipitation and sunshine hours had a positive correlation with the incidence of pulmonary tuberculosis. The results of principal component analysis show that seven meteorological factors, such as monthly precipitation, are the main factors affecting the incidence of tuberculosis in Shanxi Province.

3. The DBN early warning model of tuberculosis-meteorological factors established in this study is compared with BN and SVM model. It can be seen that for the panel data with spatial and temporal characteristics DBN is superior to other models in classification and recognition, which can better predict the future and provide new methods for the decision-making of tuberculosis prevention and control in Shanxi Province.

Abbreviations

TB: tuberculosis; BN: Bayesian Network; DBN: Dynamic Bayesian Network; SVM: Support Vector Machine.

Declarations

Ethics approval and consent to participate

Because of the study only used the surveillance data and did not contain any personal or health information. Therefore, ethics approval was not required.

Consent for publication

Not applicable.

Availability of data

Please contact corresponding authors for data requests.

Competing interests

Authors declare that they have no competing interests.

Funding

This work was supported by the National Natural Science Foundation of China (No. 81973155); Key Research and Development (R&D) Projects of Shanxi Province (201803D31066).

Authors' contributions

ML, ZZ and HR conceived and developed the idea and research. ML and ZZ performed the statistical analysis and built the warning model. ML, DQ and WS wrote the first draft of the manuscript and all other authors discussed results and edited the manuscript. YF JG and LF collected and preprocessing tuberculosis data. HR collected and compiled demographic and meteorological data. All authors read and approved the final manuscript.

Acknowledgements

The authors gratefully acknowledge the staff involved in TB surveillance at all participating levels in ShanXi province, China.

References

1. Hu H, Chen J, Sato K, Zhou Y, Jiang H, Wu P, et al. Factors that associated with TB patient admission rate and TB inpatient service cost: a cross-sectional study in China. Infect Dis of Poverty. 2016;5:4.
2. Wang T, Xue F, Chen Y, Ma Y, Liu Y. The spatial epidemiology of tuberculosis in Linyi City, China, 2005-2010. BMC Public Health. 2012;12:885.
3. Sileshi T, Tadesse E, Makonnen E, Aklillu E. The Impact of First-Line Anti-Tubercular Drugs' Pharmacokinetics on Treatment Outcome: A Systematic Review. Clin Pharmacol-Adv A. 2021;13:1-12.
4. Xia T, Chen J, Rui J, Li J, Guo Y. What affected Chinese parents' decisions about tuberculosis (TB) treatment: Implications based on a cross-sectional survey. PloS one. 2021;16(1):e0245691.

5. World Health Organization (2020). Global Tuberculosis Report [Available from: https://apps.who.int/iris/bitstream/handle/10665/336069/9789240013131-eng.pdf.
6. Rao H, Zhang X, Zhao L, Yu J, Ren W, Zhang X, et al. Spatial transmission and meteorological determinants of tuberculosis incidence in Qinghai Province, China: a spatial clustering panel analysis. Infect Dis of Poverty. 2016;5(1):45.
7. Wang X, Yin S, Li Y, Wang W, Du M, Guo W, et al. Spatiotemporal epidemiology of, and factors associated with, the tuberculosis prevalence in northern China, 2010-2014. BMC Infect DIS. 2019;19(1):365.
8. Long Q, Qu Y, Lucas H. Drug-resistant tuberculosis control in China: progress and challenges. Infect Dis of Poverty. 2016;5:9.
9. WHO. End TB Strategy; 2015. [Available from: https://www.who.int/tb/strategy/en/.
10. Li X, Wang L, Zhang H, Du X, Jiang S, Shen T, et al. Seasonal variations in notification of active tuberculosis cases in China, 2005-2012. PloS one. 2013;8(7):e68102.
11. Zuo Z, Wang M, Cui H, Wang Y, Wu J, Qi J, et al. Spatiotemporal characteristics and the epidemiology of tuberculosis in China from 2004 to 2017 by the nationwide surveillance system. BMC Public Health. 2020;20(1):1284.
12. Wang W, Guo W, Cai J, Guo W, Liu R, Liu X, et al. Epidemiological characteristics of tuberculosis and effects of meteorological factors and air pollutants on tuberculosis in Shijiazhuang, China: A distribution lag non-linear analysis. Environ Res. 2020:110310.
13. Pan J, Rao H, Zhang X, Li W, Wei Z, Zhang Z, et al. Application of a Tabu search-based Bayesian network in identifying factors related to hypertension. Medicine. 2019;98(25):e16058.
14. Koller D, Friedman N. Probabilistic Graphical Models: Principles and Techniques: MIT Press; 2009.
15. Li Z, Xu T, Gu J, Dong Q, Fu L. Reliability modelling and analysis of a multi-state element based on a dynamic Bayesian network. R Soc Open Sci. 2018;5(4):171438.
16. Brandherm B, Jameson A. An Extension of the Differential Approach for Bayesian Network Inference to Dynamic Bayesian Networks. Int. J. Intell. Syst. 2004;19(8):727–48.
17. Liu Y, Li X, Wang W, Li Z, Hou M, He Y, et al. Investigation of space-time clusters and geospatial hot spots for the occurrence of tuberculosis in Beijing. Int J Tuberc Lung D. 2012;16(4):486-91.
18. Kumar V, Singh A, Adhikary M, Daral S, Khokhar A, Singh S. Seasonality of tuberculosis in delhi, India: a time series analysis. Tuberc Res Treat. 2014;2014:514093.
19. Zhang Z, Zhang J, Wei Z, Ren H, Song W, Pan J, et al. Application of tabu search-based Bayesian networks in exploring related factors of liver cirrhosis complicated with hepatic encephalopathy and disease identification. Sci Rep. 2019;9(1):6251.
20. Wang R, Jiang Y, Michael E, Zhao G. How to select a proper early warning threshold to detect infectious disease outbreaks based on the China infectious disease automated alert and response system (CIDARS). BMC Public Health. 2017;17(1):570.

21. Wang R, Jiang Y, Guo X, Wu Y, Zhao G. Influence of infectious disease seasonality on the performance of the outbreak detection algorithm in the China Infectious Disease Automated-alert and Response System. J Int Med Res. 2018;46(1):98-106.
22. Petousis P, Han S, Aberle D, Bui A. Prediction of lung cancer incidence on the low-dose computed tomography arm of the National Lung Screening Trial: A dynamic Bayesian network. Artif Intell Med. 2016;72:42-55.
23. Trifonova N, Karnauskas M, Kelble C. Predicting ecosystem components in the Gulf of Mexico and their responses to climate variability with a dynamic Bayesian network model. PloS one. 2019;14(1):e0209257.
24. Yuan Z, Zhuo K, Zhang Q, Zhao C, Sang S. Probabilistic assessment of visual fatigue caused by stereoscopy using dynamic Bayesian networks. Acta Ophthalmol. 2019;97(3):e435-e41.
25. Yan W, Nie S, Xu B, Dong H, Palm L, Diwan V. Establishing a web-based integrated surveillance system for early detection of infectious disease epidemic in rural China: a field experimental study. BMC Med Inform Decis Mak. 2012;12:4.
26. Alakus T, Turkoglu I. Comparison of deep learning approaches to predict COVID-19 infection. Chaos Solitons Fractals. 2020;140:110120.
27. An Q, Wu J, Yao W. Comparison of the effects of the automatic early-warning information system of infectious diseases and spatial-temporal clustering analysis in predicting rubella outbreaks in Jinzhou district, Dalian city. Disease Surveillance. 2010;25(07):577-9.
28. Wang Y, Xu C, Zhang S, Wang Z, Yang L, Zhu Y, et al. Temporal trends analysis of tuberculosis morbidity in mainland China from 1997 to 2025 using a new SARIMA-NARNNX hybrid model. BMJ Open. 2019;9(7):e024409.
29. Narula P, Sihota P, Azad S, Lio P. Analyzing seasonality of tuberculosis across Indian states and union territories. J Epidemiol Glob Health. 2015;5(4):337-46.
30. Li Z, Pan H, Liu Q, Song H, Wang J. Comparing the performance of time series models with or without meteorological factors in predicting incident pulmonary tuberculosis in eastern China. Infect Dis of Poverty. 2020;9(1):151.
31. Xiao Y, He L, Chen Y, Wang Q, Meng Q, Chang W, et al. The influence of meteorological factors on tuberculosis incidence in Southwest China from 2006 to 2015. Sci Rep. 2018;8(1):10053.
32. Wang W. Progress in the impact of polluted meteorological conditions on the incidence of asthma. J Thorac Dis. 2016;8(1):E57-61.
33. Shi H, Critto A, Torresan S, Gao Q. The Temporal and Spatial Distribution Characteristics of Air Pollution Index and Meteorological Elements in Beijing, Tianjin, and Shijiazhuang, China. Integr Environ Assess Manag. 2018;14(6):710-21.
34. Zhang C, Zhang A. Climate and air pollution alter incidence of tuberculosis in Beijing, China. Ann Epidemiol. 2019;37:71-6.
35. Agrahari R, Foroushani A, Docking T, Chang L, Duns G, Hudoba M, et al. Applications of Bayesian network models in predicting types of hematological malignancies. Sci Rep. 2018;8(1):6951.

Figures

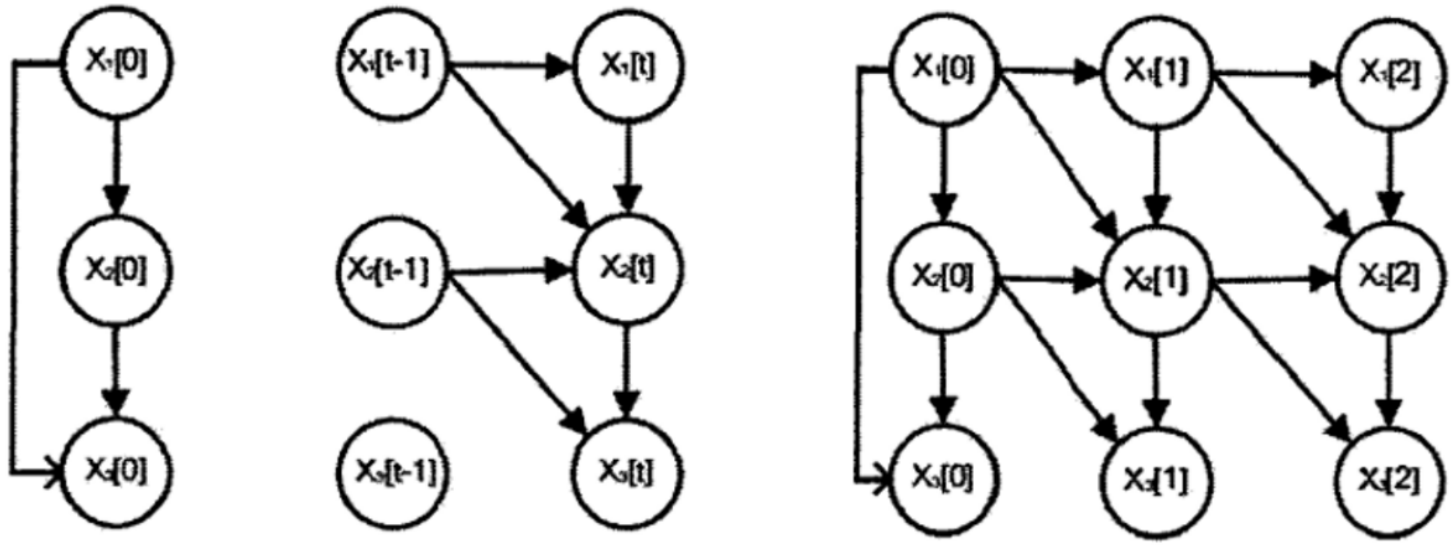

Figure 1

The example of DBN.

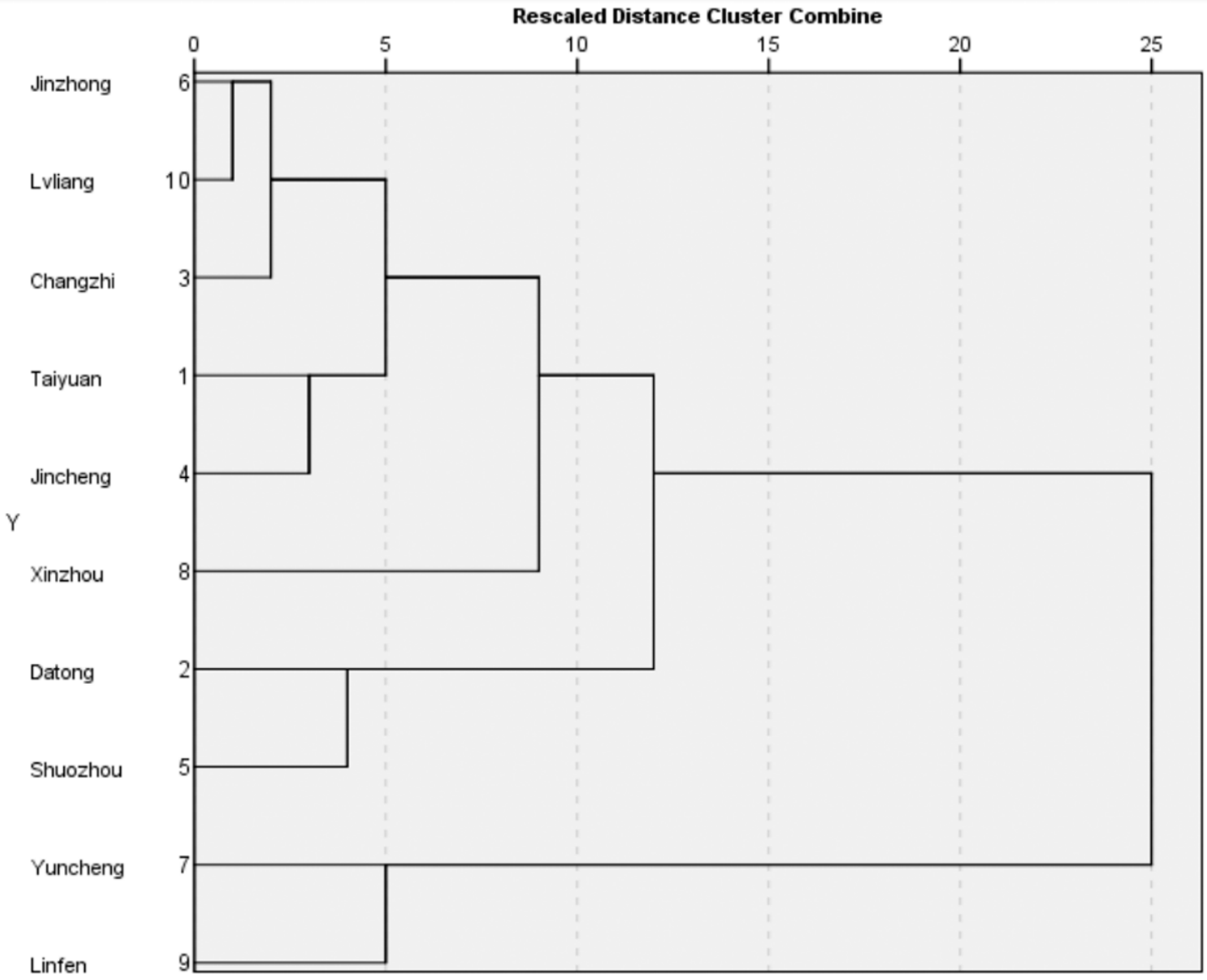

Figure 2

The results of Panel data clustering in each city of Shanxi Province

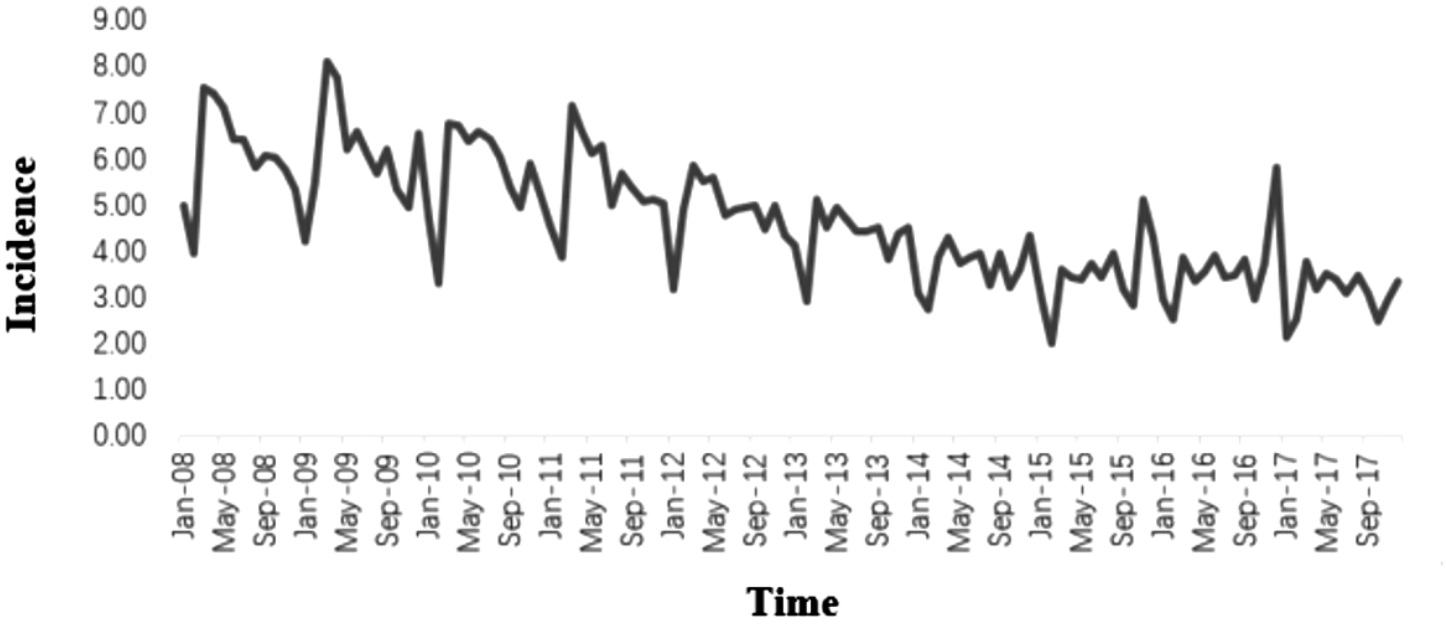

Figure 3

Tuberculosis incidence in Shanxi Province from 2008 to 2017

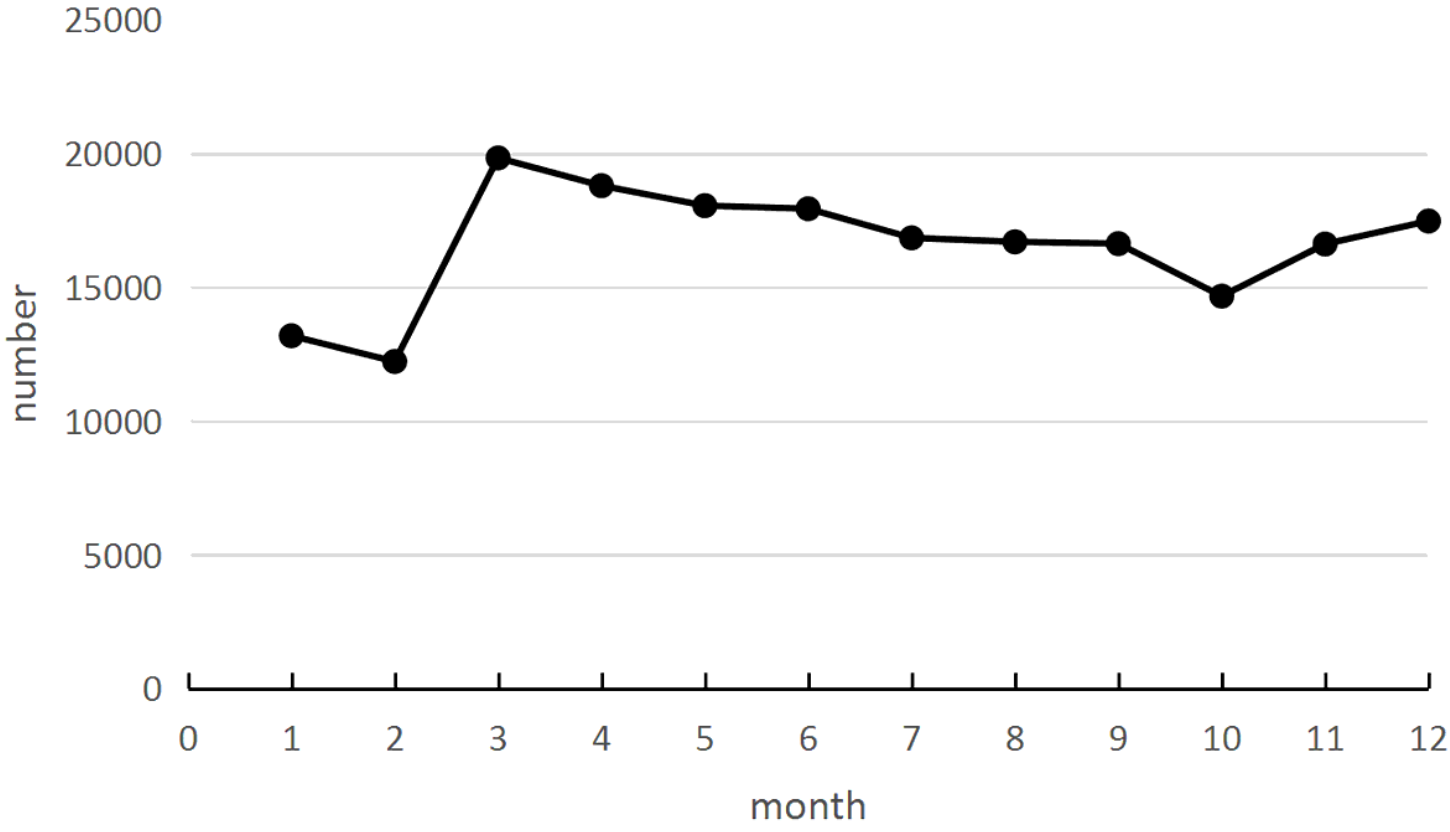

Figure 4

The total numbers of tuberculosis in Shanxi Province from 2008 to 2017

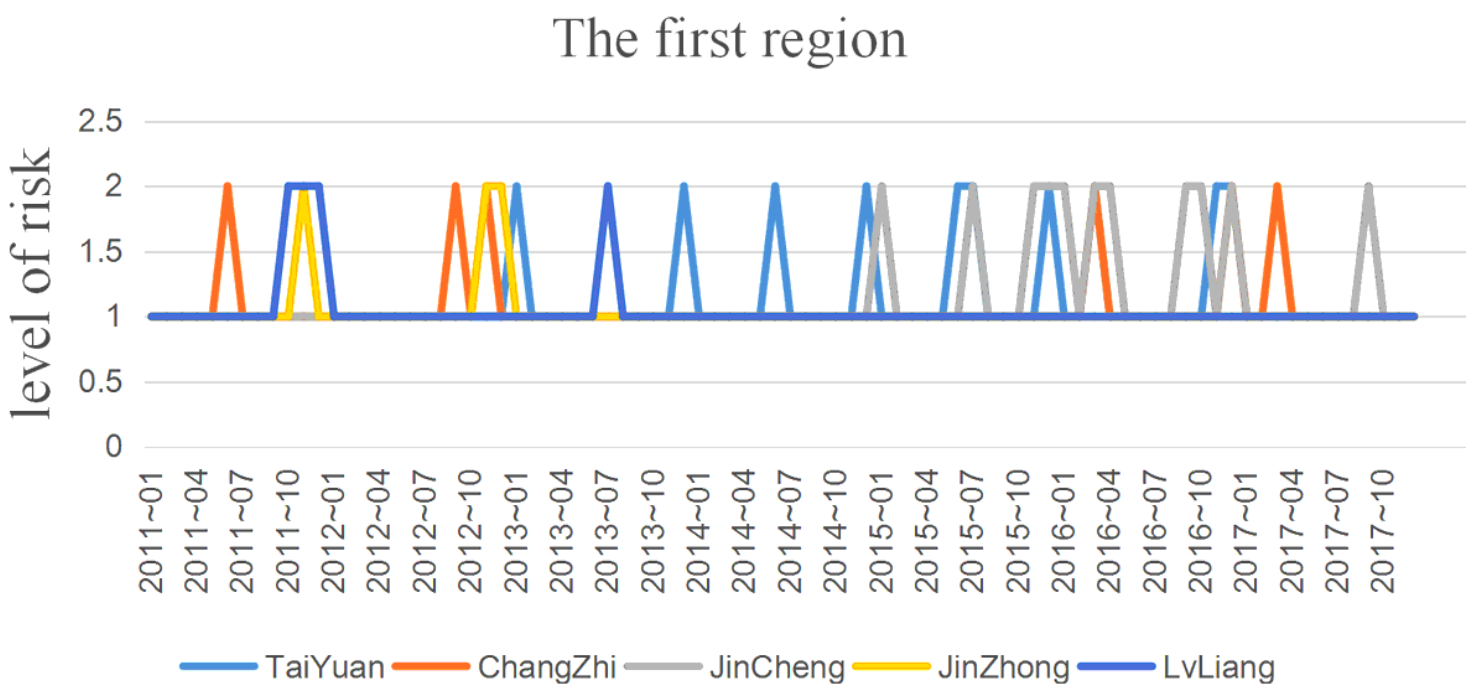

Figure 5

Results of TB risk grade classification in the first region.

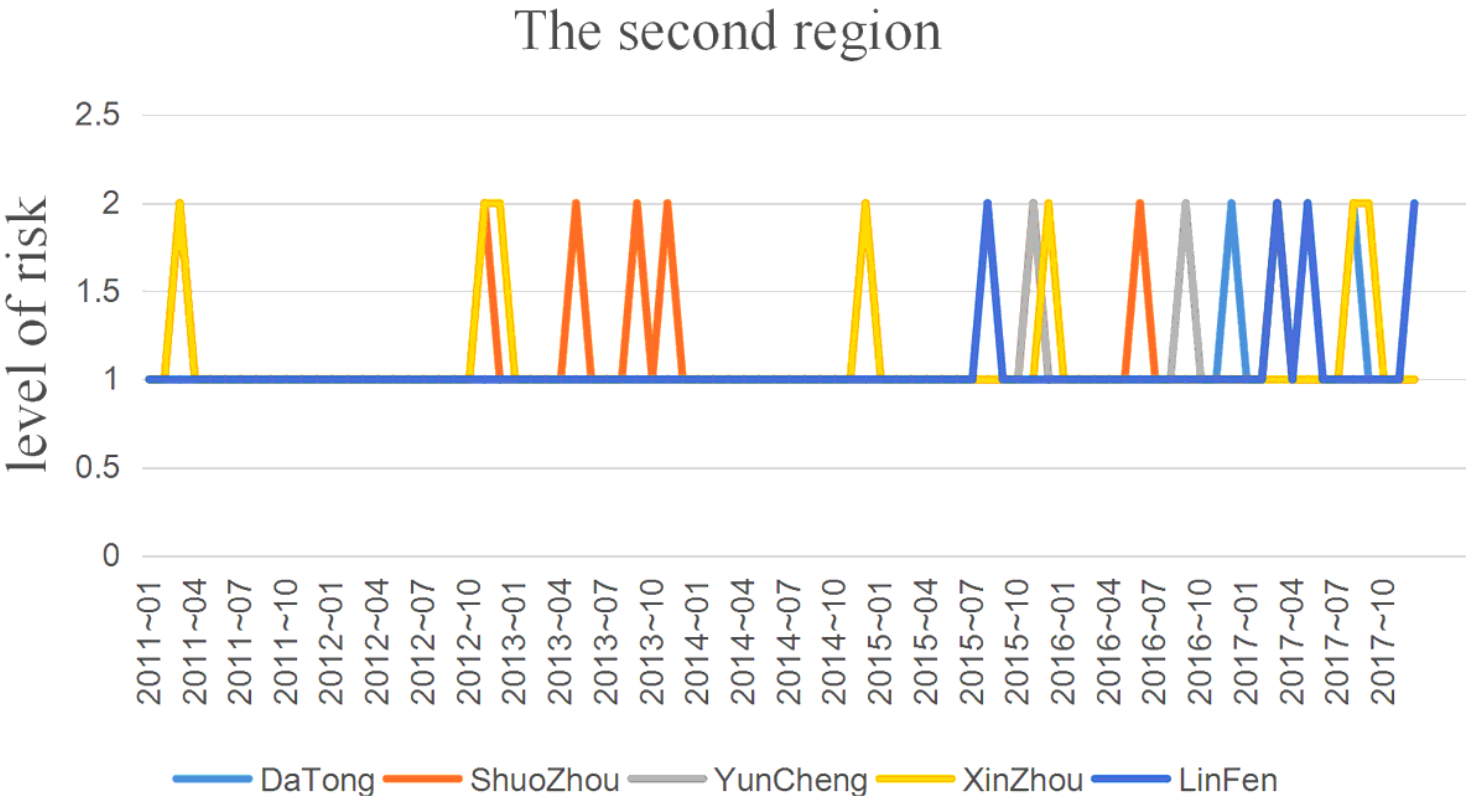

Figure 6

Results of TB risk grade classification in the second region.

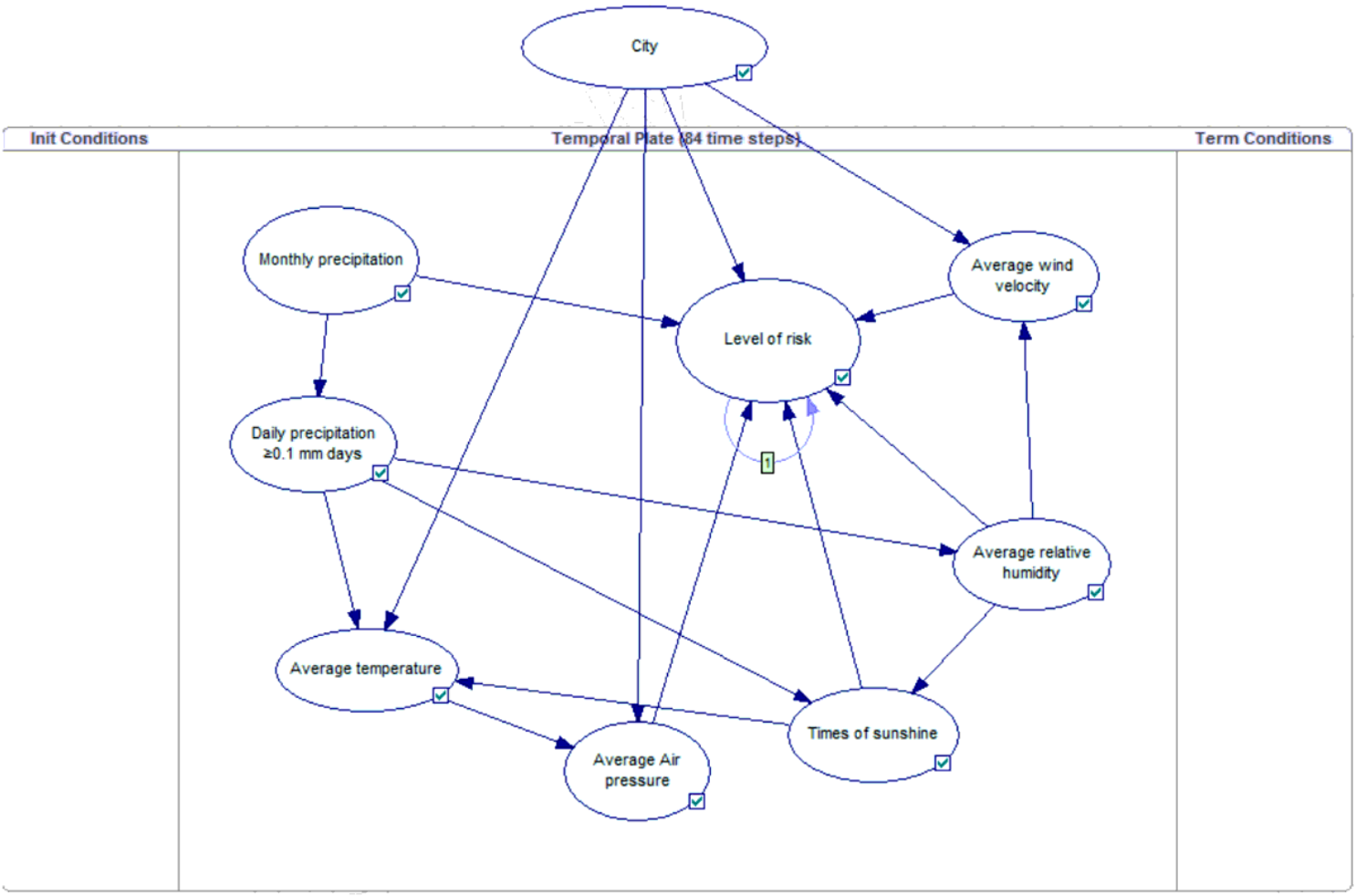

Figure 7

The structure learning of Bayes network in the first region

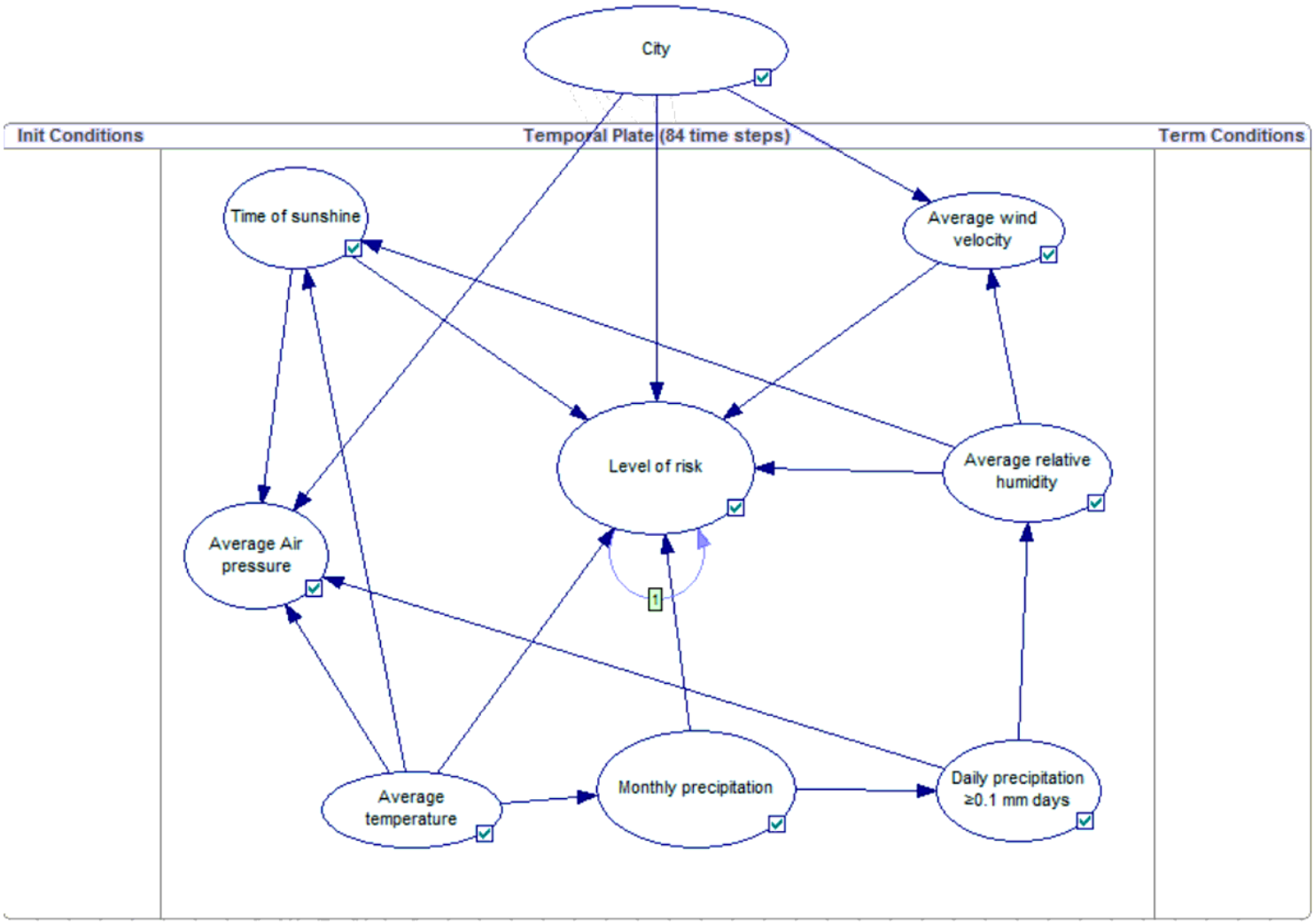

Figure 8

The structure learning of Bayes network in the second region

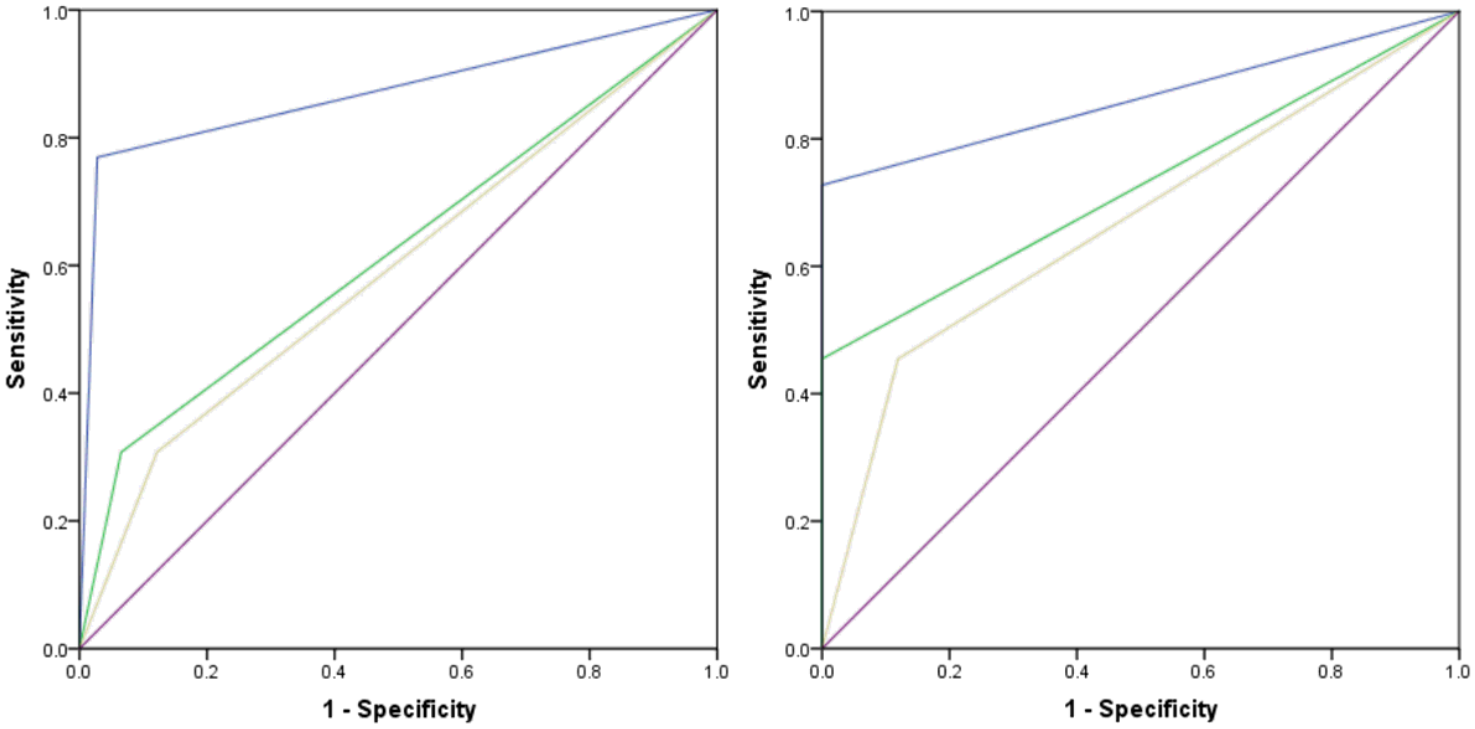

Figure 9

ROC curves of three models in the first region and second region

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.docx