Stoch Environ Res Risk Assess
DOI 10.1007/s00477-009-0303-5

ORIGINAL PAPER

Risk assessment of human neural tube defects using a Bayesian belief network

Yilan Liao · Jinfeng Wang · Yaoqin Guo · Xiaoying Zheng

Abstract Neural tube defects (NTDs) constitute the most common type of birth defects. How much risk of NTDs could an area take? The answer to this question will help people understand the geographical distribution of NTDs and explore its environmental causes. Most existing methods usually take the spatial correlation of cases into account and rarely consider the effect of environmental factors. However, especially in rural areas, the NTDs cases have a little effect on each other across space, whereas the role of environmental factors is significant. To demonstrate these points, Heshun, a county with the highest rate of NTDs in China, was selected as the region of interest in the study. Bayesian belief network was used to quantify the probability of NTDs occurred at villages with no births. The study indicated that the proposed method was easy to apply and high accuracy was achieved at a 95% confidence level.

Keywords Neural tube birth defects · Bayesian belief network · Data discretization · Case–effect relationship

Y. Liao · J. Wang (✉) · Y. Guo
Institute of Geographical Sciences and Nature Resources Research, Chinese Academy of Sciences, 100101 Beijing, People's Republic of China
e-mail: wangjf@lreis.ac.cn

Y. Liao
e-mail: liaoyl@lreis.ac.cn

Y. Guo
e-mail: guoyq@lreis.ac.cn

X. Zheng
Institute of Population Research, Peking University, 100871 Beijing, People's Republic of China

1 Introduction

Birth defects, as defined by the March of Dimes Birth Defects Foundation, refer to any anomaly (functional or structural) presented in infancy or later in life and induced by events preceding birth (whether inherited, or acquired). Varying from minor cosmetic irregularities to life-threatening disorders, birth defects are the major cause of infant mortality and a leading cause of disability (Carmona 2005). NTDs (neural tube birth defects) constitute one of the most common forms of birth defect, often occurring between the third and fourth weeks of gestational age. They result in structural defects that occur anywhere along the neuroaxis from the developing brain to the sacrum and often result in the exposure of neural tissue (Frey and Hauser 2003).

Birth defects have a substantial public health impact on mortality, morbidity, disability, and to the cost of healthcare provision. Fortunately, they can often be prevented and early intervention is an important component in the minimization of their consequences. However, for the vast majority of birth defects, the etiology is still unknown. Advances in geographical information systems (GIS) and risk assessment methods now provide opportunities for people to quickly analyze spatial relationships and disease risk factors to facilitate policy planning and implementation (Wiwanitkit 2008; Canales and Leckie 2007). It is used to visualize spatial patterns in the geographical distribution of disease, usually for explorative and descriptive purposes, as well as to provide information for further studies. It can also be used to gain important clues about the etiology of a disease (Sankoh et al. 2002).

Numerous risk assessment methods of birth defects from simple to complicated ones have been proposed. Initially an epidemiological measure is often used as a measure of relative risk, and in particular the standardized morbidity ratio.

Published online: 12 February 2009

However, the calculation of disease rate can not be calculated when the birth data are missed or there is no birth in many of the geographical areas. Then some approaches such as empirical Bayesian estimation (Wu et al. 2004; Ismaila et al. 2007; Hemmi 2008) and spatial filtering method (Rushton and Lolonis 1996; Chi et al. 2007) were proposed to overcome this problem and applied the spatial correlation of cases to assess the disease risk in each area. But in many areas birth defects cases affect each other little across space (this is true especially in rural areas because they are low probability events). These statistical or geo-statistical techniques which do not involve mechanisms that account for the relevant physical and/or laws and are not based on deductively valid principles often led to nonsensical inferences and uninformative maps (Christakos 2002). In addition, some researchers (Ritz et al. 2002; Carmichael et al. 2003) used knowledge of a number of environmental factors that are related to birth defects development to accurately assess the risk of birth defects. This estimation, however, may be complicated by the fact that there is often local risk variations that cannot easily be accounted for by the known covariates. There is, therefore, a need to develop suitable techniques for assessing the risk of birth defects in different human groups.

Bayesian belief network (BBN) is a form of artificial intelligence that incorporates uncertainty through probability theory and conditional dependence (McCabe et al. 1998). It is a graphical model that presents probabilistic relationships among a set of variables by determining the dependence relationships among them (Heckerman 1997). The variables of interest and the relationship between those variables make up a belief network. Since its development in the 1970s, BBN has provided a powerful framework for modeling complex problems involving uncertain knowledge and impacts of causes. Accordingly, in this study we used BBN to assess birth defects risk. We first constructed a belief network with risk factors and disease incidence, and then used training data to calculate the joint probability distribution among these variables. Then, the risk of birth defects in unknown areas could be derived from this probability distribution. Different birth defects may be caused by different risk factors, so we limited our research to NTDs. The experimental results indicated that this method is simple to apply, has better fault-tolerance and greatly enhances calculation accuracy.

2 Materials and methods

2.1 Description of study site

China has the largest population in the world, and also has the highest incidence of birth defects, with the highest levels compared with the national average being observed in the Shanxi Province. Shanxi also has the highest rate of NTDs in the world (Hu 2003). So we selected Heshun, a county in Shanxi province as the study area in research. Heshun is also one of the pilot regions for the national birth defect intervention project launched the State Family Planning Commission.

Heshun lies in the Tai Hang Mountain Region, at 37o03′E and 113o05′N (see Fig. 1). It consists of 326 administrative villages, and has an area of 2,250 km^2. Most of the people in this county are farmers and their living environments seldom change. Furthermore, there have been no large-scale movements of people in the history of this region. The inherited and congenital causes of birth defects are similar among the people in this region, and these factors explain only a small fraction of all of NTD cases seen. Furthermore, most types of birth defects designated by the WHO can be found in Heshun, and NTDs predominate (Wu et al. 2004). During 1998–2005 period, there were 7,880 births in Heshun but 187 caught NTDs.

2.2 Data sources

For this analysis, we included all live and still births occurring in Heshun from January 1, 1998 to December 31, 2005, born to women at the hospital or at home, and who were residents of the county during that time period. We also included all therapeutic abortions in residents in that area where the estimated date of delivery fell in the time period of interest. All NTD cases, regardless of pregnancy outcome, were verified by doctors in the hospitals. Records of NTD cases were collected from the local family planning department. The NTDs in the study included anencephaly, spina bifida, encephalocele, holoprosencephaly, and hydrocephalus, among others.

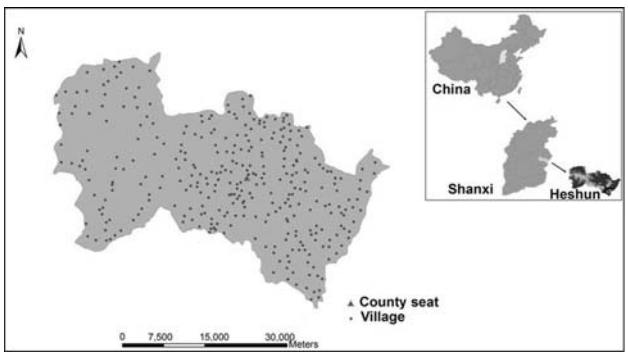


Fig. 1 Location of Heshun

The local planning department declined to provide identifiers to link substantiated NTD cases to births, so we were unable to conduct the study at an individual level; instead, we conducted an ecological study. That is, we used the relationship among NTDs and the environmental characteristics of the villages in which the patients' mothers lived to assess the disease risk. There were 326 villages and one town in the study area. Since the main object of this study was to estimate the disease risk based on the relationships between environmental risk factors and NTDs, the town was not included as the environmental factors there are somewhat complex. In addition, birth defect registers in the town were not included.

In the study, the environmental factors of various villages were classified into socioeconomic and geographical factors (Table 1). The socioeconomic factors reported useful information on medical conditions (the number of doctors), the per-capita incomes (per-capita net incomes), the agricultural chemical exposures (the use of fertilizers and pesticides), and the crop yields (vegetable and fruit productions) of every village during the 1998–2005 period. Taking the spatial interaction into account, we also collected the socioeconomic factors of neighboring villages within a specified distance. All socioeconomic data were provided by the Heshun statistical bureau. The geographical factors included elevation, vegetations coverage (normalized difference vegetation index), access condition (distance to main roads), pollution risk (influence of coal mines and distance to factories) and the geological background (distance to faults) of the villages. The normalized difference vegetation index (NDVI) is an index that provides a standardized method of comparing vegetation greenness between satellite images and can be used as an indicator of relative biomass and greenness. The source of the NDVI dataset used in the study is the VITO (Flemish Inst. Technological Research, Belgium), http://www.vgt.vito.be. Furthermore, Li et al. (2006) found that the occurrence ratio of NTDs in Heshun had a significant negative correlation with increasing-distance from faults. So we chose the distance between villages and faults as an important influence factor. All of the data were discretized before being input BBN.

Geographic information system (GIS) is a computer package used to store, manage, analyze, and map geographical data. It plays a significant role in data processing and has insuperable advantages over traditional methods. GIS allows the addition of relevant layers which can be used for analyzing the spatial relationships among selected factors. Also, it offers database capabilities that can handle attributes data effectively. Attribute calculations are simple and relatively accurate. We used ARCGIS 9.0i as the GIS platform to locate the 326 villages and to quantify the selected factors.

2.3 Bayesian belief network

The idea of BBN is derived from Bayes's theorem (after Reverend Thomas Bayes, 1702–1762). In contrast to a regression model which can only represent the dependency of one outcome variable on one or more predictor variables, it can represent the mutual and hierarchal relationships among many variables using probabilistic rules and thus, in many instances, is more appropriate for prognostic and diagnostic applications (Sebastiani et al. 2008). BBN has been applied widely in the domain of epidemiology, such as disease diagnosis (Aronsky and Haug 2000; Burnside et al. 2000), risk analysis (Maglogiannis et al. 2006; Maskery et al. 2008), classification of cytological findings (Hamilton et al. 1995), nursing research (Lee and Abbott 2003). In this study, we used Bayesian network to model the relationships between environmental variables and the risk of NTDs. The course of assessing NTDs risk based on BBN comprises three steps: structural learning, parameter learning, and network validation. The BBN software tool used in this study was BN PowerSoft, which was developed by Jie Cheng.

2.3.1 Structural learning

The structural learning of BBN, so-called qualitative analysis, is the graphical representation of independence holding among variables and has the form of an acyclic directed graph (Lee et al. 2008). The purpose of this phase is to identify significant environmental factors being applicable to assess the NTDs risk. There are two methods for structural learning using data. One is a Bayesian approach based on scoring and searching, the other is a constraint-based approach based on independence test. A Bayesian approach finds the optimal model structure from data after a BBN is constructed by the user's priori knowledge, and a constraint-based approach finds the optimal model structure from conditional dependences in each pair of variables. However, a constraint-based approach is commonly used due to its computational simplicity compared to the Bayesian approach (Lee et al. 2008). So in the study we applied constraint-based approach to construct the network. A set of factors was initially identified from the published literature and expert advices. A bivariate correlation analysis was then applied to explore the relationship among different factors and NTDs. Finally some factors correlated with NTDs were remained: the number of doctors, the use of pesticides and fertilizer, the production of vegetable and fruit, per-capita

Table 1 Input data of BBN

Variables	Values	Numbers	Marginal percentage	Variables	Values	Numbers	Marginal percentage
NTDs risk (unit: 1/1,000)	0	217	68.89	Vegetable (unit: ton)	1–20.0	31	9.84
	0.1–50	50	15.87		20.1–100	85	26.98
	50.1–100	26	8.25		100.1–200	69	21.90
	More than 100	22	6.98		200.1–500	61	19.37
Doctor (unit: person)	1–2	38	12.06		500.1–1,000	35	11.11
	3–4	94	29.84		More than 1,000	34	10.79
	5–6	70	22.22	Net-income (unit: yuan)	1,000–5,000	88	27.94
	7–8	63	20.00		5,001–8,000	91	28.89
	9–10	27	8.57		8,001–10,000	50	15.87
	More than 10	23	7.30		10,000–20,000	68	21.59
Fruit (unit: ton)	0–0.5	89	28.25		More than 20,000	18	5.71
	0.6–1.5	77	24.44	Factory (unit: m)	0–2,000	73	23.17
	1.6–5	81	25.71		2,001–4,000	93	29.52
	5.1–10	22	6.98		4,001–6,000	47	14.92
	More than 10	46	14.60		6,001–10,000	58	18.41
Pesticide (unit: ton)	0–0.5	107	33.97		More than 10,000	44	13.97
	0.6–1	67	21.27	Coal mines (unit: ton/m)	0	203	64.44
	1.1–2	60	19.05		1–2,000	35	11.11
	2.1–5	66	20.95		2,001–4,000	31	9.84
	More than 5	15	4.76		More than 4,000	46	14.60
NDVI	150–500	79	25.08	Fault buffer (unit: m)	0–2,000	116	36.83
	501–700	66	20.95		2,001–4,000	86	27.30
	701–1,000	125	39.68		4,001–6,000	42	13.33
	More than 1,000	45	14.29		6,001–8,000	28	8.89
Road buffer (Unit: M)	0–2,000	189	60.00		8,001–10,000	19	6.03
	2,001–4,000	67	21.27		10,001–12,000	12	3.81
	4,001–6,000	38	12.06		12,001–14,000	10	3.17
	6,001–8,000	12	3.81		14,001–16,000	2	0.63
	8,001–10,000	8	2.54	Elevation (unit: m)	1,100–1,250	23	7.30
	10,001–12,000	1	0.32		1,251–1,300	82	26.03
Fertilizer (unit: ton)	0–100	65	20.63		1,301–1,400	106	33.65
	101–200	133	42.22		1,401–1,500	65	20.63
	201–300	80	25.40		More than 1,500	39	12.38
	301–400	37	11.75				

net incomes, elevation, NDVI, road and fault buffer, influence of coal mines and distances to the nearest factory.

Seen from Fig. 2, BBN presents graphically so that each variable is presented as a node with the directed links forming arcs between them. The information content of each variable is presented as one or several probability distributions. If a variable has no incoming arcs and is hence not dependent on any other variables in the model universe (i.e., has no parents), it has one probability distribution, and if it has parents, it has one probability distribution per each combination of possible values of the parents.

2.3.2 *Quantitative analysis*

The parameter learning part of BBN, the so-called quantitative analysis, finds dependence relations as joint conditional probability distributions among variables using cause and consequence relationships from the qualitative part and data of variables (Lee et al. 2008). The dependence relations can be assigned by expert

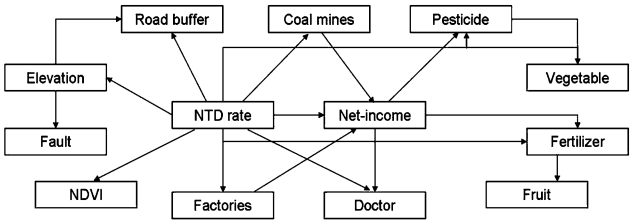


Fig. 2 BBN in the NTDs risk assessment

knowledge. Alternately, by inducing a learning algorithm, they can be learned from data. These methods can also be combined, which may strengthen the performance of a model.

In our study, the estimation of parameters was simple and was done just by calculating (counting and dividing) the prior or conditional probabilities. For each node of BBN, a conditional probability formula or table is supplied that represents the probabilities of each value of this node, given the conditions of its parents (i.e., all the nodes that have arcs pointed to this node) (Lin and Haug 2008). Meanwhile the distributions of the parents can be calculated given the values of their children. That is, one can proceed not only from causes to consequences, but also deduce the probabilities of different causes given the consequences (Uusitalo 2007). The algorithm which calculates the joint probability distribution of BBN can be expressed as:

$$P(X_1 = x_1, \ldots, X_n = x_n) = P(x_1, \ldots, x_n) = P(x_n) \prod_{i=1}^{n} P(x_i|\text{Parents}(X_i)) \quad (1)$$

$P(x_1,\ldots,x_n)$ means the probability of a specific combination of values $x_1,\ldots,x_n$ from the set of variables $X_1,\ldots,X_n$ and Parent (X_i) means the set of X_i's immediate parent nodes. Thus, $P(X_i|\text{Parent}(X_i))$ means the conditional probability, which is related to the node X_i considering its parent nodes.

2.3.3 Network validation

The parameter learning step is accomplished with a randomly assigned set of raw data designated as the "training" set. The next step of the network validation phase is to validate a trained network on new cases in a test set. The assigned test set is comprised of the remaining village data (those not used to actually estimate the parameters in the first place) in the overall dataset. These data are considered "unseen", and thus, performance measures should be generated from a test set results, which given some insights into the usefulness of the models (Lee and Abbott 2003). In this study, data of 237 villages (75% of the villages where there were babies born during the study period) were randomly selected as training set to build the BBN. The performance of BBN was defined as the percentage of correct predictions on the test sets (i.e., using a 0–1 loss function).

3 Results and discussion

3.1 BBN for estimating NTDs risk

In the study, the NTDs rate in each village where there were babies born was regarded as its NTDs risk. Figure 2 illustrates the network structure of BBN used for estimating the risk of NTDs in the study. For the etiology of NTDs is unknown, we selected variable "NTD rate" as parent node of various factors. We considered that there may be some common natural and socioeconomic characters in those villages with NTDs cases. Table 2 lists the posterior probabilities of factors in different grades given various levels of disease. Seen from Table 2, the affect to NTDs varies among the selected factors.

Heshun is a state-level poverty-stricken county and the local income mainly depends on coal production. Relatively well-off villages there often appear at the area near the coal mines ($r = 0.705$, $P = 0.000$) or high-polluting enterprises ($r = -0.617$, $P = 0.000$). Although medical condition in those villages is obviously better than in others ($r = 0.723$, $P = 0.000$), serious environmental pollution has affected the health of local residents. Table 2 shows that the nearer pregnant women live to high-yielding coal mines or high-polluting enterprises, the higher the NTD risk. So, following with the increase of doctor number and net-income, the risk of NTDs appears to become higher. The phenomena mean that in addition to launching the birth defects intervention program, the government should strengthen its local environmental governance.

Chemical exposures may have an impact on NTDs there. Similar to the conclusion of Heeren's study (Heern et al. 2003), the NTDs risk suddenly increases when the use of pesticide is more than five ton. Meanwhile, the corresponding posterior probabilities shows that the excessive use of fertilizers may increase the NTDs risk. Though there is no data suggesting that any particular fertilizer causes birth defects. The medical and public health department should strengthen the publicity of the knowledge about minimizing the exposure to chemical objects.

It is interesting that with the production of vegetable increase, the NTDs risk also on the rise. In Heshun, most of farmers plant vegetable, mainly for consumption. According to experts concerned, in poor areas of Heshun, local people often eat old sprouted potatoes. The previous study found that sprouting blighted potato tuber is lack of zinc

Table 2 The posterior probabilities of factors in different grades given various levels of disease

Factors	Values	NTD rate (unit: ‰)				Factors	Values	NTD rate (unit: ‰)			
		0	0.1–50	50.1–100	More than 100			0	0.1–50	50.1–100	More than 100
Doctor (unit: person)	1–2	0.842	0	0.079	0.079	Vegetable (unit: ton)	1–20.0	0.839	0	0.097	0.065
	3–4	0.787	0.064	0.053	0.096		20.1–100	0.800	0.082	0.071	0.047
	5–6	0.657	0.243	0.071	0.029		100.1–200	0.696	0.145	0.087	0.072
	7–8	0.683	0.159	0.095	0.063		200.1–500	0.639	0.148	0.131	0.082
	9–10	0.519	0.185	0.185	0.111		500.1–1,000	0.657	0.229	0.029	0.086
	More than 10	0.348	0.522	0.087	0.043		More than 1,000	0.382	0.471	0.059	0.088
Net-income (unit: yuan)	1,000–5,000	0.807	0.057	0.057	0.080	Pesticide (unit: ton)	0–0.5	0.804	0.056	0.075	0.065
	5,001–8,000	0.791	0.077	0.066	0.077		0.6–1	0.791	0.090	0.015	0.104
	8,001–10,000	0.68	0.200	0.080	0.14		1.1–2	0.533	0.283	0.133	0.050
	10,000–20,000	0.500	0.294	0.118	0.103		2.1–5	0.636	0.182	0.121	0.061
	More than 20,000	0.333	0.444	0.167	0.389		More than 5	0.267	0.600	0.067	0.067
Fertilizer (unit: ton)	0–100	0.831	0.015	0.046	0.108	NDVI	150–500	0.772	0.063	0.089	0.076
	101–200	0.752	0.098	0.098	0.053		501–700	0.758	0.106	0.061	0.076
	201–300	0.563	0.263	0.100	0.075		701–1,000	0.640	0.216	0.072	0.072
	301–400	0.486	0.405	0.054	0.054		More than 1,000	0.578	0.244	0.133	0.044
Elevation (unit: m)	1,100–1,250	0.565	0.348	0.087	0	Factories (unit: m)	0–2,000	0.452	0.37	0.137	0.041
	1,251–1,300	0.598	0.268	0.049	0.085		2,001–4,000	0.688	0.172	0.065	0.075
	1,301–1,400	0.670	0.142	0.123	0.066		4,001–6,000	0.809	0.064	0.085	0.043
	1,401–1,500	0.815	0.062	0.077	0.046		6,001–10,000	0.759	0.034	0.069	0.138
	More than 1,500	0.795	0.026	0.051	0.128		More than 10,000	0.864	0.045	0.045	0.045
Road buffer (unit: m)	0–2,000	0.624	0.222	0.106	0.048	Coal mines (unit: ton/m)	0	0.778	0.094	0.069	0.059
	2,001–4,000	0.731	0.104	0.045	0.119		1–2,000	0.714	0.114	0.086	0.086
	4,001–6,000	0.842	0	0.079	0.079		2,001–4,000	0.613	0.258	0.097	0.032
	6,001–8,000	0.750	0.083	0	0.167		More than 4,000	0.326	0.413	0.130	0.130
	8,001–10,000	1.000	0	0	0						
	10,001–12,000	1.000	0	0	0						

which is significant for development of fetus. Long-term consumption of zinc-deplete, blighted potato tuber by pregnant woman could be potentially teratogenic birth of a baby with NTDs (UIman et al. 2005). Wang et al. (2008) also used geographical detectors-based health risk assessment to improve that basic nutrition (food) was more important than man-made pollution in the control of the spatial NTD pattern of Heshun. Besides, the increasing NTDs risk following with the crease of distance to roads inflects poor health status of people who live in remote areas. The inequities, quality and accessibility in services available to poor people remain to be improved.

Heshun lies in the Tai Hang Mountain Region and the natural environment there is complex. So we took some natural factors into account in the study. Elevation does not seem to affect NTDs risk apparently. The analysis result indicates that posterior probabilities of elevation in 1,100–1,500 m given NTDs rate equal to 0 change from 0.565 to 0.815. However, the corresponding probability value decreases to 0.795 when elevation is more than 1,500 m. At the same time, the probability that NTDs rate are more than 100‰ suddenly increases to 0.128. NDVI, the selected other natural factor, was found to have negative effects on the NTDs rate. The more NDVI of residence, the easier fetuses catch NTDs. We also discovered that the areas near to coal mines usually are covered with dense vegetation ($r = 0.500$, $P = 0.000$). How does NDVI influence NTDs is a problem to be explored in the future.

3.2 NTDs risk mapping

Figure 3 illustrates the spatial distribution of estimated NTDs risk based on BBN. From the figure, the higher NTDs risk mainly distributes along roads. It is because that most of population in Heshun inhabit there. There are three clusters with high NTDs risk on the map: west, middle and south east regions. The middle region gathers a majority of coal mines and high-polluting factories of the county and is

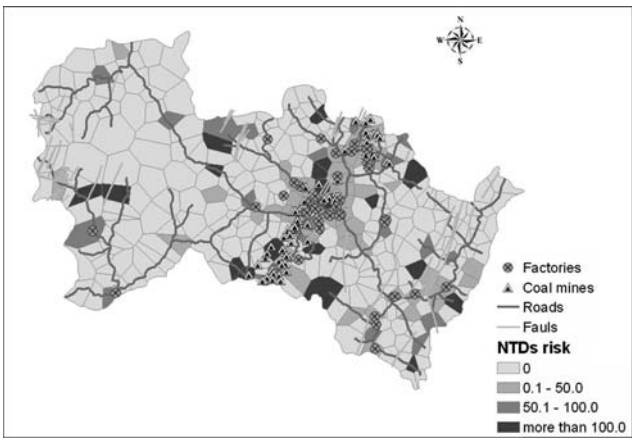


Fig. 3 NTDs risk map based on BBN

polluted seriously. So the government should pay more attention to reduce the effect of environment pollution there. Mountain areas of Heshun mainly locate in the east and west regions and geological background there is relatively complex. Another two clusters in the regions are obviously near to faults. In 2006, Li et al. verified that people residing near a geological default have risk to have a baby with NTDs. Which microelements release from faults may give an impact to NTDs? How they did?These problems are urgent to be solved in the process of reducing the NTDs risk there. The risk estimation result suggests that the government needs to adapt the intervention measures according to local conditions.

3.3 Accuracy analysis

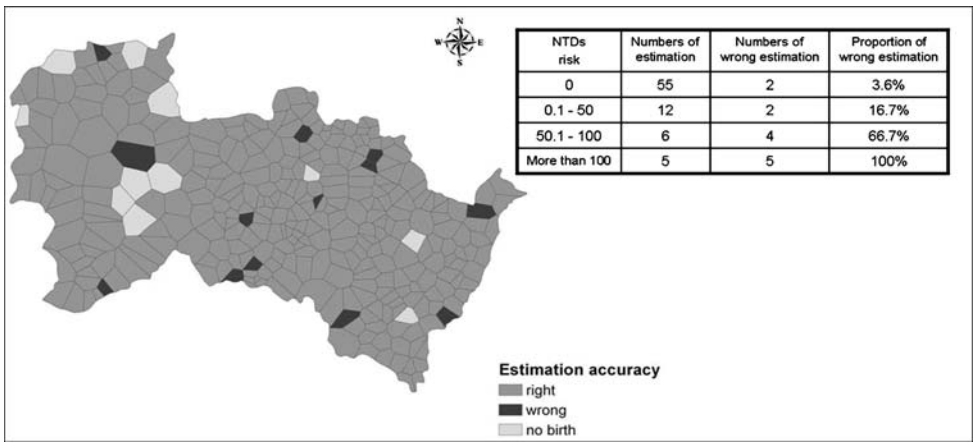
As mentioned before, data of 78 villages (25% of the villages where there were babies born during the study period) were randomly selected as test dataset to ensure the accuracy of the method. The NTDs risks in 65 villages were classified correctly. The estimate prediction accuracy is 83.33 ± 8.27% at 95% confidence level. The error for estimating zero NTDs risk achieved 0.0054 while there was no risk of more than 100‰ can be estimated correctly. It is because that the number of villages with NTDs risk of more than 100‰ is too small to find the interaction rule between factors and NTDs. The accuracy of the method can be improved by proper data discretization method. In addition, the estimation error is randomly distributed in space except the villages with no birth (Fig. 4).

4 Conclusions

A number of disease risk estimation methods have been applied to analysis the size, behavior and spatial distribution of NTDs, which is useful for understanding the etiology of the disease. However, which factors actually impact on the NTDs risk in a specific region? How do they affect it? None of these proposed methods performed very well to solve these problems. After reviewing issues related to the existing disease risk estimation methods, we creatively introduced a BBN combining various GIS derived variables to estimate NTDs risk. The results showed that the BBN can spatially model NTDs risk with a moderate level of accuracy.

There are many advantages of using BBN to estimate NTDs risk: (1) BBN provides great flexibility in their capacity for accepting input and providing output. (2) BBN may be developed using expert opinion instead of requiring historical data. (3) BBN gains insight into relationships among variables of the process due to its graphical display (Lee et al. 2008). (4) Properly designed BBN can provide a valid output when any subset of the modeled variables is present. Although the proposed method has been tested only in Heshun, a rural region in the north of China, it can be used as a tool to solve NTDs, even another diseases risk problems on any scales

Fig. 4 Estimation accuracy map and table



NTDs risk	Numbers of estimation	Numbers of wrong estimation	Proportion of wrong estimation
0	55	2	3.6%
0.1 - 50	12	2	16.7%
50.1 - 100	6	4	66.7%
More than 100	5	5	100%

as BBN is general and provides a single, unified method for addressing a variety of seemingly different problems in a variety of areas.

At the same time, there are some problems in our method for future studies. BBN is not good at deal with continuous data, and such data generally needs to be discretized, which may bring about information loss. Moreover, how to construct a proper network of variables and disease risk is crucial to improve the accuracy of the method. This problem is often difficult to be solved although there are many network construction algorithms. In addition, the study only focus on environmental factors, which may decrease the likelihood for identifying the contribution of other factors to the risk of NTDs. Incorporating information on genotypic variation into epidemiological studies of environmental exposures may potentially leads to improvements in risk estimation (Lammer 1998).

Acknowledgments This work was supported by the Project of the National Natural Science Foundation of China (70571076&40471111), the Hi-Tech Research and Development Program of China (2006AA12Z15), the National Basic Research Priorities Program (2001CB5103) of the Ministry of Science and Technology of the People's Republic of China,and Knowledge Innovation Program of the CAS (KZCX2-YW-3-8).

References

Aronsky D, Haug PJ (2000) Automatic identification of patients eligible for a pneumonia guideline. Proc/AMIA Annu Symp 12–16

Burnside E, Rubin D, Shachter R (2000) A Bayesian network for mammography. Proc/AMIA Annu Symp 106–110

Canales RA, Leckie JO (2007) Application of a stochastic model to estimate children's short-term residential exposure to lead. Stoch Environ Res Risk Assess 21:737–745

Carmichael SL, Nelson V, Shaw GM, Wasserman CR, Croen LA (2003) Socio-economic status and risk of conotruncal heart defects and orofacial clefts. Paediatr Perinat Epidemiol 17:264–271

Carmona RH (2005) The global challenges of birth defects and disabilities. Lancet 366:1142–1144

Chi WX, Wang JF, Li XH, Zheng XY, Liao YL (2007) Application of GIS-based spatial filtering method for neural tube defects disease mapping. J Wuhan Univ (Nat Sci Ed) 12(6):1125–1130

Christakos G (2002) On the assimilation of uncertain physical knowledge bases: Bayesian and non-Bayesian techniques. Adv Water Resour 25:1257–1274

Frey L, Hauser WA (2003) Epidemiology of neural tube defects. Epilepsia 44:4–13

Hamilton PW, Montironi R, Abmayr W, Bibbo M, Anderson N, Thompson D, Bartels PH (1995) Clinical applications of Bayesian belief networks in pathology. Pathologica 87(3):237–245

Heckerman D (1997) Bayesian networks for data mining. Data Min Knowl Disc 1(1):79–119

Heern GA, Tyler J, Mandeya A (2003) Agricultural chemical exposures and birth defects in the Eastern Cape Province, South Africa A case–control study. Environ Health 2(1):11–18

Hemmi I (2008) Bayesian estimation of the incidence rate in birth defects monitoring. Congenit Anom 28(2):103–109

Hu HT (2003) Methods to prevent Shanxi birth defect. China.org.cn. Available viable by http://www.china.org.cn/english/2003/sep/74927.htm

Ismaila AS, Canty A, Thabane L (2007) Comparison of Bayesian and frequentist approaches in modeling risk of preterm birth near the Sydney Tar Ponds, Nova Scotia, Canada. BMC Med Res Methodol 7:39–52

Lammer EJ (1998) Gene-environment analyses in human birth defects research. Neurotoxicol Teratol 25(3):351

Lee SM, Abbott PA (2003) Bayesian networks for knowledge discovery in large databasets: basics for nurse researchers. J Biomed Inform 36:389–399

Lee E, Park Y, Shin JG (2008) Large engineering project risk management using a Bayesian belief network. Expert Syst Appl doi:10.1016/j.eswa.2008.07.057

Li XH, Wang JF, Liao YL, Meng B, Zheng XY (2006) A geo-analysis for the environmental cause of human birth defects. Toxicol Environ Chem 88(3):551–559

Lin JH, Haug PJ (2008) Exploiting missing clinical data in Bayesian network modeling for predicting medical problems. J Biomed Inform 41:1–14

Maglogiannis I, Zafiropoulos E, Platis A, Lambrinoudakis C (2006) Risk analysis of a patient monitoring system using Bayesian network modeling. J Biomed Inform 39:637–647

Maskery SM, Hu H, Hooke J, Shriver CD, Liebman MN (2008) A Bayesian derived network of breast pathology co-occurrence. J Biomed Inform 41:242–250

McCabe B, AbouRizk SM, Goebel R (1998) Belief networks for construction performance diagnostics. J Comput Civil Eng ASCE 12(2):93–100

Ritz B, Yu F, Fruin S, Chapa G, Shaw GM, Harris JA (2002) Ambient air pollution and risk of birth defects in southern California. Am J Epidemiol 155(1):17–25

Rushton G, Lolonis P (1996) Exploratory spatial analysis of birth defect rates in an urban population. Stat Med 15:717–726

Sankoh OA, Berke O, Simboro S, Becher H (2002) Bayesian and GIS mapping of childhood mortality in rural Burkina Faso. Control of Tropical Infectious Diseases, Uni-Heidelberg Discussion Paper

Sebastiani P, Nolan VG, Baldwin CT, Abad-Grau MM, Wang L, Adewoye AH, McMahon LC, Farrer LA, Taylor JG, Kato GJ, Gladwin MT, Steinberg MH (2008) A network model to predict the risk of death in sickle cell disease. Blood 41(3):432–441

UIman C, Taneli F, Oksel F, Hakerlerler H (2005) Zinc-deficient sprouting blight potatoes and their possible relation with neural tube defects. Cell Biochem Funct 23:69–72

Uusitalo L (2007) Advantages and challenges of Bayesian networks in environmental modeling. Ecol Model 203:312–318

Wang JF, Li XH, Christakos G, Liao YL, Zhang T, Gu X, Zheng XY (2008) Geographical detectors-based health risk assessment and its application in the neural tube defects study of the Heshun Region, China. Int J Geogr Inf Sci (in press)

Wiwanitkit V (2008) Estimating cancer risk due to benzene exposure in some urban areas in Bangkok. Stoch Environ Res Risk Assess 22:135–137

Wu JL, Wang JF, Meng B, Chen G, Pang LH, Song XM, Zhang KL, Zhang T, Zheng XY (2004) Exploratory spatial data analysis for the identification of risk factors to birth defects. BMC Public Health 4:23–33