International Journal of Geographical Information Science
Vol. 24, No. 1, January 2010, 107–127

Research Article

Geographical Detectors-Based Health Risk Assessment and its Application in the Neural Tube Defects Study of the Heshun Region, China

JIN-FENG WANG*†, XIN-HU LI‡, GEORGE CHRISTAKOS§, YI-LAN LIAO†, TIN ZHANG¶, XUE GU¶ and XIAO-YING ZHENG**

†Institute of Geographic Sciences & Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China

‡Institute of Urban Environment, Chinese Academy of Sciences, Xiamen 361003, China

§Department of Geography, San Diego State University, San Diego, CA 92182-4493, USA

¶Beijing Institute of Paediatrics, Beijing, 100012, China

**Institute of Population Science, Peking University, Beijing 100871, China

(*Received 1 July 2008; in final form 27 August 2008*)

Physical environment, man-made pollution, nutrition and their mutual interactions can be major causes of human diseases. These disease determinants have distinct spatial distributions across geographical units, so that their adequate study involves the investigation of the associated geographical strata. We propose four geographical detectors based on spatial variation analysis of the geographical strata to assess the environmental risks of health: the risk detector indicates where the risk areas are; the factor detector identifies factors that are responsible for the risk; the ecological detector discloses relative importance between the factors; and the interaction detector reveals whether the risk factors interact or lead to disease independently. In a real-world study, the primary physical environment (watershed, lithozone and soil) was found to strongly control the neural tube defects (NTD) occurrences in the Heshun region (China). Basic nutrition (food) was found to be more important than man-made pollution (chemical fertilizer) in the control of the spatial NTD pattern. Ancient materials released from geological faults and subsequently spread along slopes dramatically increase the NTD risk. These findings constitute valuable input to disease intervention strategies in the region of interest.

Keywords: Geographical detectors; Disease; Determinants; Spatial consistence; Birth risk

1. Introduction

Environmental health is a modern yet critically important scientific discipline, which is generally concerned with the study of connections between environmental attributes that can affect the state of human health and the processes describing this state (e.g. Shields 1990). Ambient air pollutants, groundwater chemical contaminants, acid precipitation and net radiation are examples of environmental attributes. Coronary heart disease cases, breast cancer incidence and mesothelioma mortality

*Corresponding author. Email: wangjf@igsnrr.ac.cn

International Journal of Geographical Information Science
ISSN 1365-8816 print/ISSN 1362-3087 online © 2010 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/13658810802443457

are examples of health effects. Rigorous concepts and quantitative techniques have been developed for the purpose of assessing human exposure to environmental hazards and their various effects on the population in a space-time domain under conditions of multi-sourced uncertainty (Christakos and Hristopulos 1998, Kentel and Aral 2005, Robson and Toscano 2007).

Certain environmental attributes have the potential to become health determinants. In general, an environmental determinant of health is any attribute (biological, chemical, physical, social or cultural) that can be causally linked to a change in health state, e.g. known determinants include particulate matter and ozone concentrations, tobacco smoke, radiation, biological agents in the water and soil, persistent organic pollutants, greenhouse gases, various forms of waste, and contaminants transported via the food chain. The corresponding health effects include respiratory diseases, malaria, diarrheal diseases, cancers, tuberculosis, pneumonia, and cardiovascular disease (Flaum *et al.* 1996, Christakos and Vyas 1998, Chen *et al.* 2004). In many cases, the exposures are closely related in space-time to the health effects they produce, whereas in some others the exposures are far apart in space-time from the witnessed effects (Maxwell and Kastenberg 1999, Breslow 2002, Tamerius *et al.* 2007).

The residents of a geographical region (city, county, state, etc.) have the right to be informed about potential environmental determinants that are active in their region; policy-makers are interested about the geographical variation of environmental health risk in a population; and decision-makers developing disease intervention strategies need to differentiate between independent and interconnected health factors across space and time (Morgan 2002, Gu *et al* 2007, Mutshinda *et al.* 2007). The above considerations are all linked to spatial information technology matters that may be summarized into four questions:

(1) What is the geographical domain of the health risk?
(2) Which environmental parameters are responsible for the risk?
(3) What is the relative importance of each risk factor?
(4) Do the risk factors operate independently or they are interconnected?

Over the years, a number of quantitative tools have been developed to address questions such as the above. Among them, hotspots statistics techniques address question (1) above by testing the statistical significance of high *in situ* disease incidence ratios compared to the surrounding areas (Anselin 1995, Kulldorff 1997). Stochastic analysis accounts for the spatial dependence of disease characteristics and generate detailed maps of non-homogeneous disease variations together with a measure of map accuracy (Christakos and Lai 1997). Spatial linear regression and conditional logistic regression techniques identify the risk factors by means of the t-value of the regression coefficients (Haining 2003). In cases where the uncertainties of the relevant health risk variables cannot be presented in terms probabilities, they can be handled through fuzzy membership functions (Heng *et al.* 2008).

In this study, we address the four risk-related questions above by means of spatial variance analysis (SVA), which can be used for both measurable and categorical variables. The basic idea of SVA is to compare the spatial consistency of health risk distribution (e.g. disease cases) versus the geographical strata (e.g. climate, soil, water, population, ethnicity, culture and lifestyle, poverty, nutrition and land use) in which potential health risk factors exist. We first explore the relationship between environmental determinants of disease and their geographical variations. Then, we

introduce the concept of the power of determinant to the potential health effect, followed by a review of data preparation techniques. Four geographical detectors are defined and their basic features are examined. Finally, we obtain valuable insight by means of the neural tube defects (NTD) study (Heshun County, China).

2. Disease determinants and geographical space

As we mentioned above, there is a number of environmental attributes that can be determinants of disease development and evolution (Christakos and Hristopulos 1998, Collins 2004, Burton 2005). These attributes are characterized by their distinct spatial variations, in which case the geographical space can provide the interface to uncover possible determinants. In recent years, the geographical information system (GIS) technology offers a versatile operational platform to host spatial information, whereas quantitative spatial analysis provides powerful tools to complement the objective.

Rapid advances in earth observation technology made geographical data much more accessible and visible (Gewin 2004). In fact, many parts of the world have already been completely evaluated, digitally surveyed and are currently under real time surveillance. In many areas of the world, the geographical strata of health related factors have already been stored in GIS or, at least, they are available at coarse spatial resolutions, thus providing the relevant datasets and operational tables needed to investigate environmental health risk determinants.

If an environmental attribute leads to a disease, this disease would exhibit a spatial distribution similar to that of the environmental factor, the spatial 'factor-disease' consistency indicating the *in situ* existence of causal factors. According to table 1, that the occurrence of factor increases the disease risk is equivalent to the fact that the absence of the factor decreases the risk; conversely, that the occurrence of the factor decreases the disease risk is equivalent to the fact that the risk would increase if the factor is absent.

There are, at least, four kinds of relationships between health determinants and geographical space, as follows:

- Physical determinants of health are spatially distributed: Potential health hazards include surface and subsurface water contaminated by insufficiently oxygenized ancient geological media; also, radiation emissions from certain rocks or along faults (Boulding 1995, Burton 2005).
- Man-made pollution is spatially distributed: Hazards of this kind include pesticides and chemical fertilizers spread over crop fields; also, polluted air and water emission from workshops and electromagnetic radiation in workplaces (Shields 1990, Morgan 2002).
- Nutrition processes are spatially distributed: e.g. nutrition strongly depends on the spatially varying residential income; hence, it is usually proportional to the

Table 1. Occurrence (absence) of an environmental attribute increases (decreases) the disease risk.

	Area with environmental attribute	*Area without environmental attribute*
Area with disease cases	Yes (No)	No (Yes)
Area without disease cases	No (Yes)	Yes (No)

GDP that is regularly surveyed across space and published in the government's annual statistics/census reports (Jamison 1986, Berdanier 2002).

- Heredity and habits are spatially distributed: Ethnic groups have specific genetic and food habits and behavioral patterns, some of which are hazardous to health (McMichael 2001). Health determinants may be detected when the disease cases and the ethnic features share similar spatial features, e.g. that the shape and size of spatial disease clusters is consistent with that of citizens' daily activities suggests that heredity is relevant to the neural tube defects in the region (Wu *et al.* 2004).

3. Power of determinant

We will demonstrate the concept of 'power determinant' with the help of a simple yet illustrative example. In figure 1 we consider a study region A and a health effect

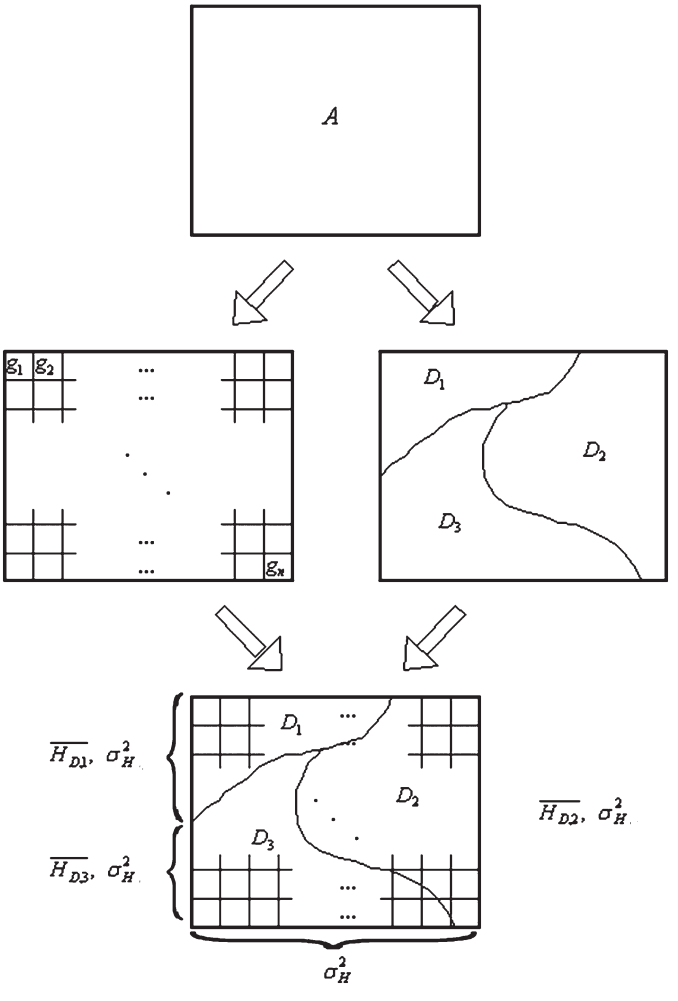

Figure 1. Division of the A study region A, the grid system $G=\{g_i;\ i=1, 2, \ldots, n\}$ and the geographical stratum $D=\{D_i;\ i=1, 2, 3\}$; the overlaid A, G and D with the corresponding statistical parameters are also shown.

H (heart disease rate, cancer mortality ratio, etc.) recorded on a grid system G consisting of units g_i ($i=1, 2, \ldots, n$) covering A. Assume that D_i ($i=1, 2, 3$) are the attributes associated with the geographical stratum of a suspected health determinant (e.g. climate or pollution), denoted as $D=\{D_i\}$. In many environmental health studies one needs to carefully examine the spatial relationship between the effect H and the stratum associated with the attributes D_i.

For this purpose, we first overlay the distribution of the health effect H over the geographical stratum of the suspected determinant $D=\{D_i\}$ (see figure 1). The mean value and the dispersion variance of H over the sub-regions of the attributes D_i are denoted as $\overline{H_{D,i}}$ and $\sigma^2_{H_{D,i}}$ ($i=1, 2, 3$), respectively. The significance of the variation of the mean values of H over the respective sub-regions i can be statistically tested. The $D=\{D_i\}$ is often suspected as a disease determinant, when the $\sigma^2_{H_{D,i}}$ of each sub-region is small, whereas the variances between sub-regions is large (which means that such a division explains most or even all of the spatial H variation). Note that $\sigma^2_{H_{D,i}} \to 0$, in the ideal case of a perfect division of the region.

Let $n_{D,i}$ be the number of samples in the sub-region i of the determinant D_i and let n be the total number of samples over the entire region A of interest, i.e. $n=\sum_{i=1}^{3} n_{D,i}$. The power of determinant $D=\{D_i\}$ to the health effect H is given by

$$P_{D,H} = 1 - \frac{1}{n\,\sigma^2_H} \sum\nolimits_{i=1}^{3} n_{D,i}\sigma^2_{H_{D,i}} \tag{1}$$

where the second term in the right hand side of equation (1) denotes the ratio of the $n_{D,i}$ weighted divisional variations $\sigma^2_{H_{D,i}}$ over the global variance σ^2_H of the health effect H in the study region. Note that, if it is a perfect division and $\sigma^2_H \neq 0$, then $P_{D,H}=1$. The $P_{D,H}$ concept is also considered in equation (10) below.

The health effect dispersion variance, the mean and its variance are the building blocks of geographical detectors and have to be estimated in the presence of spatial dependence. Consider a geographical region divided into environmental attribute units z. The mean and variance of the mean health effect $\overline{H_z}$ are given by, respectively, $\overline{\overline{H_z}}=\mu_z$ (the super-population mean of H in unit z) and $\sigma^2_{\overline{H_z}} = \frac{1}{n}\sigma^2_H + \frac{2}{n^2}\sum_{i<j} c_H(i,j)$, where c_H denotes the covariance of the H across space. The σ^2_H is estimated using (Haining 1988, Arlinghaus 1996),

$$\left.\begin{aligned} s^2 &= \tfrac{1}{n-1}\textstyle\sum_{i=1}^{n}\left(H_i - \overline{H}\right)^2 \\ \overline{s^2} &= \sigma^2_H - \tfrac{2}{n(n-1)}\textstyle\sum_{i<j} c_H(i,j) \end{aligned}\right\} \tag{2}$$

where H_i is the value of the sample $i=1, 2, \ldots, n$. Next, we describe the various steps of the population health risk assessment approach under real-world conditions. Then, we apply this approach in the study of NTD occurrences in the Heshun region (China).

4. Data preparation

4.1 *Downscaling from the r-framework to the z-framework*

In many studies there is a scale difference between the disease count unit r and the environmental attribute unit z, in that a disease is conveniently assigned to units selected by the human agent (postal, census, administrative), whereas the environmental attributes and geographic layers are naturally formed. Each

environmental attribute unit z is clipped by the disease count unit r, which divides it into a number of sub-units (polygons). There are, at least, five approaches of downscaling from the r-framework to the z-framework:

The first approach is based on *even spatial discretization*: The estimated population in geographical zone z is

$$\Pi_z^{Est} = \sum\nolimits_{r=1}^{N_{zr}} \Pi_{zr}^{Est} \tag{3}$$

where the estimated population in the unit r of z is $\Pi_{zr}^{Est} = \frac{\Pi_r^{Obs}}{A_r^{Obs}} A_{zr}^{Obs}$; Π_r^{Obs} is the observed population in unit r, A_r^{Obs} is the observed area of r, A_{zr}^{Obs} is the observed area of r that lies within z, and N_{zr} is the number of report units r in z. Accordingly, the estimated health effect (e.g., mortality or morbidity) in unit r of z will be $H_{zr}^{Est} = \frac{H_r^{Obs}}{\Pi_r^{Obs}} \Pi_{zr}^{Est}$, where H_r^{Obs} is the observed health effect in unit r; then, the effect in geographical unit z will be

$$H_z^{Est} = \sum\nolimits_{r=1}^{N_{zr}} H_{zr}^{Est} \tag{4}$$

Another approach is in terms of *monitoring classification*: Let H_p be the health effect at the sampling point p and z_p be the geographical unit covering p. Consider the conditional mean

$$\overline{H_p | z_p} = a + b\, z_p \tag{5}$$

where a and b are suitable coefficients, the bar denotes statistical expectation, and the vertical line denotes statistical conditional. By letting $\overline{H_p | z_p} = H_{zr}^{Est}$, one can use equation (4) to obtain the health effect H_{zr}^{Est} in the geographical unit z and the corresponding population $\left(\Pi_z^{Est}\right)$.

Other approaches include *Gaussian process regression* (using the sample to regress dependent variables to explanation variables in a spatial kernel and employ genetic program to form the function); *smoothing values in kernel* (i.e. setting the health effect value in z by smoothing the values in r in a distance kernel) and *point by area* (i.e. setting the effect value in z by taking the average of values in r); see (Gibbs 1997, Liao *et al.* 2008, Huang *et al.* 2008).

Monitoring classification and *Gaussian process regression* require measurable or scale quantities. For categorical variables, such as the geographical strata considered in the paper, the *spatial discretization*, *smoothing values in kernel*, and *point by area* approaches are applicable. The first two approaches approximate the real surface by area weighting and distance weighting, respectively, and are both preferable to the third approach. For the 'kernel' approach certain distance parameters have to be artificially set up, which is why we have chosen to use the *spatial discretization* approach in the present study.

4.2 *Measures of risk*

Let D denote an environmental determinant (climate, watershed, soil, poverty, etc.) that is a potential health hazard. We define certain forms of health risk measure,

say R, as

$$\left.\begin{aligned} R_r &= \frac{H_r^{Obs}}{\overline{H_r}},\ R_{z\,r} = \frac{H_{z,r}^{Est}}{\overline{H_{z,r}}} \\ \overline{R_r} &= \frac{H_z^{Est}}{\overline{H_z}},\ \overline{R_H} = \frac{H_D^{Est}}{\overline{H_D}} = 1 \end{aligned}\right\} \tag{6}$$

where $\overline{H_r} = \frac{\sum_{r=1}^{N_{z\,r}} H_r^{Obs}}{\sum_{r=1}^{N_{z\,r}} \Pi_r^{Obs}} \Pi_r^{Obs}$, $\overline{H_{z,\,r}} = \frac{H_z^{Est}}{\Pi_z^{Est}} \Pi_{z\,r}^{Est}$, $\overline{H_z} = \frac{H_z^{Est}}{\Pi_z^{Est}} \Pi_z^{Est}$ and $\overline{H_D} = H_D^{Obs}$. If necessary, the values above would be adjusted by Bayesian neighbors and age (Beaglehole *et al.* 1993, Haining 2003).

4.3 *Exploratory data analysis and Bayesian adjustment*

Since the proposed geographical detectors were based on variance addition, when a factor has many sub-regions (statistical units) of small area, the contribution of the factor could be underestimated or overestimated because the variances of such sub-regions are easily disturbed by the influence of other micro-exceptional elements. In this case, some outlier records could disguise the real spatial consistency between the factors and the disease and the real significance rank orders of the factors in causing a disease. Exploratory data analysis is necessary to filter out whenever outliers may appear, whereas the small sample problem can be further alleviated by the Bayesian adjustment technique (Haining 2003).

5. Geographical detectors

We propose four geographical detectors in order to address the following four questions:

(1) Where is the geographical location under environmental health risk R?
(2) Which determinants D_i are responsible for the risk?
(3) What is the risk difference, $\Delta R = R_i - R_j$, between the geographical regions i and j?
(4) Are the risk determinants D_i independent or interconnected as measured by the conditional probability $P[D_i|D_j]$?

The conceptual framework of the four geographical detectors (risk, factor, ecological and interaction) to be used in this study is given in table 2; a detailed technical discussion of these detectors follows.

Table 2. Conceptual framework the geographical detectors.

Detector	*Main ideas*
Risk detector	Compares the difference of average values between sub-regions; the bigger the difference, the greater the danger to the population health of the sub-region.
Factor detector	Compares the accumulated dispersion variance of each sub-region with the dispersion variance of the entire study region; the smaller the ratio, the stronger the disease contribution of the stratum.
Ecological detector	Compares the variance calculated from each sub-region divided according to one determinant with that divided according to another determinant.
Interaction detector	Compares the sum of the disease contribution of two individual attributes vs. the contribution of the two attributes when taken together.

5.1 *Risk detector*

The risk detector is used in the search for areas of potential health hazard. At the moment, it is assumed that disease occurrences are independently and identically distributed over space (p-independent; Brus and Gruijter 1997). According to the central limit theorem (Grimmett and Stirzaker 1992), the mean disease occurrence tends to be normally distributed. From the superpopulation perspective, the mean is a single realization of an underlying process (Griffith *et al.* 1994), therefore the difference between two superpopulation means was tested by student t (Press *et al.* 1992). The superpopulation mean was estimated by the observed mean using the ergodic assumption (Haining 1988). We compared the difference of the superpopulation means in paired regions.

The geographical zones z are ordered in descending order of risk R_z. The difference between the means of two geographical zones could be due to either sample random variation or fundamental differences of superpopulations, which have different meanings in epidemiology. Such a discrimination could be made by means of the statistics for measuring the significance of difference between the means of two distributions having an unequal variance (t-test; Press *et al.* 1992):

$$t_{\overline{R_{z=1}}-\overline{R_{z=2}}}=\frac{\overline{R_{z=1}}-\overline{R_{z=2}}}{\left[\frac{1}{n_{z=1,p}}\sigma^2_{\overline{R_{z=1}}}+\frac{1}{n_{z=2,p}}\sigma^2_{\overline{R_{z=2}}}\right]^{1/2}} \tag{7}$$

where $n_{z,p}$ refers to the number of sample units p in zone z. This statistic is distributed approximately as Student's t with a number of degrees of freedom equal to

$$df=\frac{\left[\frac{1}{n_{z=1,p}}\sigma^2_{\overline{R_{z=1}}}+\frac{1}{n_{z=2,p}}\sigma^2_{\overline{R_{z=2}}}\right]}{\frac{1}{n_{z=1,p}-1}\left[\frac{1}{n_{z=2,p}}\sigma^2_{\overline{R_{z=1}}}\right]^2+\frac{1}{n_{z=2,p}-1}\left[\frac{1}{n_{z=2,p}}\sigma^2_{\overline{R_{z=2}}}\right]^2} \tag{8}$$

To test the null hypothesis $H_0:\overline{R_{z=1}}=\overline{R_{z=2}}$, we give a significant level α (usually 5%), and find t_α by checking the student-t distribution table. If $\left|t_{\overline{R_{z=1}}-\overline{R_{z=2}}}\right|>t_{\alpha/2}$, reject H_0; hence, we conclude that there is a significant difference between the health risks of zone 1 and 2. This suggests examining the underlying natural mechanisms (chemical processes, interface between soil and people, the soil layer consists of different types distributed in different zones etc.) in order to explain the significant difference above.

5.2 *Factor detector*

Is a geographical stratum responsible for an observed spatial disease pattern? This can be measured by the difference between the dispersion variance $\sigma^2_{D,p}=\frac{1}{n_{D,p}}\sum_{p=1}^{n_{D,p}}\left(R_{D,p}-\mu_D\right)^2$ and the stratified population dispersion variance $\sigma^2_{D,z}=\frac{1}{n_{D,p}}\sum_{z=1}^{n_{D,z}}\sum_{p=1}^{n_{D,p}}\left(R_{z,p}-\mu_z\right)^2$, where $\mu_D=\overline{R_{D,p}}$, $\mu_z=\overline{R_{z,p}}$. We define $s^2_{D,p}=\frac{1}{n_{D,p}}\sum_{p=1}^{n_{D,p}}\left(H_{D,p}-\overline{H_D}\right)^2$ and $s^2_{D,p}=\frac{1}{n_{D,p}}\sum_{p=1}^{n_{D,p}}\left(R_{D,p}-\overline{R_D}\right)^2$, then $s^2_{D,p}=\sigma^2_{D,p}\left(1-\rho_{D,p}\right)$ and $s^2_{D,z}=\sigma^2_{D,z}\left(1-\rho_{D,z}\right)$, where ρ denotes the spatial autocorrelation coefficient, which is assumed to be second order spatially stationary, i.e. $\rho_{D,p}=\rho_{D,z}$.

To find out whether the two data sets have variances that are significantly different, the *F*-test is

$$F = \frac{\sigma^2_{D,\,p}}{\sigma^2_{D,\,z}} = \frac{m\,s^2_{1\,m}(n-1)}{n\,s^2_{2\,n}(m-1)} \tag{9}$$

This statistic is distributed approximately as $F(m-1, n-1)$, with degrees of freedom equal to $df=(m-1, n-1)=(n_{D,p}, n_{D,p})$ (Grimmett and Stirzaker 1992).

To test the null hypothesis $H_0 : \sigma^2_{D,\,p} = \sigma^2_{D,\,z}$, we give a significant level α (usually 5%), and find the f_α by checking the $F(m-1, n-1)$ in the distribution table. We find the lowest and highest $f_{1\alpha}$ and $f_{2\alpha}$; if $F^{Obs}(m-1, n-1)>f_{2\alpha}$, we reject H_0. We conclude that $\sigma^2_{D,\,p}$ is significantly bigger than $\sigma^2_{D,\,z}$; hence, geographical zonation causes significant differences in health results. The power of determinant of z to the disease, $P_{D,H}$, is defined as

$$P_{D,\,H} = 1 - \frac{\sigma^2_{D,\,z}}{\sigma^2_{D,\,p}} \tag{10}$$

Note that $P_{D,H}=1$ means that the geographical stratum completely explains the spatial pattern of the disease, whereas $P_{D,H}=0$ implies a completely random spatial occurrence of the disease.

5.3 *Ecological detector*

Is a geographical stratum (associated with one suspected determinant) more significant than another one (associated with another suspected determinant) in controlling the spread of the disease in space? e.g. if soil pollution (Layer 2) is more likely (or more significant) than water pollution (Layer 1) to cause a disease in the study area, we would expect the water population dispersion variance $\sigma^2_{D_1,\,z} = \frac{1}{n_{D_1,\,p}} \sum_{z=1}^{n_{D_1,\,z}} \sum_{p=1}^{n_{z,\,p}} \left(R_{D_1,\,p} - \overline{R_{D_1,\,p}}\right)^2$ to be larger than the soil population dispersion variance $\sigma^2_{D_2,\,z} = \frac{1}{n_{D_2,\,p}} \sum_{z=1}^{n_{D_2,\,z}} \sum_{p=1}^{n_{z,\,p}} \left(R_{D_2,\,p} - \overline{R_{D_2,\,p}}\right)^2$. The test for this is

$$F = \frac{n_{D_1,\,p}\left(n_{D_1,\,p}-1\right)\sigma^2_{D_1,\,z}}{n_{D_2,\,p}\left(n_{D_2,\,p}-1\right)\sigma^2_{D_2,\,z}} \tag{11}$$

This statistic is distributed approximately as $F\left(n_{D_1,\,p}-1, n_{D_2,\,p}-1\right)$, with degrees of freedom equal to $df = (m-1, n-1) = \left(n_{D_1,\,p}, n_{D_2,\,p}\right)$.

Again, to test the null hypothesis $H_0 : \sigma^2_{D_1,\,p} = \sigma^2_{D_2,\,z}$, we give a significant level α (usually 5%), and find f_α by checking $F(m-1, n-1)$ in the distribution table. We find the lowest and highest $f_{1\alpha}$ and $f_{2\alpha}$; if $F^{Obs}(m-1, n-1)>f_{2\alpha}$, we reject H_0. We conclude that $\sigma^2_{D_1,\,p}$ is significantly bigger than $\sigma^2_{D_2,\,z}$; hence, the soil 2 is a more significant health determinant than the water 1.

5.4 *Interaction detector*

Do two health determinants D_i (i=1, 2) when taken together weaken or enhance each another, or are they independent in developing a disease? To answer this kind of question we define the interaction detector, as follows

$$\left.\begin{array}{ll}
\textit{Enhance}: & P_{D,H}(D_1 \cap D_2) > P_{D,H}(D_1) \text{ or } P_{D,H}(D_2) \\
\textit{Enhance, bi-}: & P_{D,H}(D_1 \cap D_2) > P_{D,H}(D_1) \text{ and } P_{D,H}(D_2) \\
\textit{Enhance, nonlinear-}: & P_{D,H}(D_1 \cap D_2) > P_{D,H}(D_1) + P_{D,H}(D_2) \\
\textit{Weaken}: & P_{D,H}(D_1 \cap D_2) < P_{D,H}(D_1) + P_{D,H}(D_2) \\
\textit{Weaken, uni-}: & P_{D,H}(D_1 \cap D_2) < P_{D,H}(D_1) \text{ or } P_{D,H}(D_2) \\
\textit{Weaken, nonlinear-}: & P_{D,H}(D_1 \cap D_2) < P_{D,H}(D_1) \text{ and } P_{D,H}(D_2) \\
\textit{Independent}: & P_{D,H}(D_1 \cap D_2) = P_{D,H}(D_1) + P_{D,H}(D_2)
\end{array}\right\} \quad (12)$$

where the symbol '$\cap$' denotes the intersection between D_1 and D_2. The '$P_{D,H}(D_1 \cap D_2) > P_{D,H}(D_1)$ or $P_{D,H}(D_2)$ (*Enhance*)' is not equivalent to '$P_{D,H}(D_1 \cap D_2) > P_{D,H}(D_1)$ and $P_{D,H}(D_2)$(*Enhance, bi–*)', e.g. assume that $P_{D,H}(D_1)=0.2$, $P_{D,H}(D_2)=0.5$ and $P_{D,H}(D_1 \cap D_2)=0.3$; then '0.3 $(D_1 \cap D_2)$>0.2 (D_1) or 0.5 (D_2)' is true, but the '0.3 $(D_1 \cap D_2)$>0.2 (D_1) and 0.5 (D_2)' is not valid. It could be '0.3 $(D_1 \cap D_2)$<0.2 (D_1)+0.5 (D_2) (*Weaken*)' and also '0.3 $(D_1 \cap D_2)$>0.2 (D_1) (*Enhance*)', in which case the conclusion is that the D_1 and D_2 joint risk (0.3) enhances the D_1 single risk (0.2) but is smaller than the two individual risks added together (0.2+0.5). Similarly, other pair equations in equation (12) could be used together.

Model (12) can be easily implemented in GIS environment. The two layers D_1 and D_2 are overlaid and their attributes were combined $(D_1 \cap D_2)$ as a new attribute C. The power of determinants of D_1, D_2 and C are calculated respectively using equation (10), then are put into equation (12) for judgement.

The approach is feasibly extendable to more determinants. In the case of three determinants one finds:

$$\left.\begin{array}{ll}
\textit{Enhance}: & P_{D,H}(D_1 \cap D_2 \cap D_3) > P_{D,H}(D_1) \text{ or } P_{D,H}(D_2) \text{ or } P_{D,H}(D_3) \\
\textit{Enhance, tri-}: & P_{D,H}(D_1 \cap D_2 \cap D_3) > P_{D,H}(D_1) \text{ and } P_{D,H}(D_2) \text{ and } P_{D,H}(D_3) \\
\textit{Enhance, nonlinear-}: & P_{D,H}(D_1 \cap D_2 \cap D_3) > P_{D,H}(D_1) + P_{D,H}(D_2) + P_{D,H}(D_3) \\
\textit{Weaken}: & P_{D,H}(D_1 \cap D_2 \cap D_3) < P_{D,H}(D_1) + P_{D,H}(D_2) + P_{D,H}(D_3) \\
\textit{Weaken, uni-}: & P_{D,H}(D_1 \cap D_2 \cap D_3) < P_{D,H}(D_1) \text{ or } P_{D,H}(D_2) \text{ or } P_{D,H}(D_3) \\
\textit{Weaken, nonlinear-}: & P_{D,H}(D_1 \cap D_2 \cap D_3) < P_{D,H}(D_1) \text{ and } P_{D,H}(D_2) \text{ and } P_{D,H}(D_3) \\
\textit{Independent}: & P_{D,H}(D_1 \cap D_2 \cap D_3) = P_{D,H}(D_1) + P_{D,H}(D_2) + P_{D,H}(D_3)
\end{array}\right\}$$

The interpretation of these findings may become more complicated when the number of factors involved increase, because too many factors vary simultaneously.

6. The NTD Study of the Heshun County, China

Birth defects, as formally defined by the 'March of Dimes Birth Defects Foundation', refer to any anomaly (functional or structural) that is present in infancy or later in life and is caused 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 (Berry *et al.* 1999, Carmona *et al.* 2005, http://www.cdc.gov/ncbddd/bd/default.htm).

Laboratory experiments and epidemiological surveys reveal that chemical, biological, heredity, nutrimental materials and other un-identified factors, and/or their

interaction are relevant to birth defects (Cabrera *et al.* 2004, Detrait *et al.* 2005, Gu *et al.* 2007, Li *et al.* 2005, Meijer *et al.* 2005, Mitchell 2005, Wald 2004, Wu *et al.* 2004).

The Heshun county of Shanxi province in northern China is among the regions with the highest prevalence of neural-tube birth defects in the world. During the past 10 years, the Chinese government has taken measures to prevent the occurrence of birth defects, which has reduced the prevalence to some extent. But, because the underlying causes of many birth defects cases still remain unknown, the ratio is still high (Gu *et al.* 2007). Both physical and man-made environmental exposures as well as genetic predisposition are thought to contribute to birth defects (Wu *et al.* 2004), but the relative importance of the various factors needs to be quantitatively identified, so intervention could be more targeted and effectively implemented.

6.1 *Study area*

The Heshun county is located inside the Taihang mountain range in the eastern part of the Shanxi province (figure 2(a) and (b)), with an area of 2250 km^2, 330 administrative villages, 440 residents' aggregations and a population of 134,522 people (60 persons/km^2). The geomorpology is mountainous and hilly with nine

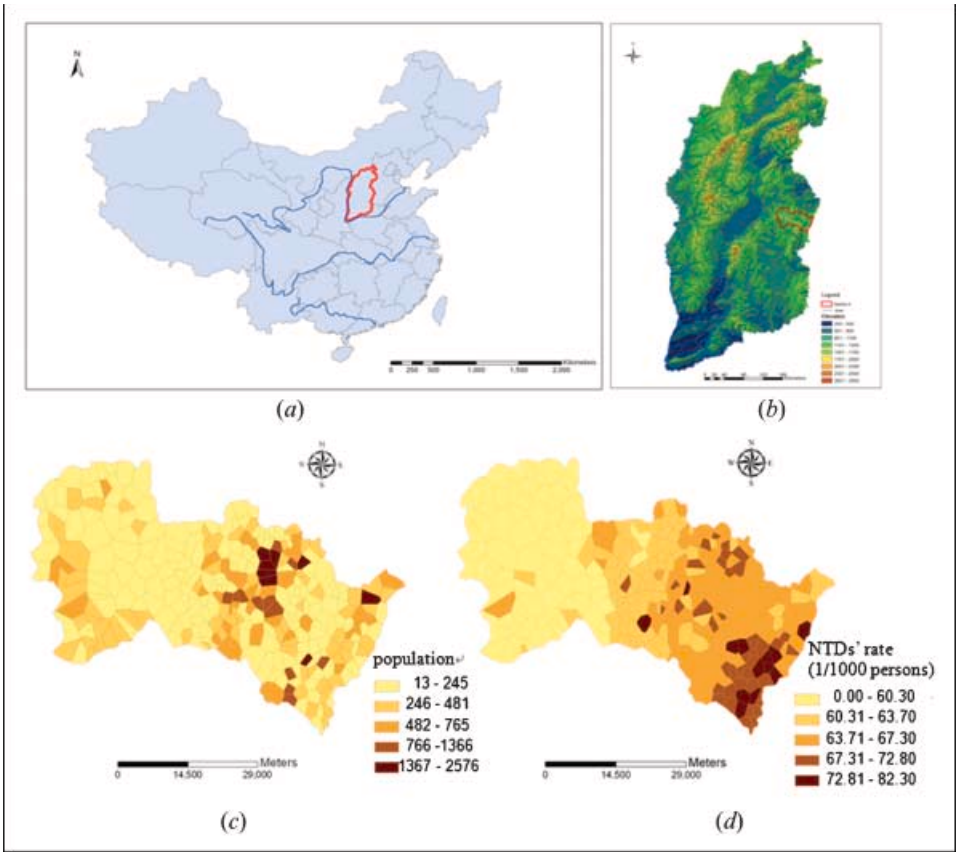

Figure 2. Maps of (a) China, (b) Shanxi, (c) Population and (d) rate of NTDs in Heshun County.

watersheds. The relative height difference is about 300–500 m with a mean elevation of 1300 m. The area has a continental climate with gusty and windy springs, warm and rainy summers, cold and rainy autumns, and long and chilly winters. The annual mean temperature is 6.3°C; the mean temperatures in January and July are 9.2 and 19.8°C, respectively. Most of the rainfall occurs in July-August. The annual mean precipitation of 593 mm mainly occurs during 54% of each year. Agriculture is the major human activity, with some coal-mining, forestry and livestock production. Both the physical environment and economic conditions are poor.

6.2 *Area population*

The population in the study area is 134522 (figure 2(c)), including 38,538 fertility-age women of whom 29,375 are married. In the year 2001, 41 of the 1200 newborns had birth defects, including 33 neural tube defects and 29 birth defect natural abortions. Using ultrasonic B-wave, the local clinic examined 970 pregnant women, and found 42 dysplastic embryos. Abortions were artificially induced in 33 of these cases; accordingly, the true birth defect prevalence was 6.17% (Heshun Birth Planning Committee 2002, Wu *et al.* 2004) (figure 2(d)). As one of the areas with the highest prevalence of neural-tube birth defects (NTD), inspection branches were well organized in this county. Birth defect records for 9 years (1998–2006) were acquired based on hospital registers and investigation in villages. These cases were divided into neural-tube birth defects and other birth defects by organ system.

6.3 *Suspect determinants of NTD*

The ordination scenarios of the social and environmental factors are displayed in figures 3–5.

6.4 *Bayesian adjusted prevalence rates*

For rare events, their rate values would subject to be high variance if they would be calculated based on short term observations, because the events may occur or not in the short period, much biased from the true superpopulation, the relationship established in the long term interaction between the environmental determinants and the neural birth defects *in situ*. In addition, the variances of the rates of rare diseases increase if the population of census units, which is the denominate of the rate, are small. Two efforts can reduce the variance of the rate values: collecting cases for as long as possible and the Bayesian technique. The latter borrows strength or information of both numerator (number of cases) and denominator (population) from spatial neighbors to increase the sample size *in situ* to reduce the variance of the ratio values. Ghosh and Rao (1994), Haining (2003) introduces the Bayesian adjustment technique; Wu *et al* (2004) apply the technique to the NTD.

Because birth defect is a low probability event (Rushton 1996), to reduce the prevalence rate variation, we collected NTD cases during seven years (1999–2005) in order to estimate the disease prevalence rate at each village. The prevalence rate variation arising from small sample size was further reduced by the Bayesian adjustment technique. In order to assess potential environmental determinants of the disease, the NTD prevalence H counted in the villages (figure 2) were transcribed into a geographical layer (D) composed of zones (z) (Mugglin and Carlin 1998, Liao *et al.* 2008).

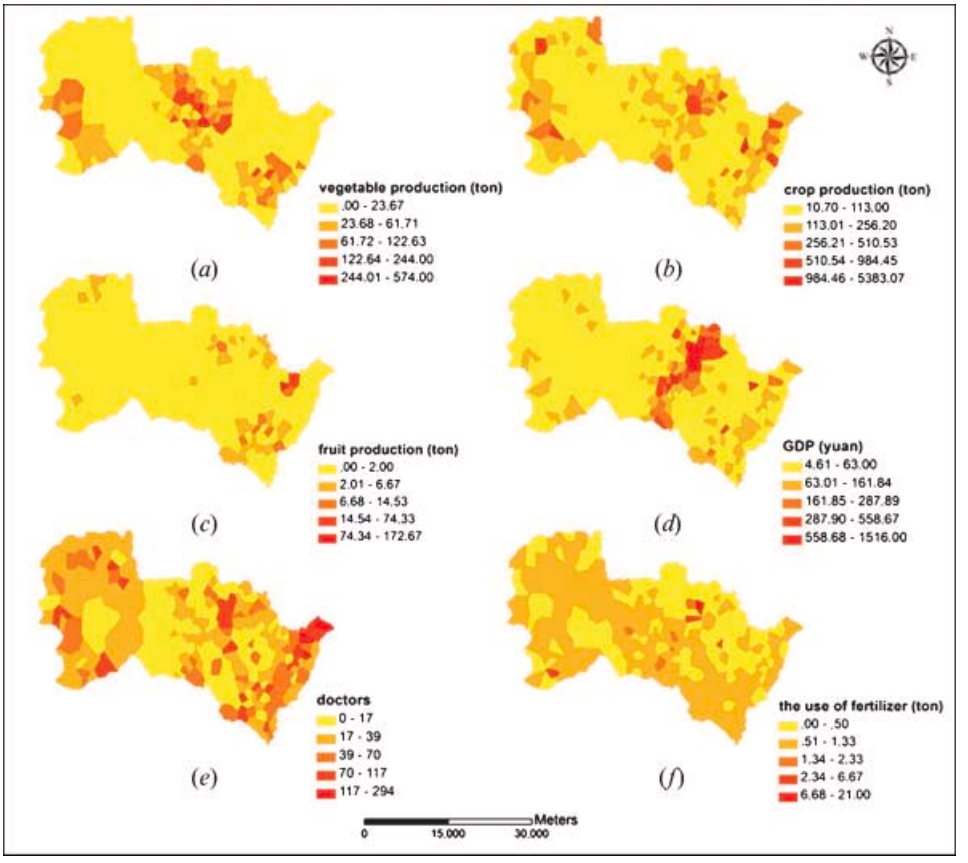

Figure 3. Map of suspect social strata of NTD.

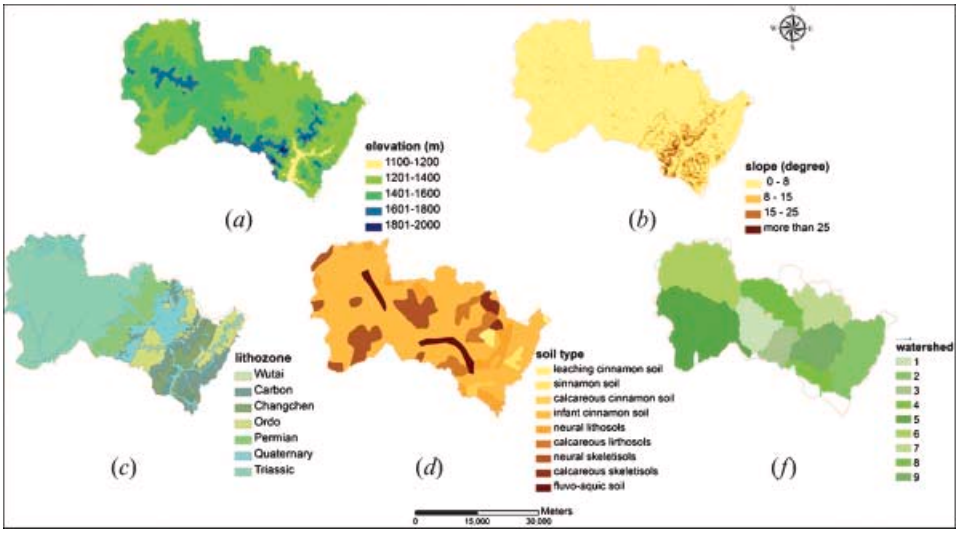

Figure 4. Map of suspect environmental strata of NTD.

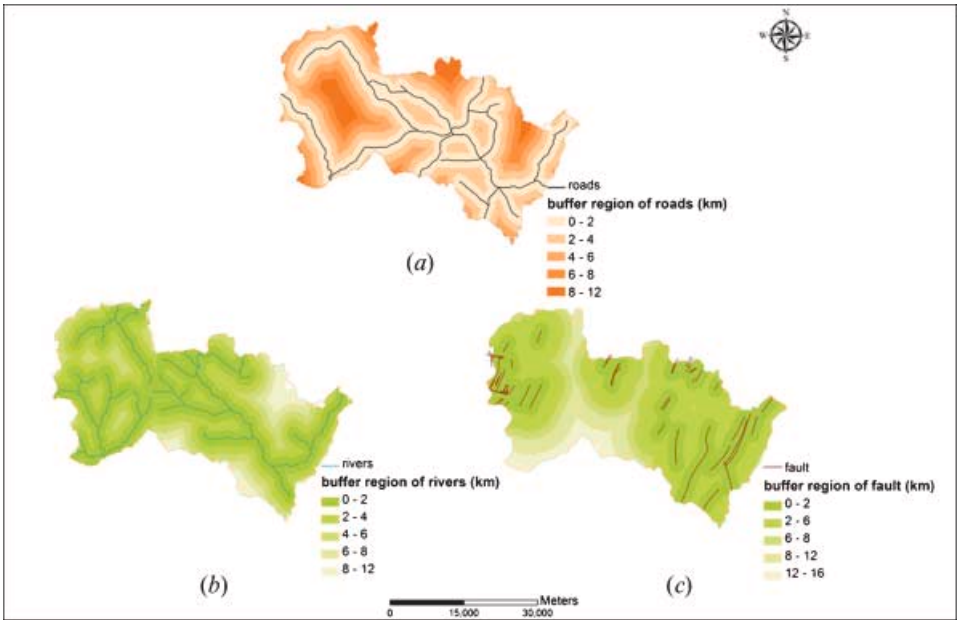

Figure 5. Map of suspect linearity type factors of NTD.

6.5 *Findings*

The risk detector answers the question of the geographical location (z) under environment health risk. Table 3 ordered watersheds by their risk R_z and compared the difference of the risk between the watersheds. Similarly, for other geographical strata.

The factor and ecological detectors disclosed that the geographical layers are ranked by their influence ($P_{H,D}$) on NTD occurrence in the following order:

> Watershed (47%)> lithozone (39%)> soil (24%)> fault (19%)> river buffer (13%)> elevation (10%)> slope (9%)> road buffer (7%)

Table 3. Statistical significance of the risk difference between nine watersheds.

Stat Sig Diff	2	4	7	9	3	8	1	5	6
2									
4	N								
7	Y	N							
9	Y	N	N						
3	Y	Y	Y	Y					
8	Y	Y	Y	Y	Y				
1	Y	Y	Y	Y	Y	Y			
5	Y	Y	Y	Y	Y	Y	Y		
6	Y	Y	Y	Y	Y	Y	Y	Y	

Note: The numbers stands for the codes of watersheds (please refer to figure 4(f)), Y means the risk difference between the two watersheds is significant with the confidence of 95%, and N means not.

The following conclusions have been drawn concerning the above: The last two factors (slope and road buffer) are statistically insignificant, and the $P_{H,D}$ difference between the first four geographical factors is also statistically insignificant. The variance of NTD rates within watersheds is the smallest, and the values in the upper, middle and lower courses in watersheds are not significantly different, which indicates that the NTD rates are relatively homogeneous within the watershed area. Water is the better medium than other physical factors to homogenize the spatial distribution of chemical and biological factors; a relatively closed watershed topology tends to assimilate the physical environment, human culture and their interaction. In addition, the variance of NTD rates in lithozone, soil and fault are small as well, showing that the primary natural environment strongly controls the NTD occurrence in the study area. The variances of the NTD rates within each buffer distance from the river, slope classifications of elevation and slope, and road buffer are bigger, reflecting their lower contributions to the disease (spatial pattern). River, elevation and slope are external drivers for the displacement of chemical elements. We used the risk detector to find that the NTD occurrence in the Carbon and Changchen lithozones were significantly higher than that in other zones, whereas NTD occurrence in the Permain and Triassic lithozones were significantly lower than others (see figure 6).

In addition to the above primary physical determinants, we also tested the impact of man-made environment and socioeconomic factors on NTD occurrence. Using the factor detector, the human factors were ranked according to their effect on NTD in the following order:

Crop production (17.5%)> vegetable production (11.6%)> GDP (11.3%)> number of doctors (1.3%)> fertilizer use (0.9%)

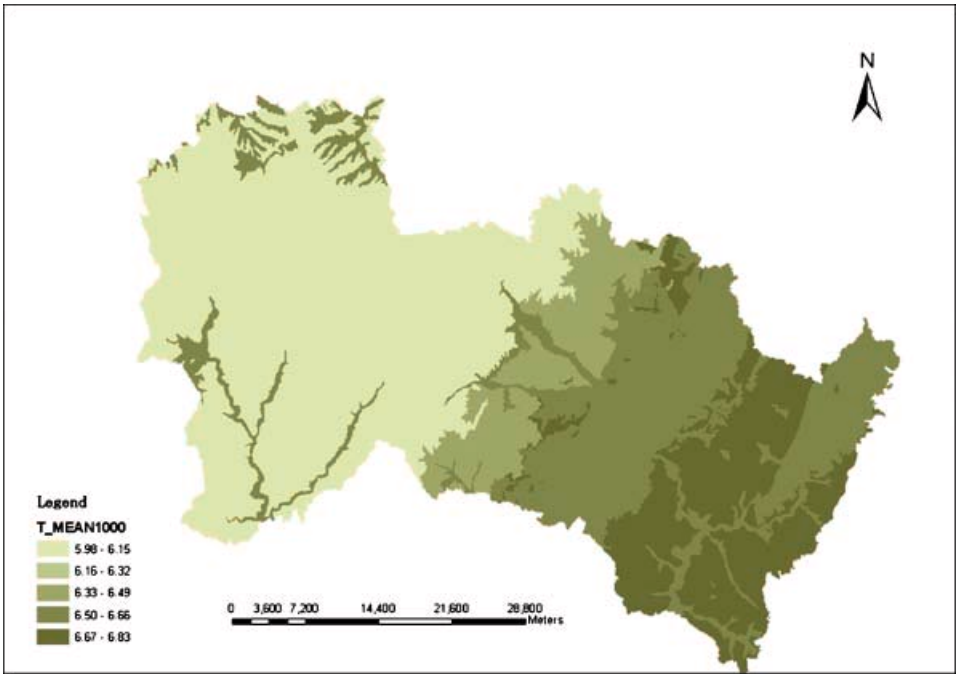

Figure 6. Prevalence rate of NTD over lithozone in Heshun county.

The food consumption is proportional to the local production. This series revealed that basic nutrition (food) rather than artificial pollution (fertilizer) controls the occurrence of NTD.

In view of the above considerations, we draw the conclusion that the power of human factors is much lower than that of physical factors in generating NTD cases in the study area.

The interaction detector was used to check whether two NTD determinants work independently or not. The findings are tabulated in table 4. Geological faults and slopes were found to enhance each other to increase the NTD risk (fault∩slope =0.86>0.28=fault (0.19)+slope (0.09)). Faults are often generated by intense deformation of rock stratum causing the release of very ancient materials such as gas or radiation, whereas slope is an external force that facilitates the spread of the materials. Lithozone was found to disrupt the watershed's control of NTD (lithozone∩watershed=0.45<0.86=lithozone (0.39)+watershed (0.47)).

We also investigated the joint impact of physical and human factors on NTD rates. Ranked by power of determinant ($P_{D,H}$) it was found that:

> Lithozone∩fruit (51.6%)> lithozone∩fertilizer (45.5%)> lithozone∩vegetable (40.3%)> lithozone+GDP (39.3%); soil∩vegetable (28.5%)> soil∩fruit (28.1%)> soil∩fertilizer (24.9%)> soil∩GDP (24.7%)> soil∩doctor (24.6%); fault∩vegetable (29.3%)> fault∩fruit (28.2%)> fault∩doctor (24.2%)> fault∩fertilizer (24.1%)> fault∩GDP (23.3%)

Table 4. Interaction between pairs of physical factors in introducing NTD.

$L=C=A\cap B: 1-\frac{\sigma^2_{L,z}}{\sigma^2_{L,p}}$	$A+B: \sum_{L=A,B}\left(1-\frac{\sigma^2_{L,z}}{\sigma^2_{L,p}}\right)$	*Conclusion*	*Interpretation*
soil∩slope	=0.1<0.33=Soil(0.24)+slope(0.09)	$C<A$	slope↘soil
lithozone∩slope	=0.39<0.48=lithozone (0.39)+slope (0.09)	$C=A$; $C<A+B$	slope↘lithozone
lithozone∩fault	=0.45<0.58=lithozone (0.39)+fault(0.19)	$C>A, B$; $C<A+B$	lithozone↑↑fault
lithozone∩watershed	=0.45<0.86=lithozone(0.39)+watershed(0.47)	$C>B$	lithozone↘watershed
lithozone∩soil	=0.51<0.63=lithozone (0.39)+soil(0.24)	$C>A, B$; $C<A+B$	lithozone↑↑soil
soil∩elevation	=0.56>0.34=soil (0.24)+elevation(0.10)	$C>A+B$	soil⥮elevation
fault∩elevation	=0.66>0.29=fault (0.19)+elevation(0.10)	$C>A+B$	fault⥮elevation
fault∩watershed	=0.71>0.66=fault (0.19)+watershed (0.47)	$C>A+B$	fault⥮watershed
fault∩soil	=0.78>0.43=fault (0.19)+soil (0.24)	$C>A+B$	fault⥮soil
stratum∩elevation	=0.84>0.49=lithozone (0.39)+elevation (0.10)	$C>A+B$	lithozone⥮elevation
fault∩slope	=0.86>0.28=fault (0.19)+slope (0.09)	$C>A+B$	faul⥮slope

Notes: $A\searrow B$ denotes A weakens B; $A\nearrow B$ denotes A enhances B; $A\uparrow\uparrow B$ denotes A and B enhance each other when $C>A, B$; $A\downarrow\downarrow$B denotes A and B weaken each other. $A\leftrightarrows B$ denotes A and B are not independent in leading to disease; $A\square B$ denotes A and B are independent in leading to disease; $A⥮B$ denotes nonlinear enhancement of A and B; $A⥯B$ denotes nonlinear weakening of A and B.

The interactions of the two types of factors, listed in table 5, shows that there is not much difference between $P_{D,H}(D_1 \cap D_2)$ and $P_{D,H}(D_1) + P_{D,H}(D_2)$ which means that human factors have only a slight impact on physical factors in controlling the NTD spatial patterns. Note that the SVA conceptual framework is general and can be used in the study of other environment-related diseases, as well.

7. Conclusions and discussion

The causes of many diseases are complicated (Christakos *et al.* 2005) and the health resources are limited in developing countries. So, tools are extremely welcome that are relatively cheap and easy to implement in determinant detection for priority prevention and disease intervention. The four geographical detectors have been developed in this paper as a response to this objective.

Many diseases can only be partially explained by genetic, environment, nutrition or other single factors; actually, they are often comprehensive consequences of mixture and interaction of multiple factors (Texas Department of State Health Service 2008). Clinical and laboratory work are concerned about a single patient and a single sample of a disease, whereas epidemiology uses population-based surveys to investigate the factors common to patients.

The prevalence of environment-related diseases often exhibits wide variations over geographic locations, in which case it may offer new perspectives about disease epigenetics. A single patient (the object of clinic) always exists at specific spatial sites, where all environmental and social determinants of the disease play a role. The determinants vary over geographical strata (soil, water, climate, poverty nutrition, pollution or any other spatial zone) associated with populations of patients. The spatial correspondence between disease determinants and their geographical storage units enable us to investigate the determinants through the geographical strata. Once a geographical stratum is tested to be statistically significant in controlling the spatial pattern of the occurrence of a disease, the search for disease determinants focuses on factors related to the geographical stratum. These factors could be genetic, dietary, infectious, occupational, chemical, physical, biological, social or interactions therein. In this study, we proposed four novel geographical detectors to

Table 5. Interaction between physical and human factors in introducing NTD.

$L = C = A \cap B : 1 - \frac{\sigma^2_{L,z}}{\sigma^2_{L,p}}$			$A + B : \sum_{L=A,B}\left(1 - \frac{\sigma^2_{L,z}}{\sigma^2_{L,p}}\right)$		Conclusion	Interpretation
lithozone∩GDP	=0.39	<	0.50	= lithozone (0.39) + GDP (0.113)	$C = A$; $C < A + B$	lithozone↓↓GDP
lithozone∩vegetable	=0.40	<	0.51	= lithozone (0.39) + vegetable (0.116)	$C > A, B$; $C < A + B$	lithozone↓↓vegetable
lithozone∩fertilizer	=0.45	>	0.40	= lithozone (0.39) + fertilizer (0.009)	$C > A + B$	lithozone⬆fertilizer
lithozone∩fruit	=0.52	<	0.56	= lithozone (0.39) + fruit (0.175)	$C > A, B$; $C < A + B$	lithozone↑↑fruit

Notes: $A \searrow B$ denotes A weakens B; $A \nearrow B$ denotes A enhances B; $A \uparrow\uparrow B$ denotes A and B enhance each other when $C > A, B$; $A \downarrow\downarrow B$ denotes A and B weaken each other. $A \leftrightarrows B$ denotes A and B are not independent in leading to disease; $A \cap B$ denotes A and B are independent in leading to disease; $A \diamondsuit B$ denotes nonlinear enhancement of A and B; $A \diamondsuit B$ denotes nonlinear weakening of A and B.

filter out and differentiate the relative importance of the determinants based on spatial variation information.

The four detectors proposed in this paper were used to identify the environmental risk of NTD in the pilot area. The primary physical environment strongly controls NTD occurrence. Basic nutrition was found to be more important than chemical fertilizers in controlling the spatial NTD pattern. Ancient materials released from faults and spread along slopes dramatically increase the NTD risk. These findings provide valuable information for disease intervention in the region.

The theory proposed in the study is based on spatial variance analysis of the spatial consistency of health risk distribution with suspect geographical strata. The validation of the results is evaluated by statistical significance test. One limitation of the geographical detectors is that they are statistical and are not causality, but the geographical detectors can filter out highly suspect factors of health for further confirmation by biological experiments. Another limitation is that some health hazards may do not present spatial patterns, or probably the study domain is too small to display a geographical strata, therefore, our theory is not sufficient to detect out all risks, field sampling survey for suspect factors is necessary to find out the health hazards that have a weak spatial pattern. The power of determinant is affected by geographical strata homogeneity, e.g. snail is an indispensable intermediate host in the schistosomiasis transmission process and varies in some very limited areas of marshland, which is heterogeneously distributed in lake region. Therefore, an optimal zonation identified by both optimal classification algorithms and prior knowledge of diseases (Wang *et al.* 1997, Li *et al.* 2007) would raise the power of determinant's efficiency.

Diseases are preventable at the genetic, personal and population levels, but especially at the geographical level for environment-related and communicable diseases (Keeling *et al.* 2003, Wang *et al.* 2006, Wang *et al.* 2008). Environmental risk assessment is also meaningful in clinical practice, because the environmental factors of a geographical zone with specified disease prevalence are most probably the health related factors for patients coming from that zone. The probability of an individual's good health over time is generally similar to that of the population's good health in an area with similar features (Beaglehole *et al.* 1993). Disease intervention could be conducted based on the findings of geographical detection (Jacquez *et al.* 2005), e.g. if the primary physical environment (watershed or soil) is responsible for a disease in the area, it is quite hard to intervene. If man-made pollution is responsible for the disease, the risk can be controlled by removing the pollution sources. If basic nutritional deficiency is the disease cause, nutritional supplements can reduce prevalence. Actually, nutrition intervention had been conducted in some Heshun province and the effects are remarkable (Chen *et al.* 2008). The genetic factor is not concluded in environmental models but is left in model residuals. Accordingly, exploring the residuals of the models may reveal the genetic susceptibility and heterogeneity of the population distributed over a space.

Acknowledgements

The Natural Science Foundation of China (40471111, 70571076), The Ministry of Science and Technology of China (2001CB5130, 2007CB5119001, 2006AA12Z215, 2007AA12Z241, 2007DFC20180), and Chinese Academy of Sciences (KZCX2-YW-308) sponsored this study. Partial support for this work was also provided by a grant from the California Air Resources Board, USA (Grant No. 55245A).

References

ANSELIN, L., 1995, Local indicators of spatial association–LISA. *Geographical Analysis*, **27**, pp. 93–115.

ARLINGHAUS, S.L. (Ed.), 1996, *Practical book of spatial statistics* (New York: CRC Press).

BEAGLEHOLE, R., BONITA, R. and KJELLSTROM, T., 1993, *Basic epidemiology* (Geneva: World Health Organization).

BERDANIER, C.D. (Ed.), 2002, *Handbook of nutrition and food* (Boca Raton, FL: CRC Press).

BERRY, V., LI, Z., ERICKSON, J., LI, S., MOORE, C., WANG, H., MULINAR, J., ZHAO, P., YANG, L. and WONG, C., 1999, Preventing neural-tube defects with folic acid in China. *N Engl J Med*, **341**, pp. 1485–1490.

BOULDING, J.R., 1995, *Practical handbook of soil, vadose zone, and groundwater contamination* (Boca Raton, FL: CRC Lewis Publ).

BRESLOW, L. (Ed.), 2002, Environmental determinants of health. In *Encyclopedia of public health* (Farmington Hills, MI: Gale Group, Inc.).

BRUS, D.J. and GRUIJTER, J.J., 1997, Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). *Geoderma*, **80**, pp. 1–44.

BURTON, I., 2005, *Integration of public health with adaptation to climate change: Lessons learned and new directions* (New York: Taylor & Francis).

CABRERA, R.M., ETHEREDGE, A.J. and FINNELL, R.H., 2004, Investigations into the etiology of neural tube defects. *Birth Defects Research (Part C)*, **72**, pp. 330–344.

CARMONA, R.H., 2005, The global challenges of birth defects and disabilities. *Lancet*, **366**, pp. 1142–1144.

CHEN, G., SONG, X.M., JI, Y., ZHANG, L., CHEN, J.P., LIU, J.F., PEI, L.J., LI, C.F. and ZHENG, X.Y., 2008, Prevention of neural tube defects with peri-conceptional multivitamin supplementation containing folic acid in China. *Birth Defects Research Part A*, **82**, pp. 592–596.

CHEN, C-C., WU, K-Y. and CHANG, M.J.W., 2004, A statistical assessment on the stochastic relationship between biomarker concentrations and environmental exposures. *Stochastic Environmental Research and Risk Assessment*, **18**, pp. 377–385.

CHRISTAKOS, G. and HRISTOPULOS, D.T., 1998, *Spatiotemporal environmental health modelling: A tractatus stochasticus* (Boston, TX: Kluwer Academic).

CHRISTAKOS, G. and LAI, J., 1997, A study of the breast cancer dynamics in North Carolina. *Soc Scie & Medicine D-Medical Geography*, **45**(10), pp. 1503–1517.

CHRISTAKOS, G. and VYAS, V., 1998, A novel method for studying population health impacts of spatiotemporal ozone distribution. *Social Science & Medicine D: Medical Geography*, **47**(8), pp. 1051–1066.

CHRISTAKOS, G., OLEA, R.A., SERRE, M.L., YU, H.L. and WANG, L.-L., 2005, *Interdisciplinary public health reasoning and epidemic modelling: The case of black death* (New York: Springer-Verlag).

COLLINS, F., 2004, The case for a US prospective cohort study of genes and environment. *Nature*, **429**, pp. 475–477.

DETRAIT, E.R., GEORGE, E.M., ETCHEVERS, H.C., GILBERT, G.R., VEKEMANSA, M. and SPEER, M.C., 2005, Human neural tube defects: Developmental biology, epidemiology, and genetics. *Neurotoxicology and Teratology*, **27**, pp. 515–524.

GEWIN, V., 2004, Mapping opportunities. *Nature*, **427**, pp. 376–377.

GHOSH, M. and RAO, J.N.K., 1994, Small area estimation: an appraisal (with discussion). *Statistical Science*, **9**, pp. 55–93.

FLAUM, J.B., RAO, S.T. and ZURBENKO, I.G., 1996, Moderating the influence of meteorological conditions on ambient ozone concentrations. *J. Air & Waste Manage Assoc*, **46**, pp. 35–46.

GIBBS, M.N., 1997, *Bayesian Gaussian processes for classification and regression*, PhD thesis, University of Cambridge, Cambridge, UK.

GRIFFITH, D.A., HAINING, R. and ARBIA, G., 1994, Heterogeneity of attribute sampling error in spatial data sets. *Geographical Analysis*, **26**, pp. 300–320.

GRIMMETT, R.R. and STIRZAKER, D.R., 1992, *Probability and Random Processes* (Oxford: Oxford Science Publications).

GU, X., LIN, L.M., ZHENG, X.Y., ZHANG, T., SONG, X.M., WANG, J.F., LI, X.H., LI, P.Z., CHEN, G., WU, J.L., WU, L.H. and LIU, J.F., 2007, High prevalence of NTD in Shanxi Province: A combined epidemiological approach. *Birth Defects Research Part A–Clinical and Molecular Teratology*, **79**, pp. 702–707.

HAINING, R., 1988, Estimating spatial means with an application to remote sensing data. *Communication Statistics*, **17**(2), pp. 537–597.

HAINING, R., 2003, *Spatial data analysis: Theory and practice* (Cambridge: Cambridge Univ Press).

HENG, L.L., HUANG, G.H. and ZOU, Y., 2008, An integrated fuzzy-stochastic modeling approach for assessing health-impact risk from air pollution. *Stochastic Environmental Research and Risk Assessment*, **22**, p. 789.

HESHUN BIRTH PLANNING COMMITTEE, 2002, *Report on the birth defect intervention project.* Government Document.

HUANG, B., PIERRE, F., XUE, L.L. and WANG, Y.J., 2008, Seeking the pareto front for multiobjective spatial optimization problems. *International Journal of Geographical Information Science*, **22**, pp. 507–526.

JACQUEZ, G.M., GREILING, D.A. and KAUFMANN, A.M., 2005, Design and implementation of a space-time intelligence system for disease surveillance. *Journal of Geographical System*, **7**, pp. 7–23.

JAMISON, D.T., 1986, Child malnutrition and school performance in China. *J. of Development Economics*, **20**(2), pp. 299–309.

KEELING, M., WOOLHOUSE, M., MAY, R., DAVIES, G. and GRENFELL, B., 2003, Modeling vaccination strategies against foot-and-mouth disease. *Nature*, **421**, pp. 136–142.

KENTEL, E. and ARAL, M.M., 2005, 2D Monte Carlo versus 2D fuzzy Monte Carlo health risk assessment. *Stochastic Environmental Research and Risk Assessment*, **19**, pp. 86–96.

KULLDORFF, M., 1997, A spatial scan statistic. *Communications in Statistics: Theory and Methods*, **26**, pp. 1481–1496.

LI, L.F., WANG, J.F., CAO, Z.D., FENG, X.L., ZHANG, L.L. and ZHONG, E.S., 2007, An information-fusion method to identify pattern of spatial heterogeneity for improving the accuracy of estimation. *Stochastic Environmental Research and Risk Assessessment*, **22**, pp. 689–704.

LI, W.Z., REN, A.G., ZHANG, L., GUO, Z.Y., LI, S., YE, R.W., ZHAI, R.Q., JIA, L.T., XIAO, Y.P., ZHAO, M.H., LI, Y.Z., ZHU, X., ZHOU, M.X. and LI, X., 2005, Prevalence of major external birth defects in high and low risk areas in China. *Chin J Epidemiol*, **26**, pp. 252–257.

LIAO, Y.L., WANG, J.F., MENG, B. and LI, X.H., 2010, Integration of GP and GA for mapping population distribution. *International Journal of Geographical Information Science*, **24**, pp. 47–67.

MAXWELL, R.M. and KASTENBERG, W.E., 1999, A model for assessing and managing the risks of environmental lead emissions. *Stochastic Environmental Research and Risk Assessment*, **13**, pp. 231–250.

MCMICHAEL, T., 2001, *Human frontiers, environments and disease: Past patterns, uncertain futures* (Cambridge: Cambridge University Press).

MEIJER, W.M., DE WALLE, H.E.K., KERSTJENS-FREDERIKSE, W.S. and DE JONG-VAN, L.T.W., 2005, Folic acid sensitive birth defects in association with intrauterine exposure to folic acid antagonists. *Reproductive Toxicology*, **20**, pp. 203–207.

MITCHELL, L.E., 2005, Epidemiology of neural tube defects. *American Jour of Medical Genetics Part C*, **135C**, pp. 88–94.

MORGAN, M.T., 2002, *Environmental health* (Belmont, CA: Brooks Cole).

MUGGLIN, A.S. and CARLIN, B.P., 1998, Hierarchical modelling in geographic information systems: population interpolation over incompatible zones. *Jour of Agricultural, Biological and Environmental Statistics*, **3**, pp. 111–130.

MUTSHINDA, C.M., ANTAI, I. and O'HARA, R.B., 2007, A probabilistic approach to exposure risk assessment. *Stochastic Environmental Research and Risk Assessment*, **22**, pp. 441–449.

PRESS, W., TEUKOLSKY, A., VETTERLING, W. and FLANNERY, B., 1992, *Numerical Recipes in C*, Second edition (New York: Cambridge University Press).

ROBSON, M.G. and TOSCANO, W.A., 2007, *Risk assessment for environmental health* (New York: Jossey-Bass).

RUSHTON, G. and LOLONIS, P., 1996, Exploratory spatial analysis of birth defect rates in an urban population. *Stat Med*, **15**, pp. 717–726.

SHIELDS, J., 1990, *Environmental health, new directions* (Princeton, NJ: Princeton Sci. Publ. Co., Inc.).

TAMERIUS, J.D., WISE, E.K., UEJIO, C.K., MCCOY, A.L. and COMRIE, A.C., 2007, Climate and human health: synthesizing environmental complexity and uncertainty. *Stochastic Environmental Research and Risk Assessment*, **21**, pp. 601–613.

Texas Department of State Health Service, 2008, Birth Defects: Epidemiology & Surveillance. Available online at: http://www.dshs.state.tx.us/birthdefects/default.shtm.

WALD, N., 2004, Folic acid and the prevention of neural-tube defects. *N Engl J Med*, **350**, pp. 101–103.

WANG, J.F., CHRISTAKOS, G., HAN, W.G. and MENG, B., 2008, Data-driven exploration of 'spatial pattern-time process-driving forces' associations of SARS epidemic in Beijing, China. *Journal of Public Health*, **30**, pp. 234–244.

WANG, J.F., MCMICHAEL, A.J., MENG, B., BECKER, N.G., HAN, W.G., GLASS, K., WU, J.L., LIU, X.H., LIU, J.Y., LI, X.W. and ZHENG, X.Y., 2006, Spatial dynamics of a SARS epidemic in an urban area. *Bulletin of World Health Organization*, **84**, pp. 965–968.

WANG, J.F., WISE, S. and HAINING, R., 1997, An integrated regionalization of earthquake, flood and drought hazards in China. *Transactions in GIS*, **2**, pp. 25–44.

WU, J.L., WANG, J.F., MENG, B., CHEN, G., PANG, L.H., SONG, X.M., ZHANG, K.L., ZHANG, T. and ZHENG, X.Y., 2004, Exploratory spatial data analysis for the identification of risk factors to birth defects. *BMC Public Health*, **4**, p. 23.