Using Spatial Prediction Model to Analyze Driving Forces of the Beijing 2008 HFMD Epidemic

JiaoJiao Wang1,* , ZhiDong Cao2, QuanYi Wang3, XiaoLi Wang3, and HongBin Song4

1 College of Geoscience and Surveying Engineering, China University of Mining & Technology (Beijing), No. 11, XueYuan Road, HaiDian District 100083 Beijing, China
2 Key Laboratory of Complex Systems and Intelligence Science, Institute of Automation, No.95, ZhongGuanCun East Road, HaiDian District 100190 Beijing, China
3 Beijing Center for Disease Control and Prevention, 100013 Beijing, China
4 PLA Institute of Disease Control and Prevention, 100071 Beijing, China
{JiaoJiao Wang,ZhiDong Cao,QuanYi Wang,XiaoLi Wang, HongBin Song,LNCS}@Springer.com

Abstract. Based on the spatial units of community, village and town in Beijing, the relationship between HFMD morbidity and the potential risk factors has been examined. According to the 6 selected risk factors (namely population density, disposable income of urban residents, the number of medical and health institutions, the number of hospital beds, average annual temperature and average annual relative humidity) significantly related to HFMD morbidity, the prediction performance of Classical Linear Regression Model(CLRM) and Spatial Lag Model(SLM) has been compared. The results showed that SLM achieved better effect and R square reached 0.82. It was showed that spatial effect played the crucial role in the HFMD morbidity prediction and its contribution attained 88%. However, CLRM showed low prediction accuracy and bias estimation. It was demonstrated that including spatial effect item into CLRM could greatly improve the performance of HFMD morbidity prediction model.

Keywords: Hand-foot-mouth disease(HFMD) morbidity, risk factor, Beijing, Classical Linear Regression Model, Spatial Lag Model

1 Introduction

Hand-foot-mouth disease (HFMD) prevalence has been reported in a majority of the countries and regions all over the world. HFMD is caused by the intestinal virus and
easily infectious among the crowds. HFMD cases are mostly made of the infants and the children. Adults HFMD cases unusually appreared, but they could take and transmit the virus. HFMD occurred in China in 1981 and became prevalence in most of the provinces all over the country. Recently, the scale of HFMD outbreak has gradually increased and greatly threatened the health security of the nation. In 2008, the administration of HFMD was strengthened further[1].

So far, most research on HFMD has been focused on the fields such as molecular biology[2], clinical medicine[3], pathogeny[4] and epidemiology[5]. Although some scholars investigated the factors that could cause HFMD prevalence from the view of physical, social, economic and humanity, the epidemiology analysis was carried out mostly based on classical statistics. At the end of 1960s, spatial correlation was firstly used in the research on pathogeny analysis[6]. In 1973, Cliff and Ord[7] put forward the concept and framework of spatial autocorrelation. Since then, a series of exploration and study on the theory and methodology about spatial data analysis has come forth[8-10]. Over the recent years, the technology of Geographical Information System(GIS) in which spatial analysis is the key point has developed very rapidly and been increasingly used in the field of public health[11-14].

2 Data

2.1 HFMD Morbidity

HFMD prevalence can be measured by HFMD morbidity which means the ratio of the number of HFMD cases in a year to the number of total population in the same region.

Beijing Center for Disease Control and Prevention (CDC) received 18,445 observed HFMD cases covering 309 towns and villages all over the 18 administrative counties in Beijing City, with the period from October 12, 2007 to October 31, 2008[15].

2.2 Risk Factors

There are many potential risk factors to influence HFMD epidemic, such as environmental, demographic, socioeconomic and human factors. The preceding study
has indicated that health service could play an important role in HFMD prevalence[15]. According to the data obtained, we divided the 16 potential factors into the 6 labels as Table 1 shows.

**Table 1.** Potential risk factors of 2008 Beijing HFMD prevalence. The calculation of population density was detailed in literature [15]. The climate materials were from meteorological observation stations under China Meteorological Administration including 137 stations located in Beijing City and the surrounding areas 250 km away from Beijing (detailed in literature [16]). The data source of the other factors was from Beijing Regional Statistical Yearbook 2009.

| Label          | Factor                                      | Note                                                                 |
|---------------|---------------------------------------------|-----------------------------------------------------------------------|
| Urbanization   | agricultural land                           | the ratio of the area of every kind of land to the total area of the land (%) |
|               | cultured land                              |                                                                       |
|               | constructive land                          |                                                                       |
|               | unused land                                |                                                                       |
| Socioeconomic | GDP per capita                              | the ratio of total regional GDP to resident population at the end of a year (yuan) |
|               | unit GDP energy consumption                 | the ratio of total GDP to comprehensive energy consumption (tce/million) |
|               | disposable income of urban residents        | the ratio of total income of the sampled residents to the number of those residents (yuan/per capita) |
| Health service | healthcare organizations                    | the ratio of the number of healthcare organizations to that of the residents at the end of a year (1/10,000) |
|               | medical practitioners                       | the ratio of the number of medical practitioners to that of the residents at the end of a year (1/1,000) |
|               | registered nurses                           | the ratio of the number of registered nurses to that of the residents at the end of a year (1/1,000) |
|               | beds in hospital                            | the ratio of the number of beds in hospital to that of the residents at the end of a year (1/1,000) |
| Environmental | green coverage                              | the ratio of the area of forest land, shrubs and all around tree to the total area of the land (%) |
| Demographic   | kindergarten                                | the ratio of kindergartens to the number of the children enrolled in kindergartens (1/1,000) |
|               | population density                          | the ratio of the residents at the end of a year to the total area of the land (persons/square km) |
| Climate       | annual average temperature                  | the ratio of the sum of daily temperature in a year to the number of days in a year (°C) |
|               | annual average relative humidity            | the ratio of the sum of daily relative humidity in a year to the number of days in a year (%) |

In order to select the factors which would be relatively significantly associated with HFMD morbidity, Pearson correlation analysis and Stepwise Regression method was adopted to remove the redundant factors. Finally we got the 6 factors as follows, disposable income of urban residents (Dis Inc), healthcare organizations (Hea Org), beds in hospital (Hos Bed), population density (Pop Den), annual average temperature (Temp) and annual average relative humidity (Rel Hum). In order to preserve the same spatial scale of the different variables, Kriging interpolation method [16,17] was adopted to realize spatial data scale conversion.
3 Methodology

Classical Linear Regression Model (CLRM)[17] assumes explanatory variable and dependent variable is independent from each other. CLRM is given by,

$$Y = a + bX + e.$$  \hspace{1cm} (1)

Where $a$ and $b$ are parameters, $e$ is a vector of i.i.d. error terms. $Y$ is a vector of observations on the dependent variable. $X$ is a matrix of observations on the explanatory variable.

Unlike CLRM approach which considers OLS (Ordinary Least Squares) Estimation, Spatial regression model usually adopts Maximum Likelihood Estimation. Spatial Error Model (SEM) includes a spatial autoregressive error term and Spatial Lag Model (SLM) includes a spatially lagged dependent variable[18]. SEM is given by,

$$Y = cX + e.$$  \hspace{1cm} (2)

$$e = dW e + u.$$  \hspace{1cm} (3)

Where $W$ is the spatial weights matrix, $e$ is a vector of spatially autocorrelated error terms, $u$ is a vector of i.i.d. errors, $c$ and $d$ are parameters.

SLM uses eigenvalues of the weights matrix and is well suited to the estimation in situations with very large data sets. Formally, SLM is given by,

$$Y = fWY + gX + e.$$  \hspace{1cm} (4)

Where $WY$ is a spatially lagged dependent variable for weights matrix $W$, $f$ and $g$ are parameters.

Fig. 2. Relationships among variables in CLRM, SEM and SLM

Lagrange Multiplier test statistics are used to suggest the alternative specification, where LM-Lag and Robust LM-Lag pertain to the spatial lag model as the alternative and LM-Error and Robust LM-Error refer to the spatial error model as the alternative. The important issue is to only consider the Robust versions of LM statistics when the standard versions (LM-Lag or LM-Error) are significant. The spatial regression specification is summarized as Fig.3. Moran’s I statistic is used to detect spatial autocorrelation and misspecifications in the model.

There are three classic specification tests[18-20] on the spatial autoregressive coefficient, where LM-Lag test is based on OLS residuals, the Likelihood Ratio (LR)
test is one of the three comparing the null model (the classic regression specification) to the alternative SLM, and the Wald test is the square of the asymptotic t-value (or, z-value).

Log-Likelihood, AIC (Akaike information criterion) and SC (Schwarz criterion) are the proper measures of fit, which are based on an assumption of multivariate normality and the corresponding likelihood function for the standard regression model. The higher the log-likelihood, the better the fit. For the information criteria, the direction is opposite, and the lower the measure, the better the fit.

![Fig. 3. Spatial regression decision procedure][18], where sig. means significant

### 4 Results

Even though LM-Lag ($z=219.3059$, $p=0$) and LM-Error ($z=183.4574$, $p=0$) statistics were both highly significant, Robust LM-Lag statistic ($z=36.1088$, $p=0$) was highly significant while Robust LM-Error statistic ($z=0.2603$, $p=0.6099$) was clearly not, SLM was finally chosen to predict HFMD morbidity according to spatial regression decision process and achieved high R square (0.82). For CLRM, Moran’s I statistic for residual spatial autocorrelation was positive and highly significant (0.4787, $p<0.01$), while for SLM the statistic was not significant (0.0297, $p>0.05$). There was still multicollinearity, abnormality of error distribution, heteroscedasticity and low R square (0.50) caused by CLRM.

Log-Likelihood (-1696.71) obtained by SLM was higher than that (-1820.62) by CLRM, AIC (3409.42) and SC (3439.29) obtained by SLM was lower than that (AIC=3655.24, SC=3681.37) by CLRM, which proved the better fitness of SLM.

Besides, the relationship among the three classic tests was as below, Wald test (1027.1350) > LR (247.8147) > LM-Lag (219.3059), indicating that LM Estimation was significant and the prediction got by SLM was reasonably effective. In SLM equation obtained,

$$ Y=0.879WY+175.619-3.675X_1-0.001X_2+0.004X_3-6.516X_4-3.665X_5-0.374X_6 $$

Where $Y$ was HFMD morbidity, $X_{1-6}$ were the selected 6 risk factors, the parameter associated with the spatial lag reached 0.879 which indicated that communities would be expected to have higher HFMD morbidity if on average, their neighbors had high HFMD morbidity. The observation and prediction map of HFMD morbidity showed as Fig. 4.
Fig. 4. Comparison of spatial distribution of the Beijing 2008 HFMD morbidity. The left plot is observed value, the middle plot is predicted value by CLRM and the right plot by SLM.

5 Discussion

The results above suggested that spatial effect included in SLM could greatly improve the performance of prediction model. Spatial correlation should be especially considered in epidemiology study. There are more and more applications indicating the prevalence of many infectious diseases such as SARS, H1N1, Bird Flu, HFMD and AIDS presented spatial correlation. Ignoring spatial effect would result in bad prediction precision and bias estimation which would both mislead the control and prevention of epidemic.

To some extent, even though Kriging interpolation method adopted in data preprocessing caused error and uncertainty in the experiment, the reliability of the results would not be pulled down. Generally speaking, the limitation will not weaken the precious value of the study, which well contributed to the application of spatial prediction model in the field of public health and infectious disease informatics.

References

1. He, Y.: Hand-foot-mouth disease (HFMD) prevention and control manual (2008 edition). J. China Medical News 23, 20–22 (2008)
2. Chen, S.C., Chang, H.L., Yan, T.R., et al.: An eight-year study of epidemiologic features of enterovirus 71 infection in Taiwan. J. The American Society of Tropical Medicine and Hygiene 77, 188–191 (2007)
3. Tseng, F.C., Huang, H.C., Chi, C.Y., et al.: Epidemiological survey of enterovirus infections occurring in Taiwan between 2000 and 2005: analysis of sentinel physician surveillance data. Journal of Medical Virology 79, 1850–1860 (2007)
4. Chang, G.-h., Lin, L., Luo, Y.-j., et al.: Sequence analysis of six enterovirus 71 strains with different virulences in humans. J. Virus Research 151, 66–73 (2010)
5. Ali, M., Emch, M., Yunus, M., et al.: Modeling spatial heterogeneity of disease risk and evaluation of the impact of vaccination. J. Vaccine 27, 3724–3729 (2009)
6. Mantel, N.: The detection of disease clustering and a generalized regression approach. J. Cancer Research 27, 209–220 (1967)
7. Cliff, A.D., Ord, J.K.: Spatial autocorrelation. Pioneer, London (1973)
8. Anselin, L.: Spatial econometrics: methods and models. Kluwer Academic Publishers, Dordrecht (1988)
9. Griffith, D.A.: Advanced spatial statistics. Kluwer Academic Publishers, Dordrecht (1988)
10. Haining, R.P.: Spatial data analysis in the social and environmental sciences. Cambridge University Press, London (1993)
11. Crighton, E.J., Elliott, S.J., Moineddin, R., et al.: A spatial analysis of the determinants of pneumonia and influenza hospitalizations in Ontario (1992-2001). J. Social Science & Medicine 64, 1636–1650 (2007)
12. Zeng, D., Yan, P., Li, S.: Spatial regression-based environmental analysis in infectious disease informatics. In: Zeng, D., Chen, H., Rolka, H., Lober, B. (eds.) BioSecure 2008. LNCS (LNBI), vol. 5354, pp. 175–181. Springer, Heidelberg (2008)
13. Wu, P.-C., Lay, J.-G., Guo, H.-R., et al.: Higher temperature and urbanization affect the spatial patterns of dengue fever transmission in subtropical Taiwan. J. Science of The Total Environment 407, 2224–2233 (2009)
14. Cao, Z., Zeng, D., Zheng, X., et al.: Spatio-Temporal Evolution of Beijing 2003 SARS Epidemic. J. Science in China Series D: Earth Sciences 53, 1017–1028 (2010)
15. Cao, Z., Zeng, D., Wang, Q.: An epidemiological analysis of the Beijing 2008 Hand-Foot-Mouth epidemic. J. Chinese Sci. Bull. 55, 1142–1149 (2010)
16. Cao, Z., Zeng, D., Wang, F.: Weather Conditions and Spatio-Temporal Spreading Risk of the Beijing 2009 Influenza A (H1N1) Epidemic. J. Science & Technology 28, 26–32 (2010)
17. Gujarati Damodar, N.: Basic Econometrics. McGraw-Hill, New York (1995)
18. Anselin, L.: Exploring Spatial Data with Geoda: A Workbook (2005)
19. Anselin, L., Florax, R. (eds.): Small sample properties of tests for spatial dependence in regression models: Some further results. Springer, Berlin (1995)
20. Anselin, L., Bera, A.K., Florax, R., et al.: Simple diagnostic tests for spatial dependence. J. Regional Science and Urban Economics 26, 77–104 (1996)