CHAPTER 13

Spatial Analysis and Statistical Modeling of 2009 H1N1 Pandemic in the Greater Toronto Area

Frank Wen and Dongmei Chen

Department of Geography, Faculty of Arts and Science, Queen's University, Kingston, ON, Canada

Anna Majury

Public Health Ontario Laboratories (Eastern Ontario), Toronto, ON, Canada Department of Biomedical and Molecular Sciences, School of Medicine, Faculty of Health Sciences, Queen's University, Kingston, ON, Canada

Au: The chapter title has been changed to "Spatial Analysis and Statistical Modeling of 2009 H1N1 Pandemic in the Greater Toronto Area" as per the Table of Contents. Please check if this is OK

13.1 INTRODUCTION

The 2009 H1N1 pandemic caused serious concerns worldwide because of the novel biological features of the influenza A virus strain, which carried genes from multiple species and resulted in high mortality/morbidity rate for youth. In 2009, Canada experienced two pandemic waves: the early-spring and the early-fall waves. A total of 40,185 laboratory-confirmed H1N1 infection cases (Standing Senate Committee on Social Affairs, Science and Technology, 2009) were reported to the Public Health Agency of Canada (PHAC), among which there were 8678 (21.6%) cases requiring admission to hospital and 428 (4.9%) deaths (Helferty et al. 2009). The number of reported cases is undoubtedly underestimated because not all infected cases were in contact with a physician nor confirmed by laboratory analysis (Standing Senate Committee on Social Affairs, Science and Technology, 2009). Based on a recent study, the newly estimated global mortality rates are more than 15 times higher

Analyzing and Modeling Spatial and Temporal Dynamics of Infectious Diseases, First Edition.
Edited by Dongmei Chen, Bernard Moulin, and Jianhong Wu.

Au: Pls check shorten RH.

than the number of laboratory-confirmed deaths reported to the WHO (Dawood et al. 2012).

In the past two decades, many methods and modeling approaches have been developed to understand the epidemic dynamics of infectious diseases on different geographic scales. Research can be found on a large or international scale (Balcana et al. 2009; Khan et al. 2009; Colizza et al. 2007; Li et al. 2011; Tatem et al. 2006), and in urban or small areas (Bian and Liebner 2007; Mao and Bian 2010; Eubank 2005; Carley et al. 2006; Xia et al. 2004; Mugglin et al. 2002). This implies that the modern transportation systems connecting densely populated cities facilitate the infection process across different geographical scales. At the urban scale, the general interested topics of these studies include clustering analyses, diffusion analyses, and disease surveillance studies. In studying these topics, GIS (Geographic Information System) and spatial statistics are combined with assisting qualitative and quantitative analysis (Clarke et al. 1996; Stevenson et al. 2008).

On the urban scale, ~~spatial~~ dynamics exhibited in ~~a~~ ~~disease's diffusion~~ in small areas include clustering effects and spatial randomness, which are largely due to rapid human movement, complex social contact patterns, and ethnic diversity. Spatial epidemiological approaches help investigate clustering effects, rule out randomness, and/or extract patterns from randomness to get a qualitative and/or quantitative view of spatial effects of disease diffusion. In such approaches, spatial influences can be interpreted in a variety of ways, depending on the particular research scenarios. For example, the spatial influence in gravity models (Xia et al. 2004) is measured by the distance-based attractiveness between the origin location and the destination location, while it is measured by the adjacency matrix in a spatial regression model to incorporate the random spatial influence (Mugglin et al. 2002). This descriptive difference implies a diversity of data-driven epidemiological studies that have different requirements for data and other prior knowledge. In practice, data quality and availability remains a bottleneck in research processes. On the urban scale, given the limited data and large number of unreported cases, spatial statistical methods and models have advantages; they provide statistical inference and they depict clustering effects and their causes regardless of the lack of information about underlying diffusion processes.

Urban scale is crucial for studying influenza pandemics, through which spatial dynamics estimation can lead to the discovery of clustering effects as well as their causes in small areas. Small area disease risk estimation for the urban scale has been an active research topic in spatial epidemiological studies (Cressie 1993; Walter 2000; Eubank et al. 2004; Lawson 2008; Martínez-Beneito et al. 2008). The degree of challenge in exploring spatial dynamics is exaggerated by the scarcity of data and by the substantial randomness of infection on the urban scale. There is a high demand for an ~~immediately~~ applicable methodology that can make the required estimations. To date, a methodology that can estimate the spatial dynamics and randomness for an acute infectious disease using very scarce data is either absent or incomplete.

This study intends to use stepwise spatial statistical analysis to analyze the spatial clustering effect and to estimate the impact of spatial randomness and develop spatial statistical-based modeling methods that can incorporate the impact of spatial autocorrelation and spatial randomness when the disease data are limited. In particular, this study seeks to discover whether the spatial Generalized Linear Mixed Model (GLMM) has better predictability (i.e., modeling fitting results) than a nonspatial GLMM by incorporating the random spatial effect~~s~~ in the modeling.

13.2 STUDY AREA AND DATA

The ~~GTA~~, with a population of 5.5 million, is one of the most multicultural regions in the world. This results in a regional diversity with a complex ethnic pattern. Based on the data released by Statistics Canada in 2006, there are 108 ethnic origins and more than 20 distinct predominant home languages in the GTA.

According to the 2006 census data, the GTA includes 1003 census tracts (CTs). These CTs are contiguous inland administrative regions except the Toronto Islands (~~as marked in purple in~~ Figure 13.1) where no H1N1 infections were recorded. Among 1003 CTs, there are five CTs covered with the conservatory land; thus, their population is zero. ~~The GIS data used in this study include~~ ~~the~~ administrative boundaries of the CTs~~.~~ ~~The boundary polygon data~~ are used to build the neighborhood weight matrix that is required by the ICAR model. The geometry centers of the polygons are used to generate a separate point layer. Individual disease records are aggregated to each point in this point layer for each CT. The demographic information (i.e., sex and age) is extracted from the Census 2006 database of the GTA and added to the point data file.

There were two H1N1 epidemic waves experienced in 2009. The first wave took place from April to August, followed by a second wave from September to December

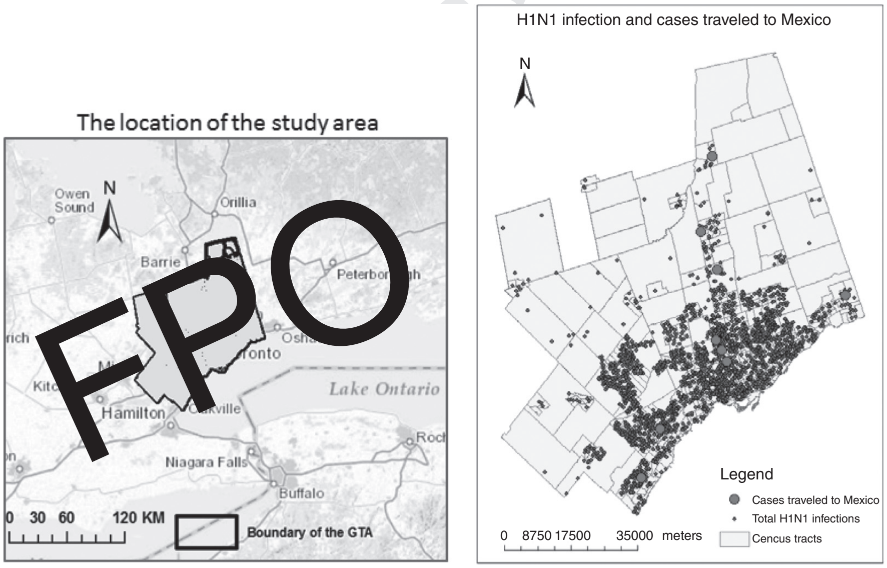


Figure 13.1 The study area (left) and the distribution of H1N1 infection (right) in GTA

in 2009. The amplitude of the peak of the first wave is larger than that of the second wave, implying that the first wave was more severe. The disease surveillance data used in this study were derived from testing data through Public Health records through a data-sharing agreement. The data set contains 3722 individual infection records from April 1, 2009, to December 31, 2009, and involves all the CTs except the Toronto Islands and the five census tracts without residents. The attributes provided in the individually based H1N1 data include spatial locations in postcode, infection time, demographic information including sex and age, city names, and information on traveling to Mexico, etc. The data do not have social contact information and ethnic characteristics of the infected individuals. There are missing values for almost every attribute. For example, there are 57 missing values (marked as unknown in the data) for the sex attribute. In the age–sex stratification calculation conducted later, these records are excluded from the calculation because, within the context of this study, there are no quantitative methods that can be used to retrieve the missing values for this categorical data. For the missing values of age, this study uses the mean of the ages of males or females.

The original data used in this study are individually based H1N1 infection records, where the location information is recorded as six-digit postcodes. By linking the individually based disease data to the postcode-coordinate lookup table in ArcGIS, the geographic coordinates of infected persons are obtained. After the geocoding, the accumulated cases are aggregated as count attributes of the 1003 census tracts. Thus, each CT has an aggregated count that presents the basic severity status of the CT in the pandemic. The individual-based disease data includes nine records of infections of those who have ever traveled to Mexico. These records appeared in April, 2009, when the H1N1 pandemic began. After April, there are no recorded cases that traveled to Mexico. One fact of influenza is that the virulence of the flu is extremely strong at the beginning of the pandemic, and it ~~will~~ become weaker as the genetic sequence and mutations carry on for the disease. Therefore, it is reasonable to have a suspicion that the nine infections may have had stronger transmissibility than other infections, which may have contributed to clustering of the disease diffusion locally and globally. The locations of the nine infections are mapped in Figure 13.1 as a layer over all the other records to observe the pattern.

13.3 ANALYSIS METHODS

In order to develop appropriate models to estimate the H1N1 risk, a two-step analysis process was used. First, H1N1 data and its potential demographic covariates were explored to understand the disease profile and the level of spatial autocorrelation. The second step is to estimate the disease prevalence risk using modeling approaches based on the results from the previous step.

13.3.1 Initial Data Exploratory Analysis

The histogram statistic and Quantile–Quantile (Q–Q) plots were created to test whether the disease and demographic data are normally distributed with the

HISTOGRAM and QQPLOT statements provided in the Proc UNIVARIANT procedure in SAS. A histogram is a representation of a frequency distribution based on the selected interval given the range of values being analyzed. In UNIVARIANT, a histogram can be generalized by the Kernel Density Function (KDF) that draws the smoothed data distribution lines. Q–Q plots provide graphs and test statistics suggesting the likely distribution of both continuous and discrete data. These plots are made by plotting the count values or their logarithmic values against their normal percentiles calculated by the Proc UNIVARIANT procedure.

The GTA area includes 19 sub-cities. The city name, indicating where the infection took place, is one of the attributes recorded in the H1N1 individual based data. This information can be used to estimate whether or not there are spatial clustering effects. Scatter plotting the aggregated counts, age, and gender against the cities provides an initial perception of the clustering patterns.

Global spatial autocorrelation measures the second-order spatial clustering effects caused by the spatial processes. Local spatial autocorrelation measures the first-order spatial clustering effects caused by the similar infection rate in adjacent local neighborhoods (Ord and Getis 2001). In previous studies, the connection between the model selection and the existence and significance of spatial autocorrelation is less discussed or absent. To address this issue, this study illustrates the existence and significance of both global and local spatial autocorrelations before the decision making on model selection. In addition, this requirement also comes from the context of the study area and data. In metropolitan areas such as the GTA, there are different factors or randomness that may create pseudo spatial autocorrelations. Thus, the results, that is, levels and significances of spatial autocorrelation tests, need to be reviewed with the prior knowledge of the study.

In spatial autocorrelation tests, the null hypothesis is first made that observations are generated from random processes for both global and local spatial autocorrelations (Cressie 1993). The levels and significance of the spatial autocorrelation are used to test the null hypothesis by examining the threshold value. The spatial autocorrelation functions include statistical significance tests that validate whether or not the null hypothesis can be rejected. The Moran's *I* statistic (Moran 1950), the most common statistic for estimating global spatial autocorrelations, was used in this study. The global Moran's *I* scores and their *Z* scales were calculated in GeoDa (Anselin et al. 2006).

13.3.2 Modeling the H1N1 Infection Risk

Based on results gained in the exploratory data analysis and spatial autocorrelation analyses, the GLMM model incorporating the ICAR model for the random effect is used to model the H1N1 infection risk. In order to estimate the H1N1 infection risk in each CT, the expected cases should be modeled. The statistical estimation of relative risks can be seen as a departure from the expected cases. This study uses a deterministic indirect standardization method to calculate the expected H1N1-infected cases in each of the CT using a stand-alone program written in the Python script language. The program first processes the data and calculates the expected infection number for each age–sex group, then computes the expected cases for each CT as specified in the previous chapter.

In the implementation of this method, first it takes 5 years as the interval to divide age groups of both males and females. The maximum age involved in the estimation is 90, so the division results in 18 age groups for both male and female, a total of 36 age–sex groups. Next, for each age–sex group, it calculates: (1) the total infected number using the individual-based records, (2) the total population in the GTA, and (3) the total population in each CT. The infection ratio for each age–sex group can be computed using the total infected number in the age–sex group divided by the population of the group. Then the expected cases in each CT can be calculated as the sum of the expected cases of each age–sex group in the CT:

$$E_j = \sum_{i=1}^{36} P_{ij} \frac{\sum_{j}^{1002} O_{ji}}{\sum_{J}^{1002} P_{ji}},$$

where

i and j stands for the age–sex group and census tract area respectively;
E_j is the expected cases in the j census tract;
P_{ij} is the population number of the i group in the j census tract;
O_{ji} is the observed numbers of H1N1 in the j census tract and the i group.

The ICAR model is used to model the random spatial effect. In this study, through testing the hypothesis, incorporating the randomness component into the modeling is expected to improve the modeling predictability. The area-specific spatial randomness effect is modeled by the ICAR model employed in the GLMM.

The basic working mechanism of the GLMM is given here as a preparatory introduction in order to understand the implementation of the model provided in this chapter. Given the observations z_i, with the associated vector γ of random effects, a GLMM model can be denoted as

$$E([z_i]) = l^{-1}(X\beta + Z\gamma + e)$$

where

$l(\cdot)$ is the link function and $l^{-1}(\cdot)$ is its inverse function;
X is the covariate matrix for fixed effects;
β is the coefficients of the covariates;
Z is the covariate matrix for random effects;
γ represents the random effect;
e is the stand-alone randomness.

In a GLMM, random effects γ have a normal distribution with a mean of 0 and a variance matrix of G. The error e has a normal distribution with a mean of 0 and a variance of R. The Z matrix and covariance matrix G need to be specified to model the G-side randomness effects. The R-side randomness can be modeled by specifying the independent covariance structure R. In practice, model programmers need to specify the columns in the fixed effects matrix and random effect matrix Z, and construct the corresponding covariance matrixes G and R.

13.4 THE IMPLEMENTATION OF THE GLMM AND ICAR

The ICAR and GLMM models were implemented using the Statistical Analysis System (SAS) programming language (SAS 2005, 2008). The implementation starts with an initialization step that is the neighboring information extraction for each CT in the administrative boundary polygons. The extraction produces a neighbor matrix table that is used in the ICAR model implementation.

The primary task in implementing the ICAR model is actually to construct the precision matrix $(I - W)D^{-1}$ that is the reverse matrix of the covariance matrix $(I - W)^{-1}D$ ~~introduced in the foregoing literature review~~. The neighboring weight matrix W in the ICAR model is confined by the neighboring structure defining the areas that share a common border with the area of interest. After the precision matrix is built, it then can be incorporated in the GLMM implementation as a random component. The last step is modeling fitting using the PROC GLIMMIX procedure in SAS.

Au: "Proc GLIMMIX" has been edited as "PROC GLIMMIX" and we hope this is OK

The neighbor matrix extraction is done by a stand-alone data-processing tool implemented using C++ programming language. The neighbor extraction tool works as follows: (1) reads the input administrative boundary polygon file; (2) parses every geographical unit in the file; (3) creates the multiple arrays that include the CT and its neighbors that share boarder(s) with it, (4) outputs the arrays in the neighbor matrix table that is a text file in the format required by building the precision matrix. The condition for being a neighbor polygon of each other is that they share at least a border line, not a single point. This condition is made according to the specification of the Poly2nb () function in R. The results generated from this tool are verified by comparing with manually selected neighbors of multiple CTs.

Using the neighbor conjunction structure of the administrative boundaries of the CTs, the neighbor matrix is constructed. To estimate the spatial area-specific random effect b_i, the focus is on estimating the spatial conditional variances σ^2. In the precision matrix $(I - W)D^{-1}$, since the diagonal matrix D has entries $1/n_i$, the precision matrix yields a rather simple form in the calculation, which has the number of neighbors n_i on the diagonal and -1 for the entries of the neighbors of CTi.

The interactive matrix language (IML) of SAS is used to build the precision matrix $(I - W)D^{-1}$. The GINV() function provided in the IML package is used to generate the inverse precision matrix, that is, the Moore–Penrose generalized inverse matrix that satisfies the criterion of the precision matrix in the ICAR model being semi-definite. In addition, IML does not have a built-in function to produce the rank of the precision matrix. Instead, the function round() is used to rank the entries of the matrix.

In the stated methods above, the expected cases of H1N1 infections in each CT are calculated based on sex and age stratification. In statistical models used to estimate relative risks of the diseases, the estimation of relative risks is commonly treated as a departure from the expected cases, which is actually adapted in this study and explained in the following discussion.

In a CTi, if O_i is denoted as the observed count, λ_i is denoted as the relative risk of infecting H1N1, and E_i is the expected number of the H1N1 infections, conditional on λ_i, the observed counts of H1N1 infections are independent Poisson variables with mean $E_i\lambda_i$, which implies:

$$O_i \sim Poisson(E_i\lambda_i)$$

where

$$\lambda_i = \exp\{\alpha + b_i\}$$

exp{ } is the exponential expression;
α is the base (log) nonspatial random risk of being infected; and
b_i is an area-specific spatial random effect capturing the overall spatial variance of the relative risk (log) of the disease in census tract i.

A spatial GLMM includes the intercept that is the log value~~s~~ of the expected H1N1 infection number E_i, the non-spatial random risk α, and the spatial random effect b_i. This spatial form can be configured into a nonspatial model that just incorporates the intercept and the non-spatial random effect. The nonspatial and spatial GLMMs are fitted, respectively, to decide which one can provide a better predictability.

The Proc GLIMMIX procedure provided in SAS can fit the statistical models with data that have correlations or nonconstant variability to estimate spatial and temporal trends. In the absence of random effects, it treats the model as GLMM and fits the model using the GENMOD Proc that provides the maximum likelihood function estimation. In the presence of structured random effects (correlations), the Proc GLIMMIX applies a residual pseudo-likelihood function to estimate the model fitting.

The two types of random effects, namely G-side and R-side random effects, are distinguished by the RANDOM statement in the Proc GLIMMIX. The RANDOM statement specifies the R-side random effect with either _RESIDUAL_ or RESIDUAL options. Similarly, it specifies the G-side random effect with a range of options including LDATA, TYPE, and LIN (q). The LDADA option specifies a data set that has the coefficient matrices. The TYPE option specifies the covariance structure. The LIN (q) option specifies a general linear covariance structure with q parameters. In the implementation SAS code, the LDATA specifies the input data set that is the covariance matrix created before the model fitting. The TYPE specifies the structure of the covariance matrix, and the LIN (1) specifies the parameter for the covariance structure, which is the conditional variance σ^2 presented in the ICAR model.

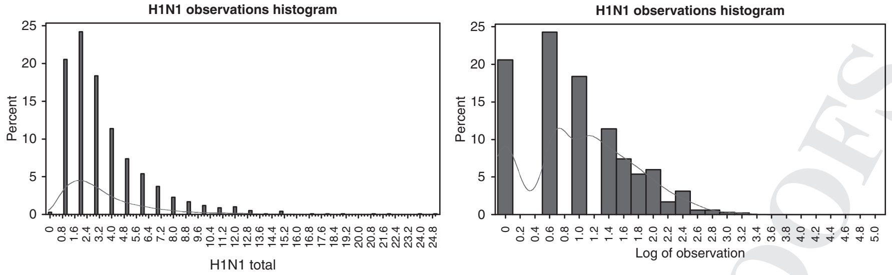


Figure 13.2 H1N1 infections (left) and its logarithm (right) histograms

13.5 RESULTS

13.5.1 Initial Exploratory Data Analysis

As shown in the frequency histogram (Figure 13.2), the distribution of the H1N1 aggregated counts shows a skewed shape with a long right tail. In the Quantile–Quantile (Q–Q) plot (Figure 13.3 left), the distribution of the H1N1 aggregated accumulated counts have an obvious departure from a normal distribution, and there is a clear staircase pattern of plateaus and gaps, suggesting that the data are discrete. Compared to the Q–Q plot in Figure 13.3 (right), the Q–Q plot created using logarithmic values of the H1N1 counts shows an improved approximation of the plotted plateaus to the reference line. By observing the two Q–Q plot figures in Figure 13.3, it can be concluded that the discrete exponential distribution fits the data more accurately. The most commonly used discrete exponential distributions

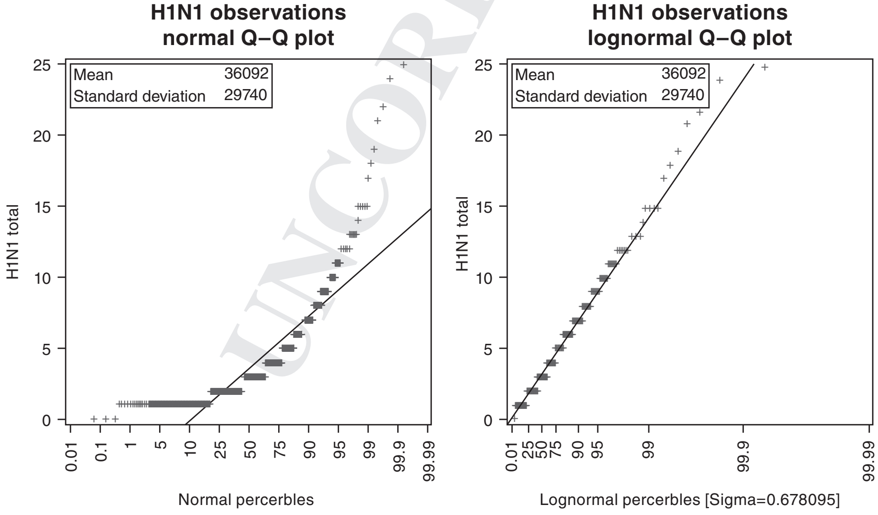
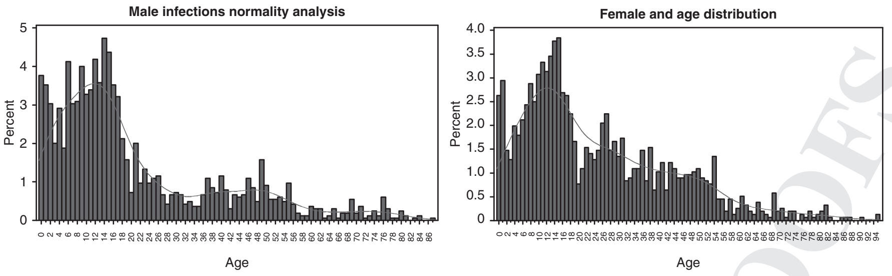


Figure 13.3 H1N1 infections Q–Q plot for normal and lognormal distributions



Au: The artwork will appear in black and white. Please modify the artwork and the caption to avoid use of colours.

Figure 13.4 Male (left) and female (right) infections histogram with kernel density estimation (~~red~~ line)

are Poisson distribution and Negative Binomial (NB) distribution. According to the previous studies' suggestion that the Poisson distribution should be the starting point for data analysis, the Poisson distribution is chosen to be used in the modeling.

Histograms are also created with count values to illustrate the distribution of male and female age groups respectively. In both histograms, the kernel density lines indicate departures from normality in the age distributions of both groups. For the male infections (Figure 13.4), the most infectious age group is the groups under the age of 18. The other noticeable infectious age groups appear between the ages of 26 and 50. Seniors (above 60) exhibit very low infectious possibility. For females, infection levels are relatively strong under the age of 55. The most obvious difference between male and female infections is that between the ages of 20 and 50, the female infection rate is far more severe than male. This may imply that the women in the indicated age range had more regular social contact than men, which needs to be studied further. The results above clearly indicate that a stratification method is required to incorporate the heterogeneity in different age–sex groups in the study area.

Figure 13.5 is the scatter plot of the aggregated counts, age, and gender for each sub-city of GTA. It is evident from Figure 13.5 that high disease count values are around Scarborough, Oakville, King City, and Richmond Hill, which means that the HINI prevalence potentially has either a local or a global clustering effect that needs to be identified with further analysis.

The global Moran's I function in GeoDa yields a relatively low Moran's I index of 0.21, indicating a very weak global spatial autocorrelation. Therefore, the global autocorrelation impact is excluded in the consideration of modeling selection.

The local spatial autocorrelation estimation involves the calculation of the local cluster index, that is, *z*-score, and the significance, that is, *p*-score, for each geographical unit. For a particular geographical unit, when the local cluster index is higher than the threshold and statistically significant, the null hypothesis can be rejected for the geographical unit. The *z*-scores generated from the Getis–Ord Gi* function is shown in Figure 13.6.

Local spatial autocorrelation results from the Getis–Ord Gi* function reveals the existence of the local spatial clustering effects. Additionally, the hot spots produced in Figure 13.6 seem irrelevant to the cases of travelers to Mexico (Figure 13.1 right) suggesting that the cause of the infection clusters is not the geographical distance

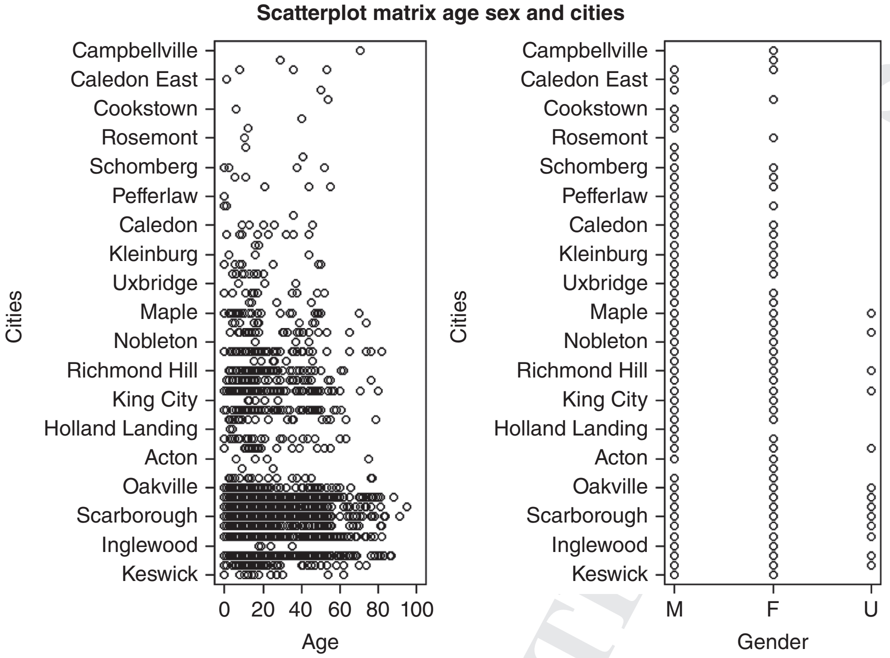


Figure 13.5 City, age, and gender plot of H1N1 infections of 2009 in the GTA

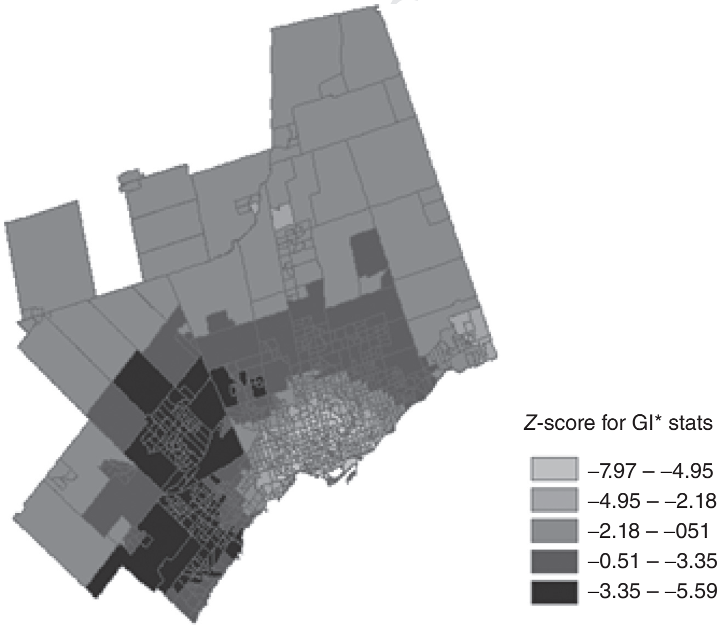


Figure 13.6 Score of local spatial autocorrelation measure Gi*

Table 13.1 Estimated Parameters and Fit Statistics of the Nonspatial GLMM and Spatial GLMM

	Nonspatial GLMM	Spatial GLMM
−2 Res Log Pseudo-Likelihood	2657.73	2556.43
Generalized chi-square	1579.76	1501.27
Fix intercept	0.1002	0.08587
α	0.1382	0.07703
σ^2		0.1610

to initial H1N1 sources, but other factors that require further research. Since the traveling behavior can be considered as the randomness in the context of this study, the potential global spatial autocorrelation may be created by the randomness. The overall local spatial diffusion pattern is likely related to other factors in the study area, such as the distribution of the transportation system and population.

Based on the results, the local spatial autocorrelation is the factor that is further considered in later statistical modeling. The spatial modeling intends to quantitatively study the cause of the clustering. The most interested variable in explaining the cause in this study is the latent spatial variable that models the influence of the spatial structure of the GTA as a random effect.

Au: Please clarify "Res Log" in its first instance of use, if deemed necessary.

13.5.2 Model Fitting Results

Table 13.1 summarizes the estimated parameters and fit statistics of the nonspatial GLMM and spatial GLMM models incorporated the ICAR model. As presented in Table 13.1, comparing the −2 ~~Res Log~~ Pseudo-Likelihood statistics (the lower, the better) of the spatial and nonspatial models, the result of the spatial model shows mild, but clear improvement of the overall modeling performance, which suggests that the spatial model provides a better fit for the H1N1 data. At the same time, the comparison also shows that the CAR model is ~~an optimal~~ solution for the local spatial autocorrelation caused by the first order spatial effects. The parameter values (Table 13.1) indicate that the fix intercept, that is, the nonspatial randomness variance parameter α, is reduced (smoothed) after the spatial randomness variance σ^2 is used in the spatial GLMM model, which implies that the overall uncertainty of the nonspatial GLMM is explained by the spatial GLMM with the incorporation of the CAR model.

The results of the spatial autocorrelation tests have revealed the existence of the local spatial autocorrelation, which confirms the hypothesis. Model fitting results showed that the spatial GLMM has better predictability compared with that from the nonspatial GLMM. The above results verified the validity of the procedures undertaken in the data processing and exploratory analysis.

13.6 DISCUSSIONS AND CONCLUSION

The spatial dynamics of the 2009 H1N1 pandemic comprise the local clustering effect and the random effect inherited from the geographical contiguous structure of the GTA. The clustering effect and the randomness contributed by the geographical contiguous structure are confirmed respectively by the spatial autocorrelation analysis and the spatial statistical modeling. According to the spatial autocorrelation results, the local spatial autocorrelation presented the significant impact of the pandemic in the study area, while the global spatial autocorrelation has no significant impacts. The causes of the local spatial autocorrelation are partially explained by model fitting results of the spatial GLMM that incorporated the ICAR model. The better predictability obtained from the spatial GLMM suggests that the spatial structure and the associated underlying population processes contribute to the spatial dynamics presented in the study area.

Au: In the sentence "As presented in (. . .) non spatial GLMM", "Table 4" has been changed to "Table 13.1." Please check if it is fine.

As presented in Table 13.1, the −2 Res Log Pseudo-Likelihood estimation is 2556.43 from the spatial GLMM, and is 2657.73 from the nonspatial GLMM. The lower value of the −2 Res Log Pseudo-Likelihood estimation indicates a better fitting. Therefore, it is clear that the spatial GLMM has better predictability. This modeling improvement can theoretically be attributed to the Gauss–Markov properties of the ICAR model, which specifies the mean and variances ~~described in $\{Z_i|Z_{-i}, \sigma^2, i = 1, 2, \ldots, n\}$~~ that captures the local spatial homogeneity. As the GLMM incorporates the ICAR model in a linear relative risk model λ_i, it provides a statistical framework that can be used in investigating the causes of the spatial autocorrelation by expanding the linear model λ_i.

The primary limitation of the study was reflected in the degree of the improvement made in the model fitting. As shown in Table 13.1, the ~~pseudo-likelihood estimation result is 2556.43 for the spatial model, and 2657.73 for the nonspatial model. The~~ mild difference between the two results indicates that the improvement made by applying the spatial model is not significantly large. The small improvement is most likely due to the fact that the Poisson distribution designated in the modeling cannot provide the most suitable distribution for count data, which commonly results from data over-dispersion. The problem of over-dispersion of count data typically occurs when the observations are correlated or collected from clustered regions (Wakefield 2007). Modified models may resolve the issue by providing more generalized parameterization for Poisson distributions. For example, in SAS, Proc GMOD and Proc MIX provide over dispersed Poisson distributions. However, these methods do not incorporate spatial randomness effects. Current spatial statistical packages are rarely seen with integral solutions taking into account both the data over-dispersion and spatial dependency effects. In practice, the negative binomial (NB) distribution was also tried out in this study intending to reduce the over-dispersion effect. However, with the NB distribution, the GLIMMIX encountered problems of nonconvergence, which resulted in an unworkable path.

Another limitation of the methodology designed in this study is that it does not provide the capability of estimating the temporal characteristic of the pandemic. The local clustering effect of this study is a result based on accumulated counts. This result is not verified in the temporal domain. Namely, in a discrete time frame, the local spatial clustering may not be detectable. This topic should be addressed in a future study.

To solve the data scarcity and over-dispersion problems that impact the model performance, it is recommended to use Bayesian statistic models to carry out a comparative study. The CAR model can be incorporated into a Bayesian model as a prior probability distribution for the post-distribution of the disease predictor (Thomas et al. 2004). In Winbug, the CAR model is available as a built-in function. There is an argument over whether or not prior distributions used in Bayesian models have efficiency and validity. However, comparative studies on Bayesian analysis and conventional statistical inference suggest that Bayesian methodology is an advanced alternative to the conventional frequentist approach for epidemiological studies. Frequentist theory often depends on a large number of repetitions; in contrast, epidemiological data are commonly seen as having much fewer replications. Bernardinelli et al. (1995) state that when the disease intensity records are rare and/or the geographical grid is fine, Poisson variation applied to the area-specific small counts may cause biased estimation.

This study suggests that changing the neighbor matrix in the CAR model may improve the overall performance further. The conventional spatial regression model uses spatial weight matrices to account for spatial interactive effects. The spatial weight matrix can be calculated based on contiguous neighbors, as in this study, or the neighbors included in a given distance. In future studies, the model predictability may be improved by altering the neighborhood definition of the spatial weight matrix. Tools for processing and calculating different weight matrices are available in various software packages. For example, the spatial matrix function POLY2NB in R can be used to build such a matrix with different neighboring strategies.

Another improvement that can be made to this methodology in future studies is enhancing the models with analysis capability in the temporal domain. The GLMM model used in this study can be extended to a spatial temporal model by incorporating a temporal auto-regressive model that is available in the GLIMMIX, which forms a more comprehensive model to investigate temporal influences.

The spatial epidemiological models are data driven, therefore, their applicability largely depends on data availability and quality. To date, the surveillance systems and associated technologies are not capable of providing fully satisfied data quality required by the models. Missing values and data scarcity are the primary factors that may cause bias in analysis results. In scenarios with limited data, it is found that combining spatial exploratory analysis and spatial statistical modeling is an effective approach that can incorporate the most relevant prior knowledge to yield better results. Exploratory analyses extracte~~d~~ demographic information from the data, and generate~~d~~ expected cases for each of the CTs. A complete spatial exploratory analysis reveale~~d~~ both global and local spatial cluster effects and identifie~~d~~ their comparative significances in the context of traveling information contained in the data. Using GIS data that describe the contiguous neighbor structure of the study area, the spatial GLMM shows a clear overall modeling improvement over its nonspatial version.

REFERENCES

Anselin L., Syabri I., and Kho Y. (2006). GeoDa: an introduction to spatial data analysis. *Geographical Analysis*, 38(1):5–22.

Balcana D., Colizzac V., Goncalvesa B., Hud H., Ramascob J. J., and Vespignani A. (2009). Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences of the United States*, 106:51.

Bernardinelli L., Clayton D. and Montomoli C. (1995). Bayesian estimates of disease maps: how important are priors? *Statistics in Medicine*, 14(21–22):2411–2431.

Bian L. and Liebner D. (2007). A network model for dispersion of communicable diseases. *Transactions in GIS,* 11(2):155–173.

Carley K. M., Fridsma D. B., Casman E., Yahja A., Altman N., Chen L. C., Kaminsky B., and Nave D. (2006). BioWar: scalable agent-based model of bioattacks. *IEEE Transactions on Systems, Man and Cybernetics A*, 36(2):252–265.

Clarke K. C., McLafferty S. L., and Tempalski B. J. (1996). On epidemiology and geographic information systems: a review and discussion of future directions*. Emerging infectious diseases*, 2(2):85–92.

Colizza V., Barrat A., Barthelemy M., Valleron A.-J., and Vespignani A. (2007). Modeling the worldwide spread of pandemic inuenza: baseline case and containment interventions. *Public Library of Science (PLoS) Medicine*, 4(1):e13+.

Cressie N. (1993). *Statistics for Spatial Data*. New York, NY: John Wiley & Sons, Inc.

Dawood F. S., Iuliano A. D., Reed C., Meltzer M. I., Shay D. K., Cheng P., Bandaranayake D., Breiman R. F., Brooks A., Buchy P., Feikin D. R., Fowler K. B., Gordon A., Hien N. T., Horby P., Huang S., Katz M. A., Krishnan A., Lal R., Montgomery J. M., Mølbak K., Pebody R., Presanis A. M., Razuri H., Steens A., Tinoco Y. O., Wallinga J., Yu H., Vong S., Bresee J., Widdowson M. (2012). Estimated global mortality associated with the first 12 months of 2009 pandemic influenza A H1N1 virus circulation: a modelling study. *The Lancet Infectious Diseases*, 12(9):687–695.

Eubank S., Guclu H., Anil Kumar V. S., Marathe M. V., Srinivasan A., Toroczkai Z., and Wang N. (2004). Modelling disease outbreaks in realistic urban social networks. *Nature*, 429(6988):180–184.

Eubank S. (2005). Network based models of infectious disease spread. *Japanese Journal of Infectious Diseases*, 58(6):9–13.

Helferty M., Vachon J., Tarasuk J., Rodin R., Spika J., and Pelletier L. (2009). Incidence of hospital admissions and severe outcomes during the first and second waves of pandemic (H1N1) 2009. *Canadian Medical Association Journal*, 182(18):1981–1987.

Khan K, Arino J, and Hu W et al. (2009). Spread of a novel influenza A (H1N1) virus via global airline transportation. *The New England Journal of Medicine*, 361:212–214.

Lawson A. B (2008). *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology*. Boca Raton: The Chemiclal Rubber Company (CRC) Press.

Li X., Tian H., Lai D., and Zhang Z. (2011). Validation of the gravity model in predicting the global spread of influenza. *International Journal of Environmental Research and Public Health*, 8:3134–3143.

Mao L. and Bian L. (2010). Spatial-temporal transmission of influenza and its health risks in an urbanized area. *Computers, Environment and Urban Systems*, 34(3):204–215.

Martínez-Beneito M. A., Conesa D., López-Quílez A., and López-Maside A. (2008). Bayesian Markov switching models for the early detection of influenza epidemics. *Statistics in Medicine*, 27(22):4455–4468.

Moran P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2): 17–23.

Mugglin A. S., Cressie N., and Gemmell I. (2002). Hierarchical statistical modelling of influenza epidemic dynamics in space and time. *Statistics in Medicine*, 21(18):2703–2721.

Ord J. K. and Getis A. 2001. Testing for local spatial autocorrelation in the presence of global autocorrelation. *Journal of Regional Science*, 41(3):411–432.

SAS Institute Inc. (2005). *SAS® 9.1.3 Language Reference: Concepts*, 3rd edition. Cary, NC: SAS Institute Inc.

SAS Institute Inc. (2008). *SAS/STAT® 9.2 User's Guide*, Cary, NC: SAS Institute Inc.

Stevenson M., Stevens K. B., Rogers D. J., and Clements A. C. A. (2008). *Spatial Analysis in Epidemiology*, 1st edition. Oxford University Press.

Tatem A. J., Rogers D. J., and Hay S. I. (2006). Global transport networks and infectious disease spread. *Advances in Parasitology*, 62:293–343.

Thomas A., Best N., Lunn D., Arnold R., and Spiegelhalter D. (2004). GeoBUGS User Manual Version 1.2. Medical Research Council Biostatistics Unit, Cambridge University.

Wakefield J. (2007). Disease mapping and spatial regression with count data. *Biostatistics*, 8(2):158–183.

Walter S. D. (2000). Disease mapping: a historical perspective. In: Elliott P Wakefield J. C. Best N. G. and Briggs D. (editors) *Spatial Epidemiology: Methods and Applications*. Oxford University Press. pp. 223–239.

Xia Y. C., Bjornstad O. N., and Grenfell B. T. (2004). Measles metapopulation dynamics: a gravity model for epidemiological coupling and dynamics. *American Naturalist*, 164(2):267–281.