WestVirginiaUniversity
THE RESEARCH REPOSITORY @ WVU

2002

A Spatial Logit Association Model for Cluster Detection

Ge Lin

Follow this and additional works at: https://researchrepository.wvu.edu/rri_pubs

Part of the Regional Economics Commons

Digital Commons Citation
Lin, Ge, "A Spatial Logit Association Model for Cluster Detection" (2002). *Regional Research Institute Working Papers*. 141.
https://researchrepository.wvu.edu/rri_pubs/141

This Working Paper is brought to you for free and open access by the Regional Research Institute at The Research Repository @ WVU. It has been accepted for inclusion in Regional Research Institute Working Papers by an authorized administrator of The Research Repository @ WVU. For more information, please contact researchrepository@mail.wvu.edu.

Ge Lin

A Spatial Logit Association Model for Cluster Detection

In this paper, I propose to set out a logit spatial association model for binary spatial events and develop a scan algorithm to search for spatial associations. I extend the traditional logit model with a spatial autocorrelated component so that the model includes not only known risk factors, but also spatially autocorrelated regions as control or explanatory factors. The case study of West Virginia lung cancer shows that the model effectively captures cool and hot spots in lung cancer mortality.

1. INTRODUCTION

Categorical models in spatial association analysis differ from traditional spatial association models in several important ways. First, categorical data analysis is based on frequencies or counts; hence, the number of observations in each cell in a multiway table is an important factor in determining goodness-of-fit statistics. In calculating Moran's I for crime rates, for example, it does not matter if an observed rate is 1/1000 or 100/100,000 as long as the rate is one per thousand, but these rates are likely to make some difference in a categorical statistic, such as the chi-squared test, because it evaluates the deviance between the observed and expected frequencies (Agresti 1990; Raftery 1995). Second, traditional spatial statistics are typically affected by values from adjacent or nearby areas, and they often ignore the diagonal elements in the spatial weight matrix, suggesting that deviations between observed and expected values in each of the regions are ignored (Rogerson 1999). A spatial categorical model, by the nature of model estimation, would always include spatial (adjacent areas) and non-spatial (diagonal) components. Third, since multiway frequency tables are the basis for categorical data analysis, covariates, such as sex or age groups can be incorporated as parts of a categorical model testing process (Diggle 2000). The traditional spatial autocorrelation statistical tests, in contrast, are not designed to include covariates. It is, therefore, desirable to have a set of spatial statistical tests for categorical data analysis.

The author would like to acknowledge the generosity of Dr. Daniel Griffith, who provided a list of references that helped to guide this research project. The helpful comments from Peter Rogerson, James LeSage, as well as comments and editorial suggestions from anonymous reviewers are greatly appreciated.

Ge Lin is an assistant professor in the Department of Geography, West Virginia University. E-mail: glin@wvu.edu.

Geographical Analysis, Vol. 35, No. 4 (October 2003) The Ohio State University
Submitted: June 4, 2002. Revised version accepted: February 25, 2003.

Even though statistical analysis for categorical data was introduced to geography in the late 1970s, its applications to spatial associations and cluster analyses are few. The logit model—a special case of the categorical model—is a case in point. Besag (1972) proposed a theoretical model of space-time autologistic regression for point data, in which a logit of a given attribute at a spatial point is modeled as a conditional probability of point attributes nearby. Haining (1982; 1983) introduced this model to geography by reanalyzing some classical data in Hagerstrand (1967). However, due to the complicated estimation method at the time and the lack of a goodness-of-fit statistic, Besag's conditional autologistic regression, to my knowledge, has never been used in practical applications in geography. Moreover, this model cannot be directly applied when specific point locations are unknown, in which case spatial analysts often resort to dummy variables at an aggregated spatial unit to capture regional effects (Senior, Williams, and Higgs 1998; Lin 2000).

Early work by Fingleton (1983a; 1983b) sparked an opportunity for extending the traditional logit model to spatial logit models using aggregated data. With an emphasis on introducing loglinear models, Fingleton fitted the classical Lansing Woods point data for the presence and absence of oak or hickory trees on a grid. In the model fitting process, Fingleton employed the concept of spatial autocorrelation and included the observed and expected numbers of oak trees within a number of grid cells to adjust for the chi-squared test. However, since the adjustment is only made at the global level, the test cannot be used to identify spatial clustering. In addition, in order to properly adjust for the chi-squared statistics, the model requires pre-testing the spatial relationship between distance range and the existence of a given tree species similar to determining distance range in variogram estimation. Nevertheless, the test represented a significant advance in attempting to bring spatial information into the modeling process. Comparing the small but significant step of Fingleton and relatively steady progress in the development of traditional spatial association models (e.g., autocorrelation), Wrigley (1985, 307) commented on the need to narrow the apparent knowledge gap between traditional autocorrelation research and new methods for categorical data analysis.

Recently, there has been some progress in bridging this gap. Dubin (1995; 1997) specified and developed an estimation routine for a logistic model with spatial dependence. To model firms' behaviors for adopting innovations, Dubin used spatial distance between firms as an information matrix to capture the spatial dispersion process. Like a linear regression with a spatial lag, Dubin's model is likely to correct model bias when explaining firms' behaviors, but it cannot reveal location-specific spatial association. More recently, Leyland et al. (2000) tested spatial effects of event data (deaths) within a multilevel modeling framework (see Best, Ickstadt, and Wolpert 2000 for a Bayesian version of the model). Based on a log-linear model controlling for age and sex, these researchers modeled mortality due to neoplasm at the postal code level while controlling spatial association effects by the row-standardized spatial weight matrix at an aggregated geographical level (postal code group). The spatial multilevel model includes potential covariates (e.g., age, sex) and the spatial context (e.g., neighbors), thus representing a significant advance to modeling multivariate spatial events. However, similar to the models of Dubin (1995) and Fingleton (1983b), this model accounts for spatially correlated events rather than the explicit revelation of location-specific spatial association or clustering.

In this paper, I set out a spatial logit association model for a binary spatial event at an aggregated spatial unit. I extend Fingleton (1983b), Dubin (1995) and Leyland et al. (2000) to capture spatial clustering. I adopt the definition of spatial clustering as either significantly high or low rates of clustering (Marshall 1991), a definition differing from local spatial association (Sokal, Oden, and Thomson 1998). The latter includes potential negative spatial association or the juxtaposition of high next to low values, while the former does not. In the following sections, I first describe the logit and spatial logit models and then demonstrate their utility using West Virginia lung cancer data. Finally, I offer some concluding remarks with regard to the applications and limitations of the model.

2. LOGIT MODEL FOR LOCAL SPATIAL ASSOCIATIONS

The logit model is often used to assess the effect of relative risk or the odds of success versus failure. Here the success (case) is a generic term to describe an event, which can be an undesirable event such as disease or crime. Fienberg (1977) sets out a general formulation for the logit model as a Generalized Linear Model (Nelder and Wedderburn 1972; McCullagh and Nelder 1989). For a categorical variable of mortality (dead = 1, alive = 2) by race (e.g., white = 1, black = 2, Hispanic = 3), for example, the saturated logit model for the expected number of cases or deaths (M_{r1}) versus noncases or alive (M_{r2}) for each specific race group r can be written as:

$$\log(M_{r1}/M_{r2}) = C + C_{r1} \tag{1}$$

where C is the grand mean, and r is the index of race. C_{r1} pertains to race-specific effects relative to the grand mean of log-odds. Suppose that, instead of racial groups, we treat this Poisson realization over a total of n regions indexed by i, the numbers of observed cases (M_{i1}) and noncases (M_{i2}) for spatial unit i can be described by the saturated logit model:

$$\log(M_{i1}/M_{i2}) = C + C_{i1} \tag{2}$$

Like equation (1), C is a constant for the overall effect of cases and C_{i1} parameters are marginal effects for $n - 1$ regions. C and C_{i1} are subjected to ANOVA-like normalization constraints. Unlike equation (1), the category of C_{i1} is based on region (i = 1 to n) rather than race or any other potential non-spatial categorical variables. This model has a very close connection to the commonly seen logistic regression or linear logit model (Wrigley 1985). If all the frequencies in (2) are disaggregated to individual observations, we have $Y_1, Y_2, \ldots Y_N$ independent binary random variables indexed by g with the probability of having a case $\Pr(Y_g = 1) = P_g$ for the gth individual. The logit, which is the logistic transformation of the probability of having a success, is:

$$\text{Logit } (P_g) = \log(P_g/(1 - P_g)) = \Sigma\beta_i X_{gi} \tag{3}$$

where X_{gi} denotes the ith region categorical covariate and β_i is the corresponding coefficient measuring the regional effect. Equation (3) is constrained in such a way that the only covariates are $n - 1$ regional dummy variables. It can be easily shown that equations (2) and (3) are equivalent. Based on model (3), Diggle (2000) added a spatial component $S(x)$ to account for unexplained spatial variation. Similarly, a spatial component can be added in equation (2) to make it a spatial logit model.

However, equation (2) has already been saturated, and it does not have any statistical power, except in that it fully describes the relative strength and magnitude for cases and noncases in each region relative to C. For this reason, statisticians often start with the independence model, where region specific coefficients (C_{i1}) in (2) are dropped. The resultant model, which is often used as the null hypothesis (H_0), tests whether all the C_{i1} parameters in (2) are zero. If they are, cases among various spatial units are independent, i.e.,

$$\log(M_{i1}/M_{i2}) = C \tag{4}$$

As demonstrated in Appendix A, this test can be achieved through the evaluation of the goodness-of-fit statistic via the likelihood ratio test statistic (L^2) for a given number of degrees of freedom (df).

When the independence model does not fit, known risk factors can be included:

$$\log(M_{i1}/M_{i2}) = C + \Sigma\beta_k x_k \tag{5}$$

where β_k are the coefficients for potential risk factors (x_k), or ecological variables pertaining to each region. We have seen this type of model in revealing region- or metropolitan-specific mobility, where region-specific risk factors could be push factors, such as economy, crime, and amenity factors. If an autocorrelated regional effect is suspected, while not knowing any potential risk factors, the spatial association terms can be used in place of risk factors in (5):

$$\log(M_{i1}/M_{i2}) = C + \Sigma\beta_i w_i \tag{6}$$

where β_i is the coefficient for spatial effect indexed by the *ith* region and its neighbors covered by w_i in a spatial weight matrix W. If $w_i = 1$, the region is adjacent to i inclusive (i.e., including the *i*th region itself), and 0 otherwise. Equation (6) is the so-called spatial-autocorrelated logit model, because the only explanatory variables are adjacent areas. In both equations (5) and (6), the number of risk factors and the number of regional vectors cannot exceed $n - 1$, the maximum number of parameters for the saturated model. Hence, we have $n - 1$ potential degrees of freedom for (6). Regions covered by each w_i would be spatially associated if its inclusion improves the model fit substantially. In a more general case, we can include both risk factors and regional autocorrelated terms by combining (5) and (6):

$$\log(M_{i1}/M_{i2}) = C + \Sigma\beta_k x_k + \Sigma\beta_i w_i \tag{7}$$

Again, the sum of the risk factors k and spatial associations i in (7) cannot exceed the maximum number of degrees of freedom $n - 1$, or $(k + i) < (n - 1)$.

In all the above equations, additional sample categories can be included by subdividing the sample into several categories (e.g., age, sex, time). For example, when an additional category (e.g., sex) is included in (7) we have an independence model not only between spatial units, but also between the control groups (e.g., male/female). The corresponding spatial logit model is:

$$\log(M_{ip1}/M_{ip2}) = C + C_{s1} + \Sigma\beta_k x_k + \Sigma\beta_i w_i \tag{8}$$

where the p subscript indexes person specific characteristics, which in this case is sex (s). C_{s1} represents marginal effects for subgroups indexed by s, which can interact with both potential risk factors and spatial associations:

$$\log(M_{ip1}/M_{ip2}) = C + C_{s1} + \Sigma\beta_k x_k + \Sigma\beta_i w_i + \Sigma\beta_{ks} x_k + \Sigma\beta_{is} w_i \tag{9}$$

In equation (9), $\beta_{ks} x_k$ and $\beta_{is} w_i$ are parameters for interaction terms indexed by *kp* and *ip* respectively. Again, the sum of *k, i, ks, si* terms should be less than ($n^*s - 1$). The task is to identify potential local spatial associations using the remaining number of degrees of freedom.

When the traditional autocorrelation tests are compared with the spatial logit association model, they complement each other. Moran's I and Getis-Od G test for global autocorrelation whereas the LISA and G_i (G_i^*) test for local association. Although these test statistics serve their purposes well, the result from a local test is separated from the corresponding global test. In other words, a significant result from a local test does not necessarily translate into the significant P-value for the corresponding global test (Anselin 1995). In some situations, it is desirable to evaluate the results from both the global and the local tests when making statistical inference about local associations (Sokal et al. 1998; Ord and Getis 2001; Tiefelsdorf 2002). The spatial logit model is complementary to these traditional test statistics, because it is a model-based test, and there is no separation between local and global tests. When a model is rejected by the likelihood ratio test at the global level, at least one statistically significant parameter must exist at the local level. Conversely, if there is a region covered by a statistically significant local association parameter, the likelihood ratio test must be significant. The evaluation of local parameter estimates, which could be assisted by various spatial test statistics, is an integral part of model testing process.

3. SPATIAL ENUMERATION ALGORITHM FOR LOCAL ASSOCIATIONS

When little is known about potential risk factors, searching for local spatial associations is a way to identify hot or cool spots on one hand, and to uncover potential risk factors on the other. Here, a hot spot refers to spatial clustering of high values or elevated events, while a cool spot refers to clustering of low values or less frequent events (see Figure 1). For cluster detections, equation (6) can be applied to test the strength of spatial logit associations. In this case, we have $n - 1$ potential spatial associations to be tested using some of the $n - 1$ degrees of freedom—the number of logits (n) minus the number of linearly independent parameters ($2 - 1 = 1$). Since the likelihood ratio chi-squared test can compare two alternative models with a nested parameter structure (Appendix I), we can design an enumeration algorithm to sequentially search for a potential set of spatial associations by retaining one significant association while searching and testing for the next one.

First, the existence of the independence model is tested. If this model is rejected, w_i (for $i = 1$ to $n - 1$) can be sequentially entered into the model to test for the exis-

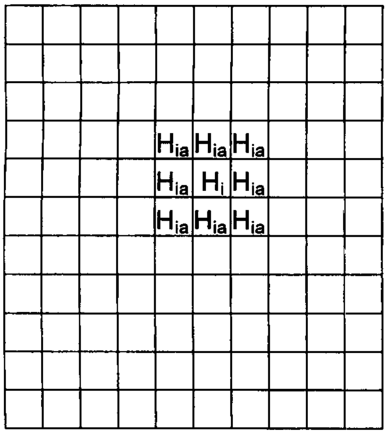

FIG. 1. Hot Spot Illustration. NOTE: Cases and noncases are generated randomly shown in empty grids. H_i is a center grid with elevated cases, and H_{ia}s are high value grids adjacent to H_i.

tence of local associations. A significant spatial cluster captured by β_i will reduce the likelihood ratio chi-squared statistic (G^2) with one additional degree of freedom. During the first step of searching, there might be multiple significant local associations, and only the one with the largest G^2 reduction is selected. Once the most significant β_i is selected, the w_i vector corresponding to this term is dropped from the next round of search for the second most significant local association term. This process is analogous to the forward stepwise search in a linear regression model. The only difference here is that in stepwise regression, the search is for significant independent variables, whereas in this enumeration algorithm the goal is to search for autocorrelated regions using a pre-specified structure. Suppose that β_1, which corresponds to adjacent regions covered by w_1, is selected in the first round of $n - 1$ searches. Then $w_{i \neq 1}$ is constrained and the search process for w_2 to w_{n-1} is repeated for $n - 2$ *times for the second most significant local association.* Once the second local association is identified, its corresponding w_i is dropped from the search set and the $n - 3$ searches proceed in the third round. This search process will stop when an additional β_i will no longer improve the model fit, which will be reflected by an insignificant parameter and weak contribution to the reduction of G^2 for one df. The theoretical upper limit of the number of searches could reach $(n - 1)!$.

In reality, $n - 1$ local associations are unlikely, as they would represent $n - 1$ pockets of distinct hot or cool spots, which is equivalent to fitting the saturated model. Another way to look at this is through the interpretation of parameter estimates. A positive and significant β_i means that the *i*th region and its neighbors are a hot spot or in an excessive pocket of cases as opposed to noncases. The inclusion of this parameter will significantly improve the model fit relative to the independence model. A negative and significant β_i, in contrast, means that the inclusion of the *i*th region and its neighbors as a "cool" spot improves the model fit significantly. Since significant associations tend to be clustered, once the most significant pocket of regions is included, some potential pockets adjacent to the most significant one are not likely to be significant. Suppose that in Figure 1, H stands for high values while empty cells are values that are randomly distributed. If the pocket of excessive values centered at the *i*th region is the most significant, then a pocket centered at a region adjacent to the *i*th region (H_{ia}) is less likely to be *significant, even though it would have been significant* if the *i*th region were not included. Hence, the number of local spatial associations are not likely to be more than the number of units (n) divided by the number of neighbors.

To illustrate this point, I randomly generated population-at-risk and sample events based on the 10 by 10 lattice in Figure 1 with the population and sample means being 2,000 and 50, respectively. The distance between each grid cell centroid is one mile, and the weight matrix is based on the queen's rule for the 0−1 adjacent matrix. For this random sample, the likelihood ratio chi-squared statistic is 111.16 with 99 df, which is not enough to reject the independence hypothesis between sample events. I then randomly raised the number of events by 50 to 60 percent around a non-boundary 3 by 3 grid. In other words, if $\lambda = 50$, then an elevated risk would correspond to $\lambda = 75$ to 80 for a hot spot. This time, the G^2 increases to 208.78 with 99 df, a significant deviation from the independence model. To search for this local association, I used the enumeration algorithm described above. In this particular case, the largest likelihood ratio chi-squared test is found at the center (H_i) for $G^2 = 108$ with 98 df. With w_{Hi} being included, any additions of w_i centered at H_{ia} are not significant. The sole hot spot being identified is encircled by the 3 by 3 cells around the center H_i.

Since the autocorrelated spatial logit model is a special case of the well-defined standard logit model, *it may not be necessary to do the standard power test.* However, to double-check if the H_0 also applies spatially, I used the same lattice from Figure 1 to randomly generate cases from 1 to 3 with an exposure or at-risk population of around 1,000. This simulation was repeated 1,000 times, and each time the enumeration algorithm was invoked to search for potential logit spatial associations. The simulations did not reject any independence model when cases and noncases were generated randomly.

4. EMPIRICAL EXAMPLE

To apply the logit spatial association model, I selected the age-adjusted number of deaths due to lung cancer for the fifty-five counties in West Virginia. The data set, which was originally compiled by the National Cancer Institute, covers a twenty-five-year period (1970 to 1994) and is disaggregated by sex, number of deaths, and at-risk population. The standardization is over the twenty-five-year period with the number of alive equal to the number of population-at-risk minus the number of deaths.

As a part of the exploratory analysis, I started with the spatial and sex independence model, in equation (9), without any known risk factors: $\log(M_{is(1)}/M_{is(2)}) = C_{s(f)} + \Sigma\beta_i w_i + \Sigma\beta_{is} w_i$ with all β terms being set to zero. In this model, s indexes sex, and $c_{s(f)}$ is the parameter for sex with male as the reference category. This model says that lung cancer cases are spatially independent, and the difference in sex is proportional to the spatial pattern by a factor of $c_{s(f)}$ for females. This model was rejected at $G^2 = 215$ with 108 df (P-value < 0.0001). To include spatial neighbors as an autocorrelated component, the enumeration algorithm was invoked to search for all significant β_i. Since it is possible that local spatial associations exist for both males and females, or for just one of the sex groups, the search should include the interaction terms between sex and w_i, and the likely outcomes are that β_i is significant for both males and females or for males or females only.

Table 1 lists results starting with the independence model. Note that in fitting a model, the smaller the deviance or G^2 the better. To evaluate each model, I compare the reduction of G^2 for a given number of df. With Model I being the baseline, Model II to Model V each improves the fit of the model significantly at the $P < 0.01$ level for one degree of freedom over the previous model. For example, Model II improves G^2 by 31 percent ([215.25 − 148.90]/215.25) over the rejected independence model. Model III, which includes Lincoln County and its adjacent counties, improves the model to 109 $G^2 = 109$ with 106 df. At this level, the hypothesis that the model with one local association term does not fit the data cannot be rejected. However, it does

TABLE 1
Results of Model Fitting Using a Forward Stepwise Searching Procedure

	$C_{s(f)}$	β_{33}	β_{44}	β_{48}	β_{17}	β_{21}	G^2	df	$G_i^2 - G_{i-1}^2$
Model I	0.322						215.24	108	
Model II	0.322	0.654					148.9	107	63.34**
Model III	0.322	0.68	1.254				108.79	106	41.11**
Model IV	0.322	0.683	1.231	1.178			99.05	105	9.74**
Model V	0.322	0.671	1.14	1.171	0.894		92.05	104	7**
Model VI	0.322	0.726	1.128	1.167	0.883	0.891	86.71	103	5.31*
Model VII*	0.317	0.727	1.131	1.168	0.867	0.893	86.68	103	5.34*
W_j counties		Pendleton	Lincoln	Logan	Doddridge	Grant			

NOTE: **P-value at the 0.01 significant level; *P-value at the 0.05 significant level. β_{17} in model VII applies to males only while other βs apply to both males and females. All other models (I to VI) are for both sexes.

[1] I used Bayesian information criteria (BIC) to check all of the models in Table 1 (Raftery 1986). All of the models had negative BIC values, suggesting that they are generally acceptable in terms of sample size.

not mean that the model fits very well. Hence, the subsequent models are not only designed to improve model fit, but also to reveal other potential local associations.[1] Indeed, the goodness-of-fit statistics for Models VI and VII are almost identical ($P <$ 0.021) compared to Model V, but the β_{17} term in Model VII applies only to males. In other words, Model VII attributing the cool spot at Doddridge County to males only is equivalent to Model VI attributing it to both males and females. However, since Model VII narrows the covariate to males and gains more information than it does from VI, Model VII could be preferred. To be conservative, however, I chose Model V with $P < 0.01$ as the final model, which captures four local association terms in two clusters: one low rate cluster (cool spot) around Pendleton and Grant Counties, and one excessive rate cluster around Lincoln and Logan Counties (see Figure 2).

The interpretations of parameters follow the typical logit model. All βs and Cs are in exponential terms or odds ratios. $C_{s(f)}$ is the odds ratio for females as opposed to males. Females are about one-third as likely as males to die from lung cancer. Those in counties adjacent to Lincoln (covered by w_{44}) and Logan (covered by w_{48}) Counties are respectively 1.128 and 1.167 times as likely to die from lung cancer as those in the grand mean, or the reference group. In other words, everything else being equal, a person regardless of sex is 12.8 percent ([1.128 − 1]*100) more likely to die from the disease than a person from counties not covered by any clusters. Similarly, one can evaluate cool spots centered on Grant and Pendleton Counties.[2]

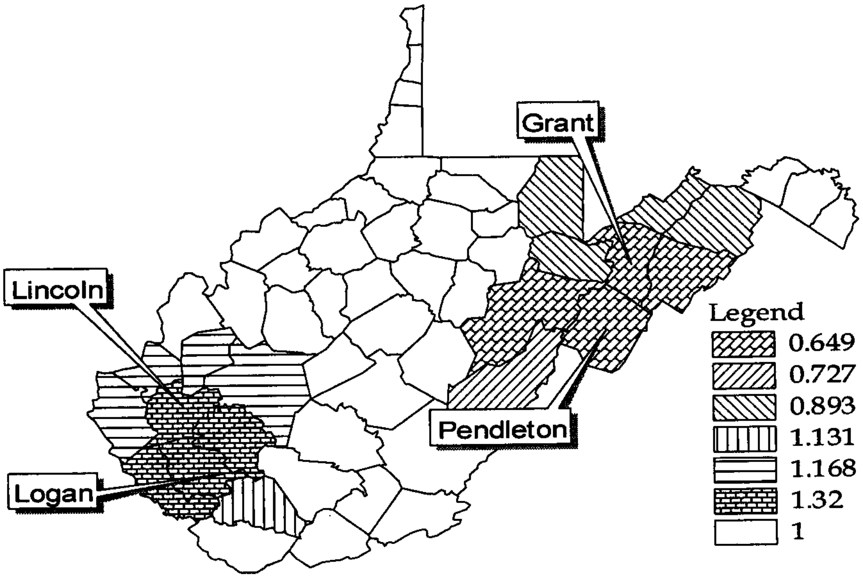

FIG. 2. Odds Ratios of West Virginia Lung Cancer Hot and Cool Spots

[2] If a single county contributes most to a hot or cool spot, the effect could be from an outlier, a spatial unit that independently appears in the study area but its value is extremely high or low. Although formal testing methods require a new test statistic, an analytical treatment of spatial outliers can be developed based on a deleted residual method by comparing two likelihood ratio tests. Under a spatial logit modeling framework, if a single region contributes most to the elevated risk within a pocket of high value regions, the addition of this region is likely to result in a significantly large deviance pulling away from the clustered area as a whole. I used this method to systematically evaluate diagonal units, and found that none of them reduce the deviance so such that the identified hot/cool spots became insignificant. All of the models were implemented with S+. Program codes and data are available upon request.

In addition, Logan and Lincoln Counties are two overlapping hot spots, meaning that counties covered by both w_{44} and w_{48} have even greater odds—1.316 (1.128 * 1.167). Likewise, counties covered by both w_{33} (centered at Pendleton County) and w_{21}(centered at Grant County) are overlapping cool spots with an odds-ratio of 0.641 (0.726*0.883). The overlapping nature of parameterizations is quite helpful because it does not treat counties within a clustered area equally. This feature is distinctly different from other local spatial statistics. For example, the Getis and Ord (1992) G_i^* test is able to identify local associations, but the values of G_i^* cannot be inferred jointly even if the two are adjacent. In the case of the male sample, the local G_i^*s are positive and significant for Logan and Boone Counties, suggesting a cluster around these two counties that cover roughly the same area as the Lincoln-Logan cluster. However, the magnitude of the excess cannot be evaluated for counties covered by both G_i^*-Boone and G_i^*-Logan. As for slight differences in terms of cluster coverage, we should remember that G_i^* is a rate-based statistic test (occurrence versus at-risk population), while β_i is an odds ratio-based statistical test for a given model. In addition, G_i^* is calculated for males only as opposed to males and females in β_i.

As indicated in equation (7), potential risk factors can be included in the model. For exploratory purposes, I examined potential relationships between excessive deaths due to lung cancer and coal mine activities, because areas around Logan are in the cluster of intense coal mining activities. I included the total number of coal mine workers in 1990, and this variable was not significant in Model VII. This result is expected, as the employment variable only reflects a particular year, while the number of deaths covers the twenty-five-year period. People die from lung cancer long after they are exposed to carcinogenic air-particulates. In addition, coal mine workers tend to retire to nearby counties, and the spatial association could capture the spatial lag due to migration (Sabel et al. 2000). Also explored was population density as a proxy for human activities, or air quality, and this variable was not significant. In the absence of county-level data with potential risk factors, such as smoking patterns and air quality, the autocorrelated spatial logit model fits the data fairly well.

5. CONCLUSIONS

This paper sets out a spatial logit model that accounts for both spatial associations and potential risk factors. The model is autocorrelated because the explanatory variables of regional logits are neighboring regions, and it is a logit model because it can incorporate likely covariates within the subgroups and known risk factors similar to the conventional logit model. The parameter estimates from the model explicitly reveal spatial clusters in terms of odds ratios. These odds ratios account for sample size in each spatial unit and overcome some of the potential problems associated with sample size in LISAs and Getis-Ord G, as well as the global Moran's I (Besag and Newell 1991; Oden 1995; Bao and Henry 1996). Like local G_i^* and spatial chi-squared (R_i) tests, the spatial logit model includes the diagonal elements in the spatial weight matrix, which is able to reveal spatial clustering by including all the spatial units within the cluster. Since traditional spatial test statistics such as Moran's I and LISA are exploratory in nature, while the spatial logit model is a model-based test that relies on exploratory data analyses, the two approaches are complementary. As spatial statistics make inroads from geography into other social and biological sciences, the need for the logit spatial association model is likely to increase.

As pointed out by Diggle (2000), when a spatial stochastic model broadly fits the available data, any autocorrelated spatial structure represents the unexplained variation. In this regard, although the logit spatial association model is designed to identify "hot" or "cool" spatial logits, it can be useful in identifying the cause of spatial autocorrelation of unknown origin. In the case study, the cluster of excessive deaths due to lung cancer was not really centered at the counties with the highest mining activities. However, this effect is, perhaps, more meaningful. Rather than die directly from the long-term exposure to air-particulates, people often die from lung cancer long after they are exposed to air-particulates (Pope et al. 2002). When the time series-air quality and mining data are lacking or cannot reflect the spatial dispersion process of moving to a nearby county, the spatial associations uncovered by the logit spatial association model make more sense.

Even though the spatial logit model is used to detect clustering of spatial events, the model can also be used to account for the unobserved autocorrelated effect. For example, to model mobility over various spatial units, such as towns, one may find that individuals from certain towns are more likely to move than those from other towns even though various individual mobility factors such as age, education, income, and number of siblings are controlled. What might be at work are the network effects that are often unobservable from survey data. To account for spatial autocorrelated logits in terms of moving and staying, we can use the enumeration algorithm to search for pockets of spatially autocorrelated movers as additional controls in a logit or logistic model. This process will not only remove some potential bias due to the spatial lag (Anselin 1988), but also explicitly reveal spatial clusters of movers. The latter, in turn, will assist researchers to unravel additional explanatory variables.

Any statistical model is based on a set of assumptions. First, the logit spatial association model is based on a Poisson realization for spatial events with binary outcomes. When the sample size is too large, any events, even the spatially random ones, tend to be significant, because the likelihood ratio chi-squared test only evaluates the deviance between a model and the observed data. In the case of a large sample, an additional term often improves the model fit, but such an improvement may be trivial in terms of the information gained. One may need to consider Bayesian information criteria and a variety of goodness-of-fit criteria (Raftery 1995). Second, like G_i^* and local R, this logit model does not differentiate between the contributions of the reference unit and its adjacent area units. In some cases, a single outlier might cause a significant local effect. An algorithm accounting for the effect of outliers or single regions effect is more appropriate in a loglinear model than the logit model (Lin 1999). Third, the reference surface for the spatial logit spatial association model is the grand mean based on the independence model, and it is very similar to G_i^*. However, Ord and Getis (2001) demonstrated that the reference surface, similar to the univariate t-test for two groups, might be more desirable in some situations. Fourth, even though it is possible to use a distance matrix along with a distance function (e.g., exponential or power), the spatial adjacency weight matrix was used in part to avoid repeat or multiple testing problems (Tango 2000). Future research needs to investigate the effects of different weight matrices and different reference surfaces. Finally, the current version of the logit spatial association model only includes two outcomes, but this association can be extended to multinomial logit models with three or more outcomes.

APPENDIX I

According to Agresti (1990, 95–96), when the total N is large, $-2L^2$ follows chi-squared distribution with the degrees of freedom being the difference in the dimensions of parameter spaces under two alternative hypotheses (e.g., H_0 and H_a). Since the predicted logits and expected frequencies can be derived from parameter estimates, the likelihood ratio chi-squared statistic can be easily calculated:

$$G^2 = 2\sum(\text{Observed})\log(\text{Observed/Expected})$$

where the summation is over all cells, and the expression is -2 times the logarithm of the likelihood (L^2). In the case of n regions with m_{ij} and $\hat{m}_{ij}$ denoting the observed and expected frequencies respectively for *ith* unit ($i = 1$ to n) and *jth* category ($j = 1$ or 2), we have

$$G^2 = 2\Sigma\Sigma\, m_{ij} \log(m_{ij} / \hat{m}_{ij})$$

with the degrees of freedom being the number of logits minus the number of linearly independent parameters. For the saturated logit model (2), the df = 0; for the independence logit model ($\log[M_{i1}/M_{i2}] = C$), the df $= n - 1$. The likelihood ratio chi-squared statistic is:

$$-2[L^2(H_a\text{-Independent model}) - L^2(H_0\text{-saturated model})].$$

Since the $2L^2$ for the saturated model = 0, this expression equals G^2 for the independence model with $n - 1$ df. The task of comparing the two models is, therefore, equivalent to evaluating the difference in G^2 for the two models. This model comparison strategy applies to any two alternative models with a nested parameter structure. In this particular case, the independence model is the alternative one, and it is nested under the saturated model with $n - 1$ fewer parameters than the saturated model.

LITERATURE CITED

Agresti, A. (1990). *Categorical Data Analysis.* New York: Wiley.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models.* Boston: Kluwer Academic

———. (1995). Local Indicators of Spatial Association—LISA. *Geographic Analysis* 27, 93–115.

Bao, S., and M. Henry (1996). Heterogeneity Issues in Local Measurements of Spatial Association. *Geographic Systems* 3, 1–13.

Besag, J. (1972). Nearest Neighbour Systems and the Autologistic Model for Binary Data. *Journal of the Royal Statistical Society B* 34, 75–83.

Besag, J., and J. Newell (1991). The Detection of Clusters in Rare Diseases. *Journal of Royal Statistic Society A* 154, 143–55.

Best, N., K. Ickstadt, and R. Wolpert (2000). Spatial Poisson Regression for Health and Exposure Data Measured at Disparate Resolutions. *Journal of the American Statistical Association* 95, 1076–88.

Diggle, P. (2000). Overview of Statistical Methods for Disease Mapping and Its Relationship to Cluster Detection. In *Disease and Exposure Mapping,* edited by P. Elliot, J. Wakefield, N. Best, and D. Briggs. Oxford: Oxford University Press.

Dubin, R. (1995). Estimating Logit Models with Spatial Dependence. In *New Directions in Spatial Econometrics,* edited by L. Anselin and R. Flrax. Berlin: Springer-Verlag.

———. (1997). A Note on the Estimation of Spatial Logit Models. *Geographical Systems,* 4, 181–93.

Fienberg, S. (1977). *The Analysis of Cross-Classified Categorical Data.* Cambridge, Mass.: MIT Press.

Fingleton, B. (1983a). Independence, Stationarity, Categorical Data and the Chi-square Test. *Environment and Planning A* 15, 483–99.

———. (1983b). Loglinear Models with Dependent Spatial Data. *Environment and Planning A* 15, 801–13

Getis, A., and J. K. Ord (1992). The Analysis of Spatial Association by Use of Distance Statistics. *Geographical Analysis* 24, 189–206.

Hagerstrand, T. (1967). *Innovation Diffusion as a Spatial Process.* Chicago: University of Chicago Press.

Haining, R. P. (1982). Interaction Models and Spatial Diffusion Processes. *Geographical Analysis* 14, 95–108.

———. (1983). Spatial and Spatial-Temporal Interaction Models and the Analysis of Patterns of Diffusion. *Trans. Inst. Br. Geogr.* N.S. 8, 158–86.

Leyland, A., I. Lanford, J. Rasbash, and H. Goldstein (2000). Multivariate Spatial Models for Event Data. *Statistics in Medicine* 19, 2469–78.

Lin, G. (1999). Assessing Structural Change in U.S. Migration Pattern: A Log-Rate Modeling Approach. *Mathematical Population Studies* 7, 1–17

———. (2000). The Geographic Assessment of Elderly Disability in the U.S. *Social Science & Medicine* 50, 1015–24

Marshall, R. (1991). A Review of Methods for the Statistical Analysis of Spatial Patterns of Disease. *Journal of the Royal Statistical Society Series A* 154, 421–41

McCullagh, P., and J. A. Nelder (1989). *Generalized Linear Models* (2d ed). London: Chapman and Hall.

Nelder, J. A., and R. Wedderburn (1972). Generalized Linear Models. *Journal of Royal Statistical Society Series A* 135, 370–84.

Oden, N. (1995). Adjusting Moran's I for Population Density. *Statistics in Medicine* 14, 17–26.

Ord J. K., and A. Getis (2001). Testing for Local Spatial Autocorrelation in the Presence of Global Autocorrelation. *Journal of Regional Science* 41, 411–32

Pope, C., R. Burnett, M. Thun, E. Calle, D. Krewski, K. Ito, and G. D. Thurston (2002). Lung Cancer, Cardiopulmonary Mortality, and Long-Term Exposure to Fine Particulate Air Pollution *JAMA* 287, 1132–41.

Raftery, A. (1986). A Note on Bayes Factors for Log-Linear Contingency Table Models with Vague Prior Information. *Journal of the Royal Statistical Society Series B* 48, 249–50.

———. (1995). Bayesian Model Selection in Social Research. In *Sociological Methodology,* edited by Peter Marsden, pp. 111–63. Washington, D.C.: The American Sociological Association.

Rogerson, P. A. (1999). The Detection of Clusters Using a Spatial Version of the Chi-Square Goodness-of-Fit Statistics. *Geographical Analysis* 31, 130–47.

Sabel, C., A. Gatrell, M. Loytonen, P. Maasilta, and M. Joelainen (2000). Modeling Exposure Opportunities: Estimating Relative Risk for Motor Neurone Disease in Finland. *Social Science and Medicine* 50, 1121–37.

Senior, M., H. Williams, and G. Higgs (1998). Spatial and Temporal Variation of Mortality and Deprivation 2: Statistical Modeling. *Environment and Planning A* 30, 1815–34.

Sokal, R., N. Oden, and B. Thomson (1998). Local Spatial Autocorrelation in a Biological Model. *Geographic Analysis* 30, 331–51.

Tango, T. (2000). A Test for Spatial Disease Clustering Adjusted for Multiple Testing. *Statistics in Medicine* 19, 191–204.

Tiefelsdorf, M. (2002). The Saddlepoint Approximation of Moran's I's and Local Moran's I_i's Reference Distributions and Their Numerical Evaluation. *Geographical Analysis* 34, 187–206.

Wrigley, N. (1985). *Categorical Data Analysis for Geographers and Environmental Scientists.* New York: Longman.