2005

Loglinear Residual Tests of Moran’s I Autocorrelation: An Application to Kentucky Breast Cancer Data

Ge Lin

Tonglin Zhang

Loglinear Residual Tests of Moran’ I Autocorrelation: An Application to Kentucky Breast Cancer Data

By

Ge Lin and Tonglin Zhang

RESEARCH PAPER 2005-6

Ge Lin
Department of Geology and Geography, Regional Research Institute
West Virginia University

Tonglin Zhang
Department of Statistics, Purdue University

Abstract: Spatial regressions have been widely used, but their use with the permutation tests of residuals either in linear or loglinear models is rarely seen. In the present study, we have linked the Cliff-Ord permutation test of Moran’s I on linear regression errors to loglinear regression residuals under asymptotic normality. We devised both Pearson residual Moran’s IP R and deviance residual Moran’s IDR tests and applied them to a set of log-rate models for early stage and late-stage breast cancer together with socioeconomic and access-to-care data in Kentucky. The results showed that socioeconomic and access-to-care variables were sufficient to account for spatial clustering of early stage breast carcinomas with breast cancer screening and number of primary care providers being more persistent than county median family income. For late-stage carcinomas, in contrast, the late-stage incidence rate was negatively associated with breast cancer screening level. This result confirmed our expectation: a high screening level is associated with high incidence rate of early stage disease, which in turn reduces late-stage incidence rates. In addition, we located four late-stage breast cancer clusters that cannot be explained by socioeconomic and access-to-care variables.

1 Introduction

Linear or loglinear spatial regressions are common in spatial epidemiology [4]. A set of ecological variables are often fitted with disease rates or counts within a given area unit. After a final model is derived, one can also visually inspect residuals on a map for spatial clusters. For a linear model, a residual test of Moran's I for spatial autocorrelation can also be performed to detect spatial clustering for the unexplained regression errors. However, there is no corresponding spatial residual test of clustering for loglinear or Poisson regressions on count data, which was the challenge for the study we reported herein. We were interested in the spatial patterns of breast cancer incidence in Kentucky counties according to the development stage of disease at diagnosis. Breast cancer staging is known to be associated with socioeconomic conditions, mammography screening services and other variables [22]. Since socioeconomic variables are often spatially autocorrelated (e.g., poor areas tend to be clustered), we expect clustering of breast cancer to occur, at least for the incidence rates for early stage disease. If there is no significant environmental cause of breast cancer, the clustering tendency should disappear once we introduce area socioeconomic variables.

One way to test for the existence of spatial clustering is to set up a spatial autocorrelation test, such as Moran's I, for Guassian or continuous data by using the permutation test of residuals for Moran's I in a linear regression, as Cliff and Ord [6 (P. 197)] or Tiefelsdorf have suggested [20]. Converting incidence to rate, however, is often less appealing than retaining the original count of each in spatial data analysis. In addition, the Moran's I test assumes that attribute values (e.g., disease prevalence) are either in equal probability among all the geographic units or from a single parent distribution. These assumptions are often violated in the permutation test of Moran's I in disease data due to heterogeneous regional populations [2, 21] and large variation in sparsely populated areas [3, 15].

Alternatively, one could use a spatial logit association model, which tests the numbers of cases and noncases in each region and its adjacent regions as a spatial logit together with potential explanatory variables [13]. This method will be able to identify high-value and low-value clusters by searching for significant local spatial associations. It does not, however, have a global measure of spatial clustering that would complement the modeling process for local spatial logit associations.

Finally, Jacqmin-Gadda and Commenges proposed a homogeneity score test of a generalized linear model [10]. The test is based on response residuals in generalized linear models, a design that its authors claimed to correspond to the Cliff-Ord permutation test of linear regression errors. Although the score test is a valuable addition to the statistical repertoire, its null hypothesis is not spatial independence, as one would assume when applying the Moran's I test. In addition, the test generally requires several steps to navigate to a proper null hypothesis. The authors pointed out the need to extend the permutation test of residuals of linear regressions to generalized linear models. The latter would allow spatial analysts to directly apply log-likelihood (or deviance) residuals or Pearson residuals of loglinear models.

In this paper, we show that permutation tests are applicable to Pearson and deviance residuals of loglinear models in the same way that the traditional permutation test of regression residuals for Moran's I is applicable. In the next section, we review the permutation test of Moran's I by regression residuals and then reformulate it in the context of Poisson data by using the Pearson and deviance residuals of a loglinear model. In section 3, we examine breast cancer incidence in Kentucky counties by their stage at the time of diagnosis by using newly derived Pearson- and deviance-residual tests. In the final section, we offer some concluding remarks.

2 From linear to loglinear residual tests of Moran's I

Let us consider a study area that has m regions indexed by i. Let y_i be variable of the interest in region i. Moran's I [16] is expressed as:

$$I = \frac{\sum_{i=1}^{m} \sum_{j \neq i} w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{[\sum_{i=1}^{m} (y_i - \bar{y})^2/m][\sum_{i=1}^{m} \sum_{j \neq i} w_{ij}]}, \tag{1}$$

where $\bar{y} = \sum_{i=1}^{m} y_i/m$, w_{ij} is an element of a spatial weight matrix W with 1 being adjacent to region i and 0 otherwise. Under the assumption of homogeneity, the moments of Moran's I can be computed under either the normality or randomization assumptions; the former assumes that observations are independently generated from a normal population, and the latter assumes that the observations are generated from a set of random permutations of the observed values. In either case, the significance of Moran's I can be determined by the equation $I_{std} = [I - E(I)]/\sqrt{Var(I)}$ under the asymptotic normality assumption. A significant and positive value of Moran's I or positive autocorrelation (i.e., $I_{std} > z_{\alpha/2}$) usually indicates the existence of either high-value or low-value clustering. A significant and negative autocorrelation (i.e., $I_{std} < -z_{\alpha/2}$) usually indicates a tendency toward the juxtaposition of high values with low values. If there is no spatial dependence, I is often close to $-1/m$ or to 0 if m is large.

In order to account for covariates, Cliff and Ord [6 (P. 198)] suggest to take e_i as the variable of the interest in region i, where e_i is the i-th residual of a linear regression model written as $Y = Xb + e$, where Y is the m-dimensional response vector, X is the $(m \times p)$ designed matrix, b is the estimated values of a p dimensional parameter vector β and e is the m-dimensional of residuals for the error term ϵ. Moran's I in equation (1) then becomes:

$$I = \frac{\sum_{i=1}^{m} \sum_{j \neq i} w_{ij} (e_i - \bar{e})(e_j - \bar{e})}{[\sum_{i=1}^{m} (e_i - \bar{e})^2/m][\sum_{i=1}^{m} \sum_{j \neq i} w_{ij}]}. \tag{2}$$

Notice $\sum_{i=1}^{m} e_i = 0$ for the regression residuals, the definition of I could be reduced by removing $\bar{e}$ terms in both the numerator and denominator.

If there is no clustering in the observed data, the residuals are considered to be permutation equivalent. In a random permutation, the respective mean and variance of Moran's I are:

$$E(I) = -\frac{1}{m} \tag{3}$$

and

$$V(I) = \frac{m[(m^2-3m+3)S_1 - mS_2 + 3S_0^2] - b_2[(m^2-m)S_1 - 2mS_2 + 6S_0^2]}{(m-1)(m-2)(m-3)S_0^2} - E^2(I) \tag{4}$$

where $S_0 = \sum_{i=1}^{m}\sum_{j=1}^{m}(w_{ij}+w_{ji})/2$, $S_1 = \sum_{i=1}^{m}\sum_{j=1}^{m}(w_{ij}+w_{ji})^2/2$, and $S_2 = \sum_{i=1}^{m}(w_{i\cdot}+w_{\cdot i})^2$ with $w_{i\cdot} = \sum_{j=1}^{m} w_{ij}$, and $b_2 = m\sum_{i=1}^{m}(e_i-\bar{e})^4/[\sum_{i=1}^{m}(e_i-\bar{e})^2]^2$ (see Cliff and Ord [6 (P. 21)] for details). For the regression residuals, $b_2 = m\sum_{i=1}^{m} e_i^4/(\sum_{i=1}^{m} e_i^2)^2$ since $\sum_{i=1}^{m} e_i = 0$.

As above, let $I_{std} = [I - E(I)]/\sqrt{V(I)}$. By assuming that y_i are independently observed from a random variable with finite mean and variance in (1), so that $e_i = y_i - \bar{y}$ in (2), Sen [19] showed that I_{std} is approximately $N(0,1)$ as $m \to \infty$ if a) $w_{ij} = w_{ji}$, b) $\sum_{j=1}^{m} w_{ij}$ is uniformly bounded, and c) the limit of $\lim_{m\to\infty}\sum_{i=1}^{m}\sum_{j=1}^{m} w_{ij}^2 = \gamma^2$ exists and is positive. Under the same conditions for w_{ij}, Schmoyer [18] proved that I_{std} is also asymptotic $N(0,1)$ if e_i is the i-th residual of a linear model, and the error terms are assumed to be iid with a finite mean and variance. For a state such as Kentucky, where rural and urban populations vary substantially, breast cancer incidence rates are likely neither to be iid nor to have iid errors. The asymptotic normality of I_{std} under random permutations, therefore, is often violated [2, 21].

To account for potentially heterogeneous populations, we can devise a log-rate model that closely resembles a linear regression model. We intend to apply the Cliff-Ord permutation test based on the asymptotic normality, so that the test of Moran's I based on log-rate residuals is analogous to the one based on regression residuals. In his synthesis of previous studies, Agresti showed that the Pearson and log-likelihood (deviance) residuals of loglinear models are asymptotically multivariate normal with mean 0 and the variance-covariance matrix a projection matrix [1 (P. 431)] if the model captures true information. Consequently, the projection matrix in loglinear models also satisfies the asymptotic normality employed by Cliff and Ord for the permutation test similar to Schmoyer's projection matrix in a linear regression [18]. We can, therefore, apply the residuals of loglinear models to the permutation tests of residuals for Moran's I.

As an example, let n_i be the observed counts for a Poisson random variable N_i at region i, $i = 1, \cdots, m$, and let ξ_i be the i-th regional population. In a loglinear model, we assume N_i to be independent Poisson random variables, eath with parameter λ_i. Suppose that a set of explanatory variables $(x_{i,1}, \cdots, x_{i,q-1})$ are observed together with n_i. A log-rate model can then be expressed as

$$\log(\hat{n}_i/\xi_i) = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_{q-1} x_{i,q-1}. \tag{5}$$

In equation (5), $\hat{n}_i$ and $\log(\hat{n}_i/\xi_i)$ are, respectively, the expected count and the expected rate for region i, β_0 is the grand mean, and the other βs are parameters of explanatory variables. In the independence model, the true proportional vector of rates is $\pi = (\pi_1, \cdots, \pi_m)^t$, where π is the disease rate (probability) for region i while controlling for a set of explanatory variables, that is,

$$\pi_i(\beta) = \frac{\lambda_i}{\sum_{i=1}^m \lambda_i}, \tag{6}$$

with $\beta = (\beta_0, \cdots, \beta_{q-1})$, and

$$\lambda_i = \lambda_i(\beta) = \xi_i e^{\beta_0 + \sum_{j=1}^{q-1} x_{i,j}\beta_j}. \tag{7}$$

The conventional *Pearson residual* for region i is defined as

$$r_{i,p} = \frac{n_i - \hat{n}_i}{\hat{n}_i^{1/2}}. \tag{8}$$

The conventional *deviance residual* [1, pp 452] for region i is defined as

$$r_{i,d} = 2sign(n_i - \hat{n}_i)[n_i \log(n_i/\hat{n}_i) - n_i + \hat{n}_i]^{1/2}, \tag{9}$$

where $sign(a)$ is 1 if $a > 0$, is 0 if $a = 0$ and is -1 if $a < 0$.

We can test the residuals in equation (5) in the same way as Cliff and Ord did by replacing error terms in equation (2) with either Pearson residuals or deviance residuals. When e_i is replaced with $r_{i,p}$, Moran's I becomes Pearson residual Moran's I and we label it I_{PR}; when e_i is replaced with $r_{i,d}$, Moran's I becomes deviance or log-likelihood residual Moran's I and we label it I_{DR}. The explanatory variables can be estimated in the same way as the log-rate model, while the corresponding mean and variance of residual Moran's I can still be computed as Cliff and Ord did under the randomization assumption according to Schmoyer's exposition. Since $\sum_{i=1}^m r_{i,P}$ and $\sum_{i=1}^m r_{i,d}$ are generally not 0, Moran's I cannot be reduced to a simpler form as in linear regression residuals.

3 Application

Data and variables. The county-level breast cancer and at-risk population data were obtained from the Kentucky cancer registry for the years 1996-2000. The data set reports breast cancer cases according to their developmental stage as follow: 0, a benign tumor; 1 an in-situ tumor; 2, localized tumor; 3, a regional tumor; 4, a distant metastatic tumor; and 5, usually used to code patients who died with later stage disease without an autopsy report on file. For the purpose of our analysis, we deleted the stage 0 cases. Following the U.S. Surveillance, Epidemiology, and End Results (SEER) Program definitions, we regrouped the in-situ and localized tumors as early stage and the regional, distant and unknown tumors as late stage. Generally, early stage breast carcinomas are confined to the breast and can often be treated successfully, whereas late-stage carcinomas tend to spread beyond the breast and are often fatal.

On average, there were 96.7 per 100,000 women diagnosed at an early stage and 36.9 per 100,000 at a later stage. If the breast cancer incidence rate is constant across counties, the early stage breast cancer rate should, theoretically, be negatively related to the late-stage breast cancer rate. Figures 1 and 2 show the early and late stage cancer incidence rates in six quantiles (equal number of observations among 6 groups). We observed that a strip of counties along the boundary between Appalachian and non-Appalachian regions had an elevated early stage breast cancer rate, as did the westernmost counties. With regard to late-stage breast cancer, there was a cluster of counties with a high incidence rate around the northeastern Appalachian area; counties along the southern border of the state also had a higher incidence rate.

Since screening for early stage breast cancer tends to be associated with socioeconomic conditions and access to health care within a geographic area [7], we included additional county variables while testing spatial clustering for both early stage and late stage breast cancer rates. We obtained county socioeconomic data from the 2000 U.S. Census, which reports the percentage of the population (age 25 and over) with a college education (COLLDG), the percentage of population living under the poverty level(PVTRATE), median family income (MEDFINC), median housing value (MEDHVAL), et cetera, in 1999. The socioeconomic conditions in a county are expected to be related to breast cancer in two ways[5]. On one hand, breast cancer is more prevalent among white women or those with higher socioeconomic status, so counties that have higher median family income are expected to have greater breast cancer incidence rates. On the other hand, women who have a higher socioeconomic status tend to be more aware of and more able to afford breast cancer screening than are those who have a lower socioeconomic status. Consequently, although counties that have better socioeconomic conditions may have a higher incidence rate of early stage disease, they may not necessarily have a higher incidence rate of late-stage disease than do counties with worse socioeconomic conditions[17].

We relied on several other data sources for access-to-care measures. We obtained breast cancer screening rates from the 1997 and 1998 Behavioral Risk Factor Surveillance Systems(BRFSS)and divided rates into tertiles of high (H-screening:$>$ 70%), middle(M-screening 65-70%), and low (L-screening:$<$ 65%). A higher breast cancer screening level (primarily by mammography) is expected to be associated with higher early stage incidence rates, and negatively associated with late stage rates. In the preliminary analysis, we found that the differences between low and middle tertiles were minimal, and we grouped them together in the final analysis. We also obtained the 1998 population-to-primary care physician ratio (POP/PMD) in 1998 from the Kentucky Department of Public Health; a lower ratio indicates that a physician can give more attention to each patient. Since breast cancer screening is most frequently recommended in a primary care setting, having a greater number of primary care physicians should help to reduce the incidence of all stages of breast cancer. For this reason, we expected the population-to-physician ratio to be negatively associated with both early stage and late-stage breast cancer rates.

Finally, we used data from a geographic information system to derive geographic access measures for Kentucky counties using GIS data. We used the TIGER file from the 2000 U.S. census to derive a measure of access to major highways, that is, whether or not a major national highway (HWY) passes through a county. It was expected that highway access would increase access to health-care facilities and reduce breast cancer incidence rates. We also divided counties into within and outside the Appalachian region (APAREA). Counties within the Appalachian region generally are economically distressed and medically underserved, and the all-cause mortality rate tends to be much higher in counties within the region.

Analysis. To test Moran's I for spatial clustering, we first used Pearson residuals I_{PR} and deviance residuals I_{DR} for the null model without any covariates. I_{PR} and I_{DR} in the null loglinear model correspond to the traditional Moran's I without any covariates. We then introduced explanatory variables in the so-called ecological model. In the preliminary analysis, we found that college education, poverty rate, and median family income were highly correlated and were all associated with the incidence of the early stage breast cancer. We used median family income (MEDFINC) in the final analysis because it was the most significant variable in terms of the likelihood ratio test. We expected that the null model would indicate some spatial clustering through significant spatial autocorrelation and that the correlation should be weakened or disappear once the explanatory variables were introduced.

If the autocorrelation was found to persist or could not be explained by the ecological model, our task was to locate spatially clustered counties and provide our findings to epidemiologists and cancer specialists for further identification of the etiologies associated with breast cancer clusters. We used a spatial mixed model to search for high-value and low-value spatial clusters by including additional local spatial association terms, as demonstrated by

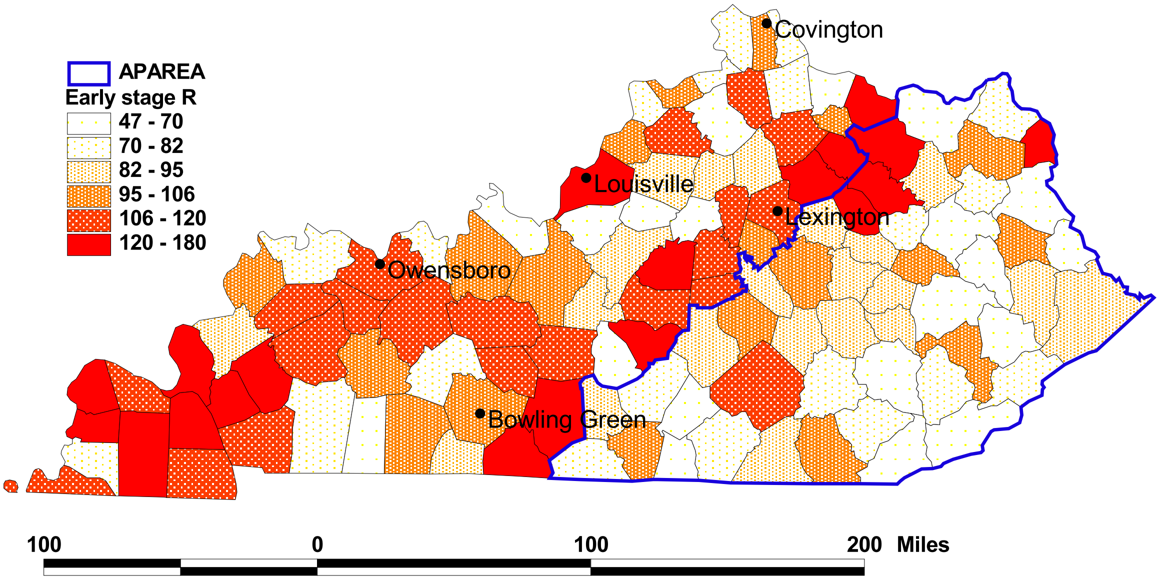

Figure 1: Early stage cancer incidence rate per 100,000 in Kentucky

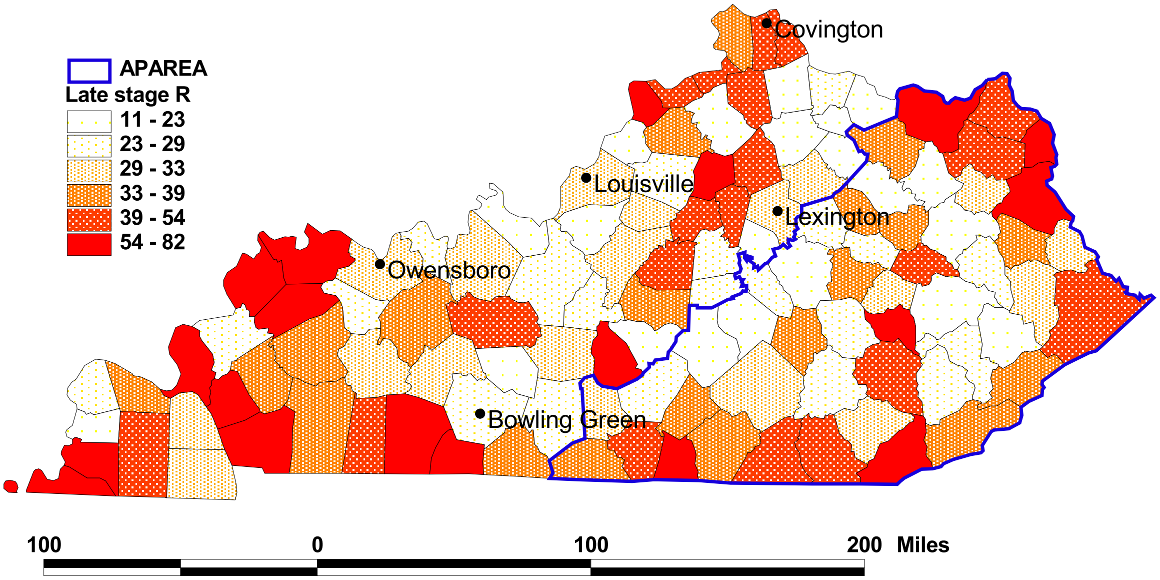

Figure 2: Late stage cancer incidence rate per 100,000 in Kentucky

Lin [13]. The method makes use of the vector of the spatial weight matrix, with 1 being adjacent to i inclusive (i.e., including the i-th region itself), and 0 being otherwise. If a cluster of counties associated with the $i-th$ vector could significantly reduce the log-likelihood (deviance), it indicates a local association or cluster centered around the $i-th$ county. If the association is positive, it suggests a high-value cluster. If the association is negative, it suggests a low-value cluster[14]. After controlling for pockets of high-value and low-value clustering and ecological covariates, we would then expect an insignificant residual Moran's I.

Results. Table 1 lists coefficients and T-ratios for Moran's I and ecological variables from the early stage log-rate models. In the null model, Moran's I from both Pearson residuals I_{PR} and deviance residuals I_{DR} were positively significant, suggesting a clustering tendency. Once the explanatory variables were introduced into the ecological model, however, I_{PR} and I_{DR} both became insignificant based on their corresponding residuals. Hence, a spatial clustering tendency in the null model reflected spatial patterning of socioeconomic status and access to care. In particular, counties with a high level of breast cancer screening or with easy highway access were associated with higher detection rates of early stage breast cancer, whereas the population-to-primary care physician ratio and location in the Appalachian region were negatively associated with detection rates. These results were all consistent with our expectations and the existing literature. In addition, county median family income was negatively associated with the early breast cancer detection rate. While it has been reported widely that women in less-developed counties, such as those in the Appalachian area of Kentucky, are less likely to have early breast cancers or to have their breast cancer diagnosed early [9], it seemed counterintuitive to us that a higher county median family income would be associated with a lower county rate of early stage breast cancer. When we added only MEDFINC to the null model, the coefficient for MEDFINC was positive and significant. We concluded, therefore, that when the level of breast cancer screening and other access-to-care variables were taken into account, a better county socioeconomic status, as indicated by family median income, was associated with a lower rate of early stage disease.

Turning to the results from the late-stage log-rate models (Table 2), we found that Moran's I_{PR} and I_{DR} were significant for both the null and ecological models and that the explanatory variables in the ecological model were insufficient to account for the clustering tendency of the late-stage incidence rates. The only two coefficients that remained significant were population-to-physician ratio and high screening level. The coefficient for the population-to-physician ratio remained consistent with the early stage model. However, the late stage rate was negatively associated with a high screening level. This result was consistent with our theory that a high screening level leads to high incidence rates of early stage disease and low rates of late stage disease.

Table 1: Early stage breast cancer log-rate models

Models	Null Coeff.	T-ratio	Ecology Coeff.	T-ratio
(Intercept)	-6.876	-703.68	-6.435	-91.35
MEDFINC			-0.017	-5.66
HWY			0.097	3.69
Appalachian			-0.269	-8.29
High screening			0.098	8.62
POP/PMD			-0.056	-6.52
Summary Stat	$G^2 = 597$	119	$G^2 = 339$	114
Moran's I_{PR}	0.177	3.41	0.070	1.41
Moran's I_{DR}	0.179	3.44	0.072	1.44

Table 2: Late-stage breast cancer log-rate models

Models	Null Coeff.	T-ratio	Ecol. Coeff.	T-ratio	Mixed Coeff.	T-ratio
(Intercept)	-7.949	-476.00	-7.558	-62.44	-7.803	-226.35
MEDFINC			-0.007	-1.41		
HWY			-0.052	-1.23		
Appalachian			-0.086	-1.59		
H-screening			-0.122	-6.25	-0.138	-7.35
POP/PMD			-0.100	-6.51	-0.091	-6.22
Barran-cluster					-0.299	-3.40
Greenup-cluster					0.308	3.86
Marshall-cluster					0.298	4.06
Union-cluster					0.347	3.63
Summary Stat	$G^2 = 392$	119	$G^2 = 303$	114	$G^2 = 252$	113
Moran's I_{PR}	0.172	3.25	0.141	2.69	0.077	1.56
Moran's I_{DR}	0.182	3.42	0.144	2.74	0.071	1.43

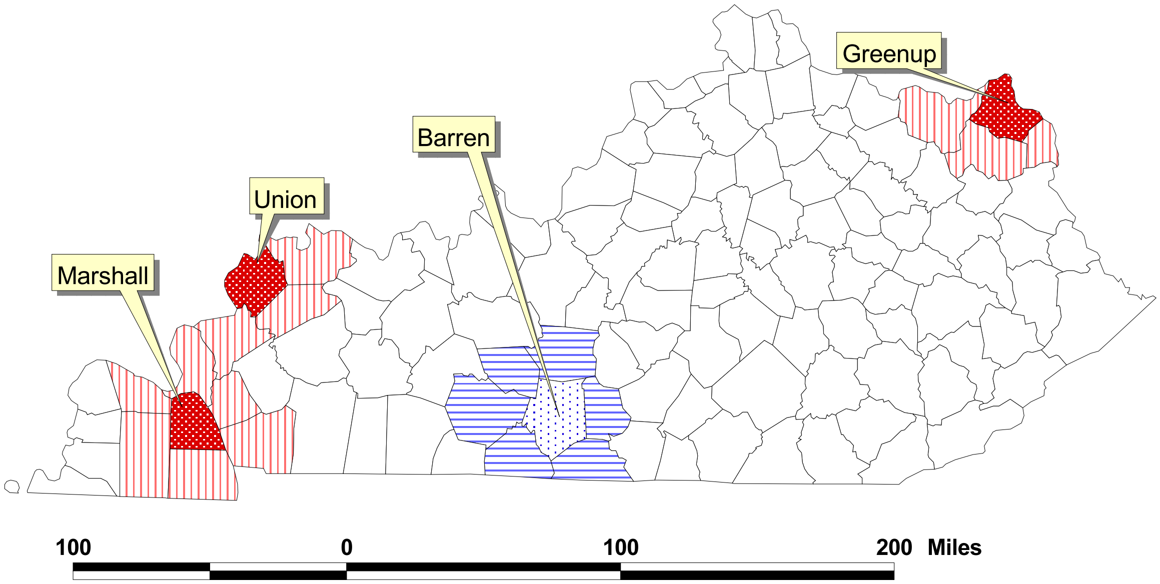

Figure 3: high-value and low-value clusters of late-stage breast cancer in Kentucky

To further pinpoint the core of clustered counties unexplained by the ecological model, we deleted insignificant variables from the ecological model and applied a stepwise regression to the column vectors of the W matrix with $w_{ii} = 1$. We identified four local association terms, none of which overlapped geographically. Each core county and its adjacent counties constituted a cluster, and including the core county only would not significantly reduce the clustered effect. Except for a cool spot around Barran County, the other three core counties represented the centers of three elevated late-stage clusters (Figure 3). For instance, the rate in Union County and its adjacent counties was 1.415 times the rates of other counties not included in the clusters. By including these clusters in the final model, both I_{PR} and I_{DR} were found to be insignificant, suggesting the disappearance of the clustering tendency once these clusters were accounted for in the model. It is worth mentioning that Jefferson County, where Louisville is located, had a very low late-stage rate, but its adjacent counties each had a relatively high rate. Although including Jefferson County in the final model would reduce the log-likelihood ratio, counties around Jefferson County would not constitute a cluster.

4 Conclusions

Although the asymptotic validity of permutation tests have been demonstrated in the literature and loglinear residuals are asymptotically normal, no one has applied them in the spatial context. In the current study, we have made the connection between the Cliff-Ord permutation test of Moran's I on linear regression errors to loglinear regression residuals under asymptotic normality and have devised I_{PR} and I_{DR} tests. We tested both based on a set of log-rate models for early state and late-stage breast cancer incidence data together with socioeconomic and access-to-care data in Kentucky. The results showed that socioeconomic and access-to-care variables were sufficient to account for spatial clustering of early stage breast carcinomas with access-to- care measures, such as breast cancer screening and number of primary care providers being more persistent than county median family income. After controlling for access-to-care measures and regional distress factors in the Appalachian counties, the purported positive association between higher socioeconomic and early stage breast cancer could be substantially weakened or reversed.

For late-stage carcinomas, two salient and persistent factors were level of breast cancer screening and population-to-primary care physician ratio. In contrast to the finding that a high screening level was associated with a high incidence rate of early stage breast cancer, the late-stage incidence rate was negatively associated with breast cancer screening level. This result confirmed our theory: a high screening level leads to high incidence rate of early stage disease, which in turn reduces late-stage incidence rates. When the two access variables failed to reduce the spatial clustering tendencies from late-stage breast cancer, we searched for a local spatial association based on the likelihood ratio test. We located four clusters, one low-value cluster around Barran County, and three high-value clusters. Two of the high-value clusters formed a single cluster region near the western corner of Kentucky. These unexplained clusters provided the basis for further investigation of the etiology of late-stage breast cancer in Kentucky.

Spatial regressions have been widely used, but their use with the permutation tests of residuals either in linear or loglinear models is rarely seen. An advantage of the loglinear residual permutation test over the linear residual permutation test is that the former can account for potential spatially heterogeneous populations, which makes it a viable alternative in the log-rate modeling of disease rates, as demonstrated in our study. The method can complement some spatial cluster tests, such as the spatial scan statistic[11] and G statistic [8]. In addition, the ability to show spatial clustering in I_{PR} and I_{DR} is complementary to disease mapping, which intends to display true disease risks while controlling for heterogeneous populations and regional risk factors[12]. Finally, we only applied loglinear model residuals in our permutation tests according to general properties set out in Agresti[1], Cliff and Ord[6], and Schmoyer[18]. The validity of permutation tests for generalized linear models remains to be established.

References

[1] Agresti, A. (1990). *Categorical Data Analysis.* Wiley, New York.

[2] Assuncao, R. and E. Reis(1999). A new proposal to adjust Moran's I for population density. *Statistic in Medicine*, **18**, 2147-2162.

[3] Besag, J. and J. Newell(1991). The detection of clusters in rare diseases. *Journal of Royal Statistic Society*, A, **154**, 143-55.

[4] 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**, 1076-1088.

[5] Bradley, CJ. Given CW, Robert C. (2001) Disparities in cancer diagnosis and survival *Cancer* 91, 178-188

[6] Cliff, A. D. and Ord, J. K. (1981). *Spatial Processes: Models And Applications*, Pion, London.

[7] Freeman, H.P (1989). Cancer and the socioeconomically disadvantaged. *Ca-A cancer Journal for clinicians* **39**:263-295.

[8] Getis, A., and Ord, J. (1992). The analysis of spatial association by use of distance statistics. *Geographical Analysis*, **24**, 189-206.

[9] Gregorio DI, Kulldorff M, Barry L, Samociuk H. (2002) Geographic differences in invasive and in situ breast cancer incidence according to precise geographic coordinates, Connecticut, 1991-95. Int J Cancer 100:194-198.

[10] Jacqmin-Gadda, H., Commenges, D., Nejjari, C. and Dartigues J. (1997). Testing of geographical correlation with adjustment for explanatory variables: an application to dyspnoea in the elderly. *Statistics in medicine*, **16**, 1283-1297.

[11] Kulldorff, M.(1997). A spatial scan statistic. *Communications in Statistics, Theory and Methods*, **26**, 1481-1496.

[12] Lawson, A.B. and A. Clark(2002). Spatial mixture relative risk models applied to disease mapping. *Statistics in medicine*, **21**, 359-370.

[13] Lin, G. (2003). A spatial logit association model for cluster detection. *Geographical Analysis*, **35**, 329-340.

[14] Lin, G. and T. Zhang (2004). A method for testing low-value spatial clustering for rare disease. *ACTA Tropica*, **91**, 279-289.

[15] Lin, G. (2004). Comparing three spatial clusters tests from rare to common diseases. *Computers, Environment and Urban Systems* 28: 691-699

[16] Moran, P. A. P. (1950). A test for the serial independence of residuals. *Biometrika*, **37**, 178-181.

[17] Roche L.S., Skinner, R., and Weinstein, RB (2002) Use of a geographic information system to identify and characterize areas with high proportions of distant stage breast cancer. *J Public Health Management Practice* 8, 26-32

[18] Schmoyer, R. L. (1994). Permutation tests form correlation in regression errors. *Journal of American Statistical Association*, **89**, 1507-1516.

[19] Sen, A. (1976). Large sample-size distribution of statistics used in testing for spatial correlation. *Geographical analysis*, **9**, 175-184.

[20] Tiefelsdorf, M. and Boots, B. (1995). The exact distribution of Moran's I. *Environment and Planning A*, **27**, 985-999.

[21] Waldhor, T.(1996). The spatial autocorrelation coefficient Moran's I under heteroscedasticity. *Statistic in Medicine*, **15**, 887-92.

[22] Yabroff. YR. And Gordis, L. (2003). Does stage at diagnosis influence the observed relationship between socioeconomic status and breast cancer incidence, case-fatality, and mortality? Social *Sciences and Medicine*. **57**:2265-2279