ELSEVIER

Statistics & Probability Letters 58 (2002) 245–251

STATISTICS & PROBABILITY LETTERS

www.elsevier.com/locate/stapro

A spatial filtering specification for the auto-Poisson model ☆

Daniel A. Griffith

Department of Geography and Interdisciplinary Statistics Program, Syracuse University, 144 Eggers Hall, N.Y. 13244-1020, USA

Received March 2001; received in revised form January 2002

Abstract

The auto-Poisson model describes georeferenced data consisting of counts exhibiting spatial dependence. Its conventional specification is plagued by being restricted to only situations involving negative spatial autocorrelation, and an intractable normalizing constant. Work summarized here accounts for spatial autocorrelation in the mean response specification by incorporating latent map pattern components. Results are reported for seven empirical datasets available in the literature.

Keywords: Auto-Poisson; Spatial autocorrelation; Georeferenced counts

0. Introduction

"The central role of the Poisson distribution with respect to the analysis of counts is analogous to the position of the normal distribution in the context of models for continuous data" (Upton and Fingleton, 1989, p. 71). Accordingly, when spatial data comprise counts, especially for rare events, the probability model that first comes to mind for describing these data is one based upon an auto-Poisson specification, which may be written in the form of approximations. The auto-log-Gaussian approximation circumvents the auto-Poisson's intractable normalizing factor, and both it and the auto-logistic approximation circumvent the auto-Poisson's restriction to only situations involving negative spatial autocorrelation, a restriction at odds with the real world where most all georeferenced data exhibit positive spatial autocorrelation. The intractable normalizing constant can be resolved using Markov Chain Monte Carlo (MCMC) procedures. The negative autocorrelation restriction can be resolved through Windsorizing (Kaiser and Cressie, 1997), whose primary drawback is that "the k most extreme observations on each end of the ordered sample are replaced by the nearest retained

☆ This research was supported by the National Science Foundation, research Grant #BCS-9905213.
E-mail address: griffith@maxwell.syr.edu (D.A. Griffith).

0167-7152/02/$ - see front matter
PII: S0167-7152(02)00099-8

observation" (Tietjen, 1986, p. 91) strictly to accommodate a model requirement. This paper departs from these practices by demonstrating that spatial autocorrelation can be accounted for in the mean response specification of an auto-Poisson model, extending the spatial filtering concept promoted by Getis and Griffith (2002).

1. Poisson regression

Poisson regression assumes independent counts, say n_i, taken at locations $i = 1, 2, \ldots, n$, where each of these counts is from a Poisson distribution, and these counts can be described by a set of explanatory variables denoted by matrix $\mathbf{X}_i$, a $1 \times p$ vector of covariate values for location i. The expected value of these data is given by $\mu_i(\mathbf{X}_i) = n_i(\mathbf{X}_i)\exp(\mathbf{X}_i\boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is the vector of non-redundant parameters, and the Poisson rates parameter is given by $\lambda_i(\mathbf{X}_i) = \mu_i(\mathbf{X}_i)/n_i(\mathbf{X}_i)$; the rates parameter $\lambda_i(\mathbf{X}_i)$ is both the mean and the variance of the Poisson distribution for location i.

Pairwise-only spatial dependence often is assumed when specifying auto-models, which renders the estimation problem here of evaluating the log-probability mass function term

$$\sum_{i=1}^{n} \alpha_i n_i - \sum_{i=1}^{n} \log(n_i!) + \rho \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij} n_i n_j, \tag{1}$$

where α_i is the parameter capturing large-scale variation (and hence could be specified in terms of vector $\mathbf{X}_i$), ρ is the spatial autocorrelation parameter, and c_{ij} is the geographic configuration weight associated with the pair of locations i and j. The n^2 set of c_{ij} values form a geographic connectivity or weights matrix, $\mathbf{C}$; if these c_{ij} values are binary 0–1, then $\mathbf{C}$ is the matrix used to calculate a Moran Coefficient (MC) and a Geary Ratio (GR) index of spatial autocorrelation.

The proposition promoted in this paper is that by including variables in matrix $\mathbf{X}$—the $n \times p$ concatenation of the n $\mathbf{X}_i$ vectors—that account for the spatial autocorrelation observed in the associated geographic distribution of counts, n_i, the third term in Eq. (1) can be dispensed with; in other words, spatial dependence effects are shifted from a small-scale variation term to the mean response term. This shift can occur by introducing synthetic variables into matrix $\mathbf{X}$ that are the eigenvectors of matrix $(\mathbf{I} - \mathbf{1}\mathbf{1}'/n)\mathbf{C}(\mathbf{I} - \mathbf{1}\mathbf{1}'/n)$, where $\mathbf{I}$ is the identity matrix, $\mathbf{1}$ is an $n \times 1$ vector of ones, $'$ denotes the operation of matrix transpose, and n is the number of areal units; this matrix expression appears in the numerator of the MC. These eigenvectors may be interpreted in the context of map pattern as orthogonal sets of geographically distributed numerical values that sequentially maximize the MC value, beginning with the largest positive MC value possible and ending with the largest negative MC value achievable. Hence, these n eigenvectors describe the full range of all possible mutually orthogonal map patterns. In the presence of positive spatial autocorrelation, then, an analysis can employ those eigenvectors depicting map patterns exhibiting consequential levels of positive spatial autocorrelation (e.g., $MC > 0.25$).

Given this result, the research problem becomes one of determining for expression (1) whether $\sum_{i=1}^{n} \alpha_i n_i$ can be replaced by $\sum_{i=1}^{n} \mathbf{E}_i\boldsymbol{\beta}_i n_i$, dispensing with $\rho \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij} n_i n_j$ by shifting spatial dependence effects to the large-scale variation term represented by $\mathbf{E}_i\boldsymbol{\beta}$.

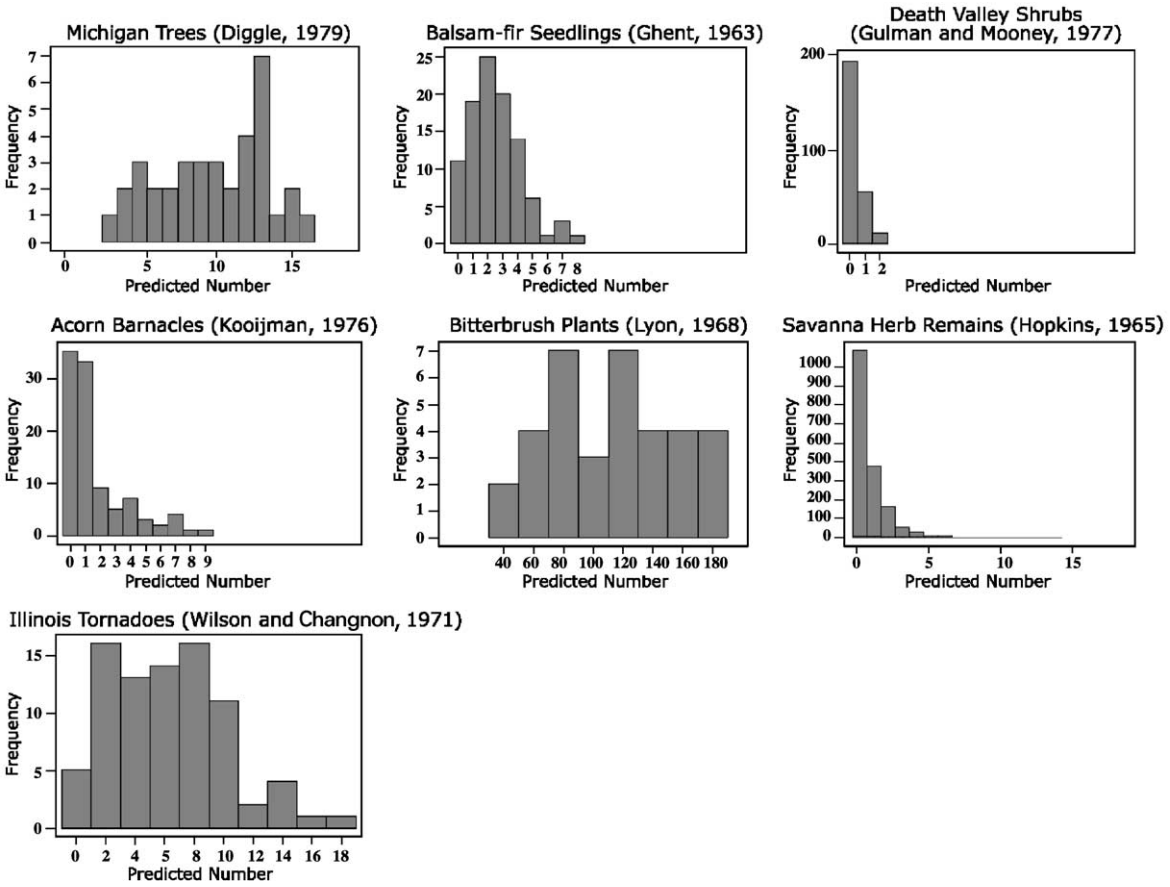

Fig. 1. Histograms for counts of the presence of phenomena in areal units for selected empirical data sets found in the literature.

2. Selected empirical demonstrations

Seven datasets available in the literature, ranging in size from $n = 35$ to $n = 1600$, are explored here: the number of bitterbrush plants contained in 100×100-foot quadrats forming a rectangular 5×7 region (Lyon, 1968); the distribution of hickory trees in Lansing Woods, Michigan (Diggle, 1979) post-stratified into a 6×6 square region; tornado occurrences in Illinois, for $n = 83$ 70×70-mile quadrats superimposed upon the state and "touch downs" counted (Wilson and Changnon, 1971); the number of balsam-fir seedlings in a 10×10 square region covered by five-foot square quadrats (Ghent, 1963); grid counts of acorn barnacles on the side of a ship for a 10×10 set of quadrats (Kooijman, 1976); the number of *Atriplex hymenelytra* across a 40 m $\times$ 40 m region of the floor of Death Valley partitioned into a 16×16 grid of quadrats (Gulmon and Mooney, 1977); and, the herb remains in a 20 m $\times$ 20 m region of the Nigeria savanna partitioned into a 40×40 grid of quadrats (Hopkins, 1965). Histograms for these different sets of counts appear in Fig. 1; aggregate Poisson distribution evaluations of them, using $e^{-\lambda}\lambda^k/k!$, appear in Table 1. The simplest evidence, ignoring

Table 1
Quantitative assessments of conformity with a Poisson frequency distribution

Dataset	χ^2 goodness-of-fit	χ^2 probability under H_0	$\frac{\hat{\mu}}{\sigma^2}$	Counts		Log-counts		Probability for S–W
				MC	GR	MC	GR	
Michigan hickory trees	48.810	0.06	0.777	0.381	0.551	0.325	0.638	0.0054
Balsam-fir seedlings	131.334	0.01	0.839	0.289	0.721	0.302	0.703	0.0012
Death Valley shrubs	215.094	0.97	1.004	0.064	0.958	0.067	0.953	≈ 0
Acorn barnacles	240.358	≈ 0	0.375	0.401	0.566	0.335	0.637	≈ 0
Bitterbrush plants	549.549	≈ 0	0.065	0.081	0.877	0.067	0.907	0.0811
Savanna herb remains	2339.708	≈ 0	0.521	0.198	0.703	0.200	0.837	≈ 0
Tornado touch downs	240.025	≈ 0	0.369	0.248	0.718	0.266	0.666	0.0953

latent spatial autocorrelation in these data, suggests that the frequency distribution of Death Valley shrubs closely conforms to, the frequency distribution of Michigan hickory trees and seedlings nearly conforms to, and the remaining frequency distributions markedly differ from a Poisson distribution.

2.1. Initial assessment of latent spatial autocorrelation in the georeferenced counts

The link between a Poisson-distributed variable and its log-normal distribution approximation permits georeferenced log-counts to be analyzed with the auto-normal model. MC and GR spatial autocorrelation indices for the count and log-count variables appear in Table 1, as do corresponding Shapiro–Wilk (S–W) results evaluating conformity to a normal frequency distribution. These statistics reveal that none of the logarithmic transformations sufficiently align the empirical frequency distributions with a bell-shaped curve, although two come close, all of the spatial autocorrelation detected is positive, and geographic distributions of both the Death Valley shrubs and the bitterbrush plants essentially contain only trace amounts of spatial autocorrelation.

Results for auto-log-Gaussian approximations estimated with each dataset appear in Table 2, and corroborate the presence of weak-to-moderate positive spatial autocorrelation. The residuals from these simultaneous autoregressive (SAR) models better display conformity with a normal distribution for all but the Michigan hickory trees dataset; especially the Death Valley shrubs frequency distribution continues to markedly deviate from a bell-shaped curve. The percentage of variance accounted for by latent spatial autocorrelation ranges from 1.6 to 29.42.

2.2. Spatial autocorrelation filters

Eigenvectors were extracted from matrix $(\mathbf{I} - \mathbf{11}'/n)\mathbf{C}(\mathbf{I} - \mathbf{11}'/n)$ for all seven datasets. Only eigenvectors depicting map patterns displaying moderate-to-strong positive spatial autocorrelation are considered, since the georeferenced data exhibit only positive spatial autocorrelation.

The log-Gaussian, logistic, and Poisson specifications often identify sets of eigenvector covariates that are very similar: three for the Michigan hickory tree data, with MC values ranging between 0.50761 and 0.91468; five for the balsam-fir seedlings data, with MC values ranging between 0.69815

Table 2
Selected SAR model and filtered Poisson regression model results

Dataset	SAR model			Filtered Poisson regression model		
	$\hat{\rho}$	pseudo-R^2	S–W	No. of prominent eigenvectors	Pseudo-R^2	MC for residuals
Michigan hickory trees	0.50686	0.2942	0.88467	5	0.613	−0.09814
Balsam-fir seedlings	0.45637	0.2281	0.98771	5	0.355	−0.03159
Death Valley shrubs	0.12092	0.0158	0.63798	6	0.201	−0.07597
Acorn barnacles	0.51386	0.2914	0.95646	8	0.465	0.00868
Bitterbrush plants	0.17477	0.0388	0.94654	7	0.375	−0.14760
Savanna herb remains	0.26189	0.0704	0.82012	36	0.210	0.00746
Tornado touch downs	0.49221	0.2673	0.98676	7	0.432	−0.05933

and 1.00041; six for the Death Valley shrubs data, with MC values ranging between 0.28653 and 1.02157; six for the acorn barnacles data, with MC values ranging between 0.69815 and 0.92578; three for the bitterbrush plants data, with MC values ranging between 0.25579 and 0.53265; ten for the savanna herb remains data, with MC values ranging between 0.61415 and 1.01813; and, seven for the tornadoes data, with MC values ranging between 0.41932 and 0.86602. These eigenvectors, uncovered as good filters, tend to represent map patterns with moderate-to-strong levels of positive spatial autocorrelation.

MC and GR values calculated for the filtered Poisson regression residuals, reported in Table 2, are very close to their corresponding expected values, indicating the presence of only trace spatial autocorrelation. In other words, the eigenvector filters capture almost all spatial autocorrelation in the mean response terms. The bitterbrush plants data, which in their raw form exhibit no spatial autocorrelation, appear to be slightly overcorrected by the filtering; but the z-score for the MC is only -0.9, indicating the absence of significant induced spatial dependence.

Scatterplots of observed-versus-predicted values appearing in Fig. 2 reveal conspicuous trends. The Death Valley shrubs scatterplot is the poorest, with relatively little trend; the savanna herb remains scatterplot has several outliers that may be increasing its pseudo-R^2 value. Meanwhile, the acorn barnacles scatterplot has one or two leverage points that may be suppressing its pseudo-R^2 value. Overall, though, the spatial filtering predicted values appear to be good.

3. Conclusions

The spatial filtering methodology outlined in this paper furnishes an alternative, successful way of capturing spatial dependency effects in the mean response term of a Poisson regression, both avoiding the complication of an intractable normalizing factor, and allowing positive spatial autocorrelation to be accounted for. This result is obtained by specifying a Poisson process as being heterogeneous across a geographic landscape. This heterogeneity is patterned through a set of fixed effects regressors that define both the mean and the variance of the Poisson process; these are fixed effects because the eigenvectors remain the same regardless of the realization being studied for a given geographic

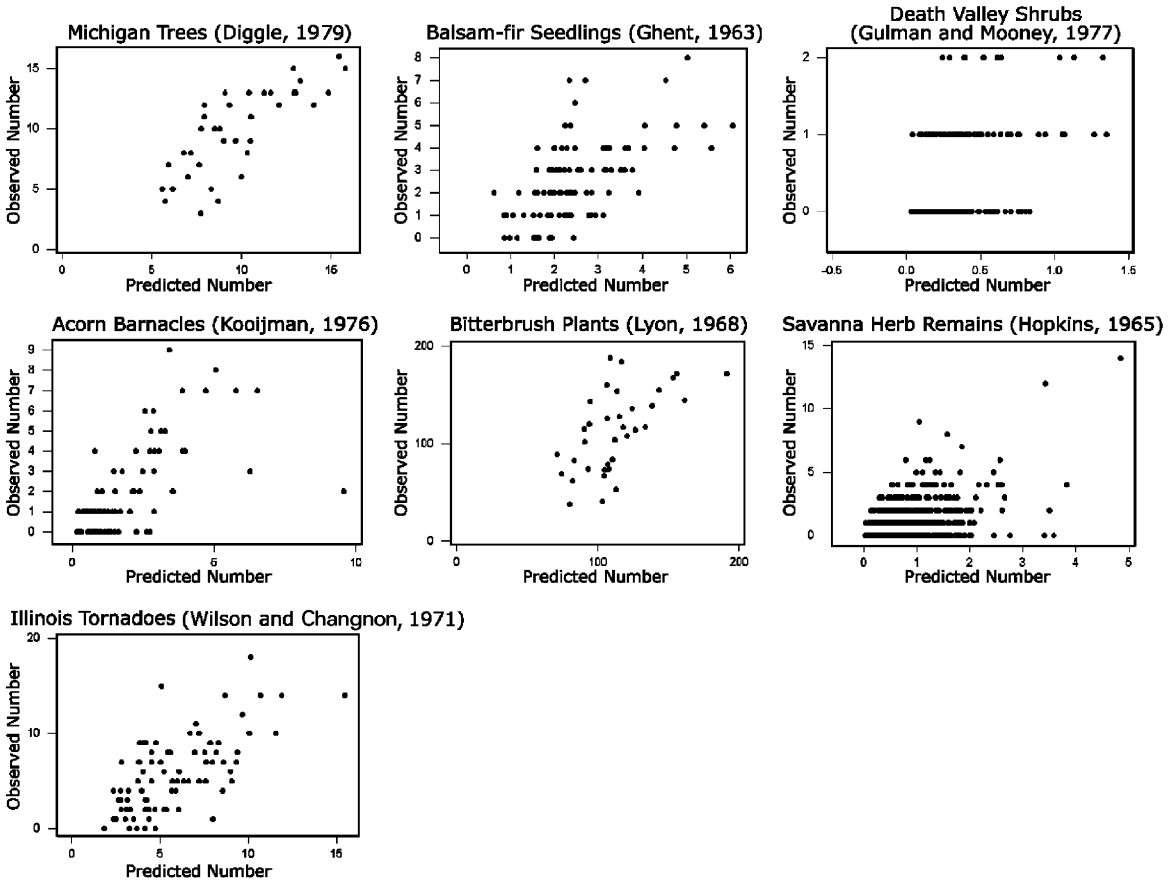

Fig. 2. Scatterplots of observed vs. predicted counts of the presence of phenomena in areal units for selected empirical data sets found in the literature, where the predicted counts are obtained with spatial filter Poisson regression models.

surface partitioning. This result is important for spatial statistics because the Poisson distribution is the cornerstone of quadrat-count modeling.

Acknowledgements

Computational and graphics preparation assistance by Ms. Claire Saint-Rossy is gratefully acknowledged.

References

Diggle, P., 1979. Statistical methods for spatial point patterns in ecology. In: Cormack, R., Ord, J. (Eds.), Spatial and Temporal Analysis in Ecology. International Co-operative Publishing House, Fairland, MD, pp. 95–150.

Getis, A., Griffith, D., 2002. Comparative spatial filtering in regression analysis. Geogr. Anal., 34, 130–140.

Ghent, A.W., 1963. Studies of regeneration of forest stands devastated by the spruce budworm. Forest Sci. 9, 295–310.
Gulmon, S., Mooney, H., 1977. Spatial and temporal relationships between two desert shrubs *Atriplex hymnelytra* and *Tidestromia oblongifolia* in Death Valley, California. J. Ecol. 65, 831–838.
Hopkins, B., 1965. Observations on Savanna burning in the Olokemeji Forest Reserve, Nigeria. J. Appl. Ecol. 2, 367–381.
Kaiser, M., Cressie, N., 1997. Modeling Poisson variables with positive spatial dependence. Statist. Probab. Lett. 35, 423–432.
Kooijman, S., 1976. Some remarks on the statistical analysis of grids especially with respect to ecology. Ann. Systems Res. 5, 113–132.
Lyon, L., 1968. An evaluation of density sampling methods in a shrub community. J. Range Movement 21, 16–20.
Tietjen, G., 1986. A Tropical Dictionary of Statistics. Chapman & Hall, New York.
Upton, G., Fingleton, B., 1989. Spatial Data Analysis By Example, Vol. 2. Wiley, New York.
Wilson, J., Changnon, S., 1971. Illinois Tornadoes. Illinois State Water Survey Circular #103.