Chapter 4
Frequency Distributions for Simulated Spatially Autocorrelated Random Variables

4.1 Introduction

Often quantitative data analysis begins with an inspection of attribute variable histograms. Ratio scale demographic variables, such as population density (which has a natural, meaningful absolute 0 value), are expected to conform, at least approximately, to a normal probability distribution. Frequently this conformity requires that these variables be subjected to a symmetricizing, variance stabilizing transformation, such as the Box-Cox class of power functions or the Manley exponential function. Counts (i.e., aggregated nominal measurement scale) data used to construct ratios, such as the crude fertility rate (i.e., number of births per number of women in the child bearing age cohort), are expected to conform to a Poisson probability distribution. And, counts data that constitute some subset of a total, such as the percentage of people at least 100 years of age or the percentage of a population that is the women in the child bearing age cohort, are expected to conform to a binomial probability distribution. Until the advent of implemented generalized linear models (GLMs), these latter two categories of data also were subjected to variable transformations in order to secure normal probability distribution approximations. Various scholars today argue that GLM technology has made the use of such previously popular variable transformations as the square root for Poisson counts, or the arcsine for percentages, obsolete.

Most spatial statistical work to date addresses impacts of spatial autocorrelation (SA) on parameter estimates, with the general conclusion that positive SA tends to have little or no impact on first moment types of parameter estimates, while inflating their respective standard errors. SA also tends to improve model prediction capabilities, serving remarkably well as a surrogate for missing covariates displaying particular geographic map patterns. This result implies that as SA in a random variable (RV) increases, its tails should become heavier and its center should become flatter. Dutilleul and Legendre (1992) appear to be about the only researchers to systematically investigate this topic, although they do so in a rather pseudo-geographic context.

As is widely acknowledged, positive SA is a source of variance inflation for normal RVs, and a source of overdispersion (i.e., excess variance) for Poisson and binomial RVs. But how does this increased variation impact upon a variable's

D.A. Griffith, J.H.P. Paelinck, *Non-standard Spatial Statistics and Spatial Econometrics*, Advances in Geographic Information Science 1,
DOI 10.1007/978-3-642-16043-1_4, © Springer-Verlag Berlin Heidelberg 2011

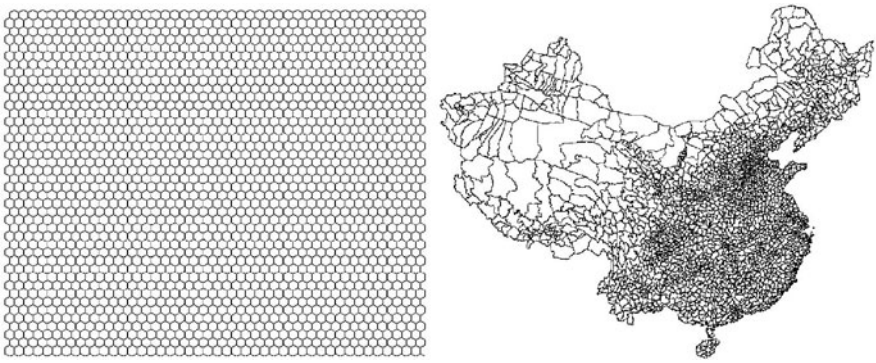

Fig. 4.1 Surface partitionings used for simulation work. *Left* (**a**): a 44-by-54 regular hexagonal tessellation forming a rectangular region. *Right* (**b**): the China county outline map

histogram? This is the question addressed in this chapter. Intuitively speaking, variance increases as increasingly extreme values (i.e., outliers) appear in a histogram. SA-generated heavy tails in a normal distribution are consistent with this data feature. But a binomial RV cannot have extreme values, since its values are constrained by given totals, so that percentages always are contained in the closed interval [0, 100]. A Poisson RV can have extreme large counts; its extreme small counts, however, can only become excessive zeroes. In other words, is some of the quite bothersome noise in or potential dirtiness of data researchers routinely encounter simply a manifestation of SA?

This chapter demonstrates positive SA impacts upon histograms with illustrations based upon simulated data. These data are generated both with autoregressive and with spatial filter (SF; see Sect. 2.5) models (Griffith, 2000a, 2002a, 2004a). Autoregressive models more explicitly focus on SA arising from spatial interaction, whereas SF models more explicitly focus on SA arising from missing variables with specific map patterns—here these map patterns have been selected to represent global, regional, and local spatial effects (Borcard and Legendre, 2002; Borcard et al., 2004). The primary difference is between a variance and a mean response specification that captures SA effects. Furthermore, SF models enable much greater degrees of SA to be explored, primarily because autoregressive models tend to encounter such problems as phase transitions when positive SA becomes excessively strong (Guyon, 1995). The simulated data, which is for an ideal 44-by-54 [n = 2,376; maximum Moran Coefficient (MC_{max}) of 1.02239] regular hexagonal tessellation (Fig. 4.1a), also is supplemented by simulations for the irregular China county surface partitioning (Fig. 4.1b).

4.2 The Normal Probability Model

Haining et al. (1983) outline a technique, in keeping with normal theory in multivariate statistics, to simulate spatially autocorrelated normal RVs with, for example, the simultaneous autoregressive (SAR; Cliff and Ord, 1973) model. A recent approach

sharing a number of the features of their procedure is furnished by Gneiting et al. (2005). Goodchild (1980) offers an alternative procedure that involves permuting independent and identically distributed (iid) values over a map until a prespecified level of SA is attained. Goodchild's method is employed here to remove any conspicuous spurious SA from the simulated data. However, it cannot be used to explore SA impacts upon histograms because histograms are completely insensitive to the locational arrangement of values, simulated or actual, on a map. Furthermore, the resulting observed map would need to have its underlying iid counterpart uncovered in order to explore SA effects.

In keeping with linear statistical models theory, eigenvector-based spatial filtering offers a striking alternative mechanism for simulating spatially autocorrelated normal RVs (see Boots and Tiefelsdorf, 2000, p. 327; Griffith, 2000, p. 146). This technique still begins with a set of iid values.

4.2.1 Simulating Spatially Autocorrelated Normal RVs

The simulated iid normal RV, say n-by-1 vector $\boldsymbol{\varepsilon}$, displays ideal properties (see Fig. 4.2 and Table 4.1). All levels of SA have been embedded into this RV.

Consider a surface that is partitioned into n mutually exclusive and collectively exhaustive areal units. Here these units are regular hexagons forming a 44-by-54 rectangular region (see Fig. 4.1a), or the counties into which China is divided (see Fig. 4.1b). The n-by-n binary geographic connectivity matrix $\mathbf{C}$ contains the elements $c_{ij} = 1$ if areal units (e.g., hexagons, counties) i and j share a common boundary, and $c_{ij} = 0$ otherwise; $c_{ii} = 0$ by construction (i.e., an areal unit cannot be spatially autocorrelated with itself). This definition of matrix $\mathbf{C}$ highlights the reason for selecting a hexagonal surface partitioning as the ideal surface, namely the

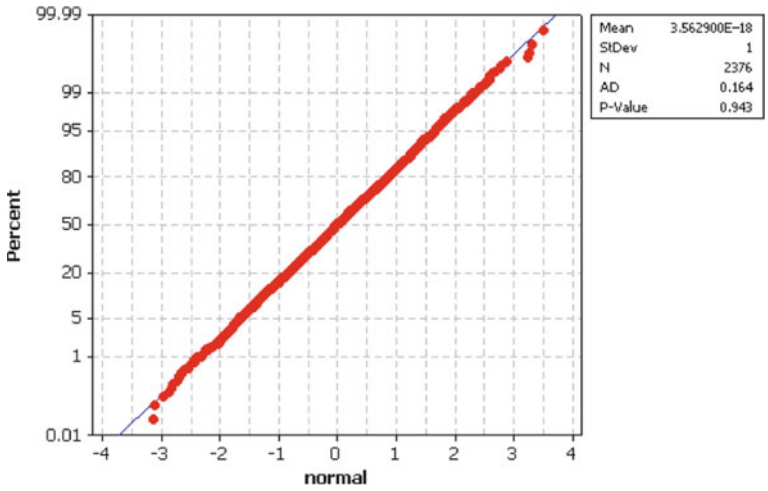

Fig. 4.2 Normal quantile plot for the simulated iid normal RV $\boldsymbol{\varepsilon}$ values

Table 4.1 Descriptive statistics for the SAR model-based simulated data and the hexagonal tessellation geographic configuration

Variable autocorrelation	MC/MC_{max}	GR	$\bar{y}$	s_y	$\lvert z_{skewness} \rvert$ [a]	$\lvert z_{kurtosis} \rvert$ [a]
None (i.e., iid)	–0.01	1.00	–0.000	1.000	0.20	0.30
Weak	0.11	0.89	–0.000	1.025	0.40	0.30
Low-moderate	0.40	0.59	–0.001	1.281	1.19	0
High-moderate	0.60	0.37	–0.006	1.711	1.59	1.29
Strong	0.90	0.07	–0.092	4.707	5.97	4.08

[a]The mean of skewness and kurtosis is 0; the respective standard errors, which can be established using the moment generating function $e^{\mu t+(\sigma^2/2)t^2}$, respectively are $\sqrt{6/n}$ and $\sqrt{24/n}$

lack of areal units sharing only a common point (i.e., a non-zero length boundary)—the difference between rook's and queen's adjacencies, using analogies with chess moves, in the spatial weights matrix literature.

Next, following Haining et al. (1983), matrix $\mathbf{C}$ was converted to its row-standardized version, matrix $\mathbf{W}$, by dividing each c_{ij} value by its row sum $\left(\text{i.e., } \sum_{j=1}^{n} c_{ij}\right)$. Then spatially autocorrelated variables were constructed with the simultaneous autoregressive (SAR)-based equation

$$\mathbf{Y}_j = (\mathbf{I} - \rho_j \mathbf{W})^{-1} \boldsymbol{\varepsilon}, \tag{4.1}$$

where $\mathbf{I}$ is the n-by-n identity matrix, and the SA parameter ρ_j was assigned the values 0.30, 0.73, 0.88, and 0.987 (i.e., $j = 1, 2, 3, 4$) in order to secure the relative MC and Geary Ratio (GR) values reported in Tables 4.1 and 4.3.

Finally, following especially Griffith (2000), the eigenvectors were extracted from matrix

$$(\mathbf{I} - \mathbf{i}\mathbf{i}^T/n)\mathbf{C}(\mathbf{I} - \mathbf{i}\mathbf{i}^T/n), \tag{4.2}$$

where T denotes matrix transpose, and $\mathbf{i}$ denotes an n-by-1 vector of ones. This matrix expression appears in the numerator of the MC. Each eigenvector represents a distinct map pattern with a level of SA indexed by its corresponding eigenvalue. These eigenvectors, and hence map patterns, are both orthogonal and uncorrelated. Then, employing the same random iid vector $\boldsymbol{\varepsilon}$ used to generate the SAR-induced spatially autocorrelated variates, spatially autocorrelated variables were constructed with the SF-based equation

$$\mathbf{Y}_j = \alpha_j \left(\sqrt{n-1}\right) \frac{(a\mathbf{E}_G + b\mathbf{E}_R + c\mathbf{E}_L)/\sqrt{n-1}}{\sqrt{a^2+b^2+c^2}} + \beta\boldsymbol{\varepsilon}, \tag{4.3}$$

where $\mathbf{E}_G/\sqrt{p_k(n-1)}$, $\mathbf{E}_R/\sqrt{p_k(n-1)}$ and $\mathbf{E}_L/\sqrt{p_k(n-1)}$ respectively denote the z-score versions of p_k ($k > 0$) summed global (G), regional (R), and local (L)

eigenvectors, coefficients a, b, and c are weights that enable a particular level of SA to be induced (Boots and Tiefelsdorf, 2000; Griffith, 2000), and here coefficients $\beta = 1$ and $\alpha_j = \sqrt{\frac{MC_j - MC_\varepsilon}{MC_{eigenvectors} - MC_j}}$, for some target value of MC (i.e., MC_j) for variate Y_j, where $MC_{eigenvectors} = \frac{a^2MC_G + b^2MC_R + c^2MC_L}{a^2+b^2+c^2}$ denotes the MC value for a given eigenvector sum. The formula for coefficient α_j assumes that the random error variate and the eigenvectors are uncorrelated.[1] Judiciously selected eigenvectors allow global, regional, and local spatial effects (this interpretation is from Borcard and Legendre, 2002; Borcard et al., 2004) to be simulated with a SF model. The relative MC and GR values obtained with this simulation method are reported in Tables 4.2 and 4.4.

4.2.2 Simulation Results for an Ideal Regular Hexagonal Surface Partitioning

Summary descriptive statistics appear in Table 4.1 for the SAR-induced simulated spatially autocorrelated data. These statistics confirm that the mean essentially is unaffected, while the variance is inflated, by SA. Within the moderate SA range, representing a preponderance of empirical studies to date, variance inflation is problematic, increasing as much as nearly 300%. The histograms appearing in Fig. 4.3 confirm the expectations that low levels of SA have little effect, whereas high levels of SA thicken the tails and squash the center of a normal frequency distribution; this trend is less noticeable with SAR models (see Fig. 4.4). But this latter outcome primarily is because of variance inflation.

Somewhat less noticeable skewness and kurtosis features are better portrayed by inspecting standardized normal curves. Z-score test statistics reported in Table 4.1 reveal that skewness and kurtosis increase as positive SA increases. Skewness becomes more problematic because, similar to a product moment correlation coefficient, the SAR SA parameter is restricted to be < 1, causing a truncation effect in the distribution of values. A surprising outcome is best seen when $MC = 0.90$: as SA becomes strong, not only do the tails become thicker, but values become more concentrated about 0 (the mean), resulting in a relative decrease in the number of intermediate values (the histogram columns are shrinking away from the normal curve outline toward the horizontal axis). This squashing toward the center of the distribution increases kurtosis. Moreover, SA produces more extreme and more near-zero values.

[1] The correlations between the simulated random normal variate and the sum of two eigenvectors representing global map, two representing regional, and two representing local map patterns used to construct Table 2 respectively are 0.031, 0.018 and –0.004—essentially 0.

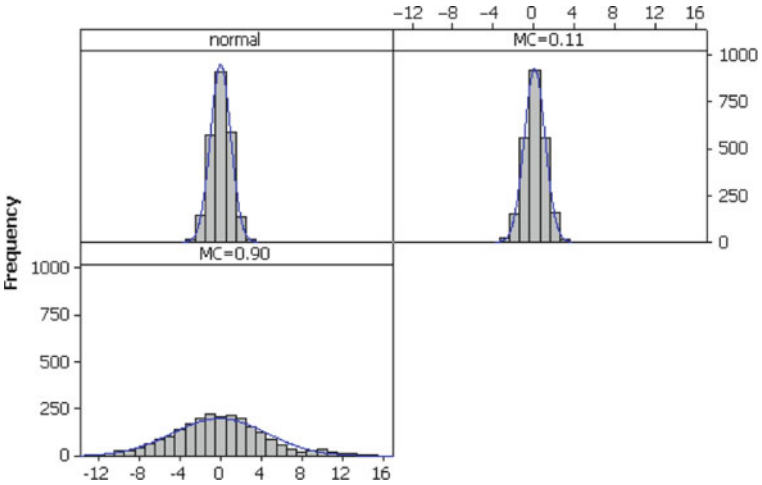

Fig. 4.3 Histograms for iid and two SAR model-induced extreme levels of SA

Because impacts beyond that of variance inflation are difficult to detect visually in the histograms themselves, normal quantile plots also can be inspected (see Fig. 4.5). These plots help highlight the noted changes in the histograms.

Summary descriptive statistics appearing in Table 4.2 for the SF-simulated spatially autocorrelated data (see Fig. 4.6 for maps of the global, regional, and local map patterns employed; as SA decreases in strength, the map patterns become more fragmented) distributed over the hexagonal surface partitioning corroborate findings gleaned from Table 4.1. These statistics confirm that the mean essentially is unaffected, while the variance is inflated, by SA. Again variance inflation is problematic within the moderate SA range, increasing anywhere from 1- to 10-fold, depending upon the mixture of global, regional, and local map patterns. The quantile plots appearing in Fig. 4.7 confirm the expectations that low levels of SA have little effect, whereas high levels of SA thicken the tails and squash the center of a normal frequency distribution—as before, SA produces more extreme and more near-zero values—with this trend being less noticeable with moderate levels of SA. The central tendency concentration is more conspicuous with the global SF-based results than with the preceding SAR-based results. To some degree, local sources of SA seem to dampen more extreme impacts of regional sources, whereas local and regional sources of SA seem to dampen more extreme impacts of global sources. However, a mixture of map patterns—the more common case in practice—appears to produce a more marked impact on variance inflation for moderate levels of SA, but without noticeably affecting kurtosis. Once again, z-score test statistics reported in Table 4.2 reveal that as positive SA becomes marked, kurtosis—but not skewness—increases, with global sources of SA causing the most significant change in kurtosis.

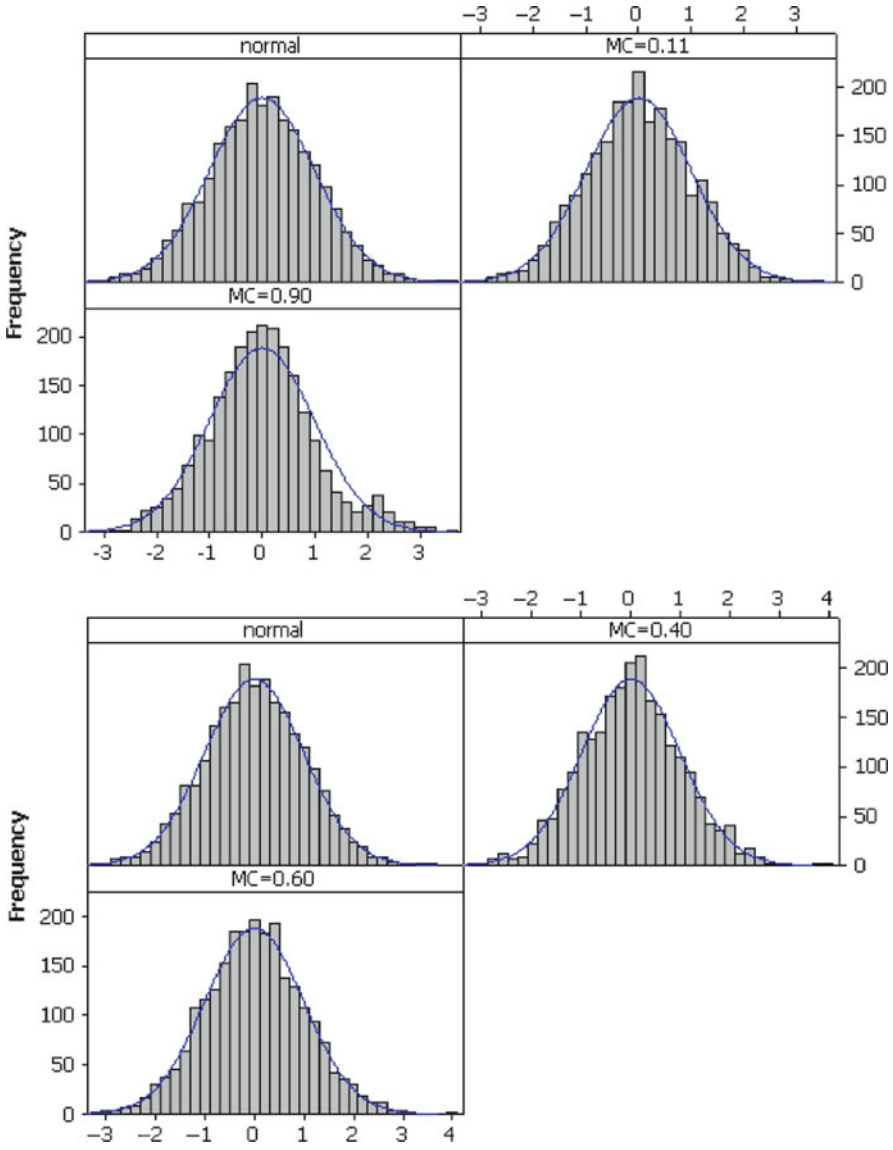

Fig. 4.4 Standard normal deviate histograms for iid and four SAR model-induced levels of SA. *Left* (**a**): induced extreme levels. *Right* (**b**): induced moderate levels

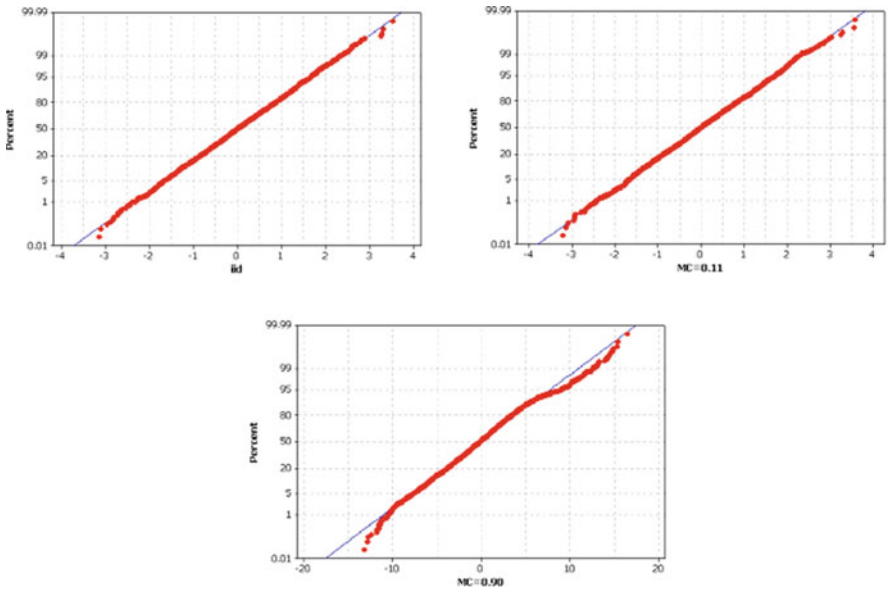

Fig. 4.5 Normal quantile plots for SAR model-induced levels of SA. *Left* (**a**): iid. *Middle* (**b**): weak SA. *Right* (**c**): strong SA

Table 4.2 Descriptive statistics for the SF model-based simulated data and the hexagonal tessellation geographic configuration

Variable autocorrelation	MC	GR	$\bar{y}$	s_y	$\lvert z_{skewness} \rvert$	$\lvert z_{kurtosis} \rvert$
None (i.e., iid)	–0.01	1.00	–0.000	1.000	0.20	0.30
Global map pattern-base results						
Weak (using $0.35\mathbf{E}_G$)	0.11	0.89	–0.000	1.060	0.40	0.50
Low-moderate (using $0.85\mathbf{E}_G$)	0.42	0.59	–0.000	1.306	0.60	0.20
High-moderate (using $1.25\mathbf{E}_G$)	0.62	0.40	–0.000	1.600	0.80	0.90
Strong (using $3.00\mathbf{E}_G$)	0.92	0.10	–0.000	3.181	0.40	3.58
Global + regional map pattern-base results						
Weak (using $0.50\mathbf{E}_R$)	0.10	0.89	–0.000	1.126	1.19	0.70
Low-moderate [using $0.75(\mathbf{E}_G+\mathbf{E}_R)$]	0.40	0.60	–0.000	1.452	1.99	1.39
High-moderate [using $1.33(\mathbf{E}_G+\mathbf{E}_R)$]	0.60	0.41	–0.000	2.123	1.79	0.20
Strong [using $2.50(3\mathbf{E}_G+\mathbf{E}_R)$]	0.91	0.11	–0.000	4.112	0.80	3.08
Global + regional + local map pattern-base results						
Weak (using $0.85\mathbf{E}_L$)	0.10	0.90	–0.000	1.297	0.20	0.10
Low-moderate [using $1.80(1.5\mathbf{E}_R+\mathbf{E}_L)$]	0.40	0.60	–0.000	3.403	0.40	0.99
High-moderate [using $1.10(1.5\mathbf{E}_G+\mathbf{E}_R+\mathbf{E}_L)$]	0.61	0.41	–0.000	2.444	1.59	0.50
Strong [using $1.00(5\mathbf{E}_G+2\mathbf{E}_R+\mathbf{E}_L)$]	0.90	0.12	–0.000	5.624	0.80	3.18

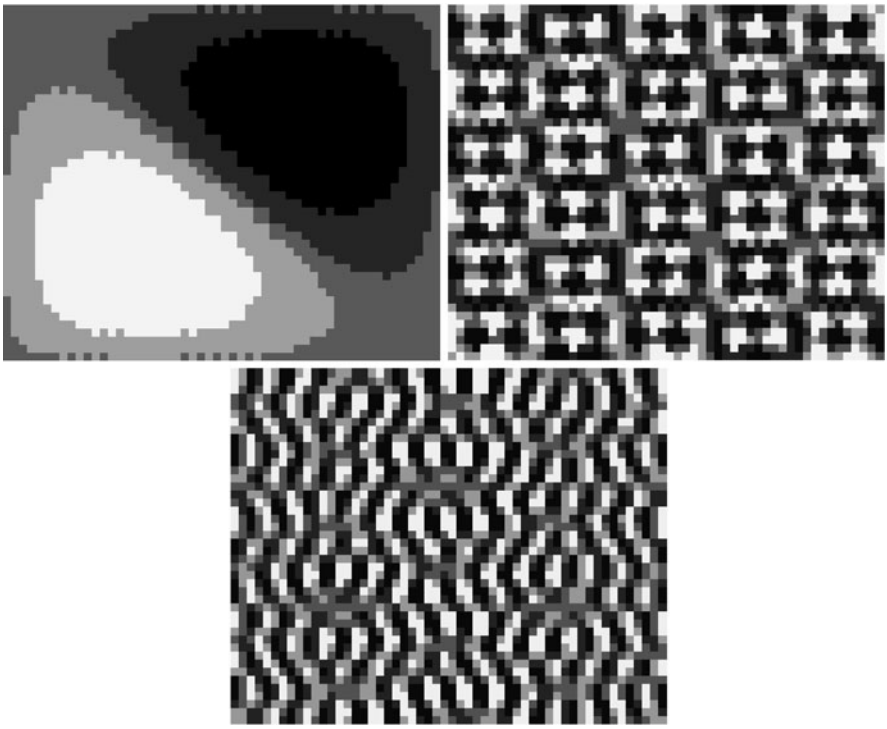

Fig. 4.6 SF map patterns. *Left* (**a**): global map pattern [the sum of eigenvectors # 1 and #2 ($\div\sqrt{2}$)]. *Middle* (**b**): regional map pattern [the sum of eigenvectors #365 and # 366 ($\div\sqrt{2}$)]. *Right* (**c**): local regional map pattern [the sum of eigenvectors a #1532 and #1533 ($\div\sqrt{2}$)]

4.2.3 Simulation Results for the China County Geographic Configuration

The simulated data coupled with the China county geographic configuration (see Fig. 4.1b) includes the 2,376 values used for the regular hexagonal tessellation simulation together with three additional values that were carefully selected so that the descriptive statistics appearing in Table 4.1 and the normal quantile plot appearing in Fig. 4.2 essentially remain unchanged. Once again the SA parameter ρ_j has taken on the values 0.30, 0.78, 0.93, and 0.986 (at this point a phase transition is encountered), rendering the relative MC and GR values reported in Table 4.3. The goal here is to explore impacts in terms of an irregular lattice surface partitioning.

Summary descriptive statistics appear in Table 4.3 for the SAR-induced simulated spatially autocorrelated data. These statistics confirm that the mean essentially

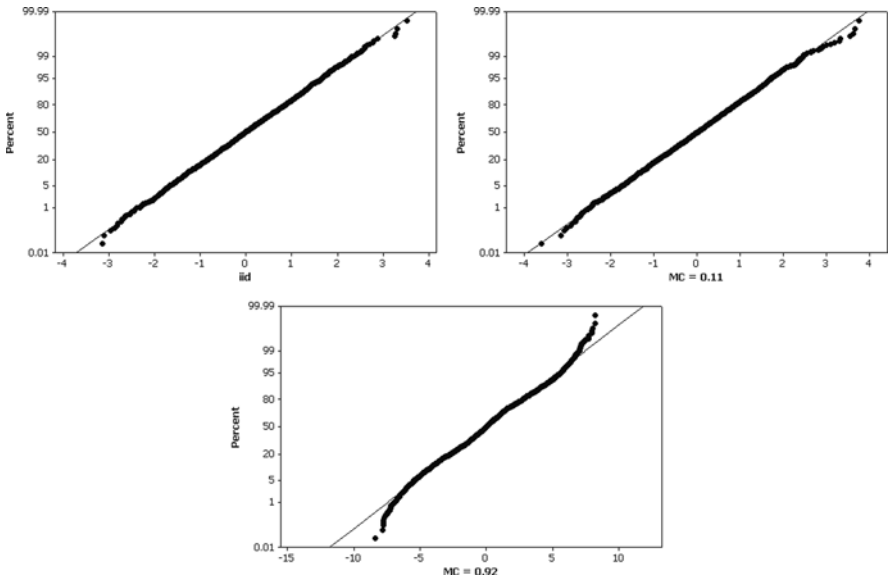

Fig. 4.7 Normal quantile plots for global map pattern SF model-induced levels of SA. *Left* (**a**): iid. *Middle* (**b**): weak SA. *Right* (**c**): strong SA

Table 4.3 Descriptive statistics for the SAR model-based simulated data and the China county geographic configuration

Variable autocorrelation	MC/MC_{max}	GR	$\bar{y}$	s_y	$\lvert z_{skewness} \rvert$	$\lvert z_{kurtosis} \rvert$
None (i.e., iid)	–0.00	1.00	0.000	1.000	0.20	0.30
Weak	0.11	0.86	–0.001	1.031	0.20	0.20
Low-moderate	0.41	0.49	–0.008	1.426	0.00	0.50
High-moderate	0.61	0.24	–0.037	2.253	1.00	3.78
Strong	0.73	0.07	–0.222	4.760	10.75	61.33

is unaffected, while the variance is inflated, by SA. Variance inflation is problematic within the moderate range, increasing as much as nearly 500%. The normal quantile plots appearing in Fig. 4.8 again confirm the expectations that low levels of SA have little effect, whereas high levels of SA thicken the tails and squash the center of a normal frequency distribution. In this case, pronounced levels of SA interact with the irregularness of the surface partitioning to result in the generation of rather dramatic extreme values.

As mentioned previously, less noticeable skewness and kurtosis features are better portrayed by inspecting standardized normal curves. Z-score test statistics reported in Table 4.3 reveal that as positive SA increases, so do skewness and

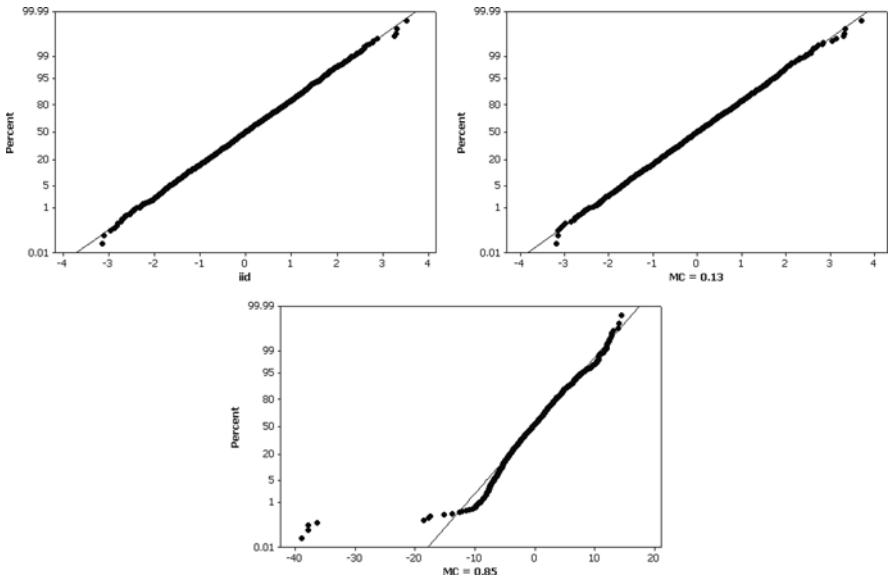

Fig. 4.8 Normal quantile plots for SAR model-induced levels of SA. *Left* (**a**): iid. *Middle* (**b**): weak SA. *Right* (**c**): strong SA

kurtosis. In part, skewness becomes more problematic because of the irregularness of the underlying county geographic configuration. As before, relatively strong levels of SA are accompanied by not only thicker tails, but also values that are more concentrated about 0 (the mean), resulting in a relative decrease in the number of intermediate values (the histogram columns shrink away from the normal curve outline toward the horizontal axis), with this squashing toward the center of the distribution increasing kurtosis.

SF induced SA coupled with the China county geographic configuration, based upon mixtures of global, regional (two levels, R-1 and R-2), and local map pattern eigenvectors render the MC and GR values reported in Table 4.4 (see Fig. 4.9 for maps of the global, two regional, and local map patterns employed here). A lack of impact upon the mean as well as variance inflation continue to characterize these variables. But histogram distortions affiliated with the underlying histogram for the global trend dominate the skewness and kurtosis modifications arising from positive SA. The example normal quantileplots appear in Fig. 4.10 (histograms portray a situation of more extreme values materializing under strong positive SA; a denser concentration about the mean still occurs). Here histogram distortions already become quite apparent at moderate levels of positive SA. Again these tendencies are more apparent visually by inspecting the corresponding standard normal curves.

Table 4.4 Descriptive statistics for the global SF model-based simulated data and the China county geographic configuration

Variable autocorrelation	MC	GR	$\bar{y}$	s_y	$\lvert z_{skewness} \rvert$	$\lvert z_{kurtosis} \rvert$
None (i.e., iid)	–0.00	1.00	0.000	1.000	0.20	0.30
Global map pattern-base results						
Weak (using $0.33\mathbf{E}_G$)	0.11	0.91	0.000	1.052	1.19	0.20
Low-moderate (using $0.75\mathbf{E}_G$)	0.40	0.67	0.000	1.254	8.56	6.17
High-moderate (using $1.10\mathbf{E}_G$)	0.61	0.50	0.000	1.484	15.73	16.13
Strong (using $2.10\mathbf{E}_G$)	0.90	0.26	0.000	2.296	28.47	38.33
Global + regional map pattern-base results						
Weak (using $0.55\mathbf{E}_{R-1}$)	0.10	0.90	0.000	1.136	0.20	0.30
Low-moderate [using $1.00(\mathbf{E}_{R-1}+\mathbf{E}_{R-2})$]	0.41	0.63	0.000	1.758	0.40	0.00
High-moderate [using $1.10(\mathbf{E}_G+\mathbf{E}_{R-1}+\mathbf{E}_{R-2})$]	0.60	0.48	0.000	2.141	4.18	4.48
Strong [using $2.20(\mathbf{E}_G+\mathbf{E}_{R-1})$]	0.90	0.23	0.000	6.229	9.96	17.03
Global + regional + local map pattern-base results						
Weak [using $0.55(\mathbf{E}_{R-1}+\mathbf{E}_L)$]	0.10	0.87	0.000	1.25	1.39	0.70
Low-moderate [using $1.80(\mathbf{E}_{R-1}+\mathbf{E}_{R-2}+\mathbf{E}_L)$]	0.40	0.59	0.000	3.30	0.60	1.10
High-moderate [using $1.375(1.25\mathbf{E}_G+\mathbf{E}_{R-1}+\mathbf{E}_{R-2}+\mathbf{E}_L)$]	0.60	0.46	0.000	3.10	5.77	5.38
Strong [using $1.30(3\mathbf{E}_G+\mathbf{E}_{R-1}+\mathbf{E}_{R-2}+\mathbf{E}_L)$]	0.90	0.25	0.000	4.60	23.10	30.27

4.2.4 Implications

The conceptual discussions allow expectations to be posited with regard to impacts of SA on histograms of normal RVs. In absolute terms, variance inflation generated by positive SA makes a histogram appear flatter. Positive SA also encourages more extreme values (thickening of the tails) to materialize.

The principal implications for normal RVs are: (1) positive SA generates variance inflation, which flattens a frequency distribution; (2) kurtosis tends to be dramatically altered when positive SA becomes very strong; and, (3) tail thickening and

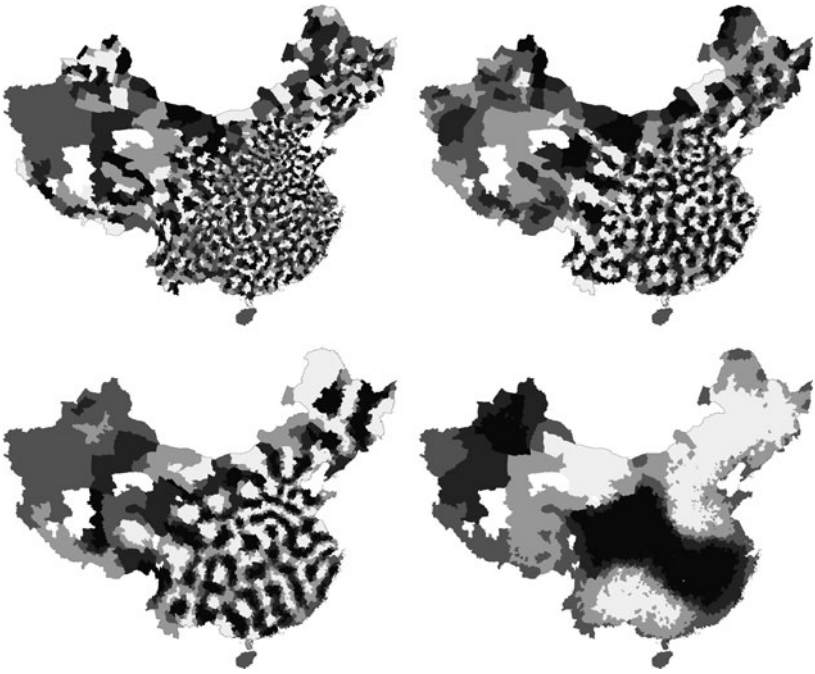

Fig. 4.9 SF map patterns. *Top left* (**a**): local (MC = 0.11). *Top right* (**b**): regional (MC = 0.47). *Bottom left* (**c**): regional (MC = 0.73). *Bottom right* (**d**): global (MC = 1.11)

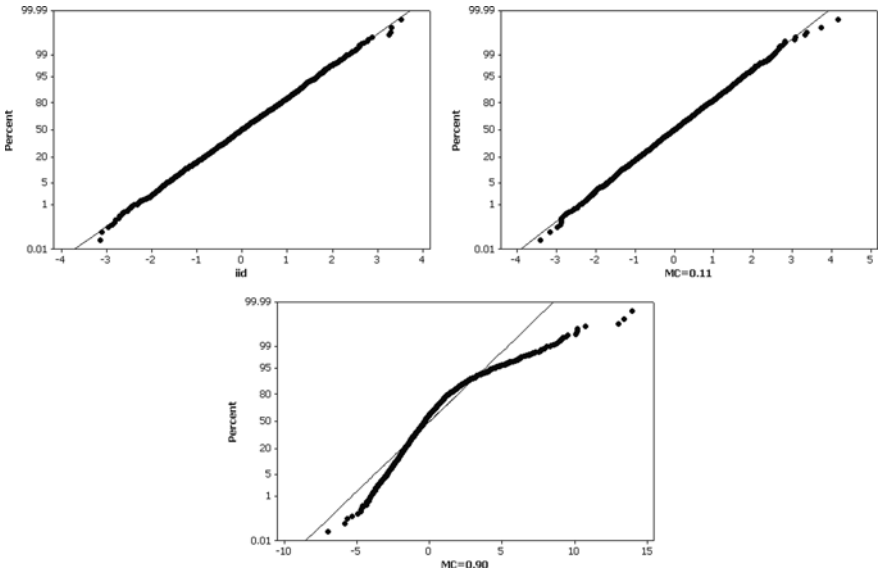

Fig. 4.10 Normal quantile plots for global map pattern SF model-induced levels of SA. *Left* (**a**): iid. *Middle* (**b**): weak SA. *Right* (**c**): strong SA

variance inflation are problematic in the moderate positive SA range that often is encountered in real world data.

4.3 The Poisson Probability Model

Little is known about the impacts of SA on Poisson RVs. Because this type of RV is a member of the exponential family of statistical distributions, just like for a normal RV, positive SA should induce variance inflation in Poisson RVs, too. This expectation is further supported by the close similarity between a normal and a Poisson frequency distribution when the latter's mean, μ, becomes very large. Thus, one should expect that positive SA will create extra-Poisson variation, a notion consistent with discussions in the overdispersion literature. But what happens to the mean of a Poisson RV?

One way that a Poisson RV differs from a normal RV is that its lower tail is truncated at 0. A Poisson RV describes counts of rare events, which naturally yields many zeroes as the event in question becomes increasingly rarer. Accordingly, the best way for Poisson variance to increase, then, is for extremely large counts to materialize, and/or perhaps for an over-concentration of zero or near-zero values to occur (i.e., excessive zeroes) to balance very large values in order to preserve μ. But what happens to the kurtosis of a Poisson RV? And, because it is a discrete RV (whereas a normal RV is continuous over the entire real number line), what happens to its modal value?

A bivariate regression tool for evaluating the Poissonness of a distribution, which is analogous to a normal quantile plot, is the Poissonness plot (Hoaglin, 1980; Hoaglin and Tukey, 1985). The ideal line for this plot, which can be estimated with ordinary bivariate linear regression techniques, is given by

$$\ln(n_k) + \sum_{j=1}^{k} \ln(j) - \ln\left(\sum_{i=1}^{n} y_i\right) = \beta_0 + \beta_1 k\,,\ k = 0, 1, 2, \ldots, \tag{4.4}$$

where k denotes the discrete non-negative values taken on by some Poisson RV Y, n_k is the count for discrete value k in a dataset, and $\beta_0 = -\mu$ and $\beta_1 = \ln(\mu)$, where μ is the mean of Y; of note is that the term $\sum_{j=1}^{k} \ln(j)$ disappears for $k = 0$ (which corresponds to a $0! = 1$ term, whose logarithm is 0). The left-hand side of this equation is referred to as the metaparameter. The Ord plot (Ord, 1967) furnishes an additional assessment tool. For this second regression analysis, which involves weighted least squares (WLS) estimation, the equation is given by

$$\frac{k\, n_k}{n_{k-1}} = \beta_0 + \beta_1\, k, \tag{4.5}$$

where $\beta_0 = \mu$ and $\beta_1 = 0$ for a Poisson distribution, and the weights are $\sqrt{n_k - 1}$. As a benchmark, a judiciously selected ideal set (n = 2,376) of independent and

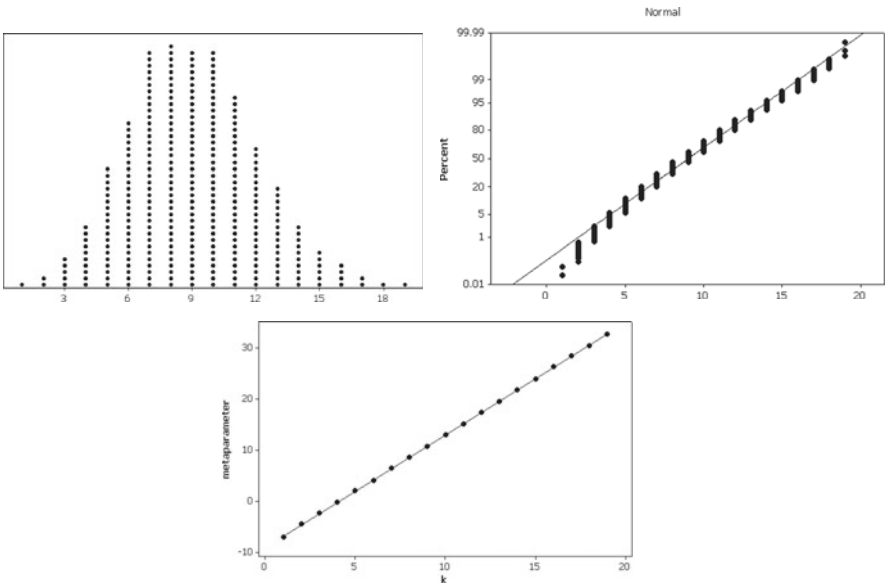

Fig. 4.11 Graphical diagnostic tools for a Poisson RV. *Left* (**a**): a dot-plot histogram. *Middle* (**b**): a normal quantile plot. *Right* (**c**): a Poissonness plot

identically distributed (iid) Poisson RVs with $\mu = 9$ was simulated; the mean and standard deviation for this set are $\bar{y} = 8.99790$ and $s_y = 3.00701$. Graphics portraying it appear in Fig. 4.11. Normal curve theory states that as μ increases beyond some sufficiently large value (e.g., 1,000), a Poisson probability distribution increasingly resembles a normal probability distribution, a feature that already is becoming visible in Fig. 4.11a. But with a mean of only 9, the normal quantile plot (Fig. 4.11b) confirms that the frequency distribution for this simulated Poisson RV deviates substantially from mimicking the form of a bell-shaped curve, particularly in its tails. A Poissonness plot (Fig. 4.11c) confirms that this is a Poisson RV. Its accompanying regression equation yields $\hat{\beta}_0 = -9.01112$ and $e^{\hat{\beta}_1} = 9.00932$. Meanwhile, the Ord plot results in bivariate linear WLS estimates of $\hat{\beta}_0 = 9.00614$ and $\hat{\beta}_1 = 0.00273$. All of these Poisson diagnostics confirm that this is a Poisson RV. In addition, when distributed across the 44-by-54 regular hexagonal surface partitioning employed in this study, this RV yields MC = 0.00143 ($z_{MC} = 0.15$) and GR = 0.99698; at most it contains only a trace amount of positive SA.

One spatial autoregression[2] theoretical statistical difficulty is that the auto-Poisson model can handle only negative SA; this drawback is problematic both because most georeferenced Poisson-distributed data contain positive SA, and because the normal and binomial approximations to an auto-Poisson model can

[2]Auto- models have values of the response variable, Y, on both sides of the equation. The right-hand side, which relates to a probability model, contains a linear combination of values of Y for other than the observation in question.

account for this positive SA. But when avoiding specification error, neither a normal nor a binomial approximation to a Poisson RV is desirable. Fortunately, Kaiser and Cressie (1997) and Griffith (2002) suggest two different ways that positive SA can be accommodated—and hence simulated—in a Poisson RV. The first of these methods truncates the auto-Poisson distribution and employs Markov chain Monte Carlo (MCMC) techniques, whereas the second employs spatial filtering techniques. A distinct difference between these two specifications is that the truncated auto-Poisson version can capture at most weak-to-moderate levels of positive SA (e.g., also see Augustin et al., 2004), whereas the SF version can capture even very strong levels (e.g., see Haining et al., 2009).

4.3.1 Simulating Spatially Autocorrelated Poisson RVs

Kaiser and Cressie (1997) circumvent the negative SA limitation of an auto-Poisson specification by Winsorizing counts to a finite set of integers, which sets an upper limit on the largest count that can occur. This adjustment yields an approximation whose probabilities sum to slightly less than 1, rather than to exactly 1, and allows the following auto-Poisson mean specification to be posited, using matrix notation:

$$\ln(\boldsymbol{\mu}) = \left[\alpha - \ln\left(\frac{1}{n}\sum_{i=1}^{n} e^{\rho\sum_{j=1}^{n} w_{ij}(y_j - e^{\alpha})}\right) \Big/ K\right]\mathbf{i} + \rho\mathbf{W}(\mathbf{Y} - e^{\alpha}\mathbf{i}), \tag{4.6}$$

where α is the regression intercept term ($\mu = e^{\alpha}$ is the mean of the Poisson RV in question), ρ is the spatial autoregression parameter, and this second term corrects for artificial inflation of the intercept term (i.e., an adjustment for trend)—K takes on the value of 1 until the mean begins to explode (see Augustin et al., 2004), at which point it increases to further compensate for this explosion. Equation (4.6) has a functional form very similar to an SAR model; here because matrix $\mathbf{W}$ mathematically is required to be symmetric, its (i, j) entry is defined as $w_{ij} = c_{ij} \Big/ \sqrt{\left(\sum_{i=1}^{n} c_{ij}\right)\left(\sum_{j=1}^{n} c_{ij}\right)}$.

Explosion of the mean occurs as the autoregressive parameter ρ becomes relatively large. This same outcome can be observed with the SAR model just as ρ approaches the boundary of its feasible parameter space (e.g., see Tables 4.1 and 4.3). Winsorizing the auto-Poisson probability distribution does not control for this explosion of the mean value; rather, it seeks to avoid entering a transition phase of instability, which tends to coincide with this explosion. However, because SA encourages relatively large counts to materialize (with the resulting contrasts with nearby values leading to local negative SA), the truncation point becomes critical. If it is too low, impacts on the mean and variance become more a function of it than of positive SA; if it is too high, phase transitions can be encountered. The two lowest levels of positive SA simulated for Sect. 4.3.2 employed a truncation point

of $3y_{max}$, where y_{max} denotes the maximum count from each MCMC initial set of iid randomly generated counts (i.e., because the maximum count by chance when $\mu = 9$ is approximately 29, this truncation point is 26 deviations above the mean); no truncations had to be performed during chain generation. In contrast, for the highest level of positive SA, this truncation point was set to $6y_{max}$ (i.e., 55 deviations above the mean), resulting in roughly 20 million truncations being performed during chain generation.

4.3.1.1 MCMC Map Simulation

A Markov chain is a stochastic process consisting of a finite number of *states (i.e.*, for a Poisson RV, a vector of length n containing integer-valued counts corresponding to n locations) and known transition probabilities of moving from state i to state j at each computational iteration. Here, the matrix of transition probabilities, **M**, is defined by a Winsorized auto-Poisson model probability mass function. An important part of Markov chain theory is based on the *Ergodicity Theorem*, which requires **M** to be irreducible (i.e., any state can be reached from any other state)—the geographic weights matrix used is irreducible—recurrent non-null (the average return time to a given state is finite), and aperiodic (a state cannot be returned to repeatedly after a specific finite number of transition steps)—each areal unit in a hexagonal tessellation has at most 6 neighbors. If a Markov chain is ergodic, then a unique steady state distribution exists, say $\mathbf{M}^*$, which is independent of the initial state. This steady state distribution is given by $\mathbf{M}^* = \lim_{k \to \infty} \mathbf{M}^k$, where k represents transition steps. Monte Carlo simulation is a technique for obtaining realizations of the limiting steady state distribution of a stochastic process through the use of a Poisson random number generator.

MCMC provides a mechanism for taking *dependent* samples from probability distributions in situations where the usual sampling is difficult, if not impossible. A case in point is where the normalizing constant for a joint probability distribution is either too difficult to calculate or analytically intractable. This is exactly the case for the auto-Poisson model. MCMC is used to simulate from some joint probability distribution **p** known only up to a constant factor, C. That is, $\mathbf{p} = C\mathbf{q}$, where **q** is known but C is unknown and an intractable mathematical expression (see Cressie, 1991, p. 428, for a mathematical statement of C for the auto-Poisson model). MCMC sampling begins with conditional (marginal) probability distributions, and with parameter estimates for the auto-Poisson model that can be obtained in practice using pseudo-likelihood estimation. This exercise involves estimating α and ρ as though observations are independent. MCMC outputs a sample of values for each parameter drawn from the joint probability distribution. Gibbs sampling is a MCMC scheme for simulation from **p** where the Markov chain transition matrix (i.e., **M**) is defined by the n *conditional* probability distributions of **p**. It is a stochastic process that returns a different result with each execution, a method for generating a joint empirical distribution of several variables from a set of modeled conditional distributions for each variable when the structure of data is too complex

to implement mathematical formulae or directly simulate. It is a recipe for producing a Markov chain that yields simulated data that have the correct unconditional model properties, given the conditional distributions of those variables under study (Robert and Casella, 1999). The principal idea behind it is to convert a multivariate problem into a sequence of univariate problems, which then are iteratively solved to produce a Markov chain. The following Gibbs sampling algorithm description (see Haining et al., 2009; Augustin et al., 2004) for a Winsorized auto-Poisson model begins with pre-specified values of the parameters α and ρ (e.g., pseudo-likelihood parameter estimates in the ensuing China data analysis):

Step 1: initialize a map ($\tau = 0$, where τ denotes the number of iterations) by taking $i = 1, \ldots, n$ independent random samples $\{y_{i,\tau=0}\}$ from a Poisson probability distribution and determine y_{max};

Step 2: obtain new values (initially $\tau = 1$) $y_{i,\tau}$ by sequentially moving from one location (i) to another (j) on the initial map and randomly sampling from the Winsorized auto-Poisson probability distribution [i.e., Eq. (4.6) coupled with a truncation value that is a function of y_{max}] using pre-specified parameter values—site selection for this process of obtaining $\{y_{i,\tau=1}\}$ from $\{y_{i,\tau=0}\}$ can follow random permutations of location sequences or simply a systematic sweep across a map;

Step 3: obtain new values (initially $\tau = 2$) $y_{i,\tau+1}$ by sequentially moving from one location to another on the τ^{th} map, again randomly sampling from the Winsorized auto-Poisson distribution, and immediately updating the value at each location; and,

Step 4: repeat Step 3 for iterations $\tau = 3, 4, 5, \ldots$, until convergence of the sufficient statistics of the parameters of interest occurs.

Once a Markov chain transition matrix is constructed, a sample of (correlated) drawings from a target distribution can be obtained. This is done by *simulating* the Markov chain a large number of times (say, 100,000) and recording its sufficient statistics after removing a burn-in set (e.g., the first 25,000) of iterations. Convergence needs to be monitored (e.g., time series plots and correlograms need to be inspected), and hence the sufficient statistics need to be recorded. This recording should be done after each iteration. A suitable burn-in period is needed in order to generate $\mathbf{M}^*$, and hence before collecting statistics, and because iteration outcomes may well be correlated, the chain needs to be weeded (e.g., only every hundredth iteration result is retained).

The sufficient statistics for the estimators of the simple auto-Poisson model parameters here are $1 \times \sum_{i=1}^{n} y_i$ and $\sum_{i=1}^{n} y_i \sum_{j=1}^{n} c_{ij} y_j$; this first statistic is required for a Poisson model intercept term, whereas this second statistic is required for an auto-Poisson model autoregressive parameter term. Once convergence has been attained (e.g., the accompanying trend line for a time series plot is flat, and the accompanying correlogram displays no significant serial autocorrelation), the last map in the chain is the simulated Winsorized auto-Poisson realization.

4.3.1.2 SF Map Simulation

Meanwhile, SF versions of the Poisson model involve specifying a geographically heterogeneous mean and variance in order to capture positive SA. This implementation requires the usual set of covariates, $X_1,\ldots,X_p$, to be replaced by the eigenvectors of matrix expression (4.2) in order to embed SA in a response counts variable. Compared with the auto- models, spatial dependence effects are shifted to the mean, resulting in the spatial autoregressive parameter [i.e., ρ in Eq. (4.6)] being forced to 0. Accordingly, a realization can be obtained by sampling from a Poisson distribution with mean

$$\mathrm{LN}(\boldsymbol{\mu}) = \left[\alpha - \ln\left(\frac{1}{n}\sum_{i=1}^{n} e^{\sum_{k=1}^{K}\mathbf{E}_k\beta_k}\right)\right]\mathbf{1} + \sum_{k=1}^{K}\mathbf{E}_k\beta_k, \tag{4.7}$$

where $\mathbf{E}_k$ denotes eigenvector k of matrix expression (4.2), β_k is its relative weight (somewhat similar to ρ in the autoregressive specification), and this second term corrects for artificial inflation of the intercept term α (i.e., an adjustment for trend) due to the presence of covariates. An additional adjustment for α in the third term is unnecessary here because the mean of each $\mathbf{E}_k\beta_k$ is zero, by construction.

4.3.2 Simulation Results for an Ideal Regular Hexagonal Surface Partitioning

Figure 4.12, which characterizes all three chains, furnishes strong evidence that the generated MCMC chains converged, rendering useful maps with positive SA embedded in them. The time series plots exhibit random stability. For example, for $\rho = 0.06$—the maximum positive SA that was successfully embedded into simulated data here—the trend line has not converged within the burn-in set of iterations, but does converge long before the end of the chain; here this situation is acceptable since only the last map of the chain is used here. Meanwhile, the correlograms reveal that virtually no serial autocorrelation is present in the three chains.

The Poissonness plots for the autoregressive model results appear in Fig. 4.13. These plots begin to exhibit slight but detectable tail disturbances beginning with low-weak positive SA. Moderate positive SA results in a complete deterioration of linearity.

Summary descriptive statistics appear in Table 4.5 for the Winsorized auto-Poisson simulated data containing positive SA. These statistics confirm that the (controlled for trend) mean essentially is unaffected, while the variance is inflated (i.e., overdispersion), quite noticeably by moderate positive SA. Corresponding histograms confirm the expectations that low levels of positive SA have little effect, whereas moderate levels tend to stretch the right-hand tail and shift the concentration of values toward 0. This same pattern is displayed by: the maximum values, the mode, and kurtosis. Plot diagnostic statistics begin detecting deviation from

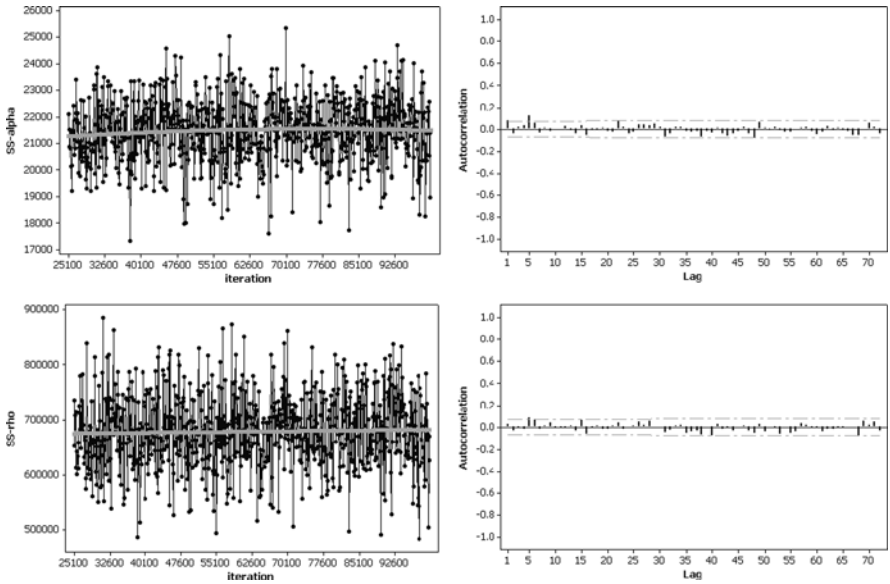

Fig. 4.12 MCMC time series plot and correlogram diagnostic graphics based on the ideal hexagonal surface partitioning when $\rho = 0.06$ (*bottom*). *Left* (**a**): for the intercept term. *Right* (**b**): for the autoregressive term

a Poisson RV at weak levels of positive SA; the Ord plot statistics emphatically detect deviation at the low-moderate level, whereas both sets of these statistics unambiguously detect deviation at the moderate level, of positive SA.

The Poissonness plots for the SF model-embedded positive SA results are illustrated in Fig. 4.14. Not only do these results confirm those found for the Winsorized auto-Poisson model, but the SF model, because it is able to capture much stronger levels of positive SA, extends the autoregressive findings. Furthermore, the corresponding summary descriptive statistics, which appear in Table 4.6, corroborate those trends detected in Table 4.5. Overall, as positive SA increases in a Poisson RV, variance increases, both near-zero and extreme values become more likely, kurtosis increases, and the Ord plot bivariate regression parameter estimates provide a very good diagnostic of its presence, one that furnishes superior diagnostics to those associated with the Poissonness plot.

4.3.3 Simulation Results for the China County Geographic Configuration

As before, MCMC simulation of Winsorized auto-Poisson model-based maps employing the China county irregular surface partitioning at most could embed only moderate positive SA. A bifurcation point appears to be present because of

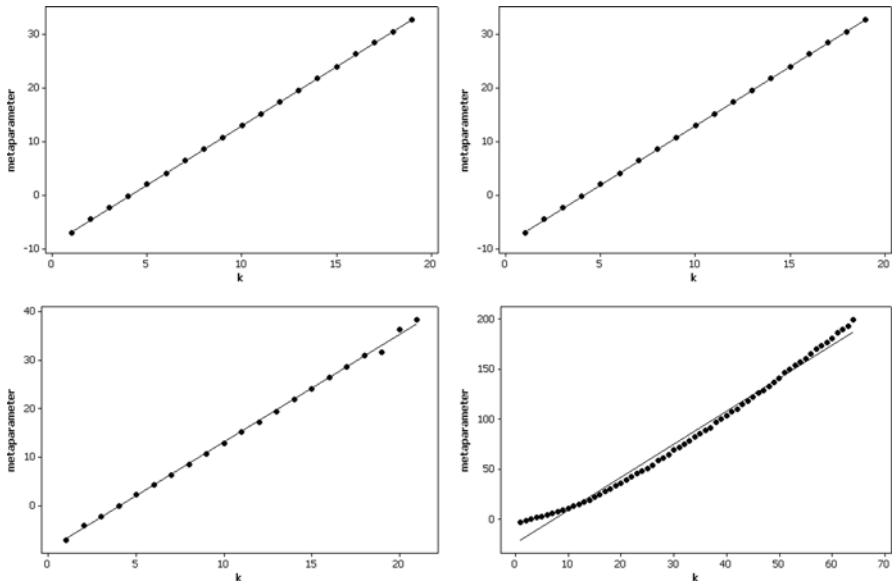

Fig. 4.13 Poissonness plots for the Winsorized auto-Poisson-model induced levels of positive SA based upon the ideal hexagonal surface partitioning. *Left* (**a**): iid with no SA. *Middle left* (**b**): weak positive SA. *Middle right* (**c**): low-moderate positive SA. *Right* (**d**): moderate positive SA

the irregularness of the surface partitioning; MCMC simulation produces maps containing either weak or moderate positive SA, without a transition between them. Nevertheless, graphical diagnostics indicate that the resulting maps are properly generated. In addition, summary descriptive statistics reported in Table 4.7 are consistent with those appearing in Tables 4.5 and 4.6: overdispersion is induced, outliers are generated, relatively small values become more likely, and kurtosis is affected

Table 4.5 Descriptive statistics for the Winsorized auto-Poisson model-based simulated data and the hexagonal tessellation geographic configuration

Variable autocorrelation	MC	GR	$\bar{y}$	s_y	y_{max}	mode	$\lvert z_{kurtosis} \rvert$ [a]	Poissonness plot		Ord plot	
								$-\hat{\beta}_0$	$e^{\hat{\beta}_1}$	$\hat{\beta}_0$	$\hat{\beta}_1$
None (i.e., iid)	0.00	1.00	9.00	3.007	19	8	0.04	9.01	9.01	9.01	0.00
Weak	0.11	0.89	8.99	3.050	21	8	0.19	9.29	9.27	9.88	–0.04
Low-moderate	0.19	0.82	8.90	3.210	21	8	0.23	9.08	9.19	2.48	0.85
moderate	0.47	0.55	9.05	16.477	123	4	67.77	–3.32	1.05	–6.83	2.26

[a]The mean of kurtosis is $1/\mu = 1/9 = 0.111$; the standard error, which can be established using the moment generating function $e^{\mu(e^t - 1)}$, is $\sqrt{151.23594 / n}$ for $\mu = 9$

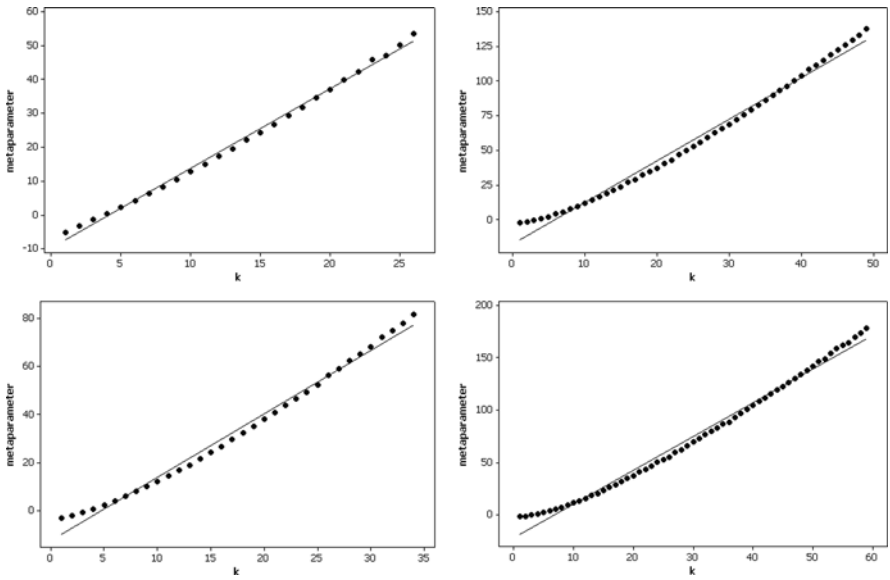

Fig. 4.14 Poissonness plots for SF-model induced positive SA using an ideal hexagonal surface partitioning and a mixture of global, regional and local map patterns. *Left* (**a**): weak positive SA. *Middle left* (**b**): low-moderate positive SA. *Middle right* (**c**): high-moderate positive SA. *Right* (**d**): strong positive SA

by even moderate amounts of positive SA. Meanwhile, corresponding histograms once more confirm the expectations that low levels of positive SA have little effect, whereas moderate levels tend to stretch the right-hand tail and shift the concentration of counts toward 0.

Poissonness plots for the SF model-embedded positive SA results appearing in Fig. 4.15 reveal that the irregularness of the China county surface partitioning introduces additional skewness into count distributions; the upper tail becomes increasingly separated from the middle and lower tail as positive SA increases. In other words, positive SA and the irregularness of a geographic configuration appear to interact. Meanwhile, Tables 4.7 and 4.8 exhibit the same histogram trends detectable in Tables 4.5 and 4.6: low levels of positive SA have little effect, whereas moderate and strong levels tend to stretch the right-hand tail and shift the concentration of values toward 0 (i.e., the mode tends to decrease), while maximum values and kurtosis increase with increasing positive SA. Plot diagnostic statistics begin detecting deviation from a Poisson RV at weak levels of positive SA, again with the Ord plot statistics being more sensitive to the presence of positive SA. A distinction between Tables 4.7 and 4.8 is that SF-induced positive SA can cover the entire range of SA, while a Winsorized auto-Poisson model encounters difficulties and phase transition problems at moderate levels. SF simulations also do not encounter a bifurcation point, and because they lack truncation, they allow much larger counts to materialize.

Table 4.6 Descriptive statistics for the SF auto-Poisson model-based simulated data and the hexagonal tessellation geographic configuration

Variable autocorrelation	MC	GR	$\bar{y}$	s_y	y_{max}	mode	$\lvert z_{kurtosis} \rvert$	Poissonness plot		Ord plot	
								$-\hat{\beta}_0$	$e^{\hat{\beta}_1}$	$\hat{\beta}_0$	$\hat{\beta}_1$
None (i.e., iid)	0.00	1.00	9.00	3.007	19	8	0.04	9.01	9.01	9.01	0.00
Global map pattern-base results											
Weak (using $0.125\mathbf{E}_G$)	0.11	0.90	9.08	3.149	21	9	0.12	9.23	9.31	6.28	0.35
Low-moderate (using $0.285\mathbf{E}_G$)	0.40	0.61	8.96	3.749	26	9	1.98	9.18	9.68	3.53	0.58
High-moderate (using $0.4\mathbf{E}_G$)	0.60	0.41	9.03	4.629	30	6	2.29	10.90	11.89	1.67	0.79
Strong (using $0.9\mathbf{E}_G$)	0.91	0.11	9.06	8.990	48	2, 4	10.14	16.80	19.24	–2.05	1.14
Global + regional map pattern-base results											
Weak (using $0.17\mathbf{E}_R$)	0.10	0.90	8.97	3.336	29	8	2.06	9.29	9.50	6.47	0.29
Low-moderate [using $0.225(\mathbf{E}_G+\mathbf{E}_R)$]	0.41	0.60	9.08	4.317	30	7	2.53	9.96	10.76	1.82	0.77
High-moderate [using $0.5(\mathbf{E}_G+\mathbf{E}_R)$]	0.61	0.41	8.98	7.626	65	4	28.42	13.83	15.59	–3.45	1.35
Strong [using $0.3(4\mathbf{E}_G+\mathbf{E}_R)$]	0.90	0.12	9.02	12.803	99	1	34.72	18.84	21.58	–1.54	1.10
Global + regional + local map pattern-base results											
Weak (using $0.3\mathbf{E}_L$)	0.11	0.88	8.99	4.082	29	7	1.98	9.80	10.49	3.40	0.59
Low-moderate [using $0.1(9\mathbf{E}_R+\mathbf{E}_L)$]	0.39	0.62	9.02	9.586	67	3	25.68	17.79	20.37	–1.94	1.15
High-moderate [using $0.25(2\mathbf{E}_G+\mathbf{E}_R+\mathbf{E}_L)$]	0.61	0.40	8.97	6.485	50	4, 5	13.67	12.39	13.89	–0.05	0.93
Strong [using $0.15(9\mathbf{E}_G+2\mathbf{E}_R+\mathbf{E}_L)$]	0.89	0.13	9.02	14.705	150	0	60.72	21.73	25.36	–4.65	1.50

Table 4.7 Descriptive statistics for the Winsorized auto-Poisson model-based simulated data and the China county geographic configuration

Variable autocorrelation	MC	GR	$\bar{y}$	s_y	y_{max}	mode	$\lvert z_{kurtosis} \rvert$ [a]	Poissonness plot		Ord plot	
								$-\hat{\beta}_0$	$e^{\hat{\beta}_1}$	$\hat{\beta}_0$	$\hat{\beta}_1$
None (i.e., iid)	0.02	0.98	8.99	3.032	20	8	0.64	9.05	9.04	8.97	0.02
Very weak	0.13	0.87	9.10	3.249	23	8, 9	0.40	9.05	9.27	6.76	0.23
Low-moderate	0.53	0.56	9.79	14.990	120	5	79.99	40.47	48.23	– 10.67	2.32

[a]The mean of kurtosis is $1/\mu = 1/9 = 0.111$; the standard error, which can be established using the moment generating function $e^{\mu(e^t - 1)}$, is $\sqrt{151.23594 / n}$ for $\mu = 9$

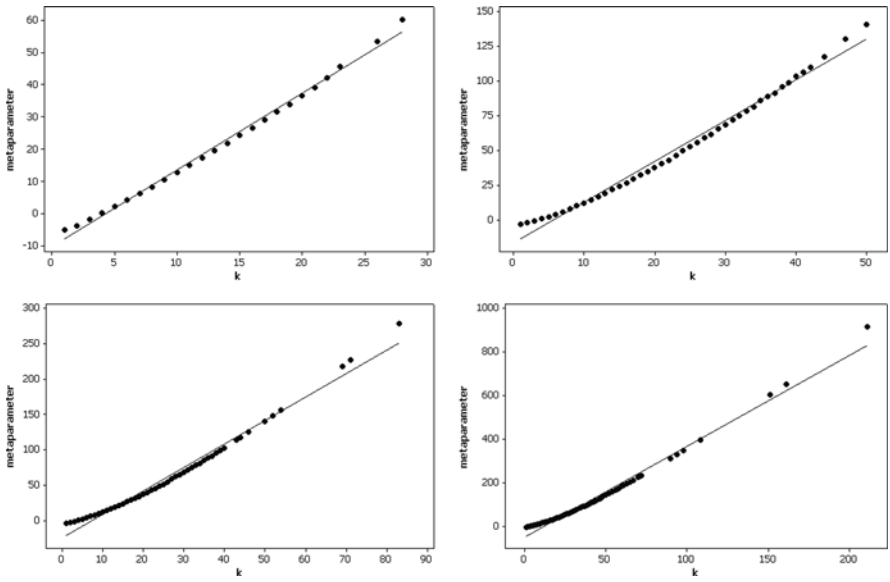

Fig. 4.15 Poissonness plots for SF-model induced levels of positive SA using the China county surface partitioning for a mixture of global, regional, and local map patterns. *Left* (**a**): weak positive SA. *Middle left* (**b**): low-moderate positive SA. *Middle right* (**c**): high-moderate positive SA. *Right* (**d**): strong positive SA

4.3.4 Implications

In conclusion, numerical results reported in this section suggest the following implications about a georeferenced Poisson RV:

(1) by controlling for trend in data when estimating a mean, positive SA has no impact upon the resulting estimated mean value;
(2) positive SA increases the chances of much larger counts materializing;

Table 4.8 Descriptive statistics for the SF auto-Poisson model-based simulated data and the China county geographic configuration

Variable autocorrelation	MC	GR	$\bar{y}$	s_y	y_{max}	mode	$\lvert z_{kurtosis} \rvert$	Poissonness plot		Ord plot	
								$-\hat{\beta}_0$	$e^{\hat{\beta}_1}$	$\hat{\beta}_0$	$\hat{\beta}_1$
None (i.e., iid)	0.02	0.98	8.99	3.032	20	8	0.64	9.05	9.04	8.97	0.02
Global map pattern-base results											
Weak (using $0.07\mathbf{E}_G$)	0.11	0.92	9.03	3.103	24	8	0.75	9.60	9.65	4.07	0.61
Low-moderate (using $0.145\mathbf{E}_G$)	0.41	0.67	8.94	3.639	36	8, 9	11.74	11.80	12.41	4.31	0.51
High-moderate (using $0.19\mathbf{E}_G$)	0.62	0.53	9.03	4.676	57	7	55.84	18.10	19.80	–2.48	1.32
Strong (using $0.35\mathbf{E}_G$)	0.90	0.46	8.92	10.175	190	5	424.57	48.67	63.28	–6.81	1.73
Global + regional map pattern-base results											
Weak (using $0.12\mathbf{E}_{R\text{-}1}$)	0.11	0.90	8.90	3.474	23	7	0.55	9.19	9.57	5.60	0.35
Low-moderate [using $0.25(\mathbf{E}_{R\text{-}1}+\mathbf{E}_{R\text{-}2})$]	0.41	0.64	8.91	5.687	52	6	15.62	16.12	17.94	0.01	0.94
High-moderate [using $0.20(\mathbf{E}_G+\mathbf{E}_{R\text{-}1}+\mathbf{E}_{R\text{-}2})$]	0.61	0.52	9.04	6.536	83	7	56.16	22.61	25.13	–6.29	1.63
Strong [using $0.09(4\mathbf{E}_G+\mathbf{E}_{R\text{-}1})$]	0.91	0.43	9.06	10.550	201	6	340.97	47.65	59.64	–7.88	1.89
Global + regional + local map pattern-base results											
Weak [using $0.125(\mathbf{E}_{R\text{-}1}+\mathbf{E}_L)$]	0.11	0.88	8.97	3.780	28	5	2.34	10.09	10.71	3.67	0.56
Low-moderate [using $0.15(2\mathbf{E}_{R\text{-}1}+2\mathbf{E}_{R\text{-}2}+\mathbf{E}_L)$]	0.42	0.61	9.01	6.816	50	5	16.38	16.68	18.84	–4.99	1.47
High-moderate [using $0.15(1.5\mathbf{E}_G+\mathbf{E}_{R\text{-}1}+\mathbf{E}_{R\text{-}2}+\mathbf{E}_L)$]	0.60	0.57	9.00	6.357	83	4	78.72	24.49	27.75	–1.32	1.09
Strong [using $0.05(7\mathbf{E}_G+\mathbf{E}_{R\text{-}1}+\mathbf{E}_{R\text{-}2}+\mathbf{E}_L)$]	0.90	0.45	9.07	10.544	211	6	404.94	49.50	64.12	–6.12	1.68

(3) especially strong positive SA increases the chances of counts toward 0 materializing;
(4) as positive SA increases, a histogram moves toward the exponential distribution in form;
(5) strong positive SA increases kurtosis;
(6) as positive SA increases, the linearity of a Poissoness plot deteriorates, especially in the tails of an empirical distribution;
(7) the Ord-plot appears very sensitive to the presence of positive SA, and appears to out-perform the Poissonness plot as a diagnostic tool in this context;
(8) a particular mixture of eigenvectors in a SF plays an important role in terms of the impacts of positive SA that materialize (see Table 4.6); and,
(9) the Winsorized auto-Poisson model is unable to capture more than weak-to-moderate positive SA.

In other words, even modest amounts of positive SA do make a difference!

The general importance of these findings concerns data analysis problems, such as excessive zeroes and outliers, that spatial scientists frequently encounter with real world data. These implications should cause a spatial researcher to think more earnestly about the georeferenced nature of his/her data when faced with such problems. In addition, particularly results for the SF-model-based simulations presented here demonstrate that georeferenced Poisson RVs are capable of containing markedly high levels of positive SA.

4.4 The Binomial Probability Model, N > 1

As with Poisson RVs, little is known about the impacts of SA on binomial RVs.[3] Because these RVs also are a member of the exponential family of statistical distributions, just like the normal and Poisson RVs, positive SA should induce variance inflation in them, too. This expectation is further supported by the close similarity between a normal and a binomial frequency distribution when the binomial probability of an event occurring is $p = 0.5$, and the number of events N becomes very large. Thus, one should expect that positive SA will create extra-binomial variation, a notion consistent with discussions in the overdispersion literature.[4] But what happens to the mean of a binomial RV?

One way that a binomial RV differs from both a normal and a Poisson RV is that its values are restricted to the range [0, N], where N is the maximum number of items that can occur at a location. In other words, it is a count with both a lower and

[3]More work has been done on the Bernoulli, vis-à-vis the autologistic model, than on the general binomial RV.

[4]This is not the case for binary 0–1 Bernoulli RVs, which by their very nature cannot exhibit extra variation. The concept of extra variation in a logistic regression has to be teased out of data by, for example, grouping values in order to have an $N > 1$.

an upper bound. The best way for binomial variance to increase is for the relative frequencies of 0 and N to increase when p = 0.5, or for the frequency of 0 to increase when p > 0.5, or of N when p < 0.5. The restricted range should help preserve the mean, μ.

The Ord plot (Ord, 1967) also can be used here for diagnostic purposes. In this context the slope parameter, β_1, becomes negative (< 0). Through the Poisson approximation of a binomial distribution when p is very small (or by symmetry, very large) and Np < 5, the preceding Poisson analysis reveals impacts of SA on binomial histograms when p becomes very small; hence, only the case of p = 0.5 is treated here. So that more direct comparisons can be made with the preceding findings, N is set to 18 (i.e., $\mu = 18/2 = 9$). The simulated iid values have the following descriptive statistics:

	Mean	*Standard deviation*	*Skewness*	*Kurtosis*
theoretical	9	2.1232	0	–0.11111
observed: n = 2,376	8.9933	2.1220	−0.01	–0.10
observed: n = 2,379	8.9975	2.1196	0.00	–0.11

The MCs and GRs for the simulated data are as follows:

n = 2,376: MC = 0.00431, GR = 0.99488
n = 2,379: MC = –0.00168, GR = 1.00731

In other words, these simulated binomial RVs display the necessary characteristics of iid.

Illustrative graphic portrayals of these values appear in Fig. 4.16. Of note is that weighted least squares regression estimation yields b = –0.9995 for the theoretical Ord plot, and b = –0.9981 and –1.0178 for the two simulation data Ord plots, confirming that the values are for binomial RVs. These slope parameter estimates can be converted to their corresponding binomial probability estimates with the equation $p = \frac{\hat{\beta}_1}{\hat{\beta}_1 - 1}$, respectively yielding 0.49988 for the theoretical binomial data, and 0.50441 and 0.49953 for the simulated data; the true value is 0.5.

Bernoulli RVs (i.e., N = 1) are not be treated in this section, since their histograms tend to be too simple to display conspicuous impacts of SA.

4.4.1 Simulating Spatially Autocorrelated Binomial RVs

The simulation of either multivariate binary or multivariate binomial georeferenced data has not been given as much attention in the literature as has the simulation of spatially autocorrelated normal or Poisson RVs. Dolan et al. (2000), for example, simulate a spatially autocorrelated log-normal RV and then do a back-transformation, an approach not endorsed here. Heagerty and Lele (1998), for

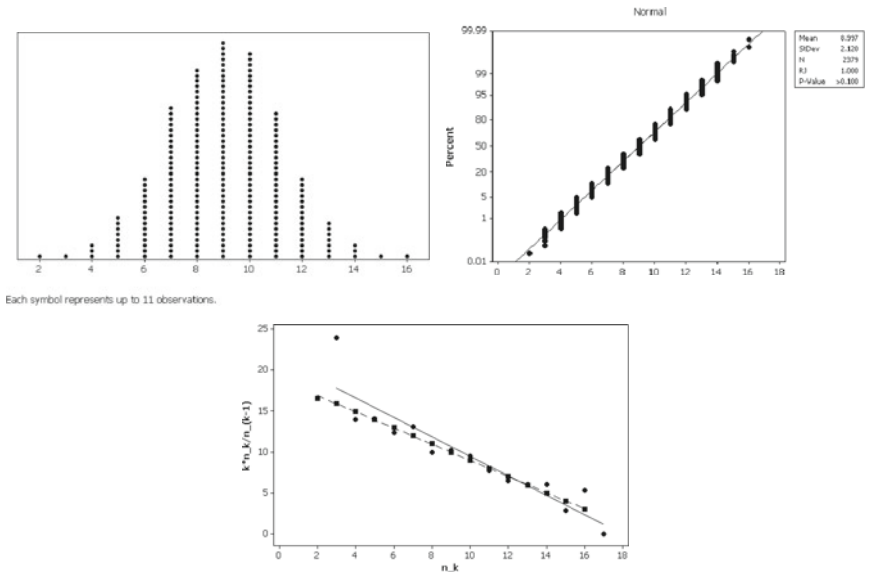

Fig. 4.16 Graphical diagnostic tools for the iid binomial RVs; $n = 2{,}379$. *Left* (**a**): a dot-plot histogram. *Middle* (**b**): a normal quantile plot. *Right* (**c**): an Ord plot with its trend line (*solid line with solid circles*) together with and Ord plot of the theoretical counterpart (*broken line with solid squares*)

instance, promote the use of a generalized linear mixed model coupled with a geostatistical perspective for binary georeferenced data. And, Augustin et al. (1998), for example, promote the use of the autologistic model. As in the Sects. 4.2 and 4.3, auto-binomial model RVs are simulated here with MCMC and SF techniques. The autoregressive equation employed with MCMC is given by

$$\begin{aligned} &P(Y_i = y|\alpha_i, \mathbf{C}_i\mathbf{Y}) \\ &\quad = \exp\left(\alpha_i + \rho\sum_{j=1}^{n} c_{ij}(y_j - \bar{y})\right) \Big/ \left[1 + \exp\left(\alpha_i + \rho\sum_{j=1}^{n} c_{ij}(y_j - \bar{y})\right)\right], \end{aligned} \tag{4.8}$$

where y is contained in the interval [0, N], and including subtraction of the mean $\bar{y}$ in parallel with Kaiser and Cressie's (1997) specification for the Winsorized auto-Poisson model specification. Meanwhile, the SF equation employed is given by

$$P\left(Y_i = y|\mathbf{E}_{i,K}\right) = \exp(\alpha + \mathbf{E}_{i,K}\boldsymbol{\beta})/[1 + \exp(\alpha + \mathbf{E}_{i,K}\boldsymbol{\beta})], \tag{4.9}$$

where $\mathbf{E}_K$ is the n-by-K matrix of SF eigenvectors. The procedural steps for using these equations to simulate geographic distributions are exactly the same as those outlined in the preceding section for Poisson RVs, except that the Poisson probability model is replace with the binomial probability model.

4.4.2 *Simulation Results for an Ideal Regular Hexagonal Surface Partitioning*

MCMC map simulation can exploit the particular relationship between the intercept and autoregressive parameter when $p = 0.5$, namely asymptotically $\alpha = -3\rho$, which was done here. As with Poisson RVs, phase transitions tend to be encountered beyond moderate SA. Consequently, only weak and low-moderate SA have been simulated for analysis purposes.

Figure 4.17 furnishes strong evidence that the generated MCMC chains converged, rendering useful maps with positive SA embedded in them. The time series plots exhibit random stability. Meanwhile, the correlograms reveal that virtually no serial autocorrelation is present in the three chains. As with the Poisson RVs simulated in the preceding section, only the last map of a chain is used here.

Summary descriptive statistics appear in Table 4.9 for the auto-binomial simulated data containing positive SA. These statistics confirm that the (controlled for trend) mean essentially is unaffected, while the variance is inflated (i.e., overdispersion). Dot plot versions of histograms appearing in Fig. 4.18 confirm the expectations that low levels of positive SA have little effect, whereas low-moderate levels already tend to redistribute counts to the two tails. In addition, the Kolmogorov-Smirnov (K-S) statistic quantifies a movement away from the corresponding theoretical binomial distribution as positive SA increases. Unfortunately,

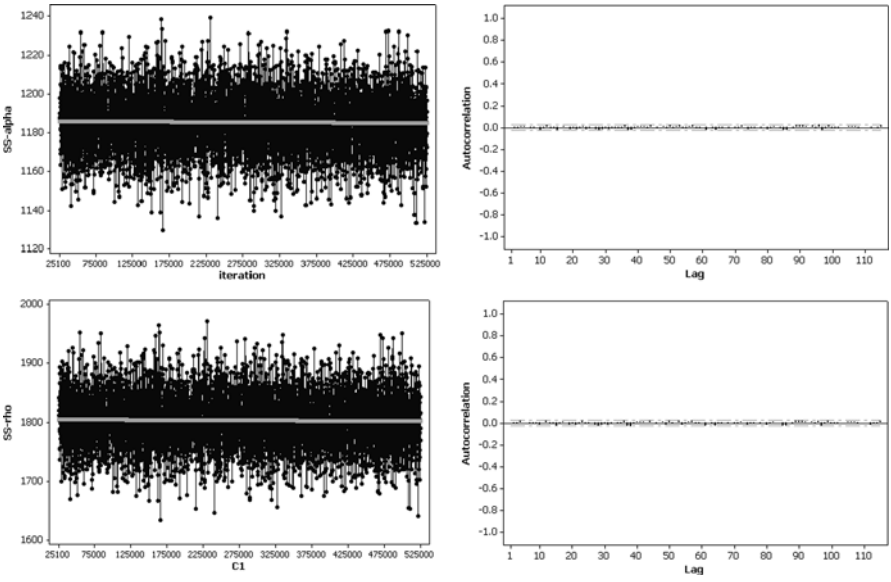

Fig. 4.17 MCMC time series plot and correlogram diagnostic graphics based on the ideal hexagonal surface partitioning when $\rho = 0.60$ (*bottom*). *Left* (**a**): for the intercept term. *Right* (**b**): for the autoregressive term

Table 4.9 Descriptive statistics for the auto-binomial model-based MCMC simulated data and the hexagonal tessellation geographic configuration

Variable autocorrelation	MC	GR	$\bar{y}$	s_y	y_{min}	y_{max}	Skewness	Kurtosis	K-S[a]
None (i.e., iid)	0.00	0.99	8.99	2.122	2	16	–0.01	–0.10	0.0038
Weak	0.10	0.89	8.81	2.212	2	15	–0.02	–0.19	0.0400
Low-moderate	0.39	0.55	9.07	2.759	0	18	–0.24	–0.07	0.0700

[a]K-S denotes the Kolmogorov-Smirnov statistic, used here to index deviation from the theoretical binomial distribution for which N = 18 and p = 0.5

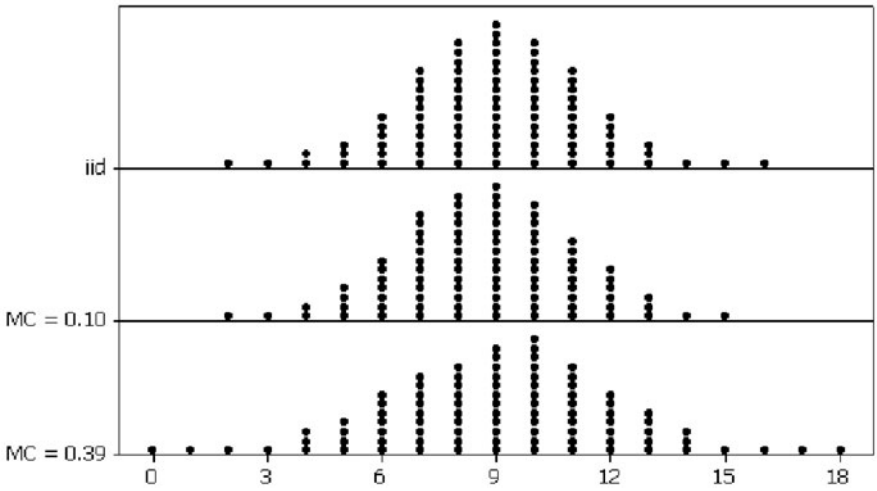

Fig. 4.18 Dot plot versions of histograms for the MCMC auto-binomial simulated data. *Top* (**a**): iid. *Middle* (**b**): weak positive SA. *Bottom* (**c**): low moderate positive SA

because strong positive SA cannot be embedded with MCMC techniques, its impacts cannot be assessed in terms of an auto-binomial model.

Summary statistics for the SF model-embedded positive SA results appear in Table 4.10; the corresponding dot plot versions of histograms appear in Fig. 4.19 for a global map pattern. These results both confirm and extend those found for the auto-binomial. Figure 4.19a includes the dot plot for extremely strong positive SA to complete the trend being revealed by these illustrative results: as positive SA approaches its maximum, the binomial histogram increasingly resembles that for a sinusoidal RV—this is the reason for change in the kurtosis statistic. Overall, as positive SA increases in a binomial RV, variance increases, and the center of a histogram flattens, converging first on a uniform distribution in appearance, and then on a near-dichotomous 0/N frequency distribution.

Dot plot versions of histograms for SFs constructed with global and regional map patterns appear in Fig. 4.19b. As with the global map pattern results, the mean remains unaffected, variance is inflated and kurtosis is impacted upon by positive

Table 4.10 Descriptive statistics for the auto-binomial model-based SF simulated data and the hexagonal tessellation geographic configuration

Variable autocorrelation	MC	GR	$\bar{y}$	s_y	y_{min}	y_{max}	Skewness	Kurtosis	K-S[a]
None (i.e., iid)	0.00	0.99	8.99	2.122	2	16	−0.01	–0.10	0.0038
Global map pattern-base results									
Weak (using $0.0035\mathbf{E}_G$)	0.10	0.90	9.01	2.239	2	16	−0.02	–0.19	0.0187
Low-moderate (using $0.008\mathbf{E}_G$)	0.39	0.61	9.03	2.673	1	18	0.00	–0.24	0.0625
High-moderate (using $0.0125\mathbf{E}_G$)	0.65	0.37	9.02	3.278	0	18	0.02	–0.47	0.1164
Strong (using $0.03\mathbf{E}_G$)	0.90	0.13	8.97	5.165	0	18	0.03	–1.07	0.2351
Global + regional map pattern-base results									
Weak (using $0.005\mathbf{E}_R$)	0.09	0.91	9.09	2.309	2	16	−0.04	–0.24	0.0413
Low-moderate [using $0.008(\mathbf{E}_G+\mathbf{E}_R)$]	0.42	0.59	8.94	3.061	1	18	0.06	–0.43	0.1063
High-moderate [using $0.015(\mathbf{E}_G+\mathbf{E}_R)$]	0.61	0.40	8.97	4.321	0	18	0.01	–0.92	0.2036
Strong [using $0.0095(4\mathbf{E}_G+\mathbf{E}_R)$]	0.90	0.12	9.01	5.763	0	18	0.01	–1.27	0.2802
Global + regional + local map pattern-base results									
Weak (using $0.008\mathbf{E}_L$)	0.10	0.90	8.98	2.710	1	17	0.01	–0.26	0.0633
Low-moderate [using $0.003(9\mathbf{E}_R+\mathbf{E}_L)$]	0.44	0.56	8.99	4.937	0	18	−0.02	–1.11	0.2457
High-moderate [using $0.007(2\mathbf{E}_G+\mathbf{E}_R+\mathbf{E}_L)$]	0.61	0.40	9.04	3.880	0	18	−0.01	–0.71	0.1643
Strong [using $0.005(9\mathbf{E}_G+2\mathbf{E}_R+\mathbf{E}_L)$]	0.91	0.11	8.95	6.147	0	18	0.01	–1.38	0.3008

[a]K-S denotes the Kolmogorov-Smirnov statistic, used here to index deviation from the theoretical binomial distribution for which $N = 18$ and $p = 0.5$

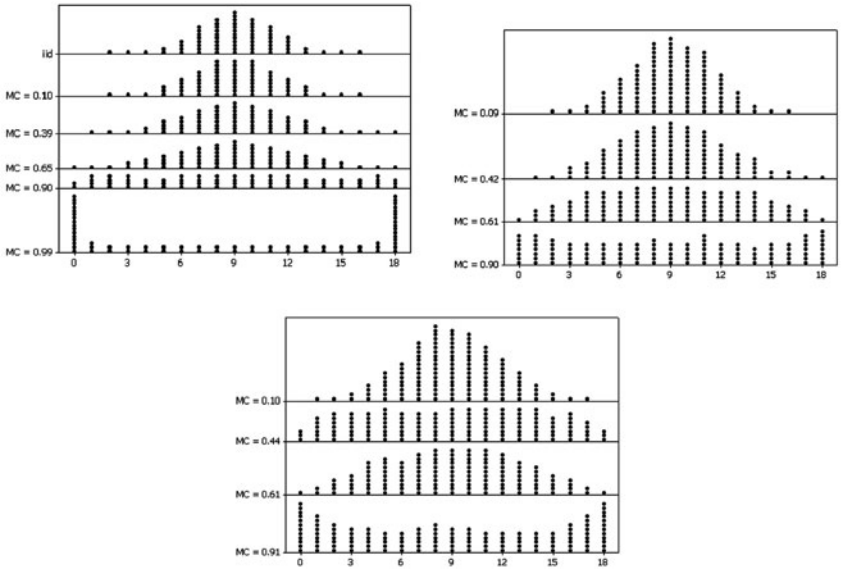

Fig. 4.19 Dot plot versions of histograms for the SF binomial simulated data using the regular hexagonal surface partitioning. *Left* (**a**): SF results from a global map pattern. *Middle* (**b**): SF results from a global combined with a regional map pattern. *Right* (**c**): SF results from a global combined with a regional and a local map pattern

SA. Again, these results both confirm and extend those found for the auto-binomial. Furthermore, the tendency toward a sinusoidal RV shaped histogram already is becoming apparent here for MC = 0.90. The same histogram patterns appear for SFs constructed with global, regional and local map patterns. As with the Poisson case, Fig. 4.19c (as well as its corresponding part of Table 4.10) indicates that the mixture of map patterns constituting a SF, rather than only the level of positive SA, plays an important role, too.

4.4.3 Simulation Results for the China County Geographic Configuration

As with the hexagonal surface partitioning, MCMC simulation of auto-binomial model-based maps employing the China county irregular surface partitioning at most could embed only moderate positive SA. The graphical diagnostics appearing in Fig. 4.20 indicate that the resulting maps are properly generated. In addition, summary descriptive statistics reported in Table 4.11 are consistent with those appearing in Table 4.10: overdispersion is induced, and the distribution appears more uniform in shape. Meanwhile, dot plot versions of histograms appearing in Fig. 4.21 once more confirm the expectations that low levels of positive SA have little effect,

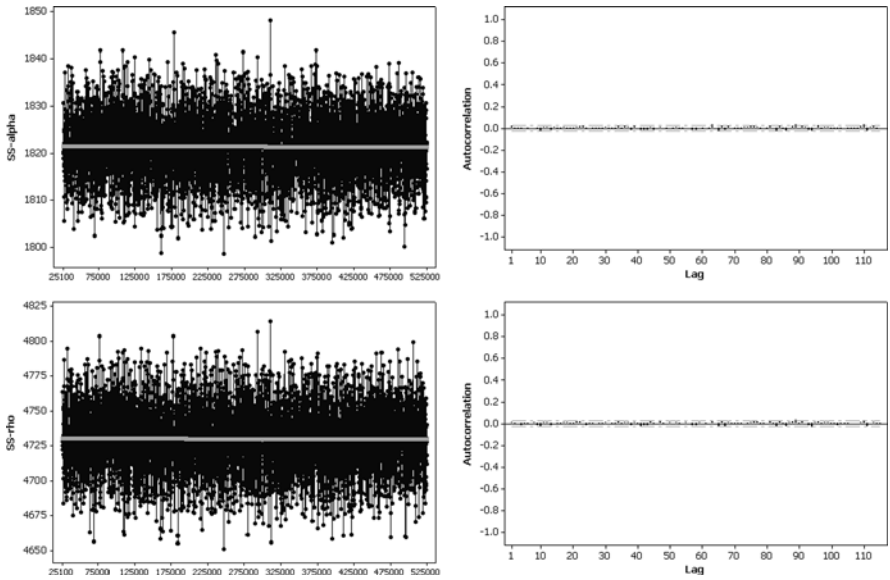

Fig. 4.20 MCMC time series plot and correlogram diagnostic graphics based on the China county surface partitioning when $\rho = 1.00$ (*bottom*). *Left* (**a**): for the intercept term. *Right* (**b**): for the autoregressive term

Table 4.11 Descriptive statistics for the auto-binomial model-based MCMC simulated data and the China irregular county geographic configuration

Variable autocorrelation	MC	GR	$\bar{y}$	s_y	y_{min}	y_{max}	Skewness	Kurtosis	K-S[a]
None (i.e., iid)	0.00	0.99	8.99	2.122	2	16	–0.01	–0.10	0.0038
Weak	0.12	0.86	9.18	2.511	1	16	–0.11	–0.27	0.0697
Low-moderate	0.40	0.55	11.96	4.167	0	18	–0.73	–0.73	0.5007
moderate	0.51	0.25	13.79	5.128	0	18	–1.37	0.61	0.6963

[a]K-S denotes the Kolmogorov-Smirnov statistic, used here to index deviation from the theoretical binomial distribution for which $N = 18$ and $p = 0.5$

whereas moderate levels tend to squash the center of a distribution and thicken its tails—in this case the irregularity of the surface partitioning distorts this tail thickening by skewing it to one side of its distribution. Of note is that the irregular surface partitioning introduces some trend in the mean, indicating that the relationship between α and ρ most likely needs to be more carefully articulated for irregular surface partitionings. Furthermore, the Kolmogorov-Smirnov statistics reported in Table 4.11 are indexing this deviation from $p = 0.5$ as much, if not more, than the change in the shape of the histogram. Skewness distortion with increasing positive SA appears in both the MCMC auto-binomial and the SF model-based simulation data.

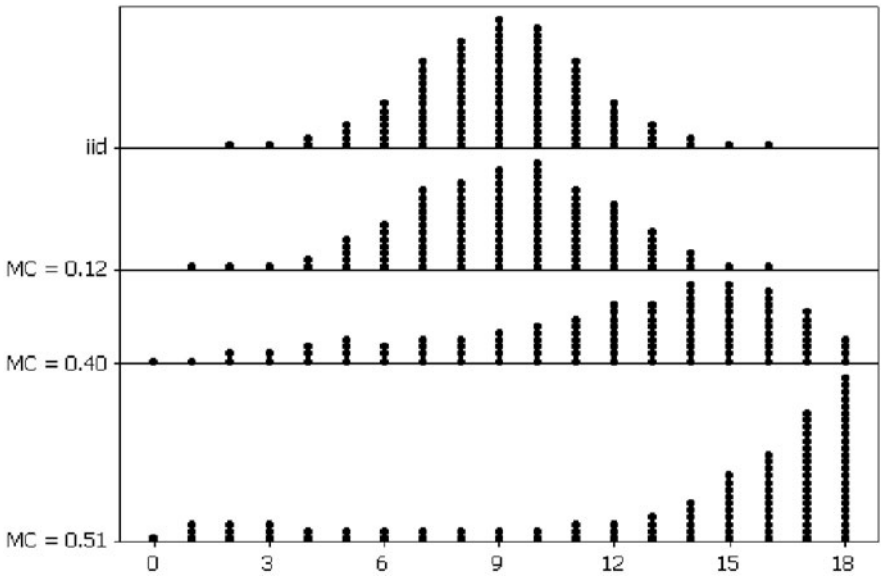

Fig. 4.21 Dot plot versions of histograms for the MCMC auto-binomial simulated data. *Top* (**a**): iid. 2nd from *top* (**b**): weak positive SA. 2nd from *bottom* (**c**): low moderate positive SA. *Bottom* (**d**): moderate positive SA

As is also seen with the Poisson RV analysis, one conspicuous difference between Figs. 4.18 and 4.19, and Figs. 4.21 and 4.22, is the interaction effect between positive SA impacts and the irregularness of the geographic configuration. One outcome of this interaction is that the flattening of a binomial histogram is followed by less of a sinusoidal RV shape as positive SA approaches its maximum value.

4.4.4 Implications

In conclusion, numerical results reported in this section suggest the following implications about a georeferenced binomial RV:

(1) by controlling for trend in data when estimating a mean (apparently this only needs to be done with MCMC simulation, not with SF simulation), positive SA has no impact upon the resulting estimated mean value;
(2) positive SA increases the chances of a histogram resembling that for a uniform distribution, and in the extreme, for a sinusoidal distribution;
(3) especially strong positive SA increases the chances of most counts being only 0 or N;
(4) as positive SA increases, the Kolmogorov-Smirnov test statistic tends to increase;

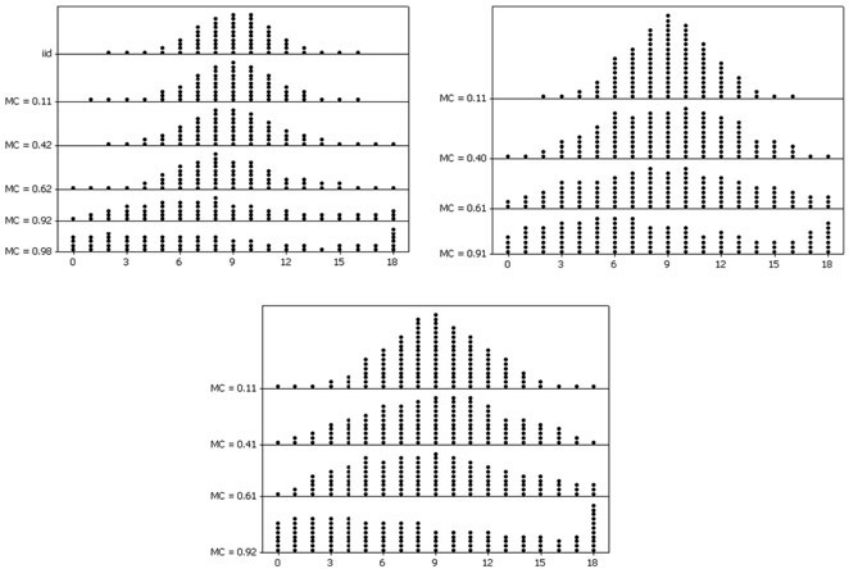

Fig. 4.22 Dot plot versions of histograms for the SF binomial simulated data using the China irregular surface partitioning. *Left* (**a**): SF results from a global map pattern. *Middle* (**b**): SF results from a global combined with a regional map pattern. *Right* (**c**): SF results from a global combined with a regional and a local map pattern

(5) a particular mixture of eigenvectors in a SF plays an important role in terms of the impacts of positive SA that materialize (see Tables 4.10 and 4.12);
(6) an interaction effect appears to occur between SA and the irregular nature of a surface partitioning; and,
(7) the conventional auto-binomial model is able to capture only weak-to-moderate positive SA.

In other words, just as with a Poisson RV, even modest amounts of positive SA do make a difference!

4.5 Discussion

This chapter indicates what a spatial scientist should expect from commonly encountered levels of SA when inspecting histograms constructed with georeferenced data. Regardless of whether a RV is normal, binomial, or Poisson in nature, its variance will tend to be inflated, with inflation increasing as positive SA increases. This is the single most common impact, which results in histograms being flatter than they would otherwise be if the data observations were iid. It leads to heterogeneity for normal RVs, excessive 0 s and extreme values (i.e., overdispersion) for

Table 4.12 Descriptive statistics for the auto-binomial model-based SF simulated data and the China irregular county geographic configuration

Variable autocorrelation	MC	GR	$\bar{y}$	s_y	y_{min}	y_{max}	Skewness	Kurtosis	K-S[a]
None (i.e., iid)	0.00	0.99	8.99	2.122	2	16	−0.01	−0.10	0.0038
global map pattern-base results									
Weak (using $0.112\mathbf{E}_G$)	0.11	0.91	8.98	2.256	1	16	−0.00	−0.14	0.0212
Low-moderate (using $0.275\mathbf{E}_G$)	0.42	0.63	9.06	2.632	2	18	0.23	0.03	0.0536
High-moderate (using $0.42\mathbf{E}_G$)	0.62	0.47	8.92	3.030	0	18	0.48	0.09	0.1002
Strong (using $1.10\mathbf{E}_G$)	0.92	0.19	8.40	4.644	0	18	0.47	−0.59	0.2674
global + regional map pattern-base results									
Weak (using $0.18\mathbf{E}_{R-1}$)	0.11	0.91	9.00	2.379	2	16	−0.02	−0.07	0.0345
Low-moderate [using $0.35(\mathbf{E}_{R-1}+\mathbf{E}_{R-2})$]	0.40	0.64	9.05	3.505	0	18	0.01	−0.55	0.1405
High-moderate [using $0.46(\mathbf{E}_G+\mathbf{E}_{R-1}+\mathbf{E}_{R-2})$]	0.61	0.45	8.93	4.416	0	18	0.07	−0.80	0.1991
Strong [using $0.35(4\mathbf{E}_G+\mathbf{E}_{R-1})$]	0.91	0.18	8.19	5.272	0	18	0.40	−0.91	0.3212
global + regional + local map pattern-base results									
Weak [using $0.2(\mathbf{E}_{R-1}+\mathbf{E}_L)$]	0.11	0.87	9.02	2.701	0	18	0.02	−0.17	0.0648
Low-moderate [using $0.2(2\mathbf{E}_{R-1}+2\mathbf{E}_{R-2}+\mathbf{E}_L)$]	0.41	0.63	9.05	3.846	0	18	−0.03	−0.64	0.1560
High-moderate [using $0.3(2\mathbf{E}_G+\mathbf{E}_{R-1}+\mathbf{E}_{R-2}+\mathbf{E}_L)$]	0.61	0.44	8.84	4.276	0	18	0.19	−0.74	0.2073
Strong [using $0.25(7\mathbf{E}_G+\mathbf{E}_{R-1}+\mathbf{E}_{R-2}+\mathbf{E}_L)$]	0.92	0.17	8.00	5.692	0	18	0.39	−1.07	0.3617

[a]K-S denotes the Kolmogorov-Smirnov statistic, used here to index deviation from the theoretical binomial distribution for which $N = 18$ and $p = 0.5$

Poisson RVs, and overdispersion for binomial RVs. Positive SA corrupts quantile plots when assessing normality, Poissonness plots, and other goodness-of-fit test, even when only its most commonly encountered moderate levels are present.

SF model specifications furnish an efficient and effective way of capturing SA effects, and render simulation results that are consistent with those obtained with the more conventional auto- model specifications. Because these models are constructed with stepwise regression techniques when an empirical analysis is being undertaken, they signal that Gaussian approximations actually are not obsolete. The role of these approximations is to supply a first glimpse of SA, as well as a first screening of a large number of candidate eigenvectors when constructing a SF.

Finally, the lessons to be learned from this chapter may be summarized as follows: *caution SA at work!* Cursory initial graphical inspections of empirical data can be misleading when SA is present. Spatial scientists need to heed this warning.

Acknowledgment This research was completed while the author was a visiting scientist at the Max Planck Institute for Demographic Research, Rostock, Germany, 2005.