[AM-03-033] Spatial Filtering Models

Abstract

The spatial statistical analysis history is fraught with spatial autocorrelation (SA) ponderings, mostly questioning the nature and degree of this observational correlation type's impacts. Tacit awareness of its impending complications emerged in the early 1900s, with time series methodology spuriously guiding thought and practice. Initial debates cast SA as a nuisance, encouraging its expulsion from geospatial data. The first spatial filtering models strove to do this. However, value ultimately bestowed upon SA by such procedures as spatial interpolation (e.g., geostatistical kriging) moderated this excessive action, with a suite of spatial autoregressive models emerging that fostered spatial filtering engendering isolation-but-retention of global SA in data analyses. Next, Getis refocused this effort on local SA statistics to devise an alternative spatial filtering model whose capabilities include response variable and covariate decompositions into disjunct spatial and aspatial components, suggestive of the spatial Durbin specification. One weakness of these models is their strict normal curve theory reliance. More recent MESF model formulation and articulation transcends this drawback. Consequently, today, spatial analysts can tap a wide variety of spatial filtering conceptualizations, the subject matter this article reviews. In doing so, it presents an original spatially autocorrelated gamma variate empirical example, itself a novel literature contribution.

Keywords: Cochrane-Orcutt, Getis, spatial autocorrelation, spatial regression and econometrics

Author & citation

Griffith, D. A. (2024). Spatial Filtering Models. The Geographic Information Science & Technology Body of Knowledge (2024 Edition). John P. Wilson (Ed.).
DOI: [10.22224/gistbok/2024.1.10](https://doi.org/10.22224/gistbok/2024.1.10)

Explanation

1. Introduction
2. Spatial Filtering Models Scope
3. Illustrative Applications
4. The Future of Spatial Filtering Models

1. Introduction

Two correlation types can burden empirical data with impediments, namely attribute and observational. Conventional multivariate statistics addresses the former within a multicollinearity context that univariate statistics naturally spawned as it matured. The latter has a more nuanced and complicated history (e.g., Griffith, 2020). The statistical filtering notion closely relates to its chronicle trajectory, in a shift from specialized independent and identically distributed (iid) to more general correlated/dependent data analytic frameworks beleaguered by redundant information. Laplace inaugurated envisaging this generic data category in the early 1800s, investigating daily barometric pressure serial correlation (i.e., time series). Next, in the early 1900s, Student (aka Gossett) enlarged its domain to embrace spatial autocorrelation (SA) impacts on the Pearson correlation coefficient, an issue preoccupying some statisticians into the end of the last millennium. Unfortunately, mathematical statistical theory for both varieties lacked momentum for a half-century or more, a period in which Hotelling conceptualized paired observations correlated data (e.g., repeated measures, dependent samples with before-after formats being time series predecessors), a form more akin to multicollinearity, an innovative idea that immediately flourished (e.g., pre-post intervention matched data). Recognizing analytical and inferential complications attributable to observational correlation, statisticians began devising techniques to filter it from their data in a quest to return their analyses to ones at least mimicking iid, the first and simplest being differencing partnered observation duos to compute equality-of-two-means tests, a straightforward arithmetic operation that filters out correlation between paired observations (halving a sample size in doing so). Thus, they initiated a subtraction strategy that eventually extended to time series and SA situations.

Although Yule and Walker began applying autoregressive specifications to time series data in the 1920s and 1930s, filtering serial correlation from such dependent data had to await the 1949 arrival of the Cochran-Orcutt tactic, which again involved the existing differencing protocol, but modified so that subtraction is of a time-lag term. Next, Cliff and Ord's popularizing transferal of temporal data analytics to spatial series (circa 1970), promoting the formative Whittle-Bartlet-Besag spatial statistics tradition, inspired Griffith to spatialize this Cochran-Orcutt autoregressive-based approach to filter georeferenced data (circa 1980), conserving the differencing protocol by substituting a spatial- for a time-lag subtraction. Eventually, its inflexibilities prompted him to craft Moran eigenvector spatial filtering methodology (MESF; circa 1996), further adapting the differencing practice to subtract a SA component in a way that transports it to its variable's mean, a breakthrough shortly thereafter (circa 2002) independently replicated by Legendre and his colleagues in statistical ecology that loosely links to geostatistics. In tandem, Getis (circa 1995) autonomously concocted a spatial filtering apparatus whose differencing routine subtraction exploits his Getis-Ord local SA hot/cold spot index, and also is reminiscent of a geostatistics viewpoint. In conclusion, the key message this introductory narrative advocates is that spatial filtering is part of a long, robust, pioneering praxis encompassing cognate disciplines that transforms data ladened with unorthodox correlation into synthetic numerical content imitating the classical iid scenario for inferential and interpretation purposes through differencing processes.

2. Spatial Filtering Models Scope

Spatial filtering's goal is an informative model-based management of SA in geospatial data that accounts for it while enabling legitimate statistical analysis undertakings. Accordingly, by contextualizing it within its boundaries and extent, targeted disciplinary domains, given application utilities, pertinent georeferenced attribute variables, spatial statistical methodologies, and limitations and vulnerabilities, it may be defined as follows (e.g., Getis, 1995; Borcard et al., 2002; Griffith, 2008, 2010; Griffith and Getis, 2016; Griffith, Chun, and Li, 2019; Griffith and Chun, 2022):

a conceptually anchored mathematical expression (i.e., model statement) isolating and/or removing SA, a fundamental latent structured stochastic feature lurking in geospatial data, to make an efficient and effective handling of its

implicit and explicit effects less problematic when spatial analysts engage in data analytic work.

Spatial filtering's overarching objective is to enable a creative and unhindered use of classical statistical techniques, particularly linear and generalized linear model (GLM) regression, when analyzing geotagged data, a brand of correlated/dependent information, which constitutes its importance. Its mechanics are of two types: removal or transferal/isolation. The Cochrane-Orcutt (1949) mechanism genre purges SA. In contrast, its autoregressive counterparts sequester and preserve its response variable constituent, often via its regression residual manifestations, with the autoregressive response (AR) specification ignoring SA in covariates, the simultaneous autoregressive (SAR) specification artificially filtering its exact same nature and degree from all covariates (i.e., most likely just partial filterings), and the spatial Durbin specification—somewhat emulating Cochrane-Orcutt—estimating an individual variable-specific SA term for each covariate (i.e., spatially lagged explanatory variates), frequently without discarding any of them. MESF (e.g., Borcard and Legendre, 2002; Griffith and Peres-Neto, 2006; Griffith et al., 2019), in its elementary mode, extracts SA from regression residuals (i.e., essentially focusing on SA in Y), and shifts it to a (usually regression) equation's intercept, converting it from a constant to a varying term. In some instances, this algebraic manipulation constructively substitutes SA surrogates/proxies for potentially relevant omitted variables sharing their same map patterns (Griffith and Chun, 2016; Paez, 2019). More generally, it can furnish a portal for the usage of instrumental variables (Le Gallo and Paez, 2013), a practice capable of addressing bias arising from not only omitted variables, but also measurement error (i.e., errors in variables) and simultaneous causality. Meanwhile, duplicating the Cochrane-Orcutt manoeuvre, each covariate also can undergo this decomposition. Getis (e.g., 1995) does a bit of both, basically removing SA from covariates in order to better segregate its response variable Y component. He argues that one always should retain the response variable spatial factor, and exercise discretion with regard to including or excluding each of the covariate spatialized elements. Getis and Griffith (2002), Griffith and Peres-Neto (2006), and Thayn and Simanis (2013) furnish enlightening side-by-side comparisons of a number of these formulations.

A second invaluable spatial filtering advantage eclipses its staunchest competitor, spatial autoregression, and Besag's family of auto- models, which, as already noted, separate but usually keep at least some SA in an equation, again concentrating on such correlation latent in Y. Regardless, these are the basis of the Cochrane-Orcutt SA elimination scheme. Besag's auto-Poisson, negative binomial (NB), exponential, and gamma creations are incapable of accommodating positive SA, generally (although perhaps erroneously) viewed as the sole SA nature in practice. Consequently, their publicized successful treatment utilizes Box-Cox transformation inputs into normal approximations, and hence frequently confront serious specification error obstacles (e.g., a la back-transformations). In contrast, MESF negotiates this possibility with ease, whereas Getis's method does so strictly for normal/Gaussian geospatial random variables.

3. Illustrative Applications

Besag dismisses auto-exponential and gamma probability model inventions as being unwieldly and lacking any intuitive appeal to spatial analysts, even though their minimal aim is to seclude and then deposit into a spatial lag term SA latent in Y. However, exponential distributions provide elucidatory descriptions of wait times until a certain event occurs, such as Puerto Rican municipio urban area electricity restoration after Hurricane Maria devastated the island in 2017. This case study encompasses positive SA, a condition defied by auto-gamma models, via its geographic diffusion of recovery efforts pursuing a contiguous expansion rebuilding of a dramatically damaged electric grid from 22 coastal power plants, coupled with the United States Federal Emergency Management Agency's (FEMA's) aid entrance through an eastern seaport (i.e., the defunct Roosevelt Roads naval base) to all island neighborhoods (the last one being reconnected 328 days after the hurricane). This empirical example furnishes a novel and insightful spatial filtering exemplification involving a spatially autocorrelated (positive SA) gamma variate.

Covariation exists between re-electrification wait times for Puerto Rican urban places (see https://earthobservatory.nasa.gov/images/144371/night-lights-show-slow-recovery-from-maria) and several salient substantive attributes (Figure 1): population density (i.e., intensely inhabited coastal lowland built-up areas received high, whereas sparsely inhabited interior mountainous rural regions received low, priority); elevation (i.e., coastal generating station locations privileged nearby cities, towns, and other human settlements); and, an east-west gradient (i.e., the island's elongated/elliptical shape conflated with FEMA's entry point) channeled relief aid movements (e.g., circumnavigating la Cordillera Central, the island's interior mountain range). One question spatial filtering models help answer here asks about conspicuous SA effects on measured associations for this battery of correlates. The Cochrane-Orcutt device, requiring estimation of, for example, multiple spatial AR models, supplies the simplest reply, producing a contrast between outcomes rendered by the presence and absence of SA in a linear regression analysis. It begins with a Box-Cox power transformation of variables (i.e., attaching a γ exponent to measurements): γ = –0.43 for wait times (Y); γ = –0.17 for 2020 population density (X_1); γ = 0.27 for mean elevation [X_2; extracted from a digital elevation model (DEM)]; and, γ = 1 (i.e., no Box-Cox transformation; but converted to z-scores to ensure numerical stability) for longitude (X_3). These three covariates account for roughly 50% of the geographic variation (a la R^2, adjusted-R^2, predicted-R^2 and GLM pseudo-R^2 indices) across the island in urban wait times, and yield extremely favorable residual diagnostics.

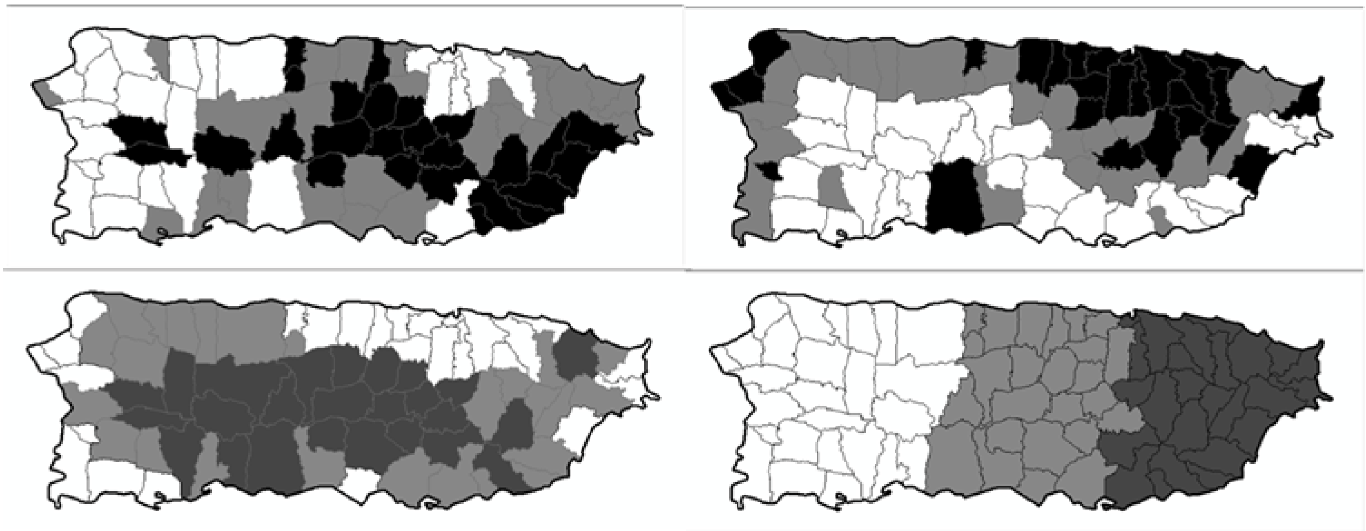

Figure 1. Tertile choropleth maps portraying the geographic distributions of Box-Cox power transformed georeferenced attribute variables: white, gray, and black respectively denote relatively low, intermediate, and high values. Top left (a): urban wait times (MC = 0.314, GR = 0.710). Top right (b): 2020 population density (MC = 0.600, GR = 0.490). Bottom left (c): mean DEM elevation (MC = 0.511, GR = 0.495). Bottom right (d): longitude (MC = 0.871, GR = 0.065). Source: author.

Tables 1-4 respectively tabulate conventional linear and gamma regression (always using a log-link function—capturing positive-only wait time data with positively-skewed errors, sustaining some degree of consistency with the normal approximations, and depicting an underlying multiplicative process) results ignoring the presence of SA in any of the georeferenced variables (i.e., no spatial filtering), isolating and retaining SA in the response variable only (i.e., implicit spatial filtering of Y), in the response variable plus the covariates (i.e., implicit spatial filtering of Y as well as the covariates), and SA removal from all variates (i.e., spatial filtering as though SA is a nuisance, expunging it from all georeferenced variates). One principal finding concerns null hypothesis Type I errors (i.e., false positives, or decision errors of commission). Another entails multicollinearity affiliated with non-intercept (i.e., covariate) parameter estimates.

Table 1. SA-ladened response variate and covariate null hypothesis probabilities

spatial filter	intercept	X_1: population density		X_2: elevation		X_3: longitude		λ_1 / λ_3	(pseudo)-R^2
		p-value	VIF	p-value	VIF	p-value	VIF		
Box-Cox supported normal curve approximation cases									
none	< 0.0001	< 0.0001	1.58	0.0556	1.47	< 0.0001	1.11	4.2	0.5368
	GLM theory: a gamma random variable								
none	< 0.0001	< 0.0001		0.0539		< 0.0001		4.3	0.5090
NOTE λ_1/λ_p: 1-10 expected; 10-100 multicollinearity concern; 100+ multicollinearity problem									

Table 2. SA-isolated response variate and SA-ladened covariate null hypothesis probabilities						
spatial filter	intercept	X_1: population density	X_2: elevation	X_3: longitude	$\frac{\lambda_1}{\lambda_p}$	(pseudo)-R^2
Box-Cox supported normal curve approximation cases						
spatial AR†	< 0.0001	< 0.0001	0.1598	< 0.0001	4.8 (p = 4)	0.5613
Getis	< 0.0001	< 0.0001	0.2232	0.0003	5.5 (p = 4)	0.5611
MESF	< 0.0001	< 0.0001	0.0103	< 0.0001	5.3 (p = 6)	0.5871
GLM theory: a gamma random variable						
MESF	< 0.0001	< 0.0001	0.0129	< 0.0001	5.4 (p = 6)	0.5999
† denotes the spatial autoregressive response (aka spatial econometrics spatial lag) model in which ρ estimated for response variable Y only rescales isolated SA in it.						
NOTE λ_1/λ_p: 1-10 expected; 10-100 multicollinearity concern; 100+ multicollinearity problem						

Table 3. SA-isolated response variate as well as covariate null hypothesis probabilities						
spatial filter	intercept	X_1: population density	X_2: elevation	X_3: longitude	$\frac{\lambda_1}{\lambda_p}$	(pseudo)-R^2
Box-Cox supported normal curve approximation cases						
spatial SAR†	< 0.0001	< 0.0001	0.1598	< 0.0001	3.5 (p = 4)	0.5613
spatial Durbin	< 0.0001	< 0.0001	0.1552	0.0178	107.6 (p = 7)	0.5970
Getis	< 0.0001	< 0.0001	0.3431	0.1840	9.5 (p = 7)	0.6256
MESF	< 0.0001	< 0.0001	0.6929	0.0151	6.5 (p = 8)	0.6536
GLM theory: a gamma random variable						
MESF	< 0.0001	< 0.0001	0.6727	0.0079	8.0 (p = 9)	0.6616
† denotes the spatial simultaneous autoregressive (aka spatial econometrics spatial error) model in which ρ estimated for Y rescales isolated SA in it as well as filters all covariates						
NOTE λ_1/λ_p: 1-10 expected; 10-100 multicollinearity concern; 100+ multicollinearity problem						

The remarkable unambiguous revelations documented by Tables 1-4 for this exclusive case study are that the detected weak-to-moderate positive SA encumbering the Puerto Rican geospatial urban wait times data—disregarded by Besag's auto-gamma model—can mask and/or distort true geospatial relationships. Ignoring SA (Table 1) implies that all three covariates are statistically significant, multicollinearity is not troublesome, a normal approximation (prior to its back-transformation) and a more appropriate gamma probability model furnish equally good data descriptions, and the three covariates together account for roughly half of the geographic variation in urban electricity restoration wait times. Completely eliminating SA in the data (Table 4) tells a different story: population density is the single important covariate—with evidence also signalling distance from the federal relief entrance point to the island being prominent—the normal approximation and gamma specifications do not perform equivalently, multicollinearity remains undisruptive, and the electricity restoration process potentially is more random than map patterned. Preserving SA latent in wait times (Tables 2 and 3), which is both sensible and satisfying given the incremental contiguous expansion of the destroyed electric grid's restoration, not only maintains, but also bolsters a spatial statistical data description. Elevation appears to be a questionable covariate. In addition, explicitly exposing SA in covariates suggests that it can exacerbate multicollinearity issues (e.g., the spatial Durbin and possibly the Getis filtering cases in Table 3). One overall implication is that normal approximations can be misleading; their results also may be overly optimistic because they still require back-transforming for final decision-making purposes. Therefore, this empirical example nicely illustrates the entire spectrum of spatial filtering perspectives and options currently available to spatial analysts.

Table 4. SA-free response variate as well as covariate null hypothesis probabilities

spatial filter	intercept	X_1: population density		X_2: elevation		X_3: longitude		$\frac{\lambda_1}{\lambda_3}$	(pseudo)-R^2
		p-value	VIF	p-value	VIF	p-value	VIF		
Box-Cox supported normal curve approximation cases									
Cochrane-Orcutt	< 0.0001	< 0.0001	1.24	0.2236	1.22	0.0094	1.03	2.6	0.4745
Getis	< 0.0001	< 0.0001	1.14	0.5356	1.14	0.0941	1.00	2.1	0.4287
MESF	0.0005	0.0034	1.96	0.0009	1.68	< 0.0001	1.25	5.9	0.5199
GLM theory: a gamma random variable									
MESF	< 0.0001	0.0018		0.0004		< 0.0001		5.9	0.1230

NOTE λ_1/λ_p: 1-10 expected; 10-100 multicollinearity concern; 100+ multicollinearity problem

4. The Future of Spatial Filtering Models

This narrative synthesizes a definition of contemporary spatial filtering models, lesser known but very powerful instruments for handling SA that already are in the spatial analysis toolbox, casting them in terms of spatial autoregressive, Getis-Ord statistic decomposition, and MESF models. As such, this construct spans the full range of spatial statistics/econometrics, from Cliff-Ord/Besag type spatial auto-, through Getis's geostatistically informed, to Griffith's Moran eigenvector-based, conceptualizations, all supporting full eradication from, or sundry degrees of partial isolation of SA in, geospatial data. As more and more present-day quantitative spatial analysts become convinced that SA is a fundamental, rather than menacing, georeferenced data property, partial solutions that clarifyingly shift SA around in an equation will replace full SA jettisonings as the modus operandi norm. MESF offers a clear advantage in this arena because it not only is capable of efficiently and effectively handling positive-negative SA mixtures, a theme aligned with this article whose widespread existence increasingly is becoming commonly recognized, but also easily integrates with any non-normal probability model having a regression implementation, avoiding the historically dominant normal curve theory approximations and their attendant specification error.

References

Borcard, D., and Legendre, P. (2002). All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrixes. Ecological Modeling, 153: 51-68.

Cochrane, D., and Orcutt, G. H. (1949). Application of least squares regression to relationships containing auto-correlated error terms. Journal of the American Statistical Association, 44 (245): 32-61.

Getis, A. (1990). Screening for spatial dependence in regression analysis. Papers of the Regional Science Association, 69: 69-81.

Getis, A. (1995). Spatial filtering in a regression framework: examples using data on urban crime, regional inequality, and government expenditures. In: L. Anselin and R. Florax (eds.), New Directions in Spatial Econometrics, pp. 172-188. Berlin: Springer.

Getis, A., and Griffith, D. (2002). Comparative spatial filtering in regression analysis. Geographical Analysis, 34:130-140.

Griffith, D. (2008). Spatial filtering, in K. Kemp (ed.), Encyclopedia of Geographic Information Science. Thousand Oaks, CA: SAGE, pp. 413-415.

Griffith, D. (2010). Spatial filtering). Chapter B.5 in A. Getis and M. Fischer (eds.), Handbook of Applied Spatial Analysis. Berlin: Springer-Verlag, pp. 301-318.

Griffith, D. (2020). A Family of Correlated Observations: From Independent to Strongly Interrelated Ones. Stats, 3: 166-184.

Griffith, D. A. and Getis, A. (2016). Spatial filtering. In Shekhar, S., H. Xiong, and X. Zhou, Encyclopedia of GIS (2nd ed.). Cham, Switzerland: Springer, pp. 1-14.

Griffith, D. A., and Chun, Y. (2016). Evaluating eigenvector spatial filter corrections for omitted georeferenced variables. Econometrics, 4: 29.

Griffith, D., and Peres-Neto, P. (2006). Spatial modeling in ecology: the flexibility of eigenfunction spatial analyses. Ecology, 87: 2603-2613.

Griffith, D., Chun, Y., and Li, B. (2019). Spatial Regression Analysis Using Eigenvector Spatial Filtering, Cambridge, MA: Elsevier.

Juhl, S. (2021). spfilteR: An R package for semiparametric spatial filtering with eigenvectors in (generalized) linear models. R Journal, 13(2): 450-459.

Le Gallo, J., and Páez, A. (2013). Using synthetic variables in instrumental variable estimation of spatial series models. Environment and Planning A, 45(9): 2227-2242.

Murakami, D. (2020). spmoran: Moran Eigenvector-Based Scalable Spatial Additive Mixed Modeling. R package version 0.2.0-2.

Paez, A. (2019). Using spatial filters and exploratory data analysis to enhance regression models of spatial data. Geographical Analysis, 51(3): 314-338.

Thayn, J., and Simanis, J. (2013). Accounting for spatial autocorrelation in linear regression models using spatial filtering with eigenvectors. Annals of the Association of American Geographers, 103(1): 47-66.