Estadística Española • Volumen 54, número 177 / 2012, pp. 7-34

Space, time, and space-time eigenvector filter specifications that account for autocorrelation

Daniel A. Griffith
University of Texas at Dallas

Abstract

Currently the development of eigenvector spatial filtering theory and methodology focuses on spatial autocorrelation. An overview of this development is summarized, and extended to serial correlation in time series, and space-time autocorrelation structures. Temporal comparisons are made with ARIMA model results. Illustrations are with linear and generalized linear model descriptions for selected datasets.

Keywords: serial correlation, spatial autocorrelation, space-time autocorrelation, eigenvector.

JEL classification: C5, R0

AMS classification: 62-07, 62P25

Especificaciones de filtrado espacial, temporal y espacio-temporal mediante autovectores para casos de autocorrelación

Resumen

Actualmente el desarrollo de la teoría y la metodología de filtrado espacial mediante autovectores está centrada en la autocorrelación espacial. En este artículo se resume dicho desarrollo y se extiende tanto a la correlación serial en el campo de las series temporales como a las estructuras de correlación espacio-temporales. Las comparaciones se llevan a cabo con modelos ARIMA. Se aportan ilustraciones de la metodología propuesta mediante modelos lineales y modelos lineales generalizados para una serie de bases de datos debidamente seleccionadas.

Palabras clave: correlación serial, autocorrelación espacial, autocorrelación espacio-temporal, autovector.

Clasificación JEL: C5, R0

Clasificación AMS: 62-07, 62P25

1. Introduction

Classical mathematical statistics avoids correlation amongst observations by invoking the assumption of independent observations, which sets the covariance term to 0 in expressions such as the variance of a linear combination of random variables. Consequently, joint probability distributions are the products of marginal probability distributions. In contradistinction, many datasets comprise a collection of observational units that are related to each other in some way, such as being adjacent in a time sequence (e.g., repeated measures, time series), being linked members in a subgroup of a population (e.g., social networks), or being neighbors in a geographic distribution (e.g., spatial series). Consequently, joint probability distributions are the products of conditional probability distributions. Autoregressive model specifications furnish one popular modification to classical mathematical statistics to capture these observational dependences. In a regression context, these specifications result in the response variable, Y, being on both sides of an equation: the left-hand side contains Y, and the right-hand side contains some linear combination of observed y_is, the n realizations of Y, such that each y_i is not a function of itself (i.e., its coefficient is 0 in the linear combination). In classical mathematical statistics, all of the weights for this linear combination are zero. Another popular approach deals with the inter-correlations among observations (e.g. geostatistics, spectral analysis). A third approach, whose formulation and on-going development is more recent, is eigenvector filtering.

Eigenvector filtering is a general data analysis methodology that uses a set of synthetic proxy variates, which are based upon some articulation that ties observations together, as control variables in a model specification. These control variables identify and isolate stochastic dependencies among the observations, thus allowing modeling to proceed as if these observations are independent. The purpose of this paper is to present an overview of this eigenvector filtering concept, extending it from spatial to both time series and space-time data. Background discussion conceptualizes it for time series, and then focuses on its initial development that has taken place in terms of eigenvector spatial filtering. Results are illustrated with selected empirical examples.

2. Eigenvector temporal filtering

The Box-Jenkins ARIMA (autoregressive integrated moving average) modeling is well developed for time series data. Dependency is one directional and one dimensional, which simplifies the handling of within variable correlation. One problematic complication arises when observations in time are not uniformly spaced. Perhaps data were not collected at uniform intervals through time, or a data generating process does not produce values that are equally spaced through time. R is one of the few software environments currently supporting analysis of irregularly space time series data (e.g., approx.irts). Other software environments support an autoregressive error model with missing values, which are retained in order to maintain uniform spacing of observations.

The former allows forecasting, whereas the latter allows imputation of past values (i.e., interpolation within the context of backcasting).

Filtering methodology for time series data is in keeping with the goals of ARIMA modeling—and is in the spirit of the Cochrane-Orcutt (1949) pre-whitening perspective— while exploiting the strategy of autoregressive error modeling. It uses a set of temporal proxy variables, which are extracted as eigenvectors from a modified (i.e., doubly centered) binary temporal relationship matrix that ties time series observations together, and adds these vectors as control variables into a model specification. These control variables identify and isolate the stochastic temporal dependencies among the time-indexed observations, thus allowing model building to proceed as if the observations are independent. The eigenvectors involved relate to the Durbin-Watson (DW) statistic, whose matrix version for T points in time is a T-by-1 response vector **Y** regressed on a T-by-(p+1) covariate matrix **X** (i.e., p covariates and a vector of 1s for the intercept), which is given by

$$\frac{\mathbf{Y}^T [\mathbf{I} - \mathbf{X} (\mathbf{X}^T\mathbf{X})^{-1} \mathbf{X}^T] \mathbf{A} [\mathbf{I} - \mathbf{X} (\mathbf{X}^T\mathbf{X})^{-1} \mathbf{X}^T] \mathbf{Y}}{\mathbf{Y}^T [\mathbf{I} - \mathbf{X} (\mathbf{X}^T\mathbf{X})^{-1} \mathbf{X}^T] \mathbf{Y}}, \qquad [1]$$

where superscript T denotes matrix transpose, and T-by-T matrix **A** is defined as follows:

$$\begin{pmatrix} 1 & -1 & 0 & \cdots & 0 & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 & 0 \\ 0 & -1 & 2 & \cdots & 0 & 0 & 0 \\ 0 & 0 & -1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{pmatrix}.$$

Matrix **A** is asymptotically equal to matrix $2\mathbf{I} - \mathbf{C}_T$, where **I** is a T-by-T identity matrix and $\mathbf{C}_T$ is a T-by-T binary geographic weights matrix for a linear landscape (i.e., it has 1s in its upper and lower off-diagonals, and 0s elsewhere). Basilevsky (1983) reports the analytical i^{th} element for eigenvector j for this matrix to be $e_{ij} = \frac{1}{\sqrt{T+1}} \text{SIN}\left(\frac{ij\pi}{T+1}\right)$.

When no covariates are present (i.e., p = 0), then $\mathbf{X} = \mathbf{1}$, a T-by-1 vector of ones, and the T-by-T projection matrix ($\mathbf{I} - \mathbf{11}^T/n$) results in the first eigenvector being replaced by a vector proportional to **1** whose corresponding eigenvalue is 0. It also forces each of the remaining T–1 eigenvectors to have a zero mean, and hence to be mutually orthogonal and uncorrelated. These are the previously mentioned synthetic control variates. These eigenvectors can be easily approximated by generating the set of j = 2, 3,… T analytical eigenvectors, and then subjecting these vectors to a factor analysis. Forecasts can be generated by constructing the analytical eigenvectors for a T + Q matrix $\mathbf{C}_T$, and then imputing the Q future values (i.e., forecasting them).

Substituting each eigenvector into equation (1) results in a DW value that equals its corresponding eigenvalue. In other words, the DW values range from $2-2\text{COS}\left(\frac{2\pi}{T+1}\right)$ to $2+2\text{COS}\left(\frac{T\pi}{T+1}\right)$, or approximately 0 to 4 (2 indicates zero temporal autocorrelation). The number of positive and of negative eigenvalues equals (T – 1)/2 if T is odd; when T is even, the number of positive eigenvalues is (T – 2)/2 and the number of negative eigenvalues is (T – 1)/2. Meanwhile, because substituting the eigenvectors into equation (1) results in a Rayleigh quotient, with vector $\mathbf{E}_1$ maximizing the expression, these eigenvectors can be interpreted as follows:

the first eigenvector, say $\mathbf{E}_1$, is the set of real numbers that has the largest DW achievable by any set for the temporal arrangement defined by the time-series connectivity matrix **A**; the second eigenvector is the set of real numbers that has the largest achievable DW by any set that is orthogonal and uncorrelated with $\mathbf{E}_1$; the third eigenvector is the third such set of real numbers; and so on through $\mathbf{E}_T$, the set of real numbers that has the largest negative DW achievable by any set that is orthogonal and uncorrelated with the preceding (T – 1) eigenvectors.

As such, these eigenvectors furnish distinct temporal pattern descriptions of latent serial correlation in time series variables.

2.1 Annual sugarcane production in Puerto Rico: an exploratory interpolation experiment

Griffith (2008a) describes annual sugarcane production, in 1,000s of tonnes, for the time series 1828-1996 using an ARIMA model specification. He subjects these data to a logarithm transformation to stabilize variance, and introduces first differencing to account for the dominant trend, and an indicator variable to differentiate between the Spanish and the United States (U.S.) control of the island (the transition occurred in 1899). These data furnish a useful time series for experimental purposes because the series is of a reasonable length, is complete, and already has been described with an ARIMA model.

These data could be described by an eigenvector temporal filter (ETF), or by a semi-variogram model (Figure 1). The basic time structure may be accounted for with a set of four covariates: the year (centered; which relates to the ARIMA model differencing), the pre-U.S. control indicator variable, and second- and third-powers of the centered year covariate (to capture the concave sections of the time series plot). These four variables were converted to orthogonal synthetic variates with factor analysis. Three of the factors account for roughly 78% of the variability in the transformed sugarcane data; the residuals from this regression are not normal.

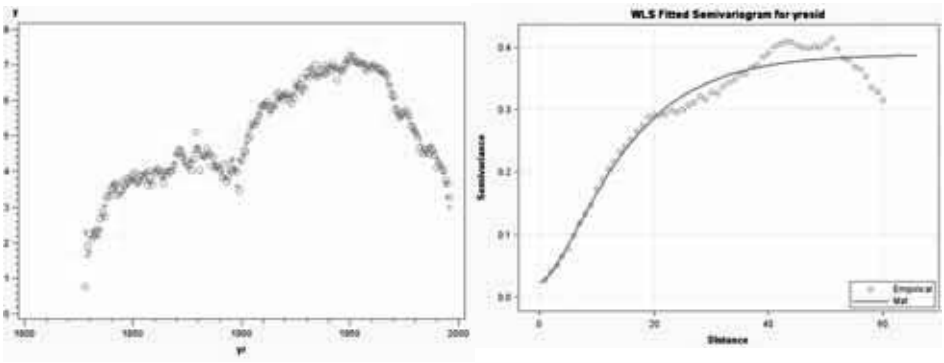
Twenty-four eigenvectors extracted from matrix $\mathbf{[I-X(X^TX)^{-1}X^T]A[I-X(X^TX)^{-1}X^T]}$ appearing in the numerator of expression (1) account for another roughly 11% of the

variation in the transformed sugarcane data (Table 1)[1]. The residuals from this augmented model still are not normal, but very little evidence exists suggesting the presence of heteroscedasticity, and no serial correlation appears to remain in the residuals (DW = 1.69). Figure 1.a portrays the time series plot for the data as well as the predicted values from the ETF model. The filter, which can be constructed without using the year indices, also replicates the time series reasonably well.

One advantage of this type of specification is that if annual values are missing at random from a time series, it still can be well-described (Table 1). One disadvantage is that it involves only fixed effects (i.e., the time-trend terms and the eigenvectors, all of which do not change unless the time horizon changes). A set of simulation experiments involving random suppression of response values (i.e., missing at random), and based upon 10,000 replications, reveals that stepwise regression selects all of the highly significant time trend terms and eigenvectors, and yields R^2 values that are equivalent to that for the complete data time series, even with 50% of the values suppressed (Table 1). Occasionally the stepwise procedure selects a few incorrect eigenvectors. In other words, the ETF model furnishes an excellent tool for the interpolation of irregular time series version of these data.

Figure 1

Annual sugarcane time series data. Left (a): log-transformed sugarcane yield time series plot. Right (b): semi-variogram plot of log-transformed sugarcane yield time series data, with a superimposed Bessel function trend line



[1] Because eigenvectors $\mathbf{E}_{49}$, $\mathbf{E}_{50}$ and $\mathbf{E}_{70}$ are highly multicollinear with the four time trend factors, they were removed from the candidate set of eigenvectors.

Table 1

Puerto Rico sugarcane ETF parameter estimates and simulation results

(Continue)

Variable	Complete time series data			% inclusion (10,000 replications) for % missing values				
	Parameter Estimate	Standard Error	Pr > F	10	20	30	40	50
Intercept	5.0120	0.0176	<0.0001	100.00	100.00	100.00	100.00	100.00
Factor 1	0.1276	0.0239	<0.0001	100.00	100.00	100.00	99.98	100.00
Factor 2	1.0858	0.0201	<0.0001	100.00	100.00	100.00	100.00	100.00
Factor 3	-0.7546	0.0178	<0.0001	100.00	100.00	100.00	100.00	100.00
Factor 4	0.3601	0.0249	<0.0001	100.00	99.99	99.97	99.99	99.98
E_1	-0.4193	0.0222	<0.0001	100.00	100.00	100.00	100.00	100.00
E_2				0.02	0.02	0.05	0.06	0.03
E_3				0.02	0.01	0.03	0.03	0.03
E_4	-0.0830	0.0179	<0.0001	99.99	100.00	99.97	99.98	99.96
E_5				0.01	0.03		0.03	0.04
E_6	-0.0785	0.0177	<0.0001	100.00	100.00	100.00	99.98	99.96
E_7				0.01	0.01	0.02	0.04	0.03
E_8	-0.0535	0.0179	0.0033	99.98	99.99	99.97	99.96	99.94
E_9				0.01	0.01		0.04	0.03
E_{10}						0.01	0.01	0.03
E_{11}						0.01	0.01	0.01
E_{12}					0.01	0.01	0.02	0.01
E_{13}	-0.0535	0.0177	0.0030	100.00	100.00	99.97	99.94	99.93
E_{14}	0.0427	0.0177	0.0175	99.97	99.93	99.96	99.96	99.93
E_{15}				0.01		0.02	0.04	0.02
E_{16}							0.04	0.02
E_{17}					0.01			0.01
E_{18}								0.01
E_{19}							0.01	0.01
E_{20}	-0.0443	0.0177	0.0135	99.99	99.95	99.97	99.94	99.94
E_{22}				0.01		0.03	0.02	
E_{23}								0.01
E_{24}						0.01	0.03	
E_{25}								0.01
E_{27}	-0.0375	0.0177	0.0360	99.95	99.94	99.93	99.90	99.91
E_{29}						0.01	0.02	0.01
E_{30}							0.02	
E_{31}							0.01	
E_{32}						0.01		
E_{34}				0.02		0.02	0.01	
E_{35}	-0.0667	0.0182	0.0004	99.99	99.98	99.94	99.96	99.95
E_{36}	-0.0808	0.0180	<0.0001	99.99	100.00	99.97	99.98	99.96

Table 1

Puerto Rico sugarcane ETF parameter estimates and simulation results

(Conclusion)

Variable	Complete time series data			% inclusion (10,000 replications) for % missing values				
	Parameter Estimate	Standard Error	Pr > F	10	20	30	40	50
E_{37}				0.01			0.01	0.01
E_{38}				0.01	0.01	0.01		0.01
E_{39}						0.02	0.01	
E_{40}					0.02	0.01		0.01
E_{41}	-0.0980	0.0180	<0.0001	100.00	100.00	99.99	100.00	99.97
E_{42}				0.01	0.01	0.01	0.02	0.01
E_{43}							0.03	0.01
E_{44}	-0.0607	0.0181	0.001	99.99	99.95	99.93	99.96	99.95
E_{45}	-0.1589	0.0191	<0.0001	100.00	100.00	100.00	100.00	99.99
E_{46}						0.01	0.01	
E_{47}	-0.0616	0.0177	0.0007	99.99	100.00	99.96	99.95	99.93
E_{48}						0.01	0.01	0.03
E_{51}	-0.1408	0.0199	<0.0001	99.99	99.99	99.97	99.99	99.96
E_{52}	-0.0645	0.0178	0.0004	100.00	99.96	99.96	99.97	99.94
E_{53}					0.01	0.02		
E_{54}	-0.1411	0.021	<0.0001	99.99	99.99	99.94	99.97	99.96
E_{55}	-0.0437	0.0177	0.0148	99.96	99.96	99.94	99.92	99.93
E_{56}	-0.0420	0.0177	0.0192	99.98	99.94	99.95	99.93	99.92
E_{57}					0.01	0.01		0.02
E_{58}							0.01	0.01
E_{59}	-0.0974	0.0185	<0.0001	100.00	100.00	99.97	100.00	99.97
E_{60}	-0.0982	0.0187	<0.0001	99.99	99.99	99.93	99.95	99.96
E_{61}				0.02	0.07	0.07	0.06	0.05
E_{62}	-0.0402	0.0177	0.0245	99.94	99.97	99.92	99.93	99.91
E_{63}							0.01	0.01
E_{64}	-0.0419	0.0177	0.0194	99.94	99.95	99.93	99.92	99.90
E_{66}	-0.0407	0.0179	0.0240	99.92	99.93	99.94	99.94	99.92
E_{67}							0.01	
E_{68}				0.01	0.01	0.01	0.03	
E_{69}					0.01		0.04	
R^2		0.9782		0.9782 (0.0002)	0.9782 (0.0002)	0.9782 (0.0004)	0.9783 (0.0004)	0.9782 (0.0005)

Semi-variogram models were developed for two-dimensional data, and with more recent extensions to three-dimensional data (e.g., Heuvelink and Griffith, 2010). But they also can be employed with one-dimensional data such as a time series. Interpolation is irrelevant when a complete time series exists; it is relevant when an

irregular time series exists. Figure 1b portrays the K-Bessel (or Matérn) function description of the residuals from a regression of the transformed sugarcane figures on the time trend terms. Table 2 summarizes semi-variogram estimation simulation results for randomly suppressed values. Not surprisingly, the principal impact is on the standard deviation (s.d.). Estimates for the nugget are very stable, in part because its lower bound is restricted to be 0. In an attempt to capture as much structural trend as possible, 100% of the regressions include all four of the time trend factors.

Table 2

Puerto Rico sugarcane temporal semi-variogram model estimates and simulation results

Parameter	Complete data		% inclusion (10,000 replications) for % missing values									
			10		20		30		40		50	
	esti-mate	s.e.	esti-mate	s.d.	esti-mate	s.d.	esti-mate	s.d.	esti-mate	s.d.	esti-mate	s.d.
Nugget	0.0123	0.0117	0.0123	0.0003	0.0123	0.0002	0.0123	0.0003	0.0123	0.0003	0.0123	0.0004
Partial sill	0.3794	0.0163	0.3794	0.0008	0.3794	0.0014	0.3794	0.0021	0.3794	0.0017	0.3794	0.0035
Range	9.3180	0.8968	9.3182	0.0464	9.3181	0.0435	9.3179	0.0636	9.3174	0.0572	9.3188	0.1100
Nu	0.9411	0.0802	0.9412	0.0039	0.9412	0.0049	0.9412	0.0055	0.9411	0.0062	0.9411	0.0113

2.2 Daily atrazine levels (ppb) in drinking water: an exploratory non-normal probability model experiment

Atrazine is a principal herbicide used extensively by farmers during the last half century to control weeds in the production of, especially, sugarcane, corn, and sorghum. Human health risk concerns arise from this chemical entering community water systems (CWSs) through watershed runoff after rainfall events following its application to farm fields. Concentrations in runoff surface water are a function of such factors as hydrology, meteorology, soil type, agronomic and land use practices, herbicide application rates, timing and methods, and environmental fate properties of atrazine. Consequently, the government monitors finished drinking (i.e., treated) water for levels of the chemical. One goal of this monitoring program is to minimize the costly and time-consuming collection of daily samples for monitoring purposes. In other words, government agency personnel want to be able to effectively and efficiently analyze irregular time series during each annual growing season, recognizing that atrazine levels essentially are zero outside of a growing season. Irregular time series are to be converted to regular time series by interpolation.

From May 15 (three days before a rainfall event) to June 30, 2011, Syngenta Corporation Protection, LLC, collected and assayed daily water samples from a CWS supported by a 4,055-square-mile watershed, and whose water source is a river[2]. These

[2] CWS-45 designates this system. See Table 2 in the attachment to entry ID EPA-HQ-OPP-2011-0399-0042 that can be accessed on page 6 of 7 at http://www.regulations.gov/#!searchResults;dct=SR%252BFR%252BPS;a=EPA;dkt=N;pd==07%257C01%257C11-08%257C24%257C11;rpp=10;po=50;s=atrazine.

45 consecutive days of data were subjected to a logarithmic transformation—LN(atrazine+0.13)— similar to that for the preceding sugarcane data. But one notable feature of these data is that they are better described by a 3-parameter Weibull distribution (Figure 2). One difference between this atrazine and the preceding sugarcane time series is that the latter portrays a serpentine curve with a single frequency spike, whereas the atrazine time series portrays a serpentine curve with three consecutive frequency spikes having different amplitudes—a much more complex time series.

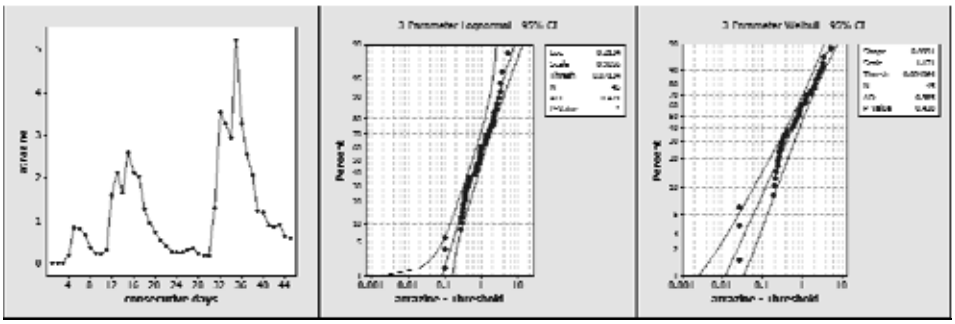
These data are well described by the following ARIMA model (unit root tests imply the need for a first differencing):

$$\hat{y}_t = 0.0361 + 1.4016y_{t-1} - 0.5754y_{t-2} + 0.1738y_{t-3},$$

which accounts for roughly 80% of the variation in log-atrazine—roughly 65% of the variation in atrazine after back-transformation—and whose residuals contain only trace serial correlation but fail to conform closely to a normal distribution (Figure 3a).

Figure 2

Daily atrazine time series data. Left (a): time series plot. Middle (b): log-normal quantile plot. Right (c): Weibull quantile plot

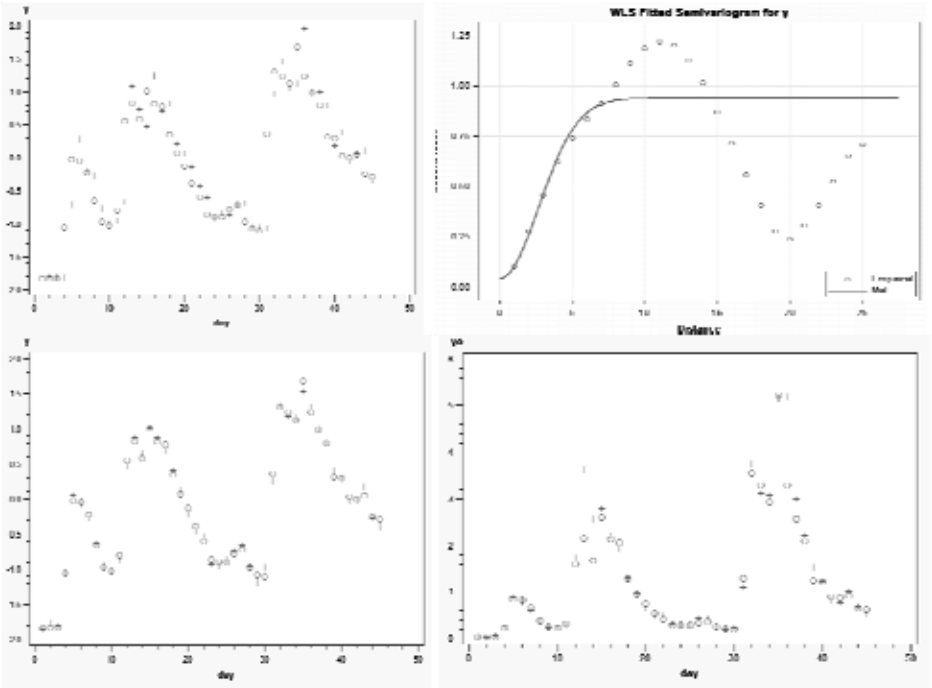


The ARIMA model findings suggest that the Bessel function is the most appropriate semi-variogram model. But Figure 3b reveals that it furnishes a poor description of the autocorrelation structure in the log-atrazine data (the description improves little by convoluting a Bessel with a wave hole function to try to capture the conspicuous periodicity). Meanwhile, the ARIMA model accounts for considerably less variance than is accounted for by the ETF model (roughly 99% in both its log-transformed and back-transformed versions; it contains 22 eigenvectors), whose residuals contain only trace serial correlation and conform very closely to a normal curve. In other words, the ETF model appears to furnish a superior description for these data in a conventional context. Its additional advantage is that the ETF model can furnish a description in terms of a Weibull distribution, which currently is not possible with the ARIMA and geostatistical models. With this distributional assumption, the ETF specification accounts for roughly 93% of the variability in untransformed atrazine measures (Figure 3d). Its residuals conform closely to

an extreme value distribution. This latter modeling exercise furnishes a proof of concept demonstration.

Figure 3

Daily atrazine time series data. Top left (a): log-transformed atrazine time series plot with ARIMA fit superimposed. Top right (b): semi-variogram plot of log-transformed atrazine time series data, with a superimposed Bessel function trend line. Bottom left (c): log-transformed atrazine time series plot with normal ETF fit superimposed. Bottom right (d): atrazine time series plot with Weibull ETF fit superimposed



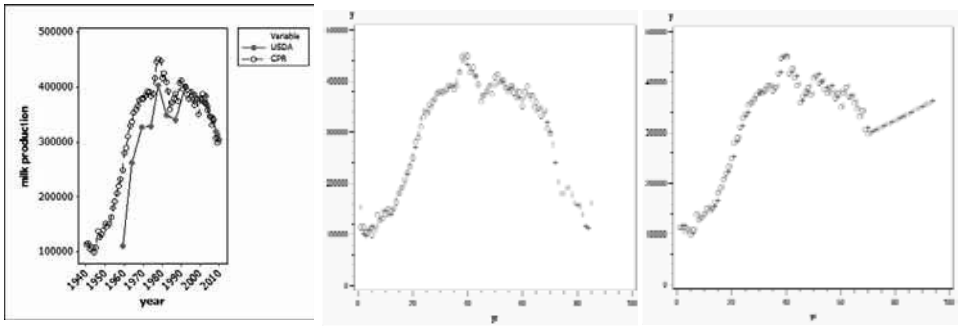
2.3 Annual milk production in Puerto Rico: an exploratory forecasting experiment

The previous two experiments focus on interpolation. The experiment summarized in this section involves the more common practice of forecasting with a time series. The Department of Agriculture, Commonwealth of Puerto Rico (CPR) reported annual milk production in Puerto Rico, for the period 1939/40-2008/09. Unlike the preceding two, this time series does not disappear; rather, it is ongoing. The 2009/10 annual forecast is compared with the CPR estimate. Two features of these data are noteworthy. First, the trend is roughly the same as that for the sparser U.S. Department of Agricultural (USDA) data, but the USDA figures tend to be less than the CPR figures through about

2000—most likely because the two government agencies defined farm slightly differently, and used a different definition of production year until the CPR became more closely affiliated with USDA through the National Agricultural Statistics Service (NASS). Second, the trend beyond 2002 is for deceasing milk production.

Figure 4

Milk production in Puerto Rico. Left (a): time series plot of reported quantities. Middle (b): ETF desciption of the CPR statistics concatenated with forecasts. Right (c): ARIMA description of the CPR statistics concateneted with forecasts



The ARIMA model equation describing this time series is given by:

$$\hat{y}_t = 2667 + y_{t-1},$$

which accounts for roughly 98% of the variation in 1,000s of quarts of milk (the bivariate regression equation relating these two variates is: $Y = 10845 + 0.9655\,\hat{Y} + e$), and whose residuals contain only trace serial correlation and conform to a normal distribution. But the ARIMA forecasts predict a turn-around growth in milk production. This model yields a mean squared prediction error (MSPE) of 138.9×10^3, for the CPR prediction of 288,964 thousands of quarts of milk for 2009/10. A larger MSPE of 11.6×10^3 results from a comparison with USDA reported tons of milk produced in 2010 (which translate into roughly 304,151 thousands of quarts[3]). Although these MSPEs are larger than their ETF counterparts for a one-year-ahead forecast, the ARIMA 15-years-ahead forecast trend line is inconsistent with existing data.

Meanwhile, the ETF model employs eigenvectors that span both the observed CPR time series and the forecast period (i.e., T = 72). Its candidate set includes 35 eigenvectors portraying positive serial correlation. Because quarts of milk are counts, the response variable is treated as a negative binomial random variable (i.e., a Poisson random variable with overdispersion). The constructed ETF consists of 13 eigenvectors, and accounts for roughly 99% of the variation in the milk production figures (the bivariate

[3] The average cow produces 53 pounds of milk a day, which on average converts to 6.2 gallons of milk.

regression equation relating these two variates is: Y = -2679.0240 + 1.0092^Y + e). The dispersion parameter estimate is 0.0031, indicating little initial overdispersion; the resulting deviance statistic is 1.25. The ETF forecasts predict a continued downturn in milk production. This model yields a MSPE of 170.8×10^3, for the CPR prediction of 288,964 thousands of quarts of milk for 2009/10. A smaller MSPE of 798.3×10^3 results from a comparison with USDA reported tons of milk produced in 2010. Although these MSPEs are larger than their ARIMA counterparts for a one-year-ahead forecast, the ETF 15-years-ahead forecast trend line is consistent with existing data.

2.4 A summary of findings for temporal filtering

Constructed ETFs perform very well in the selected empirical examples. This specification outperforms geostatistical models when interpolating, and it shows promise of outperforming an ARIMA model when forecasting. It also is capable of accommodating non-normal statistical distributions, such as the Weibull and the Poisson (or its negative binomial overdispersion counterpart). Because of its relative newness, the ETF's most conspicuous weakness is its lack of a more fully developed mathematical statistics theoretical basis. Both geostatistical and Box-Jenkins models have such bases.

3. Eigenvector spatial filtering

The preceding discussion describes how eigenvector filtering can be applied to time series data, essentially treating them as one-dimensional geographic data. But the initial formulation and development of this filtering methodology was for two-dimensional geographic data (see Griffith, 2000, 2002, 2003, 2004; Tiefelsdorf and Griffith, 2007). In this context, eigenvector spatial filtering uses a set of spatial proxy variables, which are extracted as eigenvectors from an n-by-n modified (i.e., doubly centered) binary spatial relationship matrix, $\mathbf{C}_s$—which has 1s in the cells whose row and column locations are neighbors, and 0s elsewhere—that ties geographic observations together[4] (Figure 5), and adds these vectors as control variables into a model specification. Like before, these control variables identify and isolate the stochastic spatial dependencies among the location-indexed observations, thus allowing model building to proceed as if the observations are independent. The eigenvectors involved usually relate to the Moran Coefficient (MC)[5], whose matrix version for n-by-1 response vector $\mathbf{Y}$ adjusted only for its mean is given by

$$\frac{\mathbf{Y}^T(\mathbf{I}-\mathbf{1}\mathbf{1}^T/n)\mathbf{C}_s(\mathbf{I}-\mathbf{1}\mathbf{1}^T/n)\mathbf{Y}}{\mathbf{Y}^T(\mathbf{I}-\mathbf{1}\mathbf{1}^T/n)\mathbf{Y}}. \qquad [2]$$

[4] The entries of matrix $\mathbf{C}_s$ are c_{ij}, which equals 1 if areal units i and j are geographic neighbors, and 0 otherwise. Neighborness often is defined as whether or not two areal unit polygons share a common boundary (non-zero length sharing is called the rook's case; zero and non-zero length sharing is called the queen's case), or are within a specified distance of each other (frequently with distance measured between areal unit centroids).

[5] It also can be based upon the Geary Ratio (GR).

Unless the set of polygons comprising a surface partitioning forming a regular square tessellation, analytical eigenvectors for expression (2) are unknown.

Figure 5

Left (a): the partitioning of Puerto Rico into municipalities. Gray lines denote the topological network, which connects the centroids of municipalities with common boundaries. Right (b): the initial part of the 73-by-73 connectivity matrix for Puerto Rico

$$\mathbf{C} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ & & & \vdots & & & \end{bmatrix} \end{array} \cdots$$

Again, because substituting the eigenvectors into equation (2) results in a Rayleigh quotient, with vector $\mathbf{E}_1$ maximizing the expression, these eigenvectors can be interpreted as follows:

the first eigenvector, say $\mathbf{E}_1$, is the set of real numbers that has the largest MC achievable by any set for the geographic arrangement defined by the spatial connectivity matrix **C**; the second eigenvector is the set of real numbers that has the largest achievable MC by any set that is orthogonal and uncorrelated with $\mathbf{E}_1$; the third eigenvector is the third such set of real numbers; and so on through $\mathbf{E}_n$, the set of real numbers that has the largest negative MC achievable by any set that is orthogonal and uncorrelated with the preceding (n – 1) eigenvectors.

As such, these eigenvectors furnish distinct map pattern descriptions of latent spatial autocorrelation in geographically distributed variables. An eigenvector spatial filter (ESF) is constructed as some linear combination of a subset of these eigenvectors, called the candidate set.

The candidate set begins with all eigenvectors portraying the same nature (i.e., positive or negative) of spatial autocorrelation as is measured in a response variable. Next, those eigenvectors representing inconsequential levels of spatial autocorrelation are removed from this candidate set. For positive spatial autocorrelation situations, Griffith and Chun (2009) suggest the following formula for establishing the threshold eigenvector MC value defining a candidate set of these eigenvectors for constructing an ESF:

$$\mathrm{MC}_{\mathrm{eigenvector}} \geq 2.9970 - \frac{2.8805}{1 + e^{-0.6606 - 0.2525\, z_{\mathrm{MC}}}}, \qquad [3]$$

where z_{MC} denotes the z-score of the MC for the response variable Y (or its transformed version, if used). Finally, a stepwise regression procedure can be used to select those eigenvectors that account for the spatial autocorrelation in the response variable. This

stepwise selection can be based upon the conventional R^2-maximization criterion, or a residual MC minimization criterion (see the R package spdep).

3.1 2010 population density across Puerto Rico: a comparison of autoregressive and ESF results

Population density (i.e., total population count divided by total area) is a continuous variable with a lower bound of 0. A log-transformed version of it often conforms closely to a normal distribution. This is the version commonly described with a Cliff-Ord spatial simultaneous autoregressive (SAR) model, frequently employing the row-standardized version of spatial relationship matrix **C**, namely **W**. This specification supports the use of spatial autoregressive theory to account for spatial autocorrelation with an SAR model specification (Ord, 1975), as well as residual spatial autocorrelation statistical distribution theory for the ESF specification.

The following logarithmic transformation aligns the 2010 geographic distribution of population density across Puerto Rico by municipio with a bell-shaped curve:

LN(population/area – 75) .

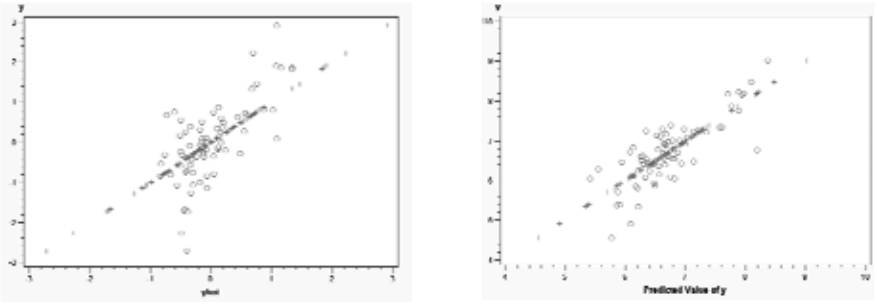
The probability of the Shapiro-Wilk (S-W) statistic increases from < 0.0001 to 0.2133. The MC for the geographic distribution of this variable is 0.5255 (z_{MC} = 7.1145). The normal approximations presented in this section utilize this transformed variable. The estimated SAR model with no covariates is

$$\hat{\mathbf{Y}} = 0.6929\mathbf{WY},$$

with the spatial lag term accounting for roughly 49% of the variation in the log-transformed population density (Figure 6a) and produces residuals that are approximately normally distributed [PR(S-W) = 0.2445].

Figure 6

Scatterplots of observed (y) versus normal-approximation (yhat/predicted value of y) predicted log-transformed population density, with superimposed ideal line. Left (a): estimated SAR model results. Right (b): estimated ESF model results



Meanwhile, equation (3) indicates that the candidate set of eigenvectors should be only those with a MC of at least 0.34385; the candidate set contains 16 eigenvectors. The linear regression constructed ESF includes 10 eigenvectors, has a MC of 0.88616, accounts for roughly 65% of the variation in the log-transformed population density (Figure 6b), and produces residuals that are approximately normally distributed [PR(S-W) = 0.0610] and have only trace spatial autocorrelation remaining (z_{MC} = 0.4512).

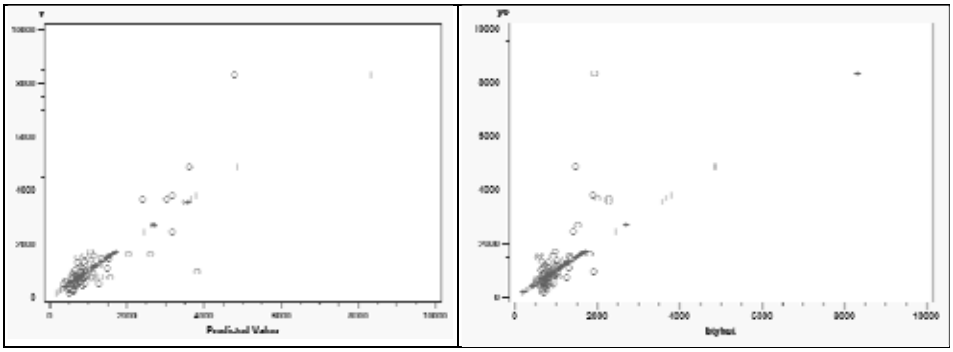
In this particular data analysis, the ESF specification essentially outperforms the SAR specification. It renders a better distribution of predicted values, whereas the SAR specification produces residuals that better conform to a bell-shaped curve.

3.2 2010 population density across Puerto Rico: an extension to a Poisson model specification

Population figures are counts, and as such should be treated as a Poisson random variable. Population by municipio exhibits excess Poisson variation, requiring it to be treated as a negative binomial random variable. The ESF constructed with this new distribution assumption includes all but one of the eigenvectors selected for the normal approximation ESF. The percent of variance accounted for increases slightly, to roughly 71%, with this more appropriate specification. The overdispersion parameter is 0.1609, indicating the presence of considerable extra Poisson variation. A comparison between this result and the back-transformed normal approximation SAR result (Figure 7b) reveals serious distortions by the employment of a Box-Cox power transformation. Now the SAR model results account for only roughly 52% of the variation in population density.

Figure 7

Scatterplots of observed (y) and predicted (predicted value/btyhat) population density, with a superimposed ideal line. Left (a): estimated negative binomial ESF model results. Right (b): back-transformed normal approximation SAR model results



The improvement by removing specification error attributable to using a normal approximation for a Poisson probability model is small but noticeable. The ESF

specification is superior to an auto-Poisson model specification here because the auto-Poisson model is incapable of capturing any positive spatial autocorrelation effects, which are the only ones present in the analyzed geographic distribution of population density.

3.3 Spatially varying coefficients: an alternative to geographically weighted regression (GWR)

Griffith (2008b) outlines methodology for employing eigenvector spatial filtering to construct geographically varying regression coefficients. In effect, the eigenvector filtering discussed in this paper results in a spatially varying intercept term, similar to the spatial lag or spatial autoregressive response model. This type of specification can be extended to the coefficients of covariates by introducing interaction terms—Hadamard products of eigenvectors and covariates—into a model specification. Estimates of the regression coefficients can be pooled for products with a common covariate; factoring out this common covariate yields the spatially varying coefficients as linear combinations of eigenvectors. This specification is more in keeping with the SAR model specification.

3.4 A summary of findings for eigenvector spatial filtering

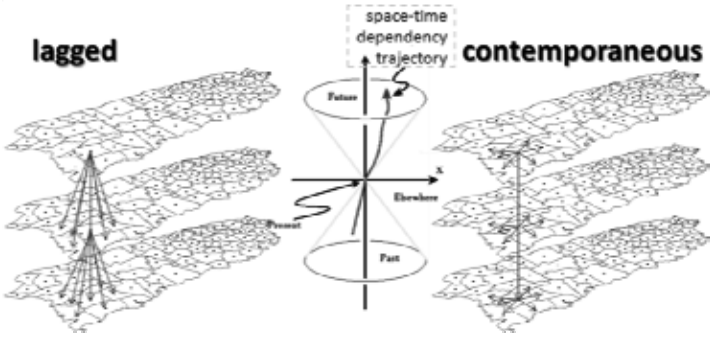
Constructed EFSs perform very well in the selected empirical examples. ESF specifications outperform autoregressive model specifications in terms of goodness-of-fit assessments. They also are capable of accommodating non-normal statistical distributions, such as the Poisson (or its negative binomial overdispersion counterpart).

4. Eigenvector space-time filtering

The preceding discussion describes how eigenvector filtering can be applied to either time series (indexed with t) or spatial series [indexed with (u, v)] data; this section synthesizes these two approaches into a space-time [indexed with (u, v, t)] methodology.

Figure 8

The linkages for the two space-time conceptualizations of dependency



Space-time autocorrelation can be accounted for in such data with a set of eigenvectors extracted from a modified (i.e., doubly centered) space-time adjacency matrix. This section summarizes an evaluation of two principal space-time dependency structure specifications (Figure 8):

– location (u, v, t) links to the preceding *in situ* location as well as the preceding neighboring locations, a lagged specification; and,
– location (u, v, t) links to the preceding *in situ* location as well as the instantaneous neighboring locations, a spatially contemporaneous specification.

This work is an extension of research summarized in Griffith (2004), Griffith (2010), and Griffith and Heuvelink (2010).

Matrix versions of the conceptualizations portrayed in Figure 8 are as follows:

space-time lagged specification:

$$(\mathbf{I} - \mathbf{1}\mathbf{1}^{\mathrm{T}}/\mathrm{n})[\mathbf{C}_{\mathrm{T}} \otimes (\mathbf{C}_{\mathrm{S}} + \mathbf{I}_{\mathrm{S}})](\mathbf{I} - \mathbf{1}\mathbf{1}^{\mathrm{T}}/\mathrm{n}), \text{ and} \quad [4]$$

space-time contemporaneous specification:

$$(\mathbf{I} - \mathbf{1}\mathbf{1}^{\mathrm{T}}/\mathrm{n})(\mathbf{I}_{\mathrm{T}} \otimes \mathbf{C}_{\mathrm{S}} + \mathbf{C}_{\mathrm{T}} \otimes \mathbf{I}_{\mathrm{S}})(\mathbf{I} - \mathbf{1}\mathbf{1}^{\mathrm{T}}/\mathrm{n}), \quad [5]$$

where: $\otimes$ denotes Kronecker product, and $(\mathbf{I} - \mathbf{1}\mathbf{1}^{\mathrm{T}}/\mathrm{n})$ is an nT-by-nT projection matrix. Griffith (1996) discusses eigenfunctions for these matrices. These are the connectivity matrices for space-time versions of the MC (Cliff and Ord, 1981; Griffith, 1981).

Analogous to the preceding discussions, eigenvector space-time filtering uses a set of synthetic proxy variables, which are extracted as eigenvectors from either a modified space-time lagged or a space-time contemporaneous connectivity matrix that ties geographic objects together in space and time, and adds these vectors as control variables into a model specification. Like before, these control variables identify and isolate the stochastic space-time dependencies among the (u, v, t)-referenced observations, thus allowing model building to proceed as if the observations are independent. The eigenvectors involved relate to the spaced-time MC, whose matrix version for nT-by-1 response vector **Y** adjusted only for its mean is given by

$$\frac{\mathbf{Y}^{\mathrm{T}}(\mathbf{I} - \mathbf{1}\mathbf{1}^{\mathrm{T}}/\mathrm{n})\mathbf{A}(\mathbf{I} - \mathbf{1}\mathbf{1}^{\mathrm{T}}/\mathrm{n})\mathbf{Y}}{\mathbf{Y}^{\mathrm{T}}(\mathbf{I} - \mathbf{1}\mathbf{1}^{\mathrm{T}}/\mathrm{n})\mathbf{Y}}, \quad [6]$$

where matrix **A** is either expression (4) or (5).

Similar to the time only and space only cases, because substituting the eigenvectors into equation (6) results in a Rayleigh quotient, with vector $\mathbf{E}_1$ maximizing the expression, these eigenvectors can be interpreted as follows:

The first eigenvector, say $\mathbf{E}_1$, is the set of real numbers that has the largest space-time MC achievable by any set for the areal unit articulation defined by the space-time connectivity matrix **A**; the second eigenvector is the set of real numbers that has the largest achievable space-time MC by any set that is orthogonal and uncorrelated with $\mathbf{E}_1$; the third eigenvector is the third such set of real numbers; and so on through $\mathbf{E}_n$, the set of real numbers that has the largest negative space-time MC achievable by any set that is orthogonal and uncorrelated with the preceding (n – 1) eigenvectors.

As such, these eigenvectors furnish distinct map patterns that may change through time in some structured way (i.e., the Kronecker products are time specific values by which geographic map pattern values are multiplied), and in doing so furnish descriptions of latent space-time autocorrelation in correlated variables. A eigenvector space-time filter (ESTF) is constructed as a linear combination of a subset of these eigenvectors, called the candidate set.

4.1 The Puerto Rico urbanization experiment: some background about island-wide urbanization

Puerto Rico has a long Western European-based urban history, covering roughly half a millennium. From its initial settlements of San Juan and San Germán in the early 1500s, the island has developed an urban dimension that, today, appears to be undergoing absorption by the San Juan metropolitan region. The evolution of this urban dimesnion on a remote island has helped to isolate it over the centuries from many contagion-type geographic effects, but not external effects via pronounced migration and port-to-world linkages, allowing the undertaking of studies in a more isolated natural experiment context. The purpose of this section is to summarize Puerto Rican urbanization during the period 1899-2000, for which U.S. federal census records are available. This landscape transformation is characterized by positive temporal and spatial autocorrelation, dependencies that produced conspicuous map patterns over time. Urban population may be cast as a percentage of total population in an area, resulting in it being a binomial random variable.

To date, the trajectory of in/decreases in Puerto Rican population follows an exponential growth curve (Figure 9a). This same trend characterizes the increase in per cent urban population on the island (Figure 9b). These trends may be described, respectively, by the following two equations:

$$\text{population: } p_t \sim P\left(e^{15.1506-0.0855LN\{1+e^{322.17-216.96[(t-1493)/100]^{0.2461}}\}}\right), \text{ pseudo-}R^2 = 0.9985,$$

$$\text{and, per cent urban: } U_t \sim b\left(p_t, \frac{1}{1+e^{17.7854-0.0382(t-1493)}}\right), \text{ pseudo-}R^2 = 0.9635,$$

where p_t denotes the population at time t, U_t denotes the urban population at time t, P denotes a Poisson distribution, and b denotes a binomial distribution.

Figure 10a portrays the historical urban centers, revealing a tendency for urban population to concentrate on the island's coastal lowlands through time. Figure 10b

portrays the digital elevation model (DEM), displaying geographic variation in elevation across the island, and revealing why the coastal areas were sensible selections for establishing urban centers. The interior city of Caguas is located in the conspicuous valley that is visible in Figure 10.b. This DEM figure suggests that elevation may be a useful covariate in describing the space-time growth of urban population across Puerto Rico.

Figure 9

Left (a): the growth of population in Puerto Rico over time. Right (b): the increase in percentage urban population in Puerto Rico over time. Asterisk (*) denotes an actual value; open circle (o) denotes a predicted value

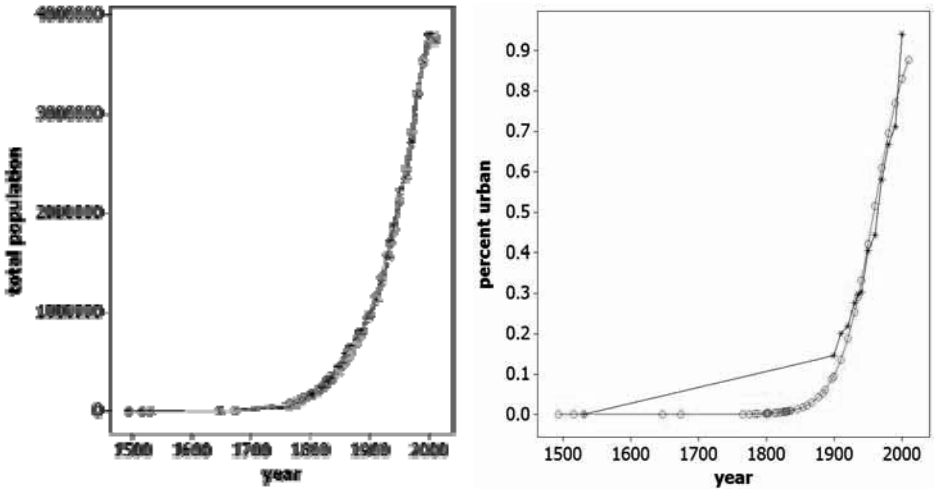
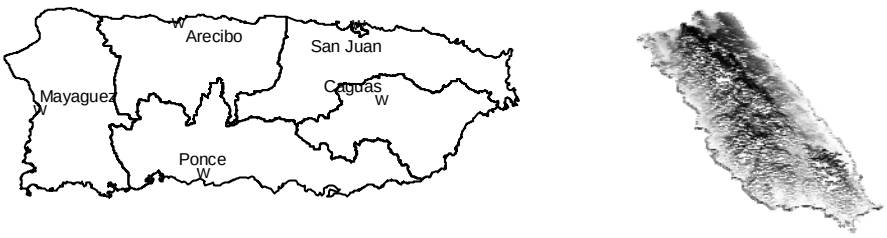


Figure 10

Left (a): historical principal cities of Puerto Rico. Right (b): DEM for Puerto Rico



4.2 Space-time urbanization in Puerto Rico, 1899-2000: the lagged specification

The lagged structure portrayed in Figure 8 casts the value at some location (u_i, v_i, t) as a function of those values at (u_i, v_i, t-1) and its neighbors (u_j, v_j, t-1), where $a_{ii,t,t-1} = 1$ and

$a_{ij,t,t-1}$ = 1in expression (4). The equation describing this formulation for a binomial percentage $\frac{1}{1+e^{\alpha_{u,v,t}}}$ is as follows:

$$\alpha_{u_i\, v_i\, t} = \beta_0 + \beta_1 t + \beta_2 e_{u_i\, v_i} + \sum_{k=1}^{K} \gamma_k E_{k\, u_i\, v_i\, t} + \left(\sum_{k_S=1}^{K_S} \alpha_{k_S} F_{k_S\, u_i\, v_i} + \xi_{u_i\, v_i} \right) \quad [7]$$

where $e_{u_i\, v_i}$ denotes the elevation at location (u_i, v_i), E_k denotes a space-time eigenvector, F_k denotes a spatial eigenvector, and the sum in the parentheses is a random effects term [$\sum_{k_S=1}^{K_S} \alpha_{k_S} F_{k_S\, u_i\, v_i}$ denotes spatially structured (SSRE; Figure 11a), and $\xi_{u_i\, v_i}$ denotes spatially unstructured (SURE; Figure 11b), random effects]. The candidate set for the ESTF contains 99 (of 876) eigenvectors; the candidate set for the SSRE contains 18 (of 73) eigenvectors. While the SURE contains only trace levels of spatial autocorrelation, the spatial autocorrelation indices for the SSRE are: MC = 0.865, GR = 0.253. The ESTF contains 15 eigenvectors. This set of covariates accounts for roughly 86% of the time-invariant variation in the percentage of location-specific urban population across the century[6] (elevation alone accounts for about 4%). The specification still results in considerable extra-binomial variation. But the random effects term reduces this overdispersion by roughly 63%, and accounts for an additional roughly 26% in the variation of percentage urban population. The random effects term has a mean of 0.0010, a variance of 1.0105, and conforms closely to a normal curve [P(S-W) = 0.2717]. The SSRE component highlights the San Juan metropolitan region (Figure 11a).

Figure 11

Components of the random effects term. Left (a): SSRE. Right (b): SURE. The magnitude of the numbers is directly proportional to the darkness of the grayscale

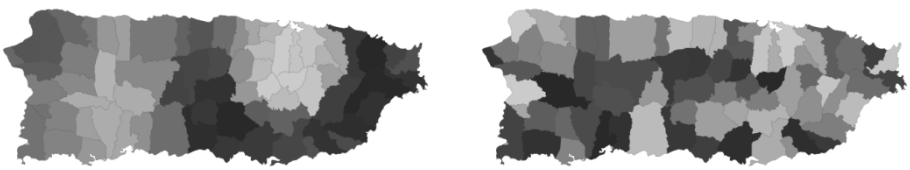


Table 3 summarizes the spatial filter eigenvectors appearing in the Kronecker products of the selected space-time eigenvectors. Autocorrelation latent in these space-time data becomes increasingly complex with the passing of time, which is indicated by an increase in the number of spatial eigenvectors describing the ESTF: the island evolves

[6] A large number of 0 percentages during the first half of the century is one reason why this percentage is not larger.

from essentially a rural to an urban society. In addition, this autocorrelation increasingly is dominated by spatial structure, which is indicated by an increase in the percentage of ESTF variance accounted for with the passing of time. Figures 12a-12e portray the five eigenvectors common to all points in time in the ESTF. Eigenvector E_3 visualizes a contrast between the west coast (dominated by the Mayaguez metropolitan region) and the San Juan urban area. Eigenvector E_4 visualizes a north-south contrast (relating to elevation variation portrayed in Figure 10b).

4.3 Space-time urbanization in Puerto Rico, 1899-2000: the contemporaneous specification

The lagged structure portrayed in Figure 8 casts the value at some location (u_i, v_i, t) as a function of those values at (u_i, v_i, t-1) and its neighbors (u_j, v_j, t), where $a_{ii,t,t-1}$= 1 and $a_{ij,t,t}$ = 1in expression (5). The equation describing this formulation is the same form as for equation (7); but the space-time eigenvectors are different. With this specification, the candidate set increases to 204, of which 21 are used to construct the ESTF. Figure 13a portrays the SSRE, and Figure 13b portrays the SURE, for this specification. While the SURE contains only trace levels of spatial autocorrelation (i.e., the ESTF captures virtually all of the spatial structure latent in the space-time data series), the spatial autocorrelation indices for the SSRE are: MC = 0.689, and GR = 0.349. This set of covariates accounts for roughly 88% of the variation in the time invariant percentage of location-specific urban population across the century (again, elevation alone accounts for about 4%). This specification still results in considerable extra-binomial variation. But the random effects term reduces this overdispersion by roughly 64%, and accounts for an additional roughly 28% in the variation of percentage urban population. The random effects term has a mean of 0.0067, a variance of 1.0166, and conforms closely to a normal curve [P(S-W) = 0.1287]. The SSRE component highlights a north-south contrast (again, relating to the elevation variation portrayed in Figure 10b).

Table 3

Results for the lagged specification: the ESTF as a function of the geographic eigenvectors

(Continue)

Vector	MC	1899	1910	1920	1930	1935	1940	1950	1960	1970	1980	1990	2000
1	1.09			X	X	X	X	X		X	X	X	X
2	1.05				X	X	X	X	X			X	X
3	0.91	***X***	***X***	***X***	***X***	***X***	***X***	***X***	***X***	***X***	***X***	***X***	***X***
4	0.85	***X***	***X***	***X***	***X***	***X***	***X***	***X***	***X***	***X***	***X***	***X***	***X***
5	0.83				X	X	X	X	X	X	X		X
6	0.81			X	X	X	X	X	X	X	X	X	X
7	0.72												
8	0.70						X	X					
9	0.61					X	X	X					X
10	0.57	***X***	***X***	***X***	***X***	***X***	***X***	***X***	***X***	***X***	***X***	***X***	***X***
11	0.54					X	X	X	X				X
12	0.50	***X***	***X***	***X***	***X***	***X***	***X***	***X***	***X***	***X***	***X***	***X***	***X***
13	0.47					X	X	X			X		X

Table 3

Results for the lagged specification: the ESTF as a function of the geographic eigenvectors

(Conclusion)

Vector	MC	1899	1910	1920	1930	1935	1940	1950	1960	1970	1980	1990	2000
14	0.43												X
15	0.40												
16	0.37	X		X	X	X		X		X	X	X	X
17	0.33	***X***	***X***	***X***	***X***	***X***	***X***	***X***	***X***	***X***	***X***	***X***	***X***
18	0.28							X		X	X	X	X
R^2		0.647	0.666	0.792	0.887	0.991	0.992	0.975	0.802	0.878	0.932	0.854	0.998

Note: X denotes selected ESF eigenvector. Bold italic X (***X***) denotes ESF eigenvectors common to all years.

Table 4 summarizes the spatial filter eigenvectors appearing in the Kronecker products of the selected space-time eigenvectors. Autocorrelation latent in these space-time data maintains about the same degree of complexity with the passing of time as for the lagged specification. Now only three eigenvectors are common to all points in time, revealing that the autocorrelation complexity is less well structured. Figure 12f portrays the additional eigenvectors common to all points in time in this ESTF.

Figure 12

Eigenvectors from expression (2) spanning all time periods in the ESTF. Top left (a): E_3. Top middle (b): E_4. Top right (c): E_{10}. Bottom left (d): E_{12}. Bottom middle (e): E_{17}. Bottom right (f): E_6. The magnitude of the numbers is directly proportional to the darkness of the grayscale

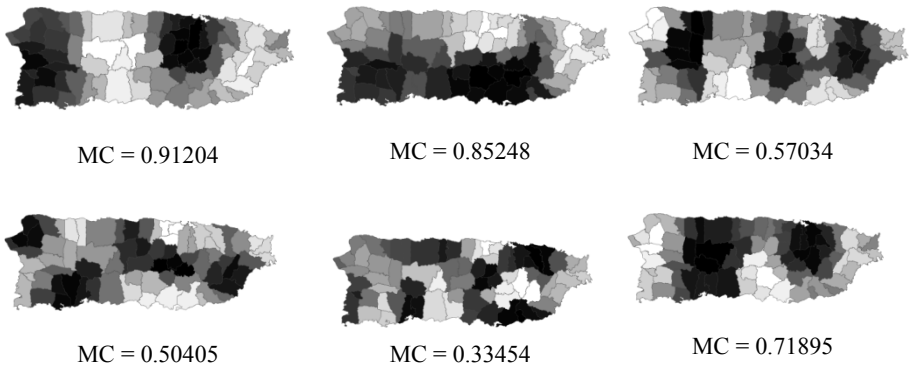


Table 4

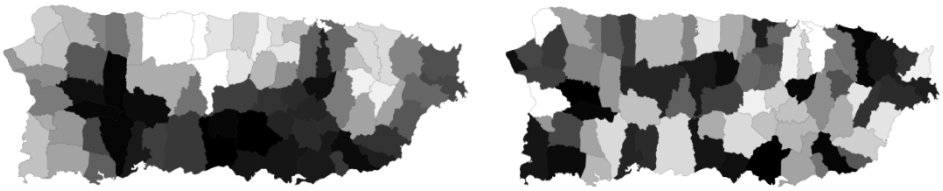
Results for the lagged specification: the ESTF as a function of the geographic eigenvectors

Vector	MC	1899	1910	1920	1930	1935	1940	1950	1960	1970	1980	1990	2000
1	1.09	X	X	X	X	X		X	X	X	X	X	X
2	1.05		X	X		X	X		X		X	X	
3	0.91	***X***	***X***	***X***	***X***	***X***	***X***	***X***	***X***	***X***	***X***	***X***	***X***
4	0.85	***X***	***X***	***X***	***X***	***X***	***X***	***X***	***X***	***X***	***X***	***X***	***X***
5	0.83	X	X	X		X	X	X	X		X	X	X
6	0.81	***X***	***X***	***X***	***X***	***X***	***X***	***X***	***X***	***X***	***X***	***X***	***X***
7	0.72												
8	0.70				X	X	X	X	X	X	X	X	
9	0.61												
10	0.57	X	X	X	X	X		X	X	X	X	X	X
11	0.54	X			X	X			X	X	X	X	
12	0.50	X	X	X	X		X	X	X	X	X	X	X
13	0.47	X			X	X	X	X		X	X	X	X
14	0.43												
15	0.40												
16	0.37					X	X	X	X	X	X	X	
17	0.33	X	X	X	X		X	X	X	X	X	X	X
18	0.28								X				
R^2		0.803	0.888	0.914	0.935	0.971	0.970	0.966	0.939	0.953	0.974	0.975	0.954

Note: X denotes selected ESF eigenvector. Bold italic X (***X***) denotes ESF eigenvectors common to all years.

Figure 13

Components of the random effects term. Left (a): SSRE. Right (b): SURE. The magnitude of the numbers is directly proportional to the darkness of the grayscale



4.4 A summary of comparative findings for two eigenvector space-time filter conceptualizations

Both ESTF conceptualizations result in roughly the same amount of variance being accounted for in the binomial random variable; the contemporaneous conceptualization is better by only 2%. Table 5 summarizes articulation differences between them for the ESF eigenvectors. The ESTF for the lagged conceptualization contains more common geographic structure, whereas that for the contemporaneous conceptualization contains more geographic structure in general. In a mixed model specification, both reduce excess binomial variation by almost exactly the same amount; unfortunately, considerable overdispersion still remains. The SSRE for the lagged conceptualization captures more common spatial autocorrelation across time. In addition, the lagged conceptualization has information concentrated into fewer eigenvectors (15 of 99 versus 21 of 204).

Table 5

Differences between Tables 3 and 4

Vector	MC	1899	1910	1920	1930	1935	1940	1950	1960	1970	1980	1990	2000
1	1.09	T4	T4				**T3**		T4				
2	1.05		T4	T4	**T3**			**T3**			T4		**T3**
3	0.91												
4	0.85												
5	0.83	T4	T4	T4	**T3**					**T3**		T4	
6	0.81	T4	T4										
8	0.70				T4	T4			T4	T4	T4	T4	
9	0.61					**T3**	**T3**	**T3**					**T3**
10	0.57						**T3**						
11	0.54	T4			T4		**T3**	**T3**		T4	T4	T4	**T3**
12	0.50					**T3**				T4		T4	
13	0.47	T4			T4								**T3**
14	0.43												
16	0.37	**T3**		**T3**	**T3**		T4		T4				**T3**
17	0.33					**T3**							
18	0.28							**T3**	T4	**T3**	**T3**	**T3**	**T3**
$\|R^2$ difference$\|$		0.156	0.222	0.122	0.048	0.02	0.022	0.009	0.137	0.075	0.042	0.121	0.044

Note: **T3** denotes that the eigenvector appears in Table 3 only; T4 denotes that it appears in Table 4 only

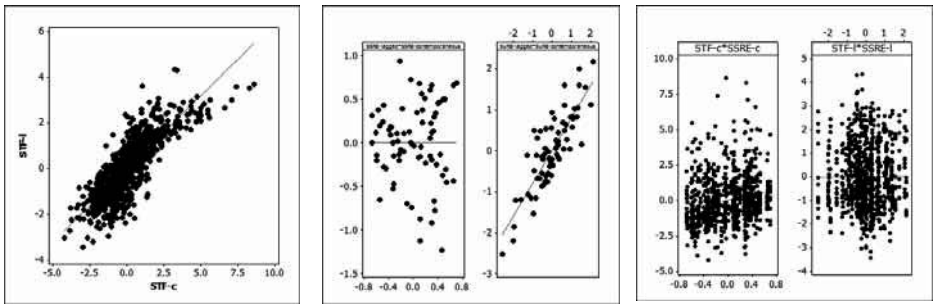
Figure 14a furnishes a graphical comparison of the two ESTFs. Although very similar (they have roughly 68% common variance), they are not the same; the scatterplot (Figure 14a) suggests that the relationship between them is slightly nonlinear. Figure 14b furnishes a graphical comparison of the two random effects terms. The SSRE are completely different (they have no common variance). This finding may seem surprising, because these two variates share two eigenvectors. But these two eigenvectors respectively account for only 5.3% of the variance in the contemporaneous, and 2.7% of the variance in the lagged, random effects term. The respective ESFs account for 13.5% and 28.5% of the variance in, respectively, the contemporaneous and lagged random effects terms. Presumably these components are compensating for relatively inconsequential spatial structure deficiencies in their respective ESTFs; as expected, they have little relationship to their respective ESTFs (Figure 14c). And, the SURE are very similar (they have roughly 71% common variance).

In summary, eigenvector space-time filtering successfully accounts for spatial and temporal autocorrelation in space-time data. It highlights that spatial autocorrelation changes through time, and that a random effects term captures heterogeneity with little spatial structure (i.e., the ESTF effectively captures virtually all of the spatial structure). The contemporaneous conceptualization appears to outperform the lagged

conceptualization on selected criteria, but its filter is less parsimonious. Meanwhile, with regard to percentage of urban population in space and time, large amounts of overdispersion are attributable to: space-time dependencies, random effects (i.e., heterogeneity), and some unknown source (i.e., overdispersion still remains).

Figure 14

A comparison of the lag and the contemporaneous ESTFs. Left (a): the total filters. Middle (b): SSRE and SURE components of the filters. Right (c): the total filters versus the time invariant SSREs. STF-l denotes the lagged ESTF; STF-c denotes the contemporaneous ESTF



5. Implications and conclusions

Eigenvector filtering offers a flexible and powerful tool for dealing with correlated data. It is competitive with ARIMA modeling for time series, and geostatistical and autoregressive modeling for spatial series. Once a matrix is posited that articulates the connectivity structure of linkages tying observations together so that they can become correlated, eigenvectors extracted from a modified version of this matrix (i.e., pre- and post-multiplied by a projection matrix) offer control variables that can be introduced into a model specification to identify and isolate the stochastic dependencies among observations, allowing model building to proceed as if the observations are independent.

This approach offers advantages over more traditional approaches. Foremost, it allows a much wider range of non-normal random variables to be employed in analysis. Second, it allows spatial and temporal components to be better isolated and described. Third, it supports mixed modeling, furnishing eigenvectors as fixed effects. Fourth, it functions well in both interpolation and extrapolation situations. Fifth, it relates to conventional multivariate analysis vis-à-vis principal components analysis. And, sixth, it enables model specifications that are infeasible within an autoregressive formulation (e.g., the auto-Poisson specification).

The single most conspicuous weakness of filtering is that the eigenvectors need to be selected with a stepwise regression procedure. Drawbacks of stepwise multiple regression include: parameter estimation bias, inconsistencies among model selection algorithms, a need to adjust for multiple hypothesis testing, and an obsession with identifying a single best model (e.g., see Derksen and Keselman, 1992). The first of

these is minimized with eigenvector filtering because the eigenvectors involved are mutually orthogonal and uncorrelated. The last also is less relevant, because frequently those eigenvectors selected last account for very little of the variance, and may or may not need to be included, depending upon the behavior of other model properties. Table 1 suggests that the second and third concerns may not be very relevant to eigenvector filtering, although these topics merit serious future research attention. Of note is that these concerns cannot be minimized by increasing sample size. Moreover, as sample size increases, filters tend to include larger numbers of eigenvectors, in part because the standard errors asymptotically go to 0.

REFERENCES

BASILEVSKY, A. (1983). «Applied Matrix Algebra in the Statistical Sciences.» *North-Holland, New York*.

CLIFF, A. AND ORD, J. (1981). «Spatial and temporal analysis: autocorrelation in space and time», in N. WRIGLEY and R. BENNETT (eds.), *Quantitative Geography: A British View*. Routledge & Kegan Paul, London, 104–110.

COCHRANE, D. AND ORCUTT, G.H. (1949). «Application of least squares regression to relationships containing autocorrelated error terms». *Journal of the American Statistical Association*, 44, 32–61.

DERKSEN, S. AND KESELMAN, H. (1992). «Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables». *British Journal of Mathematical & Statistical Psychology*, 45, 265–282.

GRIFFITH, D. (1981). «Interdependence in space and time: numerical and interpretative considerations», in D. GRIFFITH and R. MaCKINNON (eds.), *Dynamic Spatial Models*. Sijthoff and Noordhoff, Alphen aan den Rijn, 258–287.

GRIFFITH, D. (1996). «Spatial statistical analysis and GIS: exploring computational simplifications for estimating the neighborhood spatial forecasting model», in P. LONGLEY and M. BATTY (eds.), *Spatial Analysis: Modelling in a GIS Environment*, Longman GeoInformation, 255–268.

GRIFFITH, D. (2000). «A linear regression solution to the spatial autocorrelation problem». *J. of Geographical Systems* 2, 141–156.

GRIFFITH, D. (2002). «A spatial filtering specification for the auto-Poisson model». *Statistics & Probability Letters*, 58, 245–251.

GRIFFITH, D. 2003. «Spatial Autocorrelation and Spatial Filtering: Gaining Understanding through Theory and Scientific Visualization». *Springer-Verlag, Berlin.*

GRIFFITH, D. (2004). «A spatial filtering specification for the auto-logistic model», *Environment & Planning A*, 36, 1791–1811.

GRIFFITH, D. (2008A). «A comparison of four model specifications for describing small heterogeneous space-time datasets: sugar cane production in Puerto Rico», 1958/59-1973/74. *Papers in Regional Science*, 87, 341–356.

GRIFFITH, D. (2008B). «Spatial filtering-based contributions to a critique of geographically weighted regression (GWR)». *Environment & Planning A*, 40, 2751–2769.

GRIFFITH, D. 2010. «Modeling spatio-temporal relationships: retrospect and prospect». *J. of Geographical Systems* 12, 111–123.

GRIFFITH, D. AND CHUN, Y. (2009). «Eigenvector selection with stepwise regression techniques to construct spatial filters». Paper presented at the annual Association of American Geographers meeting, Las Vegas, NV, March 25.

GRIFFITH, D. AND HEUVELINK, G. (2010). «Deriving space-time variograms from space-time autoregressive (STAR) model specifications». *Keynote presentation, IGU, Honk Kong.*

HEUVELINK, G. AND GRIFFITH, D. (2010). «Space-time geostatistics for geography: a case study of radiation monitoring across parts of Germany». *Geographical Analysis* 42, 161–179.

ORD, J. (1975). «Estimating methods for models of spatial interaction». *J. of the American Statistical Association*, 70, 120–26.

TIEFELSDORF, M. AND GRIFFITH, D. (2007). Semi-parametric filtering of spatial autocorrelation: the eigenvector approach. *Environment & Planning A*, 39, 1193–1221.