Paper SAS535-2017

Big Value from Big Data: SAS/ETS® Methods for Spatial Econometric Modeling in the Era of Big Data

Guohui Wu and Jan Chvosta, SAS Institute Inc.

ABSTRACT

Data that are gathered in modern data collection processes are often large and contain geographic information that enables you to examine how spatial proximity affects the outcome of interest. For example, in real estate economics, the price of a housing unit is likely to depend on the prices of housing units in the same neighborhood or nearby neighborhoods, either because of their locations or because of some unobserved characteristics that these neighborhoods share. Understanding spatial relationships and being able to represent them in a compact form are vital to extracting value from big data. This paper describes how to glean analytical insights from big data and discover their big value by using spatial econometric methods in SAS/ETS® software.

INTRODUCTION

Spatial data have become increasingly popular in the past few decades, covering a wide array of areas from marketing research and risk analysis to health care and finance, just to name a few. Thanks to technological advances, spatial data can be collected faster than ever before and at a relatively low cost through the use of modern sensors and smart devices. These data contain geographic information such as spatial coordinates, enabling you to plot the data on a map for better visualization and thus glean more insights from the data. However, the geographic nature of spatial data can complicate their modeling and analysis, making these data more challenging to work with than their nonspatial counterpart.

Among the many disciplines that study spatial data is spatial econometrics, a subfield of econometrics that concentrates on the econometric modeling and analysis of spatial data. At the core of spatial econometric modeling is dealing with spatial interaction and spatial heterogeneity of spatial data in the regression setting (Anselin 1988). The key idea behind spatial econometric modeling resonates with the first law of geography: “Everything is related to everything else, but near things are more related than distant things" (Tobler 1970). According to this principle, there can be spatial dependence in the data, and the strength of this dependence is determined by the neighborhood structure of regions in space.

One vital component of spatial econometric modeling is appropriately accounting for spatial dependence in the data. In general, spatial dependence can arise from three sources: endogenous interaction, exogenous interaction, and correlated error terms (Elhorst 2013). Endogenous interaction refers to the case where the value of the dependent variable in one region is correlated with those in other regions. In comparison, exogenous interaction means that the value of the dependent variable in one region depends on the values of explanatory variables in other regions. There could also be spatial dependence in the error terms when the error in one region is affected by errors in other regions; this might be caused by some form of unobserved heterogeneity.

Another important part of spatial econometric modeling is the choice of a spatial weights matrix. In a broad sense, spatial weights matrices are used to describe the spatial neighborhood structure among all geographic regions and how much influence one region exerts over a nother. Because of the use of a spatial weights matrix and the computational cost, there are some practical challenges for spatial econometric modeling of big data.

In the era of big data, data are characterized by velocity, variety, volume, and value. At high velocity, many of the wide variety of data that are generated by modern sensors and smart devices can reach a large volume. However, the widespread accessibility of big data can present challenges for spatial econometric modeling. First, the amount of memory that is required to store the spatial weights matrix can be massive. For example, you need about 31.5 gigabytes of memory to store a full spatial weights matrix for the 64,999 tracts in the 2000 United States Census. Also, computation can become extremely intensive, partly because of the need to calculate the determinant of a large matrix. As a result, having the capability to represent a spatial weights matrix in a compact form in order to ease the computational burden is key to extracting big data’s value by using spatial econometric modeling.

This paper discusses several spatial models and the challenges that commonly arise when you fit a spatial model in various scenarios. In particular, the paper focuses on problems that large spatial data sets can pose and memory requirements that typical computational resources cannot accommodate. To overcome these problems, the paper introduces compact representation of the spatial weights matrix in order to reduce the memory requirement and computational algorithms based on approximations in order to provide a stable scalable solution.

The paper is organized as follows. First, an overview of a wide range of commonly used spatial econometric models and features available in the SPATIALREG procedure is presented. These models are fitted to a data set that contains data on population growth for 3,098 US counties. To demonstrate the power of the SPATIALREG procedure and its scalability to large problems, a simulated data set for the US Census tracts that contains 64,999 observations is considered. If the full representation of the matrix of spatial neighbors were used, the data size in memory would be over 30 GB, and solving a problem of this magnitude would be computationally very demanding. However, using the compact representation of the spatial weights matrix and approximation algorithms that are introduced in PROC SPATIALREG makes the problem computationally feasible. The paper ends with a summary and concluding comments.

AN OVERVIEW OF THE SPATIALREG PROCEDURE

The SPATIALREG (spatial regression) procedure analyzes spatial econometric models for cross-sectional data whose observations are spatially referenced or georeferenced. For example, data that are collected from counties and census tracts from the 48 contiguous US states fall into the category of spatially referenced data. Unlike nonspatial regression models, spatial econometric models are capable of handling spatial interaction and spatial heterogeneity in a regression setting.

The SPATIALREG procedure supports the following models:

- spatial autoregressive (SAR) model

$$\mathbf{y} = \rho\mathbf{W}_1\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- spatial Durbin model (SDM)

$$\mathbf{y} = \rho\mathbf{W}_1\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{W}_1\mathbf{Z}\boldsymbol{\theta} + \boldsymbol{\epsilon}$$

- spatial error model (SEM)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad \mathbf{u} = \lambda\mathbf{W}_2\mathbf{u} + \boldsymbol{\epsilon}$$

- spatial Durbin error model (SDEM)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}_2\mathbf{Z}\boldsymbol{\theta} + \mathbf{u}, \quad \mathbf{u} = \lambda\mathbf{W}_2\mathbf{u} + \boldsymbol{\epsilon}$$

- spatial moving average (SMA) model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad \mathbf{u} = \boldsymbol{\epsilon} - \lambda\mathbf{W}_2\boldsymbol{\epsilon}$$

- spatial Durbin moving average (SDMA) model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}_2\mathbf{Z}\boldsymbol{\theta} + \mathbf{u}, \quad \mathbf{u} = \boldsymbol{\epsilon} - \lambda\mathbf{W}_2\boldsymbol{\epsilon}$$

- spatial autoregressive moving average (SARMA) model

$$\mathbf{y} = \rho\mathbf{W}_1\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad \mathbf{u} = \boldsymbol{\epsilon} - \lambda\mathbf{W}_2\boldsymbol{\epsilon}$$

- spatial Durbin autoregressive moving average (SDARMA) model

$$\mathbf{y} = \rho\mathbf{W}_1\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{W}_1\mathbf{Z}\boldsymbol{\theta} + \mathbf{u}, \quad \mathbf{u} = \boldsymbol{\epsilon} - \lambda\mathbf{W}_2\boldsymbol{\epsilon}$$

- spatial autoregressive confused (SAC) model

$$\mathbf{y} = \rho\mathbf{W}_1\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad \mathbf{u} = \lambda\mathbf{W}_2\mathbf{u} + \boldsymbol{\epsilon}$$

- spatial Durbin autoregressive confused (SDAC) model

$$\mathbf{y} = \rho\mathbf{W}_1\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{W}_1\mathbf{Z}\boldsymbol{\theta} + \mathbf{u}, \quad \mathbf{u} = \lambda\mathbf{W}_2\mathbf{u} + \boldsymbol{\epsilon}$$

- linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- spatial lag of X (SLX) model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}_1\mathbf{Z}\boldsymbol{\theta} + \boldsymbol{\epsilon}$$

In the preceding equations, $\mathbf{y} = (y_1, y_2, \ldots, y_n)'$ and y_i is the value of the continuous dependent variable that corresponds to region i for $i = 1, 2, \ldots, n$, where n is the number of observations in the data. $\mathbf{X}$ and $\mathbf{Z}$ are $n \times p$ and $n \times q$ matrices of regressors, respectively, and $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)'$, where $\epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_\epsilon^2)$ for $i = 1, 2, \ldots, n$. In addition, $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are a $p \times 1$ parameter vector and a $q \times 1$ parameter vector, respectively. For models that require two spatial weights matrices, $\mathbf{W}_1$ and $\mathbf{W}_2$, these matrices are assumed to be identical in SAS/ETS 14.2.

Of particular importance to spatial econometric modeling is how to choose a model that can describe your data. Although the choice of a model is often problem-specific, there are general guidelines that you can follow. Among the preceding models, you use SAR models to account for endogenous interaction effects. SDM models can be used to address both endogenous and exogenous interaction effects. You use SEM and SMA models to account for spatial dependence in the error terms. To account for both exogenous interaction effects and spatial dependence in the error terms, you can use SDMA and SDEM models. SARMA and SAC models can be used to address both endogenous interaction effects and spatial dependence in the error terms. In the case where endogenous interaction effects, exogenous interaction effects, and spatial dependence in the error terms coexist, you can use SDARMA and SDAC models. You use SLX models to account for exogenous interaction effects and linear models for purely linear regression.

After you choose a model, you can use PROC SPATIALREG to fit the model. You can issue a simple call to PROC SPATIALREG by using the following statements:

```
proc spatialreg data=dat wmat=W NONORMALIZE
  APPROXIMATION=(Taylor NMC=100 ORDER=10 SEED=2017);
  model y=x1 x2/type=SAR;
  spatialeffects z1 z2;
  spatialid SID;
run;
```

You supply two data sets—the primary data set and a spatial weights matrix—by using the DATA= option and the WMAT= option, respectively. The primary data set contains the dependent variable, the independent variables, and possibly the spatial ID variable. The spatial weights matrix, $\mathbf{W}$, can take two different forms. First, you can provide a full spatial weights matrix. In this case, the data set that you specify in the WMAT= option has n rows. The number of columns can be n+1 if you need matching observations in the primary data set and the spatial weights matrix. In such a case, you use the SPATIALID statement to specify a spatial ID variable for the purpose of matching observations. If no matching is needed, the number of columns should be n.

When the spatial weights matrix is sparse, an alternative is to use a compact form, as described in the section "COMPACT REPRESENTATION OF A SPATIAL WEIGHTS MATRIX" on page 5. With this compact form, you must use the SPATIALID statement to specify a spatial ID variable that enables you to match observations. By default, the spatial weights matrix that you specify in the WMAT= option is row-standardized unless you specify the NONORMALIZE option. The NONORMALIZE option requests that the spatial weights matrix be used "as is" rather than be row-standardized.

The APPROXIMATION= option enables you to approximate the Jacobian term in SAR and SDM models instead of computing the term exactly. This is advantageous especially for big data because exact computation of the Jacobian term can be infeasible. To control the approximation, you use the keyword TAYLOR to request that Taylor approximation be used. By default, Chebyshev approximation is used. The NMC= option specifies the number of standard random normal draws to be used for Monte Carlo simulation. The ORDER= option enables you to specify

the order of series in Taylor approximation and Chebyshev approximation. You use the SEED= option to specify a seed for the random number generator for Monte Carlo simulation, enabling you to reproduce your analysis.

In the MODEL statement, you specify the dependent variable y and regressors x. You use the TYPE= option to specify the type of model to be fit, selecting one of the following values: SAR, SEM, SMA, SARMA, SAC, and LINEAR. For example, you specify TYPE=SAR to fit a SAR model and TYPE=LINEAR to fit a linear regression model. You use the SPATIALEFFECTS statement to specify exogenous interaction effects. For example, you specify TYPE=SAR together with the SPATIALEFFECTS statement to fit an SDM model.

SPECIFYING THE SPATIAL WEIGHTS MATRIX

The spatial weights matrix $\mathbf{W}$ plays a vital role in spatial econometric modeling. Among the many different ways to create the $\mathbf{W}$ matrix, two common ones are k-order binary contiguity matrices and k-nearest-neighbor matrices (Elhorst 2013).

You often start with a spatial contiguity matrix $\mathbf{C}$. For a k-order binary contiguity matrix, a value of 1 for the (i, j)th entry in $\mathbf{C}$ indicates that the two regions i and j are k-order neighbors to each other, and 0 indicates otherwise. The neighbor relationship is often defined based on sharing a common boundary. For example, a first-order binary contiguity matrix might look like the following:

$$\mathbf{C} = \begin{pmatrix} \text{SID} & \text{L1} & \text{L2} & \text{L3} & \text{L4} \\ \text{L1} & 0 & 1 & 0 & 1 \\ \text{L2} & 1 & 0 & 0 & 0 \\ \text{L3} & 0 & 0 & 0 & 1 \\ \text{L4} & 1 & 0 & 1 & 0 \end{pmatrix}$$

The diagonal elements of $\mathbf{C}$ are zeros because, in general, a region is not considered to be a neighbor of itself. The two regions L2 and L4 are neighbors of L1; L2 has L1 as its only neighbor; L3 has L4 as its only neighbor; and L4 has L1 and L3 as its neighbors. The spatial weights matrix $\mathbf{W}$ is often row-standardized. To standardize rows, you divide entries in each row of $\mathbf{C}$ by the sum of that row. The spatial weights matrix $\mathbf{W}$, which is the row-standardized version of $\mathbf{C}$, is as follows:

$$\mathbf{W} = \begin{pmatrix} \text{SID} & \text{L1} & \text{L2} & \text{L3} & \text{L4} \\ \text{L1} & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \text{L2} & 1 & 0 & 0 & 0 \\ \text{L3} & 0 & 0 & 0 & 1 \\ \text{L4} & \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{pmatrix}$$

For k-nearest-neighbor matrices, you create a spatial contiguity matrix based on a distance metric. Let d_{ij} denote the distance between the two regions i and j, which might be the Euclidean distance between the centroids (geographic centers) of the two regions. For example, let $(\text{lon}_i, \text{lat}_i)$ and $(\text{lon}_j, \text{lat}_j)$ be the centroids of regions i and j, where $1 \le i, j \le n$, and lon and lat denote the longitude and latitude, respectively. Under the Euclidean distance metric, the distance d_{ij} between regions i and j is

$$d_{ij} = \sqrt{(\text{lat}_i - \text{lat}_j)^2 + (\text{lon}_i - \text{lon}_j)^2}$$

After computing the distance between region i and other regions under a certain metric, you sort d_{ij} in ascending order; for example, $d_{ij_1} \le d_{ij_2} \le \cdots \le d_{ij_k} \le \cdots \le d_{ij_{n-1}}$. For a given k, let $N_k(i) = \{j_1, j_2, \ldots, j_k\}$ be the set that contains the indices of k-nearest neighbors of region i; then the (i, j)th entry of the contiguity matrix $\mathbf{C}$ is defined as

$$C_{ij} = \begin{cases} 1 & \text{if } j \in N_k(i) \\ 0 & \text{otherwise} \end{cases}$$

The (i, j)th entry of the corresponding row-standardized matrix $\mathbf{W}$ is $W_{ij} = C_{ij} \left\{ \sum_{j \in N_k(i)} C_{ij} \right\}^{-1}$.

Unlike the k-order binary contiguity matrix, which is often symmetric by construction, k-nearest-neighbor matrices can be asymmetric. To obtain symmetric k-nearest-neighbor matrices, you can define the (i, j)th entry of the spatial contiguity matrix $\mathbf{C}$ as follows:

$$C_{ij} = \begin{cases} 1 & \text{if } j \in N_k(i) \text{ or } i \in N_k(j) \\ 0 & \text{otherwise} \end{cases}$$

In addition to the Euclidean distance metric, you can use other distance metrics as appropriate. A variant of k-nearest-neighbor matrix $\mathbf{C}^*$ that is used in some empirical studies defines its (i, j)th entry as

$$C_{ij}^* = \begin{cases} 1 & \text{if } d_{ij} \leq d_{\text{cutoff}} \\ 0 & \text{otherwise} \end{cases}$$

where d_{cutoff} is a prespecified threshold distance.

COMPACT REPRESENTATION OF A SPATIAL WEIGHTS MATRIX

When the number of observations n increases, the amount of memory that it takes to store a full spatial contiguity matrix $\mathbf{C}$ or spatial weights matrix $\mathbf{W}$ increases dramatically. To circumvent the storage issue, PROC SPATIALREG enables you to provide a compact representation of $\mathbf{W}$ (or $\mathbf{C}$) when appropriate (for example, when the spatial contiguity or weights matrix is sparse). For the compact matrix representation, you provide a data set that contains three variables by using the WMAT= option. The first two variables identify the row r and column c of $\mathbf{W}$ (or $\mathbf{C}$), and (r, c) can be expressed either as numerical indices or as values of the variable specified in the SPATIALID statement. The third variable contains the nonzero value of $\mathbf{W}$ (or $\mathbf{C}$) for row r and column c. For the compact representation, the number of observations in the data set that you specify in the WMAT= option equals the total number of nonzero entries in $\mathbf{W}$ (or $\mathbf{C}$). You must use a SPATIALID statement when you use the compact representation of a spatial weights matrix.

For the spatial weights matrix $\mathbf{W}$ shown previously, its compact representation might be as follows:

```
data Ws;
  input SID $2 cSID $2 Weight;
  datalines;
  L1 L2 0.5
  L1 L4 0.5
  L2 L1 1.0
  L3 L4 1.0
  L4 L1 0.5
  L4 L3 0.5
  ;
run;
```

EXAMPLES OF ECONOMETRIC MODELING WITH THE SPATIALREG PROCEDURE

The two examples in this section demonstrate how to use PROC SPATIALREG for spatial econometric modeling. In the first example, the data set to be used is **US48_CountyG**, which contains the data for population growth between 1980 and 1990 in counties of the 48 contiguous US states. In the second example, a simulated data set, **CensusTr_SimData**, is used to show how PROC SPATIALREG can handle big data. For illustration purposes, the computation times presented in these examples are based on a Windows desktop with Intel Core i7—4770 3.40 GHz CPU and 16 GB RAM.

POPULATION GROWTH OF COUNTIES IN THE UNITED STATES

To demonstrate spatial econometric modeling by using PROC SPATIALREG, a real data set, **US48_CountyG**, is considered for the purpose of illustration. This data set is adapted from the data set provided by Wheeler (2003) at `http://qed.econ.queensu.ca/jae/2003-v18.1/wheeler/`. Four variables in the **US48_CountyG** data set are selected: **FIPS**, **PopGR**, **CollRate**, and **EducSh**. The variable **FIPS** represents a five-digit Federal Information Processing Standard (FIPS) code for counties in the United States in 1980. The dependent variable **PopGR** represents the county-level population growth rate from 1980 to 1990. The two independent variables **CollRate** and **EducSh** refer, respectively, to the county-level proportion of the adult population in 1980 with at least a bachelor's degree and the share of local government spending on education in 1982 (Wheeler 2003). The map data set **US48_CountyM** contains the mapping information for the 48 contiguous US states.

Table 1 shows the summary statistics for three of the selected variables in the **US48_CountyG** data set. According to this table, the **US48_CountyG** data set contains 3,098 observations. Each observation corresponds to one of 3,098 counties.

Table 1 Summary Statistics for Selected Variables in **US48_CountyG** Data Set

Variable	N	Mean	Std Dev	Minimum	Maximum
PopGR	3,098	0.029	0.143	–0.385	0.967
CollRate	3,098	0.114	0.054	0.016	0.478
EducSh	3,098	0.516	0.120	0	0.997

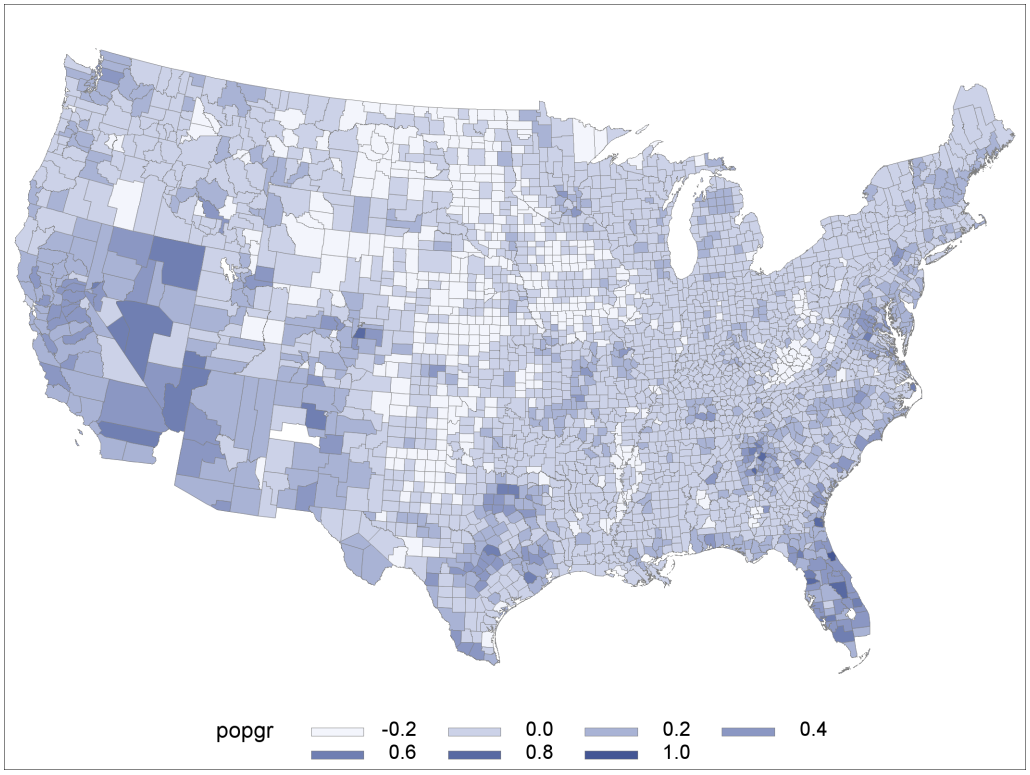
To visualize the data, you can plot **PopGR** on the map by using the following statements:

```
libname SGF2017 'U:\SGF';

ods graphics on;
proc gmap map=SGF2017.us48_countym data=SGF2017.us48_countyg;
  id FIPS;
  choro popgr/coutline=gray midpoints=(-0.2 to 1.0 by 0.2);
run;
```

Figure 1 graphically presents the population growth rate in each county. According to this map, Elko and Nye Counties had the highest population growth rate in Nevada. Among all US counties, Flagler County in Florida had the highest population growth rate (0.967), followed by Douglas County in Colorado (0.8759). You can also see that neighboring counties seem to demonstrate similar population growth, suggesting that there is spatial dependence in the data.

Figure 1 Population Growth Rate for US Counties between 1980 and 1990



Because of spatial dependence in the data, it might be reasonable for you to fit a SAR model. In doing so, you need to create a spatial weights matrix for the 3,098 counties in the data. The data set **US48County_WList** contains a compact representation of a border-contiguity matrix; it is adapted from the data set **USCBContig** retrieved from the "Spatial Econometrics Library" link at http://www.spatial-econometrics.com/. You can fit a SAR model by issuing the following statements:

```
proc spatialreg data=SGF2017.us48_countyg wmat=SGF2017.us48county_wlist;
  model popgr=collrate educsh/type=SAR;
  spatialid FIPS;
run;
```

The estimation of the SAR model takes about 260 seconds. Figure 2 shows parameter estimation results from the model. In this figure, _rho and _sigma2 are the internal names for ρ and σ^2, respectively. The coefficient ρ is estimated to be 0.71, and it is significant at the 0.05 level, indicating that there is a positive spatial dependence in the data.

Figure 2 Parameter Estimates for SAR Model for **US48_CountyG** Data Set

The SPATIALREG Procedure

Model: MODEL 1
Dependent Variable: popgr

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept	1	-0.089119	0.009646	-9.24	<.0001
collrate	1	0.490390	0.034126	14.37	<.0001
educsh	1	0.079508	0.015072	5.28	<.0001
_rho	1	0.710687	0.013343	53.26	<.0001
_sigma2	1	0.009588	0.000250	38.34	<.0001

For comparison, the same SAR model can be fit using Chebyshev approximation, as follows:

```
proc spatialreg data=SGF2017.us48_countyg wmat=SGF2017.us48county_wlist
  approximation=(order=10 NMC=100 seed=1);
  model popgr=collrate educsh/type=SAR;
  spatialid FIPS;
run;
```

According to the specification in the APPROXIMATION= option, a 10th-order Chebyshev polynomial is used. Estimating the SAR model by using Chebyshev approximation takes about 1.2 seconds, which is about 0.5% of the time used for the exact computation. Comparing Figure 2 and Figure 3, you can conclude that the parameter estimates in these two tables are very similar. Because the computational speed of the approximation is so much faster and the parameter estimates for the two methods are so similar, you can see the great potential for an application to even bigger data that would not fit in memory if you used the full representation of the spatial weights matrix and the computational methods in the first part of the example.

Figure 3 Parameter Estimates for SAR Model Using Chebyshev Approximation for **US48_CountyG** Data Set

The SPATIALREG Procedure

Model: MODEL 1
Dependent Variable: popgr

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept	1	-0.089108	0.009645	-9.24	<.0001
collrate	1	0.490175	0.034123	14.37	<.0001
educsh	1	0.079508	0.015070	5.28	<.0001
_rho	1	0.711155	0.013348	53.28	<.0001
_sigma2	1	0.009586	0.000250	38.34	<.0001

You can fit multiple models by using only one call to PROC SPATIALREG. This can be helpful if you want to identify a model that best describes your data by using a criterion such as Akaike's information criterion (AIC). For example, you can fit all 12 models except the linear and SLX models by issuing the following statements:

```
proc spatialreg data=SGF2017.us48_countyg wmat=SGF2017.us48county_wlist;
  model popgr=collrate educsh/type=SAR;
  model popgr=collrate educsh/type=SAR;
  spatialeffects collrate educsh;
  model popgr=collrate educsh/type=SEM;
  model popgr=collrate educsh/type=SEM;
  spatialeffects collrate educsh;
  model popgr=collrate educsh/type=SMA;
  model popgr=collrate educsh/type=SMA;
  spatialeffects collrate educsh;
  model popgr=collrate educsh/type=SAC;
  model popgr=collrate educsh/type=SAC;
  spatialeffects collrate educsh;
  model popgr=collrate educsh/type=SARMA;
  model popgr=collrate educsh/type=SARMA;
  spatialeffects collrate educsh;
  spatialid FIPS;
run;
```

Table 2 shows the AIC values for the 10 models being considered. Among these models, the SDARMA model is identified as the best model because its AIC value is lowest. The parameter estimation results for this model are shown in Figure 4.

Table 2 Model Selection Using AIC for **US48_CountyG** Data Set

Model	AIC
SAR	–5214
SDM	–5247
SEM	–5246
SDEM	–5214
SDM	–5245
SMA	–4806
SDMA	–4813
SAC	–5439
SDAC	–5464
SARMA	–5327
SDARMA	–5480

```
proc spatialreg data=SGF2017.us48_countyg wmat=SGF2017.us48county_wlist;
  model popgr=collrate educsh/type=SARMA;
  spatialeffects collrate educsh;
  spatialid FIPS;
run;
```

Figure 4 Parameter Estimates for the SDARMA Model for **US48_CountyG** Data Set

The SPATIALREG Procedure

Model: MODEL 1
Dependent Variable: popgr

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept	1	-0.002240	0.004965	-0.45	0.6518
collrate	1	0.625991	0.040105	15.61	<.0001
educsh	1	0.106733	0.016310	6.54	<.0001
W_collrate	1	-0.587719	0.045488	-12.92	<.0001
W_educsh	1	-0.109303	0.018216	-6.00	<.0001
_rho	1	0.971452	0.006254	155.34	<.0001
_lambda	1	0.714975	0.026332	27.15	<.0001
_sigma2	1	0.008119	0.000214	37.90	<.0001

SIMULATED DATA FOR US CENSUS TRACTS

To illustrate scalability of the SPATIALREG procedure to large data sets, this example uses the simulated data set **CensusTr_SimData**, which contains 64,999 observations based on the 2000 US Census tracts. The shapefile that contains the geographic information needed to create the matrix of spatial neighbors for these census tracts is obtained from `ftp://ftp2.census.gov/geo/tiger/TIGER2010/TRACT/2000/`. The **CensusTr_SimData** data set contains nine variables, consisting of the dependent variable **y**; seven independent variables, **x1**–**x7**; and **CTIDFP00**, which contains the census tract identification code. This example is intended to demonstrate the scalability of the SPATIALREG procedure in both the number of observations and the number of regressors. Compared to the earlier example, which uses the county-level data, the number of observations increased approximately 20 times and the number of regressors more than tripled. If the full matrix of spatial weights that is needed for this problem were stored in memory, it would require over 30 GB.

The data generating process for the simulation is represented as

$$\mathbf{y} = 0.3\mathbf{W}\mathbf{y} + 1.8 - 0.9\mathbf{x}_1 + 1.2\mathbf{x}_2 + 0.7\mathbf{x}_3 + 0.6\mathbf{x}_4 - 1.0\mathbf{x}_5 + 0.8\mathbf{x}_6 - 0.6\mathbf{x}_7 + \boldsymbol{\epsilon}$$

where $\mathbf{x}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,64999})'$ and $x_{i,j} \sim \mathrm{N}(0,1)$ for $i = 1, 2, \ldots, 7$ and $j = 1, 2, \ldots, n = 64{,}999$. Moreover, $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)'$, with $\epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, 0.5)$ for $i = 1, 2, \ldots, n$.

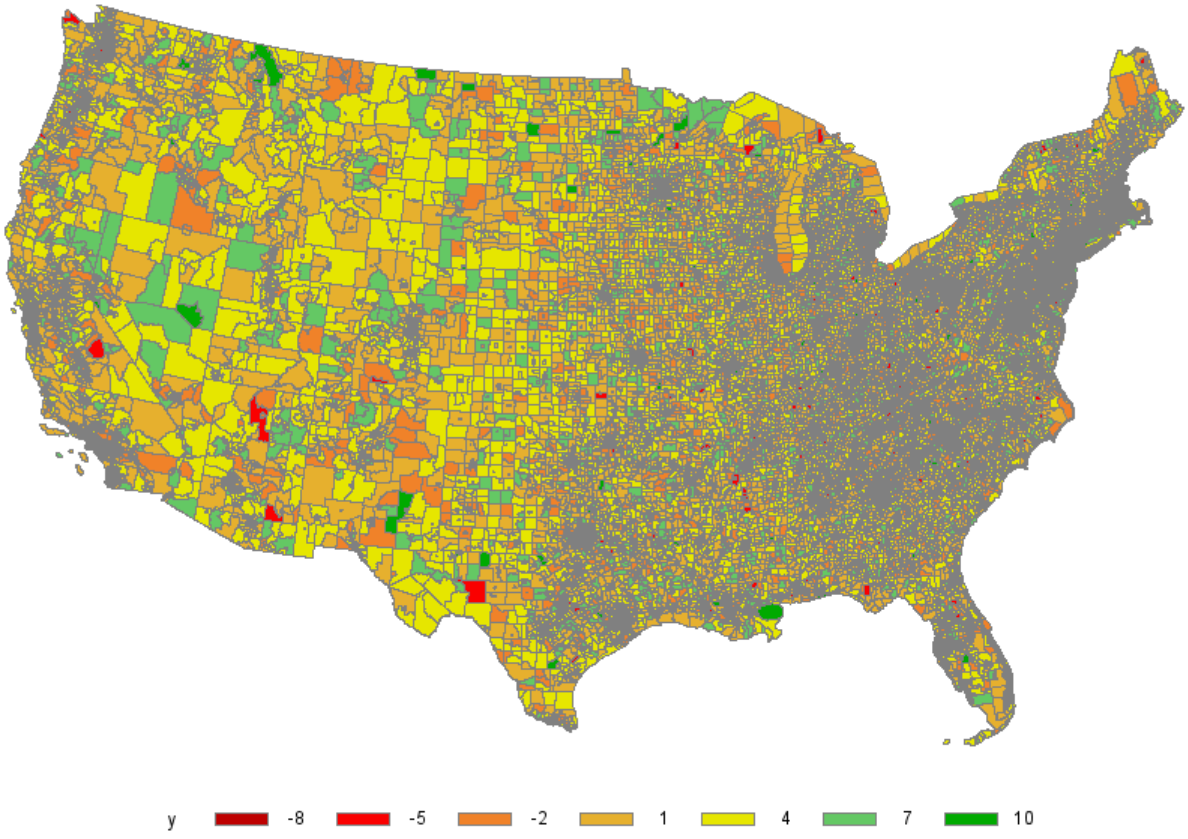
The spatial weights matrix $\mathbf{W}$ is created based on three nearest neighbors. In particular, the Euclidean distance between the centroids (that is, longitude and latitude) of two census tracts is used as a distance measure. The compact representation of $\mathbf{W}$ is given in the **CensusTr_WList** data set. The map data set **US_CensusTr** contains the mapping information for the 2000 US Census tracts.

You can plot the values of the dependent variable y on the map by using the following statements:

```
proc gmap map=us_censustr data=SGF2017.censustr_simdata;
  id ctidfp00;
  choro y/coutline=gray midpoints=(-8 to 10 by 3);
run;
```

Figure 5 graphically presents the values of y for each census tract. According to this map, similar values of y tend to cluster together, indicating a positive spatial dependence in the data.

Figure 5 Observed Values of y for Census Tracts



Because the **CensusTr_SimData** data set is so large, it becomes infeasible to use the exact computation. However, you can use Chebyshev and Taylor approximations in PROC SPATIALREG. The following statements show you how to fit the SAR model to the **CensusTr_SimData** data set by using a 10th-order Chebyshev polynomial:

```
proc spatialreg data=SGF2017.censustr_simdata wmat=SGF2017.censustr_wlist
  approximation=(NMC=100 order=10 seed=1);
  model y=x1-x7/type=SAR;
  output out=sar_censustr pred=pred xbeta=xbeta;
  spatialid SID;
run;
```

Estimating the SAR model by using the 10th-order Chebyshev polynomial takes about 36 seconds. Figure 6 shows the parameter estimation results, which are close to the true values that are used for data generation.

Figure 6 Parameter Estimates for SAR Model Using Chebyshev Approximation for **CensusTr_SimData** Data Set

The SPATIALREG Procedure

Model: MODEL 1
Dependent Variable: y

Parameter Estimates

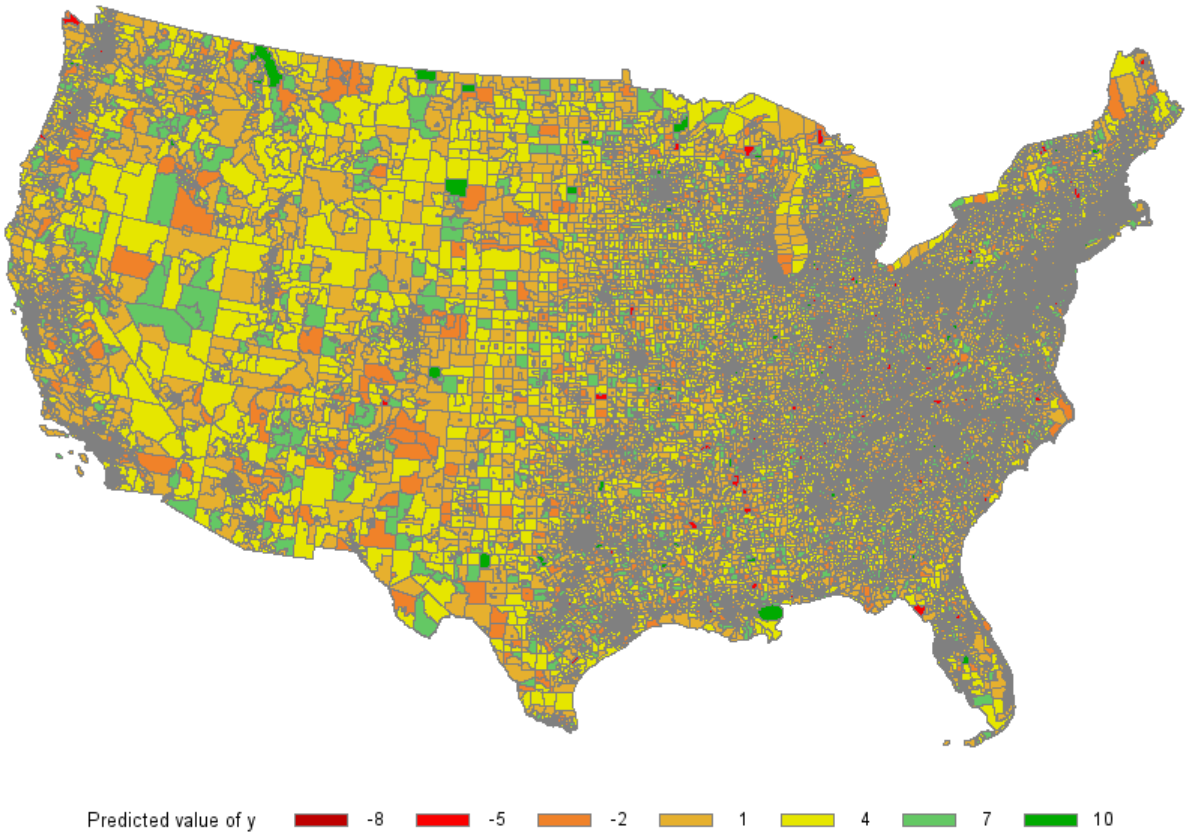
Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept	1	1.793634	0.005620	319.13	<.0001
x1	1	-0.899255	0.002788	-322.54	<.0001
x2	1	1.202600	0.002808	428.23	<.0001
x3	1	0.704044	0.002789	252.47	<.0001
x4	1	0.601804	0.002796	215.22	<.0001
x5	1	-0.996427	0.002786	-357.66	<.0001
x6	1	0.799476	0.002794	286.14	<.0001
x7	1	-0.600595	0.002795	-214.85	<.0001
_rho	1	0.303034	0.001910	158.68	<.0001
_sigma2	1	0.504769	0.002805	179.95	<.0001

To plot the predicted values of y on the map, you can submit the following statements:

```
proc gmap map=us_censustr data=sar_censustr;
  id ctidfp00;
  choro pred/coutline=gray midpoints=(-8 to 10 by 3);
run;
```

Figure 7 graphically presents the predicted values of y for each census tract. Comparing Figure 7 to Figure 5, you can see that the predicted values from the SAR model that you create using Chebyshev approximation seem to capture the overall pattern in the data.

Figure 7 Predicted Values of y for Census Tracts Using Chebyshev Approximation



You can also submit the following statements to fit the SAR model by using a 50th-order Taylor polynomial:

```
proc spatialreg data=SGF2017.censustr_simdata wmat=SGF2017.censustr_wlist
  approximation=(Taylor NMC=100 order=50 seed=1);
  model y=x1-x7/type=SAR;
  output out=sart_censustr pred=pred xbeta=xbeta;
  spatialid SID;
run;
```

Estimating the SAR model by using a 50th-order Taylor polynomial takes about 42 seconds. Figure 8 shows the results of this model, which are almost identical to the results in Figure 6. The results in both Figure 6 and Figure 8 are close to the true values that are used for data generation.

Figure 8 Parameter Estimates for SAR Model Using Taylor Approximation for **CensusTr_SimData** Data Set

The SPATIALREG Procedure

Model: MODEL 1
Dependent Variable: y

Parameter Estimates

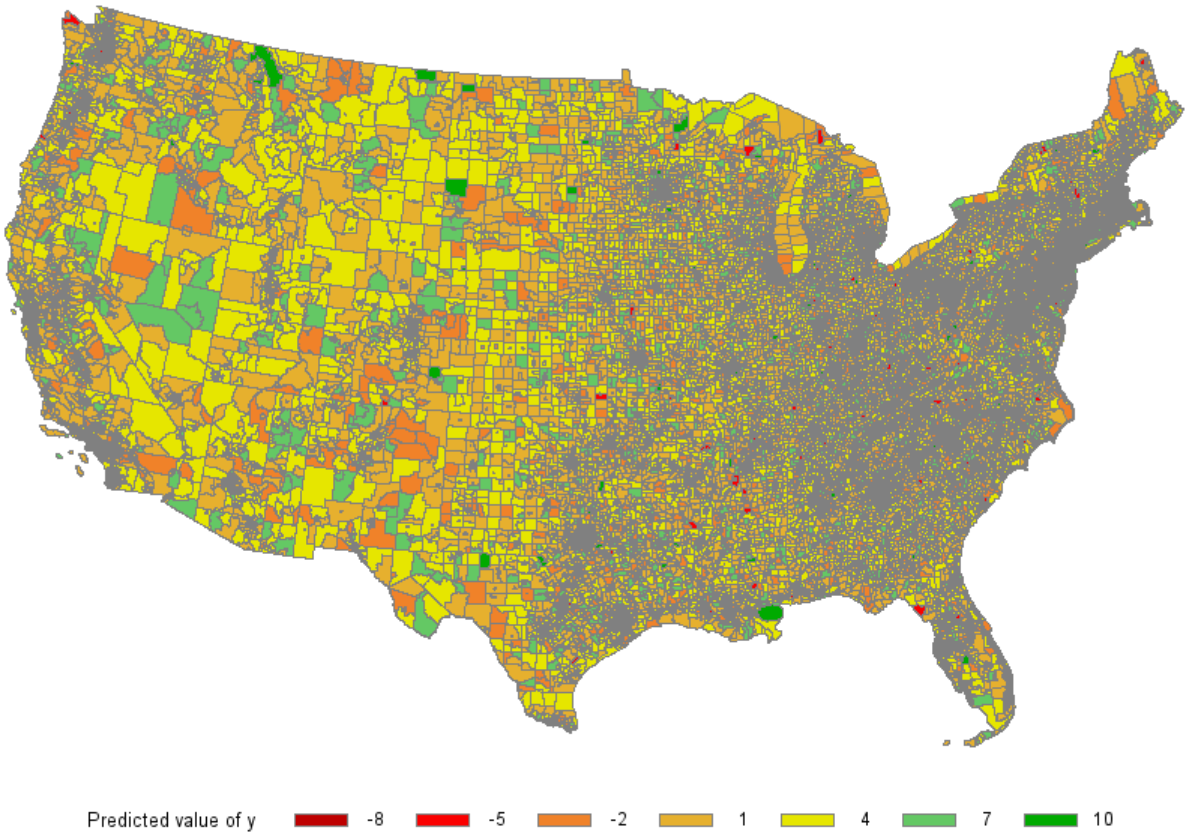
Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept	1	1.793634	0.005620	319.13	<.0001
x1	1	-0.899255	0.002788	-322.54	<.0001
x2	1	1.202600	0.002808	428.23	<.0001
x3	1	0.704044	0.002789	252.47	<.0001
x4	1	0.601804	0.002796	215.22	<.0001
x5	1	-0.996427	0.002786	-357.66	<.0001
x6	1	0.799476	0.002794	286.14	<.0001
x7	1	-0.600595	0.002795	-214.85	<.0001
_rho	1	0.303034	0.001910	158.68	<.0001
_sigma2	1	0.504769	0.002805	179.95	<.0001

Similarly, you submit the following statements to plot the predicted values of y:

```
proc gmap map=us_censustr data=sart_censustr;
  id ctidfp00;
  choro pred/coutline=gray midpoints=(-8 to 10 by 3);
run;
```

Figure 9 graphically presents the predicted values of y for each census tract. Comparing Figure 9 to Figure 5, you can see that the predicted values from the SAR model that you create using Taylor approximation seem to capture the pattern in the data.

Figure 9 Predicted Values of y for Census Tracts Using Taylor Approximation



CONCLUSION

This paper introduces the new SPATIALREG procedure for spatial econometric modeling released in SAS/ETS 14.2. This procedure enables you to fit a wide array of spatial econometric models to account for a variety of spatial interactions in the data. The main challenge of analyzing spatial data is often the design of the spatial matrix and its resulting size. When the data are large, the cost of forming, storing, and processing this matrix, along with other increased computational requirements imposed by the size of the problem, can be prohibitive. To circumvent the storage issue, PROC SPATIALREG enables you to use the compact form of the spatial weights matrix when appropriate. To alleviate the computational burden, two approximation techniques—Taylor and Chebyshev approximation—are available. Using the compact form of the spatial weights matrix and the two approximations enables the SPATIALREG procedure to provide spatial modeling capabilities for very large data sets that would otherwise be difficult to fit in memory, thus making the modeling problem computationally feasible.

REFERENCES

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Amsterdam: Springer.

Elhorst, J. P. (2013). *Spatial Econometrics: From Cross-Sectional Data to Spatial Panels*. Berlin: Springer.

Tobler, W. (1970). "A Computer Movie Simulating Urban Growth in the Detroit Region." *Economic Geography* 46:234–240.

Wheeler, C. H. (2003). "Evidence on Agglomeration Economies, Diseconomies, and Growth." *Journal of Applied Econometrics* 18:79–104.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors:

Guohui Wu and Jan Chvosta
SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513
Guohui.Wu@sas.com or Jan.Chvosta@sas.com