How Do My Neighbors Affect Me? SAS/ETS® Methods for Spatial Econometric Modeling

Guohui Wu and Jan Chvosta, SAS Institute Inc.

ABSTRACT

Contemporary data-collection processes usually involve recording information about the geographic location of each observation. This geospatial information provides modelers with opportunities to examine how the interaction of observations affects the outcome of interest. For example, it is likely that car sales from one auto dealership might depend on sales from a nearby dealership either because the two dealerships compete for the same customers or because of some form of unobserved heterogeneity common to both dealerships. Knowledge of the size and magnitude of the positive or negative "spillover" effect is important for creating pricing or promotional policies. This paper describes how geospatial methods are implemented in SAS/ETS® and illustrates some ways you can incorporate spatial data into your modeling toolkit.

INTRODUCTION

Spatial data are common in modern data science because of the wide availability of global positioning system (GPS) and geographical information system (GIS) data collection devices. These devices enable you to collect data that include GIS attributes such as spatial coordinates and topology. The GIS attributes enable the data to be mapped and thus categorized as spatial data. One important characteristic of spatial data is that neighborhood structure can be defined for observation units in space via certain metrics. This characteristic makes modeling and analysis of spatial data different from that of nonspatial data. For spatial data analysis, geospatial methods are advantageous and have become the de facto option.

The motivation behind spatial data analysis is the first law of geography; that is, "everything is related to everything else, but near things are more related than distant things" (Tobler 1970). According to the first law of geography, the unique characteristic about spatial data is that they are often correlated. Moreover, the strength of such a correlation between observations depends on their closeness in space, as measured by certain metrics. For example, it is likely that housing prices in one county might depend on the prices in a nearby county either because of the latter's economic growth or because of some form of unobserved heterogeneity common to both counties.

Spatial dependence in spatial data can arise from three different forms: endogenous interaction effects, exogenous interaction effects, and the interaction effects among the error terms (Elhorst 2013). The endogenous interaction effects exist when the value of the dependent variable in one location is affected by its values in other locations. The exogenous interaction effects, however, refer to the fact that the value of the dependent variable in one location is affected by the values of independent explanatory variables in other locations. Spatial dependence can also occur in the error terms. In this case, the error in one location is correlated with errors in other locations. To a large extent, understanding the form of spatial dependence is crucial for analyzing spatial data and your choice of models.

The past few decades have seen the increasing popularity of spatial data in econometrics and the demand for analytical software and tools that are well suited for spatial econometric modeling. Although you might be tempted to analyze spatial data by using tools that are designed exclusively for nonspatial data, you run the risk of erroneous conclusions. For example, you might have either biased or inefficient parameter estimates, depending on the discrepancy between your model and the underlying true model (LeSage and Pace 2009). The purpose of this paper is to introduce the methods available for spatial econometric modeling and some spatial features in SAS/ETS procedures. In particular, it demonstrates spatial econometric modeling methods that apply both to spatial data that contain a discrete dependent variable and to spatial data that contain a continuous dependent variable.

This paper is organized as follows: First, an overview of models for count data in the COUNTREG procedure is provided. The spatial features in PROC COUNTREG enable you to consider a wide array of regular and zero-inflated count regression models that include a spatial lag of X effect. Second, a set of spatial models in the SPATIALREG procedure is reviewed. These models can be used for data that contain a continuous dependent variable and exhibit endogenous and exogenous interaction effects in addition to spatial dependence in the error terms. In particular,

PROC SPATIALREG enables you to fit models such as the spatial autoregressive (SAR) model, spatial error model (SEM), spatial autoregressive confused (SAC) model, spatial moving average (SMA) model, and so on. Examples are provided for each procedure to demonstrate its utility. Finally, discussion and concluding comments are provided for spatial econometric modeling.

SPATIAL ECONOMETRIC MODELING

SAS/ETS provides two procedures you can use for spatial econometric modeling: PROC COUNTREG and PROC SPATIALREG. Your choice of procedure depends on the type of the dependent variable in your data. If you have count data, you can use the COUNTREG procedure for your analytical needs. If you have continuous data, the SPATIALREG procedure enables you to fit a range of widely used spatial econometric models.

COUNT REGRESSION FOR SPATIAL DATA WITH A DISCRETE DEPENDENT VARIABLE

The COUNTREG procedure enables you to analyze regression models for count data. The procedure includes a rich set of regression models such as Poisson regression, Conway-Maxwell-Poisson regression, negative binomial regression, and their respective zero-inflated variations. The zero-inflated Poisson, zero-inflated Conway-Maxwell-Poisson regression, and zero-inflated negative binomial regression models are suitable if your data include an excessive number of zero counts.

In the most general form, count regression models in the COUNTREG procedure can be described as

$$y_i|\psi_i, \mu_i, \nu_i \sim \psi_i 1(y_i = 0) + (1 - \psi_i)g(y_i|\mu_i, \nu_i), i = 1, 2, \ldots, n$$

where ψ_i is the probability of zero-inflation for the ith observation and n is the sample size. Moreover, μ_i and ν_i are parameters specific to the distribution function g. You can model μ_i, ψ_i, and ν_i as

$$\mu_i = \exp(\mathbf{x}_i'\boldsymbol{\beta}_1), \ \psi_i = F(\mathbf{z}_i'\boldsymbol{\gamma}_1), \ \nu_i = \exp(-\mathbf{g}_i'\boldsymbol{\delta}_1)$$

where $\mathbf{z}_i$, $\mathbf{x}_i$, and $\mathbf{g}_i$ are the the vectors of ith observation on covariates.

For the distribution function g, you can specify a Poisson, Conway-Maxwell-Poisson, or negative binomial distribution. For the zero-inflation link function F, you can choose either a probit function or a logistic function. In the vector form, count regression models in the COUNTREG procedure are represented as

$$\boldsymbol{\mu} = \exp(\mathbf{X}_1\boldsymbol{\beta}_1), \ \boldsymbol{\psi} = F(\mathbf{Z}_1, \boldsymbol{\gamma}_1), \ \boldsymbol{\nu} = \exp(-\mathbf{G}_1\boldsymbol{\delta}_1)$$

where $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)'$, $\boldsymbol{\psi} = (\psi_1, \psi_2, \ldots, \psi_n)'$, and $\boldsymbol{\nu} = (\nu_1, \nu_2, \ldots, \nu_n)'$. Moreover, $\mathbf{X}_1$, $\mathbf{Z}_1$, and $\mathbf{G}_1$ are the design matrices whose rows consist of $\mathbf{x}_i'$, $\mathbf{z}_i'$, and $\mathbf{g}_i'$, respectively.

The syntax of a simple call to the COUNTREG procedure has the following form:

```
proc countreg data=carsale;
  model y=x1 x2/dist=POISSON;
  zeromodel y~z1 z2/link=normal;
  dispmodel y~g1 g2;
run;
```

The MODEL statement enables you to specify the dependent variable and the covariates for μ_i. You use the DIST= option to specify the type of distribution function (g). Depending on your analytical needs, both the ZEROMODEL and DISPMODEL statements are optional. If you want to consider a zero-inflated model, you can use the ZEROMODEL statement to specify covariates for ψ_i. In this case, you can specify LINK=NORMAL (or LINK=LOGISTIC) if you want to use a probit (or logistic) zero-inflation link function. If you specify the DIST=COMPOISSON option, you can use the DISPMODEL statement to specify covariates for ν_i.

Now suppose your data are spatial and contain some GIS attributes, such as spatial coordinates. These GIS attributes provide you some extra information about the data. For example, you can measure proximity between two geographical locations according to certain metrics. In the spatial econometrics context, the proximity is often represented by a spatial weights matrix $\mathbf{W}$. In practice, some requirements are imposed on $\mathbf{W}$ (Anselin 2001; Elhorst 2013). First, $\mathbf{W}$ is an $n \times n$ matrix with nonnegative entries. Second, the diagonal elements of $\mathbf{W}$ should be zeros. When $\mathbf{W}$ consists of zeros and ones, a value of 1 for the (i, j)th entry often indicates that locations i and j are neighbors of each other. Moreover, $\mathbf{W}$ is often row-normalized for ease of interpretation. In such a case, the (i, j)th entry of $\mathbf{W}$, W_{ij}, quantifies the impact of location j over location i.

For spatial count data, the COUNTREG procedure enables you to fit what is often called the spatial lag of X (SLX) model (LeSage and Pace 2009; Elhorst 2013). The SLX model in the COUNTREG procedure can be presented in vector form as

$$
\begin{aligned}
\boldsymbol{\mu} &= \exp(\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{W}\mathbf{X}_2\boldsymbol{\beta}_2) \\
\boldsymbol{\psi} &= F(\mathbf{Z}_1\boldsymbol{\gamma}_1 + \mathbf{W}\mathbf{Z}_2\boldsymbol{\gamma}_2) \\
\boldsymbol{\nu} &= \exp(-\mathbf{G}_1\boldsymbol{\delta}_1 - \mathbf{W}\mathbf{G}_2\boldsymbol{\delta}_2)
\end{aligned}
$$

Three spatial lag terms, $\mathbf{W}\mathbf{X}_2$, $\mathbf{W}\mathbf{Z}_2$, and $\mathbf{W}\mathbf{G}_2$, are included to account for exogenous interaction effects among the independent variables. In practice, you can choose $\mathbf{X}_2$, $\mathbf{Z}_2$, and $\mathbf{G}_2$ to be the same as $\mathbf{X}_1$, $\mathbf{Z}_1$, and $\mathbf{G}_1$, respectively.

A simple call to an SLX model in the COUNTREG procedure can take the following form:

```
proc countreg data=carsale wmat=W;
  model y=x1 x2/dist=poisson;
  zeromodel y~z1 z2/link=normal;
  dispmodel y~g1 g2;
  spatialeffects x3 x4;
  spatialzeroeffects z3 z4;
  spatialdispeffects g3 g4;
  spatialid County;
run;
```

The WMAT= option enables you to provide a spatial weights matrix $\mathbf{W}$ of your choice. In addition, you use SPATIALEFFECTS, SPATIALZEROEFFECTS, and SPATIALDISPEFFECTS statements to include variables in $\mathbf{X}_2$, $\mathbf{Z}_2$, and $\mathbf{G}_2$, respectively. You can specify the SPATIALDISPEFFECTS (or SPATIALZEROEFFECTS) statement only if the DISPMODEL (or ZEROMODEL) statement is present. The SPATIALID statement enables you to specify a variable to match observations in the two data sets provided by the DATA= option and the WMAT= option.

To illustrate the usage of SLX models in the COUNTREG procedure, synthetic data that include car sales in 100 counties in North Carolina are created using the SAS® code in the Appendix. In the simulated **CarSale** data, **WeeklySale** and **DailySale** refer respectively to the number of weekly and daily car sales for each car dealership. These two variables are counts and are used to illustrate count regression models. In comparison, the continuous variable **Revenue** refers to the total revenue from car sales for each car dealership and is used to illustrate methods for spatial data that contain a continuous dependent variable. Three variables, **X1**, **X2**, and **X3**, are assumed to be predictors that affect car sales: for example, **X1** might represent median household income, **X2** might represent population size, and **X3** might represent crime rate in each county. Table 1 shows the summary statistics for the variables in the **CarSale** data. According to Table 1, the data contain 100 observations, and the minimum and maximum numbers of both the weekly and daily car sales are 0 and 8, respectively.

Table 1 Summary Statistics for Selected Variables in the **CarSale** Data Set

Variable	N	Mean	Std Dev	Minimum	Maximum
WeeklySale	100	0.570	1.451	0	8
DailySale	100	0.350	1.095	0	8
Revenue	100	4.138	1.775	−0.521	7.569
X1	100	0.043	1.001	−2.268	3.174
X2	100	0.124	0.877	−1.900	2.247
X3	100	0.020	1.028	−2.511	3.140

The following statements fit a Poisson regression model with an SLX effect for the simulated **CarSale** data:

```
libname SGF2016 'U:\SGF2016';
proc countreg data=SGF2016.carsale wmat=SGF2016.W;
  model weeklysale=x1 x2 x3/dist=poisson;
  spatialeffects x1 x2 x3;
  spatialid County;
run;
```

Here the spatial weights matrix W provides neighborhood information for counties in the state of North Carolina. Two counties are defined to be neighbors of each other if they share a common border. The variables **X1**, **X2**, and **X3** are linked to the mean parameter of the Poisson distribution. Figure 1 shows parameter estimation results, where W_x1, W_x2, and W_x3 denote the spatial lag of **X1**, **X2**, and **X3**, respectively. Moreover, you can see from Figure 1 that **X1**, **X2**, and **X3** are significant at the 5% level.

Figure 1 Parameter Estimates for Poisson Regression with an SLX Effect

The COUNTREG Procedure

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept	1	-1.945692	0.295041	-6.59	<.0001
x1	1	1.225100	0.150402	8.15	<.0001
x2	1	-0.498799	0.171867	-2.90	0.0037
x3	1	0.465696	0.169787	2.74	0.0061
W_x1	1	1.169532	0.387681	3.02	0.0026
W_x2	1	0.810996	0.327123	2.48	0.0132
W_x3	1	-0.797009	0.424839	-1.88	0.0607

If you want to test whether spatial lag effects are needed in the model, the TEST statement in PROC COUNTREG enables you to perform hypothesis testing. The COUNTREG procedure supports three different tests: likelihood ratio (LR), Wald, and Lagrange multiplier (LM). For example, if your null hypothesis is H_0: W_x1=W_x2=W_x3=0 for the preceding Poisson regression model that has an SLX effect, you can issue the following statements to conduct hypothesis testing:

```
proc countreg data=SGF2016.carsale wmat=SGF2016.W;
  model weeklysale=x1 x2 x3/dist=poisson;
  spatialeffects x1 x2 x3;
  spatialid County;
  test_poisson_slx: test W_x1=0,W_x2=0,W_x3=0/all;
run;
```

In the TEST statement, you specify the null hypthesis and the type of tests. The ALL option in the TEST statement requests all available tests. If you want to request only one of three tests, you replace ALL with LM, LR, or Wald. Figure 2 shows the test results from the three tests. According to Figure 2, you can conclude that H_0 should be rejected at the 5% significance level.

Figure 2 Test Results for H_0: W_x1 = W_x2 = W_x3 = 0 in Poisson Regression with an SLX Effect

The COUNTREG Procedure

Test Results				
Test	Type	Statistic	Pr > ChiSq	Label
TEST_POISSON_SLX	Wald	16.57183	0.0009	W_x1 = 0, W_x2 = 0, W_x3 = 0
TEST_POISSON_SLX	L.R.	19.02948	0.0003	W_x1 = 0, W_x2 = 0, W_x3 = 0
TEST_POISSON_SLX	L.M.	17.27478	0.0006	W_x1 = 0, W_x2 = 0, W_x3 = 0

The variable **DailySale** in the CarSale data contains excess zeros and can be used to illustrate the zero-inflated models in the COUNTREG procedure. To compute the proportion of zero counts, you can use the FREQ procedure. Table 2 shows you the results: 81 out of 100 car dealerships have zero car sales on a particular day. As a result, you might want to consider a zero-inflated regression model.

Table 2 Proportion of Dealers Selling Each Observed Number of Cars on a Day

DailySale	Frequency	Percent	Cumulative Frequency	Cumulative Percentage
0	81	81.00	81	81.00
1	14	14.00	95	95.00
2	2	2.00	97	97.00
3	3	3.00	98	98.00
6	1	1.00	99	99.00
8	1	1.00	100	100.00

You can fit a zero-inflated regression model with an SLX effect to the **CarSale** data by using the following statements:

```
proc countreg data=SGF2016.carsale wmat=SGF2016.W;
  model dailysale=x1 x2 x3/dist=poisson;
  zeromodel dailysale~z1 z2/link=logistic;
  spatialeffects x1 x2 x3;
  spatialzeroeffects z1 z2;
  spatialid County;
run;
```

Here **DailySale** is the daily car sales for each dealership. The two additional variables, **z1** and **z2**, are covariates to model the probability of zero inflation. Figure 3 shows parameter estimation results for the zero-inflated Poisson model with an SLX effect. In Figure 3, Inf_W_z1 and Inf_W_z2 denote the spatial lag of **z1** and **z2**, respectively, in the ZEROMODEL statement.

Figure 3 Parameter Estimates for Zero-Inflated Poisson Regression with an SLX Effect

The COUNTREG Procedure

			Standard		Approx
Parameter	DF	Estimate	Error	t Value	Pr > |t|
Intercept	1	-1.622217	0.408357	-3.97	<.0001
x1	1	0.902000	0.263661	3.42	0.0006
x2	1	-0.373442	0.220280	-1.70	0.0900
x3	1	0.388271	0.222023	1.75	0.0803
W_x1	1	1.533449	0.587689	2.61	0.0091
W_x2	1	0.484590	0.362219	1.34	0.1809
W_x3	1	-1.004189	0.505264	-1.99	0.0469
Inf_Intercept	1	0.616017	1.054285	0.58	0.5590
Inf_z1	1	4.767497	3.426538	1.39	0.1641
Inf_z2	1	-5.544391	4.003067	-1.39	0.1660
Inf_W_z1	1	6.456906	4.128179	1.56	0.1178
Inf_W_z2	1	8.445003	4.802757	1.76	0.0787

The zero-inflated count regression models account for the excess zeros in comparison to the regular count regression models. However, for the **CarSale** data, it might be of interest to determine whether you should include any SLX effect terms in the model for the probability of zero counts, ψ_i. You can use the TEST statement for this purpose. First, you formulate the null hypothesis H_0: Inf_W_z1 = Inf_W_z2 = 0. To formally test the hypothesis by using the LR test, you then issue the following statements:

```
proc countreg data=SGF2016.carsale wmat=SGF2016.W;
  model dailysale=x1 x2 x3/dist=poisson;
  zeromodel dailysale~z1 z2/link=logistic;
  spatialeffects x1 x2 x3;
  spatialzeroeffects z1 z2;
  spatialid County;
  test_zip_slx: test Inf_W_z1=0,Inf_W_z2=0/LR;
run;
```

Similarly, you specify the null hypothesis and the type of test in the TEST statement. The parameter names that relate to the SPATIALZEROEFFECTS statement have Inf_W_ as a prefix. Figure 4 shows the test results from the LR test. According to Figure 4, you can conclude that H_0 should be rejected at the 5% significance level.

Figure 4 Test Results for H_0: Inf_W_z1 = Inf_W_z2 = 0 in the Zero-Inflated Poisson Regression with an SLX Effect

The COUNTREG Procedure

		Test Results		
Test	Type	Statistic	Pr > ChiSq	Label
TEST_ZIP_SLX	L.R.	16.04272	0.0003	Inf_W_z1 = 0, Inf_W_z2 = 0

METHODS FOR SPATIAL DATA WITH A CONTINUOUS DEPENDENT VARIABLE

Although count data are common in spatial econometric modeling, you might have spatial data that contain a continuous dependent variable. For example, revenues of each car dealership from each county do not need to be integers. In this and many other cases, the variable of interest is continuous. For spatial data that contain a continuous dependent variable, you can use the SPATIALREG procedure for spatial econometric modeling.

The following type of models are widely used in spatial econometric modeling (Anselin 1988; LeSage and Pace 2009), and they can be formulated as shown:

- spatial autoregressive (SAR) model

$$\mathbf{y} = \rho\mathbf{W}_1\mathbf{y} + \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon} \tag{1}$$

- spatial error model (SEM)

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{u}, \quad \mathbf{u} = \lambda\mathbf{W}_2\mathbf{u} + \boldsymbol{\epsilon} \tag{2}$$

- spatial moving average (SMA) model

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{u}, \quad \boldsymbol{u} = \boldsymbol{\epsilon} - \lambda\mathbf{W}_2\boldsymbol{\epsilon} \tag{3}$$

- spatial autoregressive moving average (SARMA) model

$$\mathbf{y} = \rho\mathbf{W}_1\mathbf{y} + \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{u}, \quad \boldsymbol{u} = \boldsymbol{\epsilon} - \lambda\mathbf{W}_2\boldsymbol{\epsilon} \tag{4}$$

- spatial autoregressive confused (SAC) model

$$\mathbf{y} = \rho\mathbf{W}_1\mathbf{y} + \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{u}, \quad \boldsymbol{u} = \lambda\mathbf{W}_2\mathbf{u} + \boldsymbol{\epsilon} \tag{5}$$

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 location i for $i = 1, 2, \ldots, n$. Moreover, $\mathbf{X}_1$ is an $n \times p$ design matrix, whereas $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)'$ with $\epsilon_i \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$ for $i = 1, 2, \ldots, n$. The two n-by-n spatial weights matrices, $\mathbf{W}_1$ and $\mathbf{W}_2$, define spatial configuration of n spatial locations. The key difference between the SEM and SMA models and between the SARMA and SAC models lies in a different covariance structure for $\mathbf{u}$. In some cases, the two matrices $\mathbf{W}_1$ and $\mathbf{W}_2$ in SARMA and SAC models might be identical; that is $\mathbf{W}_1 = \mathbf{W}_2$.

You use a SAR model to account for endogenous interaction effects. You use either an SEM or SMA model to account for spatial dependence in the error terms. If both endogenous interaction effects and spatial dependence in the error terms need to be accounted for, you can use a SARMA or SAC model.

The SPATIALREG procedure in SAS/ETS enables you to fit the preceding spatial econometric models. You can make a simple call to PROC SPATIALREG by issuing the following statements:

```
proc spatialreg data=carsale wmat=W;
  model y=x1 x2/type=SAR;
  spatialid County;
run;
```

As in the COUNTREG procedure, you can supply your data and a spatial weights matrix by using the DATA= option and the WMAT= option, respectively. The TYPE= option enables you to specify the type of spatial econometric model. You can specify one of the following six values: SAR, SEM, SMA, SARMA, SAC, or LINEAR. You use TYPE=SAR if you want to fit a SAR model. You can request a SARMA or SAC model with a common spatial weights matrix by specifying TYPE=SARMA or TYPE=SAC, respectively. When you specify TYPE=LINEAR, a linear regression model is considered. You use the SPATIALID statement to specify a variable that can be used to match observations in the two data sets that are specified in the DATA= and WMAT= options.

You can include an SLX effect into the set of models that are described in preceding equations. In this scenario, the equations that involve $\mathbf{y}$ become as follows for a SAR, SARMA, or SAC model:

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

For an SEM or SMA model, the equations that involve $\mathbf{y}$ become

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_1 + \mathbf{W}_1\mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$$

As in the COUNTREG procedure, you use the SPATIALEFFECTS statement to specify covariates whose spatial lag effects should be included in the model. For example, a call to a SAR model with an SLX effect might look like the following:

```
proc spatialreg data=carsale wmat=W;
  model y=x1 x2/type=SAR;
  spatialeffects x3 x4;
  spatialid County;
run;
```

The visualization of spatial data plays an important role in spatial data analysis, because the plot can help identify the pattern in your data. For the **CarSale** data, you can visualize the data on the map by using the following statements:

```
ods graphics on;
goptions reset=global gunit=pct border colors=(pink salmon red orange cyan blue);
proc gmap data=SGF2016.carsale map=SGF2016.nccounty_map;
  choro revenue/ coutline=gray levels=6 midpoints=(0.5 to 7 by 1.2);
  id County;
run;
```

Figure 5 graphically presents car sale revenues of car dealers in each county of North Carolina. You can see that similar values of revenue seem to cluster together, suggesting that there is spatial dependence in your data.

Figure 5 Plot of Car Sales Revenue

| revenue | 0.5 | 1.7 | 2.9 | 4.1 | 5.3 | 6.5 |
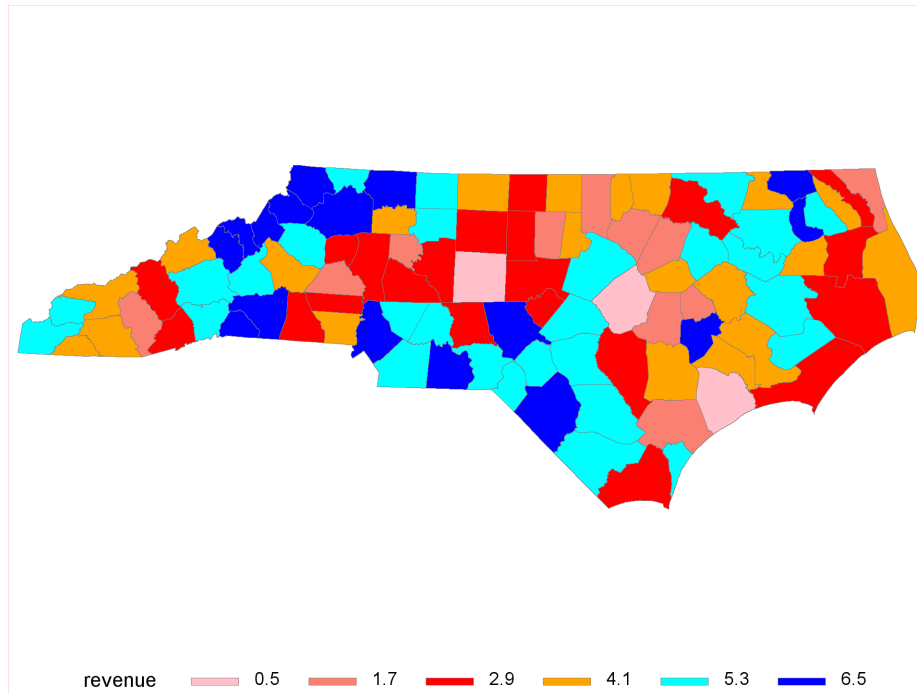
Because spatial dependence and a clustering pattern are present in the data, it might be reasonable for you to fit a SAR model. You can fit a SAR model by issuing the following statements:

```
proc spatialreg data=SGF2016.carsale wmat=SGF2016.W;
  model revenue=x1 x2 x3/type=SAR;
  spatialid County;
run;
```

Figure 6 shows parameter estimation results from a SAR model. In Figure 6, _rho and _sigma2 are the internal names for ρ and σ^2, respectively. The coefficient ρ is estimated to be 0.75, indicating that there could be a strong positive correlation.

Figure 6 Parameter Estimates for a SAR Model

The SPATIALREG Procedure

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept	1	1.303514	0.255510	5.10	<.0001
x1	1	-1.143304	0.069000	-16.57	<.0001
x2	1	0.416048	0.076724	5.42	<.0001
x3	1	-0.583617	0.065577	-8.90	<.0001
_rho	1	0.682037	0.058804	11.60	<.0001
_sigma2	1	0.438933	0.063444	6.92	<.0001

The following statements include an SLX effect into the preceding SAR model, test the significance of the SLX effect, and examine the predicted values and residuals:

```
proc spatialreg data=SGF2016.carsale wmat=SGF2016.W;
  model revenue=x1 x2 x3/type=SAR;
  spatialeffects x1 x2 x3;
  spatialid County;
  test_sar_slx: test W_x1=0,W_x2=0,W_x3=0/all;
  output out=SGF2016.sarm2 pred=pred resid=resid;
run;
```

Figure 7 shows parameter estimation results from the SAR model with an SLX effect. You see that the coefficient ρ is estimated to be 0.63, implying a strong positive spatial correlation. Figure 8 shows three different test results for the null hypothesis H_0: W_x1 = W_x2 = W_x3 = 0. According to Figure 8, you can conclude that H_0 should be rejected at the 5% significance level.

Figure 7 Parameter Estimates for a SAR Model with an SLX Effect

The SPATIALREG Procedure

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept	1	2.535573	0.350036	7.24	<.0001
x1	1	-1.304547	0.050391	-25.89	<.0001
x2	1	0.341334	0.050785	6.72	<.0001
x3	1	-0.446264	0.045435	-9.82	<.0001
W_x1	1	-0.952952	0.184334	-5.17	<.0001
W_x2	1	-0.712082	0.101350	-7.03	<.0001
W_x3	1	0.687763	0.099703	6.90	<.0001
_rho	1	0.420353	0.081515	5.16	<.0001
_sigma2	1	0.177390	0.025306	7.01	<.0001

Figure 8 Test Results for H_0: W_x1 = W_x2 = W_x3 = 0 in the SAR Model with an SLX Effect

Test Results				
Test	Type	Statistic	Pr > ChiSq	Label
TEST_SAR_SLX	Wald	152.717	<.0001	W_x1 = 0, W_x2 = 0, W_x3 = 0
TEST_SAR_SLX	L.R.	99.78587	<.0001	W_x1 = 0, W_x2 = 0, W_x3 = 0
TEST_SAR_SLX	L.M.	67.39617	<.0001	W_x1 = 0, W_x2 = 0, W_x3 = 0

The following statements enable you to visualize the predicted values and the corresponding residuals:

```
proc gmap data=SGF2016.sarm2 map=SGF2016.nccounty_map;
  choro pred/ coutline=gray levels=6 midpoints=(0.5 to 7 by 1.2);
  choro resid/ coutline=gray midpoints=(-1 to 0.9 by 0.35);
  id County;
run;
```

Figure 9 graphically presents the predicted car sales revenue on the map. You can see from Figure 9 that the plot of predicted values captures the trend of revenue as shown in Figure 5.

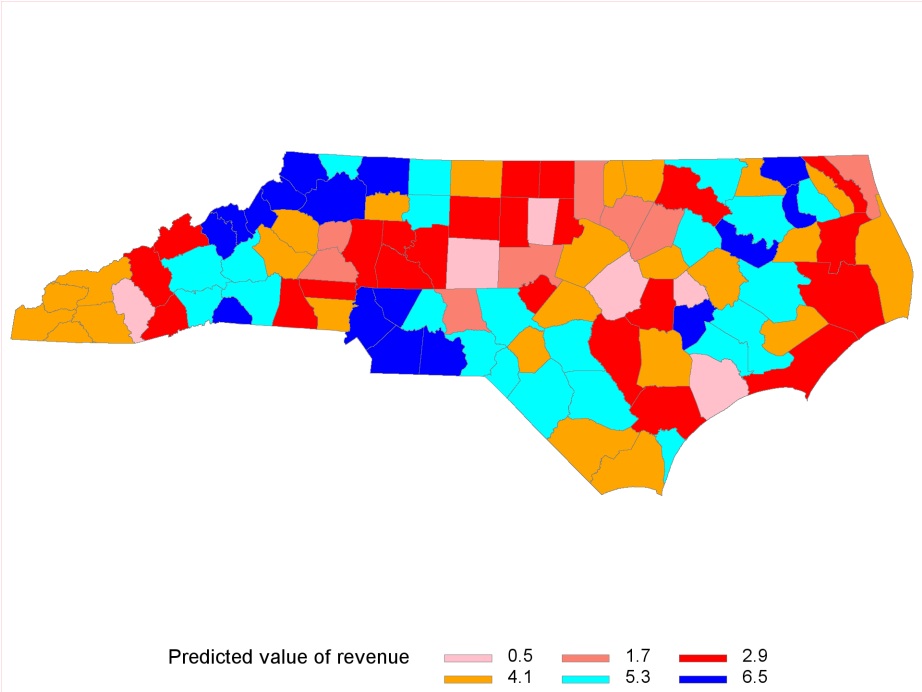**Figure 9** Plot of Predicted Car Sales Revenue in the SAR Model with an SLX Effect

Figure 10 presents the residuals on the map. You can see that similar values of residuals seem to cluster together. As a result, you might want to fit a SARMA model and test whether there is spatial dependence in the error terms because of unobserved heterogeneity.

Figure 10 Plot of Residuals in the SAR Model with an SLX Effect
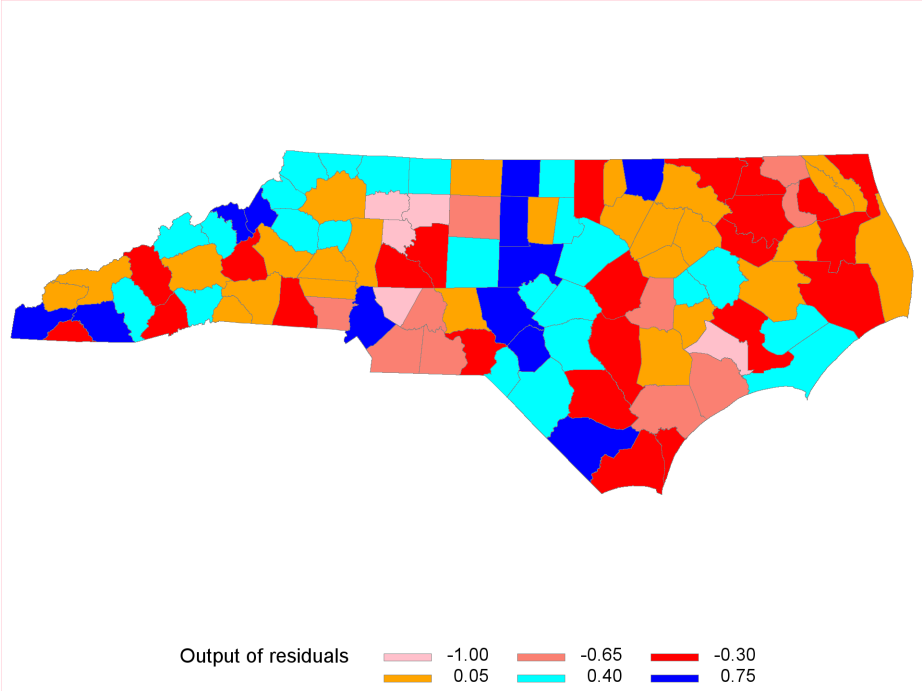
You can fit a SARMA model and formally test the hypothesis $H_0 : \lambda = 0$ by issuing the following statements:

```
proc spatialreg data=SGF2016.carsale wmat=SGF2016.W;
  model revenue=x1 x2 x3/type=SARMA;
  spatialeffects x1 x2 x3;
  spatialid County;
  test_sarma: test _lambda=0/all;
run;
```

Figure 11 shows the parameter estimation results for the preceding SARMA model, where _lambda is the internal name for the parameter λ. Comparing Figure 7 and Figure 11 demonstrates that the estimates for $\boldsymbol{\beta}$ are quite similar between these two models. For the test of $H_0 : \lambda = 0$, results in Figure 12 suggest that you cannot reject the null hypothesis at the 0.05 significance level under the three different tests being considered. This leads to the conclusion that the SAR model with an SLX effect is appropriate for the **CarSale** data.

Figure 11 Parameter Estimates for the SARMA Model

The SPATIALREG Procedure

Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept	1	2.788970	0.468000	5.96	<.0001
x1	1	-1.325725	0.056423	-23.50	<.0001
x2	1	0.331901	0.052845	6.28	<.0001
x3	1	-0.437798	0.047529	-9.21	<.0001
W_x1	1	-1.066450	0.223070	-4.78	<.0001
W_x2	1	-0.712831	0.108136	-6.59	<.0001
W_x3	1	0.674398	0.106492	6.33	<.0001
_rho	1	0.360436	0.109680	3.29	0.0010
_lambda	1	-0.182727	0.199827	-0.91	0.3605
_sigma2	1	0.179277	0.026018	6.89	<.0001

Figure 12 Test Results for $H_0 : \lambda = 0$ in the SARMA Model

Test Results

Test	Type	Statistic	Pr > ChiSq	Label
TEST_SARMA	Wald	0.836168	0.3605	_lambda = 0
TEST_SARMA	L.R.	0.906299	0.3411	_lambda = 0
TEST_SARMA	L.M.	0.827282	0.3631	_lambda = 0

CONCLUSION

This paper discusses the use of SAS/ETS procedures to analyze two types of spatial data that arise from spatial econometric modeling: spatial data that contain a count dependent variable and spatial data that contain a continuous dependent variable.

For spatial data that contain a count dependent variable, this paper illustrates how PROC COUNTREG enables you to fit Poisson regression, negative binomial regression, and Conway-Maxwell-Poisson regression models, in addition to their zero-inflated versions. By including the SLX effects into the model, you can account for direct impacts of spatial externalities in the explanatory variables.

For spatial data that contain a continuous dependent variable, a wide array of spatial econometric models are available in the SPATIALREG procedure. These spatial econometric models enable the spatial dependence in the data to be properly addressed. With PROC SPATIALREG, you can fit commonly used spatial econometric models such as SAR, SEM, SMA, SARMA, SAC, and so on.

REFERENCES

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

Anselin, L. (2001). "Spatial Econometrics." In *A Companion to Theoretical Econometrics*, edited by B. H. Baltagi, 310–330. Oxford: Wiley-Blackwell.

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

LeSage, J., and Pace, R. K. (2009). *Introduction to Spatial Econometrics*. Boca Raton, FL: CRC Press.

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

APPENDIX

This section contains SAS code that simulates the CarSale data.

```
libname SGF2016 'U:\SGF2016';
proc iml;
  /*Read in the list of neighbors for counties in North Carolina.*/
  use SGF2016.NCCounty_Neighbor_list nobs nobs;
  read all;

  /*Construct the spatial weights matrix Wmat.*/
  nloc=100;
  Wmat=repeat(0,nloc,nloc);

  do i=1 to nobs;
     rid=Row[i];
     cid=Col[i];
     Wmat[rid,cid]=1;
     Wmat[cid,rid]=1;
  end;

  /*Row-standardize Wmat.*/
  Wmat_rowstd=Wmat/Wmat[,+];

  /*Read in county names.*/
  use SGF2016.NCcounty_name;
  read all;
  County=compress(County);
  cSID=County`;

  /*Create the data set W from Wmat.*/
  create SGF2016.W from Wmat[rowname=County colname=cSID];
  append from Wmat[rowname=County];
  close SGF2016.W;

  /*Simulate county-level count data for car sales in North Carolina.*/
  call randseed(123456);
  beta_true={1.2,-0.4,0.5,0.6,0.7,-0.9};
  p=nrow(beta_true);
  x1=j(nloc,p/2);
  call randgen(x1,"Normal");

  ycount=j(nloc,1);
```

```
   x2=Wmat_rowstd*x1;
   x=x1 || x2;
   u=-1.8+x*beta_true;
   m=exp(u);
   call randgen(ycount,'poisson',m);


   /*Simulate county-level zero-inflated count data.*/
   gamma_true={-.6,1.2,-0.8,0.9,0.5};
   q=nrow(gamma_true)-1;
   yzero=j(nloc,1);
   z1=j(nloc,q/2);
   call randgen(z1,"Normal");
   z2=Wmat_rowstd*z1;
   z=j(nloc,1) || z1 || z2;
   psi=1/(1+exp(-z*gamma_true));
   B=j(nloc,1);
   call randgen(B,"bernoulli",1-psi);
   yzero=ycount#B;



   /*Simulate county-level revenue data for car sales in North Carolina.*/
   sig2_true=.2;
   rho_true=.6;
   e=j(nloc,1);
   call randgen(e,"Normal",0,sqrt(sig2_true));
   In=I(nloc);
   ycont=solve(In-rho_true*Wmat_rowstd,-u-e);

   ColID={"x1","x2","x3","z1","z2","weeklysale","revenue","dailysale"};
   simdata=x1 || z1 || ycount || ycont || yzero;
   create SGF2016.carsale from simdata[rowname=County colname=ColID];
   append from simdata[rowname=County];
   close SGF2016.carsale;
quit;
```

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