

Estimating Regression Models with Spatially Correlated Errors Using the MIXED Procedure.

Sam Field

University of Texas at Austin

Abstract

The REPEATED statement in the MIXED procedure offers a wide range of correlation structures for the error term in a mixed model. This paper discusses a subset of these structures applicable to spatial data and demonstrates their use using an example data set. First, the problem of spatial correlation is discussed as it relates to the assumption of independent errors in regression and ANOVA models. Second, a method of identifying the appropriate spatial correlation structure for a set of data is discussed and demonstrated on the example data using the VARIOGRAM procedure. Third, information from the identification stage is used to estimate a spatial correlation model for the example data using the MIXED procedure. The paper concludes with a discussion of some common methods of comparing alternative models of spatial correlation for a given set of data. The application of these methods is demonstrated using the sample data set.

Introduction

Spatial autocorrelation refers to the tendency of observations that are closer together in space to be more similar to one another than are observations further apart. One of the consequences of spatial autocorrelation in regression or ANOVA analysis is that estimates of model error are biased downward. This is because typical regression or ANOVA models assume that the data from which the mean response of the dependent variable is estimated, consists of a random sample of n independent observations of the dependent variable. However, when there exists dependencies among the n observations in a data set, a certain amount of duplication of information about the mean response of the dependent variable occurs. The consequence of this

duplication of information is that the sample contains less information than it would otherwise have, had the data come from an independent random sample of observations. An analogous situation would exist if a researcher increased the size of his/her sample by simply duplicating the observations already in the sample. Intuitively, we would not expect the additional cases to provide any additional information concerning the mean response of the dependent variable since the values of the new cases are wholly dependent on observations already in the sample. Nonetheless, conventional estimators of model error used in regression and ANOVA analysis would treat these additional cases as independent and the estimated sampling variability of model parameters would be reduced with the increase in the sample size.

If the spatial autocorrelation in a data set can be accounted for by a set of underlying explanatory variables, then the residuals of a regression or ANOVA model may no longer be auto-correlated and no adjustment to the model needs to be made. However, in practice it is often the case that the researcher does not know or cannot measure all of the relevant variables affecting the dependent variable and there remains some residual autocorrelation among the residuals due to the omission of these variables. This situation can lead to biased estimates of the effect parameters in a model (e.g. treatment means, regression coefficients, etc) as well as model error. One approach to this problem is to remove the effects of spatial correlation by incorporating a model of spatial correlation into the analysis. This paper briefly describes the spatial correlation models available in the Mixed Procedure in SAS, and demonstrates their use using an example data set.

Spatial Correlation Models in the Mixed Procedure

The models of spatial correlation available in the Mixed procedure have their origins in the field of geostatistics. In these models, the covariance between the error term of any two observations at different spatial locations (e_i, e_j) is taken to be a function of the distance between

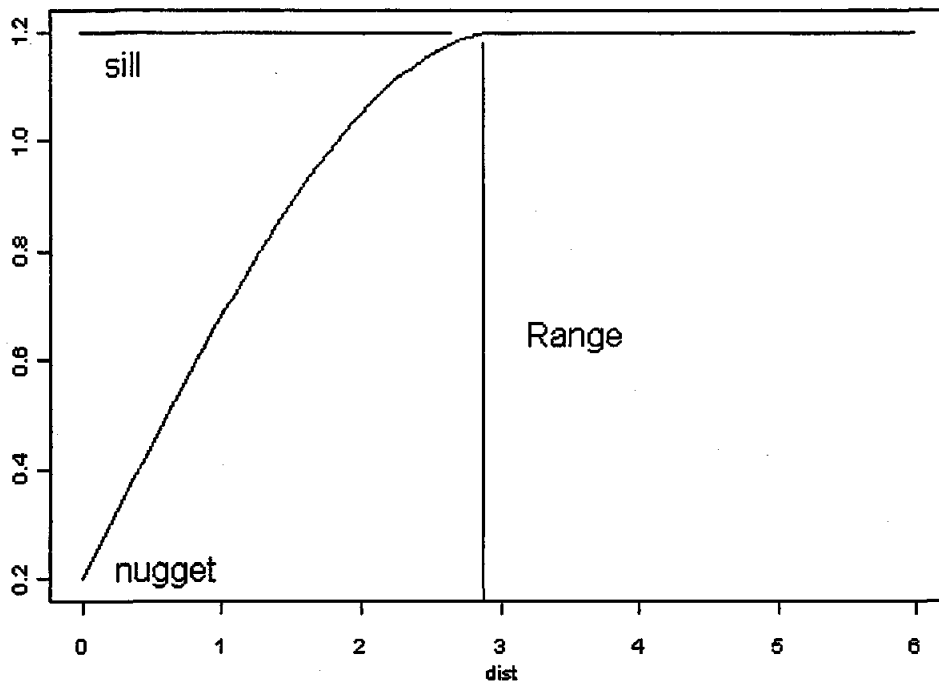
them. Without going into too much mathematical detail, its easy to represent each of the models available in the mixed procedure as a semivariogram. The semivariogram is a tool developed in geostatistics that is used to measure spatial variability as a function of distances between observations in a sample. The semivariogram is estimated using the formula below;

$$\gamma(h) = (1/2m)\sum(y_i - y_j)^2 \quad (1)$$

where m is the number of pairs of observations a distance, h , apart. The figure below presents an idealized graph of the semivariogram

Figure 1.

semivar



The figure above depicts how, in the presence of spatial autocorrelation, the similarity between any two observations decreases with the distance between them. The particular semivariogram depicted above is an example of the *spherical* model. Notice how the variance between observations rises and plateaus sharply with distance. Other commonly used models with different shapes are also available in PROC MIXED. These include the *exponential*,

gaussian, *linear*, *linear log*, and *power* models. The *nugget*, *sill*, and *range* are spatial covariance parameters of the semivariogram that can be estimated in PROC MIXED. The *nugget* represents the variance of the independent error term of the model. In other words it is that portion of the variance of e which is not spatially autocorrelated. The *range* represents the distance at which observations cease to be autocorrelated and the *sill* is the level of the variance at that distance. Although PROC MIXED does not compute the empirical semivariogram for a given set of data, knowledge about the empirical semivariogram is useful when fitting spatial correlation models in PROC MIXED. This point will be demonstrated below.

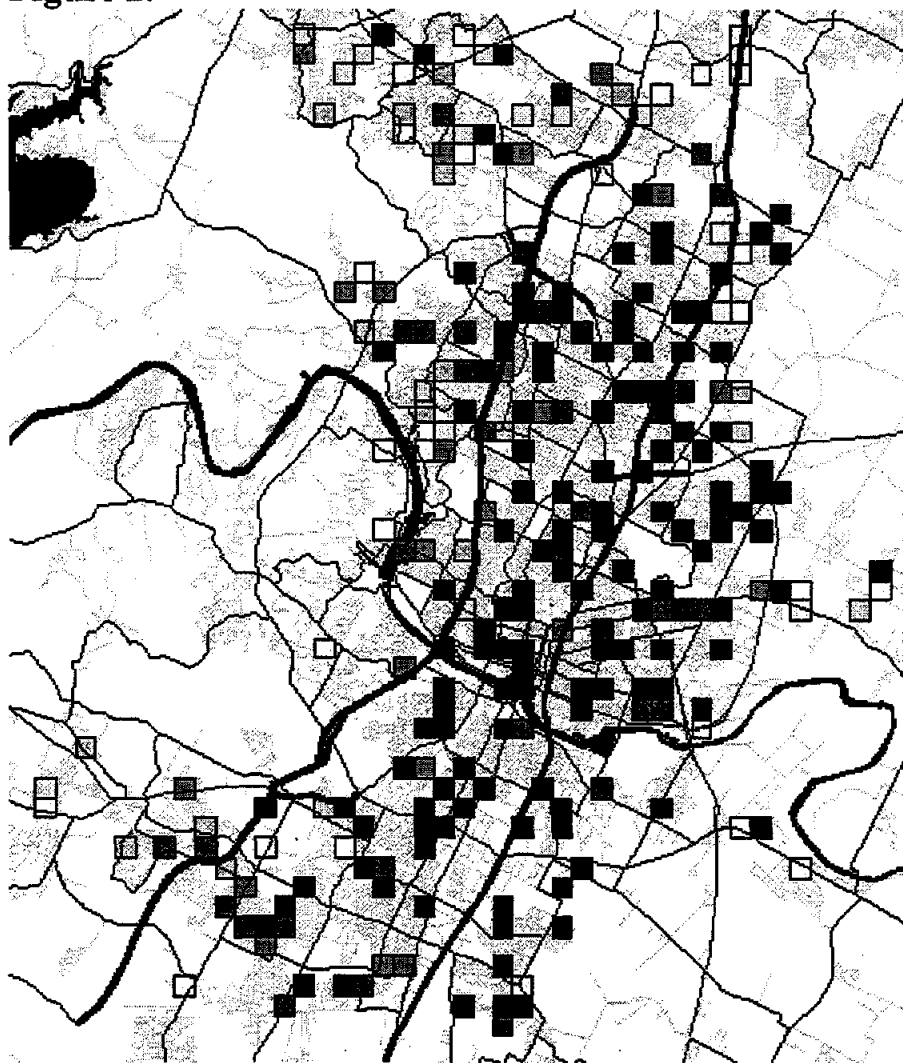
Estimating Spatial Correlation Models using the Mixed Procedure

As with many statistical procedures, the estimation of spatial correlation models using PROC MIXED is as much an art as it is a science. The analysis proceeds in three stages that are analogous to those that appear in time series analysis. First, there is the *identification stage* in which the researcher computes the empirical semivariogram using PROC VARIOGRAM and identifies a plausible correlation structure for the data. Second, information from the identification stage informs the selection of spatial correlation models (e.g. spherical, exponential, linear, etc) and plausible ranges of model parameters in the *model estimation stage*. In the final stage, *the model checking stage*, the estimated model is checked for how well it fits the data. In this stage, the results from several different models can be compared and one of them selected based on how well it fits the data. Using an example data set, the rest of this paper demonstrates each of these stages.

The example data comes from the Austin Police department and consists of counts of burglary events for the year 1999 (i.e. burglary of residence, burglary of auto, burglary of business) in the city of Austin aggregated up to square grid tiles. The data also contains counts of public order offending (i.e. public intoxication, panhandling, and disorderly conduct) for the

same areas and year. The objective of the analysis that follows is to examine the relationship between these two types of criminal events. A map of the city depicting the grid tiles and the spatial distribution of burglary events in Austin is shown below. Darker shaded grid tiles indicate areas of high rates of burglary (calculated as the number of events per mile of street network). The spaces in between the grid tiles indicate areas of the city that were not sampled.

Figure 2.



Stage 1: Identification

The first step in estimating spatial correlation models using the Mixed procedure is to compute the empirical semivariogram using the Semivariogram procedure. The semivariogram procedure computes the pairwise distance between all pairs of data points in a data set and then groups the pairs of observations into distance bins of a user-specified width. It then computes $\gamma(h)$ from equation 1 for each distance bin defined. The semivariogram for the burglary data was created by first extracting the residuals from an OLS regression of the log-transformed burglary rate on the log-transformed rate of public order offending. The syntax for the regression analysis is shown below.

```
proc reg data = burgdata;
  model log_PERCRIME = log_po/ partial ;
  output out = residuals
    p = yhat
    r = yresid ;
run;
```

The parameter estimates of the model specified in the syntax above indicate as significant and positive relationship between public order offending and burglary. These estimates along with the ANOVA results are shown below.

Analysis of Variance

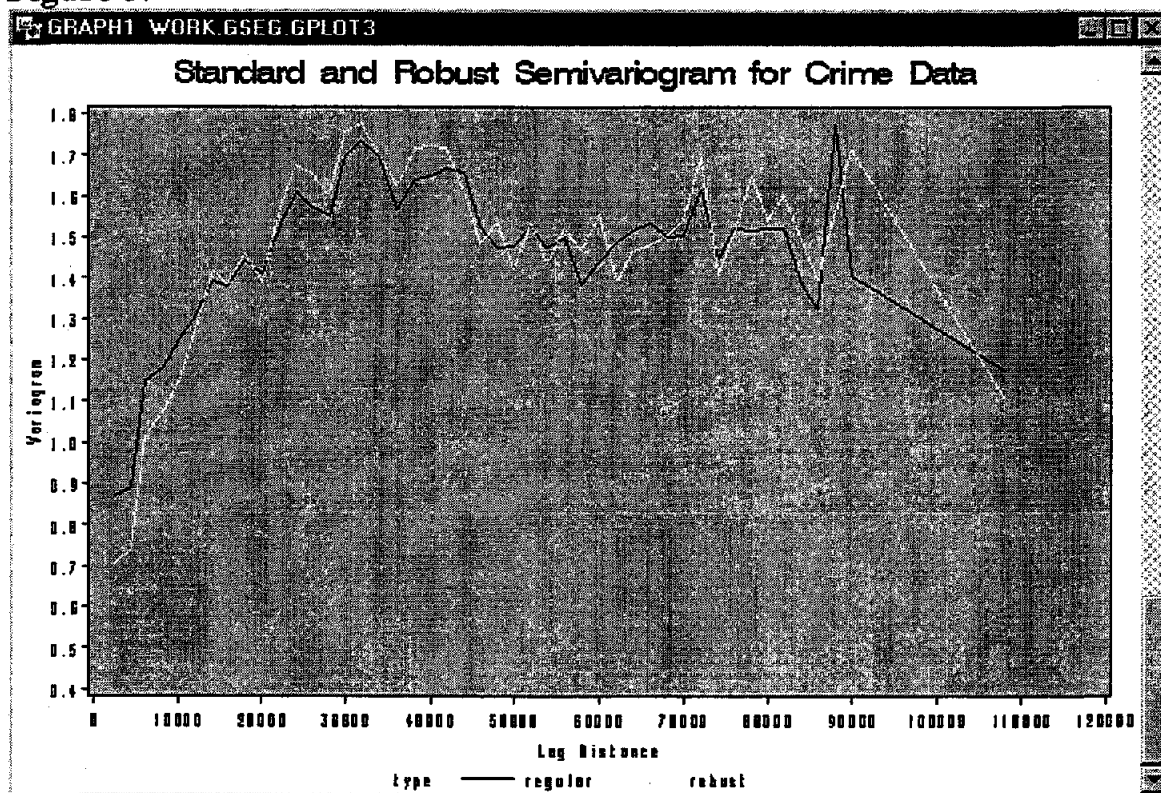
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	267.85439	267.85439	176.21	<.0001
Error	270	410.42417	1.52009		
Corrected Total	271	678.27857			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	0.60138	0.07801	7.71	<.0001
log_PO	1	0.66542	0.05139	12.95	<.0001

For this example, we are concerned about the effect that accounting for spatial autocorrelation in the model has on the parameter estimates. We therefore extract the residuals from this model and examine them for evidence of spatial autocorrelation. We are particularly interested in the structure of autocorrelation present in the data. In order to examine this structure, we estimate a semivariogram for the residuals of the model above using the Semivariogram procedure. The semivariogram for this data is shown below.

Figure 3.



The empirical semivariogram shown above is readily comparable to the semivariogram depicted in Figure 1. The steep rise and then subsequent sharp leveling of the variance across pairwise distance suggests that the data follow a spherical model of spatial correlation. We can also infer a range of plausible spatial covariance parameter values from the semivariogram above. The range of the semivariogram appears to be somewhere between 15,000 and 40,000 ft (approximately 3 to 8 miles), the sill is in the vicinity of 1.3 to 1.7, and the nugget is between .7

and .9. The lack of any significant difference between the robust and standard semivariogram estimates suggests that there are few outliers in the data.

The syntax for producing Figure 2. is shown below:

```
proc variogram data=residuals outv=outv;
compute lagd=2000 maxlag=60 robust;
coordinates xc=x_coord yc=y_coord;
var yresid;
run;

title 'OUTVAR= Data Set Showing Sample Semivariogram Results';
proc print data=outv label;
var lag count distance variog rvario;
run;

data outv2; set outv;
vari=variog; type = 'regular'; output;
vari=rvario; type = 'robust'; output;
run;

title 'Standard and Robust Semivariogram for Crime Data';
proc gplot data=outv2;
plot vari*distance=type / frame cframe=ligr vaxis=axis2
haxis=axis1;
symbol1 i=join l=1 c=blue /* v=star */;
symbol2 i=join l=1 c=yellow /* v=square */;
axis1 minor=none
label=(c=black 'Lag Distance') /* offset=(3,3) */;
axis2 order=(.4 to 1.8 by .1) minor=none
label=(angle=90 rotate=0 c=black 'Semivariogram')
/* offset=(3,3) */;
run;
```

The width of the bins for the empirical semivariogram is specified in the compute statement. The option lagd= specifies the width of each distance bin and the maxlag= option specifies the number of bins created. In practice, these values have to be chosen carefully in order to insure that there are enough pairs (approximately 30) in each distance bin to obtain reliable estimates of the semivariogram. The effect small sample size has on the estimation of the semivariogram can be seen in the values of the semivariogram at large distances. At these distances, the number of available pairs in the data declines precipitously as is evidenced by the erratic behavior of the semivariogram in the right hand side of the graph in Figure 2.

Stage 2. Estimation

Once a plausible model for the spatial correlation structure has been identified using the semivariogram, analysis can proceed to the estimation stage. In this stage, information from the identification stage is used to estimate the spatial correlation model in PROC MIX. In particular, we know from the analysis above that the residuals appear to follow a spherical model of spatial correlation and that parameters of this model fall within a certain range. We then input this information into the mixed procedure in the following way:

```
proc mixed data = disk.a scoring = 50;  
model log_percrime = log_po /solution;  
parms (1.3 to 1.7 by .1) (20000 to 40000 by 1000) (.1 to .8 by .1);  
repeated /subject = intercept local type= sp(sph) (x_coord y_coord);  
run;
```

The type= option in the repeated statement specifies the spherical correlation model for this data. The model statement specifies the fixed effects of the model and, with the exception of the solution option, is identical to the model statement in the previous PROC REG syntax above. The parms statement specifies a range of starting values for the REML procedure. The use of the parms statement is highly recommend by the authors of *SAS System for Mixed Models* in order to get the REML procedure started in a range of plausible values for the sill, range, and nugget (Littell et al., 1996). Otherwise, the REML procedure can converge to local maxima and provide grossly unreasonable estimates. The values chosen in this example correspond our previous interpretation of the semivariogram depicted in figure 2. The output for the spatial correlation model specified above is shown below.

The Mixed Procedure

Model Information

Data Set	DISK.A
Dependent Variable	log_PERCRIME
Covariance Structure	Spatial Spherical
Subject Effect	Intercept
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Between-Within

Dimensions

Covariance Parameters	3
Columns in X	2
Columns in Z	0
Subjects	1
Max Obs Per Subject	272
Observations Used	272
Observations Not Used	17
Total Observations	289

Parameter Search

CovP1	CovP2	CovP3	Variance	Res Log Like	-2 Res Log Like
1.3000	20000	0.1000	0.2304	-418.6647	837.3293
1.3000	20000	0.2000	0.3583	-415.0787	830.1573
1.3000	20000	0.3000	0.4448	-413.1903	826.3806
1.3000	20000	0.4000	0.5091	-412.0828	824.1657
.					
.					
.					
1.6000	40000	0.4000	0.5984	-408.5710	817.1420
1.6000	40000	0.5000	0.6454	-408.0083	816.0166
1.6000	40000	0.6000	0.6829	-407.6588	815.3175
1.6000	40000	0.7000	0.7137	-407.4449	814.8898
1.6000	40000	0.8000	0.7397	-407.3228	814.6456

Iteration History

Iteration	Evaluations	-2 Res Log Like	Criterion
1	2	813.15684179	0.00133169
2	3	813.13212884	0.00046017
3	3	813.12249579	0.00014018
4	3	813.11823442	0.00004362
.			
.			
.			
27	2	813.11185692	0.00000004
28	3	813.11185595	0.00000001
29	3	813.11185553	0.00000000

Convergence criteria met.

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
Variance	Intercept	1.5008
SP(SPH)	Intercept	45336
Residual		0.7610

Fit Statistics

-2 Res Log Likelihood	813.1
AIC (smaller is better)	819.1
AICC (smaller is better)	819.2
BIC (smaller is better)	829.9

PARMS Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
2	1.40	0.4971

The Mixed Procedure

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > t
Intercept	0.1297	0.4365	0	0.30	.
log_PO	0.4439	0.04932	270	9.00	<.0001

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
log_PO	1	270	80.99	<.0001

The spatial covariance parameters can be found directly below the iteration history portion of the output. The first estimate, 1.5008, is the sill; the second estimate, 45336, is the range; and the final estimate is the nugget or residual variation. Comparing these results to the semivariogram depicted in Figure 2, it can be seen that the REML estimates of these parameters differ slightly from what was suggested by the empirical semivariogram. In particular, the sill appears to be a little lower than what is suggested by the empirical semivariogram and the range seems much larger than what was observed. This suggests that the REML estimates may not fit the data particularly well and, perhaps, should be disregarded. The authors of *SAS System for Mixed Models* caution against using REML estimates of spatial covariance parameters when there is reason to suspect that they may not be producing accurate results. The accuracy of REML estimates are especially questionable when applied to large data sets. The reason for this can be seen in the right half of the empirical semivariogram depicted in Figure 2. When estimating spatial covariance, most spatial variability experts discount the information provided by this portion of the semivariogram because it is often estimated with too few pairs of observation for reliable estimation. Instead, estimation techniques are performed on an edited empirical semivariogram with pairs of observation at large distances omitted from the data. Unfortunately, REML estimation uses information equivalent to the full semivariogram, thus

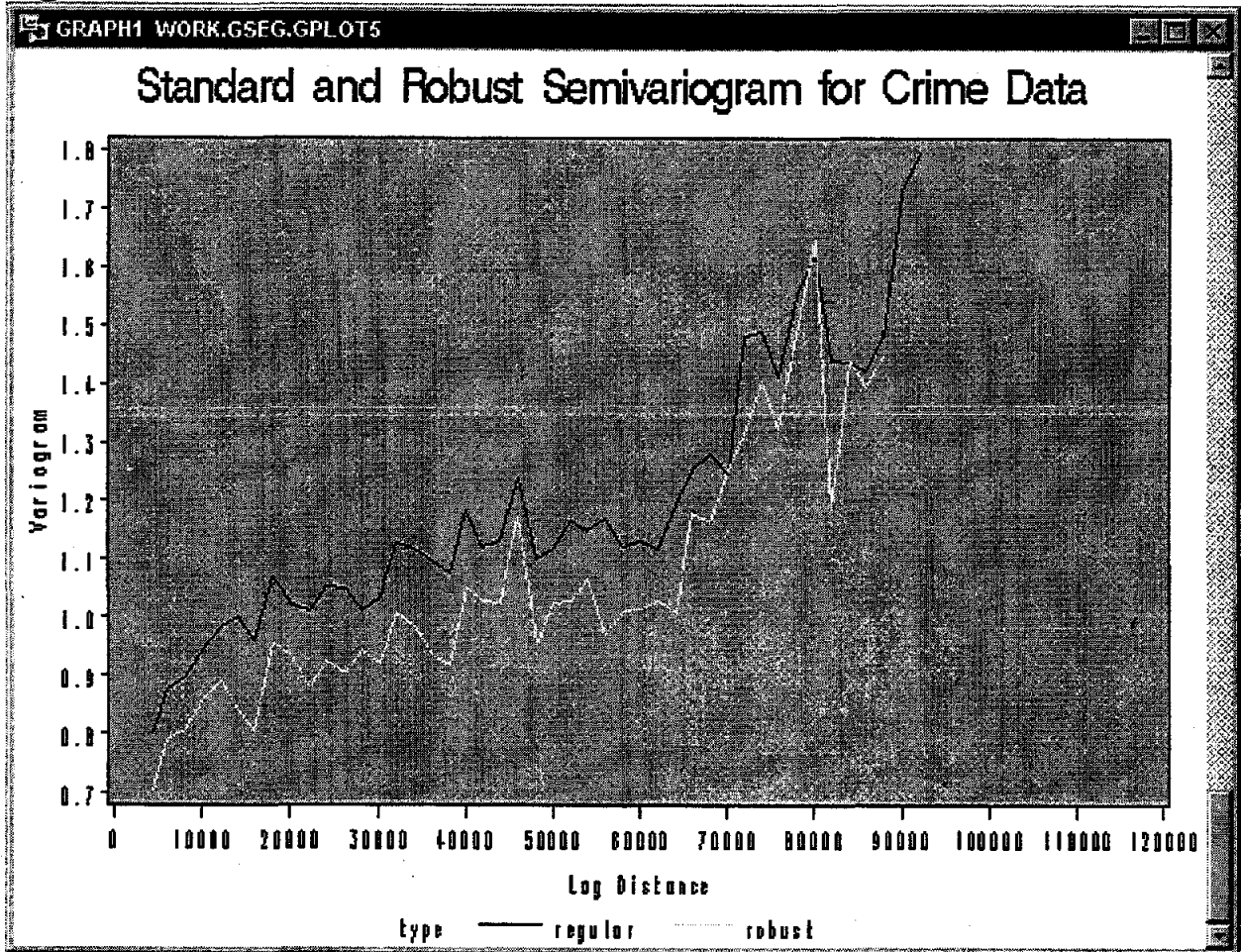
tending to weight distant pairs too heavily. We will examine the adequacy of the current model estimates in the next stage of the analysis.

The results for the fixed effects in the model are shown below the spatial covariance parameters. These results indicate that account for spatial variability in the data reduces the observed association between public order crime and burglary by approximately 33%. Additionally, the F statistic for the single parameter model also changed from 176 in the nonspatial model to 81 in the spatial model. This is a reflection of the loss of information due to spatial autocorrelation discussed in the introduction. One can also compare the mean squared error (MSE) of the non-spatial model to the estimated nugget effect of the spatial model since both of these values represent the level of variation of the independent error term. Accounting for spatial variability in the data reduced this value from 1.52 in the non-spatial model to .76 in the spatial model.

Stage 3: Model Checking

Given the questionable reliability of REML estimates under certain conditions, it is advisable to check that the model estimated fits the data adequately. There are a variety of ways of accomplishing this. The authors of *SAS System for Mixed Models* suggests fitting a semivariogram to the residuals of the model fit by REML in PROC MIXED, and verifying that it appears reasonable. This semivariogram is shown in the figure below.

Figure 4.

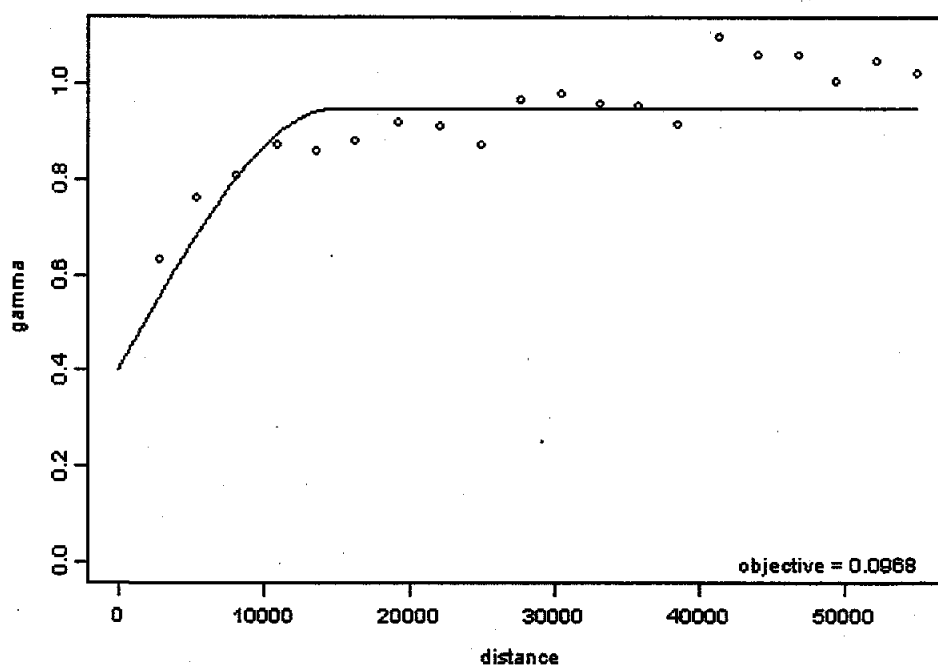


The semivariogram above tell a significantly different story then the semivariogram depicted in Figure 2. This is due to the fact that this semivariogram was estimated on a different set of residuals. A quick glance at the above semivariogram immediately reveals a problem with REML estimates of the spatial covariance parameters. The estimated sill of 1.5 is much to high for this data and this is probably due to the upward spike at the right half of the semivariogram.

If the REML estimates appear to be inaccurate, the solution suggested by Littell et. al. (1996) is to obtain external estimates of the spatial covariance parameters and enter them into the model as known parameter values. SAS does not currently have the usual optimization techniques spatial variability experts use to fit empirical semivariograms. Thus, the user can either obtain "eye ball" estimates from the empirical semivariogram computed using PROC VARIOGRAM, or utilize another software package which have the appropriate procedures.

The figure below was a semivariogram of the same data generated by the software package S+. The procedure used to create the semivariogram allows the user to enter different parameter values for the spatial correlation model interactively and fit an appropriate model visually. In the figure below the range for the spherical model is set to 15000; the sill is 5.5; and the nugget is .4.

Figure 5.



These values can now be entered into the Mixed procedure using the following syntax;

```
proc mixed data = a noprofile;
  model log_percrime = log_po /solution;
  parms (5.5) (15000)(.4) / noiter;
  repeated /subject = intercept local type= sp(sph) (x_coord y_coord);
run;
```

The results for the fixed effects of this model are shown below:

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > t
Intercept	0.2741	0.3700	0	0.74	.
log_PO	0.3948	0.07216	270	5.47	<.0001

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
log_PO	1	270	29.94	<.0001

The new specification of the model resulted in an even lower estimated level of association between public order crime and more serious crime than was obtained from previous models. However, before placing too much confidence in these findings, the fit of this model should also be checked using the method demonstrated above.

Conclusion

Assuming one can obtain reasonably accurate estimates of the spatial covariance parameters, using the Mixed procedure offers a good way of incorporating spatial effects into standard statistical models. However, the lack of reliability of the estimates when using large data sets is somewhat troubling and may limit the use of the Mixed procedure to situations in which the researcher possesses some prior knowledge of the correlation structure in the data. Also troubling is the interaction between the fixed effects in the model and the spatial effects. The interaction between these two components of the model suggests the need to fit spatial models iteratively. One first decides on an initial covariance structure for the data, and then checks the fit of this structure to the data. If the model doesn't fit, then a new covariance structure is proposed based on the residuals of the previous model and the process starts over again until the analysts converge on a solution.

References

Littell, R.C., Milliken, G.A., Stroup, W.W., & Wolfinger, R.D. (1996). SAS system for mixed models. Cary, NC: SAS Institute, Inc.