Paper PO-05-2014

Prediction Improvement from Geostatistical Methodology in Regression Modeling For Census Tract Data

Robert G. Downer, Grand Valley State University, Allendale, MI

ABSTRACT

Ordinary least squares is often sufficient and justified in utilizing a linear regression model for the prediction of new observations. For data which also involves spatial co-ordinates, incorporating spatial correlation through geostatistical techniques can adjust for underlying spatial dependence and may improve prediction. Relative position of locations as well as the selected predictor variables could impact whether such an adjustment leads to prediction improvement. Utilizing a Toledo, Ohio census tract housing dataset, twenty simulations were performed which considered the following factors: variable selection stopping criterion (Mallows Cp or SBC through stepwise selection), test data set size, and fraction of the test points away from the center or northeast corner of the region. Ordinary least squares, residual kriging and spatial error regression were compared with respect to prediction performance. For this data set and simulation scenarios, PROC MIXED showed superior predictive ability in sixteen out of twenty simulations. PROC SURVEYSELECT, PROC GLMSELECT, PROC VARIOGRAM, PROC KRIGE2D and are also utilized within the methodology for this investigation. Some regression methodology is assumed as background for this paper.

INTRODUCTION

In most non-spatial contexts, least squares regression will be a suitable linear modeling method for considering the impact of p predictors on a continuous response y. The coefficients $\beta_1,.....\beta_p$ are estimated by least squares and the errors ε_i are assumed to be independent of each other and normally distributed. With an intercept as part of the model, this gives a prediction equation for observation i

$$\hat{Y}_i = \hat{\beta}_o + \sum_{j=1}^{p} \hat{\beta}_{ji} X_{ji} + \hat{\varepsilon}_i$$

The fitted value $\hat{Y}_i$ is the estimated mean of the response for the combination of covariates $X_1 \ldots X_p$ for observation i. What's remaining in the estimated residual has not been yet accounted for in modeling the mean or trend. If this observation is actually geographically near others in the data set, the underlying independence assumption for these errors may be violated. In other words, these errors may be spatially correlated; their values will be more similar if they are geographically closer together than if they are further apart.

Geostatistical and regionalized methods (Matheron, 1971, Journel and Huijbregts, 1978) were developed under concepts where a spatially referenced variable is random and can be modeled probabilistically. A cornerstone of this modeling is the characterization of the spatial correlation through a variogram. Various types of distance based models are possible; in this paper we are assuming that trend has been removed by a multiple regression. Variogram models are investigated through consideration of correlation among all possible points at inter-point

distances. Incorporating the accepted correlation function into a set of equations leads to a (kriging) solution in which a new unknown can be predicted as a weighted function of its neighbors. When a regression residual is the random variable utilized as input to a set of kriging equations, this methodology is often referred to as residual kriging or ordinary kriging of residuals. In this framework, the residual prediction at a new location q becomes:

$$\hat{Y}_{qrk} = \hat{Y}_q + \sum_{g=1}^{h} w_g \hat{\varepsilon}_g$$

where the residual kriging prediction $\hat{Y}_{qrk}$ has utilized the kriged weighted combination of h available residuals and the estimated expected value $\hat{Y}_q$ from the regression equation and alternative more recent methodology estimates the variogram simultaneously (with regression coefficients) via maximum likelihood or restricted maximum likelihood (REML). This methodology incorporates the correlations into the R matrix of a spatial correlation model. This correlated errors model formulation can be fit through the PROC MIXED in SAS as described in detail in Moser(2004). Other geostatistical methodology and their application using SAS/STAT statistical procedures can be found in SAS papers by Downer (2000), Kolovos(2010), and Massengill (2012).

DATA SET

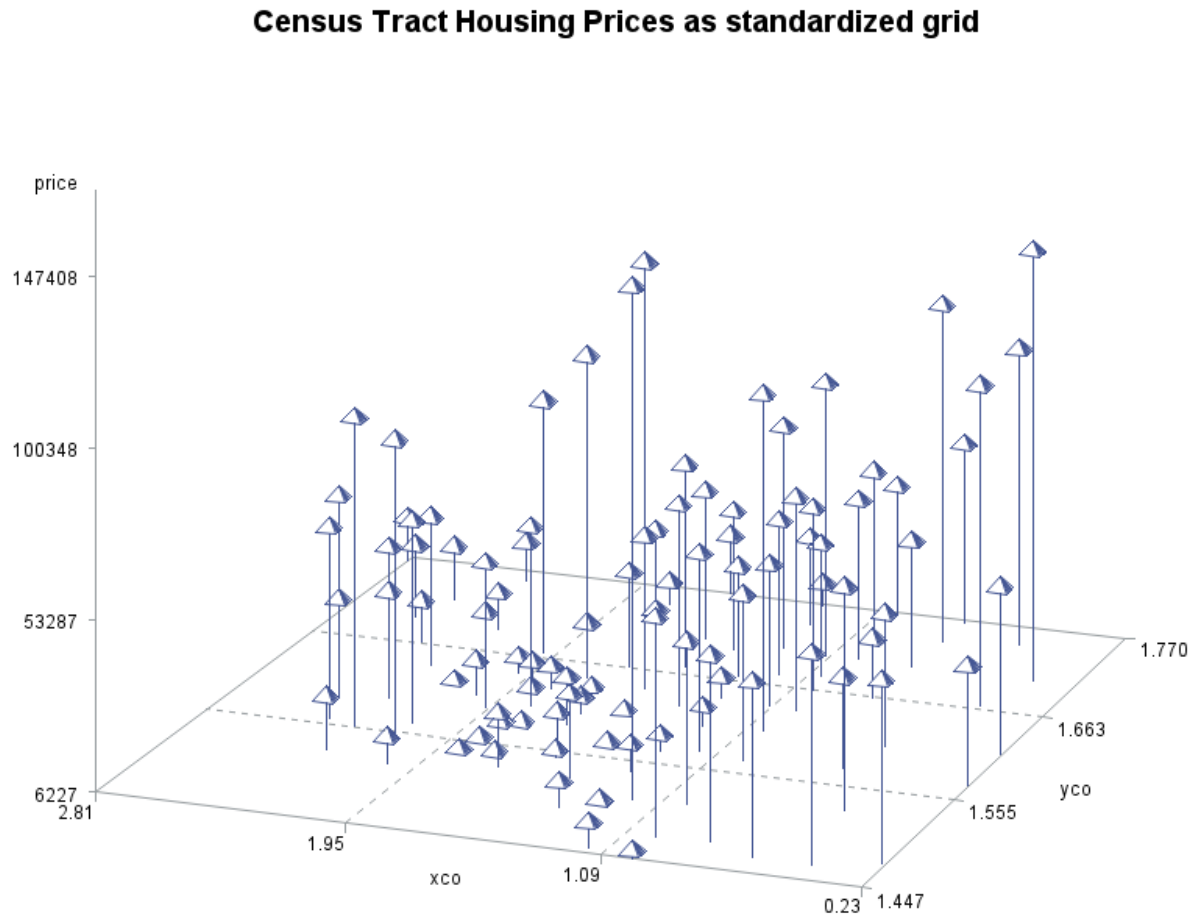
The Toledo census tract data set used in this paper can be found within the example econometrics data sets available at www.spatial-econometrics.com. See Le Sage and Pace (2002) and Le Sage and Pace (2009) for other spatial modeling methodology involving housing and applications at the census tract level. The data set of this paper consists of average housing sale values and average housing characteristics for 98 census tracts in the Toledo area.

The focus of this paper was to compare prediction methods and understand the associated SAS applications. Inference regarding housing price modeling was not emphasized. As a result, only the response (average selling price of houses in the census tract) and five (tract average) predictors variables were included: lot size, living area square footage, number of bathrooms, presence of air conditioning (an indicator variable) and year built (the 'youngest' of which is 1974).One census tract was removed as it was an extreme outlier in all initial regressions involving the full data set. Natural log of housing price was ultimately used as the response in all models to properly satisfy the homogeneity of variance residual assumption. For ease of understanding and display, the six digit original values for the x co-ordinate (longitude) and y co-ordinate (latitude) were standardized by subtracting and dividing by the approximate center of the map for each of the directions.

INITIAL ANALYSIS AND VARIOGRAM INVESTIGATION

Initial exploration of the housing prices geographically showed some pockets of transect price similarity (and possible spatial autocorrelation) but the distance between similar observations appeared to be small. After establishing a mesh of points through PROC G3GRID, PROC G3D gives Figure 1 through a SCATTER statement (default options).

FIGURE 1



PROC GLMSELECT was used fit a model with all 5 predictors and all 97 observations. As discussed in the introduction, predicted values are the estimated mean fitted value for the given set of predictors. The 'de-trended' residuals are put into an output data set REGRESID for exploration of remaining spatial autocorrelation.

PROC VARIOGRAM has a variety of options for exploring the spatial autocorrelation prior to considering variogram models. While utilizing a novariogram option on the compute statement and ODS GRAPHICS on, the following code produces the displays in Figure 2 and Figure 3 below with the regression residuals (regresid) and spatial co-ordinates xco, yco as input information.

```
proc variogram data = regresid  plots=pairs;
compute novariogram ;
coordinates xc = xco yc = yco ;
var res ;
run;
```

FIGURE 2

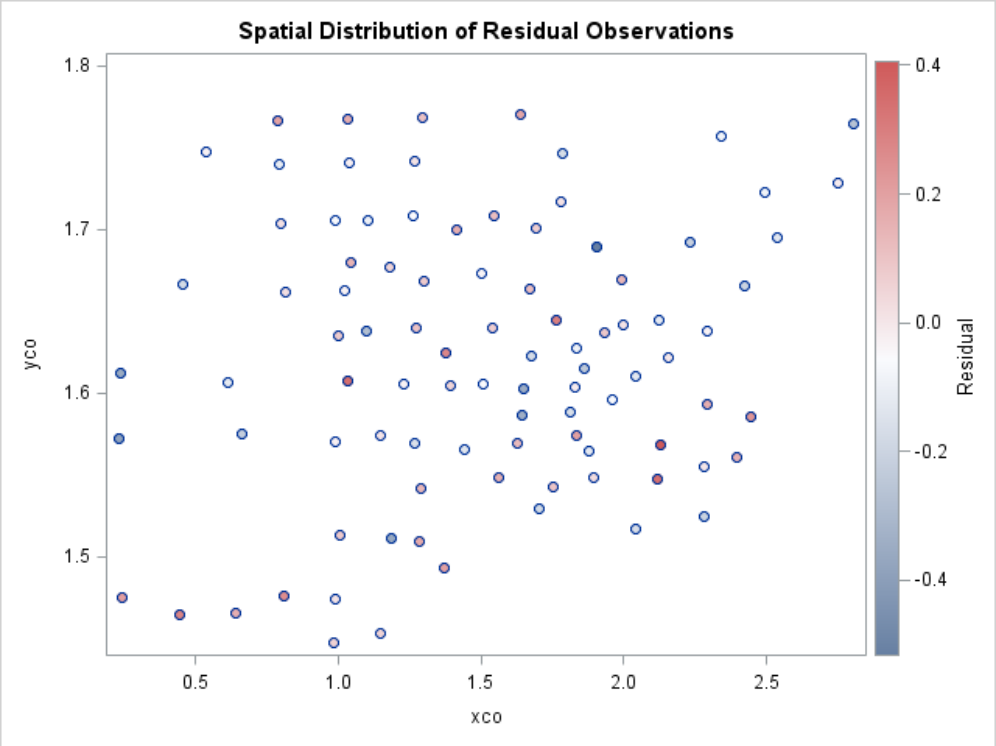
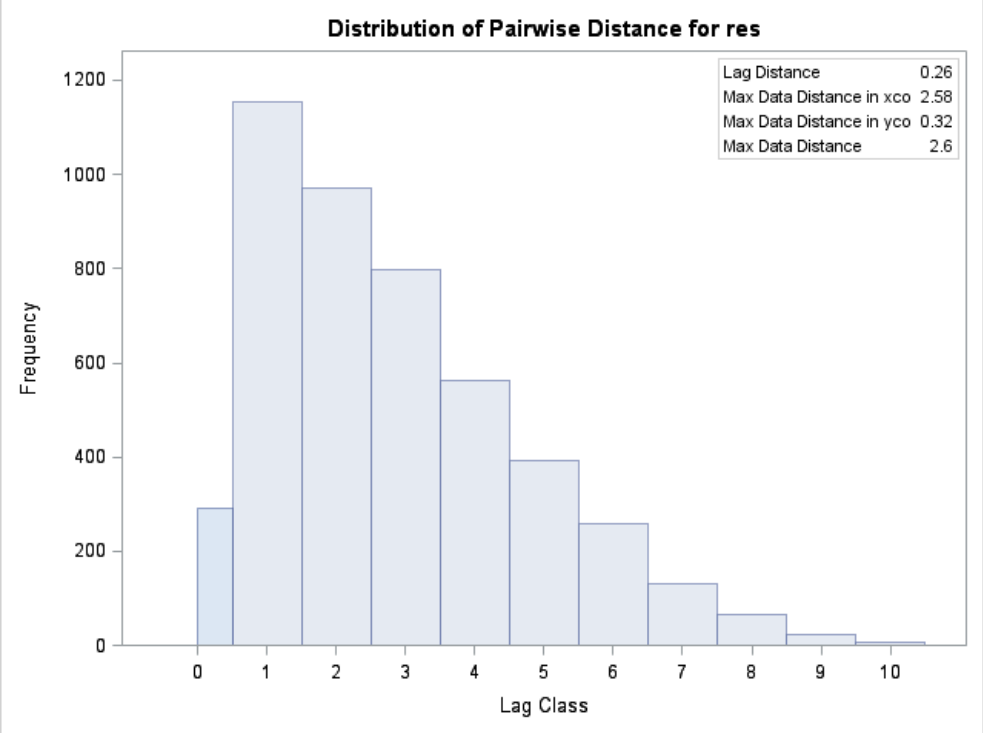


FIGURE 3

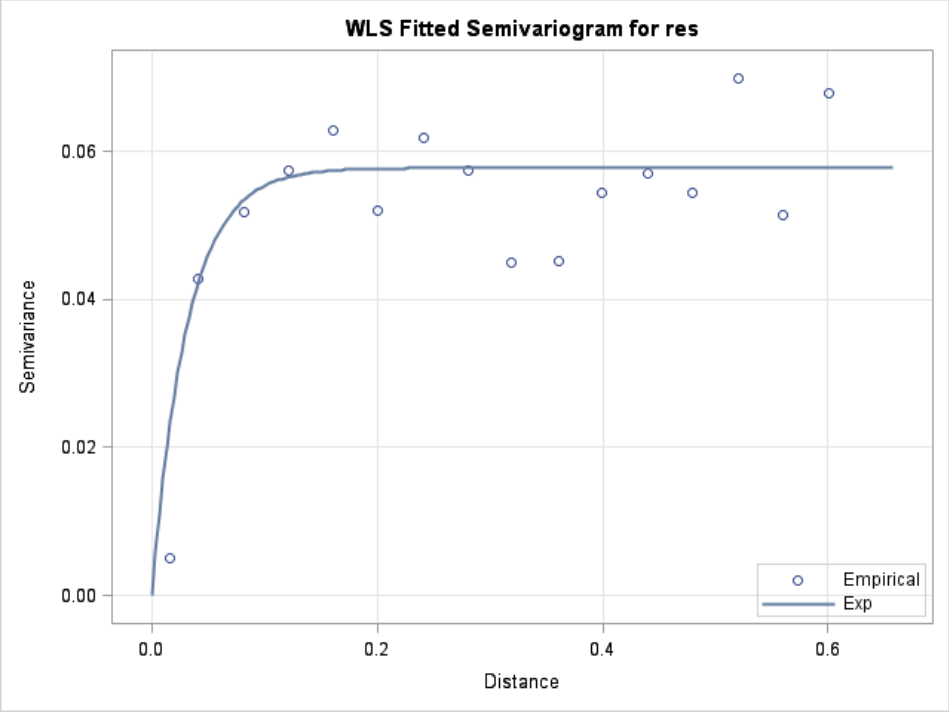


In the simulations to follow, variograms will be estimated for a training set (using either 89 or 93 observations from the total of 97 observations). As can be seen in Figure 3 (and also from further empirical variogram exploration), it was deemed necessary to greatly reduce the lag distance in order to have more pairs of observations at shorter distances. This would allow for efficient and effective variogram modeling for each repetition in the simulation since the range of remaining spatial autocorrelation appeared to be a very small fraction of the grid. In the simulations, variograms would be fit using either 89 or 93 observations from the set of 97. Trial and error with random subsets of the data revealed that lag distances of .03 and a maximum of 20 lags would allow for reasonable variogram estimation in general and an exponential theoretical variogram model would fit well for the remaining training set observations

The following code reflects this framework for estimated variograms and produces Figure 4. The LOWERB and UPPERB options on the parms statement allowed entry of a non-zero lower bound for the estimated range of spatial autocorrelation and an upper bound for the range which seemed appropriate from exploratory work. See Kolovos (2010) for more detail on code related to confidence intervals for estimated variogram points and other theoretical models.

```
proc variogram data= regresids ;
   compute lagd  = 0.03 maxlags = 20;
   coordinates xc=xco yc=yco;
   var res;
    MODEL FORM= exp  ;
      parms (.) (.) (.) / lowerb= (.,.,0.002) upperb = (.,.,0.6) ;
run;
```

Figure 4



SIMULATION SET-UP AND EVALUATION

Two stratification scenarios were created in which there were 83 observations in stratum one and 16 observations in the other. In these scenarios, the test set of 4 or 8 observations would be one quarter or one half of the second stratum, In the first scenario, the 16 observations of stratum two were the 16 observations furthest from the approximate center of the map. In the other scenario, stratum two was defined by the 16 observations closest to the northeast corner of the map. The median selling price of these 16 observations in this northeast corner stratum was approximately twice the median of the rest of the map but yet the range of transect average prices was still over $100,000. It was deemed to be a likely challenge for the methods within the 20 simulations.

Factors

Test set size (2 levels, n=4 or n=8)

Spatial scenario:

- General: test set observations anywhere in the region
- Test set observations all away from center
- 50% of test set observations away from center
- Test set observations all in NE corner
- 50% of test set observations in NE corner

Test set size (2 levels, n=4 or n=8)

Model stopping criterion (2 levels, SBC or Mallows Cp via stepwise selection in GLMSELECT)

Evaluation:

Firstly, a practical ranking of methods was of interest. Overall, which method was most frequently closest to the true price of the test set transects throughout the simulation? Since the same number of predictions was made by each method for a given replication, standardization by the test set sample size or price wasn't necessary. On the natural log scale, with n as either 4 or 8 the prediction sum of squares $\sum_{i=1}^{n}\left((predicted)_i - (trueprice)_i\right)^2$ was calculated for each of 500 repetitions. For a given repetition, the ORDINAL function was then utilized to record which

of the 3 methods was lowest, second and highest for this sum Out of 500 repetitions, the frequency of each method in the high and low position was recorded (using PROC FREQ).

Since the magnitude of prediction error also made sense, the test set sample standard deviation of the absolute errors $|(predicted)_i - (trueprice)_i|$ was also recorded for each repetition. The values and ordinality for the replicate were recorded, as well as the median value of this sample standard deviation over the entire simulation (using PROC MEANS).

RUNNING THE SIMULATIONS

The spatial scenario of a simulation and the training/test set situation of each data set in the simulation itself was quite straightforward using PROC SURVEYSELECT. For example, for modeling with 89 observations and using a test of 8 observations in which 4 of the 8 test set (unused) observations are in the north east corner (while the remaining 4 can be anywhere on the map), the strata statement allows one to specify that the random sample is to be split as 4 and 4.:

```
proc surveyselect method = srs n=(4 4) rep = &numr outall
out = outsel noprint;
strata strne ;
run;
```

The strata variable above *strne* were established as ‘not northeast corner’ and ‘northeast corner’ prior to the run of PROC SURVEYSELECT. .The macro variable *numr* (number of replications) is set as 500 early in the program. The outall option produces an indicator variable defined by selection which is part of the output data. A stacked data set of 500 replicates is created. Missing responses are created for the observations in the test set while the true selling price is known and stored for these observations in each of the 500 replications.

The stacked data set of 500 replications was then sent to PROC GLMSELECT for model selection and formation of residuals by replicate.

```
proc glmselect  ;
model lgpr = lotsize asqft baths airc age /selection=stepwise(choose =
cp)  ;
output out = regresids predicted = yhat residual = res  ;
by replicate ;
ods output selectedeffects = effout;
run;
```

For each replicate, the selected variables from the ods table ‘effout’ are removed as a character string through use of the SUBSTRN function and appended to the string of the previous replicate to create a stacked data set of selected variables (via PROC APPEND).

PROC VARIOGRAM receives the stacked data set of residuals and estimates a set of 3 geostatistical parameters (nugget,sill, range) for an exponential variogram as discussed in the variogram estimation section. The estimates from all replicates are output through the ods table ‘parameterestimates’ and stored in the same (replicate stacked) output data set.

A macro is now called which takes the stacked data sets as input. A loop within the macro indexes from 1 to &numr. For each value of the index, only the index replicate subset is processed.

The variogram estimates, and the residual data set of a given replicate are passed to PROC KRIGE2D. The training part of the replicate is in trainrep and the test set while the locations of the test set are input through testrep in the GRID statement The predicted residuals (for those with missing responses in the test set) are output through the OUTEST option of the KRIGE2D statement and later added back in to the fitted values (to form the predicted log price)

```
proc krige2d data=trainrep outest=predout noprint ;
   coordinates xc=xco yc=yco;
   predict var=res ;
   model scale=&sillrep range=&rngrep nugget = &nuggrep form=exp;
   grid gdata =testrep xcoord = xco ycoord = yco;
run;
```

PROC MIXED fits a spatial correlated errors model simultaneously to the data of a given replicate with log(price) as the response. The variogram estimates for the residuals of this replicate are used only as starting values (with the discussed lower and upper bounds). The predictor variables in &varlist (below) were externally chosen by GLMSELECT for that replicate are utilized by PROC MIXED without any further model reduction. The input data *seluse* is the replicate subset of the raw data with missing values appropriately placed for the test set observations.

```
proc mixed data = seluse   noitprint  ;
model lgpr = &varlist / outp = tol2spat ;
repeated / subject=intercept local  type=sp(exp)(xco yco);
parms / lowerb=.,.,.002 upperb = .,.,0.6 ;
```

For all three methods, the true log(price) values of the observations in the test set are known and hence the prediction error could be found and tabulated for each test set sample and over the entire set of replicates of the simulation.

RESULTS

Based on the ranking of the prediction sum of squares (of the test set over the entire simulation), the spatial correlated errors model performed by PROC MIXED was closest to in 16 of the 20 simulations. The simulation standard deviation of the prediction error was lowest for this method in each of these same 16 simulation simulations.

In these 16 simulations, ordinary least squares performed the worst with respect to how often it was the closest of the 3 methods with prediction sum of squares as the criterion. In these same 16 simulations, ordinary least squares also had the highest median variance in absolute error.

The only spatial test set scenario which gave a different ranking of methods was the scenario in which all 4 or all 8 test set observations were in the northeast corner. Given that there are only 16 observations in this area, it would not make sense to attempt to separately model spatial correlation in this corner. However, the wide range of sale prices and erratic changes in average selling price for neighboring transects would suggest that spatial autocorrelation in this region would be much less than the entire region if not zero. It would also make sense that ordinary least squares would do better than the other 2 methods and this was the case in 3 of the 4 simulations. In the fourth situation, the methods were very similar with each of the three being best according to one criterion.

Model selection stopping criterion had no noticeable effect on the prediction performance of the methods

SUMMARY DISCUSSION

Due to the superior performance of PROC MIXED, this empirical investigation gives strong prediction performance support for simultaneous estimation of geostatistical parameters and regression parameters. However, the spatial scenarios were limited and we're only considering one data set. Further definitive research would investigate many more spatial scenarios, sample sizes and other prediction techniques. Simulating multivariate spatial data sets (perhaps with some incorporation of PROC SIM2D) would allow for more general concrete recommendations.

Non-geostatistical linear model adjustments such as thin-plate splines or geographically weighted regression are other possible approaches for similar data. Non-modeling interpolation techniques such as inverse distance weighting may also perform well in similar situations. These methods have had considerable popularity within geographical research and GIS applications.

Some of the SAS processing logic associated with the linear modeling within this paper could be utilized to compare prediction performance in non-spatial contexts. Model selection criterion was not an emphasis of this paper but could be investigated in much more detail through the many options available in PROC GLMSELECT. If model selection is the emphasis, the variety of cross-validation options available within this procedure may also override the need for a macro to compare techniques. Repeated Sampling through PROC SURVEYSELECT is always a viable option to perform modeling by replicate without unnecessary looping.

REFERENCES

Downer, R.G. (2000) "A Primer in Spatial Simulation with PROC SIM2D" *Proceedings of the SAS Users Group International 2000 Conference*. Cary, NC: SAS Institute Inc.

Journel, A.G. and Huijbregts, C.J (1978), Mining Geostatistics, New York: Academic Press

Kolovos, Alexander (2010) "Everything in Its Place: Efficient Geostatistical Analysis with SAS/STAT Spatial Procedures" *Proceedings of the SAS Global Forum 2010 Conference*. Cary, NC: SAS Institute Inc.

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

LeSage, J.P and Pace, R. K (2002) "Semiparametric maximum likelihood estimates of spatial dependence", *Geographical Analysis*, 34, 76-90.

Massengill, D. (2012) Together at Last: Spatial Analysis and SAS Mapping" *Proceedings of the SAS Global Forum 2012 Conference*. Cary, NC: SAS Institute Inc.

Matheron, G. (1971) The theory of regionalized variables and its applications. Les Cahiers du Centre de Morphologie. *MatheÂmatique de Fontainebleau*. Fontainebleau:CMMF.

Moser, E.B. (2004) “Repeated Measures Modeling With PROC MIXED’. *Proceedings of the SAS Users Group International 2004 Conference*. Cary, NC: SAS Institute Inc.

CONTACT INFORMATION:

Your comments and questions are valued and encouraged. Contact the author at:

Robert G. Downer, PhD.
Biostatistics Director & Professor
Department of Statistics
Grand Valley State University
1 Campus Drive
Allendale, MI 49401
downerr@gvsu.edu