ANALYSIS OF SPATIAL VARIABILITY USING PROC MIXED

David B. Marx

Walter W. Stroup

ANALYSIS OF SPATIAL VARIABILITY USING PROC MIXED

David B. Marx and Walter W. Stroup
Department of Biometry
University of Nebraska-Lincoln
Lincoln, NE 68583-0712

ABSTRACT

Many data sets in agricultural research have spatially correlated observations. Examples include field trials conducted on heterogeneous plots for which blocking is inadequate, soil fertility surveys, ground water resource research, etc. Such data sets may be intended for treatment comparisons or for characterization. In either case, linear models with correlated errors are typically used. Geostatistical models such as those used in "kriging" are often used to estimate the error structure.

SAS PROC MIXED allows the estimation of the parameters of mixed linear models with correlated errors. Fixed and random effects are estimated by generalized least squares. Variance and covariance components are estimated by restricted maximum likelihood (REML).

The purpose of this presentation is to show how PROC MIXED can be used to work with spatial data. Several examples will be presented to illustrate how various analyses could be approached and some of the pitfalls users may encounter.

1. INTRODUCTION

Statistical methods traditionally used in agricultural research have emphasized designs and analyses which assume that variation among experimental units is either (i) homogeneous or (ii) can be controlled by blocking. In many field situations, however, variation is more likely to be characterized by smooth, localized, irregular trends - variation which is neither homogeneous nor necessarily well-controlled by blocking.

Figure 1 illustrates the distinction among these three general classes of variability. Rows and columns represent field plots in a rectangular arrangement and the y-variable is some response of interest. The first case represents variation across the field when we have "homogeneous" experimental units, the nominal situation for which the completely random design its associated analysis of variance are appropriate. These data were produced by a normal random number generator. The second case represents experimental units with obvious homogeneous subsets - an idealized case for the use of a blocked design and its associated analysis of variance. The third case is a visual characterization of spatial variability - the variation obviously has pattern, but the pattern, while smooth, is irregular and an implementable criterion for blocking is not obvious.

In the third case, the pattern of variability can often be characterized by a linear model with spatially correlated errors. Mixed linear model methods (Henderson, 1975; Harville, 1976, 1977; McLean, Sanders, and Stroup, 1991) are therefore useful tools to analyze such data.

The general form of the mixed linear model is as follows.

$$\mathbf{y} = X\beta + Z\mathbf{u} + \mathbf{e}, \qquad [1]$$

where $\mathbf{y}$ is a vector of observations;
X is a matrix of constants (describing regression or design structure) for the fixed effects;

ß is a vector of fixed effects parameters;
Z is a matrix of constants for the random effects;
u is a vector of random effects; and
e is a vector of residuals.

For the random effects, **u** and **e**, assume E(**u**) = E(**e**) = **0**, Var(**u**) = G, Var(**e**) = R, and Cov(**u**,**e**') = 0. In "traditional" models, e.g. standard analysis of variance models for completely random and randomized block designs, $R = I\sigma^2$ is assumed, but mixed model theory places no requirements on R or G - both can be general.

Inference with mixed linear model has three basic building blocks. The first is the *mixed model equation*, used to estimate ß and **u**, given as follows:

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix} \begin{bmatrix} ß \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} X'R^{-1}\mathbf{y} \\ Z'R^{-1}\mathbf{y} \end{bmatrix} \qquad [2]$$

In most cases, the variance and covariance components of G and R are unknown, and estimates must be used. Estimates are typically obtained using restricted maximum likelihood (REML), although other methods can be used.

The second building block of inference is the *predictable function*, K'ß + M'**u**, which is *predictable* is K'ß is *estimable*. Adjusted marginal treatment means (a.k.a. "Least Squares means"), treatment differences, and contrasts are typical predictable functions of interest to researchers. The marginal means or treatment comparisons of interest guide the choice of K. In most applications, M will be a matrix of zeros, corresponding to what McLean, et. al. (1991) call the "broad inference space." Other M matrices may be selected to restrict the inference space or to obtain *best linear unbiased predictors* (BLUP). See McLean, et. al. (1991) for further detail.

The third building block of interest is the "standard error," or, more precisely, the square root of the prediction error of the estimated predictable function. The standard error is given by the formula $\sqrt{L'CL}$, where L' = [K' M'] and C is the generalized inverse of

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix}$$

When L is a vector - e.g. when a single marginal mean, treatment contrast, or BLUP is of interest - the ratio estimate/(standard error) is a t-statistic and be used as such. In the more general case, where L is a matrix - e.g. when multiple degree of freedom hypotheses, such H_0: all treatment effects equal, are of interest,

$$\mathbf{\theta}'L(L'CL)^{-1}L'\mathbf{\theta}/\text{rank}(L), \text{ where } \mathbf{\theta}' = [ß'\ \mathbf{u}']$$

is an approximate F-statistic. The numerator degrees of freedom equal rank(L). The denominator degrees of freedom are more complicated. A naive approach is to use the degrees of freedom implied by the appropriate error term in a analysis of variance table. In more complex models, e.g. for which the implied error from the ANOVA term involves more than one mean square, or the errors are correlated, Satterthwaite's approximation, or other alternatives, may be used. Jeske and Harville (1988) consider this issue in some detail, and the interested reader is referred to their article. PROC

MIXED (SAS Institute, 1992) uses the naive, ANOVA-analog approach to determine error degrees of freedom.

Several spatial correlation models are potentially useful for agricultural data. Typically, spatial correlation refers to variability among experimental units in a single location (e.g. among plots in a field) and is thus modeled through the covariance of the residual vector, **e**, that is, the matrix R. Zimmerman and Harville (1991) discuss several alternative structures for R. Many of these were originally developed for applications in geostatistics (Journel & Huijbregts, 1978). The basic idea is as follows.

In a typical field trial with spatial variability, responses of plots close together are highly correlated, whereas plots farther apart are less correlated. At some critical distance, responses of plots that distance or farther apart are essentially uncorrelated. In geostatistics, the *semivariogram* is used to characterize spatial variability. The *semivariance* defined as

$\Gamma(h) = \frac{1}{2}\text{Var}(\text{difference between pairs of observations } h \text{ units apart})$

The semivariogram is a plot of $\Gamma(h)$ versus h. A typical semivariogram is given in Figure 2, below. The key features of the semivariogram are the *range*, defined as the critical distance above which observations are uncorrelated, the *sill*, the semivariance of uncorrelated observations (equal to the error variance, it can be shown), and the *nugget*, defined as the semivariance at distance zero. The nugget describes abrupt changes and was originally intended to model data from searches for diamonds, where probes a very short distance apart could find either nothing or a very high concentration of diamonds. Such abrupt variation is uncommon in agricultural field trials, so the nugget is frequently assumed to be zero.

The semivariance is related to the R matrix in the mixed model as follows:

$\text{Cov}(2 \text{ observations } h \text{ units apart}) = C(h) = C(0) - \Gamma(h).$

Thus, C(0) corresponds to the diagonal elements of R and the C(h), where h>0, are the off-diagonal elements of R. Typical semivariance models for the mixed model are

Spherical

$$C(h) = \sigma^2[1 - (3h/2r) + (h^3/2r^3)], \quad \text{if } h < r$$
$$= 0, \quad \text{otherwise}$$

Exponential

$$C(h) = \sigma^2[\exp(-h/r)]$$

Gaussian

$$C(h) = \sigma^2[\exp(-h^2/r^2)]$$

Linear

$$C(h) = \sigma^2[1-hr], \quad \text{if } h < 2/r$$
$$=0, \quad \text{otherwise}$$

For all of the above models, two parameters, σ^2 and r, corresponding to the error variance and range, respectively, must be estimated.

Each of the above semivariance models describes a different pattern of

correlation among neighboring experimental units. In the examples given in sections 3 and 4, we will show how the problem of selecting an appropriate semivariance can be approached. The important concept to keep in mind is that the use of linear models with correlated errors to analyze field data is strongly indicated when irregular, local gradients, as portrayed in Figure 1, case 3, are present. The correlation model is basically a model of the "surface" resulting from such variability.

PROC MIXED (SAS Institute, 1992) permits the user to specify mixed models whose errors are correlated. Error correlation models include all of the semivariance models given above (spherical, gaussian, exponential, and linear) as well as other correlation models not as commonly of interest in agricultural field trials. **ß** and **u** are estimated using the mixed model equations [2], and σ^2 and **r** are estimated using REML. In the following sections, we will present examples of basic applications of mixed models with spatial correlation. We will present the basic PROC MIXED programming requirements, highlights of the output of particular interest, and our experiences with problems and pitfalls users should anticipate.

2. BASIC *PROC MIXED* PROGRAMMING AND OUTPUT

Consider the simplest mixed model with spatially correlated errors,

$$y_{ij} = \mu + f_i + r_j + e_{ij},$$

where μ and the f_i's are fixed, the vector **u** of r_j's is distributed $N(\mathbf{0}, I\sigma_R^2)$, and the vector **e** of e_{ij}'s is distributed $N(\mathbf{0}, R)$, where R is some spatial covariance matrix. To analyze data using this model using PROC MIXED, the following input statements are minimally required:

```
DATA a;
INPUT fix_eff rand_eff row col y;
```

In describing program statements, we will refer to words that are mandatory *verbatim* in the program using capital letters and words that are mandatory but the specific word is user's choice using lower case. The variables "fix_eff" and "rand_eff" name the fixed and random factors in the model, "row" and "col" locate the observation in space, and "y" is the observed response. The number and specific names of the fixed and random factors will depend on the particular data set. Some data sets may have none, e.g. if estimating the semivariogram is the only objective (see section 3). Others may have only a fixed effect, e.g. treatment. Others may have several fixed and random effects, e.g. factorial treatments designs (factor A, B, etc.), and blocks, locations, etc. *ALL* data sets must have a "col" and "row" variable, corresponding, for example, to the longitudinal or east-west and the latitudinal, or north-south location of the observation, respectively.

The following is the basic SAS program. Variables in *italics* are not mandatory, but we have found them to be useful options.

```
PROC MIXED SCORING=n;
CLASS fix_eff rand_eff;
MODEL y=fix_eff;
PARMS (σR²) (nugget) (range) (sill);
RANDOM rand_eff;
REPEATED / SUBJECT=rand_eff LOCAL TYPE=SP(SPH) (row col);
```

The *SCORING* option forces the REML algorithm to use a scoring procedure at least *n* times per iteration. We have found this option to be very helpful in obtaining convergence to reasonable solutions for the range and sill. The values in parenthesis in the PARMS statement are initial numeric values for the variance and covariance parameters. Typically, the nugget is assumed to

be zero, so the *nugget* option often will not be used in the PARMS statement. σ_R^2 will only be specified is one wishes to include the random effect in the model (which, note, is included in a separate RANDOM statement, not in the MODEL statement *a la* SAS-GLM). Our experience is that it is essential to specify initial estimates of the range and sill; PROC MIXED's default initial values frequently lead to grossly unreasonable estimates of the sill and range.

REPEATED specifies the structure of the covariance matrix, R, of the **e** vector. If there is a random effect in the model, use SUBJECT=rand_eff; otherwise use SUBJECT=INTERCEPT. *LOCAL* is used if the nugget is assumed to be non-zero. TYPE specifies the covariance (or semivariance) model to be estimated. The example above is for the spherical semivariance model. Other options include EXP (exponential), GAU (gaussian), and LIN (linear). Consult the SAS manual for other options.

3. EXAMPLE 1 - NO TREATMENT EFFECTS, NUGGET ZERO

In this example, we present the SAS statements required to estimate the sill and range of a rectangular array of spatially correlated data. The data are given in Table 1. Please note that only the location (LAT and LNG, i.e. "row" and "col") and response (X) variables from Table 1 are used in this example. The model is

$y_i = \mu + e_i$, where **e** is distributed N(**0**,R).

The program:

```
DATA A;
 INFILE 'KSUTALK DATA A';
 INPUT REP BLOC TRT LAT LNG Y X;
PROC MIXED SCORING=5;
 MODEL X= ;
 PARMS (5.0) (4.0);
 REPEATED / SUBJECT=INTERCEPT TYPE=SP(SPH) (LAT LNG);
 TITLE 'SPHERICAL COVARIANCE MODEL';
```

The PARMS statement contains initial estimates of the range and sill. No value for σ_R^2 is included (since there is no RANDOM statement) and no value for *nugget* is included (i.e. we assume it is zero). Selected output:

SPHERICAL COVARIANCE MODEL

Parameter Search

COL1	COL2	Variance	REML_LL	-2REML_LL	Objective
5.0000	4.0000	2.9556	-125.342	250.6835	134.8973

REML Estimation Iteration History

Iteration	Evaluations	Objective	Criterion
1	2	112.71950815	0.00280361
.			
.			
.			
8	1	111.71705126	0.00000000

Scoring stopped after iteration 5.

Convergence criteria met.

Covariance Parameter Estimates (REML)

Cov Parm	Ratio	Estimate	Std Error	Z	Pr > |Z|
DIAG SP(SPH)	2.71198105	8.84278813	2.77131363	3.19	0.0014
Residual	1.00000000	3.26063787	0.65576061	4.97	0.0000

Model Fitting Information for X

Description	Value
Observations	64.0000
Variance Estimate	3.2606
Standard Deviation Estimate	1.8057
REML Log Likelihood	-113.752
Akaike's Information Criterion	-115.752
Schwarz's Bayesian Criterion	-117.895
-2 REML Log Likelihood	227.5033
PARMS Model LRT Chi-Square	23.1802
PARMS Model LRT DF	1.0000
PARMS Model LRT P-Value	0.0000

The REML iteration history tracks the progress of the sill and range estimates. The important line is "convergence criteria met." The range and sill estimates are, respectively, 2.712 and 3.261 (rounded to 3 decimal places). They are given under the "ratio" for DIAG SP(SPH), and the "estimate" for "residual," respectively. This output was produced by Version 6.07 of SAS. Later versions have the actual estimate of the range under the "estimate" column; thus a Version 6.08 PROC MIXED output would have 0.832 and 1.000, respectively in the "ratio" column and 2.712 and 3.261 respectively in the "estimate" column.

The model fitting information can be useful in comparing plausible models. The REML log likelihood, and two related criteria, Akaike's Information and Schwarz's Bayesian, which are adjusted for various model characteristics, can be used for likelihood ratio tests when comparing models which are subsets of one another, or simply interpreted as "the higher, the better" for different models with the same number of parameters. These data were also fit using the *exponential*, *gaussian*, and *linear* semivariance models. The linear produced a warning "Scoring did not stop. Stopped because of infinite likelihood," which is typical - the *linear* semivariance model works poorly with the mixed model and REML. For the others, the model fitting output was

EXPONENTIAL COVARIANCE MODEL

```
REPEATED / SUBJECT=INTERCEPT TYPE=SP(EXP)(LAT LNG);
```

Akaike's Information Criterion	-118.287
Schwarz's Bayesian Criterion	-120.430
-2 REML Log Likelihood	232.5740

GAUSSIAN COVARIANCE MODEL

```
REPEATED / SUBJECT=INTERCEPT TYPE=SP(GAU)(LAT LNG);
```

Akaike's Information Criterion	-118.056
Schwarz's Bayesian Criterion	-120.199
-2 REML Log Likelihood	232.1112

Thus, the *spherical* model would be preferred, because the Akaike and Schwarz criteria are higher and the -2 REML Log Likelihood is lower.

The "PARMS model LRT" (likelihood ratio test) tests H_0: range=0 if NO parms option is used, but it test the hypothesis that the difference between the initial value and the estimated value is zero when the PARMS statement is used. This is not a particularly useful test.

4. EXAMPLE 2 - COMPLETELY RANDOM DESIGN WITH CORRELATED ERRORS

The second example contains 64 observations on 16 treatments, laid out in an 8 x 8 array of field plots. Data appear in Table 1. The 64 observations were subdivided into four squares, corresponding to replications of a 4 x 4 lattice design, but, for now, we will analyze them according to the model

$y_{ij} = \mu + \tau_i + e_{ij}$, where **e** is distributed N(**0**,R).

The SAS program:

```
DATA A;
 INFILE 'KSUTALK2 DATA A';
 INPUT REP BLOC TRT LAT LNG Y X;
PROC MIXED SCORING=5;
 CLASS TRT;
 MODEL Y=TRT;
 PARMS (5.00) (4);
 REPEATED / SUBJECT=INTERCEPT TYPE=SP(SPH)(LAT LNG);
 LSMEANS TRT;
 ESTIMATE 'T1 VS T2' TRT 1 -1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0;
 ESTIMATE 'T1 VS T3' TRT 1 0 -1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0;
 ESTIMATE 'T1 VS T8' TRT 1 0 0 0 0 0 0 -1 0 0 0 0 0 0 0 0 0 0;
 CONTRAST 'T1 VS T2' TRT 1 -1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0;
 CONTRAST 'T1 VS T3' TRT 1 0 -1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0;
 CONTRAST 'T1 VS T8' TRT 1 0 0 0 0 0 0 -1 0 0 0 0 0 0 0 0 0 0;
 TITLE 'SPHERICAL COVARIANCE MODEL - TRT EFFECT IN MODEL';
```

Selected output:

SPHERICAL COVARIANCE MODEL - TRT EFFECT IN MODEL

Covariance Parameter Estimates (REML)

Cov Parm	Ratio	Estimate	Std Error	Z	Pr > |Z|
DIAG SP(SPH)	3.13277873	11.17435804	3.25978411	3.43	0.0006
Residual	1.00000000	3.56691583	0.77855187	4.58	0.0000

Tests of Fixed Effects

Source	NDF	DDF	Type III F	Pr > F
TRT	15	48	12.96	0.0000

ESTIMATE Statement Results

Parameter	Estimate	Std Error	DDF	T	Pr > |T|
T1 VS T2	-0.03705715	0.78250253	48	-0.05	0.9624
T1 VS T3	0.81100587	0.80309159	48	1.01	0.3176
T1 VS T8	3.69534390	0.81410883	48	4.54	0.0000

CONTRAST Statement Results

Source	NDF	DDF	F	Pr > F
T1 VS T2	1	48	0.00	0.9624
T1 VS T3	1	48	1.02	0.3176
T1 VS T8	1	48	20.60	0.0000

Least Squares Means

Level	LSMEAN	Std Error	DDF	T	Pr > |T|
TRT 1	12.49765196	0.74954746	48	16.67	0.0000
TRT 2	12.53470911	0.75986106	48	16.50	0.0000
TRT 3	11.68664609	0.76719779	48	15.23	0.0000

The data are actually the same as example 1 except that treatment effects were added according to the lattice design. The estimates of range and sill are 3.133 and 3.567, different, but not substantially so, from the previous example. Note that PROC MIXED only gives F-values rather than a full ANOVA (sums of squares and mean squares do not have any conventionally useful meaning in mixed models). Also the standard errors for least squares means and differences are not the same. Although the number of observations per treatment are equal, the observations lie at different points on the local gradients resulting from the spatial correlation and at different average distances from one another. Assuming the spatial correlation is accurately estimated, the variability estimates for means and differences are thus more realistic.

These data could alternatively be analyzed using conventional ANOVA models. In particular, the traditional RCBD model or the Lattice model could be evaluated. Selected results are as follows:

RCBD ANOVA

```
CLASS REP TRT;
MODEL Y=TRT;
RANDOM REP;
```

model $y_{ij} = \mu + \tau_i + r_j + e_{ij}$

$\mathbf{r} \sim N(0, I\sigma_R^2)$, $\mathbf{e} \sim N(0, I\sigma^2)$

Covariance Parameter Estimates (REML)

Cov Parm	Ratio	Estimate	Std Error	Z	Pr > |Z|
REP	0.00000000	0.00000000	.	.	.
Residual	1.00000000	3.05074506	0.62273073	4.90	0.0000

LATTICE ANOVA

```
CLASS REP BLOC TRT;
MODEL Y= TRT;
* RANDOM REP BLOC(REP);
RANDOM BLOC;
```

model $y_{ijk} = \mu + \tau_i + r_j + b(r)_{jk} + e_{ijk}$

assumptions as RCBD + $\mathbf{b(r)} \sim N(0, I\sigma_{BR}^2)$

Covariance Parameter Estimates (REML)

Cov Parm	Ratio	Estimate	Std Error	Z	Pr > |Z|
BLOC	0.73832005	1.31435334	0.70964068	1.85	0.0640
Residual	1.00000000	1.78019457	0.44073910	4.04	0.0001

SPHERICAL COVARIANCE MODEL - TRT EFFECT IN MODEL

Akaike's Information Criterion	-99.9192
Schwarz's Bayesian Criterion	-101.790
-2 REML Log Likelihood	195.8385

RCBD ANOVA

Akaike's Information Criterion	-107.969
Schwarz's Bayesian Criterion	-109.840
-2 REML Log Likelihood	211.9373

LATTICE ANOVA

Akaike's Information Criterion	-104.109
Schwarz's Bayesian Criterion	-105.981
-2 REML Log Likelihood	204.2189

The model fitting criteria indicate that the RCBD model is inferior to the alternatives. The Lattice model is an improvement over the RCBD, but is not as good as the spatial covariance model. This result is common in the analysis of field data (see also Stroup, Baenziger, and Mulitze, 1994). Complete block designs should be avoided when spatial variability is suspected. Incomplete block designs are often only partial fixes. We now turn our attention to some pitfalls we have encountered using PROC MIXED to analyze data with spatial variability.

5. HOW TO USE PROC MIXED IN DESIGNED EXPERIMENTS

In using PROC MIXED it is important to realize that the definitions of parameters has changed between versions. For example, in version 6.07 the parameter DIAG SP(SPH) estimates the range times the sill, and consequentially the ratio of the DIAG SP(SPH) to the residual estimates the range. In version 6.08 the parameter estimate of DIAG SP(SPH) is itself the estimate of the range. When assuming a spatial structure, say spherical, without a nugget effect only one parameter, the range, will effect the value of the log likelihood. Since SAS uses an iterative procedure, it is important to grid the parameter space for initial evaluations of the likelihood function so that local minimums can be identified. A typical plot of the range against the -2 REML log likelihood looks like figure 3. If the range were assumed larger than six then the procedure would converge to a local minimum.

Often the procedure will not converge. If this is the case using SCORING = 10 will often help. Another practice which should be followed when using PROC MIXED in spatially correlated models is to graph the empirical semivariogram with the modeled semivariogram from PROC MIXED. A simple procedure to do this is to use the fixed effects derived from PROC MIXED and calculate the residuals. Use these residuals in a geostatistical package, such as GEOEAS, and have that package automatically obtain the empirical semivariogram and then input the parameters of the modeled semivariogram from PROC MIXED. A caution is noted here: sometimes GEOEAS will not use the larger lag distances and hence the resulting graph may be misleading. In figure 4 the maximum lag distance was 5 and it looks as if PROC MIXED has underestimated the sill. However, if all the lags are used, the modeled semivariogram looks much better as in figure 5.

PROC MIXED can be used to help determine which spatial model is appropriate for a set of data. This can be done by first assuming a particular spatial model, say spherical, and obtaining the model fitting

criteria (Akaike's, Schwarz's, and -2 REML LL), then repeating the process for several other models, say gaussian and exponential. The fitted models can be graphed and the fitting criteria compared. For example, in figure 6 the graph indicates that the spherical fitted model looks better than either the exponential or gaussian. However, even though the sill seems a little too high for the spherical, the gaussian clearly underestimates the sill and the exponential overestimates it.

Finally, the likelihood is not appropriately calculated for some starting values in the gaussian model. In figure 7, at large ranges, the value for -2 REML LL will abruptly change between a reasonable value (around 1200) and an unreasonable large number (say 1.789E308). Hence there are a number of local minimums as seen in the gaussian graph in figure 7. For this experimental data set, the three different models used provided the following statistics (Table 2):

Table 2. Statistics associated with one simulation run (#1) corresponding to Figure 7.

Model	Range	Sill	Akaike's	Schwarz's	-2 REML LL
Spherical	2.916	5.161	-128.22	-130.36	252.44
Exponential	1.895	6.054	-129.25	-131.39	254.49
Gaussian	1.005	4.530	-131.22	-133.36	255.44

These three models are graphed in figure 8, and as one can see the models are all very similar. Again it is important to look at the graphs as well as the fitting criteria since the best fit model may be totally inadequate.

A simple simulation study was run using 100 simulations of 64 spatially correlated observations each in an 8 by 8 grid. For each run the original spatially correlated data had treatment effects added to the observations for the 16 treatments. The treatment effects were 13, 13, 14, 14, 15, 15, 16, 16, 16, 16, 17, 17, 18, 18, 19, and 19. The spatial model used was spherical with no nugget effect, a range of 3.50 and a sill of 4.00. First we tried to determine if PROC MIXED would be able to correctly handle the treatment effects. The results indicted that PROC MIXED would estimate the fixed effects very accurately in that the average of the simulation treatment effects were very close to the true treatment effect and had little variability from run to run. As an example the true treatment effect for treatments 1 and 2 were both 13. The average for the 100 simulation effects given by PROC MIXED were 12.94 and 13.03 respectively with standard deviations of 0.77 and 0.76. The results of the simulation indicated that PROC MIXED does a fairly good job of estimating the sill as seen in Table 3. However, PROC MIXED also tends to underestimate the range, although with only 100 simulations the underestimation is not statistically significant ($p>.05$). The true spatial structure was spherical and the model assumed by PROC MIXED was spherical as well. The results are:

Table 3. Comparison of Simulation Data 100 Simulations of 64 Observations

	parameter	mean	maximum	minimum	std dev
No treatment Effect	range	3.08	5.30	1.55	0.63
	sill	4.03	7.35	2.09	0.98
Treatment Effect	range	3.09	6.26	1.68	0.75
	sill	3.99	8.21	1.94	1.07
Difference	range	-0.01*	1.55	-2.60	0.50
	sill	0.05*	2.64	-3.56	0.72

* not significant (p>.05)

One common (Cressie, 1991) method of removing "drift" of fixed effects from spatially correlated data is by the use of median polish. The original data are "polished" and the residuals from the polished data are then assumed to be free of drift and can then be subjected to the usual geostatistical estimation procedures. Median polish is preferred to mean polishing in that it is suspected that mean polishing also polishes out some of the spatial correlation. Thus the simulated data were polished to remove drift (here a linear drift) and drift was removed by PROC MIXED. These were compared to the original data before the drift was added in Table 4. Both the range and the sill are significantly (p<.05) for the polished data.

Table 4. Simulated Results for Polished Data

	mean	maximum	minimum	std dev
No Drift				
range	3.08	5.30	1.55	0.63
sill	4.03	7.35	2.09	0.98
Drift Removed by PROC MIXED				
range	3.09	5.33	1.28	0.77
sill	4.06	7.75	1.70	1.22
Drift Removed by Polishing				
range	2.91	6.13	1.09	1.01
sill	3.49	7.24	0.98	1.41
Difference in PROC MIXED and Polishing				
range	0.18*	3.30	-3.44	0.88
sill	0.57*	4.22	-2.69	1.12

* both range and sill are significantly smaller (p < .05) for polished data

To determine how well the exponential and gaussian models fit the spherical data PROC MIXED was used assuming those spatial structures. Since the data were simulated using a spherical structure we would not expect the exponential or gaussian models to estimate the range or sill as well. These results (Table 5) indicate that the average range (sill) is different with the spherical (exponential) model being the largest and the gaussian (gaussian) being the smallest. Note how the gaussian underestimates the range and the exponential overestimates the sill. Actual values should be close to the simulation parameters of the range = 3.50 and the sill = 4.00.

Table 5. Simulation Results for Spherical Exponential and Gaussian Models

	mean	maximum	minimum	std dev
Spherical				
range	3.08	5.30	1.55	0.63
sill	4.03	7.35	2.09	0.98
Exponential				
range	2.29	8.14	0.59	1.41
sill	5.37	16.08	2.12	2.75
Gaussian				
range	1.02	1.29	0.76	0.09
sill	3.49	5.22	2.02	0.75

To determine if PROC MIXED had a preference for one spatial structure over another we simulated 100 exponential data sets each with 64 observations as well as 100 gaussian data sets. All data sets were structured as in the previous simulation with treatment effects, 8 by 8 grid and same nugget, range and sill. However, the spatial structure was exponential or gaussian rather than spherical. Combining the two simulation runs we now have a data set which consists of 100 runs of each of spherical, exponential and gaussian. PROC MIXED was used to compute Akaike's and Schwarz's criteria and -2 REML LL. Using these fitting criteria the "best" of either spherical, exponential, or gaussian models was chosen. These results (Table 6) indicated that the model chosen was generally correct. If the data were actually exponential, the exponential model fit best over 60% of the time with the other two models about splitting the remaining data sets. If the data were gaussian the gaussian model always fit best. This seems to be because of the s-shaped nature of the gaussian model which neither the spherical nor the exponential can adequately represent. The spherical model was correctly identified over 75% of the time when the original data were spherical.

Table 6. Simulation Result for Spherical Exponential and Gaussian Data

	Model Chosen		
Data Structure	Spherical	Exponential	Gaussian
Spherical	77	12	11
Exponential	21	64	15
Gaussian	0	0	100

In conclusion, we would recommend that the researcher always use SCORING=10 so that many of the problems of failing to converge will be alleviated. Secondly, we would grid the parameter space fairly densely. Look at the map of the gridded values to see if local minimums exist. Try several models including the spherical, exponential, and gaussian. Also try a model with and without the nugget effect. Remove "drift" if necessary and compare the results when drift was assumed to be absent. Finally, compare the final model with the empirical semivariogram. It will take a great amount of time to use PROC MIXED with spatially correlated data effectively, but the results are worth the effort.

In the future, we would like to see SAS implement nested structures and anisotropic models in PROC MIXED. A test for isotropy would be extremely helpful and allow a more effective way of choosing between an isotropic or anisotropic model. The scoring problem needs to be rectified so that SCORING=10 would not have to be included every time. Although we realize that there is a great tendency to misuse multiple comparisons, it would be extremely convenient if a multiple comparison were easily available in PROC MIXED.

REFERENCES

Cressie, N. 1991. Statistics for Spatial Data. John Wiley: New York

Harville, D.A. 1976. Extension of the Gauss-Markov Theorem to include the estimation of random effects. Ann. Statist. 4:384-395.

Harville, D.A. 1977. Maximum likelihood approaches to variance component estimation and to related problems. J. Amer. Statist. Assoc. 72:320-340.

Henderson, C.R. 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics 31:19-28.

Jeske, D.R. and D.A. Harville. 1988. Prediction-interval and (fixed-effects) confidence-interval prediction for mixed linear models. Commun. Statist.-Theory Meth. 17(4): 1053-1087.

Journel, A.G., and C. Huijbregts. 1978. Mining geostatistics. Academic Press. London.

McLean, R.A., W.L. Sanders, and W.W. Stroup. 1991. A unified approach to mixed linear models. Amer. Statistician. 45:54-63.

SAS Institute. 1992. SAS technical report P-229, SAS/STAT software: changes and enhancements, release 6.07. SAS Institute Inc., Cary, NC.

Stroup, W.W., P.S. Baenziger, and D.K. Mulitze. 1994. Comparison to methods to account for spatial variation in wheat yield trials. Crop Science. in press - to appear in January-February, 1994 issue.

Zimmerman, D.L., and D.A. Harville. 1991. A random field approach to the analysis of field plot experiments and other spatial experiments. Biometrics 47:223-240.

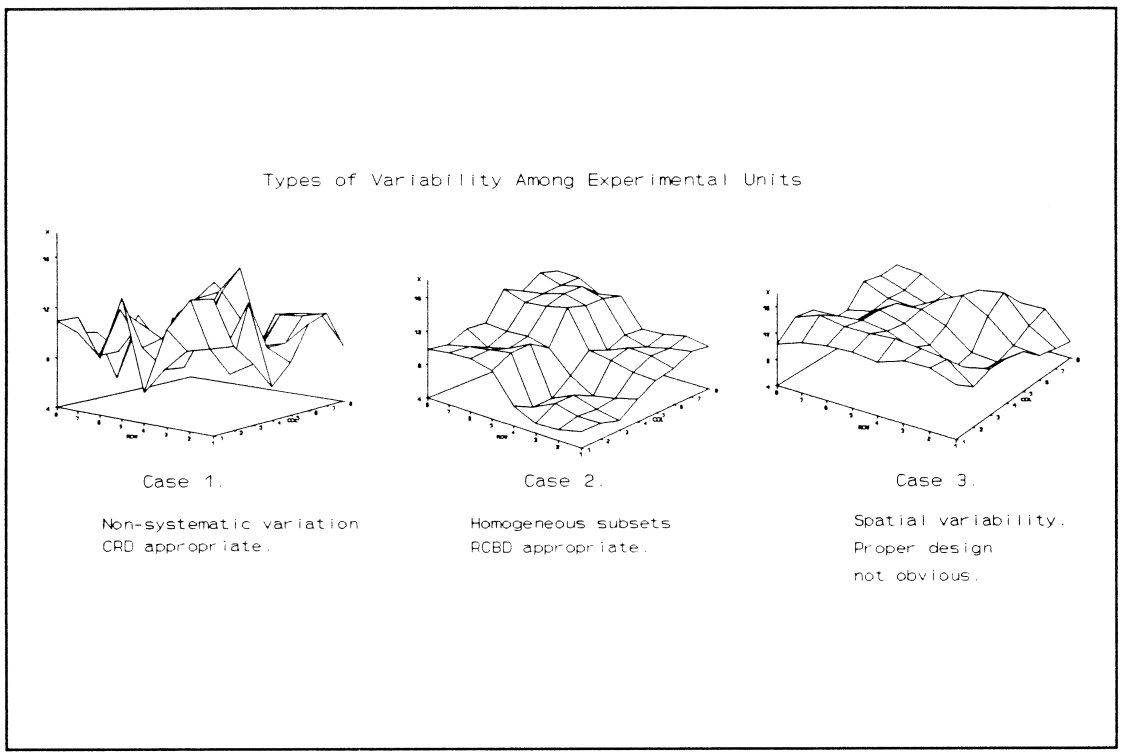

Figure 1

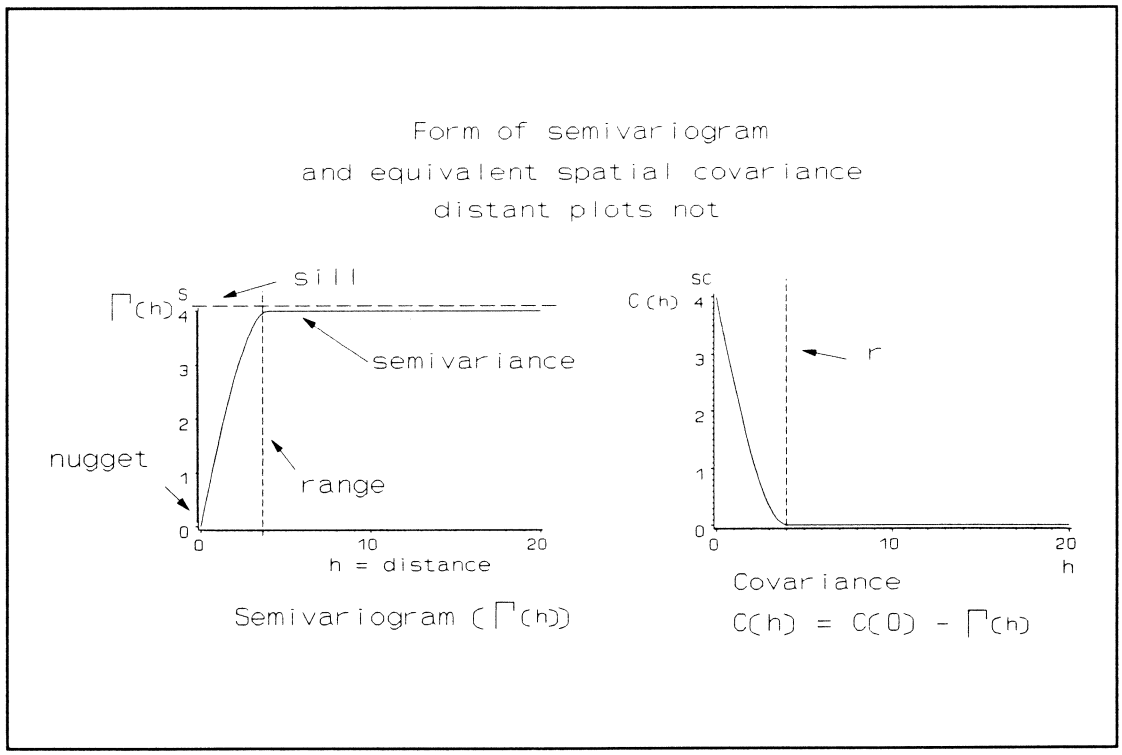

Figure 2

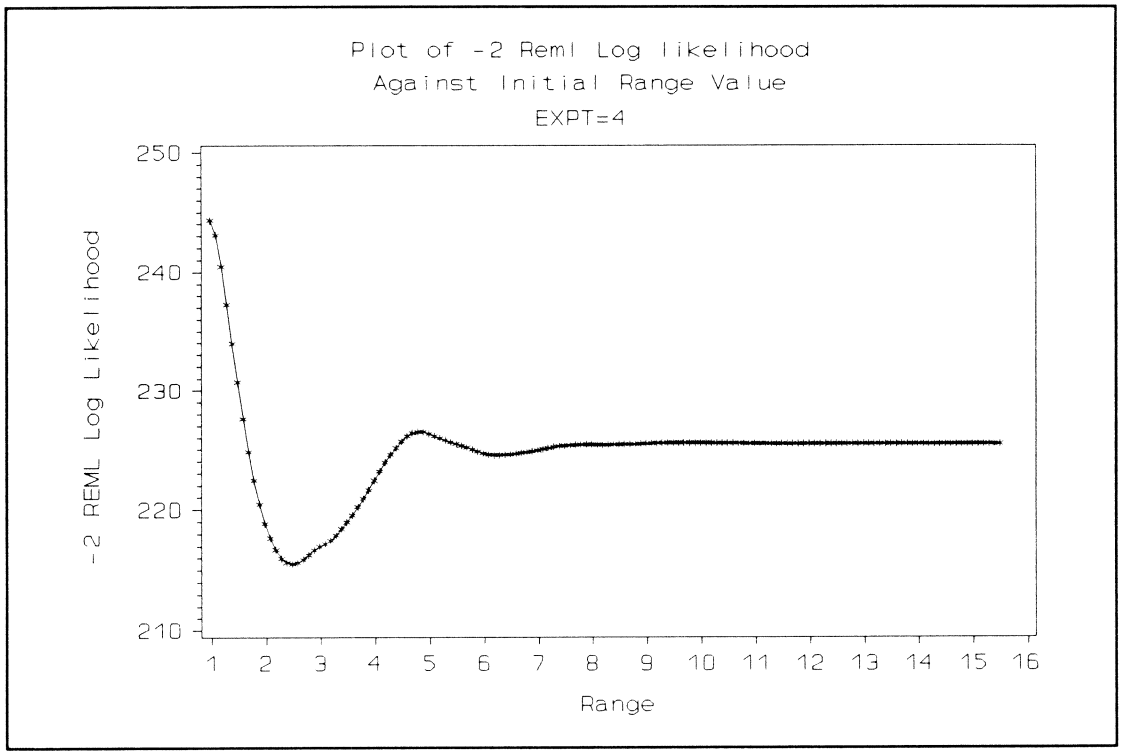

Figure 3

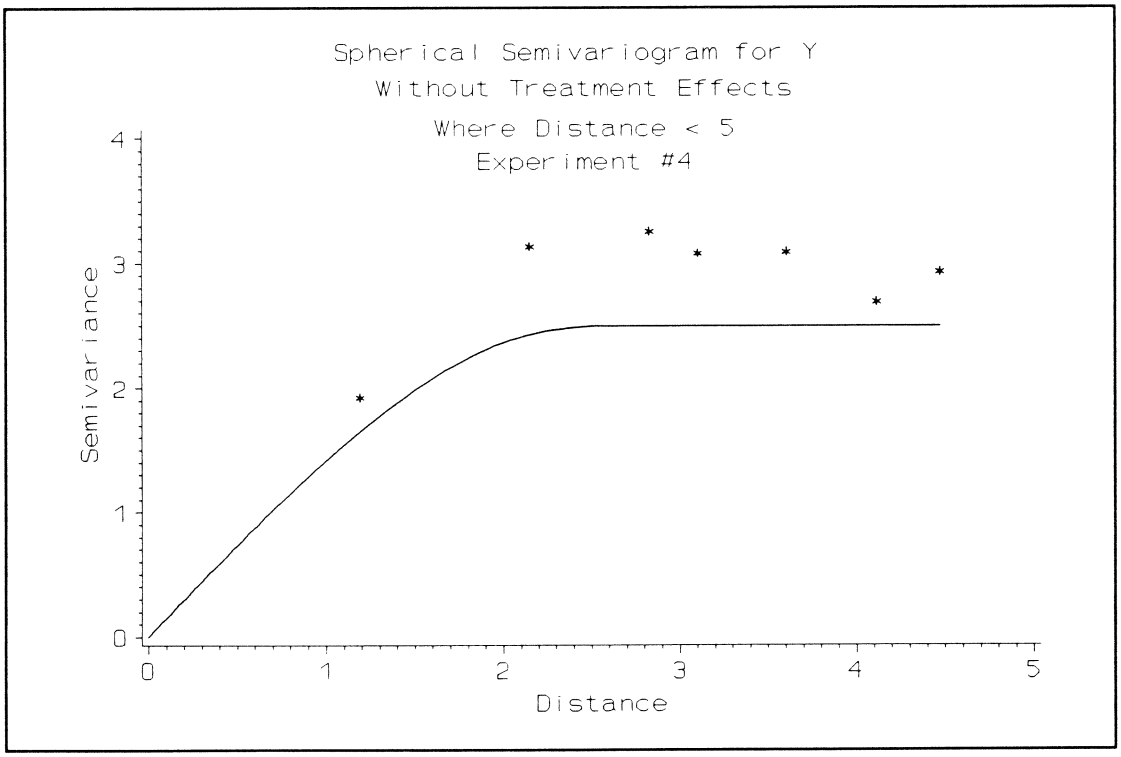

Figure 4

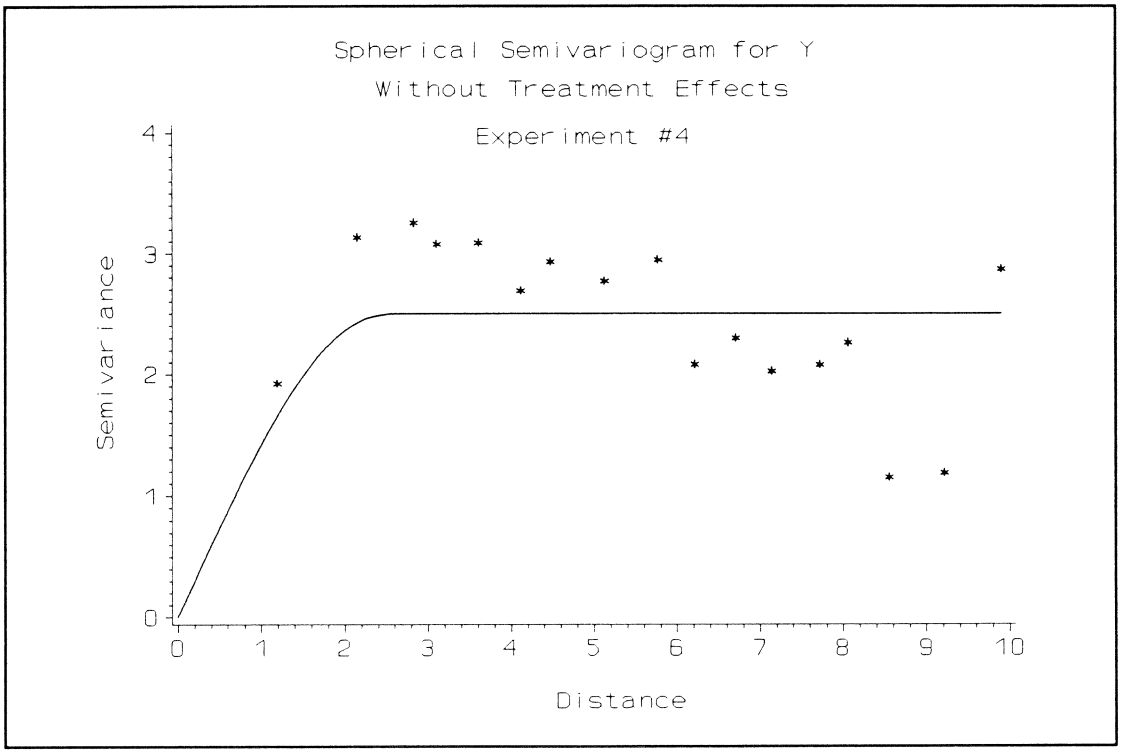

Figure 5

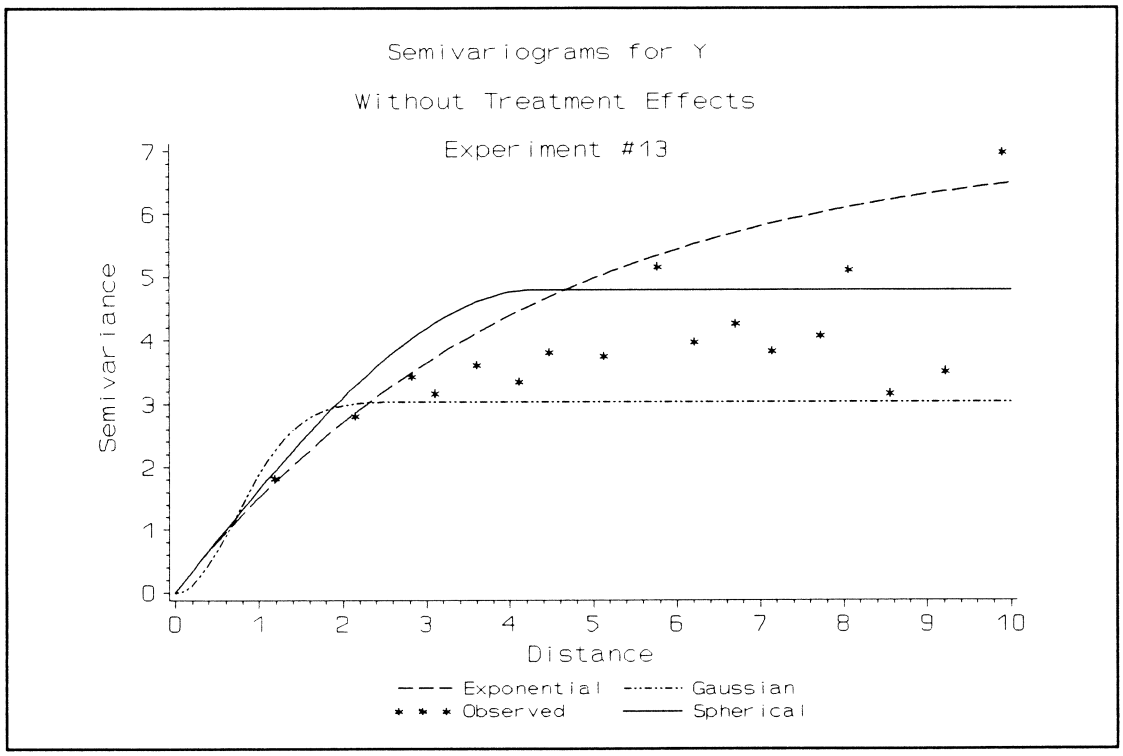

Figure 6

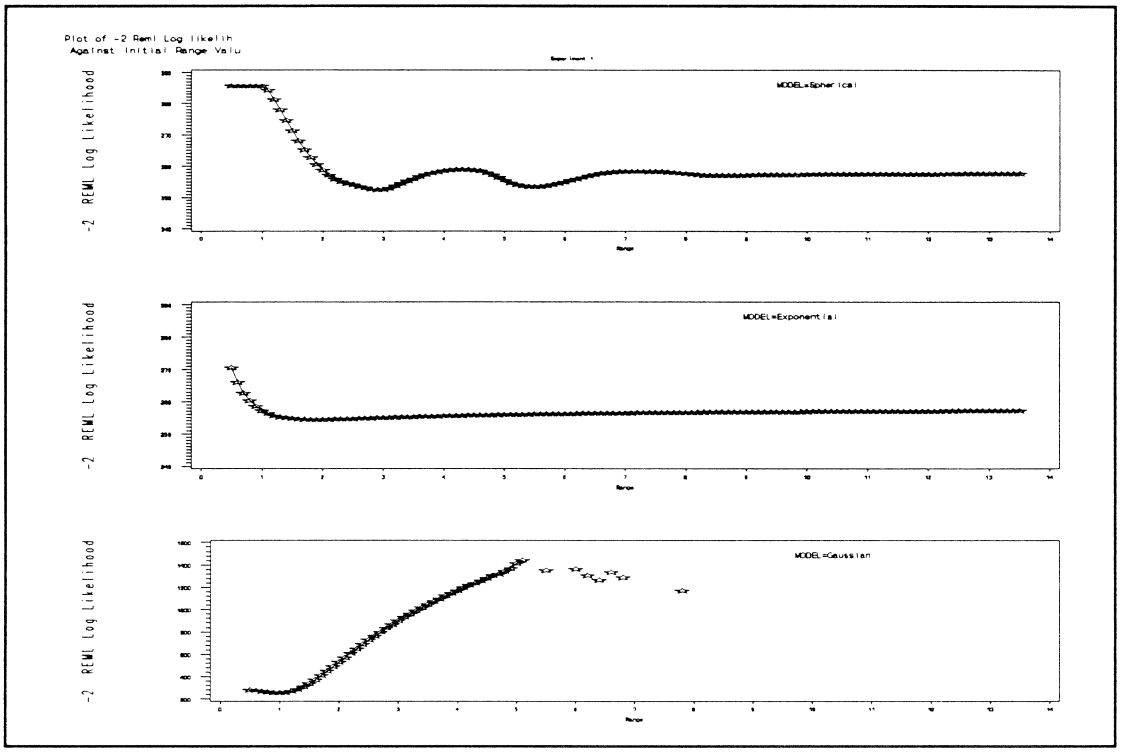

Figure 7

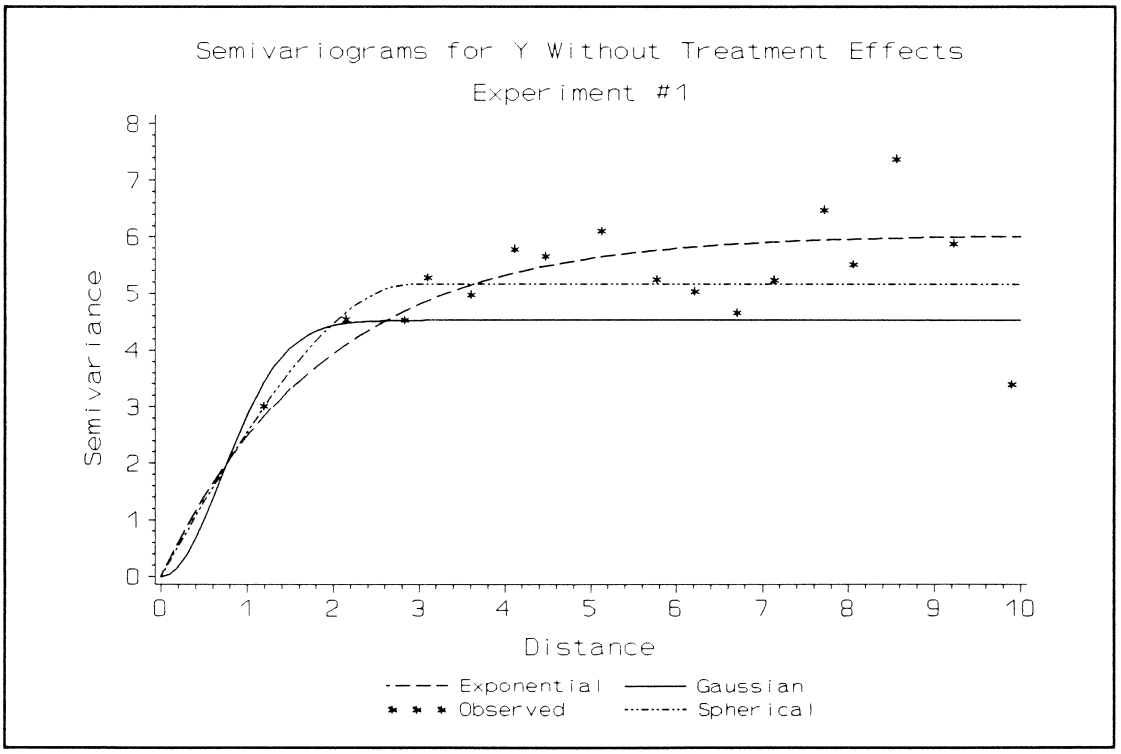

Figure 8

Table 1. Data set used in examples 1 and 2. X is response variable used in example 1.Y is response variable in example 2.

OBS	REP	BLOC	TRT	LAT	LNG	Y	X
1	1	4	14	1	1	8.5411	10.5411
2	1	4	16	1	2	5.5806	8.5806
3	1	2	7	1	3	11.2790	11.2790
4	1	2	6	1	4	13.4344	12.4344
5	1	4	13	2	1	8.3416	10.3416
6	1	4	15	2	2	8.3103	11.3103
7	1	2	8	2	3	9.0282	9.0282
8	1	2	5	2	4	10.7985	9.7985
9	1	3	11	3	1	9.4939	10.4939
10	1	3	12	3	2	10.2576	11.2576
11	1	1	2	3	3	10.3720	7.3720
12	1	1	3	3	4	8.0833	6.0833
13	1	3	10	4	1	9.8869	9.8869
14	1	3	9	4	2	8.2849	8.2849
15	1	1	4	4	3	9.2836	7.2836
16	1	1	1	4	4	11.0018	8.0018
17	2	7	15	5	1	7.3349	10.3349
18	2	7	3	5	2	11.9135	9.9135
19	2	6	10	5	3	8.1662	8.1662
20	2	6	2	5	4	13.7679	10.7679
21	2	7	11	6	1	11.1580	12.1580
22	2	7	7	6	2	11.0230	11.0230
23	2	6	14	6	3	7.2912	9.2912
24	2	6	6	6	4	10.1392	9.1392
25	2	5	5	7	1	14.1097	13.1097
26	2	5	13	7	2	8.0121	10.0121
27	2	8	12	7	3	7.2482	8.2482
28	2	8	16	7	4	4.3975	7.3975
29	2	5	9	8	1	11.0226	11.0226
30	2	5	1	8	2	13.7690	10.7690
31	2	8	8	8	3	6.2206	6.2206
32	2	8	4	8	4	8.5696	6.5696
33	3	12	7	1	5	11.1944	11.1944
34	3	12	13	1	6	5.9737	7.9737
35	3	11	8	1	7	5.8400	5.8400
36	3	11	14	1	8	4.9580	6.9580
37	3	12	4	2	5	12.2561	10.2561
38	3	12	10	2	6	9.8180	9.8180
39	3	11	3	2	7	12.3009	10.3009
40	3	11	9	2	8	7.4719	7.4719
41	3	9	6	3	5	11.1148	10.1148
42	3	9	1	3	6	12.6252	9.6252
43	3	10	15	3	7	5.7800	8.7800
44	3	10	12	3	8	10.2786	11.2786
45	3	9	11	4	5	6.9548	7.9548
46	3	9	16	4	6	3.1100	6.1100
47	3	10	5	4	7	9.6507	8.6507
48	3	10	2	4	8	12.2237	9.2237
49	4	16	9	5	5	10.3129	10.3129
50	4	16	4	5	6	9.3161	7.3161
51	4	13	12	5	7	7.6394	8.6394
52	4	13	1	5	8	10.8669	7.8669

Table 1. continued.

OBS	REP	BLOC	TRT	LAT	LNG	Y	X
53	4	16	15	6	5	6.0250	9.0250
54	4	16	6	6	6	8.2483	7.2483
55	4	13	14	6	7	8.0104	10.0104
56	4	13	7	6	8	10.0473	10.0473
57	4	14	13	7	5	5.0507	7.0507
58	4	14	8	7	6	11.1225	11.1225
59	4	15	3	7	7	14.0253	12.0253
60	4	15	10	7	8	10.4298	10.4298
61	4	14	2	8	5	10.3220	7.3220
62	4	14	11	8	6	9.5104	10.5104
63	4	15	5	8	7	13.6808	12.6808
64	4	15	16	8	8	7.4482	10.4482