A Comparison of the GLM, IML, and MIXED procedures for use in Spatial Statistical Analyses

William J. Price¹, University of Idaho
Bahman Shafii², University of Idaho

ABSTRACT

Incorporation of spatial information into statistical analyses can greatly enhance the estimation process. However, this may not be easily realized due to intensive computational requirements and a lack of a comprehensive software package. One proposed method of analysis is the SAS® procedure PROC MIXED. Use of this procedure has certain drawbacks and potential problems when compared to more familiar SAS® products such as PROC IML and PROC GLM. Limitations as well as advantages associated with all three products are discussed with reference to a previously published data set exhibiting spatial dependencies.

INTRODUCTION

Classical statistical analyses are based on the assumption of independent and identically distributed data in which observations are assumed to have a common variance with no correlation. This simplistic arrangement does not often occur in practice. Specifically, violations of the assumptions of common variability and independence lead to inconsistent parameter estimates, inflated error estimates and an overall lack of resolution in the estimation process. A modification of the distributional assumptions may be made which gives a second possible error structure allowing for different variances (heterogeneous data), while retaining the assumption of no correlation. The heteroskedastic error structure may still result in inflated error estimates in problems where spatial dependency is likely. A third approach uses established spatial statistical techniques, when appropriate, to model the departures from independence, providing an error structure which accounts for correlation as well as heteroskedasticity. Without the availability of adequate hardware and software however, such models are difficult to implement due to problems involved with parameter estimation and matrix inversion. The SAS® procedure PROC MIXED attempts to address these difficulties by fitting a spatial model to the error structure or having the user specify the structure directly as part of the estimation process. While this is a step in the right direction, several problems exist for using PROC MIXED on spatially correlated data. This paper explores these problems and compares the ability of PROC MIXED, PROC IML and PROC GLM to handle the three error structures outlined above. Results are demonstrated with reference to a published data set having a simple treatment design.

METHODS

PROC MIXED is based on the general linear mixed model which may be written as

\[ Y = X\beta + Zu + \epsilon \]  

where

- \( Y \) = n x 1 vector of observations,
- \( X \) = n x p design matrix of known fixed-effects,
- \( \beta \) = p x 1 vector of unknown fixed-effects parameters,
- \( Z \) = n x q design matrix of known random-effects,
- \( u \) = q x 1 vector of unknown random-effects parameters, and
- \( \epsilon \) = n x 1 vector of unknown random errors.

The variance of \( Y \) under this general assumption is:

\[ \text{Var}(Y) = ZGZ' + R \]

where,
Restricted (modified) Maximum Likelihood (REML) is used to provide estimates of unknown variance components which can subsequently be used to construct generalized least square equations if estimates of $\beta$ and $u$ are needed. Incorporation of spatial information into the estimation process is accomplished by modeling the variance structure of either the random components, $G$, or the error term, $R$. The techniques described here will only consider the variance structure of $R$.

The classical fixed effects model can be constructed as a special case of the general linear mixed model by setting $G = 0$ and $R = \sigma^2 I$. This leads to the ordinary least squares estimator of:

$$\hat{\beta} = (X'X)^{-1}X'Y.$$  

(3)

If spatial structure is present, a more appropriate estimator is the generalized least squares estimator:

$$\hat{\beta} = (X'R^{-1}X')^{-1}X'R^{-1}Y$$  

(4)

where $R = \text{Var}(\epsilon)$ is a $n \times n$ symmetric nonnegative definite matrix whose elements are determined according to the underlying spatial structure of the data.

Let $W$ be an $n \times n$ positive definite weight matrix. Three different structures may be assumed for $R$:

i) $W = \sigma^2 I$, Classical model,

ii) $W = \text{Diag}(R)$, Heteroskedastic model, and

iii) $W = R$, Full spatial model.

Determination of the elements of $W$ in the last two cases utilizes geostatistical techniques to characterize the spatial dependency through estimation and modeling of the sample variogram using one of several variogram models (Matheron, 1971; Cressie, 1991). This estimation may be accomplished independently or it may be directly incorporated into the estimation process.

Final estimation of $\beta$ which accounts for spatial variability may be obtained by applying the estimated weight matrix in the generalized least squares estimator (4).

**DEMONSTRATION**

Agricultural data often display spatial characteristics due to the nature of their design and layout. The examples used here are adapted from Cressie (1991, chap. 4, pp 230-243) and are related to soil-water infiltration rates under four soil tillage treatments (moldboard plow, paraplow, chisel, and no till). Three parallel strips (columns) were laid out within each treatment and eight measurements were taken sequentially along each column giving 24 measurements per treatment. The data points are adjusted with a square root transformation to achieve distributional assumptions. Furthermore, tillage treatments are assumed to be independent.

All programs were written and executed under SAS® Windows 6.08 on a 486DX2/66 PC.

The statistical model designates a simple treatment design of TILLAGE with a nested term of COL(TILLAGE). The later term is considered a random component while TILLAGE is considered fixed. Of the procedures covered here, only PROC MIXED can directly deal with mixed models through the estimation of fixed, random, and error components. Although this raises the possibility of applying covariance structure to random model components, this paper only deals with the covariance structure of the error term.

Program 1 illustrates the simple covariance structure where $W = I$ (Table 1). PROC GLM handles the error structure correctly by default, although direct estimation of the random component is not possible. Both PROC IML and PROC MIXED handle the simple structure easily. Estimation of the random component in PROC IML is possible with additional coding (not shown in example) while PROC MIXED estimates the component directly via the RANDOM statement and maximum likelihood estimation. It should also be noted that PROC MIXED correctly tests the TILLAGE effect with the random
component by default, while PROC GLM would require an extra test statement. PROC IML could do the same with proper coding.

Program 2 and Table 2 list the codes and results, respectively, for the next hierarchial step in covariance structure, \( W = \text{diag}(R) \). PROC GLM accomplishes this by specifying the diagonal covariance elements as a variable in a weight statement. This requires prior calculation of the diagonal elements, which correspond to the values of estimated variogram models at infinite lag. Similarly, PROC IML can predefine the elements in a diagonal matrix for subsequent estimation. PROC MIXED on the other hand, can use the weight statement (similar to PROC GLM), or the values can be incorporated through the addition of PARMS and REPEATED statements. Note that to reproduce the results of Cressie for this example, the PARMS statement must be restricted to no iterations (NOITER option), and the REPEATED statement requires an error structure of TYPE=TOEP(1) which is Toeplitz-diagonal. In addition, this method requires the user to specify a value for the random component, COL(TILLAGE), in the PARMS statement as well, thus preventing its estimation (NOITER option). In this case (and the one that follows), the value supplied to PROC MIXED was that obtained from the previous estimation under the simple error structure. Although the correct tests of hypotheses are performed with the specified value, it is not possible to completely define the error structure and simultaneously estimate a random model component. Therefore, any advantages to using PROC MIXED (i.e. random component estimation and construction of appropriate tests) are lost, since results are just as easily obtained from PROC GLM or a generalized PROC IML routine.

It can also be seen that PROC MIXED results in Table 2 differ slightly from those of PROC GLM and PROC IML. This is due to the random component estimation problem combined with the construction of an ANOVA table from PROC MIXED results. PROC MIXED reports only random component estimates and test values for fixed effects. Although these are sufficient extrapolation to sums of squares, some round off errors are unavoidable.

The final step in covariance structures is the full specification of the weight matrix. At this point, PROC GLM lacks the ability to handle such structures and therefore cannot be considered, leaving the user with PROC IML or PROC MIXED. The programs and results for these cases are given in Program 3 and Table 3, respectively. As seen earlier, PROC IML needs the correct matrix specification to do the analysis (it should be noted however, that PROC IML may be limited by computing ability for any of the above structures if the weight matrix becomes too large). The covariance structure in this case arises from Cressie's variogram model estimations for the data. Each tillage treatment follows a unique variogram model. The sample variogram of the first tillage treatment, moldboard plow, is fitted to a spherical model while the paraplow, chisel, and no-till treatments are fit with linear models. Such flexibility is both desirable and necessary for this type of analysis and PROC IML can accommodate this given sufficient computing power. In contrast, PROC MIXED is designed to handle this type of spatial correlation with its estimation options, but is deficient in some aspects. While PROC MIXED can attempt the variogram estimations, it is limited to one functional form at a time. For this example the procedure would be forced to estimate only one variogram model for all treatments which could result in under- or over-parameterization. There are further issues regarding the handling of spatial data in PROC MIXED. The documentation is not clear as to the exact parameters being estimated in the variogram models. More importantly, information regarding assumptions related to spatial modeling such as stationarity and isotropy do not seem to be addressed. Diagnostic information on the
variogram fits such as correlation and significance of parameter estimates are also not available. With these issues unclear, the user may opt for fitting the variograms separately (i.e. PROC REG, PROC NLIN, etc.) with the intent of supplying the results to PROC MIXED as was done with the diagonal model. In this case, the user must determine the required structure for the REPEATED statement. If completely "unstructured" is chosen, it will result in lack of degrees of freedom for the residuals. Under these circumstances, PROC MIXED will use degrees of freedom for covariance parameters, even though they have been fixed with the NOITER option. In order to circumvent this, it may be noted that the weight matrix, $W$, is of a special form, which is banded and can be described with the Toeplitz structure in the REPEATED statement as TYPE=TOEP(8), so fewer parameters need be specified. This option is illustrated in Program 3 and Table 3. Even under this technique, the problem of estimating a random component while specifying an error structure still prevails. No satisfactory analysis has been found in PROC MIXED for this type of fully specified error structure problem.

One advantage of PROC MIXED is the reporting of general model fitting information. Comparison of these statistical criteria among appropriate models is useful for model selection. Table 4 lists three of these criteria for each of the error structures run under PROC MIXED. Larger values are considered 'better' for these measures. The log likelihood values indicate the best fit with the full spatial structure. The other criteria, Akaike’s Information and Schwartz’s Bayesian, show a better fit for the diag($R$) model, but have very small values for the full $R$ model. This occurs because PROC MIXED is assuming 1 df for each covariance parameter and these measures take into account the number of parameters estimated. If these criteria are adjusted for the appropriate number of parameters, i.e. random component and variogram parameters only, they will also indicate the same results (see Table 4).

Although PROC MIXED attempts to address useful aspects of spatial analyses, justification for its use is limited due to poor documentation and restrictive or nonexistent statement options. In this context, questions regarding the adequacy and flexibility of spatial modeling should be addressed. Since similar results may be obtained through other procedures, the potential use of PROC MIXED will be dependent upon future enhancements and additions.

**REFERENCES**


---

1 William J. Price, Research Associate, Statistical Programs, College of Agriculture, University of Idaho, Moscow, ID 83844-2337

INTERNET: bprice@uidaho.edu

2 Bahman Shafii, Director, Statistical Programs, College of Agriculture, University of Idaho, Moscow, ID 83844-2337

bshafii@uidaho.edu

SAS is a registered trademark or trademark of SAS Institute Inc in the USA or other countries. * indicates USA registration.
Table 1. Analysis of variance results for the three SAS procedures assuming a classical model, \( W = I \).

<table>
<thead>
<tr>
<th>Source</th>
<th>PROC GLM</th>
<th>PROC IML</th>
<th>PROC MIXED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>df</td>
<td>SS</td>
<td>MS</td>
</tr>
<tr>
<td>TRT</td>
<td>3</td>
<td>114.267</td>
<td>38.089</td>
</tr>
<tr>
<td>COL(TRT)</td>
<td>8</td>
<td>43.540</td>
<td>5.442</td>
</tr>
<tr>
<td>Error</td>
<td>84</td>
<td>72.831</td>
<td>0.867</td>
</tr>
<tr>
<td>C. Total</td>
<td>95</td>
<td>230.638</td>
<td>2.306</td>
</tr>
</tbody>
</table>

Table 2. Analysis of variance results for the three SAS procedures assuming a heteroskedastic model, \( W = \text{diag}(R) \).

<table>
<thead>
<tr>
<th>Source</th>
<th>PROC GLM</th>
<th>PROC IML</th>
<th>PROC MIXED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>df</td>
<td>SS</td>
<td>MS</td>
</tr>
<tr>
<td>TRT</td>
<td>3</td>
<td>115.542</td>
<td>38.514</td>
</tr>
<tr>
<td>COL(TRT)</td>
<td>8</td>
<td>52.708</td>
<td>6.588</td>
</tr>
<tr>
<td>Error</td>
<td>84</td>
<td>69.986</td>
<td>0.833</td>
</tr>
<tr>
<td>C. Total</td>
<td>95</td>
<td>238.236</td>
<td>2.382</td>
</tr>
</tbody>
</table>

Table 3. Analysis of variance results for the three SAS procedures assuming a full spatial model, \( W = R \).

<table>
<thead>
<tr>
<th>Source</th>
<th>PROC GLM</th>
<th>PROC IML</th>
<th>PROC MIXED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>df</td>
<td>SS</td>
<td>MS</td>
</tr>
<tr>
<td>Model</td>
<td>11</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>TRT</td>
<td>3</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>COL(TRT)</td>
<td>8</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Error</td>
<td>84</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>C. Total</td>
<td>95</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

* Not possible in PROC GLM

Table 4. Model fitting information for the three error structures under PROC MIXED.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Error Structure W = I</th>
<th>Error Structure W = diag(R)</th>
<th>Error Structure W = R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log Likelihood</td>
<td>-137.683</td>
<td>-127.424</td>
<td>-120.194</td>
</tr>
<tr>
<td>Akaike's Information</td>
<td>-139.683</td>
<td>-132.424</td>
<td>-153.194 (-127.194)*</td>
</tr>
<tr>
<td>Schwartz's Bayesian</td>
<td>-142.205</td>
<td>-138.728</td>
<td>-194.803 (-136.169)*</td>
</tr>
</tbody>
</table>

* Adjusted for 7 df.
Program 1. Implementation of the classical model where the weight matrix \( W = I_n \).

```
PROC GLM DATA=SOIL;
   CLASS TRT COL;
   MODEL Y=TRT COL(TRT)/SS3;
   TEST H=TRT E=COL(TRT);
TITLE 'Classic Model on Cressie Data';
PROC MIXED DATA=SOIL;
   CLASS TRT COL;
   MODEL Y=TRT;
   RANDOM COL(TRT);
TITLE 'Mixed Classic Model on Cressie Data';
PROC IML;
   USE SOIL VAR{Rijk};
   READ ALL VAR{Rijk} INTO Y;
   X=I(4)*I(3)*I(8,1,1);
   V=I(96);
   B=INV(V*V)*X*V*Y;
   SSE=Y'*(V-V*X*INV(X'*V*X)*X'*V)*Y;
   SSM=INV(J(NROW(Y),96,1)*V*J(96,1,1))*
       (Y'*V*J(96,1,1)*J(1,96,1)*V*Y);
   SSMOD=B'*X'*V*Y - SSM;
   SSTOT=Y'*V*Y - SSM;
   Htrt=(I(3)*I(1,1,1)/I(3,1,1))*Hcol=(I(2,1,1)/I(2,1,1));
   SSttrt=(Htrt*B)'*INV(Htrt*INV(X'*V*X)*Htrt')*
       (Htrt*B);
   SScol=(Hcol*B)'*INV(Hcol*INV(X'*V*X)*Hcol')*
       (Hcol*B);
   SOURCE={'Model '
        | 'Treatment '}
       | 'Col(Trt) '}
       | 'Residual '
       | 'C. Total ')
   DF=(NROW(B)-1)/NROW(Htrt)/NROW(Hcol)/
       (NROW(Y)-NROW(B))/NROW(Y)-1);
   SS=SSMOD/SStrt/SScol/SSSE/SSTOT;
   MS=SS/DF;
   L_SOURCE=('Source');
   L_DF=('df ')
   L_SS=('SS ')
   L_MS=('MS ');
   L_B='Estimate ' Std Err';
PRINT 'GLS Analysis of Variance',
   SOURCE DF SS MS L_B;
```

Program 2. Implementation of the heteroskedastic model where the weight matrix \( W = \text{diag}(R) \).

```
PROC GLM DATA=SOIL;
   CLASS TRT COL;
   WEIGHT WT;
   MODEL Y=TRT COL(TRT)/SS3;
   TEST H=TRT E=COL(TRT);
TITLE 'Diagonal Model';
DATA PARAM;
   ARRAY COLUMN{34} COL1-COL34;
   DO I=1 TO 34;
      COLUMN{I }=0.0;
   END;
   COL1=.674348; COL2=3.03080; COL3=1.66200;
   COL4=.28810; COL5=-.2881; COL6=1.0;
PROC MIXED DATA=SOIL;
   CLASS TRT COL COLUMN;
   MODEL Y=TRT;
   PARMS/PARMSDATA=PARAM NOlTER;
   RANDOM COL(TRT);
   REPEATED/GROUP=TRT SUB=COLUMN TYPE=TOEP(8)
   TITLE 'Mixed Full Model';
PROC IML;
   START MAKE(M);
      DO I=1 TO 8;
         DO J=1 TO 8;
            H=ABS(I-J);
            DO -
               M(I,J) = 3.0308*3.0308
               IF (H) >= 6 THEN DO:
                  M(I,J,5) = 3.0308*3.0308
               END;
            END
            M(I,J) = 3.0308;
         END;
      END;
      M=J(8,8,3.0308);
      RUN MAKE(M);
      DO I=1 TO 8;
         DO J=1 TO 8;
            N[I,J] = 3.0308*(1.5*(3*H)
            /17.2980) - (.5*(3*H)/17.2980)**3;
            M[I,J] = 3.0308*(1.5*(3*H)
            /17.2980) - (.5*(3*H)/17.2980)**3;
         END;
      END;
      IF (H) >= 6 THEN DO;
      N[I,J]=3.0308;
      M[I,J]=3.0308;
      END;
      FINISH;
      T1=J(8,8,0);
      RUN MAKE(T1);
      T2=1(8)*1.662;
      T34=1(8)*.2881;
      USE SOIL VAR(Rijk);
      READ ALL VAR(Rijk) INTO Y;
      X=I(4)*I(3)*I(8,1,1);
      V=I(96);
      B=INV(V*V)*X*V*Y;
      SSE=Y'*(V*V*INV(X'*V*X)*X'*V)*Y;
```

Program 3. Implementation of the full spatial model where the weight matrix \( W = R \).

```
DATA PARAM;
   ARRAY COLUMN{34} COL1-COL34;
   DO I=1 TO 34;
      COLUMN{I }=0.0;
   END;
   COL1=.674348; COL2=.3.03080; COL3=1.66200;
   COL4=.28810; COL5=-.2881; COL6=1.0;
PROC MIXED DATA=SOIL;
   CLASS TRT COL COLUMN;
   MODEL Y=TRT;
   PARMS/PARMSDATA=PARAM NOlTER;
   RANDOM COL(TRT);
   REPEATED/GROUP=TRT SUB=COLUMN TYPE=TOEP(8);
TITLE 'Mixed Diagonal Model';
PROC IML;
   START MAKE(M);
      DO I=1 TO 8;
         DO J=1 TO 8;
            H=ABS(I-J);
            DO -
               M(I,J) = 3.0308*3.0308
               IF (H) >= 6 THEN DO:
                  M(I,J,5) = 3.0308*3.0308
               END;
            END
            M(I,J) = 3.0308;
         END;
      END;
      M=J(8,8,3.0308);
      RUN MAKE(M);
      DO I=1 TO 8;
         DO J=1 TO 8;
            N[I,J] = 3.0308*(1.5*(3*H)
            /17.2980) - (.5*(3*H)/17.2980)**3;
            M[I,J] = 3.0308*(1.5*(3*H)
            /17.2980) - (.5*(3*H)/17.2980)**3;
         END;
      END;
      IF (H) >= 6 THEN DO;
      N[I,J]=3.0308;
      M[I,J]=3.0308;
      END;
```