R^2 STATISTICS FOR MIXED MODELS

Matthew Kramer
Biometrical Consulting Service, ARS (Beltsville, MD), USDA

Abstract

The R^2 statistic, when used in a regression or ANOVA context, is appealing because it summarizes how well the model explains the data in an easy-to-understand way. R^2 statistics are also useful to gauge the effect of changing a model. Generalizing R^2 to mixed models is not obvious when there are correlated errors, as might occur if data are georeferenced or result from a designed experiment with blocking. Such an R^2 statistic might refer only to the explanation associated with the independent variables, or might capture the explanatory power of the whole model. In the latter case, one might develop an R^2 statistic from Wald or likelihood ratio statistics, but these can yield different numeric results. Example formulas for these generalizations of R^2 are given. Two simulated data sets, one based on a randomized complete block design and the other with spatially correlated observations, demonstrate increases in R^2 as model complexity increases, the result of modeling the covariance structure of the residuals.

1 Introduction

While statisticians tend not to have much interest in R^2, researchers in other disciplines find it useful as a way of describing how well a statistical model fits the observed data (a measure of goodness of fit). Researchers familiar with regression and ANOVA are often surprised when they do not see the familiar R^2 statistic provided with the output from running a mixed model. Mixed models are often suggested by consulting statisticians and can now be estimated by many major statistics packages. A review of the literature reveals many formulas for R^2 statistics that might be adapted for mixed models. All yield the same value for ordinary regression and ANOVA (assuming an intercept term is in the model) but involve different philosophies or assumptions about what an R^2 statistic should represent. When these different philosophies are applied to mixed models, for the same data and mixed model, different R^2 values can result.

Kvålseth (1985) proposed the following requirements for a general R^2:

1. R^2 must possess utility as a measure of goodness of fit and have an intuitively reasonable interpretation,
2. R^2 ought to be dimensionless,
3. $0 \leq R^2 \leq 1$, where $R^2 = 1$ corresponds to a perfect fit, and $R^2 \geq 0$ for any reasonable model specification,
4. Applicable to (a) any type of model, (b) whether effects are fixed or random, and (c) regardless of the statistical properties of the model variables,
5. R^2 should not be confined to any specific model-fitting technique,
6. Values for different models fit to the same data set are directly comparable,
7. Generally compatible with other acceptable measures of fit, and
8. Positive and negative residuals weighted equally.

One could add additional properties, such as those proposed by Cameron and Windmeijer (1996) for count data,

1. R^2 does not decrease as regressors are added,
2. R^2 based on the residual SS (sum of squares) coincides with R^2 based on the explained SS,
3. There is a correspondence between R^2 and a significance test on all slope parameters and between changes in R^2 as regressors are added and significance tests, and
4. R^2 has an interpretation in terms of information content of the data.

There are many formulas that produce the familiar R^2 for regression and ANOVA (Kvålseth, 1985), probably the most commonly seen is $1 - \text{SSE/SST}$, where SSE is the sum of squares of the residuals (error) and SST is the sum of squares of the mean adjusted observations. In the following sections I discuss some of the ways R^2 may be extended into the mixed models framework. Yu (2003) discusses several R^2 measures for a specific type of mixed model commonly used in panel studies (hierarchical models with clustered observations). A Bayesian R^2 for these same kinds of models is described by Gelman and Pardoe (2004). This paper differs from others in extending the concept of R^2 to mixed models with random (block) effects and spatially correlated residuals, the kinds of models commonly used by researchers in designed experiments.

2 R^2 for mixed models

To produce a consistent R^2 statistic in the mixed models framework, additional properties would have to be agreed upon. In particular, whether one perceives random effects and autocorrelated errors as noise or as an important part of the model will greatly affect one's choice of R^2. Different problems necessarily emphasize the importance of different parts of a model—this is a fundamental part of modeling a process and cannot be resolved mathematically. Thus, there can be no general definition of R^2 for mixed models that will cover every model, which is problematic for software developers.

One can think of random effects and autocorrelated errors as noise, that is, effects that mask or distort the true relationship between the predictors and the dependent variable. In this case, one wants an R^2 where these effects have been "removed". This can be accomplished by conditioning on them, so that this R^2 measures a "pure" between-variables relationship. Pierce (1979), in a time series context, suggested the following form: $R^2_* = (\sigma^2_{y|y_*} - \sigma^2_{y|y_*,x})/\sigma^2_{y|y_*}$, where y_* denotes past y (note: the ordinary R^2 in regression can be written as $(\sigma^2_y - \sigma^2_{y|x})/\sigma^2_y$). Then, $\sigma^2_{y|y_*}$ represents the average variance of an observation conditioned on its past and $\sigma^2_{y|y_*,x}$ represents the average variance of an observation conditioned on both its past and on explanatory variables in the model. If one can calculate $\mathrm{SS}_{y|y_*}$, where SS represents a sum of squares, then there is an easy way to go from R^2 calculated the traditional way (ignoring the autocorrelated errors) to R^2_*, as $R^2_* = 1 - V(1 - R^2)$, where $V = \mathrm{SS}_y/\mathrm{SS}_{y|y_*}$. Nelson (1976) gives examples of how to estimate $\mathrm{SS}_{y|y_*}$ for some time series models (these could be adapted for geostatistical data or for random effects). Note that, for random effects, one cannot simply subtract the sum of squares due to random effects from both SST and SSM (sum of squares of the model) and get $R^2_* = \mathrm{SSM}_{\mathrm{adj}}/\mathrm{SST}_{\mathrm{adj}}$, since that would be treating the random effects as if they were fixed.

While obtaining a "between-variables only" R^2 is reasonable for some problems, most of the researchers we work with want to know how R^2 changes if one allows for autocorrelated residuals or random effects, or the difference between an R^2 when blocks are considered as fixed effects rather than as random effects. They are more interested in R^2 as a measure of goodness of fit of the model to the data than as a "pure" measure of between-variables relationship.

Again, there are several ways such an R^2 can be constructed. Magee (1990) suggested generalized R^2 measures based on Wald and likelihood ratio test statistics. One could also base them on score statistics (see Jaffrézic, et al., 2003, and Smyth, 2003, for using the score test to construct a goodness-of-fit measure) and possibly other functions of the

data. Buse (1973) derives a modified R^2 from a Wald statistic as

$$R^2_W = 1 - \frac{\hat{\mathbf{u}}'\mathbf{V}^{-1}\hat{\mathbf{u}}}{(\mathbf{Y}-\bar{\mathbf{Y}})'\mathbf{V}^{-1}(\mathbf{Y}-\bar{\mathbf{Y}})},$$

where $\hat{\mathbf{u}} = \mathbf{Y} - \hat{\mathbf{Y}}$, $\mathbf{Y}$ is the vector of observations, $\hat{\mathbf{Y}}$ is the vector of in-sample predictions from the model, $\mathbf{V}$ is the variance-covariance matrix of the residuals, and $\bar{\mathbf{Y}} = \bar{y}\mathbf{1}$, where $\bar{y}$ is the mean of the data. A likelihood ratio test R^2 (Magee, 1990) is

$$R^2_{LR} = 1 - \exp(-\frac{2}{n}(\text{logL}_M - \text{logL}_0)),$$

where logL_M is the log-likelihood of the model of interest (which would include fixed and random effects and a correlated error structure), logL_0 is the log-likelihood of the intercept-only model, and n is the number of observations. These two different R^2 statistics do not produce identical values for a mixed model, though they do for ordinary regression; other formulas developed from other perspectives would undoubtably also produce different values. Thus, there is an element of choice and no guidance from the literature as to which is the best for a given situation.

The likelihood ratio R^2 is attractive for a number of reasons. For one, it is computationally easy since most mixed models software outputs the maximum log-likelihood of the model. Since it is based on likelihoods, there is a direct relationship with the Kullback-Liebler distance, "information", and information gain, $\text{IG} = -\log(1 - R^2)$ (see Kent, 1983). Third, it can be used when generalizing ordinary regression in other ways, so it provides a coherent strategy for producing an R^2 statistic, given that there is an intercept model, and that maximum log-likelihoods can be calculated for the intercept model and the model of interest. Nagelkerke's (1991) procedure for adjusting R^2 in cases where the maximum attainable R^2 is less than one could also be incorporated (e.g., for some generalized linear mixed models). The "unified" approach for an R^2 given by Huh, et al. (1991), is also related to a likelihood test statistic (they describe it as a likelihood distance measure), but is does not reduce to the usual R^2 in ordinary regression.

Mixed models software provides other measures that are useful for evaluating and comparing models, such as varioius information criteria (e.g., AIC). Information criteria are used to compare models based on the principle of parsimony (the smallest number of parameters to adequately capture the structure of the data). Using, say, AIC, the best fitting model will be neither overparameterized nor underparameterized. In contrast, R^2 will increase (or, at least, not decrease) as parameters are added. While information criteria can help decide which candidate model is best, they do not give one an idea of whether the model explains most or explains little of the variation in the dependendent variable. The best model in a group of models, judged using AIC, may have a low R^2,

though presumably others in that group would also have low R^2 values. The generalized χ^2, also output by some software packages, is a measure of model adequacy based on the distribution of the residuals, and is useful for certain generalized linear models. The generalized χ^2 provides different information about the model than either information criteria or the R^2 statistic.

3 Examples

Two examples will be discussed, both developed from generated data, allowing the comparison of R^2_W and R^2_{LR} with different models. The first example is a random coefficients model in a randomized complete block design. The data were generated from the model $y_{ijk} = \beta_0 + \tau_i + \gamma_j + \alpha_j \delta_{ijk} + \epsilon_{ijk}$, where i indexes the three treatments, j indexes the four blocks ($\text{var}(\gamma) = 4$), δ_{ijk} represents the covariate value for observation ijk ($\text{var}(\alpha) = 1$, and $k = 1, 2$), and ϵ_{ijk} is normally distributed error ($\text{var}(\epsilon) = 1$). Figure 1 gives a realization of data generated from this model, where the relationship between the covariate and Y is depicted for each block. Examination of this figure reveals that the treatment effect is small (true values were 1, 2, 3), that the block effect is modest (there is some difference among the block means), and the effect of the covariate differs among blocks (slope parameters for the covariates are positive in the first three blocks but close to zero in block 4).

The number of parameters, the maximum log-likelihood, R^2_W and R^2_{LR} estimated for various models are given in Table 1. The maximum log-likelihood can be used to calculate AIC or other information criteria (e.g., AIC is $-2 \times$ maximum log-likelihood $+ 2 \times$ the number of parameters in the model). The maximum log-likelihoods were calculated using the *nlme* package (Pinheiro and Bates, 2000) in the statistical software program, **R** (Ihaka and Gentleman, 1996, freely available at `www.r-project.org`); these are standard log-likelihood values (not REML estimates, which adjust for uncertainly due to estimating fixed effects; the REML function does not produce the usual R^2 if used for an ordinary regression model).

As expected, models with more estimated parameters have larger R^2 values. The intercept only model has two parameters, a mean and a variance, and $R^2 = 0$ using either of the expressions for calculating R^2 given above. The model that includes the fixed treatment effect (two additional parameters) produces only a small increase in the log-likelihood value, and an R^2 of only 0.07. However, including the block effect as a fixed effect (three additional parameters) greatly improves the fit; R^2 increases to 0.54. Considering block as a random effect (following the way the data were generated) yields a smaller increase in the log-likelihood and R^2 values. This is expected, since only one

(variance) parameter is used to estimate the block effect, and estimates are shrunk towards zero and away from their fixed estimates. Note that R^2_{LR} is not as high as R^2_W. Since these two R^2 values are based on different premises, there is no reason to expect them to yield the same value.

Similar results occur when comparing full models, all effects fixed (the model that includes the block $\times$ covariate interaction) versus random block and covariate effects (the latter matching the model generating the data). The R^2 value for the all effects fixed model (11 parameters) exceeds either of the R^2 values for the corresponding mixed model (seven parameters).

One might conclude, based on R^2 values, that the all fixed effects model "fits" the data better than the all random effects model. While, in the R^2 sense, this is true, the assumptions underlying the models differ, so a comparison is not enlightening. The model with all effects fixed assumes that the inference space is only to those blocks used in the model, not a population of blocks from which a sample of blocks was drawn. Under each set of assumptions, the log likelihood was maximized, so parameter estimates are "optimal" for their respective models. A better comparison for the all random effects model would be with the model containing no block effect but just the covariate, so that one can see the improvement by adding in a block effect. We see that the five parameter model (treatment and covariate) produces an R^2 of 0.36, adding in the block effect as two variances (seven parameters) produces an R^2_{LR} of 0.84 and an R^2_W of 0.93, a large improvement.

The second example is based on data generated with spatially autocorrelated residuals, where the correlation is given by $\rho = \exp(-\mathrm{d}_{i,j}/2)$, where $\mathrm{d}_{i,j}$ is the distance between an observaton at location i and an observation at location j. There are two levels, so this data set could represent observations on some characteristic of two different species distributed in a field with observations near each other more similar than those further apart, perhaps due to unmeasured local microhabitat or soil conditions. The data are displayed in Fig. 2, the two levels (species) are represented by different symbols, and the magnitude of the observed characteristic is represented by topographic colors, blue the lowest values, green to yellow, middle values, and brown the highest values. That the observations are autocorrelated can be readily seen by examining the semivariograms for each level (Fig. 3).

Table 2 gives maximum log-likelihood and R^2_{LR} values for an intercept only model (two parameters), a model that also includes a level (species) effect (three parameters), and a model that additionally allows for autocorrelated residuals (four parameters). Allowing for autocorrelated residuals (one additional parameter) increases R^2_{LR} by about 16%. Note that, while we are saying the model captures more information about the data, at the

same time we are saying there is less information in the data because observations are not independent.

4 Summary and Conclusions

There are various R^2's that can be developed for mixed models, all produce the same value for ordinary regression, so would satisfy the properties set forth by Kvålseth (1985) and Cameron and Windmeijer (1996). An R^2 based on the likelihood ratio test is easy to calculate from standard mixed models output and has a connection to information theory. Examples were shown demonstrating increases in R^2 when adding random effects or correlated errors to the model. Because philosophies about what R^2 should measure can differ in the mixed models framework, there will be no universally acceptable R^2 value for a mixed model. However, R^2_{LR} is easy to calculate using generally available statistical software, so it can serve as a measure of goodness of fit of the model to the data.

5 Acknowledgments

I thank Mary Camp, Bryan Vinyard, and an anonymous referee for critically reviewing the manuscript.

6 References

Buse, A. 1973. Goodness of fit in generalized least squares estimation. Amer. Stat. 27, 106–108.

Cameron, C. and F.A.G. Windmeijer. 1996. R-squared measures of count data regression models with applications to health-care utilization. J. Bus. Econ. Stat. 14, 209–220.

Gelman, A. and I. Pardoe. 2004. Bayesian measures of explained variance and pooling in multilevel (hierarchicial) models. `www.stat.columbia.edu/~gelman`. 21 pp.

Huh, M.H., J.H. Lee, J.W. Jung. 1991. Unified approach to coefficient of determination R^2 using likelihood distance. Korean J. Applied Stat. 4, 117-127. (In Korean.)

Ihaki, R., and R. Gentleman. 1996. R: A language for data analysis and graphics. J. Comput. Graph. Stat. 5, 299–314.

Jaffrézic, F., I.M.S. White, and R. Thompson. 2003. Use of the score test as a goodness-of-fit measure of the covariance structure in genetic analysis of longitudinal data. Genet. Sel. Evol. 35, 185–189.

Kent, J.T. 1983. Information gain and a general measure of correlation. Biometrika 70, 163–173.

Kvålseth, T.O. 1985. Cautionary note about R^2. Amer. Stat. 39, 279–285.

Magee, L. 1990. R^2 measures based on Wald and likelihood ratio joint significance tests. Amer. Stat. 44, 250–253.

Nagelkerke, N.J.D. 1991. A note on a general definition of the coefficient of determination. Biometrika 78, 691–692.

Nelson, C.R. 1976. The interpretation of R^2 in autoregressive-moving average time series models. Amer. Stat. 30, 175–180.

Pierce, D.A. 1979. R^2 measures for time series. J. Amer. Stat. Assoc. 74, 901–910.

Pinheiro, J.C. and D.M. Bates. 2000. Mixed-effects models in S and S-PLUS. Springer, N.Y. 528 pp.

Smyth, G.K. 2003. Pearson's goodness of fit statistic as a score test statistic. *In* Goldstein, D.R. (ed.) Science and statistics: A Festschrift for Terry Speed. IMS Lecture Notes—Monograph Series, Volume 40, Inst. of Math. Stat., Hayward, CA.

Xu, R. 2003. Measuring explained variation in linear mixed effects models. Statist. Med. 22, 3527–3541.

7 Tables

Table 1. Maximum log-likelihood, and R^2_{LR} and R^2_W values for various models for data generated from a randomized complete block design with a covariate (see text for details), (f) = fixed effect, (r) = random effect.

model	num. parms.	log-likelihood	R^2_{LR}	R^2_W
intercept only	2	-64.45	0	0
trt	4	-63.55	0.07	0.07
trt + cov (f)	5	-59.10	0.36	0.36
trt + block (r)	5	-60.60	0.27	0.32
trt + block (f)	7	-55.18	0.54	0.54
trt + block (r) + cov (r)	7	-42.74	0.84	0.93
trt + block (f) + cov (f)	8	-40.15	0.87	0.87
trt + block (f) + cov (f) + block×cov (f)	11	-24.63	0.96	0.96

Table 2. Log-likelihood and R^2_{LR} values for models for data generated with two levels and spatially autocorrelated residuals (see text for details).

model	log-likelihood	R^2_{LR}
intercept only	-495.94	0
level	-389.68	0.51
level + correlated residuals	-225.27	0.67

8 Figures

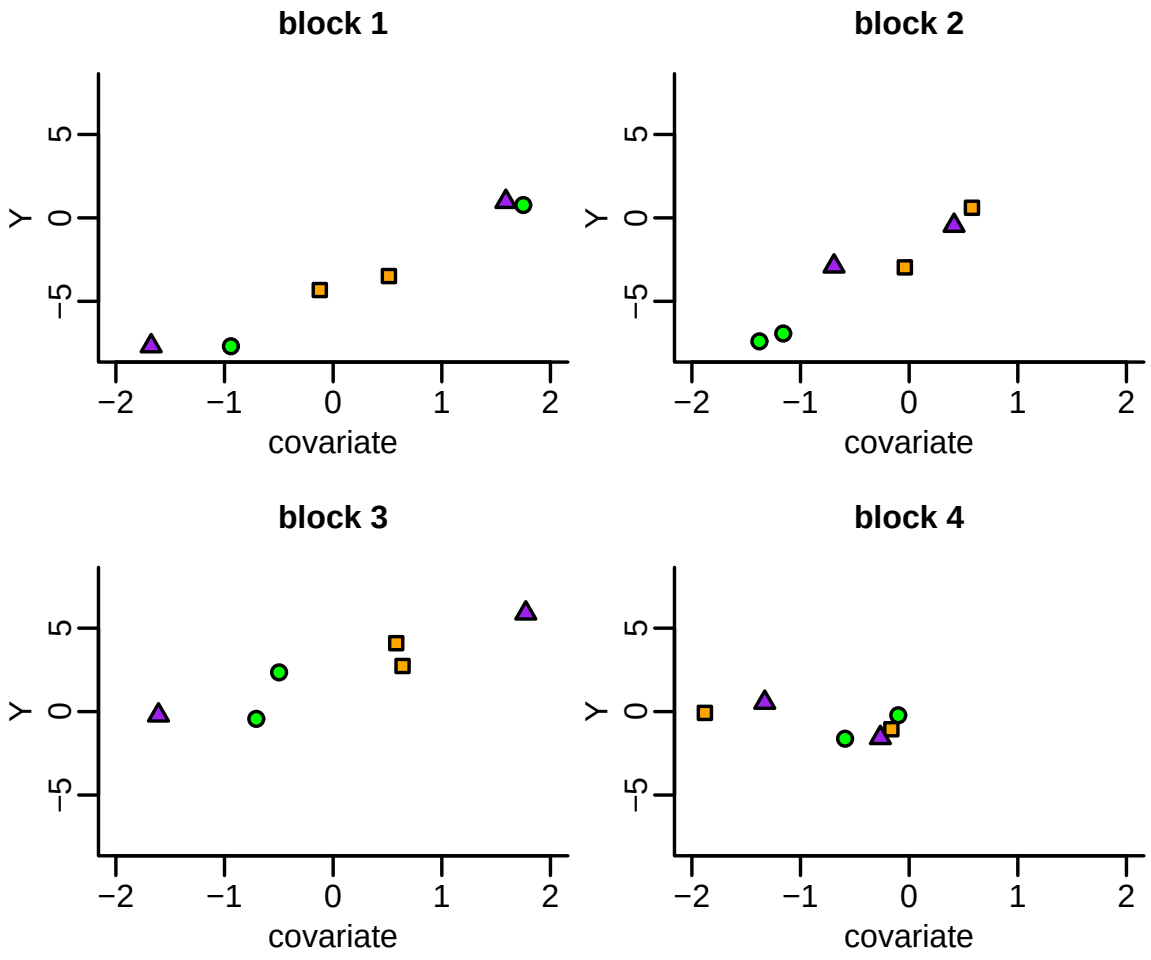

Figure 1. Relationship between the dependent variable and covariate for each block for the simulated data of Example 1. The three treatments are distinguished by symbol and color.

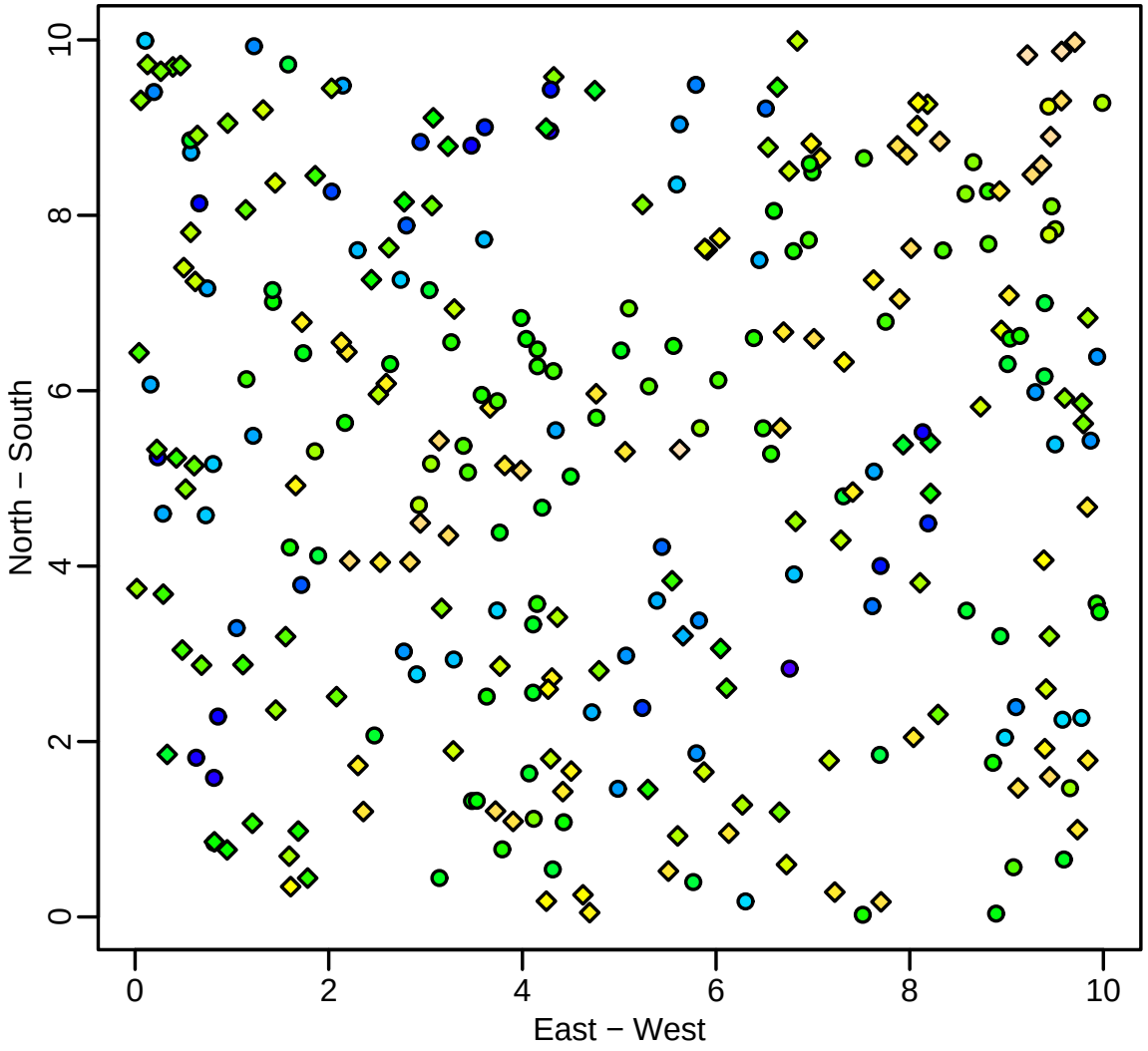

Figure 2. Spatial distribution of simulated data of Example 2. The two levels (species) are distinguished by symbol type. Magnitude of the response variable is indicated by topographic color, low is blue, medium is green to yellow, high is brown.

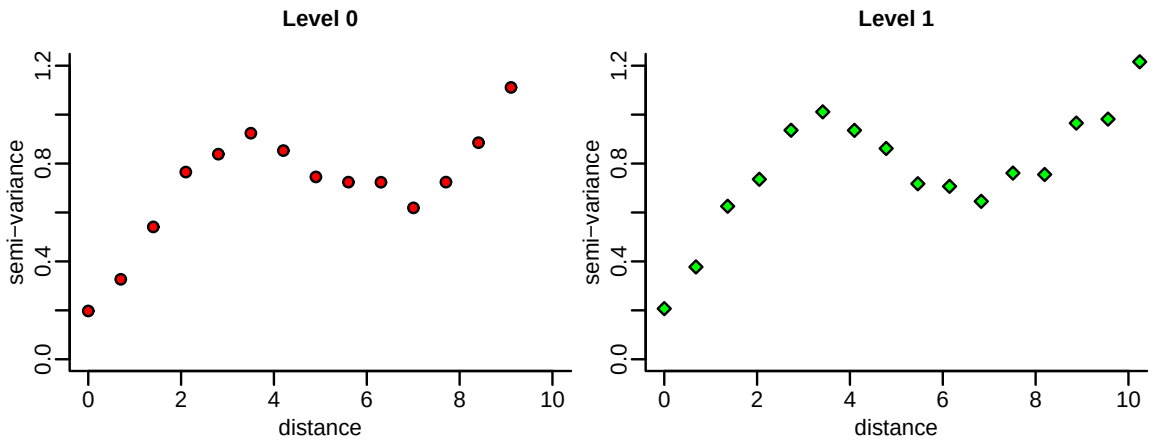

Figure 3. Semivariogram for each of the levels (species) of Example 2, depicting the nature of the spatial autocorrelation.