MODELLING OF CORRELATED SOIL ANIMALS COUNT DATA

Legesse Kassa Debusho[*1] and Gudeta W. Sileshi[2]

[1]Department of Statistics, University of Pretoria, Private Bag X20,
Hatfield 0028, South Africa

[2]World Agroforestry Centre (ICRAF), Southern Africa Regional Programme,
Chitedze Agricultural Research Station, P.O. Box 30798, Lilongwe, Malawi

ABSTRACT

Ecological studies naturally result in correlated data. Ignoring these correlations can result in biased estimation of ecological effects jeopardizing the integrity of the scientific inference. Mixed effects models are likely to appeal to ecologists for handling correlated data (e.g. Sileshi, 2008), however careful consideration must be given to the interpretation of the parameter estimates from generalized linear mixed effects models with non-identity link functions. The objective of this study was to compare the generalized estimating equations (GEE) under different correlation structures and suggest appropriate models to describe the relationship between soil animal counts and covariates. The GEE with independence, exchangeable and AR1 correlation structures were compared using count data set of ants from soils under the agroforestry systems in eastern Zambia. The GEE model with AR1 correlation structure gave a better description of the data than did the independence and exchangeable correlation structures.

1. INTRODUCTION

The most common analyses used for soil animals count consisted of either non-parametric tests (Jabin et al., 2004), log-normal least squares regression (e.g. ANOVA), or generalised linear model (GLM) with Poisson distribution (Sileshi, 2008). However, such analyses violate the independence assumption if the response variable, the number of soil animals, is measured repeatedly over time at the same site. Statistical methods that assume independence among observations result in optimistic estimates of uncertainty when applied to correlated data, which are ubiquitous in applied ecological research (Fieberg et al., 2009). The log-normal regression is generally inappropriate for modelling a discrete process. When testing for

* To whom correspondence should be addressed: E-mail: legesse.debusho@up.ac.za

habitat, land-use or treatment effects, the distributional assumptions made about the response variable can have a critical impact on the conclusions drawn. Often data do not support only one model as clearly best for analysis (Johnson and Omland, 2004). This raises the issue of using appropriate models to assess which ones are adequate for the data and which one could be chosen as the basis for interpretation, prediction, or other subsequent use. Liang and Zeger (1986) introduced a generalized estimating equation (GEE) approach based on a working correlation matrix to obtain efficient estimators of regression parameters in the class of generalized linear models for repeated measures data. While GEE allows for specification of a working matrix for modelling within-subject correlations, it is very important to correctly specify the variance and correlation structure for efficient estimation (Leung et al., 2009; Wang and Carey, 2003; Wang and Lin, 2005). According to Wang and Lin (2005) correct specification of the variance function can improve the estimation efficiency even if the correlation structure is misspecified. According to Leung et al. (2009) the efficiency of a GEE estimate can be seriously affected by the choice of the working correlation model. Therefore, the objective of this study was to compare the generalized estimating equations (GEE) under different correlation structures and suggest appropriate models for the analysis of soil animal count data.

2. MATERIALS and METHODS

2.1. Sources of data

The data used in this study were collected from agroforestry systems in eastern Zambia. These were reported elsewhere (Sileshi and Mafongoya, 2007; Sileshi, 2008). The data collected were used to quantify temporal variations in macrofauna in relation to different land-use categories (Sileshi and Mafongoya, 2006). A total of 356 soil samples were collected from maize grown using leguminous agroforestry species and continuous monoculture maize four times between December 2003 and February 2005 at Msekera and Kalunga sites. A stratified-random sampling procedure was followed when sampling the agroforestry according to tree species, which differed in the quality and quantity of their organic inputs. Five treatments were compared in the agroforestry system: maize monoculture, maize grown after pure species fallows of four legume species, namely, Gliricidia sepium, Acacia anguistissima, Leucaena collinsi and Calliandra calothyrsus. The treatments were replicated three times. The samples were collected using a soil monolith (25 cm × 25 cm and 25 cm

depth) placed over a randomly selected spot (Swift and Bignell, 2001), and driven into the soil to ground level using a metallic mallet. From each soil monolith, macrofauna were hand-sorted to a family or order level and numbers recorded. The data collected contains six different soil animals, however in this paper only the results of ants will be discussed. Because repeated observations were recorded on the same site, one might expect observations from the same site to be more similar than observations from different sites.

2.2. Generalized estimating equations

Since the seminal publication of Liang and Zeger (1986) several approaches have been developed to improve the technique. The literature on GEE is extensive and the basic ideas can be found in Ziegler et al. (1996), Greene (1997), Hardin and Hilbe (2002), Fitzmaurice et al. (2004) and (Molenberghs and Verbeke, 2005, Chapter 8) .

Let Y_{ij} be the average number of ants in site i at month (i.e. time) j, $j = 1,\ldots,m_i; i = 1,\ldots,n.$ Assume $Y_{ij} \sim$ Poisson (μ_{ij}), and therefore the mean and variance of Y_{ij} are equal to μ_{ij}. Following Fitzmaurice et al. (2004), in the matrix notation the systematic part of the model is given by

$$\boldsymbol{\eta}_i = \mathbf{X}_i\boldsymbol{\beta} \tag{1}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression parameters and $\mathbf{X}_i$ is an $m_i \times p$ matrix holding the ith individual covariate values at each of m_i response times. The relationship between the conditional mean with respect to the explanatory variables, $\mathbf{X}_i$ (i.e. the expected value of Y_{ij} for a given $\mathbf{X}_i$), and the systematic component has the same form as in GLM models and hence we use

$$E(Y_{ij} \mid \mathbf{X}_i) = \mu_{ij} \text{ and } g(\mu_{ij}) = \mathbf{X}_i\boldsymbol{\beta}, \tag{2}$$

where $g(.)$ is the link function. The conditional variance structure of Y_{ij} is given by

$$Var(Y_{ij} \mid \mathbf{X}_i) = \phi \times v(\mu_{ij}), \tag{3}$$

where $v(.)$ is the variance function and ϕ is the scale parameter or overdispersion parameter which we need to estimate. Choosing $\phi = 1$ and $v(\mu_{ij}) = \mu_{ij}$ gives the variance structure of Poisson GLM. The next step for GEE analysis is to specify an association structure between

Y_{ij} and Y_{ik}, where j and k are two different sampling months on the same site. Depending upon the type of data, continuous, count or binary, there are many ways of defining the structure such as unstructured correlation, auto-regressive correlation and exchangeable correlation structure.

The estimates for regression parameters in $\boldsymbol{\beta}$ are obtained by solving

$$\sum_{i=1}^{n} \mathbf{D}_i \boldsymbol{\Sigma}_i^{-1}(\alpha)(\mathbf{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}, \qquad (4)$$

where the $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{im_i})'$ denotes the response vector for subject (site) i (at each of m_i observation times) and contains all the longitudinal data from site i, $\mathbf{D}_i = \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta}$ is an $m_i \times p$ matrix of first-order derivatives of the $\boldsymbol{\mu}_i$ (= $f(\mathbf{X}_i, \boldsymbol{\beta})$ represents the mean response as a function of covariates) with respect to $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}_i(\alpha) = \mathbf{A}_i^{1/2} \mathbf{R}_i(\alpha) \mathbf{A}_i^{1/2}$ is the variance-covariance matrix of subject i and $\mathbf{A}_i$ are diagonal matrices containing the variances. The working correlation matrix $\mathbf{R}_i(\alpha)$ describes within-subject dependencies. The $\boldsymbol{\Sigma}_i(\alpha)$ is usually modelled by adopting a common form for the variance based on an appropriate member of the exponential family. Therefore, when the distribution of the response variable is a member of the exponential family we do not need all the details of the probability distribution in order to estimate the regression parameters $\boldsymbol{\beta}$, only its mean and variance. However, we have to determine what form $\mathbf{R}_i(\alpha)$ takes, this means that we choose a correlation structure (e.g. exchangeable correlation or auto-regressive correlation) that closely describes what is observed in the response data.

The equation in (4) is solved numerically by iteration that consists of the following steps:

Step 1: For a given ϕ and α (and therefore $\hat{\boldsymbol{\Sigma}}_i(\alpha)$, an estimate of $\boldsymbol{\Sigma}_i(\alpha)$), obtain an estimate for the regression parameters.

Step 2: Given the regression parameters, update ϕ and α (and therefore $\hat{\boldsymbol{\Sigma}}_i(\alpha)$).

Step 3: Iterate between steps 1 and 2 until convergence.

At convergence, the estimated regression parameters are nearly equal to the population parameters, i.e. consistent, and identically normally distributed with mean $\boldsymbol{\beta}$ (see Molenberghs and Verbeke, 2005, page 159) and has a covariance matrix

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \mathbf{B}^{-1} \times \mathbf{M} \times \mathbf{B}^{-1}, \qquad (5)$$

where $\mathbf{B} = \sum_{i=1}^{n} \mathbf{D}_i \times \boldsymbol{\Sigma}_i^{-1}(\alpha) \times \mathbf{D}_i$ and $\mathbf{M} = \sum_{i=1}^{n} \mathbf{D}_i \times \boldsymbol{\Sigma}_i^{-1}(\alpha) \times \mathrm{Cov}(\mathbf{Y}_i) \times \boldsymbol{\Sigma}_i^{-1}(\alpha) \times \mathbf{D}_i$. The empirically corrected estimate for $\mathrm{Var}(\hat{\boldsymbol{\beta}})$ is obtained from replacing in (5) $\boldsymbol{\Sigma}_i(\alpha)$ by its estimate $\hat{\boldsymbol{\Sigma}}_i(\alpha)$ and $\mathrm{Cov}(\mathbf{Y}_i)$ by the covariance matrix $(\mathbf{Y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})(\mathbf{Y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})'$. The chosen correlation structure $\mathbf{R}_i(\alpha)$ is then used in the covariance matrix, resulting in the sandwich estimate or the robust variance estimate. The diagonal element of this matrix can be used to obtain the standard errors of the regression parameter estimates. For many statistical packages this is the default estimate that is displayed in the standard error column of the summary table of the model.

A correlation matrix $\mathbf{R}_i(\alpha)$ can be selected as follows. Fit a series of models that differ only in the choice of correlation matrix $\mathbf{R}_i(\alpha)$; all of the remaining features of these models are the same. For each model compare the sandwich variance estimates with the model-based variance estimates. The model whose sandwich variance estimates most closely resembles its model-based variance estimates is the one with the best correlation matrix $\mathbf{R}_i(\alpha)$.

The analyses were carried out using the R language and environment for statistical computing.

3. RESULTS and DISCUSSION

We have assumed the relationship between the mean μ_{ij} and the covariates is given by

$$\mathrm{E}(\mathrm{Y}_{ij}) = \mu_{ij} = e^{\beta_0 + \beta_1 \times \mathrm{Month}_{ij} + \beta_2 \times \mathrm{Month}_{ij}^2 + \beta_3 \times \mathrm{Treatment}_{ij}}.$$

Further, we used the variance of the observed data as $\mathrm{Var}(\mathrm{Y}_{ij} \mid \mathbf{X}_i) = \phi \times \mu_{ij}$. We used three different working correlation assumptions: independence (equivalent to GLM), exchangeable (equal correlation among all observations from the same site) and an auto-regressive with order 1 (AR1) structure (i.e. serial dependence). The three GEE models yielded similar conclusions regarding the effect of covariates and also agreed well with the relationship

estimated from the generalized linear mixed model (GLMM). The parameter estimates from the fitted GEE models were relatively insensitive to the assumed correlation structure as the correlations were close to 0 (Table 1).

Table 1. Estimates of regression parameters and standard errors (SE) for models fit to the ants average count data. Generalized estimating equation (GEE) regression models were fitted using independence, exchangeable and AR1 working correlation structures (GEE-IND, GEE-EXC and GEE - AR1, respectively) and the GLMM model was fit using a random intercept for site.

Parameter‡	GEE-IND	GEE-EXC	GEE-AR1	GLMM
Intercept	0.384 (0.862)	0.394 (0.861)	0.440 (0.865)	0.369 (0.401)
Treatment 2	-1.039 (0.048)	-1.035 (0.046)	-1.046 (0.042)	-1.039 (0.220)
Treatment 3	-1.039 (0.752)	-1.035 (0.742)	-1.005 (0.723)	-1.039 (0.220)
Treatment 4	-0.257 (0.178)	-0.256 (0.176)	-0.252 (0.175)	-0.257 (0.170)
Treatment 5	-0.927 (0.135)	-0.923 (0.135)	-0.878 (0.145)	-0.927 (0.211)
Month	1.103 (0.432)	1.098 (0.431)	1.037(0.443)	1.103 (0.341)
Month2	-0.268 (0.062)	-0.267 (0.063)	-0.255 (0.065)	-0.268 (0.070)
Correlation	NA	-0.012	-0.036	
CIC	2.59	2.57	2.50	
C.crit	9.04	8.80	8.53	

‡ Treatment 1 = maize, Treatment 2 = Gliricidia sepium, Treatment 3 = Acacia anguistissima, Treatment 4 = Leucaena collinsi Treatment 5 = Calliandra calothyrsus, and Month2 = Month × Month.

The naïve standard errors of the parameter estimates are slightly greater than the sandwich estimates in all three models (the results are not given here). To compare the models we use the correlation information criterion (CIC) of Hin and Wang (2009). As was discussed above we can also use the naive (i.e. model-based) and robust (i.e. sandwich) variance estimates to select a correlation model. The model whose robust variance estimates most closely resembles its naive variance estimates is the better correlation model. To obtain a single summary statistic for this comparison we use the entire parameter covariance matrix and sum the absolute differences between the naive and robust covariance matrices (C.crit in Table 1). The sum of the absolute differences between the naive and robust covariance estimates is smallest for the AR1 correlation structure. This is consistent with the CIC conclusion.

As the AIC of the GLM (i.e. Poisson) regression model (results are not given here) is greater than that of the GLMM regression model, it indicates that GLMM should be preferred. Thus we have evidence for observational heterogeneity, a conclusion that is consistent with the GEE model selection results that favoured an AR1 correlation structure over independence.

REFERENCES

Fieberg, J., Randall H. Rieger, Michael C. Z. & Jonathan S.S. 2009. Regression modelling of correlated data in ecology: subject-specific and population averaged response patterns. *Journal of Applied Ecology* 46(5): 1018–1025.

Fitzmaurice, G.N., Laird N.M. & Ware J. 2004. Applied longitudinal analysis. Wiley-IEEE.

Greene, W.H. 1997. Econometric analysis, Third edition. Prentice-Hall, Inc, Upper Saddle River, N.J.

Halekoh, U., Søren H. & Jun Y. 2006. The R package geepack for generalized estimating equations. *Journal of Statistical Software* 15(2).

Hardin, J.W. & Joseph M.H. 2002. *Generalized Estimating Equations*. Chapman & Hall/CRC Press: Boca Raton, FL.

Hin, L.Y. & You-Gan W. 2009. Working-correlation-structure identification in generalized estimating equations. *Statistics in Medicine* 28: 642–658.

Jabin, M., Mohr, D., Kappes, H. & Topp, W. 2004. Influence of deadwood on density of soil macro-arthropods in a managed oak–beech forest. *For. Ecol. Manage*. 194, 61–69.

Johnson, J.B. & Omland, K.S., 2004. Model selection in ecology and evolution. *Trends Ecol. Evol*. 19, 101–108.

Leung, DH, Wang, YG & Zhu, M. 2009. Efficient parameter estimation in longitudinal data analysis using a hybrid GEE method. Biostatistics Jul;10(3):436-45.

Liang, K.Y. & Zeger, S.L. 1986. Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13–22.

Molenberghs, G. & Verbeke, G. 2005. Models for discrete longitudinal data. Springer, NewYork,NY.

R Development Core Team. 2012. R: A language and environment for statistical computing. R Foundation for statistical computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Sileshi, G. & Mafongoya, P.L. 2006. Variation in macrofaunal communities under contrasting land use systems in eastern Zambia. *Appl. Soil Ecol*. 31, 49–60.

Sileshi, G. & Mafongoya, P.L. 2007. Quantity and quality of organic inputs from coppicing leguminous trees influence abundance of soil macrofauna in maize crops in eastern Zambia. *Biol. Fertil. Soils* 43, 333–340.

Sileshi, G. 2008. The excess-zero problem in soil animal count data and choice of models for statistical inference. *Pedobiologia* 52, 1-17.

Swift, M. & Bignell, D. 2001. Standard methods for assessment of soil biodiversity and land use practice – Lecture Note 6b. International centre for research in agroforestry, Bogor, Indonesia.

Wang, YG & Carey, V. 2003. Working correlation structure misspecification, estimation and covariate design: Implications for generalised estimating equations performance. *Biometrika* 90*(*1*):* 29-41.

Wang, YG & Lin, X. 2005. Effects of variance-function misspecification in analysis of longitudinal data. *Biometrics* 61: 413-421.

Ziegler, A. Kastner C., Gromping U. & Blettner M. 1996. The generalized estimating equations in the past ten years: An overview and a biomedical application fttp://ftp.stat.uni-muenchen.de/pub/sfb386/paper24.ps.Z