

Choosing a Mixed Effects Model for Complex Time Series¹

Scelta di un Modello a Effetti Misti per Serie Storiche Complesse

Ornella Giambalvo

Università degli Studi di Palermo, e-mail: giambi@unipa.it

Antonella Puglia

Istat-Ufficio Regionale per la Sicilia, e-mail: antonellapuglia@libero.it

Riassunto: L'obiettivo del lavoro è modellare la complessa struttura di covarianza che caratterizza le serie storiche a misure ripetute dei dati di monitoraggio della concentrazione del nitrato nelle acque dell'estuario del *Clyde* (Scozia). L'effetto-tempo è stato esplicitato riproducendo i *pattern* osservati attraverso modelli a effetti misti nella loro formulazione estesa. I risultati dei processi di selezione per una versione lineare e una non lineare di tali modelli sono stati messi a confronto.

Keywords: mixed effects, nonlinear, environmental time series, correlated data.

1. Introduction

The statistical eco-physiological analysis on patterns of environmentally dangerous chemical components (*nutrients*) in water is particularly complex when the system investigated is a *transitional* water basin. This work is an attempt to individuate, with respect to the *Firth of Clyde*, which of several models - fixed, casual or mixed effects models and, further, linear or nonlinear forms - better matches relationships, effects and patterns in *nitrate* (NO₃) concentration data; this nutrient has been preferred to others because its presence/absence is directly imputable to human activity (being originated by domestic, industrial and, particularly, agricultural waste). The dataset it has been referred to is made of complex series developed in time and space: available data, irregularly collected from 1982 to 1996 by SEPA (*Scottish Environment Protection Agency*), register nutrients concentration in 30 areas (called stations) of the *Firth*. At each station and for each survey, variables are measured at consecutive depths (*Depth*). This design reproduces the typical structure of longitudinal data: each survey defines a group of correlated and ordered measurements. Given this clustered-like structure in data, it has been chosen to shape the complex covariance structure in nitrate (NO₃) series via mixed effect models, *extended* version. The modelling here developed, deals with data collected in *Garerloch North* (running 67 surveys and 7 depths), as this station is placed in the *Firth* area covering the river *estuary*; two seasonal environmental factors are involved as covariates: water temperature (*Temp*) and salinity (*Sal*).

¹ This paper is the result of the common work and reflections of the two authors. Nevertheless O. Giambalvo wrote section 2 and A. Puglia section 3.

2. Mixed effects models

Let $i = 1, 2, \dots, M$ denote the groups (surveys) and $j = 1, 2, \dots, n_i$ the repeated measures (depths); mixed effects models would reproduce the single response profile $\mathbf{y}_i$, whose generic component is y_{ij} , as follows:

$$\mathbf{y}_i | \mathbf{b}_i \sim F_i(\mathbf{v}, \mathbf{b}_i) \quad \mathbf{b}_i \sim G(\boldsymbol{\Psi}) \quad (1)$$

i.e., conditional on $\mathbf{b}_i$, $\mathbf{y}_i$ follows a pre-specified distribution F_i , depending on the selected covariates (the factors *Temp* and *Sal*) and parameterized through a vector $\mathbf{v}$ of unknown parameters, common to all subjects. The q -dimensional vector $\mathbf{b}_i$ is the *random* effects vector, a vector of group-specific unknown parameters; it is supposed distributed according to G and a vector of unknown parameters $\boldsymbol{\Psi}$ (Aerts *et al.*, 2002). On the basis of this general formulation, a class of mixed effects models can be obtained from introducing the random effects vector $\mathbf{b}_i$ (random intercepts and slopes), in the multivariate **linear** regression model with $\boldsymbol{\beta}$ as p -dimensional vector of *fixed* effects, i.e. intercepts and slopes common to the whole population (Laird and Ware, 1982). In the formal expression for it (2), $\mathbf{X}_i$ and $\mathbf{Z}_i$ are $n_i \times p$ and $n_i \times q$ -dimensional full rank design matrices for fixed and random effects, respectively. The **nonlinear** mixed effects model (3) extends the previous (2), in the extent it reproduces nonlinear relationships between the response vector and its covariates (Lindstrom and Bates, 1990). The *link* function f , real valued and differentiable, depends on the covariates vector $\mathbf{v}_i$ and is nonlinear in, at least, one of the components of the r -dimensional vector of group-specific parameters $\boldsymbol{\varphi}_i$. Same definitions are used for fixed and random effects, while new design matrices, $\mathbf{A}_i$ and $\mathbf{B}_i$, are defined: these group-specific $r \times p$ and $r \times q$ -dimensional matrices have values depending on covariates.

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\varepsilon}_i \quad i = 1, \dots, M \quad (2)$$

$$\mathbf{y}_i = f_i(\boldsymbol{\varphi}_i, \mathbf{v}_i) + \boldsymbol{\varepsilon}_i \quad \boldsymbol{\varphi}_i = \mathbf{A}_i \boldsymbol{\beta} + \mathbf{B}_i \mathbf{b}_i \quad i = 1, \dots, M \quad (3)$$

In both (2) and (3), the error components are two: $\boldsymbol{\varepsilon}_i$ is the n_i -dimensional vector of intra-group errors, supposed independently distributed from the second one, i.e. the random effects $\mathbf{b}_i$. The *extended* version of mixed effects models overcomes the strong assumptions on structure and distribution of intra-group errors characterizing the *general* formulation, by making possible to specify correlated and eteroscedastic errors. In fact, specifying the variance components, $\mathbf{b}_i \sim N(\mathbf{0}, \boldsymbol{\Psi})$ and $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \sigma^2 \boldsymbol{\Lambda}_i)$ where $\boldsymbol{\Lambda}_i = \mathbf{V}_i \mathbf{C}_i \mathbf{V}_i$, through properly selected *correlation* and *variance structures*, $cor(\boldsymbol{\varepsilon}_{ij}, \boldsymbol{\varepsilon}_{jk}) = [\mathbf{C}_i]_{jk}$ and $Var(\boldsymbol{\varepsilon}_{ij}) = \sigma^2 [\mathbf{V}_i]_{jj}^2$ ($j, k = 1, \dots, n_i$) respectively, one may reproduce in the model the variation sources suggested by Diggle *et al.* (1994) for longitudinal data (or similarly structured): random effects and coefficients, serial/spatial correlation and measurement error.

3. Results

The increasing-complexity driven model fitting flow has involved models generated both by integrating the **linear** relationships set and by choosing a proper **nonlinear** link function (3) among *close-to-linear*, two coefficients and one variable candidates. For each relationships set hypothesized, model building has been developed through a four steps process (model preliminary structures selection, assessment, modification and reduction). This process led for each model to the definition of proper:

- *intergroup* variability: evaluating which of the coefficients would consist of mixed or casual effects instead of purely fixed ones.
- *intragroup* variability: evaluating both the opportunity of modelling a possible dependence among observations made in the same occasion through a proper correlation structure, and the opportunity of modelling the empiric evidence for eteroscedastic measurement errors through a proper variance function.

Alternative models considered from time to time have been estimated through ML and compared through LRT and information criteria AIC and BIC ($\alpha=0.05$). Models eventually chosen, expressing patterns of NO3 as a depth (the experimental factor determining the grouped structure) function, are (see Puglia, 2004):

$$\begin{aligned} NO3_{ij} &= (\beta_0 + b_{0i}) + (b_{1i})Depth_{ij} + (\beta_2 + b_{2i})Temp_{ij} + (b_{3i})Sal_{ij} + \varepsilon_{ij} \\ Var(\varepsilon_{ij}) &= \sigma^2 \exp(2\delta NO3_{ij}) \end{aligned} \quad (4)$$

$$\begin{aligned} NO3_{ij} &= (\beta_1 + b_{1i} + \gamma_1 Temp_{ij} + \delta_1 Sal_{ij})Depth_{ij}^{(b_{2i})} + \varepsilon_{ij} \\ Var(\varepsilon_{ij}) &= \sigma^2 \exp(2\delta NO3_{ij}) \end{aligned} \quad (5)$$

See table 1 for parameters estimates. Though group responses follow a unique functional form, the NO3 patterns' seasonality, highlighted in explorative analysis, makes the random effects structure significant in both models; this structure determines through variable parameters all coefficients, and, in this way, both surface concentration and concavity of the relationship. The seasonal environmental factors, *Temp* and *Sal*, contribute in both models to defining survey-specific effects' heterogeneity. Estimates concerning common ($\beta_2, \gamma_1, \delta_1$) and survey-specific effects (b_{2i} and b_{3i} Std Dev) contributing to factors slopes would be balanced in order to determine negative coefficients so reproducing the physical fact of negative relationships between such factors and NO3. This complex definition of between-group variance allows to better defining the time effect (it makes possible to separate the effect of day-specific events-“survey effect”-from the effects of the period in the year-“season effect”) but binds global covariance decomposition. Practically, when introducing the previous mentioned seasonal factors, the covariance decomposition trade-off would be predominated by the random effects structure. This subtracts the necessity to hypothesize among variance sources a space-varying stochastic process generating serial correlation between measurements in the same unit: probably the grouping structure originates dependence among these measurements by the within-group correlated values seasonal variables take on. As a consequence, the residual covariance

would depend just on one source: the measurement errors. The evident heterogeneity of within-group errors variance, particularly marked in interseasonal comparisons, has been modelled in both (4) and (5) by the use of a variance function common to all groups: an increasing (see δ estimates) exponential of fitted values thus confirming the explorative evidence for greater intragroup variability in colder seasons. The choice between the two models cannot be based on estimates: classical analytical instruments for models comparison (criteria based on the log-likelihood) cannot be used on not-nested models, the variance heterogeneity does not allow a direct comparison between estimates of residuals standard deviations and graphical residuals analysis does not discriminate sufficiently. It can be argued that this choice would rather became a preference between the greater simplicity and the less computational efforts characterizing the selection process in the linear approach and the greater *flexibility* (nitrate-depth relationship features nonlinear patterns and alternate concavity reproducible through an exponential function with variable curvature parameter), *parsimony* (since the two models share the same variance function, the two sets of not known parameters would differ just on the number of effects declared in the main expression of the model) and *interpretability* (it is clearer expressed the role of time in determining each profile mean level and variability: surface concentration would depend on both “survey effect” and “seasonal effect”; relationship concavity would just be influenced from day events) in the nonlinear one.

Table 1: *Parameters estimates for linear model (4) and nonlinear model (5)*

	Fixed effects (Std Dev)		Random effects Std Dev				Res Std Dev	Var fun
Linear	β_0	β_2	b_{0i}	b_{1i}	b_{2i}	b_{3i}	σ	δ
	19.843	-0.868	70.204	0.069	0.964	2.289	0.701	0.052
	(1.778)	(0.184)						
Non linear	β_1	γ_1	δ_1	b_{1i}	b_{2i}		σ	δ
	42.81	-1.346	-0.509	7.861	0.141		1.063	0.04
	(2.769)	(0.133)	(0.099)					

References

- Aerts M., Geys H., Molenberghs G., Ryan L.M. (2002) *Topics in modelling of clustered data*, Chapman & Hall.
- Diggle P.J., Liang K., Zeger S.L. (1994) *Analysis of longitudinal data*, Oxford University Press.
- Laird N.M., Ware J.H. (1982) Random-effects models for longitudinal data, *Biometrics*, 38, 963-974.
- Lindstrom, M.J., Bates D.M. (1990) Nonlinear mixed-effects models for repeated measures data. *Biometrics*, 46, 673-687.
- Puglia A. (2004). *Modellazione di serie storiche ambientali complesse. Uno studio sulla concentrazione del nitrato nelle acque del fiume Clyde*, Ph.D. thesis in Applied Statistics-Università degli Studi di Palermo.