A study of the spatial variability of soil water retention by mixed effects linear models with a spatial continuous autoregressive correlation structure

Barbara Cafarelli[1], Annamaria Castrignanò[2], Nunzio Romano[3]

[1]Dipartimento di Scienze Economiche, Matematiche e Statistiche, Università degli Studi di Foggia, Piazza IV Novembre I, 70100 Foggia, Italy, b.cafarelli@unifg.it,
[2]Istituto Agronomico Sperimentale di Bari, Via Celso Ulpiani 5, 70125 Bari, Italy, annamaria.castrignano@tin.it,
[3]Dipartimento di Ingegneria Agraria Università di Napoli Federico II, Via Università 100, 80055 Portici (NA),Italy, nunzior@unina.it

Abstract: The knowledge of hydraulic properties of soil is necessary in many environmental applications and land planning. These properties, however, are difficult to determine and often they demand high labour costs, for which the tendency is to estimate them on the base of other more easily measurable or already available soil data. The level of detail reached using this method is not always satisfactory for some applications to basin scale, where variables to measure the morphologic property of the landscape are required. This study is proposed to characterize the spatial distribution of the water retention of a soil on wide scale using data relative to the physical, topographical and chemical characteristics of the soil within a model based approach

1 Introduction

The accurate evaluation of soil hydraulic characteristics is necessary for the reliable application of spatial distribution models, for the simulation of the soil water content in environmental studies and land planning (Romano and Santini 1997). This has also both economical and environmental important consequences, the former related to production costs and productivity enhancement, the latter to social costs and soil degradation. As the afore mentioned hydraulic characteristics are generally quite difficult to be determined and require high labour costs, they are often predicted by the use of other data readily available in soil surveys (such as soil texture, organic matter content and bulk density). In the standard agronomical practice this kind of data is generally treated by general linear models with uncorrelated errors. However soil hydraulic characteristics are known to be related to the landscape position at the catchment scale. In order to consider ex-

plicitly the spatial variability of hydraulic caractheristcs we have analysed this kind of data within a mixed effects linear models framework (Diggle et al. 1998). Infact these models, generally used to analyse correlated observations, represent a flexible tool to show the erraticity often influencing georeferentiated data separately from the spatial correlation and the systematic component of a spatial process. Then the estimated model was compared to those obtained without considering the spatial dependence.

2 Spatial linear mixed models

Let $\mathbf{Y} = \{\mathbf{Y}(\mathbf{u}) : \mathbf{u} \in D\}$ be a spatial stochastic process, where $\mathbf{u}$ is a spatial sampled location and $D \subset \mathbf{R}^2$. A mixed effects linear model is given by:

$$\mathbf{Y}(\mathbf{u}) = \mathbf{X}(\mathbf{u})\boldsymbol{\beta} + \mathbf{S}(\mathbf{u}) + \boldsymbol{\varepsilon}(\mathbf{u}) \tag{1}$$

where $\mathbf{X} = \mathbf{X}(\mathbf{u})$ is the matrix of spatially referenced non random predictors, $\boldsymbol{\beta}$ is the fixed effects parameter vector, $\mathbf{S} = \{S(\mathbf{u}) : \mathbf{u} \in D\}$ is a latent second order stationary Gaussian spatial process with $E(\mathbf{S}) = \mathbf{0}$ and variance covariance matrix $Var(\mathbf{S}) = \sigma_S^2 \mathbf{H}_{11}(k, \phi)$ defined by the continuous autoregressive correlation function of order one $h(s, \phi) = \phi^s$, where $s \geq 0$ is the distance between two observations on the transect and ϕ is the correlation parameter ($\phi \geq 0$) (Pinheiro and Bates 2000). The Gaussian random error vector $\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}(\mathbf{u})$ has $E(\boldsymbol{\varepsilon}) = 0$ and $Var(\boldsymbol{\varepsilon}) = \sigma_\varepsilon^2 \mathbf{I}$ and is uncorrelated with random effect **S**. The independence between **S** and $\boldsymbol{\varepsilon}$ causes the response variable components to be independent and normally distributed conditionally on **S**. So, if **y** stands for an *n*-dimensional finite realization of **Y**, its marginal model is given by

$$\mathbf{y} \sim \mathbf{N}_n(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}_{11}) \tag{2}$$

where $\text{var}(\mathbf{y}) = \text{var}(\mathbf{S}) + \text{var}(\boldsymbol{\varepsilon}) = \sigma_S^2 \mathbf{H}_{11}(k, \phi) + \sigma_\varepsilon^2 \mathbf{I} = \boldsymbol{\Sigma}_{11}$ (Pollice and Bilancia 2003). According to expression (2) the marginal likelihood for the trend parameter vector $\boldsymbol{\beta}$ and covariance structure parameters $\varpi = (\sigma_S^2, \phi, \sigma_\varepsilon^2)$ is the same as that of a general linear model with correlated errors. So, if we assume $\boldsymbol{\Sigma}_{11}$ as known, the Aitken GLS estimator of $\hat{\boldsymbol{\beta}}$ is a BLUE. However, this assumption obviously leads to the underestimation of the variability of $\hat{\boldsymbol{\beta}}$. A way to get around this problem is to consider the REML estimators which involves the maximisation of a likelihood based on residuals $\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ with respect to the spatial correlation parameters and allows to achieve unbiased estimates of the latter.

3 Spatial predictions and measurement error

Suppose to predict the behaviour of the process at *m* unsampled locations and let $\mathbf{y}_0$ be the corresponding unknown random vector. Under the same modelling assumptions $\mathbf{y}_0 \sim \mathbf{N}_m(\mathbf{X}_0\boldsymbol{\beta}, \boldsymbol{\Sigma}_{00})$, where $\mathbf{X}_0$ and $\boldsymbol{\Sigma}_{00}$ are the known design matrix and the variance covariance matrix associated with unsampled locations. The joint distribution of the response variable at sampled and unsampled locations is then given by:

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{y_0} \end{pmatrix} = \mathbf{N}_{n+m}\left(\begin{bmatrix} \mathbf{X}\boldsymbol{\beta} \\ \mathbf{X}_0\boldsymbol{\beta} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{10} \\ \boldsymbol{\Sigma}_{01} & \boldsymbol{\Sigma}_{00} \end{bmatrix}\right) \tag{3}$$

where $\boldsymbol{\Sigma}_{10} = \sigma_S^2\mathbf{H}_{10}(k,\phi)$ and $\boldsymbol{\Sigma}_{00} = \sigma_S^2\mathbf{H}_{00}(k,\phi) + \sigma_\varepsilon^2\mathbf{I}$, where the covariance between observed and unobserved locations $\boldsymbol{\Sigma}_{10}$ is not supposed to be influenced by the measurement error (Christensen, 1991). The linear mixed model predictor is

$$\hat{\mathbf{y}}_0 = \mathbf{X}_0\hat{\boldsymbol{\beta}} + \boldsymbol{\Sigma}_{01}\boldsymbol{\Sigma}_{11}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) \tag{4}$$

where $\hat{\boldsymbol{\beta}}$ and $\hat{\varpi}$ are estimated by REML. An analogous predictor $\hat{\mathbf{y}}$ can be used to obtain a noiseless prediction at sampled locations by substituting $\mathbf{X}_0\hat{\boldsymbol{\beta}}$ with $\mathbf{y}$ and $\boldsymbol{\Sigma}_{01}$ with $\sigma_S^2\mathbf{H}_{11}(k,\phi)$ in expression (4). Such predictor allows to obtain residuals as the difference between observed and fitted values $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$, to be used as a further check of the fitted model in the following case study (Pollice and Bilancia 2003).

4 Case study: the distribution of the water content on a transect sample

The study area is located on a hill slope of the Sauro river catchment in Southern Italy. The soils develop mainly in xeric moisture and mesic temperature regimes, and parent materials consist mostly of clayey components. From the geomorphological point of view the area is quite heterogeneous with slope gradients ranging from -0.13 to 0.44 and elevation from 458 m to 1,073 m. Soil samples were collected from topsoil along a transect at 100 locations 50 m apart. Standard laboratory measurements were used to determine the following soil properties: texture (%), bulk density (g cm^{-3}), particle density (g cm^{-3}), porosity, organic carbon content (%). The measured topographic variables were elevation (m) and slope. The volumetric water content at –0.1 m pressure head was measured with a suction table. Exploratory data analysis showed that water retention data had a linear trend in respect of the position by the transect side (Figure 1) and a marginal bell shaped distribution (Figure 2). Both considerations lead to analyse the possible relation between water retention data and the topographical (elevation and slope), physical

(bulk density, particle density, porosity, clay, silt and sand contents) and chemical (organic carbon) variables by linear mixed effects models with an autoregressive correlation structure to express the dependance among the observed values collected along the transect. The model assumptions were assessed by the comparison of the residuals and the estimated random effects with the quantilies of standard normal (figure 3 and 4). The estimates of the model correlation parameters are reported in table 1. The inspection of table 2 shows that bulk density and porosity are negatively linked to water retention. On the contrary, the clay and silt content and particle density are positively linked, which means that textural classes may significantly affect the soil's hydraulic properties. The dependence of the -0.1 m water retention on topographic variables resulted significantly only for altitude. This dependence probably exists because the –0.1 m water retention exhibits a dependence on soil textural components and because substantial differences in the soil structure were encountered along the transect at the different elevations (Romano and Palladino 2002). The estimated model was then compared with a general linear model with uncorrelated errors, such as those generally used for this kind of data in the standard agronomical practice. The comparison of the two models showed that the residuals of the fitted mixed model were clearly smaller than those obtained by a general linear model (figure 5) and that the model residuals were a white noise at a significative level of 0.05 (figure 6). As a consequence of these considerations, the fitted values of the linear mixed model with the autoregressive correlation structure were clearly much more close to the observed values then which fitted by the linear model (figure 7).

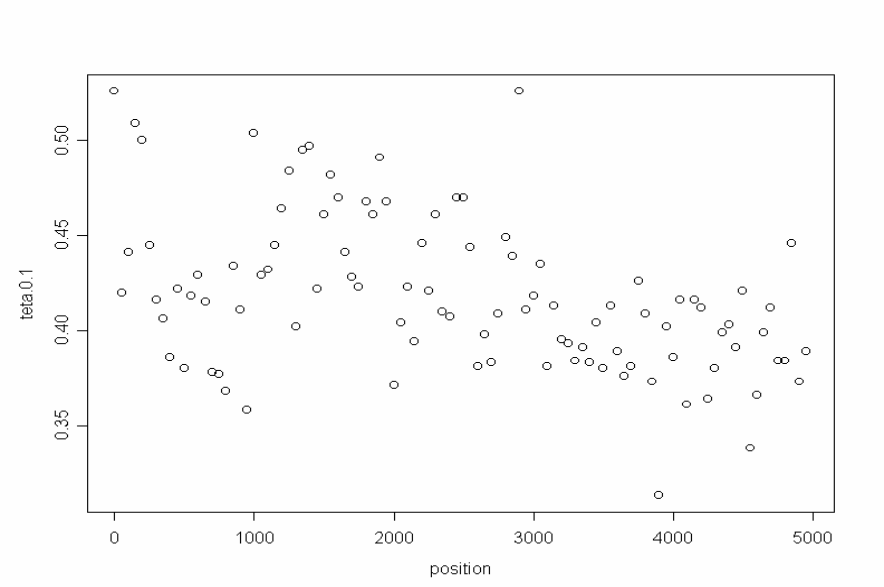

Fig. 1. Spatial distribution of the soil water retention respect to the 100 sites

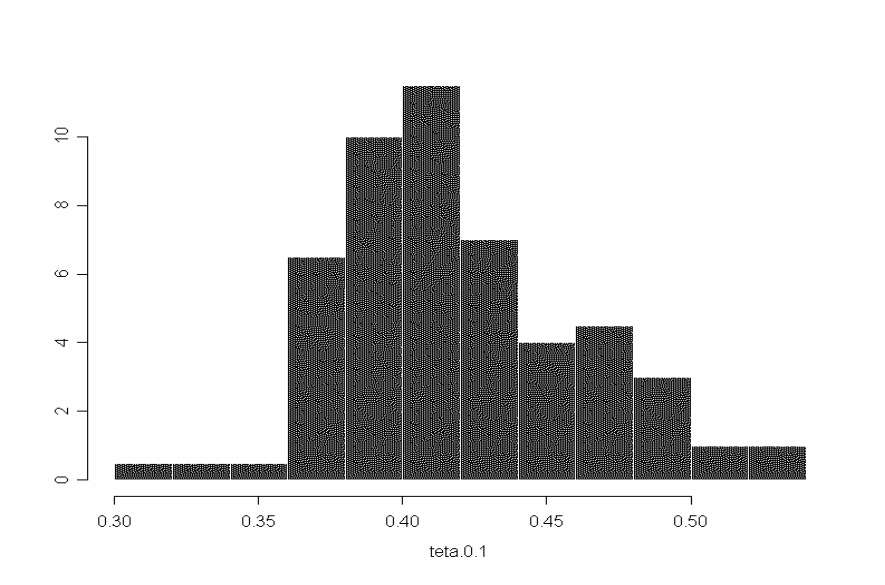

Fig. 2. Marginal distribution of soil water retention

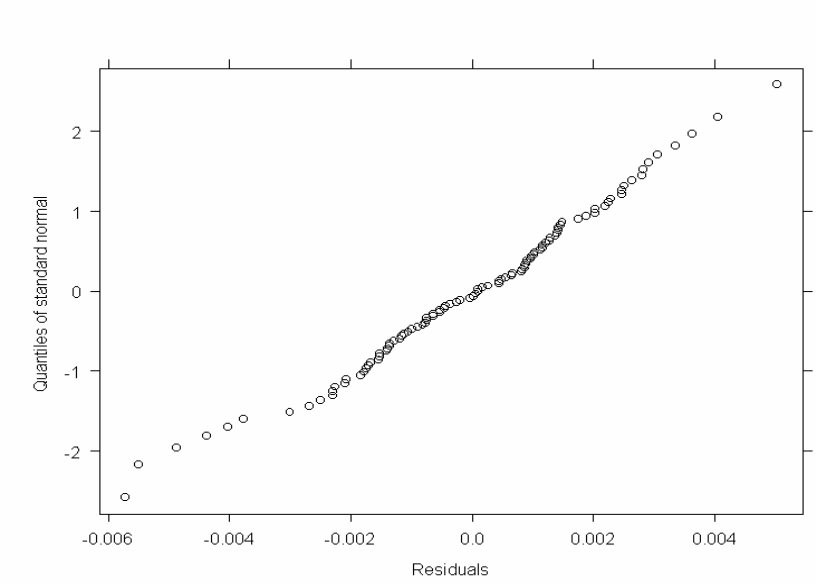

Fig. 3. Normal plot of residuals for the fitted linear mixed model

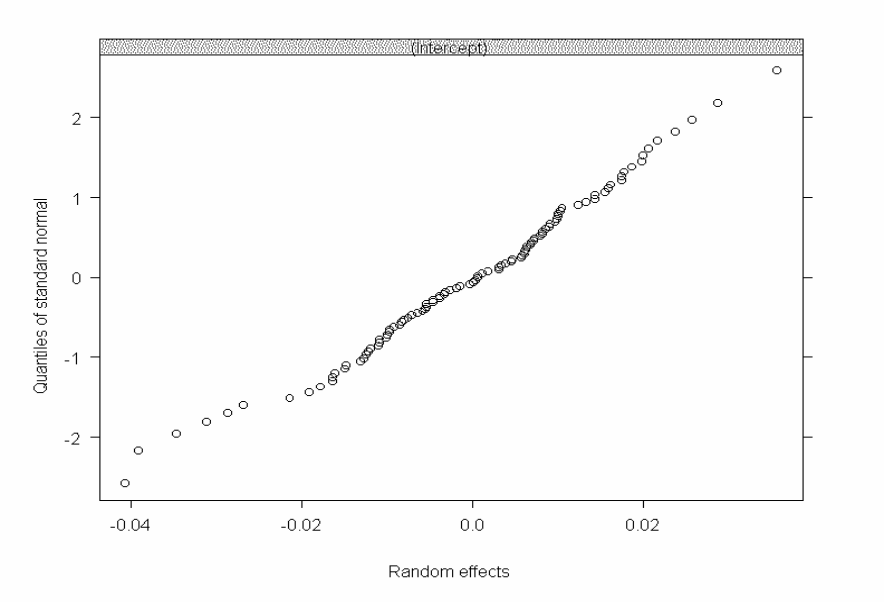

Fig. 4. Normal plot of random effects for the fitted linear mixed model

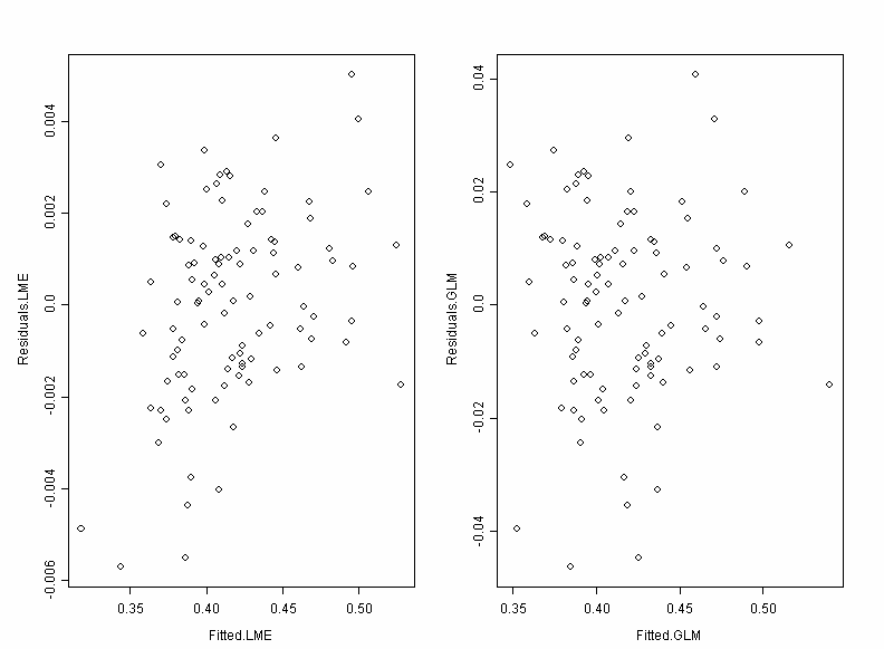

Fig. 5. Left: The scatter plot of the fitted mixed model residuals versus fitted values. Right: The scatter plot of the fitted general linear model residuals versus fitted values.

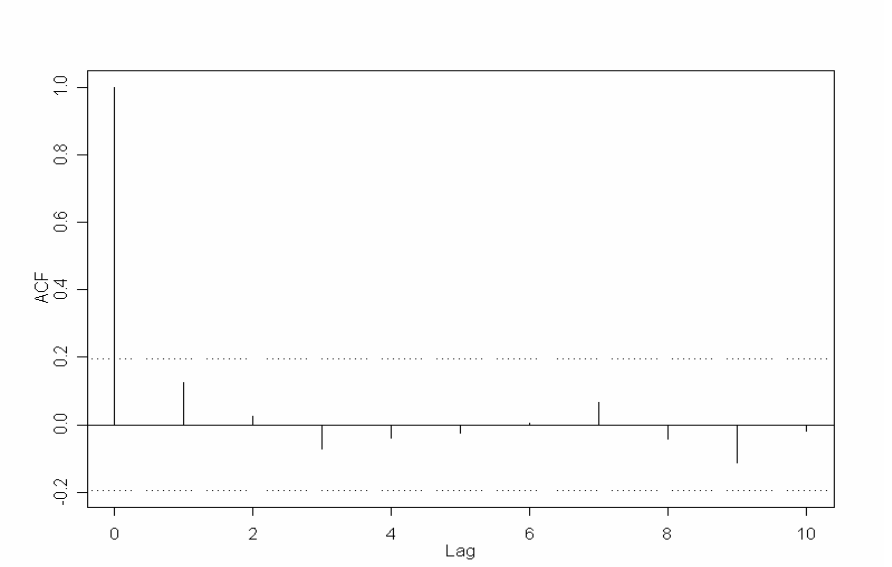

Fig. 6. Empirical autocorrelation function of the normalized residuals for the spatially correlated model with the significance level for critical bounds (alpha=0.05).

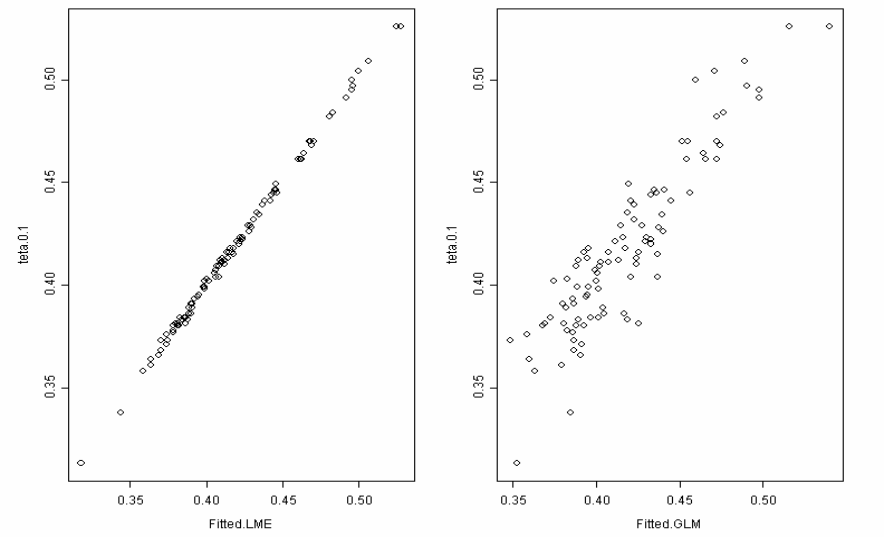

Fig. 7. Left: The plot of the observed values versus mixed model fitted values. Right: The plot of the observed values versus general linear model fitted values.

Table 1. CAR correlation parameters and 95% approximate confidence bounds

	lower	estimates	upper
σ_S	0.00233	0.01153	0.05700
ϕ	0.32901	0.39293	0.46073
σ_ε	0.02740	0.02989	0.03261

Table 2. Significance of tests on fixed effects

Fixed Effects	Value	Std. Error	p-value
Intercept	1,505711	0,573861	0,0102
Elevation	0,000052	0,000018	0,0041
Bulk density	-1,565164	0,506142	0,0026
Particle density	0,974541	0,28129	0,0008
Porosity	-3,251124	1,278185	0,0127
Clay	0,000708	0,000214	0,0014
Silt	0,000862	0,000279	0,0027

Conclusions

The proposed approach results to be a quick and effective method to predict soil hydraulic characteristics by other more easily available variables. The paper shows an application of the linear mixed model approach to the estimation of the volumetric water content at –0.1 m pressure head. Different types of statistical tests have proved the necessity to take into account error correlation structure and terrain parameters in fitting regression model. Therefore, a wider use of linear mixed models is recommended in hydrology for characterising the relationship of water content to matric soil water pressure.

References

Christensen R (1991) Linear models for multivariate, time series and spatial data. Springer–Verlag, New York

Diggle PJ, Moyeed RA and Tawn JA (1998) Model-based Geostatistics. Applied Statistics, 47: 299-350

Romano N, Santini A (1997) Effectiveness of using pedo-transfer functions to quantify the spatial variability of soil retention characteristics. Journal of Hydrology, 202: 137-157

Romano N, Palladino M (2002) Prediction of soil water retention using soil physical data and terrain attributes. Journal of Hydrology, 265: 56-75

Pinheiro JC, Bates DM (2000) Mixed-Effects Models in S and S-Plus. Statistics and Computiung. Springer, New York

Pollice A, Bilancia M, (2003) Kriging with mixed effects models, Statistica, LXII, 3, 405-429