Fitting soil-water characteristic curves by truncated normal nonlinear regression models

Carolina C. M. Paraíba[1], Carlos A. R. Diniz[2], Aline H. N. Maia[3] and Lineu N. Rodrigues[4]

1 Introduction

In the present paper, we propose an alternative approach for estimating SWCC based on generalized nonlinear models, assuming that the response variable follows a truncated normal distribution. The parameters of the curve are estimated by maximum likelihood method. Simulation studies are provided to assess the quality of estimates for the proposed regression model and to assess the behavior of least squares procedures when applied to simulated truncated data sets. A real data set is analyzed using the proposed methodology. We also propose a diagnostic analysis to check the underlying model assumptions, outliers, and influent observations for the proposed truncated normal nonlinear regression model. Thus, the standardized residuals and the Pearson residuals (Cook & Weisberg, 1982), as well as two metrics based on the principle of case-deletion first proposed by Cook (1977) are used for outliers and influent case detection and to check for model adequacy.

2 Soil-water characteristic curves

SWCCs are fitted considering pairs, (y, x), which are obtained by applying different tensions, x, to the a given soil sample, and observing the remaining soil-water content. In studies to determine SWCCs, the analytical expressions considered are nonlinear functions of the type $y = \eta(x, \boldsymbol{\beta})$ where $\boldsymbol{\beta}$ is the vector of parameters of the curve. The van Genuchten-Mualem (van Genuchten, 1980; Mualem, 1976) expression for SWCCs is given by

$$y = \theta_r + \frac{\theta_s - \theta_r}{\left[1 + (\beta_1 x)^{\beta_2}\right]^{1-\frac{1}{\beta_2}}}, \tag{1}$$

In van Genuchten (1980), the author highlights that θ_s is easily obtained experimentally, being available most of the times, whereas θ_r is defined as the soil-water content at $x = -15atm$.

3 Truncated normal nonlinear regression model

Let Y be a normal random variable (r.v.) with mean μ and standard deviation σ. If, in addition, $a < Y < b$, then Y is a truncated normal r.v. (Greene, 2003) with density function

[1]PhD Student at Departamento de Estatística, UFSCar. e-mail: *carolina_paraiba@yahoo.com.br*
[2]Departamento de Estatística, UFSCar
[3]Embrapa, Meio Ambiente
[4]Embrapa, Cerrados

given by

$$f\left(y\,|a<y<b\right)=\frac{1}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right)\left[\Phi\left(\frac{b-\mu}{\sigma}\right)-\Phi\left(\frac{a-\mu}{\sigma}\right)\right]^{-1}\mathrm{I}_{(a,b)}\left(y\right), \tag{2}$$

where ϕ is the standard normal distribution and Φ the standard normal cumulative distribution function. We shall denote a truncated normal distribution by $TN(\mu,\sigma,a,b)$.

Assume that $\mu=\eta\left(\boldsymbol{x},\boldsymbol{\beta}\right)$ and $\sigma=g\left(\boldsymbol{x}_q,\boldsymbol{\lambda},\sigma\right)$, where $\boldsymbol{x}=\left(x_1,\ldots,x_p\right)'$ is a vector of p covariates, $\boldsymbol{x}_q$ is a subset of x, $\eta\left(\cdot\right)$ is a continuous and twice differentiable function with respect to $\boldsymbol{\beta}$, $f\left(\cdot\right)$ is a continuous and twice differentiable function with respect to $\boldsymbol{\lambda}$ and σ, and $\boldsymbol{\theta}=\left(\boldsymbol{\beta},\boldsymbol{\lambda},\sigma\right)$ is the vector of indexing parameters. Let $\boldsymbol{y}=\left(y_1,\ldots,y_n\right)'$ be a vector of observed values of $\boldsymbol{Y}=\left(Y_1,\ldots,Y_n\right)'$, where $Y_i\sim TN\left(\eta\left(\boldsymbol{x}_i,\boldsymbol{\beta}\right),g\left(\boldsymbol{x}_{qi},\boldsymbol{\lambda},\sigma\right),a,b\right)$, $i=1,\ldots,n$. Then, given the data set $D=\left(n,\boldsymbol{y},\boldsymbol{x}\right)$, the log-likelihood function for $\boldsymbol{\theta}$ is written as

$$\begin{aligned}l\left(\boldsymbol{\theta}\right)=&-\frac{n}{2}\log 2\pi-\sum_{i=1}^{n}\log g\left(\boldsymbol{x}_{qi},\boldsymbol{\lambda},\sigma\right)-\frac{1}{2}\sum_{i=1}^{n}\left[\frac{y_i-\eta\left(\boldsymbol{x}_i,\boldsymbol{\beta}\right)}{g\left(\boldsymbol{x}_{qi},\boldsymbol{\lambda},\sigma\right)}\right]^2\\&-\sum_{i=1}^{n}\log\left[\Phi\left(\frac{b-\eta\left(\boldsymbol{x}_i,\boldsymbol{\beta}\right)}{g\left(\boldsymbol{x}_{qi},\boldsymbol{\lambda},\sigma\right)}\right)-\Phi\left(\frac{a-\eta\left(\boldsymbol{x}_i,\boldsymbol{\beta}\right)}{g\left(\boldsymbol{x}_{qi},\boldsymbol{\lambda},\sigma\right)}\right)\right].\end{aligned} \tag{3}$$

The maximum likelihood estimates (MLEs) of $\boldsymbol{\theta}$ can be obtained by direct nonlinear optimization of (3).

4 Diagnostic analysis

In regression analysis, diagnostic procedures are aimed to check if the underlying assumptions of a proposed model are reasonable enough and to detect evidences of possible model misspecification. The regression model constructed in Section 3 was based in the following underlying assumptions: (i) the response variable, $\boldsymbol{y}$ follows a $TN\left(\eta\left(\boldsymbol{x}_i,\boldsymbol{\beta}\right),g\left(\boldsymbol{x}_{qi},\boldsymbol{\lambda},\sigma\right),a,b\right)$ distribution, $i=1,\ldots,n$; (ii) the observations are mutually independent. Model assumptions may be checked by visual inspections of several residual plots, such as the standardized residuals and the Pearson residuals, which are also useful to detect outliers and possible influent observations. We also consider the generalized Cook distance and the likelihood distance to detect influent observations.

5 Simulation study

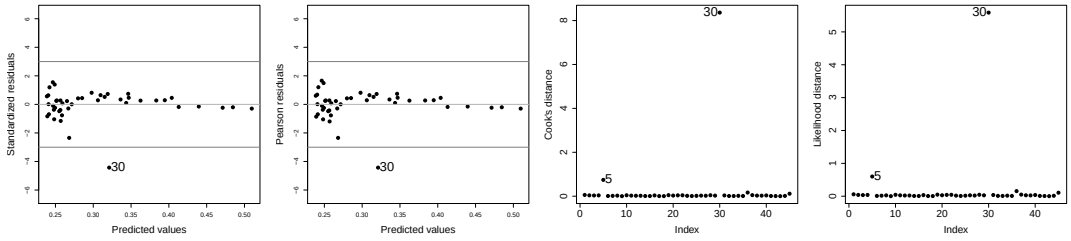
A simulation study was conduced to assess the quality of MLEs for the proposed model under different sample sizes. Our main goal is to assess the quality of MLEs and study its frequentist properties. Data sets were generated mimicking the characteristics of the real data set to be analyzed later. Thus, we consider a explanatory variable, x, with $k=9$ tension levels ranging from 0,01 to $15atm$, and $r=\{1;3;5;55\}$ replications which give us $n=\{9;27;45;495\}$. For each replication we set the residual soil-water content (the lower truncation limit), θ_r, as

a value generated from a $Uniform\,(0,20;0,25)$, and the soil-water content at saturation (the upper truncation limit), θ_s, as a value generated from a $Uniform\,(0,40;0,65)$.

We shall consider $\eta\,(x,\boldsymbol{\beta})$ as the van Genuchten-Mualem model given in (1), and $f\,(x_q,\lambda,\sigma) = \sigma x^{\lambda}$. Thus, $Y_i \sim TN\left(\eta\,(x_i,\boldsymbol{\beta})\,,\sigma x_i^{\lambda},y_r,y_s\right)$, for $i=1,\ldots,n$. The model parameters are subjected to the following restrictions: $\beta_1>0$, $\beta_2>1$ and $\sigma>0$.

From the results obtained in the simulation study (not shown) we notice that as sample size increases, both the bias and MSE decreases and coverage probabilities approaches the expected nominal one of 95%. In order to illustrate the consequences of disregarding the effects of truncation, we also considered the usual nonlinear regression fitted by LS to the same simulated data sets. The obtained results (not shown) revealed parameter estimates as highly biased and inaccurate, and the estimated coverage probability was far from 95%. Moreover, as sample size increases, we notice that coverage probabilities approaches zero. This phenomenom occurs since the standard error of parameters are very small for large samples; thus producing very small confidence intervals. We can also observe that parameter estimates for β_1 are more imprecise than those obtained for β_2.

For model diagnostic we take one data set with $n=45$ where two observations, 5 and 30, were deliberately transformed into atypical ones by recalculating x_5 and x_{30} by adding 10 times the standard deviation of x to their original values. The main idea is to assess the effectiveness of the different model diagnostics considered and to illustrate their ability to detect influential outliers observation when the truncated normal nonlinear regression model is fitted to a data set. Figures 1a-1b show that observation 30 is identified as an outlier. From Figures 1c-1d it is possible to see that both the approximated generalized Cook's distance and approximated likelihood distance detect disturbed cases, 5 and 30, as influent observations.



(a) (b) (c) (d)

Figure 1: Perturberd simulated data set: (a) standardized residuals; (b) Pearson residuals; (c) approximated generalized Cook distance; (d) approximated likelihood distance.

6 Real data set analysis

In this section we analyze the soil profile data set selected from a database collected in the Buriti Vermelho River Basin, located in the eastern part of the Federal District in Brazil (Rodrigues & Maia, 2011). The data set consists of soil samples of $0-5cm$, $15-20cm$, and

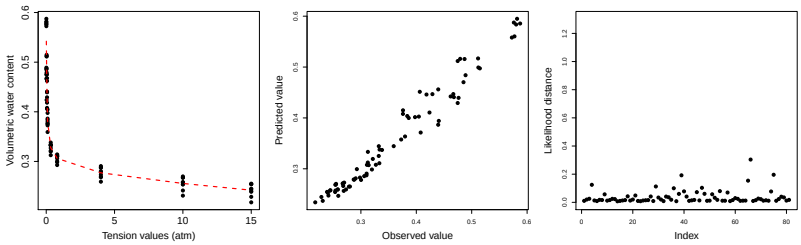
$60 - 65cm$ deep measured in $k = 9$ tension levels with $r = 3$ replications per level, giving a total of 27 soil water content measurements for a total of 17 soil profiles.

We shall consider the location parameters $\eta(x, \boldsymbol{\beta})$ as the Van Genuchten-Mualem model given in (1). Soil water content at saturation, θ_s, were calculated by weighing the soil profile samples directly. The residual soil water content, θ_r, were calculated by submitting the soil samples to a tension of 1500 kPa.

Model fit summary provided in Table 1 indicate all parameters in the heteroscedastic truncated normal van Genuchten-Mualem regression model as statistically significant with a 95% confidence. From the estimated SWCC presented in Figure 2a, it is possible to see that the van Genuchten-Mualem model is a good choice for the representation of the relationship between soil-water content and matric potencial for the analyzed soil profile. Predicted against observed values are depicted in Figure 2b, indicating that the predicted values are reasonably close to the observed values of the response variable. Moreover, the standadized residual plots show the residuals as randonly distributed around zero with no outlier observations. We also note that no influent observation was depicted by Cook's generalized distance in Figure 3c and by the likelihood distance in Figure 3d.

Table 1: Model fit summary for the heteroscedastic van Genuchten-Mualem truncated normal regression model adjusted to soil profile 204 data.

Parameter	Estimate	St. Dev.	95% C.I.	
β_1	49,9135	4,9812	40,1504	59,6766
β_2	1,5077	0,0238	1,4611	1,5544
σ	0,0159	0,0015	0,0128	0,0189
λ	-0,1495	0,0391	-0,2262	-0,0728

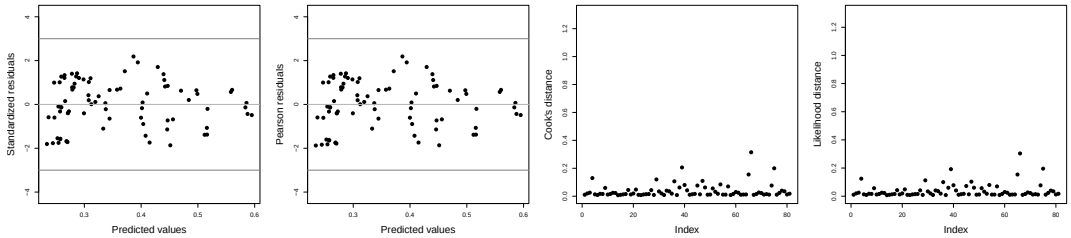


(a) (b) (c)

Figure 2: Profile 204 data: (a) estimated SWCC; (b) Observed y against predicted values of y.

7 Conclusions

We have proposed and illustrated an alternative approach to model SWCCs based on truncated normal nonlinear regression models, which take truncation into account, an important



(a) (b) (c) (d)

Figure 3: Profile 204 data: (a) standardized residuals; (b) standardized Pearson residuals; (c) approximated Cook's generalized distance; (d) approximated likelihood distance.

feature of the data. The ML estimation procedure have successfully been applied and a simulation studies was provided to assess the quality of estimates for the proposed nonlinear regression model. Moreover, diagnostic analysis tools were used to check the model assumptions and for outlier and influent observations detection. When comparing the proposed methodology and the usual nonlinear least squares procedure based on simulation results, it was verified that LS method does not provide precise estimates for the model parameters, thus leading to an inaccurate estimation of the SWCC. Nevertheless, we acknowledge that the truncated normal is one of many truncated distributions that can be considered to model soil-water retention data, such as the truncated beta and truncated inverted beta distributions. Also, we could consider the truncated version of some recently propose skewed distributions.

References

Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics*, **19**(1), 15–18.

Cook, R. D. & Weisberg, S. (1982). *Residuals and influence in regression*. Chapman & Hall, New York.

Greene, W. H. (2003). *Econometric Analysis*. Prentice Hall, New Jersey, fifth edition.

Mualem, Y. (1976). A new model for predicting the hydraulic conductivity of unsaturated porous media. *Water Resources Research*, **12**, 593–622.

Rodrigues, L. N. & Maia, A. H. N. (2011). Funções de pedotransferência para estimar a condutividade hidráulica saturada e as umidades de saturação e residual do solo em uma bacia hidrográfica do cerrado. In *XIX Simpósio brasileiro de recursos hídricos*.

van Genuchten, M. T. (1980). A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Science Society of America Journal*, **44**, 892–898.