Truncated beta nonlinear regression models for soil-water characteristic curves estimation

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

Abstract

We propose an alternative approach for estimating soil-water characteristic curves based on truncated beta nonlinear regression models. Thus, assuming that the response variable follows a truncated beta distribution. Maximum likelihood estimator of the curve parameters are obtained by direct maximization of the likelihood function and diagnostic analysis tools are considered to check for model adequacy under the two considered models. A soil profile from the Buriti Vermelho River Basin database is analyzed using the proposed methodology.

Key-words: Truncated normal, truncated beta, nonlinear generalized models, soil-water characteristic curves.

1 Introduction

A soil-water characteristic curve (SWCC) is a useful graphical tool which describes the amount of water remaining in the soil (water volume content) as a function of the soil water tension (matric potential). These curves are usually estimated by nonlinear regression models fitted to data sets obtained from laboratory experiments or from pedotransfer functions. The most widely used method for estimating the parameters of a SWCC is the nonlinear least squares (LS) method (Yates *et al.*, 1992; Dourado-Neto *et al.*, 2000; Cornelis *et al.*, 2005; Silva *et al.*, 2006; Chao *et al.*, 2008). However, given the nature of the SWCC data, it is known that the observed water content at a matric potential will be such that it is not less than the residual soil-water content, and no more than the saturated soil-water content; therefore, the data is subjected to a phenomenon known in statistics as truncation. As argued in Maddala (1983), it is important to account for truncation in regression analysis since usual least squares estimates (LSEs) can be biased,

[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

inefficient, and inconsistent, which can seriously affect the estimated curve and prediction based on it.

As argued in Greene (2003), truncation is a characteristic of the probability distribution from which the sample data are drawn. Therefore, to account for the truncated nature of the observed data we must consider a truncated probability distribution, which is the part of a distribution that is above, below or between some specified value. In common truncated regression the effects of truncation arise when inferences about a population are based on a sample drawn from a subset of the population. For example, Hausman & Wise (1977) present a study of income where only households with income below a certain poverty line are part of the sample, and A'Hearn (2004) analyzes historical height data drawn from military records which are truncated from below since armies imposed a minimum height requirement. On the other hand, truncation in water retention data occurs naturally since, given the residual soil-water content and the soil-water content at saturation, the soil-water content at any other matric potential will be truncated between this two values.

In the context of linear truncated regression models, a variety of contributions is available in the literature and a number of different approaches have been considered to estimate the model parameters. Heckman (1976) proposes a corrected least squares estimator, where the bias produced from applying the LS procedure to truncated regression models is characterized as a specification error or an omitted variable problem. Thus, the corrected LS estimator is constructed by including the omitted variable as a regressor. Since usual LS estimator are known to be biased, a popular choice for the estimation of such models are the method of maximum likelihood or likelihood based methods. In Hausman & Wise (1977), the authors propose a maximum likelihood procedure and provide a Newton-type algorithm to obtain the parameter estimates. In Amemiya (1985), the method of maximum likelihood is used to estimate the model parameters. In A'Hearn (2004) a restricted maximum likelihood estimator is proposed and applied to height samples data. This restricted ML estimator imposes an a priori value on the standard deviation of the response variable while estimating its mean freely. The author also uses simulation results to show that his proposed methodology behaves as the restricted ordinary least squares. There is also a great deal of contributions in semiparametrics and nonparametrics estimation of truncated linear regression model (e.g. Powell (1986), Lee (1992), Lee (1993), Newey

(2004), Cosslett (2004), and Chen & Zhou (2012)).

Moreover, since the statistical analysis is conditioned on the probability model considered and can be misleading if the assumed model is not plausible enough, it is important to conduct diagnostics to assess model adequacy. As argued in Ritz & Streibig (2008), it is known that substantial departures from model assumptions could result in biased and inaccurate estimates and distorted standard errors. Thus, we propose a diagnostic analysis to check the underlying model assumptions, outliers, and influent observations for the proposed truncated beta nonlinear regression model. Following the diagnostic methodology of generalized linear models (GLMs) and usual regression analysis, we consider the standardized residuals (Cook & Weisberg, 1982) for outliers detections and to check for model adequacy. We also consider two metrics for influent observations detections based on the principle of case-deletion first proposed by Cook (1977).

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 beta distribution. The parameters of the curve are estimated by maximum likelihood method and diagnostic analysis are conduced to check for model adequacy. To illustrate the proposed methodology, we analyze a soil profile from the Buriti Vermelho River Basin database presented by Rodrigues & Maia (2011). The Buriti Vermelho River Basin is located in the eastern part of the Federal District in Brazil and the data sets consists of samples from 17 soil profiles collected in three different soil depth for which soil-water content were measured at nine tension level ranging from $0,01atm$ to $15atm$, with three replications per level. The soil profiles represent the sites were the soil sample were collected along the considered region.

The paper is organized as follows. In Section 2, we give a brief introduction on SWCCs and describe the van Genuchten model with the Mualem restriction. In Section 3, we present the truncated beta regression model, which shall be considered for estimating SWCCs taking into account the truncated nature of soil-water retention data. In Section 5, we present the analysis of a real data set, providing the inferences about the parameters and the diagnostic analysis of the fitted model. Finally, in Section 6, we give a few brief concluding remarks.

2 The van Genuchten-Mualem soil-water characteristic curve

When constructed in laboratory using observed data, SWCCs are fitted considering pairs, (y, x), which are usually obtained by applying different tensions, x, to the a given soil sample, and observing the water content, y, remaining in the sample after application of each tension level considered. Thus, a SWCC relates a variable response, y, with a regressor variable, x. 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 relationship between water volume content and matric potential is not trivially modeled and several analytical expressions have been proposed in the literature for representing the SWCC. Among the most widely used expressions for SWCCs are the ones proposed by Gardner (1958), Brooks & Corey (1964), van Genuchten (1980), and Fredlund & Xing (1994). These expressions are preferred since they provide a good approximation of the relationship between the amount of water in the soil and soil suction. We refer to Leong & Rahardjo (1997) and Sillers *et al.* (2001), for a revision of different expression proposed to model SWCCs. In this paper, we consider the model proposed by van Genuchten (1980) combined with the relation given in Mualem (1976) - hereafter van Genuchten-Mualem model.

The van Genuchten expression is given by

$$y = \theta_r + \frac{\theta_s - \theta_r}{\left[1 + (\beta_1 x)^{\beta_2}\right]^{\beta_3}}, \tag{1}$$

where β_1 is related to the inverse of the air entry value, β_2 is related to the pore-size distribution of the soil and β_3 is related to the asymmetry of the model.

Mualem (1976) proposed a fixed relationship between β_2 and β_3 given by

$$\beta_3 = 1 - \frac{1}{\beta_2}, \tag{2}$$

since $\beta_3 > 0$, β_2 must be greater than 1.

In van Genuchten (1980), the author highlights that y_s is easily obtained experimentally, being available most of the times, whereas y_r is defined as the soil-water content at $x = -15atm$ (van Genuchten, 1980), or as a fitting parameter equal the soil-water content where the first derivative of y with respect to x, dy/dx, equals zero (van Genuchten & Nielsen, 1985).

3 Truncated beta nonlinear regression model

The truncated beta nonlinear model is constructed based on the beta regression model of Ferrari & Cribari-Neto (2004), where the beta distribution is reparameterized in terms of a mean and a dispersion parameter.

As in Ferrari & Cribari-Neto (2004), we consider Y be a $Beta\left(\alpha,\gamma\right)$ r.v. with probability density distribution given by

$$f\left(y\right) = \frac{\Gamma\left(\alpha+\gamma\right)}{\Gamma\left(\alpha\right)\Gamma\left(\gamma\right)} y^{\alpha-1}(1-y)^{\gamma-1}\mathrm{I}_{(0,1)}\left(y\right),$$

Letting $\mu = \alpha/\left(\alpha+\gamma\right)$ and $e^{\phi} = \alpha+\gamma$, it follows that

$$E\left(Y\right) = \frac{\alpha}{\alpha+\gamma} = \mu, \tag{3}$$

and

$$Var\left(Y\right) = \frac{\alpha\gamma}{\left(\alpha+\gamma\right)^2\left(\alpha+\gamma+1\right)} = \frac{\mu\left(1-\mu\right)}{1+e^{\phi}}. \tag{4}$$

Therefore, μ is the mean parameter and e^{ϕ} is the dispersion parameter, and the probability density distribution of Y can be rewritten as

$$f\left(y\right) = \frac{\Gamma\left(e^{\phi}\right)}{\Gamma\left(\mu e^{\phi}\right)\Gamma\left(\left(1-\mu\right)e^{\phi}\right)} y^{\mu e^{\phi}-1}(1-y)^{(1-\mu)e^{\phi}-1}\mathrm{I}_{(0,1)}\left(y\right), \tag{5}$$

with $0 < \mu < 1$ and $\phi \in \mathbb{R}$.

If Y is truncated to a known interval (a,b), then the distribution of Y given $a < Y < b$,

$0 < a < b < 1$, denoted by $TB(\mu e^{\phi}, (1-\mu)\, e^{\phi}, a, b)$, is written as

$$f(y) = \frac{\Gamma\left(e^{\phi}\right)}{\Gamma\left(\mu e^{\phi}\right)\Gamma\left((1-\mu)\, e^{\phi}\right)} y^{\mu e^{\phi}-1}(1-y)^{(1-\mu)e^{\phi}-1}$$
$$\left[I\left(b; \mu e^{\phi}, (1-\mu)\, e^{\phi}\right) - I\left(a; \mu e^{\phi}, (1-\mu)\, e^{\phi}\right)\right]^{-1} \mathrm{I}_{(a,b)}(y), \tag{6}$$

where $I(t; \kappa, \tau) = B(t; \kappa, \tau)/B(\kappa, \tau)$, with $B(\kappa, \tau) = \int_0^1 y^{\kappa-1}(1-y)^{\tau-1}dy$ the beta function, and $B(t; \kappa, \tau) = \int_0^t y^{\kappa-1}(1-y)^{\tau-1}dy$ the incomplete beta function.

The expectation and variance of a truncated beta r.v., under the reparameterized distribution, are given by

$$E(Y) = \frac{I\left(a; \mu e^{\phi}+1, (1-\mu)\, e^{\phi}\right) - I\left(b; \mu e^{\phi}+1, (1-\mu)\, e^{\phi}\right)}{I\left(a; \mu e^{\phi}, (1-\mu)\, e^{\phi}\right) - I\left(b; \mu e^{\phi}, (1-\mu)\, e^{\phi}\right)}, \tag{7}$$

and

$$Var(Y) = \frac{I\left(a; \mu e^{\phi}+2, (1-\mu)\, e^{\phi}\right) - I\left(b; \mu e^{\phi}+2, (1-\mu)\, e^{\phi}\right)}{I\left(a; \mu e^{\phi}, (1-\mu)\, e^{\phi}\right) - I\left(b; \mu e^{\phi}, (1-\mu)\, e^{\phi}\right)}$$
$$- \left[\frac{I\left(a; \mu e^{\phi}+1, (1-\mu)\, e^{\phi}\right) - I\left(b; \mu e^{\phi}+1, (1-\mu)\, e^{\phi}\right)}{I\left(a; \mu e^{\phi}, (1-\mu)\, e^{\phi}\right) - I\left(b; \mu e^{\phi}, (1-\mu)\, e^{\phi}\right)}\right]^2 \tag{8}$$

respectively.

We assume that $\mu = \eta(\boldsymbol{x}, \boldsymbol{\beta})$, where $\boldsymbol{x} = (x_1, \ldots, x_p)'$ is a vector of p covariates, $\boldsymbol{x}_q$ is a subset of x, $\eta(\cdot)$ is a continuous and twice differentiable function with respect to $\boldsymbol{\beta}$. On the other hand, the parameter related to the dispersion of the response variable, ϕ, is left unmodeled. Therefore, we denote the vector of indexing parameters by $\boldsymbol{\theta} = (\boldsymbol{\beta}, \phi)$.

It is worthwhile to mention that in the truncated beta nonlinear regression model the response variable is asymmetric and heteroscedastic, whereas in the truncated normal nonlinear regression model the response variable is symmetric and it allows for both a homoscedastic or a heteroscedastic structure for the variance of the response variable.

Suppose $Y_1, \ldots, Y_n$ are independent random variables such that each Y_i, $i = 1, \ldots, n$, follows a $TB\left(\eta(\boldsymbol{x}_i, \boldsymbol{\beta})\, e^{\phi}, (1-\eta(\boldsymbol{x}_i, \boldsymbol{\beta}))\, e^{\phi}, a, b\right)$ distribution. Let $\boldsymbol{y} = (y_1, \ldots, y_n)'$ be a vector of observed values of $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$. Then, given the data set $D = (n, \boldsymbol{y}, \boldsymbol{x})$, the log-likelihood function for $\boldsymbol{\theta} = (\boldsymbol{\beta}, \phi)$, the vector of unknown parameters to be estimated, is written as

$$
\begin{aligned}
l(\boldsymbol{\theta}) =& \sum_{i=1}^{n} \log\Gamma\left(e^{\phi}\right) - \sum_{i=1}^{n} \log\Gamma\left(\eta\left(\boldsymbol{x}_i, \boldsymbol{\beta}\right) e^{\phi}\right) - \sum_{i=1}^{n} \log\Gamma\left(\left(1-\eta\left(\boldsymbol{x}_i, \boldsymbol{\beta}\right)\right) e^{\phi}\right) \\
&+ \sum_{i=1}^{n} \left[\eta\left(\boldsymbol{x}_i, \boldsymbol{\beta}\right) e^{\phi} - 1\right] \log\left(y_i\right) + \sum_{i=1}^{n} \left[\left(\eta\left(\boldsymbol{x}_i, \boldsymbol{\beta}\right) - 1\right) e^{\phi} - 1\right] \log\left(1-y_i\right) \\
&- \sum_{i=1}^{n} \log \left\{ I\left(b; \eta\left(\mathbf{x}_i, \beta\right) e^{\phi}, \left(1-\eta\left(\mathbf{x}_i, \beta\right)\right) e^{\phi}\right) \right. \\
&\left. - I\left(a; \eta\left(\mathbf{x}_i, \beta\right) e^{\phi}, \left(1-\eta\left(\mathbf{x}_i, \beta\right)\right) e^{\phi}\right) \right\} \qquad (9)
\end{aligned}
$$

The MLEs of $\boldsymbol{\theta}$ can be obtained by direct nonlinear optimization of (9).

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. Therefore, model diagnostic analysis procedures can provide a guidance whether the regression model being fitted is plausible and whether the conclusions based on it are correct. The regression model constructed in Section 3 was based in the following underlying assumptions: (i) the response variable, $\boldsymbol{y}$ follows a $TB\left(\eta\left(\boldsymbol{x}_i, \boldsymbol{\beta}\right) e^{\phi}, \left(1-\eta\left(\boldsymbol{x}_i, \boldsymbol{\beta}\right)\right) e^{\phi}, a, b\right)$ distribution, $i = 1, \ldots, n$; (ii) the observations are mutually independent. These assumptions may be checked by visual inspections of several residual plots, 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.

4.1 Residual analysis

Let $\hat{y}_i$ be the predicted value of the i^{th} observation defined as

$$
\hat{y}_i = E\left(Y_i \middle| a < Y_i < b, \boldsymbol{x}_i, \hat{\boldsymbol{\theta}}\right), \qquad (10)
$$

where the expectation corresponds to (7) and $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$ obtained by maximizing (9).

We consider the standardized residuals and the standardized Pearson residuals to check for model adequacy, outliers and influent observations. These residuals are given by

$$r_i^s = \frac{y_i - \hat{y}_i}{\sqrt{Var\left(Y_i \middle| a < Y_i < b, \boldsymbol{x}_i, \hat{\boldsymbol{\theta}}\right)}}, \tag{11}$$

and

$$r_i^P = \frac{y_i - \hat{y}_i}{\sqrt{Var\left(Y_i \middle| a < Y_i < b, \boldsymbol{x}_i, \hat{\boldsymbol{\theta}}\right)(1 - h_{ii})}}, \tag{12}$$

respectively.

In both (11) and (12), y_i is the observed value of the i^{th} case and $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$. The variance, $Var\left(Y_i \middle| a < Y_i < b, \boldsymbol{x}_i, \hat{\boldsymbol{\theta}}\right)$ is as given in (8). For the Pearson residuals in (12), h_{ii} is the i^{th} diagonal element of the Hat matrix defined as $H = X(X'X)^{-1}X'$.

Standardized residuals can be interpreted as how much predicted values deviate from real values. Thus, we can set limits for the standardized residuals based on the amount of deviation that we are willing to tolerate between the real value and the predicted value. In this paper, we shall consider an observation as an outlier if its standardized residual is larger than 3 or smaller than -3.

4.2 Influence measures

Let $\hat{\boldsymbol{\theta}}_{(-i)}$ be the MLE of $\boldsymbol{\theta}$ with the i^{th} case deleted. Case-deletion diagnostic metrics were first proposed by Cook (1977) and they rely on the principle that the influence of a given observation can be assessed by comparing the difference between parameters estimates obtained fitting the considered model to the complete data, D, and parameters estimates obtained for the model fitted to the data with the i^{th} observation deleted, $D_{(-i)}$. If the deletion of the i^{th} observation influences the estimates, then $\boldsymbol{\theta}_{(-i)}$ is far from $\boldsymbol{\theta}$ and the i^{th} case is considered influent.

We shall consider Cook's generalized distance and the likelihood distance for influent observations detections. These two metrics are given by

$$GC_i = \left(\hat{\boldsymbol{\theta}}_{(-i)} - \hat{\boldsymbol{\theta}}\right)' I\left(\hat{\boldsymbol{\theta}}\right)\left(\hat{\boldsymbol{\theta}}_{(-i)} - \hat{\boldsymbol{\theta}}\right), \tag{13}$$

and

$$LD_i = 2\left\{l\left(\hat{\boldsymbol{\theta}}\,|D\right) - l\left(\hat{\boldsymbol{\theta}}_{(-i)}\,|D\right)\right\}, \tag{14}$$

respectively.

In (13), $I\left(\hat{\boldsymbol{\theta}}\right)$ is the observed Fisher information matrix defined as

$$I\left(\hat{\boldsymbol{\theta}}\right) = -\left.\frac{\partial^2 l\left(\boldsymbol{\theta}\,|D\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}.$$

Since both (13) and (14) require $\boldsymbol{\theta}_{(-i)}$ for each $i = 1, \ldots, n$, which can be computationally demanding if n is large, Cook & Weisberg (1982) provided the approximation

$$\hat{\boldsymbol{\theta}}_{(-i)} = \hat{\boldsymbol{\theta}} + \left[I\left(\hat{\boldsymbol{\theta}}\right)\right]^{-1} U_{(-i)}\left(\hat{\boldsymbol{\theta}}\right),$$

where $U_{(-i)}\left(\hat{\boldsymbol{\theta}}\right)$ is the score vector defined as

$$U_{(-i)}\left(\hat{\boldsymbol{\theta}}\right) = -\left.\frac{\partial l\left(\boldsymbol{\theta}\,|D_{(-i)}\right)}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}.$$

In practice, a case is considered influent if its GC_i or LD_i value is large. In Cook & Weisberg (1982), the authors suggest that GC_i and LD_i can be compared to critical values of a Chi-squared distribution with the degrees of freedom equal to the number of parameters, p. However, this method may fail to detect influent cases for moderate and large values of p since the critical value of the χ^2_p may be large enough to prevent the detection of cases that are indeed influent.

5 Real data set analysis

In this section we analyze a 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 refer the reader to Rodrigues & Maia (2011) for a more detailed description of the experimental procedures applied to collect the soil samples and the laboratory procedures used to measure soil-water content.

We notice that in common soil-water data analysis, for each soil profile and for each depth, a different SWCC is fitted. In this paper, we choose to fitted a single SWCC to each soil profile with all depths. Thus, the data set features a total of 81 observations. Soil water content at saturation, θ_s, were calculated by weighing the soil profile samples directly. The residual soil water content for each soil profile sample were calculated by submitting the soil samples to a tension of 1500 kPa.

We shall consider the location parameters $\eta\left(x,\boldsymbol{\beta}\right)$ as the Van Genuchten-Mualem model given in (1) and (2), and the truncation limits are $a=\theta_r$ and $b=\theta_s$. Therefore, we shall fit the truncated beta nonlinear regression model where $Y_i \sim TB\left(\eta\left(\boldsymbol{x}_i,\boldsymbol{\beta}\right)e^{\phi},\left(1-\eta\left(\boldsymbol{x}_i,\boldsymbol{\beta}\right)\right)e^{\phi},\theta_r,\theta_s\right)$, $i=1,\ldots,n$.

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. We notice that the estimated standard deviation of β_1 is quite larger than would be prefered.

From the estimated SWCC presented in Figure 1a, 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 1b, 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 randomly distributed around zero with no outlier observations. We also note that no influent observation was depicted by Cook's generalized distance in Figure 2c and by the likelihood distance in Figure 2d.

Table 1: Model fit summary for the truncated beta van Genuchten-Mualem regression model adjusted to soil profile 214 data.

Parameter	Estimate	St. Dev.	95% C.I.	
β_1	61,8415	5,4828	51,0951	72,5878
β_2	1,4324	0,0191	1,3950	1,4697
ϕ	6,8356	0,1714	6,4996	7,1716

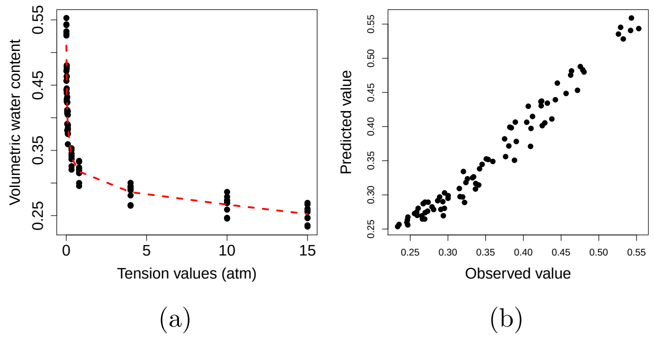

Figure 1: Profile 204 data: (a) estimated SWCC; (b) Observed y against predicted values of y.

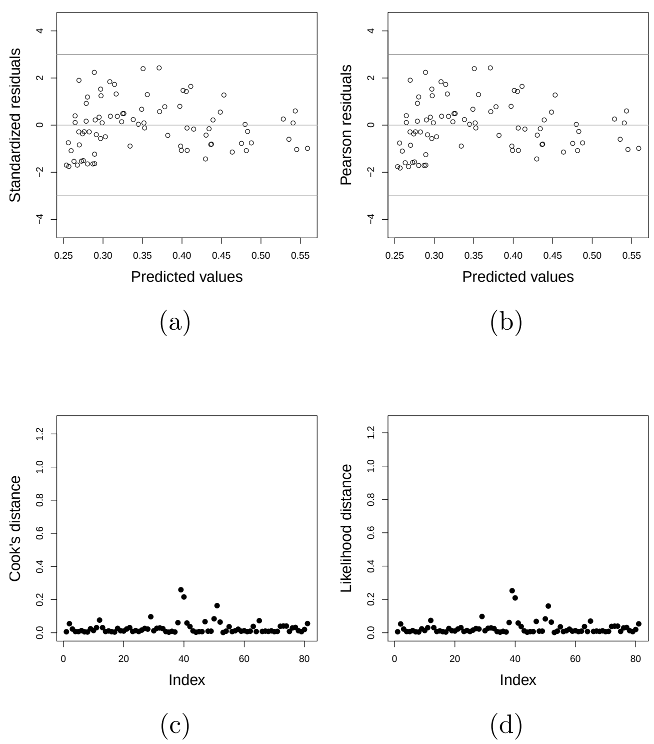

Figure 2: Profile 204 data: (a) standardized residuals; (b) standardized Pearson residuals; (c) approximated Cook's generalized distance; (d) approximated likelihood distance.

6 Conclusions

In this paper, we have proposed and illustrated an alternative approach to model SWCCs based on the truncated beta nonlinear regression model, which take truncation into account, an important feature of the data. Moreover, diagnostic analysis tools were used to check the model assumptions and for outlier and influent observations detection.

We acknowledge that the truncated beta is only one of many truncated distributions that can be considered to model soil-water retention data. Also, we could consider the truncated version of some recently propose skewed distributions. Moreover, it is important to take into account the sample depth information an extend the proposed model under the generalized nonlinear mixed models framework. Another key point that could be improved is the estimation procedure which could be performed under the Bayesian perspective, thus taking into account prior information about the model parameters. Those are issues to be considered in a future work.

In summary, we present a novel methodology to model and study the SWCC which are important to study the relationship between soil and water, a physical phenomenon that affects soil use in many different purposes. The proposed truncated normal and truncated beta nonlinear regression models are simple models, that can be estimated with known implemented procedures. Yet, the model improves the quality of parameters estimates as it encompass important features of the data.

References

A'Hearn, A. (2004). A restricted maximum likelihood estimator for truncated height samples. *Economics and Human Biology*, **2**, 5–19.

Amemiya, T. (1985). *Advanced econometrics*. Harvard University Press, Cambridge, MA.

Brooks, R. H. & Corey, A. T. (1964). Hydraulic properties of porous media. *Hydrology Paper*, **3**(3).

Chao, K., Nelson, J. D., Overton, D. D. & Cumbers, J. M. (2008). Soil water retention curves for remolded expansive soils. In *Unsaturated Soils: Advances in Geo-Engineering, Proceedings of the 1st European Conference, E-UNSAT*.

Chen, S. & Zhou, X. (2012). Semiparametric estimation of a truncated regression model. *Journal of Econometrics*, **167**, 297–304.

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.

Cornelis, W. M., Khlosi, M., Hartmann, R., van Meirvenne, M. & de Vos, B. (2005). Comparison of unimodal analytical expressions for the soil-water retention curve. *Soil Science Society of America Journal*, **69**, 1902–1911.

Cosslett, S. (2004). Efficient semiparametric estimation of censored and truncated regressions via smoothed self-consistency equation. *Econometrica*, **72**, 1277–1293.

Dourado-Neto, D., Nielsen, D. R., Hopmans, J. W., Reichardt, K. & Bacchi, O. O. S. (2000). Software to model soil water retention curves (swrc, version 2.00). *Scientia Agricola*, **57**(1), 347–354.

Ferrari, S. L. P. & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, **31**(7), 799–815.

Fredlund, D. G. & Xing, A. (1994). Equations for the soil-water characteristic curve. *Canadian Geotechnical Journal*, **31**(3), 521–532.

Gardner, W. R. (1958). Some steady state solutions of unsaturated moisture flow equations with application to evaporation from water table. *Soil Science*, **85**, 228–232.

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

Hausman, J. A. & Wise, D. A. (1977). Social experimentation, truncated distributions, and efficient estimation. *Econometrica*, **45**(4), 919–938.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, **5**(4), 475–492.

Lee, L. F. (1992). Semiparametric nonlinear least-square estimation of truncated regression models. *Econometric Theory*, **8**, 52–94.

Lee, M. (1993). Quadratic mode regression. *Journal of Econometrics*, **57**, 1–19.

Leong, E. C. & Rahardjo, H. (1997). Review of soil water characteristic curve equations. *Journal of Geotechnical and Geoenvironmental Engineering*, **123**(12), 1106–1117.

Maddala, G. S. (1983). *Limited dependent and qualitative variables in econometrics*. Cambridge, New York.

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

Newey, W. (2004). Efficient semiparametric estimation via moment restrictions. *Econometrica*, **72**, 1877–1897.

Powell, J. (1986). Symmetrically trimmed least squares estimation for tobit models. *Econometrica*, **54**, 1435–1460.

Ritz, C. & Streibig, J. C. (2008). *Nonlinear regression with R*. Use R! Springer, New York.

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*.

Sillers, W. S., Fredlund, D. G. & Zakerzadeh, N. (2001). Mathematical attributes of soil-water characteristic curves models. *Geotechnical and Geological Engineering*, **19**, 243–283.

Silva, E. M., Lima, E. F. W., Azevedo, J. A. & Rodrigues, L. N. (2006). Valores de tensão na determinação da curva de retenção de água de solos do cerrado. *Pesquisa Agropecuária Brasileira*, **41**(2), 323–330.

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.

van Genuchten, M. T. & Nielsen, D. R. (1985). On describing and predicting the hydraulic properties of unsaturated soils. *Annales Geophysicae*, **3**(5), 615–628.

Yates, S. R., van Genuchten, M. T. & Warrick, A. W.; Leij, F. J. (1992). Analysis of measured, predicted, and estimated hydraulic conductivity using the retc computer program. *Soil Science Society of America Journal*, **56**, 347–354.