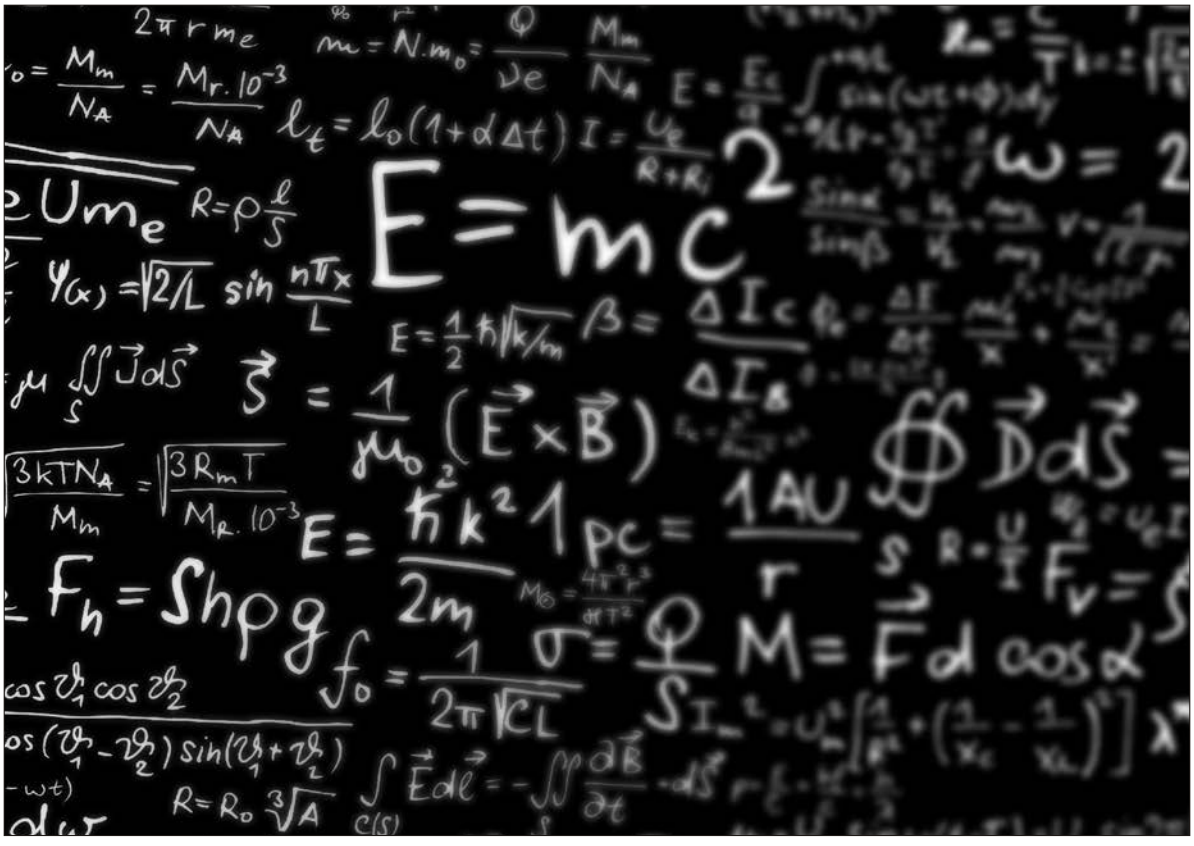

Asian Journal of Mathematics & Statistics

ISSN 1994-5418

science alert

ANSInet
an open access publisher
http://ansinet.com

Asian Journal of Mathematics and Statistics 1 (1): 1-13, 2008
ISSN 1994-5418

On the Selection of Models in Nonlinear Regression

S.A. El-Shehawy
Department of Mathematics, College of Science, Qassim University,
P.O. Box 237, Buriedah 81999, Kingdom of Saudi Arabia

Abstract: This study discusses the selection of parametric models of the moisture retention characteristic MRC as nonlinear regression models from a mathematical and statistical viewpoint. Simulation studies and some measures of nonlinearity are given. A comparison is introduced between the used famous models of the moisture retention characteristic (van Genuchten and King) which are presented early and their variants. Following the considered simulation study and the used measures of nonlinearity in nonlinear regression the author find that although a variant of van Genuchten model with four parameters is an optimal model but it has a strong nonlinear parameter. Moreover, a possible appropriate reparametrization with respect to this nonlinear parameter is proposed.

Key words: Nonlinear regression models, measure of nonlinearity, model selection, retention function

INTRODUCTION

The moisture retention characteristic MRC, which is the relation between water content (the volumetric water content Θ) and pressure head (the soil capillary pressure h), can be measured for natural soils. It is referred to as the soil water retention curve. This relationship is primarily based on the soil pore structure and the pore size distribution. The MRC $\Theta(h)$ is typical for a given soil having its particular status of consolidation, geometrical arrangement of particles and aggregation and other chemical and biological feature. The retention curve graphically displays a continuously differentiable (smooth) S-shaped curve between the saturated and residual water contents (Θ_s and Θ_r, respectively). Knowledge of the MRC is indispensable in describing soil water processes (flow). The MRC can be determined either directly in the field or in the laboratory on undisturbed core samples. MRC data from Vereecken *et al*. (1989) are used. These data sampled the soil horizons of forty important Belgian soil series and measured their MRC.

Several of parametric models have been proposed to describe MRC and all these models are nonlinear regression models (NLR-models). Most of the proposed models are curve-fitting equations and they are able to describe the typical S-shaped behavior of MRC and represent NLR-models which can be fitted to give data of MRC. The model-parameters can be estimated by applying algorithms minimizing a least squares (LS) object function (Bates and Watts, 1988).

Mathematical statistics have developed several methods for model selection in nonlinear regression. In this study, the researcher studies problems of model selection for MRC from a statistical and mathematical viewpoint. This study is carried out through the comparison between the used MRC-model proposed by van Genuchten (1980), a MRC model developed by King (1965) and its variants. A main task of the analysis of experimental data is the estimation of the parameters in a NLR-model (Bates and Watts, 1988; Ratkowsky, 1983; Ratkowsky, 1990;

Corresponding Author: Shebin El-Kom, Department of Mathematics, Faculty of Science, Menoufia University, Egypt

Seber and Wild, 2005). Usually it is not known whether a proposed regression model describes the unknown true regression function sufficiently well. But the results of the statistical analysis may heavily depend on the chosen model.

Modelers probably have one of the three purposes in mind when they wish to fit a NLR-model to set of data. First, they are interested in obtaining a good fit to the data. That means, they are primarily interested in the representation of the relationship between the independent and dependent variables by the chosen model. Secondly, they focus on the prediction of values of the dependent (response) variable for given values of the independent (regressor) variable. Moreover in some situations the modelers wish to make inference based upon interpretation of the parameter estimates of the corresponding model. The current study concerns with the second and the third aspects.

A more general NLR-model may be written:

$$y_{ij} = f(x_i, \beta) + \varepsilon_{ij}, \quad i = 1, 2, \ldots, k, \; j = 1, 2, \ldots, n_i, \quad N = \sum_{i=1}^{k} n_i \tag{1}$$

where, y_{ij} is the values of the response variable Y at fixed values x_i of exploratory variables (nonrandom design points). Here, the real valued function $f(.,\beta)$ is expectation function and it is known up to a P-dimensional vector of parameters $\beta = (\beta_1, \beta_2,\ldots,\beta_p) \in R^P$ and this function is twice continuously differentiable in β. Moreover, the random perturbations ε_{ij} (the measurement errors) are uncorrelated random variables with zero mean and unknown variance, σ^2 which do not depend on x_i. Given the date

$$(x_i, y_{ij}), \quad i = 1, 2, \ldots, k, \; j = 1, 2, \ldots, n_i, \quad N = \sum_{i=1}^{k} n_i \tag{2}$$

the estimation of the regression function f reduces to estimate its parameters. The most popular approach to estimate $\beta = (\beta_1, \beta_2,\ldots,\beta_p)$ is to employ the ordinary least squares estimator $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)$, which is a minimizer of the sum of squares

$$S(\beta) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - f(x_i, \beta))^2 \tag{3}$$

The linearity of the regression model (1) may be produced if the expected response $f(x, \beta)$ is a linear function of the parameter vector $\beta = (\beta_1, \beta_2,\ldots,\beta_p)$.

The solution $\hat{\beta}$ of the corresponding minimum problem

$$S(\hat{\beta}) = \min_{\beta} S(\beta) \tag{4}$$

with respect to a linear regression model (LR-model) will be given by an explicit algebraic expression. This solution, which is the least squares estimator $\hat{\beta}$ of β, is a linear function of the data vector of y_{ij}, asymptotically normally distributed, unbiased and it has a minimum variance (Bates and Watts, 1988; Haines *et al*., 2004).

For the NLR-models the situation is different. There is no explicit expression for $\hat{\beta}$. One gets the least squares estimator $\hat{\beta}$ for β only by an iterative procedure starting from some assumed value of

β. In general the least squares estimators of the parameters of a NLR-model are biased, non-normally distributed (skewed) with variance exceeding the minimum possible variances in the corresponding linear model (Ratkowsky, 1983). The extend of bias, non-normality and excess of variances differs widely from model to model. The researcher looks for MRC models which, with respect to their estimation behavior, are closely related to LR-models.

NLR-models differ in their estimation properties from linear regression models. Given the usual assumption of independent and identically distributed measurement errors, the parameter (LS) estimators for linear models are linear, unbiased, at least asymptotically normally distributed, minimum variance estimators. On the other hand, NLR-models tend to do so only when the sample size becomes very large.

In this study, the purpose is the exploration and the comparison of the properties of estimators for some different MRC models in situations having sample sizes typically obtained in practice. The MRC models with close-to-linear estimation behavior are looked. The extend of nonlinear estimation behavior (e.g., bias) depends upon the nonlinearity of a model/data set combination. Therefore, this work can be carried out in the first step by using simulation studies with respect to the considering models and in the second step by computing some measure of nonlinearity.

The aim of the presented study is to select an optimal MRC model, to determine the cause of the nonlinearity in this selected model and finally to diminish (or avoid) this defect through a suggested reparameterization.

THE CLASS OF FUNCTIONS USED FOR MODEL SELECTION

van Genuchten (1980) presented a widely used class of functions for parametrizing measured soil water characteristics

$$\Theta(h) = \Theta_r + \frac{(\Theta_s - \Theta_r)}{(1+(\alpha.h)^n)^m} \tag{5}$$

where α, n and m are positive parameters defining the MRC's shape and this model is denoted by (VG1). This function is a MRC model of five parameters α, n, m beside, Θ_s, Θ_r. Moreover, King (1965) suggested the following model

$$\Theta(h) = \Theta_s \times \left[\frac{\cosh((h/h_0)^b + \varepsilon) - \gamma}{\cosh((h/h_0)^b + \varepsilon) + \gamma}\right] \tag{6}$$

with five real parameters Θ_s, h_0, b, ε and γ where b having negative values and this model denoted by (K1). The Table 1 indicates to special cases (VG2), (VG3), (VG4) and (VG5) of van Genuchten model (VG1) and also the model (K2) as a special case of the model (K1) of King, which are models with fewer parameters:

Each of the models (VG2), (VG3) and (VG4) has four parameters Θ_s, Θ_r, α, n but (VG5) is the model with only three parameters Θ_s, α, n. Also, the model (K2) has four parameters and Θ_s, h_0, b and γ.

Table 1: Some variants of van Genuchten model (VG1) and King model (K1)

Model	(VG2)	(VG3)	(VG4)	(VG5)	(K2)
Case	m = 1-1/n	m = 1-2/n	m = 1	$\Theta_r = 0$, m = 1	ε = 0

USING METHODS TO SELECT A NLR-MODEL FOR MRC

Two methods to select a NLR-model for MRC are used in this study. These methods are described in the following two subsections.

Description of the Simulation Experiments

Simulation studies (experiments) are probably the most direct and best way to enable the modeler to study the sampling properties of the LS estimator. Data are generated using a set of predetermined values of the parameters, allowing only the values of the measurement error to change randomly or pseudo-randomly form set to set. By this means, many sets of simulated data produced and each set provides a LS estimates of the parameters of the model under consideration. These estimated parameters may be examined for their bias, variance and other distributional properties.

Throughout this study, the reported results of simulation are based on 100 sets of simulated data. Recall that the NLR-model is given by (1), where the ε_{ij} are uncorrelated identically distributed random variables. The measurement errors are generated under the considering assumptions (independent identically normally distributed measurement errors with zero mean and constant variance σ^2). If one assumes such a type of distribution of the measurement error, then there is a possibility to choose an appropriate value of the variance. The considering selection of σ is based on MRC data from Vereecken *et al*. (1989) (Table 2). The variance is model independent estimated using the repeated measurements (pure error, intra sample estimates). The calculated value $\hat{\sigma} = 0.015905$ is a typical mean value with respect to the other data sets of Vereecken with repeated measurements. Now, a simulation study for two different soils (sand and loam) is illustrated using this estimated value $\hat{\sigma}$ and true parameters for the model (VG2).

Table 3 contains on the true parameters of the model (VG2) for two different soils sand and loam (Carsel and Parrish, 1988).

Every simulated data set has the structure of Vereecken MRC data. A simulated value is the sum of the expected value corresponding to the value of pressure and a random error, which is normally distributed with zero mean and variance $\hat{\sigma}^2$. Table 2 shows the measured MRC data of a Belgian soil (Bates and Watts, 1988), the theoretically expected values of sand and the first simulated data set.

Each set of simulated data is then fitted by least squares. That means, the vector of parameters β of the considering models (VG1), (VG2), (VG3), (VG4), (VG5), (K1) and (K2) is estimated.

Table 2: Measured MRC-data, theoretically expected (MODEL) of sand and the first simulated MRC-data for sand (SIMUL 1)

h	Θ (h)	MODEL	SIMUL 1
1.00	0.512	0.429	0.431
1.00	0.545	0.429	0.376
3.16	0.510	0.403	0.391
3.16	0.544	0.403	0.382
10.00	0.504	0.216	0.232
10.00	0.537	0.216	0.193
31.62	0.481	0.075	0.071
31.62	0.506	0.075	0.068
100.00	0.411	0.050	0.046
100.00	0.422	0.050	0.040
199.53	0.329	0.046	0.029
199.53	0.348	0.046	0.072
630.96	0.293	0.045	0.053
630.96	0.287	0.045	0.040
2512.59	0.141	0.045	0.039
2513.30	0.139	0.045	0.034
15848.92	0.149	0.045	0.071
15848.92	0.158	0.045	0.036

Table 3: True parameters of van Genuchten model (VG2)

Texture class	Θ_r	Θ_s	α	n	m = 1-1/n
Sand	0.045	0.43	0.145	2.68	0.620
Loam	0.078	0.43	0.036	1.56	0.359

Table 4: Estimated vector of parameters of the model (VG1) corresponding to the first simulated data set for sand

Parameter	Θ_r	Θ_s	α	n	m
Estimated values	0.045431	0.404116	0.129247	2.869584	0.678016

Different algorithms are used, for instance the Levenberg/Marquardt method (Bates and Watts, 1988). In the case of model (VG1) and (VG2), the true (simulated) vector of parameters is considered as an initial estimate. Table 4 contains the estimated vector of parameters $\hat{\beta}$ of the model (VG1) corresponding to the first simulated data set for sand.

For every model under consideration, 100 estimates are produced. Taking each parameter separately, univariate statistics are calculated and the corresponding distribution is graphed. In addition, it is possible to examine the multivariate behavior of the LS estimator by calculating bivariate statistics and using two or three dimensional plots. This means that the researcher can see how close to linear in its behavior the LS estimator is. An acceptable model leads to LS estimations for its parameters with small biases, with distributions close to normal distribution and variances close to the minimal possible variances.

Description of the Measures of Nonlinearity

Another possible way to analyze the nonlinear behavior of a model data set combination is the calculation of so called measures of nonlinearity (e.g., curvature, bias and skewness). Using differential geometry-concepts, the measures of nonlinearity based on the notion of curvature were developed (Bates and Watts, 1988; El-Shehawy, 2001; El-Shehawy and Karawia, 2006). These measures are independent of scale changes in both the data and parameters. They can be used to compare different models with different parameterizations combined with different data set.

The N-vector $\eta(\beta)$ with the components

$$\eta_i(\beta) = f(x_i,\beta), \quad i = 1,\ldots,N, \quad \beta \in R^P \tag{7}$$

defines a P-dimensional surface, the so called expectation surface or solution locus in the n-dimensional response space. The LS estimate $\hat{\beta}$ corresponds to that point $\eta(\hat{\beta})$ on this surface which is closest to the measured N-dimensional vector of response y. The parameter vector β is assumed in the neighborhood of its LS estimate $\hat{\beta}$.

If the quadratic term in the second order Taylor approximation of $\eta(\beta)$ is neglected, one will have for β in the vicinity of $\hat{\beta}$ the linear approximation

$$\eta(\beta)-\eta(\hat{\beta}) \approx \dot{V}\,(\beta-\hat{\beta}), \tag{8}$$

where $\dot{V}$ is the (N, P) Jacobi-matrix $\frac{d\eta(\beta)}{d\beta}$ at $\beta = \hat{\beta}$ (Ratkowsky, 1983). The range of the matrix $\dot{V}$ is the tangent plane to the expectation surface at the point $\hat{\beta}$ and the linear approximation (8) amounts to approximating the expectation surface in a neighborhood of $\hat{\beta}$ by this plane. Using (8) the following form is got

$$\left\|\eta(\beta)-\eta(\hat{\beta})\right\|^2 \approx (\beta-\hat{\beta})^T\dot{V}^T\dot{V}\,(\beta-\hat{\beta}) \tag{9}$$

From some statistical properties of the linear models (Bates and Watts, 1988; Haines *et al*., 2004), the following ellipsoid with center $\hat{\beta}$ is considered as an approximation of an 100 (1-a)% confidence region for $\beta = (\beta_1, \beta_2, ..., \beta_p) \in R^P$:

$$(\beta - \hat{\beta})^T \dot{V}^T \dot{V} (\beta - \hat{\beta}) \leq P s^2 F(P, N - P; a) \tag{10}$$

where $s^2 = \frac{S(\hat{\beta})}{N - P}$ is the residual mean square, is the upper a quantile for the F-distribution with P, N-P degrees of freedom.

From (7), (8) and (10), the expectation surface $\eta(\beta)$ lies approximately within the intersection of tangent plane and a sphere with center $\eta(\hat{\beta})$ and radius $\sqrt{P s^2 F(P, N - P; a)}$.

A LR-model has a linear solution locus, which means a hyper-plane for $p \geq 3$ and (9) holds exactly. In addition, lines (parameter lines) on the solution locus representing constant values of β_r, r = 1, 2, ..., P, are straight, parallel and equally spaced for equal increments of β_r.

For a nonlinear solution locus the situation is different. The solution locus is a curved surface and the parameter lines on this surface (or the projections of these lines onto the tangent plane) are, in general, neither straight, parallel nor equi-spaced.

The extent of the curvature of the solution locus has been called intrinsic nonlinearity, since this nonlinearity cannot be altered by reparametrization. It is an inner geometrical property of the surface. The extent of the curvature of the parameter lines, their lack of parallelism and equi-spacedness has been called parameter-effects nonlinearity, since it is determined by the way in which the parameters appear in the model, that means, it depends on the parameterizations (Bates and Watts, 1988).

The validity of the tangent plane approximation (8) will depend on the magnitude of the quadratic term in the Taylor expansion of $\eta(\beta)$ relative to the linear term. In making this comparison it is helpful to split the quadratic term into two orthogonal components, the respective projections onto the tangent plane and the component normal to the tangent plane. These components were compared with the linear term and got two measures of nonlinearity, the parameter-effects curvature and the intrinsic curvature. Standardizing the model and the data leads to scale independent quantities. Both measures depend on the direction $(\beta - \hat{\beta})$. Root Mean Square (RMS) curvatures are used, which are the square root of the average over all directions of squared curvatures. These measures are denoted by c^{IN} (intrinsic curvature) and c^{PE} (parameter-effects curvature). The symbol $1/\sqrt{F}$ refers to the inverse of the radius of the (standardized) sphere (10).

A convenient scale of reference can be established by comparing the RMS curvature with that of the (scaled) confidence disk (10) at a specified level (1-a), a>0. That means, we compare the radius of curvature 1/c ($c = c^{IN}$ or $c = c^{PE}$) with the radius of the confidence disk $\sqrt{F}$. An RMS curvature will be considered as small if it is much less than the curvature of the (1-a) confidence disc, that is if $c << 1/\sqrt{F}$, where F = F (P, N-P; a).

Following some earlier several research study s, an expectation surface with radius of curvature 1/c is considered and the deviation of the tangent plane from the surface or the deviation of the parameter line from the straight line at a distance $\sqrt{F}$ is determined. This deviation, expressed as a percentage of the radius of the confidence disk, is $100 (1 - \sqrt{1 - c^2 F}) / (c\sqrt{F})$ %, so that:

- A value of $c^{IN}\sqrt{F} = 0.2$ causes the surface to deviate by 10% of the radius of the confidence at the edge of the confidence disk;
- A value of $c^{IN}\sqrt{F} = 0.2$ causes the surface to deviate by 27% of the radius of the confidence at the edge of the confidence disk;
- A value of $c^{IN}\sqrt{F} = 1$ causes the surface to deviate by 100% of the radius of the confidence at the edge of the confidence disk and so on.

If c^{IN} is replaced by c^{PE} then a corresponding rule about the deviation of a parameter line from a straight line will be get.

In almost all cases known from several works the intrinsic curvature is very much smaller than the parameter-effects nonlinearity. Since c^{PE} depends on the parameterization it is possible to reduce the parameter-effects nonlinearity using an appropriate reparametrization for the model under consideration.

Another practicable way to study the nonlinear behavior of a model data set combination is the calculation of estimation for the bias in the LS estimates. A corresponding formula for bias was presented by Box (1971) and El-Shehawy (2001). The approximate bias for each component of the estimate $\hat{\beta}$ is calculated separately.

Although the bias can be used as a measure of the extent to which parameter estimates may exceed or fall short of the true parameter values yet it cannot be used to compare parameters in two different parameterizations (Ratkowsky, 1990). This comparison is possible with another measure of nonlinearity, the measure of skewness (of the distribution of the estimated parameter) due to Hougaard (1985) and El-Shehawy (2001). Following Ratkowsky (1990), it is possible to use a rule-of-thumb for asserting whether the estimator $\hat{\beta}_r$ of the rth component β_r of the parameter vector β, as assessed by the measure of skewness Sk_r, is close-to-linear (nearly symmetrically distributed) or contains considerable nonlinearity:

- If $Sk_r < 0.1$, the estimator $\hat{\beta}_r$ of β_r will be very close-to-linear in behavior;
- If $0.1 \leq Sk_r < 0.25$, the estimator will be reasonably close-to-linear in behavior;
- If $Sk_r \geq 0.25$, the skewness is very apparent and
- If $Sk_r > 1$, indicates considerable nonlinear behavior.

RESULTS

Simulation Experiments

Table 5 shows the mean value of the residual sum of squares $S(\hat{\beta})$ for all models under consideration with respected to the simulation of sand.

If the retention models with respect to the ability to fit the simulated data are compared, the models with five parameters (VG1) and (K1) will be most flexible. The difference between these best fitting models and the models with four parameters is marginal, only the model (VG5) with three parameters is not flexible enough. These properties of the retention models neither depend on the type of the simulated soil nor on the used optimization algorithm.

Consider the distribution of the estimated parameters of the models (VG1) and (VG4). Some descriptive statistics of the 100 estimated parameters of the model (VG1) for the simulated data sets of sand are given in Table 6.

Table 5: Residual sum of squares for the simulated data sets of sand

Model	(VG1)	(VG2)	(VG3)	(VG4)	(VG5)	(K1)	(K2)
$S(\hat{\beta})$	0.0035	0.0038	0.0039	0.0040	0.0223	0.0035	0.0038

Table 6: Descriptive statistics for the distribution of the estimated parameters for the model (VG1) and the simulated data sets of sand

Parameters	True value	Mean	Min.	Max.	St. dev.	Skewness	Sign.
Θ_r	0.045	0.045	0.024	0.065	0.007	-0.104	0.196
Θ_s	0.430	0.431	0.391	0.485	0.015	0.431	0.001
α	0.145	0.143	0.040	0.322	0.047	0.440	0.000
n	2.680	7.187	1.409	296.331	29.697	9.523	0.000
m	0.620	0.787	0.003	4.206	0.641	2.179	0.000

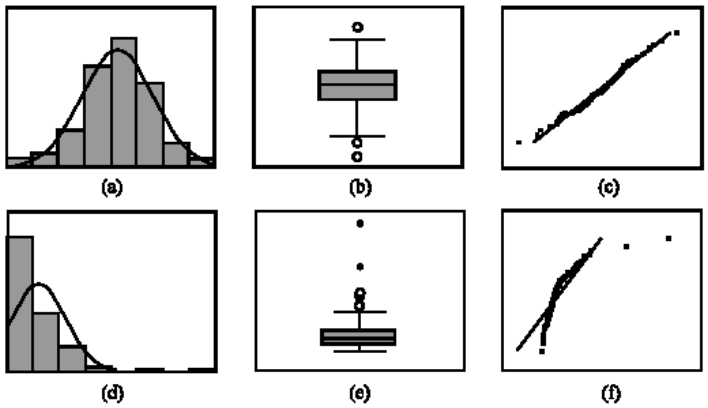

Fig. 1: The shape of the distributions of the estimated parameters Θ_r ((a), (b), (c) figures) and m ((d), (e), (f) figures) of the model (VG1) and the simulated data sets of sand (Histogram with density, Pox-Plot and Q-Q Plot)

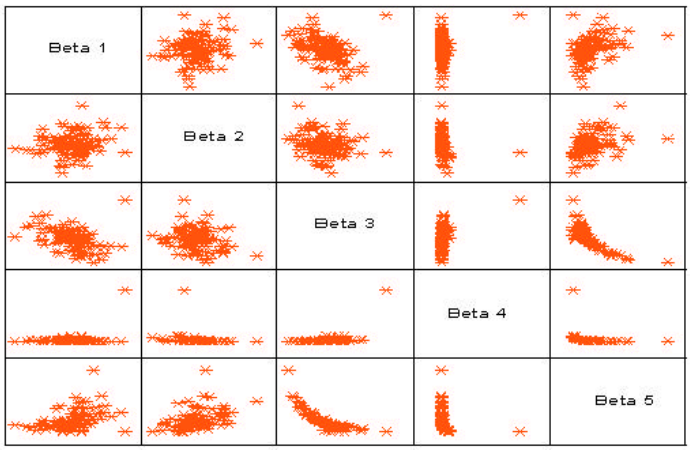

Fig. 2: The scatter-plot matrix for the estimated parameters of the model (VG1) and the simulated data sets of sand

The symbol Sign. refer to the p-values for the goodness of fit test (Kolmogorov-Smirnov test) for the fit of a normal distribution. Only the distribution of the linear parameter Θ_r is similar to a normal distribution (p-value>0.05). In case of all other parameters the hypotheses of normal distribution has to be rejected (since p-value<0.05). These parameters are nonlinear parameters. The estimation of these parameters is biased and the distribution of the estimations is non-symmetric with a great variability. Especially, the parameters n and m are skewed (right tail) with some very extreme values. Figure 1 describes the shape of the distributions and the similarity to the normal distribution of the linear parameter Θ_r and of the nonlinear parameter m.

Consider dependencies between the estimated parameters of the model with five parameters (VG1). The scatter-plot matrix in Fig. 2 shows strong relationships especially between the nonlinear parameters a, n, m. In the Fig. 2, Beta 1, Beta 2, Beta 3, Beta 4 and Beta 5 denote to Θ_r, Θ_s, α, n and m, respectively.

Table 7: Descriptive statistics for the distribution of the estimated parameters for the model (VG4) and the simulated data sets of sand

Parameters	Mean	Min.	Max.	St. dev.	Skewness	Sign.
Θ_r	0.047	0.026	0.064	0.006	-0.205	0.200
Θ_s	0.435	0.401	0.747	0.014	0.182	0.067
α	0.115	0.088	0.134	0.008	0.485	0.003
n	2.283	1.529	4.589	0.469	1.506	0.015

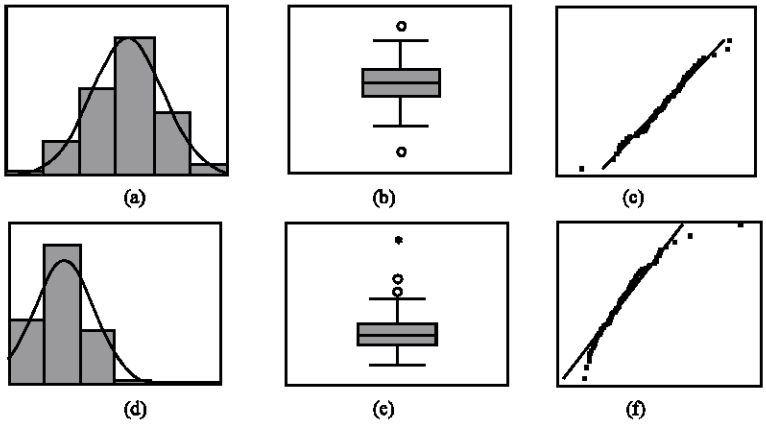

Fig. 3: The shape of the distributions of the estimated parameters Θ_r ((a), (b), (c) figures) and n ((d), (e), (f) figures) of the model (VG4) and the simulated data sets of sand (Histogram with density, Pox-Plot and Q-Q Plot)

Spearman's rank correlation coefficient (a measure on monotone dependence (Johnson and Bhattacharyya, 1992) for a pair α, m is equal to -0.917. Within the three dimensional space the estimated values α, n, m describe a nonrandom point could. The points are concentrated along space curves. The corresponding constellations of parameters allow an equal fit of the same data. That means, with the model (VG1) over fitting is possible, estimated parameters may depend on each other and therefore it is impossible to identify a (simulated or real) soil using the estimated vector of parameters.

If the model (VG4) with four parameters and m = 1 is considered, the results for the simulated data sets of sand are given in Table 7.

It is obvious, that the estimation properties of the parameters of the model (VG4) are much more better than the corresponding properties of the model (VG1). The strong nonlinear parameter m of the model (VG1) is fixed. Although the model (VG4) is flexible yet it is also stiff enough (Fig. 3).

The scatter-plot matrix in Fig. 4 shows strong relationships especially between the nonlinear parameters α and n.

The parameter n of the model (VG4) has the worst estimation properties in comparison with the other parameters Θ^r, Θ^s and α. The distribution of the estimated values of this parameter has a great variability and is skewed (Fig. 3).

If the interest is considered in the model (VG4), an attempt to improve these problematic qualities will be occurred by a reparametrization. As an example, consider the replacement of the parameter n by $1/\tilde{n}$. This reparametrization of the model (VG4) is denoted by the model (RVG4). The distributional properties of this new parameter $\tilde{n}$ are studied. Table 8 contains descriptive statistics for the parameter $\tilde{n}$ for the model (RVG 4) and the simulated data sets of sand.

Table 8 and Fig. 5 show these the distributional properties in comparison with the corresponding properties of the model (VG4). The results with respect to the other parameters do not change.

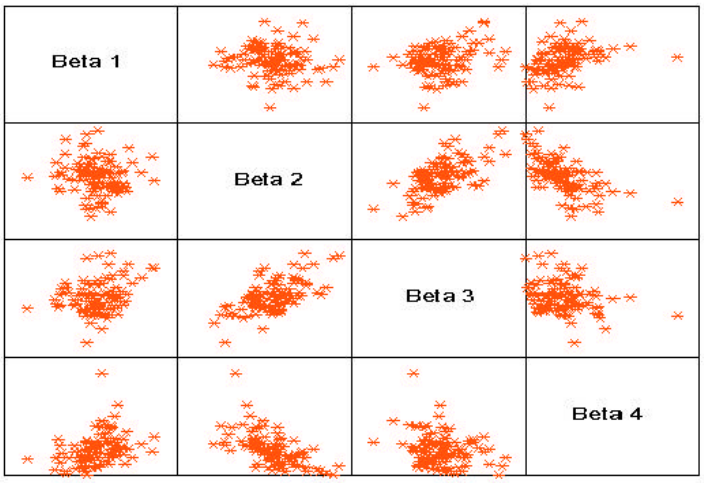

Fig. 4: The scatter-plot matrix for the estimated parameters of the model (VG4) and the simulated data sets of sand

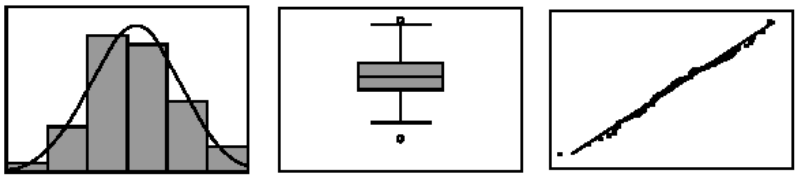

Fig. 5: The shape of the distributions of the estimated parameter ñ of the model (RVG4) and the simulated data sets of sand (Histogram with density, Pox-Plot and Q-Q Plot)

The statistical properties of the model (VG4) are better than the estimation properties of the other models of the van Genuchten type with four parameter (VG2) and (VG3). These observed properties do not depend on the simulated soil.

Consider the model (K1) of King with five parameters and its variant (K2), a similar situation will be obtained. The model (K1) is very flexible but not stiff enough. Fix the strong nonlinear parameter ε, the model (K2) with four parameters will be obtained. The distributional properties of the estimated parameters of this model are much better than the properties of the corresponding model (K1).

Some Calculations for Measures of Nonlinearity

Throughout this subsection some results of nonlinearity measures are discussed. These results relate to the current and previously studied models in combination with some simulated data sets of sand and loam.

Firstly, consider the calculated RMS curvature measures for the models (VG1), (VG2), (VG3), (VG4), (VG5), (K1) and (K2). Table 9 indicates to the scaled RMS curvatures (the intrinsic curvature and the parameter-effects curvature) relative to a 95% confidence disk radius (i.e., a = 0.05) for the studies models under consideration and the first simulated data set SIMUL_1 of sand.

Table 8: Descriptive statistics for the distribution of the estimated parameters for the model (RVG4) and the simulated data sets of sand

Parameters	Mean	Min.	Max.	St. dev.	Skewness	Sign.
$\tilde{n}$	0.454	0.218	0.654	0.084	0.143	0.200

Table 9: Scaled RMS curvatures relative to 95% confidence disk radius for the considered models and the first simulated data set of sand

Model	(VG1)	(VG2)	(VG3)	(VG4)	(VG5)	(K1)	(K2)
$c^{IN}\sqrt{F}$	0.2462	0.2582	0.2920	0.2167	0.3671	0.4730	0.4280
$c^{PE}\sqrt{F}$	10.0923	1.9284	2.7480	1.0743	1.6242	3.2690	1.2680

Table 10: Calculated parameter-effects curvature in the direction corresponding to the parameters of the van Genuchten models in combination with the first simulated data set of sand

Model	(VG1)	(VG2)	(VG3)	(VG4)	(VG5)
c_{α}^{PE}	0.21	0.19	0.21	0.16	0.41
c_{n}^{PE}	2.30	3.00	4.30	1.60	1.70
c_{m}^{PE}	18.00	----	----	----	----

The intrinsic curvature only depends on the geometry (the shape) of the model data set combination. This measure is independent of the chosen parameterization of the model.

Referring to Table 9, it can be seen that the model (VG4) with four parameters possess the best value with respect to the parameter-effects curvature in comparison with the other models. This table shows also the extreme value of RMS parameter-effects curvature with respect to the model with five parameters (VG1).

Consider different data sets, different values of the curvature are obtained and the order of models is in general stable. If the number of repeated measurements increases, the curvature will be reduced. A value of $c^{IN}\sqrt{F}$ causes a moderately deviation of 10% between the expectation surface and the approximating tangent plane at the radius of confidence disk. The values of the parameter-effects curvature are much more greater. This means that there are a great difference between the parameter lines on the expectation surface and the parameter lines on the tangent plane at the edge of the confidence disk. For example, the parameter-effects value 1.00 corresponds to a deviation of a parameter curve from a straight line at the edge of 100%.

Considering for the single parameters of the studied model, which cause the corresponding value of the parameter-effects curvature, it is possible to calculate a value of $c_{\beta r}^{PE}$ in the direction of a rth component of parameter vector of the model (Bates and Watts, 1988). Table 10 contains the calculated parameter-effects curvature in the directions corresponding to the parameters α, n, m of the van Genuchten models (VG1), (VG2), (VG3), (VG4), (VG5) in combination with the first simulated data set SIMUL_1 of sand.

It is obvious, that the corresponding values for the parameters Θ_r and Θ_s are zero, since these parameters having linear behavior in the models of van Genuchten. Similarly, the linear parameter Θ_s in the models of King has a value of zero. On the other hand, the parameter-effects curvature has the largest values for the parameters n and m, which having strong nonlinear behavior. Therefore, these are strong nonlinear parameters with estimation properties, which greatly differ from the properties of linear or near linear parameters. Since these values depend on the parameterization, it is possible to reduce this nonlinear behavior by a reparametrization.

Table 11: Bias and skewness for the parameter n and its reparameterization $\tilde{n}$ in the models (VG4) and (RVG4) respectively in combination with the first simulated data set of sand.

Model	Parameter	Bias	Skewness
(VG4)	n	0.139	1.540
(RVG4)	$\tilde{n}$	-0.006	-0.284

Considering the reparametrization (RVG4) of the model (VG4), a value of $c^{PE}\sqrt{F} = 0.4908$ for the RMS curvature is got for the same simulated data set SIMUL_1 of sand for the RMS curvature but the corresponding values for the intrinsic curvature do not change. A value 0.53 of the parameter-effects curvature for the direction of the parameter $\tilde{n}$ is obtained. The parameter $\tilde{n}$ in the model (RVG4) has properties, which are better than the corresponding properties with respect to the parameter n in the model (VG4). Moreover, the value of the parameter-effects curvature for (RVG4) is smaller than the corresponding value for the model (VG4).

The values of the RMS curvatures for the models (K1) and (K2) are smaller than the corresponding values for the van Genuchten model (VG1) with five parameters, but as a rule the model (VG4) with four parameters has the best value of the parameter-effects curvature. On the other hand, the King model (K2) with four parameters has more favorable qualities in comparison with the model (K1) of King with five parameters.

The results of the calculations of the bias and the skewness measures for the parameter n and its reparametrization $\tilde{n}$ in the models (VG4) and (RVG4) respectively in combination with the simulated data set SIMUL_1 of sand are given in Table 11.

Referring to Fig. 3, 5 and Table 11, it can be seen that the distribution of the estimated values for the parameter n is skewed with heavy right tail and the distribution of $\tilde{n}$ is near symmetric. The mean value of the estimations of $\tilde{n}$ corresponds to the true value, but the mean value of the estimated values of n is greater than the true one.

Using data sets of different simulations (or soils), similar results with respect to the used measures of nonlinearity will be obtained as a rule.

CONCLUSIONS

Model choice of NLR-models depends on the main aim of the analysis. A retention model, which considers as a NLR-model, has to be able to describe the typical S-shaped retention curve. If one is interested in the best fitting model, very flexible models with more parameters should be used. As a criterion ordinary or in case of heteroscedasticity, weighted least squares can be used.

In case of a retention curve, the models (VG1) and (K1) with five parameters of van Genuchten and King respectively are optimal with respect to the goodness of fit.

If the stable estimation of the parameters of a model is of prime importance, the distributional properties of the model parameters should be studies. These can be done by simulation studies or the calculation of measures of nonlinearity. In this case one looks for models with parameters which have near linear estimation behavior. These estimation properties are dependent on the inner nonlinearity of model data set combination, which is independent of the parameterization and the nonlinearity of model data set combination, which is dependent on the used parameterization. Near linear behavior means unbiased, near symmetrically distributed estimations with a near minimal variability in comparison with an approximating linear model.

Using the list of the studied models for the retention function in combination with the considered data sets (simulated or real), the van Genuchten model (VG4) with four parameters is as an optimal model. At the same time, this model has a strong nonlinear parameter n. A possible appropriate reparametrization with respect to this parameter is the model (RVG4).

ACKNOWLEDGMENTS

I would like to thank the committee of the journal and the referees for their constructive comments.

REFERENCES

Bates, D.M. and D.G. Watts, 1988. Nonlinear Regression Analysis and its Applications, John Wiley and Sons, Inc., New York.

Box, M.J., 1971. Bias in nolinear estimation, J.R. Statist. Soc. Ser., B33: 171-201.

Carsel, R.F. and R.S. Parrish, 1988. Developing joint probability distributions of soil water retention characteristics. Water Resour. Res., 24: 755-769.

El-Shehawy, S.A., 2001. Mathematische modellierung des wasserflusses in naeturlichen boeden statistiche methoden zur auswahl und bewertung von retentionsfunctionen und zur schaetzung ihrer parameter. Ph.D. Thesis, Dresden University of Technology.

El-Shehawy, S.A. and A.A. Karawia, 2006. An alternative computational algorithm for calculating the nonlinearity of regression models with two parameters. Applied Math. Comput., (In Press).

Haines, L.M., T.E. Brien and G.P. Clarke, 2004. Kurtosis and curvature measure for nonlinear regression models. Stat. Sin., 14: 547-570.

Hougaard, P., 1985. The appropriateness of the asymptotic distribution in a nonlinear regression model in relation to curvature. J. R. Statist. Soc. Ser., B47: 103-114.

Johnson, R.A. and G.K. Bhattacharyya, 1992. Statistics: Principles and Methods. Wiley, New York.

King, L.G., 1965. Description of soil characteristics for partially saturated flow. Soil Sci. Soc. Am. Proc., 29: 359-362.

Ratkowsky, D.A., 1983. Nonlinear Regression Modeling. Marcel Dekker, New York.

Ratkowsky, D.A., 1990. Handbook of Nonlinear Regression Models. Marcel Dekker, New York.

Seber, G.A.F. and C.J. Wild, 2005. Nonlinear Regression. Jonh Wiley and Sons, New York.

van Genuchten, M.Th., 1980. A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. Soil Sci. Soc. Am. J., 44: 892-898.

Vereecken, H., J. Feyen, J. Maes and P. Darius, 1989. Estimating the soil moisture retention characteristic form texture, bulk density and carbon content. Soil Sci., 148: 389-403.