Hu et al. Environmental Systems Research (2015) 4:4
DOI 10.1186/s40068-015-0031-4

Environmental Systems Research
a SpringerOpen Journal

RESEARCH ARTICLE Open Access

Evaluation of parameter uncertainties in nonlinear regression using Microsoft Excel Spreadsheet

Wei Hu[1], Jing Xie[1], Henry Wai Chau[2] and Bing Cheng Si[1*]

Abstract

Background: Nonlinear relationships are common in the environmental discipline. Spreadsheet packages such as Microsoft Excel come with an add-on for nonlinear regression, but parameter uncertainty estimates are not yet available. The purpose of this paper is to use Monte Carlo and bootstrap methods to estimate nonlinear parameter uncertainties with a Microsoft Excel spreadsheet. As an example, uncertainties of two parameters (a and n) for a soil water retention curve are estimated.

Results: The fitted parameters generally do not follow a normal distribution. Except for the upper limit of a using the bootstrap method, the lower and upper limits of a and n obtained by these two methods are slightly greater than those obtained using the SigmaPlot software which linearlizes the nonlinear model.

Conclusions: Since the linearization method is based on the assumption of normal distribution of parameter values, the Monte Carlo and bootstrap methods may be preferred to the linearization method.

Keywords: Bootstrap; Monte Carlo; Soil water retention; Parameter uncertainty; Excel

Background

Nonlinear relationships are common in natural and environmental sciences (Wraith and Or 1998; Luo et al. 2003; Cwiertny and Roberts 2005). As a result, there are many software packages (such as SAS and MathCAD) that implement nonlinear parameter estimation. However, spreadsheet techniques are easier to learn than other specialized mathematical programs for nonlinear parameter estimation, because no programming skills are needed in spreadsheets to develop their own parameter estimation routines (Wraith and Or 1998). In addition, spreadsheets have the merits of wide accessibility and powerful computation in terms of fitting nonlinear models. For these reasons, spreadsheets such as Microsoft Excel are widely suggested to make nonlinear parameter estimation (Harris 1998; Smith et al. 1998; Wraith and Or 1998; Brown 2001; Berger 2007).

Parameter uncertainty refers to lack of knowledge regarding the exact true value of a quantity (Tong et al. 2012). Different observations are usually obtained when experiments are repeated, resulting in different values of parameters. It is usually expressed as an interval of parameter values at a certain confidence level, say, 95%. It is also expressed as the standard error of the mean by assuming normal distribution of parameter values. Parameter uncertainty can be used to judge the degree of reliability of the parameter estimates, which is important to making decisions for environmental management. For these reasons, estimation of parameter uncertainties is significant for nonlinear parameter estimates. However, relatively less work has focused on the nonlinear parameter uncertainty estimates using spreadsheet packages.

Parameter uncertainty can be obtained exactly by assuming normal distribution of a parameter in linear regression, but not in nonlinear regression. Nonlinear regression programs usually give the parameter uncertainty by calculating the standard error of the mean, and assuming linear relationship between variables in the vicinity of the estimated parameter values and normal distribution of parameter values. Furthermore, this method usually involves evaluating a Hessian matrix (a square matrix of second-order partial derivatives of a scalar-valued function to describe the local curvature of a function of many variables) or an inequality, which makes it more complicated and time demanding (Brown 2001). More general methods such as Monte Carlo and bootstrap simulation can be used to estimate the parameter uncertainties. Both methods have their

* Correspondence: bing.si@usask.ca
[1]Department of Soil Science, University of Saskatchewan, Saskatoon, SK S7N 5A8, Canada
Full list of author information is available at the end of the article

own advantages: while the Monte Carlo method is based on a theoretical probability distribution of a variable, the bootstrap method has no assumption on the probability distribution of a variable and thus has no limits on sampling size. Among numerous related applications are testing fire ignition selectivity of different landscape characteristics using the Monte Carlo simulation (Conedera et al. 2011) and estimating uncertainty of greenhouse gas emissions using the bootstrap simulation (Tong et al. 2012). However, parameter uncertainties estimation in spreadsheets using the Monte Carlo and bootstrap methods has been rarely discussed.

Both nonlinear parameter values and their associated uncertainties are important for decision making and thus should be implemented in spreadsheet program like Excel. Microsoft Excel spreadsheets have other advantages including their general facility for data input and management, ease in implementing calculations, and often advanced graphics and reporting capabilities (Wraith and Or 1998). These advantages are likely to make the use of spreadsheets to quantify parameter uncertainties more desirable.

The objective of this paper is to apply the Monte Carlo and bootstrap simulations to obtain parameter uncertainties with a Microsoft Excel spreadsheet. In addition, the influences of number of simulation on uncertainty estimates are also discussed. For this, we use as an example, a common soil physical property - soil water retention curve, which has been widely used in soil, hydrological, and environmental communities.

Results and discussion

Nonlinear regression parameters estimation

Here are the steps to estimate parameters α and n in Excel using nonlinear regression.

1. List the applied suction pressure as the independent variable in column A and measured soil water content (θ) as the dependent variable in column B (Figure 1).

2. Temporarily set the value of α as 0.1 and n as 1 in cells B19 and B20, respectively (Figure 1). It is important to set an appropriate initial value because an obviously unreasonable initial value will lead to an unanticipated value. Please refer to related document for initial value estimation (e.g., Delboy 1994). List the measured θ_r and θ_s in cells B21 and B22, respectively. Then the predicted θ value can be calculated with the van Genuchten soil water retention curve model (Eq. 5) using suction pressure and all parameter values. For example, the predicted θ in cell E2 ($\hat{\theta}_{E2}$) is calculated by the following formula:

$$\hat{\theta}_{E2} = \$B\$21 + (\$B\$22 - \$B\$21) * (1 + (\$B\$19 * \$A2)\hat{}\$B\$20)\hat{}(-1 + 1/\$B\$20) \quad (1)$$

As Figure 1 shows, all the predicted θ values are 0.395 given the initial values.

	A	B	C	D	E
1	Soil matrix potential (-cm)	Measured θ (cm^3 cm^{-3})			Predicted θ (cm^3 cm^{-3})
2	0	0.37545			0.39500
3	1	0.38246			0.39500
4	2	0.38483			0.39500
5	3	0.38529			0.39500
6	4	0.38301			0.39500
7	5	0.37299			0.39500
8	6	0.37135			0.39500
9	7	0.36688			0.39500
10	15	0.24659			0.39500
11	25	0.14189			0.39500
12	50	0.09960			0.39500
13	100	0.07646			0.39500
14	300	0.03281			0.39500
15	500	0.03189			0.39500
16	1000	0.01832			0.39500
17	15000	0.01082			0.39500
18					
19	α	0.10000			
20	n	1.00000			
21	θ_r	0.01100			
22	θ_s	0.39500			
23	SSE	0.83005			

Figure 1 Data input and initial value set for α and n in a spreadsheet.

3. Calculate the sum of squared residuals (*SSE*) using Excel function SUMXMY2 in cell B23 by entering "SUMXMY2(B2 : B17, E2 : E17)". We obtain 0.83005 for *SSE* for the given initial parameter values (Figure 1).

4. The model obtains the maximum likelihood when the *SSE* is minimized, which is the principle of least-square fitting method. The *Solver* tool in Excel can be used to minimize the *SSE* values. The *Solver* tool can be found under the Data menu in Excel. If not found there, it has to be added from File menu through the path *File- > Options- > Add-Ins- > Solver Add-in*. As Figure 2 shows, the "Set Objective" box is the value to be optimized, which is the *SSE* value in cell B23. Click "Min" to minimize the objective *SSE* by changing the values of α and n as shown in the "By Changing Variable Cells".

5. The *Solver* will then find the minimum *SSE* (in cell B23) and corresponding α (in cell B19) and n (in cell B20) values (Figure 3). The measured θ values are in agreement with the predicted θ values (Figure 4), indicating a good nonlinear curve fitting.

Using Monte Carlo method to estimate parameter uncertainty

Stepwise application of the Monte Carlo method in estimating parameter uncertainties with 200 simulations is demonstrated below:

1. Resample θ using the Monte Carlo method in different columns. Take cell L2 for example, the simulated θ (θ_{L2}) is calculated by the following formula:

$$\theta_{L2} = \$E2 + \mathrm{NORM.INV}(\mathrm{RAND}(), 0, \mathrm{SQRT}(SSE/df)) \quad (2)$$

where *$*E2 refers to the corresponding predicted θ. *SSE* is the value calculated above, which is the value in cell B23 in Figure 5. The degree of freedom (*df*) equals 14. The θ values in the other rows in column L are simulated in a similar way. The simulated values for a new dependent variable θ are demonstrated in cells L2-L17 (Figure 5). The same Monte Carlo simulations are performed from column M to column HC. Therefore, a total of 200 sets of simulated θ are obtained (Figure 5).

2. Use the same procedure as introduced before to conduct nonlinear regression for each new data set of θ. Note that the new data set of θ will change during optimization, which will result in errors in fitting. Therefore, we copy the simulated θ data to a new sheet by right-clicking "Paste Special" and selecting "Values" in the dialogue of "Paste Special". For better display, predicted θ array, all the parameter values, and corresponding *SSE* value are presented in the same column for each simulation (Figure 5). Optimization of parameters α and n is made independently for each simulation using the *Solver* tool. The initial values are set as the fitted values

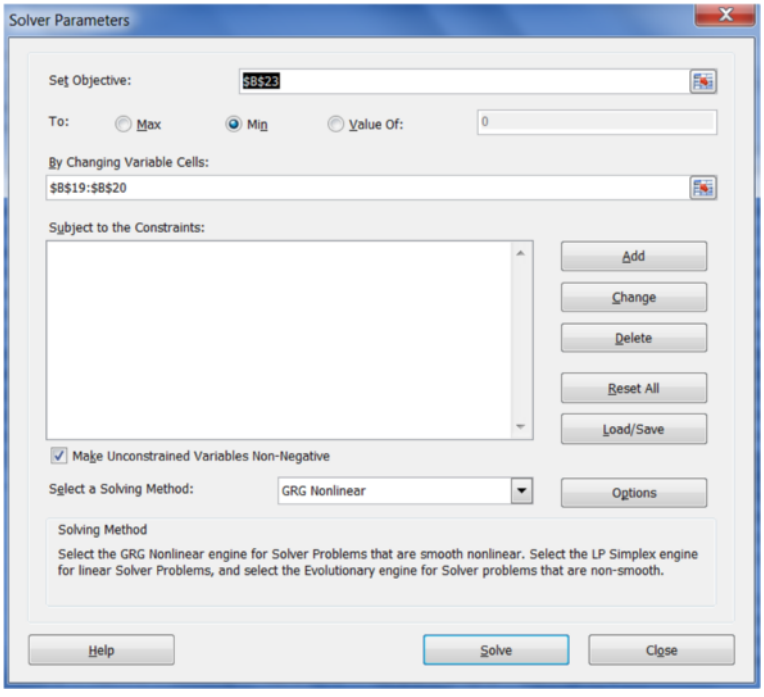

Figure 2 Solver working screen in Excel 2010.

	A	B	C	D	E
1	Soil matrix potential (-cm)	Measured θ (cm^3 cm^{-3})			Predicted θ (cm^3 cm^{-3})
2	0	0.37545			0.39500
3	1	0.38246			0.39401
4	2	0.38483			0.39080
5	3	0.38529			0.38536
6	4	0.38301			0.37788
7	5	0.37299			0.36865
8	6	0.37135			0.35804
9	7	0.36688			0.34642
10	15	0.24659			0.25062
11	25	0.14189			0.17179
12	50	0.09960			0.09245
13	100	0.07646			0.04981
14	300	0.03281			0.02266
15	500	0.03189			0.01765
16	1000	0.01832			0.01410
17	15000	0.01082			0.01116
18					
19	α	0.07988			
20	n	2.09920			
21	θ_r	0.01100			
22	θ_s	0.39500			
23	SSE	0.00319			

Figure 3 A spreadsheet for estimating nonlinear regression coefficients α and n.

obtained before for all simulations to reduce the time required during the optimization. Therefore, initial values of 0.07988 and 2.09920 are set for α and n, respectively (Figure 5). Because the maximum number of variables *Solver* can solve is 200, we can minimize the *SSE* values for 100 simulations at one time by minimizing the sum of *SSE* values of 100 simulations. For example, the parameters α and n for the first 100 simulations can be optimized by minimizing cell E19 by entering "=SUM(L18:DG18)" (Figure 5). Similarly, the parameters for the second 100 simulations can be optimized in cell E20 by entering "=SUM(DH18:HC18)". Therefore, we obtain 200 values for both parameters (α and n) as shown in cells from L19 to HC20 (Figure 6). The frequency distribution of α and n are shown in Figure 7. Visually, both of them follow a normal distribution. However, the Shapiro-Wilk test shows that the parameter α does not conform to a normal distribution, whereas parameter n does. This indicates that the fitted parameters may not necessarily be normally distributed even if the dependent variable is normally distributed, due to the nonlinear relations between them.

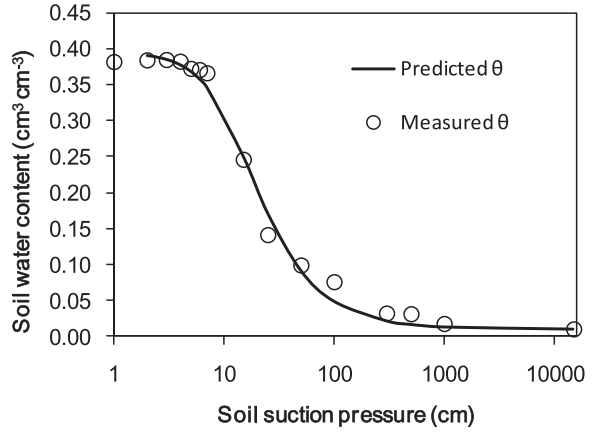

Figure 4 Measured and predicted soil water content versus soil suction pressure.

3. Calculate the 95% confidence interval of α or n values with 200 simulations. We copy all the fitted α or n values, then paste them to a new sheet by right-clicking "Paste Special" and selecting "Transpose" in the dialogue of "Paste Special" to list all the fitted α or n values in one column. Select the transposed data, and rank them in an ascending order using *Sort* tool in *Data* tab. Find the value of α and n corresponding to the 2.5 percentile and 97.5 percentile, which are the lower limit and upper limit, respectively, at a 95% confidence. The 95% confidence interval are (0.0680, 0.0939) and (1.9185, 2.3689) for α and n, respectively (Table 1). The difference between upper limit and lower limit is 0.0259 and 0.4504 for α and n, respectively.

	A	B	C	D	E	K	L	M	N	O	P
1	Soil matrix potential (-cm)	Measured θ (cm^3 cm^{-3})			Predicted θ (cm^3 cm^{-3})		Monte Carlo				
2	0	0.37545			0.39500		0.40260	0.39739	0.40217	0.40302	0.38397
3	1	0.38246			0.39401		0.41745	0.37636	0.39762	0.38644	0.38687
4	2	0.38483			0.39080		0.39070	0.40722	0.37544	0.37140	0.39730
5	3	0.38529			0.38536		0.39195	0.37067	0.37169	0.39430	0.40272
6	4	0.38301			0.37788		0.40189	0.39625	0.38612	0.36903	0.37417
7	5	0.37299			0.36865		0.38383	0.37797	0.38544	0.37240	0.35964
8	6	0.37135			0.35804		0.35263	0.36020	0.37474	0.38739	0.36548
9	7	0.36688			0.34642		0.34043	0.33859	0.34283	0.34326	0.37098
10	15	0.24659			0.25062		0.26945	0.24960	0.25453	0.24270	0.23907
11	25	0.14189			0.17179		0.15018	0.17221	0.17870	0.17457	0.15425
12	50	0.09960			0.09245		0.10468	0.08979	0.10991	0.11010	0.10472
13	100	0.07646			0.04981		0.06072	0.07586	0.03827	0.06461	0.04422
14	300	0.03281			0.02266		0.02130	-0.00292	0.05046	0.02671	0.01236
15	500	0.03189			0.01765		0.04485	0.00853	0.03129	0.02363	-0.01379
16	1000	0.01832			0.01410		0.00228	0.00753	0.01668	0.01281	0.03531
17	15000	0.01082			0.01116		0.00160	0.02626	0.01119	0.00555	-0.00526
18						*SSE*	0.00359	0.00299	0.00259	0.00223	0.00371
19	α	0.07988	1-100	SUM_SSE(L:DG)	0.34049	α	0.07988	0.07988	0.07988	0.07988	0.07988
20	n	2.09920	101-200	SUM_SSE(DH:HC)	0.34874	n	2.09920	2.09920	2.09920	2.09920	2.09920
21	θ_r	0.01100				θ_r	0.01100	0.01100	0.01100	0.01100	0.01100
22	θ_s	0.39500				θ_s	0.39500	0.39500	0.39500	0.39500	0.39500
23	*SSE*	0.00319					1	2	3	4	5
24											
25				0			0.39500	0.39500	0.39500	0.39500	0.39500
26				1			0.39401	0.39401	0.39401	0.39401	0.39401
27				2			0.39079	0.39079	0.39079	0.39079	0.39079
28				3			0.38534	0.38534	0.38534	0.38534	0.38534
29				4			0.37785	0.37785	0.37785	0.37785	0.37785
30				5			0.36861	0.36861	0.36861	0.36861	0.36861
31				6			0.35799	0.35799	0.35799	0.35799	0.35799
32				7			0.34636	0.34636	0.34636	0.34636	0.34636
33				15			0.25059	0.25059	0.25059	0.25059	0.25059
34				25			0.17181	0.17181	0.17181	0.17181	0.17181
35				50			0.09250	0.09250	0.09250	0.09250	0.09250
36				100			0.04986	0.04986	0.04986	0.04986	0.04986
37				300			0.02268	0.02268	0.02268	0.02268	0.02268
38				500			0.01767	0.01767	0.01767	0.01767	0.01767
39				1000			0.01411	0.01411	0.01411	0.01411	0.01411
40				15000			0.01116	0.01116	0.01116	0.01116	0.01116

Figure 5 Resampling dependent variable θ using Monte Carlo method and initial values set (only the first 5 simulations are shown).

Using bootstrap method to estimate parameter uncertainty

Stepwise application of the bootstrap method in estimating parameter uncertainties with 200 simulations is demonstrated as follows:

1. Resample θ using the bootstrap method in different columns (Figure 8). Take cell L2 for example, the simulated θ (θ_{L2}) can be calculated by the following formula:

$$\theta_{L2} = \$E2 + \mathrm{INDEX}(\$C : \$C, \mathrm{INT}(\mathrm{RAND}() * 16 + 2)) \quad (3)$$

where the function INDEX is used to randomly select a residue value from row 2 to row 17 in column C (the residual is calculated by subtracting predicted θ from the original θ). The θ values at other rows in column L and in other columns (column M to column HC) are simulated in a similar way. Here, 2 in the right hand side of Eq. (3) means that data start at second row.

2. Similar to the Monte Carlo method, parameters α and n for all simulations are fitted by minimizing the sum of every 100 *SSE* values using the *Solver* tool (Figures 8 and 9). Therefore, we can also obtain 200 values for both parameters (α and n) as shown in cells from L19 to HC20 (Figure 9). The frequency distribution of α and n are shown in Figure 10. They also visually follow a normal distribution. However, the Shapiro-Wilk test shows that the parameter α does not conform to normal distribution, whereas parameter n does.

3. Similar to the Monte Carlo method, the 95% confidence intervals for these two parameters are calculated. They are (0.0680, 0.0925) and (1.9172, 2.3356), for α and n respectively (Table 1). The corresponding difference between upper limit and lower limit is 0.0245 and 0.4183 for α and n, respectively.

Influences of number of simulation on parameter uncertainty analysis

Datasets of fitted values with different numbers of simulations are obtained using a similar method as demonstrated

	A	B	C	D	E	K	L	M	N	O	P
1	Soil matrix potential (-cm)	Measured θ (cm^3 cm^{-3})			Predicted θ (cm^3 cm^{-3})		Monte Carlo				
2	0	0.37545			0.39500		0.40260	0.39739	0.40217	0.40302	0.38397
3	1	0.38246			0.39401		0.41745	0.37636	0.39762	0.38644	0.38687
4	2	0.38483			0.39080		0.39070	0.40722	0.37544	0.37140	0.39730
5	3	0.38529			0.38536		0.39195	0.37067	0.37169	0.39430	0.40272
6	4	0.38301			0.37788		0.40189	0.39625	0.38612	0.36903	0.37417
7	5	0.37299			0.36865		0.38383	0.37797	0.38544	0.37240	0.35964
8	6	0.37135			0.35804		0.35263	0.36020	0.37474	0.38739	0.36548
9	7	0.36688			0.34642		0.34043	0.33859	0.34283	0.34326	0.37098
10	15	0.24659			0.25062		0.26945	0.24960	0.25453	0.24270	0.23907
11	25	0.14189			0.17179		0.15018	0.17221	0.17870	0.17457	0.15425
12	50	0.09960			0.09245		0.10468	0.08979	0.10991	0.11010	0.10472
13	100	0.07646			0.04981		0.06072	0.07586	0.03827	0.06461	0.04422
14	300	0.03281			0.02266		0.02130	-0.00292	0.05046	0.02671	0.01236
15	500	0.03189			0.01765		0.04485	0.00853	0.03129	0.02363	-0.01379
16	1000	0.01832			0.01410		0.00228	0.00753	0.01668	0.01281	0.03531
17	15000	0.01082			0.01116		0.00160	0.02626	0.01119	0.00555	-0.00526
18						*SSE*	0.00351	0.00298	0.00225	0.00195	0.00345
19	α	0.07988	1-100	SUM_SSE(L:DG)	0.29465	α	0.07698	0.07884	0.07677	0.08020	0.07722
20	n	2.09920	101-200	SUM_SSE(DH:HC)	0.30399	n	2.10742	2.10373	2.04938	2.01214	2.21869
21	θ_r	0.01100				θ_r	0.01100	0.01100	0.01100	0.01100	0.01100
22	θ_s	0.39500				θ_s	0.39500	0.39500	0.39500	0.39500	0.39500
23	*SSE*	0.00319					1	2	3	4	5
24											
25				0			0.39500	0.39500	0.39500	0.39500	0.39500
26				1			0.39410	0.39404	0.39398	0.39380	0.39428
27				2			0.39115	0.39093	0.39084	0.39023	0.39170
28				3			0.38611	0.38564	0.38565	0.38446	0.38702
29				4			0.37915	0.37834	0.37860	0.37677	0.38027
30				5			0.37050	0.36931	0.37000	0.36753	0.37164
31				6			0.36048	0.35890	0.36015	0.35710	0.36139
32				7			0.34944	0.34748	0.34939	0.34585	0.34990
33				15			0.25588	0.25237	0.25972	0.25565	0.24919
34				25			0.17630	0.17321	0.18292	0.18098	0.16458
35				50			0.09479	0.09312	0.10171	0.10239	0.08306
36				100			0.05078	0.05005	0.05588	0.05733	0.04262
37				300			0.02286	0.02269	0.02527	0.02634	0.01933
38				500			0.01774	0.01765	0.01935	0.02015	0.01547
39				1000			0.01413	0.01410	0.01504	0.01554	0.01292
40				15000			0.01116	0.01116	0.01124	0.01129	0.01107

Figure 6 Nonlinear regression fitting for resampled θ using Monte Carlo method (only the first 5 simulations are shown).

before (data not shown). According to Shapiro-Wilk test, both parameters do not follow a normal distribution with different numbers of simulations except for a few cases that the number of simulations ≤400. This may indicate that the assumption of normality in the linearization method does not hold true.

The lower limit, upper limit, and their difference change slightly with the number of simulations. However, they are almost constant beyond a certain number of simulations (Figure 11). Here, we determine the number of simulations required for both methods according to the change in difference between the upper limit and lower limit. If

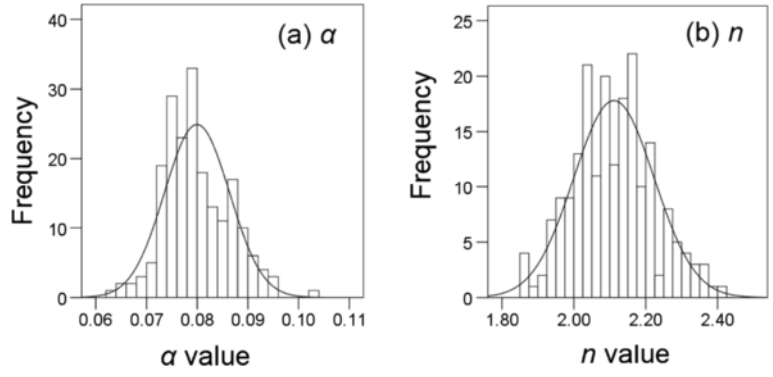

Figure 7 Frequency distribution of (a) α and (b) n obtained using Monte Carlo method with 200 simulations. The heights of bars indicate the number of parameter values in the equally spaced bins. The curve is the theoretical normal distribution.

Table 1 Comparison of parameter uncertainties calculated by different methods

Parameter		Monte Carlo	Bootstrap	Linearization
α	Lower limit	0.0680	0.0680	0.0670
	Upper limit	0.0939	0.0925	0.0928
	Upper limit-Lower limit	0.0259	0.0245	0.0258
n	Lower limit	1.9185	1.9172	1.8758
	Upper limit	2.3689	2.3356	2.3218
	Upper limit-Lower limit	0.4504	0.4184	0.4460

the relative difference (RD%) of the difference between the upper limit and lower limit under a certain number of simulation is less than 5% compared with that under 2000 simulations, the number of simulations tested is taken to be the required number of simulations. The RD% can be calculated as

$$\mathrm{RD}\% = \left|\frac{\mathrm{V}_m - \mathrm{V}_{2000}}{\mathrm{V}_{2000}}\right| * 100\% \quad (4)$$

where $\mathrm{V_m}$ and $\mathrm{V_{2000}}$ are the differences between the upper limit and lower limit under m simulations and 2000 simulations, respectively.

For the Monte Carlo method, the RD% is less than 5% for α and n when the numbers of simulation are ≥100 and 200, respectively. For the bootstrap method, the RD % is less than 5% for α and n when the numbers of simulation are ≥500 and 400, respectively. Therefore, simulation number of 200 and 500 are needed to produce reliable data at the 95% confidence interval of parameters for the Monte Carlo and bootstrap methods, respectively. In this sense, the Monte Carlo method may be better than the bootstrap method. However, the optimal number of simulation may also differ with specific situations. For example, Efron and Tibshirani (1993) stated that a minimum of approximately 1000 bootstrap re-samples was sufficient to obtain accurate confidence

	A	B	C	D	E	K	L	M	N	O	P
1	Soil matrix potential (-cm)	Measured θ (cm^3 cm^{-3})	Residue (cm^3 cm^{-3})		Predicted θ (cm^3 cm^{-3})		Bootstrap				
2	0	0.37545	-0.01955		0.39500		0.38346	0.36509	0.38903	0.38903	0.40515
3	1	0.38246	-0.01154		0.39401		0.40825	0.39914	0.40732	0.37446	0.39393
4	2	0.38483	-0.00597		0.39080		0.39514	0.38483	0.40411	0.39593	0.39501
5	3	0.38529	-0.00007		0.38536		0.40583	0.37940	0.37382	0.40583	0.40583
6	4	0.38301	0.00513		0.37788		0.38803	0.38803	0.34797	0.38301	0.40453
7	5	0.37299	0.00434		0.36865		0.37880	0.37378	0.37287	0.36858	0.38290
8	6	0.37135	0.01331		0.35804		0.37228	0.37135	0.36519	0.37228	0.38468
9	7	0.36688	0.02046		0.34642		0.35063	0.36688	0.35063	0.35357	0.35973
10	15	0.24659	-0.00403		0.25062		0.23908	0.25575	0.25575	0.26393	0.23908
11	25	0.14189	-0.02991		0.17179		0.18510	0.17692	0.16776	0.17172	0.16025
12	50	0.09960	0.00716		0.09245		0.08842	0.11291	0.09678	0.08648	0.10260
13	100	0.07646	0.02665		0.04981		0.04973	0.03026	0.05996	0.03826	0.04947
14	300	0.03281	0.01015		0.02266		0.03597	0.02232	0.01111	0.01111	0.01863
15	500	0.03189	0.01425		0.01765		0.02278	0.03189	0.02186	0.03096	0.01757
16	1000	0.01832	0.00421		0.01410		0.02741	0.02741	0.04075	0.00256	0.02126
17	15000	0.01082	-0.00034		0.01116		0.01108	0.01549	0.01108	0.01549	0.01108
18						*SSE*	0.00191	0.00297	0.00253	0.00196	0.00278
19	α	0.07988	1-100	SUM_SSE(L:DG)	0.30515	α	0.07988	0.07988	0.07988	0.07988	0.07988
20	n	2.09920	101-200	SUM_SSE(DH:HC)	0.32341	n	2.09920	2.09920	2.09920	2.09920	2.09920
21	θ_r	0.01100				θ_r	0.01100	0.01100	0.01100	0.01100	0.01100
22	θ_s	0.39500				θ_s	0.39500	0.39500	0.39500	0.39500	0.39500
23	*SSE*	0.00319					1	2	3	4	5
24											
25				0			0.39500	0.39500	0.39500	0.39500	0.39500
26				1			0.39401	0.39401	0.39401	0.39401	0.39401
27				2			0.39079	0.39079	0.39079	0.39079	0.39079
28				3			0.38534	0.38534	0.38534	0.38534	0.38534
29				4			0.37785	0.37785	0.37785	0.37785	0.37785
30				5			0.36861	0.36861	0.36861	0.36861	0.36861
31				6			0.35799	0.35799	0.35799	0.35799	0.35799
32				7			0.34636	0.34636	0.34636	0.34636	0.34636
33				15			0.25059	0.25059	0.25059	0.25059	0.25059
34				25			0.17181	0.17181	0.17181	0.17181	0.17181
35				50			0.09250	0.09250	0.09250	0.09250	0.09250
36				100			0.04986	0.04986	0.04986	0.04986	0.04986
37				300			0.02268	0.02268	0.02268	0.02268	0.02268
38				500			0.01767	0.01767	0.01767	0.01767	0.01767
39				1000			0.01411	0.01411	0.01411	0.01411	0.01411
40				15000			0.01116	0.01116	0.01116	0.01116	0.01116

Figure 8 Resampling dependent variable θ using bootstrap method and initial values set (only the first 5 simulations are shown).

	A	B	C	D	E	K	L	M	N	O	P
1	Soil matrix potential (-cm)	Measured θ (cm^3 cm^{-3})	Residue (cm^3 cm^{-3})		Predicted θ (cm^3 cm^{-3})		Bootstrap				
2	0	0.37545	-0.01955		0.39500		0.38346	0.36509	0.38903	0.38903	0.40515
3	1	0.38246	-0.01154		0.39401		0.40825	0.39914	0.40732	0.37446	0.39393
4	2	0.38483	-0.00597		0.39080		0.39514	0.38483	0.40411	0.39593	0.39501
5	3	0.38529	-0.00007		0.38536		0.40583	0.37940	0.37382	0.40583	0.40583
6	4	0.38301	0.00513		0.37788		0.38803	0.38803	0.34797	0.38301	0.40453
7	5	0.37299	0.00434		0.36865		0.37880	0.37378	0.37287	0.36858	0.38290
8	6	0.37135	0.01331		0.35804		0.37228	0.37135	0.36519	0.37228	0.38468
9	7	0.36688	0.02046		0.34642		0.35063	0.36688	0.35063	0.35357	0.35973
10	15	0.24659	-0.00403		0.25062		0.23908	0.25575	0.25575	0.26393	0.23908
11	25	0.14189	-0.02991		0.17179		0.18510	0.17692	0.16776	0.17172	0.16025
12	50	0.09960	0.00716		0.09245		0.08842	0.11291	0.09678	0.08648	0.10260
13	100	0.07646	0.02665		0.04981		0.04973	0.03026	0.05996	0.03826	0.04947
14	300	0.03281	0.01015		0.02266		0.03597	0.02232	0.01111	0.01111	0.01863
15	500	0.03189	0.01425		0.01765		0.02278	0.03189	0.02186	0.03096	0.01757
16	1000	0.01832	0.00421		0.01410		0.02741	0.02741	0.04075	0.00256	0.02126
17	15000	0.01082	-0.00034		0.01116		0.01108	0.01549	0.01108	0.01549	0.01108
18						*SSE*	0.00177	0.00246	0.00246	0.00129	0.00224
19	α	0.07988	1-100	SUM_SSE(L:DG)	0.26909	α	0.07524	0.07137	0.08117	0.07007	0.07234
20	n	2.09920	101-200	SUM_SSE(DH:MC)	0.28362	n	2.14942	2.17036	2.04523	2.29828	2.27779
21	θ_r	0.01100				θ_r	0.01100	0.01100	0.01100	0.01100	0.01100
22	θ_s	0.39500				θ_s	0.39500	0.39500	0.39500	0.39500	0.39500
23	*SSE*	0.00319					1	2	3	4	5
24											
25				0			0.39500	0.39500	0.39500	0.39500	0.39500
26				1			0.39421	0.39433	0.39385	0.39452	0.39446
27				2			0.39154	0.39201	0.39032	0.39265	0.39239
28				3			0.38688	0.38789	0.38452	0.38911	0.38852
29				4			0.38031	0.38202	0.37671	0.38381	0.38278
30				5			0.37203	0.37455	0.36724	0.37680	0.37526
31				6			0.36232	0.36568	0.35651	0.36823	0.36614
32				7			0.35151	0.35569	0.34489	0.35832	0.35569
33				15			0.25716	0.26454	0.25163	0.26211	0.25745
34				25			0.17536	0.18137	0.17552	0.17263	0.16924
35				50			0.09226	0.09483	0.09733	0.08412	0.08314
36				100			0.04849	0.04921	0.05373	0.04147	0.04145
37				300			0.02167	0.02163	0.02464	0.01836	0.01852
38				500			0.01694	0.01685	0.01900	0.01479	0.01492
39				1000			0.01368	0.01360	0.01488	0.01254	0.01262
40				15000			0.01112	0.01111	0.01123	0.01105	0.01105

Figure 9 Nonlinear regression fitting for resampled θ using bootstrap method (only the first 5 simulations are shown).

interval estimates. In order to obtain reliable confidence interval estimates, we suggest increasing the simulation times by 100 at each step, and the final results can be obtained when the values stabilize within consecutive steps.

Comparison with parameter uncertainty approximated by linear model

The values α and n are estimated to be 0.0799 and 2.0988, respectively, by SigmaPlot 10.0 (Figure 12), which are exactly the same as the estimates made by nonlinear

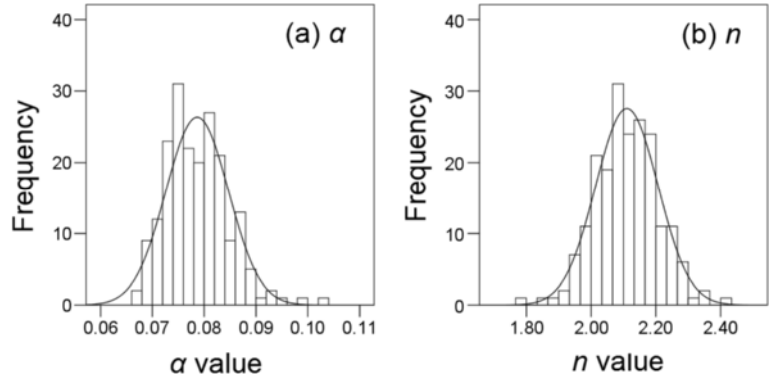

Figure 10 Frequency distribution of (a) α and (b) n obtained using bootstrap method with 200 simulations. The heights of bars indicate the number of parameter values in the equally spaced bins. The curve is the theoretical normal distribution.

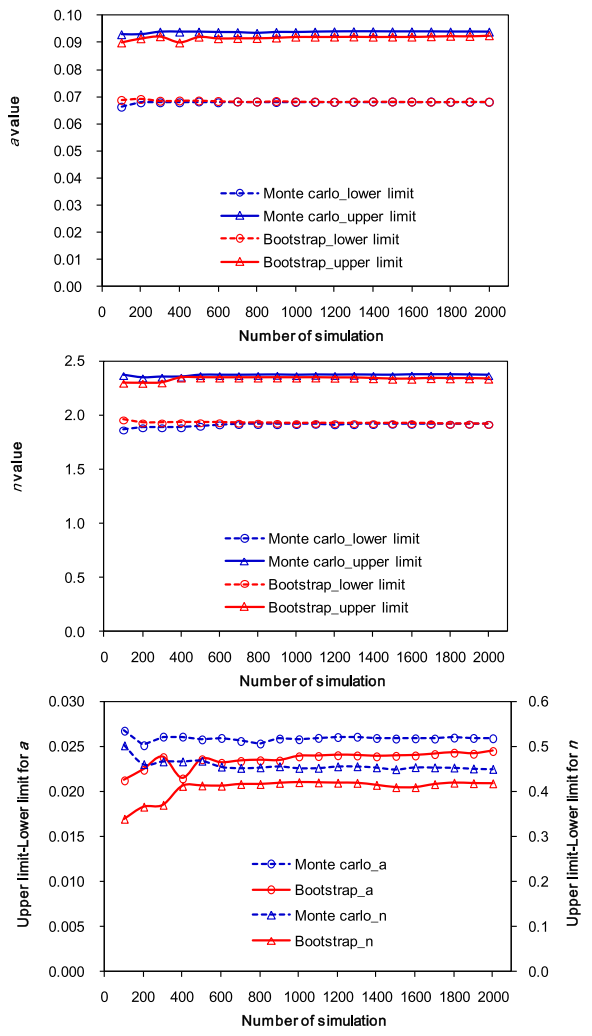

Figure 11 Influences of number of simulation on the lower limit, upper limit, and difference of upper limit and lower limit.

regression in Excel. Based on the linear model, the associated standard errors are estimated to be 0.0066 and 0.1138, respectively (Figure 12). Then the 95% confidence intervals of α and n are (0.0670, 0.0928) and (1.8758, 2.3218), respectively. As Table 1 shows, the difference between upper limit and lower limit by the Monte Carlo and bootstrap methods are comparable, although the Monte Carlo method produces a slightly greater uncertainty than the bootstrap method. The slight difference is due to the differences in re-sampling residues. While the Monte Carlo simulation generates residues based on a theoretical normal distribution, the bootstrap method randomly takes the residues with replacement and no assumption is made about the underlying distributions. They are also comparable to those approximated by the linear model obtained from the SigmaPlot software. However, by comparing the results of these three methods, the lower limit and upper limit of α and n obtained by the Monte Carlo and bootstrap methods are slightly greater than those obtained based on a linear assumption except for upper limit of α by the bootstrap method. Because the linearization method is based on the assumption of normal distribution of parameters and linearity at the vicinity of the estimated parameter value, and it is more complicated in terms of calculation, the Monte Carlo and bootstrap methods may be preferred to the linearization method to calculate the parameter uncertainties in spreadsheets. Furthermore, the Monte Carlo method may be preferred to the bootstrap method considering the less number of simulations required for the Monte Carlo method. However, if the number of measurements is too small to determine the probability distribution for Monte Carlo method, the bootstrap method may be superior.

Conclusions

This paper shows step-by-step how to use a Microsoft Excel spreadsheet to fit nonlinear parameters and to estimate their uncertainties using the Monte Carlo and bootstrap methods. Both Monte Carlo and bootstrap methods can be applied in Excel spreadsheets to resample a large number of measurements for dependent variable from which different values of parameters can be obtained. Our results clearly show that the Monte Carlo and bootstrap methods can be used to estimate the parameter uncertainties using spreadsheet methods. The main limitation is that one execution of standard Microsoft Excel Solver has a limit of 200 simultaneous optimizations. This limit can be overcome by multiple independent executions of Solver. Due to the wide accessibility of Microsoft Excel software and ease of use for these two methods, employing the

R	Rsqr	Adj Rsqr	Standard Error of Estimate
0.9960	0.9919	0.9913	0.0151

	Coefficient	Std. Error	t	P	VIF
a	0.0799	0.0066	12.0888	<0.0001	2.0984
n	2.0988	0.1138	18.4450	<0.0001	2.0984

Figure 12 Estimates of parameters (α and n) and associated standard errors using SigmaPlot 10.0.

Monte Carlo and bootstrap methods in spreadsheets is strongly recommended to estimate nonlinear regression parameter uncertainties.

In this paper, we demonstrated the methodology with the van Genuchten water retention curve. The method can be applied to any mathematical functions or models that can be evaluated by Excel. Therefore, the methodology presented in this paper has wide applicability. Further, with little modification, the Monte Carlo method or bootstrap method can be used in Microsoft Excel to estimate the uncertainty of hydrologic or environmental predictions with single or multiple input parameters under different degrees of uncertainty.

Methods

Soil water retention curve

Soil water content is a function of soil matric potential ψ under equilibrium conditions, and this relationship $\theta(\psi)$ can be described by different types of water retention curves. The soil water retention curve is a basic soil property and is critical for predicting water related environmental processes (Fredlund et al. 1994). Among various soil water retention curve models, the van Genuchten (1980) model is the most widely used one (Han et al. 2010). It is highly nonlinear and can be expressed as:

$$\theta(\psi) = \theta_r + (\theta_s - \theta_r)(1 + (\alpha|\psi|)^n)^{-1+\frac{1}{n}} \tag{5}$$

where $\theta(\psi)$ is the soil water content [$cm^3\ cm^{-3}$] at soil water potential ψ (−cm of water), θ_r is the residual water content [$cm^3\,cm^{-3}$], θ_s is the saturated water content [$cm^3\ cm^{-3}$], α is related to the inverse of the air entry suction [cm^{-1}], and n is a measure of the pore-size distribution (dimensionless). We measured $\theta(\psi)$ at 16 soil water potentials for a sandy soil using Tempe pressure cells (at soil matrix potentials ranging from 0 to −500 cm) and pressure plates (at soil water potentials of −1000 and −15000 cm). θ_s is measured using oven drying method after saturation, and θ_r is estimated as water content of soil approaching air-dry conditions (Wang et al. 2002). θ_s and θ_r are 0.395 and 0.011, respectively. Note that the soil water content (0.375) at zero matrix potential is lower than θ_s due to the soil water movement under gravity. This paper will focus on the estimation of parameters α and n and their associated 95% confidence intervals.

Parameter uncertainty estimation by linearization of nonlinear model

We express the van Genuchten model (Eq. 5) as:

$$\theta_i = f(\beta, \psi_i) + \varepsilon_i \tag{6}$$

where θ_i is the i th observation for the dependent variable $\theta(\psi)$ ($i = 1, 2, \ldots 16$), ψ_i is the i th observation for the predictor $|\psi|$. β is a vector of parameters which includes parameters α and n. ε_i is a random error, which is assumed to be independent of the errors of other observations and normally distributed with a mean of zero and variance of σ^2.

The sum of squared residuals (*SSE*) for nonlinear regression can be written as:

$$SSE(\beta) = \sum (\theta_i - f(\beta, \psi_i))^2 \tag{7}$$

The model has the maximum likelihood when the *SSE* is minimized. Namely, when the partial derivative

$$\frac{\partial SSE(\beta)}{\partial \beta} = -2\sum (\theta_i - f(\beta, \psi_i)) \frac{\partial f(\beta, \psi_i)}{\partial \beta} \tag{8}$$

is zero, parameters β are optimized. Once the optimum values of β are obtained, the parameter uncertainties can be estimated by linearizing the nonlinear model function at the optimum point using the first-order Taylor series expansion method (Fox and Weisberg 2010).

Let

$$\boldsymbol{F_{ij}} = \frac{\partial f\left(\hat{\beta}, \psi_i\right)}{\partial \beta_j} \tag{9}$$

where $\hat{\beta}$ is the optimized value, j refers to the jth of parameters ($j = 1$, 2, and $\beta_1 = \alpha$, $\beta_2 = n$).

Assume matrix $\boldsymbol{F} = [\boldsymbol{F_{ij}}]$. In our case,

$$\boldsymbol{F} = \begin{bmatrix} \frac{\partial f\left(\hat{\beta}, \psi_1\right)}{\partial \alpha} & \frac{\partial f\left(\hat{\beta}, \psi_1\right)}{\partial n} \\ \frac{\partial f\left(\hat{\beta}, \psi_2\right)}{\partial \alpha} & \frac{\partial f\left(\hat{\beta}, \psi_2\right)}{\partial n} \\ \vdots & \vdots \\ \frac{\partial f\left(\hat{\beta}, \psi_{16}\right)}{\partial \alpha} & \frac{\partial f\left(\hat{\beta}, \psi_{16}\right)}{\partial n} \end{bmatrix} \tag{10}$$

where $\frac{\partial f(\hat{\beta}, \psi_i)}{\partial \alpha}$ and $\frac{\partial f(\hat{\beta}, \psi_i)}{\partial n}$ can be calculated by the following formulae:

$$\frac{\partial f\left(\hat{\beta}, \psi_i\right)}{\partial \alpha} = (f((\hat{\alpha} + \Delta\hat{\alpha}), \hat{n}, \psi_i) - f((\hat{\alpha} - \Delta\hat{\alpha}), \hat{n}, \psi_i))/(2\Delta\hat{\alpha}) \tag{11}$$

$$\frac{\partial f\left(\hat{\beta}, \psi_i\right)}{\partial n} = (f(\hat{\alpha}, (\hat{n} + \Delta\hat{n}), \psi_i) - f(\hat{\alpha}, (\hat{n} - \Delta\hat{n}), \psi_i))/(2\Delta\hat{n}) \tag{12}$$

where $\Delta = 0.015$, $\hat{\alpha}$ and $\hat{n}$ are optimized value of α and n, respectively.

The estimated asymptotic covariance matrix (V) of the estimated parameters can be obtained by (Fox and Weisberg 2010):

$$V = \begin{bmatrix} \delta^2_{\alpha\alpha} & \delta^2_{\alpha n} \\ \delta^2_{n\alpha} & \delta^2_{nn} \end{bmatrix} = \sigma^2 \left(\boldsymbol{F}' \boldsymbol{F} \right)^{-1} \quad (13)$$

where $(\boldsymbol{F}'\boldsymbol{F})^{-1}$ is the inverse of $\boldsymbol{F}'\boldsymbol{F}$, and $\boldsymbol{F}'$ is a transpose of $\boldsymbol{F}$.

The σ^2 can be approximated by dividing the *SSE* by the degree of freedom, *df*, as in the form (Brown 2001):

$$\sigma^2 = \frac{SSE}{df} \quad (14)$$

where *df* is calculated as the number of observations in the sample minus the number of parameters. In this study, *df* equals 14 (i.e., 16 minus 2).

Therefore, $\delta^2_{\alpha\alpha}$, δ^2_{nn}, $\delta^2_{\alpha n}$ (or $\delta^2_{n\alpha}$) in Eq. (13) are the estimated variance of α, variance of n, and covariance of α and n, respectively. Specifically, $\delta_{\alpha\alpha}$ and δ_{nn} are the standard errors used to characterize the uncertainties of α and n, respectively. At 95% confidence, the intervals of α and n are $\hat{\alpha}$ ±1.96 $\delta_{\alpha\alpha}$, $\hat{n}$ ±1.96 δ_{nn}, respectively. SigmaPlot 10.0 is used to estimate the parameters and associated standard errors.

Monte Carlo method to estimate parameter uncertainty

Monte Carlo method is an analytical technique for solving a problem by performing a large number of simulations and inferring a solution from the collective results of the simulations. It is a method to calculate the probability distribution of possible outcomes.

In this paper, Monte Carlo simulation is performed to obtain residues of dependent variable θ. The residues follow a specified distribution with a mean of zero and standard deviation of $\sqrt{SSE/df}$. The simulated residues are added to the predicted θ ($\hat{\theta}$) to reconstruct new observations for dependent variable θ. The expression for obtaining new observations for dependent variable θ in Excel is:

$$\theta = \hat{\theta} + \mathrm{NORM.INV}\left(\mathrm{RAND}(), 0, \mathrm{SQRT}\left(\frac{SSE}{df} \right) \right) \quad (15)$$

where function NORM.INV gives a value which follow a normal distribution with a mean of zero and standard deviation of $\sqrt{SSE/df}$ at a probability of RAND(). Therefore, normal distribution on the θ is assumed for Monte Carlo method. Excel function RAND produces a random value that is greater than or equal to 0 and less than 1. SQRT is a function to obtain the square root of a variable.

Monte Carlo simulations are performed 2000 times. Nonlinear regression is made on the simulated θ values versus $|\psi|$ to obtain 2000 values for parameters α and n. The fitted values with different numbers (from 100 to 2000 with intervals of 100) of simulation is analyzed separately to determine the influences of number of simulation on uncertainty estimates. For each dataset, the probability distribution of α and n will be determined by the Shapiro-Wilk test using SPSS 16.0, and the 95% confidence intervals of α and n will be calculated to represent their uncertainties. For simplification, only 200 simulations are shown as an example. Readers can run different numbers of simulation by analogy.

Bootstrap method to estimate parameter uncertainty

Bootstrap method is an alternative method first introduced by Efron (1979) for determining uncertainty in any statistic caused by sampling error. The main idea of this method is to resample with replacement from the sample data at hand and create a large number of "phantom samples" known as bootstrap samples (Singh and Xie 2013). Bootstrap method is a nonparametric method which requires no assumptions about the data distribution.

Residues of θ are calculated by subtracting the $\hat{\theta}$ from the original θ measurements. Bootstrap method is used to resample the residues with replacement for each θ from the calculated residues. The re-sampled residues are added to the $\hat{\theta}$ to reconstruct new observations for dependent variable θ. The expression for obtaining new observations for dependent variable θ using bootstrap method in Excel is:

$$\theta = \hat{\theta} + \mathrm{INDEX}(\text{Range of residual}, \mathrm{INT}(\mathrm{RAND}() * \text{Row number}) \quad (16)$$

where function INDEX is used to randomly return a calculated residual from a certain array. Range of residual refers to the calculated residues. INT is a function to round a given number, which is randomly produced by RAND() multiplied by row number.

The non-parametric bootstrap method is a special case of Monte Carlo method used for obtaining the distribution of residues of θ which can be representative of the population. The idea behind the bootstrap method is that the calculated residues can be an estimate of the population, so the distribution of the residues can be obtained by drawing many samples with replacement from the calculated residues. For the Monte Carlo method, however, it creates the distribution of residues of θ with a theoretical (i.e., normal) distribution. From this aspect, the bootstrap method is more empirically based and the Monte Carlo method is more theoretically based.

Similar to the Monte Carlo method, bootstrap simulations are performed 2000 times. Distribution type and 95% confidence intervals of α and n will also be determined for fitted datasets with different numbers (from 100 to 2000 with intervals of 100) of simulation. For simplification, only 200 simulations are shown as an example.

Competing interests
The author declares that there are no competing interests associated with this research work.

Authors' contributions
WH analyzed the data and wrote the draft. JX and HC participated in the data analysis. BS designed the study. All authors read and approved the final manuscript.

Authors' information
Wei Hu is a professional research associate at the University of Saskatchewan and specialist for soil hydrology. Jing Xie is a PhD student in University of Saskatchewan who is investigating legume fertilization. Henry Wai Chau is a lecturer in environmental physics in Lincoln University. Bing Cheng Si is a full professor in University of Saskatchewan and specializes in soil physics.

Acknowledgements
The project was funded by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

Author details
[1]Department of Soil Science, University of Saskatchewan, Saskatoon, SK S7N 5A8, Canada. [2]Department of Soil and Physical Science, Lincoln University, PO Box 84Lincoln, Christchurch 7647, New Zealand.

Received: 22 January 2015 Accepted: 11 March 2015
Published online: 24 March 2015

References

Berger RL (2007) Nonstandard operator precedence in Excel. Comput Stat Data An 51:2788–2791
Brown AM (2001) A step-by-step guide to non-linear regression analysis of experimental data using a Microsoft Excel spreadsheet. Comp Meth Prog Bio 65:191–200
Conedera M, Torriani D, Neff C, Ricotta C, Bajocco S, Pezzatti GB (2011) Using Monte Carlo simulations to estimate relative fire ignition danger in a low-to-medium fire-prone region. Forest Ecol Manag 26:2179–2187
Cwiertny DM, Roberts AL (2005) On the nonlinear relationship between k(obs) and reductant mass loading in iron batch systems. Environ Sci Technol 39:8948–8957
Delboy H (1994) A non-linear fitting program in pharmacokinetics with Microsoft® Excel spreadsheet. Int J Biomed Comput 37:1–14
Efron B (1979) Bootstrap method: another look at the Jackknife. Ann Stat 7:1–26
Efron B, Tibshirani R (1993) An introduction to the Bootstrap. Champman & Hall, London, UK
Fox J, Weisberg S (2010) Nonlinear regression and nonlinear least squares in R: An appendix to an R companion to applied regression, second edition. http://socserv.socsci.mcmaster.ca/jfox/Books/Companion/appendix/Appendix-Nonlinear-Regression.pdf.
Fredlund DG, Xing AQ, Huang SY (1994) Predicting the permeability function for unsaturated soils using the soil-water characteristic curve. Can Geotech J 31:533–546
Han XW, Shao MA, Hortaon R (2010) Estimating van Genuchten model parameters of undisturbed soils using an integral method. Pedosphere 20:55–62
Harris DC (1998) Nonlinear least-squares curve fitting with Microsoft Excel Solver. J Chem Educ 75:119–121
Luo B, Maqsood I, Yin YY, Huang GH, Cohen SJ (2003) Adaption to climate change through water trading under uncertainty - An inexact two-stage nonlinear programming approach. J Environ Inform 2:58–68
Singh K, Xie M (2013) Bootstrap: A statistical method. From Rutgers University. http://www.stat.rutgers.edu/home/mxie/rcpapers/bootstrap.pdf.
Smith LH, McCarty PL, Kitanidis PK (1998) Spreadsheet method for evaluation of biochemical reaction rate coefficients and their uncertainties by weighted nonlinear least-squares analysis of the integrated Monod equation. Appl Environ Microbiol 64:2044–2050
Tong L, Chang C, Jin S, Saminathan R (2012) Quantifying uncertainty of emission estimates in National Greenhouse Gas Inventories using bootstrap confidence intervals. Atmos Environ 56:80–87
van Genuchten MT (1980) A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. Soil Sci Soc Am J 44:892–898
Wang QJ, Horton R, Shao MA (2002) Horizontal infiltration method for determining Brooks-Corey model parameters. Soil Sci Soc Am J 66:1733–1739
Wraith JM, Or D (1998) Nonlinear parameter estimation using spreadsheet software. J Nat Resour Life Sci Educ 27:13–19