Improved Models of Particle-Size Distribution: An Illustration of Model Comparison Techniques

G. D. Buchan, K. S. Grewal, and A. B. Robson

ABSTRACT

We investigated a relatively unexplored area of soil science: the fitting of parameterized models to particle-size distribution (a subject more thoroughly explored in sedimentology). Comparative fitting of different models requires the use of statistical indices enabling rational selection of an optimum model, i.e., a model that balances the improvement in fit often achieved by increasing the number of parameters, p, against model simplicity retained by minimizing p. Five models were tested on cumulative mass-size data for 71 texturally diverse New Zealand soils: a one-parameter (p = 1) Jaky model borrowed from geotechnics; the standard lognormal model (p = 2); two modified lognormal models (each with p = 3); and the bimodal lognormal model (p = 3). The Jaky and modified lognormal models have not previously been introduced into the soil science literature. Three statistical comparators were used: the coefficient of determination, R^2; the F statistic; and the C_p statistic of Mallows. The bimodal model and one modified lognormal model (denoted ORL) best fit the data. The bimodal model gave a marginally better fit, but incorporates a sub-clay mode (untestable with the present data), so we adopted the ORL model as the physically best benchmark for comparison of other models. The simple Jaky one-parameter model gave a good fit to data for many of the soils, better than the standard lognormal model for 23 soils. The model comparison methods described have potential utility in other areas of soil science. The C_p statistic is advocated as the best statistic for model selection.

A FREQUENT NEED in soil science is to fit parameterized models to data. Examples include the fitting of adjustable, analytic functions to data for the soil moisture characteristic, hydraulic conductivity function, or PSD. Often several candidate models exist, posing the problem of choice. In general, algorithms for fitting such models minimize an aggregated discrepancy between observed and model-estimated data. A lower bound to this discrepancy is set by experimental errors in the observed data. Often (though not always), increasing p in a model will improve the fit; however, increasing p may sacrifice simplicity and utility of the model, and may simply be an empirical expedient for conforming the model to fit the data. The first test for admitting an additional parameter is to check for its statistical significance. This can be done via a Student's t-test or Wald test (Gallant, 1987). Failure in this test means the additional parameter overparameterizes the model. Also, if the aggregate error produced by the model is less than random experimental error, the model is again overparameterized, though in a different sense. Selection of an optimum model from a group thus requires use of a sensitive discriminating statistic. Here, an optimum model is defined as one selected by balancing the minimization of some objective function (measuring aggregate discrepancy) against minimization of p.

We explored the application of new parametric models for soil PSD. We compared five models, using data for 71 New Zealand soils. Three of these models are, as far as we are aware, new to the soil science literature. Three model comparison techniques were compared: the coefficient of determination (R^2), the F statistic, and the C_p statistic of Mallows (1973).

Modeling of PSD is a poorly researched area in soil science, in strong contrast to sedimentology, geology, and geotechnics, where diverse model forms have been explored, ranging from the Jaky one-parameter model (Jaky, 1944) to the more recent log-hyperbolic (Bagnold and Barndorff-Nielsen, 1980) and log-skew Laplace models (Fieller et al., 1984; Flenley et al., 1987). Recently, Shiozawa and Campbell (1991) proposed a bimodal lognormal model, comparing it with a unimodal lognormal model, using R^2 as a model-selection criterion. The two methods of model comparison proposed (i.e., F and C_p) enable rational selection of an optimum model for PSD, and serve as better discriminators than R^2.

THEORY

Modeling Particle-Size Distribution

Modeling PSD is of interest from two viewpoints: fundamental pedological characterization of the soil, or as a basis for estimation of bulk soil properties, such as the water retention and hydraulic conductivity functions. There are two basic approaches to the representation of PSD: via parametric models of the full distribution (discussed below), or more simply via statistical transformation of limited three-fraction texture data (e.g., Shirazi and Boersma, 1984). Soil PSD is often assumed to be approximately lognormal (Shirazi and Boersma, 1984; Campbell, 1985). It is then represented by two parameters, a geometric mean particle diameter and a geometric standard deviation, i.e., *size mean* and *size spread*, respectively (Buchan, 1989). However, Buchan (1989) tested the applicability of the simple lognormal model. He showed that about one-half of the textures on a textural triangle cannot be represented by a lognormal distribution because they constrain excessive spread of that distribution, with tails extending beyond lower (0.02-μm) or upper (2-mm) size limits. Other two-parameter schemes for summarizing PSD were described in Buchan (1989). There remains a need, however, to investigate alternative, physically based models for soil PSD. The following five models were tested, all except the bimodal model being unimodal:

Jaky One-Parameter Model

Jaky (1944) proposed the following simple model for grain-size distribution in sediments:

$$S = \exp\left\{-\frac{1}{p^2}\left[\ln\left(\frac{d}{d_0}\right)\right]^2\right\} \qquad [1]$$

where S is the cumulative mass of particles with equivalent

G.D. Buchan and K.S. Grewal, Dep. of Soil Science, and A.B. Robson, Centre for Computing and Biometrics, Lincoln Univ., Canterbury, New Zealand. Received 27 Aug. 1991. *Corresponding author.

Published in Soil Sci. Soc. Am. J. 57:901–908 (1993).

Abbreviations: PSD, particle-size distribution; SD, standard deviation; ONL, offset-nonrenormalized lognormal; MLP, Maximum Likelihood Program; MSE, mean square error; SSE, sum of squared errors; RSS, residual sum of squares; RMS, residual mean square error.

diameter < d, p is a particle-size distribution index characterizing the stretching of the curve, and d_0 is the largest diameter, here taken as 2000 μm. The parameter S takes the sigmoid shape of the left-hand half of a Gaussian lognormal curve (see Eq. [2]).

Simple Lognormal Model

The Gauss function, $f(X)$, is represented by the expression

$$f(X) = [1/\sigma(2\pi)^{1/2}] \exp[-(X-\mu)^2/2\sigma^2] \quad [2]$$

where μ is the mean of X, and σ is the standard deviation. For a lognormal distribution, X is replaced by ln$d$. For examining distribution functions, it is more convenient to use the cumulative form of Eq. [2]:

$$F(X) = (1 + \text{erf}[(X-\mu)/\sigma\sqrt{2}])/2 \qquad (X \geq \mu)$$

$$F(X) = (1 - \text{erf}[(X-\mu)/\sigma\sqrt{2}])/2 \qquad (X \leq \mu) \quad [3]$$

where $F(X)$ is cumulative mass, and erf [ ] is the error function, defined as

$$\text{erf}[y] = (2/\sqrt{\pi}) \int_0^y \exp[-t^2]\, dt \quad [4]$$

There are good physical grounds for using the lognormal model for PSD, based on general concepts of comminution mechanisms, and fluid-borne transport processes (Buchan, 1989).

Shiozawa and Campbell Model

Shiozawa and Campbell (1991) recently proposed a bimodal model, with PSD assumed to be a weighted sum of two lognormal distributions: an upper mode and a lower (clay) mode. They represented the cumulative mass fraction $G(X)$ as

$$G(X) = \epsilon F_1(X) + (1-\epsilon)F_2(X) \quad [5]$$

where $F_1(X)$ and $F_2(X)$ are cumulative functions (Eq. [3]) centered in the clay fraction, and in the primary minerals (sand and silt), respectively, and ϵ is the mass fraction in the lower mode. Through lack of data below 1.3 μm, Shiozawa and Campbell (1991) arbitrarily set both μ_1 and σ_1 as constants in the lower mode. They chose $\mu_1 = -1.96$, representing the natural log of the geometric mean diameter (0.141 μm) for the clay fraction, assumed to have a lower size limit of 0.01 μm, and arbitrarily set $\sigma_1 = 1$. They reported that this model gave an improved fit to data for six soils, compared with a unimodal (lognormal) model. As a fitting criterion, they used R^2. However, note the following points: (i) the model imposes a lognormal lower mode, with arbitrary choice of both μ_1 and σ_1; (ii) with $\sigma_1 = 1$, this lower mode is very narrow, concentrated well below 2 μm (± 2 SD limits are 0.02 and 1.04 μm); (iii) the same distribution is assumed for all soils, differing only in its weighting, ϵ; (iv) it presumes that we can quantify the clay size distribution, despite the lack of clay (<2-μm) PSD data. The assumption of a lognormal distribution within the clay fraction is not testable for either their data or the data in this study. While physical arguments support applicability of the lognormal distribution to materials derived from fragmentation processes, there is as yet no corresponding argument to support its applicability to secondary (clay) minerals formed from weathering, dissolution, and recrystallization processes. Shiozawa and Campbell (1991) analyzed only six soils in their study, and so their approach requires further testing on a wider range of soils.

Offset-Renormalized Lognormal Model

Reasoning that PSD may be approximately lognormal across only part of the size range, Shiozawa and Campbell (1990, personal communication) have proposed the following modification of the lognormal model by introducing an "offset" or displacement parameter, ϵ:

$$G(X) = (1-\epsilon)F(X) + \epsilon \quad [6]$$

$$g(X) = (1-\epsilon)f(X) \quad [7]$$

Here, $G(X)$ is the modified cumulative function and $g(X)$ its corresponding mass distribution function. The parameter ϵ has a distinct physical interpretation: it indicates a residual fraction of soil above which the distribution may be approximated as a lognormal function. Inclusion of the factor $(1-\epsilon)$ to renormalize $F(X)$ in Eq. [6] ensures that $G(X)$ in that equation has the asymptotic behavior $G(X) \to 1$ as $X \to \infty$. However, strictly $G(X)$ does not satisfy the zero-limit condition, i.e., $G(X) \to 0$ as diameter $d \to 0$. In fact, $G(X) = \epsilon$ for $d = 0$, implying a finite mass of soil of zero diameter. By contrast, $G(X)$ in Eq. [5] does satisfy the zero-limit condition. However, taking a pragmatic viewpoint, one can justifiably match PSD data only above the lower limit of measurement, here at ~1 μm. Thus, insistence on correct limiting behavior as $d \to 0$ becomes an artificial constraint, unjustified by the absence of data below ~1 μm.

Offset-Nonrenormalized Lognormal Model

For this model, they distribution functions were modified as follows:

$$G(X) = F(X) + c \quad [8]$$

$$g(X) = f(X) \quad [9]$$

Here, c is simply an extra "offset" parameter, adjustable to give a better fit to the data. Equation [3] is normalized so that $F(X) \to 1$ as $X \to \infty$. However, soil fines (<2 mm) may represent a truncated sample of the total size distribution. If a lognormal model is appropriate to the whole soil (including gravel and stones), then it will be better to fit a truncated lognormal distribution to the fines. To achieve this, $F(X)$ was modified as in Eq. [8]. Strictly, Eq. [8] has the asymptotic behavior $G(X) \to (1 + c)$ as $X \to \infty$. However, the nonlinear fitting procedure used (described below) is designed to optimize the match between model and data, and hence should ensure $G(2\text{ mm}) \approx 1.0$. Further, a problem with use of Eq. [3] for soils with large clay fraction is that the lognormal function will tend to be distorted by the matching constraint $F(2\ \mu\text{m}) \approx$ clay fraction. This will force a large tail in the distribution, increasing σ, and losing some of the flexibility in fitting $F(X, \mu, \sigma)$ to the silt–sand mass distribution. In summary, the reasons for introducing this ONL model are: (i) it allows a truncated lognormal model, in case the soil PSD is approximately lognormal across the whole size range, extending beyond d = 2 mm; and (ii) simultaneously, the effect of the offset c at d = 2 μm will be to absorb some of the clay percentage, which may be important for soils high in clay.

Estimation of Model Parameters

The method chosen to estimate parameters depends on whether the model is linear or nonlinear in its parameters, and on the statistical assumptions made concerning measurement errors. Models that are linear (or linearizable) in their parameters are amenable to direct solution. Nonlinear models require iterative solution, by using a search algorithm to determine the minimum of an objective function, and with starting values for the

Table 1a. Coefficient of determination (R^2) values for Set 1 (Canterbury) soils from the five models (p = number of parameters). ORL and ONL are the offset renormalized and offset nonrenormalized models. The Wakanui soils spans four cultivation treatments: permanent grassland (PP), continuous arable cultivation (AC), short-term pasture (SP), and minimum tilled (MT).

Soil	Depth	Jaky $p = 1$	Lognormal $p = 2$	ONL $p = 3$	ORL $p = 3$	Shiozawa and Campbell $p = 3$
	cm					
Temuka	3–15	0.9742	0.9801	0.9841	0.9980	0.9980
	40–55	0.9761	0.9604	0.9761	0.9860	0.9880
	40–55	0.9274	0.9880	0.9900	1.0000	1.0000
	89–92	0.9293	0.9821	0.9900	0.9940	0.9960
	131–136	0.8949	0.9940	0.9960	1.0000	1.0000
Templeton	0–5	0.9860	0.9841	0.9920	0.9980	0.9980
	5–10	0.9841	0.9821	0.9900	0.9980	0.9980
	10–15	0.9860	0.9841	0.9900	0.9900	0.9980
	15–20	0.9841	0.9821	0.9880	0.9980	0.9980
	20–25	0.9860	0.9841	0.9920	0.9980	0.9980
Wakanui (PP)	0–5	0.9863	0.9821	0.9880	0.9980	0.9980
	5–10	0.9722	0.9821	0.9980	0.9980	0.9980
	10–15	0.9722	0.9821	0.9880	0.9980	0.9980
	15–20	0.9722	0.9841	0.9980	0.9980	0.9980
	20–25	0.9663	0.9821	0.9880	0.9980	1.0000
Wakanui (AC)	0–5	0.9312	0.9940	0.9960	1.0000	1.0000
	5–10	0.9351	0.9920	0.9920	1.0000	1.0000
	10–15	0.9351	0.9920	0.9940	0.9980	0.9980
	15–20	0.9351	0.9920	0.9920	0.9980	0.9980
	20–25	0.9101	0.9960	0.9960	1.0000	1.0000
Wakanui (SP)	0–5	0.9761	0.9801	0.9880	0.9980	0.9980
	5–10	0.9781	0.9821	0.9880	0.9980	1.0000
	10–15	0.9781	0.9801	0.9880	0.9980	0.9980
	15–20	0.9781	0.9841	0.9900	0.9980	1.0000
	20–25	0.9781	0.9860	0.9920	0.9980	0.9980
Wakanui (MT)	0–5	0.9604	0.9841	0.9880	1.0000	1.0000
	10–15	0.9624	0.9841	0.9880	1.0000	1.0000
	15–20	0.9624	0.9821	0.9860	1.0000	1.0000
	20–25	0.9604	0.9860	0.9880	1.0000	1.0000
Cookson	5–10	0.9880	0.9920	0.9940	1.0000	1.0000
	17–22	0.9900	0.9920	0.9960	0.9980	0.9980
	35–40	0.9860	0.9761	0.9841	1.0000	1.0000
	55–60	0.9920	0.9761	0.9880	0.9960	0.9960
	70–75	0.9880	0.9722	0.9821	0.9980	0.9980
Timpendean	5–10	0.9900	0.9742	0.9860	0.9960	0.9980
	12–17	0.9920	0.9761	0.9880	0.9980	0.9980
	55–60	0.9683	0.9487	0.9624	0.9960	0.9980
	90–95	0.9683	0.9487	0.9604	0.9821	0.9960

parameters. As all the above five models are nonlinear, a nonlinear estimation technique was used. The best-fit parameters were determined using the MLP (Ross, 1980).

Model Comparison Techniques: Available Methods

Historically, several approaches have been reported for comparison of and selection from different models. The simplest approach may be to find the "best" model, i.e., that which minimizes some measure of aggregate discrepancy, such as mean square error or raw R^2 (see below). However, a better approach is to seek an optimum model, by statistically measuring the significance of extra parameters. By omitting parameters whose contribution to the fit is not significant, the model selected will be optimum in the sense that it balances minimization of discrepancy against removal of excess parameters. First, the statistical significance of an additional parameter may be determined using either Student's t-test or a Wald test (Gallant, 1987). The following model comparison criteria are available:

Mean Square Error

The lower the MSE, the better the model represents the data. A lower bound to the MSE is set by experimental (random) errors in the data: a model that reduces the MSE below this lower bound is overparameterized.

Table 1b. Coefficient of determination (R^2) values for Set 2 (Waikato) soils from the five models (p = number of parameters). ORL and ONL are the offset renormalized and offset nonrenormalized models.

Soil	Depth	Jaky $p = 1$	Lognormal $p = 2$	ONL $p = 3$	ORL $p = 3$	Shiozawa and Campbell $p = 3$
	cm					
Horotiu	0–6	0.9980	0.9940	0.9980	0.9960	0.9960
	6–17	0.9960	0.9940	0.9980	0.9960	0.9960
	17–31	0.9960	0.9920	0.9940	0.9920	0.9920
	31–55	0.9624	0.9428	0.9841	0.9920	0.9920
	55–73	0.9801	0.9624	0.9880	0.9900	0.9920
	73–91	0.9370	0.9197	0.9761	0.9900	0.9900
	91–107	0.9801	0.9663	0.9920	0.9880	0.9880
	107–130	0.9960	0.9940	0.9980	0.9980	0.9980
Te Kowhai	0–9	0.9506	0.9960	0.9960	1.0000	1.0000
	9–22	0.9565	0.9900	0.9920	1.0000	1.0000
	32–39	0.9370	0.9960	0.9960	1.0000	1.0000
	80–93	0.9565	0.9584	0.9702	0.9920	0.9940
	97–100	0.9801	0.9702	0.9880	0.9821	0.9821
Hamilton	0–9	0.9860	0.9860	0.9920	0.9980	0.9980
	9–19	0.9900	0.9920	0.9960	0.9980	0.9980
	19–29	0.9880	0.9880	0.9920	1.0000	1.0000
	29–46	0.9841	0.9920	0.9940	1.0000	1.0000
	46–73	0.9643	0.9526	0.9604	0.9940	0.9920
	73–88	0.9900	0.9920	0.9960	0.9980	1.0000
	88–97	0.8372	0.9624	0.9663	0.9643	0.9643
	97–120	0.9409	0.9900	0.8354	0.9900	0.9920
Otorhanga	0–8	0.9742	0.9960	0.9960	0.9980	0.9980
	8–23	0.9565	1.0000	1.0000	1.0000	1.0000
	23–47	0.9781	0.9940	0.9960	0.9960	0.9960
	47–62	0.9860	0.9860	0.9900	0.9980	0.9980
	62–81	0.9683	0.9841	0.9860	0.9980	0.9980
	81–97	0.9761	0.9960	0.9960	0.9980	0.9980
	97–114	0.9841	0.9920	0.9960	0.9980	0.9980
Netherton	0–10	0.9565	0.9880	0.9880	0.9900	0.9900
	10–20	0.9467	0.9821	0.9821	0.9860	0.9841
	20–38	0.8482	0.9604	0.9624	0.9604	0.9624
	38–49	0.7832	0.9940	0.9940	0.9980	0.9980
	48–87	0.8732	0.9860	0.9880	0.9880	0.9980

Coefficient of Determination

A larger R^2 implies a better fit to the observed data, and so a model with larger R^2 is preferred over one with smaller R^2. When, however, comparison of models with differing numbers of parameters is desired, an adjusted R^2 is sometimes used, although this statistic is not entirely free from defect (Snedecor and Cochran, 1989).

F-Statistic

Green and Caroll (1978) suggested the F statistic, defined as

$$F = [(\mathrm{SSE_r} - \mathrm{SSE_f})/\mathrm{SSE_f}][d_f/(d_r - d_f)] \quad [10]$$

Here $\mathrm{SSE_f}$ and $\mathrm{SSE_r}$ are, respectively, the sums of squared errors of a full model, and a 'restricted' model with fewer parameters; d_f and d_r are the respective degrees of freedom. The SSE is calculated as

$$\mathrm{SSE}_{(j)} = \sum_{i=1}^{n} (Y_e - Y_o)^2 \quad [11]$$

where Y_o and Y_e are observed and estimated data values, j = r (restricted) or f (full) model, and n is the number of measured data points.

If the F value for a restricted model does not exceed the F value at the 0.05 significance level (obtainable from standard tables, e.g., Snedecor and Cochran, 1989), the restricted model

Table 2a. Statistical analysis of the particle-size distribution models for Set 1 (Canterbury) soils using F statistic and Mallow's C_p test. ORL and ONL are the offset renormalized and offset nonrenormalized models. The Wakanui soil spans four cultivation treatments: permanent grassland (PP), continuous arable cultivation (AC), short-term pasture (SP) and minimum tilled (MT).

Soil	Depth, cm	Jaky		Lognormal		ONL	ORL	Shiozawa and Campbell
		F	C_p	F	C_p	C_p	C_p	C_p
Temuka	3–15	77.60**	154.20	74.92**	75.92	55.46	3.00	2.13†
	40–55	3.75	6.49	13.63**	14.63	8.11	3.00	2.58
	40–55	865.07**	1729.14	194.71**	195.71	162.79	3.00	2.28
	88–92	65.78**	130.57	25.74**	26.74	17.28	3.00	2.00
	131–136	1303.96**	2606.91	93.14**	94.14	79.82	3.00	2.63
Templeton	0–5	49.07**	97.13	68.36**	69.36	36.35	3.00	2.21
	5–10	44.65**	88.29	62.67**	63.67	36.33	3.00	2.20
	10–15	43.17**	85.34	60.61**	61.61	33.40	3.00	2.33
	15–20	26.49**	51.98	42.44**	43.44	23.89	3.00	2.37
	20–25	55.65**	110.30	71.55**	72.55	39.65	3.00	2.23
Wakanui	0–5	98.74**	196.49	65.83**	66.83	45.87	3.00	2.28
(PP)	5–10	102.42**	203.84	83.21**	84.21	56.37	3.00	2.19
	10–15	62.98**	124.97	47.85**	48.85	32.84	3.00	2.36
	15–20	109.03**	217.06	83.84**	84.84	60.51	3.00	2.32
	20–25	167.12**	333.24	121.15**	122.15	84.79	3.00	2.03
Wakanui	0–5	781.06**	1561.13	79.66**	80.66	69.79	3.00	2.47
(AC)	5–10	367.46**	733.92	58.63**	59.63	51.35	3.00	2.51
	10–15	287.13**	573.26	35.87**	36.87	31.63	3.00	2.62
	15–20	261.85**	522.69	41.28**	42.28	36.19	3.00	2.60
	20–25	1778.73**	3556.45	121.50**	122.50	104.23	3.00	2.36
Wakanui	0–5	86.93**	172.85	93.31**	94.31	58.03	3.00	2.09
(SP)	5–10	111.41**	221.83	120.48**	121.48	74.35	3.00	2.07
	10–15	79.75**	158.51	91.74**	92.74	56.88	3.00	2.14
	15–20	105.04**	209.09	96.28**	97.27	58.16	3.00	2.17
	20–25	95.65**	190.30	76.37**	77.37	46.09	3.00	2.24
Wakanui	0–5	583.48**	1165.96	286.88**	287.88	218.57	3.00	1.55
(MT)	10–15	356.71**	712.42	189.42**	190.42	144.86	3.00	1.84
	15–20	237.57**	474.14	131.80**	132.80	101.61	3.00	2.00
	20–25	353.45**	705.89	157.55**	158.55	120.86	3.00	1.93
Cookson	5–10	93.01**	185.01	66.37**	67.37	35.86	3.00	2.30
	17–22	51.03**	101.05	39.53**	40.53	17.76	3.00	2.12
	35–40	340.77**	680.53	839.64**	840.64	520.16	3.00	−0.86
	55–60	3.42	5.85	27.79**	28.79	12.46	3.00	2.16
	70–75	19.01**	37.01	74.40**	75.40	42.58	3.00	2.00
Timpendean	5–10	9.92**	18.84	49.48**	50.48	25.96	3.00	1.84
	12–17	8.16**	15.33	52.24**	53.24	24.40	3.00	1.71
	55–60	28.81**	56.62	87.50**	88.50	62.01	3.00	1.03
	90–95	4.21	7.43	14.48**	15.48	11.11	3.00	−2.26

*, ** Significant at the 0.05 and 0.01 probability levels, respectively. Also indicates that model should be rejected in favor of a higher parameter model.

† Using C_p as a discriminator, the best model is that with the lowest C_p value (underlined for each soil in the table).

can be used. In the opposite case, the full or reference model should be used. Note that F does not permit intercomparison of equiparameter models fitted to the same data, i.e., models with the same number of parameters. In that case $d_f = d_r$ and F is undefined. The F-statistic has been used recently by Vereecken et al. (1989), who compared models fitted to soil moisture characteristic data.

There are two important considerations concerning the F statistic. First, for nonlinear models, Eq. [10] is approximate because the numerator and denominator no longer contain independent χ^2 distributions (Beck and Arnold, 1977). Still it is a useful criterion to judge the necessity of model parameters, through the increase of SSE. Secondly, Eq. [10] can be used only when the error term is normally independently distributed with zero mean and constant variance. It is desirable to check these conditions when fitting the model to the data. Violation of these conditions is most frequently checked through visual inspection of the residuals plotted against the independent variable. Instead of this graphical inspection, one can also examine the so-called unit normal deviate of the residuals in an overall plot. Using this latter test, no indication was found here that the conditions regarding the error term were violated.

Mallows' C_p Test

The statistic C_p given by Mallows (1973) has been recommended by Daniel and Wood (1971) as a simple criterion for comparing the relative goodness (or badness) of fit of different models. Snedecor and Cochran (1989) advocated the use of C_p to overcome the inherent defects in raw R^2 or adjusted R^2 mentioned above. The C_p statistic measures the total squared error in Y at all N data points, i.e., the sum of the squared biases plus the squared random errors. It is a simple function of the residual sum of squares from both the models being compared:

$$C_p = \frac{\mathrm{RSS}_p}{\mathrm{RMS}_f} - (N - 2p) \qquad [12]$$

where RSS_p is the residual sum of squares (total squared error, i.e., bias plus random) from a p-parameter model. If the p-parameter model contains no bias, then the RSS will reflect only random error. The RMS_f is the residual mean square error from a full or ideal model and $\mathrm{RMS}_f = \mathrm{RSS}_f/(N - p)$, where RSS_f is the residual sum of squares from the full model, and N is the number of data points. For the artificial comparison of the full or reference model with itself, the C_p value equals the number of parameters, p. In contrast to the F statistic, C_p allows comparison of equiparameter models, a key feature that we exploit below. However, our use of C_p here is different from the standard use: we use it as a relative measure between nonideal models, rather than for comparison against some "ideal" model, because an "ideal" model does not exist for soil PSD.

Table 2b. Statistical analysis of the particle-size distribution models for Set 2 (Waikato) soils using F statistic and Mallow's C_p test. ORL and ONL are the offset renormalized and offset nonrenormalized models.

Soil	Depth, cm	Jaky		Lognormal		ONL	ORL	Shiozawa and Campbell
		F	C_p	F	C_p	C_p	C_p	C_p
Horotiu	0–6	−1.81**	−4.61†	4.74	5.74	−1.16	3.00	2.73
	6–17	2.96	4.91	6.64*	7.64	0.69	3.00	2.71
	17–31	−0.70	−2.40	1.01	2.01	0.27	3.00	2.90
	31–35	16.11**	31.21	43.11**	44.11	−9.27	3.00	2.13
	55–73	5.62*	10.24	26.02**	27.02	−5.21	3.00	2.49
	73–91	24.25**	47.50	52.60**	53.60	11.73	3.00	1.91
	91–107	6.63*	12.27	16.12**	17.12	0.10	3.00	2.62
	107–130	10.28**	19.55	32.00**	33.00	1.66	3.00	2.72
Te Kowhai	0–9	485.45**	969.90	36.12**	37.12	34.25	3.00	3.35
	9–22	771.95**	1542.90	217.97**	218.97	168.31	3.00	1.41
	32–39	969.98**	1938.96	84.58**	85.58	62.73	3.00	2.36
	80–93	29.19**	57.37	37.93**	38.93	27.91	3.00	2.49
	97–100	1.65	2.29	4.41	5.41	−0.01	3.00	2.84
Hamilton	0–9	76.05**	151.09	74.33**	75.33	44.99	3.00	1.81
	9–19	56.84**	112.67	43.33**	44.33	20.39	3.00	2.05
	19–29	110.10**	219.19	107.79**	108.79	63.90	3.00	2.57
	29–46	186.33**	371.65	108.08**	109.08	69.90	3.00	2.04
	46–73	26.70**	52.40	55.21**	56.21	44.29	3.00	4.59
	73–88	35.41**	69.83	24.55**	25.55	7.48	3.00	−0.72
	88–97	16.16**	31.32	0.79	1.79	2.21	3.00	2.66
	97–120	31.21**	61.43	0.74	1.74	131.92	3.00	1.78
Otorohanga	0–8	56.47**	111.94	0.40	1.40	2.52	3.00	3.08
	8–23	388.61**	776.23	0.36	1.36	2.81	3.00	3.04
	23–47	39.10**	77.21	4.01	5.01	2.64	3.00	3.04
	47–62	71.78**	142.56	70.76**	71.76	40.98	3.00	1.23
	62–81	133.79**	266.59	78.66**	79.66	64.47	3.00	2.93
	81–97	94.95**	188.89	16.04**	17.04	9.30	3.00	2.60
	97–114	57.95**	114.89	22.66**	23.66	11.80	3.00	2.17
Netherton	0–10	21.72**	42.45	1.63	2.63	4.60	3.00	3.91
	10–20	15.56**	30.13	2.12	3.12	5.10	3.00	4.44
	20–38	14.28**	27.56	0.24	1.24	2.61	3.00	3.03
	38–49	444.74**	888.47	11.03*	12.03	12.66	3.00	2.79
	48–87	672.19**	1343.37	71.75**	72.75	65.68	3.00	2.17

*, ** Significant at the 0.05 and 0.01 probability levels, respectively. Also, indicates that model should be rejected in favor of a higher parameter model.

† Using C_p as a discriminator, the best model is that with the lowest C_p value (underlined for each soil in the table).

MATERIALS AND METHODS

Two independent data sets from two regions of New Zealand were used. Set 1 contains PSD data for 40 samples from five different texturally layered soils, collected from depths in the range 0 to 136 cm (Table 1a) from the Canterbury region in the South Island. The first three soils are representative of dominant texture types (silt and fine sandy loams) in the Canterbury Plains. The Cookson and Timpendean soils, from the North Canterbury hill country, widen the range to include soils higher in clay. Set 2 (Table 1b), containing 39 samples from the Waikato area in the North Island, was selected from the Soil Water Assessment and Measurement Programme (SWAMP) completed by staff of the New Zealand Soil Bureau (Joe and Watt, 1986). This set covers a wide range of textures, developed under different parent materials, physiographic positions, and climates.

Detailed size analysis of Set 1 (Canterbury) samples was carried out at Lincoln University. Pretreatments and dispersion were identical to the methods used by DSIR Soil Bureau, Wellington, New Zealand (similar to Day, 1965; Thomas, 1973). Hydrogen peroxide (30%) was used for organic matter removal. Both chemical and physical methods of dispersion were applied. Sodium hexametaphosphate (4%) was added to the sample as a dispersant. The sample was ultrasonicated for 5 min using a 100-W probe, then washed through a 63-μm-diam. sieve. Size analysis of the >63-μm fraction was done by dry sieving at 1000-, 500-, 250-, 1250-, and 63-μm equivalent diameters. Material <63 μm was analyzed first by sedigraph, and then by pipette measurements at 20-, 10-, 6-, and 2-μm equivalent diameters. The analysis below is based on the pipette data, as the sedigraph was found to systematically overestimate mass fractions. For the Timpendean subsoil samples (depths 35–95 cm), the suspensions showed thixotropic behavior. Thus for these samples the suspension was diluted twofold, which checked the gel formation during sampling. Particle-size analysis of Set 2 (Waikato) soils was carried out by staff of the New Zealand Soil Bureau, using the sedigraph technique, and then adjusting the data points to their equivalent pipette readings using a statistical algorithm based on a separate comparison of sedigraph and pipette methods. For full details of Set 2, see Joe and Watt (1986).

Model fitting was done using the MLP and values of R^2, F, and C_p were computed using Minitab (Minitab, 1989).

RESULTS AND DISCUSSION

The five models described above were fitted to data for all 79 Set 1 and Set 2 soils. The MLP algorithm converged for all soils except for six in Set 2 and two in Set 1, so these eight are omitted. The significance of the estimated parameters for each model and depth was tested using Student's t-test. Each parameter was found to be significant at the 0.1% level, hence all five models are compared in Tables 1 and 2. Values of R^2, for regres-

sion of predicted on observed values of cumulative mass, ranged from 0.7832 to 1.0000. See Tables 1a and 1b. For the majority of the soils (47 of the 71), the lowest values of R^2 were, as expected, obtained with the Jaky one-parameter model. For the other 24 soils, however, the Jaky model gave an R^2 value higher for 23 soils than that of the two-parameter lognormal model, and for 10 soils higher than that for the three-parameter ONL model. For all except seven of the soils (in Set 2, Table 1b), the R^2 value is higher for the ORL model than for the Jaky, lognormal, or ONL models. An illustration of the comparative fit of different models is shown in Fig. 1 to 4 for four different Set 2 soils, representing best fits for the Jaky, lognormal, ONL, and ORL models, respectively. Note that the ORL model gave a very good fit for all four soils. The Shiozawa and Campbell (1991) bimodal model yielded the same R^2 value as the ORL model for 55 of the 71 soils. The R^2 values are slightly improved for 14 soils by the bimodal model, while for two soils they are slightly decreased.

Shiozawa and Campbell (1991) used R^2 for comparison of their bimodal model with the lognormal model. However, in most cases in Tables 1a and 1b, $R^2 > 0.90$, and for the majority of cases with the two- and three-parameter models, R^2 exceeded 0.98. Further, R^2 usually (but not always) increased as the number of model parameters (p) increased. One then needs to identify whether a model with larger p significantly improves the fit to the data, or if the increase in R^2 is due merely to the addition of parameters. In the latter case, the model is overparameterized. Therefore, R^2 is not a powerful tool for relative discrimination of nonlinear models, although it does measure the absolute amount of variability accounted for by the model.

Thus we require better relative indicators of model performance, capable of discriminating statistically significant differences between models. The use of either the C_p or F statistics solves this problem. Tables 2a and 2b summarize relative model performance using both Cp and F, calculated with the ORL model ($p = 3$) as the reference model. Compared with the ORL model, the Cp values for Jaky and lognormal models are higher in most cases, indicating that the ORL model significantly improved representation of the data. Similarly, for most cases, F for the Jaky and lognormal models is higher than the F value corresponding to a 0.05 significance level, again indicating that the three-parameter model should be used. However, there are 9.9 and 15.5% of the 71 cases where the Jaky or lognormal model F values, respectively, are not significant. For those soils, these models are better than the ORL model. The C_p and F tests agree for most (83.1%) soils when comparing the Jaky ($p = 1$) and lognormal ($p = 2$) models, but disagree for 16.9% of the soils. This is to be expected, since both tests measure similar, but not identical, quantities: both utilize SSE, but compounded in different ways.

The ONL ($p = 3$) and Shiozawa and Campbell ($p = 3$) models were compared with other models using only C_p, as these models could not be compared with the ORL model using F, which cannot be calculated for equiparameter models. Tables 2a and 2b show that, in 58 (82%) of the 71 samples, the ORL model has the lowest C_p value compared with the Jaky, lognormal, and ONL models and therefore, in a relative sense, best describes

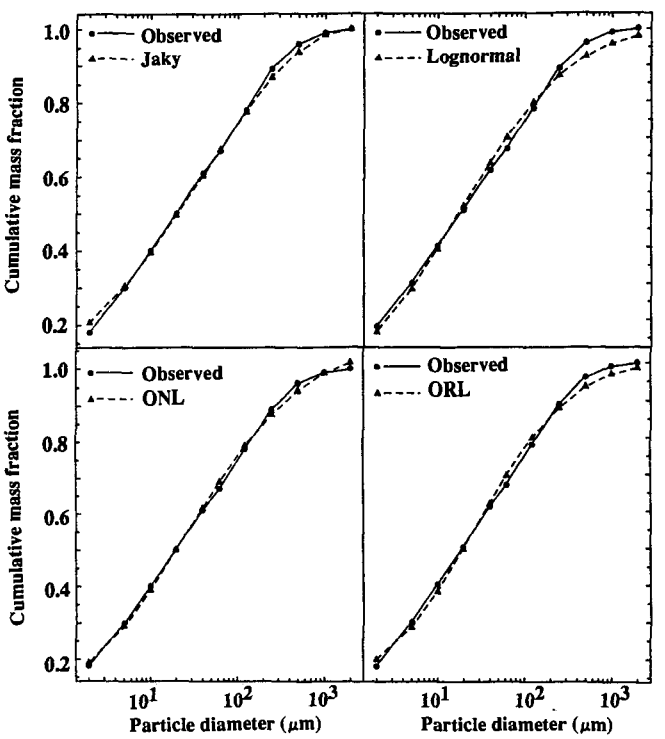

Fig. 1. Horotiu soil, 0–6 cm, one of the 18 soils for which the Jaky one-parameter model (Eq. [1] gives a better fit to particle-size distribution data than the standard lognormal model. ORL and ONL are the offset renormalized (Eq. [6]) and offset nonrenormalized (Eq. [8]) lognormal models.

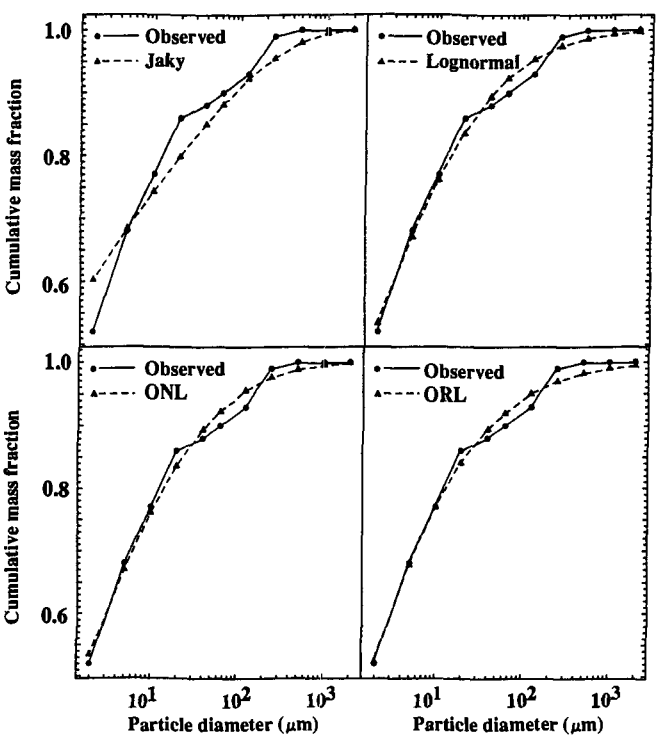

Fig. 2. Netherton soil, 0–10 cm, one of the six soils for which the two-parameter simple lognormal model (Eq. [3]) gives the best fit to particle-size distribution data. ORL and ONL are the offset renormalized (Eq. [6]) and offset nonrenormalized (Eq. [8]) lognormal models.

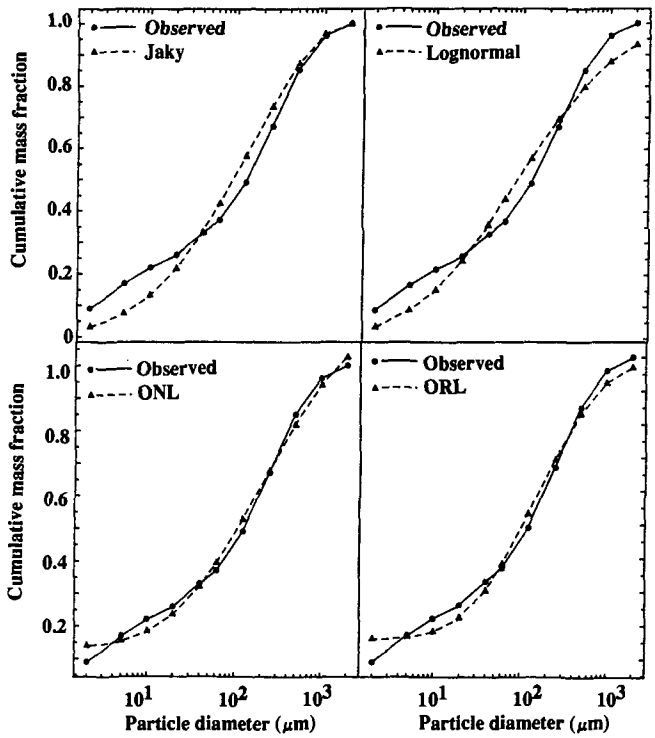

Fig. 3. Horotiu soil, 91–107 cm, one of the five soils for which the ONL model (Eq. [8]) gives the best fit to particle-size distribution data. ORL and ONL are the offset renormalized (Eq. [6]) and offset nonrenormalized (Eq. [8]) lognormal models.

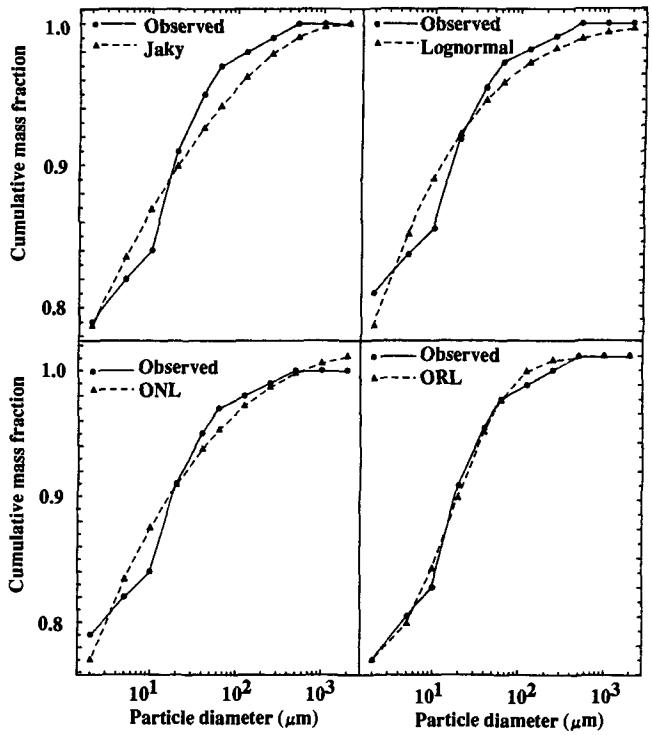

Fig. 4. Hamilton soil, 46–73 cm, one of the 58 soils for which the ORL model (Eq. [6]) gives the best fit to particle-size distribution data. ORL and ONL are the offset renormalized (Eq. [6]) and offset nonrenormalized (Eq. [8]) lognormal models.

the distribution of most of the soils studied. The ONL model performed better (as measured by lower C_p) than the simple lognormal model in 63 of the 71 cases, and best of all models in five cases. This vindicates the testing of this model: the PSD of some soils may approximate a lognormal distribution extending beyond 2 mm, but truncated by the 2-mm cutoff. The Shiozawa and Campbell (1991) bimodal model performed slightly better than the ORL model for the majority of the soils, as measured by C_p. However, the use of a narrow sub-clay mode, imposing structure in the PSD curve well below the limit of available data at ≈ 1 μm, makes this model unacceptable as a reference model.

CONCLUSIONS

We compared five models for soil PSD: the Jaky (1944) one-parameter model (the sigmoid half of a Gaussian distribution); the simple lognormal model; two new adjusted lognormal models, an ORL model, and an ONL model; and a bimodal lognormal model. The results indicate that all five models account for >90% of the variance (R^2) in the PSD of most soils. However, raw R^2 is a poor measure of relative model fit, and should not in general be used for model selection. A more valid comparison is achieved with F and C_p statistics; however, F cannot be used to compare equiparameter models. Thus, we advocate the use of C_p as the best model selection criterion: when compared with adjusted R^2, it not only represents an improvement over that statistic, but also tends to choose the model with smaller p (Snedecor and Cochran, 1989). The C_p statistic thus conforms well with Occam's razor, in that it chooses the model with fewer parameters.

Based on the R^2 and C_p tests, the ORL model performed best of the first four models for the majority (82%) of the soils studied. For 62 (87%) of the 71 soils, the model ranking obtained with R^2 agreed with the ranking obtained with C_p. For the remaining 13% of the soils, however, C_p is an essential discriminating statistic. The simple Jaky one-parameter model gives a good fit to PSD data for many of the soils, better than the simple lognormal model for 18 of the soils. Other more complex models used in sedimentology and geotechnics (e.g., log-hyperbolic and log-skew Laplace) deserve further testing for soils.

ACKNOWLEDGMENTS

We are grateful to Mr. Rob McPherson for construction of a pipette facility, to staff of DSIR Land Resources for access to their SWAMP database, and to the New Zealand government for Dr. Grewal's Commonwealth Scholarship.

REFERENCES

Bagnold, R.A., and O. Barndorff-Nielsen. 1980. The pattern of natural size distributions. Sedimentology 27:199–207.
Beck, V.J., and K.J. Arnold. 1977. Parameter estimation in engineering and science. John Wiley & Sons, New York.
Buchan, G.D. 1989. Applicability of the simple lognormal model to particle size distribution in soils. Soil Sci. 147:155–161.
Campbell, G.S. 1985. Soil physics with Basic: Transport models for soil–plant systems. Elsevier, Amsterdam.
Daniel, C., and F.S. Wood. 1971. Fitting equations to data. Wiley-Interscience, New York.
Day, P.R. 1965. Particle fractionation and particle-size analysis.

p. 545–567. *In* C.A. Black (ed.) Methods of soil analysis. Part 1. Agron. Monogr. 9. ASA, Madison, WI.

Fieller, N.R.J., D.D. Gilbertson, and W. Olbricht. 1984. A new method for environmental analysis of particle size distribution data from shoreline sediments. Nature (London) 311:648–651.

Flenley, E.C., N.R.J. Fieller, and D.D. Gilbertson. 1987. The statistical analysis of 'mixed' grain size distributions from aeolian sands in the Libyan Pre-Desert using log skew Laplace models. p. 271–280. *In* L. Frostick and I. Reid (ed.) Desert sediments: Ancient and modern. Geol. Soc. Spec. Publ. no. 35. Geol. Soc. London.

Gallant, A.R. 1987. Nonlinear statistical models. John Wiley & Sons, New York.

Green, P.E., and J.D. Caroll. 1978. Analyzing multivariate data. John Wiley & Sons, New York.

Jaky, J. 1944. Soil mechanics. Egyetemi Nyomda, Budapest. (In Hungarian.) *Cited in* Probabilistic solutions in geotechnics. Dev. Geotech. Eng. 46:157.

Joe, E.N., and J.P. Watt. 1986. Soil water characterisation studies of 6 soils in the Waikato district, New Zealand. SWAMP data sheets 1984 (1–6). N.Z. Soil Bur., Lower Hutt.

Mallows, C.L. 1973. Some comments on C_p. Technometrics 15:661–675.

Minitab. 1989. Minitab reference manual. Version 7 ed. Minitab Inc., State College, PA.

Ross, G.J.S. 1980. MLP maximum likelihood program. Lawes Agricultural Trust, Rothamstead Exp. Stn., Harpenden, England.

Shiozawa, S., and G.S. Campbell. 1991. On the calculation of mean particle diameter and standard deviation from sand, silt and clay fractions. Soil Sci.(in press).

Shirazi, M.A., and L. Boersma. 1984. A unifying quantitative analysis of soil texture. Soil Sci. Soc. Am. J. 48:142–147.

Snedecor, G.W., and W.G. Cochran. 1989. Statistical methods. Iowa State Univ. Press, Ames, IA.

Thomas, R.F. 1973. E4.B. Determination of particle-size distribution for fine-grained soils. p. E4.5–E4.9. *In* N.Z. Soil Bur. Sci. Rep. 10E: Test methods for soil engineering. Sci. Rep. 10E. N.Z. Soil Bur., Lower Hutt.

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