TITLE PAGE

Full title: ESTIMATING THE VERY FINE SAND FRACTION FOR CALCULATING THE SOIL ERODIBILITY K-FACTOR

Short title: ESTIMATING VERY FINE SAND FRACTION FOR USLE K-FACTOR

Author 1: Eva Corral-Pazos-de-Provens. Departamento de Ciencias Agroforestales. Universidad de Huelva. E-21819 Palos de la Frontera. Huelva, Spain. E-mail: eva.corral@dcaf.uhu.es

Author 2: Juan M. Domingo-Santos. Departamento de Ciencias Agroforestales. Universidad de Huelva. E-21819 Palos de la Frontera. Huelva, Spain. Email: juan.domingo@uhu.es. Telephone: +34654076402 (CORRESPONDING AUTHOR)

Author 3: Ígor Rapp-Arrarás. Departamento de Ciencias Agroforestales. Universidad de Huelva. E-21819 Palos de la Frontera. Huelva, Spain. E-mail: igor@uhu.es

Key words: erosion modelling, RUSLE, textural triangle, piecewise quantile regression, prediction intervals

ESTIMATING THE VERY FINE SAND FRACTION FOR CALCULATING THE SOIL ERODIBILITY *K*-FACTOR

Eva Corral-Pazos-de-Provens, Juan M. Domingo-Santos, and Ígor Rapp-Arrarás

Departamento de Ciencias Agroforestales, Universidad de Huelva, 21819 Palos de la Frontera, Huelva, Spain

ABSTRACT

The *K*-factor of the Universal Soil Loss Equation (USLE) is a core component in many erosion models, as a measure of soil erodibility. It can be estimated by a nomograph, where the summed fractions of silt and very fine sand (VFS) are basic inputs. Frequently, only the three broad particle-size classes of sand, silt, and clay are measured in laboratories, thus the VFS fraction must be estimated. Three models are currently available for this estimation, namely i) the RUSLE2 formula, ii) the European Soil Data Centre method, and iii) the Shirazi-Boersma theory, all three use just the sand fraction as explanatory variable. Nevertheless, their accuracy has never been assessed and this is the main purpose of this study. The data used to test the VFS estimation methods were drawn from the National Cooperative Soil Survey Soil Characterization Database, incorporating data from more than 300,000 soil horizon samples. The test results show a poor performance of the models, all of which were found to be unsuitable for 31.1% of the textural triangle, accounting for 32.3% of the soil samples. Moreover, it is demonstrated that any conceivable model based solely on the broad particle-size classes would suffer from a high degree of uncertainty. Consequently, the number of explanatory variables should be increased in order to improve the performance of models. An alternative prediction chart is provided for the first approximation of *K*-factor, based on the textural triangle.

1 INTRODUCTION

The relationship between the texture of a soil and its erodibility has been well-known for many decades (see, for example, Middleton, 1930; Bouyoucos, 1935), such that the different particle-size classes are key parameters in any soil erosion prediction model (Soil Survey Staff, 2011, Chapter 3). Among the various indexes proposed for characterising soil erodibility, Wischmeier & Smith's (1965, 1978) *K*-factor for the Universal Soil Loss Equation (USLE) is particularly notable. In fact, as Auerswald *et al*. (2014) demonstrate, the USLE and its successors, the Revised Universal Soil Loss Equation (RUSLE) version 1 (RUSLE1) (Renard *et al*., 1997) and version 2 (RUSLE2) (USDA Agricultural Research Service, 2013), are by far the most

widely used models for soil erosion predictions. A recent example of its application can be found in the development of the map of soil loss by water erosion in the European Union by the European Soil Data Centre (ESDAC) (Panagos *et al*., 2015).

Moreover, Auerswald *et al.* (2014) note that the *K*-factor has also been included in a number of USLE modifications and extensions, such as MUSLE (Williams, 1975), USLE-M (Kinnell & Risse, 1998), and dUSLE (Flacke *et al*., 1990); and has also been integrated into the more complex models using USLE/RUSLE technology to estimate erosion, such as EPIC (Williams *et al*., 1983), SWAT (Arnold *et al*., 1998), AGNPS (Cronshey & Theurer, 1998), and Watem/Sedem (Van Rompaey *et al*., 2001).

The USLE, RUSLE1, and RUSLE2 models were developed by the United States Department of Agriculture (USDA) Agricultural Research Service with strong support from USDA Natural Resources Conservation Service (NRCS) and other agencies, as well as various universities and non-profit organizations. Though developed initially based on over 10,000 plot-years of data collected within the United States (US), this family of models has been used extensively in other parts of the world for research and conservation management (Renard *et al.*, 2011). For US soils, the value of the *K*-factor can be taken from the NRCS RUSLE2 Database or calculated using the nomograph developed by Wischmeier *et al*. (1971) and, if needed, its subsequent algebraic approximations and extensions (Wischmeier & Smith, 1978; USDA Agricultural Research Service, 2013, Chapter 4; Auerswald *et al*., 2014). For non-US soils, apart from several equations recently developed for certain areas of Iran (Vaezi *et al*., 2008; Shabani *et al*., 2014; Ostovari *et al*., 2016, 2018), Italy (Bagarello *et al*., 2012), and China (Wang *et al*., 2013, 2016), the best option is to use the nomograph approach.

Use of the Wischmeier *et al*. (1971) nomograph for any particular soil requires various properties of the upper horizon to be known, namely the proportion of particles with a diameter of less than 0.002 mm (clay), the proportion of particles with a diameter between 0.002 and 0.10 mm (silt plus very fine sand), the organic matter (OM) content, and the structure, along with the permeability of the complete soil profile, all in accordance with the USDA Soil Survey Manual descriptions (Soil Survey Staff, 1951).

The inclusion of the very fine sand (VFS) fraction as an explanatory variable resulted in a significant improvement in modelling the *K*-factor, in terms of both precision and simplicity (Wischmeier *et al*., 1971). Nevertheless, because obtaining and analyzing soil samples is a laborious and time-consuming process, the laboratory texture analysis is frequently limited to the three broad particle-size classes of sand, silt, and clay.

As Auerswald *et al*. (2014) point out, a lack of information about VFS fraction can cause a large error in the calculation of the *K*-factor, in view of which it is important to find some means of providing an estimate for this fraction. Various procedures of doing so, based on the broad particle-size classes of the USDA system, are available in the literature, but their performance has never been subjected to evaluation. This study aims specifically to fill this lacuna, drawing on data from more than 300,000 soil horizons.

2 MATERIALS

2.1 Input data

The data used in this study were drawn from the National Cooperative Soil Survey (NCSS) Soil Characterization Database (NCSS, 2017). This database was chosen for three reasons: (i) it includes records for the required soil particle-size classes with the certainty that these have been precisely measured and not simply estimated, (ii) the NCSS enjoys a worldwide reputation, and (iii) the number of records exceeds one order of magnitude that of other widely used soil databases. It should be made clear here that, although the database contains records from all parts of the world, over 97% of the profiles are located within the US territory.

After a preliminary review of the database, we decided to discard those records lacking the specification for the VFS fraction, those involving non-standard methods of sample preparation (such as in the case of soils derived from volcanic material), and those displaying any kind of numerical inconsistency (such as the soil fractions totalling 99.9% or 100.1% rather than 100.0%). Once the data had been filtered in this way, we were left with 307,705 records for analysing.

2.2 Models for estimating the very fine sand fraction

An exhaustive literature review provided three ways for estimating the VFS fraction from the broad particle-size classes information: the RUSLE2 formula (USDA Agricultural Research Service, 2013, Chapter 4), the ESDAC method (Panagos *et al.*, 2014), and the Shirazi-Boersma theory (Shirazi *et al.*, 1988, 2001). It so happens that in all three cases the estimation is based on a single explanatory variable, namely the sand fraction. In this respect we would like to make it clear that, unless otherwise stated (for example, in charts), the numbers referring to the fractions of sand *d,* silt *t*, clay *y*, and VFS *v* are assumed to be between 0 and 1 (both inclusive).

2.2.1 The RUSLE2 formula

RUSLE2 estimates the VFS fraction by means of expression

$$v = (0.74 - 0.62 \cdot d) \cdot d \qquad (1)$$

(USDA Agricultural Research Service, 2013, Chapter 4). This equation was derived by regression analysis using data in the NRCS RUSLE2 Database for Lancaster County, Nebraska (Foster, 2004, Chapter 7). The corresponding graph is shown in Figure 1.

2.2.2 The ESDAC method

In developing the *K*-factor map of the European Union, Panagos *et al.* (2014) considered the VFS fraction to be 20% of the sand fraction, on the basis that VFS constitutes one of the five subfractions comprising sand: very coarse, coarse, medium, fine, and very fine. Expressing it in formal mathematical terms, it would be

$$v = \frac{1}{5} \cdot d\,, \qquad (2)$$

with the corresponding graph shown in Figure 1.

As demonstrated in Appendix A, this method is essentially equivalent to using log-linear interpolation, that is, linear interpolation in terms of the logarithm of the particle sizes.

2.2.3 The Shirazi-Boersma theory

Shirazi *et al.* (1988, 2001) maintain that the cumulative particle-size distribution for fine-earth of a soil sample can be approximated by three segments of lognormal distribution function: one for the sand fraction, another for the silt fraction, and a third for the clay fraction. According to this theory, the VFS fraction can be estimated by the following interpolating function:

$$v = \Phi\left[0.698810 + 0.812098 \cdot \Phi^{-1}(1 - d)\right] - 1 + d\,, \qquad (3)$$

where Φ denotes the standard normal distribution function and Φ^{-1} its inverse (see Appendix B). The corresponding graph is shown in Figure 1.

3 METHODS

3.1 Descriptive statistics

3.1.1 Least squares polynomial regression

In order to describe how the VFS fraction varies in relation to the broad soil separate fractions (sand, silt, and clay), we first applied least squares fits by means of polynomials, namely

$$v = \sum_{i=0}^{n} \sum_{j=0}^{n-i} a_{i,j} \cdot f_1^{\,i} \cdot f_2^{\,j} \,, \tag{4}$$

where n is the degree of the polynomial, $a_{i,j}$ represents the fitting parameters, and f_1 and f_2 represent any two of the broad soil separate fractions (including the third makes no difference, as the sand, silt, and clay fractions are linearly dependent). Note, finally, that the number p of parameters $a_{i,j}$ in the above polynomial is given by the expression $p = ½\cdot(n+1)\cdot(n+2)$ (Venables & Ripley, 2002, Chapter 15).

3.1.2 Subdivision of the textural triangle

The quantity of data to be processed in the study was so great and so widely distributed around the textural triangle (TT) that it was decided to carry out a piecewise statistical analysis, that is, to divide the TT into smaller units and deal with each in turn. To this end, we first divided it into four equilateral triangles, each side of which was the length $l/2$, where l was the length of the side of the original triangle. The coefficient of determination for the VFS fraction can then be calculated by

$$R^2 = 1 - \frac{\sum_{i=1}^{q} \sum_{j=1}^{N_i} \left(v_{i,j} - m_i \right)^2}{\sum_{i=1}^{q} \sum_{j=1}^{N_i} \left(v_{i,j} - m \right)^2} \,, \tag{5}$$

where q is the number of non-empty triangles (after this first division, $1 \le q \le 4$), N_i is the number of data in the i-th non-empty triangle, $v_{i,j}$ is the j-th VFS fraction in the i-th non-empty triangle, m_i is the average VFS fraction in the i-th non-empty triangle, and m is the average VFS fraction in the original triangle. Put another way, we carried out a piecewise least-squares regression (see for example Dagnelie, 2006, Chapter 15; James *et al.*, 2013, Chapter 7), albeit in a domain of two explanatory variables where the pieces are the q non-empty triangles.

We then subdivided each of the above triangles into 4 new equilateral triangles, thus initiating an iterative process in which r (sub)divisions generated 4^r equilateral triangles with sides of $(l/2)^r$. Logically, the coefficient of determination increases as the number of subdivisions increases, but does so at the expense of a (near) exponential growth in the number q of parameters m_i in the regression. Hence, in order to fix the order of magnitude of the number of triangles to be subjected to the analysis described in sections 3.1.3 and 3.1.4 below, we applied the minimization of Akaike information criterion (AIC) (see for example Burnham & Anderson, 2002, Chapter 2). In this instance, the criterion takes the form

$$\mathrm{AIC} = N \cdot \ln\left[\frac{\sum_{i=1}^{q}\sum_{j=1}^{N_i}\left(v_{i,j} - m_i\right)^2}{N}\right] + 2 \cdot q\,, \qquad (6)$$

where N is the total number of data, that is

$$N = \sum_{i=1}^{q} N_i\,. \qquad (7)$$

Whenever one of the data points, as a result of a subdivision, lays on a side or vertex shared with two or more adjacent triangles, the data point in question was assigned to the triangle whose centre was closest to the centroid of the TT.

Finally, given that one of the statistical tests considered below, namely the Shapiro-Wilk test, requires the sample size to be not smaller than 3, some of the triangles resulting from the final subdivision were merged. Specifically, in those cases where a triangle contained fewer than 3 data points, we reverted locally the iterative process to the next immediate level until the resultant triangle met the minimum. At the end of this process, q' triangular tiles of various sizes were produced.

3.1.3 Testing normality

In order to determine the suitability of using the classical methods of statistical inference, we applied Royston's (1992) version of the Shapiro-Wilk test of normality to the values of the VFS fraction in each of the q' tiles. It has been shown that the Royston (1992) procedure is very powerful for sample sizes up to 500 (Hain, 2010) and even up to 2000 (Razali & Wah, 2011). All calculations were performed using the 'stats' package of R software (R Core Team, 2017).

3.1.4 Conditional statistical measures

From the information contained in each of the tiles, we generated the conditional values of various statistical measures related to the VFS fraction for each point in the TT. Specifically, these were the conditional values of the mean *m*, the standard deviation *s*, Fisher's skewness coefficient *g*, the first quartile Q_1, the median or second quartile Q_2, and the third quartile Q_3, obtained by means of the procedure described below.

1. Calculation of the coordinates d_i, t_i, and y_i of the centres of mass of the N_i data points in the *i*-th tile, according to the following formulae:

$$d_i = \frac{1}{N_i} \cdot \sum_{j=1}^{N_i} d_{i,j} \,, \qquad t_i = \frac{1}{N_i} \cdot \sum_{j=1}^{N_i} t_{i,j} \,, \quad \text{and} \quad y_i = \frac{1}{N_i} \cdot \sum_{j=1}^{N_i} y_{i,j} \,, \tag{8}$$

where $d_{i,j}$, $t_{i,j}$, and $y_{i,j}$ are, respectively, the sand, silt, and clay fractions of the *j*-th record of the *i*-th tile (*i* = 1, 2,…, *q'*). These centres of mass were then used as nodes of the interpolating functions defined below. Evidently, the condition that $d_i + t_i + y_i = 1$ is fulfilled.

2. Calculation of the mean m_i of the N_i values of the VFS fraction corresponding to the *i*-th tile, where *i* = 1, 2,…, *q'*.

3. Localization on the *i*-th centre of mass (interpolation node) of the mean m_i just calculated (*i* = 1, 2,…, *q'*).

4. Generation of an interpolating function for the mean *m*, in terms of the coordinates *d*, *t*, and *y* of an arbitrary point *P* in the TT, using the expression

$$m(d,t,y) = \frac{\sum_{i=1}^{q'} w_i(d,t,y) \cdot m_i}{\sum_{i=1}^{q'} w_i(d,t,y)}, \tag{9}$$

where the weighting coefficient w_i for the *i*-th interpolation node is given by

$$w_i(d,t,y) = \frac{f(N_i)}{[D_i(d,t,y)]^u}, \tag{10}$$

$f(N_i)$ being a generic function of the number of records in the *i*-th tile, D_i the measured distance on the TT between data point *P* and the *i*-th interpolation node, and *u* a simple numeric parameter. Both the function $f(N_i)$ and the value of *u* were selected from a set of pre-established options, namely N_i, sqrt (N_i), log N_i, and 1 for $f(N_i)$; 1, 2, 3, and 4 for *u*, by means of

a least squares approach. Note that in the hypothetical case of $f(N_i) = 1$ and $u = 2$, we would find ourselves using Shepard's (1968) canonical interpolating function.

5. Repetition of steps 2-4 for each of the remaining statistical measures, keeping both the function $f(N_i)$ and the value of u originated from the calculations associated with the mean.

At the end of the process described above, we obtained the interpolating functions for the mean $m(d, t, y)$, the standard deviation $s(d, t, y)$, the skewness coefficient $g(d, t, y)$, the first quartile $Q_1(d, t, y)$, the median $Q_2(d, t, y)$, and the third quartile $Q_3(d, t, y)$; in short the mathematical expressions of the conditional statistical measures we were seeking.

It should be pointed out that performing step 2 for each quartile consists of a piecewise quantile regression in which the pieces are the *q'* tiles.

3.2 Model assessment

Preliminary performance calculations based on a single value for the whole data set have yielded very poor results for all three models. For example, in the case of the Nash-Sutcliffe efficiency coefficient (Nash & Sutcliffe, 1970), we have obtained the following values: −10.1% with the formula of RUSLE2, 3.8% with the ESDAC method, and −92.3% with the Shirazi-Boersma theory. Even so, it is very possible that such models work properly in specific areas of the TT. Starting from this consideration, we judged that a model would be acceptable for estimating the VFS fraction at a point in the TT if it provided, for the said point, a value between those determined by the interpolating function of the first quartile $Q_1(d, t, y)$ and the interpolating function of the third quartile $Q_3(d, t, y)$, both inclusive. As a result of this procedure, for each of the models we studied, we were able to establish two regions within the TT: one where the performance of the model in question was acceptable and the other where it was not.

4 RESULTS

4.1 Descriptive statistics

4.1.1 Least squares polynomial regression

Table 1 provides, for each degree of polynomial considered (first column), the number of parameters fitted (second column), and the value of the coefficient of determination (third column).

As can be seen, the value of the coefficient of determination begins to stabilise at the second degree polynomial. At the same time, it was observed that once a polynomial of a given degree had been fit, some of the terms comprising it held little relevance, leading us to repeat the fitting procedure with a reduced number of parameters. This way, we arrived at a second degree polynomial with a coefficient of determination of 32.0%, containing just 3 parameters, namely

$$v = 0.219386 \cdot d - 0.151968 \cdot d^2 + 0.437529 \cdot d \cdot t \,. \qquad (11)$$

4.1.2 Subdivision of the textural triangle

Table 2 shows, for each level of (sub)division considered in the TT (first column), the number of triangles generated (second column), the number of non-empty triangles (third column), the value of the coefficient of determination (fourth column), and the value of the Akaike information criterion (fifth column).

As can be seen, the Akaike information criterion reaches its minimum value when *r* is 5. This determined that the subsequent calculations began precisely from this level of subdivision, that is to say, from establishing and considering 1024 equilateral triangles of the same size. The merger of some of these triangles in order to obtain tiles with at least 3 records each, resulted in 976 tiles being generated, 964 with sides of 3.125 per cent, 11 with sides of 6.25 per cent, and 1 with sides of 12.5 per cent, as seen in Figure 2.

Figure 2 likewise shows the relative density of the records in each tile, that is to say, the ratio between the density of the records in each tile and the density of the records across the whole TT. In the interests of clarity, the limits of the texture classes of the USDA system have been superimposed. Thus, the highest values occur in the sand (S) class, with an absolute maximum of 14.4 in the tile whose centroid has coordinates 97.9% sand, 1.1% silt, and 1.0% clay. There is also a relative maximum of 5.0 in the silty clay loam (SICL) class, specifically in the tile whose centroid has coordinates 2.1% sand, 64.6% silt, and 33.3% clay. The texture classes loam (L), silty clay (SIC), loamy sand (LS), sandy loam (SL), and clay loam (CL) also have generally above average densities, whilst in contrast, the texture classes sandy clay (SC), silt (SI), clay (C), and sandy clay loam (SCL) generally present below average densities. Finally, the class silt loam (SIL) has values clearly above the average in its more clayey part and far below average in its less clayey part.

4.1.3 Testing for normality

The application of the Shapiro-Wilk test to the VFS data of each tile led to the rejection of the null hypothesis, with a significance level of 0.05, in 93.1% of the tiles. In view of this, we applied the transformations of variables most commonly used in such cases, namely the logarithm transformation, the square-root transformation, and, after dividing by the sand fraction, the arcsine transformation. Unfortunately, the results did not substantially improve, returning 71.1% rejection for the logarithmic transformation, 77.6% for the square root transformation, and 85.8% for the arcsine transformation.

4.1.4 Conditional statistical measures

Among the functions $f(N_i)$ and the values of the parameter u which were considered (see section 3.1.4), those which performed best in terms of least squares were $f(N_i) = \log N_i$ and $u = 3$. Figures 3 and 4 show the interpolated values of the statistical measures of interest.

In Figure 3a, it can be seen that, with the exception of the interval between approximately 0 and 20% sand, the conditional mean of the VFS fraction depends on more than one broad particle-size classes. The highest values are to be found in the less clayey areas of SL and SIL classes. Specifically, the maximum is located at 52.1% sand, 46.2% silt, and 1.7% clay (SL class), with a value of 23.5%.

In Figure 3b, it can be seen that the conditional standard deviation of the VFS fraction is highly variable, indicating that there is a clear case here of absence of homoscedasticity. This implies that polynomial regressions such as that calculated in section 4.1.1 are not really suitable for providing prediction intervals.

According to Bulmer (1979), a distribution can be considered fairly symmetrical when the absolute value of the Fisher´s skewness coefficient is below 0.5, moderately skewed when it is between 0.5 and 1, and highly skewed when it is above 1. In accordance with this rule, Figure 3c allows us to conclude that the conditional distribution of the VFS fraction is asymmetrical in 85% of the area of the TT, an area in which the median is thus preferable to the mean as a measure of location.

In Figure 4, it can be seen that the conditional first quartile, median, and third quartile of the VFS fraction follow a similar pattern to the conditional mean (Figure 3a). Hence, with the exception of the interval comprising between approximately 0 and 20% sand, these three measures do not depend solely on the sand fraction. In the case of all three quartiles, the highest values are found in the least clayey areas of SL and SIL classes. Specifically, the

maximum value of the first quartile lies at 26.1% sand, 73% silt, and 0.9% clay (SIL class), with a value of 17.7%; the median maximum is found at 39.7% sand, 58.7% silt, and 1.6% clay (SIL class), with a value of 21.5%; and the third quartile maximum lies at the same point as the maximum of the median, with a value of 32.7%.

4.2 Model assessment

As can be seen in Figure 5a, the RUSLE2 formula is acceptable in SI and SIL classes, in the most silty area of SICL class, and in the least clayey area of SL class. In quantitative terms, the acceptance region represents only 19.7% of the TT, with 22.7% of the records located within it.

With respect to the ESDAC method, Figure 5b shows that its acceptance region covers a large portion of the TT, specifically 47.5% of its area, albeit accounting for only 38.3% of the records. Broadly speaking, the method is acceptable in the band containing between 10 and 30% silt, and in those areas of the L and SIL classes with clay between 8 and 18%, provided that the silt content does not exceed 80%.

For its part, Figure 5c shows that the Shirazi-Boersma theory is acceptable in SI, SIL, and SICL classes, and also the least sandy area of SIC class. In this respect, the figure bears a strong resemblance to Figure 5a (RUSLE2). In quantitative terms, the acceptance region represents 23.3% of the total area of the TT, and accounts for 32.6% of the records.

When the three figures above were superimposed, it became evident that there are areas of overlap among the acceptance regions of the models. At those points where more than one model is acceptable, we gave preference to the one providing the value closest to the median. Figure 6 shows the result of applying this criterion. Thus, the ESDAC method is the most appropriate in 46.8% of the TT, the Shirazi-Boersma theory in 15.0%, and the RUSLE2 formula in 7.1%. This leaves 31.1% of the TT, accounting for 32.3% of the records, in which none of the models provides an acceptable estimation. It could be highlighted that the S class, having the highest density of records, falls thoroughly in this area.

5 DISCUSSION AND CONCLUSIONS

As noted in section 1, there are no published works evaluating procedures to estimate the VFS fraction of fine earth of soils on the basis of the broad particle-size classes, neither are there any antecedents for the statistical analysis of texture data drawn from hundreds of several thousands of soil samples. In quantitative terms, the most notable studies to date are Nemes *et al*. (1999), drawing on 14,584 samples, and Shangguan *et al*. (2014), with 16,349 samples, both aimed at evaluating the performance of several procedures and models for describing

cumulative particle-size distributions; the work of Auerswald *et al*. (2014), with 19,055 samples, along with Panagos *et al*. (2014), 19,969 samples, both focussing on calculating the *K*-factor; that of Levi (2017), which included 75,736 samples and tested pedotransfer functions; and that of Padarian *et al*. (2012), with 160,904 samples, aiming the conversion from Australian to USDA soil particle size classification system.

Although the data used in this study come chiefly from US soils, the scatter plots in the TT based on data from other geographical regions are consistent with the density of records shown in Figure 2. Examples of this can be found in Nemes *et al*. (1999, Fig. 3 and 8), drawing on data from the Soil Information System of the Netherlands and the Hydraulic Properties of European Soils (HYPRES) database, Minasny *et al*. (1999, Fig. 1), for Australian soils, and Hwang *et al*. (2002, Fig. 1), for South Korean soils. The consistency of the data across these studies allows us to conjecture that at least our findings can be extrapolated to a large part of the soils in Europe, Australia, and the Far East.

The claim to estimate the VFS fraction based only on the sand fraction, as the three models evaluated do, has been shown to be over-optimistic in view of Figures 3 and 4. Moreover, any conceivable model based solely on the broad particle-size classes would suffer from a high degree of uncertainty, as the coefficient of determination in Table 1 stabilizes around 33%. The only way to increase the coefficient of determination is to consider further explanatory variables. Potential candidates in this respect could be the fraction of particles with a diameter greater than 2 mm (coarse fragments) or non-granulometric information such as differentiation by type of soil horizon, by kind of parent material, or by mode of deposition in the case of sedimentary environments.

None of the three models evaluated achieved an overall satisfactory performance. Indeed, it is worth noting that, despite its modest approach, the ESDAC method came out best. With regard to the Shirazi-Boersma theory, it could be argued in mitigation that it was not specifically designed to estimate the VFS fraction, but rather describe the whole cumulative particle-size distribution. Nevertheless, the fact that it is inaccurate in this particular instance somehow calls into question its efficacy in fulfilling its original purpose. The weakest performance was by the RUSLE2 regression equation, probably because the data on which it was based (Lancaster County, Nebraska) are unrepresentative of the range of US soils.

The deficiencies of the three models can be illustrated with an example focussing on the texture class S, to which, it will be remembered, the greatest number of records pertain. It is a horizon containing 93.4% sand, 3.9% silt, and 2.7% clay. The VFS fraction would be 15.0%

according to the RUSLE2 formula, 18.7% according to the ESDAC method, and 23.4% according to the Shirazi-Boersma theory. If the actual value for this fraction were what Figure 4b (the conditional median) indicates, that is 6.0%, the relative errors would be 150.0%, 211.7%, and 290.0% respectively. Assuming that the horizon contains 2% of organic matter, has a fine granular structure and moderate permeability, we can calculate the repercussions of these errors on the *K*-factor. Taking the VFS fractions given above, the *K*-factor values calculated by the Wischmeier *et al*. (1971) nomograph are 0.146 t h MJ^{-1} cm^{-1} (RUSLE2), 0.179 t h MJ^{-1} cm^{-1} (ESDAC), and 0.222 t h MJ^{-1} cm^{-1} (Shirazi-Boersma). If we accept the value of 6% VFS given by Figure 4b, the *K*-factor would be 0.103 t h MJ^{-1} cm^{-1} and the relative errors would be 41.7% (RUSLE2), 73.8% (ESDAC), and 115.5% (Shirazi-Boersma).

In view of the foregoing, it is clear that accurate models for estimating the VFS fraction are yet to be developed. In this respect, Figure 4 could represent an appropriate starting point, to be validated by future studies.

6 APPLICATION

Evidently, each of the models evaluated can be used to estimate the VFS fraction in its corresponding region of acceptance. The main problem is that there is a third of the area of the TT where no model is acceptable.

Equation (11), obtained through a global (non-piecewise) least squares regression could be an alternative. However, the establishment of prediction intervals for this kind of regression function requires the conditional distribution of the target variable to be normal and homocedastic (Preston, 2000); these requirements, as demonstrated, are not satisfied in the case of the VFS fraction. Moreover, least squares regressions are closely linked to the concept of mean, whereas the median is preferable as a measure of location for distributions like the conditional VFS distribution, which is clearly asymmetric in most of the TT.

Unlike equation (11), the charts in Figure 4 have their origin in distinct local (piecewise) quantile regressions, leading us to a context in which it is not necessary to satisfy the requirements of normality and homoscedasticity in order to establish prediction intervals (Preston, 2000). Consequently, when it comes to estimating the VFS fraction, we recommend the adoption of the value provided by Figure 4b (conditional median) along with the 50% prediction interval delimited by the values provided by Figures 4a (first conditional quartile) and 4c (third conditional quartile). Thus, for the example in section 5, the predicted value for

the VFS fraction would be 6.0% (Figure 4b) and its 50% prediction interval would be delimited by 2.6% (Figure 4a) and 12.1% (Figure 4c).

The fact that the *K*-factor is a monotonic function of the VFS fraction allows us to produce charts of the conditional quartiles of the *K*-factor from charts of the conditional quartiles of the VFS fraction. ~~On that basis we have generated the charts for K_0 (Figure 7), that is, the first approximation of the *K*-factor for 2% OM content, by means of the RUSLE2 formulae (USDA Agricultural Research Service, 2013, Chapter 4).~~

On that basis we have generated by means of the RUSLE2 formulae (USDA Agricultural Research Service, 2013, Chapter 4) the charts for K_0 (Figure 7), that is, the so-called first approximation of *K*-factor (Auerswald et al 2014), calculated with a 2% OM.

~~setting the OM content to 2%~~

Thus, for the given example, the estimated value of K_0 would be 0.070 t h MJ^{-1} cm^{-1} (Figure 7b) and the 50% prediction interval would range between 0.043 t h MJ^{-1} cm^{-1} (Figure 7a) and 0.120 t h MJ^{-1} cm^{-1} (Figure 7c).

In order to obtain values of the first approximation of *K* for OM contents other than 2%, the K_0 values from the chart should be multiplied by the following coefficient of correction:

$$k = \begin{cases} 1.2 - 0.1 \cdot a, & \text{for } a \le 4 \\ 0.8, & \text{for } a > 4 \end{cases}, \qquad (12)$$

where *a* is the OM content expressed as a percentage.

The final value of the *K*-factor is obtained by adding the effect of the structure of the upper soil horizon and the effect of permeability of the whole profile (Wischmeier *et al*., 1971; Auerswald *et al.*,2014).

7 CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

8 ACKNOWLEDGEMENTS

The authors would like to thank Ellis Benham, from USDA/NRCS Kellogg Soil Survey Laboratory, for the advice given on the use of the NCSS Soil Characterization Database; Daniel C. Yoder from the University of Tennessee-Institute of Agriculture, for his help in defining authorship and quoting of RUSLE2 documents.

9 REFERENCES

Arnold, J.G., Srinivasan, R., Muttiah, R.S., & Williams, J.R. (1998). Large area hydrologic modeling and assessment part 1: Model development. *Journal of the American Water Resources Association, 34*(1), 73-89. doi:10.1111/j.1752-1688.1998.tb05961.x

Auerswald, K., Fiener, P., Martin, W., & Elhaus, D. (2014). Use and misuse of the K factor equation in soil erosion modeling: An alternative equation for determining USLE nomograph soil erodibility values. *Catena, 118*, 220-225. doi:10.1016/j.catena.2014.01.008

Bagarello, V., Di Stefano, V., Ferro, V., Giordano, G., Iovino, M., & Pampalone, V. (2012). Estimating the USLE soil erodibility factor in Sicily, South Italy. *Applied Engineering in Agriculture, 28*(2), 199-206. doi:10.13031/2013.41347

Bouyoucos, G.J. (1935). The clay ratio as a criterion of susceptibility of soil to erosion. *Journal of the American Society of Agronomy, 27*(9), 738-741.

Bulmer, M.G. (1979). *Principles of statistics*. New York: Dover.

Burnham, K., & Anderson, D. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York: Springer.

Cronshey, R.G., & Theurer, F.D. (1998). AnnAGNPS: Non-point pollutant loading model. In *Proceedings of the First Interagency Hydrologic Modeling Conference, April 19-23, 1998, Las Vegas, Nevada: Vol. 1* (pp. 1.9-1.16). Reston, Virginia: USGS Water Information Coordination Program.

Dagnelie, P. (2006). *Statistique théorique et appliquée, 2: Inférence statistique à une et à deux dimensions* (2nd ed.). Bruxelles: De Boeck.

Flacke, W., Auerswald, K., & Neufang, L. (1990). Combining a modified Universal Soil Loss Equation with a digital terrain model for computing high resolution maps of soil loss resulting from rain wash. *Catena, 17*(4), 383-397. doi:10.1016/0341-8162(90)90040-K

Foster, G.R. (2004). *User's reference guide: Revised Universal Soil Loss Equation, Version 2 (RUSLE2), Draft.* Accessed at https://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/16/nrcs144p2_075348.pdf, 27 April, 2018.

Hain, J. (2010). *Comparison of common tests for normality.* Master thesis. Würzburg, Germany: University of Würzburg. Accessed at http://www.statistik-mathematik.uni-

wuerzburg.de/fileadmin/10040800/user_upload/hain/da_hain_final.pdf, April 27, 2018.

Hwang, S.I., Lee, K.P., Lee, D.S., & Powers, S. (2002). Models for estimating soil particle-size distributions. *Soil Science Society of America Journal, 66*(4), 1143-1150. doi:10.2136/sssaj2002.1143

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. New York: Springer. doi:10.1007/978-1-4614-7138-7

Kinnell, P.I.A., & Risse, L.M. (1998). USLE-M: Empirical modeling rainfall erosion through runoff and sediment concentration. *Soil Science Society of America Journal, 62*(6), 1667-1672. doi:10.2136/sssaj1998.03615995006200060026x

Levi, M. (2017). Modified centroid for estimating sand, silt, and clay from soil texture class. *Soil Science Society of America Journal, 81*(3), 578-588. doi:10.2136/sssaj2016.09.0301

Middleton, H.E. (1930). The properties of soils which influence erosion. *U.S. Department of Agriculture Technical Bulletin, 178*, 1-16. doi:10.2136/sssaj1930.036159950b1120010021x

Minasny, B., McBratney, A.B., & Bristow, K.L. (1999). Comparison of different approaches to the development of pedotransfer functions for water-retention curves. *Geoderma, 93*(3), 225-253. doi:10.1016/S0016-7061(99)00061-0

Nash, J.E., & Sutcliffe, J.V. (1970). River flow forecasting through conceptual models part I: A discussion of principles. *Journal of Hydrology, 10*(3), 282-290. doi:10.1016/0022-1694(70)90255-6

NCSS (2017). National Cooperative Soil Survey. National Cooperative Soil Survey Characterization Database. Accessed at http://ncsslabdatamart.sc.egov.usda.gov/, September 25, 2017.

Nemes, A., Wösten, J.H.M., Lilly, A., & Oude Voshaar, J.H. (1999). Evaluation of different procedures to interpolate particle-size distributions to achieve compatibility within soil databases. *Geoderma, 90*(3), 187-202. doi:10.1016/S0016-7061(99)00014-2

Ostovari, Y., Ghorbani-Dashtaki, S., Bahrami, H.A., Abbasi, M., Dematte, J.A.M., Arthur, E., & Panagos, P. (2018). Towards prediction of soil erodibility, SOM and $CaCO_3$ using laboratory Vis-NIR spectra: A case study in a semi-arid region of Iran. *Geoderma, 314*, 102-112. doi:10.1016/j.geoderma.2017.11.014

Ostovari, Y., Ghorbani-Dashtaki, S., Bahrami, H.A., Naderi, M., Dematte, J.A.M., & Kerry, R. (2016). Modification of the USLE K factor for soil erodibility assessment on calcareous soils in Iran. *Geomorphology, 273*, 385-395. doi:10.1016/j.geomorph.2016.08.003

Panagos, P., Borrelli, P., Poesen, J., Ballabio, C., Lugato, E., Meusburger, K., Montanarella, L., & Alewell, C. (2015). The new assessment of soil loss by water erosion in Europe. *Environmental Science & Policy, 54*(Supplement C), 438-447. doi:10.1016/j.envsci.2015.08.012

Panagos, P., Meusburger, K., Ballabio, C., Borrelli, P., & Alewell, C. (2014). Soil erodibility in Europe: A high-resolution dataset based on LUCAS. *Science of the Total Environment, 479-480*(1), 189-200. doi:10.1016/j.scitotenv.2014.02.010

Padarian, J., Minasny, B., & McBratney, A. (2012). Using genetic programming to transform from Australian to USDA/FAO soil particle-size classification system. *Soil Research, 50*(6), 443-446. doi:10.1071/SR12139

Preston, S. (2000). Teaching prediction intervals. *Journal of Statistics Education*, *8*(3). Accessed at http://ww2.amstat.org/publications/jse/secure/v8n3/preston.cfm, April 27, 2018.

R Core Team (2017). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Razali, N.M., & Wah, Y.B. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics, 2*(1), 21-33. Accessed at https://www.nrc.gov/docs/ML1714/ML17143A100.pdf, April 27, 2018.

Renard, K.G., Foster, G.R., Weesies, G.A., McCool, D.K., & Yoder, D.C. (Coordinators) (1997). *Predicting soil erosion by water: A guide to conservation planning with the Revised Universal Soil Loss Equation (RUSLE)*. Agriculture Handbook 703. Tucson, Arizona: USDA Agricultural Research Service.

Renard, K.G., Yoder, D.C., Lightle, D.T., & Dabney, S.M. (2011). Universal Soil Loss Equation and Revised Universal Soil Loss Equation. In R.P.C. Morgan & M.A. Nearing (Eds.), *Handbook of Erosion Modelling* (pp. 137-167). Chichester, United Kingdom: Wiley-Blackwell. doi:10.1002/9781444328455.ch8

Royston, P. (1992). Approximating the Shapiro-Wilk W-test for non-normality. *Statistics and Computing, 2*(3), 117-119. doi:10.1007/BF01891203

Shabani, F., Kumar, L., & Esmaeili, A. (2014). Improvement to the prediction of the USLE K factor. *Geomorphology, 204*, 229-234. doi:10.1016/j.geomorph.2013.08.008

Shangguan, W., Dai, Y., García-Gutiérrez, C., & Yuan. H. (2014). Particle-size distribution models for the conversion of Chinese data to FAO/USDA system. *The Scientific World Journal, 2014*(Article ID 109310), 1-11. doi:10.1155/2014/109310

Shepard, D. (1968). A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the 1968 23rd ACM National Conference, August 27-29, 1968* (pp. 517-524). New York: Association for Computing Machinery. doi:10.1145/800186.810616

Shirazi, M.A., Boersma, L., & Burch Johnson, C. (2001). Particle-size distributions: Comparing texture systems, adding rock, and predicting soil properties. *Soil Science Society of America Journal, 65*(2), 300-310. doi:10.2136/sssaj2001.652300x

Shirazi, M.A., Boersma, L., & Hart, J.W. (1988). A unifying quantitative analysis of soil texture: Improvement of precision and extension of scale. *Soil Science Society of America Journal, 52*(1), 181-190. doi:10.2136/sssaj1988.03615995005200010032x

Soil Survey Staff (2011). *Soil Survey Laboratory Information Manual*. Soil Survey Investigations Report 45, Version 2.0. Lincoln, Nebraska: USDA Natural Resources Conservation Service.

Soil Survey Staff (1951). *Soil survey manual*. Agriculture Handbook 18. Washington, D. C.: USDA Agricultural Research Administration.

USDA Agricultural Research Service (2013). Science Documentation: Revised Universal Soil Loss Equation, Version 2 (RUSLE2). Accessed at https://www.ars.usda.gov/ARSUserFiles/60600505/RUSLE/RUSLE2_Science_Doc.pdf, 27 April, 2018.

Vaezi, A.R., Sadeghi, S.H.R., *Bahrami*, H.A., & Mahdian, M.H. (2008). Modeling the USLE K-factor for calcareous soils in northwestern Iran. *Geomorphology*, *97*(3), 414-423. doi:10.1016/j.geomorph.2007.08.017

Van Rompaey, A.J.J., Verstraeten, G., Van Oost, K., Govers, G., & Poesen, J. (2001). Modelling mean annual sediment yield using a distributed approach. *Earth Surface Processes and Landforms, 26*(11), 1221-1236. doi:10.1002/esp.275

Venables, W.N., & Ripley, B.D. (2002). *Modern applied statistics with S* (4th ed.). New York: Springer. doi:10.1007/978-0-387-21706-2

Wang, B., Zheng, F., & Guan, Y. (2016). Improved USLE-K factor prediction: A case study on water erosion areas in China. *International Soil and Water Conservation Research, 4*(3), 168-176. doi:10.1016/j.iswcr.2016.08.003

Wang, B., Zheng, F., & Römkens, M.J.M. (2013). Comparison of soil erodibility factors in USLE, RUSLE2, EPIC and Dg models based on a Chinese soil erodibility database. *Acta Agriculturae Scandinavica, Section B: Soil & Plant Science, 63*(1), 69-79. doi:10.1080/09064710.2012.718358

Williams, J.R. (1975). Sediment-yield prediction with Universal Equation using runoff energy factor. In *Present and Prospective Technology for Predicting Sediment Yield and Sources.* Publication ARS-S-40 (pp. 244-252). New Orleans: USDA Agricultural Research Service, Southern Region.

Williams, J.R., Renard, K.G., & Dyke, P.T. (1983). EPIC: A new method for assessing erosion's effect on soil productivity. *Journal of Soil and Water Conservation, 38*(5), 381-383.

Wischmeier, W.H., Johnson, C.B., & Cross, B.V. (1971). A soil erodibility nomograph for farmland and construction sites. *Journal of Soil and Water Conservation, 26*(5), 189-193.

Wischmeier, W.H., & Smith, D.D. (1978). *Predicting rainfall erosion losses: A guide to conservation planning.* Agriculture Handbook 537. Washington, D. C.: USDA Science and Education Administration.

Wischmeier, W.H., & Smith, D. D. (1965). *Predicting rainfall erosion losses from cropland east of the Rocky Mountains: Guide for selection of practices for soil and water conservation.* Agriculture Handbook 282. Washington, D. C.: USDA Agricultural Research Service.

10 FIGURE CAPTIONS

Figure 1. Models for estimating the very fine sand fraction.

Figure 2. Relative density of data points, expressed as the ratio between the density of the records in each tile and the density of the records across the whole textural triangle.

Figure 3. Interpolated values of the main conditional measures of the very fine sand fraction.

Figure 4. Interpolated values of the conditional quartiles of the very fine sand fraction.

Figure 5. Acceptance regions of the models for estimating the very fine sand fraction.

Figure 6. Preference regions of the models for estimating the very fine sand fraction.

Figure 7. Charts of the conditional quartiles of the first approximation of *K*-factor for 2% organic matter content, K_0.

11 TABLES

Table 1. Coefficients of determination for least squares polynomial regression as function of the polynomial degree.

n	*p*	R^2 (%)
1	3	16.8
2	6	32.0
3	10	32.3
4	15	32.5
5	21	32.7
6	28	32.8
7	36	32.9

Legend: *n* is the degree of the polynomial, *p* is the number of parameters of the polynomial and R^2 is the coefficient of determination.

Table 2. Values of the Akaike information criterion for the piecewise least squares regression as function of subdivision level of the textural triangle.

r	4^r	*q*	R^2 (%)	AIC × 10^{-6}
1	4	4	16.8	−1.57252
2	16	16	26.5	−1.61280
3	64	64	30.9	−1.63155
4	256	256	32.5	−1.63835
5	1024	1008	33.1	−1.63980
6	4096	3923	33.9	−1.63728
7	16384	14842	36.1	−1.62597

Legend: *r* is the level of subdivision in the TT; 4^r is the number of triangles generated; *q* is the number of parameters adjusted; R^2 is the coefficient of determination, and AIC is the value of the Akaike information criterion.

12 FIGURES

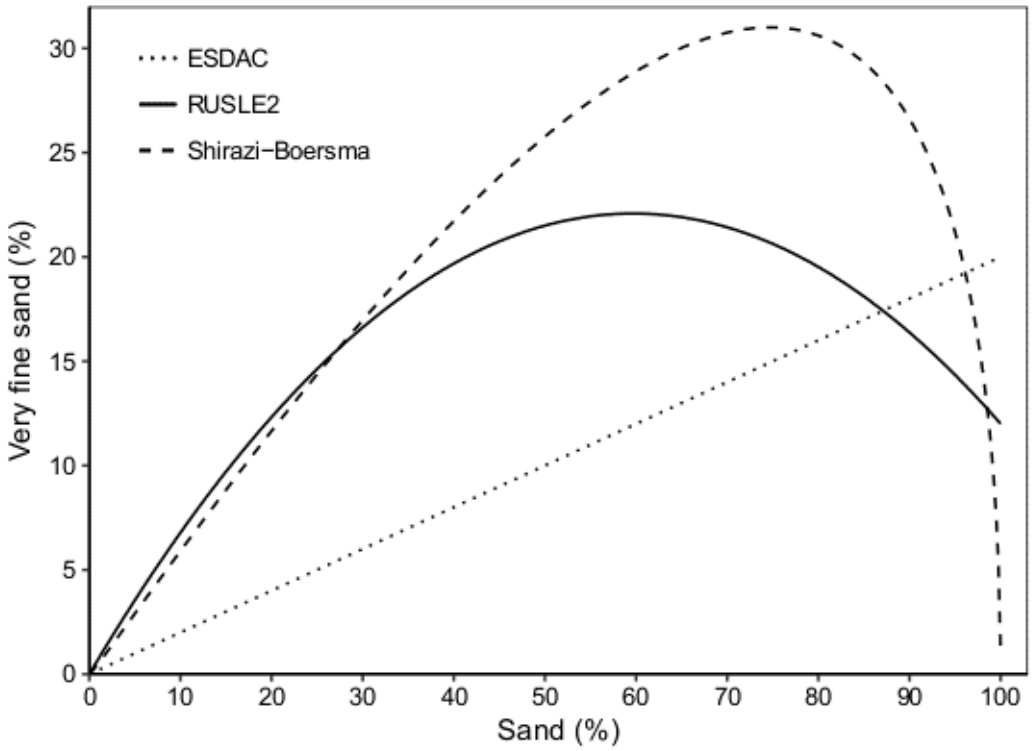

Figure 1. Models for estimating the very fine sand fraction.

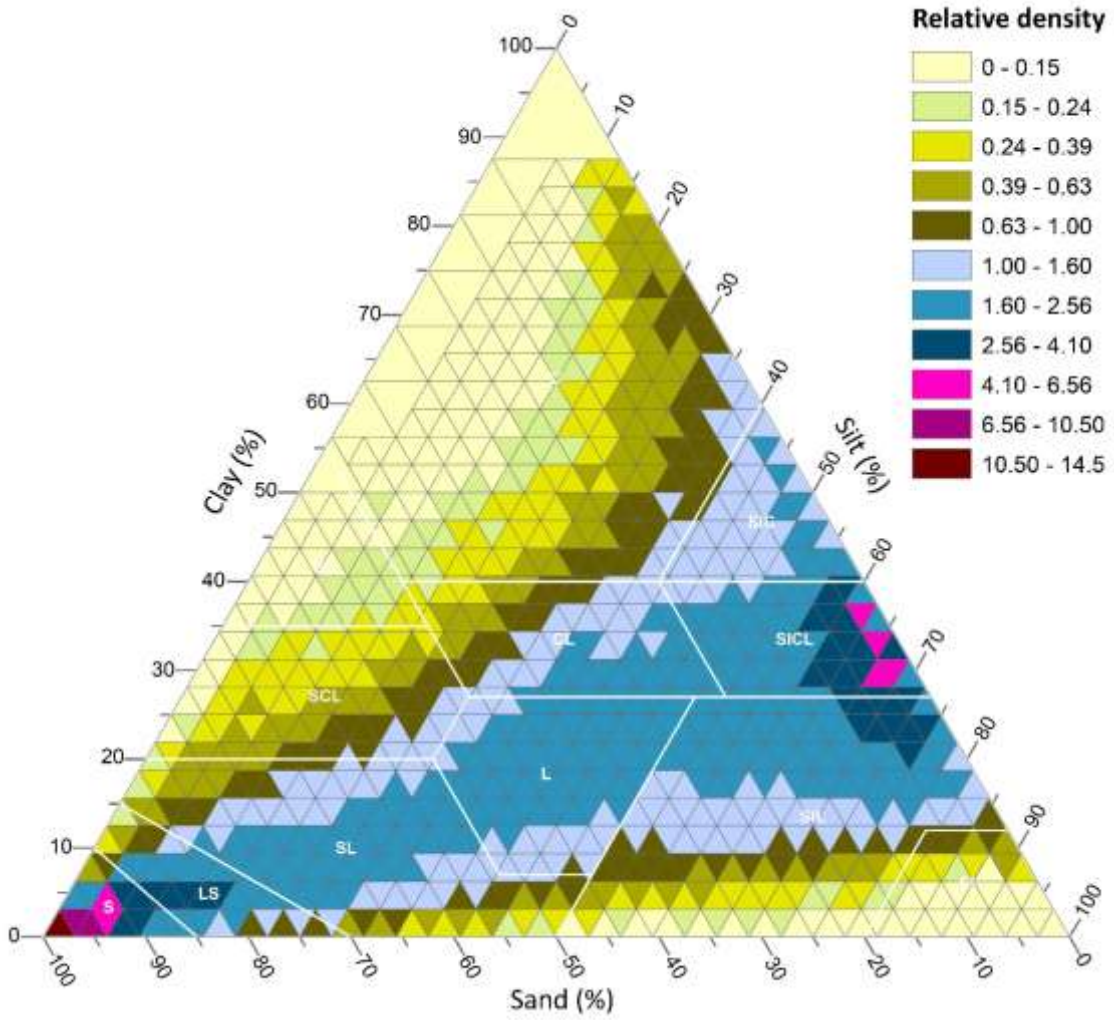

Figure 2. Relative density of data points, expressed as the ratio between the density of the records in each tile and the density of the records across the whole textural triangle.

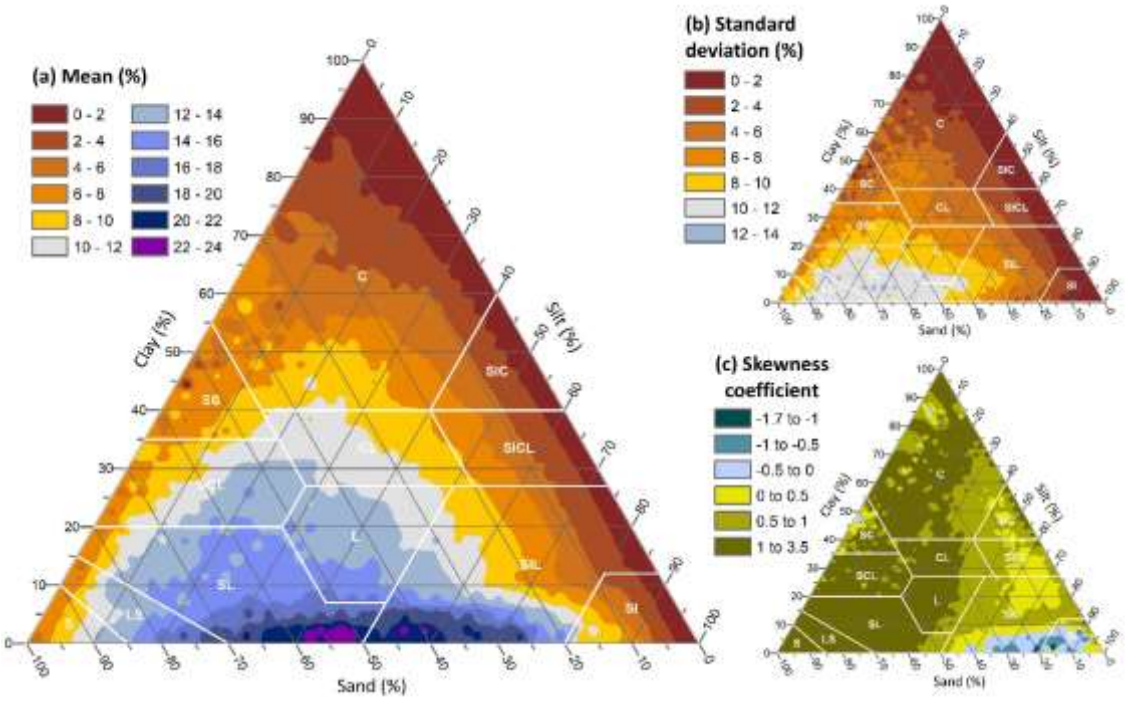

Figure 3. Interpolated values of the main conditional measures of the very fine sand fraction.

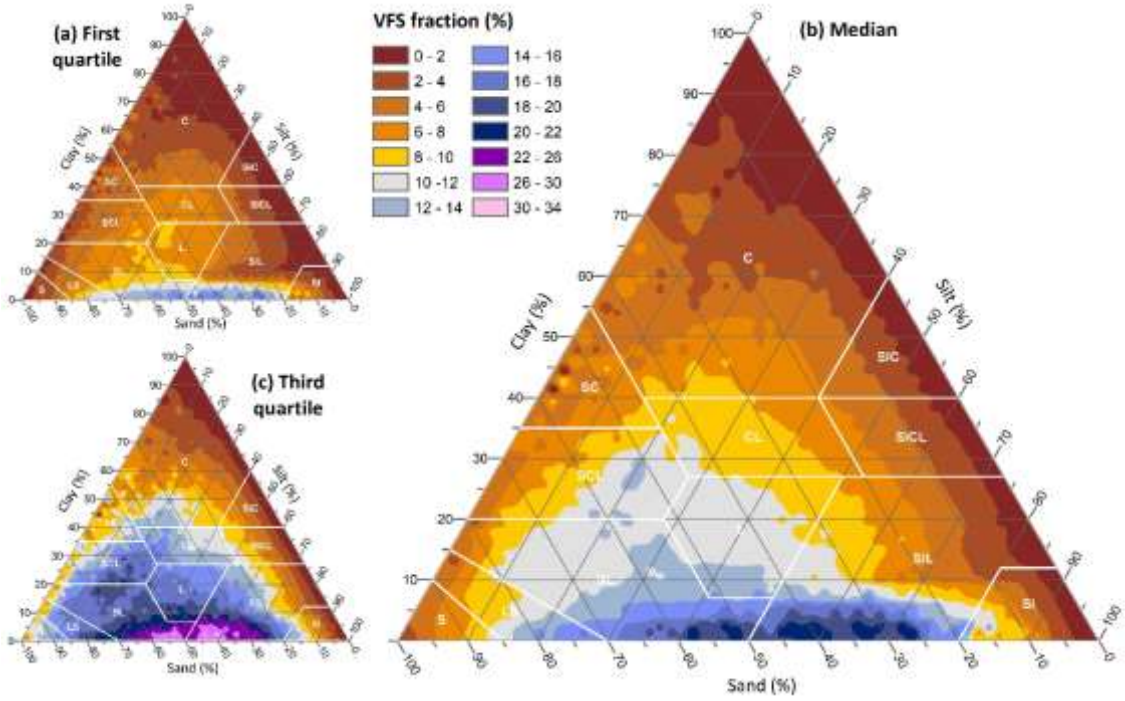

Figure 4. Interpolated values of the conditional quartiles of the very fine sand fraction.

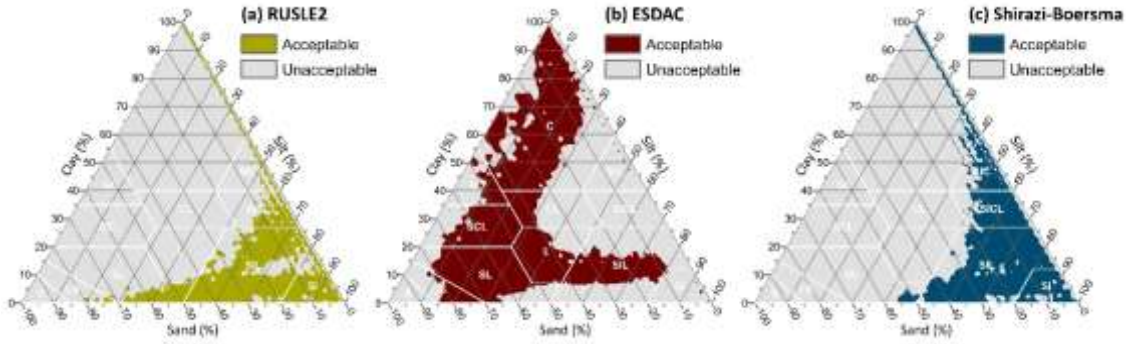

Figure 5. Acceptance regions of the models for estimating the very fine sand fraction.

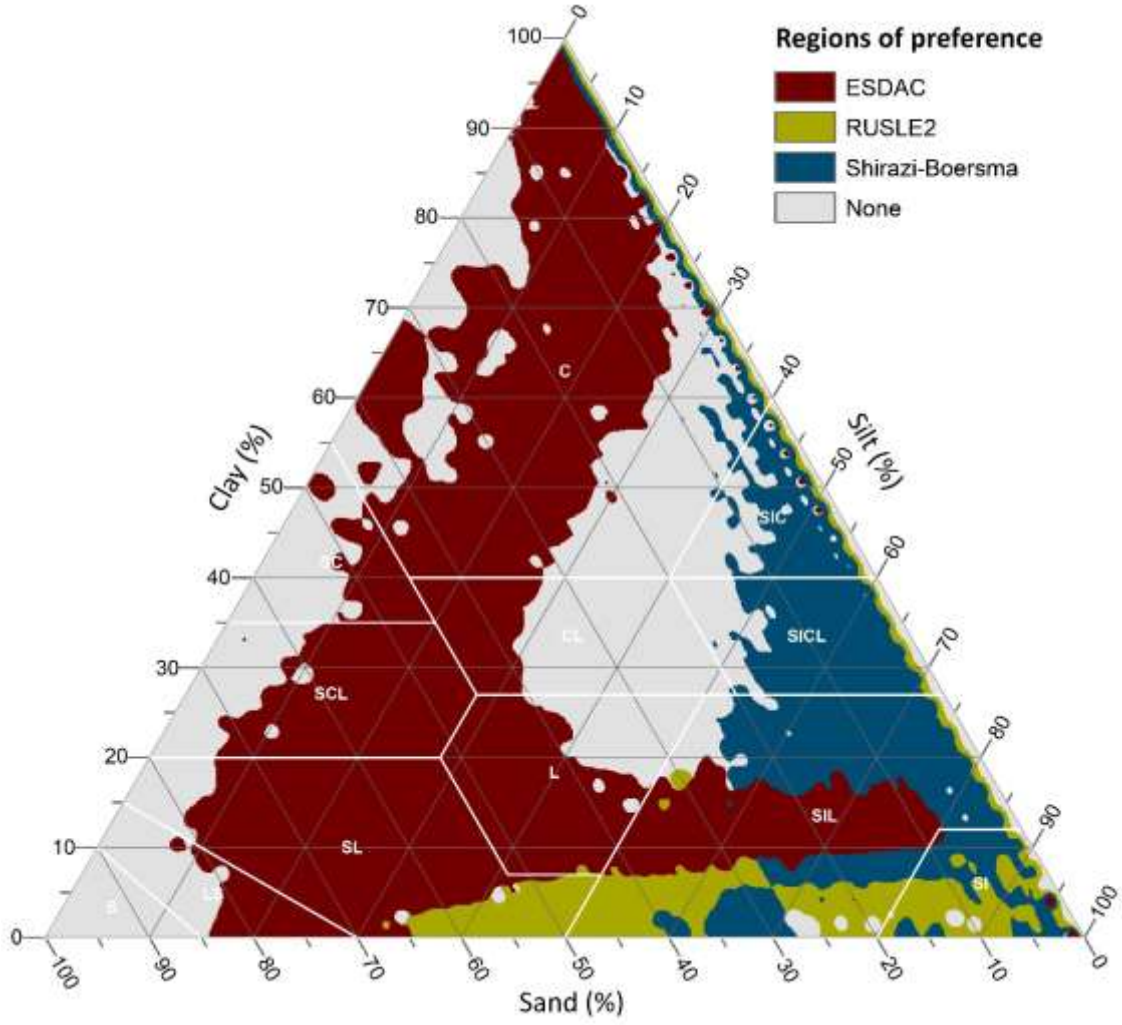

Figure 6. Preference regions of the models for estimating the very fine sand fraction.

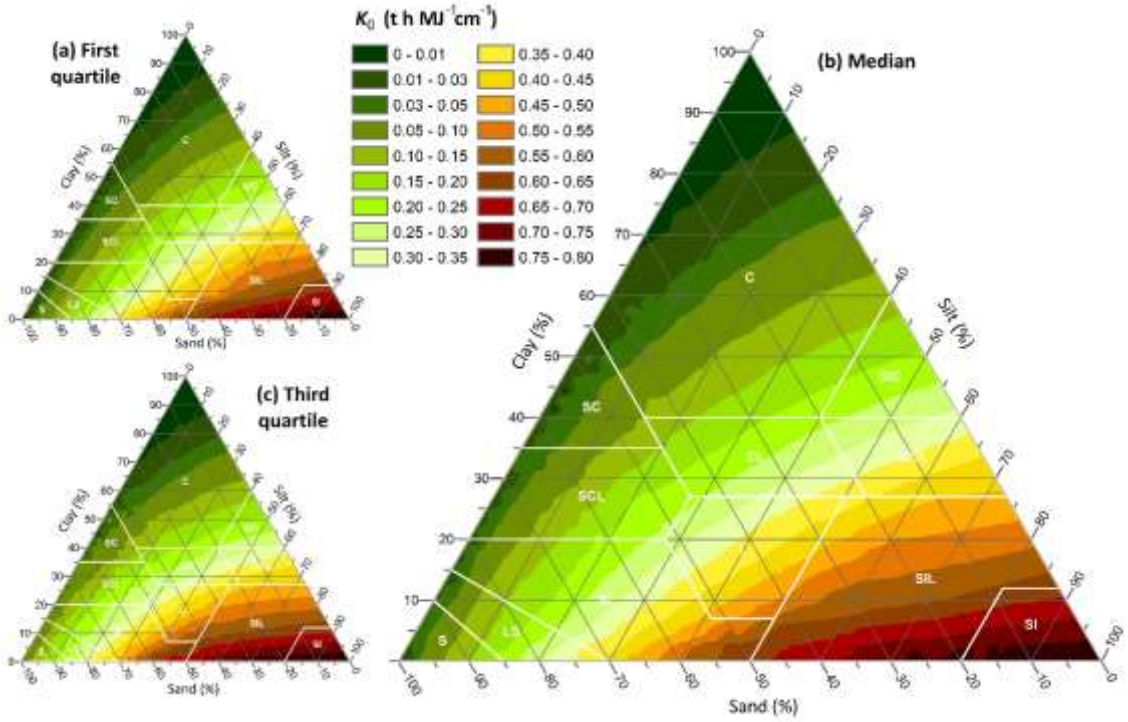

Figure 7. Charts of the conditional quartiles of the first approximation of *K*-factor for 2% organic matter content, K_0.