Received: 5 March 2021 | Revised: 28 May 2021 | Accepted: 1 June 2021

DOI: 10.1111/ejss.13137

ORIGINAL ARTICLE

European Journal of Soil Science WILEY

Statistical modelling of measurement error in wet chemistry soil data

Cynthia C. E. van Leeuwen[1,2] | Vera L. Mulder[2] | Niels H. Batjes[1] | Gerard B. M. Heuvelink[1,2]

[1]ISRIC–World Soil Information, Wageningen, The Netherlands

[2]Soil Geography and Landscape Group, Wageningen University, Wageningen, The Netherlands

Correspondence
Cynthia C. E. van Leeuwen, ISRIC–World Soil Information, PO Box 353, Wageningen, AJ 6700, The Netherlands.
Email: cynthia.vanleeuwen@wur.nl

Abstract

There is a growing demand for high-quality soil data. However, soil measurements are subject to many error sources. We aimed to quantify uncertainties in synthetic and real-world wet chemistry soil data through a linear mixed-effects model, including batch and laboratory effects. The use of synthetic data allowed us to investigate how accurately the model parameters were estimated for various experimental measurement designs, whereas the real-world case served to explore if estimates of the random effect variances were still accurate for unbalanced datasets with few replicates. The variance estimates for synthetic pH_{H_2O} data were unbiased, but limited laboratory information led to imprecise estimates. The same was observed for unbalanced synthetic datasets, where 20, 50 and 80% of the data were removed randomly. Removal led to a sharp increase of the interquartile range (IQR) of the variance estimates for batch effect and the residual. The model was also fitted to real-world pH_{H_2O} and total organic carbon (TOC) data, provided by the Wageningen Evaluating Programmes for Analytical Laboratories (WEPAL). For pH_{H_2O} , the model yielded unbiased estimates with relatively small IQRs. However, the limited number of batches with replicate measurements (5.8%) caused the batch effect to be larger than expected. A strong negative correlation between batch effect and residual variance suggested that the model could not distinguish well between these two random effects. For TOC, batch effect was removed from the model as no replicates were available within batches. Again, unbiased model estimates were obtained. However, the IQRs were relatively large, which could be attributed to the smaller dataset with only a single replicate measurement. Our findings demonstrated the importance of experimental measurement design and replicate measurements in the quantification of uncertainties in wet chemistry soil data.

Highlights

- Accurate uncertainty quantification depends on the experimental measurement design.

- Linear mixed-effects models can be used as a tool to quantify uncertainty in wet chemistry soil data.
- Lack of replicate measurements leads to poor estimates of error variance components.
- Measurement error in wet chemistry soil data should not be ignored.

KEYWORDS

accuracy, experimental design, linear mixed-effects model, replicate measurements, soil chemical data, uncertainty

1 | INTRODUCTION

A soil system's physical and chemical properties are commonly determined by the collection and subsequent wet chemistry analysis of soil samples. The results from wet chemistry measurements can be further used to, for instance, develop soil spectroscopy models (McBratney, Minasny, & Rossel, 2006) or estimate soil organic carbon stocks (Smith et al., 2020). Accurate and reliable analytical data of relevant soil properties are key to achieve accurate calibration and validation of such models (Dangal, Sanderman, Wills, & Ramirez-Lopez, 2019). The need for high-quality soil data is widely acknowledged by organizations such as the FAO's Global Soil Partnership, which established the Global Soil Laboratory Network (GLOSOLAN) in 2017 (FAO, 2019a). GLOSOLAN aims to build laboratory capacity and improve the provision of reliable and comparable soil data by harmonizing methods, units, data and information.

During the measurement process, errors can occur from field sampling (e.g., Goidts, Van Wesemael, & Crucifix, 2009), sample handling, sample transport, shipment preparation, taking a subsample for laboratory analysis and the laboratory analysis itself (Van Ee, Blume, & Starks, 1990). Errors can also occur after the laboratory analysis, for example during the data processing. Other post-analysis errors are model error, present in for instance pedotransfer functions and spectral models (Libohova et al., 2019), and interpolation error, included in digital soil mapping. In this study, we focused on the error that occurs during the laboratory analysis: the laboratory measurement error. Factors that often contribute to measurement error are the analyst, complex wet chemistry methodologies, varying measurement conditions (e.g., temperature and humidity), a variety of different sample preparation methods and the measurement instrument itself (Allchin, 2001; Libohova et al., 2019; Viscarra Rossel & McBratney, 1998). Error in soil measurements may be defined as the difference between the 'true' value of a soil property and its measured value (Hibbert, 2007). GLOSOLAN aims to reduce such method-related errors through harmonization of standard operating procedures (SOPs) for commonly used wet chemistry methods, and development of quality assurance and quality control programs. In this research, we aimed to quantify the uncertainty associated with defined analytical methods, building upon the need for high-quality soil data.

Errors in wet chemistry soil data can propagate in further applications, such as pedotransfer functions, spectral models and digital soil mapping (Heuvelink, 2018; McBratney, Minasny, Cattle, & Vervoort, 2002). Over time, the prediction accuracy of such models has improved significantly, making the relative contribution of measurement error in the calibration data larger. For example, visible (Vis), near-infrared (NIR) and mid-infrared (MIR) diffuse reflectance spectroscopy models have improved rapidly due to advances in computation, instrument manufacturing and multivariate statistics (Dangal et al., 2019; Guerrero, Viscarra Rossel, & Mouazen, 2010). Furthermore, developments in computation and statistics helped to extract useful information from the measured spectra (Viscarra Rossel et al., 2016).

The 'true' value of a soil property must be known to quantify errors in soil measurements. However, we rarely have this knowledge and therefore we are uncertain about the 'true' value. For instance, suppose a laboratory clay content measurement yields a value of 24.2%, while the 'true' clay content is 27.5%. The error is 3.3%, but we do not know it because we have only the measurement. Because we are aware that there may be a measurement error, we are uncertain about the 'true' clay content. Heuvelink, Brown, and van Loon (2007) defined uncertainty as an expression of confidence in our knowledge of the 'true' value of a specific soil property. Several methods have been developed to deal with uncertain data. Commonly, uncertainties are quantified through probability distribution functions (PDFs) (Heuvelink, 2018; Heuvelink et al., 2007). This approach allows for a complete characterization of the uncertainty, including correlations between uncertainties in soil measurements. Just as soil observations can be dependent on each other, so can the uncertainties associated with them. Furthermore, PDFs can easily be implemented in the uncertainty propagation and stochastic sensitivity analysis of environmental models (Malone, McBratney, & Minasny, 2011). Despite the availability of multiple methods, uncertainty estimates are rarely specified by providers of wet chemistry soil data, meaning that the user has little to no knowledge about the quality, or uncertainty, of these data. Furthermore, measures for uncertainty associated with compilations of such analytical datasets are seldom provided, with the exception of the WoSIS database (Batjes, Ribeiro, & van Oostrum, 2020). Hence, there is a need to quantify uncertainties in wet chemistry soil data and store detailed uncertainty information in soil databases.

In this research, we aimed to model uncertainties in synthetic and real-world wet chemistry soil pH_{H_2O} and total organic carbon (TOC) measurements through PDFs. For this, we assumed that we can represent uncertainties by normal distributions. To estimate the parameters of these distributions, that is, the variances of the error components, we applied a linear mixed-effects model approach. The use of balanced and unbalanced synthetic data allowed us to investigate how well the model parameters (i.e., the variances of the error components) can be estimated given a specific experimental measurement design. We aimed to provide guidance on the 'best', or recommended, experimental measurement design for accurate representation of the laboratory measurement error through PDFs, based on the results of the synthetic case. Furthermore, the model was applied on real-world data provided by the Wageningen Evaluating Programmes for Analytical Laboratories (WEPAL) (http://www.wepal.nl). This real-world case study served to quantify the contribution of multiple error sources to the overall measurement uncertainty and to explore if the error source variance estimates were still accurate for such unbalanced designs with only few replicate measurements.

2 | ERROR IN ANALYTICAL CHEMISTRY

Analytical measurements are subject to various error sources, such as the analyst and the instrument. Because of error, a measurement on a soil sample for a given property is considered to be only an approximation of the 'true' value. To assess the quality of a measurement, we wish to know the size of the error. In the case of repeated measurements, the measurement error can be divided into a systematic and a random component. The systematic error will be the same for all repeated measurements, whereas the random error may vary between replicates. Each measurement is the sum of the 'true' value, a systematic error and a random error:

$$Y_i = X + S + \varepsilon_i,\ \ i = 1, \cdots, n \tag{1}$$

where Y_i is the i-th measurement, X is the 'true' value of the soil property and n refers to the total number of measurements performed on the soil sample. The measured value differs from the 'true' value by a systematic component, S, and a random component, ε_i, which differs for each measurement (Figure 1). In this research, we assume that the random error follows a normal distribution, with zero mean and standard deviation σ. Because it has zero mean, it is theoretically possible to eliminate the random error by measuring the same sample an infinite number of times (Theodorsson, Magnusson, & Leito, 2014). The systematic error in Equation (1) affects the results by the same amount (i.e., we assume S to be constant). The systematic error can also be proportional to the 'true' value (Ramsey, 1998). Here, S would be dependent on X (e.g., $S = c \cdot X$, where c is a constant).

In analytical chemistry, the error in measurement results is typically addressed through applying a method validation scheme (Hibbert, 2007). Commonly used terms in method validation schemes are trueness, precision and accuracy. Trueness is defined as the closeness of agreement between the average value of a large number of measurement results and an accepted reference value (International Organization for Standardization [ISO], 1994). In other words, trueness is the systematic difference between the 'true' value, X, and the mean value of a large number of measurements, represented by S in Equation (1). When

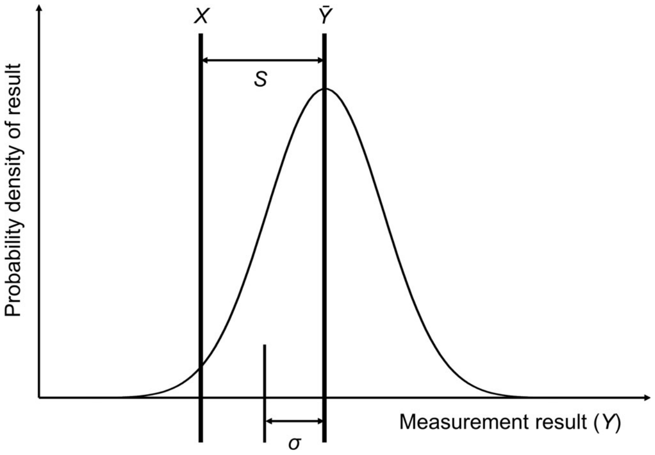


FIGURE 1 Systematic (S) and random error (with standard deviation σ) in a measurement (Y). Each measurement result deviates from the 'true' value, X, because of error. $\bar{Y}$ refers to the average of a large number of measurements on the same sample. Figure adapted from Hibbert (2007)

the trueness of measurement results is low, this means that there is a large systematic error or bias.

In method validation schemes, the random error, ε, characterizes the precision of repeated measurements. The ISO (1994) defines precision as the closeness of agreement between independent test results that were obtained under stipulated conditions. Precision includes terms such as repeatability and reproducibility (Ebentier et al., 2013; Ellison, Barwick, & Farrant, 2009). Repeatability refers to the best precision a single laboratory can obtain and can be calculated from repeated measurements in comparable conditions, that is, the same analyst, on the same day, using the same instruments. Reproducibility is a measure of agreement between repeated measurements on the same samples, obtained by the same method, under different conditions. Measurement conditions can differ within and between laboratories, resulting in within- and between-laboratory reproducibility (Ellison et al., 2009).

Both trueness and precision are used to determine the quality performance characteristics of the measurements: the accuracy (Menditto, Patriarca, & Magnusson, 2007). Accuracy can be defined as the closeness of agreement between a single measurement result and the accepted reference value, which is considered as the 'true' value (ISO,1994). Thus, accuracy relates to the sum of S and ε_i in Equation (1). The accuracy of a result is used to determine the degree of confidence that one has in that particular measurement. A high accuracy can only be achieved when trueness is high, meaning that the systematic error (S) is small, and when the precision is high, meaning that the standard deviation, σ, of the random error is small.

The systematic error can include multiple components, for instance, the matrix variation effect, batch bias, laboratory bias and method bias (Thompson, 2000). For measurements taken in the same batch, we can assume that conditions are more similar. Bias in a single batch will systematically affect all measurements in that particular batch because conditions are the same, such as the analyst and the instrument. The same applies to a laboratory. Each laboratory applies the same method but is likely to have its own interpretation of the method protocol, leading to a laboratory bias. Whenever multiple methods are applied to measure a certain soil property, each method has its own bias. In the next section we developed these error sources more formally using a statistical modelling approach.

3 | METHODS

3.1 | Linear mixed-effects model

We assume that errors in analytical data result from varying measurement conditions between batches and laboratories but also from method bias, as explained in Sections 1 and 2. We can include identified error sources in a linear mixed-effects model. Such a model partitions the total variance in the dataset into components and determines the relative contribution of each variance component (Zuur, Ieno, Walker, Saveliev, & Smith, 2009). The linear mixed-effects model includes both fixed and random effects as predictor variables. Fixed effects are group specific (i.e., the error of each group or component), whereas random effects inform on the variation between groups. Essentially, the systematic and random errors from Equation (1) are broken down into multiple components. Here, we assume that variance in measurements is associated with true soil variation (characterized by sample ID, a fixed effect), and the batch, laboratory and sample preparation method (random effects). Residual variance that cannot be attributed to these components is regarded as the random error.

We included these fixed and random effects in the following linear mixed-effects model:

$$\begin{matrix} Y_{pqrst} = X_s + B_{pqr} + L_q + M_r + \varepsilon_{pqrst}, \; p = 1, \ldots, P; \\ q = 1, \ldots, Q; r = 1, \ldots, R; s = 1, \ldots, S; t = 1, \ldots, T \end{matrix}, \qquad (2)$$

where Y_{pqrst} is the t-th measurement of the s-th soil sample of the p-th batch, performed by the q-th laboratory while using the r-th preparation method. B_{pqr} represents the random effect of the p-th batch in the q-th laboratory using the r-th preparation method, L_q represents the random effect of the q-th laboratory, M_r represents the random effects of the r-th preparation method, X_s is the 'true' value of a soil sample (modelled as a fixed effect) and ε_{pqrst} is the random error. Based on the theory of linear mixed-effects models, we assume that B_{pqr}, L_q, M_r and ε_{pqrst} are uncorrelated and normally distributed with zero mean, and standard deviations σ_{batch}, $\sigma_{\text{laboratory}}$, σ_{method} and σ_{residual}, respectively.

3.2 | Parameter estimation

In this research, the linear mixed-effects model was used to model measurement error in wet chemistry soil data. In an ideal world, wet chemistry soil datasets would be balanced and include replicate measurements for all soil samples. Having a balanced dataset structure means that the number of observations per combination of factor levels is equal. In other words, multi-laboratory soil data are balanced when each laboratory measured the same soil samples the same number of times, and included an equal number of samples in an equal number of batches. However, data delivered by multiple laboratories are seldom balanced, as the number of analyses is highly dependent on available material, time and financial resources.

Furthermore, repeated measurements on soil samples are required to accurately quantify the random effects included in the model. The batch effect can only be estimated when the same soil sample is analysed in different batches. The same applies to the laboratory effect. Furthermore, sufficient replicate measurements per batch should be available to correctly estimate the residual variance. However, replicate measurements of all soil samples are costly and are often only collected for a small subset, for example, as part of the laboratory's quality control programme.

Therefore, the accuracy of model parameter estimates is dependent on the experimental measurement design, namely the number of laboratories and batches, and the number of replicates taken. A higher number of replicate measurements means that more information is available to accurately estimate the model parameters, in particular the residual variance. Whenever insufficient information is available, that is, in case of (near-)singularity, the model will become unstable and model parameter estimates will become highly uncertain, or cannot be estimated at all.

In the case studies of this research, restricted maximum likelihood (REML) was used to estimate the variance components of the random effects, as well as the fixed effects (Lark & Cullis, 2004; Webster & Oliver, 2007; Webster, Welham, Potts, & Oliver, 2006). For Gaussian models, REML is known to produce less biased parameter estimates when compared to maximum likelihood (ML) (Harrison et al., 2018).

3.3 | Software implementation

The linear mixed-effects model was fitted on synthetic data (Section 4) and real-world data (Section 5) in R Studio (R Core Team, 2017) using the *lmer* function from the lme4 package (Bates, Mächler, Bolker, & Walker, 2015). In the *lmer* syntax, the model was expressed through a formula including both the fixed and random effects. Here, we assumed that hierarchy was present in the grouping of the observations, by the grouping variables batch and laboratory. Every laboratory groups a number of unique batches, making batches nested within laboratory. All control arguments in the *lmer* function kept their default value. The R scripts of the synthetic case were made publicly available via GitLab (van Leeuwen, Mulder, Batjes, & Heuvelink, 2021).

4 | SYNTHETIC CASE STUDY

4.1 | Data

As explained in Section 3, we can distinguish between many different error sources, which result from varying measurement conditions between batches and laboratories. However, we would also like to quantify these errors. To do so, an experimental measurement design is required that allows reliable estimation of the variance associated with each error source. In this section, we used synthetic datasets to estimate these variances. Using synthetic data allowed us to compare various experimental measurement designs. Furthermore, because we were in control of generating the data, we knew the variances associated with all error sources. Therefore, we could evaluate how well they were estimated from the data.

Five synthetic datasets were generated, consisting of $n = 20, 50, 100, 200$ and 500 sample IDs. The datasets included synthetic 'true' $\mathrm{pH_{H_2O}}$ values for each sample ID, which were drawn from a normal distribution with a mean of 6.5 and a standard deviation of 1 pH unit. Each soil sample was 'measured' in duplicate, distributed over four batches. This set-up was repeated for three hypothetical laboratories. In the case of $n = 100$ sample IDs, this experimental measurement design resulted in $600\ (100 \times 3 \times 2)$ synthetic $\mathrm{pH_{H_2O}}$ measurements in total, distributed over 12 batches (Figure 2). The batch effect was taken to have a variance of 0.01 ($\sigma_{\mathrm{batch}} = 0.1$), whereas the laboratory effect had a variance of 0.0625 ($\sigma_{\mathrm{laboratory}} = 0.25$). The residual variance was taken as 0.04 ($\sigma_{\mathrm{residual}} = 0.2$). Two thousand repetitions were made for the five datasets to study the effect of sample size and experimental measurement design on the estimation of the variance components. The 'true' $\mathrm{pH_{H_2O}}$ of a soil sample, as well as the random effects, were simulated for each of these 2000 repetitions. The *lmer* function was applied to each of the 2000 repetitions, after which the estimated variance components were stored. We thus obtained 2000 estimates of each variance component and could compare these with the 'true' variances used to simulate the data. Note that we did not generate data with different preparation methods. Therefore, M_r was excluded from the model in Equation (2). All simulations from normal distributions were performed using the *rnorm* function in R.

To analyse the effect of missing data on model parameter estimation, 20, 50 and 80% of the data were randomly removed from the original datasets, using the *sample* function from the R Base package. The effect of various percentages of missing data was studied through comparing the interquartile ranges (IQR) of the model parameter distributions. We used the IQR instead of the

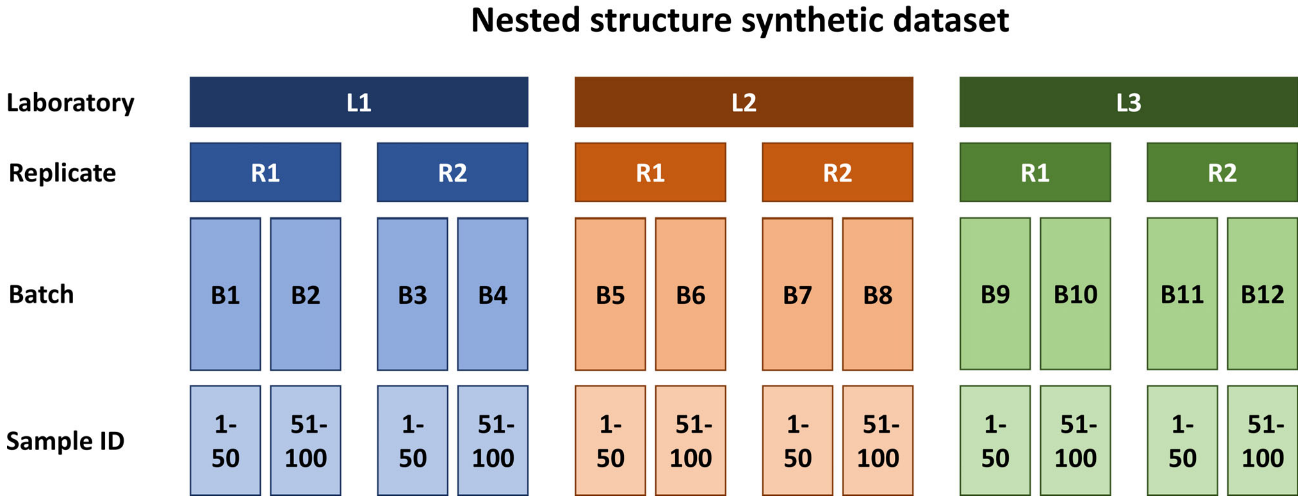


FIGURE 2 Nested structure of the synthetic dataset, here including three hypothetical laboratories, 12 batches and $n = 100$ sample IDs [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 1 Mean and interquartile range (IQR) of the batch, laboratory and residual variance estimates for different numbers of sample IDs.

No. sample IDs	Mean batch effect variance [−]	IQR batch effect variance [−]	Mean laboratory effect variance [−]	IQR laboratory effect variance [−]	Mean residual variance [−]	IQR residual variance [−]
20	0.010	0.0086	0.064	0.0736	0.040	0.0080
50	0.010	0.0077	0.065	0.0774	0.040	0.0051
100	0.010	0.0073	0.061	0.0701	0.040	0.0035
200	0.010	0.0067	0.062	0.0725	0.040	0.0024
500	0.010	0.0066	0.063	0.0702	0.040	0.0016

standard deviation to evaluate the spread of the model parameter estimates, because some of the distributions were skewed. In such a case, the IQR is easier to interpret.

4.2 | Results

4.2.1 | Balanced data

The linear mixed-effects model was fitted on the five different synthetic datasets (Table 1; Figure 3). For all sample sizes, the average of the estimated variances for the random effects was close to the 'true' values. This indicates that REML estimation of the variance components was unbiased. Some modest variation could be observed in the mean estimated laboratory effect variance, which ranged from 0.061 to 0.065. As expected, the IQR of the model parameter estimates decreased with increasing n, meaning that the estimates became more accurate as the size of the dataset increased (Table 1). However, this effect was limited for batch and laboratory effect, with a decrease in IQR of 23.3 and 4.6%, respectively, when increasing from $n = 20$ to $n = 500$. The IQR of the residual variance estimates dropped by 80% between $n = 20$ and $n = 500$. Furthermore, the IQRs of the batch and laboratory effect were almost equal to the variance estimates. For example, for $n = 100$, $\sigma^2{}_{\text{batch}} = 0.010_,$ whereas the corresponding IQR was 0.0086. In contrast, the IQR of the residual variance estimates was quite small compared to the residual variance estimate itself.

Figure 3 shows the density plots of the estimated model parameters. For batch and laboratory effect variance, the distributions are right skewed. Furthermore, the laboratory effect variance shows a large mass at zero, meaning that approximately 11% of the laboratory variances were estimated to be zero. Therefore, the laboratory effect has a mixed discrete-continuous distribution (Weld & Leemis, 2019), indicated by the discrete spike with solid disk in Figure 3. The probability density estimates for the residual variance were fairly symmetric. Their distribution became significantly narrower when

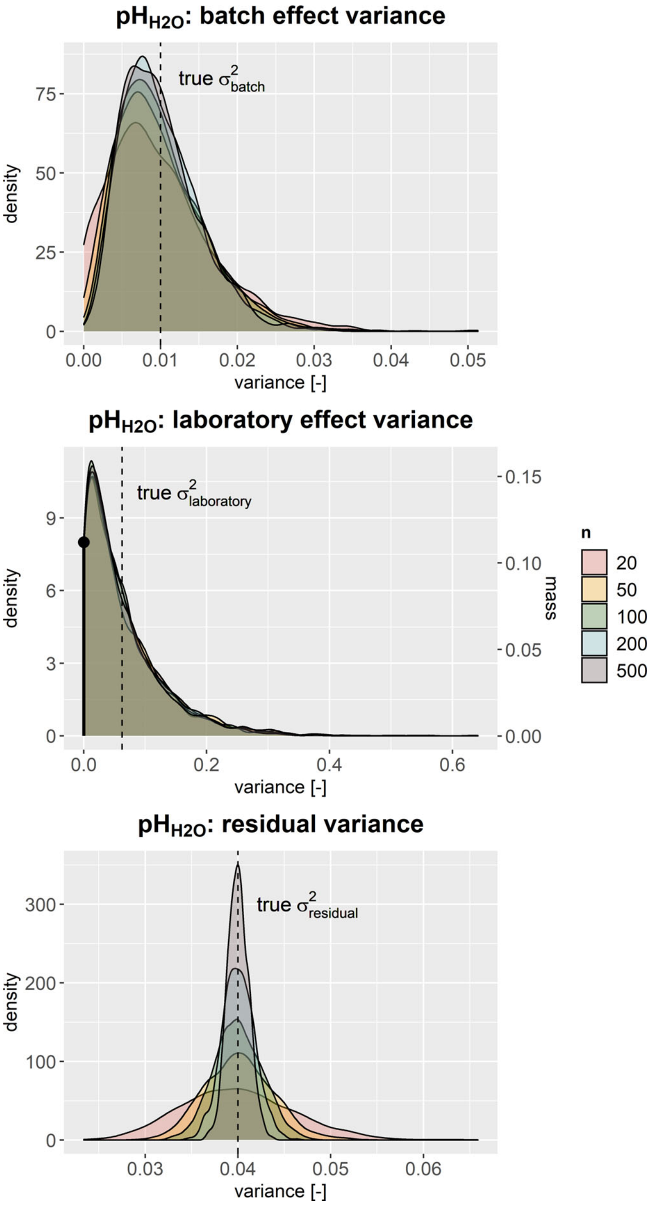


FIGURE 3 Density plots of the estimated batch, laboratory and residual variances for the five synthetic datasets, computed over 2,000 repetitions. The dashed line indicates the 'true' variance. As the laboratory effect variance had a mixed discrete-continuous distribution with a probability mass at zero, the discrete part is indicated by the black spike together with the probability mass represented by a solid disc. Note that the scales of the x- and y-axes differ [Color figure can be viewed at wileyonlinelibrary.com]

more sample IDs were included in the dataset. Correlations between estimates of individual random effects were close to zero (Appendix S1).

4.2.2 | Unbalanced data

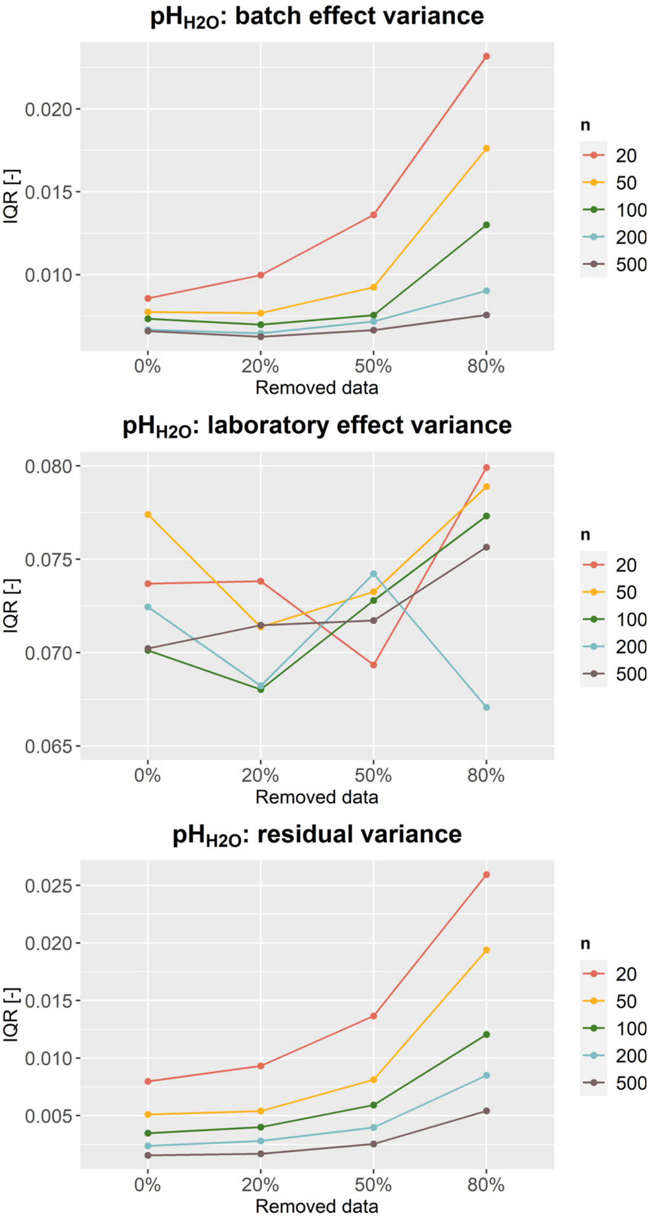
After removing 20, 50 and 80% of the data, the means of the model parameter estimates over all 2,000 repetitions were again similar to the 'true' parameter values (results not shown). The IQRs of the estimates for batch effect and residual variance clearly increased with a higher percentage of randomly removed data (Figure 4). The IQR of the laboratory effect variance behaved more randomly. In general, the IQRs of $n = 20$, 50, 100 and 500 increased when comparing 0 and 80% of removed data. However, the IQR values at 20 and 50% removed data behaved randomly. Furthermore, the IQR values based on the $n = 200$ dataset were highly fluctuating, showing no distinct relationship between the IQR and the percentage of randomly removed data.

Removing 80% of the data resulted in a larger IQR of the batch effect and residual variance for the $n = 20$ dataset compared to the $n = 500$ dataset. For example, the IQR of the batch effect variance of the $n = 20$ dataset increased by 171%, whereas the IQR of the $n = 500$ dataset increased by only 15%. The same was observed for the IQR of the residual variance. For both cases, the effect of missing data is largest for the smallest dataset, $n = 20$.

4.3 | Discussion

In this section, we explored the ability of the proposed model (Equation 2) to estimate the random effects, $\sigma^2{}_{batch}$, $\sigma^2{}_{laboratory}$ and $\sigma^2{}_{residual}$, while using different experimental measurement designs. We expected that the width of the distribution of the estimated variances for each of the random effects would decrease with increasing n, as more data were available. The distributions of the estimated $\sigma^2{}_{batch}$ and $\sigma^2{}_{laboratory}$ were both right skewed, whereas that of $\sigma^2{}_{residual}$ was symmetrical. The mean estimates of batch effect and residual variance were similar to the 'true' values, indicating unbiased model estimates for all n. The laboratory effect showed some variation in the mean variance estimates, ranging from 0.061 to 0.065, whereas the 'true' value was 0.0625. This minor variation around the 'true' value could be attributed to the finite number of repetitions (2,000) we performed. Increasing the number of repetitions, to for example 10,000, is likely to lead to more stable estimates.

The IQR of the estimated variances decreased with increasing n. However, this effect was only clear for the residual variance, where the IQR decreased by 80% when comparing $n = 20$ to $n = 500$. The large number of replicates available in the $n = 500$ dataset simply led to more precise residual variance estimates, shown by the small IQR. For batch and laboratory effect, the decrease in IQR when increasing n from 20 to 500 was less marked, with a reduction of only 23.3 and 4.6%, respectively. The limited decrease of the IQR could be explained by the


pH_{H2O}: batch effect variance

pH_{H2O}: laboratory effect variance

pH_{H2O}: residual variance

FIGURE 4 Interquartile range (IQR) of estimates for batch effect, laboratory effect and residual variance ($n = 20$, 50, 100, 200 and 500). For each n, 0, 20, 50 and 80% of the data were removed randomly, before fitting the linear mixed-effects model (2,000 repetitions) [Color figure can be viewed at wileyonlinelibrary.com]

available information to estimate the random effects. For batch effect, the datasets contained information of 12 batches, whereas for the laboratory effect, data from only three laboratories were included. In other words, increasing n did not increase the number of laboratories and batches. For example, for the $n = 500$ dataset, 500 (sample IDs) × 2 (replicates) × 3 (laboratories) yielded 3,000 measurements. However, from these data, groups of 1,000 measurements had the same laboratory error, meaning that overall little information was available to estimate the laboratory effect. As only limited laboratory information was available, the model parameter estimates were less precise.

After determining the random effects for balanced datasets, we explored how the model parameter estimates were influenced by different unbalanced experimental measurement designs. We expected that for each n, the IQR of the estimated variances for the random effects would increase with a higher percentage of missing data. Furthermore, we also expected that the IQR values would be lower and show a gentler increase in the case of a larger n. This effect could be explained by the fact that more observations remained for $n = 500$ after removing 80% of the data, compared to when the same percentage was removed from the $n = 200$ dataset. Our expectations were in line with the observed IQRs for the batch effect and residual variance, where the difference in IQR for 0 and 80% removed data was largest for $n = 20$ (171% increase in batch effect variance IQR). For comparison, the batch effect IQR for $n = 500$ increased by only 15% when 80% of the data were removed. The relation between IQR and the size of n was less distinct for the laboratory effect. Randomly removing data from all datasets led to highly fluctuating IQRs. These fluctuations are likely to be reduced by increasing the number of repetitions, for example, from 2,000 to 10,000. Thus, they are a consequence of the approximation errors caused by the numerical evaluation of IQRs. However, the size of the IQRs was, again, mainly influenced by the limited number of laboratories present in the datasets. Increasing the number of laboratories is likely to result in a smaller IQR of the laboratory effect, which would show an increase when removing data randomly.

The use of synthetic datasets to develop and test the model allowed us to study the influence of different experimental measurement designs on model parameter estimates. This application is especially interesting for laboratories that aim to quantify and reduce measurement error caused by the various error components. The model could be used by individual laboratories to determine the minimal number of replicate measurements to be included in the experimental measurement design. To illustrate this, consider a case with batch and laboratory effects removed, so that all variance in the data is captured by the residual part of the model. In this simplified case, results may also be obtained analytically (and perhaps also for more complex cases, but numerical approaches may be more attractive in such instances). Because we assumed normal distributions for the errors, the variance estimator has a chi-squared distribution (Webster & Oliver, 2007, Section 2.6.5), from which the IQR can be derived using the *qchisq* function in R (van Leeuwen et al., 2021).

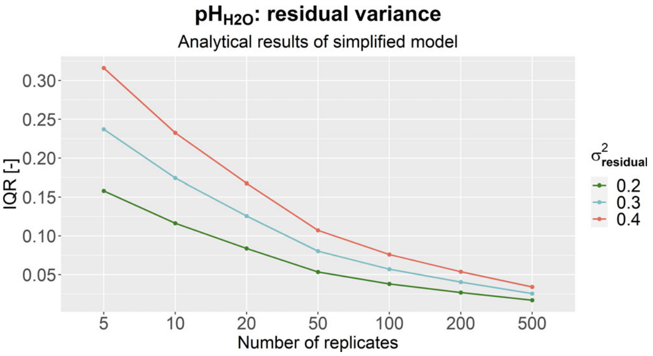


FIGURE 5 Interquartile ranges (IQRs) of residual variance estimated based on the number of available replicates (5, 10, 20, 50, 100, 200 and 500), in a case where residual error is the only source of uncertainty [Color figure can be viewed at wileyonlinelibrary.com]

Figure 5 shows the results of the analytical solution for synthetic pH_{H_2O} data with $\sigma_{\text{residual}} = 0.2$, 0.3 and 0.4. For each σ_{residual}, the IQR was determined for when 5, 10, 20, 50, 100, 200 and 500 replicates were available. The IQR decreased between 26 and 29% whenever the number of replicates was doubled. In other words, the IQR decreased approximately inversely proportional to the square root of the number of replicates included in the experimental measurement design.

In future research, the effect of different numbers of batches and laboratories could be investigated. Furthermore, in the current model (Equation 2), we assumed that only $\sigma_{\text{laboratory}}$ was systematically different for each laboratory. However, the batch and residual standard deviations may also vary between laboratories, meaning that each laboratory has its own σ_{batch} and σ_{residual}. Adding this effect would result in a more complex model with more parameters, which are more difficult to identify and hence lead to larger IQRs with the same experimental measurement design. Another effect that could be added to the model is a proportional residual variance. For instance, in soils with a high organic carbon content (e.g., forest soils) measurements tend to have a larger error. In this case, the residual variance is proportional dependent on the TOC content of the sample.

5 | REAL-WORLD CASE STUDY

5.1 | Data

For the real-world case study, measurements of soil samples from WEPAL were used. These data are part of the International Soil-analytical Exchange (ISE) programme (WEPAL, 2020). TOC measurements (dry combustion) were included in the analysis to determine if the model was also able to precisely estimate the variances of the random effects for a soil property different from pH_{H_2O}. TOC is considered to be more difficult to measure than pH_{H_2O} and its analysis is likely to yield more uncertain results (Bisutti, Hilke, & Raessler, 2004). Here, TOC is expressed in mass percentage (%).

Soil samples were selected based on their generic soil characteristics, aiming for a diverse range of materials to be included in the analysis (Table 2). Furthermore, we required that each soil sample was included in at least two rounds of the ISE programme. This condition ensured that multiple measurements of the same soil sample by the same laboratory were available.

The pH_{H_2O} and TOC data contained batch and laboratory information. Furthermore, in the proficiency testing scheme, each homogenized soil sample was sent to participating laboratories over multiple rounds. Therefore, a round effect might be present, which was added as a random effect in the linear mixed-effects model. We expected the round variance to be close to zero, as WEPAL ensures the homogeneity of soil samples over rounds. The original model (Equation 2) was rewritten into:

$$Y_{pqrst} = X_s + R_r + B_{pqr} + L_q + \varepsilon_{pqrst}, \tag{3}$$

where the preparation method effect, M, was excluded, and the round effect, R, was added. Y_{pqrst} is the t-th measurement of the s-th soil sample of the r-th round, measured in the p-th batch by the q-th laboratory. For pH_{H_2O}, the final dataset included measurements of all nine soil samples (see Table 2), which were analysed by 332 laboratories over a maximum of four rounds. The 6,749 measurement results were spread out over 6,380 batches, thus including 369 replicates. The TOC dataset contained only 1,756 observations that were analysed by 154 different laboratories. Here, only one batch contained replicate measurements. Consequently, this made quantifying the batch effect impossible, because the model could not distinguish between batch effect and residual variance. Therefore, batch effect was removed from the model (Equation 4) before fitting it to the TOC dataset:

$$Y_{qrst} = X_s + R_r + L_q + \varepsilon_{qrst}, \tag{4}$$

We also used the selected WEPAL data to test if the linear mixed-effects model (Equations 3 and 4) could accurately estimate all model parameters given the specific experimental measurement design used by WEPAL and the number of available replicates. First, the model was fitted on the real pH_{H_2O} and TOC measurements. Second, the resulting estimated variances of the random effects

TABLE 2 Description of selected certified International Soil-analytical Exchange (ISE) reference samples and summary statistics of pH_{H_2O} and total organic carbon (TOC) (dry combustion method) measurements

Sample ID	Generic soil characteristic	Sample location	No. laboratories	pH_{H_2O} Mean [−]	pH_{H_2O} *SD* [−]	pH_{H_2O} Var [−]	TOC Mean [%]	TOC *SD* [%]	TOC Var [$\%^2$]
860	Sediment	Kreekaksluizen (NL)	181	7.53	0.32	0.102	31.10	8.68	75.34
861	Calcareous clay	Logrono (ES)	171	7.74	0.32	0.102	17.31	5.91	34.93
863	Clay	Maren Kessel (NL)	165	5.97	0.28	0.078	35.18	5.97	35.64
866	Loess	Eijsden (NL)	156	5.96	0.34	0.116	10.09	2.71	7.34
867	Forest sandy soil	Unknown (NL)	196	4.00	0.38	0.144	51.08	10.15	103.02
872	Braunerde clay	Zurich (CH)	154	7.87	0.29	0.084	28.07	5.49	30.14
890	Sandy soil	Hengelo (NL)	174	5.48	0.34	0.116	18.12	3.83	14.67
921	River clay	Wageningen (NL)	158	7.48	0.34	0.116	35.14	5.36	28.73
995	Sandy soil	Droevendaal (NL)	176	6.92	0.26	0.068	26.09	4.48	20.07

Abbreviations: NL, The Netherlands; ES, Spain; CH, Switzerland

were considered as 'true' variances and were used to simulate synthetic values for the round, batch (in case of pH_{H_2O}) and laboratory effect, as well as for the residual part of the model. Third, after the synthetic WEPAL dataset was created, we fitted the model on these data to determine if model parameter estimates were similar to the 'true' variances. We repeated this procedure 500 times to assess the distributions of the model parameter estimates, similar to the approach used in Section 4.1.

5.2 | Results

5.2.1 | pH-in-water

The linear mixed-effects model was fitted on the WEPAL pH_{H_2O} measurements. The model estimated $\sigma^2{}_{\text{round}} = 0.0015$, $\sigma_{\text{batch}} = 0.047$, $\sigma^2{}_{\text{laboratory}} = 0.029$ and $\sigma^2{}_{\text{residual}} = 0.036$. The round effect variance was close to zero. Furthermore, the batch effect was greater than the laboratory effect. The estimated residual variance was smaller than the batch variance, but larger than the laboratory effect variance. The fitted values, based on the WEPAL pH_{H_2O} measurements, were used as input, or 'true', values for the random effects in the simulation study.

Figure 6 shows the distributions of the estimated model parameters, based on 500 repetitions. The variance estimates of all four parameters were fairly normally distributed and centred around the 'true' variances, which were based on the model estimates from the real WEPAL data. The IQRs of the estimated variances were 0.0006 (round effect), 0.0037 (batch effect), 0.0039 (laboratory effect) and 0.0035 (residual). For all random effects, the IQRs were relatively small, suggesting that the model was able to accurately estimate the random effects in WEPAL's pH_{H_2O} measurements.

5.2.2 | TOC

The linear mixed-effects model estimated $\sigma^2{}_{\text{round}} = 0.21\%^2$, $\sigma^2{}_{\text{laboratory}} = 7.71\%^2$ and $\sigma^2{}_{\text{residual}} = 32.90\%^2$. As for pH_{H_2O}, the round effect was negligibly small, but for TOC the residual variance was substantially larger than the laboratory effect variance. These 'true' values were again used in a simulation study (Figure 7). In 19.8% of the simulations, the round effect variance was estimated at exactly zero, indicating a mixed discrete-continuous distribution (Weld & Leemis, 2019). The IQRs of the estimated variances were $0.35\,\%^2$ (round effect), $1.80\,\%^2$ (laboratory effect) and $1.54\,\%^2$ (residual). These were relatively large compared to the estimated variances, especially for the round and laboratory effect.

5.3 | Discussion

In Section 5.2, the model was fitted to WEPAL's pH_{H_2O} (Equation 3) and TOC (Equation 4) data. For both soil properties, the model estimated the round effect variance at almost zero (pH: $\sigma^2{}_{\text{round}} = 0.0015$, TOC: $\sigma^2{}_{\text{round}} = 0.21\ \%^2$). These low values indicate that the variance as a result of heterogeneity of the sample material was minimal. In other words, the composition of the soil sample did not change over rounds, which was expected as WEPAL guarantees homogeneity of provided test material. Furthermore, for pH_{H_2O} the model estimated that the batch effect variance was larger than that of the laboratory effect ($\sigma^2{}_{\text{batch}} = 0.047$ and $\sigma^2{}_{\text{laboratory}} = 0.029$).

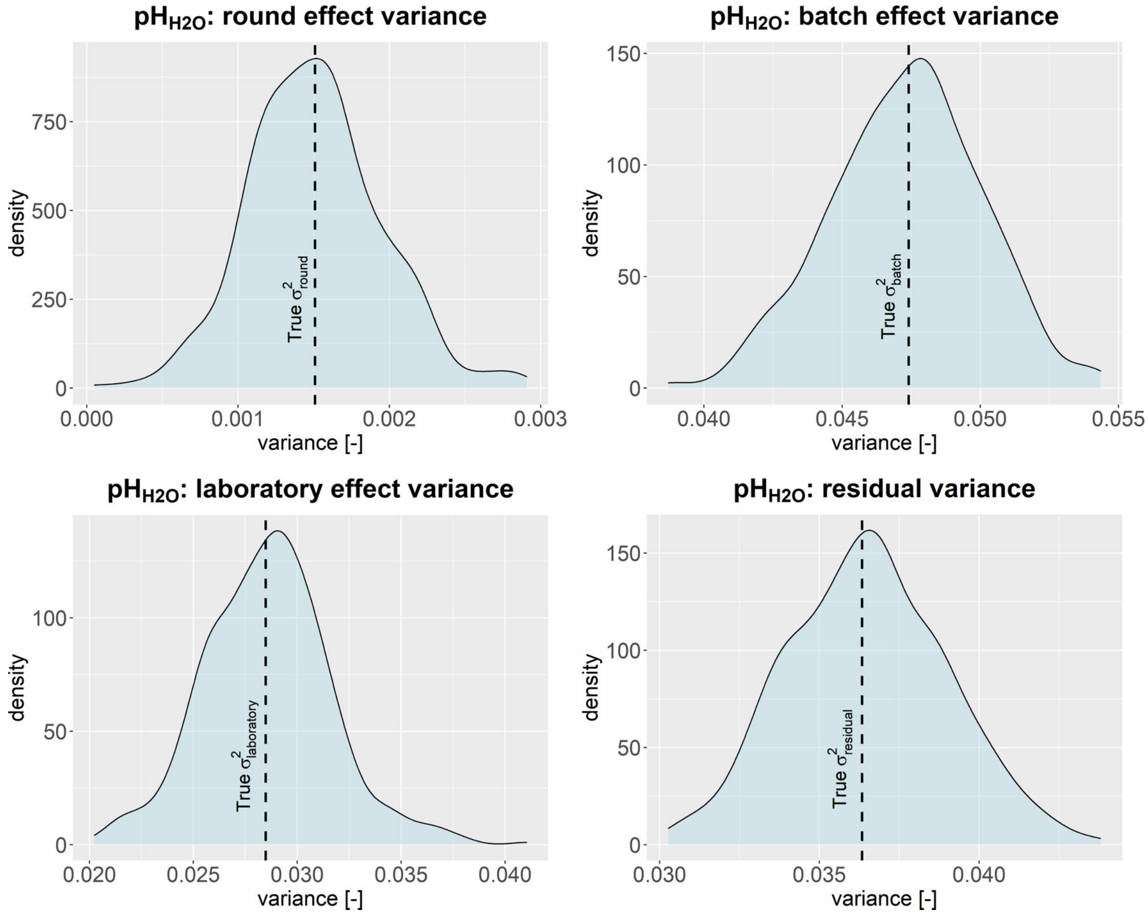
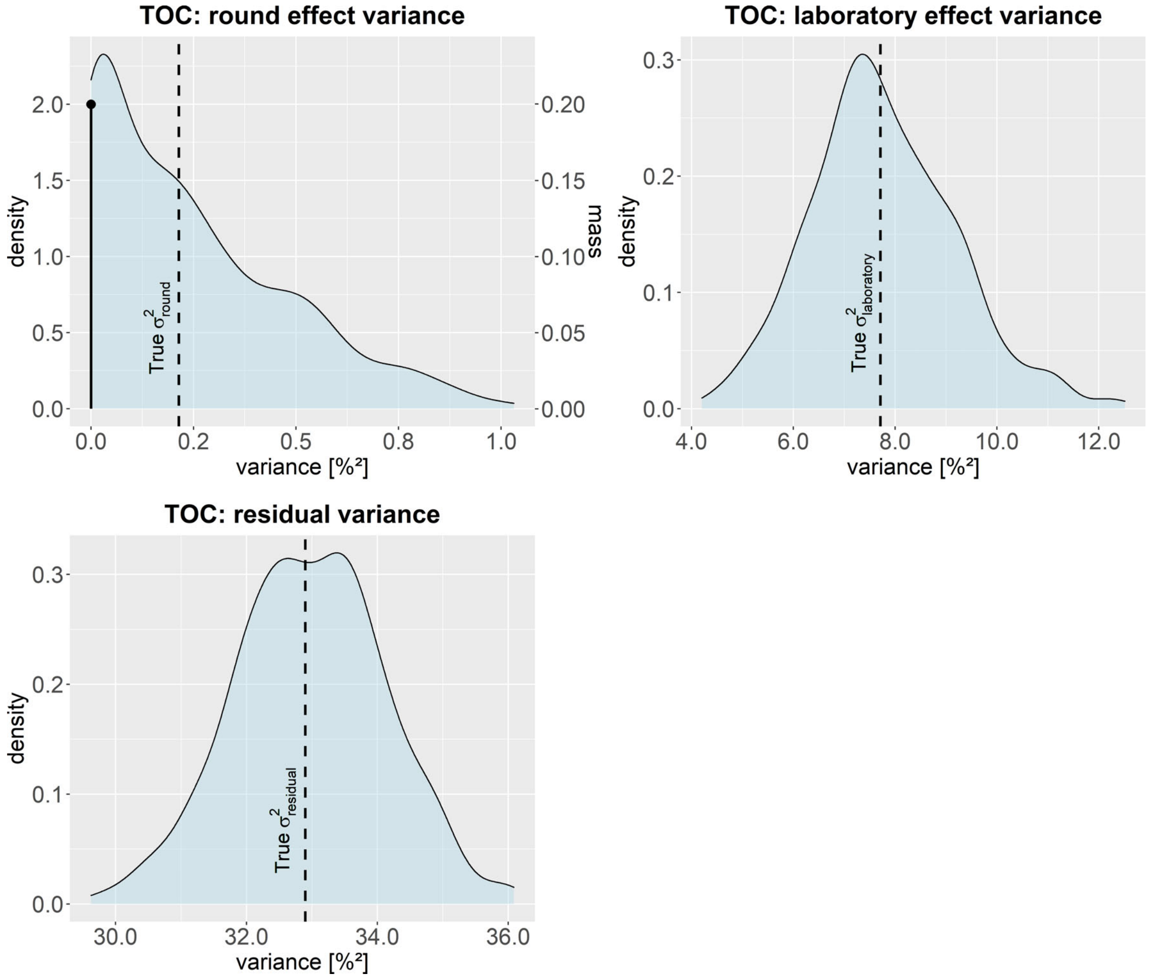


FIGURE 6 Density plots of model parameter estimates for the round and laboratory effects, and the residual variance, based on the synthetic WEPAL pH_{H_2O} dataset (based on 500 repetitions). The dashed vertical line indicates the input, or 'true', variance of the random effects. Note that the scales of the x- and y-axes differ [Color figure can be viewed at wileyonlinelibrary.com]

We expected that the batch effect would be smaller than the laboratory effect, because laboratories use their own specific procedures and within-laboratory variability is typically smaller than between-laboratory variability. Having a batch effect larger than the laboratory effect could be related to the structure of the testing scheme WEPAL applies; every soil sample is distributed over the participating laboratories between one and four times. However, not every laboratory returns all measurement results. In most cases, only a single measurement result was returned per round per laboratory. Only 369 out of 6,380 batches (5.8%) contained a replicate measurement. As a result, the model may have had difficulty in distinguishing between the batch effect and the residual part of the model, leading to a higher batch effect variance and thus an underestimated residual variance. The results from the simulation study supported this assumption. Here, batch effect and residual variance showed a strong negative correlation (−0.839), whereas correlations between the other random effects were close to zero (Figure 8). The TOC batch effect variance was captured within the residual variances, because the TOC dataset contained only one replicate measurement. Therefore, the batch effect was removed from the model. Correlations between random effect estimates were all close to zero (Appendix S1).

The model parameter estimates of the 'true' values were in agreement with the WEPAL data. For pH_{H_2O}, Table 2 showed a mean variance of 0.103 ($\sigma = 0.32$), whereas the model estimated the total variance at 0.114 ($\sigma = 0.34$). Similarly, the TOC data had a mean variance of $38.88\%^2$ ($\sigma = 6.24\%$), whereas the model estimated a total variance of $40.82\%^2$ ($\sigma = 6.39\%$). Table 2 shows that the standard deviation for TOC differed substantially between sample IDs. For example, sample 866 had a mean TOC of 10.09% and a standard deviation of 2.71%, whereas sample 867 had a mean TOC of 51.08% with a standard deviation of 10.15%. With higher mean TOC, the standard deviation increased, indicating a proportional effect. Therefore, in future research, this effect may


TOC: round effect variance

TOC: laboratory effect variance

TOC: residual variance

FIGURE 7 Density plots of model parameter estimates for the round and laboratory effects, and the residual variance, based on the simulated WEPAL TOC dataset (500 repetitions). As the round effect variance followed a mixed discrete-continuous distribution with a mass at zero, the discrete part is indicated by the black spike together with the probability mass represented by a solid disc. The dashed vertical line indicates the input, or 'true', variance of the random effects. Note that the scales of the x- and y-axes differ [Color figure can be viewed at wileyonlinelibrary.com]

be included by assuming that the residual standard deviation is proportional to the mean.

Results from the simulation study showed that the average model parameter estimates were close to the 'true' values. For pH_{H_2O}, the IQRs of the variance estimates were small, indicating that, despite the small number of replicate measurements (369 of 6,749 observations), the model was able to accurately estimate the random effects. In contrast, the relative IQRs for the model parameter estimates were larger for TOC, especially for the round and laboratory effect. This could be attributed to the larger laboratory bias in the TOC data compared to pH_{H_2O} data. Furthermore, the TOC dataset was much smaller than that of the pH_{H_2O} (1,756 and 6,749 observations, respectively), and included only a single replicate measurement, whereas the pH_{H_2O} dataset contained 369 replicates.

6 | GENERAL DISCUSSION

6.1 | Quantification of uncertainties in wet chemistry soil data

The quantification of uncertainties in wet chemistry data is of utmost importance, as errors in these data can propagate in further applications, such as pedotransfer functions or spectroscopic models (FAO, 2019b). Unfortunately, the majority of wet chemistry soil data is provided without uncertainty estimates, thereby limiting the scope for users to assess the "fitness for intended use". Most databases that are compilations of legacy data as well as spectrally derived data from various sources, do not provide quality indicators along with the data. In this study, we evaluated how measurement uncertainties can be estimated using replicates and a linear mixed-effects model approach.

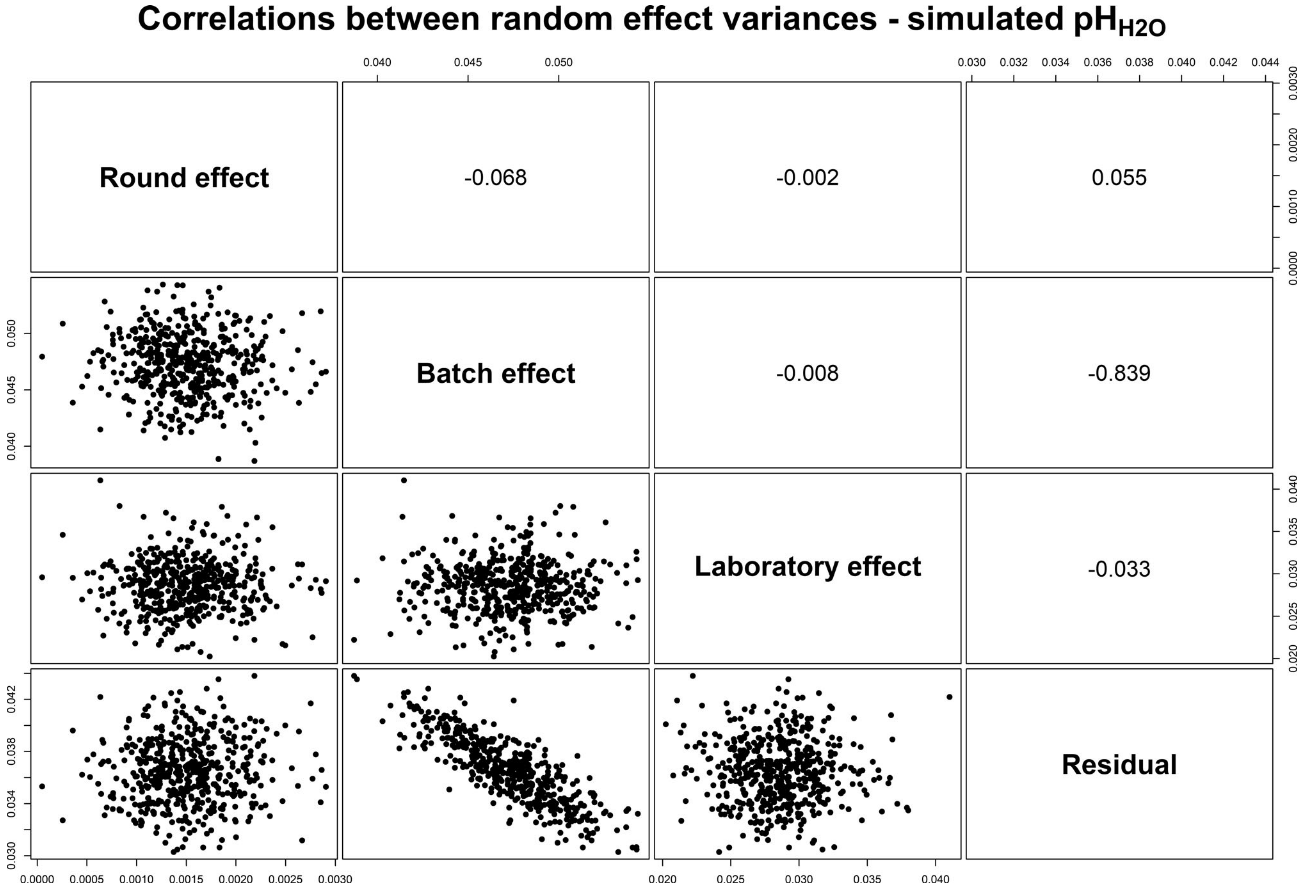


FIGURE 8 Scatter plots and Pearson correlations of the random effects variance estimates for pH_{H_2O}

We quantified uncertainties in pH and TOC measurements through a linear mixed-effects model approach. In other disciplines, such as ecology and medicine, (generalized) linear mixed-effects models are already commonly used to estimate variances from multiple sources in repeated measurements (Zuur et al., 2009). However, for wet chemistry soil data, the measurement error is often expressed by the mean squared error (MSE) or the root mean square error (RMSE) (Lagacherie et al., 2019; Libohova et al., 2019). These indicators are just summary measures, which do not provide information about the uncertainty associated with different components of the measurement procedure. Linear mixed-effects models allow determination of the contribution of each variance component to the total variance in the measurement. This information can help data providers decrease variance in their measurements by improving the limiting components in the laboratory analysis process itself.

Several studies also determined the size of the measurement error in basic soil properties, such as pH and TOC (e.g., Laslett & McBratney, 1990; Libohova et al., 2019; Pribyl, 2010). However, different approaches, alternative terminologies for the error and the use of different units made it challenging to compare our findings with those of previous studies. Furthermore, one should pay attention to what is included in the estimate of the measurement error. For example, Rawlins, Lister, and Mackenzie (2002) mention that, in practice, subsample and analytical variance cannot be separated, because the analytical variance can only be estimated by repeated analyses on the same uniform soil sample. However, as soil is never totally uniform, part of the analytical variance should actually be attributed to the subsampling error. In this research, we did not include subsampling error as a separate random effect. However, if this error source was present in the data, it was captured within the residual variance.

Libohova et al. (2019) assessed the size of errors in pH measurements, originating from the use of multiple wet chemistry methods, pedotransfer functions and spatial interpolation procedures, from different US databases. For pH_{H_2O}, Libohova et al. (2019) found a RMSE of 0.19 for the laboratory repeatability and 0.34 for the within-laboratory reproducibility. The between-laboratory reproducibility was estimated to have a RMSE of 0.50. Variance estimates were also provided for the laboratory repeatability and within-laboratory reproducibility; 0.15 and 0.06, respectively. Another study summarized analytical error variance estimates for a set of soil properties from literature (Viscarra Rossel & McBratney, 1998). A mean variance of 0.0295 for pH_{H_2O} was obtained. Both values are relatively small compared to the total variance estimated in the current research, where pH_{H_2O} had a mean total variance of 0.114 in the real-world case study. The data collected from the literature by Viscarra Rossel and McBratney (1998) originated from multiple laboratories, which all analysed the same soil sample. However, the study included only a small number of soil samples and variances were reported per soil sample. These examples from the literature show that comparison of measurement errors between studies is difficult, pointing to the need for a consistent procedure for quantifying the uncertainty in wet chemistry soil data.

6.2 | Implications for users of soil data

We applied the model to real-world soil data, provided by WEPAL. For pH_{H_2O} data, the overall measurement error was estimated at $\sigma = 0.34$, whereas $\sigma = 6.39\%$ was estimated from the TOC data. These measurement errors are relatively large, especially for TOC.

For data users, these findings are important to determine the "fitness for intended use" of the data. Negative effects can occur when using uncertain data to, for example, provide nutrient or fertilizer recommendations. For example, in the United States, applications of lime to acidic soils to raise soil pH are based on a recommendation scheme that does not take the uncertainty in measured pH into account, potentially leading to additions that are either too low or too high (Libohova et al., 2019). Target pH values in such nutrient application schemes are categorized per 0.2 pH units (Laboski, Peters, & Bundy, 2006). For these purposes, the estimated random effect standard deviations ($\sigma_{\text{batch}} = 0.22$, $\sigma_{\text{laboratory}} = 0.17$ and $\sigma_{\text{residual}} = 0.19$) could potentially lead to incorrect recommendations. TOC, a common measure for SOC (soil organic carbon), is often used to assess SOC stocks and carbon sequestration rates at multiple spatial scales (Francaviglia, Di Bene, Farina, Salvati, & Vicente-Vicente, 2019). For accurate estimates of carbon sequestration rates and SOC stock changes, precise and repeated measurements of SOC are required (Stockmann et al., 2013). Uncertainty of SOC measurements will lead to uncertainty of the SOC stock estimation. For the TOC data, the total standard deviation was estimated at 6.39% (mean TOC = 28.9%), which is a relative error of 22.1%. A relative error of 22.1% in a TOC measurement will thus lead to a relative error of 22.1% in the SOC stock estimate, besides the error that occurs from measurement error in bulk density and coarse fragments data.

6.3 | Recommendations for providers of soil data

We applied the model to real-world soil data, provided by WEPAL. Contrary to our expectations, the batch effect variance was larger than the laboratory effect variance for pH_{H_2O}. The model estimated $\sigma^2{}_{\text{batch}} = 0.047$ and $\sigma^2{}_{\text{laboratory}} = 0.029$. However, as explained in Section 5.3, the small number of batches with replicate measurements, and the strong negative correlation between the batch effect and residual variance, suggested that this experimental measurement design was not appropriate for accurately estimating the batch effect and separating it from the residual variance. Therefore, when quantifying the batch effect, more batches should contain replicate measurements.

The synthetic case study yielded insight on the influence of experimental measurement designs and number of replicates on the accuracy with which error source components could be estimated. The number of replicates influenced the accuracy of the estimate of the residual variance, as illustrated by the IQRs (Table 1). For $n = 20$, the IQR of the residual variance was 500% larger than the IQR for $n = 500$ (0.0080 and 0.0016, respectively). This change is not caused by the number of samples, but by the number of replicates, as was also demonstrated in Figure 5. The IQR was inversely proportional to the square root of the number of replicates $\left(\sqrt{500/20} = \sqrt{25} = 5\right)$. This effect was less distinct for the batch effect, where the IQR decreased from 0.0086 ($n = 20$) to 0.0066 ($n = 500$). The laboratory variance IQR remained more or less the same (0.0736 and 0.0702 for $n = 20$ and $n = 500$, respectively). As mentioned in Section 4.3, the IQR of both batch and laboratory effect variance did not decrease for larger n, as the total number of batches and laboratories remained the same. This is important to consider when using the model to assess variance in multi-laboratory datasets. The same applies to any other error source being included in the model. Whenever a limited number of batches and laboratories are included in the dataset, the model estimates for these random effects will become less accurate.

The analytical results from the simplified model demonstrated the importance of including sufficient replicates in laboratory analyses (Section 4.3). A limited number of replicate measurements led to imprecise model parameter estimates, regardless of the number of samples. The analytical tool can be used by laboratories to determine how many replicates should be included in their experimental measurement design. Of course, this decision is also based on the financial resources that are available for the analyses. Considering variance estimation before performing analyses, could greatly improve uncertainty quantification of wet chemistry measurements. These findings are relevant to initiatives such as GLOSOLAN (FAO, 2019a). The importance of having sufficient replicate measurements was also observed in the real-world case study. For pH_{H_2O}, only 5.8% of the batches included replicate measurements. The lack of sufficient replicates in each batch caused the model to have difficulties in distinguishing between the batch effect and residual variance.

Furthermore, as demonstrated in the synthetic case study with unbalanced datasets, the number of sample IDs included in the analyses indirectly affected model parameter estimates. Here, imprecise model parameter estimates were observed for the smaller datasets ($n = 20$ and $n = 50$) after 80% of the data was removed. In these smaller datasets, fewer replicates were available. When removing 80% of the existing data for $n = 20$, 96 of the total 120 measurements were removed. The remaining 24 observations contained too little information to accurately estimate the three random effects. In larger datasets more data remained available to estimate the model parameters after removing a large percentage of the data.

Additionally, as suggested by the literature review, when quantifying uncertainties in wet chemistry soil data, SOPs for measuring and reporting data are of utmost importance. Standardized methods will reduce σ_{batch}, $\sigma_{\text{laboratory}}$, σ_{method} and σ_{residual}. The positive effect of SOPs on the laboratory measurement error is illustrated by the results from the real-world case study. WEPAL applies SOPs for preparing homogeneous soil samples and subsequent storage of the material, leading to a small round effect variance, as presented in Section 5.2.

7 | CONCLUSION

We aimed to quantify uncertainties in synthetic and real-world wet chemistry soil data through a linear mixed-effects model approach. Our study showed that for balanced and unbalanced datasets, using data for three hypothetical laboratories (four batches per laboratory), the mean estimated variances of the random effects were in agreement with those for the respective random effects used to generate the synthetic datasets. In other words, there was no systematic bias in the model variance estimates.

The results from the synthetic case study also showed the importance of including sufficient batches and laboratories in the experimental measurement design when quantifying uncertainties in multi-laboratory data. Including data from only three hypothetical laboratories led to imprecise estimation of the laboratory random effect, which did not improve when more unique samples were added. A similar effect was observed for the batch effect variances, although less strong. To solve this problem, it may be better to represent the laboratory effect as a fixed effect instead of a random effect in case of data from only a few laboratories or batches. In contrast to the batch and laboratory variance estimates, the residual variance estimates did become more precise, as indicated by the decreasing IQRs. In the unbalanced case, the IQRs increased after various percentages of data were removed randomly. The effect was clearest for smaller datasets, where removing observations left only little information for the model to estimate its parameters.

The real-world case study using WEPAL data showed that the model could also accurately estimate the variances of the random effects using real unbalanced datasets. For pH_{H_2O}, the model estimated $\sigma^2{}_{\text{batch}} = 0.047$, $\sigma^2{}_{\text{laboratory}} = 0.029$ and $\sigma^2{}_{\text{residual}} = 0.036$. However, due to the small number of batches with replicate pH_{H_2O} measurements (5.8%), the model had difficulties in distinguishing between batch and residual variance. For TOC, only a single replicate measurement was available, leading to batch effect being removed from the model. The model estimated $\sigma^2{}_{\text{laboratory}} = 7.71\%^2$ and $\sigma^2{}_{\text{residual}} = 32.90\%^2$. For both pH_{H_2O} and TOC, the round effect variance was close to zero, indicating that WEPAL successfully distributes stable and homogeneous soil samples between rounds of the ISE Programme.

The results from the synthetic case also demonstrated the importance of having sufficient replicate measurements. Therefore, in laboratory performance comparisons, a greater number of batches should contain replicate measurements to successfully estimate the variance components. The analytical result of the simplified case with only residual error (Figure 5) showed that the IQR of the variance estimates decreased proportionally with the square root of the number of replicates included. Data providers can use these types of analysis to determine how many replicates should be included in their experimental measurement design, thus striking a balance between accurate measurement uncertainty quantification and financial resources.

This research demonstrated the importance of adequate experimental measurement design and sufficient replicate measurements in the quantification of uncertainties in wet-chemistry soil data. Furthermore, the results showed that the laboratory measurement error in soil data is quite large and should not be ignored by the users of the data, which, unfortunately, occurs regularly.

ACKNOWLEDGEMENTS

We thank W. van Vark from WEPAL for providing data from the ISE Programme. We would also like to thank the two anonymous reviewers for their constructive and insightful comments.

AUTHOR CONTRIBUTIONS

Cynthia van Leeuwen: Conceptualization; data curation; formal analysis; methodology; supervision; visualization; writing-original draft; writing-review & editing. **Gerard Heuvelink:** Conceptualization; formal analysis; methodology; supervision; writing-original draft; writing-review & editing. **Niels Batjes:** Conceptualization; methodology; supervision; writing-original draft; writing-review & editing. **Titia Mulder:** Conceptualization; methodology; supervision; writing-original draft; writing-review & editing.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request. Restrictions apply to the availability of the WEPAL data.

ORCID

Cynthia C. E. van Leeuwen https://orcid.org/0000-0003-3108-2136
Vera L. Mulder https://orcid.org/0000-0003-4936-0077
Niels H. Batjes https://orcid.org/0000-0003-2367-3067
Gerard B. M. Heuvelink https://orcid.org/0000-0003-0959-9358

REFERENCES

Allchin, D. (2001). Error types. *Perspectives on science*, *9*(1), 38–58. https://doi.org/10.1162/10636140152947786

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Batjes, N. H., Ribeiro, E., & van Oostrum, A. (2020). Standardised soil profile data to support global mapping and modelling (wosis snapshot 2019). *Earth System Science Data*, *12*, 299–320. https://doi.org/10.5194/essd-12-299-2020

Bisutti, I., Hilke, I., & Raessler, M. (2004). Determination of total organic carbon–an overview of current methods. *TrAC Trends in Analytical Chemistry*, *23*(10–11), 716–726. https://doi.org/10.1016/j.trac.2004.09.003

Dangal, S. R. S., Sanderman, J., Wills, S., & Ramirez-Lopez, L. (2019). Accurate and precise prediction of soil properties from a large mid-infrared spectral library. *Soil Systems*, *3*(1), 11. https://doi.org/10.3390/soilsystems3010011

Ebentier, D. L., Hanley, K. T., Cao, Y., Badgley, B. D., Boehm, A. B., Ervin, J. S., … Jay, J. A. (2013). Evaluation of the repeatability and reproducibility of a suite of qpcr-based microbial source tracking methods. *Water Research*, *47*(18), 6839–6848. https://doi.org/10.1016/j.watres.2013.01.060

Ellison, S. L. R., Barwick, V. J., & Farrant, T. J. D. (2009). *Practical statistics for the analytical scientist: A bench guide*. Cambridge: Royal Society of Chemistry. https://doi.org/10.1039/9781847559555

FAO. (2019a). Global soil laboratory network. Retrieved from http://www.fao.org/global-soil-partnership/glosolan/en/

FAO. (2019b). *Measuring and modelling soil carbon stocks and stock changes in livestock production systems: Guidelines for assessment (version 1)*. Rome, Italy: FAO.

Francaviglia, R., Di Bene, C., Farina, R., Salvati, L., & Vicente-Vicente, J. L. (2019). Assessing "4 per 1000" soil organic carbon storage rates under mediterranean climate: A comprehensive data analysis. *Mitigation and Adaptation Strategies for Global Change*, *24*(5), 795–818. https://doi.org/10.1007/s11027-018-9832-x

Goidts, E., Van Wesemael, B., & Crucifix, M. (2009). Magnitude and sources of uncertainties in soil organic carbon (soc) stock assessments at various scales. *European Journal of Soil Science*, *60*(5), 723–739. https://doi.org/10.1111/j.1365-2389.2009.01157.x

Guerrero, C., Viscarra Rossel, R. A., & Mouazen, A. M. (2010). Diffuse reflectance spectroscopy in soil science and land resource assessment. *Geoderma*, *158*, 1–2. https://doi.org/10.1016/j.geoderma.2010.05.008

Harrison, X. A., Donaldson, L., Correa-Cano, M. E., Evans, J., Fisher, D. N., Goodwin, C. E. D., … Inger, R. (2018). A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ*, *6*, e4794. https://doi.org/10.7717/peerj.4794

Heuvelink, G. B. M. (2018). Uncertainty and uncertainty propagation in soil mapping and modelling. *Pedometrics*, 439–461. Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-63439-5_14

Heuvelink, G. B. M., Brown, J. D., & van Loon, E. E. (2007). A probabilistic framework for representing and simulating uncertain environmental variables. *International Journal of Geographical Information Science*, *21*(5), 497–513. https://doi.org/10.1080/13658810601063951

Hibbert, D. B. (2007). Systematic errors in analytical measurement results. *Journal of Chromatography A*, *1158*(1–2), 25–32. https://doi.org/10.1016/j.chroma.2007.03.021

International Organization for Standardization (ISO). (1994). ISO 5725-1: 1994: Accuracy (trueness and precision) of measurement methods and results - Part 1: General principles and definitions. International Organization for Standardization. https://www.iso.org/obp/ui/#iso:std:iso:5725:-1:ed-1:v1:en

Laboski, C. A., Peters, J. B., & Bundy, L. G. (2006). *Nutrient application guidelines for field, vegetable, and fruit crops in Wisconsin*. Madison, Wisconsin: Division of Cooperative Extension of the University of Wisconsin-Extension.

Lagacherie, P., Arrouays, D., Bourennane, H., Gomez, C., Martin, M., & Saby, N. P. A. (2019). How far can the uncertainty on a digital soil map be known?: A numerical experiment using pseudo values of clay content obtained from Vis-swir hyperspectral imagery. *Geoderma*, *337*, 1320–1328. https://doi.org/10.1016/j.geoderma.2018.08.024

Lark, R. M., & Cullis, B. R. (2004). Model-based analysis using reml for inference from systematically sampled data on soil. *European Journal of Soil Science*, *55*(4), 799–813. https://doi.org/10.1111/j.1365-2389.2004.00637.x

Laslett, G. M., & McBratney, A. B. (1990). Estimation and implications of instrumental drift, random measurement error and nugget variance of soil attributes—A case study for soil ph. *Journal of Soil Science*, *41*(3), 451–471. https://doi.org/10.1111/j.1365-2389.1990.tb00079.x

Libohova, Z., Seybold, C., Adhikari, K., Wills, S., Beaudette, D., Peaslee, S., … Owens, P. R. (2019). The anatomy of uncertainty for soil ph measurements and predictions: Implications for modellers and practitioners. *European Journal of Soil Science*, *70*(1), 185–199. https://doi.org/10.1111/ejss.12770

Malone, B. P., McBratney, A. B., & Minasny, B. (2011). Empirical estimates of uncertainty for mapping continuous depth functions of soil attributes. *Geoderma*, *160*(3), 614–626. https://doi.org/10.1016/j.geoderma.2010.11.013

McBratney, A. B., Minasny, B., Cattle, S. R., & Vervoort, R. W. (2002). From pedotransfer functions to soil inference systems. *Geoderma*, *109*(1–2), 41–73. https://doi.org/10.1016/S0016-7061(02)00139-8

McBratney, A. B., Minasny, B., & Viscarra Rossel, R. A. (2006). Spectral soil analysis and inference systems: A powerful combination for solving the soil data crisis. *Geoderma*, *136*(1–2), 272–278. https://doi.org/10.1016/j.geoderma.2006.03.051

Menditto, A., Patriarca, M., & Magnusson, B. (2007). Understanding the meaning of accuracy, trueness and precision. *Accreditation and Quality Assurance*, *12*(1), 45–47. https://doi.org/10.1007/s00769-006-0191-z

Pribyl, D. W. (2010). A critical review of the conventional soc to som conversion factor. *Geoderma*, *156*(3–4), 75–83. https://doi.org/10.1016/j.geoderma.2010.02.003

R Core Team. (2017). R: A language and environment for statistical computing. Retrieved from https://www.r-project.org/

Ramsey, M. H. (1998). Sampling as a source of measurement uncertainty: Techniques for quantification and comparison with analytical sources. *Journal of Analytical Atomic Spectrometry*, *13*(2), 97–104 10.1039/A706815H

Rawlins, B., Lister, T., & Mackenzie, A. (2002). Trace-metal pollution of soils in northern England. *Environmental Geology*, *42*(6), 612–620. https://doi.org/10.1007/s00254-002-0564-5

Smith, P., Soussana, J.-F., Angers, D., Schipper, L., Chenu, C., Rasse, D. P., … Klumpp, K. (2020). How to measure, report and verify soil carbon change to realize the potential of soil carbon sequestration for atmospheric greenhouse gas removal. *Global Change Biology*, *26*(1), 219–241. https://doi.org/10.1111/gcb.14815

Stockmann, U., Adams, M. A., Crawford, J. W., Field, D. J., Henakaarchchi, N., Jenkins, M., … Jastrow, J. D. (2013). The knowns, known unknowns and unknowns of sequestration of soil organic carbon. *Agriculture, Ecosystems & Environment*, *164*, 80–99. https://doi.org/10.1016/j.agee.2012.10.001

Theodorsson, E., Magnusson, B., & Leito, I. (2014). Bias in clinical chemistry. *Bioanalysis*, *6*(21), 2855–2875. https://doi.org/10.4155/bio.14.249

Thompson, M. (2000). Towards a unified model of errors in analytical measurement. *The Analyst*, *125*, 2020–2025. https://doi.org/10.1039/B006376M

Van Ee, J. J., Blume, L. J., & Starks, T. H. (1990). *A rationale for the assessment of errors in the sampling of soils*. Las Vegas, Nevada: US Environmental Protection Agency, Environmental Monitoring Systems Laboratory.

van Leeuwen, C. C. E., Mulder, V. L., Batjes, N. H., & Heuvelink, G. B. M. (2021). Gitlab repository - statistical modelling of measurement error in wet chemistry soil data. Retrieved from https://git.wur.nl/cynthia.vanleeuwen/statistical-modelling-of-measurement-error-in-wet-chemistry-soil-data

Viscarra Rossel, R. A., Behrens, T., Ben-Dor, E., Brown, D. J., Demattê, J. A. M., Shepherd, K. D., … Ji, W. (2016). A global spectral library to characterize the world's soil. *Earth-Science Reviews*, *155*, 198–230. https://doi.org/10.1016/j.earscirev.2016.01.012

Viscarra Rossel, R. A., & McBratney, A. (1998). Soil chemical analytical accuracy and costs: Implications from precision agriculture. *Australian Journal of Experimental Agriculture*, *38*(7), 765–775. https://doi.org/10.1071/EA97158

Webster, R., & Oliver, M. A. (2007). *Geostatistics for environmental scientists*. Chichester: John Wiley & Sons. https://doi.org/10.1002/9780470517277

Webster, R., Welham, S. J., Potts, J. M., & Oliver, M. A. (2006). Estimating the spatial scales of regionalized variables by nested sampling, hierarchical analysis of variance and residual maximum likelihood. *Computers & Geosciences*, *32*(9), 1320–1333. https://doi.org/10.1016/j.cageo.2005.12.002

Weld, C., & Leemis, L. (2019). Mixed-type distribution plots. *Information Visualization*, *18*(3), 311–317. https://doi.org/10.1177/1473871618756584

WEPAL. (2020). International soil-analytical exchange programme - ise. Retrieved from https://www.wepal.nl/en/wepal/Home/Proficiency-tests/Soil/Proficiency-tests/ISE.htm

Zuur, A. F., Ieno, E. N., Walker, N., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R*. New York, NY: Springer Science & Business Media. https://doi.org/10.1007/978-0-387-87458-6

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: van Leeuwen, C. C. E., Mulder, V. L., Batjes, N. H., & Heuvelink, G. B. M. (2021). Statistical modelling of measurement error in wet chemistry soil data. *European Journal of Soil Science*, 1–17. https://doi.org/10.1111/ejss.13137