EGU21-1272
https://doi.org/10.5194/egusphere-egu21-1272
EGU General Assembly 2021

Statistical modelling of measurement error in wet chemistry soil data

Cynthia van Leeuwen[1,2], Titia Mulder[1], Niels Batjes[2], and Gerard Heuvelink[1,2]
[1]Soil Geography and Landscape Group, Wageningen University and Research, Wageningen, The Netherlands (cynthia.vanleeuwen@wur.nl)
[2]ISRIC - World Soil Information, Wageningen, The Netherlands

There is a growing demand for high quality soil data to model soil processes and map soil properties. However, wet chemistry measurements on soil properties are subjected to many error sources, such as the observer, the instrument and lack of standardised methodologies. Consequently, soil data are imperfect and uncertain because of these error sources. Uncertainties in measurements of fundamental soil properties can propagate through, e.g., pedotransfer functions, spectroscopic models and digital soil mapping algorithms. Therefore, it is important to provide detailed uncertainty information about soil measurements to potential data users. In practice, uncertainty estimates are rarely specified by providers of analytical soil data.

In this research, we aimed to quantify uncertainties in synthetic and real-world pH (1:1 soil-water suspension) and Total Organic Carbon (TOC) measurements. We assumed that uncertainty can be represented by a normal distribution. A linear mixed-effects model was applied to estimate the parameters of the normal distribution, i.e., mean and standard deviation, of both synthetic and real-world datasets. The model included 'sample ID' as a fixed effect, and 'batch' and 'laboratory' as random effects. The use of synthetic datasets allowed us to investigate how well the model parameters could be estimated given a specific experimental measurement design, whereas the real-world case served to explore if the parameter estimates were still accurate for such unbalanced datasets.

For a balanced dataset (n=20, n=100, n=200 and n=500), using synthetic pH data for three hypothetical laboratories (two batches per laboratory), the mean estimated standard deviations (σ) of the random effects were σ_{batch}=0.10, $\sigma_{laboratory}$=0.24 and $\sigma_{residual}$=0.2. These estimates were in agreement with the σ for the respective random effects used to generate the synthetic dataset, meaning that the model could accurately estimate the model parameters. Subsequently, changes were made to the experimental measurement design by randomly removing 20%, 50% and 80% of the data, resulting in unbalanced datasets. In general, the interquartile range (IQR) of σ for each random effect increased with a larger percentage of removed data. However, the increase in IQR was larger for n=20 compared to, e.g., n=200. When comparing 0% and 80% randomly removed data, the IQR for the batch effect increased with 60.3%. Conversely, for n=200 an increase of only 23.5% was observed.

Subsequently, the same model was fitted on real-world pH and TOC data, provided by the Wageningen Evaluating Programs for Analytical Laboratories (WEPAL). The unbalanced dataset structure was first reconstructed and filled with synthetically generated data, based on sample means and standard deviations derived from the measured data. The model was fitted on both datasets. For measured pH, the model yielded σ_{batch}=0.27, $\sigma_{laboratory}$=0.17 and $\sigma_{residual}$=0.10. The IQRs of the estimated σ from synthetic WEPAL data were 0.04 (batch), 0.06 (laboratory) and 0.02 (residual). The model fitted on the measured TOC data estimated σ_{batch}=5.3%, $\sigma_{laboratory}$=2.8% and $\sigma_{residual}$=2.1%. For the synthetic WEPAL data, IQRs of 1.3% (batch), 1.4% (laboratory) and 0.4% (residual) were determined for the estimated σ. These findings suggest that despite having a highly unbalanced dataset, realistic model parameter estimates can still be obtained.