Dear editor

We thank you and the reviewers for your time and valuable comments. Please find our replies displayed in blue color below each comment. A manuscript version with tracked changes is attached.

Kind regards
Mareike Ließ

Anonymous Referee #1

Review manuscript EGU SOIL; Error propagation in spectrometric functions of soil organic carbon The aim of the study was to evaluate visible and near infrared (VIS-NIR) reflectance spectroscopy to predict SOC and particularly the source of errors associated. The error propagation when assessing SOC with VIS-NIR has been overlooked in previous studies using that technique. This study is relevant because the precise monitoring of SOC with conventional methods is labour intensive and expensive. In my opinion: - The study is original and in line with EGU SOIL topics. - The manuscript is well organized and has a good study design. - This study is a sufficient contribution to knowledge to be worth publishing in EGU SOIL. - All the sections of the MS are complete, well written and interesting to read (the manuscript has apparently passed through a round of peer-reviewing). However, VIS-NIR model building process and its associated calculation of uncertainty is not directly in my line of expertise. I recommend to find other peer reviewers. Specific comment: Lines 293-294 ": : : Fehler! Verweisquelle : : : warden" this entence should be delete or translated in English. I have no other specific comment.
Reply: Adapted accordingly.

Anonymous Referee #2

This study addressed a very important issue in soil spectroscopy field. The data uncertainty is a major factor to affect calibration process. I have some general comments on this study:
1. In the introduction part, it needs more literature reviews on the data uncertainty effect on the modelling process, to enhance awareness of this issue, especially soil data uncertainty from lab chemistry analysis. Because so far not many people even noticed about it, it has been often ignored.
2. For the experimental design of this study, i suggest author should focus on effect of data uncertainty from lab chemistry analysis, how this uncertainty can effect calibration and validation. Because this is the most important part of the issue.
Reply: We have included more references on uncertainty and emphasized the various sources of uncertainty and their impact on model performance.

Consequently, authors only mentioned about this effect a bit in results and discussion part. It would be a very important study in soil spectroscopy community if author could deeply discuss this effect. Because most research project does not allow all soil samples to measure 3 replicate, due to budget limitation. Therefore, the dataset in this study is a treasure.
3. I have attached a small but very good discussion note from NIR news, hope authors can get some ideas about this issue (this note is not reviewer's publication).
Reply: Thank you. We have included a reference to the discussion news.

Anonymous Referee #3
General comments

The study by Ellinger et al. represents a under-researched topic with high practical relevance for soil spectroscopy applications. An exhaustive set of uncertainty factors contributing on the model error was examined, using a full factorial design. The chosen title is attractive and is well aligned the objective and the performed statistical analyses. Statistical model development and assessment was done using a state-of-the-art cross-validation technique. The assessment of uncertainty comprises a combination of two sampling strategies including their combination, four spectral preprocessing methods, and spectral and analytical data set combinations with different degree of averaging random noise. This discussion article is well structured and describes the technical aspects in precise and easy understandable plain language. Although covering rather fundamental spectroscopy modeling research, the spectroscopic calibration of samples from a long fertilization trial embeds the uncertainty analysis into a realistic and interesting application context (soil monitoring). This study backs up the current practice of measuring several sample or sub-sample replicate and subsequent spectral averaging, which can improve model performance in many cases.
Reply: Thank you.

Further, the data set and the error analysis is of particular interest for the soil science community because it quantifies the contribution of random analytical reference method errors on spectral predictions. This paper would comprise an even more valuable scientific contribution if the authors elaborated more on this particular uncertainty component.
Reply: We have rewritten large sections of the introduction and results section.

The inclusion of the sampling design as a factor together with the other varying uncertainty components requests more concentration from the reader to understand the experiment and read the results. In order to simplify the results and to make the message more concise, the authors could focus on the scenario with the combined “A” and “B” sampling strategies, and move the results of “A” and “B” alone to the appendix.
Reply: We understand the reviewer’s concern about the complexity of the paper. However, as the sampling design is an important aspect of this paper. We, therefore, refrain from shifting the corresponding results to the appendix. The effect of the sampling design on predictive uncertainty was quantified.

The model tuning (number of PLSR components) results would be important for the interpretation of model performance under the uncertainty scenarios, but these are missing. These together with a more detailed discussion of error patterns not following the general trend or expectations would help to validate and explain the highly variable results. Further, light scattering effects as important source of spectral model error, and their dependence on soil composition and texture could be discussed, taking into account the applied spectral preprocessing methods. Addressing the specific comments, this paper will be a scientifically sound and valuable soil spectroscopy case study.
Reply: The PLSR components vary largely in dependence on the pre-processing method. The information on the number of selected components, therefore, did not result very informative. Furthermore, it distracts the reader from the main message. We, therefore, refrain from including it in this publication. We have adapted the results section concerning the uncertainty scenarios. We have also extended the introduction section to elaborate on the various sources of uncertainty.

Specific comments

The suggested references are listed in the bottom of this review.

Abstract

Instead of giving a general description of model building, more specific details on the error propagation experiment would show the importance and the quality of the carefully conducted statistical experiment. The abstract lacks quantitative results about the contribution of tested uncertainty factors on model performance.
Reply: We have rewritten large sections of the introduction and results section to go more into detail on the error propagation experiment. Quantitative results with respect to the contribution of uncertainty factors on model performance were added to the abstract.

Introduction

The introduction reads fluently and is of adequate length. The importance of soil organic carbon for soil functioning and frequent measurements provides a good motivation to introduce the methodological relevance of the present study. The reader of this journal is likely already familiar with the role and importance of soil organic carbon an its assessment in soils; therefore, this part can be condensed. There is some statements that can be assumed common knowledge. As an example, the following sentence in l. 42–44 requires no references: "The precise monitoring of SOC on a LTFE with conventional lab analysis is labour-intensive and expensive (Adamchuk and Viscarra Rossel, 2010; Loum et al., 2016) as it requires the analysis of a rather high amount of samples." The authors clearly state the motivation of using soil spectroscopy. Shortening the paragraph on soil organic carbon, there is space to briefly explain principles that enable sensing of carbon by infrared spectroscopy (e.g. relationship between functional groups in soil constituents, absorption of electromagnetic radiation and vibration of bonds in the infrared range, and soil properties, and how these relationships were used for statistical modeling).
Reply: The section on spectral sensing was moved from the methods section to the introduction. However, we refrain from shortening the section on SOC as it is important to understand the context of the Vis-NIR application: Soil monitoring.

The theoretical foundation of uncertainty in modeling is rather sparse and short in relation to the general introduction into concepts of soil organic carbon and advantages of spectroscopy (see also comment of anonymous referee No. 2). The authors are advised to add more key concepts and terminology of uncertainty in the context of predictive modeling generally, and soil spectroscopy modeling applications specifically. It is worth mentioning that measurement and prediction errors can be separated into systematic (bias) and random errors (uncertainty, precision). There are several sources of uncertainty in the model predictions, such as predictor (spectra) measurement errors, response (chemical laboratory measurements) errors and model errors (related to model instability in model parameter estimates or model structure). For analytical data, random errors are most relevant. Further, both bias and variance contribute to the uncertainty in generalization error estimate (here RMSE), which is another source of uncertainty.
Reply: We have extended the section on the various sources of uncertainty in the introduction. However, as we do not distinguish between systematic and random errors in this study, we refrain from referring to them in this explicit way.

The authors mention and stress importance of resampling strategy, which was comprehensibly explained. To complement, a link to the conditional model error could be made (for example citing Beleites et al. (2005)).
Reply: We have included a reference to Beleites et al. and Molinaro et al. in the context of resampling.

Some more details that are more relevant for the authors' experiment can also be added to the existing section in Material and Methods. For diffuse reflection infrared spectroscopy, scattering effects are particularly important spectral noise factors, that are relevant under the chosen experimental setup and the objective. Therefore, they merit particular mention.
Reply: We have added two references concerning sensor noise and scattering in the introduction section (Schwartz et al., Pilorget et al)

For a general description of model uncertainty analysis, e.g. Jansen and Michiel (1998) provide a good reference.
Reply: Thank you. We added the reference.

l. 55: Wetterlind et al. (2013) recommend general strategies for spectral measurement and modeling; Spectral averaging is also recommend. Therefore, this would be an important message and reference to add to the existing ones. Further, Wetterlind et al. (2013) highlight the effect of the sampled area and advise to perform replicate spectral sampling for small areas.
Reply: We refer to spectral averaging and repeated scans in the introduction and discussion section. We refer to the standard measurement protocol suggested by Pimstein et al. and, therefore, refrain from citing Wetterlind et al. in this context.

Material and Methods

The model of the CN analyzer is not described.
Reply: added

l. 104f: "C measurements were taken as organic carbon due to negligibly small carbonate contents." The authors are asked to provide quantitative statement (lower than xxx % C.).
Reply: Carbonate contents were below detection limit.

There is no mention of whether or how many scanning replicate measurements were internally averaged (spectrometer setting) prior taking spectrometer readings of the sub-sample and rotation replicate spectra (further noise averaging).
What is the name or composition of the material was used as a white reference (name manufacturer if composition not known)?
Reply: The missing information was included

The description of the preprocessing techniques is a bit too long and detailed. To keep a focus on the main topic, a brief description in one or two sentences, a citation to the original publication and maybe some soil spectroscopy case study where the respective techniques was applied, suffice. Spectral preprocessing techniques have been well described and researched in chemometrics, and applied in various other scientific discipines and industry over the last decades.
Reply: As we also assess the impact of the pre-processing method on the error measure, we think a detailed description is necessary for understandability.

The chosen resampling strategy is particulary well suited when data is scarce and variability is high, but can be recommended in general. This study can serve as an exemplary resampling setup for the soil spectroscopy community (repeated nested group k-fold cross-validation for model parameter tuning and evaluation). The approach is also particulary consisely described. Repeated 10-fold cross-valiation is a

practical and widely used cross-validation strategy to reduce uncertainty in performance estimators.Therefore, adding another reference to earlier applied predictive modeling literature that study effects of resampling strategy on the estimation of model evaluation metrics is advised. For example, Molinaro et al. (2008) is a suitable reference. The description and the illustration in Fig. 6 refer to a nested or double cross-validation, where the inner resampling layer comprises parameter tuning and the outer layer is used for model evaluation, which is used to avoid selection bias in parameter tuning. The authors should mention the nested or double cross validation terminology and also reference key literature. Varma et al. (2006) is suggested as a reference here.
Reply: We specified the applied cross validation as nested approach and added Varma and Simon as well as Molinaro et al.

“Although the samples were sieved and homogenised before taking measurements, within-sample variability still had an influence on the model outcome. In general, the variability of the samples depends on the soil treatment and possibly also on the origin of the samples (e.g. agricultural or forest soils).” The influence of scattering effects and major importance of texture should be mentioned and backed up with literature in this place.
Reply: Reference to soil treatment and scattering effects was made in the introduction. We refrain from referring to soil texture as we are at within-field scale and do not have a pronounced textural variability in our dataset. A reference to sample origin is included in the discussion section.

Results and Discussion

General considerations on the reviewer’s comments: Some of the below comments may cast a rather sceptical view on the authors’ explanation of the results. However, the experimental conditions (in a statistical sense) are not necessary such that there is major weak points of the results. The authors are kindly asked to provide a response or some more results or explanations as outlined below (appendix). The intention is to challenge the author’s hypotheses. There are unfortunately no means to compare these results to other similar studies in the soil spectroscopy literature. Nevertheless, the spectroscopy community has been following the general recommendation to average noise from replicate measurements to obtain good performing and robust models. This study showcases some "surprising effects". There are no consistent patterns that clearly show beneficial effects of averaging and removing spectra under all conditions. Some inconsistent patterns many arise from interaction in contributing factors.
Reply: We have adapted the results and discussion section and also refer to these unexpected effects.

A repeated nested 10-fold cross-validation guarantees that parameter selection is unbiased, but is prone to overfitting if sample grouping is only respected in assessment and not in tuning layer, in combination of multiple spectral and/or analytical replicate data set that are used for modeling.
Reply: We have added a corresponding sentence in the conclusions: “We are aware that the consideration of stratified group CV only for model evaluation but not for tuning might impair model performance as suboptimal model parameters might be selected. This will be adapted in future studies.”

How many PLSR components were used in the final models? What was/were the frequency/frequencies of respective best number of components across the folds and repeats?
Reply: The PLSR components vary largely in dependence on the pre-processing method. The information on the number of selected components, therefore, did not result very informative. Furthermore, it distracts the reader from the main message. We, therefore, refrain from including it in this publication.

For Savitzky-Golay (method 1) and the Norris gap derivative (method 4), the RMSE in model set 101 is considerably higher than in model set 100 as well as all other sets, whereby 3 times 6 identical spectra instead of 3 times an identical average spectrum are used (Fig. 8). Could it be that this is the result of a deleterious preprocessing effect due to missing sensor jump correction (see Fig. 3)? Noise enhancement may occur under certain preprocessing strategies such as the gap derivative (see also comments in next paragraph for further interaction possibilities).
Reply: We have corrected for the sensor jump and rerun the models. Figures 8 and 9 were adapted accordingly. Still models build on dataset$_{101}$ in some cases perform worse than models build on datsets$_{100}$. We refer to this peculiarity but have no explanation. As nested repeated 5-fold CV resulted in similar model results, we have replaced 10-fold CV by 5-fold to save computation time, the same applies for the number of components tested, which was reduced to 30.

Fig. 5 shows that the spectral variation largely manifests as offset variation. Savitzky-Golay smoothing with no derivative does not remove offset errors. This is worth a discussion. Assuming identical resampling sets among preprocessing methods within a data set, there seems to be a systematic error component in the spectra in data set A for these cases. The authors are encouraged to recompute results after correcting for sensor offsets, or they should at least consider possible explanations for clearly poorer model performance in the discussion.
Reply: We have corrected for the sensor jump and rerun the models.

How do the authors explain such a drastic error increase in the data situation from 3 times 6 different averaged spectra to 3 times one final average spectrum per sample when using 3 analytical replicates? Why does this occur only when adding replicate analytical measurements to the modeling process, but not when averaged analytical measurements are used?
Reply: We have added this aspect in the results section: “It is not surprising that the dataset of 3 SOC replicate measurements with 1 averaged spectrum (Dataset$_{100}$) results in lowest model performance, as the within-sample variance concerning SOC cannot be explained by the contained predictor information, the input data uncertainty propagates through the model building process.”

The nested cross-validation can yield different optimal number of component for each fold, which can be justified such that different data situations comprise different optimal model parameters with respect to performance. The nested or double cross-validation scheme in Fig. 6. shows the model validation procedure in the outer cross-validation loop within the left partition set and the model tuning resampling procedure within the right partition set. The authors mention that k-fold cross-validation was done by assigning the entire set of replicate spectra of respective samples in either fitting or hold-out sets. The authors don’t explicitly state whether this grouping by sample is done for the both model evaluation k-fold cross-validation layer and model tuning cross-validation layer, or just in either one of them. This lets room for ambiguous interpretations of the experiment and confounding factors. The R function ‘tuneControl‘ used for the tuning resampling, but the ‘groupKFold()‘ function from the caret RÂ˘apackage – which splits data based on a grouping factor – is not mentioned. Namely, focusing on the results depicted in panel "c1"Â˘aof Fig. 8, the prediction error decreases when less sub-sample replicate spectra are included in the modeling. Assuming that simple 10-fold cross-validation without grouping, the tuning layer would suffer from a data leakage from fitting folds into the respective assessment folds when replicate spectra are present. Data leakage provides biased tuning results and can select very high number of components. The maximum number of PLSRÂ˘acomponents was set 40, which represents severe over-fit with 100 rows and so many predictors (“large p small n"" problem). The reader cannot interpret if there was variability of the number of finally selected components for the different resampling sets. The authors should therefore provide a ncomp final tuning results table in the appendix if they were variable among

the sampling sampled data set, model set and preprocessing method combinations. In worst case scenario where the outer assessment resampling is grouped, but the inner tuning resampling lacked sample grouping, presence of more replicate spectral measurements per sample can yield too high ncomp and too adaptive models, This would give an alternative explanation for poor performance in the outer assessment when all replicate spectra are used for modelling, in comparison to model set011 and set001 with increasing degree of spectral averaging and less likely too adaptive PLSRÂ˘amodels.
Reply: Only the outer CV cycle of the nested repeated k-fold CV approach considers stratified group CV (but: spectral replicate measurements and scans per sample were always assigned to the same fold!). This was specified in the corresponding methods section. Of course, considering stratified group CV also for model tuning might improve model performance. As we used the inbuilt function of the Caret package, this was not possible. We are not aware of any publication that actually implemented stratified group CV in parameter tuning. We will, however, implement this in future studies as mentioned in the conclusion section. The PLSR components vary largely in dependence on the pre-processing method. The information on the number of selected components, therefore, did not result very informative. Furthermore, it distracts the reader from the main message. We, therefore, refrain from including it in this publication.

Using the full set of analytical carbon measurement replicates can have generally lower performance. Directly addressing these aspects is an outstanding achivement of this study. The authors are highly encouraged to extend the discussion about effects of analytical uncertainty, taking into account the considerations of anonymous referee No. 2..
Reply: We have further elaborated on this aspect.

To sum up, the authors are kindly asked elaborate and comment on interactive effects between the sources of errors and the resampling, thereby also critically address confounding effects as hypothesized above in the author's response to the review. Did the authors implement the grouped k-fold procecedure for both inner an outer cross-validation layers?
The authors could move Fig. 3 to the results and further illustrate differences in preprocessed spectra for replicate spectral measurements (e.g. one example spectrum).
Further, texture might explain different model performances for data sets "A"Â˘aand "B". Sandy soils are known to confound increases in reflectance typically also found for increases in soil organic C (see e.g. Stevens et al., 2013).
Reply: These aspects were addressed in our replies to the specific comments.

The use of different preprocessing techniques has already been exhaustively discussed in the soil spectroscopy literature. The performance of different preprocessing methods depends on the data context, as stated in the last sentence of paragraph l. 347–349. Thus, such a comparison to other studies does not make sense and brings no added value to the reader. The authors are advised to remove the comparison. To mention that the effects of signal processing methods and associated parameters on model performance are study and data specific is sufficient (incl. references).
Reply: We are not aware of any study that actually quantified the effect of spectral pre-processing on model performance and, therefore, refrain from deleting it from our study.

Conclusion

"Autocorrelation between calibration and validation sets". It is valuable for the soil spectroscopy community that the present study considers and stresses data grouping effects in resampling, here repeated measures, and accounts for those in the crossvalidation procedure. Ignoring such grouping factors can result in over-optimistic estimation of the model performance due to leakage of predictive

relationships from the modeling into assessment sets. The scientific community would benefit if authors could name the strategy using the terminology "group k-fold cross-validation".
Reply: The aspect was emphasized in the conclusion section.

Technical corrections

l. 86–89: “Categorical and continuous data first entered a factor analysis with mixed data (FAMD) performed with R package FactoMineR (Lê et al., 2008) to allow for further joint analysis. For design ’A’ the LTFE plots were then grouped by a k-means cluster analysis. R package NbClust (Charrad et al., 2014) automatically determines the optimal number of clusters making use of 30 indices.” How many factors from previous factor analysis with mixed data are retained and used for k-means clustering? How much variance do these factors and continous variables explain.
Reply: The FAMD was applied to decorrelate the data. All factors were retained.

l. 127–132: The outlier analysis is in section 2.4, but is not considered as preprocessing.
This part needs either a separate section or a generic data analysis section.
Reply: We refrain from including an additional section for only two sentences. As the outlier removal is the first step before applying any spectral pre-processing, we do not see why it should not stay in this section.

l. 263f: “The violin plots of all three soil sample sets do not resembles the archive violin plot very much.” typo: "resemble"
Reply: corrected

l. 344: “The simple first derivative (d1) performs poorely because it increases noise as stated, but there is no tendency to overfit related to this particular preprocessing technique (rather consequence of improper model resampling and tuning setup in combination with adaptive models)”; typo: “poorly” .
Reply: corrected

l. 365f: “The impact of the different pre-processing methods showed clearly in this study, but it may be different when using other data sets, though.” This sentence needs to be more precise; the authors are kindly advised to use "impact the different preprocessing methods on model performance..." or mention variable performance results or similar.
Reply: The section was rewritten.

l. 239f: “In this study, the partition of the spectral data into 10 folds for 10-fold CV had to be done very carefully, as in some cases multiple spectra existed for one sample.‘
Expression ”very carefully“ needs to be rephrased in scientific language, as well as “some cases” (2/3 of the cases).
Reply: adapted

Figures

General comment for panel figures: Experimental labels are hidden within the plot area and are better placed on top left of each subplot to facilitate the reader’s visual perception.
Reply: adapted

Fig. 3: The spectra have not been joined correctly in the sensor jump region (900nm).
On the left side of the sensor shift there is a small peak which appears to be identical on the right side.
Reply: Thank you. We corrected for the sensor jump.

Fig. 5: Zooming into the range with highest replicate spectral variation would help to discriminate single replicate spectra and help the reader to visually assess the type of errors in the spectra (random noise vs. systematic offset).
Reply: Thank you. The figure was adapted, accordingly.

Fig. 9: The figure quality needs to be improved for print. Plots should be in a vector format. Increase the font sizes for axis labels. Also, mathematical expressions should be separated by full or half spacing. Use minus sign or alternatively en dash instead of hyphen for minus. R-squared and RMSEÂ˘acould be placed below each other for better readability. Hollow circles or transparency, and bigger point symbols to deal with overplotting are recommended. Letters "a/b" in panel labels should be capitalized because the sampling designs are abbreviated as "A/B" elsewhere.
Reply: adapted

References

Beleites, C., Baumgartner, R., Bowman, C., Somorjai, R., Steiner, G., Salzer, R., & Sowa, M. G. (2005). Variance reduction in estimating classification error using sparse datasets. Chemometrics and Intelligent Laboratory Systems, 79(1–2), 91–100.
https://doi.org/10.1016/j.chemolab.2005.04.008

Jansen, M. J. W. (1998). Prediction error through modelling concepts and uncertainty from basic data. Nutrient Cycling in Agroecosystems, 50(1), 247–253. https://doi.org/10.1023/A:1009748529970

Molinaro, A. M., Simon, R., & Pfeiffer, R. M. (2005). Prediction error estimation: a comparison of resampling methods. Bioinformatics, 21(15), 3301–3307.
https://doi.org/10.1093/bioinformatics/bti499

Stevens, A., Nocita, M., Tóth, G., Montanarella, L., & van Wesemael, B. (2013). Prediction of Soil Organic Carbon at the European Scale by Visible and Near InfraRed Reflectance Spectroscopy. PLoS ONE, 8(6), e66409. https://doi.org/10.1371/journal.pone.0066409

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics, 7, 91. https://doi.org/10.1186/1471-2105-7-91

Wetterlind, J., Stenberg, B., & Rossel, R. A. V. (2013). Soil Analysis Using Visible and Near Infrared Spectroscopy. In F. J. M. Maathuis (Ed.), Plant Mineral Nutrients (pp. 95–107). Totowa, NJ: Humana Press. https://doi.org/10.1007/978-1-62703-152-3_6

Error propagation in spectrometric functions of soil organic carbon

Monja Ellinger [a], Ines Merbach [b], Ulrike Werban[c], Mareike Ließ [a]

[a] Department Soil System Science, Helmholtz Centre for Environmental Research – UFZ, Halle (Saale), Germany

[b] Department Community Ecology, Helmholtz Centre for Environmental Research – UFZ, Bad Lauchstädt, Germany

[c] Department Monitoring & Exploration Technologies, Helmholtz Centre for Environmental Research – UFZ, Leipzig, Germany

Correspondence to: Mareike Ließ (mareike.liess@ufz.de)

Abstract

Soil organic carbon (SOC) plays a major role concerning the chemical, physical and biological soil properties and functions. To get a better understanding how soil management affects the SOC content, the exact monitoring of SOC on long-term field experiments (LTFE) is needed. Visible and near infrared (Vis-NIR) reflectance spectrometry ~~is~~ provides an inexpensive and fast possibility to enhance conventional SOC analysis and has often been used to predict SOC. For this study, 100 soil samples were collected at a LTFE in central Germany by two different sampling designs. SOC values ranged between 1.5 and 2.9%. Regression models were built using partial least square regression (PLSR). In order to build robust models, a nested repeated ~~10~~5-fold cross-validation was used, that comprises ~~for~~ model tuning and ~~validation procedure~~evaluation. ~~We analysed and discussed v~~Various aspects that influence the obtained error measure were analysed and discussed. Four preprocessing methods were compared in order to extract information regarding SOC from the spectra. Overall the best model performance which did not consider error propagation corresponds to a mean RMSE_{MV} of 0.12 % SOC (R^2=0.86). This model performance is impaired by $\Delta\mathrm{RMSE}_{MV}$ = 0.04% SOC while considering input data uncertainties (ΔR^2=0.09), and by $\Delta\mathrm{RMSE}_{MV}$ = 0.12 (ΔR^2=0.17) considering an inappropriate pre-processing. The effect of the sampling design amounts to a $\Delta\mathrm{RMSE}_{MV}$ of 0.02% SOC (ΔR^2=0.05). We emphasize the necessity of a~~A~~ transparent and precise documentation of the measurement protocol, the model building and validation proce~~s~~dures~~s~~, including the calculation of the error measure, ~~is necessary~~ in order to assess ~~the~~ model ~~accuracy~~ performance in a comprehensive way and allow for comparison between publications. The consideration of uncertainty propagation is essential when applying Vis-NIR spectrometry for soil monitoring. ~~This would be the first step to gain a standardized method for model building and validation procedure.~~

Keywords: Soil organic carbon, Laboratory Vis-NIR ~~reflectance~~ spectrometry, Partial least square regression, ~~repeated 10-fold~~nested cross-validation

1 Introduction

Soil is at the same time one of the most important and one of the most limited natural resources. Most of all it is needed for food production, but also for the production of energy and fibre or for the provision of fresh water (Johnson, 2008; Lorenz and Lal, 2016; Stenberg et al., 2010). All these ~~functions~~ aspects depend on the quality of the existing soil. This quality in turn is much influenced by its SOC content since it affects chemical, physical and biological soil properties and functions (Knadel et al., 2015; Lorenz and Lal, 2016). Additionally, SOC is also interesting when it comes to the global warming issue since soil is the largest terrestrial reservoir of organic carbon in the world (Conforti et al., 2015; Johnson, 2008; McBratney et al., 2014; Stockmann et al., 2013). The SOC content of soils can be increased through the sequestration of atmospheric CO_2 into long-living components of soils (Lal, 2004; McBratney et al., 2014). Thus, the SOC stock of soils could be used as a manageable sink for atmospheric carbon (Stockmann et al., 2013), achieving both food security and a strategy against the increasing CO_2-concentration in the global atmosphere (Lal, 2004; Lorenz and Lal, 2016; McBratney et al., 2014). As the SOC content of soils reacts very slowly to environmental changes (Meersmans et al., 2009), long-term field experiments (LTFE) are required to understand the impact of soil management and farming systems on the rate of SOC sequestration (Lal, 2004) as well as on yield and crop quality in the long run.

The precise monitoring of SOC on a LTFE with conventional lab analysis is labour-intensive and expensive (Adamchuk and Viscarra Rossel, 2010; Loum et al., 2016) as it requires the analysis of a rather high amount of samples. Visible and near infrared (Vis-NIR) reflectance spectrometry can facilitate this procedure. It is non-destructive, fast and economical (Mouazen et al., 2010; Tekin et al., 2014), requiring only a small number of soil samples and little sample preparation (Conforti et al., 2015). In addition, no chemicals are needed and one spectrum contains information about many different soil components (Conforti et al., 2015; Viscarra Rossel et al., 2006b). Spectral absorption features are caused by vibrational stretching and bending of structural molecule groups and electronic excitation (Ben-Dor et al., 1999; Dalal and Henry, 1986). Molecule vibrations from hydroxyl, carboxyl and amine functional groups produce soil absorption features related to soil organic matter in the mid-infrared (MIR) region of the spectra (Croft et al., 2012). In comparison, Vis-NIR spectra show only broad and unclear adsorption features related to overtone vibrations from the MIR, but instruments are less cost-intensive and available for field monitoring as well (Stenberg and Viscarra Rossel, 2010; Viscarra Rossel et al., 2006a). Furthermore, in diffuse reflectance spectroscopy, scattering properties depend on the particular wavelengths and can vary significantly over the VIS-NIR spectral range (Pilorget et al., 2016). ~~For this reason, the application of p~~Hence, pre-processing ~~methods to~~of Vis-NIR spectra is necessary in order to ~~gain~~extract soil property related information (Stenberg and Viscarra Rossel, 2010). Scattering and other effects attributed to within-sample variance can be addressed by repeated measurements of replicate samples (e.g. Pimstein et al., 2011). Alltogether, Vis-NIR soil spectrometry has been used on many occasions to build SOC prediction models (Jiang et al., 2016; Kuang and Mouazen, 2013; Nocita et al., 2013).

However, the application of Vis-NIR soil spectrometry for SOC determination involves a couple of uncertainties. The required calibration data are determined with standard lab analysis, e.g. dry combustion, with associated uncertainties. On the other hand side, the spectral measurements are affected by the sample preparation, e.g. drying, sieving, grinding (e.g. Nduwamungu et al., 2010). Furthermore, sensor noise and other spectrometer internal sources (electronic and mechanical) can affect the measurements (Schwartz et al., 2011). Finally, these two uncertain data sources are related by a regression model. And the model building procedure involves a couple of error sources itself. The development of robust models requires a resampling process to determine the model parameters and avoid overfitting; the applied resampling method impacts model performance (e.g. Molinaro et al., 2005, Beleites et al., 2005). Further aspects that impact model performance are the available dataset in concordance with the applied sampling design, the handling of outliers, spectral pre-processing, and last but not least the model evaluation procedure. ~~However within every model building process, uncertainties of the input data affect the overall model uncertainty (error propagation). Those uncertainties contain measurement errors as well as variation in the input data (Heuvelink, 1999).~~ In most studies dealing with SOC prediction from Vis-NIR spectra, no clear statement about ~~these~~ input data uncertainties or their handling is made. The reported prediction errors only refer to the model building procedure, while uncertainties from lab measurements are neglected. Commonly, only a single SOC measurement per soil sample is available. In spectral soil sensing in lab applications, ~~Up to now~~ the general approach consist~~ed~~st ~~of~~ in averaging the multiple measured spectra of one sample to one spectrum which ~~was~~ is then used for model building (Ge et al., 2011; Stevens et al., 2013; Viscarra Rossel et al., 2003). But the number of measurements used to gain one averaged spectrum differs between studies. Jiang et al. (2016), for example, averaged 10 measurements to receive one spectrum, while Volkan Bilgili et al. (2010) and Wang et al. (2014) used four measurements. This difference is also assumed to have an influence on the uncertainties implemented in the input data.

Overall, to allow for comparision between studies, in terms of predictive uncertainty in % SOC, a modelling procedure is required that deals with the propagation of the input data uncertainites. For discussion of the general concept, please refer to Jansen (1998), for applications in soil modelling compare e.g. Heuvelink et al. (1999) and Poggio and Gimona (2014). Although, the problem of the involved uncertainties in Vis-NIR spectrometry is well-known (e.g. Gholizadeh et al., 2013, Nduwamungu et al., 2010, Mortensen, 2014), implementations of uncertainty propagation in Vis-NIR spectrometric modelling are lacking. ~~Besides input data uncertainties there are many other aspects influencing the model building process such as the sampling design and the chosen pre-processing method. This study aims to investigate the impact of these aspects on the model outcome and the calculation of error measures. Also, the influence of the model tuning procedure and the data sub setting and / or resampling for the validation procedure will be discussed.~~

2 Material and Methods

2.1 The static fertilization experiment Bad Lauchstädt

The soil samples were taken at the LTFE site "Static Fertilisation Experiment" in Bad Lauchstädt in central Germany (Körschens and Pfefferkorn, 1998). Positioned at 51° 24' N, 11 ° 53' E and with an altitude of 113 m (Körschens and Pfefferkorn, 1998), the climate is characterized by a~~n~~ mean annual precipitation of 470 – 540 mm and an average mean annual temperature of 8.5 – 9.0 °C. The soil type was characterized as a haplic Chernozem developed from loess (Altermann et al., 2005) with a soil texture of 21.1 ± 1.2 % clay, 72.1 ± 1.7 % silt, and 6.9 ± 1.9 % sand (Dierke and Werban, 2013). Saturated water conductivity and air capacity are medium to high in the top soil (Altermann et al., 2005). The Static Fertilization Experiment was initialized in 1902 by Schneidewind and Gröbler and is about 4 ha in size (Merbach and Schulz, 2013) Its objective is to investigate the impact of organic and mineral fertilization on soil fertility as well as yield and quality of crops (Körschens and Pfefferkorn, 1998; Schulz, 2017). The experiment includes eight subfields with a width from 25.2 m to 28.5 m and a length of 190 m which are each divided into 18 plots that are treated with different mineral and organic fertilizer as well as planted with different crops following a crop rotation (Körschens and Pfefferkorn, 1998). The plots of subfields 4 and 5 are additionally parted into 5 smaller subplots.

2.2 Sampling design

A total of 100 soil samples were taken at the soil surface (0-10 cm) in September 2016. The exact location of the sampling points was determined by a differential GPS GNSS LEICA Viva GS08. It was decided to sample at precise point locations instead of taking samples representative for LTFE plots to allow for a direct comparison with spectrometric field measurements for area-wide regionalisation (not included in this study). The sampling points were determined beforehand by two sampling designs. Based on the LTFE treatment factors and per-plot soil archive data including C_{org}, N_{tot}, plant available P, plant available K (both with DL-Method (VDLUFA, 2012)) and pH (Fig. 1) both designs strived to select a dataset of 50 samples representative for the soil variability of the entire LTFE. Categorical and continuous data first entered a factor analysis with mixed data (FAMD) performed with R package FactoMineR (Lê et al., 2008) to allow for further joint analysis. For design 'A' the LTFE plots were then grouped by a k-means cluster analysis. R package NbClust (Charrad et al., 2014) automatically determines the optimal number of clusters making use of 30 indices. In the end, ten plots were randomly selected from each of the resulting five clusters making a total of 50 plots to be sampled. For design 'B' the Kennard-Stone algorithm was applied with R package prospectr (Kennard and Stone, 1969; Stevens and Ramirez Lopez, 2014). 50 LTFE plots were selected involving 5 repetitions of the algorithm to reduce inter-point dependence. Finally, one sampling point was randomly selected from each of the 50 LTFE plots from design A and B based on a 5 x 5 cm raster. Plot margins of 1.5 m (3 m between plots) were excluded. Fig. 2 shows the location of the so obtained 100 soil samples.

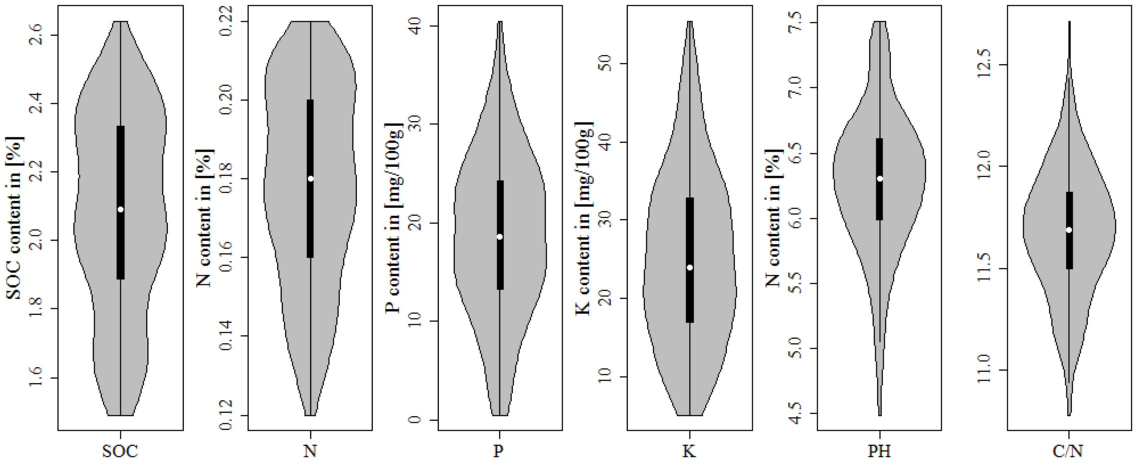

Fig. 1 Soil archive data of the LTFE measured from 2004 to 2007 (Reports of the experimental station Bad Lauschstädt 2004-2007 (unpublished)~~)~~.

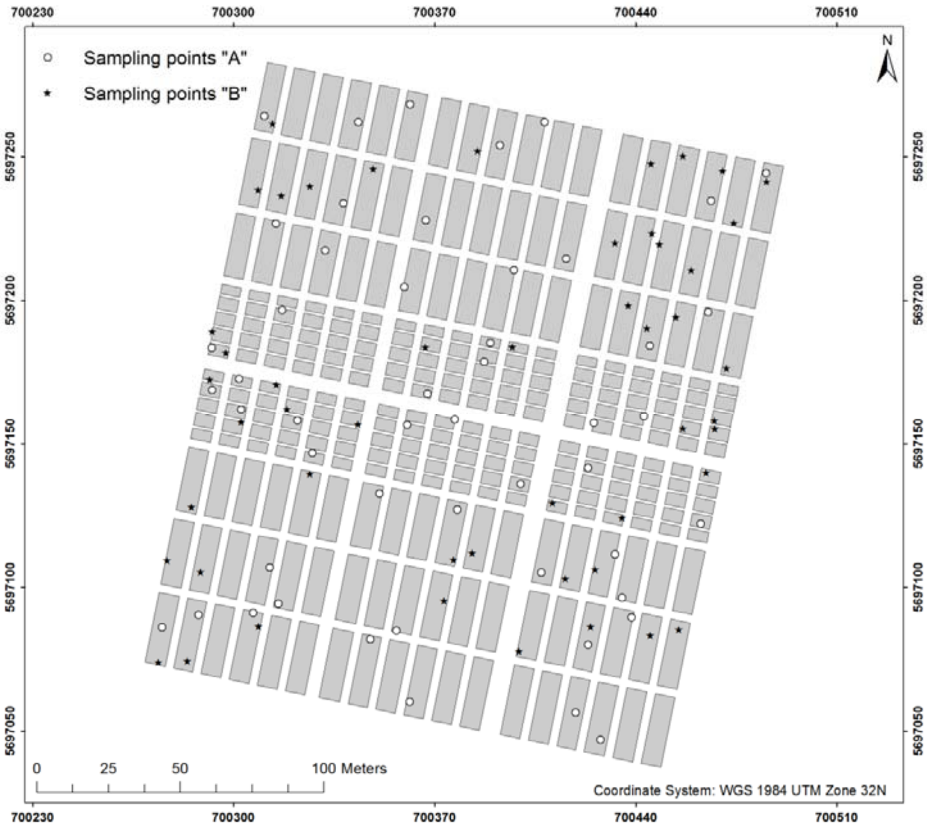

Fig. 2 Site of the S~~s~~tatic Fertilisation Experiment in Bad Lauchstädt with LTFE plots and sampling points according to design A and B. Plot margins excluded from sampling are visible as 3 m wide stripes between plots.

2.3 Laboratory measurements

The soil samples were air-dried, sieved and grinded prior to C measurements with dry combustion. A High-end Elementaranalysator vario EL Cube-CN was used. Measurements were repeated in three replicate samples. C measurements were taken as organic carbon due to negligibly small carbonate contents (below detection limit). The Vis-NIR contact measurements were performed on air-dried and sieved (2 mm) samples in July 2017, using Veris® VIS-NIR Spectrophotometer by Veris technologies, Inc. (hereinafter called Veris) containing ~~a Hamamatsu Mini-spectrometer TG series (900 to 2550 nm) and~~ an Ocean Optics USB4000 instrument (200 to 1100 nm) and a Hamamatsu Mini-spectrometer TG series (1100 to 2200 nm, resolution 6 nm). The device was warmed up for at least 20 minutes before performing measurements. All measurements were taken in a dark room to prevent daylight from affecting the outcome. The soil samples were scanned from the top. Before and between soil sample measurements, Veris was calibrated using four ~~external~~ Avian Technologies Fluorilon™ gray scale standards~~references~~. Each soil sample was divided into three sub-samples filled into petri dishes (Schott Duran petri dishes; Duran Group, Mainz, Germany). These ~~sub-~~replicate samples were not related to the three replicate samples used for C/N-measurements. For each ~~sub-~~replicate sample six spectra were gained by measuring each replicate sample three times, rotating it by 90 degrees and then measuring it three times again. This procedure resulted in 18 spectra for each soil sample. Internally the spectrometer averaged 25 scans for each spectrometer reading (spectrometer setting).

2.4 Spectral pre-processing

Veris is equipped with two spectrometers. At the beginning and end of their respective wavelength ranges noise occurs in the measurements. Therefore, the spectra between these wavelengths (1000 to 1100 nm) had to be removed. Additionally, the spectra were cut at the beginning (402 nm) and the end (2220 nm) to remove noise. A number of pre-processing methods were tested to enhance the information regarding SOC in the Vis-NIR spectra. ~~Spectral absorption features are caused by vibrational stretching and bending of structural molecule groups and electronic excitation (Ben-Dor et al., 1999; Dalal and Henry, 1986). Molecule vibrations from hydroxyl, carboxyl and amine functional groups produce soil absorption features related to soil organic matter in the mid-infrared (MIR) region of the spectra (Croft et al., 2012).Vis-NIR spectra show only broad and unclear adsorption features related to overtone vibrations from the MIR (Stenberg and Viscarra Rossel, 2010; Viscarra Rossel et al., 2006a). For this reason, the application of pre-processing methods to Vis-NIR spectra is necessary in order to gain soil property related information (Stenberg and Viscarra Rossel, 2010).~~ The spectra were tested for outliers using R package mvoutlier (Filzmoser and Gschwandtner, 2017). For this procedure a PCA is performed, using then the first two obtained PCs for outlier detection with function aq.plot. ~~According to Filzmoser (2005), the Mahalanobis distance of normally distributed data follows a chi-square distribution. Observations which lay beyond a certain quantile of this distribution are marked as outliers and removed from the data (Filzmoser and Gschwandtner, 2017). In this study, no outliers were detected.~~ Out of the tested pre-processing methods, four different combinations are shown in this study in order to demonstrate their different effects on the prediction model. Their application resulted in spectra with different wavelength ranges (Table 1) and different appearance (Fig. 3). These pre-processing techniques include the Savitzky-Golay algorithm (SG), continuum removal (CR), the standard normal variate (SNV), the first derivative (d1) and the gap-segment algorithm (gapDer). The SG algorithm fits a polynomial regression on the spectral data to find the derivative at a center point i of a defined smoothing window (w) (Rinnan et al., 2009; Savitzky and Golay, 1964; Swarbrick, 2016). CR can be seen as a spectra normalization technique which enables to compare different absorption characteristics from a mutual baseline (Kokaly, 2001;

Mutanga and Skidmore, 2003). It identifies the local reflectance spectra maximum points and connects those points to form a convex hull (Mutanga and Skidmore, 2003; Stevens and Ramirez Lopez, 2014). Calculating

$$\Phi_{\boldsymbol{i}} = \frac{\boldsymbol{x_i}}{\boldsymbol{c_i}} \tag{1}$$

for i = {1, … , p) with x_i and c_i being the initial and the continuum reflectance values at wavelength i of a set of p wavelengths then gives the continuum-removed reflectance value $\boldsymbol{\phi i}$ (Stevens and Ramirez Lopez, 2014). All other data have values between 1 and 0 (Mutanga and Skidmore, 2003; Schmidt and Skidmore, 2001). Thus the absorption peaks are enhanced (Schmidt and Skidmore, 2001). SNV is a scatter-corrective pre-processing method (Rinnan et al., 2009).

The basic formula is as follows

$$\boldsymbol{x_{corr}} = \frac{\boldsymbol{x_{org}} - \boldsymbol{a_0}}{\boldsymbol{a_1}} \tag{2}$$

with a_0 as the measured spectrum's average value which shall be corrected and a_1 being the sample spectrum's standard deviation. x_{org} are the original spectra and x_{corr} the corrected spectra after applying SNV. SNV operates row-wise, so each observation is processed on its own (Rinnan et al., 2009; Stevens and Ramirez Lopez, 2014). d1 represents the slope of the spectrum, showing peaks where the spectrum displays its maximum slope and crossing zero where the spectrum shows peaks (Leone et al., 2012). According to Knadel et al. (2015) and Smith (2002), d1 can be used to remove baseline offsets from the spectra. The estimation of d1 is done by computing the difference between two batched spectral points x_i and x_{i-1} (Eq. 3)

$$\boldsymbol{x_i}' = \boldsymbol{x_i} - \boldsymbol{x_{i-1}} \tag{3}$$

with x_i' as the value of the first derivative at the i^{th} wavelength (Rinnan et al., 2009). The downside of using derivative spectra is their tendency to over-fit the calibration model (Stevens and Ramirez Lopez, 2014). Moreover, derivatives may increase noise so that a smoothing of the data is required (Leone et al., 2012; Stevens and Ramirez Lopez, 2014). With the gapDer a smoothing is performed under a chosen segment size (s) and then a derivative follows (Stevens and Ramirez Lopez, 2014). The application of the different pre-processing methods for this study was ~~done~~ conducted using R package prospectr (Stevens and Ramirez Lopez, 2014).

Table 1 Combinations of pre-processing techniques used in this study; w = window size, s = segment size.

Pre-processing methods	Wavelength range	Abbr.
Savitzky-Golay (w =11 nm)	432 – 2201 nm	SG
Savitzky-Golay (w=11 nm) and continuum removal	432 – 2201 nm	SGCR
Standard normal variate and 1st derivative	408 – 2186 nm	SNVd1
Gap-segment algorithm (w = 11 nm, s = 10 nm)	490 – 2163 nm	gapDer

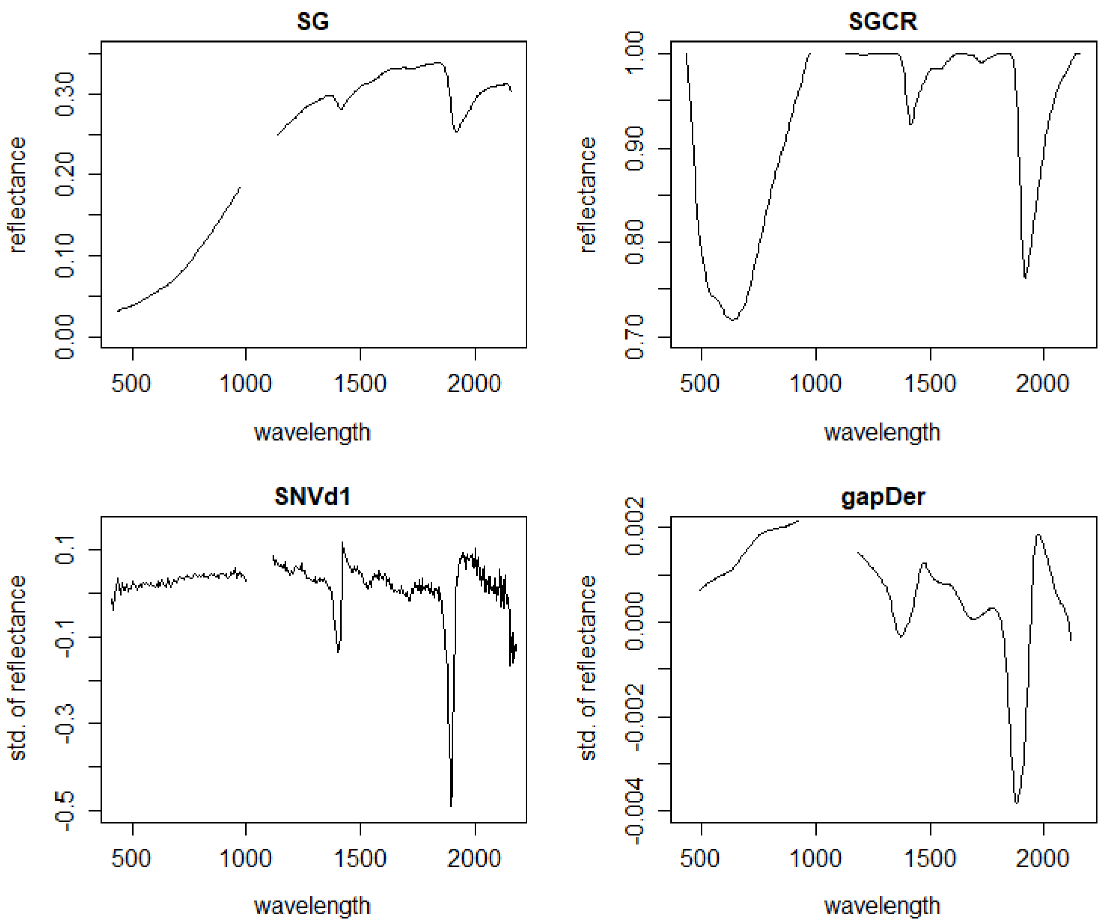

Fig. 3 Impact of different pre-processing techniques on a spectrum; SG = Savitzky-Golay, CR = continuum removal, SNV = standard normal variate, d1 = 1st derivative, gapDer = Gap-segment algorithm.

2.5 Error propagation

A problem occurring in every model building process is uncertainty propagation. ~~Where u~~Uncertainties ~~already existing in~~of the input data ~~as well as in the~~and ~~used~~ model result in uncertainties in the output (Brown and Heuvelink, 2006). Uncertainties in the input data are caused by errors in data acquisition (e.g. measurement errors) as well as variation in the data themselves (e.g. within-sample variability) (Heuvelink, 1999). For this study, there are two different sources for errors in data acquisition: the measurement of the spectral data and the measurement of the SOC content of the soil samples. In order to investigate the influence of these errors, different ~~data set~~datasets were built in this study. Fig. 4 gives an overview. From the measured Vis-NIR spectra, three different spectral data variants were created (Fig. 4, step 1). For the first variant, all 18 spectra were retained. The inclusion of all 18 spectra reveals the influence of the error implemented in the spectral measurements as well as the influence of the within-sample variability. For the second, the three measurements obtained before and after sample rotation were averaged separately resulting in 6 spectra per sample showing the influence of within-sample variability. For the third data variant, all 18 spectra were averaged to 1 mean spectrum per sample, removing the influence of the measurement error as well as the within-sample variability. The different spectra obtained through this procedure can be seen in Fig. 5. Only parts of the spectra are depicted in order to show their differences. The three different spectral data variants were then pre-processed with the different pre-processing methods from Table 1 (Fig. 4, step 2), resulting into 12 different spectral ~~data set~~datasets (Fig. 4, step 3). These were then combined with single and averaged SOC values in step 4 so that altogether 24 ~~data set~~datasets were obtained (Fig. 4, step 5). In order to compare the two sampling designs, this procedure was carried out for the 50 soil samples labelled “A” and “B” and also for the complete set of soil samples. In this way, three different soil sample sets (“A”, “B” and “all samples”) were achieved.

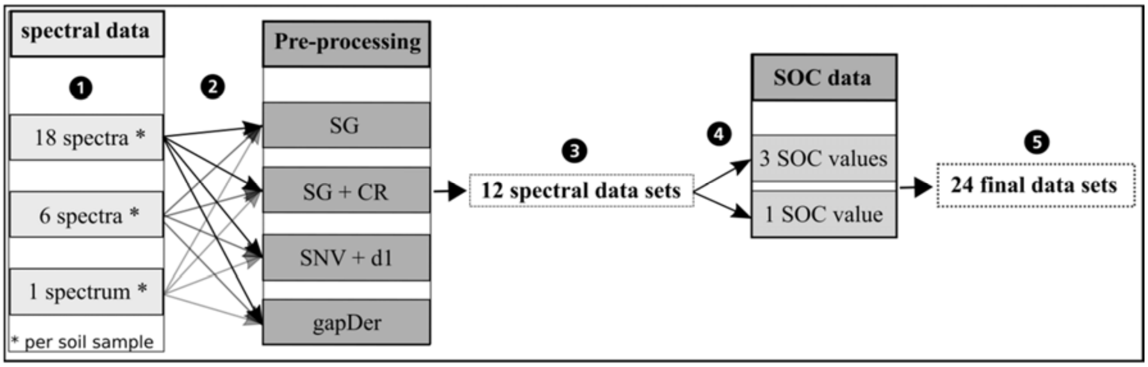

Fig. 4 ~~Data set~~Datasets to investigate the uncertainty propagation. SG = Savitzky-Golay, CR = continuum removal, SNV = standard normal variate, d1 = 1st derivative, gapDer = Gap-segment algorithm.

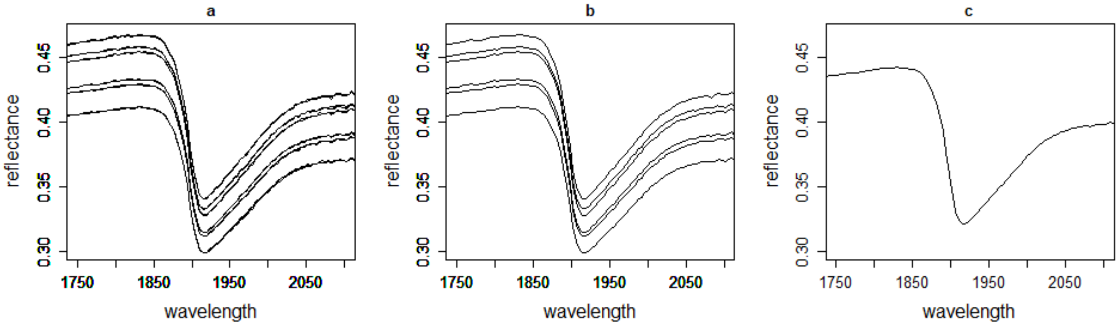

Fig. 5 ~~Different sets of spectral data per soil sample~~Zoom-in to a sample’s spectral dataset: a) 18~~ original~~ spectra comprised of 6 replicate sample measurements with 3 scans each~~1 and a2 showing 1~~, b)~~1 und b2~~ 6 ~~averaged ~~spectra related to replicate sample measurements (average of three scans each) and c) 1 averaged spectrum~~1 and c2~~.

2.6 Model building and validation

Regression models were built using partial least square regression (PLSR). Out of the many algorithms, PLSR is seen as a standard method for spectral calibration and prediction (Mouazen et al., 2010; Tekin et al., 2014; Viscarra

Rossel et al., 2006b). For applications to predict SOC from Vis-NIR soil spectra see (Conforti et al., 2015; Jiang et al., 2016; Kuang and Mouazen, 2013; Nocita et al., 2013). PLSR is described in detail by Martens and Næs (1989) and Naes et al. (2002). It incorporates characteristics from principle component analysis (PCA) and multiple regression (Abdi, 2007). The concept behind PLSR is to seek a small number of linear combinations (components or latent factors) obtained from the measured spectral data and to use them in the regression equation to predict SOC instead of the initial values (Martens and Næs, 1989; Naes et al., 2002). These components are constructed so that they account for most variance in the measured spectral data (X) and the SOC content (Y) and at the same time maximize the correlation between X and Y. In other words, PLSR leads to the covariance between X and Y being maximized (Bjørsvik and Martens, 2008; Summers et al., 2011; Tekin et al., 2014; Wehrens, 2011).

In order to receive a robust model, it is important not to include too many components in model building as this will lead to over-fitting (Hastie et al., 2009; Kuhn and Johnson, 2013). On the other hand, the inclusion of too few components comprises the risk of building an under-fitted model which is too small to cover the variability existing in the soil spectral data (Naes et al., 2002). The selection of the optimal number of components is hereinafter referred to as *model tuning*. In order to receive a robust model, resampling is commonly applied for model building and validation. But resampling can also be used for *model tuning* to receive robust tuning parameters (Guio Blanco et al., 2018; Hastie et al., 2009; Kuhn and Johnson, 2013). For small ~~data set~~datasets, k-fold cross-validation is recommended (Hastie et al., 2009).

In this study, model building, *model validation* and *model tuning* was done using a nested repeated ~~10~~5-fold cross-validation (~~10-fold~~ CV)~~, an approach applied before by (Guio Blanco et al., 2018)~~ (e.g. Varma and Simon, 2006, Guio Blanco et al., 2018). Repeated ~~10~~5-fold CV can increase the precision of the prediction while maintaining a small bias (Kuhn and Johnson, 2013). Five Repetitions of ~~10~~5-fold CV were conducted in this case.

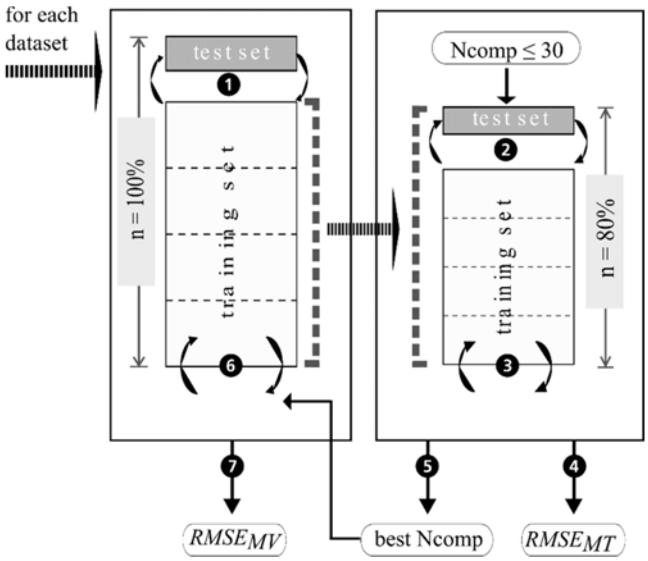

Fig. 6 M*odel tuning* and *model validation* procedure with a nested k-fold cross validation approach. The right box shows the *model tuning*, the left one the *model validation* procedure; Ncomp = number of components; adapted from Guio Blanco et al. (2018).

Fig. 6 shows the various steps of the modelling procedure involving repeated ~~10~~5-fold CV for *model tuning* (right box) and *validation* (left box). In the process the ~~data set~~dataset (n = 100%) is randomly subdivided into ~~10~~ 5 folds of equal size (step 1)~~. This was done using the sample() function in R from package dplyr (Wickham and Francois, 2017)~~. One of the ~~10~~ 5 folds is held out as test set and the other ~~nine~~ four are used as training set and partitioned again into ~~10~~ 5 folds for *model tuning* (step 2). The optimal number of components (~~*optimal*~~ *best Ncomp*) is then determined by computing a PLSR on the resampled data testing 1 to ~~40~~ 30 components (step 3) calculating the repeatedly ~~10~~5-fold cross-validated RMSE of *model tuning* (RMSE$_{MT}$) (step 4). This was implemented with the trainControl() function in package caret (Kuhn, 2017). The optimal number of components (step 5) is then used in model building ~~for validation~~ (step 6)~~ with PLSR using train() function from caret~~. The resulting model's test set RMSE of model validation (RMSE$_{MV}$) is determined in step 7. The whole procedure is repeated until all folds in the boxes ~~had~~ have once been used as ~~the~~ test set.

In this study, the ~~partition~~ subdivsision of the spectral data into ~~10~~ the folds for ~~10~~5-fold CV had to account for repeated scans and replicate measurements per sample. ~~be done very carefully, as in some cases multiple spectra existed for one sample.~~ All spectra for one sample were assigned to the same fold during k-fold CV, i.e. k-fold group CV. The folds contained always the same sample IDs for the various data variants described in Figure 4. For model validation, folds were created following a repeated stratified group CV approach (5-fold). The data were primarily subdivided into 5 equal probability strata based on their density function. The data from each of the strata were then randomly assigned to the 5 folds. Neighbouring points of ≤ 5 m distance were assigned to the same fold to avoid spatial autocorrelation and too optimistic error measures.

~~In order to reduce autocorrelation, it was especially important to ensure that those spectra and also spectra gained from samples which were located in the same treatment plot were always in the same fold during 10-fold CV, as those spectra were assumed to be very similar or even identical. Thus, it was possible to compare the results gained with the different data sets and to investigate the influence of uncertainties implemented in the input data.~~

3 Results and Discussion

3.1 Soil organic carbon content

Fig. 7 shows the distribution of the SOC content of the three soil sample sets, consisting of 50 soil samples labelled "A" and "B" and 100 soil samples referred to as "all samples". The measurement error existing in the SOC measurements, here defined as the difference between replicate measurements, ranges from 0.003 to 0.229 % SOC with a mean of 0.048 % SOC. The aim of this study was not to compare the two different sampling designs among each other, but to test whether they are representative of the SOC values existing on the LTFE. For this purpose, per-plot soil archive data from the years 2004 to 2007 are also displayed. The statistics of the data are given in Table 2. In order to compare distributions between the archive SOC data and "A", "B" and "all samples", a Mann–Whitney U test was applied to the data, testing the "A", "B" and "all samples" against the archive data, respectively. In all cases, no significant difference between the different sample sets and the archive data could be found. This shows that all soil sample sets used in this study are representative for the SOC values existing in the LTFE. Nevertheless, the SOC distribution of "A" and "B" samples differ, with the "A" samples resembling the distribution of all 100 samples more than the "B" samples. The "A" samples contained more samples representing higher SOC values, whereas the "B" samples show a higher representation for lower SOC values. This difference in the distribution of SOC values may have an influence on the prediction results of the models built with "A" and "B" samples. "A" models may be better in predicting higher SOC values, while simultaneously failing to estimate lower SOC values in an appropriate way. To the contrary, "B" models may predict lower SOC values more accurate than higher SOC values. The violin plots of all three soil sample sets do not resemble~~s~~ the archive violin plot very much. The plots for "A" and all samples show higher and lower SOC values than the archive data due to the fact that those data are obtained from compound samples for one plot. The "B" samples share the same minimum value with the archive data, but display slightly higher SOC values. This ~~shows~~ indicates that the choice of the sampling design ~~has~~ might have an influence on the model outcome, even if both designs represent the SOC values on the experimental field in an appropriate way.

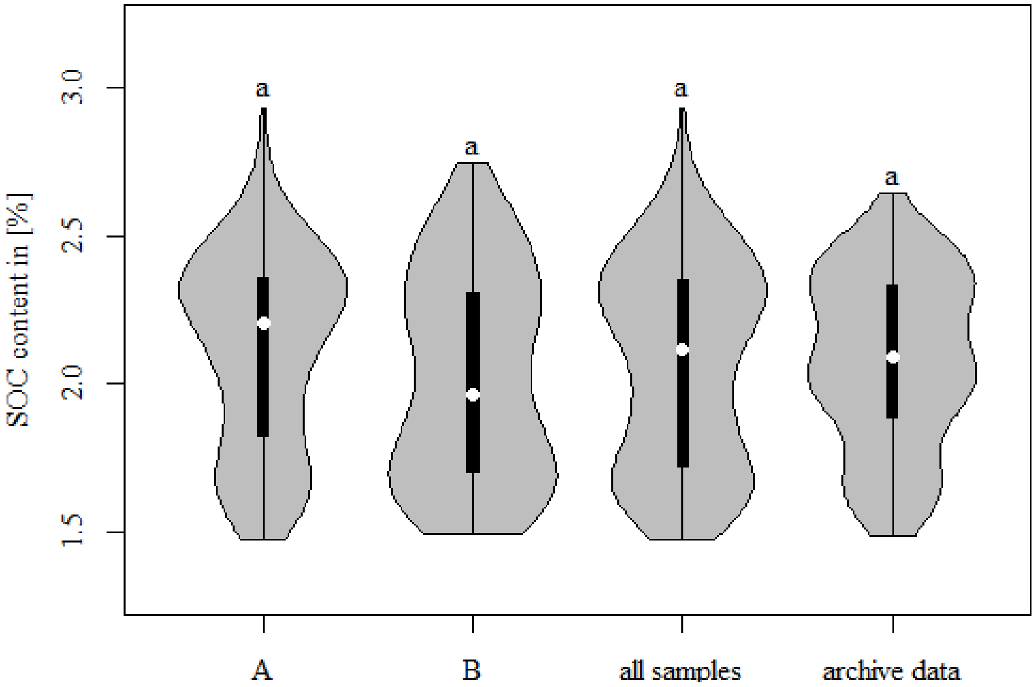

Fig. 7 Soil organic carbon (SOC) content of the three soil sample sets "A" (left), "B" (middle), and all (middle) and of archive data measured from 2004 to 2007 (right); The thin line shows the 95 % confidence interval, the bar the interquartile range and the dot the median; Mann–Whitney U test was used to compare "A", "B", and all samples to the archive data; the three soil sample sets were not compared among each other.

Table 2 Statistics of soil organic carbon in [%] for the three different soil sample sets and the per-plot soil archive data.

Samples	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
"A"	1.47	1.82	2.21	2.11	2.36	2.93
"B"	1.49	1.70	1.97	2.02	2.31	2.74
all	1.47	1.72	2.12	2.01	2.35	2.93
Archive data	1.49	1.89	2.09	2.08	2.33	2.64

3.2 Comparison of ~~data set~~datasets and pre-processing methods

According to Filzmoser (2005), the Mahalanobis distance of normally distributed data follows a chi-square distribution. Observations which lay beyond a certain quantile of this distribution are marked as outliers and removed from the data (Filzmoser and Gschwandtner, 2017). In this study, no outliers were detected.

Table 3 displays the different combinations of soil spectra and SOC values per soil sample. The resulting models are then named after the following scheme: ~~Modelset~~Dataset$_{x1\ x2\ x3}$ with the SOC measurement error (x1), the spectral measurement error (x2) and the within-sample variability (x3). The number 1 indicates that the respective error is included into the model, the number 0 shows that the error was removed beforehand by averaging the data.

Table 3 Data basis per soil sample~~: Number of SOC values, number of measured spectra and resulting spectra per soil sample are shown.~~

	Number of SOC values per sample	**Number of measured spectra per sample**	**Resulting ~~spectra~~ size of the dataset per sample ~~for model building~~**
~~Modelset~~Dataset$_{111}$	3	18	54
~~Modelset~~Dataset$_{101}$	3	6	18
~~Modelset~~Dataset$_{100}$	3	1	3
~~Modelset~~Dataset$_{011}$	1	18	18
~~Modelset~~Dataset$_{001}$	1	6	6
~~Modelset~~Dataset$_{000}$	1	1	1

Fig. 8 shows boxplots of the RMSE$_{MV}$ ~~obtained with the data sets~~. The six ~~model sets~~datasets are displayed in one panel. The models using "A" samples are shown in the 1st column (a), "B" samples in the 2nd column (b) and all samples in the 3rd column (c); ~~the number~~Figure lines~~s~~ 1 to 4 refer to the used pre-processing method. As ~~10~~5-fold CV with five repetitions was performed, five RMSE$_{MV}$ are shown ~~for each data set~~in each boxplot. The model results are now compared based on their mean RMSE$_{MV}$ and their interquartile range. It is not surprising that the dataset of 3 SOC replicate measurements with 1 averaged spectrum (Dataset$_{100}$) results in low model performance, as the within-sample variance concerning SOC cannot be explained by the contained predictor information; the input data uncertainty propagates through the model building process. This model performance is impaired in some cases by Dataset$_{101}$ which combines the 3 SOC measurements with 6 replicate spectral measurements (Figures 8b$_1$, a$_2$, a$_3$, b$_3$, c$_3$) but not always as we would expect while further uncertainty is added to the model building process. It seems that the within sample variation concerning soil spectra can somehow compensate the within sample variability concerning SOC within the model building process, although replicate measurements do not match. Considering the dataset with 18 spectra and 3 SOC measurements, model performance improves even further (Dataset$_{111}$). In contrast to this, we find the expected pattern while only 1 SOC measurement is considered: model performance results display an increase of RMSE values from from Dataset$_{000}$ to Dataset$_{001}$ to Dataset$_{011}$ due to the fact that more spectral variance is related to the same target information concerning SOC. This applies for three of the four spectral preprocessing variants (SG, SGCR, gapDer), while SNVd1 preprocessing displays an unexpected pattern with datasets including replicate measurements and multiple scans even outperforming those with averaged data. Overall SGCR resulted in the best model performance for data 'A' (Figure 8a$_2$) and all samples (Figure 8c$_2$), while SG preprocessing resulted best for data 'B' (Figure 8b$_1$). However, the latter does not apply for Dataset$_{000}$ where gapDer preprocessing resulted in the best model performance with RMSE$_{MV}$= 0.13. Comparing the mean RMSE$_{MV}$ for samples 'A' and 'B', models build on samples 'B' resulted in better model performance whith the exception of Dataset$_{100}$. The locations of the 'B' samples were determined using the Kennard-Stone-algorithm, those of the 'A' samples with k-means clustering algorithm. Fig. 7 allows an assessment of the data collected by those two sampling designs and shows no clear resemblance between the violin plots of 'A', 'B', all samples, and the archive violin plot. But the Mann–Whitney U test did not show a significant difference between the archive data and the sample sets used in this study. 'A' samples as well as 'B' samples seem to represent the LTFE SOC data in an adequate way. As already mentioned above, the difference in the distribution of SOC values of 'A' and 'B' samples might have led to a different predictive capability in certain SOC value ranges. If this difference is the reason for the better performance of the 'B' models cannot be stated with certainty.

Comparing ~~Modelset~~Dataset$_{111}$ with ~~Modelset~~Dataset$_{000}$ shows how the inclusion of all input data uncertainties affects the model results. It can be seen that a model without error propagation (~~Modelset~~Dataset$_{000}$) reaches a mean RMSE$_{MV}$ of 0.12 ~~– 0.13~~ % SOC and mean R^2 ~~0.85 –~~ 0.~~87~~86 using the pre-processing method which delivered the best results. A model with error propagation (~~Modelset~~Dataset$_{111}$) on the other hand reaches a mean RMSE$_{MV}$ of 0.16 ~~– 0.17~~ % SOC and R^2 0~~.76 – 0.78~~.77. ~~In general, Modelset$_{000}$ leads to better predictions with only two exceptions (a3 and b3).~~ This is further illustrated in Fig. 9 and could be expected, as ~~Modelset~~Dataset$_{000}$ contains no input data uncertainties, the RMSE$_{MV}$ values therefore only correspond to the model building process ~~which could influence the model outcome~~. ~~What happens when only some of the input data uncertainties are implemented shows the consideration of Modelset$_{001}$ and Modelset$_{101}$. Modelset$_{001}$, including only the within-sample variability, led to better results (R^2 0.81 – 0.82, RMSE$_{MV}$ 0.14 – 0.15 % SOC) than Modelset$_{101}$ (R^2 0.72 – 0.75, RMSE$_{MV}$ 0.17 – 0.18 % SOC) which includes both SOC measurement error and within-sample variability. Comparing Modelset$_{001}$ to Modelset$_{000}$ shows that the influence of the within-sample variability alone leads almost always to slightly poorer model results. Modelset$_{100}$, containing only the SOC measurement error, delivers in most cases~~

~~distinct poorer results than Modelset$_{000}$. This indicates that the measurement error implemented in the SOC content measurements has a high effect on the model outcome if the data basis, in this case the available spectra per soil sample, is small. When the data basis per sample is higher, as for example for Modelset$_{011}$ and Modelset$_{111}$ (see Table 3), the model may be able to process the information contained in the data set in a better way, leading to nearly similar results for Modelset$_{011}$ and Modelset$_{111}$ (R^2 0.77 – 0.79, $RMSE_{MV}$ 0.15 - 0.16 and R^2 0.76 – 0.78, $RMSE_{MV}$ = 0.16 – 0.17). These results show the high influence the treatment of the input data has on the model outcome and on the error measures.~~ Overall the best model performance which does not consider error propagation corresponds to a mean $RMSE_{MV}$ of 0.12 % SOC (R^2=0.86). This model performance is impaired by $\Delta RMSE_{MV}$ = 0.04% SOC while considering input data uncertainties (ΔR^2=0.09), and by $\Delta RMSE_{MV}$ = 0.12 (ΔR^2=0.17) considering an inappropriate pre-processing. The effect of the sampling design amounts to a $\Delta RMSE_{MV}$ of 0.02% SOC (ΔR^2=0.05).

~~The combination of the pre-processing methods SG and CR was found to be the best pre-processing variant for all models built with the “A” samples (a1 to a4) and also for the models built with all data (c1 to c4) with the exception of Modelset$_{100}$. For the models built with the “B” samples, no clear best pre-processing method could be recognized. The models in b1 to b4 though show a similar response to the pre-processing methods with SNVd1 leading to the poorest model results in almost every case.~~

~~Regarding only the models derived from “A” and “B” samples (a1 to a4 and b1 to b4) it becomes clear that the models using “B” samples led to better predictions than those using “A” samples. The locations of the “B” samples were determined using the Kennard-Stone-algorithm, those of the “A” samples with k-means clustering algorithm. Fig. 7 allows an assessment of the quality of those two sampling designs and shows no clear resemblance between the violin plots of “A”, “B”, “all samples” and the archive violin plot. The Mann–Whitney U test did not show a significant difference between the archive data and the sample sets used in this study. “A” samples as well as “B” samples seem to represent the LTFE SOC data in an adequate way. As already mentioned above, the difference in the distribution of SOC values of “A” and “B” samples might have led to a different predictive capability in certain SOC value ranges. If this difference is the reason for the better performance of the “B” models cannot be stated with certainty.~~

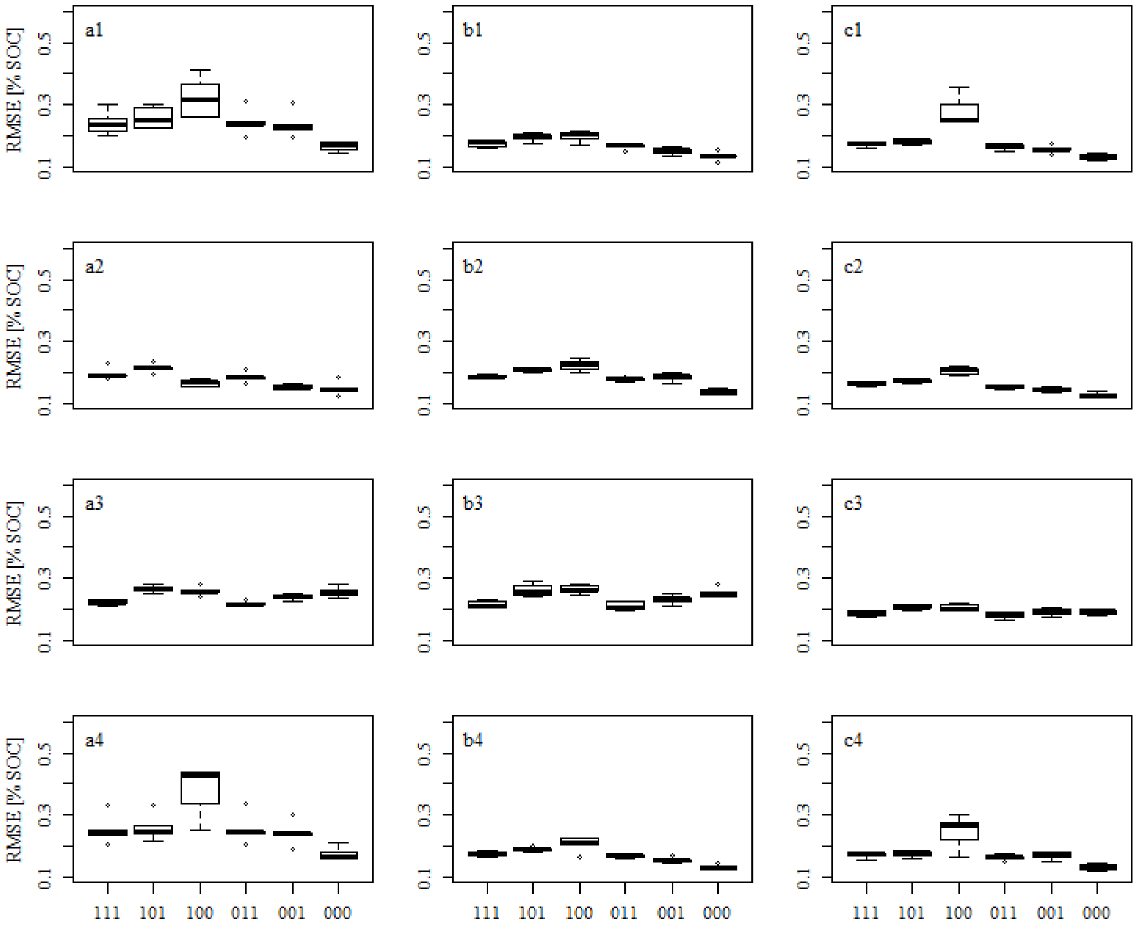

Fig. 8 Boxplots of testset RMSE$_{MV}$ (5 repetetions)~~Results~~ **obtained with the** ~~data set~~various dataset**s; Figure columns refer to** ~~data set~~dataset**s using a)** ~~“~~’A~~”~~‘ **samples, b)** ~~“~~’B~~”~~‘ **samples and c) all samples; figure rows refer to the applied pre-processing, 1 = SG, 2 = SGCR, 3 = SNVd1, 4 = gapDer.**

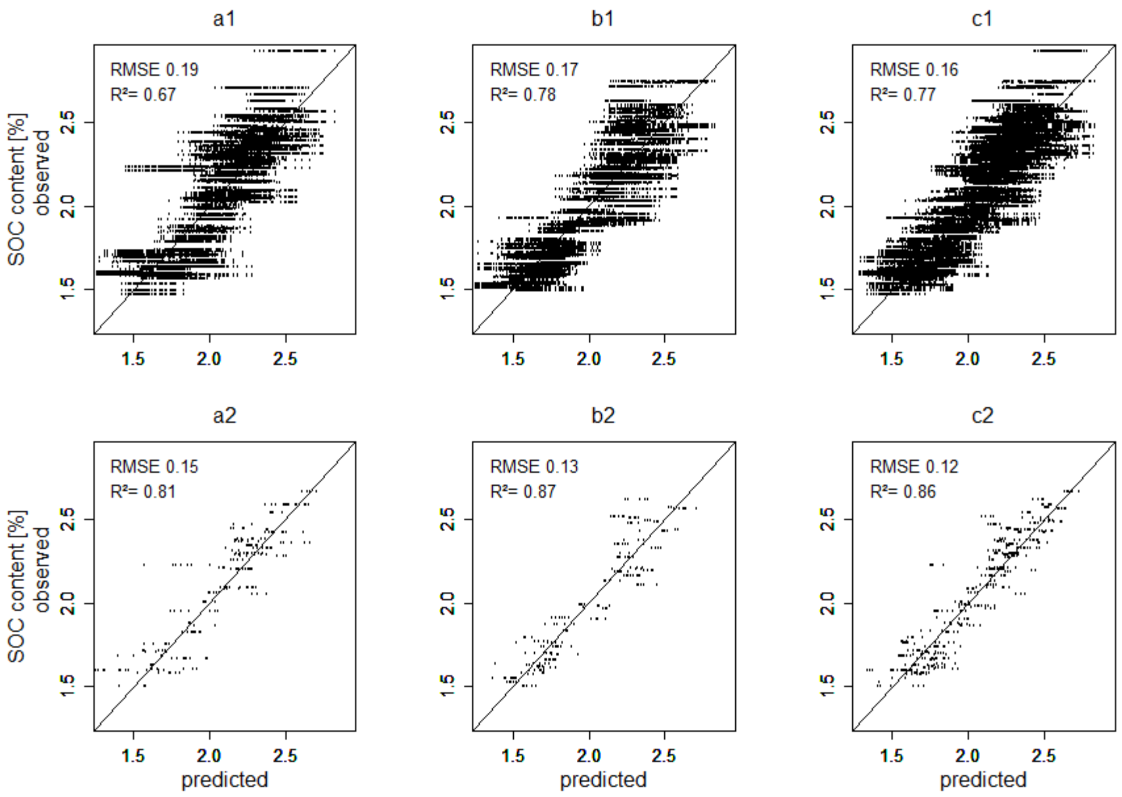

Fig. 9 Comparison of predicted and observed soil organic carbon (SOC) values for ~~Modelset~~Dataset111 (a1 to c1) and ~~Modelset~~Dataset000 (a2 to c2) for five repetitions with the corresponding best pre-processing (SGCR for data ‘A’ and all data, SG for data ‘B’); a) shows results for ~~“~~’A~~”~~‘ samples, b) for ~~“~~’B~~”~~‘ samples and c) for all samples. The depicted RMSE and R² values refer to the mean of 5 repetitions.

Examples of other studies using Vis-NIR spectra to predict SOC are listed in Table 4. Most studies used a different number of scans and replicate samples to get an averaged spectrum to predict SOC, so the error implemented in the respective input data is assumed to be different. Pimstein et al. (2011) propose a number of 3-5 replicate measurements as standard protocol for measurering Vis-NIR spectra of soil samples under laboratory conditions. Figure 5b indicated the high impact of within sample variance determined by the measurements of replicate samples, whereas the effect of the repeated scans per replicate is comparatively small (compare Figures 5a and b). We dried and sieved the samples before spectral measurements but did not grind them to fine powder. The latter might reduce spectral variance in replicate measurements, but the benefit of Vis-NIR spectroscopy as a fast and inexpensive method is reduced. One might argue that samples have to be grinded for SOC analysis, anyway. However, this requires a tiny fraction of the large amount that would have to be grinded for Vis-NIR measurements. In addition, comparison to measurements under field conditions is further distorted while grinding the samples for lab measurements. As shown in ~~section 3.1~~ Figure 8, the input data error has a major influence on the model outcome. In none of these studies the error in SOC measurements is mentioned to be considered during model building. Also, in most studies the available ~~data set~~dataset is randomly parted into calibration and validation set, using different percentages of the data for the two sets. Jeong et al., 2017 and Beleites et al. (2005) showed that different validation strategies lead to different error measures.

The overall best pre-processing methods in this study ~~was~~ were the combination of SG and CR as well as SG alone. SG was used successfully by many authors before for spectral pre-processing (Bogrekci and Lee, 2006; Nocita et al., 2013; Stevens et al., 2013; Viscarra Rossel et al., 2006a). CR was used by e.g. Viscarra Rossel et al. (2016) or Loum et al. (2016) with acceptable success. The combination of SG and CR could not be found in literature, though. SNV was applied before by other authors in order to remove baseline effects (Knadel et al., 2015; Minasny et al., 2011; Viscarra Rossel et al., 2006a). The pre-processing technique d1 was ~~often~~ found to lead to poorer model results and rather unexpected performance patterns in this study. ~~This~~ The former may have its cause in the tendency of d1 to over-fitting and the increasing of noise in the data as reported by some authors before (Leone et al., 2012; Stevens and Ramirez Lopez, 2014). We do not have an explanation for the latter, though. Leone et al. (2012) suggests the usage of SG in combination with d1 to solve the problem. For the usage of gapDer no comparison could be found in literature. As there is no standard pre-processing technique which works on all spectral data (Stenberg and Viscarra Rossel, 2010), it is recommended to always test various techniques and to choose the one which performs best for the respective data.

Table 4 R^2 from literature for soil organic carbon prediction models.

Author	number of samples	averaged scans per sample	calibration and validation set	R^2
(Reeves and Smith, 2009)	720	64	a) Cross-validation with all samples b) Independent validation set	a) 0.534 b) 0.335
(Islam et al., 2003)	161	-	Randomly selected from ~~data set~~dataset (121 / 40)	0.76
(Wang et al., 2014)	156	4	Randomly selected from ~~data set~~dataset (116 / 40)	0.67 - 0.88
(Volkan Bilgili et al., 2010)	512	100	Randomly selected from ~~data set~~dataset (70 % and 30 %)	0.80

(Kuang and Mouazen, 2013)	174	10	60 % and 40 %	-
(Jiang et al., 2016)	98	10	~~Data set~~Dataset parted into calibration and validation set	0.58 - 0.85
(Conforti et al., 2015)	201	30	-	-
(Leone et al., 2012)	374	4	Randomly selected from ~~data set~~dataset (2/3 and 1/3)	0.84 - 0.92

4 Conclusions

This study aimed to investigate the influence different aspects have on the model building process and the calculation of error measures. Those aspects included the input data uncertainties, the number of measurements per sample and the chosen pre-processing method. Furthermore, the effect of sampling design, model tuning and validation procedure was discussed. The applied nested cross validation approach that includes resampling in model tuning as well as evaluation can be recommended in general. The fact that replicate measurements and scans as well as geographically near samples have to be assigned to the same fold (k-fold group CV) in order to obtain unbiased error measures is often neglected, and, therefore, has to be emphasized. Overall the best model performance which does not consider error propagation corresponds to a mean $\mathrm{RMSE}_{\mathrm{MV}}$ of 0.12 % SOC (R^2=0.86). This model performance is impaired by $\Delta\mathrm{RMSE}_{\mathrm{MV}}$ = 0.04% SOC while considering input data uncertainties (ΔR^2=0.09), and by $\Delta\mathrm{RMSE}_{\mathrm{MV}}$ = 0.12 (ΔR^2=0.17) considering an inappropriate pre-processing. The effect of the sampling design amounts to a $\Delta\mathrm{RMSE}_{\mathrm{MV}}$ of 0.02% SOC (ΔR^2=0.05). ~~The influence of the measurement error for both SOC and spectral measurements was apparent in this study, leading almost always to worse model results. As the usage of different devices will lead to different measurement errors, other authors (Ben Dor et al., 2015; Pimstein et al., 2011) recommend the application of an internal reference to compare the measurements taken from different devices. The need of such an internal reference for spectral measurements is illustrated by the results of this study.~~The rather high within-sample variance of spectral replicate measurements of field-scale soil samples of very similar mineral composition requires a reconsideration of the number of replicate measurements per sample. 3-5 replicates as suggested for Vis-NIR soil measurements might simply not be enough.~~Although the samples were sieved and homogenised before taking measurements, within-sample variability sill had an influence on the model outcome.~~ In general, ~~the variability of the samples~~this within-sample variability depends on the soil treatment and possibly also on the origin of the samples (e.g. agricultural or forest soils). ~~The two sampling designs used in this study were tested on their statistical inference. It became apparent that both were representative of the SOC values existing in the LTFE. Nevertheless they led to models with different predictive capability, showing the influence the chosen sampling design has on the outcome. The impact of the different pre-processing methods showed clearly in this study, but it may be different when using other data sets, though. Therefore, various pre-processing techniques should be tested and compared beforehand. All in all~~Overall, this study showed that it is important to clarify which information the reported error measure contains ~~for a certain model~~. ~~In order to compare the error measures obtained with different model approaches,~~ We are aware that the consideration of stratified group CV for model evaluation but only partly for tuning (spectral replicate measurements and scans per sample were always assigned to the same fold) might impair model performance as suboptimal model parameters might be selected. This will be adapted in future studies. We emphasize the necessity of a transparent and precise documentation of the measurement protocol, the model building and validation procesdure, including the calculation of the error measure, in order to assess model preformance in a comprehensive way and allow for comparison between publications. Particularly, when Vis-NIR spectrometry is used for soil monitoring, the aspect of uncertainty propagation in the involved modelling procedure becomes essential. ~~several points have to be considered. These points include the number of sub-samples and measurements taken for one sample and the sample rotation made during measurement. It also has to be made certain that autocorrelation between calibration and validation sets is avoided during the model building process.~~

Acknowledgements

We are grateful to the support of our colleagues from the UFZ departments Soil System Science, Community Ecology, Monitoring & Exploration Technologies, and Computational Landscape Ecology for support during the field campaign and the consequent sample preparation and lab analysis.

References

Abdi, H.: Partial Least Square Regression - PLS-Regression, in Encyclopedia of Measurement and Statistics, edited by N. Salkind, ThousandOaks (CA): Sage., 2007.

Adamchuk, V. I. and Viscarra Rossel, R. A.: Development of On-the-Go Proximal Soil Sensor Systems, in Proximal Soil Sensing. Progress in Soil Science, edited by R. A. Viscarra Rossel, A. McBratney, and B. Minasny, pp. 15–28, Springer, Dordrecht., 2010.

Altermann, M., Rinklebe, J., Merbach, I., Körschens, M., Langer, U. and Hofmann, B.: Chernozem — Soil of the Year 2005, J. Plant Nutr. Soil Sci., 168, 725–740, doi:10.1002/jpln.200521814, 2005.

Ben-Dor, E., Irons, J. A. and Epema, A.: Soil Spectroscopy, in Manual of Remote Sensing, edited by A. Rencz, p. 111−188, J. Wiley & Sons, Inc., NewYork., 1999.

Ben Dor, E., Ong, C. and Lau, I. C.: Reflectance measurements of soils in the laboratory: Standards and protocols, Geoderma, 245-246, 112–124, doi:10.1016/j.geoderma.2015.01.002, 2015.

Beleites, C., Baumgartner, R., Bowman, C., Somorjai, R., Steiner, G., Salzer, R., & Sowa, M. G.: Variance reduction in estimating classification error using sparse datasets. Chemometrics and Intelligent Laboratory Systems, 79(1–2), 91–100, https://doi.org/10.1016/j.chemolab.2005.04.008, 2005.

Bird, M., Brookes, P. C., Chenu, C., Jastrow, J. D., Lal, R., Lehmann, J., Donnell, A. G. O., Parton, W. J., Whitehead, D. and Zimmermann, M.: The knowns , known unknowns and unknowns of sequestration of soil organic carbon, Agric. Ecosyst. Environ., 164, 80–99, 2013.

Bjørsvik, H.-R. and Martens, H.: Data Analysis: Calibration of NIR Instruments by PLS Regression, in Handbook of Near-Infrared Analysis, edited by D. A. Burns and E. W. Ciurczak, pp. 189 –205., 2008.

Bogrekci, I. and Lee, W. S.: EFFECTS OF SOIL MOISTURE CONTENT ON ABSORBANCE SPECTRA OF SANDY SOILS IN SENSING PHOSPHORUS CONCENTRATIONS USING UV-VIS-NIR SPECTROSCOPY, Trans. Asabe, 49(4), 1175–1180, 2006.

Brown, J. D. and Heuvelink, G. B. M.: Assessing Uncertainty Propagation through Physically Based Models of Soil Water Flow and Solute Transport, in Encyclopedia of Hydrological Sciences, edited by M. G. Anderson, pp. 1181–1195, Wiley, Chicester, UK., 2006.

Charrad, M., Ghazzali, N., Boiteau, V. and Niknafs, A.: NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set, J. Stat. Softw., 61(6), 1–36 [online] Available from: http://www.jstatsoft.org/v61/i06/, 2014.

Conforti, M., Castrignanò, A., Robustelli, G., Scarciglia, F., Stelluti, M. and Buttafuoco, G.: Laboratory-based Vis-NIR spectroscopy and partial least square regression with spatially correlated errors for predicting spatial variation of soil organic matter content, Catena, 124, 60–67, doi:10.1016/j.catena.2014.09.004, 2015.

Croft, H., Kuhn, N. J. and Anderson, K.: On the use of remote sensing techniques for monitoring spatio-temporal soil organic carbon dynamics in agricultural systems, Catena, 94, 64–74, doi:10.1016/j.catena.2012.01.001, 2012.

Dalal, R. C. and Henry, R. J.: Simultaneous Determination of Moisture, Organic Carbon, and Total Nitrogen by Near Infrared Reflectance Spectrophotometry1, Soil Sci. Soc. Am. J., 50(1), 120, doi:10.2136/sssaj1986.03615995005000010023x, 1986.

Dierke, C. and Werban, U.: Geoderma Relationships between gamma-ray data and soil properties at an agricultural test site, Geoderma, 199, 90–98, doi:10.1016/j.geoderma.2012.10.017, 2013.

Filzmoser, P.: Identification of multivariate outliers: A performance study, Austrian J. Stat., 34(2), 127–138 [online] Available from: http://www.stat.tugraz.at/AJS/ausg052/052Filzmoser.pdf, 2005.

Filzmoser, P. and Gschwandtner, M.: mvoutlier: Multivariate Outlier Detection Based on Robust Methods, 2017.

Ge, Y., Morgan, C. L. S., Grunwald, S., Brown, D. J. and Sarkhot, D. V.: Comparison of soil reflectance spectra and calibration models obtained using multiple spectrometers, Geoderma, 161(3-4), 202–211, doi:10.1016/j.geoderma.2010.12.020, 2011.

Gholizadeh, A., Boruvka, L., Sbaerioon, M., Vasat, R.: Visible, near-infrared, and mid-infrared spectroscopy applications for soil assessment with emphasis on soil organic matter content and quality: State-of-the-art and key issues. Applied Spectroscopy, 67/12, 1349-1362, 2013.

Guio Blanco, C. M., Brito Gomez, V. M., Crespo, P. and Ließ, M.: Spatial prediction of soil water retention in a Páramo landscape: Methodological insight into machine learning using random forest, Geoderma, 316(November 2017), 100–114, doi:10.1016/j.geoderma.2017.12.002, 2018.

Hastie, T., Tibshirani, R. and Friedman, J. H.: The Elements of Statistical Learning, 2nd ed., Springer, New York., 2009.

Heuvelink, G. B. M.: Propagation of error in spatial modelling with GIS, in Geographical Information Systems, edited by P. A. Longley, M. F. Goodchild, D. J. Maguire, and D. W. Rhind, pp. 207–217, New York, John Wiley & Sons., 1999.

Islam, K., Singh, B. and McBratney, A. B.: Simultaneous estimation of several soil properties by ultra-violet, visible, and near-infrared reflectance spectroscopy, Aust. J. Soil Res., 41(6), 1101–1114, doi:10.1071/SR02137, 2003.

Jansen, M.: Prediction error through modelling concepts and uncertainty from basic data. Nutrient Cycling in Agroecosystems, 50(1), 247–253. https://doi.org/10.1023/A:1009748529970, 1998.

Jeong, G., Choi, K., Spohn, M., Park, S. J., Huwe, B. and Ließ, M.: Environmental drivers of spatial patterns of topsoil nitrogen and phosphorus under monsoon conditions in a complex terrain of South Korea, PLoS One, 12(8), 1–19, doi:10.1371/journal.pone.0183205, 2017.

Jiang, Q., Chen, Y., Guo, L., Fei, T. and Qi, K.: Estimating Soil Organic Carbon of Cropland Soil at Different Levels of Soil Moisture Using VIS-NIR Spectroscopy, Remote Sens., 8(9), 755, doi:10.3390/rs8090755, 2016.

Johnson, M. G.: Soil carbon sequestration: Quantifying this ecosystem service, Present. Oregon Soc. Soil Sci. Annu. Meet. Newport, OR, Febr. 28 - 29, 2008.

Kennard, R. W. and Stone, L. A.: Computer Aided Design of Experiment, Technometrics, 11(1), 137–148, 1969.

Knadel, M., Thomsen, A., Schelde, K. and Greve, M. H.: Soil organic carbon and particle sizes mapping using vis-NIR, EC and temperature mobile sensor platform, Comput. Electron. Agric., 114(April), 134–144, doi:10.1016/j.compag.2015.03.013, 2015.

Kokaly, R. F.: Investigating a physical basis for spectroscopic estimates of leaf nitrogen concentration, Remote Sens. Environ., 75(2), 153–161, doi:10.1016/S0034-4257(00)00163-2, 2001.

Körschens, M. and Pfefferkorn, A.: Bad Lauchstädt - The Static Fertilization Experiment and other Long-Term Field Experiments, edited by UFZ – Umweltforschungszentrum Leipzig-Halle GmbH., 1998.

Kuang, B. and Mouazen, A. M.: Non-biased prediction of soil organic carbon and total nitrogen with vis e NIR spectroscopy , as affected by soil moisture content and texture, Biosyst. Eng., 114(3), 249–258, doi:10.1016/j.biosystemseng.2013.01.005, 2013.

Kuhn, M.: caret: Classification and Regression Training, 2017.

Kuhn, M. and Johnson, K.: Applied Predictive Modeling [Hardcover]., 2013.

Lal, R.: Soil Carbon Sequestration Impacts on Global Climate Change and Food Security, Science (80-.)., 304, 1623 – 1627, doi:DOI: 10.1126/science.1097396, 2004.

Lê, S., Josse, J. and Husson, F.: FactoMineR: An R Package for Multivariate Analysis, J. Stat. Softw., 25(1), 1–18, doi:10.1016/j.envint.2008.06.007, 2008.

Leone, A. P., Viscarra Rossel, R. A., Amenta, P. and Buondonno, A.: Prediction of Soil Properties with PLSR and vis- NIR Spectroscopy : Application to Mediterranean Soils from Southern Italy, Curr. Anal. Chem., 8(February 2014), 283 – 299, doi:10.2174/157341112800392571, 2012.

Lorenz, K. and Lal, R.: Soil Organic Carbon- An Appropriate Indicator to Monitor Trends of Land and Soil Degradation within the SDG Framework?, edited by S. M. Starke and K. Ehlers, Umweltbundesamt., 2016.

Loum, M., Diack, M., Ndour, N. Y. B. and Masse, D.: Effect of the Continuum Removal in Predicting Soil Organic Carbon with Near Infrared Spectroscopy (NIRS) in the Senegal Sahelian Soils, Open J. Soil Sci., 06(09), 135–148, doi:10.4236/ojss.2016.69014, 2016.

Martens, H. and Næs, T.: Multivariate Calibration, JohnWiley & Sons, Chichester, United Kingdom, UK., 1989.

McBratney, A. B., Stockmann, U., Angers, D. A., Minasny, B. and Field, D. J.: Challenges for Soil Organic Carbon Research, in Soil Carbon. Progress in Soil Science, edited by A. E. Hartemink and K. McSweeney, p. 57, Springer International Publishing Switzerland., 2014.

Meersmans, J., Wesemael, B. Van and Molle, M. Van: Determining soil organic carbon for agricultural soils : a comparison between the Walkley & Black and the dry combustion methods (north Belgium), Soil Use Manag., 25, 346–353, doi:10.1111/j.1475-2743.2009.00242.x, 2009.

Merbach, I. and Schulz, E.: Long-term fertilization effects on crop yields , soil fertility and sustainability in the Static Fertilization Experiment Bad Lauchstädt under climatic conditions 2001 – 2010, Arch. Agron. Soil Sci., 59(8), 1041 – 1057, doi:10.1080/03650340.2012.702895, 2013.

Minasny, B., McBratney, A. B., Bellon-Maurel, V., Roger, J.-M., Gobrecht, A., Ferrand, L. and Joalland, S.: Removing the effect of soil moisture from NIR diffuse reflectance spectra for the prediction of soil organic carbon, Geoderma, 167-168, 118–124, doi:10.1016/j.geoderma.2011.09.008, 2011.

Molinaro, A.M., Simon, R., Pfeiffer, R.M.: Prediction error estimation: a comparison of resampling methods. Bioinformatics, 21(15), 3301–3307, 2005

Mortensen, P.: Myth: A partial least squares calibration model can never be more precise than the reference method..., NIR news, 25/ 3, 20-22, 2014.

Mouazen, A. M., Kuang, B., Baerdemaeker, J. De and Ramon, H.: Comparison among principal component , partial least squares and back propagation neural network analyses for accuracy of measurement of selected soil properties with visible and near infrared spectroscopy, Geoderma, 158(1-2), 23–31, doi:10.1016/j.geoderma.2010.03.001, 2010.

Mutanga, O. and Skidmore, A. K.: Continuum-removed absorption features estimate tropical savanna grass quality in situ, 3rd EARSEL Work. Imaging Spectrosc. Herrsching, 542–558, 2003.

Naes, T., IsakssonT., Fearn, T. and Davies, T.: A User Friendly Guide to Multivariate Calibration and Classification, 2002.

Nieder, R. and Benbi, D. K.: Carbon and Nitrogen in the Terrestrial Environment, Springer Netherlands., 2008.

Nduwamungu, C., Ziadi, N., Parent, L.-E., Tremblay, G.F., Thuriès, T.: Opportunities for, and Limitations of, Near Infrared Reflectance Spectroscopy Applications in Soil Analysis: A Review, Can. J. Soil Sci.. 89(5): 531-541, 2009.

Nocita, M., Stevens, A., Noon, C. and Van Wesemael, B.: Prediction of soil organic carbon for different levels of soil moisture using Vis-NIR spectroscopy, Geoderma, 199, 37–42, doi:10.1016/j.geoderma.2012.07.020, 2013.

Pilorget, C., Fernando, J., Ehlmann B. et al.: Wavelength dependence of scattering properties in the VIS–NIR and links with grain-scale physical and compositional properties, Icarus, 267, 296-314, 2016.

Pimstein, A., Notesco, G. and Ben-Dor, E.: Performance of Three Identical Spectrometers in Retrieving Soil Reflectance under Laboratory Conditions, Soil Sci. Soc. Am. J., 75(2), 746, doi:10.2136/sssaj2010.0174, 2011.

Poggio, L. and Gimona, A.: National scale 3D modelling of soil organic carbon stocks with uncertainty propagation - An example from Scotland, Geoderma, 232-234, 284-299, 2014.

Reeves, J. B. and Smith, D. B.: The potential of mid- and near-infrared diffuse reflectance spectroscopy for determining major- and trace-element concentrations in soils from a geochemical survey of North America, Appl. Geochemistry, 24(8), 1472–1481, doi:10.1016/j.apgeochem.2009.04.017, 2009.

Rinnan, Å., Berg, F. van den and Engelsen, S. B.: Review of the most common pre-processing techniques for near-infrared spectra, TrAC - Trends Anal. Chem., 28(10), 1201–1222, doi:10.1016/j.trac.2009.07.007, 2009.

Savitzky, A. and Golay, M. J. E.: Smoothing and Differentiation of Data by Simplified Least Squares Procedures., Anal. Chem., 36(8), 1627–1639, doi:10.1021/ac60214a047, 1964.

Schmidt, K. S. and Skidmore, A. K.: Exploring spectral discrimination of grass species in African rangelands, Int. J. Remote Sens., 22(17), 3421–3434, doi:10.1080/01431160152609245, 2001.

Schulz, E.: Static Fertilization Experiment Bad Lauchstädt', , www.ufz.de/index.php?en=37010 , 2017.

Schwartz, G., Eshel, G., Ben-Dor, E.: Reflectance Spectroscopy as a Tool for Monitoring Contaminated Soils. In: S. Pascucci, editor. Soil Contamination. New York: InTech, Pp. 67-90. doi: 10.5772/23661, 2011.

Smith, B. C.: Quantitative spectroscopy : theory and practice, Academic Press. [online] Available from: https://www.sciencedirect.com/science/book/9780126503586 (Accessed 18 January 2018), 2002.

Stenberg, B. and Viscarra Rossel, R. A.: Diffuse Reflectance Spectroscopy for High-Resolution Soil Sensing, in Proximal Soil Sensing. Progress in Soil Science, edited by Viscarra Rossel R. A., McBratney A., and Minasny B, pp. 29–47, Springer, Dordrecht., 2010.

Stenberg, B., Viscarra Rossel, R. A., Mouazen, A. M. and Wetterlind, J.: Visible and Near Infrared Spectroscopy in Soil Science, Adv. Agron. Vol 107, 107(10), 163–215, doi:10.1016/s0065-2113(10)07005-7, 2010.

Stevens, A. and Ramirez Lopez, L.: An introduction to the prospectr package, , 1–22 [online] Available from: http://cran.r-project.org/web/packages/prospectr/vignettes/prospectr-intro.pdf, 2014.

Stevens, A., Nocita, M., Tóth, G., Montanarella, L. and van Wesemael, B.: Prediction of Soil Organic Carbon at the European Scale by Visible and Near InfraRed Reflectance Spectroscopy, PLoS One, 8(6), doi:10.1371/journal.pone.0066409, 2013.

Stockmann, U., Adams, M. A., Crawford, J. W., Field, D. J., Henakaarchchi, N., Jenkins, M., Minasny, B., Mcbratney, A. B., Remy, V. De, Courcelles, D., Singh, K., Wheeler, I., Abbott, L., Angers, D. A., Baldock, J., Summers, D., Lewis, M., Ostendorf, B. and Chittleborough, D.: Visible near-infrared reflectance spectroscopy as a predictive indicator of soil properties, Ecol. Indic., 11, 123–131, doi:10.1016/j.ecolind.2009.05.001, 2011.

Swarbrick, B.: Near-infrared spectroscopy and its role in scientific and engineering applications, in Handbook of Measurement in Science and Engineering, edited by M. Kutz, John Wiley & Sons, Inc., Hoboken, NJ, USA., 2016.

Tekin, Y., Ulusoy, Y., Tümsavas, Z. and Mouazen, A. M.: Online Measurement of Soil Organic Carbon as

Correlated with Wheat Normalised Difference Vegetation Index in a Vertisol Field, Sci. World J., 2014.

VDLUFA: Methodenbuch Band I Die Untersuchung von Böden, in Das VDLUFA Methodenbuch, VDLUFA-Verlag, Darmstadt., 2012.

Varma, S. and Simon, R.: Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics, 7, 91. https://doi.org/10.1186/1471-2105-7-91, 2006

Viscarra Rossel, R. A., Walter, C. and Fouad, Y.: Assessment of two reflectance techniques for the quantification of the within-field spatial variability of soil organic carbon, Precis. Agric., (January), 697–703 [online] Available from: <Go to ISI>://WOS:000186384100105, 2003.

Viscarra Rossel, R. A., McGlynn, R. N. and McBratney, A. B.: Determining the composition of mineral-organic mixes using UV-vis-NIR diffuse reflectance spectroscopy, Geoderma, 137(1-2), 70–82, doi:10.1016/j.geoderma.2006.07.004, 2006a.

Viscarra Rossel, R. A., Walvoort, D. J. J., McBratney, A. B., Janik, L. J. and Skjemstad, J. O.: Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties, Geoderma, 131(1-2), 59–75, doi:10.1016/j.geoderma.2005.03.007, 2006b.

Viscarra Rossel, R. A., Behrens, T., Ben-Dor, E., Brown, D. J., Demattê, J. A. M., Shepherd, K. D., Shi, Z., Stenberg, B., Stevens, A., Adamchuk, V., Aichi, H., Barthès, B. G., Bartholomeus, H. M., Bayer, A. D., Bernoux, M., Böttcher, K., Brodský, L., Du, C. W., Chappell, A., Fouad, Y., Genot, V., Gomez, C., Grunwald, S., Gubler, A., Guerrero, C., Hedley, C. B., Knadel, M., Morrás, H. J. M., Nocita, M., Ramirez-Lopez, L., Roudier, P., Campos, E. M. R., Sanborn, P., Sellitto, V. M., Sudduth, K. A., Rawlins, B. G., Walter, C., Winowiecki, L. A., Hong, S. Y. and Ji, W.: A global spectral library to characterize the world's soil, Earth-Science Rev., 155(February), 198–230, doi:10.1016/j.earscirev.2016.01.012, 2016.

Volkan Bilgili, A., van Es, H. M., Akbas, F., Durak, A. and Hively, W. D.: Visible-near infrared reflectance spectroscopy for assessment of soil properties in a semi-arid area of Turkey, J. Arid Environ., 74(2), 229–238, doi:10.1016/j.jaridenv.2009.08.011, 2010.

Wang, Y., Lu, C., Wang, L., Song, L., Wang, R. and Ge, Y.: Prediction of Soil Organic Matter Content Using VIS / NIR Soil Sensor, Sensors & Transducers, 168(4), 113–119, 2014.

Wehrens, R.: Chemometrics with R - Multivariate Data Analysis in the Natural Sciences and Life Sciences, edited by G. P. Robert Gentleman, Kurt Hornik, Springer-Verlag Berlin Heidelberg., 2011.

Wickham, H. and Francois, R.: dplyr: A Grammar of Data Manipulation, 2017.