**Response to topical editor and referees: Miniaturised visible and near-infrared spectrometers for assessing soil health indicators in mine site rehabilitation by Shen et al.**

We thank the topical editor and referees for reviewing our manuscript and for contributing to the public discussion. Below we provide our responses indicating the changes we made in the revision, in separate sections for each reviewer and (in blue text, preceded by **Authors:**).

# Topical Editor: Comments to the author

Please, provide an improved version of the manuscript following all the critics and comments of reviewers, in line with your responses, to better assess the suitability for publication. Reviewers have been highly critical of the manuscript.

**Authors:** We have revised the manuscript according to the comments made in the open discussion with the referees. In the public discussion, we uploaded our author comments, which address, point-by-point, all of the reviewer's comments and criticisms. When we thought that a comment was unfair or inaccurate, or when we disagreed with it, we explained why and provided reason, justification, and where possible, evidence for our argument. For all other comments and criticism, we proposed how we might revise. In our responses here, we do not repeat points that we have already addressed in the public discussion. Here, we only address to the comments, suggestions and criticisms that we proposed to revise.

# Referee 1

We addressed the most critical comments from referee 1 in the public discussion. Below, we address comments that relate directly to our revision. They can be summarised as follows: (1) emphasise the research gap, originality and clarify aims, (2) improve description of the sampling and experimental design, (3) provide a rationale for using the different algorithms for the modelling, (4) enrich the discussion by describing the limitations of the predictive models, (5) improve overall readability of the manuscript. Below we describe the revisions made.

## Research gap, aims and originality

**Authors:** We improved the Introduction to clarify and highlight the research gap as follows (P3 L84): "As a result, there is growing interest in using miniaturised spectrometers for characterising soil properties (Tang et al., 2020; Ng et al., 2020; Sharififar et al., 2019). They provide a unique opportunity to develop a portable and cost-efficient soil health diagnostic capacity for mine site rehabilitation. However, these spectrometers have narrower spectral ranges and lower resolutions than portable or research-grade spectrometers, which are more commonly used to model various soil properties (Stenberg et al., 2010; Soriano-Disla et al., 2014). A narrower spectral range and reduced spectral resolution might detriment a miniaturised spectrometer's capacity for estimating soil properties and developing a cost-effective soil health diagnostic solution. Therefore, we must thoroughly assess these spectrometers, the repeatability of their measurements and their capacity to accurately estimate a wide range of soil physical, chemical, and biological properties deemed to be indicators of soil health. We have not found any published reports on such assessments in the literature."

We revised the last paragraph in the Introduction to more clearly state our aims, as follows (P4 L93): "Thus the objective of our study was to assess various commercially available miniaturised visible and NIR spectrometers (both independently and in combination) and one portable research-grade spectrometer for the capacity to estimate a wide range of soil chemical, physical and biological properties typically used to assess soil health in Australian mine site soil rehabilitation. Specifically, the aims were to:

– evaluate the repeatability of the spectroscopic measurements with each spectrometer,

– evaluate the modelling of 29 soil physical, chemical, and biological properties using the spectra from each spectrometer and combinations of the visible and a NIR spectrometer with seven statistical and machine learning algorithms,

– quantify the accuracy of the model estimates and the effect of repeatability on the estimates, and

– derive an overall score for each spectrometer and combinations tested, which accounts for their accuracy and repeatability.

The many soil properties, spectrometers and the assessment of repeatability and accuracy from modelling with various algo- rithms provide a comprehensive evaluation of spectroscopy for mine site soil rehabilitation."

To further clarify that the aims and scope of our study does not pertain to direct assessments of soil health, but rather the assessment of spectrometers for their capacity to estimate soil properties that are indicators of soil health in mine site rehabilitation, we revised the Discussion as follows (P22 L351): "We used miniaturised spectrometers to estimate 29 soil physical, chemical, and biological soil properties. Amongst these, soil organic carbon, K, P, pH, EC, B, Cu, Zn, Mn, Fe, Ca, Mg, Na, Al, sand, silt, clay, soil respiration ($CO_2$), available N (ammonium and nitrate) and BD are tier 1 soil health indicators in routine soil testing programs (Karlen et al., 2019; Bünemann et al., 2018). Although less commonly included in soil health assessment frameworks, microbial community composition is an under-utilised yet valuable metric for soil health assessment (Fierer et al., 2021). Changes in soil microbial communities are associated with processes that are important for soil health, such as changes in P availability (Hermans et al., 2017), soil pH (Delgado-Baquerizo et al., 2018), labile organic carbon pools (Ramírez et al., 2020), and soil moisture levels (Isobe et al., 2020). Our aim was not to derive detailed interpretations of the soil health indicators or an overall soil health index or score. Instead, we tested 29 soil properties commonly used as key indicators of soil health and used the spectra from a range of miniaturised and portable spectrometers to model them. The spectroscopic models accurately estimated 24 out of 29 soil properties (Table 5). Therefore, the spectra from the miniaturised spectrometers can accurately estimate soil properties that are key indicators of soil health. These estimates can could then inform the assessment and diagnosis of soil health. These findings suggest using miniaturised spectrometers to develop a cost-effective soil health diagnostic capacity for mine site soil rehabilitation. In a future study, we might investigate the derivation of soil health index based on the spectra or the predicted soil properties."

The originality of our study pertains to (i) the testing of many miniaturised spectrometers in the context of mine site soil rehabilitation. Other studies consider mainly agricultural soils (Ng et al., 2020; Sharififar et al., 2019; Tang et al., 2020); (ii) we extended the assessment to include the visible range and combined miniaturised a visible spectrometer with the NIR systems because the visible range provides unique spectral information; (iii) we considered the repeatability of the spectroscopic measurements and its effect on the models' estimates of the soil properties; (iv) we evaluated the inaccuracy (both bias and imprecision) of the model estimates, using seven statistical and machine learning algorithms, and derived an overall performance score considering repeatabiliy and accuracy of the estimates. In the Discussion we summarised the originality as follows (P23 L382): "Some studies have evaluated miniaturised NIR spectrometers for characterising soil properties (Tang et al., 2020; Ng et al., 2020; Sharififar et al., 2019) in agricultural and natural environments. However, our study presents the most comprehensive assessment. We tested a diverse range of soil properties, considered key indicators of soil health; many spectrometers (one visible, three NIR, three combinations and a portable spectrometer); a range of statistical and machine learning algorithms; evaluated the repeatability of the measurements and their effect on the models' estimates, and the accuracy of the models. Our study also extends the application of miniaturise spectrometers for mine site soil rehabilitation."

## Sampling and experimental design

**Authors:** The soil sampling design aimed to characterise different mining contexts, commodity types, climates, soil types, and vegetation assemblages, not ‘vertical spatial variability’. Therefore, we collected stockpiled and undisturbed reference soil samples from seven mine sites (Table 1), with widely ranging soil property values (Table 3). Note that vertical variability in the stockpiles is less important because these are disturbed soils. We have revised Section 2.1 to better explain the rationale for the sampling design as follows (P5 L112). “We designed the sampling effort to cover a broad range of mining contexts. We sampled top-layer stockpiled and undisturbed reference soil samples from seven mines with differing soil types, climates, vegetation assemblages and commodity.”, and clarified that samples collected at depth 50–70 cm also correspond to top-layer soils as follows (P6 L118): “At the youngest stockpile at each mine, five additional samples were taken from the 50–70 cm depth,, which correspond roughly to the top layer of the original soil before stockpiling.”

To better explain the experimental design, we included a new figure that summarises the data collection and experiments performed. The flow diagram provides a high-level description of the study design. See P8 Fig. 2.

## Rationale for using different algorithms

**Authors:** To clarify the use of multiple algorithms in our assessments we have added text to Section 2.4.1 as follows (P8 L174): “Multivariate modelling is fundamental for assessing the performance of spectrometers to predict soil properties. This type of modelling, and particularly machine learning, is largely dependent on the data set and there is no single ‘best’ method for all contexts and applications. Using a single algorithm could lead to inaccurate conclusions. To prevent both over- or understating the capability of these spectrometers, we used seven statistical and machine learning algorithms that have been reported for soil spectroscopic modelling (Viscarra Rossel and Behrens, 2010; Liu et al., 2016; Yang et al., 2022; Song et al., 2021). These were partial least square regression (PLSR) (Wold et al., 2001), random forest (RF) (Breiman, 2001), support vector machines (SVM) (Vapnik, 1999), Cubist (Quinlan et al., 1992) extreme gradient boosting (XGBoost) (Chen et al., 2015), and Gaussian process regression with linear (GPRL) and polynomial (GPRP) kernels (Rasmussen, 2003). They account for linear responses (PLSR) to more complex, non-linear responses (e.g. SVM), and the bases of the algorithms are fundamentally differently: statistical (PLSR), tree-based (cubist, RF, XGBoost), Gaussian process-based (GPRL and GPRR), and support vector methods (SVM). Viscarra Rossel and Behrens (2010) described these algorithms and their implementation in soil spectroscopic modelling."

## Limitations of predictive models

**Authors:** In the revision, we discussed the limitations of deriving soil health using the estimated soil properties from the predictive models and the accuracy and repeatability of the sensors as follows (P23 L366):

"Some soil properties (bacterial richness and diversity, P, and Zn and ammonium-N) could not be well estimated with any of the spectrometers or combinations (Fig. 5). Therefore, it might be necessary to investigate other methods to measure them. Although we could estimate 24 soil properties with moderate or higher accuracy ($\rho_c \geq 0.65$), the estimates are outcomes from empirical models, which possess error ((Table 5). However, the advantage of the spectroscopic method, compared to the more accurate conventional analytical measurements, is that spectroscopy is rapid and cost effective (Li et al., 2022), allowing an order of magnitude more (spatial and temporal) measurements, which on the whole might serve to better assess and monitor soil health. Another advantage of spectroscopy is that a single spectrum can be used to estimate many soil properties. These make soil spectroscopy well suited for increasingly large-scale soil rehabilitation where large volumes of observation data are needed but overly expensive and time-consuming to obtain via conventional soil analysis."

"The accuracy of the estimates of electrical conductivity and most of the soil chemical properties from the $A_{350–830}$ spectrometer was markedly poorer than the NIR spectrometers (Fig. 5), indicating that the 350–830 nm range does not hold sufficient chemical information to produce stable models for estimating those soil properties. The poor repeatability of the $A_{350–830}$ spectrometer's measurements in the 350–500 nm range (Fig. 3) also affected the precision of the spectrometer combinations (reduced Repeatability score of the combinations in Table 4). With a more repeatable visible spectrometer, the spectrometer performance score, $e$, of the combined spectrometers would improve."

### Overall readability of the manuscript

**Authors:** In the revision, we provide a new Fig. 2 (P8) to help readers better understand the experiments that we performed. We also restructured subsection 2.4 Spectroscopic modelling (P7), and section 3 Results (P11) and improved the headings to help readers better understand the experiments conducted and connect experiments to corresponding results. We have reorganised the discussion slightly and included subheadings to better guide readers. Finally, we edited the manuscript heavily to reduce wordiness, improve grammar and its readability.

## Referee 2

We responded to all the comments from referee 2 in the public discussion. Below, we describe the revisions that we made, which are summarised as follows: (1) better explain the spectra in Fig. 3 (Fig. 2 in the original manuscript), (2) improve description of the experimental design, (3) improve explanation of the validation procedure with 56 samples; (4) define Lin's concordance correlation, (5) improve the analysis of prediction errors and emphasise the implications.

## Explanation of spectra in Figure 3

**Authors:** We re-drew this figure and labelled each column for easier referencing. We also revised the caption of Fig.3 (P14) to better explain what the figure shows: "Mean and difference spectra of the spectral replicates of the data from the 56 sampling plots. (a) average reflectance spectra of the two replicates (Fig. 2). (b) difference between the two replicates calculated using Equation (1). (c) combined average reflectance spectra from (a)."

## Experimental design

**Authors:** We revised the manuscript to improve the clarity of our experimental design in response to an earlier comment from referee 1. We drew a new figure to illustrate and explain the experiments with a flow diagram, which also links the methods to our results (See P8 new Fig. 2). For further explanation please see our response to referee 1 on sampling and experimental design.

## Validation procedure

**Authors:** Our results are based on both modelling of the 56 data from plots and modelling on all 280 data from subplots. Except that the modelling on the 280 data from subplots was performed using a 10-fold-leave-plot-out cross validation and using only the best algorithm and spectrometers and combinations from modelling of the 56 data. We had shown that the best algorithm and spectrometers performed similarly on the 56 plot samples and on the 280 subplot samples (original MS, P18 L301). There are two reasons for using the aggregated 56 plot samples rather than all 280 subplot samples for the the spectroscopic modelling: 1) soil samples from a single plot were fairly similar because they were sampled from a 5 m $\times$ 5 m subplots (see Methods: Sampling Design) ; 2) the eight spectrometers and combinations, seven algorithms, 29 soil properties and the replicates, and validation with 10-fold cross validation, meant that the experiments were complex and needed significant computation. Therefore, for computational efficiency, we opted for using the aggregated plot data. Even then we needed to run the experiments on a super computer (Pawsey, `https://pawsey.org.au/`) using 100 CPU cores. We revised the description of the experimental design in Section 2.4 as follows (P7 L169): "Given the large number of soil properties, spectrometers, algorithms, and the assessment of both spectrometer repeatability and model accuracy, the experiments became extremely complex and computationally intensive. To improve the computational efficiency of the study, we aggregated the data from the 280 subplots into 56 plots by averaging the spectra and soil properties (Fig. 2). Subplot samples were similar and we assumed that aggregating them would not seriously affect the variability in the data and the modelling. Conclusions drawn from the results of the 56 plot data were compared to results from the 280 subplot data, but using only the best spectrometers and algorithm (see below, subsection 'Assessment on data from subplots')."

## Lin's concordance correlation coefficient

**Authors:** Of course, we did cite the original paper by Lin et al., but in the revision we now also include the definition of concordance correlation coefficient at its first occurrence (P9, L207) as follows: "The $\rho_c$ is a unit invariant coefficient that measures the difference between the measured and estimated values and their deviation from a 45-degree line of perfect agreement, evaluating both precision and bias. It ranges from -1 to 1, with 1 denoting perfect agreement."

## Analysis and implications of prediction errors

**Authors:** Please note we used suitable statistical metrics to evaluate the models and their estimation errors: the concordance correlation coefficient ($\rho_c$), the root mean squared error (RMSE), mean error (ME), and standard deviation of the error (SDE). We only used $\rho_c$ as the main metric for analysing the prediction errors, because $\rho_c$ is unit invariant, ranges from -1 to 1, and measures both bias and imprecision, making the comparison across the 29 soil properties (with different units) possible. We also assigned levels of accuracy based the $\rho_c$ values (poor for $\rho_c < 0.65$, moderate for 0.65

$\leq \rho_c < 0.8$, substantial for $0.8 \leq \rho_c < 0.9$, and near-perfect for $0.9 \leq \rho_c < 1.0$) and used them in the assessments (Fig.4, 5, 7 and Table 5). In the revision we further clarified the usage of $\rho_c$ as follows (P9 L210): "... The $\rho_c$ was used as the main metric for the assessments because it allows the comparison across soil properties with different units possible."

The RMSE, ME and SDE are, of course, unit variant, making direct comparisons between properties impossible. But for a single property, ME measures bias, SDE measures imprecision, and RMSE measures inaccuracy. The relationship is: $\text{RMSE}^2 = \text{ME}^2 + \text{SDE}^2$. In the revision we further explained our use of these statistics as follows (P10 L221): "Because RMSE, ME, and SDE are unit variant, they are not suitable for comparing errors across soil properties with differing units. They were used to quantify the overall inaccuracy (RMSE), the bias (ME), and the imprecision (SDE) of the estimates for each soil property." We have a table in the manuscript (Table 5) which shows the RMSE, ME, SDE and $\rho_c$ for the spectrometers (or combinations) that had the best overall performance ($e$). We now explain the overall results with relation to these metrics, as follows (P19 L317): "The inaccuracy of the soil property estimates, quantified with the RMSE, was largely due to imprecision (SDE) and not bias (ME)."

# References

Wartini Ng, Linca Anggria, Adha Fatmah Siregar, Wiwik Hartatik, Yiyi Sulaeman, Edward Jones, Budiman Minasny, et al. Developing a soil spectral library using a low-cost nir spectrometer for precision fertilization in indonesia. Geoderma Regional, 22:e00319, 2020.

Amin Sharififar, Kanika Singh, Edward Jones, Frisa Irawan Ginting, and Budiman Minasny. Evaluating a low-cost portable nir spectrometer for the prediction of soil organic and total carbon using different calibration models. Soil Use and Management, 35(4):607–616, 2019.

Yijia Tang, Edward Jones, and Budiman Minasny. Evaluating low-cost portable near infrared sensors for rapid analysis of soils from south eastern australia. Geoderma Regional, 20:e00240, 2020.