Comments to “Dataset variability and carbonate concentration influence the performance of local visible-near infrared spectral models”

2023-07-25

Contents

1 General comments

1.1 Research question and contribution

The study tries to address the following research questions:

1. How do prediction errors of local spectral models (computed with samples from individual fields) differ from lab measurement errors?
2. “Does a general model computed [with data from all fields] improve the prediction on a target site with a poor local model performance?”
3. “What is the optimal variability in the dataset of a local model to achieve a good model performance?”
4. “Which field and soil characteristics (field size, soil texture, carbonate concentration) of the target site influence the performance of spectral models?”

The research questions (except question 3 if it is framed in terms of the variability of the target variable) are interesting, of methodological relevance — also in other disciplines than soil science — and fit well within the scope of SOIL. I like especially that the study synthesized data from many other studies, even though I think this analysis is misdirected. I also like that spectral modeling and validation seem to be of high quality and to have been conducted very thoughtfully (in particular because validation considers spatial autocorrelation). The manuscript is overall well written and clearly structured.

I have two major concerns:

1. **The analysis of how dataset variability controls model performance (research question 3) has major limitations and should be removed:** The first reason is that the analysis neglects that PRMSE, RPD, and R^2 are related to the CV simply by the way they are defined which means that we can learn much more about this relation through a simple simulation. The second reason is that measures of relative model performance (PRMSE, RPD, R^2) are of little practical relevance. The third reason is that model performance depends on the target variable identity, but the analysis neglects this and averages over all analyzed target variables.
2. **Data and code must be available to interpret, replicate and build upon the findings reported in the article and should be published**

I will discuss these points in the following sections.

1.2 The analysis of how dataset variability controls model performance (research question 3) has major limitations and should be removed

I completely agree that it is important to understand the controls of the predictive performance of spectral prediction models and also that it would be very useful if one could predict beforehand what predictive performance may be expected from such a model. However, I think that the analysis is misdirected for the following reasons (I use the same notation as you, i.e., RMSE is the RMSE computed on cross-validation or test data and RMSE_{cal} is the same computed on calibration data):

1. The analysis of relations between model performance (PRMSE or RPD) and target variable variability (CV) does not appropriately consider that these variables are related due to the way they are defined: PRMSE and CV are always positively related (except in degenerate cases where the standard deviation of the data is 0 or the RMSE is 0, see R code 3 below, or — of course — due to noise) because both PRMSE and CV have the target variable mean (μ) in the denominator. RPD and CV are always positively related when μ is positive (except in the degenerate cases where μ gets infinitely large or 0 or RMSE is 0, see R code 3 below) and negatively related when μ is negative (except in the degenerate cases where μ gets infinitely small or 0 or RMSE is 0). In my opinion, such a simple simulation tells us already more than an analysis of real data from actual model fits, which are also influenced by noise, and it can actually help to better

explain the results you have obtained with actual data. I will provide two examples, one for PRMSE and one for RPD:

- Example for PRMSE: The output of R code 3 shows that one can expect sudden breaks in the relation between PRMSE and CV for some combinations of RMSE, σ (standard deviation of the target variable in the dataset), and μ. Adapting the range to that observed for N (RMSE $\in (0.08, 0.3)$ (Fig. 3), $\sigma \in (0.2, 0.5)$ (Tab S1), $\mu \in (1.7, 2.9)$ (Tab S1)), one can see that the value ranges for N should not produce such breaks (See R code 4).

- Example for RPD: It is visible in Fig. 5, upper right panel (RPD vs CV), that the relation between RPD and CV differs between target variables — for instance, for pH, the slope RPD vs CV is steeper than for total C (whereas the CV range is smaller for pH and larger for total C). This can easily be explained by the small absolute RMSE for pH (see Fig. 3) which causes RPD to increase much stronger with CV than is the case for total C where a larger absolute RMSE (see Fig. 3) results in a smaller slope.

2. This also applies to R^2 because R^2 is directly related to the variance of the target variable: $\mathrm{R}^2 = 1 - \frac{SS_{err}}{SS_{tot}}$ and $\sigma^2 = \frac{1}{n} SS_{tot}$ (See R code 3 below; SS_{err} is the residual sum of squares of the model, SS_{tot} the total sum of squares, n is the sample size) and because R^2 is just a different way to express RPD: $\mathrm{R}^2 = 1 - \frac{SS_{err}}{SS_{tot}} = 1 - \frac{n \cdot \mathrm{RMSE}^2}{SS_{tot}} = 1 - \frac{n \cdot \mathrm{RMSE}^2}{n \cdot \sigma^2} = 1 - \frac{1}{\mathrm{RPD}^2}$. So the analysis for RPD and R^2 use the same variable, just transformed differently.

3. Model performance is strongly dependent on the target variable identity because the target variable identity defines the ranges for μ, σ, and RMSE (for example RMSE for pH will hardly ever be larger than 14, whereas RMSE for POXC, for example, can easily be larger than 14). Point 1 in this section has shown that these ranges are strong controls on the relations between PRMSE and CV and RPD and CV (see for example the relation of RPD vs CV for pH, as described above). The analysis neglects target variable identity; the computed models simply average over values for different target variables. This causes large prediction uncertainties and shows that it does not make sense to compute models which average relative estimates for model performance over all target variables.

4. I think that measures of relative model performance (PRMSE, RPD, R^2) are of little practical relevance to users of spectral prediction models, in contrast to estimates of absolute model performance (e.g., RMSE) because only absolute estimates (and consequently absolute uncertainty estimates) can be used in subsequent analyses or decision processes (e.g., to decide whether to apply fertilizer and how much, or to model nutrient limitations of crop plants). Estimates of relative model performance are indeed useful to compare how the relative performance differs, but this use case and variable is interesting only to better understand *why* the models differ

in predictive accuracy, not as a standalone measure of predictive accuracy.

5. Assuming there are no measurement errors, the performance of a spectral prediction model will depend on how strongly spectral variables are related to the target variable, how strong interference by other spectral features is, and how much the model needs to extrapolate to make predictions for given samples.

Putting the previous points together, I think that in order to learn something about the controls of the predictive performance of spectral prediction models, we need to:
1. Focus on estimates of absolute predictive performance.
2. Consider target variable identity.
3. Understand the factors which cause correlations between spectral variables and a target variable and the factors which confound or mask these relations.
4. Understand what errors are caused by extrapolation.

I think that all of these are really hard problems which can impossibly be addressed by one single study alone and I do not suggest that you should or can address all above points using the dataset at hand and knowledge which is available today. In any case, I think that your study gets stronger if all analyses related to section 3.5 (and corresponding sections in the discussion: sections 4.4 and 4.5) are removed and research question 3 is removed.

Instead, I think it would be optionally much more useful to expand the analysis on which spectral properties or other sample properties control RMSE (currently section 3.6; this analysis exactly addresses the four points I think we need to address mentioned above). This analysis makes up only a small part of the results and discussion at the moment. I know how daunting it is if a reviewer requests additional analyses which probably could amount to another manuscript and I want to make clear that I suggest this only to provide constructive criticism: not only telling you what I think is misdirected and why, but how a useful direction would look like in my opinion. In light of this, please consider my suggestion to extend this analysis as optional.

1.3 Data and code must be available to interpret, replicate and build upon the findings reported in the article and should be published

I strongly encourage you to publish data and code of your analysis in a repository for the following reasons:

1. In the previous section, I argued that to better understand what controls model predictive performance, we need to understand the factors which cause correlations between spectral variables and a target variable and the factors which confound or mask these relations. Such an analysis will only have high enough accuracy if data from many local models are available, many more than can be provided by a single study.
2. The code (with package versions) is necessary to fully understand and reproduce (or replicate on new data) the spectral preprocessing and modeling.

3. Studies have repeatedly shown that unless data are published along the original publication, the probability that they won't be accessible is rather high (e.g., Tedersoo et al. (2021)).
4. There should be no legal reason prohibiting to publish the data and code (per the current data availability statement, p. 30).

2 Specific comments

1. l. 13 to 15: "… general models (combining all fields) for organic carbon, total carbon, total nitrogen, permanganate oxidizable carbon and pH using partial least squares regression. 24 out of 30 local models showed an accurate or even excellent performance (ratio of performance to deviation (RPD) > 2) …". I think the statement does not provide any useful information because it is not used to compare performance of models, but to make a standalone statement for one specific model (compare with point 4 in section "The analysis of how dataset variability controls model performance (research question 3) has major limitations and should be removed" of my review).

2. l. 15 to 16: "… and the root mean square errors (RMSE) of prediction were, except for pH, maximum five times higher than the lab measurement error." Whilst this statement is true (and RMSE is useful as a measure of model performance for standalone statements on model performance), I am not sure whether framing the comparison of model and lab RMSE only in terms of a ratio is useful (this also is the case in the results and discussion sections, for example) because one reason why the RMSE for pH is five times higher than the lab measurement error is that the pH lab measurement error is so small (see Fig. 3). Of course, all this makes only sense when one knows, what absolute error is tolerable.

3. l. 68 to 70: "However, it remains unclear what an optimal variability of a soil property in a project of local extent would be to achieve a high measurement accuracy in absolute values (low RMSE) but also relative to the range of the data (RPD)." Just to link this to my comments in section "The analysis of how dataset variability controls model performance (research question 3) has major limitations and should be removed": I think this question is misguided because it does not depend on the variability of the target variable how good a predictive model is (except that the variability of the target variable to predict should be larger than measurement errors), but on how well spectral variables are related to the target variable and masked by others. Of course, often a larger variability of the target variable implies more masking of spectral variables related to the target variable by other variables, but this larger variability of the target variable is not the cause.

4. l. 84: I think it makes sense to replace "measurement accuracy" by "measurement error".

5. l. 124 to 126: "To estimate the measurement error of SOC we took the

sum of the standard deviation of the total C measurement with the standard error of the inorganic C measurement because inorganic C measurement were for all samples done in triplicates." Shouldn't this rather be $\sqrt{\sigma_1^2 + \sigma_2^2}$ (instead of $\sigma_1 + \sigma_2$), where σ_1 is the standard deviation of the total C measurement and σ_2 the standard error of the inorganic C measurement?

6. l. 145 to 147: "The pre-processing techniques ... led to around 100 meaningful combinations that were tested in model building and the final pre-processing option was selected based on the lowest RMSE." Have you selected the model with the lowest computed average RMSE or have you applied the one-sigma-rule also when comparing RMSE between different preprocessing workflows? If it's the former, redoing this analysis while applying the one-sigma-rule would make sense to avoid overfitting and to consider the uncertainty in the RMSE estimate when discussing which preprocessing methods performed well. In particular, I expect that within one standard deviation multiple preprocessing workflows performed equally well and one should decide that not enough information is available to pick prefer a specific one of those.

7. l. 149 to 151: "Since all soil properties showed a limited skewness (see Table S1 in the supplementary material) that was always in the range of -2 $\leq$ skew $\leq$ 2 which was proposed as acceptable to assume a normal univariate distribution (George and Mallery, 2010) we consider the application of PLSR appropriate, especially since it is robust to minor deviations from a normal distribution (Goodhue et al., 2012)." Some of the statements are not appropriate: (1) The skewness limits are a bad way to test whether a distribution is a normal distribution (See R code 1 below where I simulate data from a distribution which obviously is not normal and nevertheless has a skewness within the specified boundaries). (2) The data itself do not have to follow a normal distribution for PLSR to provide valid results (See R code 2 below where I simulate data from the same non-normal distribution and show that a PLSR model with good performance can be computed on these data).

8. l. 166: "... we calculated the mean Euclidean distances between all samples ...". I assume the Euclidean distance was computed using the preprocessed spectral variables?

9. l. 174: "Since we used a cross-validation approach on the field scale, all models showed a very small bias." Please provide evidence for this claim, e.g. in Tab. 2 or in the supporting information.

10. l. 175 to 177: The sentence "RMSE was calculated according to Equation 1 where $\hat{y}_i$ is the prediction of the spectral model and y_i the actual measured value in the laboratory." may be better understandable if it is replaced by (additions in bold): "RMSE was calculated according to Equation 1 where $\hat{y}_i$ is the prediction of the spectral model **for sample** i and y_i the actual measured value in the laboratory **for the same sample**."

11. l. 184 to 185: "… RPD is the best parameter to compare models of different scales." Actually, it is not, but it is one option to do so. For example, as described in point 2 of section "The analysis of how dataset variability controls model performance (research question 3) has major limitations and should be removed", RPD is just a transformation of R^2.

12. l. 204 to 205: "The VIP method can deal with multicollinearity and is therefore suitable for the interpretation of spectral models (Baumann et al., 2021)." Baumann et al., 2021 did not analyze to what extent VIP can deal with multicollinearity.

13. l. 217: I assume "independence" should be "in dependence".

14. l. 221 to 222: "… we analyzed the influence of mean carbonate concentrations, soil texture and field size on the model metrics but only the mean carbonate concentration showed effects and is therefore presented in the results." This description needs to be much more detailed: (1) How did you analyze this (regression, correlation, in case of regression, which distribution was assumed for the target variable? (2) Which models were computed (what variables were included)? (3) How (based on what criteria) were variables included/excluded? (4) Was this analysis performed separately for different target variables (as suggested in section "The analysis of how dataset variability controls model performance (research question 3) has major limitations and should be removed" above)? (5) Is the uncertainty in the estimated RMSE small enough to not affect your results and have you checked this?
    In addition, please provide results for all models which were computed to support your claim that "only the mean carbonate concentration showed effects" (this can be shown in the supporting information).
    If this is an option for you, I strongly encourage you to expand this analysis, perhaps combining this with the analysis of the variable importance, because this is in my opinion the right direction to better understand the controls of model predictive performance (compare with my comments in section "The analysis of how dataset variability controls model performance (research question 3) has major limitations and should be removed").

15. l. 225 to 226: Please specify the versions of the R packages as this can be important to reproduce your analyses.

16. l. 246: "There was no pre-processing combination that proved to be suitable for all models of the same field (Fig. S2 in the supplementary material) or of the same soil property (data not shown)." Please provide evidence for this claim, e.g. in the supporting information.

17. l. 268 to 269 and elsewhere: "The field-specifically calculated R^2". Maybe "field-specific R^2" is more concise?

18. l. 283 to 284: "bit higher for POXC and substantially higher for pH

(Fig. 3)." Do yo mean a bit higher for pH and substantially higher for POXC? The figure does not seem to fit to the text, at least if this difference is meant in terms of absolute values (~0.15 to 0.25 pH units versus >20 POXC units) (compare also to point 2 in this section of my review).

19. Section 3.5: Are uncertainties in the PRMSE and RPD negligible for the results of your analyses? (Please note my comments in section "The analysis of how dataset variability controls model performance (research question 3) has major limitations and should be removed" above. I just wanted to mention that this point is not clear).

20. All analyses of the relation of R^2, PRMSE, RPD versus some other variable: You've described that you used a Weibull distribution for R^2 since R^2 must be in [0, 1]. However, prediction intervals shown in Fig. 5 and 7 contain values larger than 1 and thus do not comply with this assumption. Similarly, PRMSE and RPD cannot have negative values, but prediction intervals for PRMSE in Fig. 5 and 7 contain negative values.

21. l. 410 to 412: "Even though the local models of fields A and F had the lowest performance among all local models, they still showed, with exception of pH, approximate results which, depending on the research question might still provide useful information." I assume that "approximate" refers to the classification of RPD? In that case, I think the statement does not provide any useful information because it is not used to compare performance of models, but to make a standalone statement for one specific model (compare with point 4 in section "The analysis of how dataset variability controls model performance (research question 3) has major limitations and should be removed" of my review).

22. l. 423 to 424: "... we did not find a better model performance with increasing mean clay content ..." Please provide evidence for this claim, e.g. in the supporting information.

23. l. 556: The URL "https://doi.org/10.2136/sssaj2002.6400" is wrong. It should be "https://doi.org/10.2136/sssaj2002.6400a".

24. l. 698 to 699: "Lastly, we would like to thank the two anonymous reviewers for their constructive comments and suggestions." Thank you ;)

25. Fig. 1: Why are no density plots for carbonate concentration and soil texture variables shown here (or in the supporting information)?

26. Fig. 3: R^2, RMSE, RPD, and the lab measurement error are estimates and have uncertainties which should be added as error bars to the figures.

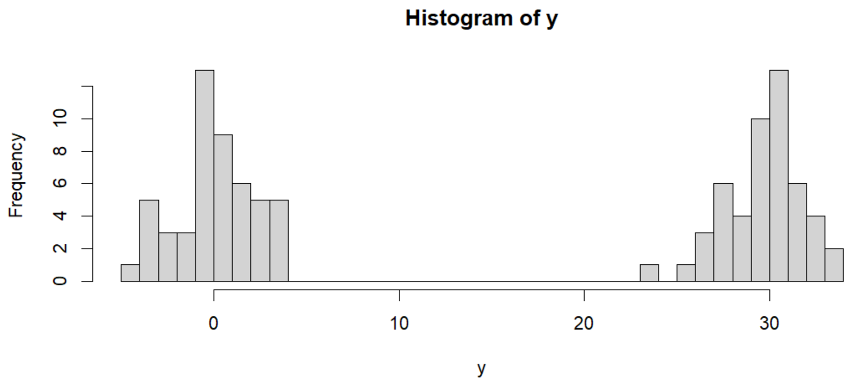
27. All figures (also in the supporting information) would benefit from (1) a higher image resolution, (2) larger legend keys, (3) thicker symbol lines or different symbol shapes such that it is possible to recognize better which points have which color.

3 R code

3.1 R code 1

Histogram of a non-normally distributed variable with -2 < skew < 2.

```
### simulate a distribution which is non-normal and nevertheless ###
### has -2 < skew < 2 ###
library(moments)
set.seed(54)
y <- c(rnorm(50, 0, 2), rnorm(50, 30, 2))
hist(y, breaks = 30) # clearly non-normal
```



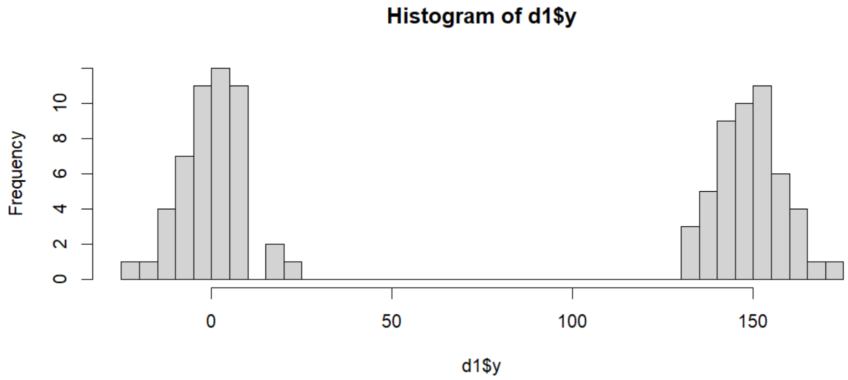
```
moments::skewness(y)
```

```
[1] -0.002923698
```

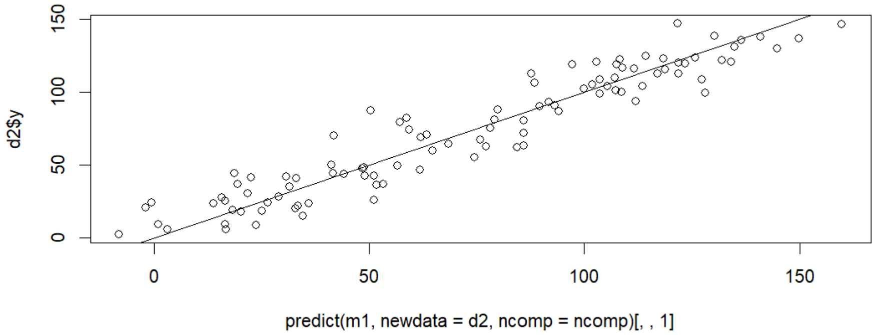
3.2 R code 2

PLSR can perform well also for a target variable which is non-normally distributed.

```
### show how PLSR performs on non-normal data ###
library(pls)
library(tibble)
library(magrittr)

set.seed(8787)

# simulate data
d1 <-
  tibble::tibble(
    x1 = c(rnorm(50, 0, 2), rnorm(50, 30, 2)),
    y = 5 * x1 + rnorm(100, 0, 1)
  ) %>%
  cbind(
    purrr::map_dfc(seq_len(100), function(i) {
      tibble::tibble(x = rnorm(100, 0, 2))
    }) %>%
      setNames(nm = paste0("x", seq_len(ncol(.)) + 1L))
  )

# show that y is non-normal
hist(d1$y, breaks = 30)
```



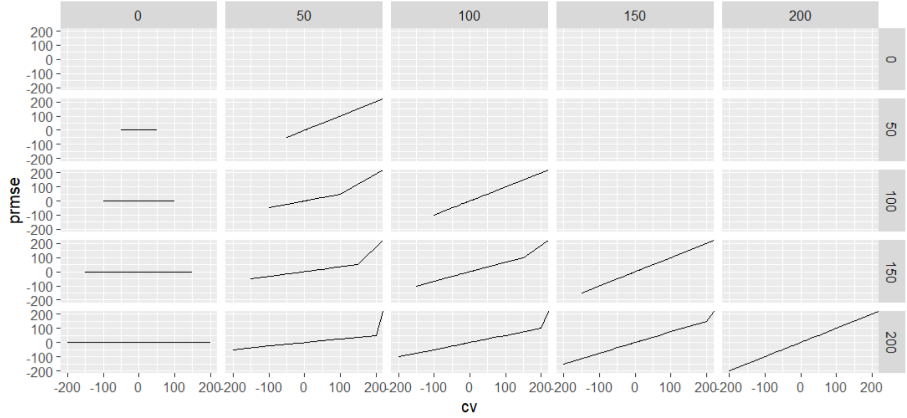
```
# fit plsr model
m1 <- pls::plsr(y ~ ., data = d1, scale = TRUE, center = TRUE, validation = "LOO")

# test on independent data
d2 <-
  tibble::tibble(
    x1 = runif(100, 0, 30),
    y = 5 * x1 + rnorm(100, 0, 1)
  ) %>%
  cbind(
    purrr::map_dfc(seq_len(100), function(i) {
      tibble::tibble(x = rnorm(100, 0, 2))
    }) %>%
      setNames(nm = paste0("x", seq_len(ncol(.)) + 1L))
  )

# select number of latent variables
ncomp <- selectNcomp(m1, method = "onesigma", plot = TRUE)

# "good" fit
plot(d2$y ~ predict(m1, newdata = d2, ncomp = ncomp)[, , 1])
abline(0, 1)
```



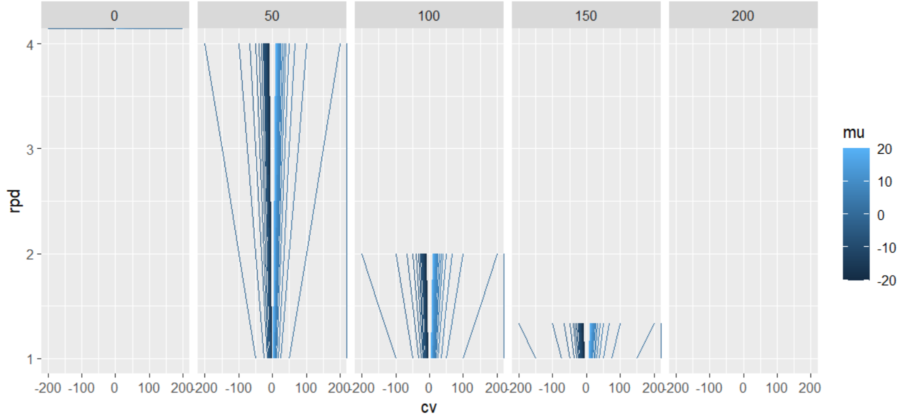
3.3 R code 3

In each plot, columns differentiate RMSE levels and rows levels for σ.

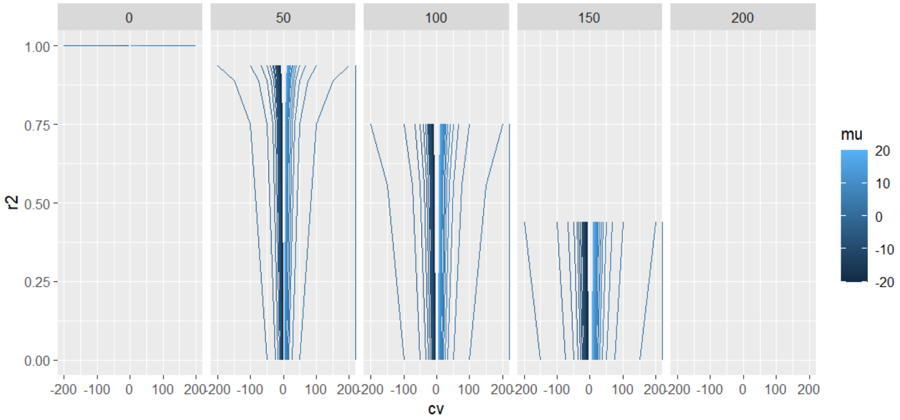
```
### simulate data to show that PRMSE, RPD, R^2 are always positively related ###
### to CV (except in degenerate cases or for R^2, RPD, when mu is negative) ###
library(ggplot2)
library(tibble)
library(magrittr)
library(dplyr)

# simulate data
d <-
  expand.grid(
    mu = seq(-20, 20, length.out = 41),
    sd = seq(0, 200, length.out = 5),
    rmse = seq(0, 200, length.out = 5)
  ) %>%
  tibble::as_tibble() %>%
  dplyr::mutate(
    cv = sd/mu,
    prmse = rmse/mu,
    rpd = sd/rmse,
    r2 = 1 - rmse^2/sd^2
  ) %>%
  dplyr::filter(rmse <= sd) # exclude cases where the model is worse
                            # than an intercept identical to the sample average

# PRMSE vs CV
d %>%
  dplyr::arrange(cv, prmse) %>%
  ggplot(aes(y = prmse, x = cv, group = paste0(sd, "_", rmse))) +
  geom_path() +
  facet_grid(sd ~ rmse)
```



```
# RPD vs CV
d %>%
  dplyr::arrange(cv, rpd) %>%
  ggplot(aes(y = rpd, x = cv, group = paste0(mu), color = mu)) +
  geom_path() +
  facet_grid( ~ rmse)
```



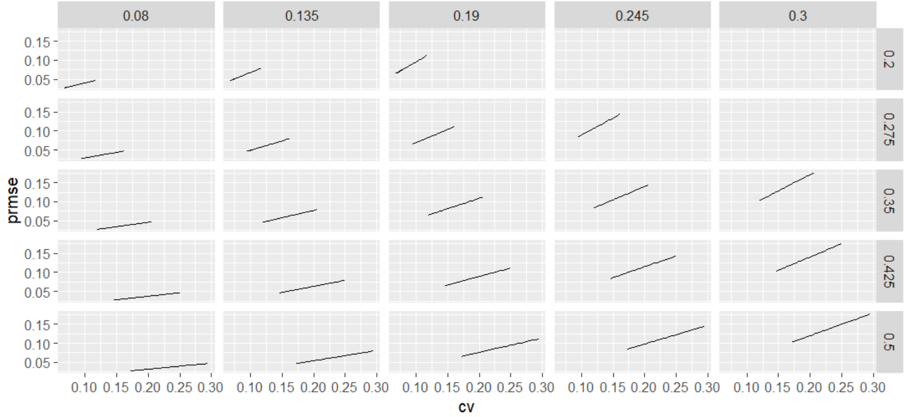
```
# R2 vs CV
d %>%
  dplyr::arrange(cv, r2) %>%
  ggplot(aes(y = r2, x = cv, group = paste0(mu), color = mu)) +
  geom_path() +
  facet_grid( ~ rmse)
```



3.4 R code 4

In the plot, columns differentiate RMSE levels and rows levels for σ.

```
### same as r code 3, but for ranges covered for N in the real data ###
library(ggplot2)
library(tibble)
library(magrittr)
library(dplyr)

d <-
  expand.grid(
    mu = seq(1.7, 2.9, length.out = 41),
    sd = seq(0.2, 0.5, length.out = 5),
    rmse = seq(0.08, 0.3, length.out = 5)
  ) %>%
  tibble::as_tibble() %>%
  dplyr::mutate(
    cv = sd/mu,
    prmse = rmse/mu,
    rpd = sd/rmse
  ) %>%
  dplyr::filter(rmse <= sd) # exclude cases where the model is worse
                            # than an intercept identical to the sample average

# PRMSE vs CV
d %>%
  dplyr::arrange(cv, prmse) %>%
  ggplot(aes(y = prmse, x = cv, group = paste0(sd, "_", rmse))) +
  geom_path() +
  facet_grid(sd ~ rmse)
```



References

Tedersoo, Leho, Rainer Küngas, Ester Oras, Kajar Köster, Helen Eenmaa, Äli Leijen, Margus Pedaste, et al. 2021. "Data Sharing Practices and Data Availability Upon Request Differ Across Scientific Disciplines." *Scientific Data* 8 (1): 192. https://doi.org/10.1038/s41597-021-00981-0.