Density of soil observations in digital soil mapping: A study in the Mayenne region, France

Thomas Loiseau, Dominique Arrouays, Anne C Richer-De-Forges, Philippe Lagacherie, Christophe Ducommun, Budiman Minasny

Density of soil observations in digital soil mapping: A study in the Mayenne region, France

Thomas Loiseau[1], Dominique Arrouays[1*], Anne Richer-de-Forges[1], Philippe Lagacherie[2], Christophe Ducommun[3], Budiman Minasny[4]

[1]INRAE, InfoSol, 45075, Orléans, France.

[2]LISAH, University of Montpellier, INRA, IRD, Montpellier SupAgro, 34060 Montpellier, France.

[3]Agrocampus-Ouest, 2, rue Le Nôtre 49045 ANGERS cedex 01, France.

[4] Sydney Institute of Agriculture, School of Life and Environmental Sciences, The University of Sydney, NSW 2006, Australia.

[*]Corresponding author: dominique.arrouays@inrae.fr

Abstract

The density of soil observations is a major determinant of digital soil mapping (DSM) prediction accuracy. In this study, we investigated the effect of soil sampling density on the performance of DSM to predict topsoil particle-size distribution in the Mayenne region of France. We tested two prediction algorithms, namely ordinary kriging (OK) and quantile random forest (QRF). The study area is a region of ~5000 km^2 with the highest density of field soil observations in France (1 profile per 0.64 km^2). The number of training sites was progressively reduced (from n = 7500 to n = 400, corresponding to 1 profile per 0.7 km^2 to 1 profile per 13 km^2) to simulate the different density of observations. For OK and QRF, we tested random subsampling for splitting the data into training and testing datasets using k-fold cross validation. For QRF we also tested conditioned Latin hypercube sampling based on the point coordinates or the covariates. The results indicated that, with increasing density of observations, OK performed as well or even better than QRF, depending on the particle-size fraction. For silt prediction, OK was systematically better than QRF. However, the prediction intervals were much larger for OK than for QRF, and OK did not seem to estimate uncertainty correctly. Overall, the performance indicators increased with the density of observations with a threshold at about 1 profile per 2 km^2 which suggests that the main limitation of DSM prediction accuracy using QRF is the amount of data collected in the field, not the type of calibration sampling strategy. Future DSM activities should focus on gathering more field observations.

Keywords: Digital Soil Mapping; topsoil particle-size distribution; sampling strategy; sampling density; prediction performance; Multiple soil classes; France

1. Introduction

It is generally accepted that a major limitation to Digital Soil Mapping (DSM) prediction performance is the density of soil observations that carry up-to-date information (e.g., Arrouays et al., 2014a, 2017; Samuel-Rosa et al., 2020). Samuel-Rosa et al. (2015) evaluated whether investing in more spatially detailed environmental covariates improves the accuracy of digital soil maps. Their conclusions showed that more detailed covariates only result in a modest increase in prediction performance and it may be more useful to spend extra resources on collecting more soil observations. Indeed, substantial discrepancies in soil observations' density exist among, and even within, national soil databases (Morvan et al., 2008; Arrouays et al., 2017). Somarathna et al. (2017) tested various machine learning algorithms to predict soil carbon and found that the density of observations was more important than the type of machine learning models. While many DSM studies focused on comparing various machine learning models (Padarian et al., 2020), less work focused on the impact of the density of observations. By using a large dataset of pseudo values of clay content obtained from hyperspectral data, Lagacherie et al. (2020) confirmed the importance of sampling density on DSM performances but showed that this importance diminished as sampling densities increased and that other spatial characteristics of the soil sampling (completeness and evenness in space) also had a strong effect.

Even within harmonized soil conventional mapping programs, the methodology and spatial coverage are far from homogeneous, even at the national scale. This is the case in France, where a national program for mapping soil-scapes at 1:250,000 is finalised (Laroche et al., 2014; Richer-de-Forges et al., 2019). In their prospective analysis on the future of the French national soil-mapping program, Voltz et al. (2020) proposed to move from conventional soil mapping to DSM and to improve the density of observed soil profiles to homogenize and enhance the accuracy of 1:250,000 soil maps for France. In practice, the French program of

1:250,000 soil mapping has been done using conventional mapping by administrative units i.e., the French "Départements" the mean area of which being about 6,000 km². Within these "Départements", the density of soil observations is highly variable, from 1 observation per 0.64 km² to 1 observation per 123 km². Three thresholds were set to indicate the quality of the maps (InfoSol, 2005): (i) at least 1 profile per 60 km² as a minimum standard designated as "operational level", (ii) more than 1 profile per 40 km² for a higher quality "advanced level" and (iii) more than 1 profile per 20 km² as "optimal level". These thresholds were set by experts and used to deliver labels of quality by the Ministry in charge of Agriculture, but their relevance was never assessed quantitatively. In particular, we never tested if increasing the density of points substantially would increase the performance of the predictions when using DSM techniques to map a "Département".

In this study, we chose the French "Département" with the highest soil profile density observations. We tested how the density of training profiles affects the prediction of topsoil particle size distribution, namely clay, silt, and sand contents. This paper has two objectives: i) to assess if we could identify thresholds of the density of soil profiles for prediction performances when applying DSM techniques to predict some basic soil properties and, ii) to compare the performances of a machine learning algorithm (quantile random forest) to pure spatial interpolation, ordinary kriging, as a function of the density of observations.

2. Material and methods

2.1.Study area and general framework of the study

We chose the "Mayenne" French "Département". (~8,100 points for a total area of 5208 km², corresponding to a density of one profile per 0.64 km²), which has the highest density of observations for a French department in the framework of the systematic conventional survey of soils at the scale of 1:250,000. Figure 1 shows the study area with the original sampling points from the soil surveyor and the elevation. Except for some urban areas, sampling points

are rather homogeneously distributed. Of course, the density of points is not regular, because conventional survey allows the surveyor to place the points for maximum information on the soil-landscape, but we can reasonably expect that most of the soil/landscape situations are represented in this sampling.

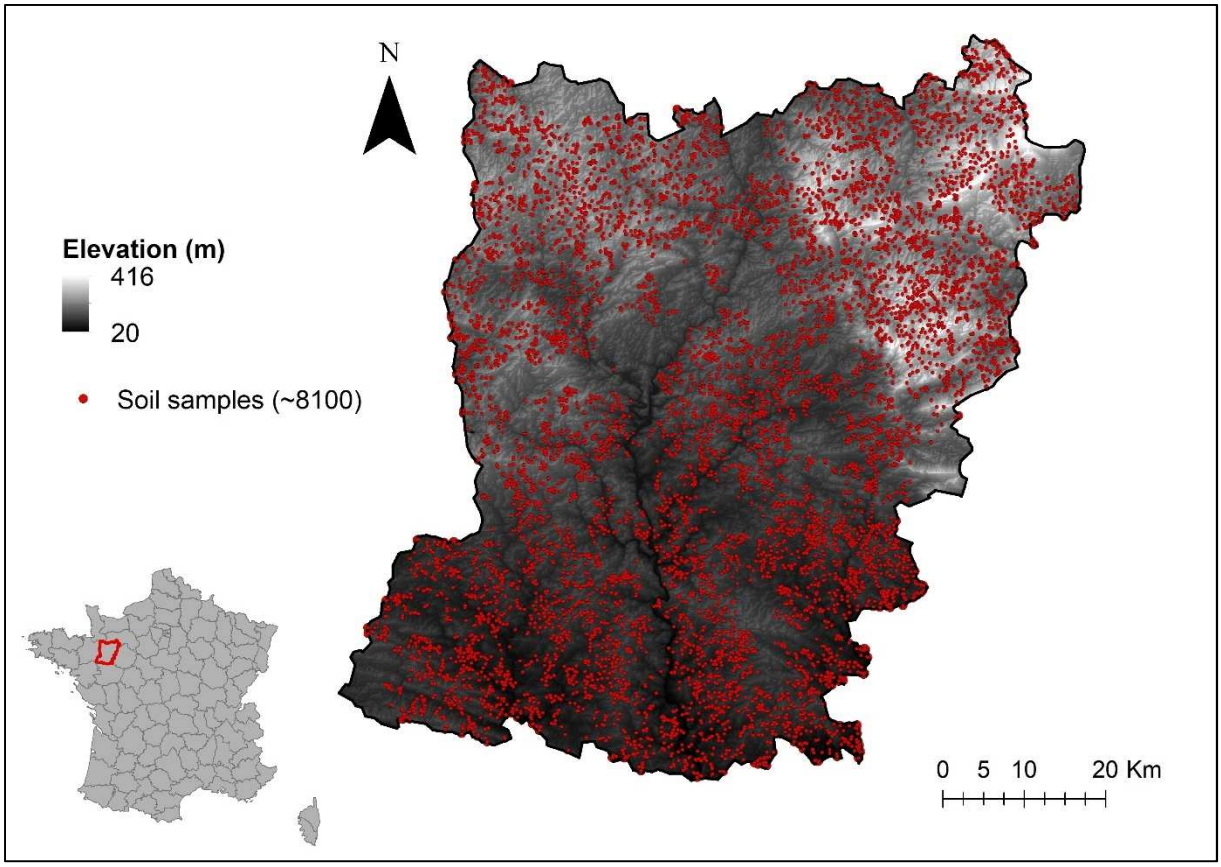


Fig. 1. Map of the study area with soil sampling points used for the conventional mapping and elevation (m).

The region is located at the boundary of sedimentary deposits (mainly calcareous and loess) and crystalline and metamorphic rocks. Thus, we expect that we would be able to see contrasting spatial structures of soil texture within this region. In addition, the area is covered by interesting spatial covariates, namely, (i) a harmonized and revised lithological map (1:50,000 scale; Vernhet, 2010, recoded by Bialkowski et al., 2019), and (ii) an airborne gamma-ray survey.

99

100 *2.2. Soil sampling and analysis*

101 Soil sampling was done in the framework of a conventional soil survey; its density was
102 adapted to the soil variations observed in the field by the soil surveyors. At each field
103 sampling location, the soils were sampled either by digging a soil pit or by augering. The
104 topsoil organo-mineral horizon (excluding O horizons when present) was sampled. For most
105 of the agricultural plots, the corresponding thickness was the ploughed layer (from 0-20 to 0-
106 30 cm, with an average thickness of 24 cm) whereas the thickness ranged from 0-5 to 0-30 cm
107 (with an average of 20 cm) and from 0-5 to 0-25 cm (with an average of 18 cm) under
108 permanent pastures and forest respectively. Samples were air-dried at 30°C to a constant mass
109 and then gently crushed to a 2 mm sieve. The particle-size analysis was performed using the
110 Robinson pipette method (Robinson, 1933).

111 As the fractions clay (< 2 µm), silt (2-50 µm), and sand (50-2000 µm) sum to 1000 g kg^{-1}, the
112 data were transformed with the additive log-ratio (alr-transform, Aitchison, 1982) for the
113 spatial prediction of compositional data (Lark and Bishop, 2007). The alr-transformed
114 variables were calculated as follows (Eqs (1 and 2)):

115 $$\boldsymbol{Clay_{alr}} = \ln\left(\frac{Clay}{Sand}\right) \quad (1)$$

116 $$\boldsymbol{Silt_{alr}} = \ln\left(\frac{Silt}{Sand}\right) \quad (2)$$

117 The alr predicted variables were then back-transformed to sand, silt, and clay through the alr
118 inverse transformation (Eqs (3-5)):

119 $$\boldsymbol{Clay}\ (\boldsymbol{g.kg^{-1}}) = \frac{e^{Clay_{alr}}}{(1 + e^{Clay_{alr}} + e^{Silt_{alr}})} \times 1000 \quad (3)$$

$$\boldsymbol{Silt}\ (\boldsymbol{g.kg^{-1}}) = \frac{e^{Silt_{alr}}}{(1 + e^{Clay_{alr}} + e^{Silt_{alr}})} \times 1000 \quad (4)$$

$$\boldsymbol{Sand}\ (\boldsymbol{g.kg^{-1}}) = \frac{1}{(1 + e^{Clay_{alr}} + e^{Silt_{alr}})} \times 1000 \quad (5)$$

To assess the confidence interval of our predictions, we calculated the variance of the back-transformed clay, sand, and silt fractions. Román Dobarco et al. (2019) showed that since the prediction and variance are independent, the Taylor analysis method is more efficient than using Monte Carlo simulation to estimate the prediction variance of each particle-size. Utilising the first-order Taylor analysis, the variance of prediction can be derived as follow (Eqs (6-8)):

- $$\boldsymbol{var}\left(\boldsymbol{Clay}\ (\boldsymbol{g.kg^{-1}})\right) = \sigma^2_{Clay_{alr}} \left(\frac{1000(e^{Clay_{alr}}(1+e^{Clay_{alr}}+e^{Silt_{alr}})-e^{2Clay_{alr}})}{(1+e^{Clay_{alr}}+e^{Silt_{alr}})^2}\right)^2 +$$

$$\sigma^2_{Silt_{alr}} \left(\frac{-1000e^{Clay_{alr}}e^{Silt_{alr}}}{(1+e^{Clay_{alr}}+e^{Silt_{alr}})^2}\right)^2 + 2cov(Clay_{alr}, Silt_{alr}) \frac{-10^6 e^{2Clay_{alr}}e^{Silt_{alr}}(1+e^{Silt_{alr}})}{(1+e^{Clay_{alr}}+e^{Silt_{alr}})^2} \quad (6)$$

- $$\boldsymbol{var}\left(\boldsymbol{Silt}(\boldsymbol{g.kg^{-1}})\right) = \sigma^2_{Silt_{alr}} \left(\frac{1000(e^{Silt_{alr}}(1+e^{Clay_{alr}}+e^{Silt_{alr}})-e^{2Silt_{alr}})}{(1+e^{Clay_{alr}}+e^{Silt_{alr}})^2}\right)^2 +$$

$$\sigma^2_{Clay_{alr}} \left(\frac{-1000e^{Clay_{alr}}e^{Silt_{alr}}}{(1+e^{Clay_{alr}}+e^{Silt_{alr}})^2}\right)^2 + 2cov(Clay_{alr}, Silt_{alr}) \frac{-10^6 e^{Clay_{alr}}e^{2Silt_{alr}}(1+e^{Clay_{alr}})}{(1+e^{Clay_{alr}}+e^{Silt_{alr}})^2}$$

(7)

- $$\boldsymbol{var}\left(\boldsymbol{Sand}(\boldsymbol{g.kg^{-1}})\right) = \sigma^2_{Clay_{alr}} \left(\frac{-1000e^{Clay_{alr}}}{(1+e^{Clay_{alr}}+e^{Silt_{alr}})^2}\right)^2 +$$

$$\sigma^2_{Silt_{alr}} \left(\frac{-1000e^{Silt_{alr}}}{(1+e^{Clay_{alr}}+e^{Silt_{alr}})^2}\right)^2 + 2cov(Clay_{alr}, Silt_{alr}) \frac{10^6 e^{Clay_{alr}}e^{Silt_{alr}}}{(1+e^{Clay_{alr}}+e^{Silt_{alr}})^2}$$

(8)

For more detailed explanations on the use of first-order Taylor analysis in spatial models, we refer to Heuvelink et al. (1989).

138

139

140

141

142 *2.3.Covariates*

143 The list of covariates used in this study is provided in Table 1.

Table 1. Spatially exhaustive co-variates used in the Mayenne department.

Covariates	Resolution	Data type	Reference
Climate			
Climate type	250 m	*Qualitative*	Joly et al., (2010)
Vegetation			
Forest type	1:25 K	*Qualitative*	Inventaire Forestier National (2006)
Land use 2016	10 m	*Qualitative*	CESBIO (2016)
Principal Component NDVI (N=3)	500 m	*Quantitative*	Loiseau et al., (2019)
Soil maps			
Clay content France (%)	500 m	*Quantitative*	Ballabio et al., (2016)
Silt Content France (%)	500 m	*Quantitative*	Ballabio et al., (2016)
Sand content France (%)	500 m	*Quantitative*	Ballabio et al., (2016)
Rate of river network development and persistence (IDPR)	1:50 K	*Quantitative*	BRGM (2014)
Topography			
Elevation	25 m	*Quantitative*	IGN (2014)
Compound topographic index (CTI)	25 m	*Quantitative*	IGN (2014)
Multi-Resolution Valley Bottom Flatness (MRVBF)	25 m	*Quantitative*	IGN (2014)
Multi-Resolution Ridge Top Flatness (MRRTF)	25 m	*Quantitative*	IGN (2014)
Roughness	25 m	*Quantitative*	IGN (2014)
Curvature	25 m	*Quantitative*	IGN (2014)
Exposition	25 m	*Quantitative*	IGN (2014)
Slope position	25 m	*Quantitative*	IGN (2014)
Slope, cosines(slope)	25 m	*Quantitative*	IGN (2014)

Gamma ray data			
Thorium (Gamma)	250 m	*Quantitative*	Bonijoly et al., (1999)
Uranium (Gamma)	250 m	*Quantitative*	Bonijoly et al., (1999)
Potassium (Gamma)	250 m	*Quantitative*	Bonijoly et al., (1999)
Th/K ratio (Gamma)	250 m	*Quantitative*	Bonijoly et al., (1999)
Th/U ratio (Gamma)	250 m	*Quantitative*	Bonijoly et al., (1999)
Simplified Lithology	1:50 K	*Qualitative*	Loiseau et al., (2020) from Vernhet (2010)

All covariates were transformed to a 90 m resolution grid according to the *GlobalSoilMap* specifications (Arrouays et al., 2014b), either through downscaling for coarser-resolution layers, or upscaling for the finer ones. The upscaling was done in two ways: 1) by the average aggregation for quantitative data, 2) by a majority vote estimation for the qualitative covariates. In addition, we transformed the classes from the qualitative covariates to binary information (0: absence and 1: presence) to assess each class's importance in our model.

2.4.Digital soil mapping modelling

In order to assess the effect of point density on the performance of the predictions, we reduced the number of training samples progressively and adopted three different strategies to predict particle-size distribution over the “Département”. To keep comparable testing sizes for large and small datasets, we selected 8100 to 600 sample points with a 1000 to 200 increment (Table 2). We used 200 testing points for cross-validation when the number of training points was larger than 2000, and 500 testing points when the number of training points was smaller or equal to 2000. This strategy aimed at keeping enough testing points for each fold, and as many folds as possible. Each splitting procedure was repeated 100 times.

Table 2. *Training and testing datasets used in this study.*

Total samples (n)	**600**	**800**	**1000**	**1200**	**1400**	**1600**	**1800**	**2000**	**3000**	**4000**	**5000**	**6000**	**7000**	**8100**
Training samples	400	600	800	1000	1200	1400	1600	1800	2500	3500	4500	5500	6500	7600

(n)														
Training points density (n/km^2)	0.08	0.12	0.15	0.19	0.23	0.27	0.31	0.35	0.48	0.67	0.86	1.06	1.25	1.46
Training points density (1/a km^2)	13	8.7	6.5	5.2	4.3	3.7	3.3	2.9	2.1	1.5	1.2	0.9	0.8	0.7
Number of points per-fold (n)	200	200	200	200	200	200	200	200	500	500	500	500	500	500
Number of k- fold	3	4	5	6	7	8	9	10	6	8	10	12	14	16

We tested four different prediction methods, summarized as follows:

1) Ordinary Kriging (OK), to assess the prediction of soil texture, only using soil sample location and measured particle-size values as information. Samples were randomly selected to form various sizes of training data.
2) Quantile Regression Forest (QRF) using the R package quantregForest (Meinshausen, 2006) with covariates listed in Table 1. We chose this technique as it has been shown to be robust, and it enabled us to derive quantile distributions instead of a single mean value (Vaysse and Lagacherie, 2017). Samples were randomly selected as in OK.
3) QRF as previously, but we chose to sample the points using the conditional Latin Hypercube Sampling (cLHS, Minasny and McBratney, 2006) method based on their coordinates to reduce the clustering of samples.
4) QRF as previously, but we chose to sample the points using cLHS applied on the covariates to ensure even coverage of the covariates.

2.5.Assessing the performance of the predictions

The performance of the predictions was evaluated through a k-fold cross validation by keeping a minimum of 200 to 500 samples apart for testing, depending on the strategy of

177 partitioning the data. We assessed the model performance through a 100 times replication of
178 the cross-validation method by reporting the medianvalue of the following indicators:

179 ➢ The coefficient of determination (R^2), which measures the adequacy between our
180 model and our observed value.

181 $$\boldsymbol{R}^2 = 1 - \left(1 - \boldsymbol{R}^2_{non-adjusted}\right)\frac{n-1}{n-p-1} \quad (9)$$

182 $$\text{where} \quad \boldsymbol{R}^2_{non-adjusted} = \frac{\sum_{i=1}^{n}(z_i^* - \bar{z})^2}{\sum_{i=1}^{n}(z_i - \bar{z})^2} \quad (10)$$

183 Here, $z_i$and $\bar{z}$ are the value of the observation for point *i* and the mean of all
184 observations, respectively; z_i^* is the value of the prediction for point *i.*

185 ➢ The root-mean-square error (RMSE), which provides information on the statistical
186 dispersion of our predictions in relation to our observations.

187 $$\boldsymbol{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(z_i - z_i^*)^2} \quad (11)$$

188 ➢ The concordance coefficient (CC; Lin, 1989), which is given as

189 $$\boldsymbol{CC} = \frac{2\rho\sigma_{z^*}\sigma_z}{\sigma_z^2 + \sigma_{z^*}^{2} + (\bar{z} - \bar{z}^*)^2} \times 100 \quad (12)$$

190 where $\sigma_z^2$and $\sigma_{z^*}^{2}$ are the observation and prediction variances, respectively, ρ
191 is their correlation coefficient, and $\bar{z}^*$ is the mean of the prediction

192 ➢ The bias of the predictions, which is the difference between the mean predicted
193 and mean observed values.

194 $$\boldsymbol{Mean\ error} = \frac{1}{n}\sum_{i=1}^{n}(z_i^* - z_i) \quad (13)$$

195 ➢ The PICP (prediction interval coverage probability). This indicator assesses the
196 uncertainty between the observed and predicted distribution. The PICP is the

probability that the target of an input pattern lies within the prediction limits (Shrestha and Solomatine, 2006). The PICP was calculated from the confidence interval of 90% with the extreme quantile 5% and 95% of the model prediction.

$$\boldsymbol{PICP} = \frac{\sum_{i=1}^{n}(z^*_{5\%} < z_i < z^*_{95\%})}{n} \times 100 \quad (14)$$

3. Results and discussion

3.1.Summary statistics

The distribution of the particle-size fractions in $g.kg^{-1}$, over the Mayenne department, is shown in Table 3. We observed a large dominance of the silt fraction over the samples. Moreover, testing the skewness and kurtosis indices, silt and sand seemed to follow a rather symmetric distribution with fewer and less extreme outliers than the normal distribution. The clay distribution was positively skewed and had a high kurtosis value, which shows its asymmetry and more outliers in the right part of its distribution.

Table 3. *Summary statistics of particle-size distribution of the samples over the Mayenne department.*

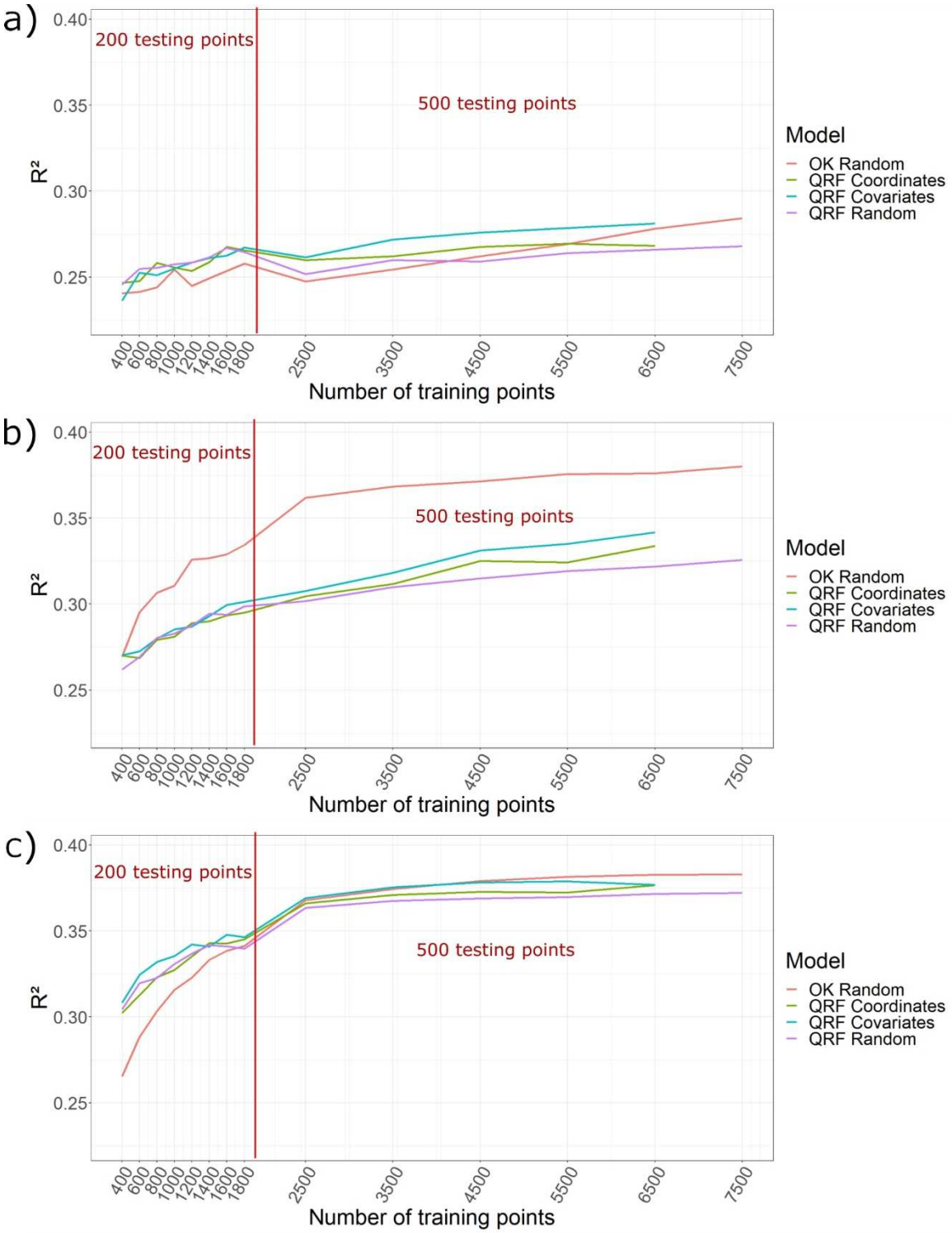
	Clay	**Silt**	**Sand**
Min	18,83	20	19,06
Q1	141	448	189
Mean	177,22	529,26	293,33
Median	164	540	267

Q3	200	624	374
Max	631	839	898
Decile 1	125,87	360,6	144
Decile 9	249	683	481,54
Skewness	1,76	-0,48	0,88
Kurtosis	6,07	-0,08	0,59

3.2.Performance of the predictions

In order to assess the performances of each sampling strategy, we compared the performance of the predictions for each sampling density and algorithms used. Figures 2 to 6 show the evolution of selected statistical indicators (R^2, RMSE, ME, CCC, and PICP) as a function of training sample densities.

The R^2 for the QRF algorithms for silt and sand showed a sharp increase with an increasing number of training points until about 1800 and 2500 respectively, followed by a small increase (Fig 2). This shape is similar to those obtained by Lagacherie et al (2020). These threshold were less evident for clay. The three QRF algorithms gave quite similar results. Interestingly the R^2 for silt was always higher when using OK than using QRF methods. This may be due to the fact that the covariates we used did not capture silt distribution effectively. Moreover, when the number of training samples was very large, OK R^2 gave similar results compared to QRF for clay and slightly better results for sand. Note, however, that OK gave poorer results than QRF for clay and especially for sand at the lowest densities of sampling and that for these fractions, the decrease of R^2 was more pronounced for low sampling densities.

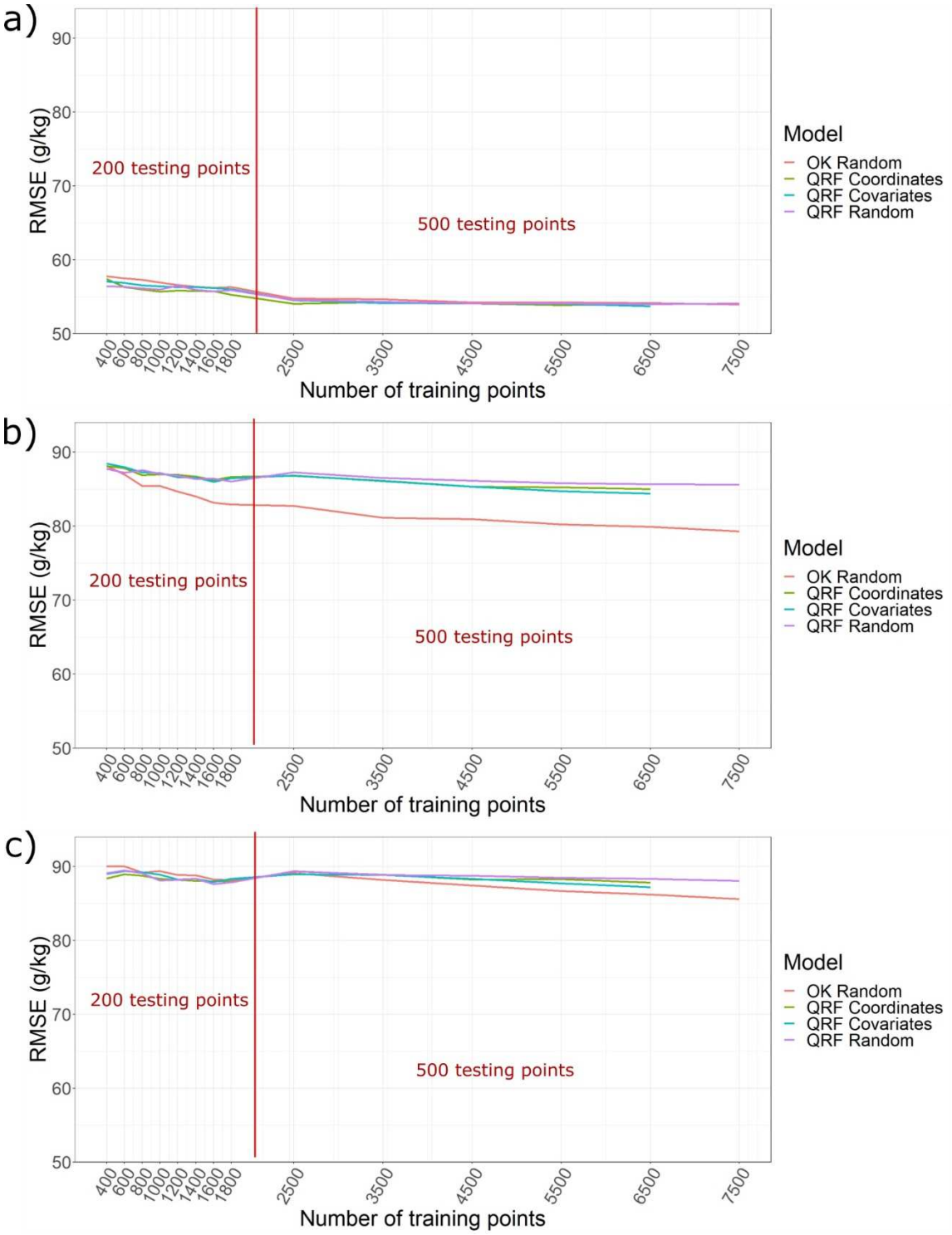


241 **Fig 2.** Median R^2 for testing: comparison for a) Clay, b) Silt and c) Sand fraction for each
242 sampling strategy with increasing training sample size over the Mayenne department. The red
243 line separates the strategies using 500 testing points from those using 200 testing points.

244 For clay, the RMSE (Fig 3) showed a substantial decrease with increasing training densities
245 until reaching a threshold of about 2500, followed by a slight decrease thereafter. No real

246 trend was observed for silt and sand. No substantial difference was observed between the
247 algorithms (Fig. 3) except for silt, in which OK always performed slightly better than the QRF
248 methods, especially at high sampling densities, which is consistent with the results of R^2. Note
249 that the RMSE values are always lower than the interquartile range (see Table 3), suggesting
250 that the mapping is effective.

251



252 **Fig 3.** Median RMSE for testing: comparison for a) Clay, b) Silt and c) Sand fractions for
253 each sampling strategy with increasing training sample size over the Mayenne department.
254 The red line separates the strategies using 500 testing points from those using 200 testing
255 points.

256 The mean error for clay seemed to be constant for all models at a very low value for clay
257 (around -2 $g.kg^{-1}$). Silt and sand exhibited much larger ME, even if the OK presented better
258 results for these fractions. For all QRF models, the ME median value stayed somewhat
259 similar. Interestingly, silt and sand showed opposite trends, increasing ME for silt until about
260 2500 points and a decreasing one for sand until the same threshold.

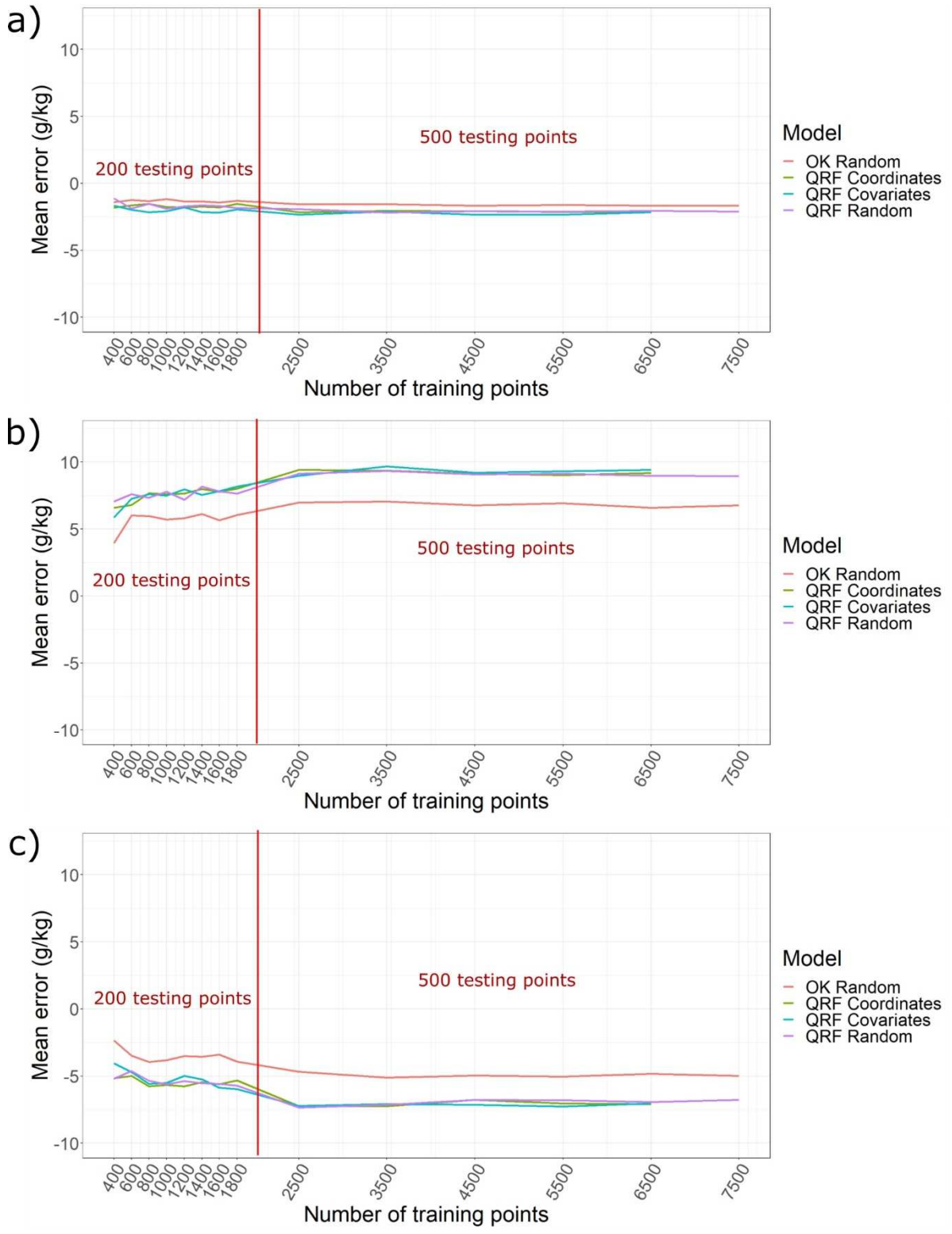


Fig 4. Median error for testing: comparison for a) Clay, b) Silt and c) Sand fraction for each sampling strategy with increasing training sample size over the Mayenne department. The red line separates the strategies using 500 testing points from those using 200 testing points.

The CCC indicator increased with increasing density for all fractions (Fig. 5). Interestingly, OK worked better than QRF for large sampling densities, and even for all densities for silt. For sand, the CCC trend was similar to R^2, with a sharp increase with the increasing density of training points and a lower increase when the number of training points became higher (> 2500). For the highest training densities, it was higher for OK than for QRF, especially for silt. On the contrary, for the lowest densities, it became lower for OK than for QRF.

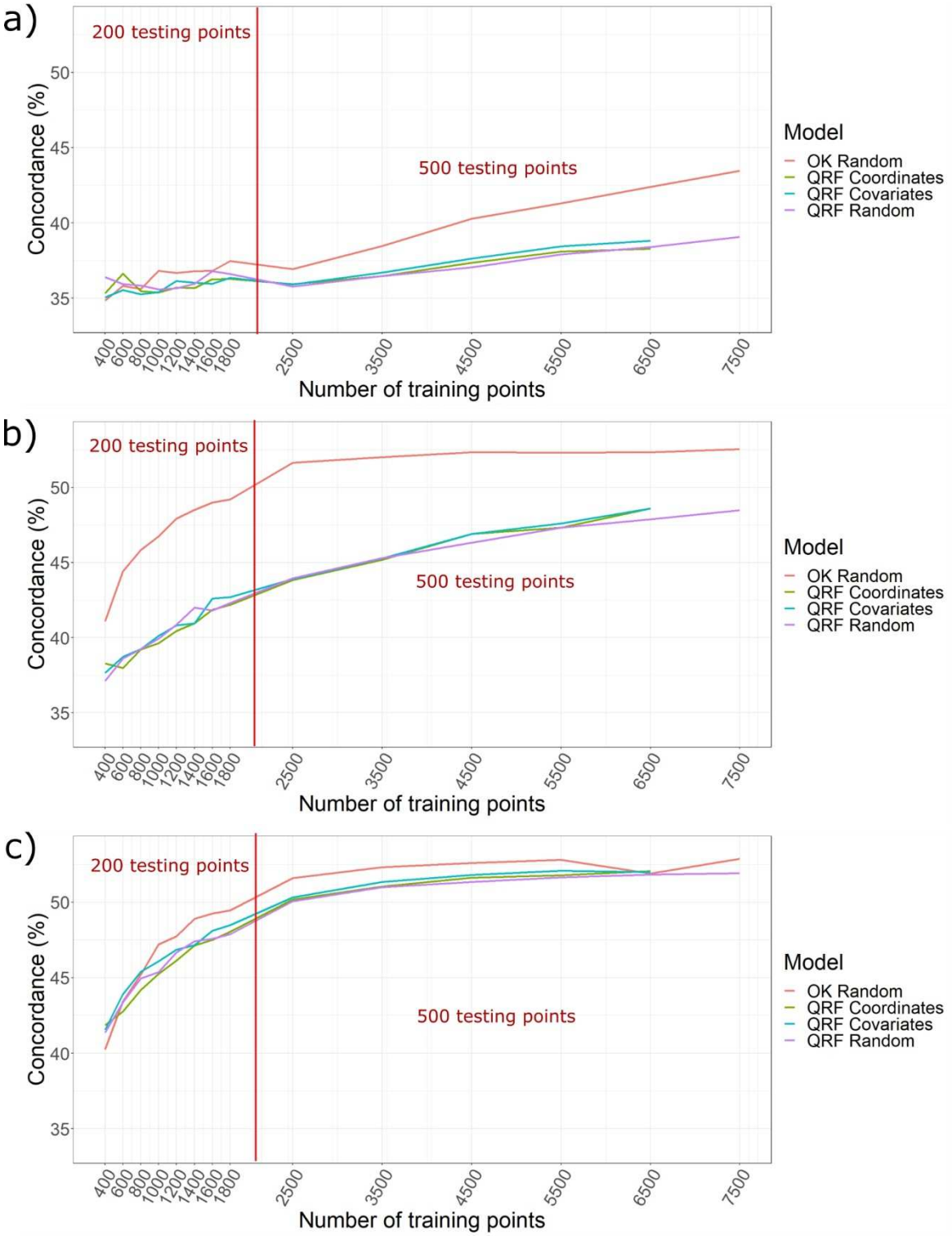
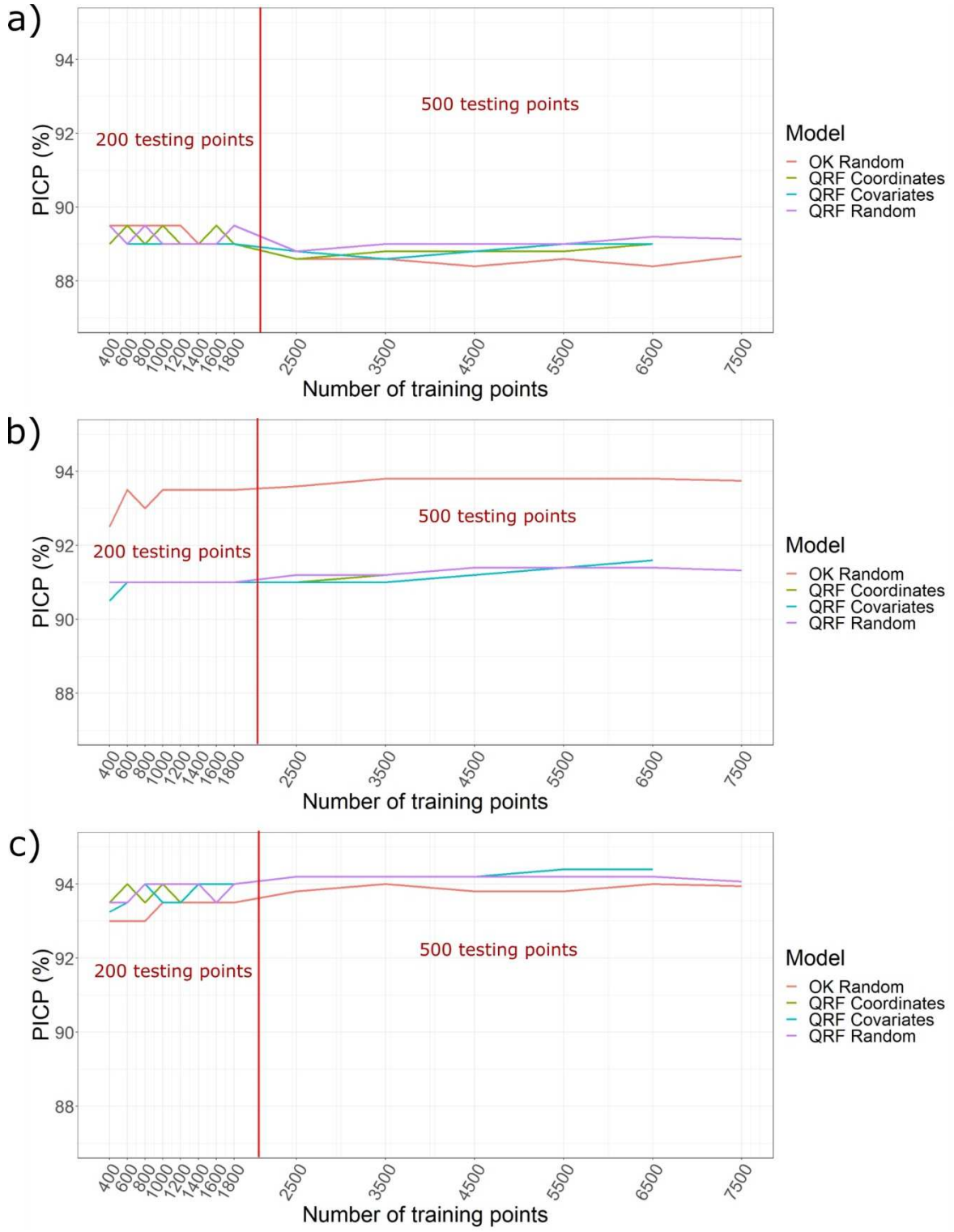


Fig 5. Median Concordance for testing: comparison for a) Clay, b) Silt and c) Sand fractions for each sampling strategy with increasing training sample size over the Mayenne department. The red line separates the strategies using 500 testing points from those using 200 testing points.

For clay (Fig. 6), the PICP was always lower than 90%, which shows that the 90% prediction intervals were slightly under-estimated. This was more evident for OK. From a practical point of view, this shows that kriging smoothed out high/low values, thus under-estimated the prediction intervals. The PICP is mostly over 90% for the prediction of silt by OK, although it shows acceptable values (close to 90%) for the QRF models. This shows that contrary to clay, OK over-estimated PIs for silt prediction, which counter-balanced the better prediction results obtained by OK. For sand prediction, all PICPs were larger than 90%. The contrast between the trends observed for clay and sand could be due to the negative correlation between these two fractions. We observed a rather erratic behavior of the PICP for the lowest calibration sample values. This is likely due to the low number of training and testing samples and the small number of k-fold cross-validations. This is consistent with the findings of Lagacherie et al. (2019), who showed that a small number of validation samples is not robust enough to assess the performance of the model predictions. However, this erratic behavior remains in a narrow range and we should recall here that we replicated 100 times the splitting between training and testing samples.



295

296 **Fig 6.** Median PICP for testing: comparison for a) Clay, b) Silt and c) Sand fraction for each
297 sampling strategy with increasing training sample size over the Mayenne department. The red
298 line separates the strategies using 500 testing points from those using 200 testing points.

299

There were no substantial differences between the 3 different sampling strategies for the QRF algorithms. This may be because all subsampling strategies that we tested were denser than the sampling strategies commonly used in France. Moreover, when the number of selected samples is large, cLHS has no advantage over other sampling designs (Wadoux et al., 2019).

When the number of training samples was equal or larger than 2500, some QRF performance indicators were slightly improved with increasing sampling density. This suggests that collecting new data or rescuing more legacy data should further improve the prediction accuracy, as indicated by Arrouays et al. (2017) and Samuel-Rosa et al. (2015; 2020). When the sampling density was very high, OK performed as well, or even better than machine learning methods, which validates the general framework proposed by Minasny and McBratney (2010) for global digital soil mapping, that was thereafter adopted by the *GlobalSoilMap* initiative (Arrouays et al., 2014b). Interestingly, the silt fraction had a particular behavior compared to other fractions, i.e., except for PICP, the performance was better when using OK for most sampling densities. This may be due to the fact that the covariates that we selected were less responsive for silt than for the other fractions. A threshold of the number of points (around 2500) was identified, under which most of the performance indicators performed worse. Over this threshold, the performance indicators remained stable or increased more slowly with the density of points, suggesting that the main limitation of the goodness of DSM predictions is the amount of data collected in the field. We note that the number of testing points was changedfrom 200 to 500 at 2500 training points, which may affect the results. To keep a minimum number of k-fold the same time, we needed to change the number of testing points. But we assessed the model performance through a 100 times replication of the cross-validation method. This avoids a random chance of under- or over-estimation of error. Moreover, our results are consistent with those obtained recently by Lagacherie et al. (2020).

It is worth noting that this threshold of 2500 points (about 1 profile per 2 km^2) corresponds to a nearly ten times larger density that the density recommended in the traditional 1:250,000 soil mapping programs, even for those considered as optimal (1 profile per 20 km^2). The conventional survey recommendation at 1:250,000 may be adequate if applied by a skilled soil surveyor, but certainly not for fine resolution DSM applications. The large majority of the French departments have a number of profile observations much less than this threshold of 1 profile per 2 km². One can, therefore, assess the enormous efforts needed if the objective is to produce fine resolution DSM products with an acceptable level of quality at the department level using only training points and available covariates. This is even more critical if we consider that the most important covariates in our case (airborne gamma-ray, Loiseau et al., 2020) are not available for the entire French territory.

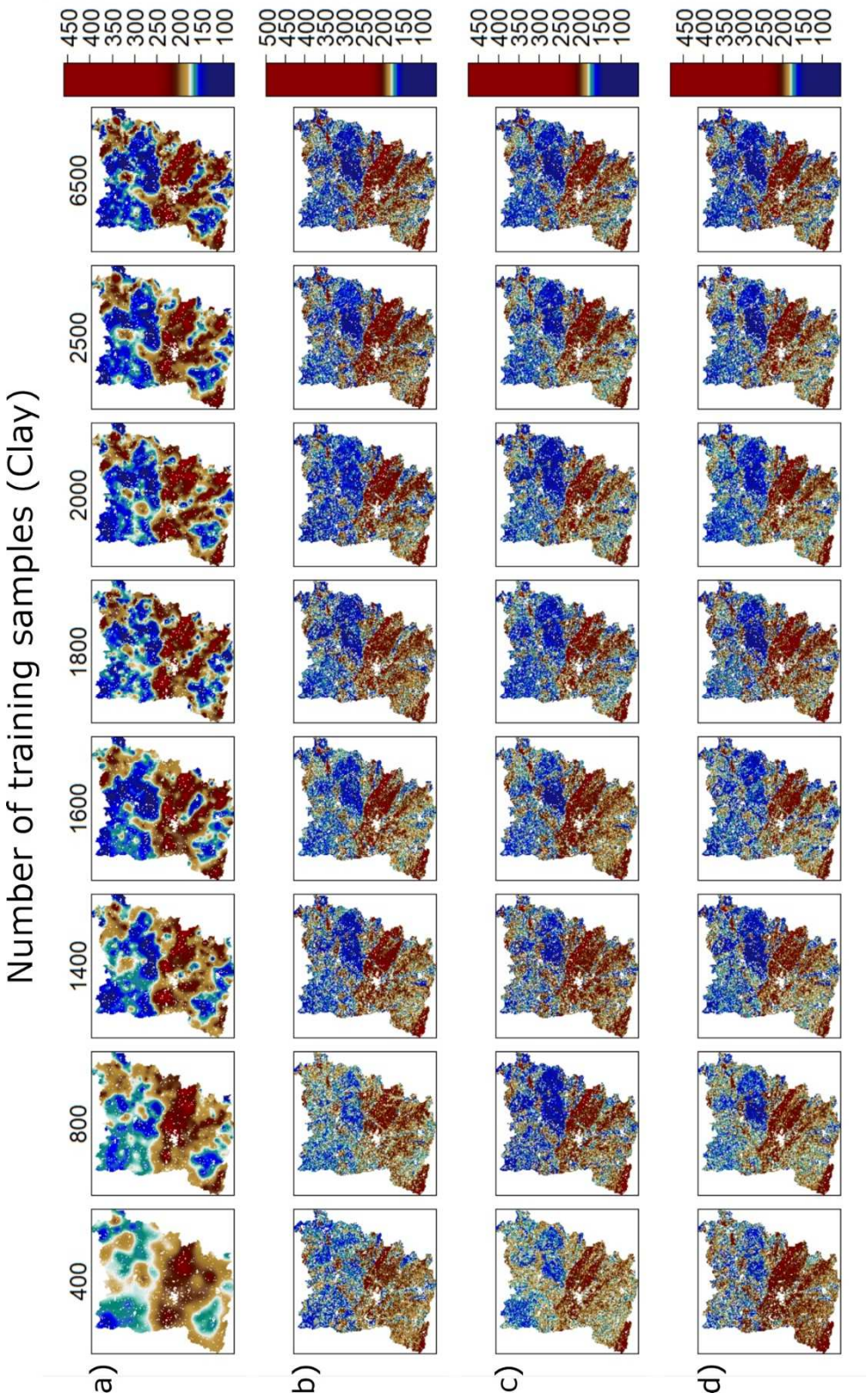
3.3.Maps of predictions

To assess the influence of increasing the number of calibration points on mapping predictions, we mapped several examples for the four trials that we tested for each particle-size fraction (Figures 7 to 9).

The trend showed in these figures is a stabilization of the spatial prediction with the increasing number of training points. This trend is also more pronounced when we focus on OK. The increasing number of calibration points allowed the models to describe more detailed spatial structures, and decreased the smoothing effect of the kriging.

When comparing QRF methodologies sampling, few differences were observed between sampling densities. However, for the lowest training densities, the predictions were more pixelated and some spatial patterns were less visible on the maps. At thresholds ranging from 1800 to 2500 training points, the general patterns seemed to remain stable for clay, silt, and sand.

For clay (figure 7), the spatial patterns with high values (over 200 g.kg^{-1}) seemed to remain stable from 1800 training points in the middle part of the department.



352 **Fig 7.** Clay predictions comparison for the four maps a) OK, b) QRF random, c) QRF
353 coordinates and d) QRF covariates, when the number of training data increases.

For silt, (figure 8), the spatial patterns remained more stable with the integration of more training data and produced more detailed patches when the number of training points exceeded 2000

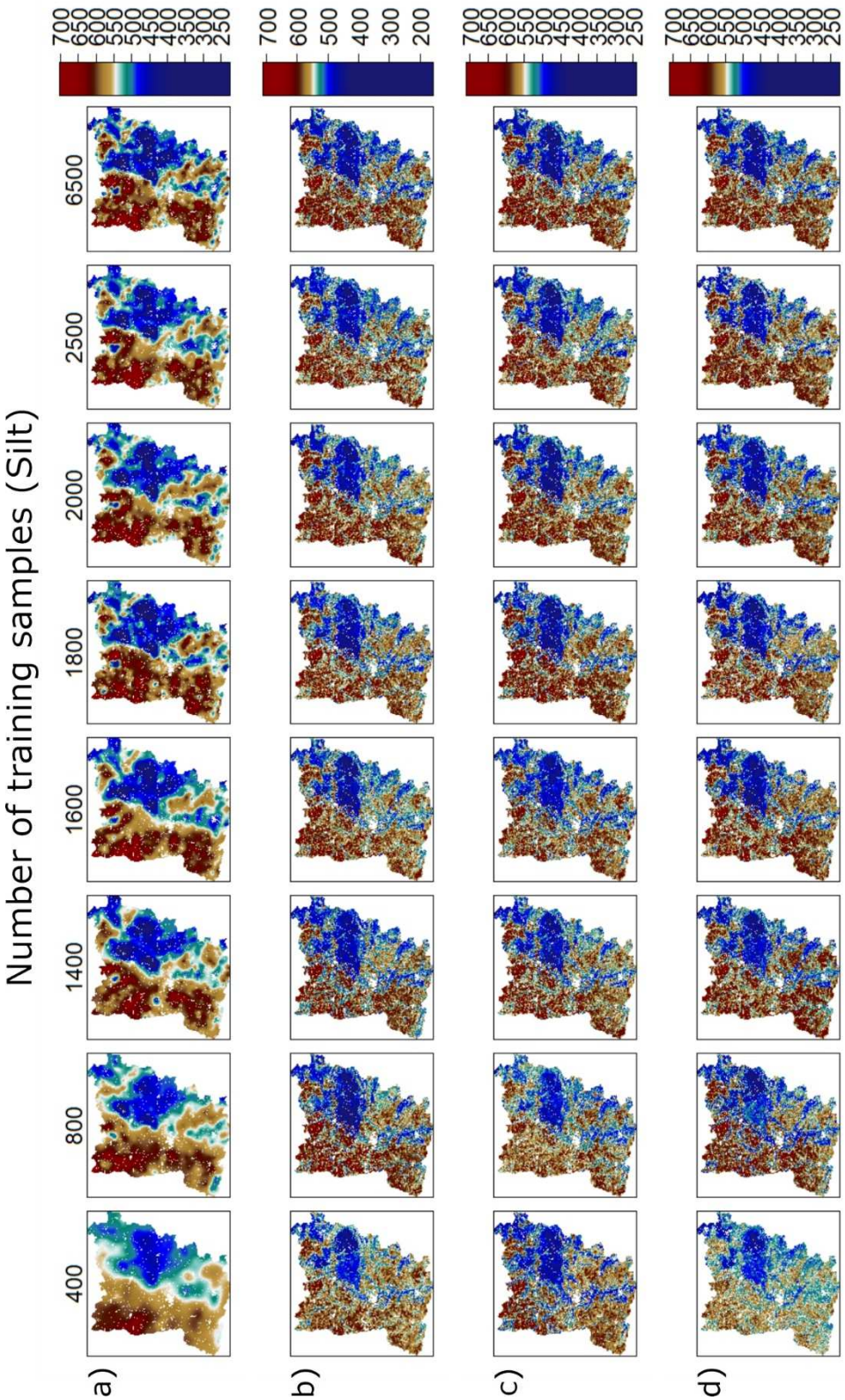
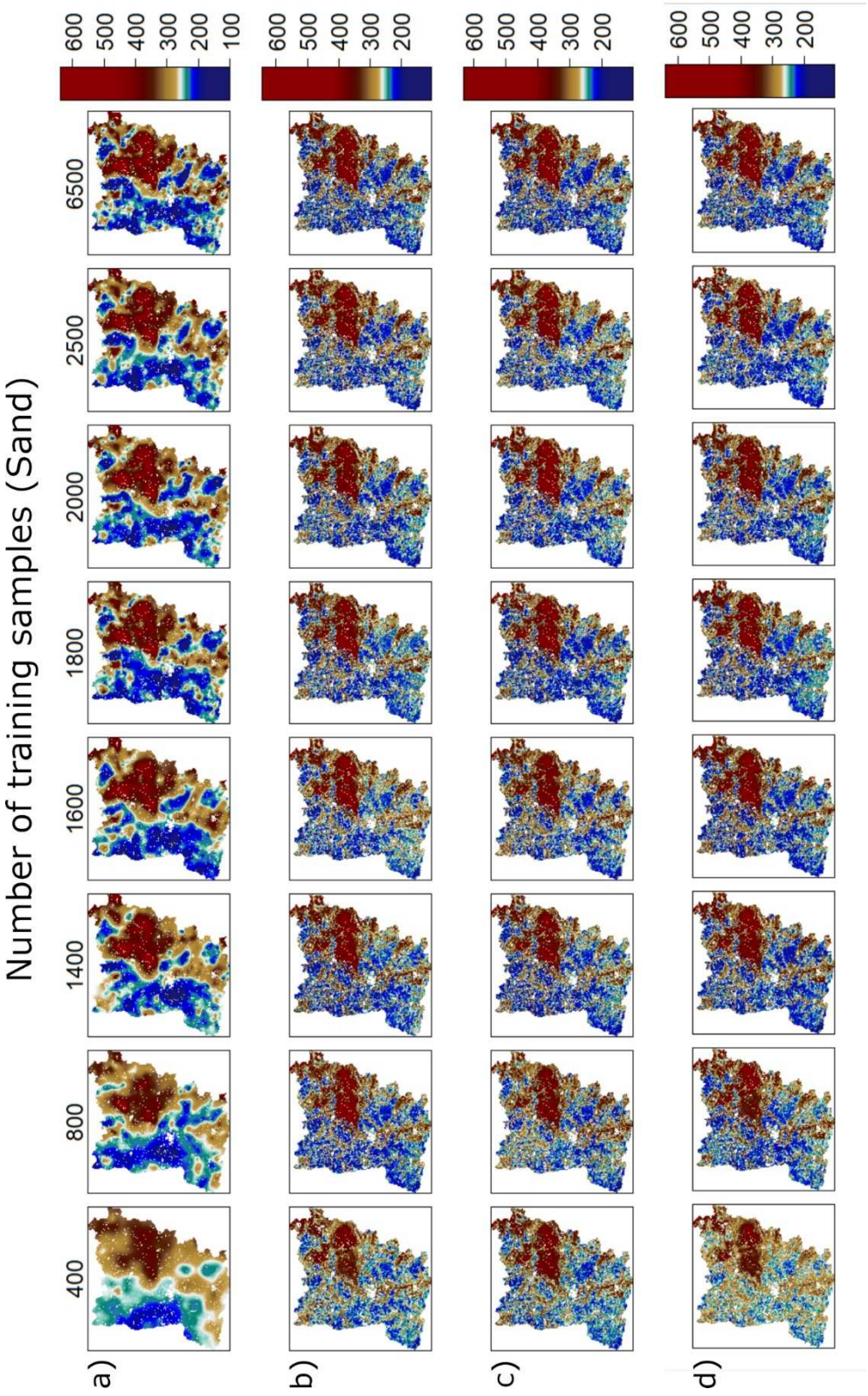


Fig 8. Silt predictions comparison for the four maps a) OK, b) QRF random, c) QRF coordinates and d) QRF covariates, when the number of training data increases.

The predictions of sand (figure 9) presented also similar variations, with a quite similar stabilization of the predictions, for all QRF models, over a threshold close to 1800 points.



361

362 **Fig 9.** Sand predictions comparison for the four maps a) OK, b) QRF random, c) QRF

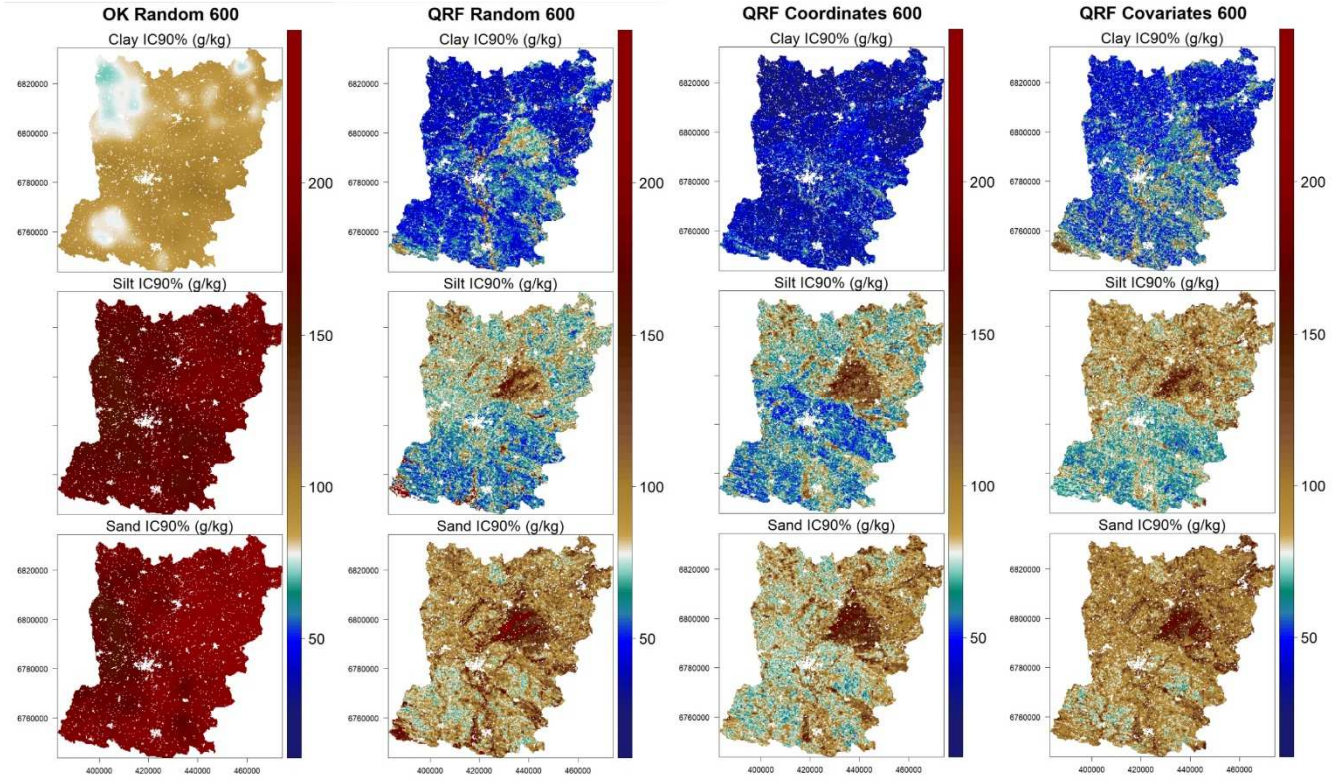
363 coordinates and d) QRF covariates, when the number of training data increases.

364

Several thresholds, ranging from 1800 to 2500 training points, can be observed for the prediction maps of particle-size over the Mayenne department. Indeed, the sampling strategies did not impact the prediction maps compared to the number of training points. However, as shown in the precedent observations, the particle size variability was not entirely explained by our selected covariates. In complex regions, more observations could explain the spatial variation of the soil in more detail. Moreover, using good quality covariates increased a lot in the details of the spatial prediction. In comparison to OK, QRF seemed to indicate the variability of the predictions closer to the observed soil properties variability and produced less smoothed predictions. There are limitations in this study; we still missed some crucial covariates and that our selection of training samples was based on observations collected by a traditional soil survey, which may not be statistically optimal.

3.4.Interval of predictions and mapping

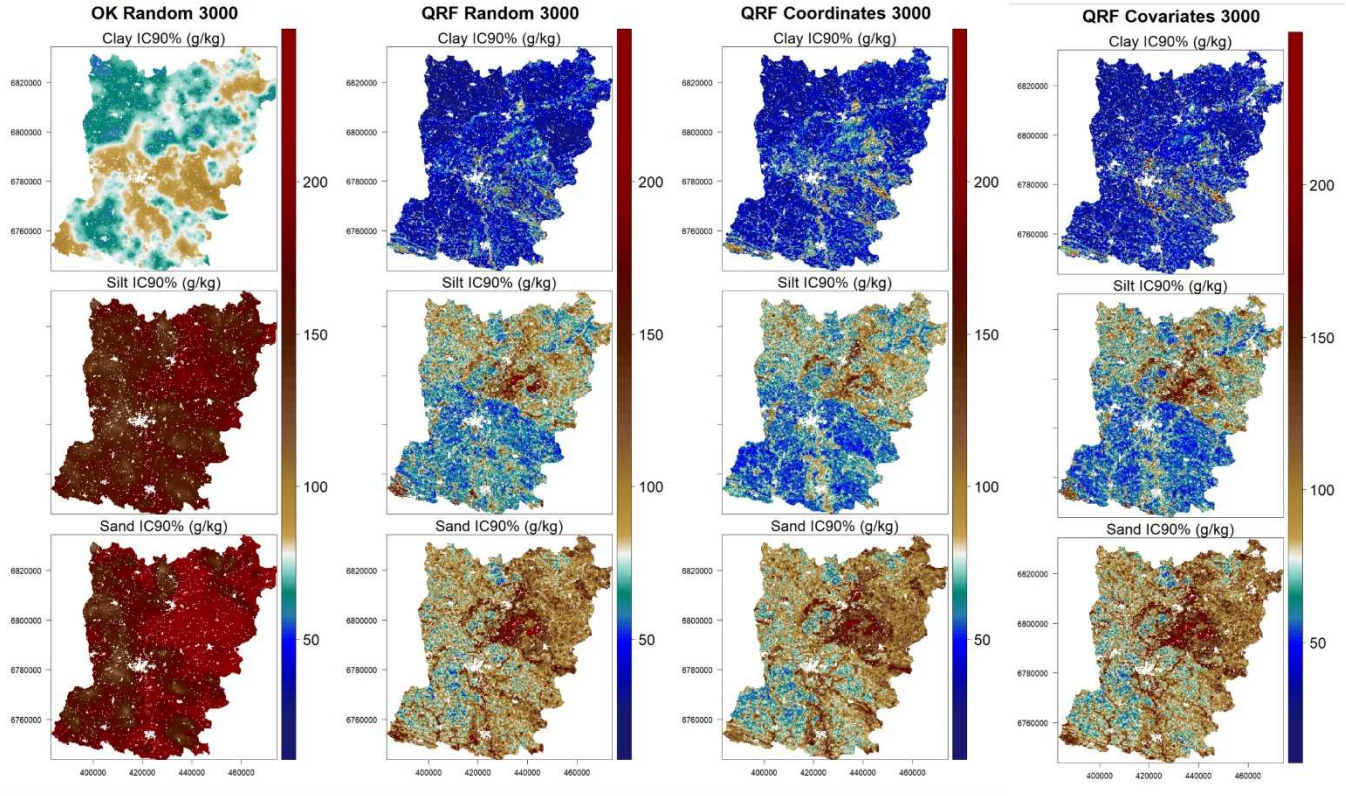
To assess the influence of the training dataset size over our textural predictions, we estimated and mapped their 90% prediction interval (PI). Figures 10 and 11 show the PI of each sampling strategy for training sets of 400 and 2500 points. We observed for each sampling, a decrease of the PI with the increase of the number of training points, except for coordinates sampling where the PI seemed relatively stable or better, depending on the particle size, with a smaller dataset (Figure 9).



384

385 **Fig 10.** Particle-size prediction interval (90%) for the four maps using 600 points (400 for
386 training and 200 for testing).

387 For clay, PIs were lower and showed the most extensive variations over the metamorphic
388 area, whereas the uncertainty was much larger with a few points for each sampling strategy.
389 Silt showed large variations in the North, but also in the South-Est. This can be due to its
390 variability and the fact that complex deposition processes could not be captured by our model
391 and covariates, even with a large set of training points (Figure 11). The global variation of
392 sand over the department resulted from the cumulative error from the back transformation and
393 presented the largest variation compared to clay and silt. Note that the PIs were always much
394 larger for OK than for QRF models and that, as expected, the PI intervals strongly decreased
395 when going from a small number to a large number of training points. The spatial patterns of
396 PIs were very different from OK to QRF as already observed by Vaysse and Lagacherie
397 (2017). Note also that the PI maps were very similar for all the QRF models, suggesting that
398 the sampling strategy is not important at this high number of sampling points.



399

400 **Fig 11.** Particle-size prediction interval (90%) for the four maps using 3000 sampling points
401 (2500 for training and 500 for testing)

402 These results were expected and supported environmental information integration in DSM
403 models for mapping against pure geostatistical interpolation. With a large set of data, the QRF
404 sampling strategies were indifferent, and the variability of prediction was more affected by
405 the number of calibration samples. However, for small size of training points, cLHS appeared
406 to be useful (i.e., for QRF, sampling capturing the maximum combinations of the values of
407 relevant covariates). We suggest that for planning soil sampling for DSM, the use of sampling
408 design that aims to capture the maximum of information is required. There are still
409 discussions about the best sampling strategy for a small number of samples: it is beneficial to
410 use methods such as cLHS or k-means clustering based on covariates when covariates exhibit
411 contrasted values, which is the case in the present study. However, if the contrast of
412 covariates is smoother, then it may be more important to get a rather regular coverage of the
413 geographical space. Overall, increasing the number of training points to more than 2500 led to

a relatively small increase in prediction performances, which suggests that acquiring more points may not be worth considering the gain in accuracy that can be achieved. This observation might be different if we consider more complex models such as deep learning (Padarian et al., 2019; Ng et al., 2020) or new (or finer resolution) covariates that can explain soil texture variability better. This proposal will be tested in future work.

4. Conclusions

We evaluated the effect of soil observations density on the digital soil mapping model's performance for topsoil particle-size distribution with four sampling strategies and with a decreasing number of training points. We demonstrated a threshold of sampling density, and that sampling with a density less than about 1 profile per 2 km^2 could lead to a substantial decrease in the performance accuracy. This result is significant, considering that the density recommended for conventional soil mapping at 1:250,000 is 1 profile per 20 km^2, which is 10 times smaller. We showed that sampling strategies based on covariates or coordinates with cLHS did not produce different results from random sampling when the number of training samples was large. We showed that increasing the number of training samples produced better improvements in the predictions maps when using QRF models compared to ordinary kriging. We conclude that the main limitations of DSM prediction accuracy are the amount of data collected in the field and high-quality covariates. Future DSM activities should focus on gathering more field observations. However, increasing the number of training points may in some cases not be worth considering the gain in accuracy that can be achieved.

Acknowledgments

The airborne gamma-spectrometric data used in this study were made available by BRGM to INRAE, under the framework of this study and the license agreement n°2019/04. The

sampling and most of the soil analyses were funded by a French Scientific Group of Interest on Soils, the "GIS Sol", involving the French Ministry for Ecology and Sustainable Development, the French Ministry of Agriculture, the French Agency for Energy and Environment (ADEME), the National Institute for Agronomic Research (INRAE), the Institute for Research and Development (IRD), the National Forest Inventory (IFN) and the French Agency for Biodiversity, and by local department funds. We thank all of the soil surveyors and technical assistants involved in sampling the sites, preparing soil samples, and producing soil analyses. This research was conducted within the "Centre d'Expertise Scientifique Cartographie Numérique des Sols" granted by the CNES TOSCA program. We are grateful to the INRAE MIGALE bioinformatics facility (MIGALE, INRAE, 2020. Migale bioinformatics Facility, doi: 10.15454/1.5572390655343293E12) for providing computing and storage resources. We thank Mercedes Román Dobarco for her help with the implementation of the Taylor first-order analysis. Thanks to Nicolas Saby for fruitful discussions about this work. The work of Thomas Loiseau was funded by the French Ministry in charge of agriculture. We thank David G. Rossiter and an anonymous reviewer for their helpful and constructive comments on the paper. D.A. is the coordinator, B.M. is a member, and A.R.d.F. and P.L. are collaborators of the research consortium GLADSOILMAP supported by LE STUDIUM Loire Valley Institute for Advanced Studies through its LE STUDIUM Research Consortium Programme.

Authors' contribution

D.A. and A.R-d-F conceptualized and supervised the work and obtained the funding for the project. T.L. did the majority of the data treatment and modelling. P.L., C.D., and B.M., substantially improved several early drafts of the paper especially for methods considerations, comments on the results and discussion. C.D. provided point data and his expertise of soil mapping in this region. T.L., D.A. and A.R-d-F wrote the final draft of the paper with a

substantial contribution from B.M., P.L., and C.D. All authors contributed equally to the revision of the paper.

References

Aitchison, J., 1982. The statistical analysis of compositional data. J. R. Stat. Soc. Ser. B Methodol. 44 (2), 139–177. https://doi.org/10.1111/j.2517-6161.1982.tb01195.x

Arrouays, D., Grundy, M.G., Hartemink, A.E., Hempel, J.W., Heuvelink, G.B.M., Hong, S.Y., Lagacherie, P., Lelyk, G., McBratney, A.B., McKenzie, N.J. Mendonça-Santos, M.D., Minasny, B., Montanarella, L., Odeh, I.O.A., Sanchez, P.A., Thompson, J.A., Zhang, G.-L., 2014a. GlobalSoilMap: towards a fine-resolution global grid of soil properties. Advances in Agronomy 125, 93-134.

Arrouays, D., McBratney, A.B., Minasny, B., Hempel, J.W., Heuvelink, G.B.M., Mac Millan, R.A., Hartemink, A.E., Lagacherie, P., McKenzie, N.J., 2014b. The GlobalSoilMap project specifications. In: Arrouays, D., McKenzie, N., Hempel, J., Richer-de-Forges, A.C., McBratney, A.B. (Eds.), GlobalSoilMap. Basis of the Global Spatial Soil Information System. Taylor&Francis, CRC Press, London, 9-13.

Arrouays, D., Leenaars, J., Richer-de-Forges, A.C., Adhikari, K., Ballabio, C., Greve, M., Grundy, M., Guerrero, E., Hempel, J., Hengl, T., Heuvelink, G., Batjes, N., Carvalho, E., Hartemink, A., Hewitt, A., Hong, S.-Y., Krasilnikov, P., Lagacherie, P., Lelyk, G., Libohova, Z., Lilly, A., McBratney, A., Mckenzie, N., Vasques, G., Mulder, V.L., Minasny, B., Montanarella, L., Odeh, I., Padarian, J., Poggio, L., Roudier, P., Saby, N., Savin, I., Searle, R., Stolbovoy, V., Thompson, J., Smith, S., Sulaeman, Y., Vintila, R., Viscarra Rossel, R., Wilson, P., Zhang, G.-L., Swerts, M., van Oorts, K., Karklins, A., Feng, L., Ibelles Navarro,

A.R., Levin, A., Laktionova, T., Dell'Acqua, M., Suvannang, N., Ruam, W., Prasad, J., Patil, N., Husnjak, S., Pásztor, L., Okx, J., Hallet, S., Keay, C., Farewell, T., Lilja, H., Juilleret ,J., Marx, S., Takata, Y., Kayusuki, Y., Mansuy, N., Panagos, P., van Liedekerke, M., Skalsky, R., Sobocka, J., Kobza, J., Eftekhari, K., Kazem Alavipanah, S., Moussadek, R., Badraoui, M., da Silva, M., Paterson, G., da Conceição Gonçalves, M., Theocharopoulos, S., Yemefack, M., Tedou, S., Vrscaj, B., Grob, U., Kozak, J., Boruvka, L., Dobos, E., Taboada, M., Moretti, L., Rodriguez, D., 2017. Soil legacy data rescue via GlobalSoilMap and other international and national initiatives. GeoRes J., 14, 1-19.

Ballabio, C., Panagos, P., Montanarella, L., 2016. Mapping topsoil physical properties at European scale using the LUCAS database. Geoderma 261, 110-123. https://doi.org/10.1016/j.geoderma.2015.07.006.

Bialkowski, A., Tourlière, B., Bernachot, I., Chêne, F., Bauer, H., Bernard, J., 2019. Carte lithologique harmonisée et hiérarchisée V0 (niveau 3) de la Mayenne à l'échelle 1:50 000. BRGM, France (personal communication).

Bonijoly, D., Perrin, J., Truffert, C., Asfirane, F., 1999. Couverture géophysique aéroportée du Massif armoricain - Rap. BRGM R 40471, 75 p., 13 fig., 12 tabl., 2 ann.

BRGM, 2014. Indice de développement et de persistance des réseaux (IDPR). Info Terre-Site cartographique de référence sur les géosciences, Orléans, France. (in French).

CESBIO, 2016. Carte d'occupation des sols 2016, http://dx.doi.org/10.3390/rs9010095

Heuvelink, G.B., Burrough, P.A., Stein, A., 1989. Propagation of errors in spatial modelling with GIS. International Journal of Geographical Information System 3(4), 303-322. https://doi.org/10.1080/02693798908941518

IGN, 2014. BD ALTI®. http://professionnels.ign.fr/bdalti

InfoSol, 2005. Référentiel Régional Pédologique : Cahier des Clauses Techniques Général, INRAE.

Inventaire Forestier National, 2006. BD Forêt®. http://inventaire-forestier.ign.fr

Joly, D., Brossart, T., Cardot, H., Cavailhes, J., Hilal, M., Wavresky, P., 2010. Types of climates on continental France, a spatial construction, Cybergeo: European Journal of Geography [Online], URL: http://journals.openedition.org/cybergeo/23155;. doi: 10.4000/cybergeo.23155.

Lagacherie, P, Arrouays, D, Bourennane, H, Gomez, C, Martin, M, Saby, N., 2019. How far can the uncertainty on a Digital Soil Map be known?: a numerical experiment using pseudo values of clay content obtained from Vis-SWIR Hyperspectral imagery. Geoderma 337, 1320-1328.

Lagacherie, P., Arrouays, D., Bourennane, H., Gomez, C., Nkuba-kasanda, L., 2020. Analysing the impact of soil spatial sampling on the performances of Digital Soil Mapping models and their evaluation : A numerical experiment on Quantile Random Forest using clay contents obtained from Vis-NIR-SWIR hyperspectral imagery. Geoderma, 375, 114503.

Lark, R.M., Bishop, T.F.A., 2007. Cokriging particle size fractions of the soil. Eur. J. Soil Sci. 58 (3), 763–774. https://doi.org/10.1111/j.1365-2389.2006.00866.x

Laroche, B., Richer-de-Forges, A.C., Leménager, A., Arrouays, D., Schnebelen, N., Eimberck, M., Chenu, J.P., 2014. Le programme Inventaire Gestion et conservation des Sols. Volet Référentiel Régional Pédologique. Etude et Gestion des Sols, 21, 125-40. (in French).

Loiseau, T., Chen, S., Mulder, V.L., Román Dobarco, M., Richer-de-Forges A.C., Lehmann, S., Bourenanne, H, Saby N.P.A., Martin, M.P., Vaudour, E., Gomez, C., Lagacherie, P., Arrouays D., 2019. Satellite data integration for soil clay content modelling, at a national

scale. International Journal of Applied Earth Observation and Geoinformation. 82, 101905. https://doi.org/10.1016/j.jag.2019.101905

Loiseau, T., Richer-de-Forges, A.C., Martelet, G., Bialkowski, A., Nehlig, P., Arrouays, D., 2020. Could airborne gamma-spectrometric data replace lithological maps as co-variates for digital soil mapping of topsoil particle-size distribution? A case study in Western France. Geoderma Regional, 22, e00295. https://doi.org/10.1016/j.geodrs.2020.e00295

Meinshausen, N., 2006. "Quantile Regression Forests", Journal of Machine Learning Research 7, 983-99.

Minasny, B., McBratney, A.B., 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Computers and Geosciences 32,1378-1388.

Minasny, B., McBratney, A.B., 2010. Methodologies for global soil mapping. In: Boettinger, J.L., Howell, D., Moore, A.C., Hartemink, A.E., Kienast-Brown, S. (Eds.), Digital Soil Mapping – Bridging Research, Environmental Application, and Operation. Progress in Soil Science, Springer, Dordrecht.

Morvan, X.P.P., Saby, N.P.A., Arrouays, D., Le Bas, C., Jones, R.J.A., Verheijen, F.G.A., Bellamy, P.H., Stephens, M., Kibblewhite, M.G., 2008. Soil monitoring in Europe: a review of existing systems and requirements for harmonisation. Sci. Tot. Env., 391, 1-12.

Ng, W., Minasny B., de Sousa Mendes, W., Melo Demattê, J.A., 2020. The influence of training sample size on the accuracy of deep learning models for the prediction of soil properties with near-infrared spectroscopy data. SOIL, 6, 565–578.

Padarian, J., Minasny, B., McBratney, A.B., 2019. Using deep learning for digital soil mapping. Soil, 5(1), 79-89.

554 Padarian, J., Minasny, B., McBratney, A.B., 2020. Machine learning and soil sciences: A
555 review aided by machine learning tools. Soil, 6, 35-52.

556 Richer-de-Forges, A.C., Arrouays, D., Bardy, M., Bispo, A., Lagacherie, P., Laroche, B.,
557 Lemercier, B., Sauter, J., Voltz, M., 2019. Mapping of Soils and Land-Related Environmental
558 attributes in France: analysis of end-users' needs. Sustainability, 11, 2940.
559 https://doi.org/10.3390/su11102940

560 Robinson, G.W., 1933. The dispersion of soils in mechanical analysis. Bur. Soil Sci. Tech.
561 Commun., 26, 27-28.

562 Román Dobarco, M., Bourennane, H., Arrouays, D., Saby, N.P., Cousin, I., Martin, M.P.,
563 2019. Uncertainty assessment of GlobalSoilMap soil available water capacity products: A
564 French case study. Geoderma, 344, 14-30. https://doi.org/10.1016/j.geoderma.2019.02.036

565 Samuel-Rosa, A., Heuvelink, G.B.M., Vasques, G.M., Anjos, L.H.C., 2015. Do more detailed
566 environmental covariates deliver more accurate soil maps? Geoderma, 243, 214-227.

567 Samuel-Rosa, A., Dalmolin, R.S.D., Moura-Bueno, J.M., Teixeira, W.G., Filippini Albad,
568 J.M., 2020. Open legacy soil survey data in Brazil: geospatial data quality and how to
569 improve it. Scienta Agricola, 77(1), e20170430. DOI: 10.1590/1678-992X-2017-0430.

570 Shrestha, D.L., Solomatine, D.P., 2006. Machine learning approaches for estimation of
571 prediction interval for the model output. Neural Networks, 19(2), 225-235.
572 https://doi.org/10.1016/j.neunet.2006.01.012

573 Somarathna, P.D.S.N., Minasny, B., Malone, B.P., 2017. More data or a better model?
574 Figuring out what matters most for the spatial prediction of soil carbon. Soil Science Society
575 of America Journal, 81(6), 1413-1426.

576 Vaysse, K., Lagacherie, P., 2017. Using quantile regression forest to estimate uncertainty of
577 digital soil mapping products. Geoderma, 291, 55-64.
578 https://doi.org/10.1016/j.geoderma.2016.12.017

579 Vernhet, Y., 2010. Carte géologique harmonisée du département de la Mayenne. Notice
580 technique. Rapport final. BRGM/RP-58050-FR, 213 p. (in French).

581 Voltz, M., Arrouays, D., Bispo, A., Lagacherie, P., Laroche, B., Lemercier, B., Richer-de-
582 Forges, A.C., Sauter, J., Schnebelen, N., 2020. Possible futures of soil-mapping in France.
583 Geoderma Regional. 23, e00334.

584 Wadoux, A.M.C., Brus, D.J., Heuvelink, G.B.M., 2019. Sampling design optimization for soil
585 mapping with random forest. Geoderma. 335, 113913.

586

587