



HAL
open science

Prediction of total silicon concentrations in French soils using pedotransferfunctions from mid-infrared spectrum and pedological attributes

Amélia Landre, Nicolas Saby, Bernard Barthès, Céline Ratié, A Guerin, A. Etayo, Budiman Minasny, Marion Bardy, Jean-Dominique Meunier, Sophie Cornu

► To cite this version:

Amélia Landre, Nicolas Saby, Bernard Barthès, Céline Ratié, A Guerin, et al.. Prediction of total silicon concentrations in French soils using pedotransferfunctions from mid-infrared spectrum and pedological attributes. *Pedometrics* 2017, Jun 2017, Wageningen, Netherlands. 298 p. hal-01602785

HAL Id: hal-01602785

<https://hal.science/hal-01602785>

Submitted on 19 Sep 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - ShareAlike 4.0 International License

1 **Prediction of total silicon concentrations in French soils using**
2 **pedotransfer functions from mid-infrared spectrum and pedological**
3 **attributes**

4
5 Co-authors: Landré, A.^a, Saby, N.P.A.^a, Barthès, B.G.^b, Ratié C.^a, Guerin, A.^c Etayo, A.^c,
6 Minasny, B.^d, Bardy, M.^a, Meunier, J.-D.^e, Cornu, S.^e

7 a) INRA, Infosol, US 1106, Orléans, France

8 b) IRD, UMR Eco&Sols, Montpellier SupAgro, 2 place Viala, 34060 Montpellier Cedex
9 2, France

10 c) INRA, Laboratoire d'Analyses des Sols US, Arras, France

11 d) School of Life and Environmental Sciences and Sydney Institute of Agriculture, The
12 University of Sydney, NSW 2006, Australia

13 e) Aix-Marseille Univ, CNRS, IRD, Coll de France, INRA, CEREGE, Aix-en-Provence,
14 France

15

16 ***Abstract***

17 Silicon (Si) is the second most abundant element of the Earth's crust, and its terrestrial cycle
18 depends on soil, vegetation, and human activities. The spatial extent of terrestrial Si
19 perturbation is poorly documented since maps of Si concentration in soils are rare. In
20 addition, Si content is rarely measured in non-paddy soil databases. Here we demonstrate that
21 pedotransfer functions based on either pedological attributes (particle size fraction, pH,
22 organic carbon, cation exchange capacity, calcium carbonate and parent material) or mid
23 infrared spectra (MIRS) can be used to accurately predict total Si concentration. In this

24 research, we utilised a unique dataset from the French monitoring network of soil quality
25 (RMQS - Réseau de Mesures de la Qualité des Sols) database. Pedotransfer functions were
26 built using a regression tree model on a subset of the data for which total Si concentration was
27 measured. To compare the relative performance of the models obtained for the two different
28 sources of data, a suite of performance indicators were calculated. Our results showed that
29 PTF based on MIR spectra produces highly accurate and precise estimates of the total Si
30 concentration for French soils. The pedological PTF is less accurate, but still provides a good
31 estimation of the Si concentration. The pedological PTF provides an alternative method when
32 only basic soil data are available, and an approximate estimation of Si concentrations is
33 sufficient. These PTFs can be readily applied at the European scale except on a few soil
34 groups not represented in France.

35 **keywords**

36 Silicon, silica, pedotransfer function, diffuse reflectance, FTIR, mid-infrared spectra,
37 regression tree, soil monitoring.

38

39 **1- Introduction**

40 Silicon (Si) is the second most abundant element of the Earth's crust, after oxygen, with an
41 average mass concentration of 28% (Wedepohl, 1995). The SiO₄ tetrahedron is the elemental
42 brick of soil minerals which constitutes the basic structure of Si in nature from solid (silicates)
43 to soluble (silicic acid) states. Through weathering of primary minerals, Si is released into soil
44 solution where it is either recombined with other elements to form secondary minerals,
45 adsorbed on oxides surfaces, absorbed by plants or exported to groundwater and rivers.
46 Indeed, despite being considered as a non-essential element, Si is encountered in most
47 terrestrial plants with concentrations highly variable, ranging from less than 0.2 to more than
48 10 % dry weight (Ma and Takahashi, 2002). Meanwhile, plant Si is a significant pool of the
49 global Si cycle as evidenced from the total annual biogenic Si retention in terrestrial plants,
50 which is estimated in the same order of magnitude as the Si fixed annually by diatoms in the
51 ocean (Carey and Fulweiler, 2012; Loucaides et al., 2010). The terrestrial Si cycle notably
52 depends on the type of vegetation (Alexandre et al., 1997; Bartoli, 1983; Blecker et al., 2006;
53 Cornelis et al., 2010) and is suggested to be perturbed by human activities through urban
54 activities, agriculture and deforestation (Conley et al., 2008; Struyf et al., 2010; Vandevenne
55 et al., 2012). However, the extent of this perturbation is not well documented. One approach
56 to address this question is a spatial analysis at the territory scale. Soil Si maps are rare to our
57 knowledge, with the notable exception of the Si map for European soils with one site every
58 2500 km² (De Vos et al., 2006; Reimann et al., 2014). Soil silica content is mostly measured
59 in rice-growing areas (Minasny et al., 2016), but rarely measured in non-paddy soil databases,
60 especially in most of the national soil monitoring networks of Europe (Arrouays et al., 2008;
61 Imrie et al., 2008; Morvan et al., 2008). The possible reasons are (1) the cost of the
62 measurement and (2) little attention paid to Si in soil and agriculture until now.

63 In the literature, several soil characteristics, generally hydraulic properties or bulk density,
64 (Baldwin et al., 2017; Van Looy et al., 2017; Wösten et al., 1999), among others), have been
65 estimated by predictive functions based on pedological characteristics (organic matter,
66 particle size distribution etc.). This is commonly known as pedotransfer function (PTFs).
67 Recently, infrared spectroscopy has been proposed as a rapid and cost-effective alternative of
68 conventional chemical analysis as one spectrum can be used to derive several soil parameters
69 (Bertrand et al., 2002; Janik et al., 1998; McCarty et al., 2002; Minasny et al., 2009; Soriano-
70 Disla et al., 2014, 2013; Viscarra Rossel et al., 2006). In their review, Viscarra Rossel et al.
71 (2006) demonstrated mid infrared spectra (MIRS) produce better accuracy over near infrared
72 spectra (NIRS) for a large number of soils properties. However, that review did not show any
73 study that predicts Si from infrared spectra. Nevertheless, Janik et al. (1995) showed that soils
74 dominated by quartz, which is made up of SiO₂, has a peak around 1700-2000 cm⁻¹. Other
75 studies showed that MIRS presented good predictions of oxalate extractable Si, a specific
76 form of Si used to assess the degree of soil weathering (Bertrand et al., 2002; Minasny et al.,
77 2009). These findings suggest that MIRS could be used for prediction of total Si as recently
78 demonstrated by Mohanty et al. (2016).

79 In this study, we developed and compared pedotransfer functions of total topsoil Si
80 concentrations based on either pedological data (particle size fraction, pH, organic carbon,
81 cation exchange capacity, calcium carbonate and parent material) or MIRS. These data were
82 derived from the French monitoring network of soil quality database, RMQS (Réseau de
83 Mesures de la Qualité des Sols (Arrouays et al., 2003; Jolivet et al., 2006). Pedotransfer
84 functions were built on the RMQS data subset on which total Si concentrations were
85 measured. Subsequently, the PTFs were applied to the whole RMQS dataset to predict topsoil
86 Si concentration for the whole of France.

87 **2- Materials and methods**

88 **2.1- Soil samples**

89 **2.1.1- The RMQS database**

90 The RMQS monitoring network encompasses 2088 sites sampled following a $16 \times 16 \text{ km}^2$
91 regular grid across the French metropolitan territory ($550\,000 \text{ km}^2$). At each site, land-use,
92 climate, soil type and parent material were described. Twenty five individual cores were
93 sampled, from the topsoil (0-30 cm), using an unaligned sampling design within a $20 \times 20 \text{ m}$
94 area (Jolivet et al., 2006). Core samples were then bulked. The resulting composite samples
95 were air-dried and sieved to 2 mm before analysis. Particle-size distribution, pH in water,
96 organic carbon (OC), cation exchange capacity (CEC), calcium carbonate (CaCO_3) were
97 analyzed according to international (ISO) or French (NF) standardized methods by the
98 accredited Soil Analysis Laboratory of INRA (Arras, France) (Table 1). This study is based
99 on the analytical results of the samples collected from 2002 to 2009 (first sampling campaign)
100 for the previous characteristics.

101 **2.1.2- MIRS data**

102 MIRS were also acquired on RMQS samples (Grinand et al., 2012). 0.5-g aliquots of < 0.2 -
103 mm ground sample were scanned from 4000 to 400 cm^{-1} (i.e., 2500 - $25,000 \text{ nm}$) at 4 cm^{-1}
104 resolution using a Nicolet 6700 Diffusive Reflectance Fourier Transform Spectrophotometer
105 (Thermo Fisher Scientific Instruments, Madison, WI, USA). Then, 32 scans per sample were
106 acquired and averaged. Spectra were recorded as absorbance.

107 MIRS were pre-processed, before statistical modelling to reduce baseline variations, enhance
108 spectral features, reduce the particle-size scattering effect, remove linear or curvilinear trends
109 of each spectrum, or remove additive or multiplicative signal effects (Boysworth and Booksh,

110 2008). The pre-processing routine consisted of an 11 bands window smoothing Savitzky–
111 Golay filter (Savitzky and Golay, 1964) using the *sgolayfilt* function from the signal R
112 package (Ligges et al., 2015) followed by a standard normal variate (SNV, Barnes et al.,
113 1989) transform.

114 **2.1.3- Selection of samples for total Si analysis**

115 To develop the pedotransfer functions, a subset of 673 samples from the 2088 RMQS
116 samples, called hereafter the Si dataset, were analyzed for total Si concentration. The subset
117 sites were selected using the following criteria: (1) one site out of four from the original grid
118 excluding Corsica, and (2) 160 sites randomly selected from the remaining sites (Figure 1).
119 This sub-sampling of the grid preserves the systematic grid sampling. This gridsampling
120 method was established by (Brus and Saby, 2016) as a flexible design for statistical soil
121 surveys leading to relatively accurate estimates of the statistical distribution of spatial
122 parameters. Total Si concentration was measured on air-dried, less than 2 mm samples by
123 inductively coupled plasma atomic emission spectrometry (ICP-AES) after sodium peroxide
124 fusion of the samples.

125 **2.2- Total Si modelling and predictions**

126 **2.2.1- Pedotransfer functions (PTFs)**

127 Pedotransfer functions predicting total Si content was established using two different types of
128 soil variables as inputs:

- 129 (1) basic pedological attributes including particle size fraction, pH, organic carbon, cation
130 exchange capacity, calcium carbonate and parent material, or
- 131 (2) pre-processed MIRS data,

132 These PTFs are termed as pedological PTF and MIRS PTF hereafter. The variables of the
133 pedological PTF were chosen from the most common soil analytical variables in soil
134 databases. For that reason, parent material was also included, while soil type was not as it had
135 only a limited influence in the model (data not shown).

136 **2.2.2- Statistical modelling approach**

137 **Regression procedure:** Quantitative prediction of total Si concentrations by the soil
138 properties and the MIRS were obtained using Cubist, a type of regression tree model
139 (Quinlan, 1992). The Cubist model is a form of regression rules that build regression trees
140 with final nodes containing linear models instead of discrete values. Cubist creates
141 interpretable rules that describe the relationships between predictive variables (in this case
142 spectral bands or soil properties) and the variable of interest (Si). Minasny et al. (2009) and
143 Minasny and McBratney (2008) demonstrated that this approach could provide higher
144 accuracy than the partial-least-squares (PLS) approach, commonly used in chemometrics.
145 Moreover, this type of approach is flexible as it can handle both quantitative and qualitative
146 variables as well as spectral data, which allows having a unique approach for both PTFs, and
147 thus their results can be fairly compared.

148 To optimize the Cubist model, two parameters can be adjusted: the number of model trees as
149 ensembles (*committees*) and the number of nearest-neighbors to adjust the prediction of the
150 rules (*neighbors*). To optimize the model parameters, we used the *train* function in the caret
151 R package. The tuning parameter 'neighbors' was held constant at a value of zero to avoid
152 shortcomings in the interpretability of the rules by local averaging. The optimal numbers of
153 'committees' was found to be 5 for the two PTFs.

154 **Calibration and evaluation steps:** The modelling approach is summarized in Figure 2. we
155 used a leave p out cross-validation approach combined with a bootstrap step (James et al.,
156 2013). The p cross-validation leaves out a p proportion of samples for validation. We used a

157 75%-25% split for calibration and validation respectively, and this procedure was repeated 10
158 times. Cross-validation allows a more robust assessment of the quality of the prediction. The
159 subdivision was performed using the conditioned Latin Hypercube Sampling (cLHS) method
160 (Minasny and McBratney, 2006). This method is a stratified random procedure that provides
161 an efficient way of sampling variables from their multivariate distributions. The bootstrap step
162 involved simulating 100 datasets by random sampling with replacement from 95% of the
163 calibration dataset (formed at the cross-validation step). This whole procedure generated 100
164 Cubist models in the calibration procedure. These outcomes were used to build the
165 distribution of the prediction. The mean prediction could be obtained from the average of the
166 100 bootstrapped models.

167 **Software:** Our modelling approach involves a large number of model calibrations and parallel
168 processing was used to handle the computational load. Parallel processing was implemented
169 in R using the packages foreach (Calaway et al., 2017b) and doParallel (Calaway et al.,
170 2017a). All data analyses were performed using the R statistical environment (R core Team,
171 2017) for descriptive statistics, spectrum pre-processing and model building. We used the
172 Cubist implementation from the Cubist package (Kuhn et al., 2016), the cLHS function
173 implemented in clhs package (Roudier, 2011).

174 ***2.3- Assessments and interpretations***

175 **2.3.1- Representativeness assessment**

176 Before evaluating the prediction ability of the two PTFs, we first checked that the Si dataset
177 used to develop the PTFs was representative of the whole RMQS dataset. Graphical and
178 numerical comparison of the statistical distribution of the basic pedological data, and the MIR
179 spectra were performed. The Wilcoxon test for the quantitative soil attributes was also
180 performed to compare the two datasets (the Si dataset and the RMQS database). In addition,

181 principal component analyses (PCA) was performed on the pedological attribute as well as the
182 MIRS variables. Using the PCA, samples with or without Si measurements can be readily
183 compared. The PCA allows comparison of the multivariate variables, while the distribution
184 analyses only consider one soil variable at a time.

185

186 **2.3.2- Accuracy assessment**

187 To compare the relative performance of the models obtained for the two PTFs, three
188 conventional performance indicators were calculated: the coefficient of determination (R^2),
189 the root mean square error (RMSE, also known as standard error of prediction, SEP) and the
190 bias, which is the mean residual of the model. In addition, we took into account the
191 probability distribution of model predictions using the continuous rank probability score
192 average (CRPS, equation 6). The CRPS represents the closeness between the prediction
193 distribution and the corresponding observations (Gneiting et al., 2007). This score is
194 commonly used in meteorological forecasts as a verification tool for (probabilistic) forecast
195 systems (Hersbach, 2000; Trinh et al., 2013). The metric is calculated using:

$$196 \quad CRPS = \int_{-\infty}^{\infty} BS(y) dy, \quad (6)$$

$$197 \quad BS(y) = \frac{1}{n} \sum_{i=1}^n \{(F_i(y) - \mathbb{1}(x_i \leq y))\}^2, \quad (7)$$

198 where $BS(y)$ denotes the Brier score (Brier, 1950) for probability forecasts of the binary
199 event at the threshold value $y \in \mathbb{R}$, x is the observation and y is the model prediction, n the
200 number of samples, F is the cumulative distribution function (CDF) of X , a random variable,
201 such as $F(y) = P[X \leq y]$ and $\mathbb{1}$ is the Heaviside step function. This function is a
202 discontinuous function where the value is zero for negative argument and unity for positive
203 argument.

204 The CRPS is a distance criterion, which is a positive value and should be close to 0. The
205 prediction is expressed in terms of a probability distribution rather than a single value. The

206 CRPS compares the cumulative probability distribution of the predicted value to the observed
207 value. In our case, we only took into account the uncertainty of the prediction and assumed
208 the uncertainty of the observation is small. The probability distribution of our observation is
209 set to equal to 1 for the observed value and null elsewhere. As a distance, the CRPS can be
210 linked to the mean absolute error used in the deterministic prediction. It uses the information
211 provided by the probabilistic prediction instead of just using the mean of the median value.
212 We used the crps function implemented in the verification package (Laboratory NCAR-
213 Research Applications, 2015).

214

215 **2.3.3- The importance of the predictors in the model**

216 In order to interpret the PTFs results, we extracted and computed the variable of importance
217 from the Cubist rulesets. The variable of importance is computed as the percentage of times
218 each variable was used in a rule condition and/or a linear model. Following our calibration
219 and validation step, we calculated the average importance of predictors over the 100 Cubist
220 models produced by bootstrap and then over the 10 iterations of the cross-validation step for
221 each PTF. Because it is an average value, the sum of the variables of importance do not sum
222 up to 100.

223 **3- Results & Discussion**

224 ***3.1- Representativeness of the Si dataset compared to the whole RMQS*** 225 ***dataset***

226 Measured Si concentrations range from 22.81 to 455.8 g kg⁻¹ over the Si dataset with a
227 median equal to 327.2 g kg⁻¹ (Figure 3). Soils with low total Si concentrations (under
228 124 g kg⁻¹, the statistical threshold for outliers in this dataset) were poorly represented (17
229 over 674; Figure 3). The corresponding samples originated from soils developed in
230 sedimentary parent materials, mostly calcareous with a carbonate concentration of greater or
231 equal to 395 g kg⁻¹.

232 The parent material distribution of both the Si dataset (n=673) and the RMQS set (n= 2088)
233 are very similar (Figure 4). For the pedological attributes, the empirical density estimates of
234 soil properties were well represented, both in the Si dataset and the RMQS (Figure 5). The
235 summary statistics of the pedological attributes were reported in Table 2. The empirical
236 density functions for both datasets overlapped. This is supported by the Wilcoxon test which
237 showed no significant difference in the distribution for the considered attributes (p-values
238 recorded in Table 2). Considering the whole dataset, the PCAs showed a good overlap
239 between the RMQS sites with and without Si measurement (Figure 6). Therefore, we can
240 consider the Si dataset to be representative of the whole RMQS.

241

242 ***3.2- Total Si prediction by the PTFs***

243 The validation of the MIRS PTF estimating Si content was excellent with an R² of 0.96.

244 Estimates from this PTF were unbiased, and their average RMSE is 15.31 g kg⁻¹ (Table 3).

245 The average CRPS was very close to the RMSE value. The validation of the pedological PTF

246 was also very good with an R^2 of 0.87. Estimates from the pedological PTF were slightly
247 biased, with an average RMSE of 26.48 g kg^{-1} . Finally, the average CRPS were larger
248 indicating higher prediction uncertainty. The results of leave p out cross-validation were also
249 used to compute indicators of the variability of the performance indicators (Table 3). The
250 standard deviation of these indicators for the MIRS PTF was small.

251 To better figure out the difference between the accuracy of the two PTFs, we plotted the
252 predicted *versus* measured Si concentrations for one iteration of the cross-validation steps
253 (Figure 7). For the MIRS PTF, the prediction and analytical uncertainties of the data are of the
254 same order of magnitude (Figure 7a) as suggested by the RMSE values. This means that this
255 PTF gives good predictions, and close to analytical measurements. In contrast, the pedological
256 PTF presents a larger prediction uncertainty than the MIRS PTF, and it also has larger
257 uncertainty compared to the analytical uncertainty (Figure 7b) as shown by RMSE results. In
258 addition, total Si concentration tends to be over-estimated by the pedological PTF at low
259 concentrations ($\leq 270 \text{ g kg}^{-1}$) and under-estimated at high concentrations ($\geq 370 \text{ g kg}^{-1}$). This
260 is further confirmed by the coefficients of the linear regression between observed and
261 predicted values of the pedological PTF, with 67.18 for the intercept and 0.80 for the slope. In
262 comparison, the MIRS PTF has an intercept of 22 g kg^{-1} and slope of 0.07. The bias maybe
263 due to the low representation in the dataset of samples having Si concentration lower than
264 12.5 %, as discussed earlier. The bias can also come from sites which were over predicted
265 where the Si concentrations are between 200 and 300 g kg^{-1} .

266 All in all, the MIRS PTF tends to show an accuracy as good as the chemical analysis when
267 considering both the prediction and analytical uncertainties. However, as predictions were
268 made on the basis of analytical measurements, the prediction uncertainty does not only come
269 from the model accuracy but also from the uncertainty of analytical measurements (Janik et
270 al., 1998). In our case, the analytical uncertainty was not taken into account in the prediction

271 uncertainty calculation, as they were not always available. Despite this, the obtained PTFs
272 show an exceptional accuracy that is rarely obtained in the PTFs literature (Minasny et al.,
273 2009; Viscarra Rossel et al., 2006). Viscarra Rossel et al. (2006) reported the accuracy of
274 MIRS PTFs for different soil properties from the literature have R^2 values ranging from 0.07
275 to 0.98, where one third of the cases an R^2 value larger or equal to 0.90 was obtained.
276 Minasny et al. (2009) built MIRS PTFs for predicting soil properties on three different
277 databases and reported R^2 values from 0.0 for total S to 0.92 for CEC and OC. They
278 concluded that basic soil organic and mineral constituents, as well as properties that are
279 related to the mineral and organic components could be well predicted. This study confirms
280 the hypothesis. PTFs based on pedological properties are generally used to predict
281 hydrological properties that are difficult to measure (e.g., Baldwin et al., 2017; Wösten et al.,
282 1999, among others) but seldom developed to predict chemical characteristics.

283 ***3.3- The Pedological significance of the calibrated PTFs***

284 Regarding pedological significance, the MIRS PTF uses mostly combination-overtone bands
285 of quartz ranging from 1800 to 2000 cm^{-1} (Table 4), to predict total Si concentration, which is
286 expected, as quartz is a mineral composed of Si and oxygen (O) atoms (SiO_2) (Figure 8a).
287 This region of the spectrum presents the peak with the most important weight (>80%) around
288 2000 cm^{-1} followed by two other peaks around 1800 and 1900 cm^{-1} (> 40%, Figure 8b, Table
289 4). The carbonate concentration also has a role in the MIRS PTF, with carbonates bands
290 ranging from 2400 to 3100 cm^{-1} , which correspond to CaCO_3 bonds (Table 4). It exhibits
291 three peaks of average weight > 20%, one around 2500 and two around 3000 cm^{-1} (Figure 8b).
292 As shown in Figure 8a, samples with low Si concentration contain carbonates while samples
293 with high Si concentration do not. This link is due to the absence of Si in carbonates (Table
294 4). Bands related to Si-O bond ranging from 1400 to 400 cm^{-1} also presents a noticeable
295 weight in the PTF (Figure 8b).

296 For the pedological PTF, the two most important predictors are the organic carbon and the
297 carbonate concentrations, with an average weight of 78.6% and 73.6% respectively.
298 Carbonates as discussed for the MIRS PTF act as a diluent for Si, which is also the case of
299 organic carbon. No significant organic carbon contribution was observed in the MIRS PTF
300 probably because, in MIRS, "organic carbon cannot be identified with clearly separated peaks
301 but as a whole spectral region with overlapping bands", as stated by Grinand et al. (2012). In
302 the pedological PTF, an important influence of the sand fraction could be expected as a
303 positive correlation between total Si and both fine and coarse sand is observed (Kendall's
304 correlation coefficient: $\tau = 0.17$ and $p\text{-value} = 3.585 \cdot 10^{-11}$; $\tau = 0.11$ and $p\text{-value} = 3.108$
305 10^{-5} , respectively). Indeed, these two variables have an average weight of 37% in the
306 pedological PTF (Figure 9). In addition, the clay fraction also has an important weight in the
307 pedological PTF (60%) with a negative correlation between the total Si concentration and the
308 clay fraction (Kendall's correlation coefficient: $\tau = -0.46$ and $p\text{-value} < 2.2e-16$).
309 As a conclusion, the two PTFs were mainly underlined by the same processes: dilution of the
310 Si concentration by carbonates, organic carbon and possibly the clay fraction to a lesser extent
311 and concentration due to the presence of quartz mainly in the sand fractions.

312 ***3.4- Domain of potential application of the developed PTFs***

313 We compared pedological PTF predictions of total Si concentration for the non-Si analysed
314 RMQS sites to that predicted by the MIRS PTF (Figure 10). The relative difference of
315 predictions between the two PTFs is small, 90% of the time, the difference between the 2
316 PTFs is less than 20%. This result highlights the consistency of the two PTFs and confirms
317 that despite less accurate, the pedological PTF gives a reasonable estimation of the soil Si
318 concentration for the whole dataset. The soil observations of this study came from a
319 systematic probability sampling which leads to good spatial coverage, i.e. the sites are
320 uniformly spread over France. This design proves to be efficient in providing accurate

321 estimates of means over the whole area and can be used to generalize the results for the whole
322 area of France (Brus and Saby, 2016).

323 We further investigated if the domain of application of our PTFs at the European scale by
324 comparing the Si concentrations statistical distribution of the Si measurements analytically
325 measured from the RMQS to that of the Geochemical Mapping of Agricultural Soils
326 (GEMAS) dataset (De Vos et al., 2006; Reimann et al., 2014). The GEMAS study provided a
327 few soil Si data over the French territory with one site every 2500 km², i.e. 214 sites for
328 France. Comparing those two distributions for France showed 1) a slight over-estimation of
329 the occurrence of soils with concentrations around 300 g kg⁻¹ ; 2) an under-estimation of soils
330 with concentrations around 400 g kg⁻¹ ; 3) a slight smaller median of the French soil Si
331 concentrations (320.7 g kg⁻¹). Nevertheless, the Mann and Whitney test shows no significant
332 difference between the two datasets (p-value = 0.5757). For the European territory, the
333 comparison is shown in Figure 11. The two datasets cover the same range of Si concentration
334 with an over-representation of the soils with Si concentration ranging from 350 to 400 g kg⁻¹
335 in French compare to other European soils, resulting in contrasted median Si-concentrations
336 of 327.2 g kg⁻¹ and 313.9 g kg⁻¹ respectively. When looking at the Si average, the Mann and
337 Whitney test also shows a small significant difference between the two datasets ($0.05 > p$ -
338 value = 0.02273 > 0.01). This result was expected since France is one of the countries
339 exhibiting the largest soil diversity in the world (Minasny et al., 2010). Thus, the established
340 PTFs can be applied at the Europeans scale to predict total soil Si concentrations at a higher
341 spatial density than that provided by the GEMAS study with the exception of some soil types
342 that are not represented in France, such as Chernozems, Kasternozem, Solonetz.

343 Finally, to better define the application range of our PTFs outside of Europe, future users can
344 determine the appropriate domain of application of a specific PTF to a new dataset using
345 distance metrics, such as the one presented by Tranter et al. (2009).

346 **4- Conclusions**

347 We developed PTFs based on either MIRS or pedological data to estimate the topsoil total Si
348 concentration. Both PTFs provide accurate estimations of the total Si concentration for French
349 soils. These PTFs are underlined by the link between Si and quartz, organic matter, and
350 carbonate contents. The PTF based on MIRS data produces a highly accurate and precise
351 estimates. Since the acquisition of MIRS data allows the estimation of a range of soil
352 properties, such as particle size fraction, major elements or chemical properties that are
353 related to surface solid characteristics like CEC, the use of MIRS PTF represents a powerful
354 tool for populating soil databases. The pedological PTF is less accurate, but still provides a
355 reasonable estimation of the Si concentration for French soils. It is an alternative method
356 when only pedological data are available and an approximate estimation of Si concentrations
357 is sufficient. This PTF can be applied to databases of legacy soil data to provide an initial
358 estimate of Si distribution.

359 Both PTFs can be readily applied at the European scale with the possible exclusion of a few
360 soil groups not represented in France. For these soil types, this study provides a pathway for
361 the development of new calibration PTFs procedure to local data.

362 This modelling approach yields very robust results with an adaptable method. Overall, this
363 work provides the first approach to estimate nation-wide topsoil total Si concentration and
364 opens the way for further works on Si in soils.

365 **5- Acknowledgements**

366 This work was performed in the frame of the French ANR BioSiSol project (ANR-14-CE01-
367 0002). RMQS soil sampling and physico-chemical analyses were supported by the GIS Sol,
368 which is a scientific group of interest on soils involving the French Ministry for ecology and
369 sustainable development and Ministry of agriculture, the French National forest inventory
370 (IFN), ADEME (Agence de l'environnement et de la maîtrise de l'énergie, which is a French
371 government agency concerned with environmental protection and energy management), IRD
372 (Institut de recherche pour le développement, which is a French public research organization
373 dedicated to southern countries) and INRA (Institut national de la recherche agronomique,
374 which is a French public research organization dedicated to agriculture s.l.). Claudy Jolivet is
375 thanked for his strong involvement in the RMQS monitoring network. Manon Villeneuve,
376 Emmanuel Bourdon, Didier Brunet, Jérôme Lourd, and Clément Robin (IRD) are thanked for
377 their skillful technical assistance and strong investment in the tedious scanning process of the
378 RMQS soil library.

379

6- References

- 381 Alexandre, A., Meunier, J.-D., Colin, F., Koud, J.-M., 1997. Plant impact on the
382 biogeochemical cycle of silicon and related weathering processes. *Geochim.*
383 *Cosmochim. Acta* 61, 677–682. [https://doi.org/10.1016/S0016-7037\(97\)00001-X](https://doi.org/10.1016/S0016-7037(97)00001-X)
- 384 Arrouays, D., Jolivet, C., Boulonne, L., Bodineau, G., Ratié, C., Saby, N., Grolleau, E., 2003.
385 Le réseau de mesures de la qualité des sols (RMQS) de France. *Etude Gest. Sols* 10,
386 241.
- 387 Arrouays, D., Morvan, X., Saby, N., Richer de Forges, A., Le Bas, C., Bellamy, P.H., Üveges,
388 B., Freudenschu\s s, A., Jones, A.R., Jones, R.J.A., others, 2008. Environmental
389 Assessment of Soil for Monitoring. Volume Ila: Inventory & Monitoring. European
390 Communities.
- 391 Baldwin, D., Manfreda, S., Keller, K., Smithwick, E.A.H., 2017. Predicting root zone soil
392 moisture with soil properties and satellite near-surface moisture data across the
393 conterminous United States. *J. Hydrol.* 546, 393–404.
394 <https://doi.org/10.1016/j.jhydrol.2017.01.020>
- 395 Barnes, R.J., Dhanoa, M.S., Lister, S.J., 1989. Standard Normal Variate Transformation and
396 De-Trending of Near-Infrared Diffuse Reflectance Spectra. *Appl. Spectrosc.* 43, 772–
397 777. <https://doi.org/10.1366/0003702894202201>
- 398 Bartoli, F., 1983. The biogeochemical cycle of silicon in two temperate forest ecosystems.
399 *Ecol. Bull.* 469–476.
- 400 Bertrand, I., Janik, L.J., Holloway, R.E., Armstrong, R.D., McLaughlin, M.J., 2002. The rapid
401 assessment of concentrations and solid phase associations of macro-and micronutrients
402 in alkaline soils by mid-infrared diffuse reflectance spectroscopy. *Soil Res.* 40, 1339–
403 1356.
- 404 Blecker, S.W., McCulley, R.L., Chadwick, O.A., Kelly, E.F., 2006. Biologic cycling of silica
405 across a grassland bioclimosequence: GRASSLAND SILICA CYCLING. *Glob.*
406 *Biogeochem. Cycles* 20, n/a-n/a. <https://doi.org/10.1029/2006GB002690>
- 407 Boysworth, M.K., Booksh, K.S., 2008. Aspects of Multivariate Calibration Applied to Near-
408 Infrared Spectroscopy, in: *Handbook of Near-Infrared Analysis, Third Edition,*
409 *PRACTICAL SPECTROSCOPY SERIES.* pp. 207–229.
- 410 Brier, G.W., 1950. Verification of forecasts expressed in terms of probability. *Mon. Weather*
411 *Rev.* 78, 1–3.
- 412 Brus, D.J., Saby, N.P.A., 2016. Approximating the variance of estimated means for
413 systematic random sampling, illustrated with data of the French Soil Monitoring
414 Network. *Geoderma* 279, 77–86.
- 415 Calaway, R., Corporation, M., Weston, S., Tenenbaum, D., 2017a. doParallel: Foreach
416 Parallel Adaptor for the “parallel” Package.
- 417 Calaway, R., Microsoft Corporation, Weston, S., 2017b. foreach: Provides Foreach Looping
418 Construct for R.
- 419 Carey, J.C., Fulweiler, R.W., 2012. The terrestrial silica pump. *PLoS One* 7, e52932.
- 420 Conley, D.J., Likens, G.E., Buso, D.C., Saccone, L., Bailey, S.W., Johnson, C.E., 2008.
421 Deforestation causes increased dissolved silicate losses in the Hubbard Brook
422 Experimental Forest. *Glob. Change Biol.* 14, 2548–2554.
- 423 Cornelis, J.-T., Ranger, J., Iserentant, A., Delvaux, B., 2010. Tree species impact the
424 terrestrial cycle of silicon through various uptakes. *Biogeochemistry* 97, 231–245.
- 425 De Vos, W., Tarvainen, T., Salminen, R., Reeder, S., De Vivo, B., Demetriades, A., Pirc, S.,
426 Batista, M.J., Marsina, K., Ottesen, R.T., others, 2006. *Geochemical Atlas of Europe:*

427 Part 2: Interpretation of geochemical maps, additional tables, figures, maps, and
428 related publications. Geological Survey of Finland.

429 Gneiting, T., Balabdaoui, F., Raftery, A.E., 2007. Probabilistic forecasts, calibration and
430 sharpness. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 69, 243–268.

431 Grinand, C., Barthès, B.G., Brunet, D., Kouakoua, E., Arrouays, D., Jolivet, C., Caria, G.,
432 Bernoux, M., 2012. Prediction of soil organic and inorganic carbon contents at a
433 national scale (France) using mid-infrared reflectance spectroscopy (MIRS). *Eur. J.*
434 *Soil Sci.* 63, 141–151. <https://doi.org/10.1111/j.1365-2389.2012.01429.x>

435 Hersbach, H., 2000. Decomposition of the Continuous Ranked Probability Score for
436 Ensemble Prediction Systems. *Weather Forecast.* 15, 559–570.
437 [https://doi.org/10.1175/1520-0434\(2000\)015<0559:DOTCRP>2.0.CO;2](https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2)

438 Imrie, C.E., Korre, A., Munoz-Melendez, G., Thornton, I., Durucan, S., 2008. Application of
439 factorial kriging analysis to the FOREGS European topsoil geochemistry database.
440 *Sci. Total Environ.* 393, 96–110. <https://doi.org/10.1016/j.scitotenv.2007.12.012>

441 James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. *An Introduction to Statistical*
442 *Learning: with Applications in R.* Springer Science & Business Media.

443 Janik, L., Skjemstad, J., Raven, M., 1995. Characterization and Analysis of Soils Using
444 Midinfrared Partial Least-Squares .1. Correlations with Xrf-Determined Major-
445 Element Composition. *Aust. J. Soil Res.* 33, 621–636.
446 <https://doi.org/10.1071/SR9950621>

447 Janik, L.J., Skjemstad, J.O., Merry, R.H., 1998. Can mid infrared diffuse reflectance analysis
448 replace soil extractions? *Aust. J. Exp. Agric.* 38, 681.
449 <https://doi.org/10.1071/EA97144>

450 Jolivet, C., Arrouays, D., Boulonne, L., Ratié, C., Saby, N., 2006. Le réseau de mesures de la
451 qualité des sols de France (RMQS). *Etat D'avancement Prem. Résultats Etude Gest.*
452 *Sols* 13, 149–164.

453 Kuhn, M., Weston, S., Keefer, C., Coulter, N., 2016. *Cubist: Rule- And Instance-Based*
454 *Regression Modeling ; R package version 0.0.19.* C code for Cubist by Ross Quinlan.
455 Laboratory NCAR-Research Applications, 2015. *verification: Weather Forecast Verification*
456 *Utilities.*

457 Ligges, U., Short, T., Kienzle, P., Schnackenberg, S., Billingham, D., Borchers, H.-W.,
458 Carezia, A., Dupuis, P., Eaton, J.W., Farhi, E., Habel, K., Hornik, K., Krey, S., Lash,
459 B., Leisch, F., Mersmann, O., Neis, P., Ruohio, J., III, J.O.S., Stewart, D., Weingessel,
460 A., 2015. *signal: Signal Processing.*

461 Loucaides, S., Behrends, T., Van Cappellen, P., 2010. Reactivity of biogenic silica: Surface
462 versus bulk charge density. *Geochim. Cosmochim. Acta* 74, 517–530.

463 Ma, J.F., Takahashi, E., 2002. *Soil, fertilizer, and plant silicon research in Japan.* Elsevier,
464 Amsterdam.

465 McCarty, G.W., Reeves, J.B., Reeves, V.B., Follett, R.F., Kimble, J.M., 2002. Mid-infrared
466 and near-infrared diffuse reflectance spectroscopy for soil carbon measurement. *Soil*
467 *Sci. Soc. Am. J.* 66, 640–646.

468 Minasny, B., Hong, S.Y., Hartemink, A.E., Kim, Y.H., Kang, S.S., 2016. Soil pH increase
469 under paddy in South Korea between 2000 and 2012. *Agric. Ecosyst. Environ.* 221,
470 205–213. <https://doi.org/10.1016/j.agee.2016.01.042>

471 Minasny, B., McBratney, A.B., 2008. Regression rules as a tool for predicting soil properties
472 from infrared reflectance spectroscopy. *Chemom. Intell. Lab. Syst.* 94, 72–79.

473 Minasny, B., McBratney, A.B., 2006. A conditioned Latin hypercube method for sampling in
474 the presence of ancillary information. *Comput. Geosci.* 32, 1378–1388.
475 <https://doi.org/10.1016/j.cageo.2005.12.009>

476 Minasny, B., Tranter, G., McBratney, A.B., Brough, D.M., Murphy, B.W., 2009. Regional
477 transferability of mid-infrared diffuse reflectance spectroscopic prediction for soil
478 chemical properties. *Geoderma* 153, 155–162.
479 <https://doi.org/10.1016/j.geoderma.2009.07.021>

480 Mohanty, B., Gupta, A., Das, B.S., 2016. Estimation of weathering indices using spectral
481 reflectance over visible to mid-infrared region. *Geoderma* 266, 111–119.
482 <https://doi.org/10.1016/j.geoderma.2015.11.030>

483 Monger, H.C., Kelly, E., 2002. Silica minerals, in: *Soil Mineralogy with Environmental*
484 *Applications.*, Soil Science Society of America Book Series. Soil Science Society of
485 America Inc., Madison, Wisconsin, USA, pp. 611–636.

486 Morvan, X., Saby, N.P.A., Arrouays, D., Le Bas, C., Jones, R.J.A., Verheijen, F.G.A.,
487 Bellamy, P.H., Stephens, M., Kibblewhite, M.G., 2008. Soil monitoring in Europe: A
488 review of existing systems and requirements for harmonisation. *Sci. Total Environ.*
489 391, 1–12. <https://doi.org/10.1016/j.scitotenv.2007.10.046>

490 Nguyen, T.T., Janik, L.J., Raupach, M., 1991. Diffuse reflectance infrared fourier transform
491 (DRIFT) spectroscopy in soil studies. *Soil Res.* 29, 49–67.
492 <https://doi.org/10.1071/sr9910049>

493 Quinlan, J.R., 1992. Learning with continuous classes, in: *5th Australian Joint Conference on*
494 *Artificial Intelligence.* Singapore, pp. 343–348.

495 R core Team, 2017. *R: A Language and Environment for Statistical Computing.* R Foundation
496 for Statistical Computing, Vienna, Austria.

497 Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., O’Connor, P., 2014. *Chemistry of*
498 *Europe’s agricultural soils, part A.*

499 Roudier, P., 2011. *clhs: a R package for conditioned Latin hypercube sampling.*

500 Savitzky, A., Golay, M.J., 1964. Smoothing and differentiation of data by simplified least
501 squares procedures. *Anal. Chem.* 36, 1627–1639.

502 Soriano-Disla, J.M., Janik, L., McLaughlin, M.J., Forrester, S., Kirby, J., Reimann, C., 2013.
503 The use of diffuse reflectance mid-infrared spectroscopy for the prediction of the
504 concentration of chemical elements estimated by X-ray fluorescence in agricultural
505 and grazing European soils. *Appl. Geochem.* 29, 135–143.
506 <https://doi.org/10.1016/j.apgeochem.2012.11.005>

507 Soriano-Disla, J.M., Janik, L.J., Viscarra Rossel, R.A., Macdonald, L.M., McLaughlin, M.J.,
508 2014. The Performance of Visible, Near-, and Mid-Infrared Reflectance Spectroscopy
509 for Prediction of Soil Physical, Chemical, and Biological Properties. *Appl. Spectrosc.*
510 *Rev.* 49, 139–186. <https://doi.org/10.1080/05704928.2013.811081>

511 Struyf, E., Smis, A., Van Damme, S., Garnier, J., Govers, G., Van Wesemael, B., Conley,
512 D.J., Batelaan, O., Frot, E., Clymans, W., 2010. Historical land use change has
513 lowered terrestrial silica mobilization. *Nat. Commun.* 1, 129.

514 Trinh, B.N., Thielen-del Pozo, J., Thirel, G., 2013. The reduction continuous rank probability
515 score for evaluating discharge forecasts from hydrological ensemble prediction
516 systems: Reduction continuous rank probability score for HEPS. *Atmospheric Sci.*
517 *Lett.* 14, 61–65. <https://doi.org/10.1002/asl2.417>

518 Vaculikova, L., Plevova, E., 2005. Identification of clay minerals and micas in sedimentary
519 rocks. *Acta Geodyn. Geomater.* 2, 163.

520 Van Looy, K., Bouma, J., Herbst, M., Koestel, J., Minasny, B., Mishra, U., Montzka, C.,
521 Nemes, A., Pachepsky, Y.A., Padarian, J., Schaap, M.G., Tóth, B., Verhoef, A.,
522 Vanderborght, J., van der Ploeg, M.J., Weihermüller, L., Zacharias, S., Zhang, Y.,
523 Vereecken, H., 2017. Pedotransfer Functions in Earth System Science: Challenges and
524 Perspectives. *Rev. Geophys.* 2017RG000581. <https://doi.org/10.1002/2017RG000581>

- 525 Vandevenne, F., Struyf, E., Clymans, W., Meire, P., 2012. Agricultural silica harvest: have
526 humans created a new loop in the global silica cycle? *Front. Ecol. Environ.* 10, 243–
527 248.
- 528 Viscarra Rossel, R.A., Walvoort, D.J.J., McBratney, A.B., Janik, L.J., Skjemstad, J.O., 2006.
529 Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for
530 simultaneous assessment of various soil properties. *Geoderma* 131, 59–75.
531 <https://doi.org/10.1016/j.geoderma.2005.03.007>
- 532 Wedepohl, K.H., 1995. The composition of the continental crust. *Geochim. Cosmochim. Acta*
533 59, 1217–1232.
- 534 Wösten, J.H.M., Lilly, A., Nemes, A., Le Bas, C., 1999. Development and use of a database
535 of hydraulic properties of European soils. *Geoderma* 90, 169–185.
536 [https://doi.org/10.1016/S0016-7061\(98\)00132-3](https://doi.org/10.1016/S0016-7061(98)00132-3)
537

538 **7- List of figures and tables:**

539 Figures:

- 540 - Figure 1: Location of the sampling sites of the French monitoring network of soil
541 quality (RMQS). Black triangles represent the sites for which Si measurements were
542 available.
- 543 - Figure 2: Schematic diagram of the modelling procedure.
- 544 - Figure 3: Distribution of the Si concentrations in the Si dataset: a- histogram; b-
545 boxplot. The dashed grey line represents the Si overall mean .
- 546 - Figure 4: Percentage of samples in different parent material classes for both the
547 French monitoring network of soil quality (RMQS) database and the Si dataset.
- 548 - Figure 5: Comparison of the empirical density estimates of the pedological properties
549 computed for all the sites (2088) of the French soil monitoring network (RMQS) and
550 those of the Si dataset (673 sample).
- 551 - Figure 6: Principal component analysis of the whole dataset (2088 sampling sites) for
552 two set of variables: Mid-infrared spectra (MIRS) (a and c) and pedological attributes
553 (b and d). Bivariate plot of the scores of PCs 1 and 2 (a and b). Most correlated MIRS
554 band are plotted in a. Score plot of first two components (c and d) (black points
555 correspond to the extrapolated dataset from the French soil monitoring network
556 (RMQS) and grey points to the Si dataset).
- 557 - Figure 7: Predicted versus measured Si concentrations (in g kg^{-1}) for the first cross
558 validation replication of (a) the mid-infrared spectra (MIRS) and (b) the pedological
559 pedotransfer Functions (PTFs). In black, the one to one line and in red, the fitted
560 regression line. Black vertical error bars represent the prediction's uncertainty and
561 blue horizontal error bars represent the analytical uncertainty.

- 562 - Figure 8: (a) Mid-infrared reflectance spectra (MIRS) of two sites randomly selected
563 among those with low Si concentrations (in grey) and high Si concentration (in black)
564 respectively; (b) average importance based on cubist models of spectral region for the
565 prediction of total Si concentration. Most important wavelengths (vertical lines) are
566 identified in Table 4.
- 567 - Figure 9: Average importance based on cubist model of the variables used for the
568 prediction of total Si concentration by pedological attributes.
- 569 - Figure 10: Scatterplot of the total Si concentration (in g kg^{-1}) predicted by the
570 pedological PedoTransfer Function (PTF) *versus* those predicted by the mid-infrared
571 spectra PTF. The samples correspond to the non-Si analysed sites of the French
572 monitoring network of soil quality (RMQS) (1407 sites).
- 573 - Figure 11: Empirical density estimate of Si concentrations obtained for the sites of the
574 French monitoring network of soil quality (RMQS) (this study) to those obtained for
575 French and European sites in GEMAS (Reimann et al., 2014). Vertical lines represent
576 the median values: 327.2 g kg^{-1} , 320.7 g kg^{-1} and 313.9 g kg^{-1} respectively.
577

578 Tables:

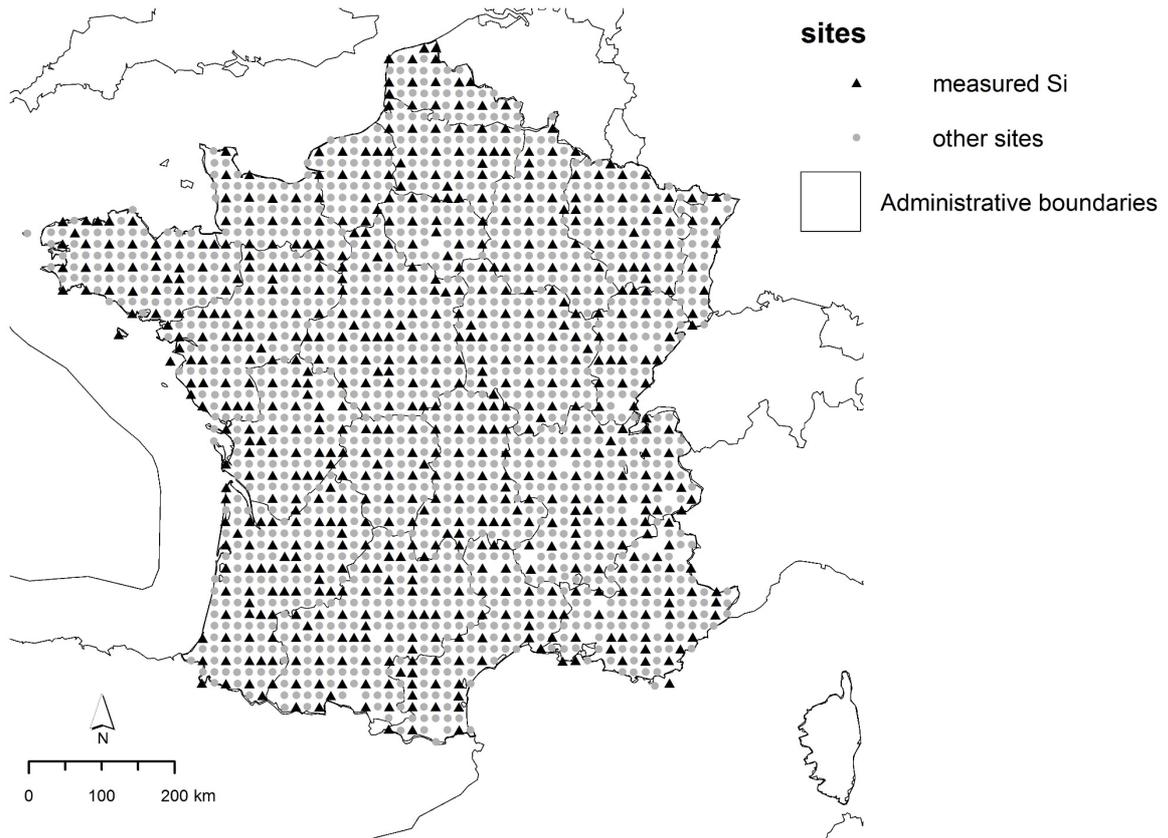
579 - Table 1 : Analytical methods employed for the selected soil properties

580 - Table 2: Summary statistics of pedological attributes for the two datasets: the RMQS
581 and Si dataset.

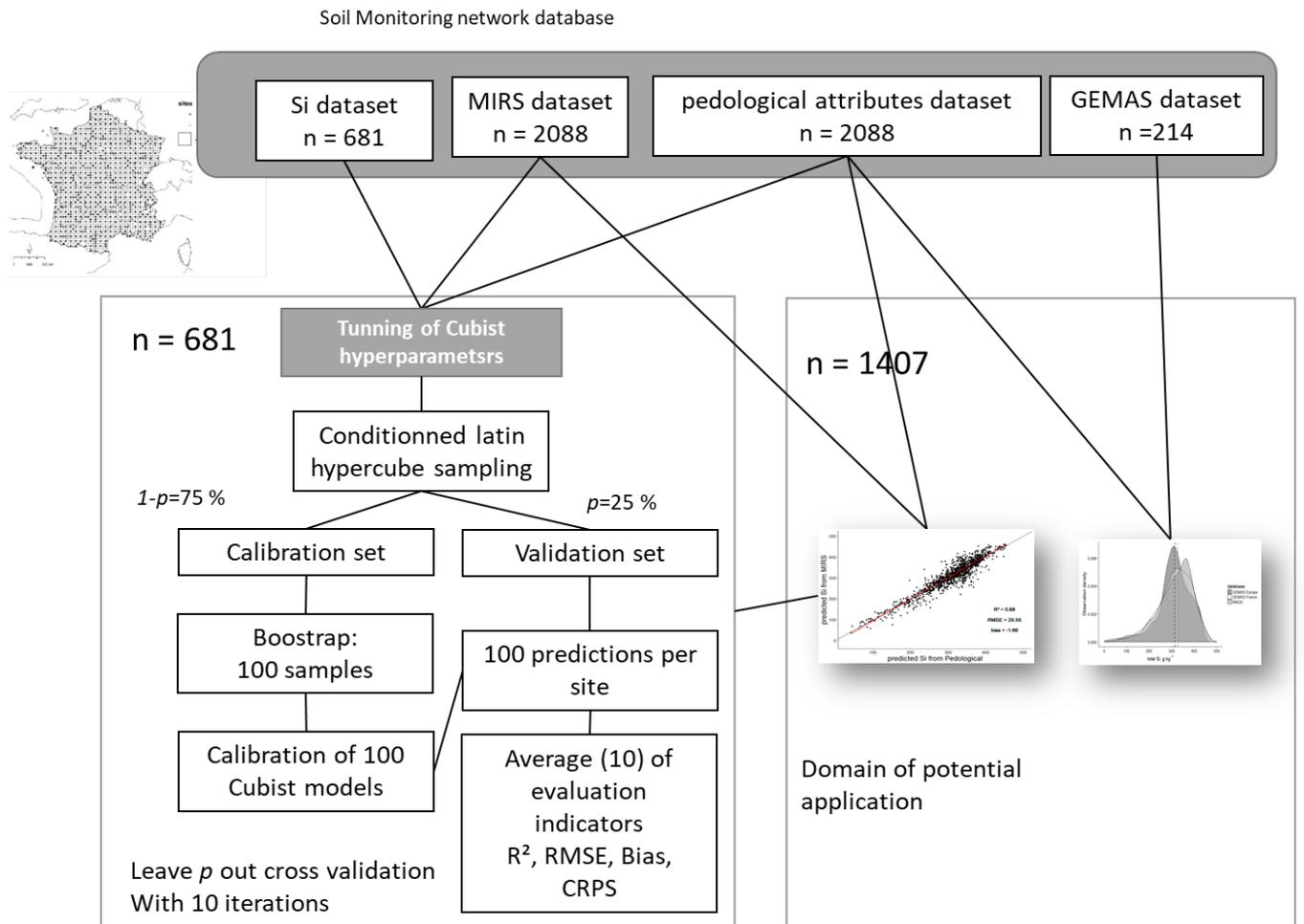
582 - Table 3: Summary statistics of performance indicators of the cross-validation for the
583 two PTF models. RMSE, bias and CRPS are in g kg^{-1} .

584 - Table 4: Important wavelengths in MIRS along with reported peaks and their
585 assignments.

586

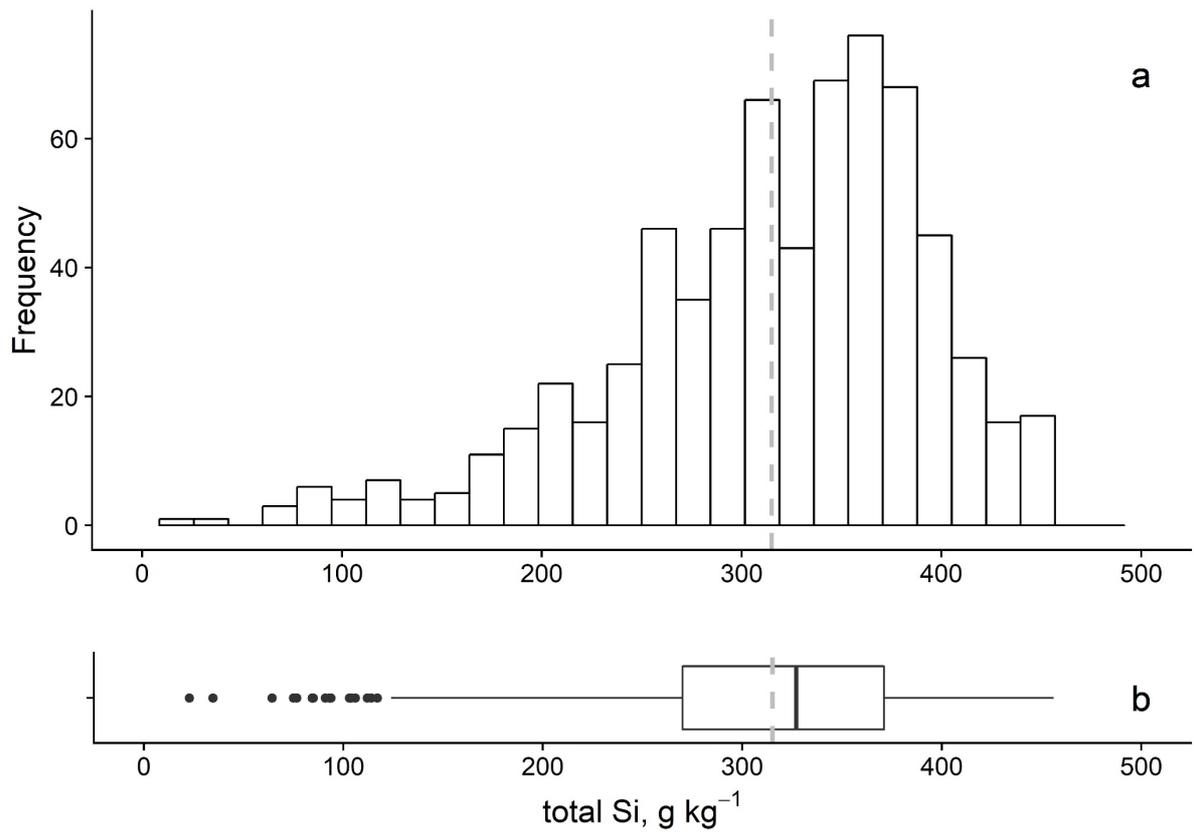


587
588 Figure 1
589

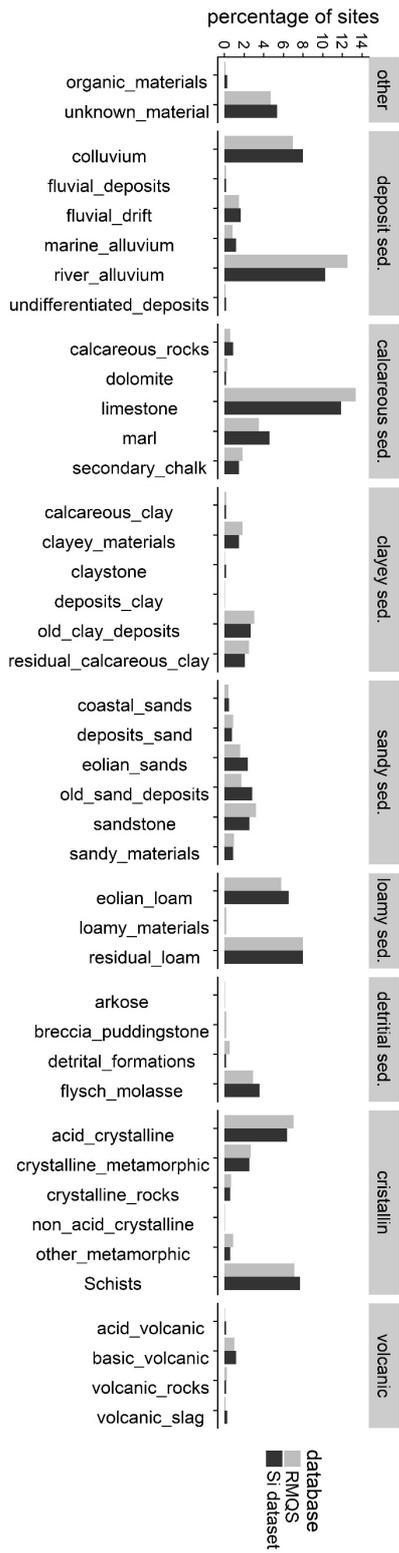


590
591
592

Figure 2

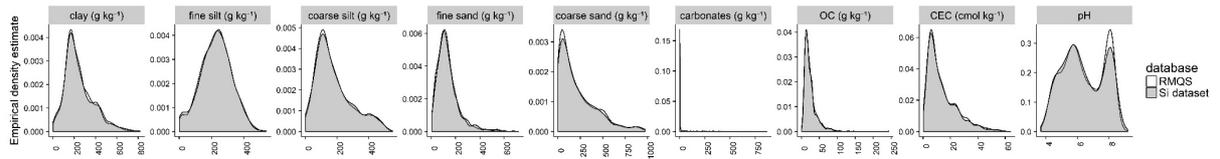


593
594 Figure 3
595

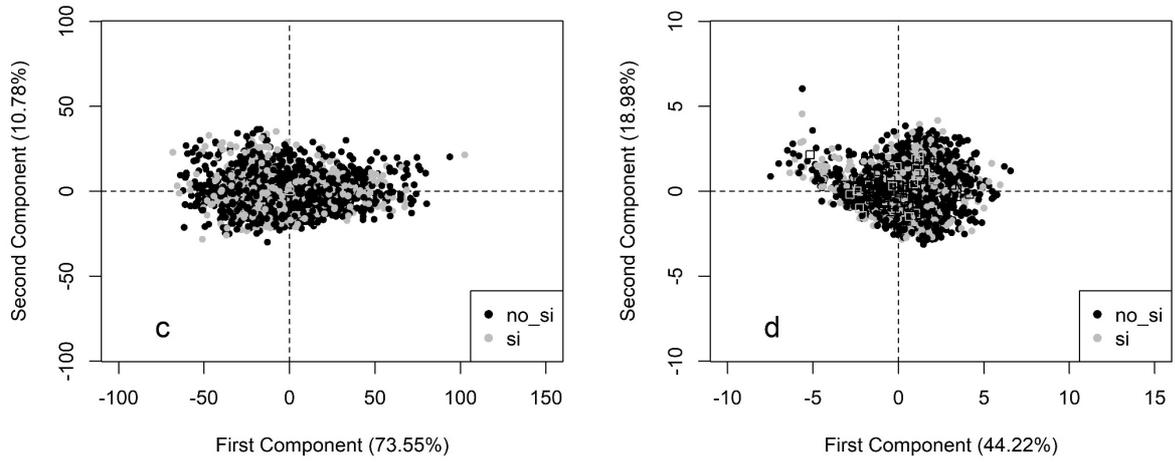


596
597
598

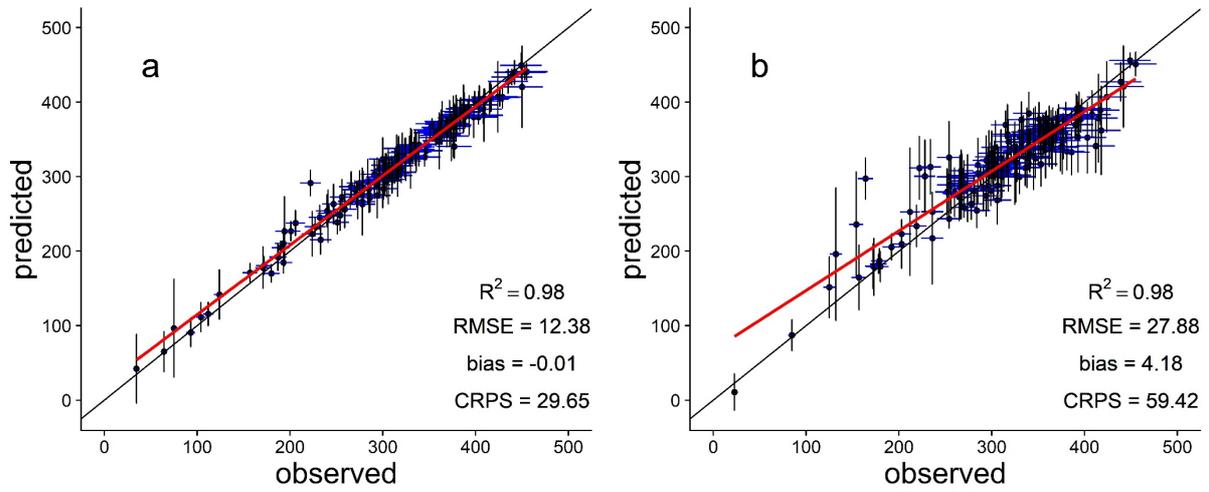
Figure 4



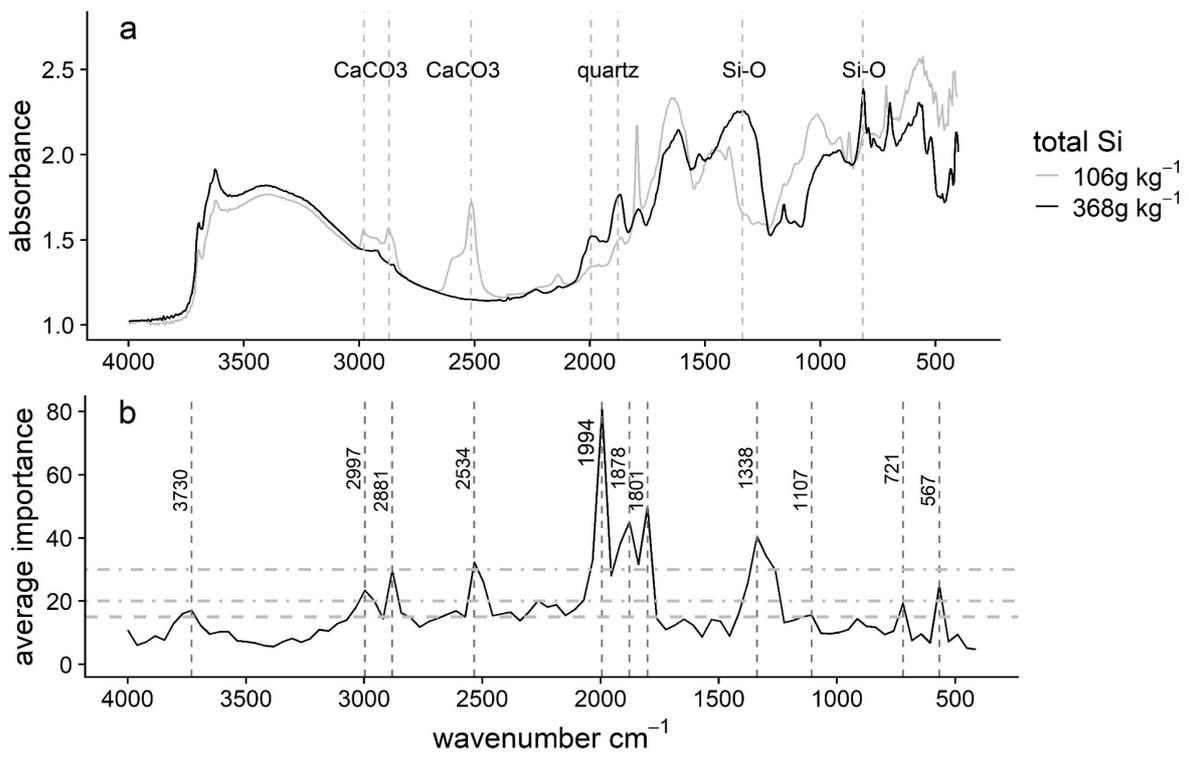
599
600 Figure 5



601
602 Figure 6
603

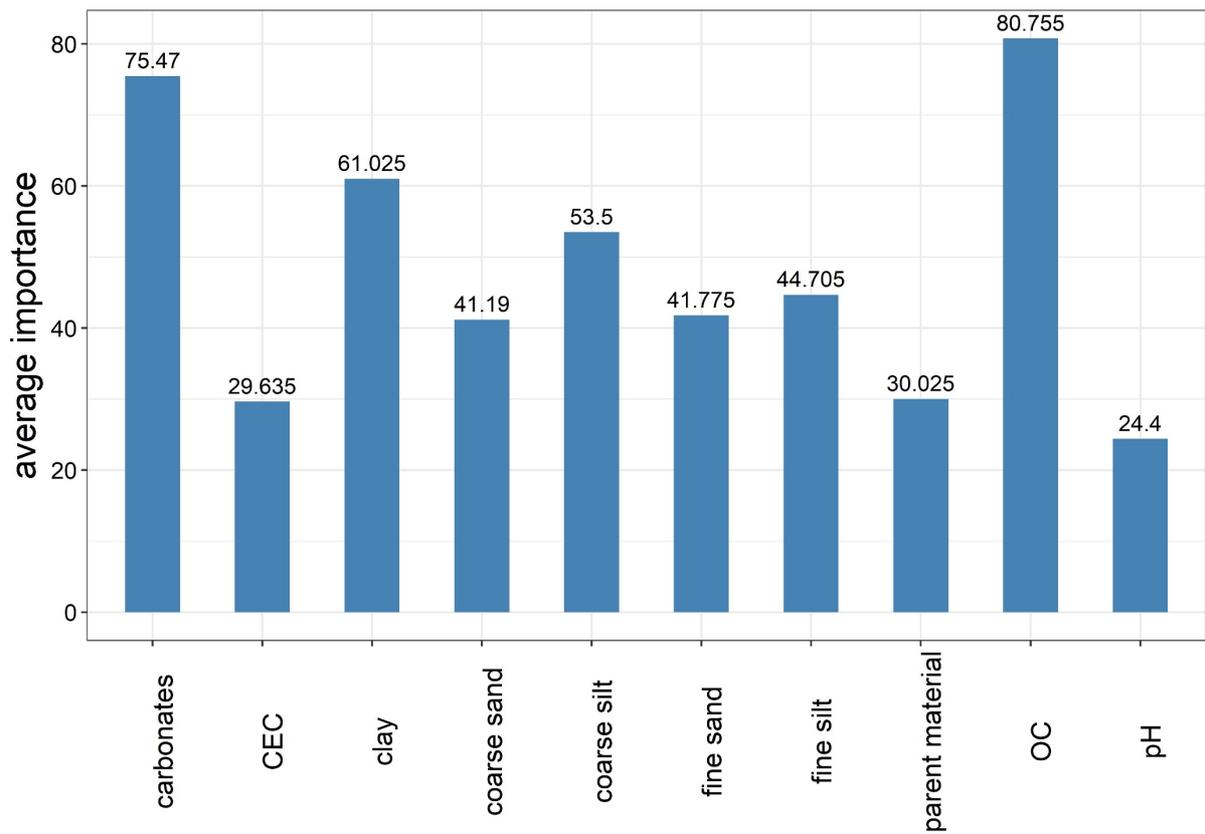


604
605 Figure 7
606



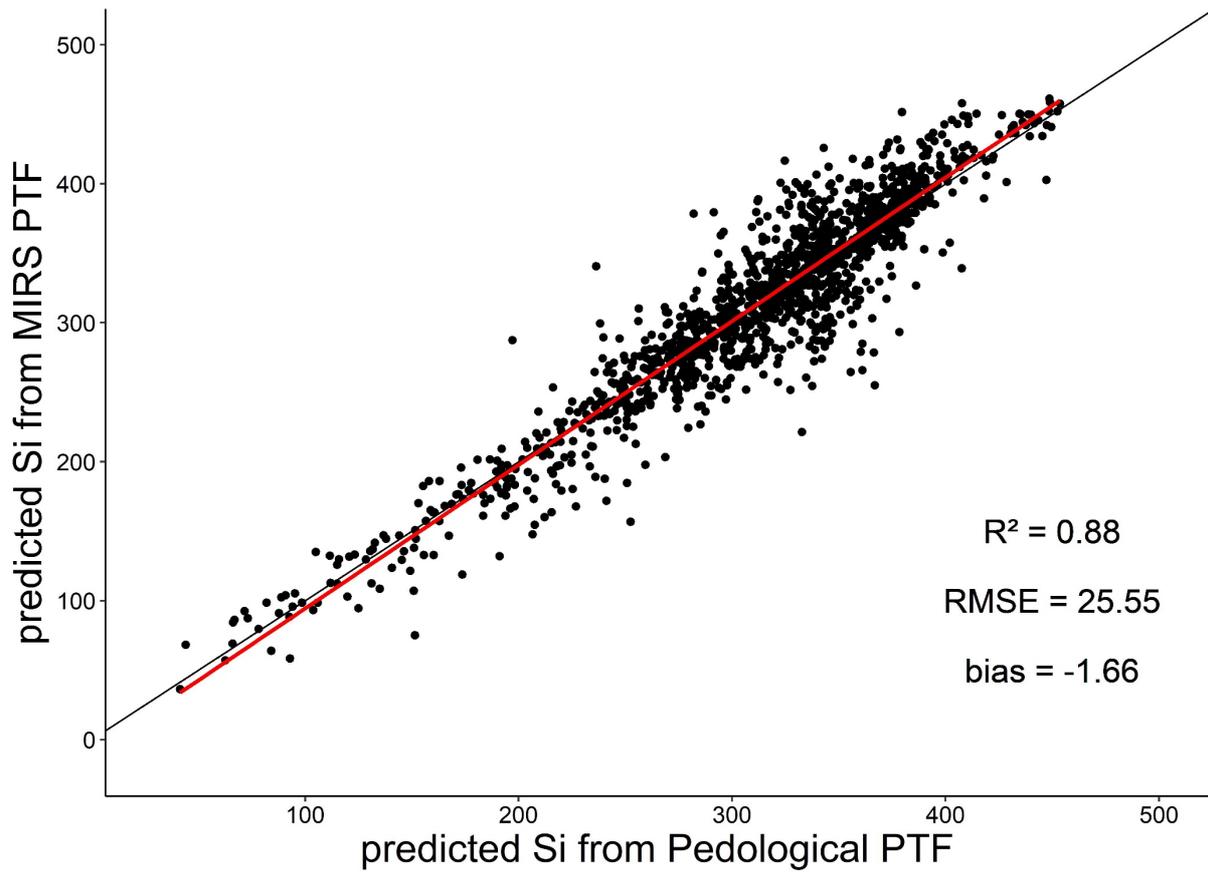
607
608

Figure 8



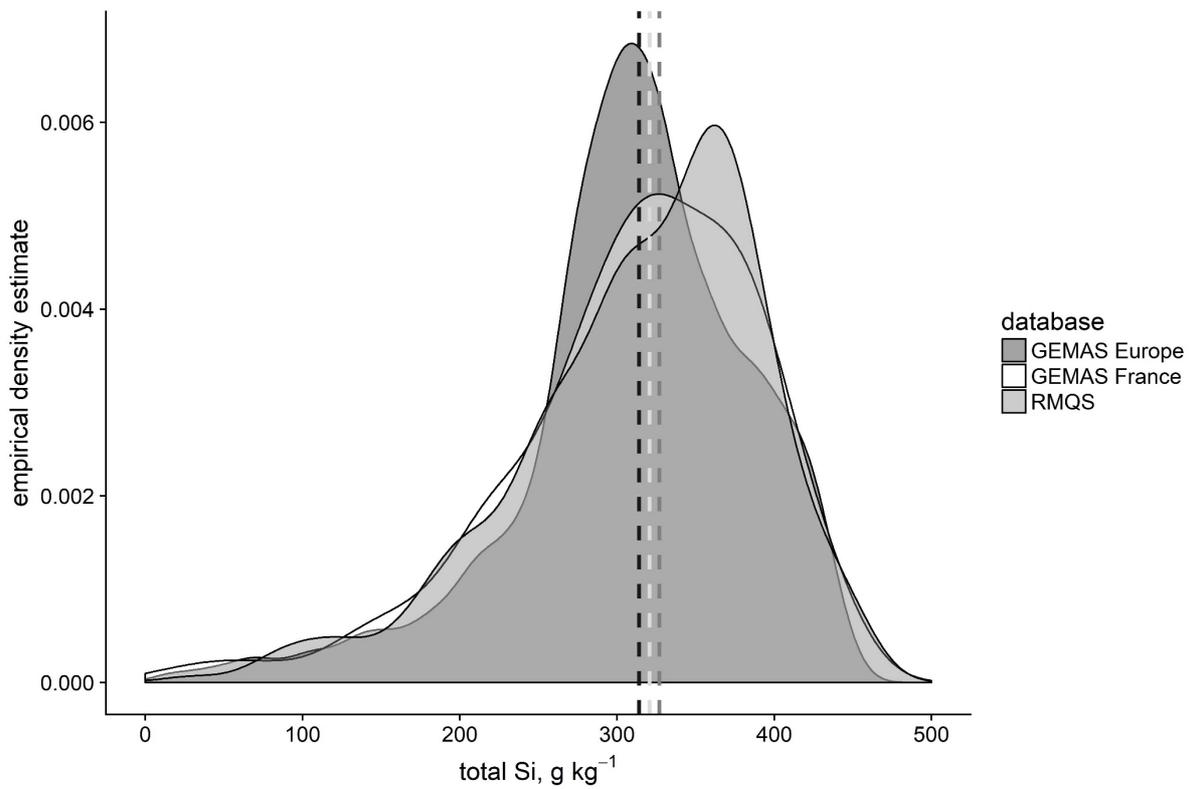
609
610
611

Figure 9



612
 613

Figure 10



614
 615
 616

Figure 11

617
618

Table 4 : Analytical methods employed for the selected soil properties

Soil properties	Method	Source
Clay, fine silt, coarse silt, fine sand, coarse sand	Robinson's pipette after water sieving (NF X 31-107)	RMQS
CEC	Cobalthexammine extraction (NF X 31-130)	RMQS
pH	Water (ISO 10390)	RMQS
Organic carbon	Dry combustion (NF ISO 10694)	RMQS
Carbonates	Volumetric method (NF X 31-106)	RMQS
Total Si	ICP-AES after sodium peroxide fusion	This study
MIR Spectra	Diffusive Reflectance Fourier Transform Spectrophotometer from 4000 to 400 cm^{-1}	(Grinand et al., 2012)

619
620

621

Table 5: Summary statistics of pedological attributes for the two datasets: the RMQS and Si dataset.

Properties	Units	RMQS, n=2088		Si dataset, n=673		Wilcox-test p-value
		range	Median	range	Median	
Clay	g kg ⁻¹	2 - 819	211	5 - 819	212	0.65
Fine silt	g kg ⁻¹	1 - 539	231	1 - 539	231	0.99
Coarse silt	g kg ⁻¹	1 - 551	146	1 - 518	144	0.81
Fine sand	g kg ⁻¹	3 - 722	115	4 - 677	118	0.43
Coarse sand	g kg ⁻¹	1 - 970	140	2 - 966	137	0.83
CEC	cmol kg ⁻¹	0.3 - 64	10	0.5 - 60	10	0.36
pH	-	3.7 - 9.2	6.2	3.8 - 8.9	6.2	0.61
OC	g kg ⁻¹	0.6 - 243	19.3	2.6 - 243	18.8	0.36
carbonates	g kg ⁻¹	0.5 - 866	0.5	0.5 - 866	0.5	0.57

623

624

625
626

Table 6: Summary statistics of performance indicators of the cross-validation for the two PTF models. RMSE, bias and CRPS are in g kg⁻¹.

Model	statistics	R ²	RMSE	Bias	CRPS
pedological	Min.	0.85	23.21	0.43	53.47
	Max.	0.90	28.57	5.63	64.88
	1st Qu.	0.86	25.50	1.19	55.62
	3rd Qu.	0.87	27.39	3.66	60.07
	Median	0.86	26.95	1.94	57.68
	Mean	0.87	26.48	2.37	58.04
	Var.	0.2 10 ⁻³	2.47	3.12	13.16
	sd.	0.01	1.57	1.77	3.63
MIRS	Min.	0.94	11.41	-0.12	17.43
	Max.	0.98	17.67	3.71	29.65
	1st Qu.	0.95	14.21	0.20	19.74
	3rd Qu.	0.96	17.07	2.06	24.59
	Median	0.95	16.04	0.64	21.25
	Mean	0.96	15.31	1.15	22.36
	Var.	0.2 10 ⁻³	4.70	1.57	14.32
	sd.	0.01	2.17	1.25	3.78

627
628
629

630 Table 7: Important wavelengths in MIRS along with reported peaks and their assignments.

631

Reported peaks (wavenumber in cm^{-1})	MIRS region from reference (wavenumber in cm^{-1})	Soil constituent and Assignment	Reference
567	600-150	clay : Si-O, Al-O bending	Vaculíková and Plevová, 2005
721	727-713	carbonates : CO_3	Vaculíková and Plevová, 2005; Nguyen et al., 1991
1107	1120-1000	clay : O-Si-O stretching	Saikia and Parthasarathy, 2010; Madejová, 2003
1300-1338	1360-1347	clay : Al-O as Si cage (TO_4)	Saikia and Parthasarathy, 2010
1801 1878 1994	2000-1650	quartz : O-Si-O stretching	Bertrand et al., 2002; Janik et al., 1998; Nguyen et al., 1991
2534	2600-2500	carbonates : CaCO_3 overtone and combinaison vibrations	D'Acqui et al., 2010; Du and Zhou, 2009; Vaculíková and Plevová, 2005; Bertrand et al., 2002; McCarty et al., 2002; Nguyen et al., 1991
2881	2880	carbonates : CaCO_3 overtone and combinaison vibrations	Vaculíková and Plevová, 2005; Bertrand et al., 2002; McCarty et al., 2002; Nguyen et al., 1991
2997	3000-2900	carbonates : CaCO_3 overtone and combinaison vibrations	Vaculíková and Plevová, 2005; McCarty et al., 2002; Nguyen et al., 1991
3730	3697 (1), 3750-3400 (2)	clay : Al---O-H stretching	Saikia and Parthasarathy, 2010 (1); Vaculíková and Plevová, 2005 (2); Nguyen et al., 1991

632

633