Do climate and land use affect the pool of total silicon concentration? A digital soil mapping approach of French topsoils

Amélia Landré, Sophie S. Cornu, J.-D Meunier, Annie Guérin, Dominique Arrouays, Manon Caubet, Céline Ratié, Nicolas P A Saby

DO CLIMATE AND LAND USE AFFECT THE POOL OF TOTAL SILICON CONCENTRATION? A DIGITAL SOIL MAPPING APPROACH OF FRENCH TOPSOILS.

Landré, A.[a,b], Cornu, S.[c], Meunier, J.-D.[c] ,Guerin A. [d], Arrouays D. [a], Caubet M. [a], Ratié C.[a], Saby, N.P.A.[a]

a) INRA, Infosol, US 1106, Orléans, France
b) INRA - INPT-ENSAT - INPT-EI-Purpan, UMR 1248 AGIR AGroécologie, Innovations, teRritoires. Centre de recherche Occitanie-Toulouse, Auzeville, France.
c) Aix-Marseille Univ, CNRS, IRD, Coll de France, INRA, CEREGE, Aix-en-Provence, France
d) INRA, Laboratoire d'Analyses des Sols US, Arras, France

Abstract

Silicon (Si) is the second most abundant element in the Earth's crust after O. Its concentration in soils is highly variable from less than 1 % to greater than 45 %. Parent material is well known to be a major parameter for explaining this variability. In this study, we proposed to analyze the impact of climate and land use on the total Si concentration in soils and to explore the link between total Si and plant available Si (PAS). To do so, we based our analysis on the French soil monitoring network considering the upper soil horizon that was thought to be the most impacted by both the effect of land use and climate and was also the most important horizon in terms of plant availability. In order to extract the impact of climate and land use and for digital mapping purposes, we stratified the database by parent material and soil-types. This stratification was based on the classification used in the 1:100 000 French soil map and 1:100 000 French soil parent material map. For non carbonated soils, we showed that Si concentrations was decreasing with annual rainfall, evidencing a climatic effect on the total Si concentration of French topsoils. No significant effect of the land used could be identified. At last, we showed that PAS (by the $CaCl_2$ method) is negatively weakly correlated to total Si concentration. This relationship is however variable among soil classes.

Keywords

Silicon, silica, RMQS, Digital Soil Mapping, France.

1 Introduction

Silicon is the second most abundant element in the Earth's crust after O, and it has an average mass concentration of 28 % (Wedepohl, 1995). The SiO_4 tetrahedron is the elemental brick that constitutes the basic structure of Si in nature from solid (silicates) to soluble (silicic acid) states. In soils, the variability of Si is large, from less than 1 wt% in Histosols to greater than 45 wt% in some Podzols (Sommer et al., 2006). Si is included in a large number of minerals issued either from the parent material as for instance quartz, feldspar, and phyllosilicates (muscovite, biotite for instance), or from the transformation of the parent material minerals, along weathering and pedogenesis, in clay minerals and amorphous silica particles, notably phytoliths for the part of Si recycled by vegetation (Sommer et al., 2006).

The soil concentration of total Si is intimately link to the concentration of Si from parent rocks (Berner and Berner, 1996; Gray et al., 2016). Gray et al. (2016 showed that sand dunes are shown to be the richest parent material (close to 100 %) while limestones are the poorest (a few percent). Among the parent material rich in Si (> 60 %), quartz and arkose sandstones, granite, rhyotite, greywacke, granolite, dacite, shale are encountered, while diorite, andesite, basalt and peridotite have Si content ranging from 40 to 60 % and laterite and bauxite from 10 to 15 % (Gray et al., 2016 and references included). In addition, weathering and pedogenesis modify the initial Si pool both in concentration and status (Lucas et al., 1993; White et al., 2012). The upper soil horizon is the most impacted by these processes being both depleted by weathering and clay translocation, that may have contradictory effect on the total Si concentration of the horizon. Weathering and pedogenesis are influenced by both climate and vegetation as demonstrated by river dissolved Si (DSi) (Bluth and Kump, 1994; Dürr et al., 2009) at the territory scale. Using DSi in the streams as an indicator of total chemical

63 weathering is however not straightforward because of the fraction of Si that is retained
64 in the soil for forming clay minerals or the one that is retained in biogenic silica in soils
65 and streams (Frings et al., 2015). Analysis of climosequences have proven to be
66 helpful for documenting some modification of the soil Si pools. On a volcanic
67 climosequence, secondary Si phases (produced by weathering) increase under a more
68 humid climate (Taboada et al., 2019). On intrusive rock climosequence, smectite
69 distribution is correlated to Si loss when rainfall (or precipitation) increases (Egli et al.,
70 2003), but the analysis of the variation of soil surface Si is not well addressed. As
71 shown by Dere et al. (2016), the presence of poorly reactive quartz in the parent rock
72 may be greater at the surface resulting from a relative enrichment during weathering.
73 Therefore, the impact of rainfall on the soil Si deserves to be more fully documented.

74 While vegetation on the weathering scale is strongly correlated to climate, human
75 activity has strongly modified the land use and thus the biological cycle of most of the
76 elements. Some of the main cultivated crop are Si accumulators and long-term
77 exportation may act on Si budget. It seems well documented that conversion of forest
78 to cropland lead to a decrease of the phytolith pools (Struyf et al., 2010; Guntzer et al.,
79 2012; Vandevenne et al. (2015). Struyf et al. (2010) suggest that the transformation of
80 the european temperate forests into cultivated land has also lead to a decrease of
81 weathering of silicate minerals. However, Yang and Zhang (2018) came to the opposite
82 conclusion (agricultural activities increase silicate weathering) based on a geochemical
83 analysis of rain and streams in subtropical China. In the subtropical southern Brazil,
84 Ameijeiras-Marino et al. (2018) share Yang and Zhang (2018)’s conclusion using the
85 Ge/Si tracer. Accordingly, the question of the intensity of the impact of land use on soil
86 Si loss remains open.

87 The study of the status of Si in soils and its biogeochemical cycle is a subject of
88 increasing interest because of the growing body of evidence showing that Si can
89 improve crop development (Coskun et al., 2019; Guntzer et al., 2012; Liang et al.,
90 2015; Rodrigues and Datnoff, 2015). Plant available Si (PAS) depends on the reserve
91 of weatherable Si-bearing minerals (Cornelis and Delvaux, 2016) and can be estimated
92 using various protocols (Narayanaswany and Prakash, 2010; Meunier et al; 2017). The
93 relationship between soil total Si and PAS is complex depending on soil type, parent
94 material and degree of weathering. On various agricultural soils of Asia, a negative
95 significant correlation between PAS and soil Si content was found (Yanai et al., 2016;
96 Meunier et al. (2017). This correlation was due to the presence of low solubility Si
97 minerals such as quartz for explaining low PAS values. On volcanic soils containing
98 no quartz, Henriet et al. (2008) founded a highly significant positive correlation (n= 6)
99 between soil Si content and Si extracted by the 0.01M $CaCl_2$ method due to a decrease
100 of the reserve of weathered Si minerals along weathering.

101 The objective of this study was to (i) determine the impact of climate and land-use on
102 the total topsoil Si concentrations (Si_{tot}) in French soils and (ii) explore the relationship
103 between Si_{tot} and PAS. Topsoil horizon was chosen as it was considered as the most
104 weathered and pedogenized horizon where the roots are mainly concentrated. The
105 French territory was chosen as (i) France is one of the countries with the largest soil
106 diversity in the world (Minasny et al., 2010), that offers notably a good
107 representativeness of European soils notably; (ii) France processes a soil monitoring
108 network (RMQS) for which Si_{tot} is available along with all the typical soil characteristics
109 (Landré et al., 2018). Indeed, soil Si is rarely analyzed at the territory scale, with the
110 notable exception of the soil Si map at the European scale with only one site every
111 2500 km² (De Vos et al., 2006; Reimann et al., 2014) and the French soil monitoring

network (RMQS) with one point at each node of 16 to 16 km grid (Jolivet et al., 2006; Landré et al., 2018).

Since geology and soil type (pedologenesis) are the main Si drivers, we first stratified the database in geological-pedological classes. Then we considered the impact of climate and land-use in both the obtained classes and at the national scale with a modern digital soil mapping (DSM) approach using the geological-pedological classes defined as a covariate. The strong input of this type of covariate in the DSM model was clearly demonstrated by Gray et al., (2016).

2 Materials and Methods

2.1 The data

Most of the data used in this study were obtained from 2111 sites from the first campaign of the RMQS sampled between 2001 and 2009 (Jolivet et al., 2006), which covered all the mainland of France (around 550 000 km^2) based on a 16 km $\times$ 16 km grid. Among these sites, two had no soil type description, 98 had no parent material description, six were anthroposols and three histosols. All these sites were not further considered for our analysis, ending with database of 2004 sites.

Composite samples of the top horizon (0-30 cm) were sampled, air dried, and sieved to 2 mm before laboratory analysis at the Soil Analysis Laboratory of INRA in Arras, France. The following parameters were measured: (i) the total organic C (OC) content measured by dry combustion (NF ISO 10694), (ii) particle size distribution by wet sieving and the pipette method (NF X 31-107), (iii) cation exchange capacity and exchangeable cations (hexamminecobalt method NF ISO 23470), (iv) pH in water (1 to 5 soil to water ratio; NF ISO 10390), (v) calcium carbonate using the volumetric

method (NF ISO 10693) ($CaCO_3$), (vi) and total P, K, Ca, Mg, Fe, and Al determined by ICP-AES after dissolution with hydrofluoric and perchloric acids (NF X 31-147).

Landré et al. (2018) analyzed a subset of 673 samples for Si_{tot}. For the sites not analyzed, we estimated Si_{tot} according to the following conceptual equation:

$$\mathrm{Si} = \mathrm{f}(\mathrm{Al, Fe, K, Na, Ca_{nc}, Mg_{nc}, P, Mn, OC, CaCO_3, residual\ water}) \quad (1)$$

where Ca_{nc} and Mg_{nc} are the fractions of Ca and Mg that are not included in carbonate minerals nor adsorbed on the exchangeable surfaces, and OC is the organic matter percentage. The total Si concentration in French topsoil was modelled using a Cubist modelling algorithm (see Supplementary File 1 for more details).

The database was completed by PAS measurement (Si_{CaCl2}) using the 0.01 M $CaCl_2$ method (Haysom and Chapman 1975) on 1986 sites. This widely used method (Meunier et al., 2017) allows estimating the pool of Si that is readily soluble. Briefly, 2g of dry soil was mixed with 20 ml of the solution and shaken during 16h in polyethylene tubes. The solutions were then filtered at 0.45 µm, and Si concentration was measured using Inductively Coupled Plasma Atomic Emission Spectroscopy (axial ICP-AES; 720 ES, Varian).

2.2 Stratification of the database in homogeneous geo-pedological classes

Since Si_{tot} in soil is known to be linked to parent material and being anti-correlated to carbonate concentration in soils (Landré et al., 2018), we first stratified the database by the type of parent material and by the content in carbonates. In French soils, carbonates were encountered in soils developed on sedimentary rocks with a limit for carbonated soils classically considered at a $CaCO_3$ concentration of 5 % (Baize, 2000).

Since a large part of the sedimentary rocks were defined on the basis of geomorphological processes (e.g. alluvions, colluvions, terraces…), we were not able for DSM purposes to relate directly parent material to carbonate content. In addition, since we were working on the topsoil horizon, depending on the stage of pedogenesis, soils developed on carbonated rocks may have been completely decarbonized (e.g. Calcisols notably). Other geological domains were sorted in igneous extrusive or intrusive and metamorphic. The classification used did not allow further classifying the parent materials in more geochemical meaning classes as made by Gray et al. (2016), since (i) a large part of the sites had no more information than igneous extrusive or intrusive or metamorphic rocks; (ii) acidic extrusive igneous rocks or basic intrusive igneous rocks were too poorly represented in the database (no more than two to three individuals). Podzols were also separated, these soils being mainly composed of quartz were supposed to have high Si_{tot}. We then sorted the soils in the different geological groups a step further with the exception of the soils on igneous intrusive rocks that were poorly represented. Soils on metamorphic parent materials and on igneous extrusive rocks were sorted by soil types, ranging from poorly differentiated soils (regosols, lithosols and rankers) to strongly differentiated soils (planosols, luvisols and podzoluvisols) for soils on metamorphic rocks. Podzols were sorted between those developed on sedimentary rock and those developed on other parent materials. For non-carbonated soils on sediment, sorting was performed both by soil types and a more detailed description of the considered sediment.

2.3 Data treatment

2.2.1. Mapping total topsoil Si concentrations in French soils

The DSM approach was based on the work of McBratney et al. (2003), who proposed a quantitative relationship between soil properties and the soil forming factors plus a spatially correlated residual element (*e*), as follows:

$$Soil = f(s,c,o,r,p,a,n) + e \quad (2)$$

where $Soil$ is a soil property. The s refers to other or previously measured properties of the soil at a point either from prior soil maps or from remote or proximal sensing data; c refers to the climatic properties of the environment at a point; o refers to organisms, including vegetation or fauna, or human activity; r refers to relief; p refers to the parent material or lithology; a refers to the soil age; and n refers to the space or spatial position. Finally, e is the locally varying, spatially dependent residuals from f.

We selected a set of available environmental covariates describing the *scorpan* factors for the whole French territory. We harmonized them at 90 m resolution (Table 1). The parental material and soil covariates were generated by the results of the PCA step. We also used a very popular spectral index, namely the normalized difference vegetation index (NDVI) (Huete et al., 2002). We focused on the yearly changes in the NDVI computed from a time series of remote sensing data to describe the photosynthetic capacity of the vegetation cover. This variable was interpreted as a vegetation growth dynamics proxy. The underlying assumption was that this information on changes in vegetation may reflect various behaviors linked to climate, land management or soil properties (see Loiseau et al., 2019).

The spatial predictive model between Si_{tot} and scorpan covariates was first constructed using the ensemble learning method Random Forests (RF) (Breiman, 2001). We used the RF implementation provided by the package randomForest in R (Liaw and Wiener,

2002; R. Core Team, 2016). Three parameters should be defined in the RF model, namely the number of trees to grow (n_{tree}), the number of variables randomly sampled as candidates at each split (m_{try}), and the minimum size of the terminal nodes (nodesize; Liaw et al., 2002). The default values were used for n_{tree} and nodesize, which were 500 and 5, respectively. The optimal value of m_{try} was set at 2 by the lowest out-of-bag error estimate.

To extract useful information from this large set (22) of potentially correlated covariates, we ran a preliminary step of variable selection using the Boruta algorithm (Kursa and Rudnicki, 2010). This algorithm was a wrapper built around the RF classification algorithm implemented in the R package randomForest, and it used a Z score computed by dividing the average loss by its standard deviation. The algorithm was used in feature selection (Kursa and Rudnicki, 2010), but was also applied to support the model establishment of the RF regression (Kursa, 2014).

The residuals of the model in equation 2) were computed as the difference between the RF predictions and the measured values at the observed location and then interpolated by ordinary kriging (Matheron, 1971). The R package gstat (Pebesma and Graeler, 2019) was used to select variogram models and perform the kriging procedure. The final predictions summed the RF predictions and the Kriging outputs (Keskin and Grunwald, 2018). The model performance was evaluated by 30-folds cross-validation.

To rank the influences of the final list of scorpan factors on Si_{tot}, we calculated covariate importance from the RF algorithm as the mean increase in accuracy (%IncMSE). This indicator was constructed by permuting the values of each variable of the validation set, recording the prediction error, and comparing the set with the un-permuted

validation set prediction of the variable (normalized by the standard error). It calculated the average increase in the squared residuals of the validation set when the variable was permuted. A higher %IncMSE value represented higher variable importance.

2.2.2. Other statistical analysis

Since Gray et al. (2016) showed that pedological parameters and Si_{tot} were highly dependent from parent material composition, the database was stratified in homogeneous subgroup defined on a combination of geological and pedological classes, in order to use the obtained classes for DSM approach. The parent material and soil type classifications used for the stratification were those used at the French territory scale that is the EUSIS classification (King et al, 1994) for the parent material and the FAO (1985) classification for the soil types.

We then performed multiple comparison tests: multiple Mann-Whitney tests, Kruskal-Wallis tests, and a post-hoc test from Siegel and Castellan (1988). These tests were performed using the medians of Si_{tot} within the geo-pedological groups identified in order to discuss the discriminating power of the obtained classification in terms of total Si topsoil concentrations.

In order to understand the meaning of the relations between total Si_{tot} and Si_{CaCl2} obtained for the different parent material domains, we ran a principal component analysis (PCA) on all the soil characteristics available in the RMQS database, with total Si was considered a passive variable.

3 Results and Discussion

3.1 Stratification of the database in homogeneous geo-pedological classes

For soils developed on sedimentary rocks, we observed as expected a strong negative correlation between Si_{tot} and carbonate contents for soils with carbonates concentration higher than 5 % (Figure 1a). This negative correlation was interpreted as a diluting effect of silicate minerals by carbonates. We then sorted geo-pedological classes for soils developed on sediment in carbonated soils developed on sediments and non-carbonated soils developed on sediment (Figure 1b) in order to use this classification in the DSM approach. Based on this separation and on other parent material classes (igneous intrusive and extrusive rocks, metamorphic rock), a boxplot analysis confirmed that this first separation was meaningful (Figure 2) with the highest concentrations in Si_{tot} for Podzol topsoils and the lowest for topsoils on igneous extrusive rocks. This low concentration obtained for igneous extrusive rocks may seem surprising as Gray et al. (2016) showed that both Si rich (rhyotite) and Si poor (basalt, andesite) rocks could be found in that group. However, in the case of France, igneous extrusive rocks that are poorly abundant (only 28 sites over 2000) consist mainly in basalt, slag, tuff and other basic volcanic rocks while rhyolite and other acid volcanic rocks are rare (only two sites over the 28 classified as igneous extrusive volcanic rocks).

At last, some parent material classes while significantly different from others, still exhibit a very large Si concentration variability, justifying the further classification performed on parent material and soil type criteria (Fig. 3). In this figure, we observe that Podzol topsoils on sedimentary parent materials exhibited significantly higher Si_{tot} (*p*-value = 8.573e-06; Fig. 3). These Podzols are located in two main areas in France,

273 namely the Landes of Gascony (in southwest France) and an area in central France
274 north of the Massif Central mountains (Figure 4). The Landes of Gascony was entirely
275 covered by aeolian sands (their texture is usually more than 95 % sand; Augusto et
276 al., 2010) that were nearly pure quartz grains (300-550 µm in size) and therefore could
277 be considered as a reference for nearly pure Si topsoils. The other area in central
278 France corresponded to an ancient delta of the Loire River, with very sandy material
279 also quartz rich originating from the erosion of the ancient Massif Central and its
280 northern foothills. Our results are in a good agreement with those of Gray et al. (2016)
281 showing that dune sands were among the richest parent material in Si.

282 On igneous extrusive rock, andosol topsoils exhibited significantly lower Si_{tot} than
283 those of other soil types (*p*-value < 0.001; Figure 3). For topsoils on metamorphic rock,
284 poorly differentiated soils (Lithosols, Regosols, and Rankers) differed from well
285 differentiated soils (Planosols, Luvisols, and Podzoluvisols), with Cambisols and
286 Fluvisols being intermediate (Figure 3). These three groups of topsoils significantly
287 differed in their Si_{tot} (*p*-value < 0.001 and post-hoc test with a *p*-value < 0.05). Soils on
288 sedimentary rocks were the more abundant, and their situation was more complex.
289 Nineteen and seventeen groups were identified either on the basis of both the parent
290 material and the soil type for the carbonated and non-carbonated soils, respectively.
291 These groups differed in their topsoil Si_{tot} (*p*-value < 0.001 for the carbonated soils and
292 *p*-value < 0.001 for the non-carbonated soils; Figure 3). As an example, very low
293 concentrations of topsoil Si_{tot} were found for poorly differentiated soils on chalk (Fig 3)
294 i.e. nearly pure $CaCO_3$ with Si_{tot} close to zero. These soils are encountered mainly in
295 two areas in France: in the Champagne area and in the Charentes (in western France)
296 where soils are locally called Champagne's soils (Figure 4).

297 Interestingly, this analysis showed that some carbonated soils could exhibit very high
298 Si_{tot} in their topsoils, despite the relative absence of Si in carbonates (e.g. calcic
299 cambisols on other sediments and solonchack; Figure 3). This classification was used
300 in the DSM approach as a covariate along with climate, vegetation using land use, and
301 relief variables.

302 3.2 Impact of environmental factors (climate and land
303 use) on the total Si concentrations of the French top
304 soils

305 The covariate importance from the RF algorithm are presented in Figure 5. Two groups
306 of covariates stood out with a threshold of around 20 %. The first group contained six
307 environmental covariables. The first covariate, called PG, corresponded as expected
308 to the pedogeological classes previously defined. This covariable made a significant
309 contribution in the model construction, as it exhibited a mean increase in accuracy of
310 55 %. Next was the elevation (“srtm”) with an importance of 33 %, followed by the NDVI
311 covariate (“PC1_NDVI”) and the type of climate (“typo”) with an importance of 27 %
312 and 24 %, respectively. Finally, the annual mean evapotranspiration (“etp_mean”) and
313 the net primary production (“NPP_max”) presented an importance of 22 % and 21 %,
314 respectively. NDVI covariate is not an independent variable. Indeed, as demonstrated
315 by Loiseau et al. (2019), beside land use, NDVI is partially linked to pedology and
316 climate. Elevation is also not an independent variable in the case of France. For
317 instance, in the case of the Massif Central, igneous extrusive rocks only located at the
318 highest altitude Elevation can be a proxy for both climate and geology in France. Thus,
319 after parent material and soil type, climate seemed to be an important factor in
320 determining the total topsoil Si concentration.

321 Figure 6 presents Si_{tot} as a function of the annual rainfall for the different parent
322 material domains. For carbonated soils on sediments, Si_{tot} was independent from
323 annual rainfall as well as non-carbonated soils developed on sedimentary parent
324 materials. A significant negative correlation was obtained for soils on other parent
325 materials, with higher Si_{tot} for lower annual rainfall. This correlation was highly
326 significant with the exception of the soils developed on igneous extrusive rocks. The
327 latter being represented by fewer individuals (n=28) than the other groups, a
328 correlation significant only at a 5 % confidence level was acceptable. Such a negative
329 correlation between the concentration in Si_{tot} and the average annual rainfall was
330 interpreted as resulting from the increasing losses of Si with an increase in weathering
331 intensity owing to the larger water flow through the soil (higher rainfall). Therefore, the
332 analysis of surface soils may be used as an alternative to the river chemistry (Bluth
333 and Kump (1994) to document the effect of climate on Si weathering.

334 The impact of land use on Si_{tot} in French topsoils was tested on the geo-pedological
335 classes defined on Figure 3. To do so, for each class sites were sorted in two land use
336 groups, namely permanent vegetation (forests and pasture) and arable land. We then
337 performed a two-way ANOVA with an interaction term between land use and geology
338 followed by a tukey HSD test. The results showed that the interaction effect was slightly
339 significant ($p = 0.012$). We investigated also the model coefficients and we found that
340 for most of them, the 95 % confidence intervals overlapped except for a few cases.
341 (see *p*-values reported in supplementary material).

342 Lastly, neither the relief nor the landuse could be clearly identify as a factor impacting
343 total Si concentration in soils at least at this scale. Concerning relief, all the proxies of
344 the landforms had a very low weight in this analysis (Figure 4d). For landuse, our data

could not document the positive or negative effect of agriculture, at this scale, on the intensity of silicate weathering that has been suggested in the literature (Struyf et al., 2010; Yang and Zhang, 2018).

3.3 Relationship between Si_{tot} and Si_{CaCl2} in French topsoils.

When considering the total dataset (Table 2), a significantly different from 0, negative correlation between Si_{CaCl2} and Si_{tot} was given but the coefficient is weak (r = -0.32). While considering the different classes, Si_{CaCl2} and Si_{tot} were also weakly negatively correlated and significantly different from 0 for soils on metamorphic rocks. The negative correlation was moderate (r = -0.41) for non-carbonated soils on sediments, very weak (r = -0.16) and poorly significant for soils on igneous intrusive rocks and non significant for soils on igneous extrusive rocks. Such a negative trend was already observed in other continents by Yanai et al. (2016) and Meunier et al. (2017). Nevertheless this negative correlation between Si_{CaCl2} by Si_{tot} was not robust enough to be used as a proxy for PAS. However, the correlation for podzol was stronger than the previous ones (r = -0.75) and for carbonated soils on sediments, a positive non-significant correlation was obtained. PCA (Figure 7) allowed disentangling the observed correlation between Si_{tot} and various soil characteristics for the different soil/parent material classes. Si_{tot} was correlated positively with sand and negatively with clay for Podzols showing that the content of sand was mainly composed of quartz, a poorly soluble mineral. Therefore, the best negative correlation given for podzols between Si_{CaCl2} and Si_{tot} was interpreted as the presence of quartz as the dominant mineral, which control the poorly soluble pool of Si_{tot}. For soils on igneous and metamorphic rocks and non-carbonated soils on sediments, sands were associated with elements such as K or Na, which may be interpreted as the presence of feldspars,

more soluble than quartz and explain the weak correlation between Si_{CaCl2} and Si_{tot}. Besides, the elements opposed to Si_{tot} in the PCA may indicate pools of minerals, which are more favorable to PAS, such as the finer fractions and Al and Fe oxides (Yanai et al., 2016). For metamorphic rocks for instance, Si_{tot} was negatively correlated to Fe and Al concentration as seen by the second axis of the PCA (21 % of the variance), but independent from sand content. This axis expressed a quartz content opposed to a secondary mineral content (represented by Al and Fe) and was independent from the sand content since loess deposits are frequent on metamorphic rocks in the Britany region, one of the largest area of metamorphic rocks in France (Lemercier et al., 2011). For carbonated soils on sediments, the PCA showed that Si_{tot} were negatively correlated to Ca (carbonates) and positively correlated to the sand content along the second axis that represented 20 % of the variability. This axis showed that Si is probably mainly contained in feldspar rather than quartz for explaining the positive trend between topsoil Si_{CaCl2} and Si_{tot}. Thus, Si_{tot} may be a good proxy for Si_{CaCl2} in the cases of podzols, but for the other soil classes a more detailed analysis of the parameters that control Si_{CaCl2} should be done.

4 Conclusions

In a territory as diverse as France, we showed that the total topsoil Si concentrations were highly variable with values ranging from: 22.8 to 456 $g.kg^{-1}$ and thus covering almost the entire range of soil Si concentration recorded in the literature so far.

The spatial variability in total topsoil Si concentrations was as expected due to the diversity of parent material and soil types. However, climate, notably through the impact of rainfall on weathering intensity, was identified as a driving factor for non-carbonated soils. No impact of land use and relief could be identified, as well as no

impact of the relief at least at this scale. Further work based on a paired site approach in different pedo-geological contexts is needed to better conclude on the impact of these factors on the total topsoil Si concentrations.

Lastly, topsoil Si_{CaCl2} tended to decrease when topsoil Si_{tot} increase. The relationship was the strongest in Podzols where Si was mainly contained in quartz. For the other soil classes, we suggest that Si_{tot} may only be considered as a proxy for bioavailable Si if the mineralogical composition is well constrained.

Acknowledgements

This work was performed in the frame of the French ANR BioSiSol project (ANR-14-CE01-0002). RMQS soil sampling and physico-chemical analyses were supported by the GIS Sol, which is a scientific group of interest on soils involving the French Ministry for ecology and sustainable development and Ministry of agriculture, the French National forest inventory (IFN), ADEME (Agence de l'environnement et de la maîtrise de l'énergie, which is a French government agency concerned with environmental protection and energy management), IRD (Institut de recherche pour le développement, which is a French public research organization dedicated to southern countries) and INRA (Institut national de la recherche agronomique, which is a French public research organization dedicated to agriculture s.l.).

1-References

Achache, J., Debeglia, N., Grandjean, G., Guillen, A., Le Bel, L., Ledru, P., Renaud, X., Autran, A., Bonijoly, D., Calcagno, P., Pluchery, E., Guennoc, P., Truffert, C., Rossi,

P., Vairon, J., Avouac, J.P., Poli, E., Senechal, G., Brun, J.P., Galdeano, A., Diament, M., Tarits, P., Mervier, J., Paul, A., Poupinet, G., Marquis, G., Bayer, R., Chautra, J.M., 1997. GEOFRANCE 3D: l'imagerie géologique et géophysique 3D du sous-sol de la France. Mém. Société Géologique 53–71.

Ameijeiras-Marino, Y., Opfergelt, S., Derry, L.A., Robinet, J., Govers, G., Minella, J.P.G., Delmelle, P., 2018. Ge/Si ratios point to increased contribution from deeper mineral weathering to streams after forest conversion to cropland. Applied Geochem. 96, 24-34.

Baize, D., 2000. Guide des analyses en pédologie, Quae. ed. INRA Editions, Paris (France).

Barão, L., Clymans, W., Vandevenne, F., Meire, P., Conley, D.J., Struyf, E., 2014. Pedogenic and biogenic alkaline-extracted silicon distributions along a temperate land-use gradient. Eur. J. Soil Sci. 65, 693–705.

Berner, E.K., Berner, R.A., 1996. Global environment: water, air, and geochemical cycles. Printice Hall, Upper Saddle River, New Jersey.

Bluth, G., J., S., Kumpr, L.R., 1994. Lithologic and climatologic controls of river chemistry. Geochim. Cosmochim. Acta 58, 2341-2359.

Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324

Cerdan, O., Govers, G., Le Bissonnais, Y., Van Oost, K., Poesen, J., Saby, N., Gobin, A., Vacca, A., Quinton, J., Auerswald, K., Klik, A., Kwaad, F.J.P.M., Raclot, D., Ionita, I., Rejman, J., Rousseva, S., Muxart, T., Roxo, M.J., Dostal, T., 2010. Rates and spatial

438 variations of soil erosion in Europe: A study based on erosion plot data.
439 Geomorphology 122, 167–177. https://doi.org/10.1016/j.geomorph.2010.06.011

440 Cornelis, J.T., Delvaux, B., 2016. The functional role of silicon in plant biology. Soil
441 processes drive the biological silicon feedback loop. Functional Ecology 30, 1298-
442 1310.

443 Coskun, D., Deshmukh, R., Sonah, H., Menzies, J.G., Reynolds, O., Ma, J.F.,
444 Kronzucker, H.J., Bélanger, R.R., 2019. The controversies of silicon's role in plant
445 biology. New Phytologist 221, 67-85.

446 De Vos, W., Tarvainen, T., Salminen, R., Reeder, S., De Vivo, B., Demetriades, A.,
447 Pirc, S., Batista, M.J., Marsina, K., Ottesen, R.T., 2006. Geochemical atlas of Europe.
448 Part 2. Interpret. Geochem. Maps Addit. Tables Fig. Maps Relat. Publ. Geol. Surv. Finl.
449 Espoo.

450 Dere, A.L., White, T.S., April, R.H., Brantley, S.L., 2016. Mineralogical transformations
451 and soil development in shale across a latitudinal climosequence. Soil Sci. Soc. Am.
452 J. 80, 623-636.

453 Dürr, H.H., Meybeck, M., Hartmann, J., Laruelle, G.G., Roubeix, V., 2011. Global
454 spatial distribution of natural riverine silica inputs to the coastal zone. Biogeosciences
455 8, 5978-620

456 Egli, M., Mirabella, A., Sartori, G., Fitze, P., 2003. Weathering rates as a function of
457 climate: results from a climosequence of the Val Genova (Trentino, Italian Alps),
458 Geoderma 111, 99-121.

459 European Environment Agency, 2007. CLC2006 technical guidelines. Publications
460 Office, Luxembourg.

461 Faroux, S., Kaptué Tchuenté, A.T., Roujean, J.-L., Masson, V., Martin, E., Moigne,
462 P.L., 2013. ECOCLIMAP-II/Europe: A twofold database of ecosystems and surface
463 parameters at 1 km resolution based on satellite information for use in land surface,
464 meteorological and climate models. Geosci. Model Dev. 6, 563–582.

465 Frings, P.J., Clymans, W., Fontorbe, G., Gray, W., Chakrapani, G.J., Conley, D.J., De
466 La Rocha, C., 2015. Silicate weathering in the Ganges alluvial plain. Earth Planet. Sci.
467 Let. 427, 136-148.

468 Gray, J.M., Bishop, T.F.A., Wilford, J.R., 2016. Lithology and soil relationship for soil
469 modelling and mapping. Catena 147, 429-440.

470 Lucas, Y., Luizao, F.J., Chauvel, A., Rouiller, J., Nahon, D., 1993. The relation between
471 biological activity of the rain forest and mineral composition of soils. Science 260, 521-
472 523.

473 Matheron, G., 1971. The Theory of Regionalised Variables and its Applications. Cah.
474 Cent. Morphol. Math. 5, 212.

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

477 Guntzer, F., Keller, C., Poulton, P.R., McGrath, S.P., Meunier, J.-D., 2012. Long-term
478 removal of wheat straw decreases soil amorphous silica at Broadbalk, Rothamsted.
479 Plant Soil 352, 173–184.

Haysom, M.B.C., Chapman, L.S.,1975. Some aspects of the calcium silicate trials at Mackay. Proc. Qld. Soc. Sugar Cane Technol. 42, 117-122.

Henriet, C., Bodarwé, L., Dorel, M., Draye, X., Delvaux, B., 2008. Leaf silicon content in banana (Musa spp.) reveals the weathering stage of volcanic ash soils in Guadeloupe. Plant Soil 313, 71.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. Int. J. Climatol. 25, 1965–1978. https://doi.org/10.1002/joc.1276

Huete, A., Didan, K., Miura, T., Rodriguez, E.P., Gao, X., Ferreira, L.G., 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sens. Environ., The Moderate Resolution Imaging Spectroradiometer (MODIS): a new generation of Land Surface Monitoring 83, 195–213. https://doi.org/10.1016/S0034-4257(02)00096-2

Institut National de l'Information Géographique et Forestière, 2006. Base de Données Forêt.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning, Springer Texts in Statistics. Springer New York, New York, NY. https://doi.org/10.1007/978-1-4614-7138-7

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

Joly, D., Brossard, T., Cardot, H., Cavailhes, J., Hilal, M., Wavresky, P., 2010. Les types de climats en France, une construction spatiale. Cybergeo Eur. J. Geogr. https://doi.org/10.4000/cybergeo.23155

Keskin, H., Grunwald, S., 2018. Regression kriging as a workhorse in the digital soil mapper's toolbox. Geoderma 326, 22–41. https://doi.org/10.1016/j.geoderma.2018.04.004

King, D., Jones, R.J.A., Thomasson, A.J., 1995. European land information systems for agro-environmental monitoring.

Kuhn, M., Steve, W., Chris, K., Nathan, C., code), R.Q. (Author of imported C., code), R.R.P.L. (Copyright holder of imported C., 2017. Cubist: Rule- And Instance-Based Regression Modeling.

Kursa, M.B., 2014. Robustness of Random Forest-based gene selection methods. BMC Bioinformatics 15, 8. https://doi.org/10.1186/1471-2105-15-8

Kursa, M.B., Rudnicki, W.R., 2010. Feature selection with the Boruta package. J Stat Softw 36, 1–13.

Landré, A., Saby, N.P.A., Barthès, B.G., Ratié, C., Guerin, A., Etayo, A., Minasny, B., Bardy, M., Meunier, J.-D., Cornu, S., 2018. Prediction of total silicon concentrations in French soils using pedotransfer functions from mid-infrared spectrum and pedological attributes. Geoderma 331, 70–80.

Le Bas, C., Barthès, S., Boutefoy, I., Fort, J., Scheurer, O., Darracq, S., Lacassin, J., Sauter, J., 2004. Utilisation des données sols d'I.G.C.S. en France : Un état des lieux. Étude Gest. Sols (2), 8.

523 Liang, Y., Nikolic, M., Bélanger, R., Gong, H., Song, A., 2015. Silicon in agriculture.
524 Dordr. Springer Doi 10, 978–94.

525 Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest 2, 6.

526 Loiseau, T., Chen, S., Mulder, V.L., Román Dobarco, M., Richer-de-Forges A.C.,
527 Lehmann, S., Bourennane, H., Saby N.P.A., Martin, M.P., Vaudour, E., Gomez, C.4
528 Lagacherie, P., Arrouays D. 2019, Satellite data integration for soil clay content
529 modelling at a national scale, International Journal of Applied Earth Observations and
530 Geoinformation, 82 , 108205

531 Magidson, J., 1981. Qualitative variance, entropy, and correlation ratios for nominal
532 dependent variables. Soc. Sci. Res. 10, 177–194. https://doi.org/10.1016/0049-
533 089X(81)90003-X

534 McBratney, A.B., Santos, M.M., Minasny, B., 2003. On digital soil mapping. Geoderma
535 117, 3–52.

536 Meunier, J.D., Sandhya, K., Prakash, N. B., Borschneck, D., Dussouillez, P., 2018.
537 pH as a proxy for estimating plant-available Si? A case study in rice fields in Karnataka
538 (South India). Plant and Soil 432, 143-155.

539 Minasny, B., McBratney, A.B., 2006. A conditioned Latin hypercube method for
540 sampling in the presence of ancillary information. Comput. Geosci. 32, 1378–1388.

541 Minasny, B., McBratney, A.B., Hartemink, A.E., 2010. Global pedodiversity, taxonomic
542 distance, and the World Reference Base. Geoderma 155, 132–139.
543 https://doi.org/10.1016/j.geoderma.2009.04.024

544 Minasny, B., Tranter, G., McBratney, A.B., Brough, D.M., Murphy, B.W., 2009.
545 Regional transferability of mid-infrared diffuse reflectance spectroscopic prediction for
546 soil chemical properties. Geoderma 153, 155–162.
547 https://doi.org/10.1016/j.geoderma.2009.07.021

548 Moriasi, D.N., Arnold, J.G., Liew, M.W.V., Bingner, R.L., Harmel, R.D., Veith, T.L.,
549 2007. Model Evaluation Guidelines for Systematic Quantification of Accuracy in
550 Watershed Simulations. Trans. ASABE 50, 885–900.
551 https://doi.org/10.13031/2013.23153

552 Mulder, V.L., Lacoste, M., Richer-de-Forges, A.C., Arrouays, D., 2016. GlobalSoilMap
553 France: High-resolution spatial modelling the soils of France up to two meter depth.
554 Sci. Total Environ. 573, 1352–1369. https://doi.org/10.1016/j.scitotenv.2016.07.066

555 Narayanaswamy, C., Prakash, N.B., 2010. Evaluation of selected extractants for plant-
556 available silicon in rice soils of Southern India. Comm. Soil Sc. Pl Anal., 41, 977-989.

557 Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part
558 I — A discussion of principles. J. Hydrol. 10, 282–290. https://doi.org/10.1016/0022-
559 1694(70)90255-6

560 NCAR-Research Applications Laboratory, 2015. verification: Weather Forecast
561 Verification Utilities.

562 Pebesma, E., Graeler, B., 2019. gstat: Spatial and Spatio-Temporal Geostatistical
563 Modelling, Prediction and Simulation.

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

566 Quintana-Seguí, P., Le Moigne, P., Durand, Y., Martin, E., Habets, F., Baillon, M.,
567 Canellas, C., Franchisteguy, L., Morel, S., 2008. Analysis of Near-Surface
568 Atmospheric Variables: Validation of the SAFRAN Analysis over France. J. Appl.
569 Meteorol. Climatol. 47, 92–107. https://doi.org/10.1175/2007JAMC1636.1

570 R. Core Team, 2016. R: A language and environment for statistical computing. Version
571 3.3. 1. 2016.

572 Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., O'Connor, P., 2014. Chemistry
573 of Europe's agricultural soils, part A.

574 Rodrigues, F.A., Datnoff, L.E., 2015. Silicon and plant diseases. Springer.

575 Roudier, P., 2017. clhs: Conditioned Latin Hypercube Sampling.

576 Siegel, S., Castellan, N.J., 1988. Nonparametric Statistics for the Behavioral Science,
577 2nd edition. ed. McGraw-Hill Education, New York; etc.

578 Sommer, M., Kaczorek, D., Kuzyakov, Y., Breuer, J., 2006. Silicon pools and fluxes in
579 soils and landscapes—a review. J. Plant Nutr. Soil Sci. 169, 310–329.
580 https://doi.org/10.1002/jpln.200521981

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

584 Taboada, T., Ferro-Vazquez, C., Stoops, G., Martinez Cortizas, A., Rodriguez Flores,
585 R., Rodriguez-Laco, L., 2019. Secondary aluminium, iron and silica phases across a

586 volcanic soil climosaquence, Galapagos Islands. Europ. J. Soil Sci., 70,
587 https://doi.org/10.1111/ejss.12788

588 USGS, 2006. Shuttle Radar Topography Mission, 1 Arc Second Scene
589 SRTM_u03_n008e004, Unfilled Unfinished 2.0. Shuttle Radar Topography Mission.

590 Vandevenne, F.O., Barao, L., Ronchi, B., Govers, G., Meire, P., Kelly, E.F., Struyf., E.,
591 2015. Silicon pools in human impacted soils of temperate zones. Global Geochemical
592 Cycles, 29, doi:10.1002/2014GB005049.

593 Viscarra Rossel, R.A., Walvoort, D.J.J., McBratney, A.B., Janik, L.J., Skjemstad, J.O.,
594 2006. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy
595 for simultaneous assessment of various soil properties. Geoderma 131, 59–75.
596 https://doi.org/10.1016/j.geoderma.2005.03.007

597 Wedepohl, K.H., 1995. The composition of the continental crust. Geochim.
598 Cosmochim. Acta 59, 1217–1232.

599 White, A. F., Vivit, D.V., Shulz, M. S., Bullen, T.D., Evett, R.R., Aagarwal, J., 2012.
600 Biogenic and pedogenic controls on Si distributions and cycling in grasslands of the
601 Santa Cruz soil chronosequence, California. Geochim. Cosmochim. Acta 94, 72-94.

602 Yanai, J., Taniguchi, H., Nakao, A., 2016. Evaluation of available silicon content and
603 its determining factors of agricultural soils in Japan. Soil Sci. Plant Nutr. 62, 511-518.

604 Yang, J.L., Zhang, G.L., 2018. Silicon cycling by plant and its effects on soiil Si
605 translocation in a typical subtropical area. Geoderma 310, 89-98.

606

607

608

Captions

Figure 1: Variability of Si_{tot}. (a) Relationship between the concentration in total Si concentration and in carbonate content in French topsoils. The different soils developed on sedimentary rocks are reported in yellow while those developed on igneous and metamorphic rocks are reported in grey; (b) Separation between carbonated soils and non-carbonated soils on sedimentary rocks based on the geo-pedological classification. All the soil groups with mean carbonate content lower than 5 % were considered as non-carbonated soils.

Figure 2: Boxplots of Si_{tot} for the main geological domains and podzols. The boxes represent the interquartile range, the bold horizontal segment the median Si_{tot}, the whiskers, the 99 % range and the black dots the outliers. Letters above boxplots represent groups with Si_{tot} significantly different and numbers, the sample size.

Figure 3: Boxplots of Si_{tot} for all the considered pedo-geological groups. The boxes represent the interquartile range, the bold horizontal segment the median Si_{tot}, the whiskers, the 99 % range and the black dots the outliers. Colors represent distinction of main geological domains plus podzols.

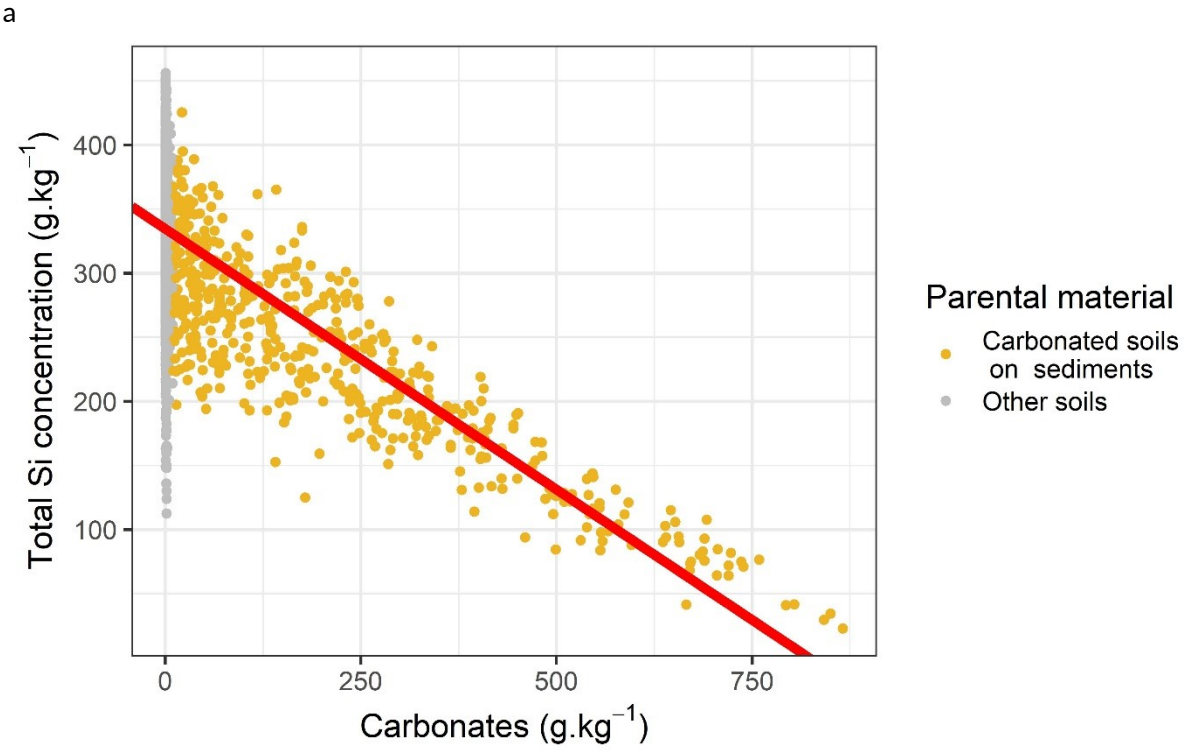
Figure 4: Si_{tot} spatial distribution from digital soil mapping approach (DSM).

Figure 5: Variable importance in the random forest Model. The X-axis is the Mean Decrease Accuracy computed as the increase percentage in mean squared error of predictions (estimated with out-of-bag-Cross Validation) as a result of variable j being permuted. The definition of the environmental covariates are listed in Table 2.

Figure 6: Plots of the total topsoil Si concentration versus the 30 year average of annual rainfall for the different parent material domains. Soils on sedimentary parent materials were sorted according to the nature of the parent material for the carbonated soils (shalk, marl, limestone and other) and of the texture of the parent material for non-carbonated soils (clay, loam, sand). The black lines represent the fitted linear regression. The Pearson correlation is also provided per parent material domains. Stars represents the level of signification of the correlation: one star being significant at a 5 % confidence level, two stars at 1 % and three stars at 1 ‰.

Figure 7: Variables correlation circle for the first two components of the Principal Component Analyzes (PCA) performed for each main geological domain and podzols. In blue, Si considered as a passive variable in these PCAs.

a


Total Si concentration (g.kg^{-1})
400
300
200
100
0
250
500
750
Carbonates (g.kg^{-1})
Parental material
Carbonated soils on sediments
Other soils

b

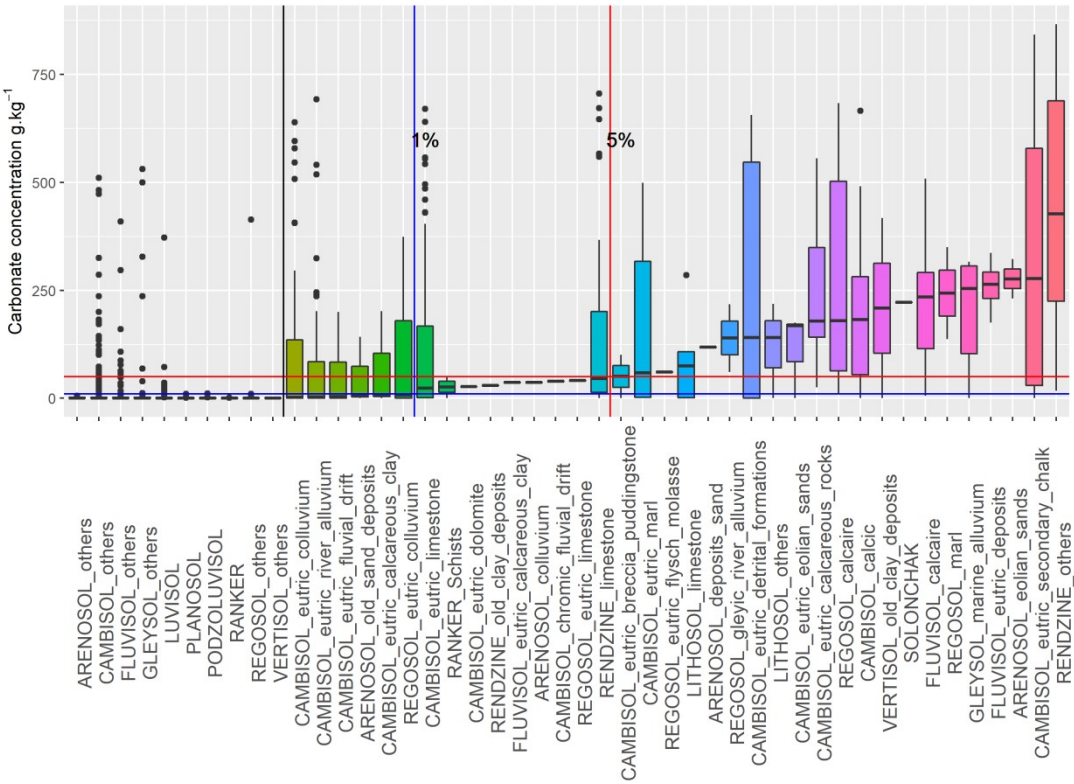

Carbonate concentration g.kg^{-1}
750
500
250
0
1%
5%
ARENOSOL_others
CAMBISOL_others
FLUVISOL_others
GLEYSOL_others
LUVISOL
PLANOSOL
PODZOLUVISOL
RANKER
REGOSOL_others
VERTISOL_others
CAMBISOL_eutric_colluvium
CAMBISOL_eutric_river_alluvium
CAMBISOL_eutric_fluvial_drift
ARENOSOL_old_sand_deposits
CAMBISOL_eutric_calcareous_clay
REGOSOL_eutric_colluvium
CAMBISOL_eutric_limestone
RANKER_Schists
CAMBISOL_eutric_dolomite
RENDZINE_old_clay_deposits
FLUVISOL_eutric_calcareous_clay
ARENOSOL_colluvium
CAMBISOL_chromic_fluvial_drift
REGOSOL_eutric_limestone
RENDZINE_limestone
CAMBISOL_eutric_breccia_puddingstone
CAMBISOL_eutric_marl
REGOSOL_eutric_flysch_molasse
LITHOSOL_limestone
ARENOSOL_deposits_sand
REGOSOL_gleyic_river_alluvium
CAMBISOL_eutric_detrital_formations
LITHOSOL_others
CAMBISOL_eutric_eolian_sands
CAMBISOL_eutric_calcareous_rocks
REGOSOL_calcaire
CAMBISOL_calcic
VERTISOL_old_clay_deposits
SOLONCHAK
FLUVISOL_calcaire
REGOSOL_marl
GLEYSOL_marine_alluvium
FLUVISOL_eutric_deposits
ARENOSOL_eolian_sands
CAMBISOL_eutric_secondary_chalk
RENDZINE_others

Figure 1

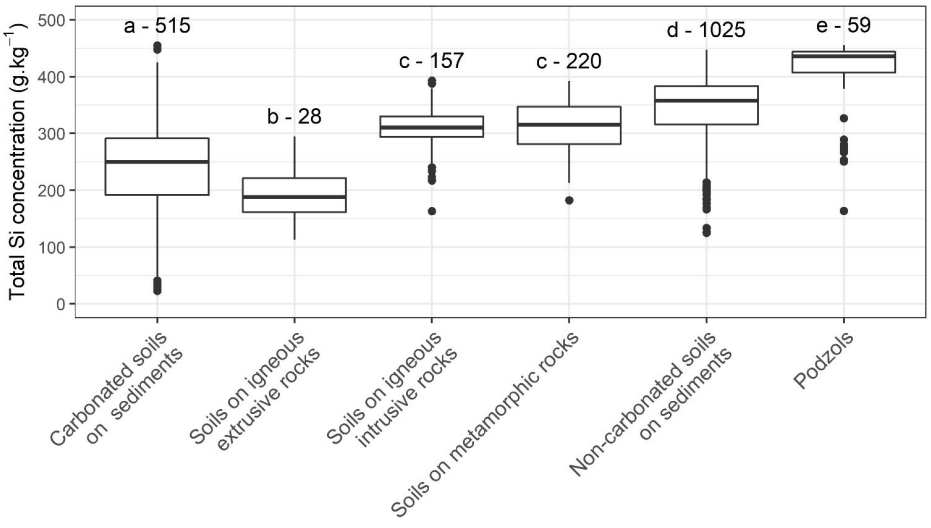

Total Si concentration (g.kg^{-1})
500
400
300
200
100
0
a - 515
b - 28
c - 157
c - 220
d - 1025
e - 59
Carbonated soils on sediments
Soils on igneous extrusive rocks
Soils on igneous intrusive rocks
Soils on metamorphic rocks
Non-carbonated soils on sediments
Podzols

Figure 2

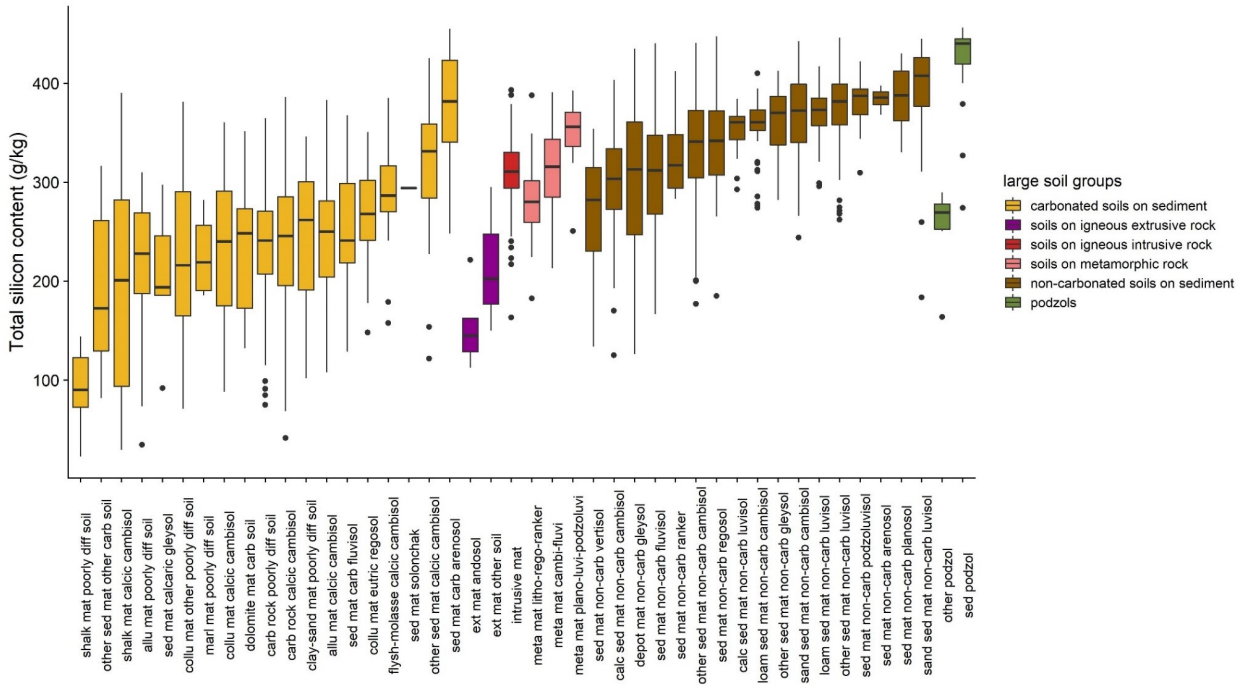

Total silicon content (g/kg)
100
200
300
400
shalk mat poorly diff soil
other sed mat other carb soil
shalk mat calcic cambisol
allu mat poorly diff soil
sed mat calcaric gleysol
collu mat other poorly diff soil
marl mat poorly diff soil
collu mat calcic cambisol
dolomite mat carb soil
carb rock poorly diff soil
carb rock calcic cambisol
clay-sand mat poorly diff soil
allu mat calcic cambisol
sed mat carb fluvisol
collu mat eutric regosol
flysh-molasse calcic cambisol
sed mat solonchak
other sed mat calcic cambisol
sed mat carb arenosol
ext mat andosol
ext mat other soil
intrusive mat
meta mat litho-rego-ranker
meta mat cambi-fluvi
meta mat plano-luvi-podzoluvi
sed mat non-carb vertisol
calc sed mat non-carb cambisol
depot mat non-carb gleysol
sed mat non-carb fluvisol
sed mat non-carb ranker
other sed mat non-carb cambisol
sed mat non-carb regosol
calc sed mat non-carb luvisol
loam sed mat non-carb cambisol
other sed mat non-carb gleysol
sand sed mat non-carb cambisol
loam sed mat non-carb luvisol
other sed mat non-carb luvisol
sed mat non-carb podzoluvisol
sed mat non-carb arenosol
sed mat non-carb planosol
sand sed mat non-carb luvisol
other podzol
sed podzol
large soil groups
carbonated soils on sediment
soils on igneous extrusive rock
soils on igneous intrusive rock
soils on metamorphic rock
non-carbonated soils on sediment
podzols

Figure 3

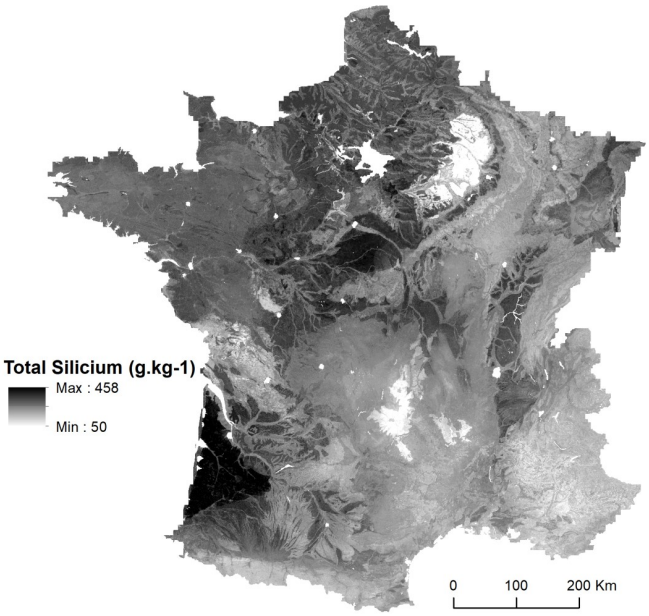


Figure 4

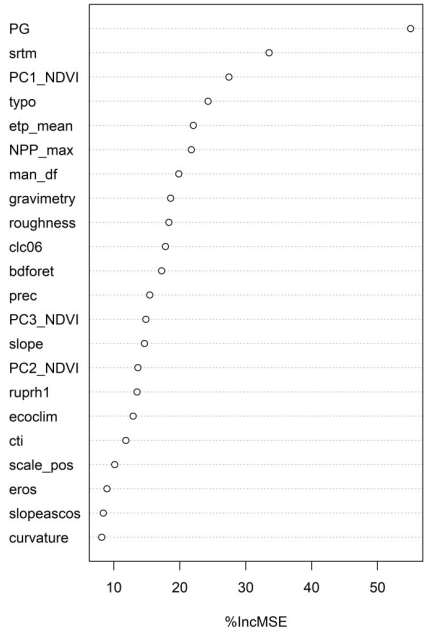

PG
srtm
PC1_NDVI
typo
etp_mean
NPP_max
man_df
gravimetry
roughness
clc06
bdforet
prec
PC3_NDVI
slope
PC2_NDVI
ruprh1
ecoclim
cti
scale_pos
eros
slopeascos
curvature
10
20
30
40
50
%IncMSE

Figure 5

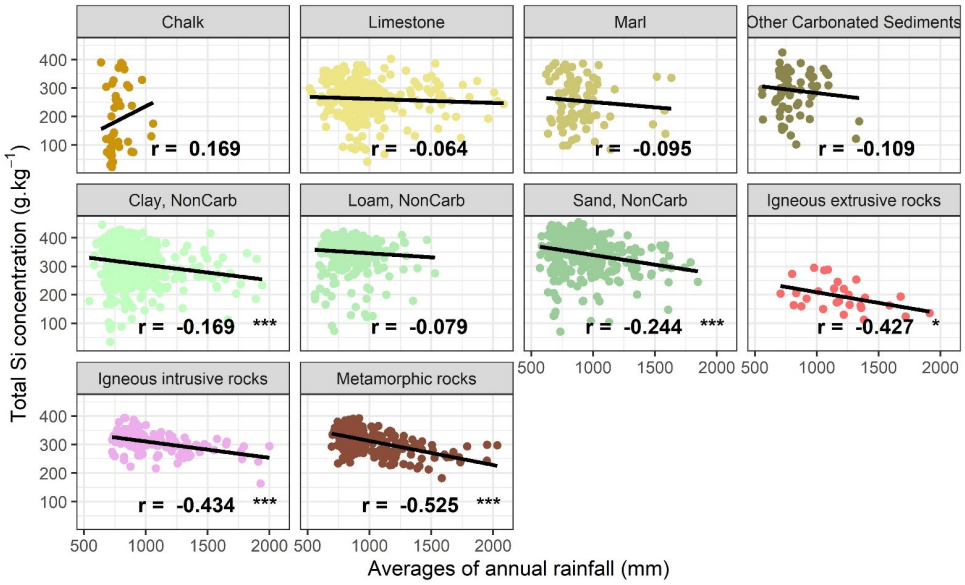
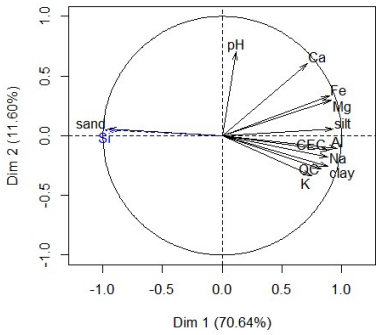
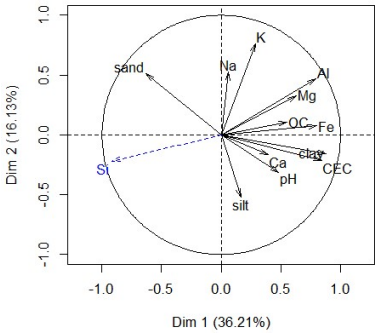

Chalk
Limestone
Marl
Other Carbonated Sediments
Clay, NonCarb
Loam, NonCarb
Sand, NonCarb
Igneous extrusive rocks
Igneous intrusive rocks
Metamorphic rocks
r = 0.169
r = -0.064
r = -0.095
r = -0.109
r = -0.169 ***
r = -0.079
r = -0.244 ***
r = -0.427 *
r = -0.434 ***
r = -0.525 ***
Total Si concentration (g.kg^{-1})
Averages of annual rainfall (mm)
100
200
300
400
500
1000
1500
2000

Figure 6

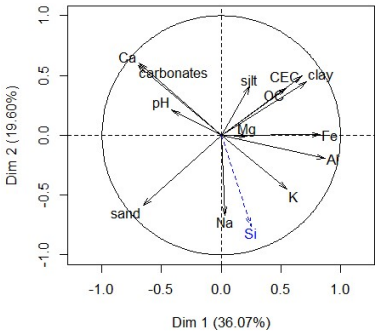
a- Podzols



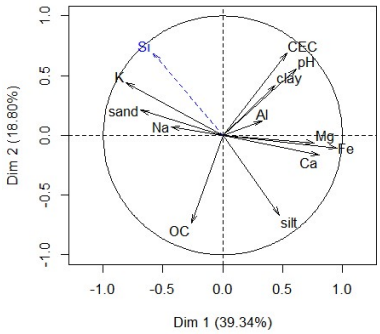
b- Non-carbonated soils on sediments



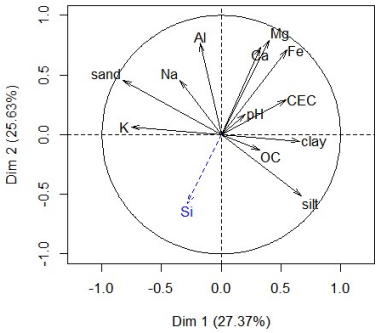
c- Carbonated soils on sediments



d- Soils on igneous extrusive rocks



e-Soils on igneous intrusive rocks



f- Soils on metamorphic rocks

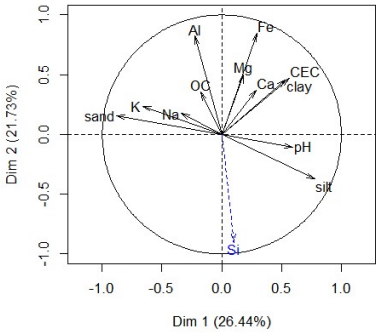


Figure 7

Table 1: Exhaustive categorical and continuous covariates used for the Digital Soil Mapping approach

Pedological factor	abreviates	variables	scale	type	Reference
Vegetation	bdforet	Forest type		category	Institut National de l'Information Géographique et Forestière (2006)
	ecoclim	Ecoclimap land use	1km	category	Faroux et al. (2013)
	NPP_max	Net Primary Production	1km	quantitative	LPDAAC (2001)
	clc06	land use (Corine Land Cover)	250m	category	European Environment Agency (2007)
	PC1_NDVI	1st component of Normalized Difference Vegetation Index	90m	quantitative	Loiseau et al. (2019)
	PC2_NDVI	2nd component of Normalized Difference Vegetation Index	90m	quantitative	Loiseau et al. (2019)
	PC3_NDVI	3rd component of Normalized Difference Vegetation Index	90m	quantitative	Loiseau et al. (2019)
Relief	cti	Compound Topographic Index (SRTM)	90m	quantitative	USGS (2006)
	roughness	roughness (SRTM)	90m	quantitative	USGS (2006)
	curvature	curvature (SRTM)	90m	quantitative	USGS (2006)
	scale_pos	slope position (SRTM)	90m	quantitative	USGS (2006)
	slope	slope (SRTM)	90m	quantitative	USGS (2006)
	slopascos	slope cosinus (SRTM)	90m	quantitative	USGS (2006)
	srtm	elevation DEM (SRTM)	90m	quantitative	USGS (2006)
	eros	Erosion rate	1 :1000000	quantitative	Cerdan et al. (2010)
Climate	prec	mean annual precipitation	~1km	quantitative	Hijmans et al. (2005)
	Typo	climate typology	250 m	category	Joly et al. (2010)
	etp_mean	mean annual potential evapotranspiration	8km	quantitative	Quintana-Segui et al. (2008)
Parent material	gravimetry	free-air Bouguer anomaly	4km	quantitative	Achache et al. (1997)
Soil	ruprh1	available water capacity	1 :1000000	quantitative	Le Bas et al. (2004)
Soil and parent material	PG	Pedological and Geological classification	1 :1000000	category	King et al. (1995)

Table 2: Pearson correlation coefficient between Si_{CaCl2} and Si_{tot} in French topsoils. Significance levels: ***: *p*-value < 0.001; **: *p*-value < 0.01; *: *p*-value < 0.05

Classes	n	r
Carbonated soils on sediments	508	0.12***
Soils on igneous extrusive rocks	28	-0.25
Soils on igneous intrusive rocks	155	-0.16*
Soils on metamorphic rocks	217	-0.26***
Non-carbonated soils on sediments	1020	-0.41***
Podzols	58	-0.75***
All data	1986	-0.32***

DO RAINFALL AND LAND USE AFFECT THE POOL OF TOTAL SI CONCENTRATION? A DIGITAL SOIL MAPPING APPROACH OF FRENCH TOPSOILS.

Landré, A.[a,b], Cornu, S.[c], Meunier, J.-D.[c] ,Guerin A. [d], Arrouays D. [a], Caubet M. [a], Ratié C.[a], , Saby, N.P.A.[a]

a) INRA, Infosol, US 1106, Orléans, France
b) INRA - INPT-ENSAT - INPT-EI-Purpan, UMR 1248 AGIR AGroécologie, Innovations, teRritoires. Centre de recherche Occitanie-Toulouse, Auzeville, France.
c) Aix-Marseille Univ, CNRS, IRD, Coll de France, INRA, CEREGE, Aix-en-Provence, France
d) INRA, Laboratoire d'Analyses des Sols US, Arras, France

Declarations of interest: none

Supplementary file: Determination of the total Si concentration (Si_{tot}) based on the concentration in major elements

Landre et al. (2018) analyzed Si_{tot} for a subset of 673 samples, among the 2007 RMQS samples. Si_{tot} for the remaining sites was modelled using a Cubist modelling approach, as follow:

$$\mathrm{Si} = \mathrm{f}(\mathrm{Al}, \mathrm{Fe}, \mathrm{K}, \mathrm{Na}, \mathrm{Ca_{nc}}, \mathrm{Mg_{nc}}, \mathrm{P}, \mathrm{Mn}, \mathrm{OC}, \mathrm{CaCO_3}, \text{residual water}) \quad \text{(S1)}$$

Where f is the cubist regression trees model,

Ca_{nc} and Mg_{nc} are the fraction of calcium and magnesium, respectively, that are not included in carbonate minerals nor adsorbed on the exchangeable surfaces. They were estimated as follow

$$Ca_{nc} = Ca_{tot} - \left(\left(\frac{CaCo_3}{10} * 0.401\right) + Ca_{exch}\right) \quad \text{with} \quad Ca_{nc} \geq 0 \quad \text{(S2)}$$

$$Mg_{nc} = Mg_{tot} - (Mg_{ech} + Ca_{excess}) \quad \text{with} \quad Mg_{nc} \geq 0 \quad \text{(S3)}$$

Where $Ca_{excess} = Ca_{nc}$ when $Ca_{nc} \leq 0$ and $Ca_{nc} = 0$ otherwise

The Cubist model is a form of regression rules that build unconventional regression trees, with final nodes containing linear models instead of discrete values (Quinlan, 1992). Cubist creates comprehensible rules that describe the relationships between predictive variables (here spectra or soil properties) and the variable of interest.

In order to calibrate and validate the models, we used a repeated cross-validation approach combined to a bootstrap step (James et al., 2013). Repeated cross-validation allows assessing the quality of the prediction. It involves randomly dividing the available set of data into two parts, a calibration set and a validation set. We repeated this operation 10 times using a split of 75%-25% for calibration and validation respectively. This subdivision was performed using the conditioned Latin Hypercube

Sampling (cLHS) method (Minasny and McBratney, 2006). This method is a stratified random procedure that provides an efficient way of sampling variables from their multivariate distributions. The bootstrap step allows assessing the uncertainty of the model predictions. For this, we simulated 100 datasets by random sampling with replacement 95 % of the calibration dataset from the cross-validation step. This procedure provided then 100 Cubist models. The final prediction is obtained by averaging the predictions of 100 bootstrapped models. Our modelling approach involves a large number of calibration and validation operations, and we thus used a parallel processing approach to overcome the computational load. It is implemented in R using the packages foreach, doParrallel or snow. We used the Cubist model implemented in the R package Cubist (Kuhn et al., 2016), the cLHS function implemented in the R package clhs (Roudier, 2011) and the crps function implemented in the R package verification (Laboratory NCAR-Research Applications, 2015).

The model was assessed using three conventional performance indicators: the coefficient of determination (R^2), the root mean square error (RMSE) also known as standard error of prediction (SEP) and the bias, which is the mean residual of the model. We also took into account the probabilistic characteristic of model predictions by using the continuous rank probability score average (CRPS, equation 6). The CRPS represents the closeness between the prediction distribution and the corresponding observations (Gneiting et al., 2007).

$$CRPS = \int_{-\infty}^{\infty} BS(y)\, dy \qquad \text{(s4)}$$

$$BS(y) = \frac{1}{n}\sum_{i=1}^{n} \{(F_i(y) - \mathbb{1}(x_i \leq y)\}^2 \qquad \text{(s5)}$$

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

The Si_{tot} prediction function is yielding R^2 greater than 0.98 with a very small variance (0.0012) among repetitions of the cross-validation, RMSE lower than 11 g kg^{-1} and CRPS lower than 7 g kg^{-1} for measured Si_{tot} ranging from 22.81 to 455.8 g kg^{-1} over the RMQS database with a median equal to 323.6 g kg^{-1}. These results are better than the results obtained by Landre et al. (2018) with their MIRS PTF and suggest that our PTF can predict with higher accuracy the Si content.

To better figure out the accuracy of our PTF, we plotted in Figure S1, the predicted *versus* measured Si_{tot} for only one iteration of the cross-validation step. In this figure, the predicted *versus* measured Si_{tot} distribution closely follow the one to one line with a smaller prediction uncertainty than the analytical uncertainty. Those results allow us considering that the predictions of this function are as good as Si measurements since the prediction uncertainties are within the range of analytical values. However, as explained by Landre et al. (2018), the analytical uncertainty was not taken into account in the prediction uncertainty calculation, as they were not always available. Despite this, the obtained PTF shows exceptional accuracy that is rarely obtained in environmental fields for PTFs (Minasny et al., 2009; Viscarra Rossel et al., 2006) (see Landre et al. (2018) for more information).

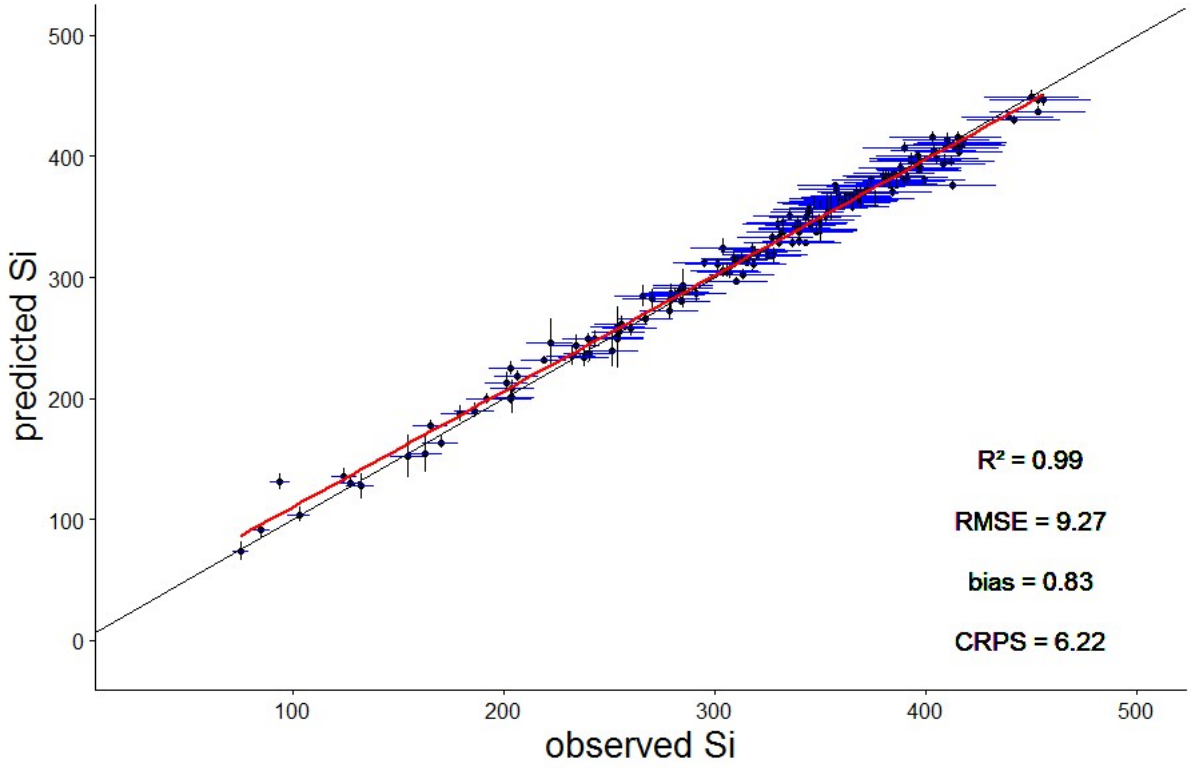


Figure S1: Predicted versus measured Si concentrations (in g kg^{-1}) for the first cross-validation replication of our Si concentration predictive function. In black, the one to one line and in red, the regression line. Black vertical error bars represent the prediction's uncertainty and blue horizontal error bars represent the analytical uncertainty.

DSM model validation

The model performance was estimated using a cross validation. four conventional performance indicators were computed: the coefficient of determination (R^2), the mean square error (MSE), the root mean square error (RMSE, also known as standard error of prediction, SEP) and the bias, which is the mean residual of the model. In addition, a fifties indicator was calculated, the concordance which is a normalized statistic that determines the relative magnitude of the residual variance compared to the measured data variance (Moriasi et al., 2007; Nash and Sutcliffe, 1970). The results of the cross validation of the SCORPAN model are in accordance with previous digital soil mapping exercise of topsoil properties (Mulder et al., 2016). We found an R^2 of 0.45, a MSE of 3647.99, a RMSE of 60.4 g kg^{-1}, an important bias of 4.70 and a high concordance of 0.59.