Efficient models for predicting durum wheat grain Cd conformity using soil variables and cultivars

Christophe Nguyen, Agathe Roucou, Guénolé Grignon, Jean-Yves Cornu, Benoit Meleard

Efficient models for predicting durum wheat grain Cd conformity using soil variables and cultivars

Christophe Nguyen[a*], Roucou Agathe[b], Guénolé Grignon[b],
Cornu Jean-Yves[a], Benoît Méléard[b]

May 27, 2020

[a] ISPA, Bordeaux Sciences Agro, INRAE. F-33140 Villenave-d'Ornon cedex, France.

[b]ARVALIS-Institut du Végétal. Station expérimentale, 91720 BOIGNEVILLE,
https://www.english.arvalisinstitutduvegetal.fr

*Corresponding author phone: +33 (0)5 57 12 25 07; fax: +33 (0)5 57 12 25 15;
e-mail: christophe.nguyen@inrae.fr

1 Abstract

Contamination of durum wheat grain by cadmium (Cd) threatens food safety and is of increasing concern because regulations concerning Cd are becoming stricter due to its toxicity. This work aimed at using soil variables and cultivar types to build models to predict whether durum wheat grain Cd will conform with current and possibly lower regulatory thresholds. We combined multiple Gaussian and logistic regressions and the random forest algorithm to take advantage of their strength. Models tested using cross-validation produced excellent performances including for the lowest regulatory threshold of 0.1 mg Cd/kg, half of the current one: 79-85% of the non-conformity cases were detected and the reliability of predictions was 69-82%. The models enabled identification of a x1.4 variability in grain Cd content between cultivars

that do not have the low Cd accumulation allele of the *Cdu1* gene. The models confirmed that for the grain Cd content, the between-cultivar variability had much less influence than the phytoavailability of Cd in soil, the critical contexts of which were characterized by the models. For farmers, these models are valuable tools to predict whether durum wheat production will conform with existing and future Cd regulation in foodstuffs.

Keywords: Cadmium; Durum wheat; Genetic variability; Models; Phytoavailability

2 Introduction

Cadmium (Cd) is a highly toxic and carcinogenic metal found naturally in soils. It is taken up by plant roots and transferred to edible plant parts and is therefore a major threat to food safety (Clemens et al., 2013). In 2009, the European Food Safety Authority (EFSA) published a scientific opinion recommending a tolerable weekly intake (TWI) for Cd of 2.5 μg kg^{-1} body weight, almost three times lower than the previous one set by the World Health Organization, which was 7 μg kg^{-1} (Alexander et al., 2009). Durum wheat concentrates more Cd in grain than bread wheat (Greger and Lofstedt, 2004). Durum wheat is a major contributor of Cd to human food intake as it is widely consumed in pasta and semolina (Clarke et al., 2010). For instance, 9 Mt of durum wheat are consumed in Europe, which is also the world's main exporter at 8 Mt (FranceAgrimer, 2020). Following the downward revision of the TWI by EFSA, the European commission also revised the directive EC1881/2006, which fixes the maximum content of Cd in some foodstuffs (DGSANCO, 2011). For durum wheat, 0.1 and 0.15 mg Cd kg^{-1} were originally considered but the project was abandoned because of the strong economic negative impact the decision would have. However, the European countries were asked to conduct research and develop practices aimed at monitoring and reducing crop contamination (EC EC 488/2014, 2014). Recently, new downwards revisions of the level of Cd in cereals including for durum wheat have again been the subject of discussions (ARVALIS-Institut du végétal, pers. comm.).

Crop uptake of Cd depends on its phytoavailability in soils, *i.e.* on the flux of Cd^{2+} at the root surface, assuming that the free ion is the main species absorbed by root cells (Clemens, 2019). In aerobic agricultural soils, Cd phytoavailability is mainly determined by sorption onto the solid organic and mineral phases (clay and oxides), by complexation with soluble ligands, generally organic compounds, and by transport to plant roots by diffusion and advection (Lin et al., 2016; Antoniadis et al., 2017; Vega et al., 2010). Soil Cd, pH, organic matter and to a lesser extent clay and oxides have been found to be the major regressors of statistical models to predict the soluble soil Cd and its accumulation by plants (Adams et al., 2004; de Vries et al., 2011; Groenenberg et al., 2010; Horn et al., 2006). Except oxides, these variables are commonly measured in soil testing, and therefore, predicting plant Cd by using statistical models based on these variables is of great interest (for example, Hough et al., 2003; Tudoreanu and Phillips, 2004; Viala et al., 2017). According to the literature, one major drawback of such models is that they have not been tested on new data (cross-validation). Their real predictive value for new data is consequently not known, which is an obstacle to their practical use in the field. Furthermore, because of the log-log relationship, the variance of the predictions inflates when the mean increases (Newman, 1993), making it difficult to rule on the conformity of critical grain samples with high Cd content. Binary classification models that predict conformity (yes/no) could be much more efficient because they concentrate on predicting the class, whereas log-log models are optimized to predict continuous variations in grain Cd. Among binary classification models, two have been shown to be particularly efficient: logistic regression and classification trees (Hastie et al., 2009). Logistic regression predicts the probability that an observation is positive (actually, the log of the odds) from a linear combination of predictors. If the predicted probability is greater than a cutoff threshold (generally p=0.5), the case is classified as positive, and conversely as negative. Classification trees are a set of hierarchical decision rules that split data into two classes based on cutoff values of the most relevant predictors. In contrast to logistic regression, classification trees are a non-parametric method and may better model a complex boundary between the two conformity classes. However, they are very sensitive to the training dataset (high variance). To cope with this problem, Breiman (2001) proposed the random forest approach, which consists in aggregating the predictions of a large number of trees (*i.e.* a forest) that are uncorrelated by bootstrapping the training dataset.

Soil conditions and particularly soil pH, strongly influence the contamination of crops by metals by controlling the phytoavailability of the latter (Kabata-Pendias, 2004). Reducing the phytoavailability of metals is often difficult, especially if the soil pH is already high. This is typically the case of calcareous soils, on which durum wheat is usually grown in France. Therefore, it is worth taking advantage of any between-cultivar variability in Cd accumulation by crops (Li et al., 1997; Li and Zhou, 2019). Due to the strong importance of durum wheat in Canada, genetic selection of cultivars that accumulate little Cd in their grain began in the 1990s when international discussions about setting Cd limits in food products started (Clarke et al., 2010). It is known that a large part of phenotypic variability in durum wheat grain Cd is linked to the *Cdu1* locus of chromosome 5B, which is involved in Cd sequestration in roots (Knox et al., 2009; Wiebe et al., 2010). One deficient allele of the gene coding for the HMA3 transporter that transfers Cd and Zn from the root cytosol into the vacuole is thought to have been selected inadvertently during breeding because it promoted growth of durum wheat in Zn-deficient soils (Maccaferri et al., 2019). As a consequence of reduced sequestration in root vacuoles due to this deficient allele, more Cd is allocated to aboveground organs, including the grain. To our knowledge, selection of low Cd durum wheat cultivars has not yet begun in Europe but thanks to some markers of the *Cdu1* locus (AbuHammad et al., 2016; Oladzad-Abbasabadi et al., 2018; Salsman et al., 2018), many common European cultivars have been assessed, and the results show that a large proportion of all cultivars are high Cd accumulators (Zimmerl et al., 2014). Therefore, as suggested by preliminary results obtained in controlled conditions (Perrier et al., 2016), it would certainly be worth characterizing the variability of grain Cd content among high Cd cultivars to discard the highest accumulators, if this is possible with respect to other agronomic performances.

Based on these elements, the present work had two goals. The first was to build sensitive and reliable models to predict durum wheat grain Cd conformity using soil analysis variables. We hypothesized that soil analysis variables combined with highly efficient statistical approaches would make it possible to obtain predictive models that are sufficiently sensitive and reliable for practical use in the field. The second goal was to identify possible solutions in the case of predicted non-conformity. To this end, we investigated the variability of Cd accumulation between cultivars that do not possess the low Cd allele of the *Cdu1* gene and we used the model simulations to identify the soil conditions that could lead to an excessive contamination of durum wheat grain by Cd.

3 Materials and Methods

3.1 Collection and analysis of paired soil and grain samples from farms and of grain samples from trials comparing cultivars

Between 2012 and 2018, 420 paired samples of soil and durum wheat grain were collected in farms and in ARVALIS-Institut du végétal trials across the regions that are representative of the French production. There were 192 different soil samples because several cultivars were grown on the same soil. The cultivars we studied had either been characterized as high Cd accumulators (Zimmerl et al., 2014) or had not yet been characterized (Table SI1, Supplementary Information). Composite soil samples were made by mixing 12 sub-samples taken from the 0-25 cm topsoil layer on a grid representative of each plot. The composite soil samples were air-dried, sieved at 2 mm before the common characteristics of agricultural soil testing were determined by a certified Inrae soil analysis laboratory (https://www6.hautsdefrance.inrae.fr/las). The total soil Cd was quantified by ICP-MS after solubilization by fluorhydric and perchloric acids (NF X 31–147). Soil pH was measured in a 1:5 soil to water solution. Total soil organic carbon (SOC) was quantified by dry combustion and corrected for carbonate content (NF ISO 10694). The Robinson pipette method (NF X 31–107) was used for soil texture (5 size classes). Total soil $CaCO_3$ content was obtained using the acid neutralization method (NF X 31–105). In France, no other oxides are usually analyzed in soil testing even though they could play an significant role in controlling Cd phytoavailability (Sun et al., 2017). The descriptive statistics of the studied soils are listed in Table 1. Grain samples, that were selected as being representative of the plot harvest, were analyzed by Capinov certified laboratory (https://www.capinov.fr/) without further drying as stated by the European regulation EC466/2001 (2001). Total grain Cd was quantified by atomic absorption spectroscopy (AAS) after wet digestion in a mixture of nitric acid (10% v:v) and hydrogen peroxide (4% v:v). Both the Capinov and Inrae laboratories are certified by Cofrac for quality controls (https://www.cofrac.fr/en/).

To rank durum wheat cultivars with respect to their capacity to accumulate Cd, additional grain samples were collected in 2016, 2017 and 2018 from experimental plots located in three ARVALIS-Institut du végétal research centers. These trials are conducted annually to assess the agronomic performances of durum wheat cultivars. The trials are located in the south-west

(Bergerac, 43° 25' 0.12''E 0° 9' 0''), south (Montesquieu-Lauragais, 43° 25' 0.12''E 0° 9' 0'') and center (Thizay, 47° 10' 0.12''E 0° 0' 0'') of France, the main French durum wheat production regions. Grain Cd content was quantified using the same procedure as that described above. No soil was collected and these samples were not used for modeling.

3.2 Modeling durum wheat grain conformity with respect to the Cd regulatory threshold

The goal of the models was to use soil characteristics to predict if a grain sample of a particular durum wheat cultivar will complie (0) or not (1) with the regulatory threshold (RT). Seven soil variables were selected as predictors and combined hierarchically based on the following ranking Cd>pH>SOC>Clay>{Fine silt, calcareous}>coarse loam. The database has a minimum of 75 positive cases for the RT of 0.2 mg Cd kg^{-1} grain. Therefore, a maximum of 7 predictors were considered, based on the guidelines proposed by Peduzzi et al. (1996), which state that at least 10 positive cases per predictor are required for a correct estimation of effects in a logistic regression. In total, nine combinations of predictors were tested, from the simplest including only the soil Cd to the full model including the 7 soil predictors. All models also include the cultivar as a categorical predictor.

Three statistical modeling approach were tested. The first approach was the mixed-effects logistic regression (MELR), which predicted the *log*-odds of the non-conformity of the grain Cd as a function of the *log* of the scaled soil variables (fixed effects) and of the cultivar (random effect for the intercept).

$$\begin{aligned} log\left(\tfrac{p}{1-p}\right) &= log(a_0) + \sum a_i log\left(\tfrac{X_i}{mean(x_i)}\right) \\ random &= log(a_0)|cultivar \\ p &= p(Cd_{grain} \geq RT) = p(Y=1) \end{aligned} \tag{1}$$

Continuous predictors were *log* (natural) transformed to i) ensure the normality of the residuals and to ii) model interactions between soil variables since the sum of *log* of variables is the *log* of the product of the variables. In order to correctly estimate the model intercept, the soil variables were first scaled by dividing them by their mean before *log* transformation. For each of the nine models, the predicted *log*-odds from the calibration dataset allowed us to calculate the probabilities of non-conformity $p(Y = 1)$. Then, each probability was individually tested as a cutoff *(ctf)* to code all predicted probabilities in conformity classes: 0 if $p \leq ctf$ and 1 if $p > ctf$. The predicted conformity classes were compared to the actual classes to calculate the confusion matrix, *i.e.* the scores for the true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). The performances of the model for a given cutoff were assessed using the Youden's J statistic ($0 \leq J \leq 1$) and the Matthews correlation coefficient MCC ($-1 \leq MCC \leq 1$):

$$J = \frac{TP}{TP + FN} - \frac{FP}{TN + FP} \tag{2}$$

$$MCC = \frac{TP.TN - FP.FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3}$$

For both statistics, a value of 1 indicates a perfect model with no FN nor FP. The MCC considers the four scores of the confusion matrix (TP, TN, FP, FN) and thus makes it possible to identify models with good performances in both the detection and reliability of predictions of actual positive and negative cases. By contrast, the Youden index concentrates on the predicted positive class (TP and FP) and favors the selection of models that have the best performances in detecting positive cases. The two best cutoffs, each maximizing either the J or the MCC statistics were identified and used to make future class predictions.

The second approach was random forest (RF) modeling (Breiman, 2001). For each of the nine models, 501 classification trees were trained with bootstrap samples that were stratified based on the (0,1) frequencies of the training dataset. The number of variables randomly sampled as candidates at each split was equal to the square root of the total number of predictors of the model. The minimum number of observations in terminal nodes was 1. For each observation of the calibration dataset, the predicted probability that the grain Cd is above the regulatory threshold was estimated by the frequency of the positive class predicted by the 501 trees for this observation. Like for logistic modeling, each predicted probability was tested as a cutoff for coding into the (0,1) classes. The Youden and MCC statistics were again used to select the best cutoffs.

The third approach was the mixed effects Gaussian multiple linear *log-log* regression (MEMR). The *log* of the grain Cd content was predicted from the *log* of the scaled soil predictors (fixed effects) with the cultivar as a random effect for the intercept.

$$\begin{aligned} log(Cd_{grain}) = log(a_0) + \sum a_i log\left(\frac{X_i}{mean(x_i)}\right) \\ random = log(a_0)|cultivar \end{aligned} \tag{4}$$

The reasons for scaling using the median and for *log* transforming the continuous variables were the same as for the MELR. After back transformation, the predicted grain Cd (Cd_{grain}) was coded as follows: 0 if $\mathrm{Cd}_{grain} \leq RT$, otherwise 1.

3.3 Ranking the predictive value of the models by 5-fold cross-validation

The predictive values of the 9 models x 3 modeling approaches (RF, MELR and MEMR) were evaluated by a 5-fold cross-validation with 400 repetitions. A single design was used per RT (splitting the database into 5 groups 400 times). One repetition consisted in randomly splitting the database into five groups of 84 observations, each group having the same proportions of (0,1) as the whole database. Each of the five groups was individually used to make predictions from the model trained on the four remaining groups that were pooled. For each of the 420 observations of the database, the 400 repetitions were bootstrapped 400 times to generate 400 combinations of the 420 predictions for the whole database. This enabled to have 400 repetitions of the J and MCC statistics to further calculate their mean and standard deviation in order to compare and group the 9 models x 3 modeling approaches at p=0.05 adjusted for the multiple comparisons (Tukey test). This was done for each regulatory threshold (RT): RT02, RT015, RT01. The best parsimonious models optimized by the J and MCC statistics were the models that had the lowest number of predictors while satisfying the following conditions: $J/J_{max} > 0.95$ or $MCC/MCC_{max} > 0.95$.

3.4 Ranking cultivars based on their grain Cd contents

The cultivars were ranked based on the random effects of the MELR best models at RT01 on the one hand, and from the field trials on the other. For field trials, grain Cd content was modeled as a function of the cultivar using a mixed-effects model with the year x location as a random effect for the intercept. The adjusted means of the grain Cd of the cultivars were grouped using the Tukey's test at p=0.05, adjusted for multiple comparisons. For the cultivars that were the same in the modeling and in the field trials, the ranking of the two approaches was compared to examine consistency. The statistical models used in this work never accounted for an interaction between the cultivars (genotypes) and the environment $i.e$ the soil variables for the modeling and the year x locations for the field trials. The interactions were tested but were not significant or led to over-parameterization of the mixed-effect models.

3.5 Characterization of the effects of soil variables and of the cultivar based on the conformity of grain Cd content

Based on the performances of the models we tested (detailed in the results section), the random forest approach where the cutoff is optimized from the MCC statistics was used to simulate the conformity of the grain Cd content from the 7 soil variables and for the cultivar "Relief", which was classified as the highest Cd accumulator and for the cultivar "Miradoux", which was classified as the lowest accumulator (ranking is detailed in the results section). A factorial experimental design crossed 6 values chosen to cover realistic ranges for each of the 7 soil variables. For a given value of a given predictor $X_i = x_{ij}$, $i \in \{1, ...7\}$, $j \in \{1, ...6\}$, the probability of non-conformity $p(Y = 1|X_i = x_{ij})$ was the frequency of the non-conformity class among the $6^6 = 46656$ predictions where $X_i = x_{ij}$ and $X_k = x_{kj}$, $k \in \{1, ...7\}$, $k \neq i$. It was therefore not possible to directly derive a standard deviation for this probability. The effect of each predictor on the probability of non-conformity was characterized for the two cultivars and for the three RTs.

3.6 Calculations and statistical software

All the data processing and statistical analyses were performed with R, version 3.6.2 (R Code Team, 2019). The following specific packages were used: *randomForest* (version 4.6-14) for the random forest modeling, *lme4* (version 1.1-21) to fit the mixed effects models MELR and MEMR and the *doParallel* (version 1.0.15) for parallel computing. The residuals of MEMR were examined to check for heteroskedasticity and non-normality. When detected, heteroskedasticity was corrected by modeling the variance with the *varPower*() function of the *nlme* package (version 3.1-143)

4 Results

4.1 Descriptive statistics of the modeling database

The ranges of variation in soil characteristics were typical of agricultural soils (Table 1). The median soil Cd content was 0.26 mg kg^{-1}, slightly higher than the French national median of 0.20 mg Cd kg^{-1} for agricultural soils (calculated from Saby et al., 2009), the contamination of which mainly originates from pedogenesis. Ninety-eight percents of soil samples had Cd contents below or equal to 1 mg Cd kg^{-1} (not shown). Most of the soils were alkaline (median pH=8.2) and only 10 had a pH of between 5.5 and 7. The soils with a pH below 7.5 did not have a significant calcareous content (Fig. SI1 in supplementary information). SOC and soil Cd were significantly correlated (Fig. SI1, $R^2 = 0.40$, $p < 0.001$, not shown). The *log-log* linear relationship between these two variables indicates that doubling the SOC is expected to correspond to a soil Cd content increased by a factor x1.36. For the French soil survey database limited to agricultural soils, the relationship is also significant but much weaker ($R^2 = 0.10$, $p < 0.001$, $n = 5201$), and the magnitude of increase is x1.12 (calculated from Saby et al., 2009). The soils in the modeling database correctly reflect the fact that in France, durum wheat is mainly cultivated in alkaline and calcareous soils in the southern-half of the country (FranceAgrimer, 2020). As evidenced by a national soil survey, these soils are slightly richer in Cd, which is partly because the calcareous of these soils are naturally rich in Cd (Saby et al., 2009, 2011).

There were 36 distinct cultivars in the modeling database, each with between 1 to 85 observations (Table 2). The cultivars with few observations were kept because they helped to estimate the fixed effects of soil variables on grain Cd conformity. By contrast, a lack of observations makes it difficult to estimates the random effect of the cultivar on the intercept of the model.

As expected, the frequency of samples that did not comply with the regulation increased with decreasing RT, namely 18% for RT02, 26% for RT015 and 48% for RT01 (Table 2). These values were higher than the estimates made in a larger French national survey: 4%, 8% and 22%, respectively. However, these values were in the range of some production areas where the phytoavailability of Cd is higher than average, for instance in the center of France (ARVALIS-Institut du végétal and France-AgriMer, unpublished statistics). Whatever the RT, there were marked variations in the frequency of non-conformity between the high Cd cultivars (Table 2).

4.2 Model performances resulting from the K=5-fold cross-validation

For RT02 and RT015, random forest (RF) was the most efficient modeling approach, followed by logistic regression (MELR) and then by Gaussian regression (MEMR) (Fig. 1). This ranking was reversed for RT01. For the three regulatory thresholds (RT02, RT015, RT01) and for both performance criteria (the Youden index J and Matthews correlation coefficient MCC), the best models in the absolute had between 5 and 7 soil predictors, and the most parsimonious ones had between 2 and 5 (Fig.1, Table 3). Regarding the parsimonious models, the number of predictors increased with lowering of the RT, indicating that more information is required to correctly predict low RTs. The parsimonious models always produced significantly lower performances than the best models, but less than 5% by definition (see materials and methods). By combining the modeling approaches and the set of predictors, it was possible to obtain models that have very good performances. Between 79% and 85% of the cases of non-conformity were detected, the sensitivity slightly decreasing with decreasing RT (True Positive Rate, TPR, Table 2). The reliability of the model predictions (Positive Predictive Value, PPV) was barely lower, between 67% and 82 % success. Reliability increased with lowering of the RT, in contrast to sensitivity. This revealed a trade-off between the detection capacity and the reliability of predictions. Concerning cases of conformity, the model performances were a little better than for non-conformity: between 83% and 92 % of successfuldetection (True Negative Rate, TNR,

Table 3) and between 81% and 97% for reliability (Negative Predictive Value, NPV), both decreased with lowering of theRT. Compared to the optimization of the models using the J statistics, optimization using the MCC increased the PPV (reliability) by around 2% at the expense of a decrease in the TPR (sensitivity) also of around 2%.

It should be noted that the reliability performances of the models depends on the actual frequencies of samples that do not comply with the RT. Hence the PPV and NPV given in Table 3 are conditioned by the percentages of samples above RT02, RT015 and RT01 in our database (18%, 26% and 48%, respectively). If the predictions are required in a context in which the frequency of non-conformity (prevalence: p) differs from that in the database used to train the models, the PPV and NPV must be corrected as follows:

$$PPV^* = \frac{pTPR}{pTPR + (1-p)(1-TNR)} \tag{5}$$

$$NPV^* = \frac{(1-p)TNR}{(1-p)TNR + p(1-TPR)} \tag{6}$$

Hence, in a prediction context where the prevalence could be lower than that of the modeling database, the reliability of the models will be reduced and conversely. TPR and TNR do not depend on prevalence.

4.3 Ranking of predictors and analysis of their effects

Fig. SI2 shows the estimated importance of the predictors for the random forest model with 7 soil variables and the cultivars. On average, the importance of the predictors increases with lowering of the RT. The most influential predictors are soil Cd and pH, the least, the cultivar, and in an intermediate position, clay, fine silt (FS) and coarse silt (CS). The relative rank of SOC and of calcareous varied greatly depending on the RT, likely because they co-varied with soil Cd and pH, respectively.

Fig. 2 shows the predicted frequencies of non-conformity as a function of the 5 most important predictors, for the three RT and for the lowest (Miradoux) and highest (Relief) Cd accumulator cultivar. The model predictions demonstrate the strong effect of soil Cd and pH followed by that of clay. The risk of non-conformity increases with soil Cd with an 'S-shaped' response.

The effect increases with lowering of the RT. For the highest soil Cd contents, depending on the cultivar, the risk of non-conformity for RT02 is low to moderate whereas it is always certain at RT01. In the case of low soil Cd, the lag phase of the risk is severely reduced at RT01. The differences between the two cultivars is predicted to be weak at low Cd at RT02, at low and high Cd at RT01 and irrespective of soil Cd at RT015.

Concerning soil pH, on average, the risk was also predicted to increase with lowering of the RT. The models predicted a slight linear decrease in the risk at RT01 with little differences between cultivars. In contrasts, at RT02 and RT015, the risk was predicted to be maximum when the pH is below 6.5 with marked differences between the two cultivars, and then, to strongly decrease between pH=6.5 and pH=7 with few changes at higher pH. The mean predicted risk of non-conformity was mapped for different combinations of soil Cd and pH (Fig. 3). At RT02, estimated risk was always low for Miradoux ($p<0.2$) whereas it was more than 40% ($p>0.4$) for soil Cd > 1.25 mg kg^{-1} and pH<6.5 for Relief. With lowering of the RT, the soil Cd and pH area of high risk logically increase. At the same time, the differences between the two cultivars decrease considerably and at RT01, even for Miradoux, the lowest Cd accumulator cultivar, the safe area was very narrow. Finally, the model predicted that at RT01, at soil Cd>0.2 mg kg^{-1}, even at high soil pH, the risk of non-conformity could be high.

For both cultivars, SOC was predicted to have little effect at RT02 and RT015 but at RT01, increasing SOC is predicted to reduce the risk of non-conformity by around 40%, more markedly below 20 g C/kg (Fig. 2).

The models did not predict marked effect of clay at RT01 (Fig. 2). At RT015, the risk was predicted to first decrease at between 50-400 g clay kg^{-1} soil and to increase with higher contents with marked differences between cultivars. At RT02, an increase above 400 g clay kg^{-1} soil was also predicted. The effects of fine silt were similar but less than the effect of clay.

The random effects for the intercept of the model fitted to RT01 data allowed us to rank the cultivars in the modeling database (Fig. 4) showing a x1.4 factor of variation in grain Cd. On the other hand, field trials allowed us to establish groups of sensitivity to grain Cd accumulation in another set of cultivars with a x2.9 factor of variation between the two extreme groups (Fig. 5). As shown in Fig. 6, for the cultivars that were used in the two approaches, ranking was consistent except for Babylone and Sculptur, which were, respectively more strongly over and under classified by the models.

5 Discussion

5.1 The choice of the modeling approach depends on the regulatory threshold

At the highest RTs (RT02 and RT015), the models concentrate on the highest phytoavailability of Cd in the soil, which, in these cases, is mainly controlled by the soil Cd and pH (Fig.2 and 3, and see the parsimonious models, Table 3). As shown in Fig.2, at these RTs, the shape of the effects of soil Cd and pH cannot be completely modeled by the power mathematical model of the MELR and MEMR approaches. Because non-parametric classification trees are more flexible for modeling complex non-linear relationships, RF models consequently performed slightly better (Fig. 1 and Table 2) as it has also been reported in other studies (Covelo et al., 2008; Qiu et al., 2016). The use of classification trees to model complex and non monotonic responses (in this case, the boundary between non-conformity and conformity of the grain Cd content) increases the risk of obtaining over-fitted models with high variance. This pitfall is counteracted by using the RF approach, which aggregates the predictions of a large number of relatively uncorrelated trees (Breiman, 2001). In this way, the errors of prediction of some trees, in particular those due to over-fitting, are offset by the remaining good predictions. This is likely the reason why RF performed better than the logistic regression at RT02 and RT015.

The cutoff optimization is one likely reason why RF and MELR performed better than MEMR at RT02 and RT015. At these two RTs, the actual frequencies of non-conformity are 18% and 26%, far from 50% (Table 2). This imbalance between the conformity and non-conformity classes biases the models if a default cutoff probability of 0.5 is used and this is the reason why the bias is corrected by optimizing the cutoff (Kuhn and Johnson, 2013), as done in the RF and MELR models. By contrast, the grain Cd content predicted by the MEMR was directly transformed into conformity classes depending on whether it was above or below the RT. This is another possible explanation why MEMR performances were clearly the worst at at RT02 and RT015 (Fig. 1). At RT01, as the actual frequency of non-conformity was 48%, cutoff optimization was less necessary. Furthermore, this higher frequency provides much more information to model the boundary between conformity and non-conformity with a parametric model, explaining why MEMR became the most efficient approach (Fig. 1)

5.2 Sensitive and precise models for field prediction of durum wheat grain Cd conformity using soil variables

Our work shows that it is possible to predict the conformity of durum wheat grain Cd content with respect to the regulatory thresholds of 0.2, 0.15 and 0.1 mg Cd kg^{-1} with high sensitivity (probability of detection) and high precision and consequently reliability (probability that a prediction is true). These models are of practical value for farmers because they only require variables already determined in soil testings plus the total soil Cd. Total soil Cd is rarely measured in France and we therefore recommend it is systematically included in all future soil analyses. On one hand, it would make it possible to use the models built in this work to predict possible risky situations and on the other hand, it would help monitor the background level of soil Cd to study in more detail should it increase as a result of agricultural practices, including fertilization with contaminated P fertilizers (Sterckeman et al., 2018; Six and Smolders, 2014). The trend in soil Cd in agricultural soil is a serious concern as shown by our results: at RT01 and for sensitive cultivars, the models predict a strong risk of non-conformity for soil Cd above 0.3 mg kg^{-1} even at high soil pH (Fig. 2 and 3). The combined analysis of soil and grain Cd contents is also advisable because the models can learn and improve from new data. The very good performances of the models at RTs ranging from 0.1 to 0.2 mg Cd kg^{-1} suggest that lowering the current regulatory threshold of 0.2 mg Cd kg^{-1} grain would not be an obstacle to the reliable detection of the great majority of cases of non-conformity. The models will also help identify possible solutions in the case of a predicted non-conformity. Mapping the risky contexts for soil variables (Fig. 2 and 3) would help decide if it is worth taking action on the phytoavailability of Cd by shifting cultivation to another location or by increasing soil pH, for instance. Ranking cultivars is also a valuable model to reduce the risk of non-conformity of grain Cd content. The models can easily be re-calibrated for a new RT, not tested in this study. If the RT is strongly revised downwards to RT01, the global performances of the models will be reduced (see *MCC* index, Fig. 1). The models will be a little less sensitive in the detection of both conformity and non-conformity cases (Table 3). As a counterpart, their reliability will increase for non-conformity but not for conformity.

From a practical point of view, the choice of the right model can be adjusted depending on priorities. If the cost of a prediction error is high, the model should be chosen to maximize reliability and therefore, the models optimized from the *MCC* statistics should be preferred (Table 3). On the other hand, if the priority is to maximize the detection of cases of non-conformity at the risk of overestimating the latter, the models optimized by the *J* index should be chosen. This decision rule does not concern MEMR models for RT01, for which the prediction class does not rely on an optimized cutoff.

5.3 Phytoavailability of Cd in soil versus between-cultivar variability to manage the conformity of grain Cd content

The marked effect of soil Cd observed in this study is due to the fact that in non-polluted agricultural soils, the Cd^{2+} concentration in the soil solution is generally low, around 0.1-1 nM (Schneider et al., 2019; Sauvé et al., 2000), compared to the capacity of roots to take up the metal (Lux et al., 2011). Therefore, the factor that limits uptake is generally the supply of Cd^{2+} to the root surface (Lin et al., 2016). The latter is mainly controlled by the pool of soil Cd that can be exchanged with the solution and by the speciation of soluble Cd. The strong effect of soil pH modeled in our work illustrates the competition between H and Cd for sorption sites and for association with soluble ligands. The threshold for the pH effect of 6.5-7.0 is unlikely to be due to the formation of complexes between Cd and OH^- or carbonates according to the stability constants of these compounds reported in the chemical databases (Tipping et al., 2011). The 6.5-7.0 threshold could correspond to several mechanisms of Cd sorption. Regarding organic matter, based on the mean *log* of the dissociation constant of the proton ($log\, K_H$) for humic substances (Tipping et al., 2011; Matynia et al., 2010), the 6.5-7.0 threshold could reflect the binding of Cd to OH groups that become increasingly deprotonated at high pH. The 6.5-7.0 threshold could also be due to the favored binding of Cd to variable charge sites of Fe, Mn and Al oxyhydroxides and organo-mineral complexes associated with clay (Violante et al., 2010; Rasmussen et al., 2018). The pH of most French soils are above 6.5-7.0, especially for durum wheat, which is frequently grown on calcareous soils. Hence, because the models predict little effect of pH above 7, (Fig. 2), this soil variable is not an important lever to manage the grain

Cd content in French durum wheat. On the other hand, because for pH above 7, Fe, Mn and Al oxides are expected to have an increasing role in binding Cd, their contents should be included in soil testing to be able to possibly improve the models.

SOC is generally found to strongly control Cd phytoavailability because it includes both solid and dissolved organic matter that sorbs and forms complexes with Cd. In our study, the correlation between Cd and SOC partly masked the effect of SOC but in agreement with the literature, the models predicted that SOC is expected to reduce the phytoavailability of Cd, particularly if the latter is low on average (RT01, Fig. 2). Adding organic matter to soil is encouraged for many reasons including improving fertility and storing carbon. Based on our study, this lever is also questionable because the effect is estimated to be moderate and considering the soil Cd-SOC correlation in our database, one may wonder if organic matter does not increase soil Cd, due to its own contamination or by sequestrating Cd deposited in soil by atmospheric fallouts or by agricultural inputs such as P fertilizers.

Finally, clay and to a lesser extent fine silts were predicted to have a moderate effect on the predicted conformity of grain Cd, but only for the highest RTs. Clay and FS are involved in the reversible exchanges of Cd with the solution (buffer capacity). On one hand clay and FS are expected to reduce the background concentration of soluble Cd due to sorption but on the other hand, they facilitate the buffering of this concentration when the roots take up the Cd. However, these antagonistic effects are unlikely to explain the slight negative effect of clay<400 g kg^{-1} soil and the positive effect above this threshold. The buffer capacity of the solid phase for Cd is all the more involved as the phytoavailability of Cd is low and therefore, the effects of clay and FS should be stronger at RT01 compared to RT02, unlike in the simulations. The effect of clay and to a lesser extent of FS are probably due to a higher abundance of soil rich in Cd for soils with high clay contents, as shown by Fig. SI1.

It is noteworthy that our models produced good performances although they use the total Cd content of soils and not the mobile Cd, suggesting that the relationship between the two pools was strong enough to allow correct prediction of the conformity of grain Cd. Furthermore, the error derived from approximating the mobile Cd by the total Cd is expected to have less negative impact when predicting a grain Cd binary class than when predicting grain Cd content.

Modeling durum wheat grain Cd showed that soil variables governing Cd phytoavailability have much more influence than the type of cultivar (Fig. SI2). This confirms previous observations (Li et al., 1997; Li and Zhou, 2019) and was probably more apparent in our work since most of the cultivars we investigated did not have the *Cdu1* low accumulation allele, thus reducing between-cultivar variability. Hence, the model estimated a x1.4 factor between the lowest and highest Cd accumulator cultivar whereas field trials, which included at least one cultivar with the *Cdu1* low Cd allele (Anvergur) showed a x3 factor of variation. The x1.4 factor of variation predicted by the models for cultivars without the low-Cd allele of the *Cdu1* gene suggests that there is also variability in some mechanisms contributing to reduce grain Cd, other than the enhanced sequestration of Cd in roots. For example, modification of the rhizosphere, including changes in pH and the release of ligands such as low molecular weight organic acids by roots can significantly differ between wheat cultivars (Cieśliński et al., 1998; Greger and Landberg, 2008). Differences in Cd sequestration in the stem and nodes of different rice cultivars has also been observed (Fujimaki et al., 2010). It was shown in sunflower and wheat, that the partitioning of plant biomass, especially aboveground biomass and plant height could partly explain the intraspecific variability in the grain Cd content (Laporte et al., 2015; Pozniak et al., 2012; Perrier et al., 2016; Álvaro et al., 2008). Hence, understanding the reasons for the variability in grain Cd contents among high Cd cultivars merits further investigations, in particular because reducing the phytoavailability of Cd in soil is not an easy task.

6 Conclusions

This work confirms the marked influence of soil Cd and pH on the transfer of Cd to durum wheat grain. For this crop, we have shown that grain Cd can exceed the current and possible lower future regulatory thresholds even in alkaline soils with moderate total Cd contents below 1 mg Cd kg^{-1} soil. Combining random forest, multiple logistic and Gaussian linear regressions enabled us to build models to predict grain Cd conformity that are both efficient and reliable. The model performances are also very good for the lowest RT that were considered by the downwards revision of the regulatory threshold. By adjusting the predicted grain Cd using the phytoavailability of Cd estimated from soil variables, the models also showed there was a x1.4 factor of variation between durum wheat cultivars that did not have the low Cd allele of the *Cdu1* gene. Because the models only require variables available from soil analyses, they are valuable tools to be able to predict possible problems of non-conformity of durum wheat production with respect to the regulation concerning Cd in foodstuffs. Further, considering the current trend towards a more restrictive regulation for food products, the approach used in this work can also be used for other heavy metals such as Ni which is currently targeted in relation with the baby foods.

7 Acknowledgments

The authors are grateful to S. Thunot, T. Robert and B. Orlando for their technical assistance. They are also grateful to FranceAgriMer and to all farmers involved in this study for their commitment. This work was funded by ARVALIS-Institut du végétal (Cadur project CR127518) and by the Agence Nationale de la Recherche (ANR 15-CE21-0001-04, Cadon). The authors declare that they have no conflict of interest.

References

AbuHammad, W. A., Mamidi, S., Kumar, A., Pirseyedi, S., Manthey, F. A., Kianian, S. F., Alamri, M. S., Mergoum, M., and Elias, E. M. (2016). Identification and validation of a major cadmium accumulation locus and closely associated snp markers in north dakota durum wheat cultivars. *Molecular Breeding*, 36(8):112.

Adams, M., Zhao, F., McGrath, S., Nicholson, F., and Chambers, B. (2004). Predicting cadmium concentrations in wheat and barley grain using soil properties rid a-8339-2008 rid b-5127-2008. *Journal of Environmental Quality*, 33(2):532–541. WOS:000220292800011.

Alexander, J., Benford, D., Cockburn, A., Cravedi, J. P., Dogliotti, E., Di Domenico, A., Férnandez-Cruz, M. L., Fürst, P., Fink-Gremmels, J., Galli, C. L., et al. (2009). Cadmium in food. scientific opinion of the panel on contaminants in the food chain. *The EFSA Journal*, 980:1–139.

Antoniadis, V., Levizou, E., Shaheen, S. M., Ok, Y. S., Sebastian, A., Baum, C., Prasad, M. N. V., Wenzel, W. W., and Rinklebe, J. (2017). Trace elements in the soil-plant interface: Phytoavailability, translocation, and phytoremediation–a review. *Earth-Science Reviews*, 171:621–645.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Cieśliński, G., Van Rees, K., Szmigielska, A., Krishnamurti, G., and Huang, P. (1998). Low-molecular-weight organic acids in rhizosphere soils of durum wheat and their effect on cadmium bioaccumulation. *Plant and Soil*, 203(1):109–117.

Clarke, J., Clarke, F., and Pozniak, C. (2010). Forty-six years of genetic improvement in canadian durum wheat cultivars. *Canadian Journal of Plant Science*, 90(6):791–801.

Clemens, S. (2019). Safer food through plant science: reducing toxic element accumulation in crops. *Journal of Experimental Botany*, 70(20):5537–5557.

Clemens, S., Aarts, M. G., Thomine, S., and Verbruggen, N. (2013). Plant science: the key to preventing slow cadmium poisoning. *Trends in Plant Science*, 18(2):92–99.

Covelo, E. F., Matías, J. M., Vega, F. A., Reigosa, M. J., and Andrade, M. L. (2008). A tree regression analysis of factors determining the sorption and retention of heavy metals by soil. *Geoderma*, 147(1):75–85.

de Vries, W., McLaughlin, M. J., and Groenenberg, J. E. (2011). Transfer functions for solid-solution partitioning of cadmium for australian soils. *Environmental Pollution*, 159(12):3583–3594. WOS:000296397300041.

DGSANCO (2011). Amendement of regulation (ec) n° 1881/2006 as regards cadmium.

EC 488/2014 (2014). Journal officiel de l'union européenne.

EC466/2001 (2001). Commission regulation (ec) no 466/2001 of 8 march 2001 setting maximum levels for certain contaminants in foodstuffs (text with eea relevance.).

FranceAgrimer (2020). Marché du blé dur. france, union européenne, monde (2019-2020). Technical Report ISSN 1779-353X.

Fujimaki, S., Suzui, N., Ishioka, N. S., Kawachi, N., Ito, S., Chino, M., and Nakamura, S.-I. (2010). Tracing cadmium from culture to spikelet: Noninvasive imaging and quantitative characterization of absorption, transport, and accumulation of cadmium in an intact rice plant. *Plant Physiology*, 152(4):1796–1806.

Greger, M. and Landberg, T. (2008). Role of rhizosphere mechanisms in cd uptake by various wheat cultivars. *Plant and Soil*, 312(1):195–205.

Greger, M. and Lofstedt, M. (2004). Comparison of uptake and distribution of cadmium in different cultivars of bread and durum wheat. *Crop Science*, 44(2):501–507. WOS:000189384600019.

Groenenberg, J. E., Romkens, P. F. a. M., Comans, R. N. J., Luster, J., Pampura, T., Shotbolt, L., Tipping, E., and de Vries, W. (2010). Transfer functions for solid-solution partitioning of cadmium, copper, nickel, lead and zinc in soils: derivation of relationships for free metal ion activities and validation with independent data. *European Journal of Soil Science*, 61(1):58–73. WOS:000273454800006.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer-Verlag, New York, 2 edition.

Horn, A. L., Reiher, W., D Ring, R.-A., and G Th, S. (2006). Efficiency of pedotransfer functions describing cadmium sorption in soils. *Water, Air, and Soil Pollution*, 170(1-4):229–247.

Hough, R. L., Young, S. D., and Crout, N. M. J. (2003). Modelling of cd, cu, ni, pb and zn uptake, by winter wheat and forage maize, from a sewage disposal farm. *Soil Use and Management*, 19(1):19 –27. WOS:000181552000003.

Kabata-Pendias, A. (2004). Soil–plant transfer of trace elements—an environmental issue. *Geoderma*, 122(2):143–149.

Knox, R. E., Pozniak, C. J., Clarke, F. R., Clarke, J. M., Houshmand, S., and Singh, A. K. (2009). Chromosomal location of the cadmium uptake gene (cdu1) in durum wheat. *Genome*, 52(9):741–747. WOS:000271100900001.

Kuhn, M. and Johnson, K. (2013). Remedies for severe class imbalance. *Applied Predictive Modeling*, pages 419–443.

Laporte, M.-A., Sterckeman, T., Dauguet, S., Denaix, L., and Nguyen, C. (2015). Variability in cadmium and zinc shoot concentration in 14 cultivars of sunflower (helianthus annuus l.) as related to metal uptake and partitioning. *Environmental and Experimental Botany*, 109(0):45–53.

Li, X. and Zhou, D. (2019). A meta-analysis on phenotypic variation in cadmium accumulation of rice and wheat: Implications for food cadmium risk control. *Pedosphere*, 29(5):545–553.

Li, Y.-M., Chaney, R., Schneiter, A., Miller, J., Elias, E., and Hammond, J. (1997). Screening for low grain cadmium phenotypes in sunflower, durum wheat and flax. *Euphytica*, 94(1):23–30.

Lin, Z., Schneider, A., Sterckeman, T., and Nguyen, C. (2016). Ranking of mechanisms governing the phytoavailability of cadmium in agricultural soils using a mechanistic model. *Plant and Soil*, 399(1):89–107.

Lux, A., Martinka, M., Vaculík, M., and White, P. J. (2011). Root responses to cadmium in the rhizosphere: A review. *Journal of Experimental Botany*, 62(1):21–37.

Maccaferri, M., Harris, N. S., Twardziok, S. O., Pasam, R. K., Gundlach, H., Spannagl, M., Ormanbekova, D., Lux, T., Prade, V. M., Milner, S. G., Himmelbach, A., Mascher, M., Bagnaresi, P., Faccioli, P., Cozzi, P., Lauria, M., Lazzari, B., Stella, A., Manconi, A., Gnocchi, M., Moscatelli, M., Avni, R., Deek, J., Biyiklioglu, S., Frascaroli, E., Corneti, S., Salvi, S., Sonnante, G., Desiderio, F., Marè, C., Crosatti, C., Mica, E., Özkan, H., Kilian, B., Vita, P. D., Marone, D., Joukhadar, R., Mazzucotelli, E., Nigro, D., Gadaleta, A., Chao, S., Faris, J. D., Melo, A. T. O., Pumphrey, M., Pecchioni, N., Milanesi, L., Wiebe, K., Ens, J., MacLachlan, R. P., Clarke, J. M., Sharpe, A. G., Koh, C. S., Liang, K. Y. H., Taylor, G. J., Knox, R., Budak, H., Mastrangelo, A. M., Xu, S. S., Stein, N., Hale, I., Distelfeld, A., Hayden, M. J., Tuberosa, R., Walkowiak, S., Mayer, K. F. X., Ceriotti, A., Pozniak, C. J., and Cattivelli, L. (2019). Durum wheat genome highlights past domestication signatures and future improvement targets. *Nature Genetics*, 51(5):885–895.

Matynia, A., Lenoir, T., Causse, B., Spadini, L., Jacquet, T., and Manceau, A. (2010). Semi-empirical proton binding constants for natural organic matter. *Geochimica et Cosmochimica Acta*, 74(6):1836–1851.

Newman, M. C. (1993). Regression analysis of log-transformed data: Statistical bias and its correction. *Environmental Toxicology and Chemistry*, 12(6):1129–1133.

Oladzad-Abbasabadi, A., Kumar, A., Pirseyedi, S., Salsman, E., Dobrydina, M., Poudel, R. S., AbuHammad, W. A., Chao, S., Faris, J. D., and Elias, E. M. (2018). Identification and validation of a new source of low grain cadmium accumulation in durum wheat. *G3: Genes, Genomes, Genetics*, 8(3):923–932.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., and Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, 49(12):1373–1379.

Perrier, F., Yan, B., Candaudap, F., Pokrovsky, O. S., Gourdain, E., Meleard, B., Bussière, S., Coriou, C., Robert, T., Nguyen, C., and Cornu, J. Y. (2016). Variability in grain cadmium concentration among durum wheat cultivars: impact of aboveground biomass partitioning. *Plant and Soil*, 404(1-2):307–320.

Pozniak, C. J., Clarke, J. M., and Clarke, F. R. (2012). Potential for detection of marker–trait associations in durum wheat using unbalanced, historical phenotypic datasets. *Molecular Breeding*, 30(4):1537–1550.

Qiu, L., Wang, K., Long, W., Wang, K., Hu, W., and Amable, G. S. (2016). A comparative assessment of the influences of human impacts on soil cd concentrations based on stepwise linear regression, classification and regression tree, and random forest models. *PLOS ONE*, 11(3):e0151131.

Rasmussen, C., Heckman, K., Wieder, W. R., Keiluweit, M., Lawrence, C. R., Berhe, A. A., Blankinship, J. C., Crow, S. E., Druhan, J. L., Hicks Pries, C. E., Marin-Spiotta, E., Plante, A. F., Schädel, C., Schimel, J. P., Sierra, C. A., Thompson, A., and Wagai, R. (2018). Beyond clay: towards an improved set of variables for predicting soil organic matter content. *Biogeochemistry*, 137(3):297–306.

Saby, N., Marchant, B., Lark, R., Jolivet, C., and Arrouays, D. (2011). Robust geostatistical prediction of trace elements across france. *Geoderma*, 162(3–4):303–311.

Saby, N., Thioulouse, J., Jolivet, C., Ratié, C., Boulonne, L., Bispo, A., and Arrouays, D. (2009). Multivariate analysis of the spatial patterns of 8 trace elements using the french soil monitoring network data. *Science of The Total Environment*, 407(21):5644–5652.

Salsman, E., Kumar, A., AbuHammad, W., Abbasabadi, A. O., Dobrydina, M., Chao, S., Li, X., Manthey, F. A., and Elias, E. M. (2018). Development and validation of molecular markers for grain cadmium in durum wheat. *Molecular Breeding*, 38(3):28.

Sauvé, S., Norvell, W. A., McBride, M., and Hendershot, W. (2000). Speciation and complexation of cadmium in extracted soil solutions. *Environmental Science & Technology*, 34(2):291–296.

Schneider, A., Nguyen, V. X., Viala, Y., Violo, V., Cornu, J.-Y., Sterckeman, T., and Nguyen, C. (2019). A method to determine the soil-solution distribution coefficients and the concentrations for the free ion and the complexes of trace metals: Application to cadmium. *Geoderma*, 346:91–102.

Six, L. and Smolders, E. (2014). Future trends in soil cadmium concentration under current cadmium fluxes to european agricultural soils. *Science of The Total Environment*, 485–486(0):319–328.

Sterckeman, T., Gossiaux, L., Guimont, S., Sirguey, C., and Lin, Z. (2018). Cadmium mass balance in french soils under annual crops: Scenarios for the next century. *Science of The Total Environment*, 639:1440–1452.

Sun, F., Polizzotto, M. L., Guan, D., Wu, J., Shen, Q., Ran, W., Wang, B., and Yu, G. (2017). Exploring the interactions and binding sites between cd and functional groups in soil using two-dimensional correlation spectroscopy and synchrotron radiation based spectromicroscopies. *Journal of Hazardous Materials*, 326:18–25.

Tipping, E., Lofts, S., and Sonke, J. E. (2011). Humic ion-binding model vii: a revised parameterisation of cation-binding by humic substances. *Environmental Chemistry*, 8(3):225–235. WOS:000291943800001.

Tudoreanu, L. and Phillips, C. J. (2004). Empirical models of cadmium accumulation in maize, rye grass and soya bean plants. *Journal of the Science of Food and Agriculture*, 84(8):845–852.

Vega, F., Andrade, M., and Covelo, E. (2010). Influence of soil properties on the sorption and retention of cadmium, copper and lead, separately and together, by 20 soil horizons: Comparison of linear regression and tree regression analyses. *Journal of Hazardous Materials*, 174(1–3):522–533.

Viala, Y., Laurette, J., Denaix, L., Gourdain, E., Méléard, B., Nguyen, C., Schneider, A., and Sappin-Didier, V. (2017). Predictive statistical modelling of cadmium content in durum wheat grain based on soil parameters. *Environmental Science and Pollution Research*, 24(25):20641–20654.

Violante, A., Cozzolino, V., Perelomov, L., Caporale, A. G., and Pigna, M. (2010). Mobility and bioavailability of heavy metals and mettaloids in soil environments. *Journal of soil science and plant nutrition*, 10(3):268–292.

Wiebe, K., Harris, N., Faris, J., Clarke, J., Knox, R., Taylor, G., and Pozniak, C. (2010). Targeted mapping of *Cdu1*, a major locus regulating grain cadmium concentration in durum wheat (*Triticum turgidum* l. var *durum*). *TAG Theoretical and Applied Genetics*, 121(6):1047–1058.

Zimmerl, S., Lafferty, J., and Buerstmayr, H. (2014). Assessing diversity in triticum durum cultivars and breeding lines for high versus low cadmium content in seeds using the caps marker usw47. *Plant Breeding*, 133(6):712–717.

Álvaro, F., Isidro, J., Villegas, D., Moral, G. d., F, L., and Royo, C. (2008). Breeding effects on grain filling, biomass partitioning, and remobilization in mediterranean durum wheat all rights reserved. no part of this periodical may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. *Agronomy Journal*, 100(2):361–370.

Table 1: Characteristics of the soils used for modeling grain Cd in durum wheat. SOC : soil organic carbon, Calc: soil calcareous, Cd: total soil Cd (mg kg^{-1}). All variables but soil Cd are in g kg^{-1}. For calcareous, the values below the limit of quantification (<0.5 g kg^{-1}) were set to 0.5 g kg^{-1}.

	Clay	Fine silt	Coarse silt	Fine sand	Coarse sand	SOC	pH	Calc	Cd
Min	34	15	18	3	2	3.0	5.5	0.5	0.05
Q25	245	209	109	54	45	10.1	8.0	9.0	0.20
Median	300	258	139	122	104	13.3	8.2	106.8	0.26
Q75	381	310	202	181	180	18.1	8.4	276.5	0.36
Max	702	467	427	499	452	45.2	8.7	655.4	1.56

Table 2: Occurrence of observed non-conformity of grain Cd for the three regulatory thresholds of 0.2 (RT02), 0.15 (RT015) and 0.1 (RT01) mg Cd kg^{-1} grain and for the cultivars used for modeling. Value are in % of total data (n). *Cdu1* column indicates if the cultivar has the high Cd allele of the *Cdu1* gene or if it is unknown.

Cultivar	RT02	RT015	RT01	Cdu1	n
ACTISUR	50	50	75	High Cd	4
ALEXIS	29	43	62	High Cd	21
ATOUDUR	8	8	31	High Cd	26
AURIS	33	33	67	High Cd	6
AVENTUR	40	40	60	Unknown	5
BABYLONE	8	8	31	High Cd	13
BIENSUR	0	0	0	High Cd	2
CLOVIS	0	50	50	High Cd	2
COUSSUR	0	33	50	High Cd	6
CULTUR	38	46	62	High Cd	13
DAKTER	17	25	42	High Cd	12
DAURUR	50	50	100	High Cd	2
FABULIS	10	20	60	High Cd	10
FLORIDOU	33	33	56	Unknown	9
ISILDUR	0	8	42	High Cd	12
JOYAU	0	0	33	High Cd	3
KARUR	20	36	64	High Cd	25
LIBERDUR	0	100	100	High Cd	1
LUMINUR	62	62	75	High Cd	8
MEMODUR	100	100	100	Unknown	1
MIRADOUX	7	14	25	High Cd	85
MURANO	0	0	100	High Cd	1
NEFER	0	0	50	High Cd	2
NEMESIS	100	100	100	Unknown	1
PESCADOU	25	32	57	High Cd	28
PHARAON	0	100	100	High Cd	1
PICTUR	17	50	67	High Cd	6
QUALIDOU	20	20	40	Unknown	15
RELIEF	12	31	69	Unknown	16
SANTUR	50	50	50	Unknown	2
SCULPTUR	17	26	43	High Cd	35
SY BANCO	40	40	80	High Cd	5
SY CYSCO	20	20	53	High Cd	15
SY ENZO	0	100	100	Unknown	1
TABLUR	20	32	60	High Cd	25
YELODUR	0	0	0	High Cd	1
All	18	26	48		420

Table 3: Performances of the absolute and parsimonious best models for the three regulatory thresholds of 0.2 (RT02), 0.15 (RT015) and 0.1 (RT01) mg Cd kg^{-1} grain. FS: fine silt, CS: coarse silt, SOC : soil organic carbon, Calc: soil calcareous, RF: random forest, MELR,: mixed-effect logistic regression, MEMR: mixed-effects multi-linear regression, J: Youden statistics, MCC: Matthew's correlation coefficient, TPR: true positive rate (% of actual positive cases that are detected), TNR: true negative rate (% of actual negative cases that are detected), PPV: positive predictive value (% of predicted positive cases that are actually positive), NPV: negative predictive value (% of predicted negative cases that are actually negative)

Regulation limit limit	Model	Modeling approach	Optimization Criteria	TPR		TNR		PPV		NPV	
RT02	Cd+H+SOC+Clay+FS+CS+Calc	RF	J	85.1	(2.23)	91.0	(0.92)	67.0	(2.33)	96.6	(0.49)
RT02	Cd+H	MELR	J	83.2	(1.66)	92.0	(0.62)	69.2	(1.72)	96.2	(0.36)
RT02	Cd+H+SOC+Clay+Calc	RF	MCC	83.3	(2.37)	92.0	(0.86)	69.1	(2.37)	96.2	(0.51)
RT02	Cd+H	MELR	MCC	83.2	(1.71)	91.9	(0.63)	68.9	(1.77)	96.2	(0.37)
RT015	Cd+H+SOC+Clay+FS+Calc	RF	J	81.8	(1.93)	89.4	(1.06)	72.8	(2.06)	93.4	(0.65)
RT015	Cd+H+SOC+Clay	RF	J	81.2	(1.76)	88.9	(1.11)	71.8	(2.11)	93.2	(0.6)
RT015	Cd+H+SOC+Clay+FS+CS	RF	MCC	81.0	(1.85)	90.0	(1.06)	73.6	(2.11)	93.2	(0.62)
RT015	Cd+H+SOC+Clay	RF	MCC	81.1	(1.8)	88.9	(1.07)	71.7	(1.99)	93.2	(0.61)
RT01	Cd+H+SOC+Clay+FS+CS+Calc	MEMR		78.7	(0.54)	84.9	(0.63)	82.4	(0.6)	81.6	(0.38)
RT01	Cd+H+SOC+Clay+FS	MEMR		78.4	(0.83)	83.5	(0.78)	81.0	(0.74)	81.2	(0.6)

1 Figure captions

Figure 1: Performances of the 9 models (y-axis) x 3 modeling approaches (different colors) for the three regulatory thresholds of 0.2 (RT02), 0.15 (RT015) and 0.1 (RT01) mg Cd kg^{-1} grain. RF: random forest, MELR: mixed-effect logistic regression, MEMR: mixed-effect multi-linear regression. The model performances are expressed by the Youden statistics (J) or Matthew's correlation coefficient (MCC), which have both 1 as maximum value for perfect models. From right to left, the dashed lines correspond to the maximum J or MCC values (best model in absolute) and to 95% of this maximum. Letters on the right of the bars are the mean grouping by the Tukey test at $p<0.05$. The orange letter is the best parsimonious model, namely the model with the least predictors while having performances greater that 95% of the best model in absolute (see materials and methods).

Figure 2: Simulations of the effect of individual soil variables on the probability of non-conformity of durum wheat grain for the three regulatory thresholds of 0.2 (RT02), 0.15 (RT015) and 0.1 (RT01) mg Cd kg^{-1} grain and for the highest (Relief) and the lowest (Miradoux) Cd accumulator cultivar of the database. The simulations were obtained from the random forest model with the following soil predictors : total soil Cd, soil pH, soil organic carbon, clay, fine and coarse silt and soil calcareous. For a given value of a soil predictor, the graphs show the frequency of non-conformity for all predictions when the other predictors vary based on a factorial design (see materials and methods).

Figure 3: Simulations of the effect of soil Cd and pH on the probability of non-conformity of durum wheat grain for the three regulatory thresholds of 0.2 (RT02), 0.15 (RT015) and 0.1 (RT01) mg Cd kg^{-1} grain and for the highest (Relief) and the lowest (Miradoux) Cd accumulator cultivar of the database. The simulations were obtained from the random forest model with the following soil predictor : total soil Cd, soil pH, soil organic carbon, clay, fine and coarse silt and soil calcareous. For a given value of a soil predictor, the graphs show the frequency of non-conformity for all predictions when the other predictors vary based on a factorial design (see materials and methods).

Figure 4: Boxplots of the predicted grain Cd content of the different cultivars of the modeling database for the regulatory thresholds of 0.1 mg Cd kg^{-1}(RT01) and by using the mixed-effects

multi-linear regression with the following soil predictors : total soil Cd, soil pH, soil organic carbon, clay, fine and coarse silt and soil calcareous. The model predicts the grain Cd content when the predictor values are set to their mean in the database and for 400 repetitions of the 5 folds cross-validation. Points are outliers extending outside 1.5 x the inter-quartile range (whiskers).

Figure 5: Adjusted least squared means $\pm$ one standard deviation for the grain Cd contents of some cultivars grown in three field trials for three years. Letters and colors correspond to mean grouping by the Tukey test at $p<0.05$. The trial location and year were considered as random effects whereas the cultivar was the fixed effect.

Figure 6: Ranking of some durum wheat cultivars for their grain Cd content by two approaches : data collected from field trials (y axis) and ranking from the mixed-effects multi-linear regression with the following soil predictors : total soil Cd, soil pH, soil organic carbon, clay, fine and coarse silt and soil calcareous. High rank values indicate high grain Cd as illustrated by the 1:1 arrow.

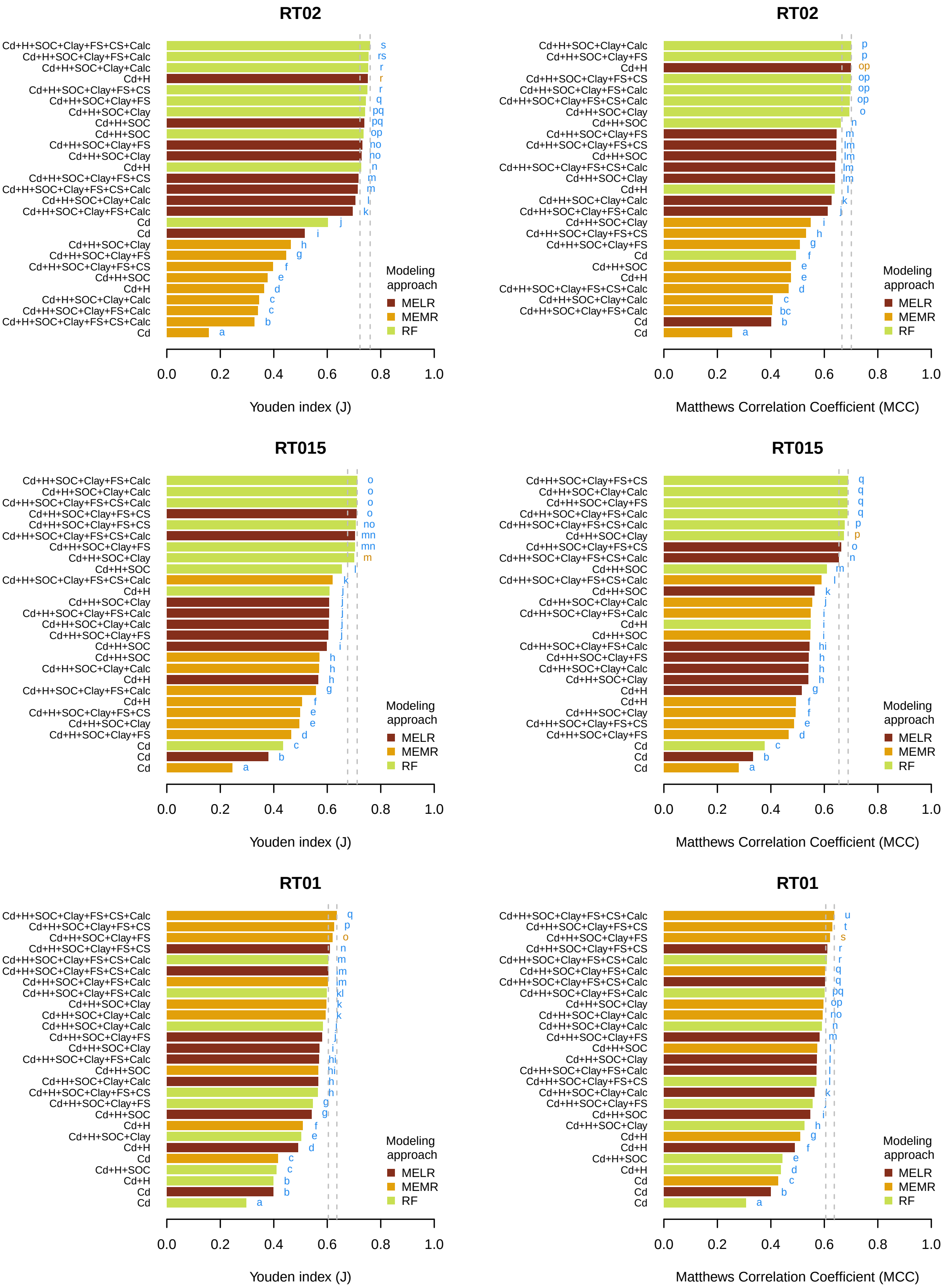

RT02
Youden index (J)
Matthews Correlation Coefficient (MCC)
RT015
Youden index (J)
Matthews Correlation Coefficient (MCC)
RT01
Youden index (J)
Matthews Correlation Coefficient (MCC)
0.0 0.2 0.4 0.6 0.8 1.0
Modeling approach
MELR
MEMR
RF

Figure 1:

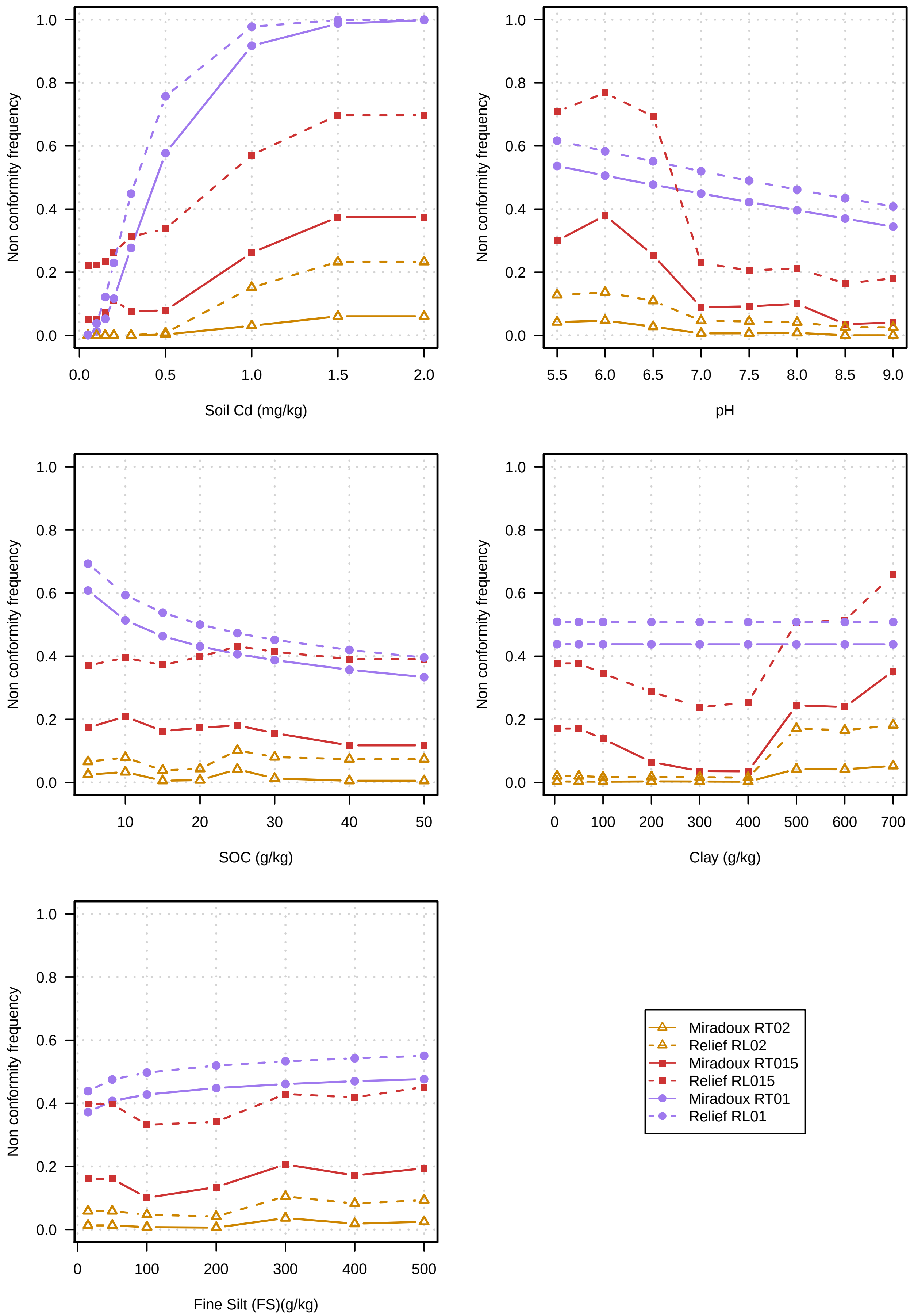

Non conformity frequency
Soil Cd (mg/kg)
pH
SOC (g/kg)
Clay (g/kg)
Fine Silt (FS)(g/kg)
Miradoux RT02
Relief RL02
Miradoux RT015
Relief RL015
Miradoux RT01
Relief RL01

Figure 2:

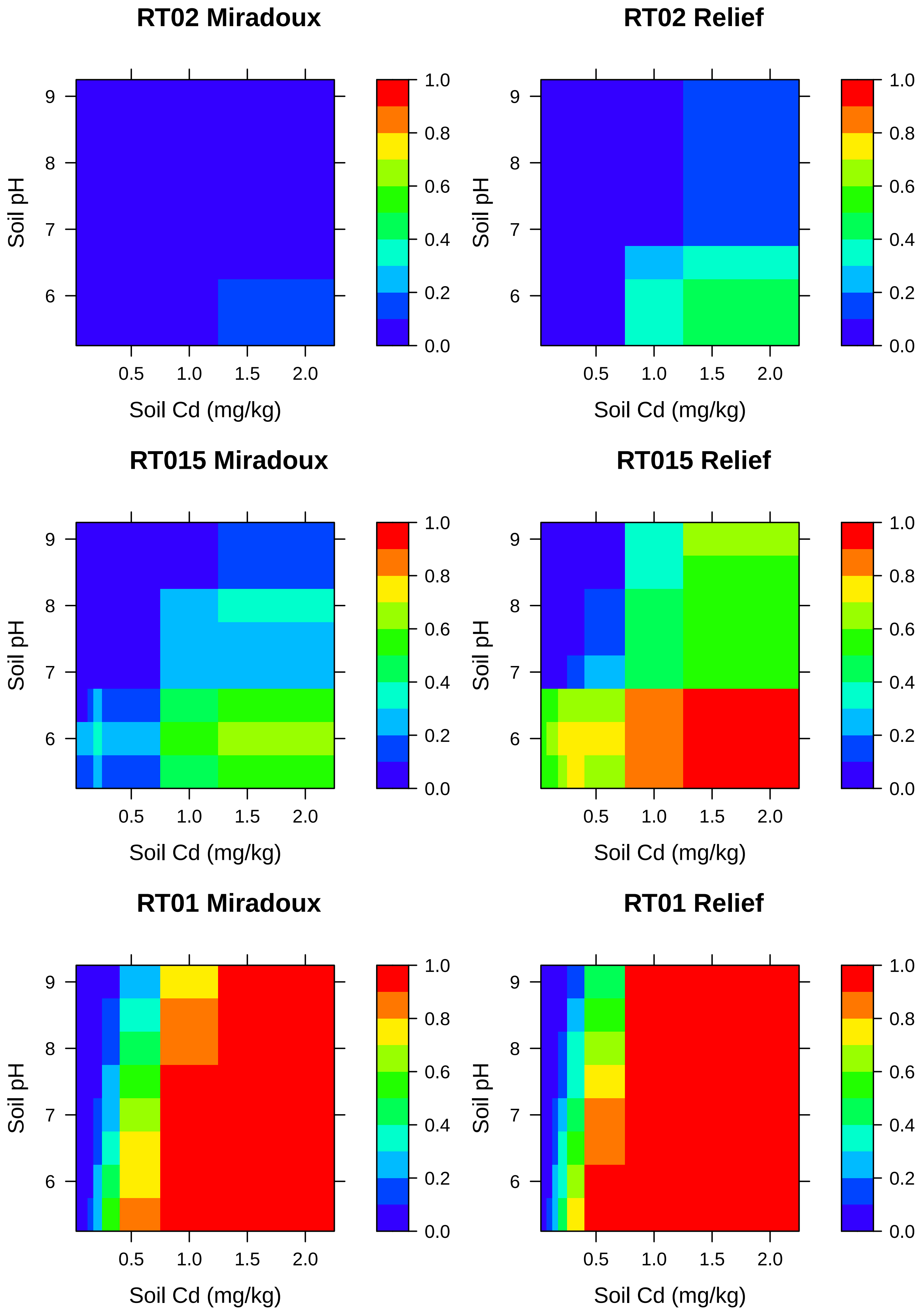
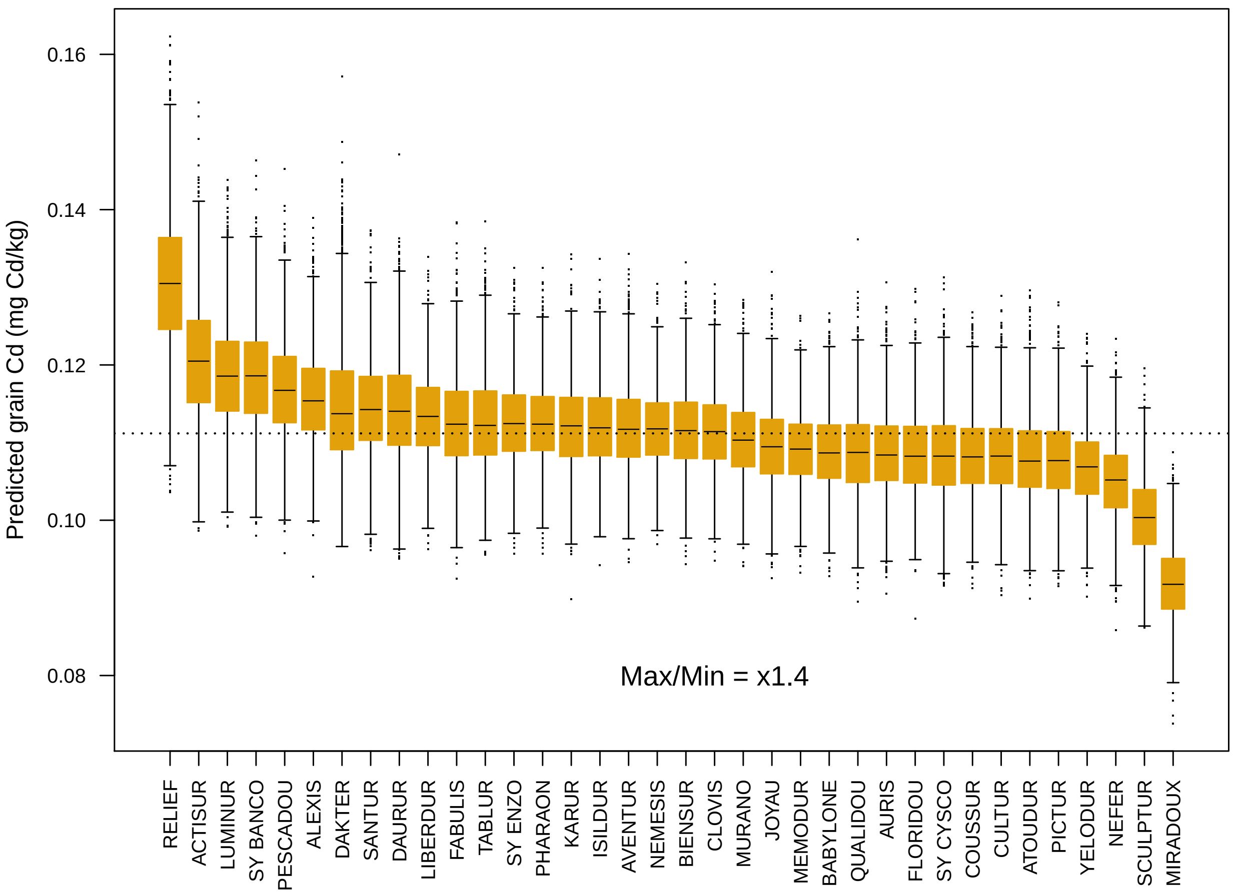

RT02 Miradoux
RT02 Relief
RT015 Miradoux
RT015 Relief
RT01 Miradoux
RT01 Relief
Soil pH
Soil Cd (mg/kg)
9
8
7
6
0.5
1.0
1.5
2.0
0.0
0.2
0.4
0.6
0.8
1.0

Figure 3:



Predicted grain Cd (mg Cd/kg)
0.16
0.14
0.12
0.10
0.08
Max/Min = x1.4
RELIEF
ACTISUR
LUMINUR
SY BANCO
PESCADOU
ALEXIS
DAKTER
SANTUR
DAURUR
LIBERDUR
FABULIS
TABLUR
SY ENZO
PHARAON
KARUR
ISILDUR
AVENTUR
NEMESIS
BIENSUR
CLOVIS
MURANO
JOYAU
MEMODUR
BABYLONE
QUALIDOU
AURIS
FLORIDOU
SY CYSCO
COUSSUR
CULTUR
ATOUDUR
PICTUR
YELODUR
NEFER
SCULPTUR
MIRADOUX

Figure 4:

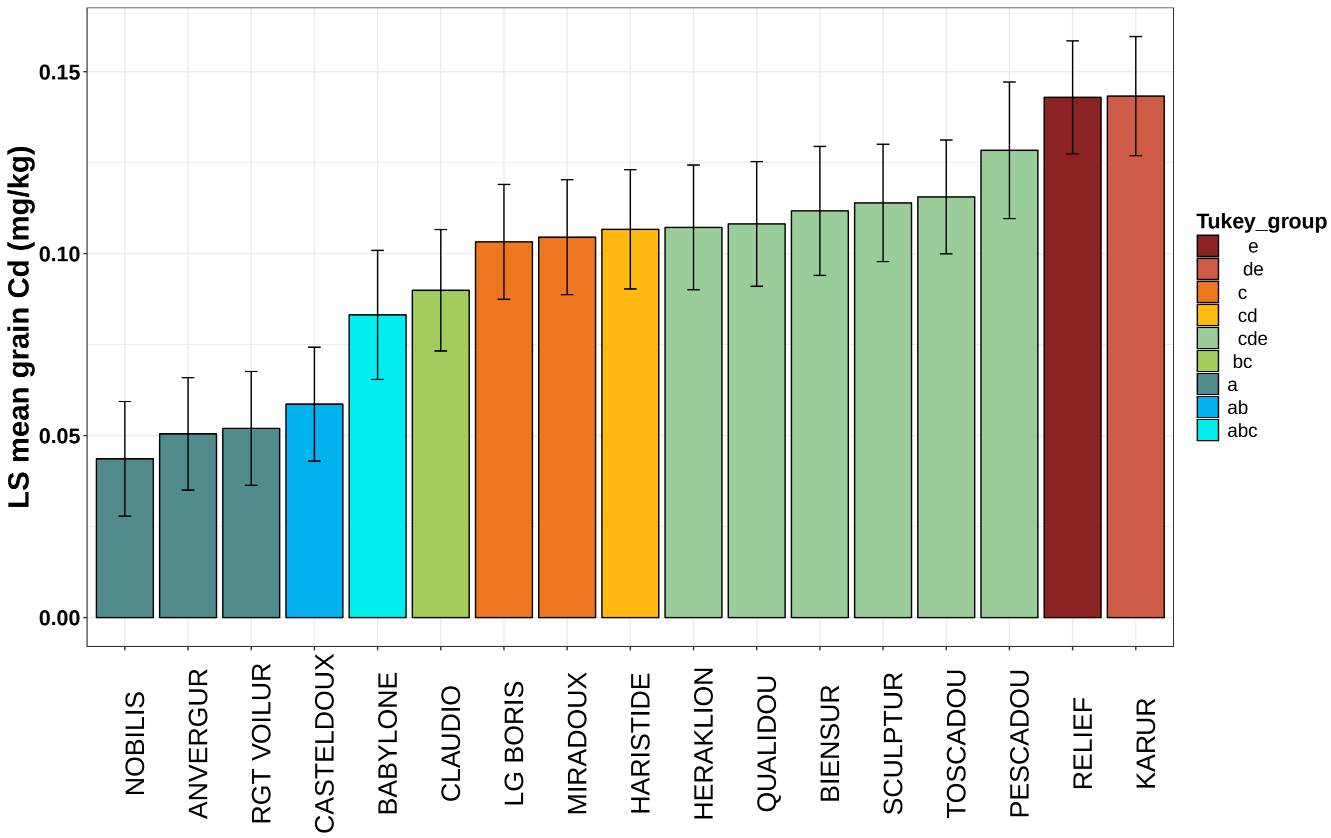

LS mean grain Cd (mg/kg)
0.15
0.10
0.05
0.00
NOBILIS
ANVERGUR
RGT VOILUR
CASTELDOUX
BABYLONE
CLAUDIO
LG BORIS
MIRADOUX
HARISTIDE
HERAKLION
QUALIDOU
BIENSUR
SCULPTUR
TOSCADOU
PESCADOU
RELIEF
KARUR
Tukey_group
e
de
c
cd
cde
bc
a
ab
abc

Figure 5:

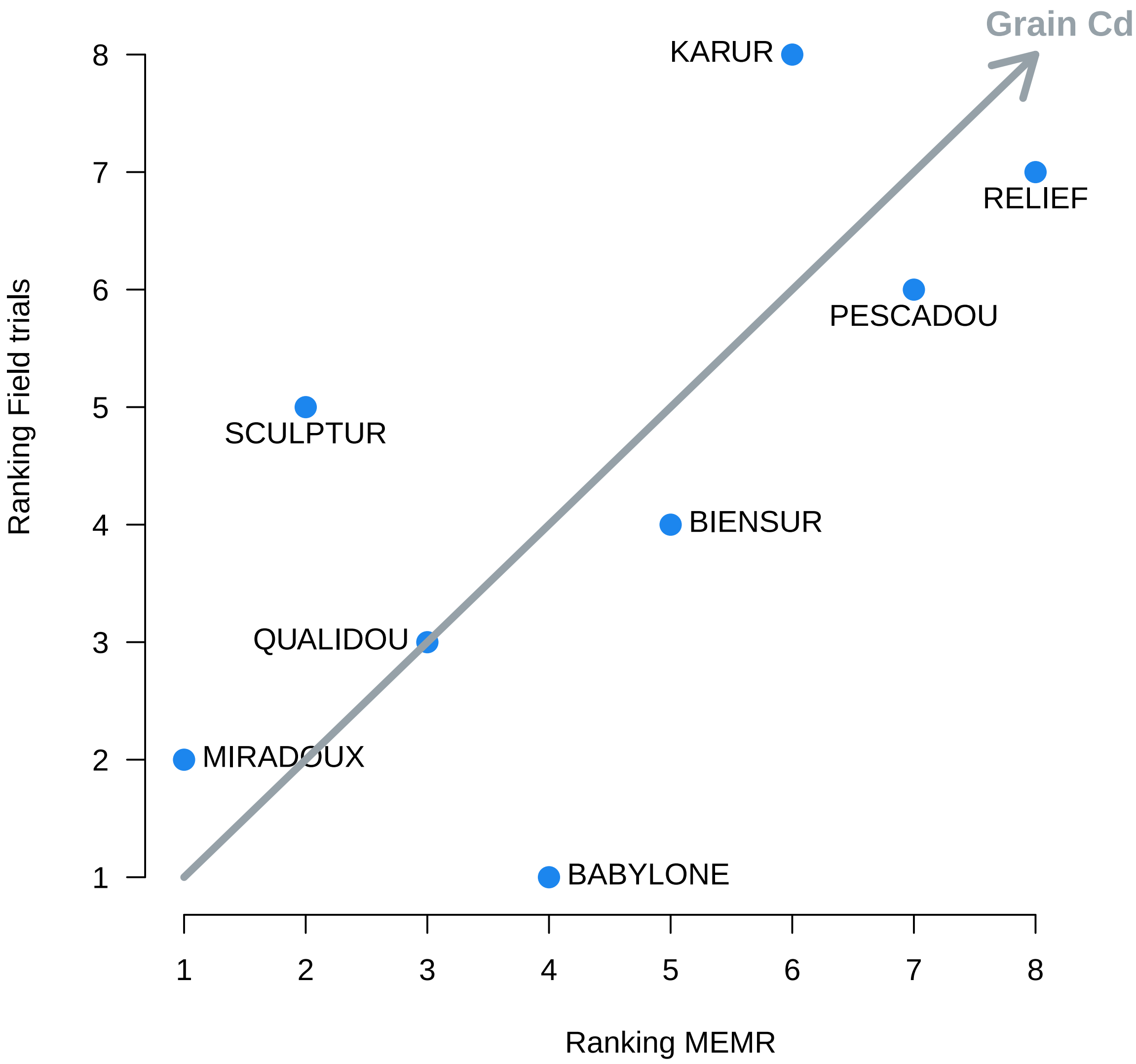

Grain Cd
KARUR
RELIEF
PESCADOU
SCULPTUR
BIENSUR
QUALIDOU
MIRADOUX
BABYLONE
Ranking Field trials
Ranking MEMR
1
2
3
4
5
6
7
8

Figure 6:

Supplementary Information

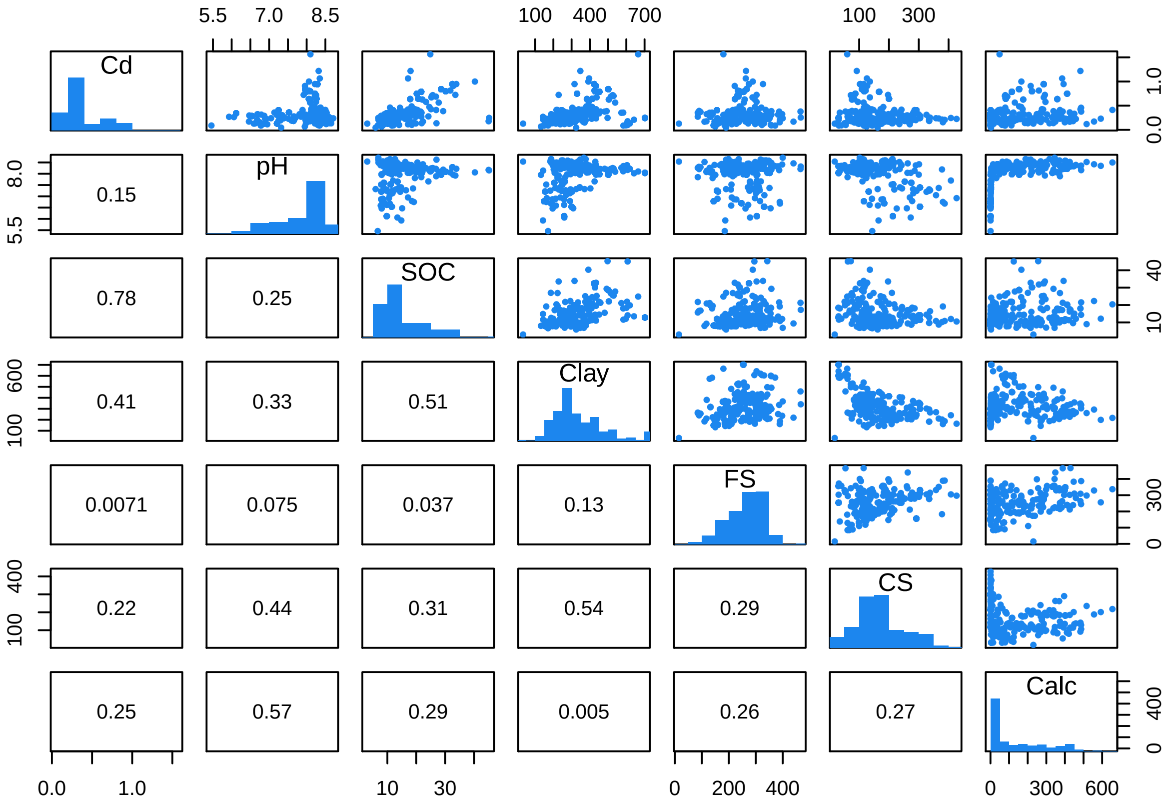


Figure SI1 : Matrix plot of correlations between the characteristics of the soils used for modeling the conformity of durum wheat grain Cd content. Cd: total soil Cd content, FS: fine silt, CS: coarse silt, SOC : soil organic carbon, Calc: soil calcareous. The bars on the diagonal show the distribution of the variable. Values below the diagonal are the Pearson correlation coefficients with two significant digits.

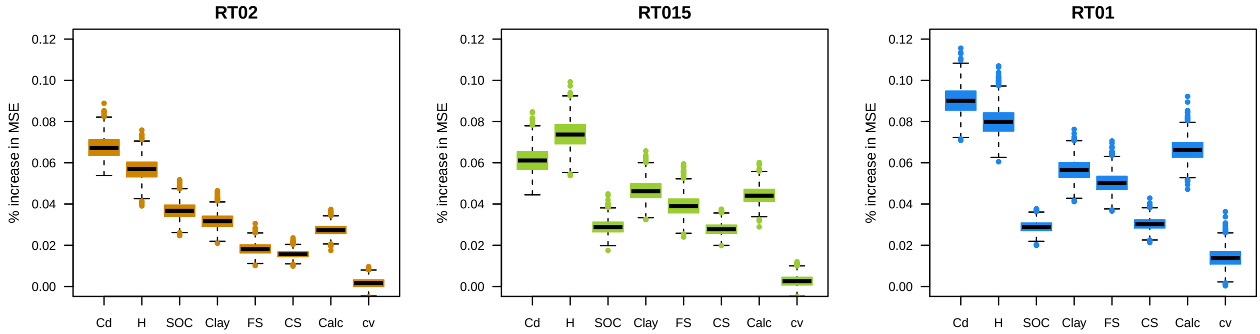


Figure SI2: Predictor importance estimated from the Random forest model with the following soil predictors : total soil Cd, soil pH, soil organic carbon, clay, fine and coarse silt and soil calcareous. The y-axis is the increase in the mean squared error of the model when the data of a given variable are shuffled while keeping other the variables unchanged.