Geoderma Regional 1 (2014) 21–30

Contents lists available at ScienceDirect

Geoderma Regional

journal homepage: www.elsevier.com/locate/geodrs

Refining a reconnaissance soil map by calibrating regression models with data from the same map (Normandy, France)

Fanny Collard [a], Bas Kempen [b], Gerard B.M. Heuvelink [b], Nicolas P.A. Saby [a], Anne C. Richer de Forges [a], Sébastien Lehmann [a], Pierre Nehlig [c], Dominique Arrouays [a,*]

[a] INRA, InfoSol Unit, US 1106, 45075 Orléans, France
[b] ISRIC-World Soil Information, PO Box 353, 6700 AJ Wageningen, The Netherlands
[c] BRGM, Direction des Géoressources, UMR 7327, BRGM-CNRS-Université d'Orléans, F-45060 Orléans, France

ARTICLE INFO

Article history:
Received 6 May 2014
Received in revised form 16 July 2014
Accepted 21 July 2014
Available online 23 July 2014

Keywords:
Digital soil mapping
Validation
Classification trees
Random forest
Multinomial logistic regression
Probability sampling
France

ABSTRACT

Reconnaissance soil maps at 1:250,000 scale are the most detailed source of soil information for large parts of France. For many environmental applications, however, the level of detail and accuracy of these maps is insufficient. Funds are lacking to refine and update these maps by traditional soil survey. In this study we investigated the merit of digital soil mapping to refine and improve the 1:250,000 reconnaissance soil map of a 1580 km^2 area in Haute-Normandie, France. The soil map was produced in 1988 and distinguishes nine soil class units. The approach taken was to predict soil class from a large number of environmental covariates using regression techniques. The covariates used include DEM derivatives, geology and land cover maps. Because very few soil point observations were available within the area, we calibrated the regression model by sampling the soil map on a grid. We calibrated three models: classification tree (CT), multinomial logistic regression (MLR) and random forests (RF), and used these models to predict the nine soil classes across the study area. The new and original maps were validated with field data from 123 locations selected with a stratified simple random sampling design. For MLR, the estimate of the overall purity was 65.9%, while that of the reconnaissance map was 55.5%. The difference between the purity estimates of these maps was statistically significant ($p = 0.014$). The significant improvement over the existing soil map is remarkable because the regression model was calibrated with the existing soil map and uses no additional soil observations.

1. Introduction

Reconnaissance soil maps at 1:250,000 scale are the most detailed source of soil information for large parts of France. The geographical coverage of 1:250,000 soil maps in mainland France is about 75% of the territory, while more detailed soil maps only cover about 35% of the country. For many environmental applications (e.g., threats to water quality, pollution of soils, soil erosion by water or wind, loss of, or damage to, rare soils, loss of terrestrial carbon store, loss of soil biodiversity; see a list of applications in France in Richer de Forges and Arrouays (2010)), however, the level of detail and accuracy of 1:250,000 maps is insufficient. Funds are lacking to refine and update these maps by traditional soil survey. This lack of detailed soil data and funds to increase resolution and accuracy through conventional soil survey is widely spread over the world (Hartemink, 2008).

Digital soil mapping (DSM) techniques (McBratney et al., 2003) have been proposed as a tool to update (Kempen et al., 2009) or disaggregate soil class maps (Häring et al., 2012; Nauman and Thompson, 2014; Subburayalu et al., 2014; Odgers et al., 2014), or to create new maps (Adhikari et al., 2014). Kempen et al. (2012a) show that DSM can be an efficient alternative to traditional soil survey for updating soil class maps. Various methods for calibration and mapping using DSM have been used, including expert based rules (Lagacherie et al., 1995; van Zijl et al., 2014), fuzzy logic systems (MacMillan et al., 2007; Zhu et al., 2001; Yang et al., 2011), neural networks (Behrens et al., 2005) and various methods of classification and regression (Carré and Girard, 2002; Grinand et al., 2008; Kempen et al., 2009; Häring et al., 2012; Adhikari et al., 2014; Nauman and Thompson, 2014; Subburayalu et al., 2014; Odgers et al., 2014).

DSM models are typically calibrated with observed point data (e.g., Häring et al., 2012; Kempen et al., 2012b; Adhikari et al., 2014). However, when resources for collecting new field point data are limited, obtaining a calibration dataset by sampling an existing soil map might be an attractive alternative, even though mapped soil properties and soil types are no substitute for real observations. This approach is taken by, for example, Lagacherie et al. (1995), Grinand et al. (2008), Debella-Gilo and Etzelmüller (2009), and more recently by Nauman and Thompson (2014), Subburayalu et al. (2014) and Odgers et al. (2014). However, some of these studies did not validate the resulting maps with independent field data (Debella-Gilo and Etzelmüller,

* Corresponding author.

http://dx.doi.org/10.1016/j.geodrs.2014.07.001
2352-0094/

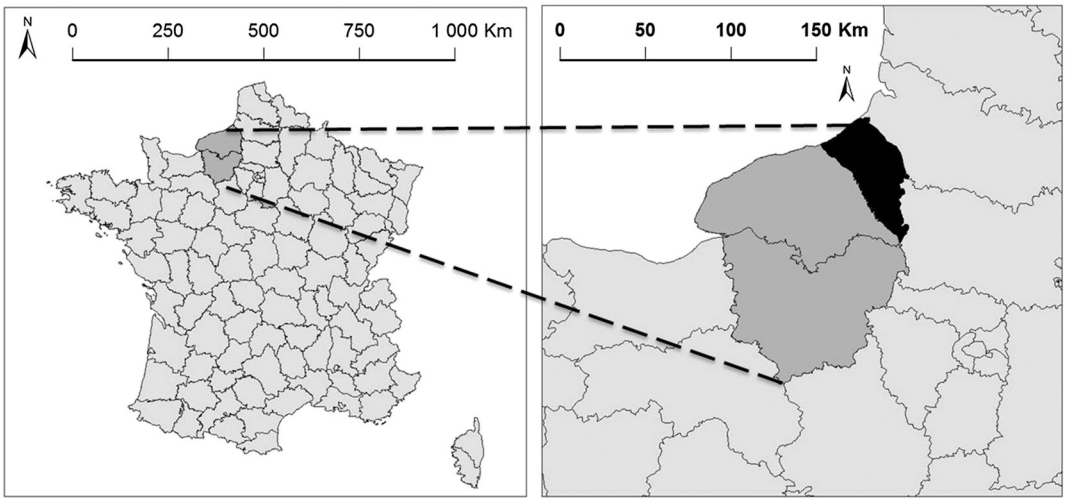

Fig. 1. Location of the study area.

2009; Grinand et al., 2008; Lagacherie et al., 1995), which makes it difficult or impossible to assess their accuracy. Others focused on a single prediction method (Grinand et al., 2008; Häring et al., 2012; Adhikari et al., 2014; Subburayalu et al., 2014; Odgers et al., 2014), and only Nauman and Thompson (2014) compared the accuracy of the digital, disaggregated soil class map with the legacy soil map. And none of these studies used independent validation collected with a probability sampling design that allows for statistically valid and unbiased accuracy assessment and model comparison.

In this paper we use multinomial logistic regression and two tree-based methods (classification trees and random forests) to investigate the merit of DSM to refine and improve the 1:250,000 reconnaissance soil map of a 1580 km^2 area in Haute-Normandie, France. We sampled the reconnaissance map and used this sample to calibrate the prediction models. Ground truth validation data were collected using probability sampling to evaluate whether i) the pedometric soil maps are more accurate than the original map, and ii) there are differences in accuracy between the three pedometric methods. This will provide insight if this method is an attractive alternative to traditional soil survey for updating and upgrading soil class maps in France.

2. Materials and methods

2.1. Study area

The study area is located in North-West France, along the Channel coast (Fig. 1). In this region, the parent materials are mainly loess deposits, chalk, sands and clays and more locally sand and gravel from alluvial deposits. Two main loess plateaus are located in the east and the north-west with elevations ranging from 200 m to 250 m and from 80 m to 180 m, respectively. Their land use is mainly intensive agriculture. Chalk soils occur mainly on steep slopes surrounding the plateau and are mostly occupied by forest. The south-eastern part is characterized by gently undulating relief. The soils are developed on sands and clays and land use is mainly permanent grassland. The climate is oceanic. The mean annual temperature is about 9 °C and the total annual precipitation is about 800 mm. A description of the nine soil classes of the 1:250,000 reconnaissance soil map of the area (Wolf et al., 1998) is given in Table 1.

2.2. Environmental ancillary data

Classical relief attributes were derived from the SRTM 90 m DEM.[1] Parent material was represented by a harmonized 1:50,000 lithological map that was synthesized from all geological surveys available for the region (Quesnel et al., 2007; Van Lint et al., 2003). Land use information was provided by the Corine Land Cover 2006 European database (Commission of the European Community, 1993) and climate information by the Ecoclimap database, a global database of land surface parameters at 1 km resolution (Masson et al., 2003). An exhaustive list of the 19 covariates used and their resolution or map scale is given in Table 2. Several of the DEM-derived covariates are (strongly) mutually correlated. Furthermore, cross-tabulating the categorical covariates with the reconnaissance soil map (from which the calibration points are derived) shows presence of zero-cell counts (Hosmer and Lemeshow, 2000). This means that the frequency distributions in the cross-table contain one or more zeros, i.e. not all combinations of predictor categories and soil classes occur. The presence of zero-cell counts causes numerical instabilities during modeling and should be avoided (Hosmer and Lemeshow, 2000). Hosmer and Lemeshow (2000) suggest combining classes of the categorical predictors in a sensible way to handle the zero-cell problem. However, this does not solve the issue about the correlated covariates. We, therefore, decided to convert the 19 covariate layers to 59 principal components (each class of the categorical covariates becomes one component after transformation), which are candidate predictors for the models.

The 1:250,000 reconnaissance soil map and the geological map were rasterized to 90 m resolution grids, corresponding to the resolution of the SRTM-derived terrain parameter grids. The Ecoclimap was resampled from 1 km to 90 m resolution.

2.3. Soil point observations

The point dataset for model calibration was obtained by sampling the reconnaissance soil map using a systematic, square grid with a random origin and 500 m grid spacing. The soil class was extracted at the grid nodes, which resulted in a sample of 6323 points.

2.4. Models

Three different methods were applied: multinomial logistic regression (MLR), classification tree modeling (CT), and random forests (RF).

2.4.1. Multinomial logistic regression

The logistic model belongs to the family of generalized linear models and is used when the response variable is categorical (Hosmer and Lemeshow, 2000). Suppose that variable Y represents the observed soil class at a sampling location, which can assume any of K categories, where K is the number of soil classes. In case K equals 2, Y has a

[1] http://srtm.csi.cgiar.org/.

Table 1
Description of the nine soil classes of the original reconnaissance soil map.

Code	French classification soil type	WRB soil type	Soil taxonomy	Description
LC	Néoluvisols	Luvic Cambisols	Hapludalf	Thick loamy soils developed on loess deposit
HLg	Luvisols rédoxiques	Haplic Luvisols Gleyic	Aquic Hapludalf	Thick redoxic loamy soils developed on loess deposit
RP	Planosols typiques sédimorphes	Ruptic Planosols	Paleudalf	Shallow loamy soils over flint clay
CL	Colluviosols limono-pierreux	Colluvic Leptosols	Eutrochrept	Shallow stony loamy soils over other materials
RL	Rendosols issus de craie	Rendzic Leptosols	Lithic Udorthent	Shallow calcareous loamy soils developed on chalk
F	Fluviosols	Fluvisols	Udifluvent	Alluvial soils
Se	Rédoxisols sablo-argileux	Stagnosols Endogleyic	Typic Haplaquept	Redoxic sandy–clayey soils
G	Réductisols argileux lourds	Gleysols	Typic Haplaquept	Reductic heavy clay soils
S	Rédoxisols argileux	Stagnosols	Humic Haplaquept	Redoxic clayey soils

Bernoulli (binomial) distribution with two possible outcomes y_1 and y_2. The probability of occurrence of y_1 is π_1 and that of y_2 is $\pi_2 = 1 - \pi_1$. Logistic regression relates probability π_1 to a set of predictors using the logit link function:

$$\text{logit}(\pi_1) = \log\left(\frac{\pi_1}{\pi_2}\right) = \log\left(\frac{\pi_1}{1-\pi_1}\right) = \mathbf{x}'\boldsymbol{\beta} \quad (1)$$

where $\mathbf{x}$ is a vector of predictors and $\boldsymbol{\beta}$ is a vector of model coefficients that are typically estimated by maximum likelihood, log is the natural logarithm. The ratio $\frac{\pi_1}{\pi_2}$ is referred to as the odds. From Eq. (1) it follows that:

$$\pi_1 = \frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1 + \exp(\mathbf{x}'\beta)}. \quad (2)$$

The binomial logistic regression model is easily generalized to the multinomial case. If there are K soil classes, then there are K probabilities of occurrence $\pi_1,\ldots,\pi_k$. One class is chosen as the reference class. Logits are formed that compare the other classes to it. Analogous to binomial logistic regression, the ratios $\frac{\pi_k}{\pi_1}$ $(k = 2,\ldots,K)$ are modeled by means of $\exp(\mathbf{x}'\boldsymbol{\beta}_k)$ where $k = 1$ is the reference class. From the constraint $\sum_{k=1}^{K}\pi_k = 1$ it follows that the probability of soil class k equals:

$$\pi_k = \frac{\exp(\mathbf{x}'\boldsymbol{\beta}_k)}{1 + \sum_{k=2}^{K}\exp(\mathbf{x}'\boldsymbol{\beta}_k)}. \quad (3)$$

We applied the MLR model using the *multinom* function of the package *nnet* (Venables and Ripley, 2002) in R (R Development Core Team, 2013). Selection of the covariate principal components was done with a manual step-wise procedure. The components were added sequentially to the MLR model, starting with the first component. A component was accepted as covariate if its selection resulted in a decrease of the Akaike Information Criterion (AIC) value (Webster and McBratney, 1989). After testing the last component, backward elimination was carried out, starting with the first component of the selected model. The elimination of a selected component was accepted if this resulted in a decrease of the AIC.

Table 2
Set of ancillary variables used for modeling and their resolution/scale.

Dataset	Resolution/map scale
Geology map (geol; 17 classes)	1:50,000
Digital Elevation Model (SRTM)	90 m
Terrain attributes	
Multiresolution Valley Bottom Flatness (MRVBF)	90 m
Slope	90 m
Aspect (EXP)	90 m
Global curvature (COURB)	90 m
Horizontal curvature (COURBL)	90 m
Transversal curvature (COURBT)	90 m
Roughness standard error on a 3 × 3 pixel window (SDT)	90 m
Roughness min–max on a 3 × 3 pixel window (PLAGE)	90 m
Topographic Position Index (TPI)	90 m
Landform Classification on a 3 × 66 pixels window (LCA)	90 m
Landform classification on a 1 × 66 pixels window (LCB)	90 m
Beven Index (BEV)	90 m
Distance to the nearest stream (DPPR)	90 m
Height to the nearest stream (HPPR)	90 m
Network persistence and development index (IDPR)	90 m
Land use	
Corine Land Cover (CLC, 21 classes)	1:100,000
Ecoclimap (ECOCL, 5 classes)	1 km

2.4.2. Classification trees

Initially developed by Breiman et al. (1984), the non-parametric classification tree algorithm partitions the training dataset, in our case the soil point dataset, recursively into increasingly homogeneous subsets (Strobl et al., 2009). The covariates are used as partitioning variables, and each binary split is chosen in such a way that it maximizes the reduction of an impurity measure, such as the Shannon entropy, the Gini index or the classification error (Hastie et al., 2009). For continuous covariates the split point is a threshold value, while categorical covariates are split by grouping the covariate classes into two groups. To avoid overfitting, trees are pruned using cost-complexity pruning (Hastie et al., 2009; Venables and Ripley, 2002). Here we used the 1-SE rule as defined by Venables and Ripley (2002) to determine the cost-complexity parameter, which was set to 0.001. Predictions at the terminal nodes of the tree are based on majority. The algorithms were run using *rpart* package (Therneau and Atkinson, 2013; Venables and Ripley, 2002). The minimum number of observations for each node to be considered for a split (*minsplit* argument) was set to 20, and the minimum number of observations in any terminal node (*minbucket* argument) was set to 10. The Gini index was used as an impurity measure for splitting (this is the default in *rpart*). Since correlated covariates and zero-cell counts are less of a problem for CT than for MLR, the non-transformed covariates were used as input.

2.4.3. Random forests

A random forest consists of an ensemble of classification or regression trees (Breiman, 2001; Strobl et al., 2009). In our case classification trees are used. Each of these trees is generated by recursive binary partitioning as described in Section 2.4.2 above. Random forests combine bootstrap sampling, aggregation (bagging) and random covariate selection for partitioning, to grow a forest of trees. Each tree in the forest is grown from a bootstrap sample drawn from the training data (in our case the soil point observations) with replacement. For each split, a random set of covariates is selected. From this set, the best covariate is chosen, i.e. the one that results in the largest reduction of the impurity index. The class prediction at a new location is based on the majority vote of the individual tree predictions (Breiman, 2001; Hastie et al., 2009). An estimate of the classification error is obtained by predicting at the training sites not included in the bootstrap sample, the so-called 'out-of-bag' (OOB) data. The OOB predictions are aggregated after which the OOB-error is computed. Since the OOB error sample is almost identical to that obtained by N-fold cross-validation (Hastie et al., 2009), no separate cross-validation is required.

Random forest modeling was applied with the *randomForest* package (Liaw and Wiener, 2002) in R. Like for CT, non-transformed covariates were used as input. To run the model, two parameters must be specified: the number of randomly selected variables to try at each split (*mtry*) and the number of trees to grow (*ntree*). For classification, the default value for *mtry* is the square root of the total number of covariates (Hastie et al., 2009). Strobl et al. (2009) recommend to use a larger number of randomly selected covariates when the covariates are (strongly) correlated. We, therefore, set *mtry* to half the total number of covariates (rounded downward). The parameter *ntree* was set to 1000. The minimum size of terminal nodes (*nodesize*) was set to 10.

2.5. Model validation

Cross-validation for the MLR and CT models was 10-fold. For RF, the OOB-error is used as an internal cross-validation measure. By default, the bootstrap sample selects 63.2% of the sampling sites for calibration. The OOB-error is computed from the remaining samples. In addition to cross-validation, the predictions were validated with an external (independent) validation sample as well.

2.5.1. Sampling design

The soil maps were validated with independent data collected using a stratified simple random sampling design (Brus et al., 2011). The nine map units of the 1988 reconnaissance soil map formed the strata. A total of 123 sampling locations were selected, with per-stratum sample sizes proportional to their surface areas. Sampling locations where permission was denied or proved otherwise impossible to sample were replaced with randomly selected locations within the same stratum. Soil type was described using auger borings. Each identified horizon was sampled and analyzed for pH, C, $CaCO_3$ and particle size distribution, and the allocation to classes was done by expert judgment on the basis of soil description and analyses, without knowing the location of the site on the pre-existing map. Seven of the 123 sampling locations were not allocated to any of the soil classes. Fieldwork took place in November 2012.

2.5.2. Estimation of map quality measures

We describe the accuracy measures for soil class maps only briefly. For an elaborate review, including the estimation of these measures and associated estimation variances, we refer to Brus et al. (2011) and de Gruijter et al. (2006).

We consider three map quality measures for the soil class maps: the overall purity, the map unit purity (user's accuracy) and class representation (producer's accuracy) (Brus et al., 2011; Stehman, 1997). Each of these properties can be derived from an error matrix (cross-tabulation of observed versus predicted soil class) and is directly interpretable in terms of a probability of occurrence of a misclassification. The 7 non-allocated sampling locations were taken into account in the measurements of map quality. The overall purity is defined as the proportion of the mapped area in which the predicted soil class, which is the soil class depicted on the map, equals the true soil class, i.e. it is the areal proportion correctly classified. To estimate the overall purity an indicator variable is created, which takes the value 1 if the observed soil class equals the predicted soil class, and 0 otherwise. For each stratum the average of this indicator is computed. The overall purity is estimated as the weighted average of the stratum purities, with weights equal to the relative areas of the strata.

The map unit purity defines the purity on the level of the map units (individual soil classes). The map unit purity for mapped soil class k is the proportion of the area of the respective map unit correctly classified. If the map units were used as the sampling strata, then the map unit purities are estimated by the strata means (in the case of the reconnaissance soil map). If the map units do not equal the strata (in the case of the MLR, CT and RF maps), then the map unit purities must be estimated by the ratio estimator. This estimator is used for purity estimates of so-called *domains* (sub-areas of interest) (Brus et al., 2011; Kempen et al., 2009). The class representation for soil class k is the proportion of the area where in reality soil class k occurs that is also mapped as class k. Class representations are also estimated by the ratio estimator.

We compared the accuracies of the soil maps by introducing variable q_{hi} defined as $y_{hi}^{m1} - y_{hi}^{m2}$, where y_{hi} is an indicator that takes value 1 if the predicted soil class at validation location i in stratum h equals the observed soil type and 0 otherwise. Superscripts $m1$ and $m2$ indicate the two maps that are being compared. The variable q_{hi} can have values of -1, 0, and 1. The mean purity difference (MPD) of two soil maps is estimated in a similar fashion as the overall purity, by summing over all strata and all locations within a stratum. To test whether the estimated MPD differs significantly from 0, we assumed that the estimated MPD follows a normal distribution.

3. Results

3.1. Prediction models and maps

For the MLR model, twelve out of 59 PCs were selected. PC1 was related to elevation (SRTM) and terrain attributes, while PC2 to PC11 all included geological classes and various terrain attributes and/or land cover classes. Some PCs were directly related to one geological class: loess deposits (PC2), colluvial loam on slopes (PC4), sedimentary clay (PC5) and alluvial deposits (PC9). PC12 was directly related to landform classification (lcb).

The CT is shown in Fig. 2. After pruning, the fitted CT had 35 splits and 36 terminal nodes. Ten out of the nineteen covariates were used to construct the tree. Geology (parent material) and elevation are the dominant splitting covariates at the upper levels of the CT, whereas DEM derivatives are used as splitting covariates at the lower levels of the CT. The ecoclimate map is used for only one split. The land cover map is not used for splitting. The first three splits (using geology twice and elevation once) increase the internal purity from 22.1% (obtained by predicting the most frequent soil class of the calibration data) to 50.4%. The CT and RF models can be compared using the calculation of the well-known variable importance (VI) which is a measure of the contribution a covariate can make in prediction In the case of CT, the reduction in the loss function (e.g. mean squared error) attributed to each variable at each split is tabulated and the sum is returned.

In the case of RF, for each tree, the prediction accuracy on the out-of-bag portion of the data is recorded. Then the same is done after permuting each predictor variable. The difference between the two accuracies is then averaged over all trees, and normalized by the standard error. To make them comparable in a unique plot, all measures of importance are scaled to have a maximum value of 1. The VI plots of CT and RF are shown in Fig. 3. They show that for both models, geology and elevation are the most important covariates followed by various DEM derivatives. Curvature seems to have more influence for RF than CT. Land cover and climate data do not have a strong influence on the spatial distribution of soil classes. Indeed, climate does not exhibit strong gradients in this area and the predominance of geology and elevation is consistent with the sedimentary origin of the soil parent materials.

The three prediction maps and the reconnaissance soil map look globally similar (Fig. 4). The shape and size of the large loamy plateau of the north-western part with Luvic Cambisols (LC) and the area with Stagnosols Endogleyic (Se) are nearly the same. On the contrary, noticeable differences occur for the western loamy plateau with Haplic Luvisols Gleyic (HLg) and for the Ruptic Planosol (RP) area. Overall, the prediction maps seem to produce smaller clusters of grid cells and a more complex spatial organization than the original reconnaissance soil map. When the prediction maps are compared, then MLR seems to predict the most intricate spatial pattern, followed by CT and then RF. The RF and CT maps are more similar to each other than to the MLR map, which is not surprising since the models are more similar to each other than to MLR. The spatial patterns in the RF map are

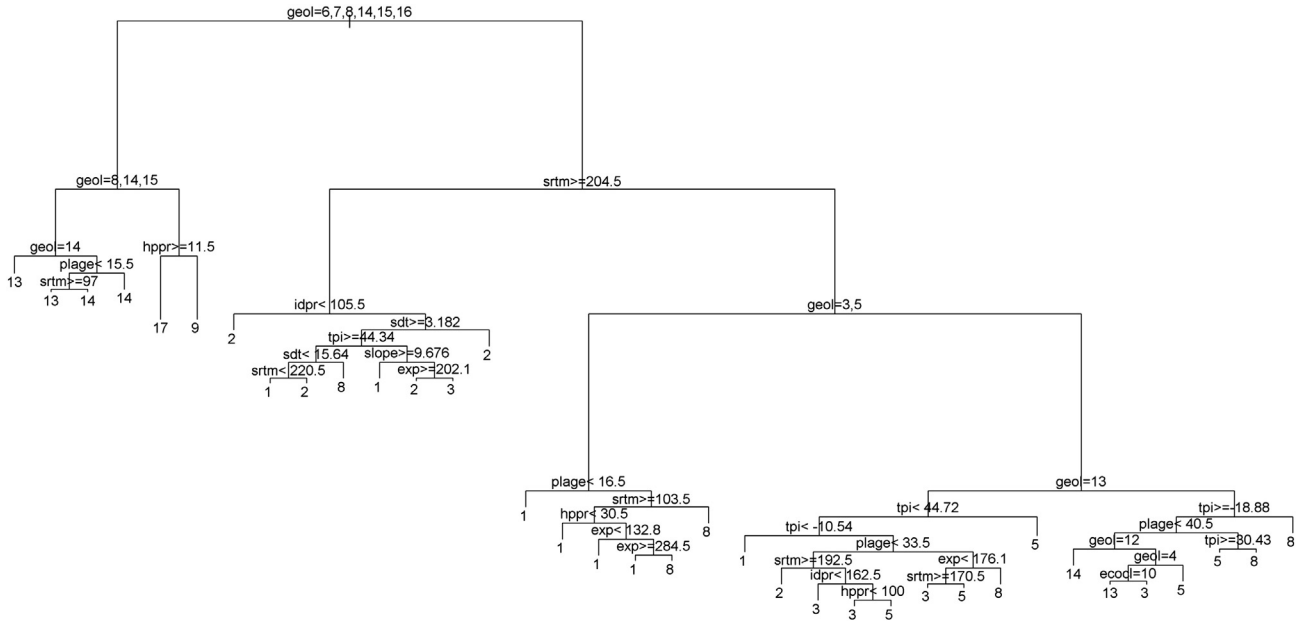

Fig. 2. Relative importance of the predictors for the random forest (RF) approach and the classification tree approach.

somewhat less intricate than for the CT map. This is likely the result of fitting a large number of trees in RF and then predicting the soil class based on a majority vote from the individual tree predictions, which filters out some of the 'noise' predicted by CT. CTs are known to be unstable and can have large variance: a small change in the data can result in a very different series of splits (Hastie et al., 2009). Despite the differences in predicted spatial patterns, the predicted areas are roughly similar (Table 3). Compared with the reconnaissance map, the models predict larger areas with RL (Rendzic Leptosols; +14.5% for RF, +16.4% for CT, +20.4% for MLR) and HLg (Haplic Luvisols Gleyic; +14.5% for RF, +16.4% for CT, +20.4% for MLR), and smaller areas with CL (Colluvic Leptosols; −26.6% for RF, −28.8% for CT, −46.0% for MLR) and RP (Ruptic Planosols; −25.3% for RF, −22.7% for CT, −23.3% for MLR). In particular, some very large polygons of Ruptic Planosols are considerably reduced in smaller clusters of grid cells.

3.2. Validation

3.2.1. Overall purity

Table 4 presents the validation results (overall purity estimates) for the reconnaissance soil map and the three DSM methods.

Based on cross-validation, RF gives the most accurate map (68.0%), followed by CT (63.6%), and MLR (58.8%). The results of the external validation show that for CT and RF these accuracy measures are over-optimistic. We had expected this for MLR as well, but strangely in this case the external purity is 7% larger than the cross-validation purity. We were unable to pinpoint the cause of this difference. The cross-validation purity falls within the 95% confidence interval of the estimated external purity so the observed purity difference might be attributed to chance.

The best validation result was obtained for MLR. The overall purity is 65.9%, which is 10.4% larger ($p = 0.014$; Table 5) than that of the

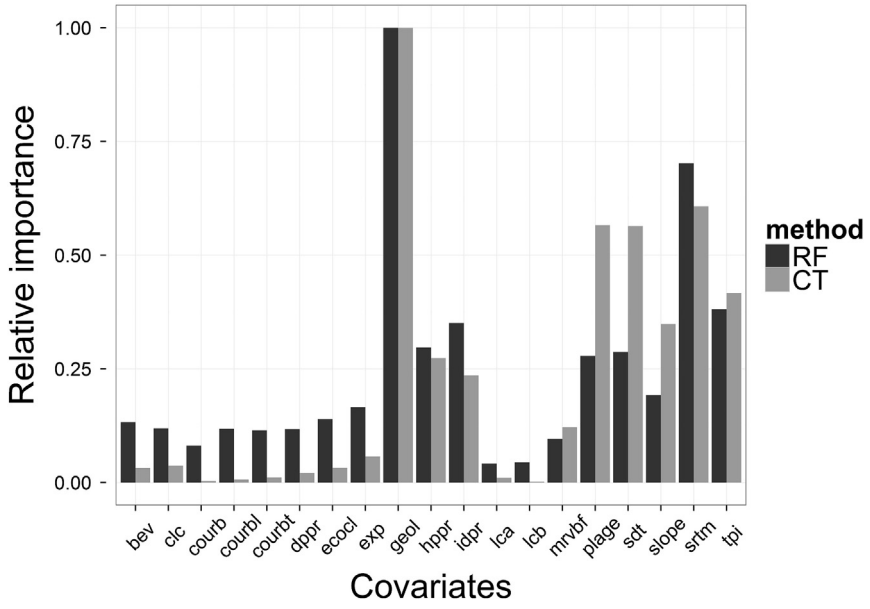

Fig. 3. The fitted classification tree model.

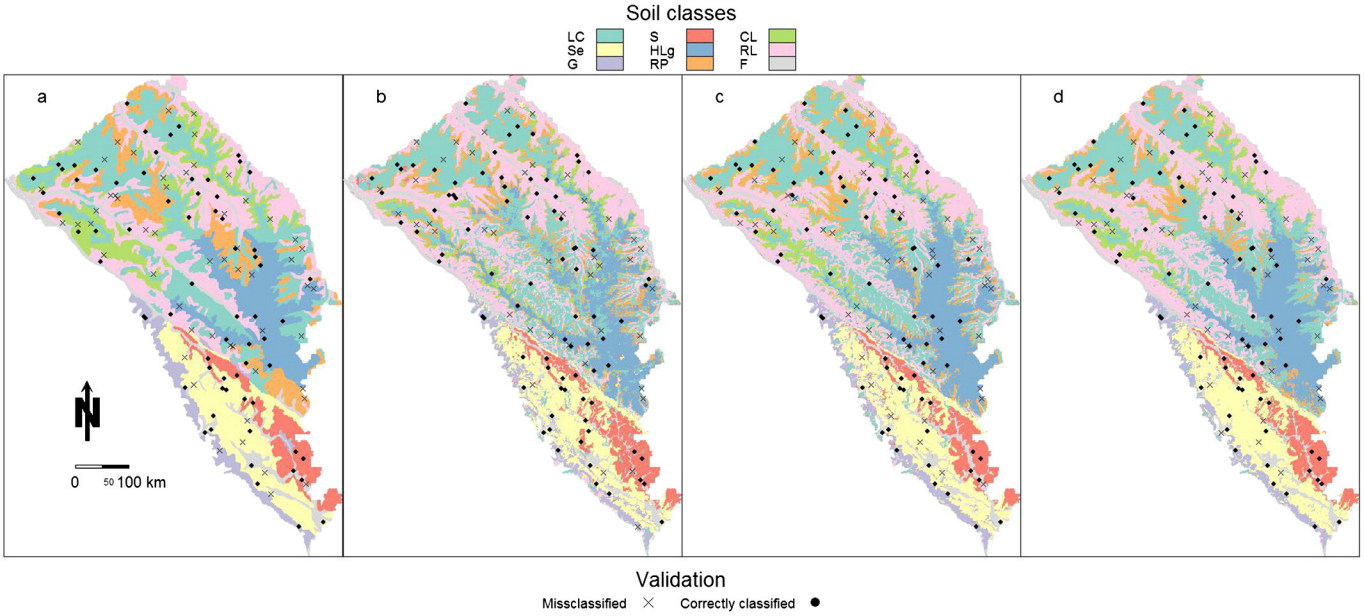

Fig. 4. Reconnaissance soil map (a) and maps of predictions obtained with the three regression models: multinomial logistic regression (b), classification tree (c) and random forests (d).

reconnaissance map. The overall purity of the RF model was 63.5% and that of the CT model 61.3%. Both CT ($p = 0.065$) and RF ($p = 0.029$) performed better than the reconnaissance map. Differences in overall purity between the DSM models were not statistically significant (Table 5).

3.2.2. Map unit purity and class representation

Table 6 presents the confusion matrices of observed versus predicted soil classes. The soil classes exhibiting a high degree of confusion with others are the Luvic Cambisols (LC) the Ruptic Planosols (RP) and the Rendzic Leptosols (RL). The observed RP are spread over seven map units for MLR and CT and six map units for RF. Conversely, the predicted RL led to confusion with five to seven other observed classes depending on the method used. The Luvic Cambisols are mainly confused with Ruptic Planosols. Indeed these soils tend to occur together in the landscape. The LC occupy the flat loamy plateau whereas RP are located on slopes at the border of the plateaus. Similarly, HLg are mainly confused with LC and RP. Both HLg and LC are located on loamy plateaus and their difference is mainly linked to small differences in topography. Both are frequently surrounded geographically by RP. Soil classes F, Se, G and S are less confused with other classes, which is consistent with the fact that these are all strictly linked to distinct geological origins.

The map unit purities and class representation measures are presented in Table 7. Using DSM, map unit purities increase for six or seven out of nine map units. The tree-based maps show somewhat less variation in map unit purities than the MLR map. The prediction performance of the pedometric models and reconnaissance map might be more easily compared using the overall purities of the sampling strata (Table 8), because in that case predictions for the same spatial entities are compared. Note that the strata are equal to the map units of the reconnaissance map. For the areas classified as Luvic Cambisols (LC), Ruptic Planosols (RP), Colluvic Leptosols (CL), Fluvisols (F) and Stagnosols (Se) on the reconnaissance map, the DSM models predict the soil class distribution considerably better than the reconnaissance map. Table 8 also shows that differences in overall purity between MLR and the tree-based methods are mainly attributed to more accurate predictions by the MLR model in strata LC, F, and Se.

Class representations increase for six (MLR) or seven (CT, RF) out of the nine soil classes compared to the reconnaissance map. Both DSM methods and the reconnaissance map have difficulty predicting the occurrence of the RP class, which has the smallest class representation. The observed RP are spread over seven predicted map units (Table 6). This indicates that the occurrence of flint clay in deep layers, that characterizes the RP class, cannot be adequately predicted. This could be due to the fact that the occurrence of RP is not well depicted by the reconnaissance map. As the reconnaissance map is used for calibration, then if RP is not accurately mapped, similar inaccuracies will result using the DSM methods. Indeed, here we reach the limits of improvements that can be obtained by calibration of a regression model with training data obtained from a reconnaissance map. Classes CL and F show the largest improvement in class representation compared with the

Table 3
Absolute (km^2) and relative (%) areas of the mapped soil classes. The 'true' (observed) areas are estimated from the validation probability sample.

	Observed			Reconn. map		MLR		CT		RF	
	km^2	%	95% CI (%)	km^2	%	km^2	%	km^2	%	km^2	%
LC	333	21.1	17.4–24.8	346	21.9	367	23.2	353	22.3	370	23.4
HLg	115	7.3	5.0–9.7	173	10.9	197	12.5	217	13.7	195	12.3
RP	283	17.9	14.4–21.4	153	9.7	117	7.4	118	7.5	114	7.2
CL	77	4.9	2.9–6.8	139	8.8	75	4.7	99	6.3	102	6.4
RL	231	14.6	11.4–17.8	325	20.6	391	24.7	378	23.9	372	23.5
F	115	7.3	4.9–9.7	122	7.7	119	7.5	107	6.8	112	7.1
Se	155	9.8	7.1–12.4	174	11.0	171	10.8	171	10.8	182	11.5
G	63	4.0	2.3–5.9	67	4.2	53	3.4	58	3.7	55	3.5
S	115	7.3	5.0–9.7	82	5.5	90	5.7	77	4.9	79	5.0
Other	90	5.7	3.6–7.8	–	–	–	–	–	–	–	–

Table 4
Results of validation of the reconnaissance soil map and the multinomial logistic regression (MLR), classification tree (CT) and random forest (RF) prediction maps, and results of internal cross-validation. The *se* is the estimated standard error of the overall purity, l-90% CI is the lower boundary value of the 90% confidence interval (CI) of the overall purity, and u-90% CI is the upper boundary.

	Reconnaissance map	MLR	CT	RF
External validation				
Overall purity (se)	55.5 (4.4)	65.9 (4.2)	61.3 (4.3)	63.5 (4.3)
l-90% CI	48.3	59.0	54.2	56.4
u-90% CI	62.7	72.8	68.3	70.6
Cross-validation				
Overall purity	–	58.8	63.6	68.0

Table 5
p-Values of the statistical test for the differences in overall purity between the reconnaissance, multinomial logistic regression (MLR), classification tree (CT) and random forest (RF) maps.

	MLR	CART	RF
Reconnaissance map	0.014	0.065	0.029
MLR		0.122	0.260
CART			0.220

reconnaissance map. This improvement is likely due to the use of the DEM derivatives, as these classes are located in specific topographic areas. The pedometric maps represent classes Haplic Luvisols Gleyic (HLg), Rendzic Leptosols (RL) and F best.

Probability sample data provide unbiased estimates of the true areal fractions of the soil classes, and the uncertainty associated with these estimates in the form of a confidence interval. These are given in Table 3. A comparison of the 'true' areas with the predicted areas indicates that the pedometric maps and the reconnaissance map strongly under-estimate the area with RP, hence the small class representation. The pedometric maps over-estimate the areas with HLg and RL, which explains the large class representations. The reconnaissance map over-estimates the areas with these soils as well, but to a smaller extent than the pedometric maps.

4. Discussion

4.1. Interest and limitations of the method

The main finding of this study is that the accuracy of the reconnaissance soil map could be improved without additional sampling, by only utilizing the relationship between soil class and covariates as calibrated on the existing soil map. For most of the soil classes, the pedometric maps gave more complex and detailed spatial patterns than the original map. The reconnaissance soil map generally delineates large polygons corresponding to the soil class that the surveyors considered to be dominant in the landscape. It seems therefore not surprising that large clusters of grid cells corresponding to plateaus look quite similar in all maps, and that more narrow clusters located in complex landscapes occur more frequently using DSM than by the original map. This indicates an improvement because these local and fragmented soil classes really occur but could not be represented in the 1:250,000 reconnaissance soil map due to map generalization principles. The largest improvements of DSM maps compared to the original soil map were obtained for map units characterized by specific topographic positions: flat plateau (LC), steep slopes (CL) and alluvial plains (F). Indeed in this case, the SRTM data proved to be more efficient to delineate these units than the conventional way that made use of an old topographic map.

The overall purity of the DSM maps ranges from 61.3% to 65.9%. These purities are comparable and consistent with previous studies (e.g., Grinand et al., 2008; Häring et al., 2012; Kempen et al., 2009, 2012b; Lemercier et al., 2012) and somewhat smaller than the recommendation by Marsman and de Gruijter (1986), who suggested 70% as an acceptable overall purity for soil maps.

All DSM methods gave a statistically significant improvement in map accuracy compared to the reconnaissance soil map. This finding is very remarkable, because no additional observations were used and the model was calibrated using a dense grid overlaid on the existing soil map. A 'perfect' regression model would calibrate such that prediction with such model would exactly reproduce the original map (note that a perfect model would not reproduce reality because we used a calibration sample obtained from a map and not from reality), which would mean no improvement. We used imperfect models that do not reproduce the calibration data exactly (as shown by the internal purity estimates), and remarkably this yielded maps with larger purities. The improvement was statistically significant so we can rule out that this happened by chance. Apparently, the 'imperfect' models rightfully replace implausible combinations of soil classes and covariates as present in the existing soil map with more plausible combinations.

Table 6
Cross-tabulation between observed and predicted soil classes at the 123 validation locations for the reconnaissance (a), the multinomial logistic regression (MLR) (b), classification tree (CT) (c), and random forest (RF) (d) maps. Bold figures are the diagonal of the cross-table.

Predicted	Observed										Total
	Reconnaissance map										
	LC	HLg	RP	CL	RL	F	Se	G	S	Other	
LC	**14**	0	6	1	2	1	0	0	0	2	26
HLg	3	**7**	3	0	0	0	1	0	0	0	14
RP	5	1	**5**	0	1	0	0	0	0	0	12
CL	3	0	3	**3**	1	0	0	0	0	1	11
RL	0	1	4	2	**14**	1	1	0	0	2	25
F	1	0	1	0	0	**5**	0	0	2	1	10
Se	0	0	0	0	0	1	**10**	1	1	1	14
G	0	0	0	0	0	1	0	**4**	0	0	5
S	0	0	0	0	0	0	0	0	**6**	0	6
Other	0	0	0	0	0	0	0	0	0	**0**	0
Total	26	9	22	6	18	9	12	5	9	7	123

Predicted	Observed										Total
	MLR										
	LC	HLg	RP	CL	RL	F	Se	G	S	Other	
LC	**17**	0	4	0	1	0	1	0	0	0	23
HLg	3	**7**	3	0	0	0	0	0	0	1	14
RP	4	0	**4**	0	0	0	0	0	0	0	8
CL	0	1	4	**4**	0	0	0	0	0	2	11
RL	2	1	5	2	**17**	0	0	1	0	3	31
F	0	0	1	0	0	**8**	0	0	0	0	9
Se	0	0	0	0	0	0	**11**	0	0	1	12
G	0	0	0	0	0	0	0	**4**	0	0	4
S	0	0	1	0	0	1	0	0	**9**	0	11
Other	0	0	0	0	0	0	0	0	0	**0**	0
Total	26	9	22	6	18	9	12	5	9	7	123

	CT										Total
	LC	HLg	RP	CL	RL	F	Se	G	S	Other	
LC	**15**	0	5	1	3	0	2	0	0	0	26
HLg	3	**8**	5	0	0	0	0	0	0	2	18
RP	4	0	**6**	1	0	0	0	0	0	1	12
CL	0	0	1	**4**	0	0	0	0	0	1	6
RL	4	1	3	0	**15**	1	0	0	0	1	25
F	0	0	1	0	0	**7**	0	0	2	1	11
Se	0	0	1	0	0	0	**9**	0	1	1	12
G	0	0	0	0	0	0	1	**5**	0	0	6
S	0	0	0	0	0	1	0	0	**6**	0	7
Other	0	0	0	0	0	0	0	0	0	**0**	0
Total	26	9	22	6	18	9	12	5	9	7	123

	RF										Total
	LC	HLg	RP	CL	RL	F	Se	G	S	Other	
LC	**16**	0	5	1	2	0	0	0	0	0	26
HLg	3	**8**	3	0	0	0	1	0	0	1	17
RP	3	0	**6**	0	0	0	0	0	0	0	9
CL	1	0	3	**3**	0	0	0	0	0	0	8
RL	3	1	4	2	**16**	1	1	0	0	4	27
F	0	0	1	0	0	**8**	0	0	1	0	23
Se	0	0	0	0	0	0	**10**	0	1	2	13
G	0	0	0	0	0	0	0	**5**	0	0	4
S	0	0	0	0	0	0	0	0	**7**	0	7
Other	0	0	0	0	0	0	0	0	0	**0**	0
Total	26	9	22	6	18	9	12	5	9	7	123

Table 7
Map unit purities and class representations of the reconnaissance soil map and multinomial logistic regression (MLR), classification tree (CT) and random forest (RF) prediction maps.

Soil class	Class representation (%)				Map unit purity (%)			
	Reconnaissance map	MLR	CT	RF	Reconnaissance map	MLR	CT	RF
LC	55.3	66.0	58.9	62.6	53.8	73.8	58.0	61.5
HLg	77.1	77.4	88.4	88.4	50.0	49.4	43.9	46.6
RP	22.5	17.8	27.1	27.1	41.7	49.8	50.2	66.8
CL	49.1	66.8	66.4	66.4	27.3	36.5	66.1	50.2
RL	77.8	94.4	83.5	83.2	56.0	54.9	60.4	55.8
F	53.9	89.2	77.5	100	50.0	89.2	64.3	75.5
Se	83.0	91.3	74.7	74.7	71.4	91.7	74.9	68.9
G	81.1	79.7	100	79.7	80.0	100	84.1	100
S	69.1	100	69.1	79.3	100	82.9	86.3	100

The results presented in this paper indicate that in situations with limited budgets for collecting field data for calibration of DSM models, using a legacy soil map to obtain calibration data can be an efficient and attractive alternative. Nevertheless, it is worth investigating if similar results can be obtained for other (larger) areas in France, areas in other parts of the world and for different mapping methods in order to draw more generic conclusions about the merit of using a legacy soil map for model calibration. Further similar work is on-going for the entire Haute-Normandy region. Also for this region, the DSM maps will be validated with independent probability sample data and compared with the reconnaissance map. It will be interesting to see if similar results with respect to predictive performance will be obtained. When successful, then the work in Normandy will serve as an example to expand the method to other regions in France.

Our results were interesting for refining a small scale reconnaissance survey that used few ground sampling. In this case, soil mapping units already correspond to well-identified landscape units that should be easily refined by adding more precise soil covariates. It will perhaps be less efficient with more detailed soil surveys in which soil class delineations are more based on the ground sampling and may therefore have a more complex landscape position.

Another interesting point is that by using a field sample for validation we found some soils (i.e., seven of the 123 sampling locations) that could not be classified in any of the soil classes defined by the original soil map. Five of these correspond to outcrops of parent material by erosion processes. More generally, this observation stresses a major limitation of the method that is linked to the quality and the completeness of the calibration data. Indeed, DSM methods can only predict the soil classes that are present in the calibration dataset, which, in our case, in turn is limited by the classes distinguished on the reconnaissance soil map.

Using validation data obtained from the soil map assumes that the map is the perfect truth, which is obviously not the case, as we showed in this study. Purities estimated from (the internal and cross-validation purities) tended to be smaller than the external purities (with the exception for MLR), i.e. using data sampled from a soil map for calibration might lead to over-optimistic estimates of the map purity, as also illustrated by Nauman and Thompson (2014). This illustrates the importance of independent validation with field data.

Table 8
Estimated stratum purities. Note that the strata coincide with the map units of the reconnaissance soil map. *n* is the number of locations sampled.

Stratum	Area (km^2)	*n*	Reconnaissance map	MLR	CT	RF
LC	346	26	53.8	73.9	57.7	61.5
HLg	173	14	50.0	50.0	44.4	47.1
RP	153	12	41.7	50.0	50.0	66.7
CL	139	11	27.3	36.4	66.7	50.0
RL	325	25	56.0	54.8	60.0	55.6
F	122	10	50.0	88.9	63.6	75.0
Se	174	14	71.4	91.7	75.0	69.2
G	67	5	80.0	100	83.3	100
S	82	6	100	81.8	85.7	100

4.2. Comparison with similar approaches

Debella-Gilo and Etzelmüller (2009) used a similar calibration method to model the relationship between thirteen WRB soil groups and terrain attributes and predict the spatial distribution of the soil groups using digital terrain analysis with multinomial logistic regression in a Norwegian county. Grinand et al. (2008) sampled a soil map and split the original dataset in two parts, one for calibration and one for validation. Lagacherie et al. (1995) proposed a quantified formulation of mapping rules derived from an existing soil map and used these within an automated soil survey procedure. The soil pattern rules of a reference area were formulated through rules which gave probabilities of soil unit presence. However, these studies did not use independent field data to evaluate their results. Another difference between these studies and the work presented here is that these studies used existing maps to extrapolate to other areas, whereas we resampled an existing map.

Disaggregation of legacy soil class maps is gaining increased interest recently (Häring et al., 2012; Subburayalu et al., 2014; Odgers et al., 2014; Nauman and Thompson, 2014). All these authors used tree-based methods. Of these, Odgers et al. (2014) and Subburayalu et al. (2014) used more advanced methods that included probabilistic resampling (Odgers et al., 2014) and a possibilistic approach (Subburayalu et al., 2014). Here we showed that also with more straightforward methods similar results in terms of map accuracy can be obtained. The aim of disaggregating (or refining) legacy maps is to increase detail and, herewith, hopefully map accuracy. So far, only Nauman and Thompson (2014) compared the accuracy of the disaggregated map with that of the legacy map. These authors found a small increase of accuracy compared to legacy map. Here we also find an improvement compared to the legacy map, and showed that, on the basis of independent validation with probability sample data, the improvement was statistically significant.

4.3. Comparison of pedometric models

The three pedometric models gave similar results in terms of map purity. MLR performed somewhat better than CT and RF, although purity differences were not statistically significant. We found this somewhat surprising since we expected that the tree-based models would outperform MLR given their greater flexibility. Tree-based methods are capable to model non-linear relationships, can handle observations with missing covariate data and situations with a large number of covariates and small number of observations, and, in the case of a categorical target variable, they do not suffer from the zero-cell problem as MLR does. Because of their flexibility, tree-based methods are becoming increasingly popular for digital soil mapping (e.g. Wiesmeier et al., 2011; Häring et al., 2012; Heung et al., 2014; Nauman and Thompson, 2014; Subburayalu et al., 2014; Odgers et al., 2014). However, studies that compare the prediction performance of linear models and tree-based

models are still very limited. To our best knowledge, this paper is the first that compares the prediction performance of a linear model with tree based model for mapping soil classes. Here we showed that a linear model as well as a tree-based model can perform.

4.4. Effects of calibration sampling

The calibration data used in this study was a dense regular grid sample from the entire study area. We might expect a further improvement in purity if the calibration sampling was limited to areas where we have more confidence in the reconnaissance map. For example, we expect that mapping errors are larger close to map unit boundaries. We tested this hypothesis by applying a 250 m buffer (1 mm on the paper map) around the map unit boundaries. Sampling locations that fell into the buffer zone were excluded from the calibration sample. The models were calibrated with the new dataset ($n = 3359$) and used to predict the soil class. Predictions were validated with the independent probability sample data. Overall purities (60.2% for MLR, 60.4% for CT, 60.3% for RF) were larger than the overall purity of the reconnaissance map but smaller than the purities obtained with the full dataset. Interestingly, the external map purities were larger than the internal and cross-validation purities. These results indicate that the effect of sampling density and the selection of sampling locations on prediction accuracy is not trivial, see for example Odgers et al. (2014) and Nauman and Thompson (2014), and merits further attention.

5. Conclusions

In this study we refined a reconnaissance soil map at scale 1:250,000 by calibrating models with data from the same map by sampling the soil map at a large number of locations. The validation of the new and old maps with independent probability sample data showed that the accuracy of the reconnaissance soil map could be improved without additional sampling by only utilizing the relationship between soil type and covariates as calibrated on the existing soil map. It could be interesting to test this method in other parts of the world where only reconnaissance soil maps are available (e.g., large parts of Africa and South-America). For France, we plan to extend this method to the whole Haute-Normandie region.

Acknowledgements

This work has been financed by the French Programme of Soil Mapping IGCS, funded by the French Ministry for Agriculture (research contract 33000186). We thank Eugénie Tientcheu, Joseph Levillain, Cyrielle Berché, Nicolas Soler-Dominguez and Solène Tonneau for technical help in the field and for sample preparation. Warm thanks are expressed to Pr Clément Mathieu for his help in Soil Taxonomy references.

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version, at http://dx.doi.org/10.1016/j.geodrs.2014.07.001. These data include Google maps of the most important areas described in this article.

References

Adhikari, K., Minasny, B., Greve, M.B., Greve, M.H., 2014. Constructing a soil class map of Denmark based on the FAO legend using digital techniques. Geoderma 214–215, 101–113.
Behrens, T., Förster, H., Scholten, T., Steinrücken, U., Spies, E.-D., Goldschmitt, M., 2005. Digital soil mapping using artificial neural networks. J. Plant Nutr. Soil Sci. 168 (1), 21–33.
Breiman, L., 2001. Random forest. Mach. Learn. 45, 5–32.
Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Chapman & Hall, New York.
Brus, D.J., Kempen, B., Heuvelink, G.B.M., 2011. Sampling for validation of digital soil maps. Eur. J. Soil Sci. 62 (3), 394–407.
Carré, F., Girard, M.C., 2002. Quantitative mapping of soil types based on regression kriging of taxonomic distances with landform and land cover attributes. Geoderma 110 (3–4), 241–263.
Commission of the European Community, 1993. CORINE Land Cover. p. 144 (European Community Publications, Bruxelles, Belgium).
de Gruijter, J.J., Brus, D.J., Bierkens, M.F.P., Knotters, M., 2006. Sampling for Natural Resource Monitoring. Springer, New York.
Debella-Gilo, M., Etzelmüller, B., 2009. Spatial prediction of soil classes using digital terrain analysis and multinomial logistic regression modeling integrated in GIS. Examples from Vestfold County, Norway. Catena 77 (1), 8–18.
Grinand, C., Arrouays, D., Laroche, B., Martin, M.P., 2008. Extrapolating regional landscapes from an existing soil map: sampling intensity, validation procedures, and integration of spatial context. Geoderma 143 (1–2), 180–190.
Häring, T., Dietz, E., Osenstetter, S., Koschitzki, T., Schröder, B., 2012. Spatial disaggregation of complex soil map units: a decision-tree based approach in Bavarian forest soils. Geoderma 185–186, 37–47.
Hartemink, A.E., 2008. Soil map density and a nation's wealth and income. In: Hartemink, A.E., McBratney, A.B., Mendonca, L. (Eds.), Digital Soil Mapping with Limited Data. Springer, Dordrecht, pp. 53–66.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning. Data Mining, Inference, and Prediction, Second ed. Springer, New York.
Heung, B., Bulmer, C.E., Schmidt, M.G., 2014. Predictive soil parent material mapping at a regional-scale: a random forest approach. Geoderma 214–215, 141–154.
Hosmer, D.W., Lemeshow, S., 2000. Applied Logistic Regression, 2nd ed. John Wiley & Sons, New York.
Kempen, B., Brus, D.J., Heuvelink, G.B.M., Stoorvogel, J.J., 2009. Updating the 1:50,000 Dutch soil map using legacy soil data: a multinomial logistic regression approach. Geoderma 151(3–4), 311–326.
Kempen, B., Brus, D.J., Stoorvogel, J.J., Heuvelink, G.B.M., De Vries, F., 2012a. Efficiency comparison of conventional and digital soil mapping for updating soil maps. Soil Sci. Soc. Am. J. 76 (6), 2097–2115.
Kempen, B., Brus, D.J., Heuvelink, G.B.M., 2012b. Soil type mapping using the generalized linear geostatistical model: a case study in a Dutch cultivated peatland. Geoderma 189–190, 540–553.
Lagacherie, P., Legros, J.P., Burrough, P., 1995. A soil survey procedure using the knowledge of soil pattern established on a previously mapped reference area. Geoderma 65 (3–4), 283–301.
Lemercier, B., Lacoste, M., Loum, M., Walter, C., 2012. Extrapolation at regional scale of local soil knowledge using boosted classification trees: a two-step approach. Geoderma 171–172, 75–84.
Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2 (3), 18–22.
MacMillan, R.A., Moon, D.E., Coupe, R.A., 2007. Automated predictive ecological mapping in a Forest Region of B.C., Canada, 2001–2005. Geoderma 140 (4), 353–373.
Marsman, B.A., de Gruijter, J.J., 1986. Quality of Soil Maps. A Comparison of Survey Methods in a Sandy Area. Soil Survey Papers, 15. Netherlands Soil Survey Institute, Wageningen.
Masson, V., Champeaux, J.-L., Chauvin, F., Meriguet, C., Lacaze, R., 2003. A global database of land surface parameters at 1 km resolution in meteorological and climate models. J. Clim. 16 (9), 1261–1282.
McBratney, A.B., Mendonça, M.L., Minasny, B., 2003. On digital Soil Mapping. Geoderma 117, 3–52.
Nauman, T.W., Thompson, J.A., 2014. Semi-automated disaggregation of conventional soil maps using knowledge driven data mining and classification trees. Geoderma 213, 385–399.
Odgers, N.P., Sun, W., McBratney, A.B., Minasny, B., Clifford, D., 2014. Disaggregrating and harmonising soil map units through resampled classification trees. Geoderma 214–215, 91–100.
Quesnel, F., Couëffé, R., Duriez, M., Lasseur, E., 2007. Carte géologique harmonisée du département de la Seine-Maritime. BRGM/RP-56185-FR (118 pp.).
R Development Core Team, 2013. R: A Language for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (URL: http://www.R-project.org).
Richer de Forges, A.C., Arrouays, D., 2010. Analysis of requests for information and data from a national soil data centre in France. Soil Use Manag. 26 (3), 374–378.
Stehman, S.V., 1997. Selecting and interpreting measures of thematic classification accuracy. Remote Sens. Environ. 62 (1), 77–89.
Strobl, C., Malley, J., Tutz, G., 2009. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychol. Methods 14 (4), 323–348.
Subburayalu, S.K., Jenhani, I., Slater, B.K., 2014. Disaggregation of component soil series on an Ohio County soil survey map using possibilistic decision trees. Geoderma 213, 334–345.
Therneau, T.M., Atkinson, E.J., 2013. An introduction to recursive partitioning using the RPART routines. Technical Report. Mayo Foundation (http://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf).
Van Lint, J., Giot, D., Callec, Y., 2003. Carte géologique harmonisée du Département de L'Eure. BRGM/RP-52766-FR (97 pp.).
van Zijl, G.M., Brouwer, D., van Tol, J.J., le Roux, P.A.L., 2014. Functional digital soil mapping: a case study from Namarroi, Mozambique. Geoderma 219–220, 155–161.
Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S, Fourth ed. Springer, New York.

Webster, R., McBratney, A.B., 1989. On the Akaike Information Criterion for choosing models for variograms of soil properties. J. Soil Sci. 40 (3), 493–496.
Wiesmeier, M., Barthold, F., Blank, B., Kögel-Knaber, I., 2011. Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. Plant Soil 340 (1), 7–24.
Wolf, C., Cintrat, J.L., Hurard, M., Lechevalier, C., Ouvry, J.F., Scheurer, O., 1998. Carte des sols de Haute-Normandie. Serda Editions, Bois-Guillaume, France.
Yang, L., Jiao, Y., Sherif, F., Zhu, A.-X., Hann, S., Burt, J.E., Qi, F., 2011. Updating conventional soil maps through digital soil mapping. Soil Sci. Soc. Am. J. 75 (3), 1044–1053.
Zhu, A.-X., Hudson, B., Burt, J., Lubich, K., Simonson, D., 2001. Soil mapping using GIS, expert knowledge, and fuzzy logic. Soil Sci. Soc. Am. J. 65 (5), 1463–1472.