Cartographie numérique d'une carte pédologique au 1/50 000 dans le Doubs, France

S. Lehmann, M. Eimberck, M. P. Martin et D. Arrouays

INRA, US 1106 INFOSOL, F-45075 Orléans, France

*: Auteur correspondant: sebastien.lehmann@orleans.inra.fr

RÉSUMÉ

Le développement récent des outils de segmentation d'image et de classification par des arbres de régression boostés offre de nouvelles perspectives pour la cartographie numérique prédictive des sols. L'accès à un grand nombre d'observations pédologiques via la base de données Donesol couplé à celui des données environnementales facilement accessibles (MNT et ses dérivées, carte géologique, carte d'occupation du sol) nous permet de produire une carte prédictive des sols avec une résolution de l'ordre du 1/50 000, validée par des observations ponctuelles. Sur le secteur de Vercel (Jura, France), la démarche originale présentée ici a permis de cartographier une zone dont ¼ seulement de la surface avait fait l'objet d'une cartographie avec des méthodes classiques de prospection sur le terrain. 2348 observations ponctuelles réparties quant à elles sur la presque totalité du 1/50 000 de Vercel ont permis d'ajuster et de valider le modèle.

Nous avons mené deux approches parallèles en prenant comme données de calibration et de validation, d'une part, des données pédologiques ponctuelles qui font référence à des types de sols, et, d'autre part, des données pédologiques surfaciques qui font référence à des unités cartographiques de sols (UCS) plus ou moins complexes. Les modèles ont été définis à partir d'un outil disponible sous R faisant appel à des arbres de régressions multiples boostés, MART (Friedmann, 2002). Les cartes de prédiction obtenues ont été lissées avec un filtre majoritaire afin de produire une carte avec des limites d'UCS bien définies.

La précision globale des prédictions donne des résultats très satisfaisants au niveau des validations internes (entre 80 et 90 %) pour les deux approches. La validation externe, conduite seulement sur l'approche surfacique avec uniquement des données surfaciques, est comprise entre 40 et 50 % selon l'ajustement. Une carte des incertitudes est également produite avec l'approche surfacique et permet de déterminer les zones où le modèle s'affaiblit. Les données ponctuelles ont alors servi à affiner la faiblesse de la prédiction par une approche d'expertise. La classification produite avec l'approche surfacique montre en effet des UCS bien définies au niveau des zones

Reçu: décembre 2012 ; Accepté: avril 2013

d'apprentissage mais relativement bruitées au niveau des zones vierges. La classification obtenue avec l'approche ponctuelle produit un résultat très pixellisé avec un mélange de classes de sols assez complexe localement mais bénéficiant d'une bonne précision globale de classification comme indiqué sur la carte des incertitudes.

Enfin, l'expertise du pédologue a permis de synthétiser les 2 cartes ainsi produites afin d'obtenir *in fine* une carte choroplèthe des unités cartographiques de sols. Les limites des unités cartographiques ont été obtenues à partir de la carte issue des données surfaciques, et leur contenu est dérivé de la carte issue des données ponctuelles. Le gain de temps par rapport à une approche conventionnelle est estimé entre 70 et 80 jours pour 36000 ha. Mais l'approche numérique seule ne suffit pas. Il reste indispensable de s'approprier le terrain avant la modélisation et de valider ses prédictions par des nouvelles observations.

Mots clés

Cartographie numérique des sols, cartographie prédictive des sols, arbre de régressions multiples boostés, analyse de modèle numérique de terrain multi-échelle, matrice de confusion.

SUMMARY

DIGITAL SOIL MAPPING OF A SOIL CLASSES MAP AT 1:50,000 SCALE IN THE DOUBS DEPARTMENT, FRANCE

Recent advances in segmentation processing and in classification using boosted regression trees enables new prospects for predictive digital soil mapping. The availability of numerous soil data through the French soil database Donesol, and to easy to access environmental co-variates (DEM and derivatives, geological maps, land cover maps…) enables to produce a predicted soil type map, at a scale of about 1:50,000, validated by point data in the area of Vercel (Jura, France). On this area, the approach we detail here led to mapping a surface ¼ of which only was previously mapped. 2348 point data scattered over the whole territory were used to calibrate and validate the model. Firstly, we produced a predictive map from point data. Half of these data was used to calibrate a model using boosted regression trees. The remaining half was used for validation. We tested 8 iterations using data integrating more and more large spatial domains. The derivates from the DEM were averaged using circular windows of growing size (diameters from 30 to 1800 m). The resulting map was affected by some noise that we removed using a filter based on the dominant classes in order to obtain a map with clear abrupt limits. Secondly, we used the same approach taking as calibration data the soil units (using a buffer around the limits in order to get "pure" pixels).

Lastly the soil surveyor expertise was used to produce a synthesis of these two predictions to obtain a chloropleth map of soil units. The time saved by this approach in comparison to a classical one is estimated to be about 70-80 days for 36,000 ha. DSM alone is not self-sufficient, knowledge of the terrain and external validation remain essential.

Key-words

Digital soil mapping, predictive soil mapping, boosted classification tree, multi-scale digital terrain analysis, confusion matrix.

RESUMEN

CARTOGRAFÍA NUMÉRICA DE UN MAPA PEDOLÓGICO AL 1/50 000 EN EL DOUBS, FRANCIA

El desarrollo reciente de herramientas de segmentación de imagen y de clasificación por arboles de regresión con "boosting" ofrece nuevas perspectivas para la cartografía numérica predictiva de los suelos. El acceso a un gran número de observaciones pedológicas vía la base de datos Donesol acoplado a los de los datos ambientales fácilmente accesibles (MNT y sus derivadas, mapa geológico, mapa de uso del suelo) nos permite producir un mapa predictivo de suelos con una resolución del orden del 1/50 000, validado por observaciones puntuales. En el sector de Vercel (Jura, Francia), el enfoque original presentado aquí permitió cartografiar una zona cuya solamente ¼ de la superficie había hecho objeto de una cartografía con métodos clásicos de prospección en el terreno. 2348 observaciones puntuales repartidas sobre la casi totalidad del 1/50 000 de Vercel permitieron ajustar y validar el modelo.

Conducimos dos enfoques paralelos tomando como datos de calibración y de validación, de una parte datos pedológicos puntuales que hacen referencia a tipos de suelos, y, de otra parte datos pedológicos surfacicos que hacen referencia a unidades cartográficas de suelos (UCS) más o menos complejas. Se definieron los modelos a partir de un herramienta disponible bajo R que utiliza arboles de regresiones múltiples con "boosting", MART (Friedmann, 2002). Los mapas de predicción obtenidos fueron suavizados con un filtro mayoritario a fin de producir un mapa con límites de UCS bien definidos.

La precisión global de las predicciones da resultados muy satisfactorios al nivel de las validaciones internas (entre 80 y 90%) para los dos enfoques. La validación externa, conducida solamente sobre el enfoque surfacico con únicamente datos surfacicos, está compren-

dida entre 40 y 50 % según el ajustamiento. Un mapa de incertidumbres está igualmente producido con el enfoque surfacico y permite determinar las zonas donde el modelo se debilita. Los datos puntuales sirvieron entonces para afinar la debilidad de la predicción con un enfoque de pericia. La clasificación obtenida con el enfoque surfacico muestra en efecto UCS bien definidas al nivel de las zonas de aprendizajes, pero relativamente con ruido de fondo al nivel de las zonas vírgenes. La clasificación obtenida con el enfoque puntual produce un resultado muy pixellizado con una mezcla de clases de suelos bastante compleja localmente pero que beneficia de una buena precisión global de clasificación como indicado en el mapa de incertidumbres.

En fin, la pericia del pedólogo permitió sintetizar los 2 mapas así producidos a fin de obtener in fine un mapa coroplético de unidades cartográficas de suelos. Los límites de las unidades cartográficas fueron obtenidos a partir del mapa resultante de los datos surfacicos, y su contenido esta derivado del mapa resultante de los datos puntuales. El ahorro de tiempo en relación con un enfoque convencional está estimado entre 70 y 80 días para 36 000 ha. Pero el enfoque numérico solo no está suficiente. Queda indispensable apropiarse el terreno antes la modelización y validar sus predicciones por nuevas observaciones.

Palabras clave

Cartografía numérica de suelos, cartografía predictiva de suelos, árbol de regresiones múltiples con "boosting", análisis de modelo numérico de terreno multiescala, matriz de confusión.

Dans le contexte actuel de disponibilité des données géographiques numériques, il est désormais possible d'avoir facilement accès à de nombreux paramètres environnementaux comme le relief à l'échelle mondiale (modèle numérique d'élévation ASTER au pas de 30 m), l'occupation du sol à l'échelle européenne, ou la géologie avec le portail Infoterre du BRGM. La directive européenne Inspire donne par ailleurs un coup d'accélérateur pour recenser et donner accès à toutes les données environnementales notamment pour les organismes de recherche publique.

En France, la disponibilité des données pédologiques numériques est centralisée au niveau de l'unité INRA InfoSol (Orléans) qui coordonne la mise en œuvre des Référentiels Régionaux Pédologiques (cartes au 1/250 000) et des cartes pédologiques à des échelles plus fines, principalement au 1/100 000. InfoSol administre également la base de donnée DoneSol (Grolleau *et al*., 2004) qui intègre les données surfaciques des RRP[1] mais aussi plusieurs dizaines de milliers de profils et de sondages issus des études pédologiques référencées à ce jour.

Dans un même temps, l'accessibilité aux variables environnementales sur l'ensemble du territoire ainsi que la présence de données pédologiques fiables et normalisées sur de nombreuses régions rend désormais possible le recours à des méthodes de cartographie numérique. En effet, les modèles implicites de distribution spatiale des sols établis par des cartographes peuvent être traduits et formalisés en utilisant des relations avec des variables environnementales (Lagacherie, 1992; Lagacherie *et al*., 1995 ; Bui *et al*., 1999 ; McKenzie and Ryan, 1999 ; Moran and Bui, 2002 ; Bui and Moran, 2003 ; Lehmann *et al*., 2007 ; Walter *et al*., 2007 ; Grinand *et al*., 2008 ; Laroche *et al*., 2011 ; Lemercier *et al*., 2012). Les relations entre le sol et les facteurs environnementaux ont été conceptualisées et formalisées par MacBratney *et al.* (2003) à travers le modèle scorpan qui fait référence à 7 facteurs de formation des sols:

- s: le sol, décrit comme tel ou certaines de ses propriétés en un point donné,
- c: les propriétés climatiques de l'environnement en un point donné,
- o: les organismes, la végétation, la faune ou l'activité humaine,
- r: le relief ou la topographie,
- p: le matériau parental ou la lithologie,
- a: l'âge ou le facteur temporel,
- n: la position dans l'espace.

Ce modèle conceptuel est largement utilisé pour prédire des types de sols, des propriétés ou des fonctions du sol en prenant en compte les facteurs de la pédogenèse et leurs interactions (Walter *et al*., 2006 ; Mendonça-Santos, 2006; Dobos et Hengl, 2009).

Les techniques de prédiction des propriétés du sol ou des types de sols s'appuient sur des méthodes d'interpolation spatiale comme les approches géostatistiques ou des méthodes d'apprentissage automatique. Les techniques d'interpolation spatiale ont montré d'excellents résultats en cartographie des sols (Bourennane and King, 2003) mais elles requièrent un grand nombre d'observations sur l'ensemble de la zone d'étude. L'apprentissage automatique (machine-learning) fait appel à un algorithme qui détermine, ou du moins explicite, les relations

1 : Référentiels Régionaux Pédologiques

entre une réponse et ses « prédicteurs » dans le cas d'une approche supervisée (Odeh *et al*., 1992).

Les méthodes de classification supervisée ont pour objectif d'extraire un modèle d'organisation spatiale du sol à partir d'un jeu de données d'apprentissage, correspondant à des situations connues. L'extrapolation de ce modèle à des situations inconnues fournit la prédiction recherchée. Ils existe de nombreuses méthodes parmi lesquelles nous pouvons citer: les réseaux de neurones (Boruvka and Penizek, 2006 ; McKenzie and Ryan, 1999) ; les machines à vecteur de support (SVM) encore peu utilisées en cartographie numérique des sols bien qu'ayant montré des résultats très prometteurs (Ballabio, 2009) et souvent supérieurs à ceux obtenus avec les réseaux de neurones ; les techniques de partitionnement utilisant une métrique pour mesurer la distance entre des classes comme par exemple la distance de Manhattan employée dans l'algorithme de CLAPAS (Robbez-Masson, 1994 ; Lehmann *et al*., 2007) ; les arbres de décisions qui permettent de reformuler un problème complexe sous la forme graphique d'un arbre, de façon à faire apparaître à l'extrémité de chaque branche les différents résultats possibles (classes de sols ou propriétés) en fonction des décisions prises à chaque étape, en fonction des facteurs environnementaux. Leur lisibilité, leur rapidité d'exécution et le peu d'hypothèses nécessaires *a priori* expliquent leur popularité actuelle (Anonyme (29 mars 2013), « Arbre de décision », < *http://fr.wikipedia.org/wiki/Arbre_de_décision* >).

Le « boosting » a été appliqué aux arbres de décisions par Friedman [(2001) d'après Freund et Schapire (1996)] pour augmenter leur performance. Lors du boosting, un grand nombre d'arbres de classification sont combinés pas à pas pour optimiser les performances de la prédiction. Les arbres individuels sont ajustés petit à petit et de façon itérative sur les données d'apprentissage pour augmenter l'importance des observations faiblement modélisées par les collections d'arbres existantes (Elith *et al*., 2008).

D'un point de vue historique, les arbres de décisions ont d'abord été employés en écologie en utilisant des données de télédétection (Lees and Ritman, 1991; Michaelsen *et al*., 1994; Pal and Mather, 2003 ; Lawrence *et al*., 2004 ; Moisen *et al*., 2006). En raison de leurs avantages et de leur potentiel pour la modélisation de la répartition spatiale des sols, les arbres de décisions sont de plus en plus employés pour prédire des types ou des classes de sols (Scull *et al*., 2005 ; Minasny and McBratney, 2007 ; Grinand *et al*., 2008; Lemercier *et al*., 2012).

Initialement développés par Breiman *et al*. (1984), les modèles basés sur de arbres de décision ont donné naissance à des modèles aujourd'hui implémentés sous R comme le modèle CART (Classification and Regression Trees) (Friedman, 2002; Friedman and Meulman, 2003). Le modèle MART (Multiple Additive Regression Tree) est une extension du modèle CART développée sous R. D'après Friedman et Meulman (2003), les arbres de décisions présentent de nombreux avantages pour la classification supervisée. En effet, ils peuvent traiter des données en entrée de n'importe quel type (qualitative, quantitative, binaire, etc.) et de la même façon. Ils sont non paramétriques, de sorte qu'aucune hypothèse n'est requise au niveau de la distribution des variables et qu'il n'est pas nécessaire de transformer les données. Les arbres ne sont sensibles ni aux valeurs manquantes, ni aux valeurs extrêmes, ni aux données sans relation avec la variable réponse. Les relations entre les facteurs déterminants sont prises en compte sans connaissance *a priori*. Ces modèles ne sont pas non plus sensibles aux différentes unités de mesures des variables prédictives. Enfin, les arbres de décisions peuvent être employés avec un jeu de données de calibration relativement limité et les résultats sont assez explicites et plus faciles à interpréter que ceux issus des autres modèles cités plus haut.

Le travail présenté ici porte sur un essai de cartographie numérique sur la feuille à 1/50 000 de Vercel (quart Nord-Est de la feuille à 1/100 000 de Besançon). Nous avons mené deux approches parallèles en prenant comme données de calibration et de validation, d'une part, des données pédologiques ponctuelles qui font référence à des types de sols et, d'autre part, des données pédologiques surfaciques qui font référence à des unités cartographiques de sols plus ou moins complexes.

Nous avons ainsi employé l'algorithme MART pour extrapoler les observations ponctuelles et surfaciques sur l'ensemble de la carte. Nous avons ensuite validé ces résultats par l'expertise du pédologue et enfin les avons synthétisés.

MATÉRIEL ET MÉTHODE

Zone d'étude

La zone d'étude correspond à la feuille de Vercel d'après le découpage au 1/50 000 des feuilles topographiques de l'IGN. Elle est située au centre du département du Doubs *(figure 1a)* et couvre 55 000 ha. Éloigné de l'influence régulatrice de l'océan, le climat du Doubs a une forte influence continentale : neige et fortes gelées l'hiver, sécheresse et chaleur l'été, ponctuées par des pluies pouvant être orageuses. La principale particularité du climat de ce département est sa grande variabilité aussi bien au cours d'une saison que d'une année sur l'autre. Les précipitations moyennes sur la carte sont comprises entre 700 et 1300 mm et la température moyenne annuelle entre 7,9 °C et 10,2 °C avec une amplitude annuelle maximale pouvant atteindre 50 °C (source: Météo France depuis 1960). Le vent dominant (27,5 %) est de secteur sud-ouest à sud/sud-ouest, c'est « le vent » venu de l'Atlantique et qui apporte la pluie. Un vent secondaire (13 %) souvent intense et desséchant, « la bise », vient du secteur nord-est.

La carte de Vercel se situe sur trois plateaux qui forment le flanc ouest du massif du Jura, séparés entre eux par des

faisceaux plissés. Du Nord-Ouest vers le Sud-Est se succèdent le plateau bisontin (180 à 300 m), le faisceau bisontin, le plateau de Montrond (300 à 380 m), le faisceau de Passavant et enfin le plateau d'Amancey-Vercel (500 à 800 m). La topographie est assez contrastée avec une alternance de reliefs faiblement ondulés sur les plateaux, localement très karstifiés avec des successions de dolines, et d'un relief beaucoup plus marqué sur les faisceaux plissés avec des pentes localement très fortes, voire de petites falaises.

Faisant partie intégrante du massif jurassien, la géologie de la zone d'étude est relativement complexe. Elle est principalement composée de calcaires durs des étages Bajocien, Bathonien, Oxfordien et Rauracien alternant avec des marnes du Bathonien supérieur. Au Quaternaire, plusieurs matériaux d'épandage sont venus recouvrir ces formations: des dépôts d'origine glaciaire, des alluvions fluviatiles, des dépôts de versants et des limons éoliens (Dreyfuss, 1965). L'origine de ces derniers, leur distribution et leur épaisseur influencent fortement les propriétés du sol (Bruckert, 1987).

Dans le contexte franc-comtois, où sont généralement juxtaposés brutalement des calcaires fissurés et des marnes, la distinction entre les sols au drainage naturel favorable sur roches perméables et les sols soumis à des épisodes d'engorgement et d'anoxie a été le premier critère pris en compte par le pédologue qui a initié les travaux de cartographie sur l'ensemble du 1/100 000 de Besançon (Bruckert, 1987). La position géomorphologique définit d'abord la situation des sols dans le paysage: position de plateau, de versant, de vallée. Elle précise ensuite la situation dans l'unité géomorphologique: zone convexe (plus sensible à l'érosion) ou dépression (pouvant recevoir des matériaux), terrasse alluviale ou surface de divagation des rivières, etc. La distribution des sols est ainsi reliée à leur situation géomorphologique et à la nature des roches sous-jacentes. Sur la feuille de Vercel, 61 unités cartographiques de sols plus ou moins complexes ont été décrites (Bruckert, 1987). Elles ont été regroupées en 32 unités qui correspondent à des types de sols identifiés à partir des observations ponctuelles. Les principales unités sont détaillées dans le paragraphe 2.2.1. « Les données pédologiques ».

Sur la zone d'étude, le mode d'occupation des sols est fortement lié aux contraintes agronomiques. On peut ainsi distinguer (Gaiffe *et al*., 2013):
- les forêts (principalement feuillues): zones de contraintes fortes pour l'agriculture (pierrosité ou hydromorphie) ;
- le bocage avec pâtures: contraintes à la mise en culture pour diverses raisons parmi lesquelles:
* sol trop superficiel, trop caillouteux ou trop séchant pour être cultivé ;
* sol oligotrophe issu de matériaux acides ;
* sol trop hydromorphe pour permettre la mécanisation ;
- le bocage composé de cultures et de prairies: contraintes circonscrites (bancs calcaires affleurant, épierrage ou dolines) ;
- les grandes cultures: peu ou pas de contraintes ; sols profonds, ou superficiels non pierreux, ou faiblement hydromorphes ;
- les prairies de fauche: sols superficiels ou faiblement hydromorphes.

Le contexte historique

La réalisation de la carte des sols à 1/100 000 de Besançon a démarré en 1980 et fut brutalement interrompue par la disparition du Professeur Sylvain Bruckert (1993) et la réorganisation du Laboratoire de Pédologie de l'Université de Besançon. A ce moment, 3 des 4 coupures à 1/50 000 composant la feuille à 1/100 000 avaient été finalisées: Besançon, Ornans et Quingey. Les levers systématiques sur celle de Vercel étaient réalisés pour 2/5 de sa surface. Afin de sauvegarder et de valoriser cet important acquis de connaissances, l'Unité InfoSol de l'INRA a décidé, dans le cadre de son programme Connaissance Pédologique de la France, de finaliser ce travail d'inventaire, en complétant la carte de Vercel à 1/50 000, puis en synthétisant l'ensemble des 4 coupures à l'échelle du 1/100 000, en accord avec la légende préparée par les auteurs. Dans la mesure où il n'était pas envisagé de réaliser une prospection systématique dans les secteurs non levés, une méthode de cartographie numérique a été mise en œuvre pour finaliser la carte pédologique au 1/50 000 de Vercel.

Les données numériques

Les données pédologiques

Deux types de données étaient disponibles: une minute de carte pédologique au 1/50 000 (levée au 1/25 000) couvrant 2/5 de la surface d'étude ainsi que 2 348 profils et sondages répartis principalement sur la moitié ouest de la carte *(figure 1a)*. Les observations ponctuelles qui ont permis d'élaborer cette carte ont été réalisées par différents cartographes de l'Université de Besançon. Elles ont été décrites selon un système de classification morpho-édaphique élaboré par Bruckert (1987). Dans un premier temps, le volume prospectable par les racines est défini par l'épaisseur de terre située au-dessus du niveau d'apparition d'un obstacle qui s'oppose à l'enracinement (profondeur rhizofonctionnelle). Dans un deuxième temps, la qualité du milieu ainsi offert aux racines est précisée, qui dépend de la porosité, de la texture et des éventuelles contraintes. Ces observations sont synthétisées sous la forme d'un sigle comportant: la texture des horizons prospectés par les racines, la nature de l'obstacle à l'enracinement et sa profondeur d'apparition et enfin les éventuelles contraintes (anoxie, pierrosité, sensibilité à la sécheresse ou à l'érosion, etc.), et rattachées à des types morphogénétiques de sol. Ensuite, ces données ponctuelles ont été géoréférencées et

Figure 1a - Carte de situation de la zone cartographiée et répartition des données ponctuelles pour l'apprentissage et la validation.
Figure 1a - *Location of the study area, training and validation area used with point data.*

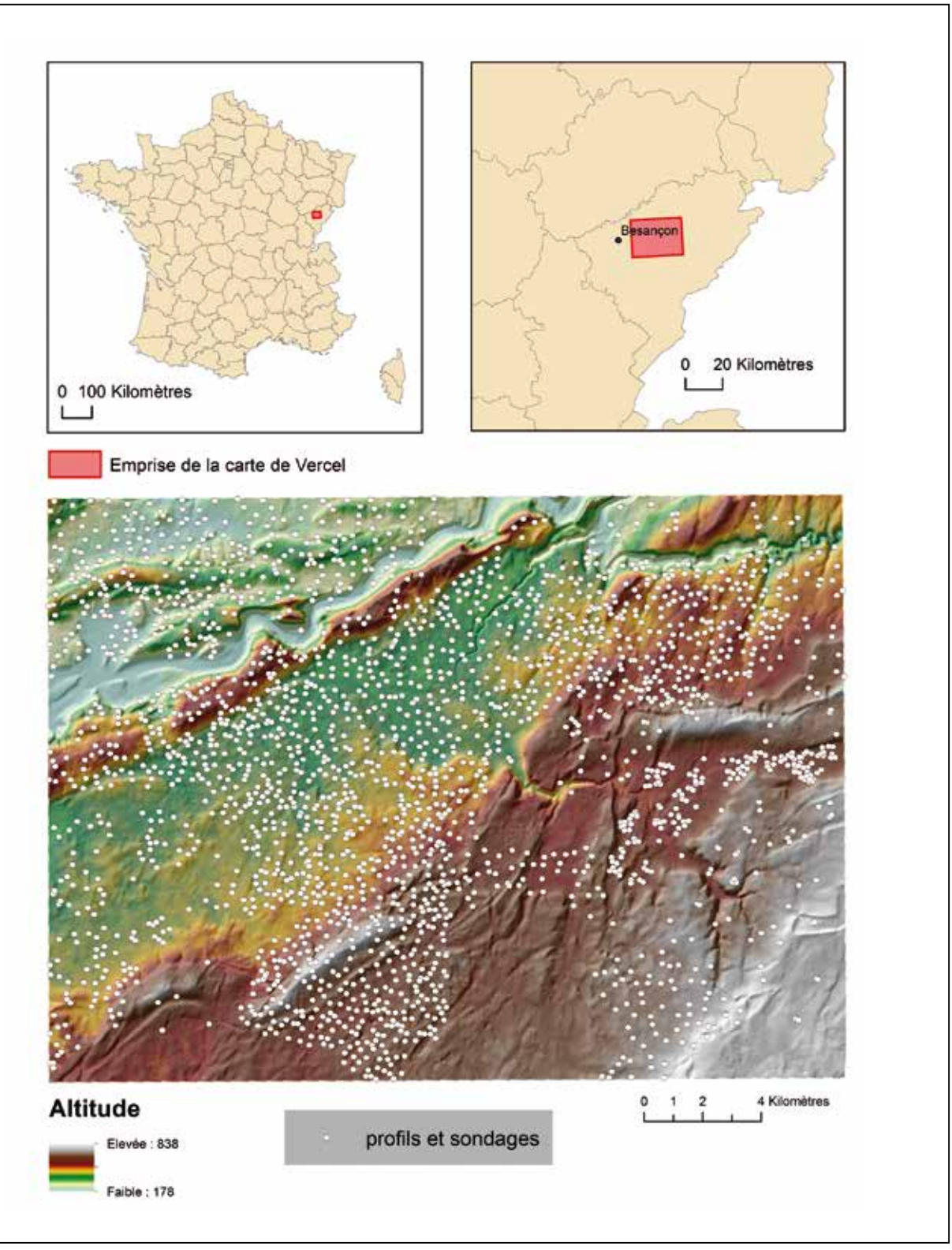
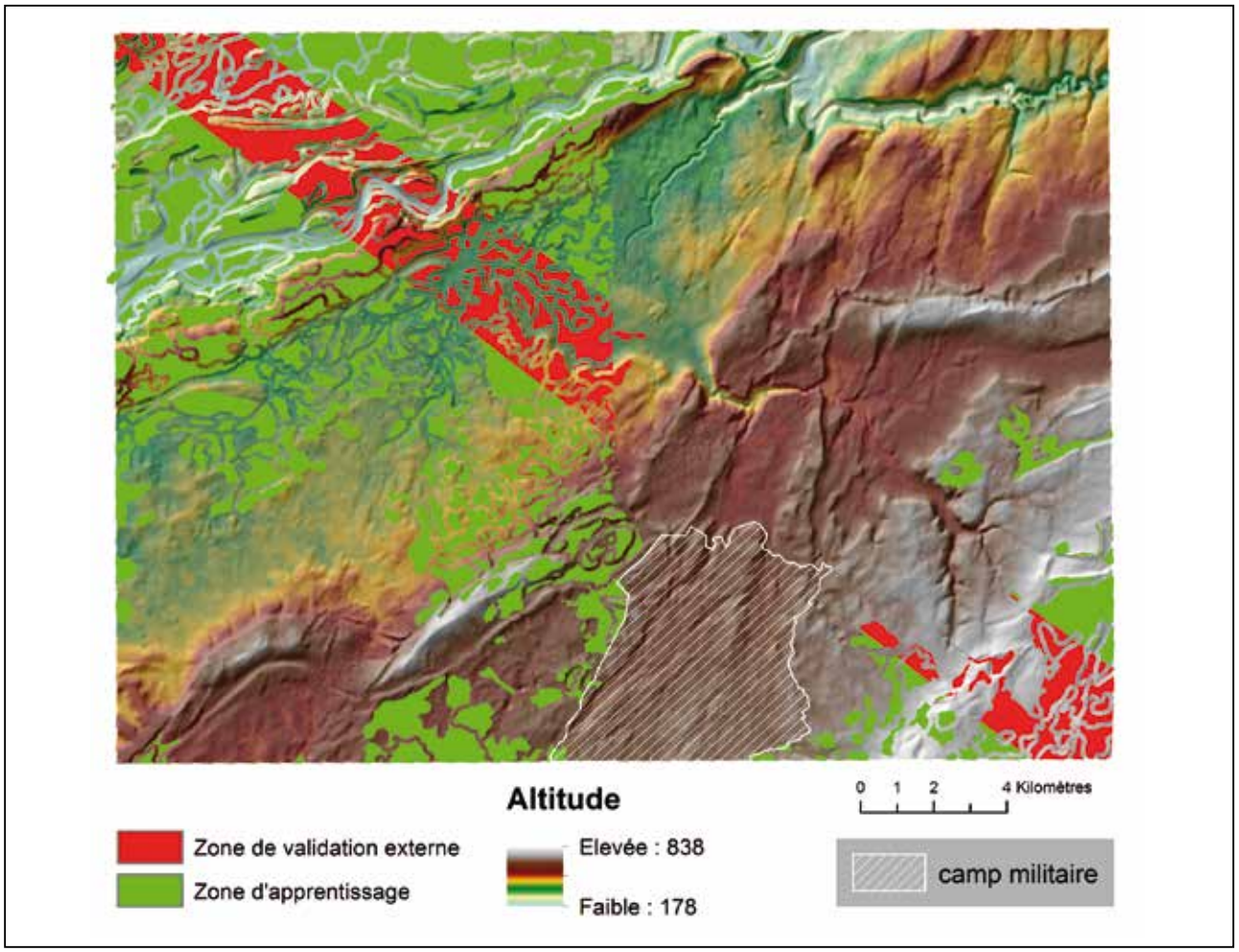


Figure 1b - Carte de répartition des données surfaciques utilisées pour l'apprentissage et la validation.
Figure 1b - *Map of the training and validation area used with surface data.*



rattachées à des unités typologiques issues des 4 cartes à 1/50 000 qui composent le 100 000e de Besançon. Les données environnementales explicatives dominantes sur un cercle de 100 m de diamètre (ce qui représente 3, 4 ou 5 pixels au pas de 30 m, selon l'emplacement) autour des observations ont ensuite été extraites.

Parmi les types de sol rencontrés, on distingue d'après le Référentiel Pédologique (Baize et Girard, 2009) :

A - Les sols aérés issus de roches fragmentées ou d'altérites autochtones non calcaires

1 - Sur les plateaux en position sommitale (formes convexes ou planes): des Calcosols, des Calcisols, des Brunisols Eutriques, souvent leptiques et mésosaturés (aussi appelés sols bruns à pellicule calcaire par les auteurs des cartes de référence), des Brunisols fersiallitiques (formations d'aspect rougeâtre que l'on trouve soit en recouvrement sur des calcaires durs soit piégées dans des fissures diaclasiques et des poches karstiques), des Brunisols bilithiques (apports de limons éoliens quaternaires).

2 - Sur les plateaux, en position dépressionnaire (formes concaves), des Colluviosols brunifiés.

3 - Sur les plateaux supérieurs, les sols sont plus superficiels et s'apparentent à des Rendosols, des Rendisols, des Calcosols et des CAlcisols leptiques. Dans les petites cuvettes, on peut trouver localement des Réductisols humiques ou des Histosols.

B - Les sols hydromorphes issus de roches non fracturées:

On retrouve le même cortège de types de sols que précédemment avec l'apparition de Rédoxisols voire de Réductisols en position topographique dépressionnaire dans les vallons à ruisseaux permanents ou temporaires.

C - Les sols bien drainés issus d'altérites acides ou de limons sur substrats rocheux fissurés

Ils se localisent au niveau des faisceaux plissés et des reculées[1]. On y observe principalement des Brunisols Dystriques et des Alocrisols Typiques.

D - Les sols des versants à pentes très fortes (>50%) et fronts de bancs:

On observe des Rendosols colluviaux, des Colluviosols hémi-organiques, humifères ou calcaires, au niveau des thalwegs et des Lithosols sur les parties érodées.

E - Les sols des terrasses du lit majeur du Doubs:

Ces sols comprennent des Fluviosols Brunifiés ou Typiques et des Colluviosols brunifiés complexes.

Ces Unités Typologiques de Sol (UTS) s'organisent de façon plus ou moins complexe au sein des unités cartographiques. Les UCS ont été rasterisées sur la grille de référence du MNE ASTER au pas de 30 m. On a ainsi produit un raster de 32 unités simples pour la prédiction réalisée à partir des données ponctuelles et un raster de 61 unités complexes pour celle réalisée à partir des données surfaciques. Les données fournies par les cartes et observations existantes ont été complétées par deux campagnes de terrain successives d'une semaine chacune, qui ont permis de s'approprier les lois de répartition des sols dans le paysage et de compléter les zones non couvertes par les observations ponctuelles existantes.

Les données prédictives

Elles sont issues de 15 variables environnementales quantitatives et qualitatives *(tableau 1)*, déclinées en 46 variables par un jeu de filtres de convolution de diamètres différents.

On peut distinguer deux grandes catégories de variables: d'une part, les données dérivées du MNE ASTER[2] au pas de 30 m, généralement quantitatives et, d'autre part, les données environnementales issues de la géologie ou de l'occupation du sol.

Les indices topographiques sont généralement des variables pertinentes pour prédire l'organisation des sols dans le paysage (Scull, 2003). Nous avons retenu cinq indices directement dérivés du MNE: l'altitude, la pente, la courbure verticale et horizontale et l'orientation. Trois autres indices ont également été pris en compte en raison de leur capacité d'intégration spatiale. Il s'agit du « focal variety », du « focal standard » et du « focal range ». La fonction « focal variety » d'arcinfo a été utilisée pour calculer, pour chaque pixel, le nombre de valeurs uniques d'altitudes à l'intérieur d'un cercle de rayon donné et le rapporte à la position du pixel correspondant. La fonction « focal range » a été utilisée pour calculer, pour chaque pixel, l'étendue des valeurs des altitudes à l'intérieur d'un cercle de rayon donné et le rapporte à la position du pixel correspondant. Cette fonction donne une bonne indication sur la rugosité du relief dans un voisinage choisi. La fonction « focal standard » est assez proche puisqu'elle renvoie l'écart type des valeurs de l'altitude à l'intérieur d'un cercle de rayon donné et le rapporte à la position du pixel correspondant.

Moran et Bui (2002), Grinand *et al*. (2008) et Behrens *et al*. (2010) ont montré l'intérêt d'employer des filtres de convolution sur des indices topographiques en plus des indices bruts pour augmenter sensiblement le résultat des prédictions en prenant en compte différentes échelles de structuration spatiale. Nous avons par conséquent appliqué un filtre de convolution d'un diamètre variable de 3 à 60 pixels, soit de 90 à 1800 m pour certains indices (altitude, courbure, pente et focal) tels que décrits dans le *tableau 1*. Les indices calculés avec des tailles de fenêtres variables sont testés de façon itérative *(tableau 1),* pour les prédictions à partir des données ponctuelles et surfaciques. Les meilleurs indices sont retenus sur la base d'une validation interne, c'est-à-dire par comparaison des valeurs prédites avec les valeurs observées sur la zone d'apprentissage. Le modèle est ensuite extrapolé à l'ensemble de la carte. Le résultat de ces prédictions étant relativement bruité, l'image produite par MART a ensuite été lissée avec un filtre de type « focal majority ».

Les données de validation

La zone utilisée pour la validation interne effectuée sur le jeu de données surfaciques est représenté sur la *figure 1b* par les plages rouges et vertes. Autrement dit, l'ensemble des données surfaciques disponibles a été utilisé pour cette validation. Nous avons pris au hasard 20 % des pixels répartis de façon aléatoire au sein de ces plages pour ajuster le modèle et nous l'avons validé avec les 80 % restants.

Le modèle MART

Les prédictions se sont dans un premier temps appuyées sur le modèle MART (Friedman and Meulman, 2003). L'algorithme fonctionne de manière itérative en affinant puis combinant des arbres de décision pour produire une classification. Pour ajuster le modèle, nous avons pris un échantillon aléatoire de 20 % des pixels de chaque classe de la zone d'apprentissage pour les données surfaciques et 50 % pour les données ponctuelles. Cette méthode garantit que même les classes peu représentées sont prises en compte (Grinand *et al*., 2008; Moran and Bui, 2002). Elith *et al*. (2008) rapportent que le paramètre le plus influent dans les arbres de classification boostés est le taux d'apprentissage. Le taux d'apprentissage ainsi que la fraction du jeu de données prise aléatoirement pour chaque itération sont les deux paramètres qui déterminent le nombre d'arbres nécessaires pour une prédiction optimale. Les résultats d'une

1 Vallée ou canyon de courte dimension, creusé profondément en bordure d'un causse et se terminant vers l'amont par un amphithéâtre (Conseil international de la langue française, 1979)
2 ASTER GDEM is a product of METI and NASA

Tableau 1 - Variables environnementales utilisées pour prédire les types de sols. * Moyenne (étendue) : valeurs sans filtre de convolution ** Moyenne (étendue) : valeurs avec un filtre de convolution de 60 pixels. *** Type: « c » pour continues et « d » pour discrètes.

Table 1 - *Environment predictors used to predict the soil types. * average: value without convolution filter. ** average: value with 60 pixels convolution filter.*

Nom	Description (unité)	Diamètre des filtres de convolution en pixels (30x30 m)	Type***	Moyenne (étendue)
Variables dérivées du MNE				
mne	Altitude (m)	3 - 6 - 9 - 12 - 15 - 20 - 30 - 60	c	495 (178; 838)*
mne90	Altitude au pas de 90 m du SRTM	Aucun	c	500 (244; 828)*
pente	Pente en %	12 - 15 - 20 - 30 - 60	c	11 (0; 170)*
curv_v	Courbure verticale du relief	3 - 6 - 9 - 12 - 15 - 20 - 30 - 60	c	0,02 (-0,943; 0,996)*
curv_h	Courbure horizontale du relief	3 - 6 - 9 - 12 - 15 - 20 - 30 - 60	c	0,02 (-0,864; 0,0765)*
fr	Focal range	20 - 30 - 60	c	185 (34; 413)**
fv	Focal variety	20 - 30 - 60	c	183 (34; 395)**
fs	Focal standard	20 - 30 - 60	c	39 (7; 110)**
dppr	Distance au plus proche réseau de drainage	Aucun	c	246 (0; 3037)
hppr	Différence d'altitude par rapport au plus proche réseau de drainage	Aucun	c	33 (0; 332)
flowacc	Logarithme népérien de l'accumulation du flux	Aucun	c	1,51 (0; 12,98)
flowdir	Direction du flux	Aucun	d	8 classes
expo	Exposition du versant	Aucun	d	10 classes
Variabales thématiques				
geol	Carte géologique au 1/50 000 (BRGM)	Aucun	d	22 classes
clc	Corine Land Cover (SOES)	Aucun	d	8 classes

analyse de la sensibilité du modèle (non montrés ici) ont conduit à choisir les paramètres suivants:
taille de l'arbre = 6, taux d'apprentissage = 20% du jeu de données, fraction du jeu de données prise aléatoirement pour chaque itération = 20%, nombre maximum d'itérations = 200.

La combinaison des approches ponctuelle et surfacique

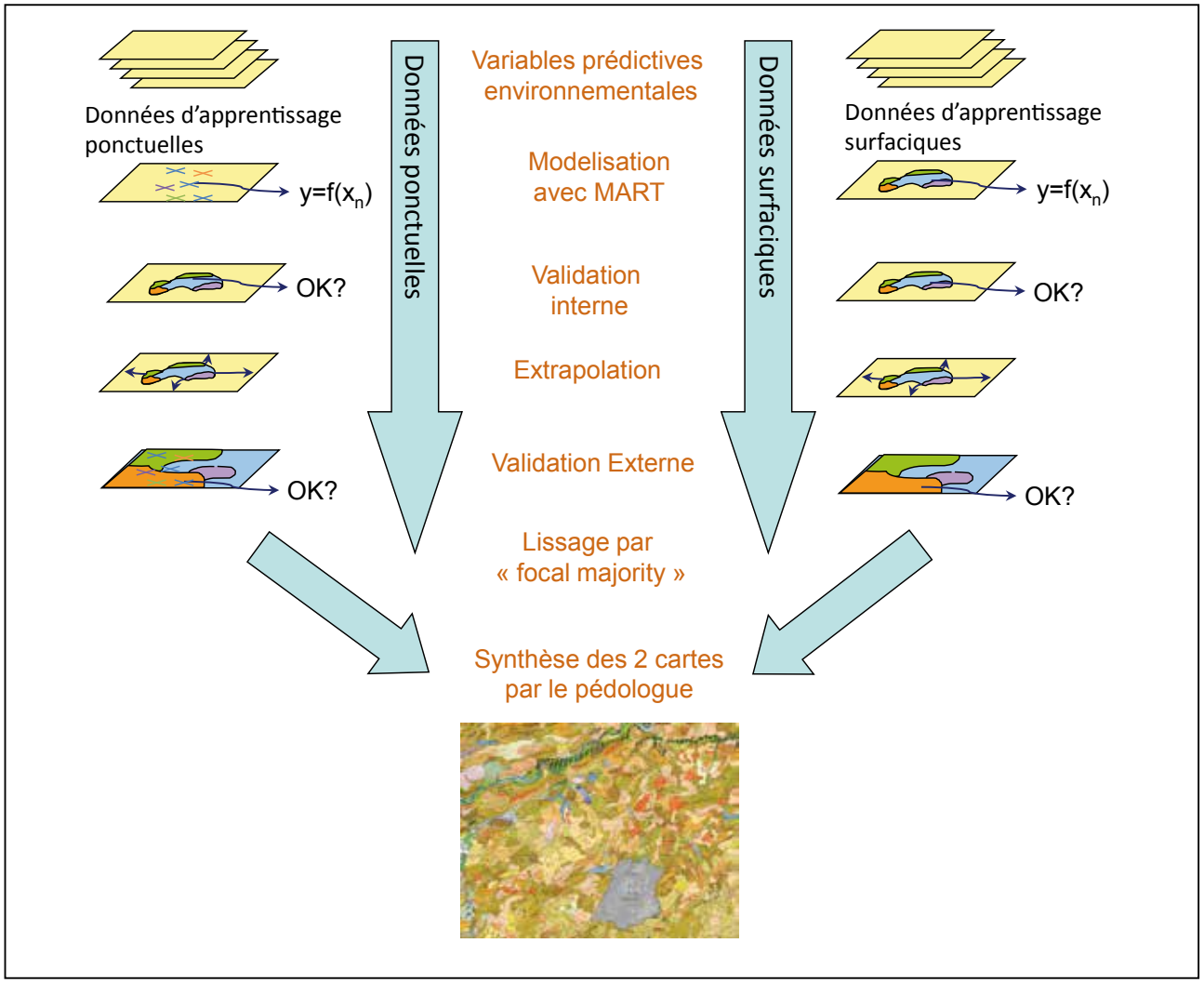
L'originalité de notre approche consiste à employer comme données d'apprentissage, d'une part, des données ponctuelles et, d'autre part, des données surfaciques. Les deux démarches sont conduites en parallèle *(figure 2).* La carte obtenue à partir de données surfaciques est la plus intéressante pour le pédologue car elle représente des unités complexes qui intègrent la diversité des UTS au sein des UCS. Le tracé des polygones, obtenu après lissage et vectorisation du résultat de la classification, doit pouvoir servir de référence pour le tracé de la carte finale. La carte des incertitudes apporte une aide importante quant à la qualité du tracé des polygones et à leur contenu. Elle met en effet en évidence des zones de forte incertitude sur lesquelles la prédiction issue de la démarche ponctuelle apporte un complément d'information précieux pour le pédologue. Il peut ainsi localement redéfinir un nouveau tracé guidé par la prédiction ponctuelle. C'est ainsi que le pédo-cartographe peut produire la synthèse des deux approches.

Le travail d'expertise a consisté à vérifier la validité de la prédiction (numéro d'UC et contours des plages cartographiques) obtenue par le croisement des deux approches.

Cette vérification a été réalisée selon plusieurs étapes:
- vérification, pour l'ensemble de la carte, de la concordance des contours et des types de sols attribués aux plages cartographiques par les deux prédictions ;

Figure 2 - Schéma organisationnel utilisé pour prédire les types de sols sur la carte de Vercel à partir des données surfaciques et ponctuelles à l'aide de l'algorithme MART.

Figure 2 - *Computational scheme used to predict soil types from surface and pontcual data with MART model on Vercel map.*



- pour les secteurs comportant des données (surfaciques et/ou ponctuelles), comparaison des contours et des types de sol des plages cartographiques prédites avec les contours et/ou les plages cartographiques dessinées sur les minutes de terrain;
- pour les « zones vierges », vérification de la cohérence des contours et des types de sols des plages cartographiques prédites, en les comparant d'une part avec leurs voisines, et, d'autre part, avec celles présentant les mêmes variables extrinsèques sur les trois autres cartes déjà réalisées par les pédologues (Besançon, Quingey et Ornans) et qui encadrent la carte de Vercel.

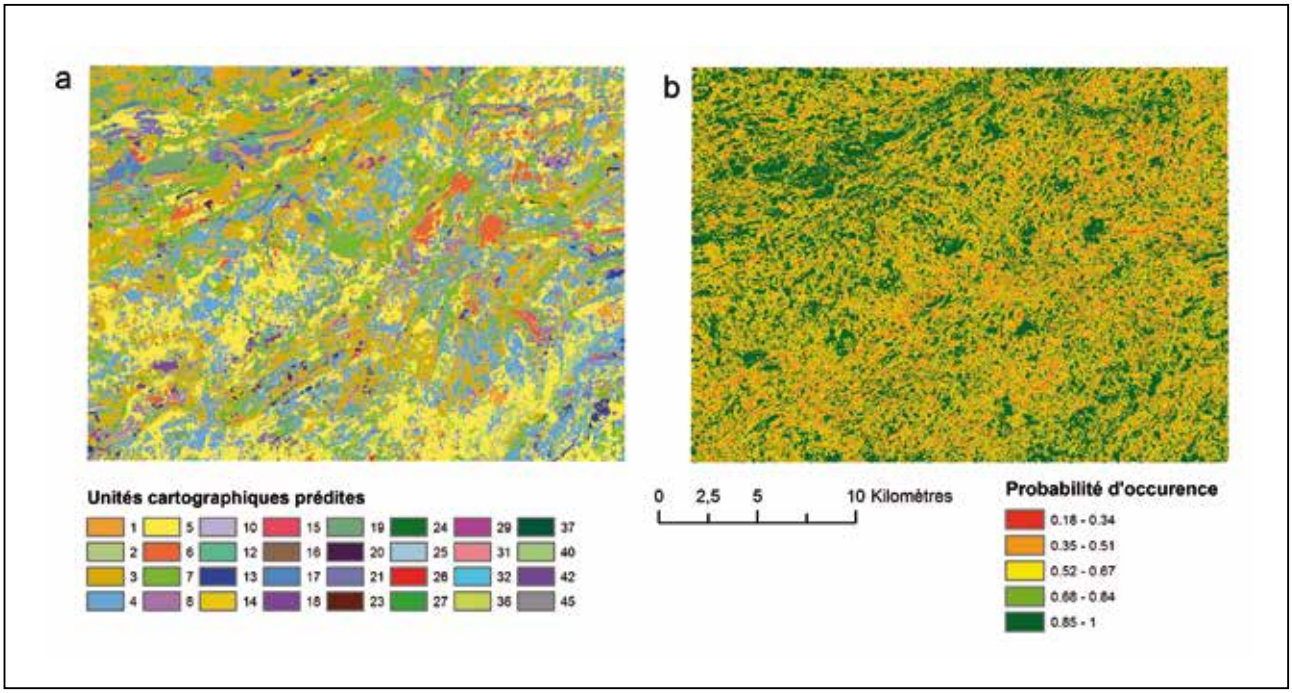
RÉSULTATS

Prédiction d'après les données ponctuelles

Evaluation du modèle

L'analyse de la contribution relative des variables pour l'ensemble des classes *(tableau 2)* a été produite à partir du jeu de données utilisé pour ajuster le modèle soit 50 % des pixels. Elle montre que deux variables se distinguent particulièrement: la géologie et l'exposition. Avec des valeurs de 100 % et 94 % de l'importance relative, elles se détachent nettement des autres variables telles que l'altitude lissée sur un rayon de 1800 m

Figure 3 - Prédictions des types de sol d'après les données ponctuelles: classes prédites (a) et cartes des probabilités d'occurrence (b).
Figure 3 - *Soil Types predictions obtained with the punctual data: predicted classes (a) and probability map (b).*



(64 %) et l'occupation du sol (59 %). La plupart des autres variables retenues par le modèle se situent entre 50 et 56 %. Il est intéressant de constater que parmi les 46 variables utilisées, la majorité des variables secondaires, dans la gamme des 50 %, concerne des indices de courbure avec des focales de tailles différentes.

La contribution relative des variables pour les 14 classes les plus représentées *(tableau 2),* montre que, pour la majorité des sols prédits, la géologie et l'exposition apportent la majorité de la contribution relative au modèle. Le taux d'erreur calculé sur les données de calibration (comparaison du type de sol observé et modélisé, sur les pixels ayant servi à ajuster le modèle) reste inférieur à 20 % pour la moitié des classes prédites.

Le taux le plus faible (5 %) est atteint pour la classe 15, correspondant aux Rédoxisols limoneux appauvris en surface qui sont assez faiblement représentés et dans des situations topographiques bien marquées. Le taux d'erreur le plus élevé (40 %) concerne la classe 13 (Brunisols Eutriques colluviaux). A part la géologie et l'exposition, les autres variables n'apportent pas suffisamment d'information pour distinguer cette classe des autres. Il convient aussi de noter que d'un point de vue fonctionnel, cette classe se rapproche fortement des unités 3, 5 et 7 (Brunisol ; Brunisol Eutrique ; Colluviosol brunifié).

Validation interne

La validation interne[2] de la prédiction effectuée à partir de 20 % des pixels pris au hasard dans le jeu des données ponctuelles *(tableau 3)* montre un indice global de prédiction de 81 % pour un indice de kappa de 0.78. Le modèle est donc globalement bien ajusté. Les erreurs d'omission[3] sont relativement faibles: de 1 à 31 % avec un taux relativement élevé pour les classes très représentées comme les classes 3, 4, 5 et 7 (Brunisol ; Quasi-Luvisol ; Brunisol Eutrique leptique et mésosaturé, Colluviosol brunifié) qui sont en effet des sols relativement semblables au regard des variables pouvant les distinguer. Les erreurs de commission[4] sont en général plus élevées que les erreurs d'omission. Ces dernières sont comprises entre 14 à 41 % et concernent, pour les valeurs les plus élevées, des classes relativement peu représentées: 13, 14 et 18 (Brunisol Eutrique colluvial; Quasi-Luvisol rédoxique ; Calcosol leptique).

La carte produite à l'issue de cette validation est représentée sur la *figure 3a*. Elle est tout à fait cohérente avec les données ponctuelles utilisées pour l'apprentissage

2 On parle de validation interne lorsque le même jeu de données sert à la fois à la calibration et à la validation de la prédiction.
3 Rapport entre le nombre total de pixels prédits pour une classe donnée et ceux correctement affectés à la classe prédite rapporté en pourcentage.
4 Rapport entre nombre total de pixels observés pour une classe données et ceux correctement affectés à la classe observée rapporté en pourcentage.

Tableau 2 - Analyse de la contribution relative des 11 variables les plus importantes pour les 14 classes les plus représentées.
Les valeurs de contribution relative des variables supérieures à 60 % figurent en gras.
Le taux d'erreur total correspond à l'erreur d'affectation calculée sur les pixels utilisés pour ajuster le modèle Unités de Sol décrites:
1 = CALCOSOL ; 2 = CALCISOL ; 3 = BRUNISOL ; 4 = QUASI-LUVISOL ; 5 = BRUNISOL EUTRIQUE, leptiques, mésosaturés ;
6 = BRUNISOL FERSIALLITIQUE ; 7 = COLLUVIOSOL BRUINFIE ; 8 = QUASI-LUVISOL et BRUNISOL ; 13 = BRUNISOL EUTRIQUE, colluvial; 14 = QUASI-LUVISOL REDOXIQUE ; 15 = REDOXISOLS limoneux appauvris en surface ; 17 = REDOXISOLS réductiques ;
18 = CALCOSOLS leptiques ; 19 = CALCISOLS leptiques.

Table 2 - *Relative contribution of the variables the 16 more represented classes. Bold type indicates variables accounting for more than 60 % in the classification. The total error rate corresponds to the affectation error calculated on the pixels used to ajust the model.*

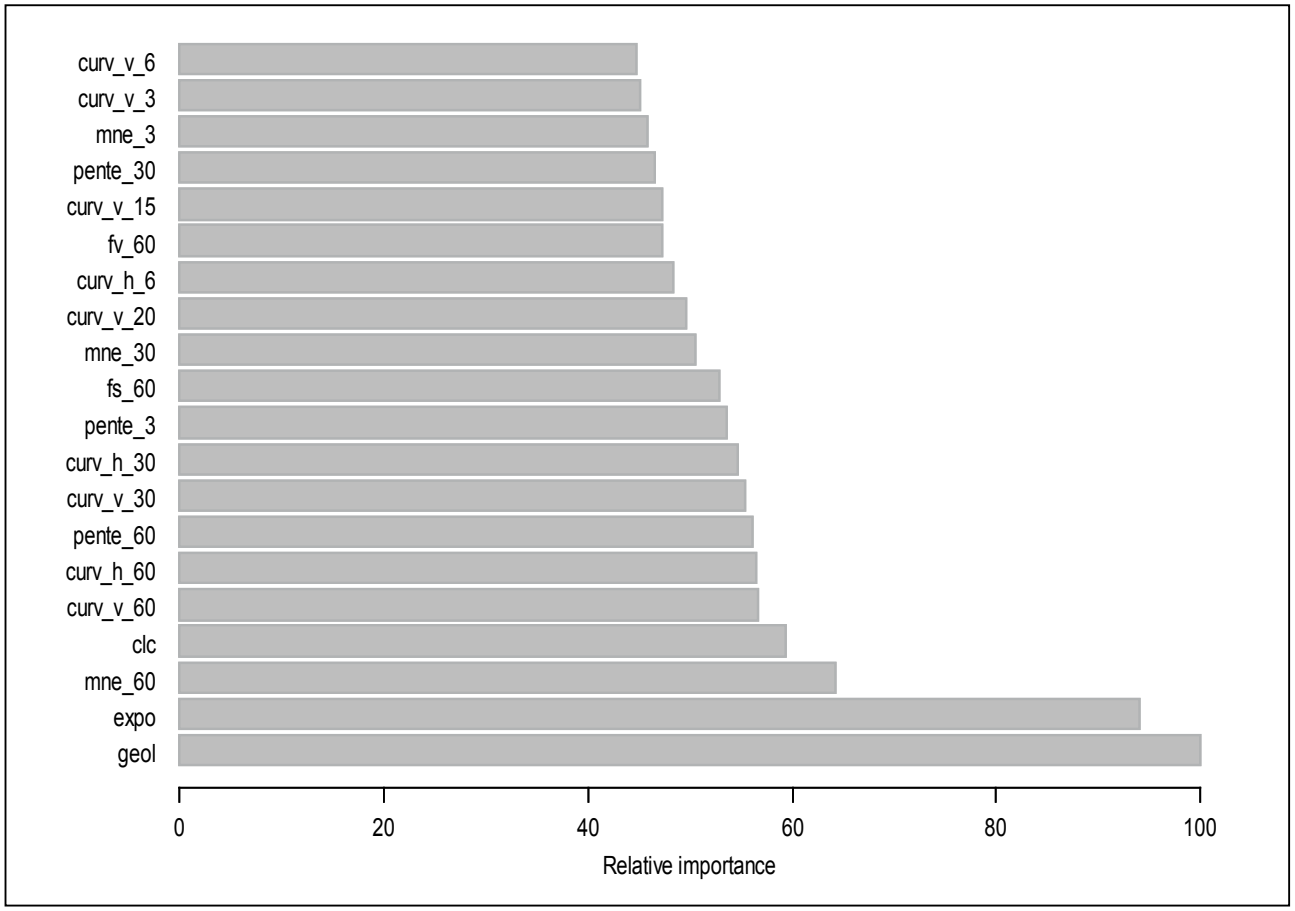
Unités de Sol	Contribution relative (%)											Taux d'erreur total (%)
	geol	expo	mne_60	clc	curvv_60	curvh_60	pente_60	curvh_30	curvv_30	pente_3	fs_60	
toutes	**100**	**94**	**64**	59	57	56	56	55	55	54	53	19
1	**100**	**98**	32	**74**	49	-	37	39	31	-	-	27
2	**96**	**100**	**77**	47	48	46	39	42	42	53	37	21
3	**99**	**100**	**88**	-	-	**75**	**68**	**72**	-	**63**	**61**	11
4	**100**	**95**	72	**63**	**66**	**82**	-	**77**	**70**	-	**69**	17
5	**100**	**86**	52	**73**	**64**	57	**79**	49	**67**	52	-	15
6	**100**	**98**	**62**	**77**	**68**	49	**65**	**64**	50	48	**67**	25
7	**100**	**92**	59	56	**74**	-	-	**70**	58	54	**60**	14
8	**80**	**100**	**63**	-	52	51	-	54	40	-	57	18
13	**92**	**100**	53	-	54	45	46	-	**77**	51	59	40
14	**100**	**72**	27	43	38	-	**60**	-	27	-	27	12
15	**100**	**74**	-	40	52	35	35	33	27	47	**61**	5
17	**98**	**100**	**92**	**67**	-	56	58	-	46	-	-	24
18	**92**	**100**	37	**64**	40	36	**68**	56	57	40	39	41
19	**100**	**81**	**69**	49	**71**	**61**	-	56	**67**	**70**	54	21

Tableau 3 - Matrice de confusion synthétique des unités de sols les plus représentées lors d'une validation interne avec les données ponctuelles.

Table 3 - *Accuracy of the most represented soil types prediction in training area with the point data.*

	Unités de sol													
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**13**	**14**	**15**	**17**	**18**	**19**
Nombre de pixels prédits	108	254	1719	1208	1552	287	1351	152	125	69	79	109	122	502
Erreur d'omission (%)	31	30	15	23	14	24	20	22	41	32	30	21	33	30
Erreur de commission (%)	1	9	31	28	29	12	29	5	10	3	4	6	2	15
Précision globale (%)	81													
Indice Kappa	0,78													

Figure 4 - Analyse de la contribution relative des variables pour toutes les classes prises dans leur ensemble pour les données ponctuelles.
Figure 4 - *Relative contribution of the main variables for all classes with the point data.*



(non présenté ici) et fait ressortir les unités présentes sur les faisceaux plissés (Colluviosols brunifiés) ainsi que celles majoritaires sur les plateaux (Quasi-Luvisol ; Brunisol Eutrique). On retrouve bien localement l'unité 6 correspondant aux Brunisols Fersiallitiques sur le quart Nord-Est. La carte des probabilités d'occurrence indique une assez bonne répartition des pixels bien prédits sur l'ensemble de la carte sans qu'il y ait véritablement de « terra incognita » selon le modèle. Les zones mal prédites *(figure 3b* (probabilité d'occurrence faible)) restent relativement clairsemées, de petite taille et entourées par des pixels bien classés. On ne peut pas y distinguer de structure spatiale proprement dite.

Prédiction d'après les données surfaciques

Importance des différentes variables

L'analyse de l'importance relative des variables utilisées pour la prédiction dite « surfacique » *(figure 6)* montre que le modèle, ajusté sur un jeu de données d'apprentissage issu de cartes pédologiques existantes, évince l'exposition qui était très importante dans la prédiction « ponctuelle » et relègue l'occupation du sol au 6e rang par ordre d'importance. Il donne en outre la primeur à l'altitude (mne_60 et mne_30) ainsi qu'à la pente (pente_60) et au « focal standard » (fs_60).

Validation interne

La matrice de confusion tirée de cette validation interne est présentée dans le *tableau 4*. Avec un indice global de 94 % et un Kappa de 0.93, on constate que le modèle est finement ajusté, voire sur-ajusté. Les erreurs d'omission et de commission se situent en moyenne en dessous de 10 %. Parmi les classes présentées ici, celle qui atteint le taux le plus élevé (32 %) correspond à l'unité la moins bien représentée, la classe 17 (Rédoxisols réductiques), avec seulement 124 pixels, alors que les autres classes en comptent généralement plusieurs milliers. La classe 10 (association de Brunisols superficiels et Colluviosols) pour laquelle l'erreur d'omission est la plus faible

Tableau 4 - Matrice de confusion synthétique de14 unités de sols (les plus fréquentes) lors d'une validation interne avec les données surfaciques.

Table 4 - *Accuracy of the most represented soil types prediction in training area with the surface data.*

	Unités de sol													
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**13**	**14**	**15**	**17**	**18**	**19**
Nombre de pixels prédits	357	1059	4555	4270	3687	4686	836	4273	7193	597	3291	697	124	1418
Erreur d'omission (%)	14	9	7	7	7	7	20	6	1	10	3	13	32	13
Erreur de commission (%)	1	4	7	9	9	9	3	10	2	1	3	1	15	8
Précision globale (%)	94													
Indice Kappa	0,93													

Figure 5 - Evolution de l'indice de classification global (G) et du Kappa (K) exprimés en % en fonction de la taille de la fenêtre de lissage des indices topographiques exprimés en nombre de pixels.

Figure 5 - *Global classification index (G) and Kappa index (K) evolution depending on the size of the focal mean window used on the topographic index in pixels.*

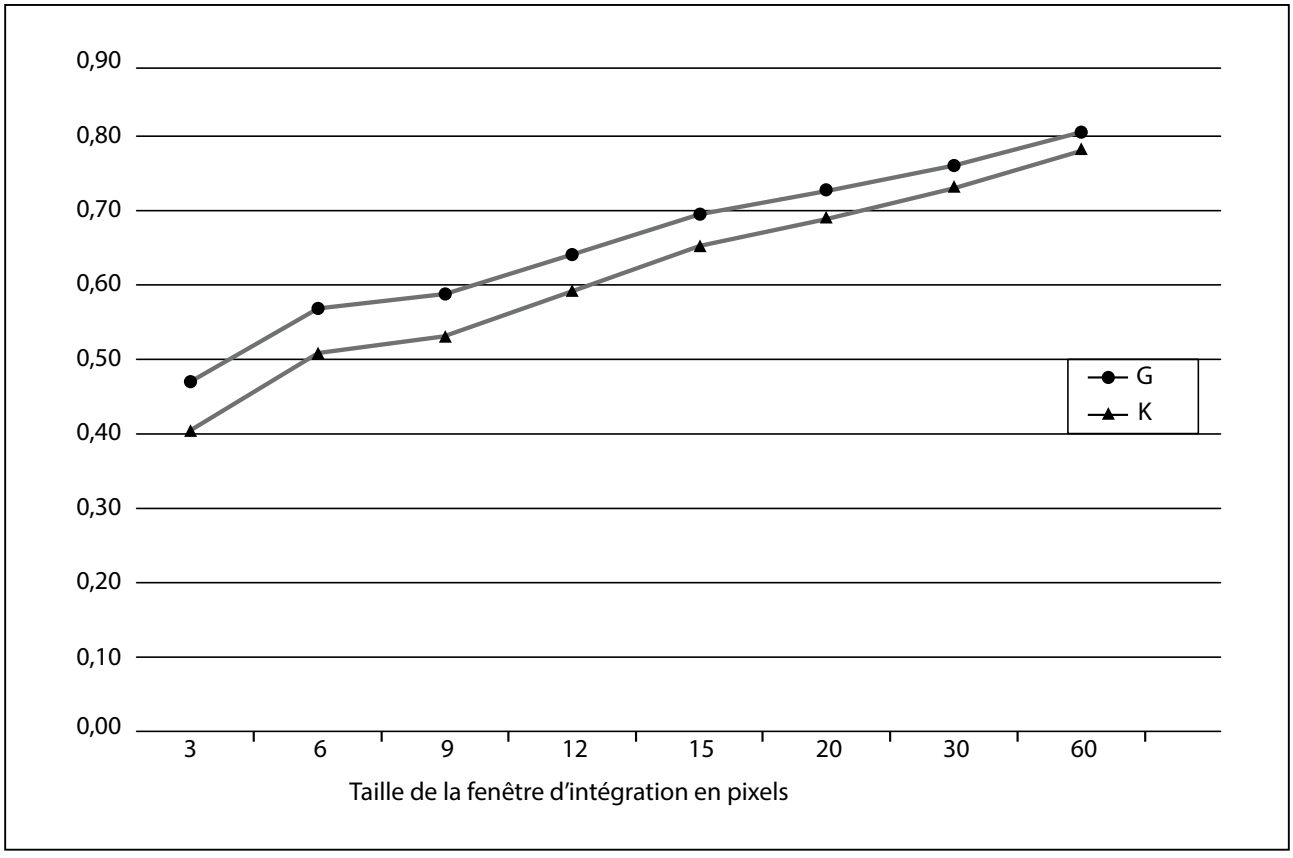
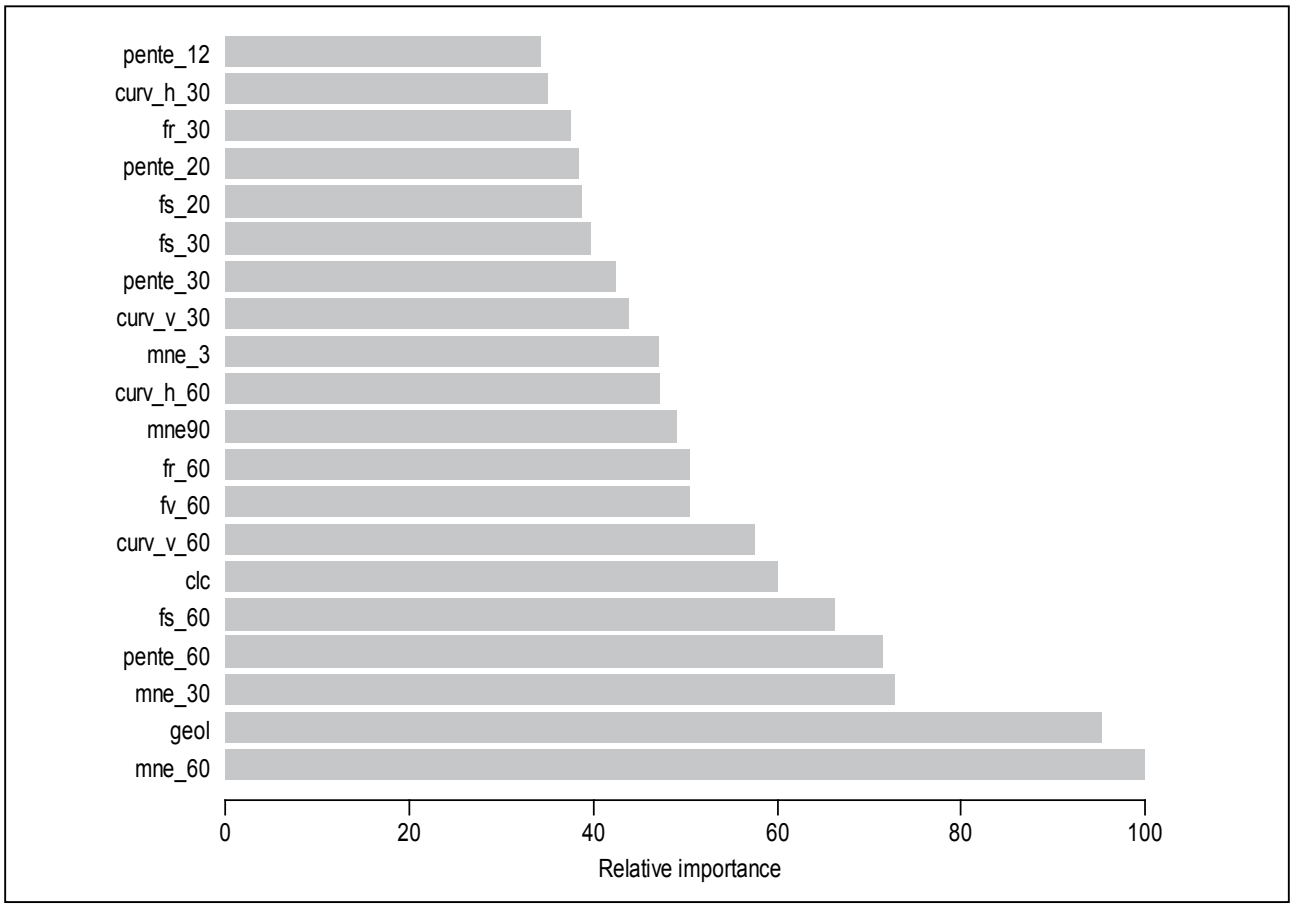


Figure 6 - Analyse de l'importance relative des variables pour toutes les classes prises dans leur ensemble pour les données surfaciques.

Figure 6 - *Relative contribution of the main variables for all classes with the surface data.*



(1 %) est également la classe la plus représentée avec 7193 pixels utilisés pour la prédiction.

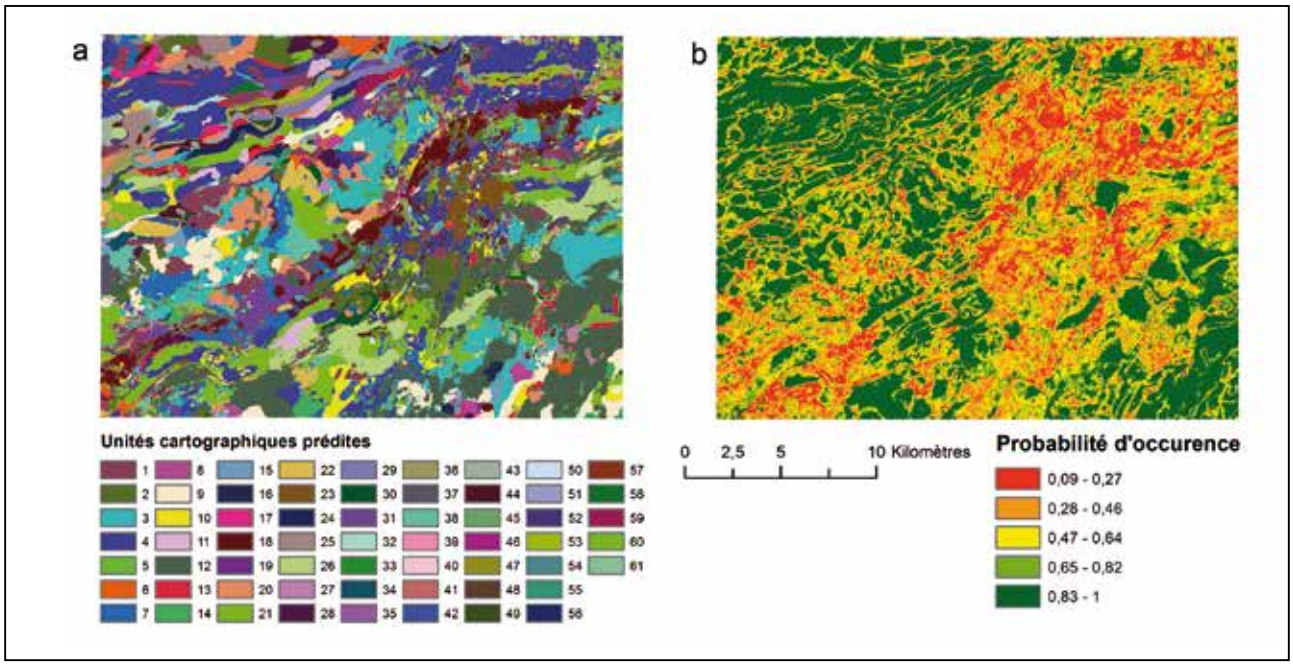
Les erreurs de commission sont encore plus faibles. Le maximum de 15 % est également atteint par la classe 17. Les classes les plus représentées conservent des taux d'erreur de commission inférieurs à 10 %.

La carte des probabilités issue de la prédiction surfacique *(figure 7b)* montre que la distribution des valeurs faibles suit une certaine organisation spatiale très différente de celle obtenue avec les données ponctuelles *(figure 3b).* On observe en effet sur la *figure 7b* une forte disparité des valeurs de probabilité entre les zones couvertes par la carte existante très bien prédites supérieures à 83 % et les zones à prédire inférieures à 23 %.

Validation externe

Nous avons simulé une validation externe au sein du jeu de données surfaciques afin de tester la sensibilité du modèle sur une zone exempte de données pédologiques. Pour ce faire, nous avons amputé le jeu de données de 20 % de sa surface le long d'une bande oblique traversant la carte du Nord-Ouest vers le Sud-Est *(cf. la zone figurée en vert sur la figure 1b).* Cette orientation perpendiculaire aux faisceaux plissés présente l'avantage de prendre en compte la grande majorité des sols présents dans le jeu de données initial de sorte que la zone d'apprentissage (représentée en vert sur la Figure 1b) et la zone de validation (représentée en rouge) comprennent exactement les mêmes unités de sols. Le modèle a cette fois été ajusté uniquement à partir des données représentées en vert et extrapolé sur les données représentées en rouge. La matrice de confusion *(tableau 5)* qui résulte de l'analyse de cette prédiction externe montre une précision globale (G) de 46 % et un Kappa de 0.31 qui pourrait être expliqué par un surajustement du modèle avec les données de calibration. Ces résultats restent cependant du même ordre de grandeur que ceux obtenus par Grinand *et al*., (2008) et Lemercier *et al*., (2012) dans des conditions comparables. En ignorant des classes extrêmement peu représentées, comme les classes 8, 13, 16 et 19, on

Figure 7 - Prédictions des types de sols d'après les données surfaciques: classes prédites (a) et cartes des probabilités d'occurrence (b).
Figure 7 - *Soil Types predictions obtained with the existing soil map: predicted classes (a) and probability map (b).*



constate que les erreurs d'omission et de commission sont, toutes choses égales par ailleurs, relativement acceptables avec des taux respectifs de 9 % et 34 %.

Lissage des données avec un filtre de convolution majoritaire

Le modèle MART nous permet d'obtenir *in fine* deux cartes ayant chacune des données d'entrées différentes, ponctuelles ou surfaciques. Les cartes ainsi obtenues et validées tel que décrit dans les paragraphes 3.1 et 3.2 ont un rendu relativement bruité, notamment sur les zones où les probabilités d'occurrence sont les plus faibles *(figures 3b et 7b).* Nous avons eu recours à un filtre de convolution retenant la valeur la plus fréquente au sein d'une fenêtre circulaire de 7 pixels de diamètre (soit 210 m) avec la fonction « focal majority » d'ArcInfo® afin de lisser l'image issue de la prédiction. La carte ainsi obtenue minimise le bruit et accentue les contours entre les unités de sols. Ce résultat est tout à fait similaire à celui obtenu avec l'algorithme CLAPAS (Robbez-Masson, 1994) en raisonnant sur des classes composées d'une seule unité de sol. Nous avons ainsi retenu la fonction « focal majority » pour des raisons de facilité d'utilisation - la fonction est directement accessible sous ArcInfo - et de temps de calcul.

Synthèse des approches ponctuelles et surfaciques établie par le pédologue

Choix des contours

La règle de base étant de s'écarter le moins possible de la cartographie déjà réalisée, nous avons privilégié les contours prédits à partir des données surfaciques, en particulier dans les secteurs dessinés non labellisés, et dans les secteurs vierges.

En revanche, pour les secteurs contenant des informations ponctuelles sans contours, le choix des contours proposés par l'une ou l'autre approche a été raisonné après synthèse des données ponctuelles.

Choix des numéros des labels (numéros des UCS)

Les labels attribués par les pédologues dans les secteurs cartographiés ont été le plus souvent conservés.

Pour les autres secteurs, le choix du label s'est avéré plus long à réaliser car il a nécessité une comparaison polygone par polygone des labels proposés par les deux prédictions. Pour trancher, il a fallu souvent remonter à la donnée de base (sondages ou profils) ; c'est donc l'approche ponctuelle qui a été privilégiée pour cette opération. Au final, le type de sol attribué à chaque polygone correspond soit au type de sol dominant dans la plage cartographique, soit à un complexe de sols.

La *figure 8* décrit synthétiquement cette démarche en trois étapes. On y voit dans la partie supérieure le résultat des prédictions surfaciques, à gauche et ponctuelles, à droite.

Figure 8 - Etapes de la modélisation à la production cartographique:
1 - Modélisation et prédiction avec MART
2 - Lissage des cartes produites pour un rendu cartographique
3 - Synthèse des prédictions par le pédologue.

Figure 8 - *Steps from modeling up to map achievement:*
1 - modeling and prediction with MART
2 - smoothing of the prediction map for a better readability
3 - synthesis of the smoothed maps made by the soil scientist.

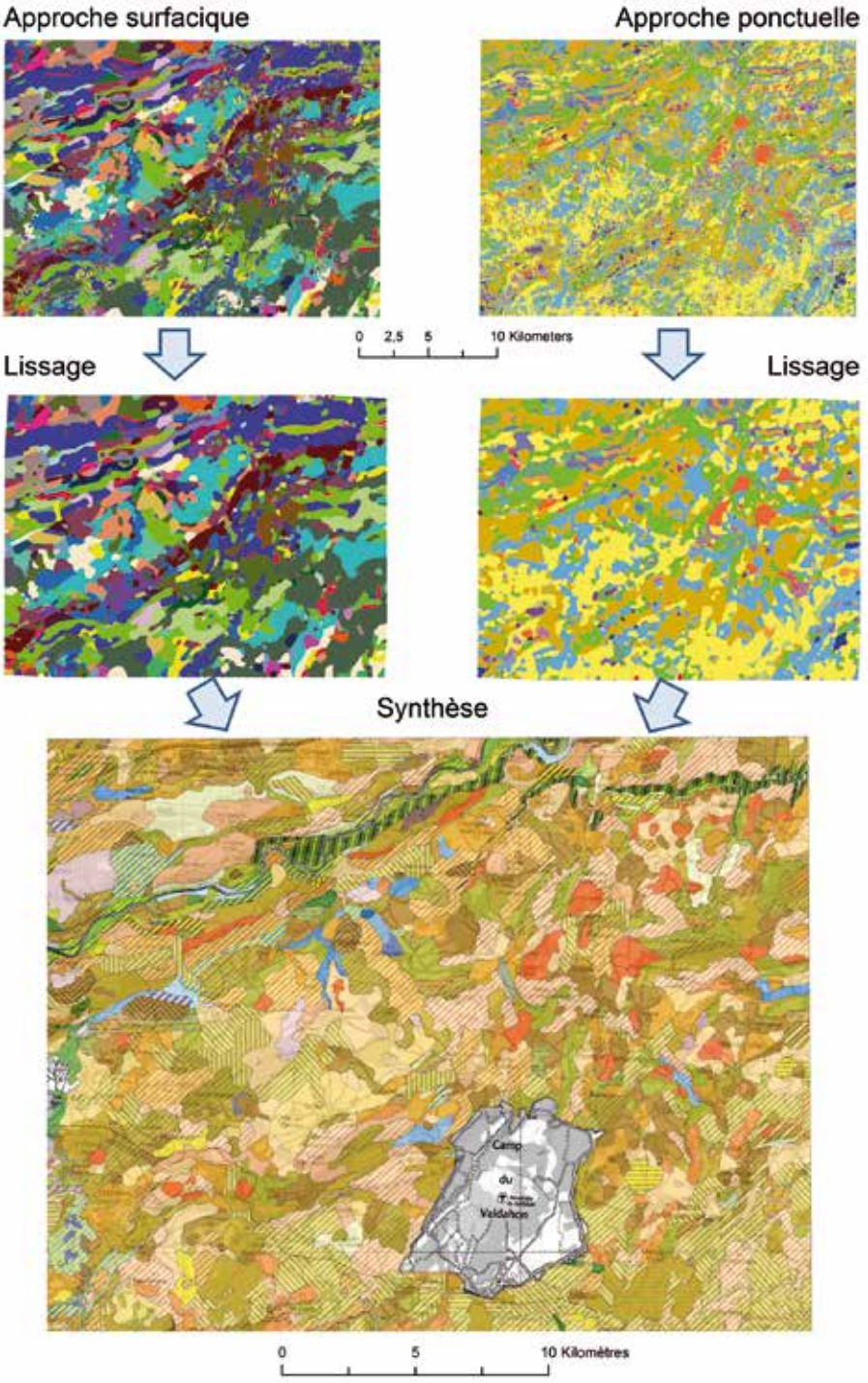


Tableau 5 - Matrice de confusion synthétique de 14 unités de sols (les plus fréquentes) lors d'une validation externe avec les données surfaciques.

Table 5 - *Accuracy of the most represented soil types prediction in validation area with the surface data.*

	Unités de sol													
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**13**	**14**	**15**	**17**	**18**	**19**
Nombre de pixels prédits	141	643	1107	1708	405	2860	36	1940	5428	1	841	16	162	163
Erreur d'omission (%)	20	53	36	23	13	42	4	48	91	0	5	0	84	12
Erreur de commission (%)	30	37	31	21	23	25	75	41	66	0	8	0	42	31
Précision globale (%)	46													
Indice Kappa	0,31													

La partie centrale expose les mêmes cartes après un lissage avec une fenêtre glissante de type « focal majority ». La carte présentée en bas est un extrait de celle publiée au 1/100 000 qui est véritablement la synthèse des deux précédentes.

DISCUSSION

Performance intrinsèque des modèles utilisés sur Vercel

Les coefficients globaux (G) de classification obtenus pour les validations internes des jeux de données ponctuelles et surfaciques, respectivement 81 % et 94 % sont tout à fait cohérents avec ceux obtenus avec des arbres de régressions boostés tels que décrit dans la littérature. Moran et Bui (2002) ont prédit des types de sols à partir de variables dérivées du MNT et issues de la télédétection avec un G de 70 %. Grinand *et al*. (2008) ont obtenu un G de 69 % sur le même type de données et 89 % en les lissant au moyen d'un filtre de convolution comparable à celui utilisé dans notre étude. Lemercier *et al* (2012) ont prédit des matériaux parentaux et classes de drainage du sol avec des taux de 73 % et 70 %. A la suite de Moran et Bui (2002), Grinand *et al*. (2008) et Behrens *et al*. (2010), on montre une fois de plus que l'emploi de filtres de convolution de plus en plus vastes spatialement augmente sensiblement la qualité de la prédiction interne *(figure 5).*

Il est important de noter que le modèle MART présente également l'avantage de gérer un nombre important de classes pour la prédiction. Sur la *figure 7a*, qui représente l'application d'un modèle ajusté sur l'ensemble des données existantes (validé sur la base d'une validation interne), nous avons pu spatialiser les 61 unités cartographiques de sols (à la fois pures et complexes). La matrice de confusion établie sur la base d'une validation interne *(tableau 4)* montre que le modèle est bien ajusté mais perd de sa précision lors d'une validation externe *(tableau 5).*

L'évolution de l'indice de classification global (G) et du Kappa (K) en fonction de la taille de la fenêtre de lissage des indices topographiques *(figure 5)*, toujours dans le cadre d'une validation interne, met en évidence l'intérêt d'utiliser des indices lissés sur une vaste surface. Pour établir ce graphique, de nombreux tests ont été réalisés avec le même jeu de données en ajoutant à chaque itération les indices topographiques calculés avec une fenêtre de taille supérieure. Ainsi, pour la première itération, tous les indices topographiques sont lissés par un « focal mean » d'un diamètre de 3 pixels ; lors de la seconde itération on remet en jeu les indices de la première itération plus les indices topographiques lissés par un « focal mean » d'un diamètre de 6 pixels et ainsi de suite jusqu'à 60. Le nombre d'indices utilisés croît ainsi rapidement et nous avons atteint les limites de calcul sous R au bout de la 8^{e} itération. Il est à noter que le modèle, après chaque itération, accorde une importance relative plus grande aux indices ayant la plus vaste fenêtre d'intégration spatiale. On retrouve ce résultat dans la *figure 4* où les premiers indices topographiques apparaissant par ordre d'importance sont: le mne_60, le curv_v60 le curv_v_h60 et le pente_60. Ces résultats sont par ailleurs tout à fait cohérents avec ceux obtenus par Behrens *et al*. (2010).

Avantages et limites de la synthèse des approches ponctuelles et surfaciques

Pour la prédiction cartographique, les contours de la prédiction obtenue avec les données d'entrées surfaciques ont été privilégiés pour deux raisons. D'une part, elles ont montré plus de robustesse et un meilleur taux de réussite notamment avec les matrices de confusion *(tableaux 3, 4 et 5)* et, d'autre part, en comparant le tracé des unités cartographiques des cartes existantes avec celui issu des prédictions, on constate qu'ils coïncident parfaitement sur les parties communes et que

la prédiction prolonge de façon cohérente le tracé existant sur les zones vierges. Les prédictions ponctuelles ont permis d'enrichir le contenu des unités cartographiques peu ou mal prédites au vu de la carte des probabilités d'occurrence *(figure 7b).*

Avantages

L'objectif de ce travail était de finaliser la carte pédologique 1/50 000 de Vercel afin de réaliser le 1/100 000 de Besançon dont elle constitue le quart nord-est. Sur 54 527 ha (surface totale de la carte de Vercel), 18605 ha avaient été levés, à raison de 1 sondage pour 17 ha en moyenne. Ceci représente un niveau de résolution intermédiaire entre les densités d'observation requises pour des cartographies au 1/25 000 et au 1/50 000, dans des régions de forte variabilité (Le Gros, 1996).

La couverture des 35990 ha restants, avec une densité d'observations comparable, aurait nécessité 2213 sondages supplémentaires. A raison de 15 à 20 sondages par jour et par personne, il faudrait compter environ 130 jours de terrain. A ce travail, il conviendrait d'ajouter le temps nécessaire à la préparation des campagnes de sondages puis à la synthèse des données et au report des contours sur les minutes de terrain soit une vingtaine de jours supplémentaires.

La modélisation cartographique que nous avons réalisée a nécessité le travail en collaboration d'un géomaticien et d'un expert pédologue, que l'on peut décomposer de la façon suivante:

- préparation des données, 10 jours,
- calibrage du modèle, 3 jours,
- validation interne et externe, 2 jours,
- concertation géomaticien / pédologue, 5 jours,
- synthèse des deux approches et validation par le pédologue, 5 jours,
- corrections et numérisation finale des contours et des labels, 5 jours,
- contrôles terrain des prédictions, 37 jours.

Soit un total de 77 jours.

On voit donc que l'utilisation de la cartographie numérique a permis de finaliser la carte de Vercel en réalisant, dans ce cas, une économie substantielle de 70 à 80 jours par rapport à une cartographie conventionnelle.

Limites

Bien que s'appuyant, de manière la plus objective possible, sur les données acquises par la prospection de terrain, l'utilisation de la cartographie numérique n'a pas permis d'aboutir à une cartographie de même précision que celle des secteurs ayant déjà fait l'objet d'une cartographie conventionnelle.

De ce fait, la carte 1/50 000 de Vercel présente, au final, une certaine hétérogénéité quant à la taille des polygones et une qualité de prédiction variable *(figures 3b et 7b)*, parfois faible, localement.

Le manque de données sur les formations superficielles (non prises en compte dans la couche de données géologiques), qui régissent en partie la distribution des BRUNISOLS, QUASI-LUVISOLS, BRUNISOLS à pellicules calcaires et les COLLUVIOSOLS brunifiés (classes 3, 4, 5 et 7), ainsi que la nécessité de synthétiser des prédictions ayant un nombre de classes différentes en sortie (32 classes pour les prédictions ponctuelles et 61 pour les prédictions surfaciques), ont abouti à une simplification inévitable de la cartographie à l'échelle du 1/50 000.

CONCLUSION

Le but de ces travaux était d'aider un pédologue à finaliser une carte des sols au 1/50 000 entamée par ses prédécesseurs sur 1/4 de sa surface. Afin de minimiser le temps de cartographie sur le terrain, nous avons choisi d'avoir recours à des techniques de cartographie numérique permettant d'extrapoler les informations existantes sur des zones comparables en terme de géologie, d'occupation du sol et de géomorphologie, toutes ces informations étant facilement disponibles. Nous avons pour cela eu recours à des arbres de régression boostés pour prédire les sols sur les 3/4 restants.

Les résultats de ces travaux montrent l'intérêt d'utiliser des variables topographiques avec différents niveaux de résolution spatiale ainsi que l'intérêt de combiner une approche surfacique et ponctuelle. Ceci est vrai notamment lorsque les données ponctuelles sont bien réparties sur l'ensemble de la carte alors que les données surfaciques n'existent que sur une portion restreinte. La combinaison des deux cartes nécessite l'expertise du pédologue qui, par sa connaissance du terrain et des lois locales de répartition des sols, peut ainsi déterminer le tracé définitif des unités cartographiques. L'approche conduite ainsi doit être considérée comme un outil d'aide à la cartographie et ne peut en aucun cas se substituer à la connaissance préalable du terrain. Comme pour toute cartographie des sols, des incertitudes demeurent mais nous sommes en mesure de les quantifier et de les localiser. La carte des sols de Vercel, dont la finalisation a été rendue possible dans le temps imparti grâce à la cartographie numérique, a permis de compléter la carte au 1/100 000 de Besançon qui a été publiée cette année.

REMERCIEMENTS

Les auteurs remercient les prospecteurs qui ont participé aux campagnes de terrain (Eugénie Tientcheu, Line Boulonne, Laurent Richard) menées dans le cadre de cette étude. La réalisation de la carte 1/50 000 de Vercel est inscrite dans le programme CPF (Connaissance Pédologique de France) qui constitue l'un des volets du programme IGCS (Inventaire, Gestion et Conservation des Sols) soutenu par le Ministère de l'Agriculture. Merci à Michèle Gaiffe pour la cession des

archives pédologiques de Sylvain Bruckert et à Marion Bardy pour sa relecture et ses conseils. Ce travail a été conduit dans le cadre de l'axe 2 du Réseau Mixte Technologique « Sols et Territoires ».

BIBLIOGRAPHIE

Baize. D., Girard. M.C. (coord.), 2009 - Référentiel pédologique 2008. Quæ Editions,

Versailles, France, 405p.

Behrens T., Zhu A-X, Schmidt K., Scholten T., 2010 - Multi-scale digital terrain analysis and feature selection for digital soil mapping. *Geoderma*, 155 (3-4), 175-185.

Ballabio C., 2009 - Spatial prediction of soil properties in temperate mountain regions using support vector regression. *Geoderma*, 151 (3-4), 338-350.

Bourennane H., King D., 2003 - Using multiple external drifts to estimate a soil variable. *Geoderma*, 114, 1-18.

Boruvka L., Penizek V., 2006 - A Test of an Artificial Neural Network Allocation Procedure using the Czech Soil Survey of Agricultural Land Data. *Developments in Soil Science*, 31, 415-424.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984 - Classification and regression trees, The Wadsworth statistics/probability series. Wadsworth International Group, Belmont, CA. 358p.

Bruckert S., 1987 - Clé dichotomique de classement des sols agricoles francs-comtois situé en climat tempéré semi-continental, *Annales scient. Univ. Besançon*, Biologie végétale, 4e série ; fasc. 7 ; 17-29.

Dobos E., Hengl T., 2009 - Soil Mapping Applications. *Developments in Soil Science*, 33, 461-479.

Dreyfuss M., 1965 - Carte géologique de la France à 1:50 000, XXXIV-23, Feuille de Vercel. Service de la carte géologique de France, BRGM, 8p.

Elith J., Leathwick J.R., Hastie T., 2008 - A working guide to boosted regression trees. *Journal of Animal Ecology*, 77 (4), 802-813.

Friedman J.H., 2002 - Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38 (4), 367-378.

Friedman J.H., Meulman J.J., 2003 - Multiple additive regression trees with application in epidemiology. *Statistics in Medicine*, 22 (9), 1365-1381.

M. Gaiffe, S. Bruckert, M. Eimberck (coll.), 2013. Carte pédologique de France à 1/100 000: Besançon. Editions Quae, Collection cartes et référentiels pédologiques, 250 p.

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.

Indyk P., Rajeev Motwani R., 1998 - Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality., *in* '98 Proceedings of the thirtieth annual ACM symposium on Theory of computing'. pp. 604-613.

Lawrence R., Bunn A., Powell S., Zambon M., 2004 - Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis. *Remote Sensing of Environment*, 90 (3), 331-336.

Lees B.G., Ritman K., 1991 - Decision-tree and rule induction approach to integration of remotely sensed and GIS data in mapping vegetation in disturbed hilly environments. *Environmental Management,* 15 (6), 823-831.

Legros JP. 1996 - Cartographies des sols. De l'analyse spatiale à la gestion des territoires. Presses polytechniques et universitaires romandes. Lausanne, Suisse. 411 p.

Lehmann S, Bégon JC, Eimberck M, Daroussin J, Wynns R, Arrouays D. 2007 - Utilisation du logiciel CLAPAS pour l'aide à la délimitation de pédopaysages. Un test sur la carte des sols de Mirande (Gers, France). *Etude et Gestion des Sols*, 14(2) : 135-151.

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, 75-84.

McKenzie N.J., Ryan P.J., 1999 - Spatial prediction of soil properties using environmental correlation. *Geoderma* 89 (1-2), 67-94.

Mendonça-Santos M.L., McBratney A.B., Minasny B., 2006 - Soil Prediction with Spatially Decomposed Environmental Factors. *Developments in Soil Science*, 31, 269-278.

Michaelsen J., Schimel D., Friedl M., Davis F., Dubayah R., 1994 - Regression tree analysis of satellite and terrain data of guide vegetation sampling and surveys. *Journal of Vegetation Science,* 5 (5), 673-686.

Minasny B., McBratney A.B., 2007 - Incorporating taxonomic distance into spatial prediction and digital mapping of soil classes. *Geoderma,* 142 (3-4), 285-293.

Moisen G.G., Freeman E.A., Blackard J.A., Frescino T.S., Zimmermann, N.E., Edwards Jr., 2006 - Predicting tree species presence and basal area in Utah: a comparison of stochastic gradient boosting, generalized additive models, and tree-based methods. *Ecological Modelling*, 199 (2), 176-187.

Moran C.J., Bui E.N., 2002 - Spatial data mining for enhanced soil map modeling. *International Journal of Geographical Information Science,* 16 (6), 533-549.

Odeh, I.O.A., McBratney, A.B., Chittleborough, D.J., 1992 - Soil patterns recognition with fuzzy-c-mean: application to classification and soil-landform interrelationships. *Soil Science Society of America Journal*, 56 (2), 505-516.

Pal M., Mather P.M., 2003 - An assessment of the effectiveness of decision tree methods for land cover classification. *Remote Sensing of Environment*, 86 (4), 554-565.

Robbez-Masson J.M., 1994 - Reconnaissance et délimitation de motifs d'organisation spatiale. Application à la cartographie des pédopaysages. Th. Doct. ENSAM, 161 p. + annexes.

Scull P., Franklin J., Chadwick O.A., 2005 - The application of classification tree analysis to soil type prediction in a desert landscape. *Ecological Modelling,* 181 (1), 1-15.

Scull P., Franklin J., Chadwick O.A. and McArthur D., 2003 - Predictive soil mapping: a review. *Progress in Physical Geography,* 27 (2), 171-197.

Walter C., Lagacherie P., Follain S., 2006 - Integrating Pedological Knowledge into Digital Soil Mapping. *Developments in Soil Science*, 31, 281-300,