

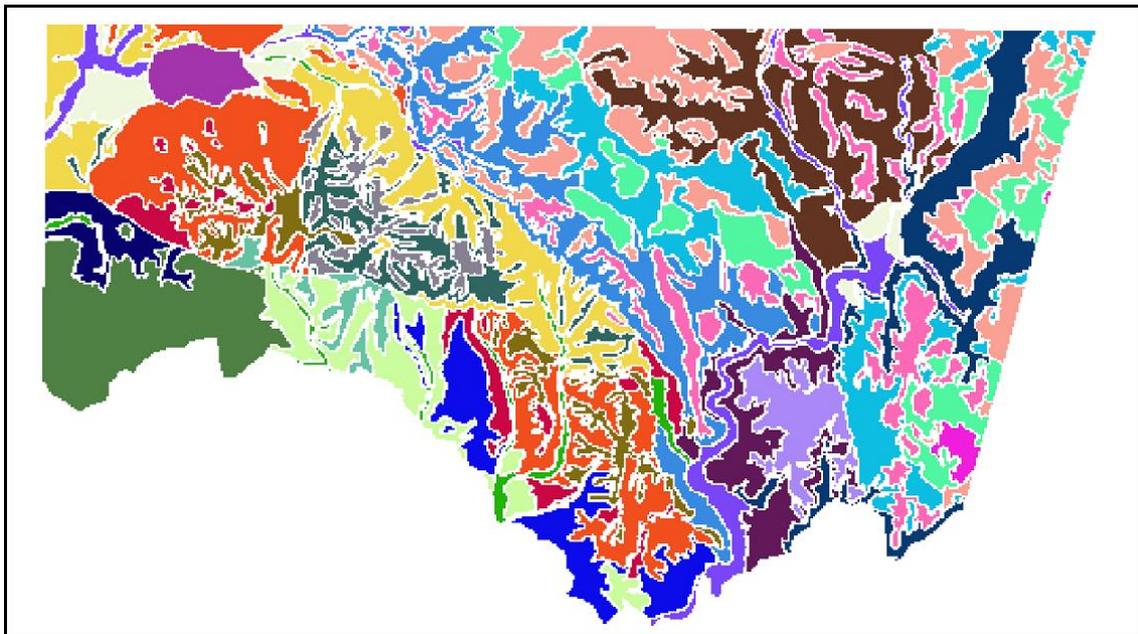


Rapport de stage

# Amélioration de CLAPAS sous R

Outil pour la cartogénèse numérique

Adeline CARNIS



## *Remerciements*

Je tiens tout d'abord à remercier Sébastien LEHMANN, pour m'avoir proposé ce stage, pour sa confiance et pour son soutien tout au long de ce travail. Ses conseils et ses réponses à toutes mes questions m'ont beaucoup aidée.

Je remercie Dominique ARROUAYS, directeur de l'unité Infosol d'Orléans, pour avoir accepté que j'effectue mon stage dans son unité et pour ses idées inépuisables et ses suggestions lors des réunions de travail.

Je voudrais également remercier les personnes qui m'ont aidée et conseillée lors de mon stage : Nicolas SABY et Manuel MARTIN pour leurs indications et leurs astuces concernant la programmation, Catherine PASQUIER pour sa compétence en Système d'Information Géographique et Bertand ORTOLLAND pour ses encouragements. Je voudrais rendre hommage à Jean-Marc Robbez-Masson, initiateur du logiciel CLAPAS. Travailler sur un logiciel comme CLAPAS a été très enrichissant.

Je tiens à remercier toute l'équipe des unité de la Carto pour m'avoir accueillie pendant ces trois mois. Avoir pu intégrer une équipe de travail aussi dynamique et conviviale m'a énormément encouragée.

Enfin, merci à mes deux collègues suédois et belge, Johanna WETTERLIND et Jeroen MEERSMANS pour m'avoir accueillie dans leur bureau européen. Merci à eux, de m'avoir permis de connaître leur culture, d'améliorer mon anglais et d'avoir partagé leurs connaissances.

# Table des matières

<b>Introduction</b>	<b>5</b>
<b>1 Présentation de l'INRA</b>	<b>6</b>
1.1 Le centre régional d'Orléans . . . . .	7
1.1.1 Les unités du centre . . . . .	9
1.2 L'Unité Infosol . . . . .	10
1.2.1 Les Missions de l'Unité . . . . .	10
1.2.2 Les Activités . . . . .	11
<b>2 Présentation de CLAPAS</b>	<b>13</b>
2.1 La cartographie numérique . . . . .	13
2.1.1 Le sol . . . . .	13
2.1.2 La cartographie conventionnelle et numérique . . . . .	14
2.1.3 CLAPAS . . . . .	16
2.1.4 Principe . . . . .	16
2.1.5 Les outils . . . . .	19
<b>3 Le stage et les résultats</b>	<b>20</b>
3.1 Objectifs du stage . . . . .	20
3.2 Les zones d'étude . . . . .	20
3.3 Travail effectué . . . . .	21
3.3.1 Amélioration de l'image d'entrée du modèle CLAPAS . . . . .	21
3.4 Adaptation de la fenêtre de forme elliptique . . . . .	29
3.4.1 Création de la fenêtre . . . . .	29
3.4.2 Création de la fonction pente . . . . .	30
3.4.3 Adaptation de la taille de la fenêtre . . . . .	31

3.4.4 Résultats des classements . . . . .	32
<b>Conclusion</b>	<b>43</b>
<b>Bibliographie</b>	<b>44</b>

## Introduction

Le recours de la cartographie numérique ou Digital Soil Mapping (DSM, alternative à la cartographie pédologique traditionnelle, basée sur les systèmes d'information géographique, les modèles statistiques et la connaissance de la formation des sols ne cesse d'augmenter. LE DSM s'oriente vers la production de cartes des sols grâce à un ensemble de recherches qui ont été menées ces vingt dernières années. Ces recherches traitent de divers sujets dont le développement des modèles numériques de prédiction des sols.

Le logiciel CLAPAS, conçu par un chercheur de l'INRA de Montpellier en 1994, s'inscrit directement dans ces recherches. Il permet la prédiction de cartes des sols dans le but de faciliter et de simplifier le travail du pédologue sur le terrain. Il fut déjà l'objet de plusieurs stages dont celui d'Olivier VITRY qui a programmé l'algorithme CLAPAS dans le langage statistique R, l'année dernière au sein de l'Unité INFOSOL du centre INRA Orléans.

Lors de mon stage, il m'a été demandé de programmer et d'automatiser la création de la carte d'entrée du modèle, paramètre indispensable lors de l'exécution de CLAPAS et également d'améliorer la fonctionnalité de la fenêtre elliptique. Ainsi, CLAPAS deviendrait un outil complet et autonome.

Dans un premier temps, je vous présenterai l'organisme d'accueil, l'Institut National de Recherche Agronomique (INRA), puis j'aborderai la cartographie numérique au sens large, qui permettra de définir l'objet de mon stage CLAPAS. Enfin, la dernière partie de ce rapport sera consacrée aux travaux réalisés pendant ces trois mois, aux problèmes rencontrés et aux résultats obtenus.

# Partie 1

## Présentation de l'INRA

L'Institut National de Recherche Agronomique (INRA) est un organisme de recherche scientifique publique, placé sous la tutelle du ministère de l'Enseignement supérieur et de la Recherche et du ministère de l'Alimentation, de l'Agriculture et de la Pêche. Ses recherches sont axées sur les domaines de l'agriculture, de l'alimentation et de l'environnement.

L'Inra a été fondé en 1946. Il est aujourd'hui, le premier institut européen de recherche agronomique et le deuxième mondial. Il mène des recherches finalisées pour une alimentation saine et de qualité, pour une agriculture compétitive et durable, et pour un environnement préservé et valorisé.

Les missions de l'INRA, à l'échelle nationale, sont nombreuses et variées :

- Produire et diffuser des connaissances scientifiques.
- Concevoir des innovations et des savoirs-faire pour la société.
- Éclairer, par son expertise, les décisions des acteurs publics et privés.
- Développer la culture scientifique et technique et participer au débat science/société.
- Former à la recherche et par la recherche.

Pour pouvoir mener à bien ses missions, l'Institut emploie 1 828 chercheurs, 2 427 ingénieurs, 4 249 techniciens et administratifs. Chaque année, il accueille également 1 784 doctorants, 1 000 chercheurs étrangers et 1 901 stagiaires. Il accroît également ses collaborations avec le monde agricole, la communauté scientifique et aussi avec

les collectivités territoriales et les entreprises.

L'INRA a perçu 745,68 millions d'euros de ressources en 2008. Ces ressources proviennent pour 80% de subventions pour charges de service public, 12% de subventions et soutiens finalisés à l'activité de la recherche, 5% de produits valorisés de l'activité de recherche et prestations de service, 2% d'autres subventions et produits. Les dépenses du personnel représentent 71% des dépenses, celles de fonctionnement 17%, les frais d'infrastructure et informatique collective 8% et l'immobilier 4%.

Étant le premier institut de recherche européen, l'INRA collabore, au sein de l'Europe, à la construction de l'Espace européen de la recherche et aussi internationalement à d'importants programmes.

Avec 14 départements scientifiques, 20 centres régionaux, près de 150 sites de recherches et d'expérimentation dans toute la France, l'INRA possède un système performant pour pouvoir répondre aux questions liées à l'agriculture, à l'alimentation et à l'environnement.

## 1.1 Le centre régional d'Orléans

Le centre d'Orléans est un des 20 centres régionaux de l'INRA. Il comprend des unités de recherche implantées à Ardon et possède également un domaine d'expérimentation animale, situé aux alentours de Bourges.

Il est placé sous l'autorité de Dominique KING, assisté du Directeur des Services d'Appui, Pierre-Jean BALARD. Les services d'appui à la recherche (SDAR) assurent sous la responsabilité de leurs directeurs la gestion des services collectifs du centre.

Sur ce site, on peut compter 200 agents titulaires dont 28 chercheurs, 53 ingénieurs et 199 techniciens et administratifs. Chaque année, il accueille environ 150 agents non titulaires. Quatre domaines sont traités au centre d'Orléans : la forêt, les insectes ravageurs forestiers, les sols et les animaux d'élevages.

L'INRA d'Orléans est situé à l'adresse suivante :

*INRA*  
*Centre de Recherches d'Orléans*  
*2163, avenue de la pomme de pin*  
*CP 40001-Ardon*  
*45075 ORLEANS CEDEX 2*



Vue aérienne janvier 2007 - Photo : © INRA J.-C. Bastien & O. Bertel

Le centre est installé sur 63 hectares dont 13 098 m<sup>2</sup> bâtis (y compris 1505.35 m<sup>2</sup> de serres) et 47 hectares de pépinière.

### 1.1.1 Les unités du centre

Les grands axes de recherche des unités du Centre INRA d'Orléans s'articulent autour de *la forêt, le sol et les animaux d'élevage*.

Six unités s'organisent autour de ces trois pôles :

#### Premier pôle : LA FORET

- L'Unité de Recherche Amélioration, Génétique et Physiologie Forestières rassemble des compétences en génétique, génomique et physiologie appliquées aux arbres forestiers.
- L'Unité de Recherche de Zoologie Forestière œuvre pour comprendre les mécanismes qui régissent les populations d'insectes en expansion sous l'effet des activités humaines et des changements environnementaux.
- L'Unité Expérimentale d'Amélioration des Arbres Forestiers est une unité d'appui à la recherche qui travaille essentiellement avec la première unité citée.

#### Second pôle : LE SOL

- L'Unité de Recherche Sciences du Sol étudie l'organisation, l'évolution et le fonctionnement des sols selon différentes échelles spatiales (horizon, solum, parcelle, bassin versant, paysage) et temporelles (événement pluvieux, saison culturale, rotation, décennie, siècle et au delà).
- L'Unité Infosol gère le Conservatoire National d'Echantillons de Sols et l'ensemble des bases de données nationales concernant les sols de France.

#### Troisième et dernier pôle : LES ANIMAUX

- L'Unité Expérimentale d'Élevage située à Bourges. Elle est un outil important dans le dispositif du Département de Génétique Animale pour répondre aux besoins des filières animales et à l'attente de la recherche dans le but de :

- Améliorer les performances des troupeaux et la qualité de leurs produits.
- Procréer des animaux de phénotypes différenciés pour l'évaluation des effets de gènes par les laboratoires de Génomique.

## 1.2 L'Unité Infosol

Le sol est l'objet d'étude central de l'Unité Infosol. Elle fut créée en 2000 par différents partenaires (le ministère en charge de l'Agriculture, le ministère chargé de l'Environnement, l'IFEN, l'ADEME et l'INRA) regroupés sous le nom de Groupe-ment d'Intérêt Scientifique (GIS Sol).

### 1.2.1 Les Missions de l'Unité

Sa mission est d'établir un système d'information sur les Sols de France et sur l'évolution de leurs qualités. Ce système d'Information doit pouvoir répondre à des demandes régionales et nationales, tout en s'inscrivant dans un dispositif européen.

Les missions du GIS Sol et d'Infosol se déclinent comme suit :

- développer des réseaux et des outils de surveillance de l'évolution de la qualité des sols de France ;
- coordonner et harmoniser l'inventaire cartographique des sols de France ;
- collecter et gérer des données concernant les pressions des sols ;
- administrer les programmes et les données ;
- diffuser l'information sur les sols ;
- répondre aux demandes institutionnelles nationales et internationales concernant les sols ;
- mettre en place et gérer un conservatoire d'échantillons de sols ;
- former, sensibiliser et communiquer en ce qui concerne l'usage des données sur les sols.

## 1.2.2 Les Activités

L'objectif de cette unité est de mettre en place les outils et les bases de données permettant une meilleure connaissance spatialisée des sols et de leurs évolutions.

Les activités de l'Unité sont organisées autour de ces trois équipes : l'équipe **Inventaire**, l'équipe **Surveillance** et l'équipe **Gestion et traitement des données**.

Nous allons maintenant exposer les activités de chaque équipe.

**L'équipe Inventaire** a pour objectif de réaliser un inventaire cartographique des sols de France. Cette équipe est sous la responsabilité de Nathalie SCHNEBELLEN. Cette équipe dirige deux programmes :

- Le programme Inventaire, Gestion et Conservation des Sols (IGCS) propose aux régions un appui pour réaliser un inventaire de leurs sols à différentes échelles.
- Le programme Base de Données Géographique des Sols de France à l'échelle du 1/1 000 000 (BDGSF) est une représentation simplifiée de la diversité spatiale de la couverture de sols de France.

**L'équipe Surveillance**, dirigée par Claudy JOLIVET, a pour objectif de mesurer et de surveiller l'évolution de la qualité des sols à l'échelle nationale. Ce groupe dirige deux programmes :

- Le programme Réseau de Mesures de la Qualité des Sols (RMQS) constitue un outil permettant d'évaluer et de suivre la qualité des sols. La mise en place de ce programme répond au besoin de disposer de données quantitatives sur l'états des sols de France. L'objectif est de détecter de façon précoce l'apparition et les tendances de dégradation du sol en mesurant un certain nombre de paramètres physiques et chimiques. Tous les échantillons de sols prélevés sont stockés dans le Conservatoire National d'Echantillons des Sols situé à l'INRA d'Orléans.
- Le programme Base de Données Analyse des Terres (BDAT) qui réalise environ 250 000 analyses de terres en France chaque année. Elles sont

généralement demandées par les agriculteurs pour gérer au mieux la fertilisation. Ces échantillons prélevés représentent une source d'information intéressante. L'objectif de ce programme est de regrouper ces résultats dans une base de données.

**L'équipe Gestion et Traitement** , dirigée par Dominique ARROUAYS, doit gérer et centraliser les données nationales des études sur le sol. Cette équipe gère trois programmes :

- Le programme Gestion des Bases de Données consiste à concevoir, implémenter, gérer, enrichir, vérifier et mettre à disposition des bases de données sur les sols. Il apporte un appui méthodologique auprès des autres programmes de l'Unité.
- Le programme Etudes et Travaux traite des données des autres équipes. Il utilise des méthodes de traitement de données spatiales et/ou temporelles (géostatistique, changement d'échelle,...), des règles d'expertise et différents modèles mathématiques (de culture, d'érosion).
- Le programme Valorisation et Communication a pour objectif de valoriser les données sur les sols afin de fournir les informations finales sur les sols de France et de les mettre à disposition.

**Mon activité au sein de l'Unité** : le stage s'inscrit dans le programme Etudes et Travaux de l'équipe Gestion et Traitement. Il m'a été confiée de contribuer à l'amélioration de l'algorithme, CLAPAS, un des différents programmes de l'équipe.

## Partie 2

# Présentation de CLAPAS

CLAPAS est utilisé pour la prédiction d'unités cartographiques des sols appelées pédopaysages. Un pédopaysage est un ensemble d'horizons pédologiques et d'éléments paysagiques : végétation, effets des activités humaines, géomorphologie, hydrologie, substrat ou roche-mère, dont l'organisation spatiale permet de définir dans son ensemble une (ou une partie d'une) couverture pédologique<sup>1</sup>. Il est donc indispensable de préciser certains éléments avant d'aborder la présentation de CLAPAS, comme la cartographie numérique et surtout l'élément principal de notre étude le sol.

## 2.1 La cartographie numérique

### 2.1.1 Le sol

Avant d'aborder la cartographie numérique et la cartographie conventionnelle, il faut définir un élément essentiel à notre travail : *le sol* ou *couverture pédologique*. En pédologie, il est déterminé comme un objet naturel dont l'existence et l'état actuel résultent de l'évolution au cours du temps<sup>2</sup>.

Les couvertures pédologiques sont formées de constituants minéraux et organiques, présents à l'état solide, liquide ou gazeux. Ces constituants sont organisés entre eux,

---

1. GIRARD 2000

2. BAIZE D., JABIOL B., 1995

formant ainsi des structures spécifiques du milieu pédologique<sup>3</sup>.

Les couvertures pédologiques montrent en général des différenciations selon un axe vertical. Ces différenciations sont appelées *horizons*. Un horizon est une couche du sol relativement homogène, perceptible à l'œil nu sur le terrain. Il est le niveau le plus pratique pour observer et pour échantillonner une couverture pédologique. L'ensemble des horizons constitue le solum. Le solum est une tranche verticale d'une couverture pédologique observée dans une fosse. Il peut-être défini comme un volume réel provenant d'une fosse, d'une largeur d'environ 60 à 100 cm, de la hauteur de la fosse et d'une épaisseur moyenne de 10 à 20 cm<sup>4</sup>.

Un autre terme est également utilisé dans le **Référentiel Pédologique** : le **profil**. Il correspond à l'identification des horizons (horizons O, H, A, E, etc.).

Après cette brève définition, on s'aperçoit que le sol est un milieu complexe. La cartographie ne peut pas être une opération facile. Elle est basée sur une connaissance théorique de l'objet à cartographier, de ses cycles, de son mode de formation, etc, et sur de lourdes campagnes de terrain et des analyses complémentaires en laboratoire. Il s'agit d'une opération longue et coûteuse, or les collectivités ne financent le plus souvent que des études ponctuelles là où le besoin spécifique se présente. La cartographie de la couverture pédologique est une démarche complexe et délicate qui s'appuie sur des observations et des connaissances du sol décrites de manière précise grâce à de diverses techniques comme les sondages, transects, profils, analyses, etc.

### 2.1.2 La cartographie conventionnelle et numérique

La cartographie pédologique traditionnelle est une approche d'expert qui consiste dans un premier temps à synthétiser les données disponibles (topographie, géologie, végétation, etc.) afin d'aboutir à des hypothèses concernant les lois de distribution de sols<sup>5</sup>. Ensuite, le travail sur le terrain et les observations obtenues permettent de vérifier ces hypothèses selon un plan d'échantillonnage correctement déterminé.

---

3. BAIZE D., GIRARD M.C, 2000

4. BAIZE D., JABIOL B., 1995

5. Baize et Jabiol 1995

Les échantillons ainsi prélevés sont analysés en laboratoire. Toutes ces informations aboutissent à la délimitation ou au traçage des contours des unités de sols et permettent au pédologue cartographe de créer une carte des sols.

La cartographie conventionnelle étant un processus long et mobilisant beaucoup de temps et de ressources matérielles et humaines, il était important de rechercher de nouvelles solutions permettant de pallier ces difficultés. La cartographie numérique ou « Digital Soil Mapping » est une alternative à la cartographie pédologique traditionnelle. Les modèles numériques de cartographie des sols ont pour objectif d'utiliser les relations qui lient le sol à son environnement. La cartographie numérique est un moyen d'améliorer le travail du pédologue. Il faut implémenter des méthodes automatiques pour obtenir une carte de prédiction des sols grâce à des variables prédictives. Le résultat de la carte n'est qu'une approche mais cela permet d'alléger le travail du pédologue<sup>6</sup>.

L'hypothèse de base de la cartographie numérique de sol est formulée sous le modèle CORPT qui explique qu'un sol peut-être prédit par d'autres caractéristiques, des variables prédictives comme le climat, les organismes vivants, le relief, le matériau parental et le temps. Jenny énonce en 1941 cette approche sous forme d'équation :

$$S = f(c, o, r, p, t)$$

Plus récemment, McBartney et al (2003) dressent un état de l'art de la cartographie numérique des sols et proposent une nouvelle version de l'équation de Jenny en y intégrant la notion de voisinage. Ce nouveau modèle s'appelle SCORPAN :

$$S = f(c, o, r, p, a, n)$$

où S correspond au sol (propriétés du sol), c au climat, o aux organismes vivants, r au relief (attributs de terrain), p au matériau parental (lithologie), a à l'âge (facteur temps) et n à l'espace (la position spatiale).

Aujourd'hui, le nombre de modèles prédictifs a augmenté fortement, grâce principalement au développement des informations spatiales numériques et des techniques de traitement de ces données. CLAPAS est un des modèles de prédiction.

---

6. McBRATNEY A., MENDONÇA SANTOS M.L, MINASNY B., 2003

### 2.1.3 CLAPAS

CLAPAS (CLAssement des PAysages et Segmentation) est un logiciel développé par Jean-Marc ROBBEZ-MASSON en 1994. Ce logiciel est un outil de segmentation d'une image numérique. L'algorithme travaille sur un voisinage cohérent de chaque point qu'il cherche à classer. Plus précisément, il s'agit de faire évoluer une fenêtre sur une image pour en extraire les données nécessaires à l'étude.

### 2.1.4 Principe

L'utilisation de CLAPAS s'effectue à partir de données et de cartes ayant subi un pré-traitement par SIG<sup>7</sup>. L'image utilisée dans la procédure de CLAPAS est une image issue de superposition de plusieurs cartes thématiques. Ces données sont donc regroupées en une seule carte, où chaque pixel est identifié par une combinaison spécifique de ces données. Cette carte est une des données d'entrée l'algorithme.

La seconde condition est de disposer de secteurs de référence (ou des unités cartographiques). Cette carte ne représente qu'une petite zone de la carte pédologique totale. Cette petite partie contient toute les informations nécessaires pour prédire la carte des sols. Á chaque type de sol est attribué un nombre entier.

Ces deux données sont primordiales pour le lancement de CLAPAS. Elles permettent de créer des histogrammes de référence en utilisant chaque secteur de référence comme masque pour extraire l'information dans la matrice des variables prédictives. Ensuite, cet algorithme parcourt la matrice des variables prédictives point par point. Sur chaque point, on étudie la composition du voisinage et on la compare avec la composition des paysages de références par le biais de la distance de Manhattan. On attribue alors à ce pixel le type de paysage dont la composition est la plus proche, dont la distance est minimale.

Les résultats fournis par CLAPAS sont des images.

---

7. Système d'Information Géographique

### **Carte des résultats de classement**

L'image des résultats de classement accorde à chaque pixel le type de paysage ou la classe pour lequel la composition du voisinage comparée avec la composition des secteurs de référence est la plus proche. Cette carte est une esquisse pour le pédologue. Elle lui permettra de réduire son temps de travail sur le terrain si les types de sol sont affectés au bon endroit. Pour vérifier cela, il y a la seconde carte produite par CLAPAS.

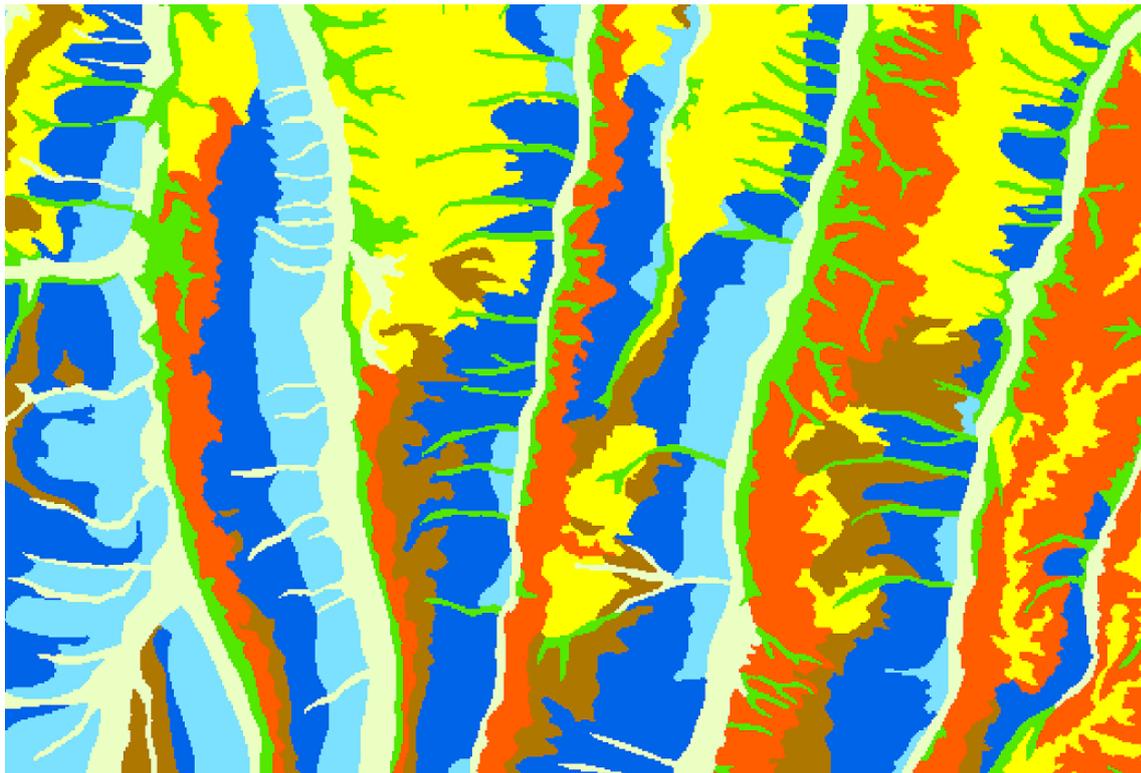
### **Carte des distances**

Cette seconde image fournie par CLAPAS indique pour chaque pixel si il est correctement classé ou au contraire si sa classification est douteuse et problématique. Elle représente donc un indicateur de la qualité du classement. Les distances proche de 0 seront affectées aux pixels très bien classés alors que les distances proches de 2 indiquent que les pixels ont été mal classés. Cette carte ne donne qu'une indication et doit être considérée avec précaution.

### **Carte de validation**

Cette carte de validation est la carte pédologique où sont répartis les différents types de sols. Cette carte va servir à la fois de référence et de validation pour les traitements à venir.

Elle va nous être utile pour extraire les secteurs de référence. Ces zones correspondent à des zones bien connues du service de cartographie des sols. Ils font partie des paramètres d'entrée indispensables du modèle CLAPAS.



**Types de sol**

- |  |  |   |
|--|--|---|
|  0            |  Rendosols  |  Luvisols dégradés           |
|  Fluviosols   |  Calciosols |  Luvisols dégradés leptiques |
|  Colluviosols |  Brunisols  |   |

FIGURE 2.1 – Carte pédologique de la feuille de Mirande (Gers).

## 2.1.5 Les outils

### Le Langage R

R est un système d'analyses statistiques et graphiques créé par ROSS IHAK et ROBERT GENTLEMAN.

R est distribué librement sous les termes de la GNU General Public Licence. Il est développé depuis plus de 10 ans par un groupe de volontaires de différents pays et sa communauté d'utilisateurs ne cesse de s'élargir. C'est un logiciel libre, développé par ses utilisateurs et modifiable par tout le monde, gratuit, très puissant grâce à sa multitude de packages. Il existe près de 1 500 packages disponibles. Ils couvrent un très large champs et permettent à R d'être très fonctionnel.

### ArcGis

ARCGIS est un logiciel permettant de traiter des données géographiques pour comprendre, interpréter et modéliser la structure, l'organisation et le fonctionnement d'un espace étudié. C'est donc un outil important pour un pédologue cartographe.

Ces deux outils étaient les outils indispensables lors du stage, R pour la programmation et arcGIS pour l'affichage et la gestion des données.

## Partie 3

# Le stage et les résultats

Après avoir présenté CLAPAS (CLassement de Paysages et Segmentation) dans la section précédente, nous allons maintenant nous intéresser aux travaux effectués durant le stage et aux résultats obtenus.

### 3.1 Objectifs du stage

L'objectif de ce stage était de faire évoluer la forme de la fenêtre, qui sert de masque lors de l'exécution de l'algorithme, en fonction de l'intensité de la pente. Au cours d'une réunion de travail, la question sur l'image d'entrée du modèle s'est posée. Peut-être contient-elle trop de données ? Et qu'elles empêchent une bonne détection des différents types de sols ? Après cette réunion, il m'a été confiée d'améliorer l'image d'entrée en utilisant les arbres de classification et d'automatiser la procédure.

Les objectifs de mon stage ont évolué au cours du temps. Pour présenter les choses de façon logique et cohérente, nous allons suivre la démarche suivante : tout d'abord la fabrication de l'image d'entrée puis l'évolution automatique de la forme de la fenêtre.

### 3.2 Les zones d'étude

Mes zones d'étude étaient le secteur de Mirande dans le Gers (23 km<sup>2</sup>) et le sud de la Corrèze, en particulier le secteur de Brive La Gaillarde (50 km<sup>2</sup>). Pourquoi

avoir travaillé sur deux zones ? Le travail sur la forme de la fenêtre correspond à la poursuite des travaux réalisés par OLIVIER VITRY qui avait exécuté CLAPAS sur les images de Mirande. Mais pour l'image d'entrée, il a été décidé de prendre un autre secteur, celui du Sud de la Corrèze. Il a déjà été utilisé lors du stage de CLOVIS GRINAND qui se servait de l'algorithme MART (Multiple Additive Regression Tree). Nous n'avons pas utilisé cet algorithme pour obtenir notre modèle d'entrée, ce qui permettra de comparer les résultats.

### 3.3 Travail effectué

L'algorithme CLAPAS a pour but de prédire la carte numérique des sols d'un secteur donné. Mais pour obtenir cette carte, il est indispensable d'avoir en entrée une carte où chaque point, chaque pixel, correspond à une combinaison de ces variables prédictives (climat, relief, matériau parental). Chaque combinaison a sa propre signification, son propre code.

#### 3.3.1 Amélioration de l'image d'entrée du modèle CLAPAS

La méthode utilisée pour créer la nouvelle carte d'entrée s'appelle CART, Classification and Regression Tree, initialement développée par Breiman et al en 1984<sup>1</sup>. Elle réalise des arbres de classification ou de régression qui sont des outils d'aide à la décision permettant de modéliser simplement, graphiquement et rapidement un phénomène mesuré plus ou moins complexe.

CART a été adapté pour la prédiction cartographique des sols à partir d'un secteur de référence par Lagacherie (1992). Le principe est de créer une segmentation binaire en découpant le jeu de données d'entrée de manière récursive en des sous-ensembles de plus en plus homogènes. Chaque division est choisie de façon à maximiser la distance interclasse entre les deux groupes formés à chaque noeud. Chaque noeud est donc défini comme un ensemble de règles conditionnelles fondées sur les variables explicatives. L'arbre créé est généralement volumineux avec des noeuds terminaux multiples, ce qui signifie que le modèle est intimement influencé par le jeu

---

1. GRINAND C., ARROUAYS D., LAROCHE B., MARTIN M., 2008

de données d'entrée. Une des caractéristiques importantes de CART est qu'il donne un aperçu quantitatif de l'ensemble de données. En outre, les variables relativement importantes peuvent être soulignées par le nombre de fois qu'elles apparaissent dans les nœuds. Cependant, les incohérences dans le jeu de données, comme le bruit et les valeurs aberrantes, peuvent grandement affecter la précision du résultat.

L'objectif poursuivi est de prédire des unités cartographiques de sol à partir d'arbres de classification.

Les données d'entrée du secteur de la Corrèze pour réaliser l'image combinée sont nombreuses. Tout d'abord, nous avons besoin des indices dérivés du MNT (modèle numérique de terrain) : la pente, l'indice de Beven, les courbures de la pente (verticale, horizontale ou moyennée), la distance hydrologique au plus proche réseau de drainage et la hauteur au plus proche réseau de drainage . Il permettent de mettre en avant certaines caractéristiques morphologiques, hydrographiques ou d'ensoleillement du terrain d'étude. Il existe beaucoup plus d'indices dérivés du MNT, plus ou moins complexes dans leur calcul, mais seulement les cinq indices cités vont être utilisés<sup>2</sup>.

- **la pente** est la première dérivée de l'altitude. C'est une variable importante car elle permet de différencier facilement les sols de plaine ou des plateaux des sols de versants. La figure 3.1 donne une vision des différentes échelles de description du relief et la variabilité des pentes dans le paysage.
- **Compound Topographic Index** calculé grâce à la formule de l'indice de **Beven** représente la saturation potentielle en eau des sols calculé avec

$$Beven = Ln(A/B)$$

A : surface drainée

B : pente locale en pourcentage

Des valeurs élevées de l'indice de Beven indiquent une accumulation potentielle

---

2. GRINAND C., 2005-2006

de l'eau dans les sols causée par des pentes faibles et par une surface drainée importante. Il représente un bon indicateur des fluviolosols et des colluviosols.

- **les courbures** représentent la convexité du terrain, soit selon le plan horizontal, soit selon le long de la ligne de plus grande pente. Ils permettent de mettre en évidence les processus d'accélération et de concentration des flux hydrologiques de surface.
- **la distance hydrologique au plus proche réseau de drainage** donne la distance de chaque pixel au talweg<sup>3</sup>. Elle permet de différencier des sols proches d'une rivière (valeurs faibles) des sols plus éloignés présents sur des terrasses ou les versants du bassin.
- **la hauteur au plus proche réseau de drainage** donne la hauteur de chaque pixel au talweg. C'est une mesure d'altitude relative qui peut-être un facteur important dans le cas de la discrimination de sols de versant.

Dans cette étude, nous avons également considéré d'autres paramètres de terrain qui sont souvent utilisés et déterminants : les bandes de Landsat ETM<sup>4</sup> et l'indice de différence de végétation normalisé (ndvi).

- **Les bandes de Landsat ETM** : le programme Landsat est un des plus anciens programmes d'observation par satellite de la surface terrestre. Les instruments embarqués sur les satellites Landsat ont permis l'acquisition de plusieurs millions d'images. Elles constituent des ressources uniques pour l'étude des changements climatiques, l'utilisation des sols, la cartographie, . . . Ces bandes sont des bandes de fréquences. Le capteur Multispectral Landsat permet de mesurer la réflectance de la Terre sur 5 canaux dont trois dans le domaine visible (canaux 1, 2, 3), un dans le proche infrarouge (canal 4), deux canaux dans le moyen infrarouge (canaux 5 et 7) à des résolutions au sol de 30 mètres et

---

3. correspond à la ligne qui rejoint les points les plus bas d'une vallée

4. Enhance Thematic Mapper

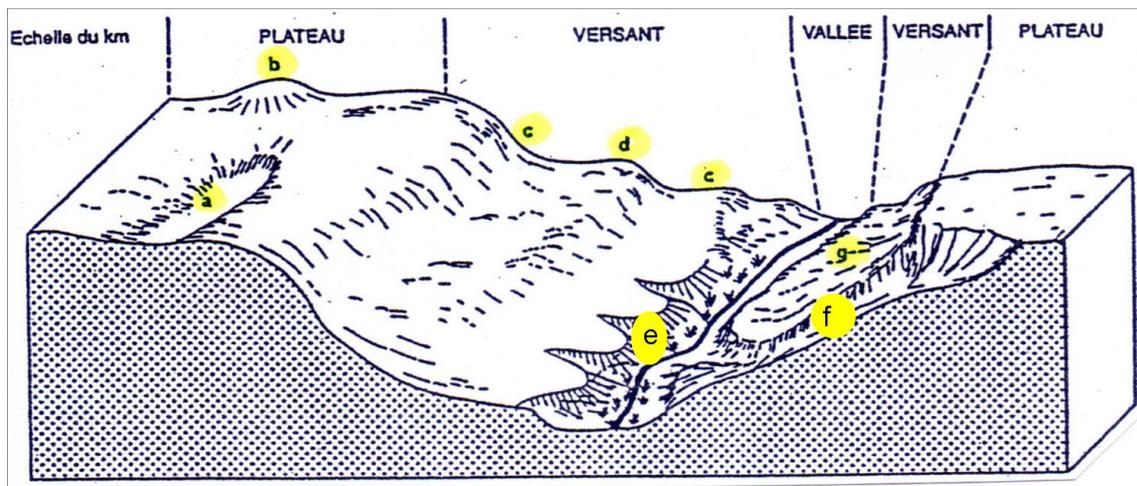


FIGURE 3.1 – différentes « échelles » de description du relief et variabilité des pentes dans le paysage (d'après M.C. Girard)

a. vallon sec au sein d'un plateau

b. bombement sur plateau

c. la pente varie constamment le long de ce grand versant

d. contre-pente

e. « rentrant » et « saillant »

f. le versant raide de ce ravin a une toute autre exposition que le versant g pris globalement

également un canal infrarouge thermique à 60 mètres (canal 6). Il y a aussi un canal panchromatique à 15 mètres, englobant les spectres du visible et du proche-infrarouge. Ce dernier n'est pas utilisé lors de notre étude car l'information qu'il renvoie est difficile à interpréter à cause de sa faible résolution spectrale.

- ***L'indice de différence de végétation normalisé*** est très courant dans de nombreuses applications de télédétection car il met en évidence la végétation, ce qui permet de cartographier les différents couverts végétaux ou de faire des estimations de biomasse. Il est calculé à partir des canaux du proche infrarouge (TM4) et du canal rouge (TM3) donné par la formule suivante :

$$ndvi = \frac{(TM4 - TM3)}{(TM4 + TM3)}$$

Dans le cas des sols, cet indice est très pertinent car la végétation peut-être un facteur pédogénétique prépondérant. De plus, le type de végétation et sa condition sont des indicateurs de certaines propriétés des sols comme par exemple la teneur en matière organique.

Toutes les variables citées sont quantitatives. Elles peuvent toutes s'exprimer en une certaine mesure (le mètre par exemple) alors que les variables qualitatives sont référencées en classes comme la géologie, l'occupation du sol et les unités cartographiques.

- ***La géologie*** est une synthèse au 1/250 000. Elle est représentée par 15 classes.
- ***L'occupation du sol ou Corine Land Cover*** fournit des informations sur la diversité et l'organisation du sol. Son échelle est de 1/100 000. 16 classes la composent et la décrivent.
- ***Les unités cartographiques du sol*** constituent la variable à prédire

grâce à toutes les autres. 25 classes représentent les pédopaysages.

Après avoir défini toutes ces variables, on peut exécuter CART grâce au package *rpart*<sup>5</sup>. Notre objectif est de récupérer les noms des variables et leurs valeurs respectives indiqués sur les branches, au niveau des nœuds de l'arbre, pour pouvoir ensuite seuiller les images dont les noms sont apparus dans l'arbre (figure 3.2).

Le package *rpart* contient une multitude de programmes pour travailler sur les arbres de classification, toutes mises à la disposition de l'utilisateur, mais aucune ne permet d'obtenir le nom des variables et leurs valeurs correspondantes.

La création du programme, permettant de récupérer les informations aux nœuds des branches puis de seuiller les images dépendant uniquement de ces informations, fut un travail motivant. Il fallait garder à l'esprit que le programme créé devait être accessible et automatique pour que tout utilisateur puisse l'exécuter facilement. Les informations recueillies sont stockées dans un tableau. Dans la première colonne, on trouve les noms des variables et dans la seconde, les valeurs respectives.

Ces informations ne sont pas triées par ordre croissant. En effet, chaque nœud définit une séparation binaire de la branche. Les renseignements qui nous sont fournis aux nœuds de chaque branche correspondent à l'information de la division du groupe de gauche. Au niveau de chaque nœud, la règle est d'abord de parcourir le côté gauche puis de revenir au nœud et ensuite parcourir de le côté droit. Par exemple, le tableau obtenu grâce à la fonction `detect.node` pour l'arbre de classification de la figure 3.3 est :

	nom	value
1	mnt	409.5
2	mnt	264.5
3	pente	6.703
4	pente	21.97
5	pente	2.026

---

5. FARAWAY J.J. 2006

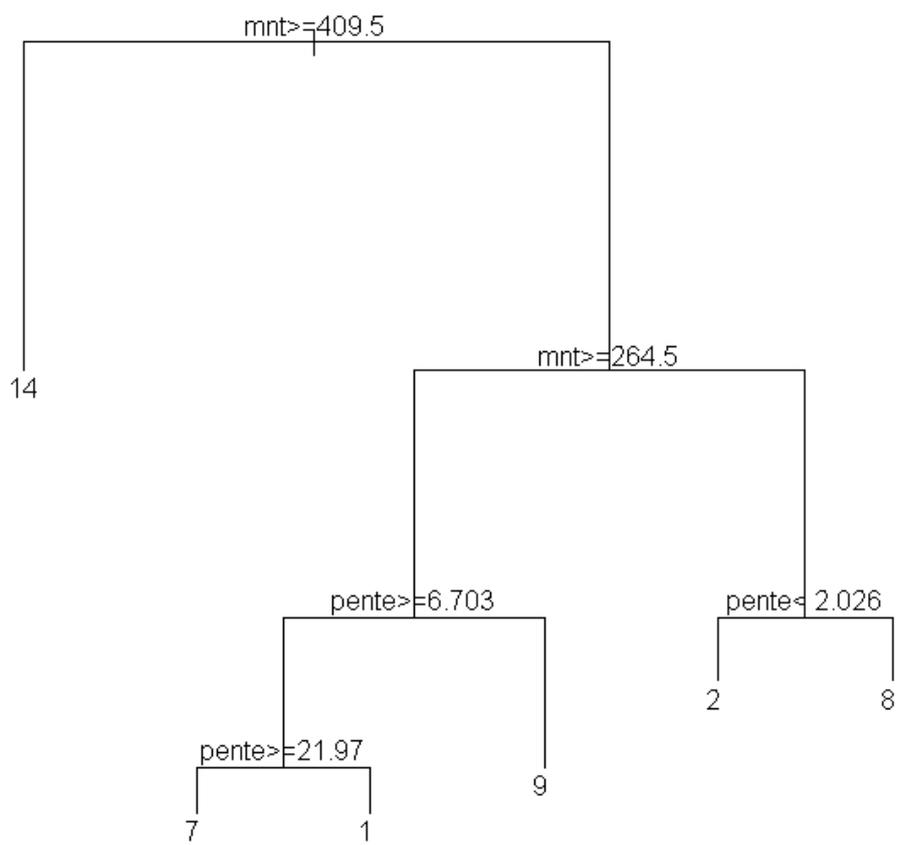


FIGURE 3.2 – Arbre de classification pour le secteur de la Corrèze

Après avoir créé ce tableau, notre objectif est maintenant de trier ces valeurs en ordre croissant en tenant compte de la variable pour pouvoir ensuite seuiller les images qui ont eu leurs noms inscrits dans le tableau. Cette étape a nécessité un important effort de synthèse dans l'automatisation du processus. Ce qui n'était pas le cas au début avec trois programmes, l'un permettait d'obtenir le tableau ci-dessus, le second donnait le même tableau mais trié, et le dernier seuillait image par image. Entre chaque fonction, l'utilisateur avait plusieurs commandes à exécuter. Il a fallu combiner les trois programmes en un. La fonction du code à laquelle il est fait référence ici est la fonction `seuillage`.

Plusieurs nouvelles images ont été créées. Elles vont nous permettre de réaliser l'image combinée. Précisons que CART a été produit uniquement avec les variables quantitatives. Les variables qualitatives comme `geol` et `clc` sont ajoutées après, lorsqu'est créé l'image combinée. Ces variables, si elles apparaissent sur l'arbre, sont difficile à lire. Elles sont définies par des combinaisons qui sont codées en lettres sur l'arbre. Plus précisément, chaque combinaison correspond à une lettre de l'alphabet.

Donc, en reprenant les images seuillées et en y ajoutant les cartes de la géologie et de l'occupation du sol, on obtient l'image combinée souhaitée (figure **3.3**).

Cette carte obtenue est le principal élément pour l'exécution de CLAPAS. À partir de ce résultat, on peut donc essayer de prédire la carte des sols qui doit être bien évidemment de mêmes dimensions que l'image combinée. Maintenant que tous les éléments sont prêts à être utilisés, nous allons pouvoir aborder le cas de la fenêtre de forme elliptique.

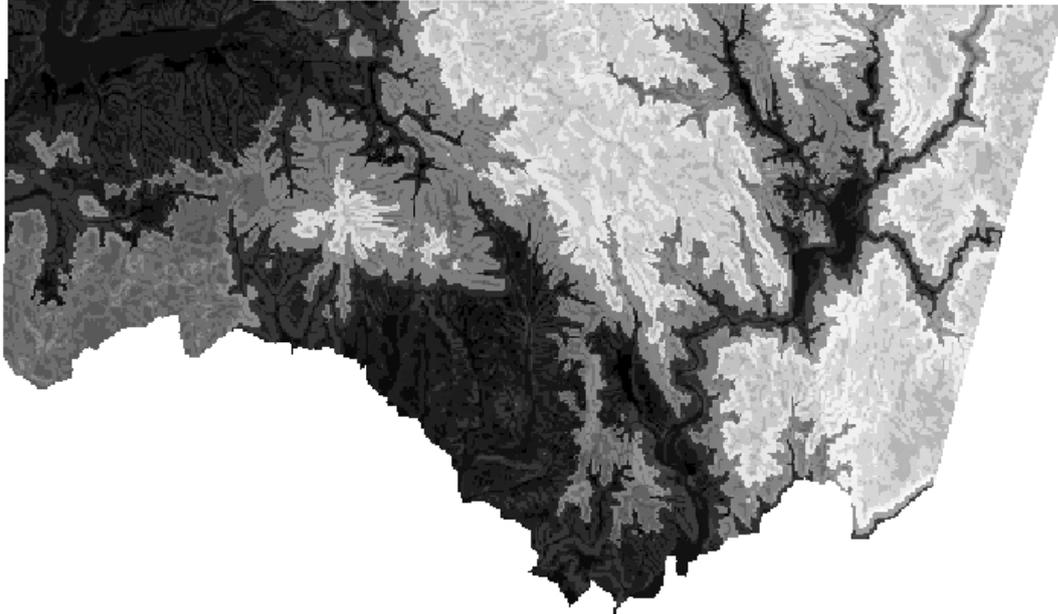


FIGURE 3.3 – Création de la carte combinant les variables prédictives (Brives)

## 3.4 Adaptation de la fenêtre de forme elliptique

### 3.4.1 Création de la fenêtre

La création d'une fenêtre de forme elliptique a été réalisée par OLIVIER VITRY. Il a fallu pouvoir générer la forme tout en gardant à l'esprit que l'objet ainsi créé doit servir de masque à placer sur l'image combinée. Il a choisi de représenter ce masque par une matrice  $K$  à valeurs booléennes, où la matrice vaut `FALSE` à l'extérieur de la fenêtre et `TRUE` à l'intérieur. La fonction correspondante est la fonction `ellipse`.

Lors de mon stage, il m'a été demandée de faire évoluer la forme de la fenêtre en fonction de l'intensité de la pente. L'intensité des pentes est directement donnée par le gradient de l'altitude. Grâce à l'application des deux masques de Sobel à la carte

de l'altitude, nous pouvons calculer le gradient.

Le masque de Sobel est très utilisé en traitement d'image pour la détection de contours. Il permet d'ignorer un bruit léger.

### 3.4.2 Création de la fonction pente

Pour calculer l'intensité de la pente, on a utilisé les masques  $\begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix}$  pour la composante principale du gradient  $\nabla_x$  et  $\begin{pmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{pmatrix}$  pour la composante verticale du gradient  $\nabla_y$ .

On a donc pris pour calculer chaque intensité de pente pour chaque pixel, une matrice de même taille que les masque c'est-à-dire 3\*3. Seuls les bords la carte de l'altitude n'ont pas pu être déterminés. Par exemple, en prenant cette matrice,

$\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}$  on ne peut calculer que l'intensité de la pente pour le pixel central. Les cartes utilisées sont échantillonnées avec un pas de 50 mètres. Donc, on multiplie cette matrice par les deux masques, on obtient,

$$dx = \frac{(c + 2 \times f + i) - (a + 2 \times d + g)}{(8 * 50)}$$

$$dy = \frac{(g + 2 \times h + i) - (a + 2 \times b + c)}{8 * 50}$$

Ces deux résultats vont être combinés pour obtenir la norme du gradient :

$$D = \sqrt{(dx^2 + dy^2)}$$

Il faut ensuite déterminer la pente soit en degrés soit en pourcentages :

$$slope\_deg = \arctan(D) \times \frac{180}{\pi}$$

$$slope\_pourcent = \arctan(D) \times 100$$

La fonction, qui permet de créer à partir de la carte de l'altitude, la carte des intensités de pente est la fonction `pente`.

### 3.4.3 Adaptation de la taille de la fenêtre

En obtenant l'intensité des pentes et en utilisant l'orientation des pentes, il m'a fallu adapter la taille et la forme de la fenêtre elliptique en fonction de ces deux paramètres. Plus la valeur de la pente est faible, plus le terrain est plat ; et inversement plus la valeur de la pente est élevée, plus le terrain est pentu. D'un point de vue pédologique, il est intéressant d'appliquer à une pente faible une ellipse de forme circulaire et au contraire à une pente élevée, une ellipse très allongée.

La fonction `c1f` prend en paramètre l'intensité et l'orientation de la pente mais également deux vecteurs. Le premier vecteur correspond aux seuils de pente à partir desquels l'ellipse change de forme. Le second, quant à lui, détermine les valeurs des demi-axes des ellipses en fonction des seuils de pente. Les seuils sont choisis par l'utilisateur et seront toujours classés par ordre croissant. La première valeur sera toujours le minimum de la matrice de pente et la dernière valeur `max(p)+1` . Entre ces valeurs, l'utilisateur pourra choisir les valeurs qu'il souhaite.

Le second vecteur contient les différentes tailles de l'ellipse. Il dépend du premier vecteur. En effet, l'expérience du pédologue montre qu'en situation plane (pente de 0 à 1°) il est recommandé d'utiliser une ellipse circulaire (les deux demi-axes ont la même valeur) et de faire évoluer l'allongement de l'ellipse en fonction de l'intensité de la pente.

### 3.4.4 Résultats des classements

Les résultats générés par CLAPAS sont des cartes. Malgré avoir travaillé pendant trois mois sur ces objets, il me paraît difficile de pouvoir valider ces résultats sans les connaissances nécessaires de pédologie. La carte des distances et les matrices de confusion vont permettre de nous renseigner sur la pertinence de la classification .

La carte de prédiction obtenue peut être comparée à la carte pédologique juste en les regardant. En effet, elle apparaît bruitée et certains types de sols sont mal détectés comme les fluvisols et les calcosols. Ceci est peut-être dû aux tailles et aux formes des ellipses qui ne sont pas adaptées pour ces types de sols et également à l'image combinée de départ qui semble contenir trop de données. Mais pour confirmer cette interprétation visuelle, il faut se ramener à des outils mathématiques.

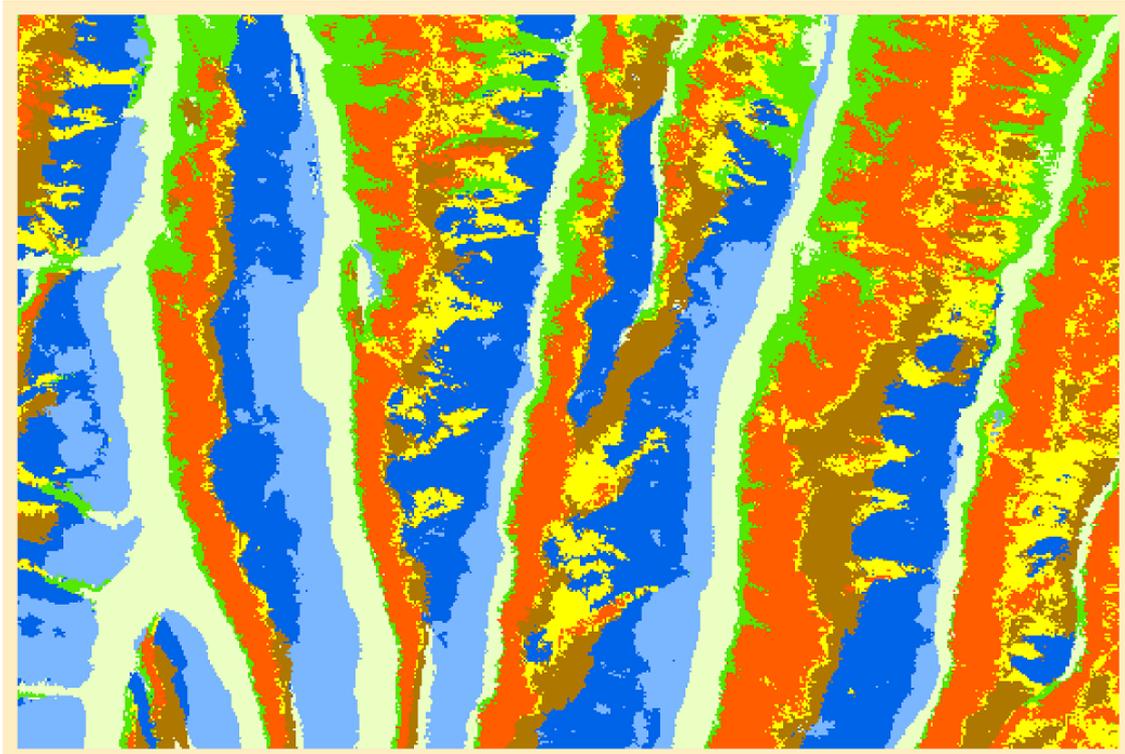
#### La carte des distance

Cette carte indique si les types de sols ont bien été classés. Les zones en vert signalent que les unités paysagères ont été bien classées alors que les zones en rouge indiquent qu'elles ont été mal classées. Mais cette carte ne donne qu'une suggestion sur la classification. Le pédologue ne peut pas s'y fier totalement.

#### La matrice de confusion

La matrice de confusion est un outil statistique permettant de mesurer la précision de classification d'une image. Elle permet d'apporter des renseignements sur le nombre de pixels de la zone d'apprentissage ou de la carte pédologique qui ont conservé leur classe après l'exécution.

Les valeurs de la diagonale de la matrice représentent le nombre de pixels correctement classés. Ces valeurs vont permettre de calculer **l'indice global**. En effet, la somme des valeur de la diagonale de la matrice correspond au nombre total de pixels bien classés. En divisant cette somme par le nombre total de pixels de l'image, on obtient l'indice global de classification exprimé en pourcentage.



### Types de sol

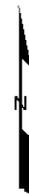
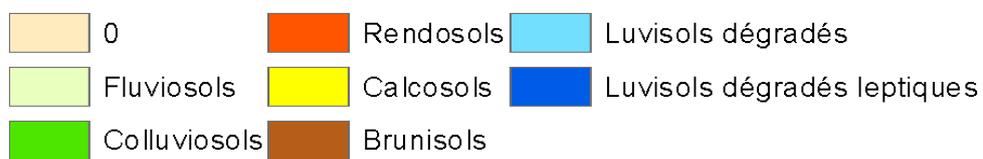
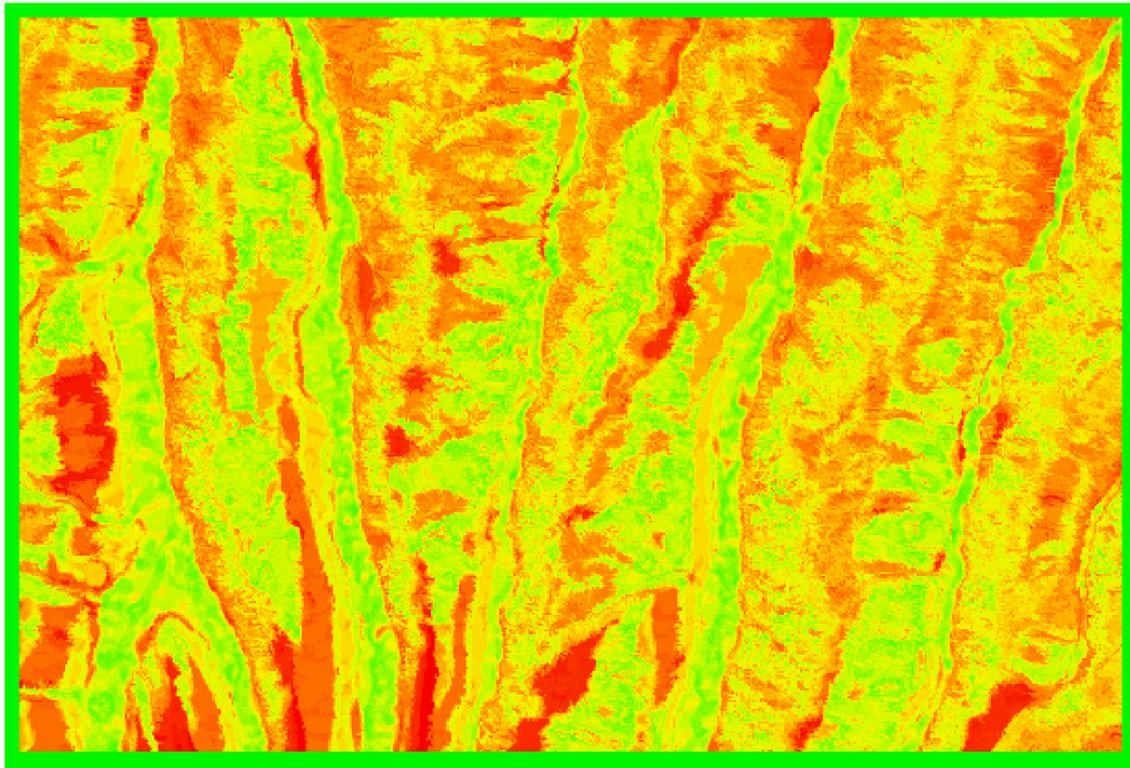


FIGURE 3.4 – Prédiction des unités cartographiques générées par CLAPAS



**Distances**



**clapas\_d**

**Valeur**

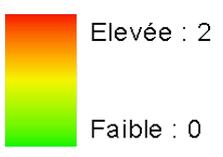


FIGURE 3.5 – Carte des distances associée à la prédiction des unités cartographiques générées par CLAPAS

L'erreur d'omission et l'erreur de commission sont calculées pour chaque classe. Pour les déterminer, il ne faut pas considérer les valeurs de la diagonale. Le total des valeurs (hors la valeur sur la diagonale) de chaque colonne représente le nombre de pixels qui ont été affectés dans d'autres classes que la classe de la colonne. Ce total est alors divisé par le nombre total de la colonne : on obtient **l'erreur d'omission**. Le total des valeurs (hors la valeur de la diagonale) de chaque ligne représente le nombre de pixels qui ont été affectés dans la classe de la ligne alors qu'ils appartenaient aux autres classes. Ce total est divisé également par le nombre total de la ligne : on obtient **l'erreur de commission**.

**L'indice Kappa** est un estimateur de la précision variant entre 0 et 1.

classes affectées	classes de référence							total aff	erreur de commission
	1	2	3	4	5	7	8		
1	17146	3000	75	199	25	4484	1539	26468	0.35
2	2143	7434	2762	4501	164	194	1424	18622	0.60
3	333	2988	25907	10178	2054	0	859	42319	0.39
4	100	1429	3511	9046	2439	99	3269	19893	0.55
5	127	922	746	5629	6211	457	4785	18877	0.67
7	1757	487	0	42	174	14292	8340	25092	0.43
8	1318	1870	228	2770	4023	4777	23310	38296	0.39
total réf	22924	18130	33229	32365	15090	24303	43526	189567	
erreur d'omission	0.25	0.59	0.22	0.72	0.59	0.41	0.46		

indice global	$G = 0.55$
---------------	------------

Les résultats obtenus (figures **3.4** et **3.5**) vont être comparées à une prédiction avec comme fenêtre un cercle de 7 pixels de diamètre<sup>6</sup>.

---

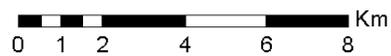
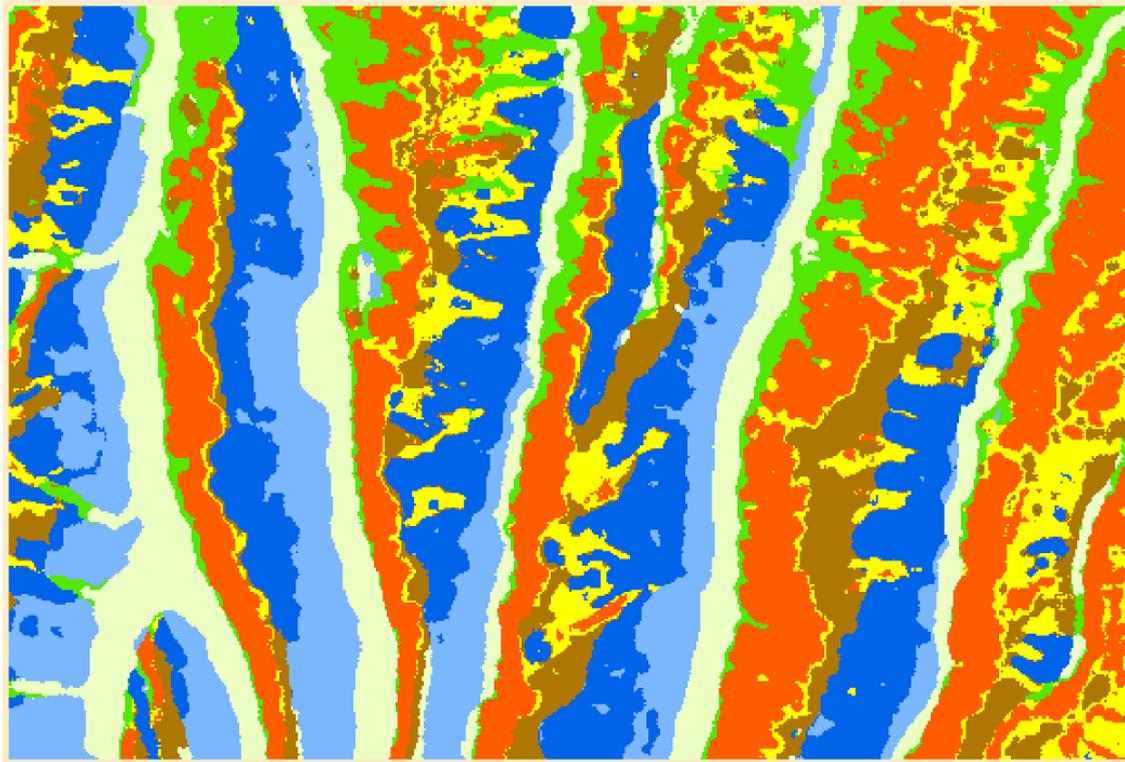
6. tous ces résultats sont générés par `clapas_niv`

Visuellement, la carte des distances (figure **3.5**) est autant marquée de rouge que celle obtenue avec comme fenêtre un cercle (figure **3.7**). Cette comparaison visuelle indique qu'il y aurait plus de pixels bien classés sur la première que sur la seconde. La matrice de confusion va nous permettre de nous faire une idée sur la classification.

classes affectées	classes de référence							total aff	erreur de commission
	1	2	3	4	5	7	8		
1	17652	2882	14	176	18	4647	1528	26917	0.34
2	2229	7141	2062	4374	133	258	1440	17637	0.60
3	565	3833	27908	11077	2085	0	834	46302	0.40
4	94	1678	4104	9355	2445	53	3220	20949	0.55
5	111	821	805	6123	6831	495	4939	20125	0.66
7	1727	511	0	26	176	15059	9049	26548	0.43
8	1380	1821	221	2960	4284	4585	23911	39162	0.39
total réf	23758	18687	35114	34091	15972	25097	44921	197640	
erreur d'omission	0.26	0.62	0.21	0.73	0.57	0.40	0.47		

indice global	$G = 0.55$
---------------	------------

L'indice global de classification est de 55% identique à celui de la matrice précédente. Ceci indique donc que la taille et la forme de l'ellipse ne sont pas optimales pour détecter tous les unités cartographiques.



**Types de sol**

 0	 Rendosols	 Luvisols dégradés
 Fluviosols	 Calcosols	 Luvisols dégradés leptiques
 Colluviosols	 Brunisols	

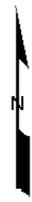
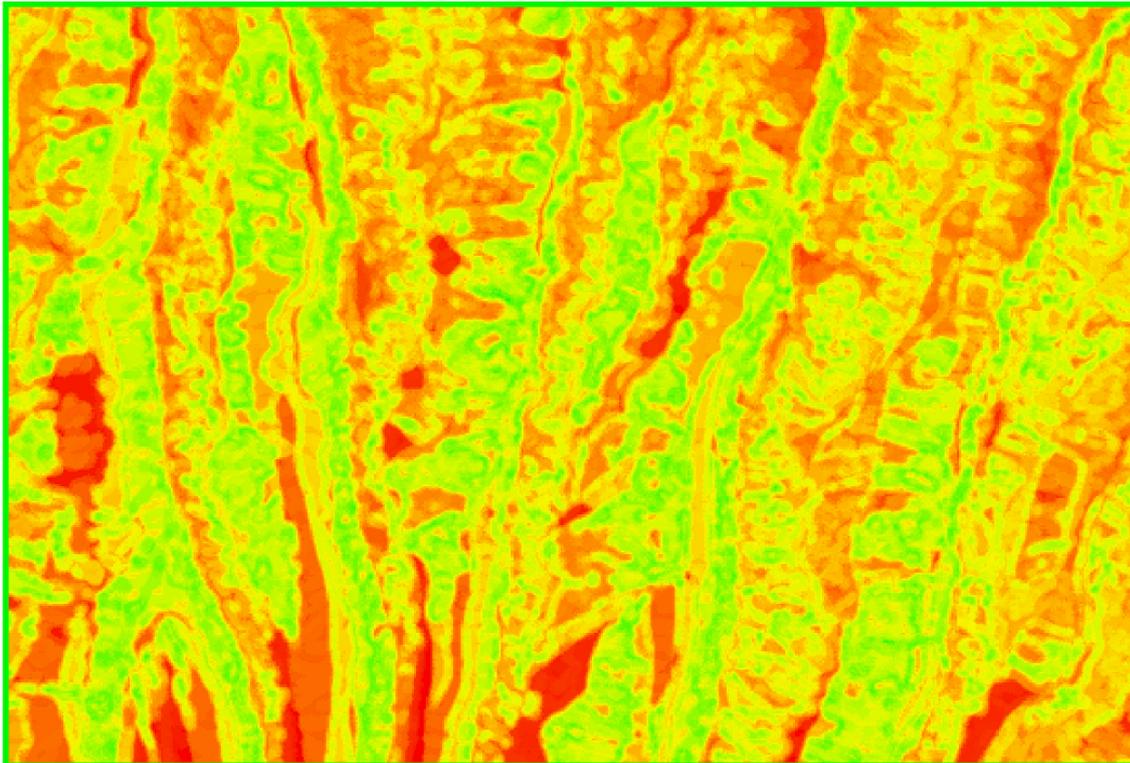


FIGURE 3.6 – Prédiction faite par CLAPAS avec comme fenêtre un cercle de 7 pixels de diamètre.



### Distances

clapas\_d

Valeur

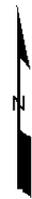
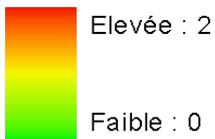


FIGURE 3.7 – Carte des distances associée à la prédiction avec un cercle de 7 pixels de diamètre.

## Nouveaux résultats avec la nouvelle image combinée de Mirande

L'image est obtenue grâce à la méthode automatisée du seuillage des images grâce aux arbres de classification. Elle va donc être utilisée lors de l'exécution de CLAPAS. Les résultats seront comparés aux résultats trouvés précédemment avec comme fenêtre un cercle de 7 pixels de diamètre.

classes affectées	classes de référence							total aff	erreur de commission
	1	2	3	4	5	7	8		
1	15342	2115	9	347	13	5029	1746	24601	0.38
2	3951	8873	2641	4714	291	779	1663	22912	0.61
3	1040	3196	25845	8713	1627	24	808	41253	0.37
4	246	1520	5743	12255	2969	59	3227	26019	0.53
5	154	725	669	5089	6763	244	5086	18730	0.64
7	2094	1288	0	994	288	14343	8084	27091	0.47
8	931	970	207	1979	4021	4619	24307	37034	0.34
total réf	23758	18687	35114	34091	15972	25097	44921	197640	
erreur d'omission	0.35	0.53	0.26	0.64	0.58	0.43	0.46		

indice global	$G = 0.55$
---------------	------------

Avant, la création de l'image combinée d'entrée résultait de la combinaison d'images seuillées sur le principe des boîtes à moustaches ou boxplot et d'une analyse en composantes principales (ACP). Maintenant, la création de l'image combinée est automatisée par le processus de choix et de seuillage des indices les plus pertinents

Le résultat de la matrice de confusion produite à partir de l'image combinée d'entrée obtenue par procédure automatisée montre que l'indice global reste identique

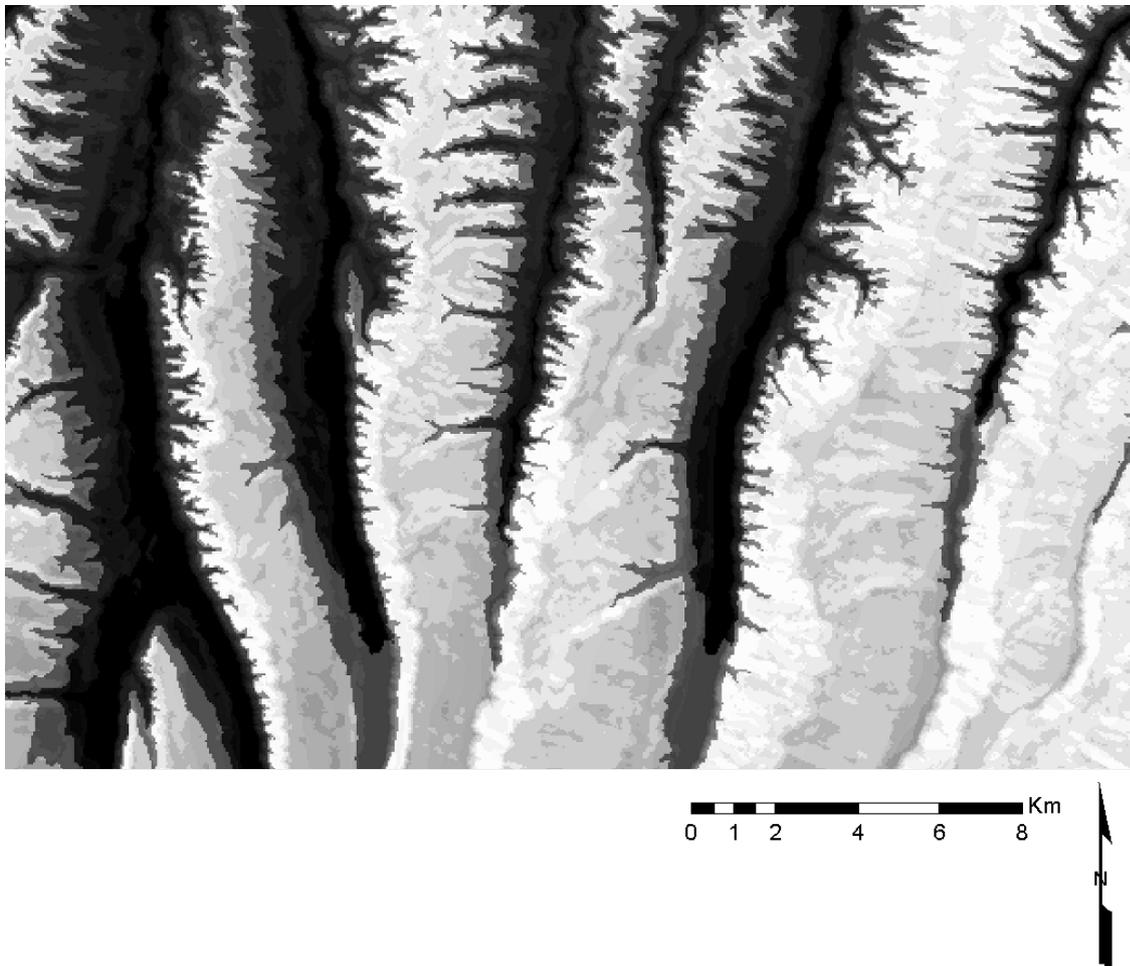


FIGURE 3.8 – Carte des variables prédictives combinées de la feuille de Mirande

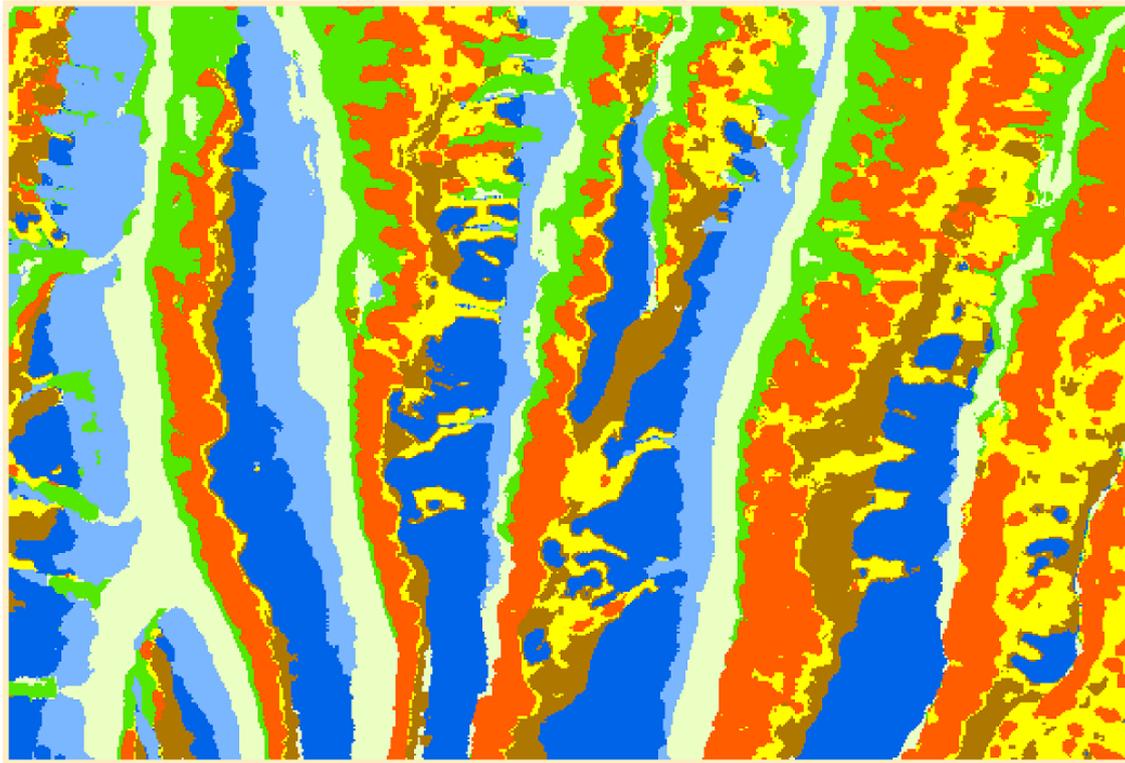
sur une validation externe<sup>7</sup> à celui de la matrice de confusion de l'image d'entrée obtenue par boxplot et ACP. Il augmente même sur la validation interne<sup>8</sup> (0.61 au lieu de 0.58).

On peut donc conclure que la procédure automatisée développée lors du stage est aussi performante, voir meilleure, que la procédure qui faisait appel à plusieurs outils mathématiques.

---

7. validation entre l'image résultat et la carte pédologique des sols

8. validation entre l'image résultat et la zone d'apprentissage (secteurs de référence)

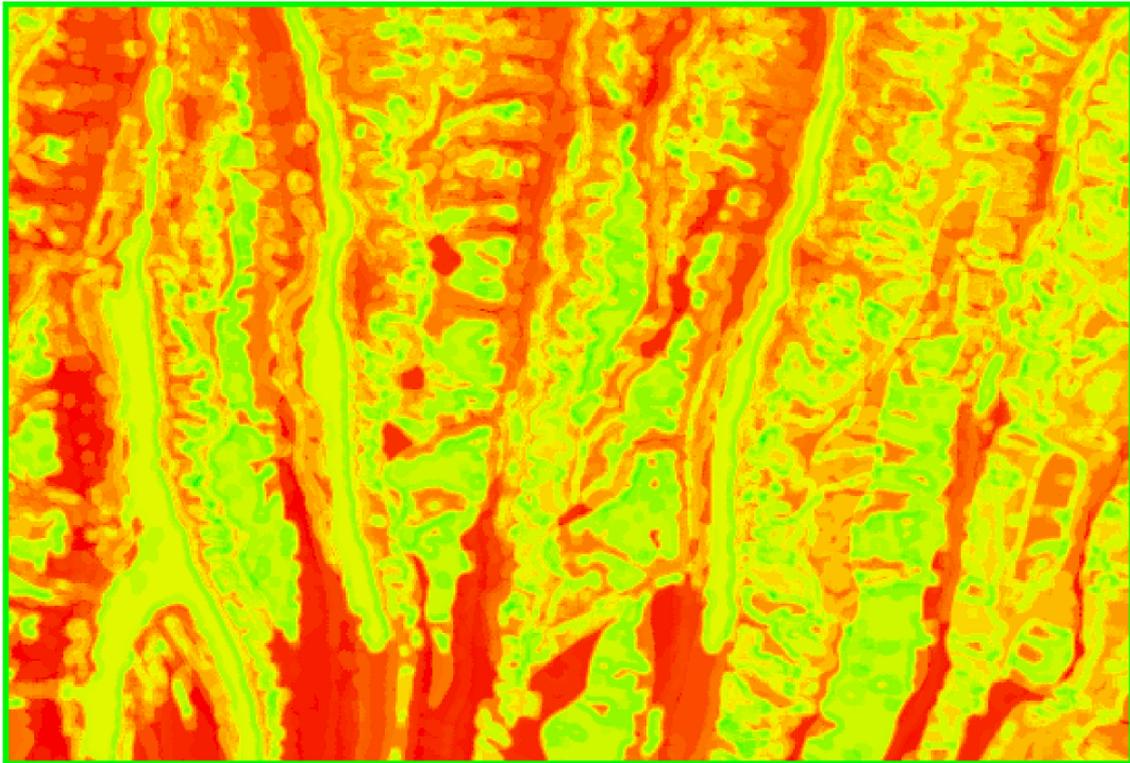


### Types de sol

	0		Rendosols		Luvisols dégradés
	Fluvisols		Calcosols		Luvisols dégradés leptiques
	Colluviosols		Brunisols		



FIGURE 3.9 – Carte des sols générée par CLAPAS avec la nouvelle image combinée



### Distances

clapas\_d

Valeur

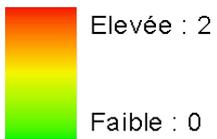


FIGURE 3.10 – Carte des distances générée par CLAPAS avec la nouvelle image combinée

## Conclusion

Ce stage fut une expérience professionnelle enrichissante qui m'a permis de mettre en application les acquis de ma formation. Améliorer un programme de recherche était un projet très motivant. Programmer et automatiser la création de la carte des variables prédictives grâce aux arbres de classification et perfectionner la forme elliptique du masque appliqué sur l'image d'entrée créée étaient les objectifs demandés et réalisés. Les résultats produits sont très satisfaisants dans la mesure où les phases d'élaboration de l'image d'entrée du modèle sont entièrement automatisées et donnent les mêmes résultats que la procédure « manuelle ».

Plusieurs pistes peuvent être envisagées pour améliorer les outils mis au point pendant mon stage.

Tout d'abord, l'algorithme CART, utilisé pour seuiller nos images, est très sensible au bruit et aux valeurs aberrantes. Les variables prises en entrée contiennent beaucoup de valeurs qui sont mal observées lors de l'exécution de CART. Il serait intéressant de travailler sur des sous-échantillons de pixels pris aléatoirement dans ces variables et de comparer les différents arbres obtenus pour voir quelles sont les variables les mieux détectées.

Par ailleurs, la taille et la forme de l'ellipse ne sont pas forcément adaptées pour détecter au mieux chaque type de sol. Il a été soumis lors d'une réunion de travail de tester pour chaque unité cartographique l'ellipse qui permettrait de les détecter correctement. Il faudrait pour réaliser ce travail ne considérer que le paysage qui nous importe et stocker les autres ensemble. On lui applique alors au fur et à mesure des formes et des tailles d'ellipse différentes pour ne retenir que la meilleure. Cette étape sera reconduite pour chaque type de sol.

Ces travaux, laissés en friche, pourront, je l'espère, être validés par une publication.

## Bibliographie

**BAIZE D., GIRARD M.C. (coord)**, Référentiel pédologique 2008, Ed. Quæ

**BAIZE D., JABIOL B.**, Guide pour la description des sols, INRA Ed. 1995

**HARTEMINK A.E., McBRATNEY A., MENDONCA-SANTOS M.L.**, Digital Soil Mapping with Limited Data, Ed. Springer 2008

**ROBBEZ-MASSON J.M.**, Reconnaissance et délimitation de motifs d'organisation spatiale, Application à la cartographie des pédopaysages, Thèse, ENSA.M-INRA 1994

**McBRATNEY A., MENDONCA SANTOS M.L, MINASNY B.**, On digital soil mapping, Geoderma 117 2003

**GRINAND C., ARROUAYS D., LAROCHE B., MARTIN M.**, Extrapolating regional soil landscapes from an existing soil map : Sampling intensity, validation procedures, and inegration of spatial context, Geoderma 143 2008

**BESSE P.**, [http://www.math.univ-toulouse.fr/~besse/pub/Appren\\_stat.pdf](http://www.math.univ-toulouse.fr/~besse/pub/Appren_stat.pdf) p.69  
[http://www.math.univ-toulouse.fr/~besse/pub/TP/appr\\_r/tp4\\_tree\\_knn\\_R.pdf](http://www.math.univ-toulouse.fr/~besse/pub/TP/appr_r/tp4_tree_knn_R.pdf)

**FARAWAY J.J.**, Extending the linear model with R : generalized linear, mixed effects and nonparametric regression models, Chapman & Hall/CRC, 2006

**GRINAND C.**, Application des techniques de cartographie numérique à la délimitation des pédopaysages en Limousin, Master 2, 2005-2006

**LEHMANN S.**, Essai de cartogénèse numérique, 2004-2005

**VITRY O.**, Programmation de CLAPAS sous R, 2008-2009