IUT de Perpignan

Département STID (Carcassonne) : 1ère Année

Modélisation de l'occupation du sol par réseau de neurones

Thomas PALMER

GEODE

Laboratoire GEODE
Université Toulouse II
(Le Mirail)

Maitre de stage : M.
Martin Paegelow

Tuteur : Mme Nathalie
Villa-Vialaneix

25/06/2010

IUT de Perpignan

Département STID (Carcassonne) : 1ère Année

Modélisation de l'occupation du sol par réseau de neurones

Thomas PALMER

Laboratoire GEODE
Université Toulouse II (Le Mirail)

Maitre de stage : M. Martin Paegelow

Tuteur : Mme Nathalie Villa-Vialaneix

25/06/2010

Remerciements

A Mme VILLA-VIALANEIX pour m'avoir présenté ce stage, puis pour son encadrement tout au long de ces quatre semaines et enfin son aide précieuse et sa disponibilité,

A Mr. PAEGELOW pour m'avoir permis d'effectuer ce stage et pour son accueil au sein du Laboratoire GEODE,

Aux membres du Laboratoire GEODE pour leur amabilité.

Sommaire

1. Introduction

1.1 Présentation synthétique du stage.

Dans le cadre de ma première année de D.U.T (Diplôme Universitaire de Technologie) S.T.I.D (Statistique et Informatique Décisionnelle), j'ai effectué un stage de quatre semaines, du 31 mai au 25 juin 2010, au sein du Laboratoire GEODE à l'Université de Toulouse II (Le Mirail). Ce stage a été proposé et encadré par Mr PAEGELOW, professeur à l'Université de Toulouse Le Mirail, mais également co-directeur du Master 2 Géomatique SIGMA, ainsi que chercheur sur la modélisation prospective de l'occupation du sol ; mais également par Mme VILLA-VIALANEIX, maître de conférences en statistique à l'Université de Perpignan (sur le site de Carcassonne) et membre de l'Institut de Mathématiques de Toulouse.

Le sujet de ce stage était donc la modélisation de l'occupation du sol grâce à un réseau de neurones. Concrètement, il va s'agir de créer un programme capable de prédire l'occupation d'une partie du territoire : une carte donc, cinq années après le dernier cliché existant de cette même carte. Pour y arriver, nous allons utiliser un modèle de calcul dont la conception est très schématiquement inspirée du fonctionnement des neurones biologiques ; d'où son nom de « réseau de neurones ». Il y a donc eu quatre étapes majeures au cours du stage. Tout d'abord, il m'a fallu comprendre le concept et la méthodologie d'utilisation appliquée au réseau de neurones. Puis il y a eu la phase « d'apprentissage » de ce dernier, afin qu'il intègre les différents paramètres à prendre en compte. Ensuite la période de tests, pour mettre en valeur la combinaison adéquate des valeurs des paramètres qui rendait la méthode optimale. Enfin, une fois cette combinaison trouvée, il ne restait plus qu'à l'appliquer au dernier cliché fourni par l'IGN (Institut Géographique National) au Laboratoire GEODE. Bien entendu, il a fallu présenter ces résultats à Mr PAEGELOW pour qu'il les compare aux programmes SIG déjà utilisés par le laboratoire.

1.2 Présentation du laboratoire.

Le laboratoire GEODE s'inscrit depuis ses origines dans une approche interdisciplinaire des relations nature/société, confrontant concepts et méthodologies issues de la géographie mais aussi de disciplines écologiques, sociales ou historiques. L'équipe, fondée en 1969 par Georges Bertrand sous le nom de CIMA (« Centre interdisciplinaire d'étude sur les milieux naturels et ruraux »), a été associée au CNRS en 1972 et est devenue UMR en 1994 sous le nom de GEODE. Elle a été pionnière dans l'élaboration et l'utilisation du géosystème et du paysage, concepts à partir desquels ont été déclinées de multiples recherches dans ce qui était alors la « géographie physique globale » et qui est devenu aujourd'hui le champ de l'environnement.

Cette approche des systèmes complexes exprimant les interactions nature/société a été très tôt enrichie par la prise en compte du temps (qu'il s'agisse de la dimension historique ou des dynamiques spatio-temporelles), démarche qui reste également une des spécificités des recherches de GEODE. Au-delà de l'analyse multi-scalaire ou de l'étude des dynamiques contemporaines, communes en géographie, dès le début des années 1980 il a été développé, dans de multiples programmes, une approche multi-temporelle intégrant la longue durée grâce à des partenariats avec archéologues et paléoécologues. L'analyse des systèmes spatio-temporels structure donc une grande partie des recherches du laboratoire, depuis l'analyse des états instantanés de l'environnement et la modélisation des dynamiques futures, jusqu'aux travaux sur la durée pluriséculaire ou plurimillénaire.

Les personnels, au 1er octobre 2009, se répartissent comme suit : 10 chercheurs CNRS (1 DR et 9 CR, dont un en détachement) et un DR émérite ; 17 enseignants chercheurs (dont 2 professeurs) et 3 professeurs émérites ; 5 personnels ITA (1 IR, 2 IE, 2 T) ; 33 doctorants. Depuis 2006, le laboratoire a intégré 4 chercheurs CNRS supplémentaires, 2 ITA et 7 enseignants-chercheurs. La pyramide des âges est équilibrée et assez jeune, avec une majorité de moins de 45 ans. GEODE est donc un laboratoire en croissance, à forte composante CNRS, mais qui souffre d'une faiblesse en nombre d'habilités à diriger les recherches : 4 personnes seulement, dont un chercheur devant partir en retraite en 2010. Des habilitations en cours devraient cependant corriger un peu ce phénomène. Le problème de l'encadrement est compensé par la pratique assez généralisée des co-directions ou co-tutorats, ainsi que par des comités de thèse.

En figure 1, on trouvera un organigramme du laboratoire.

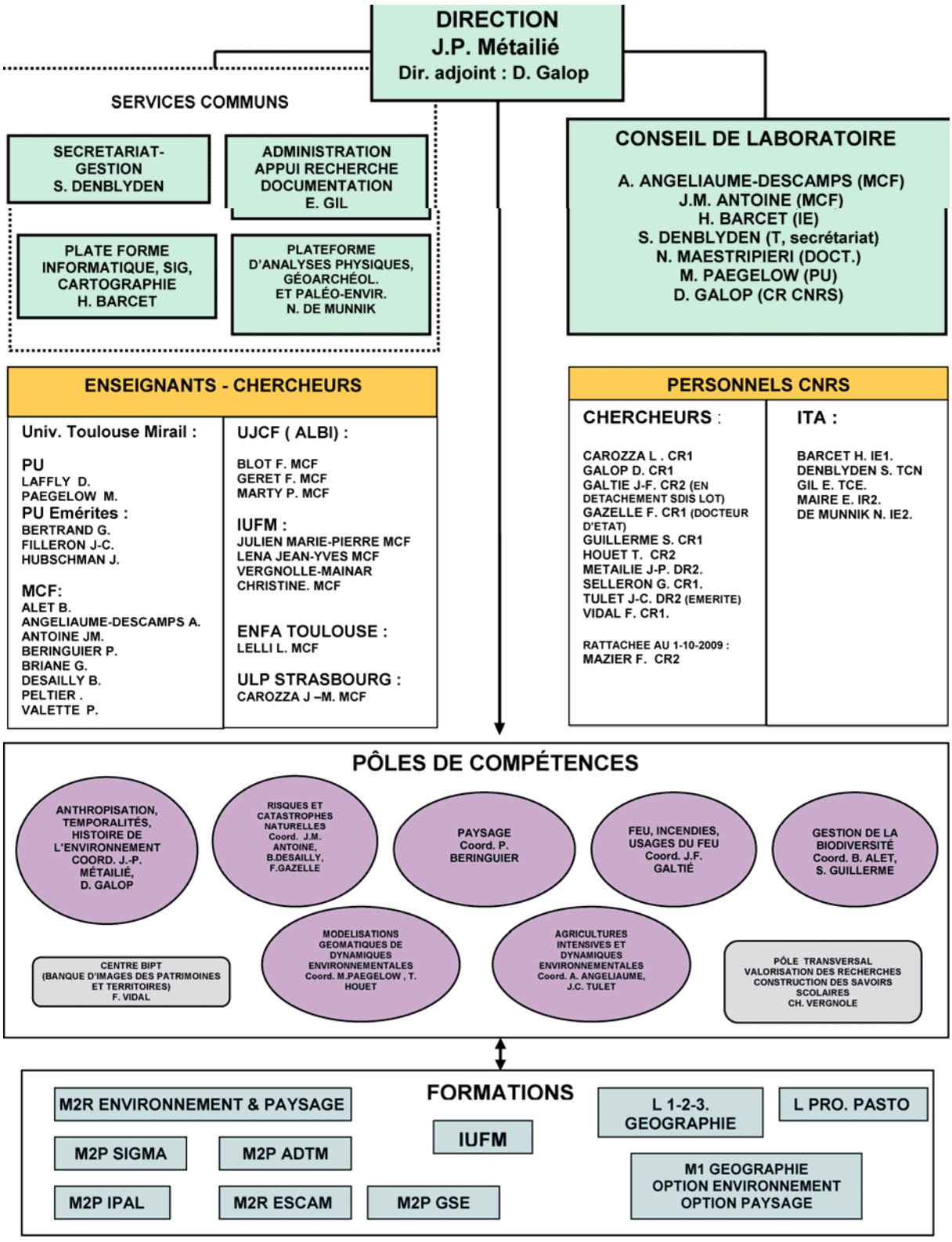


Figure 1 : Organigramme du laboratoire GEODE

2. Méthodes et outils

2.1 Description des données

La problématique de ce travail est la modélisation de l'évolution de l'occupation des sols d'une région des Pyrénées Orientales (Garrotxes). La zone géographique de l'étude a été découpée en une grille de 2005 x 2005 pixels. On connait, pour chaque pixel, l'occupation du sol à trois dates (1995, 2000 et 2004) ainsi que diverses variables explicatives. L'objectif final est de pouvoir fournir une prédiction réaliste de l'occupation du sol pour l'année 2009 à partir de la modélisation des évolutions sur les années précédentes (les trois cartes mises à notre disposition sont placées en Annexe 1).

Voici tout d'abord ces cinq variables explicatives qui influencent une autre variable, dite d'intérêt c'est-à-dire qui fait l'objet de l'étude statistique.
Il y a une seule variable qualitative : « Ecobuage » ; qui représente le type d'écobuage (débroussaillement par le feu).
Puis quatre variables quantitatives :

- « Dist-Hydro » : qui représente la distance au plus proche cours d'eau.
- « Expo » : qui représente l'exposition du terrain.
- « Mnt » : qui représente l'altitude à laquelle se trouve le terrain.
- « Pente » : qui représente la pente subit par le terrain.

Une variable explicative peut également servir à stratifier la population. Il existait également une variable supplémentaire découpant la zone en deux régions aux comportements distincts : « Région » ; mais faute de temps, elle n'a pas pu être intégrée dans mon travail.
Ces 6 variables sont représentées chacune sur une carte distincte en Annexe 2 pour l'approche concrète du terrain observé. Suite à ces cartes, nous présentons également une analyse descriptive de ces données.

Les cartes d'occupation des sols, qui sont donc des matrices de pixels, sont recodées sous forme d'une variable numérique. Concrètement, on assigne une valeur numérique de 1 à 6 pour chaque valeur que pouvait prendre un pixel, à savoir : « Bois dense», « Bois clairsemé », «Lande boisée », « Lande », « Pelouse », « Rocheux » ; respectivement devenu 1, 2, 3, 4, 5, 6. Il convient de préciser que ce que nous pourrions appeler « sol urbain » est classé dans la catégorie « Rocheux » aux yeux des géographes.

Voici une brève analyse descriptive de l'occupation des sols pour les trois cartes qui sont en notre possession :

Pourcentage de l'occupation du sol par année :

Carte 95 :

Bois dense	Bois clairsemé	Lande boisée	Lande	Pelouse	Rocheux
2,89%	5,82%	9,14%	26,67%	43,79%	11,69%

Carte 00 :

Bois dense	Bois clairsemé	Lande boisée	Lande	Pelouse	Rocheux
5,24%	5,19%	2,72%	34,46%	45,80%	6,59%

Carte 04 :

Bois dense	Bois clairsemé	Lande boisée	Lande	Pelouse	Rocheux
5,49%	9,09%	3,50%	14,77%	40,41%	26,73%

Les principaux changements sont une diminution globale de la lande boisée, une légère augmentation de la lande entre 1995 et 2000 suivie d'une forte diminution de celle-ci entre 2000 et 2004 et une légère diminution des parties rocheuses entre 1995 et 2000 suivie par une forte augmentation de celles-ci entre 2000 et 2004.

Enfin, voici les transitions entre les différentes dates :

95 / 00	Bois dense	Bois clairsemé	Lande boisée	Lande	Pelouse	Rocheux
Bois dense	49,72%	29,97%	0,49%	15,73%	3,69%	0,41%
Bois clairsemé	31,70%	27,84%	0,74%	31,46%	7,37%	0,90%
Lande boisée	9,73%	10,56%	16,36%	56,48%	6,35%	0,52%
Lande	3,41%	4,41%	2,88%	60,71%	27,24%	1,36%
Pelouse	0,35%	1,22%	0,89%	23,63%	70,13%	3,79%
Rocheux	0,08%	0,26%	0,05%	4,07%	57,42%	38,13%

00 / 04	Bois dense	Bois clairsemé	Lande boisée	Lande	Pelouse	Rocheux
Bois dense	61,18%	22,64%	0,53%	12,94%	0,80%	1,91%
Bois clairsemé	21,09%	49,92%	0,97%	17,47%	5,03%	5,52%
Lande boisée	2,22%	24,43%	33,12%	24,38%	13,11%	2,74%
Lande	2,99%	10,57%	6,85%	29,12%	39,75%	10,72%
Pelouse	0,20%	2,00%	0,34%	4,99%	53,46%	39,01%
Rocheux	0,19%	1,48%	0,09%	3,12%	23,71%	71,40%

Les tableaux obtenus donnent, pour chaque ligne, les pourcentages de transition d'une catégorie vers une autre entre les deux dates (pourcentages calculés par rapport au nombre de pixels appartenant au départ à la catégorie considérée dans la ligne). Les principaux changements observés sont :

- une transition du bois dense vers le bois clairsemé de l'ordre de 30% des pixels de bois dense entre 1995 et 2000 et de l'ordre de 23% des pixels de bois dense entre 2000 et 2004 ;
- une transition du bois clairsemé vers le bois dense de l'ordre de 32% des pixels de bois clairsemé entre 1995 et 2000 et de l'ordre de 21% des pixels de bois clairsemé entre 2000 et 2004 ;
- une transition de la lande boisée vers le bois clairsemé et la lande de l'ordre de, respectivement, 10% et 56% des pixels de lande boisée entre 1995 et 200 et de l'ordre de, respectivement, 24% et 24% des pixels de lande boisée entre 2000 et 2004 ;
- une transition de la lande vers la pelouse de l'ordre de 27% des pixels de pelouse entre 1995 et 2000 et de l'ordre de 40% des pixels de lande entre 2000 et 2004 ;
- une transition de la pelouse vers la lande de l'ordre de 24% des pixels de pelouse entre 1995 et 2000 et vers la roche de l'ordre de 40% des pixels de pelouse entre 2000 et 2004 ;
- une transition de la roche vers la pelouse de l'ordre de 57% des pixels de roche entre 1995 et 2000 et de l'ordre de 24% des pixels de roche entre 2000 et 2004.

2.2 Utilisation des réseaux de neurones

Le but de l'étude est de comprendre comment, à partir des valeurs de l'occupation du sol à une date T (où T = 1995, 2000 ou 2004) et de diverses variables explicatives annexes à cette même date, on peut prédire l'occupation du sol à la date T +1 (2000, 2004 ou 2009 qui n'est pas encore connue). Un modèle est donc défini par :

- des entrées qui sont l'occupation du sol à la date T et les variables explicatives ;
- des sorties qui sont ce que l'on cherche à prédire, à savoir l'occupation du sol à la date T +1 ;
- un « mécanisme » permettant de relier les entrées aux sorties.

Le « mécanisme » est défini de la manière suivante :

- on apprend le mécanisme à partir des entrées de 1995 et des sorties de 2000 ;
- on teste le mécanisme (sa fiabilité) à partir des entrées de 2000 et des sorties de 2004 ;
- on utilise le mécanisme à partir des entrées de 2004 pour construire un scénario réaliste pour 2009.

Les réseaux de neurones sont composés d'éléments simples (ou neurones) fonctionnant en parallèle. Ces éléments ont été fortement inspirés par le système nerveux biologique. Comme dans la nature, le fonctionnement du réseau (de neurones) est fortement influencé par la connection des éléments entre eux. On peut entraîner un réseau de neurones pour une tâche spécifique (reconnaissance de caractères par exemple) en ajustant les valeurs des connections (ou poids) entre les éléments (neurones).

En général, l'apprentissage des réseaux de neurones est effectué de sorte que pour une entrée particulière présentée au réseau corresponde une cible spécifique. L'ajustement des poids se fait par comparaison entre la réponse du réseau (ou sortie) et la cible, jusqu'à ce que la sortie corresponde à la cible. On utilise pour ce type d'apprentissage dit supervisé un nombre conséquent de paires entrée/sortie.

Nous allons utiliser dans notre cas le type de réseau « batch » ou « global » et plus particulièrement un réseau multicouche : Ils sont organisés en couches, chaque neurone prend généralement en entrée tous les neurones de la couche inférieure. Ils ne possèdent pas de cycles ni de connexions intra-classe. On définit alors une « couche d'entrée », une « couche de sortie », et n « couches cachées ». Ce type particulier de réseau de neurones est également appelé : « perceptron multi-couches ».

On peut le représenté schématiquement comme dans la Figure 2 :

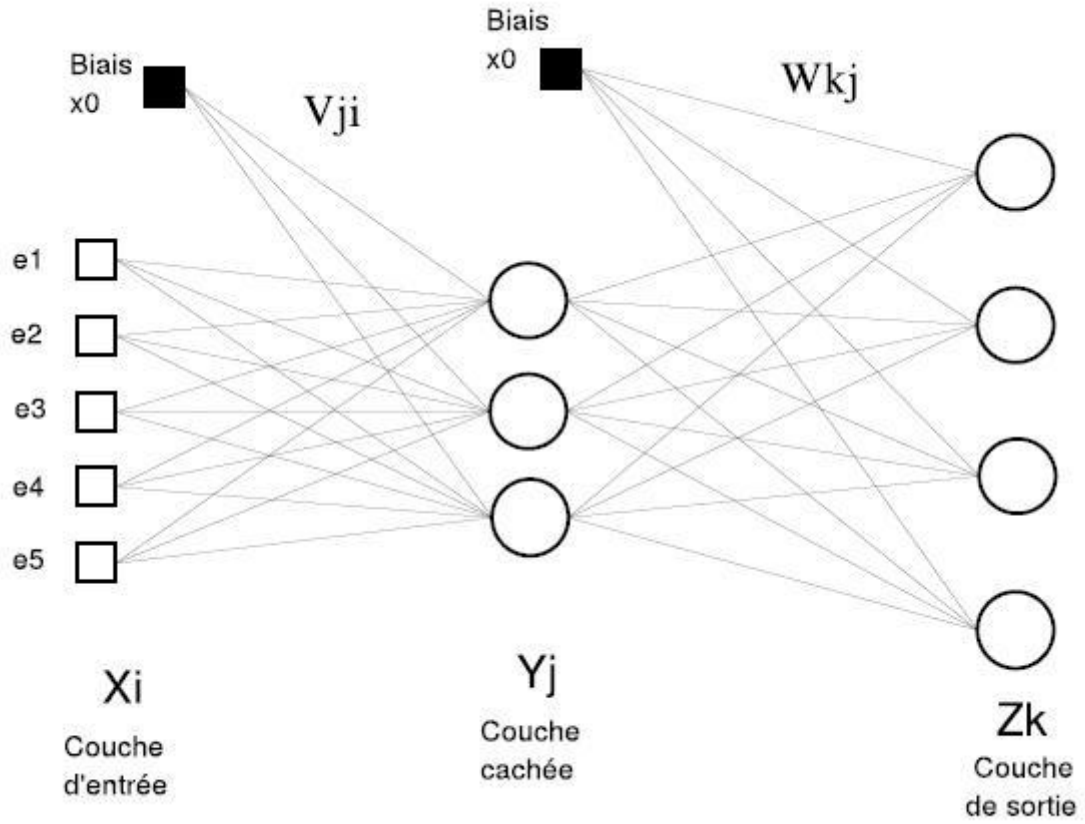


Figure 2 : Perceptron multi-couches.

On voit sur ce schéma plus clairement les « passages » entre les couches du perceptron. Dans notre cas, l'apprentissage consiste à trouver les valeurs optimales des « poids » (fixés au hasard au début de l'algorithme de recherche) qui pondèrent le système et ainsi passer d'une couche à l'autre.

Voici une formule mathématique qui concrétise les sorties des différentes couches. Tout d'abord lorsqu'on propage à la couche intermédiaire dite cachée, avec x_i correspondant aux données en entrée (e_i sur le schéma):

$$y_j = f(\sum_{i=1}^{m} x_i v_{ij} + x_0)$$

Puis de la couche cachée vers la couche de sortie :

$$z_k = f(\sum_{j=1}^{n} y_j w_{kj} + y_0)$$

Les valeurs x_0 et y_0 sont des biais : des scalaires et non des sorties de la couche précédente.

Les réseaux de neurones ont une histoire relativement jeune (environ 50 ans) et les applications intéressantes des réseaux de neurones n'ont vu le jour qu'il y a une vingtaine d'années (développement de l'informatique).

La mise en place de ce « mécanisme » va donc constituer la majeure partie de ce stage. C'est ce que nous allons voir maintenant en détaillant les trois étapes majeures : apprentissage, test et application ; avant de montrer les résultats.

2.3 Méthodologie utilisée

La première partie de ce travail à donc été l'« apprentissage », puis le « test ». Ces deux parties sont similaires, non dans la réalisation mais dans l'approche.

Comme dit précédemment, cela consiste à apprendre le mécanisme au réseau de neurones à partir des entrées de 1995 et des sorties de 2000. Néanmoins, l'intégralité des observations n'a pas été utilisée pour l'apprentissage (ainsi que pour le test) mais seule une sélection aléatoire des entrées et des sorties. Nous avons donc utilisé deux méthodes de sélection, puis avons comparé les résultats :

- de manière proportionnelle : on sélectionne des données pour chaque type de catégorie d'occupation des sols et selon que la catégorie évolue ou non entre 1995 et 2000, proportionnellement à l'effectif observé.
- de manière équilibrée : on sélectionne des données pour chaque type de catégorie d'occupation des sols et selon que la catégorie évolue ou non entre 1995 et 2000 de manière équilibrée (lorsque les catégories ont des effectifs très déséquilibrés, le rééquilibrage permet parfois de prédire les événements rares).

Pour la carte entière, une sélection de 1% des données a été effectuée sur une base proportionnelle et équilibrée.

Pour effectuer cette approche de la relation entre l'occupation du sol à la date T et celle à T+1, nous allons utiliser le perceptron multi-couches, présenté plus haut, grâce au package « *nnet* » du logiciel R. Le résultat de la simulation dépend de plusieurs paramètres :

- la taille du voisinage prise en compte pour la collecte des voisins (valeurs possibles : 1, 2 ou 3) ;
- la valeur de prise en compte pour le calcul de l'influence des voisins (valeurs possibles : 0,1 ; 0,2 ; 0,5 ; 1 ; 2 ; 5) ;
- le nombre de neurones sur la couche cachée du perceptron (valeurs à tester : 5, 10, 15, 20) ;
- la valeur du « decay » qui est une pénalisation destinée à rendre le perceptron plus robuste (valeurs à tester : 0 ; 0,1 ; 1 ; 5 ; 10).

De plus, pour chaque combinaison de paramètres et à cause des problèmes de minima locaux rencontrés lors du schéma de minimisation du perceptron, le perceptron sera appris 5 fois.

Enfin, on compare les résultats du perceptron aux données que nous avons de 2000 ; et grâce a un programme (voir partie du code en Annexe 3), nous pouvons récolter la différence entre les prédictions et le modèle réel (le meilleur des cinq essais bien entendu).

Une fois l'intégralité des simulations basées sur chacune des sélections (proportionnelle et équilibrée) effectuée, il faudra calculer les taux d'erreur sur la carte globale avec le meilleur modèle. Le « meilleur modèle » sera celui pour lequel l'erreur de test est la plus faible sur l'ensemble des combinaisons de paramètres possibles.

Pour finir, la carte de prévisions peut être générée pour chacune des deux méthodes. Nous allons donc à présent donner et commenter les résultats obtenus lors de ces différents stades du travail : l'apprentissage, le test et bien entendu la carte finale.

2.4 Logiciels et programmes

Tout d'abord, il convient de noter que ces calculs (gigantesques puisqu'il s'agit rappelons-le de matrices de 2005 colonnes sur 2005 lignes) ont tous été effectués sur un serveur distant hebergé à l'IUT de Toulouse Blagnac. Il s'agit d'un serveur Unix distant avec lequel on peut communiquer par SSH. Afin de se connecter à lui, on utilise le programme « PuTTY ».

Ensuite, pour plus de facilité sur un environnement de travail de type Windows, un programme de drag&drop (glisser-déposer) nommé « WinSCP » est utilisé. On a ainsi un accès simplifié aux répertoires locaux de Windows en même temps qu'aux répertoires du serveur Unix préalablement lancé ; le but étant de ne pas être obligé d'utiliser des programmes de l'environnement Linux sans interface graphique, pas toujours aisés à manipuler. Concrètement, on modifie une partie du code du coté Windows, puis grâce à WinSCP on le transfert directement sur le serveur distant Unix.

Enfin, tous ces calculs précédemment présentés sont effectués sur le logiciel R. Il s'agit du plus réputé et du plus complet des programmes liés à la statistique du monde libre. Il permet, en plus de très nombreux traitements statistiques, de créer des graphiques justes (contrairement à certains logiciels de Microsoft), personnalisables et surtout en très grand nombre. Ce logiciel a été choisit notamment grâce à sa capacité à traiter de nombreuses données, très intéressant compte tenu des matrices que nous avons eu à traiter ; ainsi que pour sa compatibilité avec le serveur Unix à notre disposition pour traiter ces calculs.

Nous allons maintenant présenter quelques-uns des programmes informatiques utilisés lors de ce stage. Ils ont été développés par Mme VILLA, puis nous les avons repris, modifiés et adaptés pour chacun des cas qui nous intéressaient. Ces différents programmes, d'extension .R pour être interprétés par le logiciel R, sont en Annexes 3.

Tout d'abord, il y eu les programmes (chaque programme est présenté au pluriel puisque ils sont tous modifiés et dupliqués au moins trois fois pour les années 1995, 2000 et 2004) qui ont servis à séparer la carte en plusieurs « parties » égales, afin que les traitements ne saturent pas la mémoire du serveur : *SplitCartes.R* ; puis ceux pour la recherche des « Voisins », c'est-à-dire la nature de l'occupation du sol des pixels voisins à celui traité : *Size2-Split95.R* (ou « 2 » est le paramètre de taille donc 1, 2 ou 3 et « 95 » l'année donc 95,00 et 04) ; et enfin le recollage qui réassemble les données de ces cartes : *DeSplitCartes95-01-2.R* (avec « 95 » pour l'année, « 01 » pour la valeur de gamma et « 2 » pour la taille).

Ensuite, vient le groupe de programmes les plus importants, c'est-à-dire au sujet de l'apprentissage qui sont du type : *App1_g01_n5_d0.R* (avec « 1 » pour la taille du voisinage donc 1,2 ou 3 ; « 01 » pour la valeur de gamma donc 01, 02, 05, 1, 2 ou 5 ; « 5 » pour le nombre de neurones présents dans le réseau donc 5, 10, 15 ou 20 ; et « 0 » pour le decay donc 0, 01, 1, 5 ou 10). La modification de ces programmes a été fastidieuse et a nécessité une grande concentration afin de ne pas oublier, ou du moins de mal taper, un des paramètres ; sans quoi cela aurait faussé l'ensemble des résultats. Ces programmes ont du être modifiés dans leur intégralité une deuxième fois lors de l'utilisation de la méthode dite « équilibrée ».

Enfin, pour conclure sur le sujet des programmes informatiques, les derniers ont été ceux relatifs à la phase de test : *TestProp.R* et *TestEqui.R* (respectivement pour la méthode proportionnelle puis équilibrée). Ils ont généré les tableaux d'analyse que nous verrons plus loin, ainsi que, dans un deuxième temps, la carte de prédiction de 2004 liée à chacune des méthodes.

3. Résultats

3.1 Description des résultats pour un échantillonnage proportionnel

Avant de donner les résultats de la prédiction pour 2004, il convient de donner ceux, intermédiaires, de l'apprentissage. Nous allons donc commencer la présentation des résultats avec la méthode d'échantillonnage proportionnelle, afin de suivre la chronologie du stage mais également la suite logique d'approche menée durant le stage et dans ce rapport.

Nous ne pouvons présenter l'ensemble du tableau de résultats, trop long et ne présentant que peu d'intérêt. Rappelons que ce tableau contient deux résultats : « Erreur d'apprentissage » et « Erreur de calibration », pour chaque tuple de paramètres (size, gamma, nombre de neurones et decay). Ces deux résultats sont exprimés en pourcentage ; et pour obtenir le résultat intéressant, soit le pourcentage de réussite du modèle, il faut donc effectuer le complément à 100 (voir le résultat optimal en guise d'exemple). On peut noter cependant trois caractéristiques intéressantes au niveau de l'erreur de calibration (l'erreur d'apprentissage en tant que telle étant seulement une étape menant à ce résultat) :

- Le minimum : 54.58%
- Le maximum : 58.21%
- La moyenne : 56.93%

Pour obtenir le résultat optimal, on conserve donc les paramètres associés au taux d'erreur minimum. Le taux de réussite de prévision est donc de 55,42% ; pour les paramètres suivant : size = 1 ; gamma = 1 ; nombre de neurones = 5 ; decay = 5.

Après avoir entré ces paramètres optimaux pour la méthode proportionnelle dans le programme de test, nous obtenons plusieurs tableaux statistiques. Voici donc un tableau présentant les résultats de calibration pour l'année 2004 pour notre programme, avec un pourcentage global de bonne prédiction de 45,58% :

		2004 - réalité						
		bois dense	bois clair	lande boisée	lande	pelouse	pelouse discontinue, piste	Total simul.
2004 sim. Sous régions	bois dense	**3,86**	2,27	0,03	1,41	0,10	0,27	7,94
	bois clair	0,17	**0,30**	0,00	0,08	0,02	0,02	0,58
	lande boisée	0,00	0,00	**0,00**	0,00	3,00	0,00	0,00
	lande	1,38	5,55	3,28	**10,60**	11,46	3,80	36,06
	pelouse	0,08	0,93	0,19	2,56	**28,09**	19,90	51,74
	pelouse discontinue, piste	0,00	0,06	0,00	0,13	0,75	**2,73**	3,67
	Total réalité	5,49	9,09	3,50	14,77	40,41	26,73	100,00

Les nombres en gras représentent l'occupation du sol correctement prédite et qui n'a pas changée entre les années 2000 et 2004. La colonne « Total simul. » représente les totaux par catégorie d'occupation correctement prédite dans le même intervalle. Cette colonne doit être comparée à celle intitulée « Total réalité » et non pas à 100%. Plus les deux nombres de ces colonnes sont proches et plus le modèle est fiable. Ainsi, on voit que les totaux sont très proches pour le type d'occupation « bois dense » et « pelouse ». En revanche, ils sont assez éloignés pour les autres, notamment « pelouse discontinue, piste » (équivalent de « rocheux »), et « lande boisée » n'a pas été du tout prédite (total à 0%).

On en conclue que ce modèle est relativement peu fiable, puisqu'il prend peu en compte les changements, notamment lorsqu'on le compare aux programmes existants au laboratoire GEODE tels que Dinamica ou encore le modèle de Markov.

Dans ce tableau est proposé un complément d'analyses propres aux géographes :

	Gain	Loss	Total change	Swap	Absolute value of net change	% change net
Bois dense	3,05	0,36	3,41	0,72	2,69	78,89
Bois clairsemé	0,19	4,79	4,98	0,38	4,60	92,37
Lande boisée	0,00	2,72	2,72	0,00	2,72	100,00
Lande	6,17	0,50	6,67	1,00	5,67	85,01
Pelouse	3,04	1,16	4,20	2,32	1,88	44,76
Pelouse discontinue+ rocheux/piste	0,02	2,94	2,96	0,04	2,92	98,65
TOTAL	**12,47**	**12,47**		4,46	20,48	
				2,23	**10,24**	

On peut donc voir dans la première, respectivement la deuxième, colonne pour chaque catégorie d'occupation du sol le taux de gain, respectivement de perte, de surface de cette catégorie. Puis en troisième position le cumul de ces deux changements. Le « Swap » est la valeur totale du changement sur l'ensemble moins la valeur absolue de changement, évoquée dans la colonne suivante. La colonne « Absolute change » est relative au changement net d'occupation, c'est-à-dire le taux de gain moins le taux de perte. Cette dernière tiens compte de la nature de l'occupation du sol concerné par ces pixels gagnés ou perdus, contrairement au Swap. Enfin le « Pourcentage change net » est le rapport de cette dernière colonne évoquée sur le total de changement global. Pour les géographes, les deux valeurs primordiales à la comparaison de plusieurs modèles de prédiction de cartes est le total du swap et de la valeur absolue nette de changement, écrite ici en gras sur la dernière ligne : respectivement 2,23% et 10,24%.

Enfin, la Figure 3 présente la carte obtenue avec ces paramètres, et pour la méthode d'échantillonnage proportionnelle, pour l'année 2004 (à comparer avec la Figure 8 de l'annexe représentant la carte réelle de la même année) :

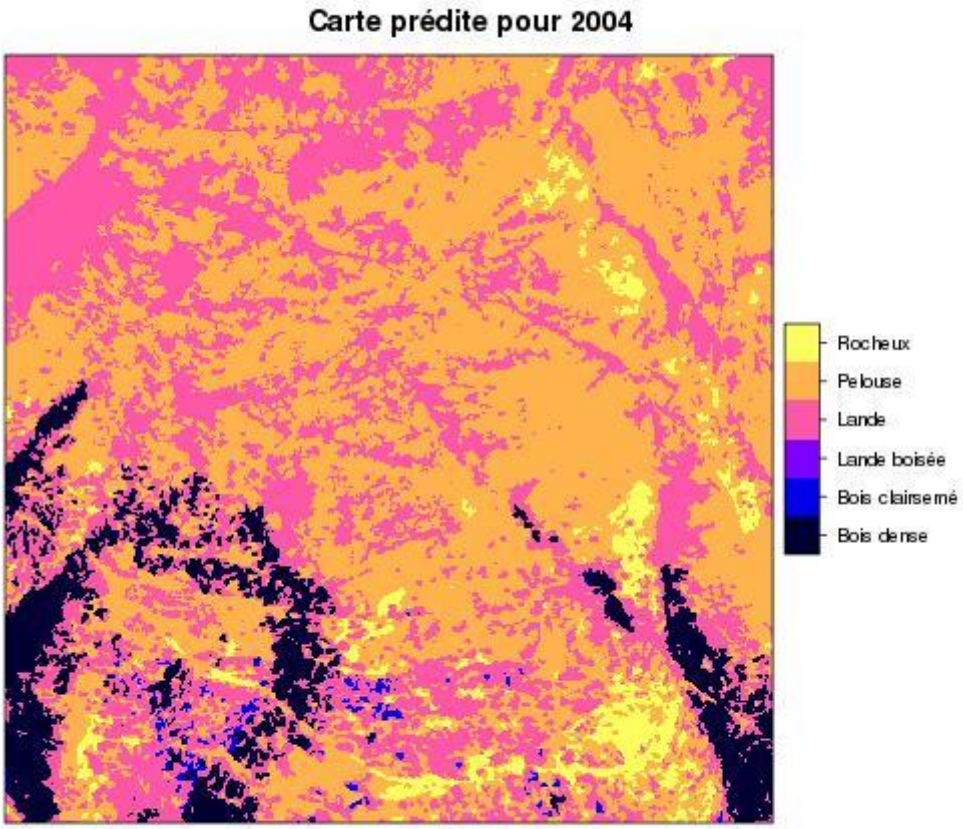


Figure 3 : Carte prédite pour 2004 avec la méthode proportionnelle.

3.2 Description des résultats pour un échantillonnage équilibré

Comme précédemment, on donne les trois caractéristiques intéressantes au niveau de l'erreur de calibration (l'erreur d'apprentissage en tant que telle étant toujours seulement une étape menant à ce résultat) :

- Le minimum : 50.70%
- Le maximum : 54.45%
- La moyenne : 52.84%

Pour obtenir le résultat optimal, on conserve donc les paramètres associés au taux d'erreur minimum. Le taux de réussite de prévision est donc de 50,70% ; pour les paramètres suivant : size = 1 ; gamma = 2 ; nombre de neurones = 5 ; decay = 10. On note immédiatement que ces résultats sont d'ores et déjà meilleurs que les précédents.

Après avoir entré les paramètres optimaux pour la méthode équilibrée dans le programme de test, nous obtenons plusieurs tableaux statistiques. Voici donc un tableau présentant les résultats de calibration pour l'année 2004 pour notre programme, avec un pourcentage global de bonne prédiction de 45,10% :

		2004 - réalité						
		bois dense	**bois clair**	**lande boisée**	**lande**	**pelouse**	**pelouse discontinue, piste**	Total simul.
2004 sim. Sous régions	**bois dense**	**3,27**	1,61	0,01	1,52	0,06	0,43	6,90
	bois clair	1,12	**2,49**	0,05	0,67	0,26	0,25	4,84
	lande boisée	0,06	0,65	**0,88**	0,61	0,33	0,07	2,61
	lande	0,95	3,24	2,39	**9,63**	13,13	6,45	35,79
	pelouse	0,08	0,98	0,16	2,14	**25,25**	15,94	44,55
	pelouse discontinue, piste	0,01	0,13	0,01	0,20	1,38	3,58	5,31
	Total réalité	5,49	9,10	3,50	14,77	40,41	26,72	100,00

Comme dans le paragraphe précédent, ce tableau s'interprète de la même façon, puisque seule la méthode a changée. Ainsi, on voit que les totaux sont proches pour quasiment tous les types d'occupation à part « lande » et « pelouse discontinue, piste » (équivalent de « rocheux »), qui sont assez éloignés.

On en conclue que ce modèle est plus fiable que le précédent, puisqu'il prend mieux en compte les changements. Il reste très performant quand on le compare aux programmes existants au laboratoire GEODE : Dinamica et le modèle de Markov.

Dans ce tableau est proposé un complément d'analyses propres aux géographes, construit comme celui de la partie 3.1 :

			Gain	Loss	Total change	Swap	Absolute value of net change	% change net
Gain								
2,27	Bois dense		2,27	0,60	2,87	1,20	1,67	58,19
0,63	Bois clairsemé		0,63	0,98	1,61	1,26	0,35	21,74
0,00	Lande boisée		0,00	0,11	0,11	0,00	0,11	100,00
6,07	Lande		6,07	4,75	10,82	9,50	1,32	12,20
5,03	Pelouse		5,03	6,28	11,31	10,06	1,25	11,05
0,37	Pelouse discontinue+ rocheux/piste		0,37	1,65	2,02	0,74	1,28	63,37
		TOTAL	**14,37**	**14,37**		22,76	5,98	
						11,38	**2,99**	

On peut donc voir que les totaux de gain et perte sont plus importants qu'avec la méthode d'échantillonnage proportionnelle, mais relativement peu au-dessus. Ensuite, on remarque aisément que les deux valeurs primordiales : le Swap et la Valeur absolue nette de changement, ont des valeurs interchangées par rapport au modèle précédent à quelques unités près : 11,38% et 2.99%.

On en conclut que cette méthode est globalement meilleure, mais surtout pour une prédiction plus détaillée et non globale.

Enfin, la Figure 4 donne la carte obtenue avec ces paramètres, et pour la méthode d'échantillonnage équilibrée, pour l'année 2004 (à comparer avec la Figure 8, comme précédemment, de l'annexe représentant la carte réelle de la même année) :

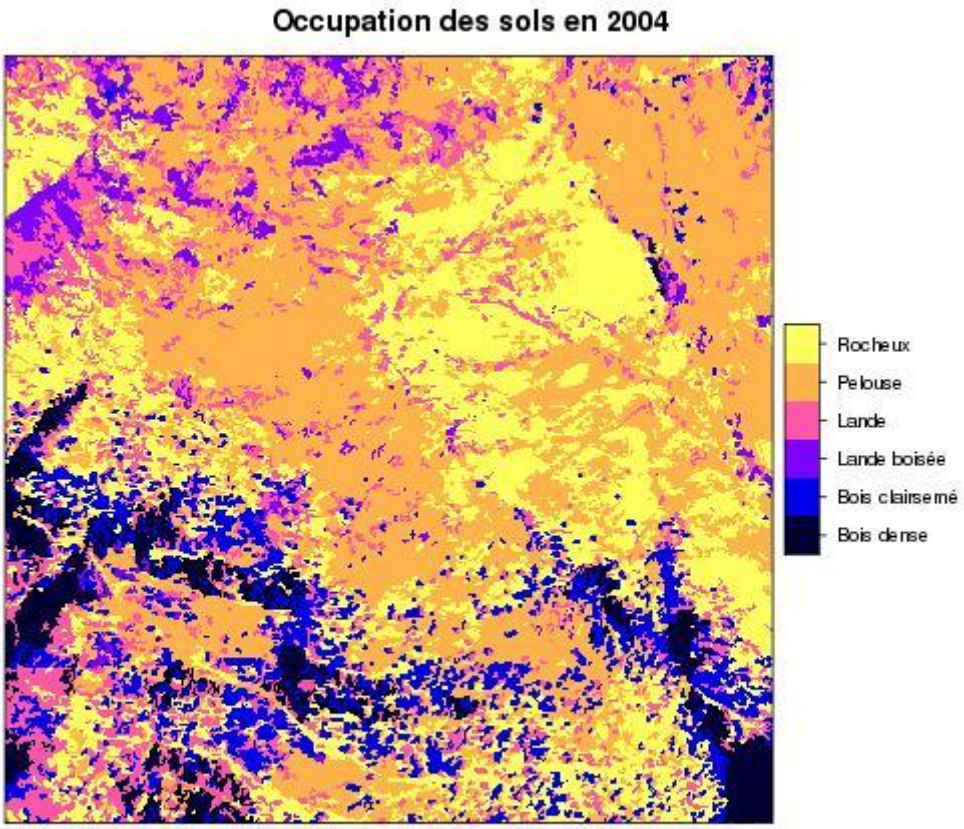


Figure 4 : Carte prédite pour 2004 avec la méthode équilibrée.

On remarque que cette carte est beaucoup plus proche du résultat réel (placé en Annexe 1) que nous étions sensés trouver. Ceci confirme bien que localement cette méthode est meilleure, même si globalement elles sont assez proches.

4. Conclusion

Finalement, après la phase d'apprentissage et de calibration, puis la phase de test, et selon les deux méthodes d'échantillonnage, il ne restait plus qu'à appliquer notre modèle aux données de 2004 afin de prédire la carte de 2009 : but ultime de ce stage. La Figure 5 présente donc cette carte finale que le laboratoire GEODE pourra comparer à la carte que va leur envoyer l'IGN au cours de la fin de l'année 2010 :

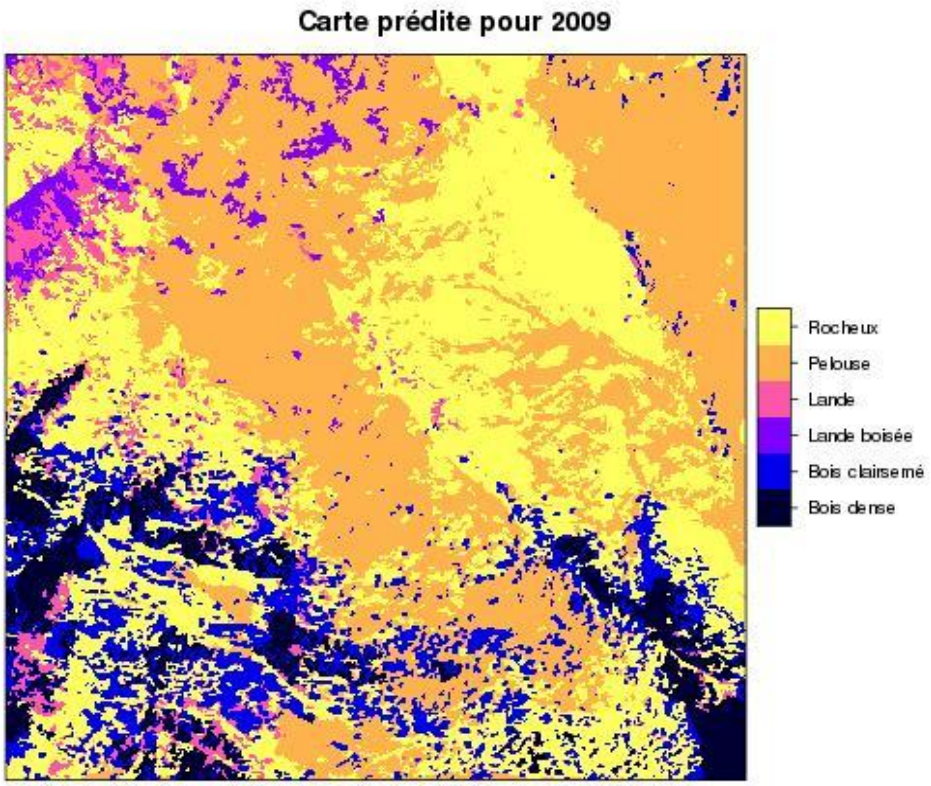


Figure 5 : Carte prédite pour 2009 avec la méthode proportionnelle.

Pour conclure, ce stage m'a fait découvrir, dans un premier temps, des fonctions du logiciel R que je ne soupçonnais pas, comme la création des cartes, présentées dans ce rapport. De plus, j'ai eu l'opportunité d'utiliser un réseau de neurones, concept assez complexe de prime abord mais facilement utilisable par la suite. Enfin, la combinaison « utilisation des statistiques » et « création (modification) de programmes informatiques », propre à la formation STID d'ailleurs, a été très enrichissante.

D'un point de vue personnel, ce stage m'a fait découvrir un environnement de travail particulier : un laboratoire de recherche, très différent de mes expériences passées mais très agréable ; ainsi que le domaine de la Géomatique et de ses nombreuses applications.

Annexes

1) Voici les trois cartes que l'ont pourrait qualifier d'« initiales », prise par l'IGN puis envoyée au GEODE et utilisées comme base du travail du stage :

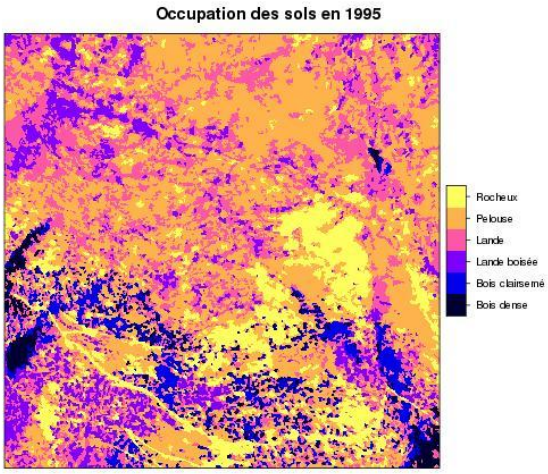


Figure 6 : Carte initiale réelle de 1995.

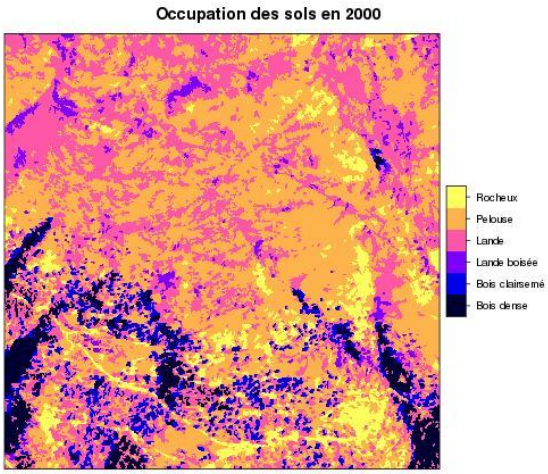


Figure 7 : Carte initiale réelle de 2000.

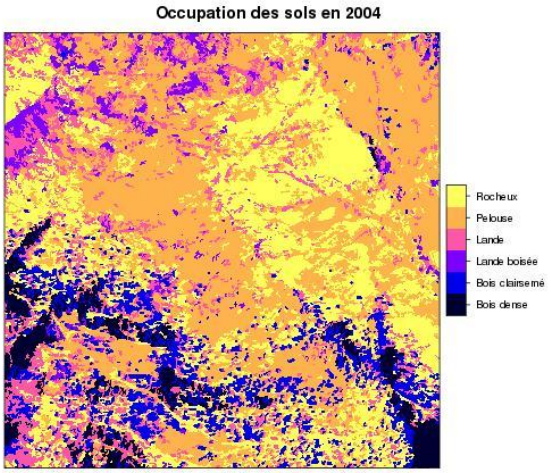


Figure 8 : Carte initiale réelle de 2004.

2) Voici les représentations des quatre variables explicatives quantitatives présentées en 2.1. Elles font parties, comme les représentations suivantes de ce deuxième paragraphe, de l'analyse des données faite en amont du stage par Mme VILLA.

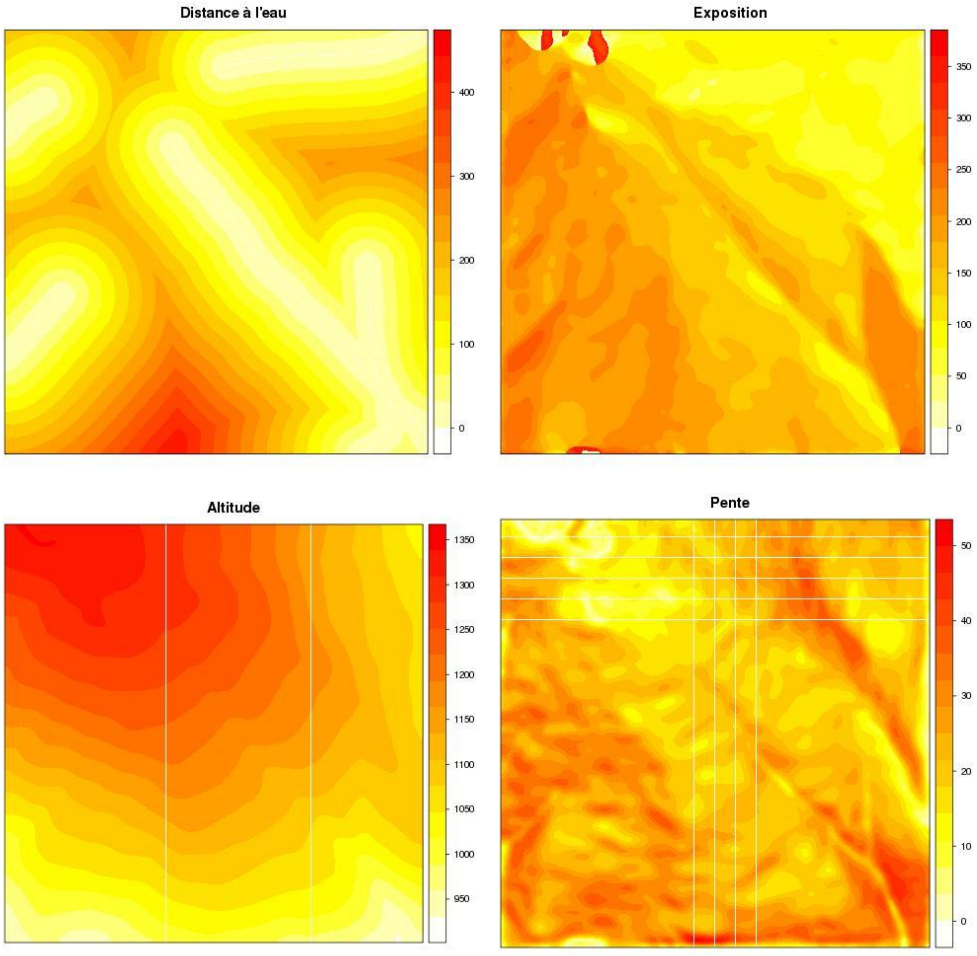
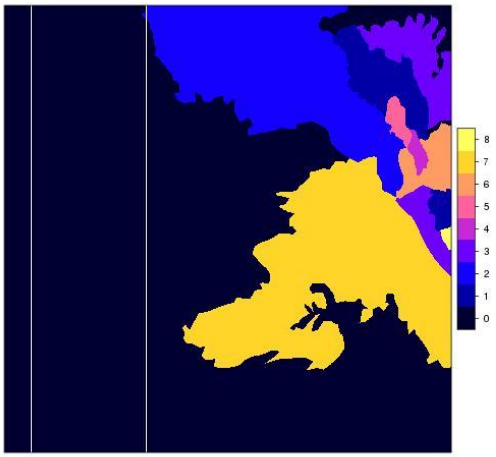
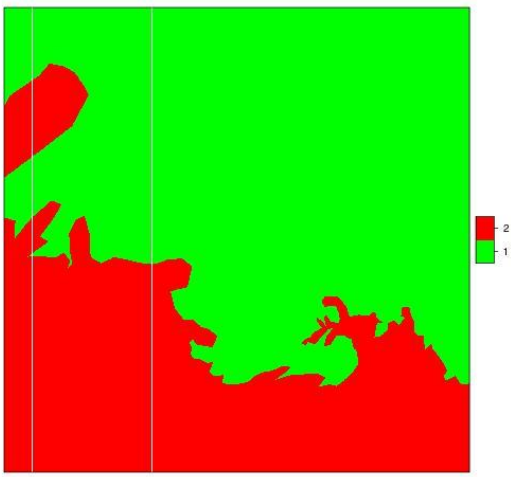


Figure 6 : Cartes des variables explicatives quantitatives.

Puis celle de la variable explicative qualitative :



Et enfin le découpage de la zone en deux régions aux comportements différents :



3) Voici le code des principaux programmes utilisés et présentés dans la partie 2.4 de ce rapport. Les paramètres sont inscrits à titre d'exemple. La méthode utilisée dans ces exemples est l'échantillonnage équilibré.

- Tout d'abord le code de l'apprentissage du réseau de neurones :

```
## Apprentissage
# Préparation des entrées et sorties
load("IndexTrainTest/SplitEquilibre-Total.RData")
load("DonneesR/CartesDisj.RData")
x.app <- carte95.disj[index.train,]
y.app <- carte00.disj[index.train,]
rm(carte95.disj,carte00.disj,carte04.disj)
```

```
load("DonneesR/Neighbors95-0.1-1.RData")
x.app <- cbind(x.app,neighbors[index.train,])
rm(neighbors)

load("DonneesR/Predictors.RData")
x.app <- cbind(x.app,predictors95[index.train,])
rm(predictors95,predictors00)

# Apprentissage (5 fois)
mlp <- list()
for (repet in 1:5)
        {
        print(paste("Train",repet,"/5"))
        mlp[[repet]] <- nnet(x.app,y.app,size=15,decay=1,maxit=1000,softmax=T,trace=F)
        }

# Prédictions
predict <- list()
error.train <- vector(length=5)
for (repet in 1:5)
        {
        predict[[repet]] <- apply(mlp[[repet]]$fitted,1,which.max)
        error.train[repet] <- sum(predict[[repet]]!=apply(y.app,1,which.max))/length(predict[[repet]])
        }

# Sauvegarde
save(mlp,predict,error.train,file="ResultatsApp/Res-Equi-0.1-1-15-1.Rdata")
```

- Ensuite celui du calcul de l'erreur de calibrage :

```
# Préparation des données
rm(x.app,y.app)
load("IndexTrainTest/SplitEquilibre-Total.RData")
load("DonneesR/CartesDisj.RData")
x.test <- carte00.disj[index.test,]
y.test <- carte04.disj[index.test,]
rm(carte95.disj,carte00.disj,carte04.disj)

load("DonneesR/Neighbors00-0.1-1.RData")
x.test <- cbind(x.test,neighbors[index.test,])
rm(neighbors)

load("DonneesR/Predictors.RData")
x.test <- cbind(x.test,predictors00[index.test,])
rm(predictors95,predictors00)

# Calcul des prédictions et erreur
predict.test <- list()
error.test <- vector(length=5)
for (repet in 1:5)
        {
        predict.test[[repet]] <- apply(predict(mlp[[repet]],x.test),1,which.max)
        error.test[repet] <- sum(predict.test[[repet]]!=apply(y.test,1,which.max))/length(predict.test[[repet]])
        }
best.mlp <- which.min(error.test)
print(paste("Best : ",best.mlp,"Apprentissage :",error.train[best.mlp],"Test :",error.test[best.mlp]))

# Sauvegarde
save(mlp,predict,error.train,best.mlp,predict.test,error.test,file="ResultatsApp/Res-Equi-0.1-1-15-1.Rdata")
```

- Puis le programme de test et la création des tableaux d'analyses :

```
# Fréquences prédites
load("ResultatsBest/ResEqui.Rdata")
print(round(table(predict.test)/length(predict.test)*100,2))
# Tableau croisé prédit/réel
load("DonneesR/Cartes.RData")
tabcroise <- table(predict.test,carte04)
print(round(tabcroise/length(predict.test)*100,2))
# Pourcentage de bien prédits
print(round(sum(predict.test==as.numeric(carte04))/length(carte04)*100,2))
# Statistiques LUCC
tabcroise2 <- table(predict.test,carte00)/length(carte00)*100
print(round(tabcroise2,2))
Gain <- apply(tabcroise2,1,sum)-diag(tabcroise2)
Loss <- apply(tabcroise2,2,sum)-diag(tabcroise2)
Total.change <- Gain+Loss
AbsNetChange <- apply(tabcroise2,1,sum)-apply(tabcroise2,2,sum)
Swap <- Total.change-AbsNetChange
PercChangeNet <- AbsNetChange/Total.change*100
print(round(as.data.frame(cbind(Gain,Loss,Total.change,Swap,AbsNetChange,PercChangeNet)),2))
# Carte
library(sp)
grille <- expand.grid(xc = 1:2005, yc = 2005:1)
predict.test[predict.test==1] <- "Bois dense"
predict.test[predict.test==2] <- "Bois clairsemé"
predict.test[predict.test==3] <- "Lande boisée"
predict.test[predict.test==4] <- "Lande"
predict.test[predict.test==5] <- "Pelouse"
predict.test[predict.test==6] <- "Rocheux"
predict.test <- factor(predict.test,levels=c("Bois dense","Bois clairsemé","Lande
boisée","Lande","Pelouse","Rocheux"),ordered=T)
grille.pred00 <- data.frame(grille,pred=predict.test)
coordinates(grille.pred00) <- ~xc + yc
gridded(grille.pred00) <- TRUE
spplot(grille.pred00,col.regions=bpy.colors(6),main="Carte prédite pour 2000")
dev.print(device=jpeg,file="ResultatsBest/Predit2000-equi.jpeg",width=500)
```