N° TSE-666

July 2016

"Land use predictions on a regular grid at different scales and with easily accessible covariates"

Raja Chakir, Thibault Laurent, Anne Ruiz-Gazen, Christine Thomas-Agnan, and Céline Vignes

Prédiction de l'usage des sols sur un zonage régulier à différentes résolutions et à partir de covariables facilement accessibles

Raja Chakir[*], Thibault Laurent[**], Anne Ruiz-Gazen[***], Christine Thomas-Agnan[***], and Céline Vignes[**]

[*]Economie Publique, AgroParisTech, INRA, Université Paris-Saclay, 78850 Thiverval-Grignon, France
[**]Toulouse School of Economics (CNRS/GREMAQ), 21 allée de Brienne, 31042 Toulouse, France
[***]Toulouse School of Economics (GREMAQ), 21 allée de Brienne, 31042 Toulouse, France

23 juin 2016

Résumé en français : Nous évaluons dans quelle mesure nous pouvons prédire l'usage des sols (urbain, agricole, forêts, prairies et sols naturels) au niveau des points de l'enquête Teruti-Lucas à partir de covariables facilement accessibles. Notre approche comporte deux étapes : la première permet de modéliser l'usage du sol au niveau des points Teruti-Lucas et la deuxième propose une méthode pour en déduire l'utilisation des sols sur un maillage défini par des carreaux. Le modèle de la première étape fournit des prédictions à un niveau fin. La deuxième étape agrège ces prédictions sur les carreaux du maillage en comparant plusieurs méthodes. Nous envisageons différents maillages réguliers du territoire en carreaux pour étudier la qualité de restitution en fonction de la résolution. Nous montrons qu'avec des variables facilement accessibles on obtient une qualité de prédiction acceptable au niveau point et que l'amélioration de la qualité est importante dès la première étape d'agrégation.

Titre en anglais : Land use predictions on a regular grid at different scales and with easily accessible covariates

Résumé en anglais : We propose in this paper models that allow to predict land use (urban, agriculture, forests, natural grasslands and soil) at the points of the Teruti-Lucas survey from easily accessible covariates. Our approach involves two steps : first we model land use at the Teruti Lucas point level and second, we propose a method to aggregate land use on regular meshes. The model of the first stage provides fine level predictions. The second step aggregates these predictions on the tiles of the mesh comparing several methods. We are considering various regular meshes of the territory to study the prediction quality depending on the resolution. We show that with easily accessible variables we have an acceptable prediction quality at the point level and that the quality of prediction is improved from the very first stage of aggregation.

Classification JEL : C21, C25, C38, Q15, R14 ;

1 Introduction

Etant donnés les coûts élevés d'accès aux données individuelles, nous proposons dans ce travail d'évaluer dans quelle mesure un modèle de prédiction à partir de variables facilement accessibles permet de donner une image de l'usage des sols comparable à celle obtenue à partir de l'information complète au niveau individuel. Nous entendons par information complète celle provenant de l'enquête Teruti-Lucas 2010 du Ministère français de l'Agriculture et par image de l'usage des sols une prédiction de cet usage sur un maillage régulier du territoire en carreaux avec la possibilité de choisir le niveau de maillage. Les problèmes soulevés par cette étude sont tout d'abord le choix d'un modèle de prédiction au niveau de la résolution de la donnée de base, c'est-à-dire le point Teruti-Lucas, ensuite le choix d'une méthode pour agréger les prédictions ponctuelles au niveau des carreaux choisis (carreaux cibles) et enfin le choix d'un critère de qualité des résultats obtenus en fonction de la taille des carreaux cibles.

Il existe un très grand nombre d'études sur la modélisation des usages des sols dans la littérature. Ces études peuvent être classées en différents groupes selon (1) les catégories de l'utilisation des sols examinées (rural *vs* urbain, agriculture *vs* forêt, et agriculture *vs* forêt *vs* urbain), (2) la résolution des données utilisées (données agrégées *vs* données individuelles), (3) la présence ou non de l'interaction spatiale, (4) la prise en compte ou non de la dimension dynamique et (5) la nature de la modélisation (statistique, économétrie, géographie, automate cellulaire) [1].

L'objectif de la modélisation économétrique/statistique en général est soit de concevoir des modèles explicatifs ou des modèles prédictifs. La distinction entre ces deux types de modélisation ainsi que les contours des deux démarches ne sont pas souvent faciles à distinguer [Shmueli, 2010]. Pour résumer, l'objectif des modèles explicatifs est de tester des résultats théoriques qui peuvent être confirmés, infirmés ou précisés par l'estimation des paramètres. L'objectif des modèles prédictifs est de donner les prédictions les plus fiables. Ceci conduit à rechercher des modèles parcimonieux c'est-à-dire avec un nombre volontairement restreint de variables explicatives. Notre objectif dans cette étude est de nous concentrer essentiellement sur la qualité prédictive des modèles et non pas sur leur caractère explicatif.

Nous nous intéressons dans ce papier à la fois à la modélisation statistique et à la modélisation économétrique des usages des sols. Les modèles statistiques se placent en général dans une optique prédictive des usages des sols [Munroe and Müller, 2007; Lambin et al., 2000; Munroe et al., 2004; Veldkamp and Lambin, 2001; Verburg et al., 2004]. Les modèles économétriques peuvent être utilisés dans une optique prédictive ou explicative. Ces derniers cherchent à expliquer les usages des sols observés à l'aide de variables suggérées par la théorie économique [2].

Nous considérons une description de l'usage des sols en cinq catégories : urbain, agricole, forêts, prairies et sols naturels. Notre approche comporte deux étapes. La première étape permet de modéliser l'utilisation du sol au niveau des points Teruti-Lucas. La deuxième étape propose une méthode pour en déduire l'utilisation des sols sur un zonage défini par des carreaux. Dans la première étape, un modèle logit multinomial (MNL) d'utilisation du sol au niveau des points est estimé à partir des données de Teruti-Lucas grâce à des variables explicatives facilement accessibles telles que la base de Corine Land Cover (avec deux nomenclatures différentes), des variables météorologiques, des variables de qualité des sols, des variables socio-démographiques, l'altitude et le prix des terres. Une alternative utilisant des arbres de classement construits à partir des mêmes variables est ensuite comparée au modèle logit multinomial. La comparaison se

1. Voir Chakir [2015] pour une revue de la littérature sur les modèles économétriques d'usage des sols.

2. Voir Irwin and Geoghegan [2001]; Irwin and Wrenn [2014] pour des revues de la littérature des modèles économétriques d'usage des sols et Chakir and Le Gallo [2013]; Chakir and Parent [2009] pour des exemples d'application dans le cas de la France.

fait en calculant des pourcentages de points bien prédits sur un sous-échantillon test, chacun des modèles ayant été ajusté sur un sous-échantillon d'apprentissage. Les résultats obtenus par les MNL et par les arbres sont très comparables.

Les modèles de la première étape permettent de fournir des prédictions à une résolution fine. Plus précisément, notre objectif est de comparer deux méthodes : la régression logistique multinomiale et l'arbre de classement. La régression logistique multinomiale est une méthode très classique dans l'étude de l'usage des sols sur données individuelles (voir McMillen [1989]; Chomitz and Gray [1996]; Nelson and Hellerstein [1997] et Chakir [2015] pour un survol) alors que l'arbre de classement est plus inédite dans ce contexte d'application. Pourtant cette méthode présente l'avantage d'être très simple et de permettre le choix automatique des variables. On considère ensuite un zonage du territoire en carreaux qui représente l'objectif de la restitution finale. La deuxième étape consiste à agréger, en comparant plusieurs méthodes, les prédictions issues de la première étape sur les carreaux du zonage. Nous proposons de mesurer la qualité d'estimation des probabilités au niveau d'agrégation considéré par un score de Brier pondéré. Enfin, nous étudions la qualité de cette estimation en fonction du niveau de résolution considéré pour la restitution de l'image. Une représentation graphique des gains en termes de score de Brier pondéré lors des agrégations successives est proposée comme une aide à la décision pour le choix du niveau de résolution minimal conduisant à une qualité d'estimation convenable. A notre connaissance, une telle comparaison entre différents niveaux d'agrégation est originale.

La section 2, portant sur les méthodes, présente la régression logistique multinomiale, les arbres de classement ainsi que la méthode du score de Brier pondéré. La section 3 décrit les données utilisées et le découpage en échantillon d'apprentissage et échantillon test. Nous y présentons également les différents découpages en carreaux considérés. Nous présentons dans la section 4 les résultats de l'ajustement des différents modèles ainsi que la comparaison de leur qualité de prédiction selon la méthode d'agrégation des probabilités estimées et selon le niveau d'agrégation spatiale.

2 Méthodes

2.1 Régression logistique multinomiale (MNL)

Afin d'estimer économétriquement les déterminants des usages des sols, le modèle théorique suggère que le propriétaire terrien maximise son utilité en comparant les bénéfices et les coûts de conversion des sols d'un usage à un autre à chaque date. Pour passer à la spécification économétrique, nous utilisons le cadre de l'approche de l'utilité aléatoire proposé par McFadden [1974], qui permet de réécrire les revenus et les coûts de conversion de l'utilisation des sols comme des fonctions des variables observées et non observées. Ainsi l'utilité U_{ik} du propriétaire de la parcelle i avec l'utilisation des sols k est la suivante :

$$U_{ik} = \beta_k x_{ik} + \epsilon_{ik} \quad \forall i = 1, \ldots, n \quad , \forall k = 1, \ldots, K, \tag{1}$$

où x_{ik} sont les variables explicatives observées, β_k est le vecteur des paramètres à estimer et ϵ_{ik} sont les termes d'erreur qui tiennent compte des variables non-observées qui pourraient influer sur l'utilité du propriétaire.

Nous supposons que le propriétaire foncier a le choix entre K catégories d'utilisation des sols pour chaque parcelle à chaque date. Le propriétaire choisit l'utilisation optimale des sols de sa parcelle en comparant les utilités associées à chaque catégorie d'utilisation des sols. Si nous notons $y_{it} = 1, 2, \ldots, K$ le choix d'utilisation des sols du propriétaire pour sa parcelle i nous avons :

$$y_i = k, \text{ si } \quad U_{ik} \geq \max_j U_{ij} \quad \forall i = 1, \ldots, n \quad , \forall k = 1, \ldots, K \quad \text{et} \tag{2}$$

Ainsi, la probabilité que la parcelle i soit allouée à l'usage k s'écrit :

$$P(y_i = k) = Pr[U_{ik} \geq \max_j U_{ij}], \quad \forall j, k = 1, \ldots, K. \tag{3}$$

McFadden [1974] identifie trois critères pour l'utilisation d'un modèle logit multinomial : l'indépendance, l'homoscédasticité et la distribution Gumbel des termes d'erreur. En supposant que ces conditions sont remplies, on peut montrer que la probabilité qu'une parcelle i est en usage k est la suivante :

$$p_{ik} = \frac{\exp(\beta_k x_{ik})}{\sum_j^K \exp(\beta_j x_{ij})} \tag{4}$$

Pour mesurer la qualité d'ajustement des modèles estimés, nous considérons plusieurs indicateurs classiques qui sont le critère d'information d'Akaike (AIC) [Akaike, 1974], le critère d'information bayésien (BIC) [Schwarz, 1978] et le R^2 de McFadden [McFadden, 1974] :

$$AIC = -2LL + 2p$$

$$BIC = -2LL + log(n) \times p$$

$$R^2 \text{ de Mc Fadden} = 1 - \frac{LL}{LL_0}$$

avec LL la log-vraisemblance, p le nombre de paramètres à estimer du modèle, n la taille de l'échantillon et LL_0 la log-vraisemblance du modèle nul (sans autre paramètre que la constante). L'AIC et le BIC permettent de pénaliser les modèles en fonction du nombre de paramètres afin de satisfaire le critère de parcimonie, la pénalité étant encore plus grande avec le BIC. Plus ces critères sont faibles, meilleur est le modèle. Le R^2 de McFadden, ou pseudo-R^2, a été construit pour ressembler au R^2 de la régression linéaire mais doit s'interpréter en termes de part de déviance et non en termes de part de variance. Bien que compris entre 0 et 1 (comme le R^2), des simulations ont montré qu'une valeur autour de 0,3 correspond à une valeur élevée de R^2 (Domencich and McFadden [1975, p. 134-135]).

A l'issue de l'ajustement d'un modèle, nous obtenons des probabilités estimées par usage ainsi qu'éventuellement des prédictions d'usage en chaque point. Concernant les probabilités estimées au niveau des points, notons que les modèles MNL fournissent directement une valeur des probabilités estimées pour chaque usage. Si l'on souhaite aller plus loin en associant un usage prédit à chaque point, il faut un processus transformant les probabilités estimées en une prédiction d'usage. Pour les modèles MNL, il est classique de prédire l'usage correspondant à la probabilité estimée maximale. Notons que ce choix correspond au classifieur de Bayes qui minimise le risque de Bayes (Hastie et al. [2009, p. 21]). On pourrait également envisager de faire pour chaque point un tirage d'une loi multinomiale ayant pour paramètres les probabilités estimées des divers usages en ce point. Cette méthode s'avère mauvaise pour les prédictions au niveau des points mais nous reprendrons cette idée ensuite lorsque nous passerons à des niveaux plus agrégés pour la comparer à la prédiction basée sur la probabilité maximale. Ces prédictions étant faites, on peut alors calculer un taux de points bien classés qui est un critère de qualité de la prédiction, en affectant à chaque observation la catégorie d'usage de sol qui maximise la probabilité prédite et en divisant le nombre d'observations correctement prédites par le nombre total d'observations.

Pour vérifier si le pourcentage d'individus bien-classés est significativement meilleur que par un classement aléatoire, nous calculons la quantité suivante [Hair, 2010] :

$$Q_{PRESS} = \frac{n(1 - \tau \times K)^2}{K - 1}$$

où n est la taille de l'échantillon test, τ est le taux de bien-classés dans l'échantillon test et K le nombre de catégories. Si le classement est aléatoire, la statistique Q_{PRESS} suit une loi de χ^2 à 1 degré de liberté.

La fonction `mnlogit` du package **mnlogit** [Zhiyu and Hasan, 2014] de R [Team, 2014] a été utilisée pour estimer les modèles de régression logistique multinomiale.

Nous avons estimé plusieurs modèles MNL, avec notamment un modèle économétrique (MNL-E) où le choix des variables est basé sur la théorie économique et un modèle statistique (MNL-Cm) où nous utilisons une procédure de sélection automatique. L'objectif de ce dernier modèle est d'arriver, à partir de l'ensemble des variables explicatives disponibles, à un modèle final qui retiendrait les variables qui s'avèrent significatives dans l'explication de la variation de la variable dépendante. Il existe plusieurs méthodes de sélection automatique de variables, parmi lesquelles la méthode pas-à-pas (en anglais « stepwise ») que nous avons retenue. Cette méthode, classique en régression [Jobson, 1999], consiste à examiner à chaque étape de la procédure à la fois si une nouvelle variable doit être ajoutée et si une des variables déjà incluse doit être éliminée. La procédure s'arrête lorsqu'aucune variable ne peut être rajoutée ou retirée du modèle selon le critère choisi (seuil de significativité, minimisation du BIC ou de l'AIC, etc.). Nous avons retenu comme critère la maximisation du taux de bien-classés dans l'échantillon test. Notons bien que ces méthodes de sélection automatique des variables ne se basent que sur des critères statistiques sans tenir compte du contexte économétrique.

2.2 Arbres de classement

Les arbres de classement constituent une alternative aux modèles de régression logistique multinomiale [Tufféry, 2010]. Ce sont des arbres de décision ou de partitionnement où la variable à prédire est une variable qualitative à deux ou plusieurs modalités qui définissent des groupes. L'avantage de cette approche est qu'elle intègre une méthode automatique de choix des prédicteurs. L'idée est d'utiliser des prédicteurs qualitatifs ou quantitatifs pour diviser l'échantillon en nœuds ou segments successifs qui correspondent à des sous-échantillons les plus homogènes possible, au sens que les observations d'un nœud appartiennent majoritairement à un groupe. On considère des arbres binaires, c'est-à-dire tels que chacune des divisions conduit à deux nœuds. Le résultat est que l'ensemble des nœuds terminaux forme une partition de l'échantillon en classes homogènes. Chaque division est associée à un prédicteur déterminé par l'algorithme et la dichotomie s'effectue selon des valeurs prises par le prédicteur : sous-ensemble de modalités lorsque le prédicteur est qualitatif et seuil de division lorsque le prédicteur est quantitatif. Pour sélectionner la meilleure division possible, un indice d'homogénéité ou de pureté d'un nœud est choisi. Les étapes successives ainsi que la partition finale sont représentées graphiquement (voir la figure 4) sous la forme d'un arbre renversé, avec la racine en haut de l'arbre qui correspond à l'ensemble de l'échantillon, les divisions successives qui forment les branches et les nœuds terminaux qui constituent les feuilles. Pour obtenir une prédiction pour un individu de l'échantillon, on suit les branches de l'arbre qui correspondent aux valeurs des prédicteurs pour l'individu en question jusqu'au nœud terminal. Puis, par exemple, on adopte la règle majoritaire qui consiste à affecter l'observation à la classe la plus fréquente dans le nœud terminal. On obtient ainsi des règles de décision comme par exemple dans la figure 4 de nos résultats.

Il existe plusieurs algorithmes d'arbres de décision dont la méthode CART [Breiman et al., 1984] qui consiste à construire une arbre maximal puis à l'élaguer. La fonction `rpart` du package **rpart** [Therneau et al., 2014] de R permet la mise en œuvre de cet algorithme pour les arbres de classement. La mesure de l'homogénéité d'un nœud prise par défaut dans `rpart` est l'indice de diversité de Gini que nous avons utilisé et qui permet une réduction de l'impureté des nœuds à chaque division. Parmi toutes les divisions binaires possibles, associées à tous les prédicteurs possibles, et

pour chacun des nœuds, la meilleure division est donc choisie au sens de l'indice de Gini. Le critère d'arrêt de l'algorithme retenu par défaut par la fonction `rpart` est la taille minimale d'un nœud terminal qui doit être de 20 observations. Par ailleurs la vitesse de l'algorithme est améliorée en prenant en compte un paramètre dit de complexité, noté *cp* et fixé par défaut à 0,01. L'idée de ce paramètre est d'introduire un critère pénalisé par le nombre de segments terminaux et de ne pas diviser tous les nœuds, même s'ils contiennent suffisamment d'observations, lorsque le gain n'est pas suffisamment important au regard de la complexité grandissante de l'arbre. Dans notre cas, ce paramètre de complexité a été modifié à la valeur cp=0,0001 pour permettre la prise en compte d'un nombre suffisant de prédicteurs et la comparaison avec les MNL.

Une fois l'arbre maximal obtenu, il est souvent conseillé de l'élaguer, c'est-à-dire de couper les branches les plus longues et les moins informatives, de façon à éviter que les dernières divisions soient instables et dépendent trop de l'échantillon d'apprentissage utilisé. L'algorithme CART utilisé pour l'élagage est composé de deux étapes. La première consiste à construire une suite de sous arbres de l'arbre maximal qui soient imbriqués et qui minimisent le paramètre de complexité. Le choix du sous arbre final dans la suite est ensuite obtenu par validation croisée. Cette procédure ne sera pas détaillée davantage ici mais est décrite en détail dans Breiman et al. [1984].

Alors que les modèles MNL nous fournissent des probabilités estimées, les arbres nécessitent un calcul supplémentaire pour estimer ces probabilités par la fréquence empirique correspondante au groupe de feuilles terminales associées à chaque usage. Par contre, la prédiction au point est directe car un usage est associé à chaque feuille terminale. A partir des prédictions, on peut calculer un taux de bien classés et comparer les résultats des arbres à ceux des MNL.

2.3 Comparaison de la qualité de prédiction en fonction du niveau d'agrégation spatiale

Nous utilisons le score de Brier [Brier, 1950] pour juger de la qualité d'estimation des probabilités ou de prédiction de l'usage des sols. Ce score, initialement proposé dans le domaine de la météorologie, est une mesure d'erreur quadratique moyenne qui reste largement utilisée de nos jours au sein d'une plus large famille d'indicateurs de qualité [Winkler et al., 1996; Buja et al., 2005; Merkle and Steyvers, 2013]. Notre objectif est de déterminer des niveaux d'agrégation qui conduisent à une bonne qualité de prédiction et de mettre en garde l'utilisateur contre l'usage de niveaux trop fins pour assurer une qualité de prédiction ou d'estimation acceptable.

Le score de Brier est défini par :

$$\text{Score de Brier} = \frac{1}{2n} \sum_{k=1}^{K} \sum_{i=1}^{n} (z_{ik} - \hat{p}_{ik})^2$$

avec

$$z_{ik} = \begin{cases} 1 & \text{si l'usage } k \text{ est observé au point } i \\ 0 & \text{sinon} \end{cases}$$

et $\hat{p}_{ik}$ est la probabilité estimée d'observer l'usage k au point i pour $k = 1, \ldots, K$ et $i \in \{1, \ldots, n\}$.

Le score de Brier varie entre 0 et 1. Plus sa valeur est faible, meilleure est la qualité. On peut calculer ce score à différents niveaux d'agrégation. Dans notre application, le niveau le plus désagrégé est le niveau ponctuel (point Teruti-Lucas) tandis que le niveau le plus agrégé est la région Midi-Pyrénées. Ainsi, pour un vecteur de probabilités estimées obtenu par le modèle multinomial logit MNL-Cm tel que détaillé en section 4, on observe un score de Brier au niveau

ponctuel (appelé score individuel) de 0,2502 tandis que le score au niveau de la région (appelé score global) est de 3,49 $\times 10^{-6}$. Intuitivement, on s'attend à une décroissance du score de Brier et donc à une amélioration de la qualité de prédiction au fur et à mesure que l'on agrège les données. On peut montrer toutefois que cette propriété n'est pas vérifiée par le score de Brier car celui-ci ne prend pas en compte la taille des groupes agrégés. Nous proposons une mesure adaptée du score de Brier, appelée score de Brier pondéré et notée B, qui permet de pallier cet inconvénient.

Le score de Brier pondéré est défini pour des groupes G_g ($g \in I_G$) formant une partition de $\{1, \ldots, n\}$ par :

$$\mathrm{B}_G = \frac{1}{2n} \sum_{k=1}^{K} \sum_{g \in I_G} \#G_g \left(\bar{z}_{gk} - \bar{\hat{p}}_{gk} \right)^2$$

avec

$$\bar{z}_{gk} = \frac{1}{\#G_g} \sum_{i \in G_g} z_{ik} \text{ est la fréquence observée de l'usage } k \text{ dans le groupe } G_g$$

$$\bar{\hat{p}}_{gk} = \frac{1}{\#G_g} \sum_{i \in G_g} \hat{p}_{ik} \text{ est la probabilité estimée de l'usage } k \text{ dans le groupe } G_g$$

Dans notre application, les groupes G_g seront par exemple les segments Teruti-Lucas ou des agrégations de segments.

Notons que le niveau individuel est nécessaire pour calculer la taille des groupes G_g dans la définition du score pondéré alors que seul le nombre de groupes est nécessaire au calcul du score de Brier.

Par ailleurs, le score de Brier pondéré global est équivalent au score de Brier global :

$$\mathrm{B}_{\text{global}} = \frac{1}{2} \sum_{k=1}^{K} \left(\bar{z}_k - \bar{\hat{p}}_k \right)^2$$

avec

$$\bar{z}_k = \frac{1}{n} \sum_{i=1}^{n} z_{ik} \text{ est la fréquence observée de l'usage } k \text{ sur l'ensemble de l'échantillon}$$

$$\bar{\hat{p}}_k = \frac{1}{n} \sum_{i=1}^{n} \hat{p}_{ik} \text{ est la probabilité estimée de l'usage } k \text{ sur l'ensemble de l'échantillon}$$

Ce score global est nul dans le cas de la régression logistique multinomiale si des constantes spécifiques aux alternatives sont introduites dans le modèle. En effet la probabilité estimée pour chaque alternative est égale à la fréquence observée de cette alternative dans l'échantillon (Train [2009, p. 62]).

La propriété suivante montre que le score de Brier pondéré agrégé à un niveau donné (G_g) s'écrit comme la somme du score pondéré agrégé à un niveau moins fin (J_j) et de la variance intra des erreurs de prédiction dans les groupes les moins fins (à un facteur multiplicatif près). La preuve de ce résultat est donnée en annexe 6.3. On en conclut que le score de Brier pondéré décroît lorsque l'on agrège des zones et que l'on moyenne les probabilités estimées sur ces zones. Cette décroissance est d'autant plus marquée que la variabilité des erreurs de prédiction est importante dans les zones qui vont être agrégées.

$$\mathrm{B}_G = \mathrm{B}_J + \frac{1}{2n}\sum_{k=1}^{K}\sum_{j\in I_J}\sum_{g\in I_{G_j}} \#G_g\Big(\bar{z}_{gk} - \bar{\hat{p}}_{gk} - \bar{z}_{jk} + \bar{\hat{p}}_{jk}\Big)^2 \tag{5}$$

où les groupes J_j ($j \in I_J$) forment une partition de $\{1,\ldots,n\}$ plus agrégée que les groupes G_g telle que les I_{G_j} ($j \in I_J$) forment une partition de I_G et $J_j = \cup_{g\in I_{G_j}} G_g$ avec $j \in I_J$; $\bar{z}_{jk} = \dfrac{1}{\#J_j}\displaystyle\sum_{i\in J_j} z_{ik}$ et $\bar{\hat{p}}_{jk} = \dfrac{1}{\#J_j}\displaystyle\sum_{i\in J_j} \hat{p}_{ik}$.

Ainsi, plus le niveau d'agrégation est fin, plus le score de Brier pondéré est élevé et pire est la prédiction ou l'estimation.

Pour choisir parmi plusieurs niveaux d'agrégation $A_1,\ldots,A_L$ où A_1 (resp. A_L) désigne le niveau le plus (resp. le moins) fin, nous proposons de comparer graphiquement les différences entre scores de Brier pondérés successifs $B_{l-1} - B_l$ divisées par le score de Brier individuel noté B_0. Ces ratios successifs mesurent le gain en termes de qualité lorsque l'on passe du niveau d'agrégation A_{l-1} à A_l. Leur somme de $l = 1$ à L vaut 1. La section 4.2 illustre la mise en œuvre de cette méthode et l'interprétation du graphique pour le choix du niveau d'agrégation dans le cadre de l'enquête Teruti-Lucas.

3 Données

Notre zone d'étude est la région Midi-Pyrénées qui est la plus vaste région française (45 348 km^2, soit 8,3 % du territoire national) et celle comportant le plus grand nombre de communes (3 020 communes au 1er janvier 2013). C'est une région plutôt rurale (cf. figure 1) qui n'abrite que 4,5 % de la population métropolitaine (2 903 420 habitants en 2011 [3]). Elle présente l'avantage d'une certaine diversité d'usage des sols avec Toulouse, le grand pôle urbain, et de vastes zones agricoles au centre, les Pyrénées au sud et des forêts et des prairies au nord (cf. figure 3).

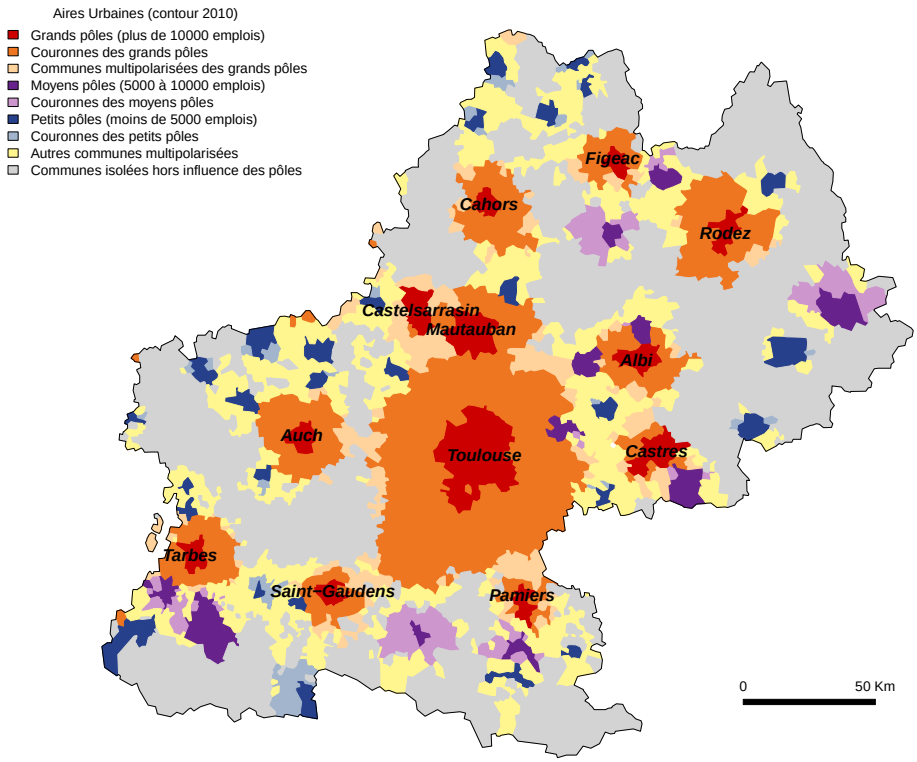
3.1 Variable d'intérêt *Usage du sol* : données Teruti-Lucas 2010

Les données d'usage des sols proviennent de l'enquête Teruti-Lucas, réalisée par le Service de la Statistique et de la Perspective (SSP) du Ministère de l'Agriculture. Cette enquête permet de renseigner l'évolution de l'occupation des sols au niveau d'un échantillon de points répartis sur l'ensemble du territoire français. L'enquête repose sur l'association de photographies aériennes et d'enquêtes de terrain. Cette enquête a débuté en 1982 mais l'échantillon de points a été entièrement renouvelé en 1991 et 1992 et la nomenclature des usages des sols a été modifiée. En 2005, l'enquête Teruti a évolué et l'échantillon de points a été totalement renouvelé pour (i) améliorer la précision de l'enquête en utilisant les progrès réalisés en matière de géo-référencement et de traitement de données cartographiques et (ii) permettre une cohérence de nomenclature et de méthode avec une enquête similaire réalisée à l'échelle européenne : l'enquête européenne LUCAS (Land Use/Cover Area frame statistical Survey). La nouvelle enquête s'appelle Teruti-Lucas.

Dans l'enquête Teruti-Lucas, chaque segment comporte 25 points alignés par 5 (cf. figure 2) et seuls les 10 points des deux premières lignes sont disponibles dans les données que nous utilisons ensuite (les autres ont été créés pour d'éventuelles extensions de l'enquête). L'échantillon pour

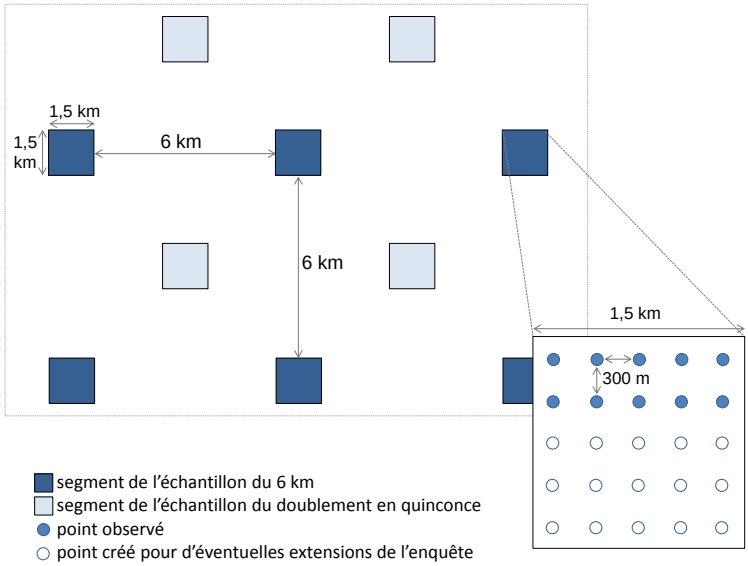
3. source : Insee, `http://www.insee.fr/fr/regions/midi-pyrenees/default.asp?page=faitsetchiffres/presentation/essentiel_mp.htm`

FIGURE 1 – Midi-Pyrénées : Zonage en aires urbaines



la période 2006-2010 est donc constitué de 309 080 points pour la France métropolitaine et de n=25 317 points pour la région Midi-Pyrénées, observés chaque année.

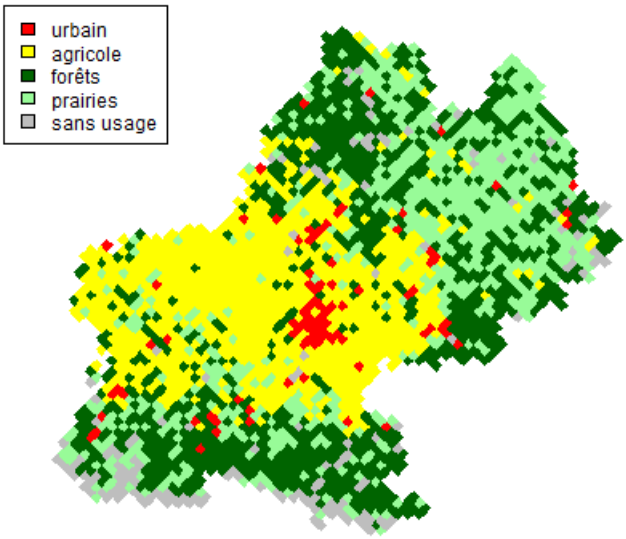
FIGURE 2 – Enquête Teruti-Lucas : segments et points



On s'intéresse uniquement à l'occupation physique des sols, et non à l'utilisation fonctionnelle qui est l'autre principale variable recueillie et qui mesure la destination socio-économique du territoire observé. L'occupation physique des sols est directement déduite de l'observation. Elle est codée selon deux nomenclatures : une nomenclature de synthèse en 57 postes, qui permet la continuité avec la version précédente de l'enquête en France (Teruti 1992-2004) et une nomenclature en 122

postes, complètement remaniée mais qui recouvre les mêmes catégories d'usage. Nous utiliserons la nomenclature en 57 postes, regroupés en cinq catégories : usage urbain, usage agricole, forêts, prairies et sols naturels. Ces cinq catégories sont pertinentes pour étudier par exemple l'impact des changements d'usage des sols sur les émissions de gaz à effets de serre et plus exactement sur la séquestration de carbone. De plus, étant données les variables explicatives dont on dispose, il serait difficile de détailler l'usage agricole par exemple. En effet l'allocation des sols entre blé, orge et colza fait partie des décisions basées sur des considérations agronomiques (rotations, système de production, ...) qu'on n'observe pas dans nos données. La figure 3 illustre l'usage majoritaire au segment et l'annexe 6.1 fournit la correspondance entre la nomenclature en 57 postes et notre variable d'intérêt, le regroupement en cinq catégories.

FIGURE 3 – Usage des sols majoritaire dans le segment



3.2 Variables explicatives

Les variables explicatives retenues proviennent de multiples bases de données disponibles à des niveaux géographiques différents. Le tableau 1 dresse une synthèse de ces bases de données et des variables en découlant. L'annexe 6.2 fournit des compléments d'information sur ces bases, notamment un résumé des aires des différentes zones (CLC, UCS, NRA, communes) dans le tableau 9 et les statistiques descriptives des variables retenues dans les modèles (tableaux 10 à 15). Les variables mesurées avec la résolution la plus fine sont l'occupation du sol au sens de Corine Land Cover et l'altitude.

Utilisant des outils d'économétrie spatiale, nous avons aussi introduit des variables spatialement décalées parmi les variables explicatives. En effet, l'occupation des sols (CLC), la densité de population, les parts d'agriculteurs et de cadres ainsi que l'altitude ont été mesurés aux points voisins. Nous considérons que les voisins d'un point Teruti-Lucas sont les points appartenant au même segment. Pour les identifier et construire les variables spatialement décalées, nous avons défini une matrice de voisinage basée sur un seuil de distance de 1 400 m. La matrice est normalisée par une standardisation en ligne. Ainsi, les variables quantitatives spatialement décalées représentent la valeur moyenne des autres points du segment. L'occupation des sols étant une variable qualitative, nous avons plutôt choisi de retenir la valeur majoritaire dans le segment et de ne pas

3. Les statistiques descriptives de l'aire de ces zones sont données dans le tableau 9 de l'annexe 6.2.

TABLE 1 – Sources des données

nom	niveau géographique	source	année	unité
usage des sols	points espacés de 6km	Teruti-Lucas	2010	-
occupation des sols (CLC)	zones (>25 ha)[3]	Corine Land Cover	2006	-
altitude	points espacés de 250m	BDAlti de l'IGN	-	mètres
nature du sol	zones UCS[3]	BGSF (© INRA, Unité INFOSOL, Orléans)	1998	-
texture dominante en surface				-
matériau de base				-
évolution de la texture du sol				-
présence d'une couche imperméable				-
météorologie	grille 25 × 25 km	JRC-MARS Agri4cast Meteorological Data Base	2010	-
température minimum (minimum annuel des températures journalières)				°C
température maximum (maximum annuel des températures journalières)				°C
température moyenne (moyenne annuelle des temp. moyennes journalières)				°C
somme annuelle des précipitations				millimètres
vitesse moyenne du vent				km/h
prix des terres et prés libres de plus de 70 ha	32 NRA[3]	Agreste	2010	€ courant/ha
données socio-économiques	communes	Insee	2010	-
densité de population				habitants/km^2
part d'agriculteurs				%
part de cadres				%
grand pôle urbain				-

utiliser la matrice de voisinage. Les données manquantes pour les variables spatialement décalées (c'est-à-dire point Teruti-Lucas seul dans le segment) sont exclues des analyses.

3.3 Echantillon d'apprentissage et échantillon test

Afin de pouvoir évaluer le pouvoir prédictif des modèles et la validité externe des résultats obtenus, nous utilisons la stratégie consistant à diviser l'échantillon d'origine en deux parties, l'une dite échantillon d'apprentissage servant à construire un modèle et l'autre dite échantillon test servant à comparer la réalité observée à sa prédiction par ce modèle. Pour cela nous divisons l'ensemble des segments Teruti-Lucas en deux parties : l'échantillon dit « du doublement en quinconce » comportant 12 657 points servant d'échantillon d'apprentissage (en clair dans la figure 2) et l'échantillon test dit « du 6 km doublé » comportant 12 660 points [4]). De ce fait la répartition des points de chaque échantillon sur le territoire est bien régulière.

Le tableau 2 donne la fréquence des différentes catégories d'usage des sols dans ces deux échantillons. Notons que pour 11 points Teruti-Lucas (tous situés dans l'échantillon test) l'usage des sols est manquant (nphys=9999 *zones interdites, photos non interprétées*).

4. voir `http://agreste.agriculture.gouv.fr/IMG/pdf/teruti2014methobsva.pdf`

TABLE 2 – Fréquence des catégories d'usage des sols en 2010 dans les deux échantillons (nombre de points Teruti-Lucas et %)

landuse	libellé	échantillon d'apprentissage		échantillon test		total	
		effectif	%	effectif	%	effectif	%
1	usage urbain	933	7,4	909	7,2	1 842	7,3
2	usage agricole	3 344	26,4	3 252	25,7	6 596	26,1
3	forêts	3 906	30,9	4 051	32,0	7 957	31,4
4	prairies	3 231	25,5	3 279	25,9	6 510	25,7
5	sols naturels	1 246	9,8	1 155	9,1	2 401	9,5
Total		12 660	100	12 646	100	25 306	100

3.4 Niveaux d'agrégation

Comme évoqué dans l'introduction et dans la section 2.3, nous considérons plusieurs zonages du territoire en carreaux de façon à étudier la qualité d'estimation et de prédiction en fonction de la résolution finale. Pour cela, nous allons définir une série de maillages emboîtés obtenus par agrégations successives des carreaux du maillage précédent.

Le niveau d'agrégation le plus fin, noté A_0, est constitué des points Teruti-Lucas tandis que le niveau le plus agrégé, noté A_7, est la région Midi-Pyrénées. Les niveaux intermédiaires sont des grilles. La grille initiale, notée A_1, a été conçue de façon à ce que chaque carreau contienne un segment Teruti-Lucas et un seul, afin de former un maillage ininterrompu du territoire. On parlera du niveau « segments ». Les carreaux de cette grille, dénommés ci-après « carreaux unités », sont centrés sur le barycentre des points d'un segment Teruti-Lucas et ont une taille de 4,2 km de côté. Cette grille comporte 2 579 carreaux et est représentée dans la figure 3.

Nous avons aussi construit des agrégations de ces « carreaux unités » afin d'obtenir un maillage de plus en plus grossier jusqu'à obtenir la région entière. A chaque étape, quatre carreaux sont réunis pour en former un seul. On obtient donc une deuxième grille (niveau d'agrégation A_2) dont chaque carreau est constitué de quatre « carreaux unités », puis une troisième (niveau d'agrégation A_3) dont chaque carreau est constitué de 16 « carreaux unités », etc. jusqu'à la dernière grille (niveau d'agrégation A_6) dont chaque carreau est constitué de 1 024 « carreaux unités »(cf. tableau 3).

TABLE 3 – Caractéristiques des grilles

Grille	Nombre de « carreaux unités » agrégés	Superficie approximative	Nombre de points par carreau	Nombre total de carreaux
A_1	1	18 km^2	1 à 10	2 579 carreaux
A_2	4	72 km^2	1 à 40	689 carreaux
A_3	16	288 km^2	4 à 160	192 carreaux
A_4	64	1 152 km^2	10 à 640	59 carreaux
A_5	256	4 608 km^2	184 à 2 559	20 carreaux
A_6	1 024	18 432 km^2	184 à 6 605	8 carreaux

Cette famille de maillages étant définie, nous devons à présent expliquer comment définir les probabilités estimées ainsi que les prédictions à chaque niveau.

Pour agréger les estimations, la méthode standard dans cette situation, que ce soit pour les MNL ou les arbres, est de calculer tout d'abord une simple moyenne des probabilités estimées au niveau point sur l'ensemble des points d'un carreau donné. Pour définir par contre les prédictions

au niveau agrégé, on a le choix entre deux solutions. La première est l'exact pendant de ce qu'on a fait au niveau des points : elle consiste, à partir des probabilités estimées agrégées, à prédire par l'usage correspondant à la probabilité estimée maximale. Nous proposons l'alternative suivante qui consiste à faire au niveau de chaque point un tirage aléatoire d'une multinomiale ayant pour paramètres les probabilités estimées au point et à en déduire ensuite au niveau du carreau les fréquences empiriques correspondant à chaque usage. Il est important de remarquer qu'il n'est plus possible, pour évaluer la qualité des prédictions, d'utiliser le taux de bien classés comme nous l'avions fait au niveau des points, car au niveau des carreaux, nous ne pouvons plus comparer la prédiction à une réalité.

4 Résultats

Dans cette partie, nous présentons les résultats d'ajustement et de prédiction des divers modèles. Il est important de noter que pour la première étape, il s'agit de l'estimation et de la prédiction au niveau des points Teruti-Lucas alors que pour la deuxième étape, il s'agit des estimations au niveau des carreaux.

4.1 Première étape : estimation et prédiction au niveau des points Teruti-Lucas

Dans cette étape, les modèles sont ajustés sur les observations aux points Teruti-Lucas et les prédictions sont calculées à ce même niveau. Les modèles ajustés sont ceux présentés à la section 2 : les modèles MNL et les arbres de classement. Nous partons d'un premier modèle multinomial logit dit « économétrique », noté MNL-E, pour lequel le choix des variables est motivé par la littérature. Selon la littérature économique empirique [Lubowski, 2002] sur l'utilisation des sols, les variables explicatives qui influent sur la décision de l'utilisation des sols incluent les rentes associées aux différents usages ou à défaut des *proxy* de ces rentes tels que les prix des inputs et des outputs, les aides publiques, la densité de la population (comme *proxy* de la rente urbaine), ainsi que d'autres variables pédoclimatiques telles que pente, altitude, qualité du sol, température, précipitations, etc.

Les résultats du modèle économétrique MNL-E présentés dans le tableau 16 montrent que les modalités de la variable couverture des sols « CLC2 »sont significatives pour déterminer les usages des sols. Le prix des terres agricoles est significatif et a un effet positif sur l'usage agricole, en forêt et prairies par rapport à l'usage urbain. La densité de la population n'a pas d'impact significatif sur les usages agricoles et forêt mais a un impact négatif et significatif sur les usages prairies et sols naturels. Un résultat qui peut paraître surprenant est la non-significativité des modalités de la variable qualité des sols. Une explication possible serait le fait que l'information contenue dans cette variable est déjà prise en compte dans la variable de couverture des sols « CLC2 »ainsi que dans la variable altitude. En ce qui concerne la qualité de prédiction du modèle MNL-E, le taux de points bien classés est de 65,12 % ce qui est assez comparable aux autres modèles MNL-Cm et MNL-S (voir tableau 4).

En cherchant à optimiser la prédiction (sélection du modèle par méthode pas-à-pas maximisant le taux de bien-classés dans l'échantillon test), on obtient un ensemble de variables explicatives pour le modèle MNL-Cm qui diffère de la sélection réalisée pour MNL-E (cf. tableau 17). En particulier les variables socio-économiques (« prixterre », « densité ») et la texture du sol ne sont pas retenues. Les variables « température moyenne »et « pluie »disparaissent au profit du « vent moyen »et de « altitude »des voisins. Les six variables retenues sont « CLC2 », « altitude », « température minimum », « vitesse du vent », « température maximum »et « altitude »dans le

TABLE 4 – Taux de bien-classés selon le modèle

Variables restantes dans le modèle	MNL	arbres
CLC2 + altitude	MNL-S : 65,15 %	TREE-S : 64,92 %
CLC2 + altitude + MAT + TEXT + prixterre + cadre	-	TREE-Ct : 64,83 %
CLC2 + altitude + tempmini + ventmoy + tempmaxi + altitude_lag	MNL-Cm : 65,37 %	-
CLC2 + altitude+prixterre+densite + tempmini + tempmaxi + tempmoy + pluie + TEXT	MNL-E : 65,12 %	TREE-E : 65,04 %

voisinage. Le taux de bien-classés est légèrement amélioré avec une valeur de 65,37 % (voir tableau 4).

Concernant les arbres de classement, on utilise l'algorithme CART et l'ensemble des variables explicatives hormis celles mesurées dans le voisinage [5]. Si on utilise les paramètres par défaut de la fonction `rpart`, le critère de complexité vaut 0,01 et on obtient un arbre très élagué qui ne fait intervenir que deux variables explicatives (« CLC2 »et « altitude »). Pour obtenir un nombre de variables comparable à celui de MNL-Cm, nous fixons le critère de complexité initial à 0,0001. Pour obtenir l'arbre TREE-Ct, nous utilisons la fonction `rpart` qui fournit un arbre maximal et la fonction `prune` qui permet d'élaguer l'arbre en utilisant la procédure décrite dans Breiman et al. [1984]. Avec cette méthode, on aboutit à un cp=0,001. L'arbre retenu est très complexe, il comporte 16 nœuds et prend en compte 6 variables (« CLC2 », « altitude », « prix des terres », « part de cadres », « matériau de base »et « texture du sol »). Les variables météorologiques n'interviennent plus mais on retrouve des variables socio-économiques (« prix des terres »et « part de cadre ») et portant sur la nature du sol (« texture », comme dans MNL-E, et « matériau de base »). En raison de sa complexité, il ne sera pas détaillé davantage. Le taux de bien-classés est de 64,83 %, soit légèrement inférieur à ceux de MNL-E et MNL-Cm (voir tableau 4).

A partir de cet arbre, on cherche à maximiser le taux de bien-classés en élaguant progressivement l'arbre (cf. tableau 5). Le meilleur taux de bien-classés, égal à 64,92 %, est obtenu avec l'arbre TREE-S qui comporte 10 nœuds et seulement deux variables : CLC2 et altitude (cf. figure 4).

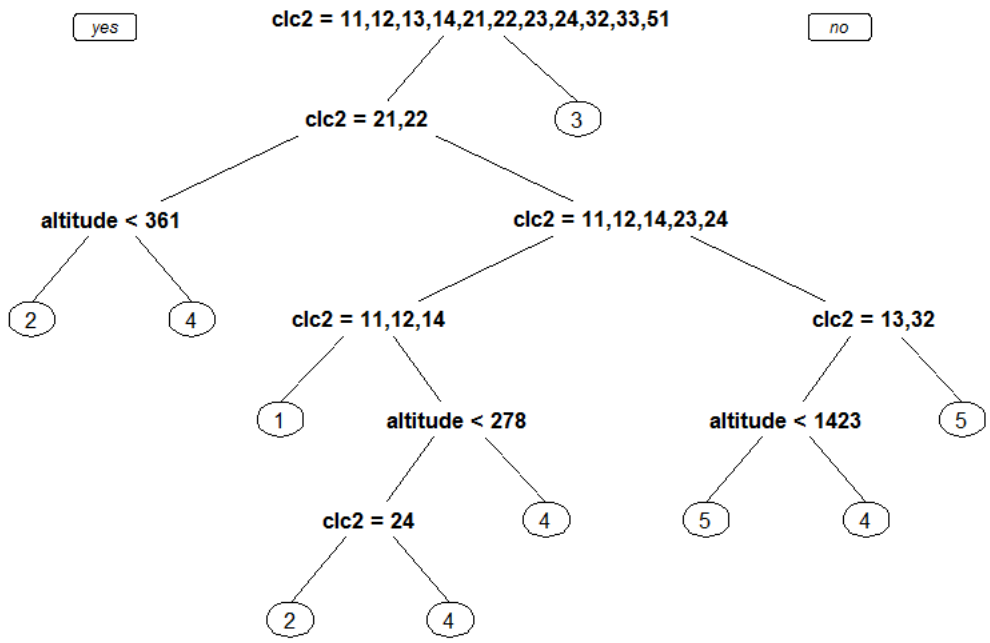
TABLE 5 – Evolution du taux de bien-classés selon le critère de complexité de l'arbre

cp	nombre de nœuds	taux de bien-classés	
0,001	16	64,83 %	(TREE-Ct)
0,002	14	64,90 %	
0,0025	13	64,87 %	
0,003	11	64,92 %	
0,005	10	64,92 %	(TREE-S)
0,007	8	64,48 %	
0,008	7	64,15 %	
0,01	6	63,47 %	

Enfin si on construit un MNL uniquement avec les variables de l'arbre TREE-S (CLC2 et altitude), modèle MNL-S (cf. tableau 18), on obtient un taux de bien-classés de 65,15 %, équivalent à celui

5. Notons que si on introduit les variables de voisinage, d'une part l'altitude des voisins remplace l'altitude et, d'autre part, la part de cadres dans le voisinage est aussi sélectionné (en plus de la variable initiale), sans que cela améliore le taux de bien-classés qui devient 64,73 %.

FIGURE 4 – Arbre TREE-S



1 urbain	$CLC2$=(« Zones urbanisées », « Zones industrielles ou commerciales et réseaux de communication » ou « Espaces verts artificialisés, non agricoles »)
2 agricole	a. $CLC2$=(« Terres arables » ou « Cultures permanentes ») et $altitude < 361$m b. $CLC2$=(« Zones agricoles hétérogènes ») et $altitude < 278$m
3 forêts	$CLC2$=(« Forêts » ou « Zones humides intérieures »)
4 prairies	a. $CLC2$=(« Terres arables » ou « Cultures permanentes ») et $altitude \geq 361$m b. $CLC2$=« Prairies » et $altitude < 278$m c. $CLC2$=(« Prairies » ou « Zones agricoles hétérogènes ») et $altitude \geq 278$m d. $CLC2$=(« Mines, décharges et chantiers », « Milieux à végétation arbustive et/ou herbacée ») et $altitude \geq 1423$m
5 sols naturels	a. $CLC2$=(« Mines, décharges et chantiers », « Milieux à végétation arbustive et/ou herbacée ») et $altitude < 1423$m b. $CLC2$=(« Espaces ouverts, sans ou avec peu de végétation » ou « Eaux continentales »)

du modèle économétrique MNL-E et à peine inférieur à celui du modèle MNL-Cm (voir tableau 4).

La principale différence entre les modèles MNL-S et MNL-E est la non significativité du coefficient de la variable altitude pour l'usage agricole dans le modèle économétrique alors que ce coefficient est significatif à 1 pour 1 000 dans le modèle simple.

Le tableau 6 regroupe des indicateurs de qualité pour les trois modèles de régression logistique multinomiale présentés. Nous voyons que les trois modèles fournissent des résultats très similaires.

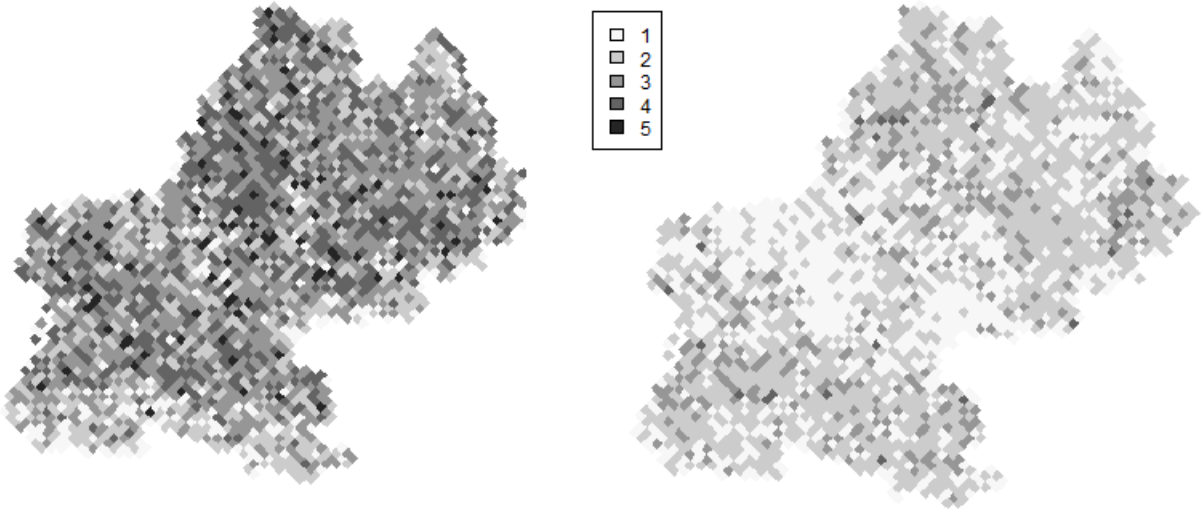
Avec près de deux tiers de points dont l'usage des sols est correctement prédit, les modèles sont très

TABLE 6 – Comparaison de la qualité des modèles de régression logistique multinomiale

Indicateur	MNL-E	MNL-Cm	MNL-S
BIC	26 255,1	26 197,3	26 197,1
AIC	25 540,3	25 661,2	25 780,2
R^2 de Mc Fadden	0,3251	0,3206	0,3166
taux de bien-classés	65,12 %	65,37 %	65,15 %

significativement meilleurs qu'un classement aléatoire, selon la statistique Q_{PRESS}. On remarque toutefois qu'on ne retrouve pas la variabilité observée d'usages de sol au niveau « segment »dans nos prédictions. Alors que 67 % des segments ont au moins trois usages observés différents, cette proportion tombe à 14,7 % pour les usages prédits. La figure 5 illustre aussi ce point.

FIGURE 5 – Nombre d'usages de sol différents par segment, usages observés (carte de gauche) ou prédits par le modèle MNL-Cm (carte de droite)



Par ailleurs, même si nous ne présentons pas les résultats, les taux de bien-classés obtenus pour les points de l'échantillon d'apprentissage et pour les points de l'échantillon test sont très similaires, ce qui montre la stabilité de nos modèles. Enfin il est intéressant de remarquer que les risques empiriques de Bayes (qui valent 1 moins la moyenne empirique des probabilités maximales estimées) sont pratiquement égaux aux taux de mal-classés, soit autour de 35 %. Il paraît donc difficile d'améliorer ces modèles au niveau ponctuel. Mais on peut essayer d'améliorer la qualité de prédiction en considérant des niveaux spatiaux plus agrégés. Cela fait l'objet de la section suivante.

4.2 Deuxième étape : estimation au niveau des carreaux

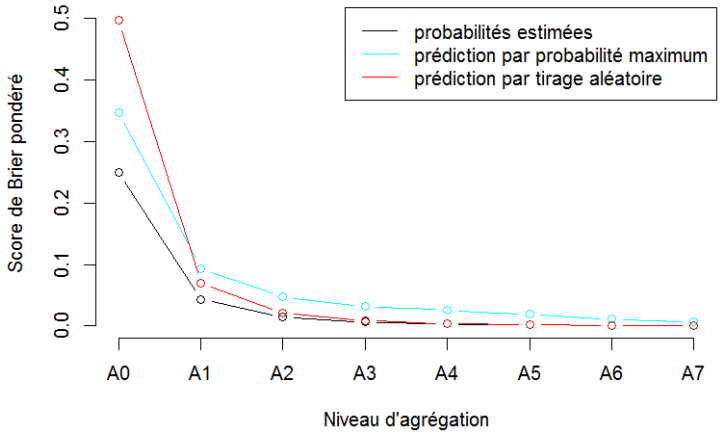
L'objectif final est d'obtenir des prédictions ou des probabilités estimées de l'usage des sols sur un carroyage du territoire que l'on utilise ensuite soit pour cartographier soit comme ingrédient dans un modèle. Selon le cas, on peut avoir à se poser la question du choix de la taille de ce carroyage. Dans tous les cas, il va falloir agréger sur les carreaux les probabilités estimées obtenues au niveau des points lors de la première étape. Nous abordons ces deux questions dans cette partie.

Pour mesurer la qualité de prédiction ou d'estimation des probabilités, nous utilisons le score de Brier pondéré défini à la section 2.3. Les niveaux d'agrégation A_0 à A_7 sont définis à la section 3.4, A_0 dénote le niveau des points Teruti-Lucas, A_1 le niveau des segments, A_2 à A_6 les grilles

successives et A_7 le niveau global de la région Midi-Pyrénées.

Les figures 6 à 8 représentent les scores de Brier pondérés en fonction du niveau d'agrégation, selon trois méthodes de prédiction ou d'estimation des probabilités pour la première figure et selon différents modèles pour les deux figures suivantes. La premières des trois méthodes que nous comparons dans la figure 6 consiste à utiliser les probabilités estimées au niveau point et à les agréger aux niveaux supérieurs. Les deux autres méthodes, détaillées à la section 3.4, sont basées sur une prédiction au niveau point et conduisent à des probabilités agrégées aux niveaux supérieurs. Sur l'ensemble de ces figures, on remarque un score plutôt élevé au niveau individuel (A_0), beaucoup plus faible au niveau « segments »(A_1), encore un peu plus faible au niveau A_2 (où un carreau contient quatre segments) et qui se stabilise en général à partir du niveau A_3 (16 segments) pour finir comme attendu à zéro au niveau de la région (A_7) pour les modèles logit.

FIGURE 6 – Scores de Brier pondérés pour le modèle MNL-Cm selon la méthode de prédiction



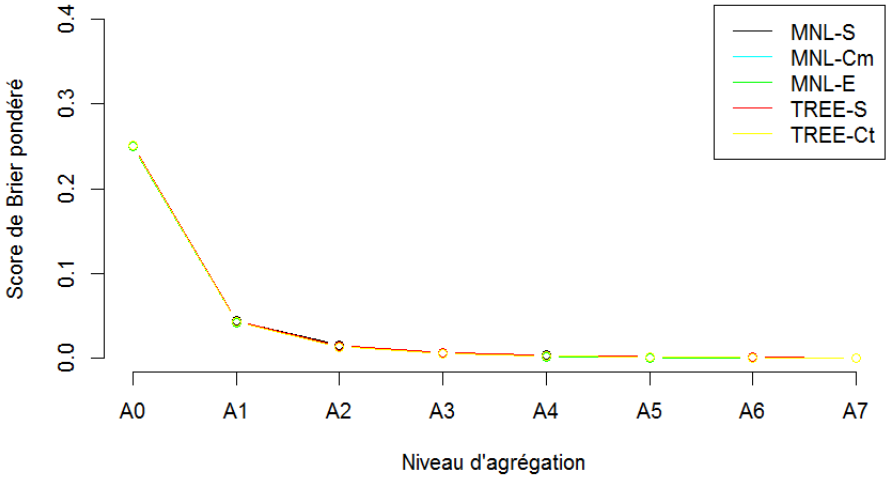
La figure 6 montre que l'utilisation des probabilités estimées conduit à un score de Brier plus faible que les deux méthodes d'agrégation des prédictions. Au niveau individuel, le tirage multinomial est particulièrement mauvais mais dépasse la probabilité maximale dès le niveau « segments »et est comparable au score obtenu avec les probabilités estimées à partir du niveau A_2.

La figure 7 montre que les modèles sélectionnés à la section 4 conduisent à des scores de Brier complètement équivalents pour chacun des niveaux d'agrégation spatiale envisagés.

La figure 8 propose une comparaison des scores de Brier pondérés entre le modèle MNL-Cm et des modèles simplifiés. On voit clairement que le modèle qui ne prend en compte que la variable CLC2 donne des résultats très similaires aux résultats du modèle MNL-Cm. Peu importe le niveau d'agrégation, en termes de qualité de prédiction, il n'est pas utile d'introduire d'autres variables que l'occupation du sol mesurée par Corine Land Cover au niveau 2 (15 catégories). De manière symétrique, les autres variables du modèle économétrique sans CLC2 mènent à une qualité de prédiction comparable à CLC2 à partir du niveau A_3. Au contraire, les Corine Land Cover au niveau 1 (5 catégories) ne permettent pas d'atteindre une qualité équivalente aux autres modèles quel que soit le niveau d'agrégation (sauf au niveau global par propriété des MNL).

Les figures 9 et 10 permettent d'analyser le gain relatif en termes de score de Brier pondéré lors des étapes d'agrégation successives. Précisément on calcule la différence entre le score aux niveaux A_{l-1} et A_l rapportée au score individuel. Ainsi pour le modèle MNL-Cm, ce ratio vaut 0,826 lorsque l'on passe du niveau individuel au niveau « segments », ce qui signifie que l'on élimine 83 % de la valeur du score de Brier pondéré en agrégeant au segment. On note que la courbe

FIGURE 7 – Scores de Brier pondérés selon le modèle (méthode : probabilités estimées)



n'est pas nécessairement décroissante. En particulier pour le modèle avec CLC1 seulement, le gain relatif augmente à la fin ce qui est dû au fait que la somme de ces ratios est égale à 1.Ce graphique est une aide à la décision pour le choix d'un niveau d'agrégation. Dans notre exemple, il montre qu'il faut agréger au niveau « segments », ou davantage, car cette première agrégation apporte un gain majeur en termes de qualité de prédiction tandis que les agrégations suivantes n'apportent pas un gain notable.

Notons qu'on peut interpréter les valeurs du ratio en termes de diminution d'erreur moyenne d'estimation des probabilités en remarquant que la racine carrée du double du score de Brier correspond à l'erreur quadratique moyenne qui est du même ordre que l'erreur moyenne en valeur absolue. Ainsi, pour un gain de 0,826, on peut calculer $\sqrt{2 \times (1 - 0,826)} = 0,59$ et dire que l'erreur moyenne en valeur absolue est diminuée de 41 % en agrégeant du point au segment.

Il est possible comme dans la figure 10 de décomposer les ratios selon les différents usages des sols. Dans notre exemple, il n'y a pas de différence notable entre les différents usages et il faut toujours agréger au niveau « segments »pour obtenir une bonne qualité de prédiction.

FIGURE 8 – Scores de Brier pondérés selon le modèle, modèles moins « bons » (méthode : probabilités estimées)

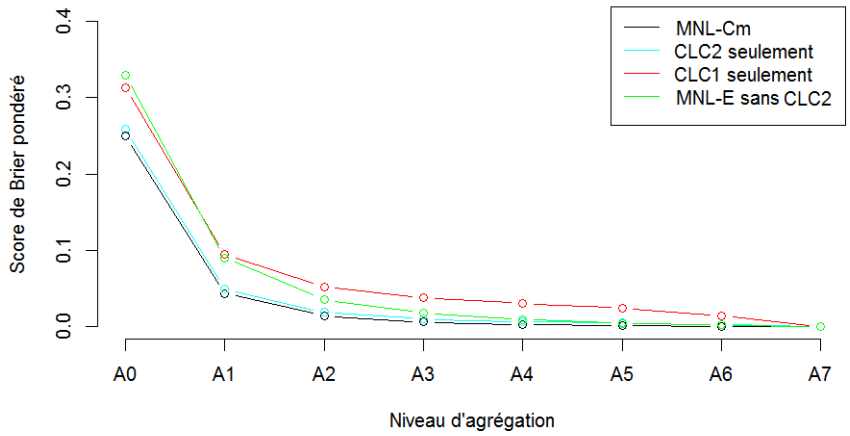
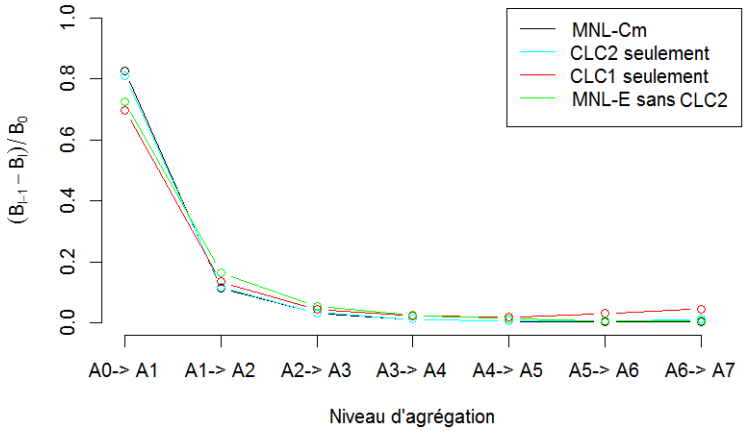


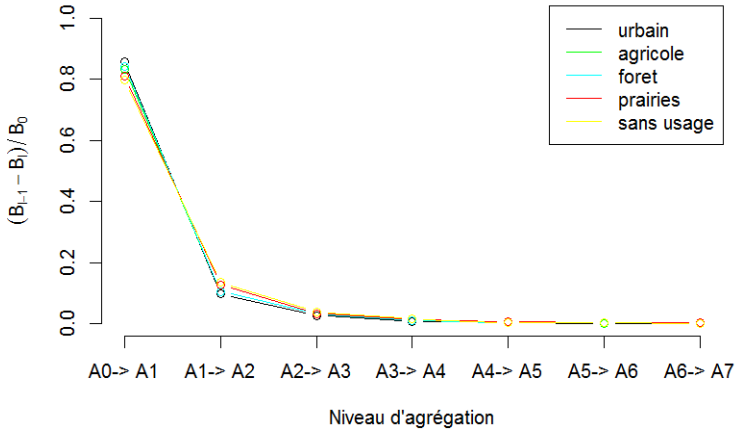
FIGURE 9 – Apport de l'agrégation selon le modèle (score de Brier pondéré total, méthode : probabilités estimées)



5 Conclusion

Notre objectif était de modéliser l'usage des sols en se restreignant à des covariables facilement disponibles, sur Internet ou sur demande auprès de fournisseurs de données. Il est reconnu qu'une des difficultés de la modélisation des usages des sols au niveau individuel est le manque fréquent de « bonnes » variables explicatives ou leur incompatibilité d'échelle, en particulier pour les variables économiques (rentes, coûts de conversion et les prix) [Chakir, 2015]. Nous avons été confrontés à ces difficultés autant pour les variables socio-économiques, qui n'étaient disponibles qu'au niveau communal ou à celui des NRA (Nouvelles Régions Agricoles), que pour les variables météorologiques (dont la résolution n'était que de 25×25 km). Nos analyses montrent le fort pouvoir explicatif de deux variables très simples, l'occupation des sols mesurée par Corine Land Cover (au niveau 2, soit en 15 catégories) et l'altitude. Ces deux variables sont retenues dans l'ensemble des modèles étudiés. Rappelons qu'à elles seules, elles conduisent à un taux de bien-classés de 65,15% (modèle de régression logistique multinomiale dit « simple »). L'ajout de

FIGURE 10 – Apport de l'agrégation selon l'usage des sols (score de Brier pondéré par usage, méthode : probabilités estimées)



variables supplémentaires (météorologiques, socio-économiques ou biophysiques) n'améliore que très marginalement les taux de bien-classés.

Deux autres résultats nous semblent importants à souligner. Tout d'abord, les modèles MNL testés donnent systématiquement des résultats légèrement meilleurs que les arbres de classement équivalents (c'est-à-dire avec le même sous-ensemble de prédicteurs) en termes de taux de bien-classés. Ensuite, nous remarquons une grande stabilité des résultats, en particulier une robustesse des arbres de classement, qui n'est pas classique dans la littérature, mais qui provient du fait que la variable CLC2 est une variable qualitative qui joue un rôle prépondérant dans les résultats.

Plusieurs pistes de recherche peuvent être envisagées à la suite de ce travail.
Premièrement, face à la difficulté de sélection de variables en présence d'un grand nombre de variables qualitatives ayant de nombreuses modalités, les méthodes de type lasso peuvent être envisagées. Ces méthodes ont l'avantage d'utiliser une pénalisation pour combiner la phase d'estimation à celle du choix d'un modèle parcimonieux dans le cas des MNL [Meier et al., 2008; Tutz et al., 2015].

Deuxièmement, la qualité de prédiction des modèles logit multinomiaux pourrait être améliorée en y introduisant explicitement l'auto-corrélation spatiale. L'estimation des modèles de choix discret avec de l'auto-corrélation spatiale reste un défi au niveau des calculs. Les méthodes d'estimation développées par Ferdous and Bhat [2013]; Sidharthan and Bhat [2012] semblent être prometteuses et présentent une bonne alternative aux méthodes bayésiennes ou aux méthodes d'estimation par simulation qui restent assez intensives en termes de calculs.

Troisièmement, nos résultats montrent que la qualité de prédiction au point s'améliore significativement lorsque les résultats sont agrégés au niveau des carreaux. Ceci peut poser la question sur la pertinence de l'estimation du modèle d'usage des sols au niveau des points Teruti-Lucas. L'utilisation du score de Brier pondéré permet toutefois de montrer qu'une agrégation au niveau des segments Teruti-Lucas est suffisante pour améliorer les résultats et que des niveaux d'agrégation supérieurs ne sont pas nécessaires. Une fois défini un niveau d'agrégation adéquat, on pourrait aussi chercher à estimer un modèle agrégé de type *land use share* [Chakir and Le Gallo, 2013]. La variable d'intérêt devient alors la proportion de chaque usage de sol dans l'unité géographique agrégée, variable de type donnée de composition (CoDa) dont le traitement nécessite une méthodologie adaptée [Aitchison, 2003].

Enfin, concernant les implications en termes de politiques économiques des usages des sols, aucune législation claire n'existe à ce jour pour la protection des sols en France. Ainsi, le sol est souvent mentionné indirectement à travers des textes sur la réglementation de l'urbanisation (loi ALUR), l'évaluation des incidences environnementales (les lois Grenelle 1 et 2) ou les réformes des aides agricoles (La loi de modernisation de l'agriculture). Ces différentes politiques sont à prendre en compte dans la modélisation des usages des sols mais leurs impacts ne sont pas facilement dissociables des politiques locales en matière de zonage ou du contexte institutionnel. Pour ce dernier élément, les travaux néo-institutionnalistes de Williamson [1975, 1985] soulignent l'importance des coûts de transaction comme une explication à l'existence de formes alternatives d'organisation des usages des sols au niveau local. La prise en compte des ces facteurs locaux (zonages, contexte institutionnel) pour expliquer les usages des sols nécessite d'avoir des données riches et précises à une résolution très fine.

Dans ce contexte, le développement d'outils de modélisation, tels que ceux proposés dans ce papier, présente une première étape de prédiction des usages des sols à partir de variables facilement accessibles. Cette première étape méthodologique ouvre de nouvelles voies de recherche pour examiner les implications des usages des sols dans le domaine de l'environnement (biodiversité, qualité de l'eau, qualité du sol, qualité de l'air) ou de l'aménagement du territoire (analyse de l'artificialisation des sols, extension des villes, suivi des espaces protégés).

Remerciements
Ce travail a été partiellement financé par l'Agence Nationale de la Recherche à travers le projet ModULand (ANR-11-BSH1-005). Nous remercions le Service de la Statistique et de la Prospective du Ministère de l'Agriculture pour avoir mis à notre disposition les données de l'enquête Teruti-Lucas dans le cadre du projet ANR ModULand, l'Unité INFOSOL de l'INRA pour les données pédologiques (BDGSF) et le Joint Research Center-MARS de la Commission Européenne pour les données météorologiques (Interpolated Meteorological Data Base). Nous remercions également Antoine Lacroix pour le travail préparatoire accompli lors de son stage de M1.

Références

Aitchison, J. (2003). A concise guide to compositional data analysis. In *2nd Compositional Data Analysis Workshop*.

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on* 19 : 716–723.

Breiman, L., Friedman, J., Stone, C. and Olshen, R. (1984). *Classification and Regression Trees*. The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review* 78 : 1–3.

Buja, A., Stuetzle, W. and Shen, Y. (2005). Loss functions for binary class probability estimation and classification : Structure and applications. *Working draft, November* .

Chakir, R. (2015). L'espace dans les modèles économétriques d'utilisation des sols : enjeux méthodologiques et applications empiriques. *Revue d'Économie Régionale & Urbaine* 1&2 : 59–82.

Chakir, R. and Le Gallo, J. (2013). Predicting land use allocation in france : A spatial panel data analysis. *Ecological Economics* 92 : 114–125.

Chakir, R. and Parent, O. (2009). Determinants of land use changes : A spatial multinomial probit approach. *Papers in Regional Science* 88 : 327–344.

Chomitz, K. M. and Gray, D. A. (1996). Roads, land use, and deforestation : a spatial model applied in belize. *World Bank Economic Review* 10 : 487–512.

Domencich, T. A. and McFadden, D. (1975). *Urban Travel Demand-A Behavioral Analysis*. North-Holland Publishing Co.

Ferdous, N. and Bhat, C. (2013). A spatial panel ordered-response model with application to the analysis of urban land-use development intensity patterns. *Journal of Geographical Systems* 15 : 1–29.

Hair, J. (2010). *Multivariate data analysis : A global perspective*. Global Edition. Pearson Education.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer.

Irwin, E. G. and Geoghegan, J. (2001). Theory, data, methods : developing spatially explicit economic models of land use change. *Agriculture, Ecosystems & Environment* 85 : 7–23.

Irwin, E. G. and Wrenn, D. H. (2014). *An assessment of empirical methods for modeling land use. Dans Duke, J. M. et Wu J. The Oxford Handbook of Land Economics*. New York : Oxford University Press. 327–351.

Jobson, J. (1999). *Applied multivariate data analysis : regression and experimental design*. Springer Texts in Statistics. Springer New York.

Lambin, E. F., Rounsevell, M. and Geist, H. (2000). Are agricultural land-use models able to predict changes in land-use intensity ? *Agriculture, Ecosystems & Environment* 82 : 321–331.

Lubowski, R. N. (2002). Determinants of land-use transitions in the United States : Econometric analysis of changes among the major land-use categories. Ph.D. thesis, Harvard University, Cambridge.

McFadden, D. (1974). *Conditional logit analysis of qualitative choice behavior. Dans Zarembka P. Frontiers in Econometrics*. New York : Academic Press.

McMillen, D. P. (1989). An empirical model of urban fringe land use. *Land Economics* : 138–145.

Meier, L., Van De Geer, S. and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)* 70 : 53–71.

Merkle, E. C. and Steyvers, M. (2013). Choosing a strictly proper scoring rule. *Decision Analysis* 10 : 292–304.

Munroe, D. K. and Müller, D. (2007). Issues in spatially explicit statistical land-use/cover change (lucc) models : Examples from western honduras and the central highlands of vietnam. *Land use policy* 24 : 521–530.

Munroe, D. K., Southworth, J. and Tucker, C. M. (2004). Modeling spatially and temporally complex land-cover change : The case of western honduras. *The Professional Geographer* 56 : 544–559.

Nelson, G. C. and Hellerstein, D. (1997). Do roads cause deforestation ? using satellite images in econometric analysis of land use. *American Journal of Agricultural Economics* : 80–88.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* 6 : 461–464.

Shmueli, G. (2010). To explain or to predict ? *Statistical science* : 289–310.

Sidharthan, R. and Bhat, C. R. (2012). Incorporating spatial dynamics and temporal dependency in land use change models. *Geographical Analysis* 44 : 321–349, p. 321–349.

Team, R. C. (2014). R : A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Therneau, T., Atkinson, B. and Ripley, B. (2014). rpart : Recursive partitioning and regression trees. R package version 4.1-8.

Train, K. E. (2009). *Discrete Choice Methods with Simulation*. Cambridge Books. Cambridge University Press.

Tufféry, S. (2010). *Data mining et statistique décisionnelle : L'intelligence des données*. Editions Technip.

Tutz, G., Pößnecker, W. and Uhlmann, L. (2015). Variable selection in general multinomial logit models. *Computational Statistics & Data Analysis* 82 : 207 – 222.

Veldkamp, A. and Lambin, E. F. (2001). Predicting land-use change. *Agriculture, ecosystems & environment* 85 : 1–6.

Verburg, P. H., Schot, P. P., Dijst, M. J. and Veldkamp, A. (2004). Land use change modelling : current practice and research priorities. *GeoJournal* 61 : 309–324.

Williamson, O. E. (1975). *Markets and hierarchies*.

Williamson, O. E. (1985). *The economic intstitutions of capitalism*. Simon and Schuster.

Winkler, R. L., Munoz, J., Cervera, J. L., Bernardo, J. M., Blattenberger, G., Kadane, J. B., Lindley, D. V., Murphy, A. H., Oliver, R. M. and Ríos-Insua, D. (1996). Scoring rules and the evaluation of probabilities. *Test* 5 : 1–60.

Zhiyu, W. and Hasan, A. (2014). mnlogit : Multinomial logit model. R package version 1.1.1.

6 Annexes

6.1 Variable à expliquer

TABLE 7: Correspondance entre la nomenclature de l'usage des sols en 57 postes (variable *NPHYS*) et le regroupement en cinq catégories utilisé ici (variable *landuse*)

landuse	NPHYS	Libellé
1	11100	Volumes construits bas
1	11200	Volumes construits hauts
1	12100	Sols de forme aréolaire revêtus ou stabilisés
1	12200	Sols de forme linéaire revêtus ou stabilisés
1	13000	Sols enherbés artificialisés
1	14200	Sols nus artificialisés
2	11300	Serres et abris hauts
2	13100	Sols enherbés liés à la production agricole, hors élevage
2	14100	Sols nus liés à une activité agricole
2	21100	Blé tendre et épeautre
2	21200	Blé dur
2	21300	Orge et escourgeon
2	21500	Avoine
2	21600	Maïs
2	21820	Triticale
2	21900	Autres céréales
2	22100	Pomme de terre
2	22200	Betterave industrielle
2	22300	Autres racines et tubercules
2	23110	Tournesol
2	23120	Colza et navette
2	23190	Autres cultures industrielles oléagineuses
2	23200	Cultures industrielles textiles
2	23320	Pois sec protéagineux et pois fourrager
2	23330	Fèves et fèveroles
2	23390	Autres cultures industrielles annuelles
2	24100	Légumes
2	24300	Fleurs, plantes ornementales et pépinières toutes espèces
2	26000	Jachère
2	27500	Vigne
2	27100	Pommiers
2	27200	Autres fruitiers
2	27900	Autres cultures permanentes
2	28000	Jardins familiaux
3	31100	Forêts de feuillus
3	31200	Forêts de résineux
3	31300	Forêts mixtes
3	31400	Peupleraies en plein
3	32000	Bosquets
3	33000	Haies et alignements d'arbres
3	34000	Sols boisés à peuplement indéterminé (coupe rase)
4	25100	Fourrages annuels

landuse	NPHYS	Libellé
4	25200	Prairies temporaires semées essentiellement de graminées
4	25300	Prairies temporaires semées essentiellement de légumineuses
4	25400	Prairies permanentes productives
4	25500	Prairies permanentes peu productives
4	25600	Alpages
5	40000	Landes, friches, maquis, garrigues, savanes
5	41000	Superficies enherbées naturelles
5	60100	Dune, plage
5	60200	Rochers, éboulis
5	60300	Sols nus naturels
5	70100	Eaux intérieures
5	70200	Plans d'eau côtiers
5	70300	Glaciers, neiges éternelles
5	70400	Zones humides
.	88888	Hors territoire
.	99999	Zones interdites, photos non interprétées

6.2 Compléments sur les variables explicatives

Occupation des sols : Corine Land Cover 2006 Nom de l'enquête : Corine Land Cover
Promoteur : Agence européenne pour l'environnement, en France : Service de l'observation et des statistiques du ministère chargé de l'environnement
Disponibilité des données : gratuit [6] après remplissage d'un formulaire
Echelle : 1 / 100 000

La base CORINE Land Cover est une base de données européenne d'occupation biophysique des sols, produite dans le cadre du programme européen CORINE de coordination de l'information sur l'environnement. Le producteur pour la France est le Service de l'observation et des statistiques du ministère chargé de l'environnement. Les données sont issues de l'interprétation visuelle d'images satellitaires, avec des données complémentaires d'appui (en particulier les BD ORTHO et CARTO de l'IGN). L'occupation biophysique du sol prévaut à son utilisation, ainsi la nature des objets (forêts, cultures, surfaces en eau, roches affleurantes...) est privilégiée par rapport à leur fonction socio-économique (agriculture, habitat...).

« L'unité spatiale au sens de CORINE Land Cover est une zone dont la couverture peut être considérée comme homogène, ou être perçue comme une combinaison de zones élémentaires qui représente une structure d'occupation. La surface de la plus petite unité cartographiée (seuil de description) est de 25 ha. » Les polygones constituant les zones CLC ont des aires comprises entre 0,25 et 3 857 km^2 en Midi-Pyrénées, avec une moyenne de 2,6 km^2. Pour plus d'informations se reporter au site Internet : `http://www.statistiques.developpement-durable.gouv.fr/donnees-ligne/t/methode-production-base-donnees.html?tx_ttnews%5Btt_news%5D=11268&cHash=88595af0806f46f2c8901fd438ea809f`

Dans cette analyse, les codages aux niveaux 1 et 2 (respectivement en 5 et 15 catégories) de cette variable sont utilisés (cf. tableau 8).

6. `http://www.statistiques.developpement-durable.gouv.fr/donnees-ligne/li/1825/1097/occupation-sols-corine-land-cover.html`

Table 8 – Codage de la variable d'occupation des sols, Corine Land Cover, niveaux 1 et 2

CLC1	CLC2
1 Territoires artificialisés	11 Zones urbanisées
	12 Zones industrielles ou commerciales et réseaux de communication
	13 Mines, décharges et chantiers
	14 Espaces verts artificialisés, non agricoles
2 Territoires agricoles	21 Terres arables
	22 Cultures permanentes
	23 Prairies
	24 Zones agricoles hétérogènes
3 Forêts	31 Forêts et milieux semi-naturels
	32 Milieux à végétation arbustive et/ou herbacée
	33 Espaces ouverts, sans ou avec peu de végétation
4 Zones humides	41 Zones humides intérieures
	42 Zones hum. maritimes
5 Surfaces en eau	51 Eaux continentales
	52 Eaux maritimes

Nature du sol : Base de données géographique des sols de France (BDGSF) Nom de l'enquête : Base de données géographique des sols de France
Promoteur : INRA, Unité INFOSOL, Orléans
Année : 1998
Disponibilité des données : disponible sur demande pour un an (sur CD), 50 euros
Echelle : 1 / 1 000 000

« La BDGSF est une représentation simplifiée de la diversité spatiale de la couverture de sol. La méthodologie utilisée pour différencier et nommer les principaux types de sol est basée sur la terminologie de la légende de la carte des sols du monde établie en 1974 par la FAO à l'échelle du 1 / 5 000 000. Cette terminologie est basée sur la distinction des processus pédologiques responsables de la différenciation des sols, c'est-à-dire la brunification, le lessivage, la podzolisation, l'hydromorphie, etc. Elle a été revue et adaptée pour prendre en compte les spécificités des paysages français. » Source : site Internet de la BDGSF `http://www.gissol.fr/programme/bdgsf/bdgsf.php`.

Les Unités Cartographiques de Sols (UCS) sont des regroupements d'Unités Typologiques de Sols (UTS) qui sont localisables dans l'espace tandis qu'il n'est pas possible de localiser et de délimiter toutes les UTS à l'échelle de travail. Les polygones constituant les UCS ont des aires comprises entre 1,8 et 4 369 km^2 en Midi-Pyrénées. Pour plus d'informations se reporter au site Internet : `http://www.gissol.fr/programme/bdgsf/contenu.php`

Climat : European Climate Database site Internet : `http://www.marsop.info` `http://marswiki.jrc.ec.europa.eu/agri4castwiki/index.php/Meteorological_data_from_ground_stations`

Nom de l'enquête : Agri4cast Interpolated Meteorological database
Promoteur : JRC-MARS (Joint Research Center, Monitoring Agricultural ResourceS (MARS) Unit, European Commission)
Années : 1975 à 2012
Disponibilité des données : enregistrement requis

Nouvelles Régions Agricoles Les Nouvelles Régions Agricoles sont des regroupements de petites régions agricoles (PRA) effectués par l'Agreste. Les régions agricoles (RA) et PRA avaient été définies en 1946 pour mettre en évidence des zones agricoles homogènes. Une RA couvre un nombre entier de communes formant une zone d'agriculture homogène ; une PRA est constituée par le croisement du département et de la RA. Dans la région Midi-Pyrénées, il y a 61 PRA et 32 NRA. L'aire des NRA est comprise entre 302,2 et 4 284 km^2, avec une moyenne de 1 426 km^2.

TABLE 9 – Statistiques descriptives de l'aire (en km^2) des polygones des différentes bases

	min	Q1	médiane	moyenne	Q3	max
CLC	0,25	0,4	0,7	2,6	1,5	3 857,0
UCS (sol)	1,8	13,6	39,9	199,8	170,9	4 369,0
NRA	302,2	779,9	1 138,0	1 426,0	1 964,0	4 284,0
communes	0,3	6,0	10,4	15,1	18,2	169,7

TABLE 10 – Répartition de la variable Corine Land Cover niveau 1

code	libellé	effectif	fréquence (%)
1	territoires artificialisés	652	2,6
2	territoires agricoles	15 385	60,8
3	forêts et milieux semi-naturels	9 148	36,2
4	zones humides	4	0,0
5	surfaces en eau	104	0,4

TABLE 11 – Répartition de la variable Corine Land Cover niveau 2

code	libellé	effectif	fréquence (%)
11	Zones urbanisées	446	1,8
12	Zones industrielles ou commerciales et réseaux de communication	122	0,5
13	Mines, décharges et chantiers	44	0,2
14	Espaces verts artificialisés, non agricoles	40	0,2
21	Terres arables	6 136	24,3
22	Cultures permanentes	388	1,5
23	Prairies	3 101	12,3
24	Zones agricoles hétérogènes	5 760	22,8
31	Forêts	6 710	26,5
32	végétation arbustive et/ou herbacée	1 800	7,1
33	Espaces ouverts, sans ou avec peu de végétation	638	2,5
41	Zones humides intérieures	4	0,0
51	Eaux continentales	104	0,4

TABLE 12 – Répartition de la variable grandpole

code	libellé	effectif	fréquence (%)
0	non	24 066	95,1
1	oui	1 227	4,9

TABLE 13 – Répartition de la variable texture du sol

code	libellé	effectif	fréquence (%)
0	Pas d'information	248	1,0
1	Grossière (argile<18% et sable>65%)	3 653	14,4
2	Moyenne (18%<argile<35% et sable>15%, ou argile<18% et 15%<sable<65%)	13 642	53,9
3	Modérément fine (argile<35% et sable<15%)	1 843	7,3
4	Fine (35%<argile<60%)	5 907	23,4

TABLE 14 – Répartition de la variable Matériau de base

code	libellé	effectif	fréquence (%)
0	Pas d'information	248	1,0
1	Dépôts alluviaux non-différenciés (ou dépôts glaciaires)	5 784	22,9
2	Roches calcaires	6 541	25,9
3	Matière argileuse	724	2,9
4	Matière sableuse	1 031	4,1
5	Matière limoneuse	177	0,7
6	Formations détritiques	4 071	16,1
7	Roches cristallines et migmatites	6 533	25,8
8	Roches volcaniques	184	0,7

TABLE 15 – Statistiques descriptives des variables explicatives quantitatives

	moyenne	écart-type	Q1	médiane	Q3	min	max
altitude (m)	492,4	467,1	198,1	317,3	608,3	59,2	3014,3
température minimum (°C)	-9,7	2,3	-10,8	-9,4	-7,8	-15,6	-6,6
température maximum (°C)	37,9	2,6	36,6	38,6	40,0	28,8	41,1
température moyenne (°C)	11,5	1,8	10,7	12,1	12,7	4,7	13,4
précipitations (mm)	759,2	160,6	611,5	715,6	793,9	537,1	1279,8
vent moyen (km/h)	10,2	2,8	8,2	9,6	12,4	3,3	16,9
densité de pop. (hab./km^2)	64,1	156,7	13,2	23,1	52,3	0,3	2807,2
part de cadres (%)	4,8	4,2	2,0	4,1	6,7	0,0	33,3
part d'agriculteurs (%)	5,8	6,1	1,2	4,0	8,5	0,0	66,7
prix des terres (€ courant/ha)	5273,5	1003,4	4510,0	5330,0	6270,0	3630,0	7310,0
altitude des voisins (m)	492,4	465,2	197,4	318,4	608,0	60,5	2866,3

6.3 Preuve de l'équation (5) page 8

Montrons que :

$$\mathrm{B}_G = \mathrm{B}_J + \frac{1}{2n} \sum_{k \in I_C} \sum_{j \in I_J} \sum_{g \in I_{G_j}} \#G_g \Big(\bar{z}_{gk} - \bar{\hat{p}}_{gk} - \bar{z}_{jk} + \bar{\hat{p}}_{jk} \Big)^2$$

Rappelons que :

$$\mathrm{B}_G = \frac{1}{2n} \sum_{k=1}^{K} \sum_{g \in I_G} \#G_g \Big(\bar{z}_{gk} - \bar{\hat{p}}_{gk} \Big)^2$$

avec

$$\bar{z}_{gk} = \frac{1}{\#G_g} \sum_{i \in G_g} z_{ik} = \text{fréquence observée de l'usage } k \text{ dans le groupe } G_g$$

$$\bar{\hat{p}}_{gk} = \frac{1}{\#G_g} \sum_{i \in G_g} \hat{p}_{ik} = \text{probabilité estimée de l'usage } k \text{ dans le groupe } G_g,$$

les I_{G_j} ($j \in I_J$) forment une partition de I_G et $J_j = \cup_{g \in I_{G_j}} G_g$ avec $j \in I_J$,
$\bar{z}_{jk}$ est la fréquence observée de l'usage k dans le groupe J_j :

$$\bar{z}_{jk} = \frac{1}{\#J_j} \sum_{i \in J_j} z_{ik} = \frac{1}{\#J_j} \sum_{i \in \cup_{g \in I_{G_j}} G_g} z_{ik} = \frac{1}{\#J_j} \sum_{g \in I_{G_j}} \sum_{i \in G_g} z_{ik} = \frac{1}{\#J_j} \sum_{g \in I_{G_j}} \#G_g \ \bar{z}_{gk}$$

$\bar{\hat{p}}_{jk}$ est la probabilité estimée de l'usage k dans le groupe J_j :

$$\bar{\hat{p}}_{jk} = \frac{1}{\#J_j} \sum_{i \in J_j} \hat{p}_{ik} = \frac{1}{\#J_j} \sum_{g \in I_{G_j}} \#G_g \ \bar{\hat{p}}_{gk}$$

Notons aussi que $\sum_{g \in I_{G_j}} \#G_g = \#J_j$.

Nous avons :

$$\begin{aligned}
\mathrm{B}_G =& \frac{1}{2n} \sum_{k \in I_C} \sum_{g \in G_g} \#G_g (\bar{z}_{gk} - \bar{\hat{p}}_{gk})^2 \\
=& \frac{1}{2n} \sum_{k \in I_C} \sum_{j \in I_J} \sum_{g \in I_{G_j}} \#G_g (\bar{z}_{gk} - \bar{\hat{p}}_{gk})^2 \\
=& \frac{1}{2n} \sum_{k \in I_C} \sum_{j \in I_J} \sum_{g \in I_{G_j}} \#G_g (\bar{z}_{gk} - \bar{\hat{p}}_{gk} - \bar{z}_{jk} + \bar{\hat{p}}_{jk} + \bar{z}_{jk} - \bar{\hat{p}}_{jk})^2 \\
=& \frac{1}{2n} \sum_{k \in I_C} \sum_{j \in I_J} \sum_{g \in I_{G_j}} \#G_g (\bar{z}_{gk} - \bar{\hat{p}}_{gk} - \bar{z}_{jk} + \bar{\hat{p}}_{jk})^2 + \frac{1}{2n} \sum_{k \in I_C} \sum_{j \in I_J} \sum_{g \in I_{G_j}} \#G_g (\bar{z}_{jk} - \bar{\hat{p}}_{jk})^2 \\
& + \frac{1}{n} \sum_{k \in I_C} \sum_{j \in I_J} \sum_{g \in I_{G_j}} \#G_g (\bar{z}_{gk} - \bar{\hat{p}}_{gk} - \bar{z}_{jk} + \bar{\hat{p}}_{jk})(\bar{z}_{jk} - \bar{\hat{p}}_{jk})
\end{aligned}$$

$$\begin{aligned}
=& \frac{1}{2n} \sum_{k \in I_C} \sum_{j \in I_J} \sum_{g \in I_{G_j}} \#G_g (\bar{z}_{gk} - \bar{\hat{p}}_{gk} - \bar{z}_{jk} + \bar{\hat{p}}_{jk})^2 + \frac{1}{2n} \sum_{k \in I_C} \sum_{j \in I_J} \#J_j (\bar{z}_{jk} - \bar{\hat{p}}_{jk})^2 \\
& + \frac{1}{n} \sum_{k \in I_C} \sum_{j \in I_J} (\bar{z}_{jk} - \bar{\hat{p}}_{jk}) \sum_{g \in I_{G_j}} \#G_g (\bar{z}_{gk} - \bar{\hat{p}}_{gk} - \frac{1}{\#J_j} \sum_{h \in I_{Gj}} \#G_h \ \bar{z}_{hk} + \frac{1}{\#J_j} \sum_{h \in I_{Gj}} \#G_h \ \bar{\hat{p}}_{hk}) \\
=& \frac{1}{2n} \sum_{k \in I_C} \sum_{j \in I_J} \sum_{g \in I_{G_j}} \#G_g (\bar{z}_{gk} - \bar{\hat{p}}_{gk} - \bar{z}_{jk} + \bar{\hat{p}}_{jk})^2 + \frac{1}{2n} \sum_{k \in I_C} \sum_{j \in I_J} \#J_j (\bar{z}_{jk} - \bar{\hat{p}}_{jk})^2 \\
& + \frac{1}{n} \sum_{k \in I_C} \sum_{j \in I_J} (\bar{z}_{jk} - \bar{\hat{p}}_{jk}) \times \left[\sum_{g \in I_{G_j}} \#G_g \ \bar{z}_{gk} - \sum_{g \in I_{G_j}} \#G_g \ \bar{\hat{p}}_{gk} - \sum_{h \in I_{Gj}} \#G_h \ \bar{z}_{hk} + \sum_{h \in I_{Gj}} \#G_h \ \bar{\hat{p}}_{hk} \right] \\
=& \frac{1}{2n} \sum_{k \in I_C} \sum_{j \in I_J} \sum_{g \in I_{G_j}} \#G_g (\bar{z}_{gk} - \bar{\hat{p}}_{gk} - \bar{z}_{jk} + \bar{\hat{p}}_{jk})^2 + \frac{1}{2n} \sum_{k \in I_C} \sum_{j \in I_J} \#J_j (\bar{z}_{jk} - \bar{\hat{p}}_{jk})^2
\end{aligned}$$

6.4 Tableaux des coefficients des MNL

Niveau de significativité des coefficients : *** : 0,001 ** : 0,01 * : 0,05.

Table 16 – Coefficients du modèle MNL-E

variable	usage agricole	forêts	prairies	sols naturels
constante	-4,089	-1,831	-5,693**	-2,808
CLC2-12	-1,267	-1,144	-1,788	0,506
CLC2-13	0,909	0,762	0,123	2,257***
CLC2-14	-18,33	0,142	-20,700	-2,143
CLC2-21	4,359***	2,223***	3,271***	1,505***
CLC2-22	3,882***	2,050***	2,479***	1,672***
CLC2-23	2,871***	3,096***	4,124***	2,063***
CLC2-24	3,155***	2,562***	3,362***	1,47***
CLC2-31	2,300***	5,459***	2,827***	2,822***
CLC2-32	1,989***	3,488***	3,605***	4,018***
CLC2-33	-13,350	1,820	1,292	3,912***
CLC2-41	2,104	20,966	19,498	20,055
CLC2-51	-15,465	3,969***	1,968	5,175***
altitude	0,0003	0,002***	0,003***	0,002***
prixterre	0,0001**	-0,0002***	0,0000	-0,0003***
densite	-0,0003	-0,0004	-0,001**	-0,001*
tempmini	0,178***	0,064	-0,095*	0,041
tempmaxi	0,161**	0,099*	0,071	0,162**
tempmoy	-0,266*	-0,215*	-0,023	-0,247*
pluie	-0,001	-0,001*	-0,001	-0,002***
TEXT-1	-0,08	0,497	-0,092	-0,246
TEXT-2	0,129	0,183	-0,233	-0,368
TEXT-3	0,413	0,406	0,100	0,178
TEXT-4	0,285	0,204	-0,075	-0,174

Table 17 – Coefficients du modèle MNL-Cm

variable	usage agricole	forêts	prairies	sols naturels
constante	-10,667***	-3,228	-7,607***	-7,356**
CLC2-12	-1,261	-1,173	-1,814	0,479
CLC2-13	0,743	0,702	0,036	2,122***
CLC2-14	-18,128	0,304	-20,643	-1,535
CLC2-21	4,461***	2,24***	3,295***	1,625***
CLC2-22	3,989***	2,152***	2,46***	1,899***
CLC2-23	2,938***	3,13***	4,125***	2,072***
CLC2-24	3,209***	2,62***	3,368***	1,542***
CLC2-31	2,357***	5,528***	2,821***	2,867***
CLC2-32	2,067***	3,604***	3,634***	4,177***
CLC2-33	-12,12	1,884	1,474	4,269***
CLC2-41	2,535	20,978	19,6	20,19
CLC2-51	-15,422	4,006***	2,001	5,262***
altitude	0,005*	0,004**	0,006***	0,006***
tempmini	0,067*	0,045	-0,09**	0,009
ventmoy	0,084***	-0,023	0,006	0,001
tempmaxi	0,209***	0,028	0,094*	0,115*
altitude_lag	-0,005*	-0,002	-0,004*	-0,003

TABLE 18 – Coefficients du modèle MNL-S

variable	usage agricole	forêts	prairies	sols naturels
constante	-1,814***	-2,688***	-3,035***	-2,813***
CLC2-12	-1,259	-1,123	-1,782	0,542
CLC2-13	0,767	0,79	0,092	2,222***
CLC2-14	-18,040	0,495	-20,440	-1,335
CLC2-21	4,513***	2,307***	3,323***	1,718***
CLC2-22	3,967***	2,246***	2,495***	2,024***
CLC2-23	2,848***	3,141***	4,221***	2,089***
CLC2-24	3,165***	2,643***	3,464***	1,586***
CLC2-31	2,271***	5,549***	2,892***	2,890***
CLC2-32	1,945***	3,661***	3,681***	4,225***
CLC2-33	-11,510	2,235*	1,535	4,560***
CLC2-41	2,612	21,006	19,870	20,264
CLC2-51	-15,582	3,989***	2,085*	5,252***
altitude	-0,001***	0,002***	0,003***	0,002***