Exploring land use prediction errors from area frame survey data

Raja Chakir
Economie Publique, INRA, AgroParisTech, Université Paris-Saclay, France
Thibault Laurent
Toulouse School of Economics, CNRS, University of Toulouse, France
Anne Ruiz-Gazen
Toulouse School of Economics, University of Toulouse Capitole, Toulouse, France
Christine Thomas-Agnan
Toulouse School of Economics, University of Toulouse Capitole, Toulouse, France
Céline Vignes
Toulouse School of Economics, CNRS, University of Toulouse, France

We consider the problem of areal level land use classification from the information provided by point level databases such as the area frame surveys (American NRI survey, EUROSTAT Lucas survey, French Teruti-Lucas survey) and easily accessible covariates. An exploratory analysis emphasizes the link between the areal level prediction error and a measure of difficulty of prediction given by the Gini-Simpson impurity index. We provide a methodology and an R code for allowing to explore the quality of an areal frame survey by generating synthetic data.

Keywords*: American NRI survey, classification tree, Gini-Simpson impurity index, land use/cover models, Teruti-Lucas survey.*

1. Introduction

In order to gather information about land use, many countries, or groups of countries devise area frame surveys. This is the case for example for the United States Department of Agriculture NRI database[1], for the EUROSTAT LUCAS survey[2] and for the French Teruti-Lucas database[3]. They consist in lists of parcels where observation of land use is performed. One common feature of these surveys is that the distribution of the sample locations is sparse and does not fill regularly the space. From this point level information, one usually tries to derive areal level predictions for a partition of the space which may correspond to administrative units or to regular meshes. Several solutions for this process are discussed in Chakir et al. (2016b).

[1] https://www.nass.usda.gov/Publications/Methodology_and_Data_Quality/Advanced_Topics/AREA%20FRAME%20DESIGN.pdf
[2] http://ec.europa.eu/eurostat/web/lucas/overview
[3] http://agreste.agriculture.gouv.fr/enquetes/territoire-prix-des-terres/teruti-lucas-utilisation-du/

This classification problem usually involves three steps: the choice of a point level classification model, the estimation of posterior probabilities for each category of land use at the areal level and the prediction of areal land use from these estimated probabilities. In Chakir et al. (2016a, 2017) a classification model is constructed to predict land use at point level in five categories (urban, farming, forests, pastures and natural land) with the Teruti-Lucas database using easily accessible covariates in the Midi-Pyrénées region. Synthetic data sets are later on simulated with this data driven model in Chakir et al. (2016b) in order to compare the prediction strategies at areal level and evaluate the quality of the area frame survey. In Chakir et al. (2016b), it is demonstrated with synthetic data sets that the prediction error at point level is essentially due to the mean distance between the observed land use and the corresponding true probability, later on called response error. The purpose of the present paper is to present a detailed exploratory analysis of the prediction errors for one of these synthetic data sets. A particular objective is to link the size of the prediction errors with the Gini-Simpson impurity index as a measure of the local difficulty of prediction and thus characterize situations where prediction is difficult. This link is explored in detail at the point level and at the areal level. By supplying our R code together with the synthetic data, another aim of this work is to provide some tools that could be used for evaluating the quality of an area frame survey design (using synthetic data) or comparing different classification methods in the framework of land use classification problems.

Section 2 presents the prediction error (mean distance between the observed and predicted land use) at point level and its decomposition into four terms. Our synthetic data set together with the data generating process are described in section 3. In Section 4, we analyze the distribution and spatial pattern of the Data Generating Process (DGP) probabilities as well as the corresponding Gini-Simpson indices. In Section 5, we first present descriptive statistics for the absolute and relative response error at point level and analyze their relationship with the corresponding Gini-Simpson indices. We then turn attention to the same questions but at areal level. We conclude in Section 6.

2. Point level prediction

The classification problem we consider consists in predicting the land use categorical variable U_i at location i (with K levels) based on the observation of this same categorical variable and a set of covariates at a given set of locations. The theoretical vector of probabilities to observe the different categories k ($k = 1, \cdots, K$) at location i is given by the vector $p_i = (p_{i1}, \cdots, p_{i5})$ where $p_{ik} = \mathbb{P}(U_i = k \mid X_i = x_i)$.
The model we get after fitting the classification tree to the initial data, $p_i = f(x_i)$, links the probability vectors of land use at location i, $p_i = (p_{i1}, \cdots, p_{i5})$, with a set x_i of covariates observed at location i. Let us define the risk $R(h)$ of a classification rule $h(X)$ to be the misclassification rate $R(h) = \mathbb{E}_{X,U}(1(h(X) \neq U)$. Given fitted probabilities $\hat{p}_i$, one could envision predicting the land use categorical variable U_i at location i by random draw from a multinomial distribution with parameter $\hat{p}_i$. A classical result shows that this is not optimal in the following sense. If, for $k = 1$ to K, momentarily dropping the location index i, $p_k(x) = \mathbb{P}(U = k \mid X = x)$ is the conditional probability of observing land use k, the risk of a prediction rule $h(X)$ satisfies

$$1 - R(h) = \sum_{k=1}^{K} \mathbb{E}_{X,U}(1(h(X) = k)1(U = k)) \quad (1)$$

$$= \sum_{k=1}^{K} \mathbb{E}_X \mathbb{E}_{U|X}(1(h(X) = k)1(U = k)) \quad (2)$$

$$= \sum_{k=1}^{K} \mathbb{E}_X(1(h(X) = k)p_k(X)) \quad (3)$$

From this last expression, it is clear that the rule that minimizes this risk, classically called the Bayes classifier, is defined by $\hat{U}^* = h^*(x) = j^*$, for $j^* = arg\max_{k=1}^{K} p_k(x)$ which means that it is the rule that assigns the category for which the probability is maximum. It is then easy to

see that the corresponding risk is given by

$$R(h^*) = 1 - \mathbb{E}_X \left(\max_{k=1}^{K} p_k(X) \right).$$

For comparison, it is easy to establish that the risk of a multinomial random draw prediction $h^r(X)$ is given by

$$R(h^r) = 1 - \mathbb{E}_X \left(\sum_{k=1}^{K} p_k(X)^2 \right).$$

It is easy to check that $\sum_{k=1}^{K} p_k(X)^2 \leq \left(\max_{k=1}^{K} p_k(X)\right) \left(\sum_{k=1}^{K} p_k(X)\right) = \max_{k=1}^{K} p_k(X)$. Note that the second risk is related with the so-called Gini-Simpson impurity index classically used in classification trees (see Section 4).

For the error computation in the sequel, instead of using the U random variable for land use, we will encode it using K dummies as follows: $d_{ik} = 1$ if land use k ($k = 1, \ldots, K$) is obtained at location i and $d_{ik} = 0$ otherwise. The corresponding optimal prediction U^* is similarly encoded by $\hat{d}_{ik}$.
Because we will focus later on areal level predictions (rather than point level), we change our error criterion and replace the misclassification rate by a classical mean square error of prediction criterion $\mathbb{E}_X \sum_{i=1}^{n} (\hat{d}_{ik} - d_{ik})^2$. Chakir et al. (2016b) show that the point level prediction error for category k measured by $\mathbb{E}_X \sum_{i=1}^{n} (\hat{d}_{ik} - d_{ik})^2$ can be decomposed into four terms:

- $\mathbb{E}_X \sum_{i=1}^{n} (d_{ik} - p_i)^2$ which represents the mean distance between the observed land use and the corresponding true probability will be called the response error,
- $\mathbb{E}_X \sum_{i=1}^{n} (\hat{d}_{ik} - \hat{p}_i)^2$, the estimated response error, which represents the mean distance between the predicted land use and the corresponding estimated probability,
- $\mathbb{E}_X \sum_{i=1}^{n} (\hat{p}_i - p_i)^2$, the estimation error, which represents the mean distance between the estimated probability and the true probability (quality of the model fit),
- a remainder term.

It is shown that the dominant term is the response error which means that even if one had the knowledge of the true probabilities, there is this incompressible error due to the fact that we observe and predict something discrete using a continuous probability as parameter. This error is going to be the focus of our present study and we will not consider further the estimated probabilities.

3. Data

3.1. *The synthetic data set*

We consider the former Midi-Pyrénées region which was, before the French territorial reform in December 2015, the largest region in France with 8.3% of the whole territory and 3020 municipalities in January 2013. It is a quite rural region with 4.5% of the metropolitan population in 2011 but presents a diversity in land uses. Toulouse is the major urban center, there are large farming areas in the middle, the Pyrénées mountains in the South, and mainly pastures and forests in the North.

The initial point level locations are the Teruti-Lucas (T-L) locations and the areal level corresponds to T-L "segments", as detailed below. Teruti-Lucas is conducted each year since 1982 by the "Service de la Statistique et de la Perspective" of the French Ministry of Agriculture. It is the French part of the LUCAS (Land Use and Coverage Area frame Survey) survey conducted by Eurostat which gathers harmonized data on land use/cover in the European Union. Since 2005, the T-L survey contains information about the land use pattern on "segments" containing 10 points used to collect the data each year (see Figure 1).

Table 1: Data sources.

Name	Geographical level	Source	Year	Unit
CLC2	zones (>25 ha)	Corine Land Cover	2006	-
Altitude	grid (250 m)	BDAlti (IGN)	-	meters
Land and empty meadow price	32 NRA	Agreste	2010	actual euros/ha
Population density	grid (200 m)	Insee	2010	inhabitants/km^2

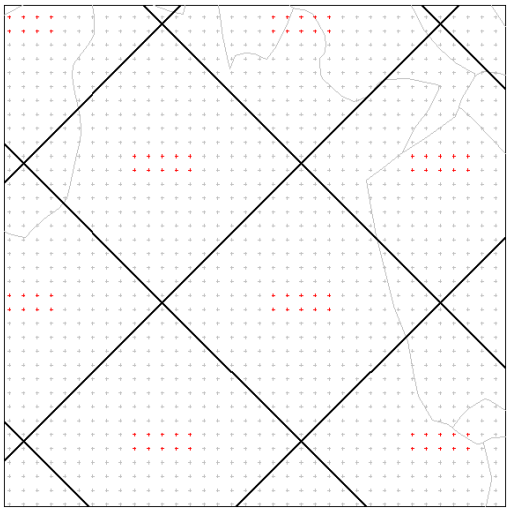

Figure 1: Areal level grid (in black) and points (Teruti-Lucas locations in red) in the Toulouse area.

Following the initial spacing of 300m between T-L points, we constructed a fine grid (point level) so that each "segment" contains 200 locations vs. 10 T-L locations, for a total of 502205 points (vs. 25317 T-L locations). The areal level is constructed so that each cell contains a unique T-L "segment" and so as to tile the Midi-Pyrénées territory. Its squares are centered at the barycenter of the 10 points of the T-L "segment" (see Figure 1) and their sides have a length of 4.2 kilometers. The areal level grid comprises 2579 such squares. Note that due to border effects, some areal units contain less than 200 locations, with a minimum of 32. The target variables d_{ik} we generate below give the land use at each location i as measured by the Teruti-Lucas "physical occupation" of the land and recoded in the following five categories: urban, farming, forests, pastures, natural land. More details about Teruti-Lucas can be found in Chakir et al. (2016b, 2017).

Covariates have been selected from free and easily accessible data bases (see Table 1) that are available at several different scales. We provide four variables: *CLC2*, *altitude*, *population density* and *land price* at each of the 502205 points. In Chakir et al. (2016b, 2017), more covariates were taken into account but they either require an online subscription (meteorological data) or are not freely available (soil composition data) and cannot be provided in the present paper.
From the four covariates, we generate the data set by using a classification tree obtained with the CART algorithm (Breiman et al., 1984) using the Gini-Simpson impurity index and a pruning step. The tree is given by Figure 2 (see Chakir et al., 2016b, 2017, for more details). Then we calculate and provide in the data base the probability vector p_i whose coordinates correspond to each of the 5 land uses (urban, farming, forests, pastures, natural land). Using a multinomial random draw at each of the 502205 locations i, with parameter p_i, we also generate and provide the values of the categorical variable U_i from which it is easy to derive the target variables d_{ik}, $k = 1, \ldots, 5$.

3.2. *The DGP probabilities*

Due to the nature of the model (classification tree), the DGP probabilities p_{ik} are discrete with 11 different values (see Figures 2, 3 and 4).
More precisely, the land use forests corresponds to one terminal node, urban, farming and natural land to two nodes and pastures to four nodes. Urban use and natural land are the

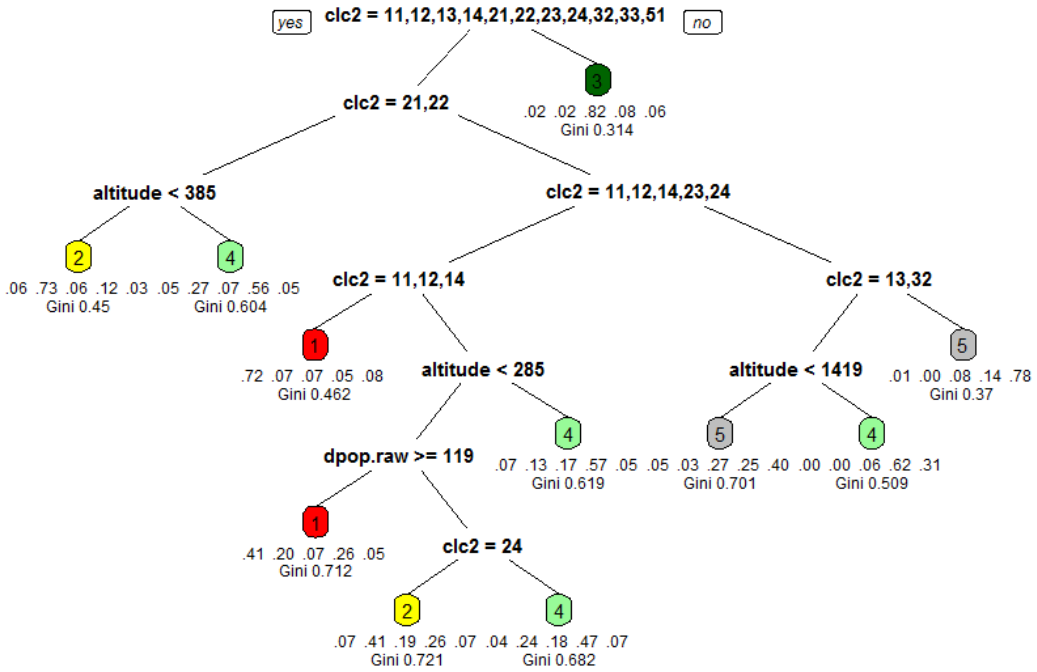

Figure 2: Classification tree chosen for the DGP.

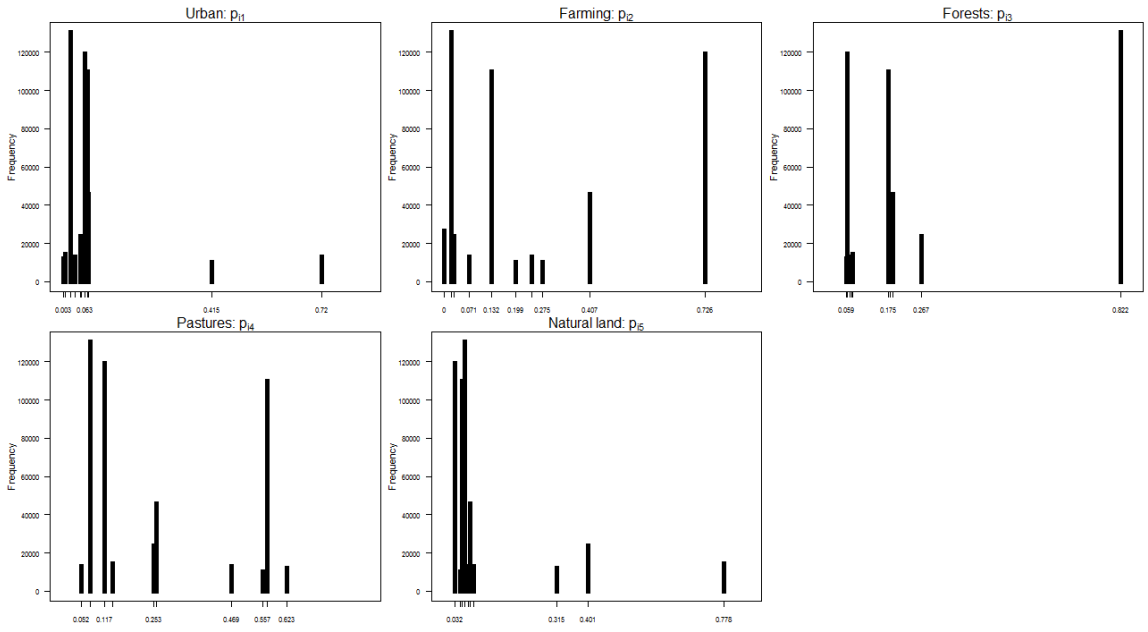

Figure 3: Bar charts of p_{ik}.

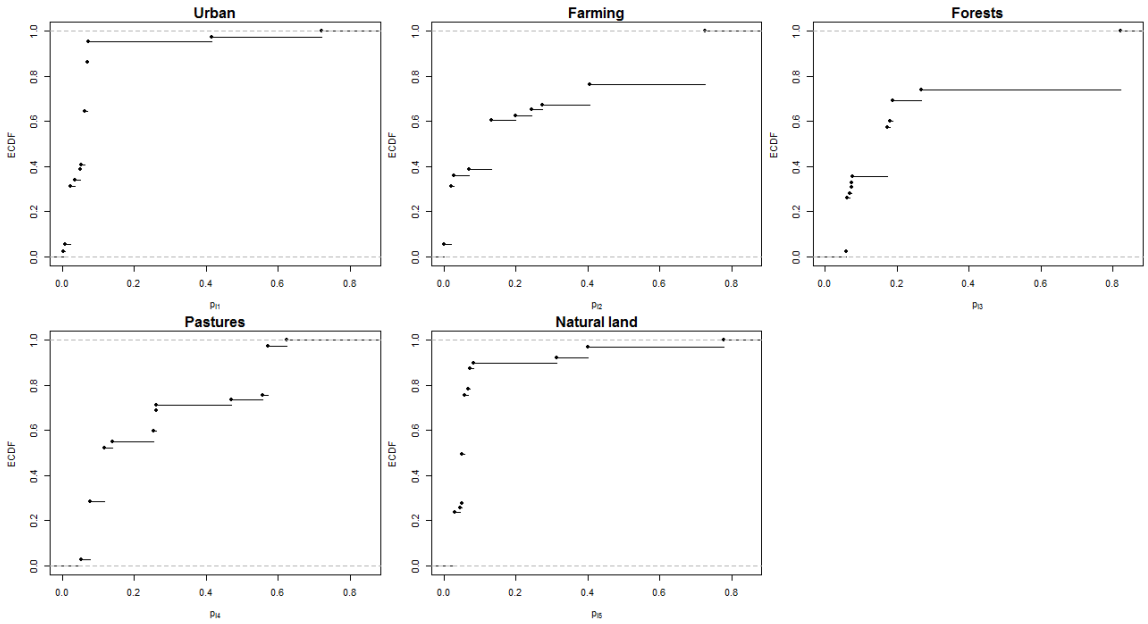

Figure 4: Empirical cumulative functions of p_{ik}.

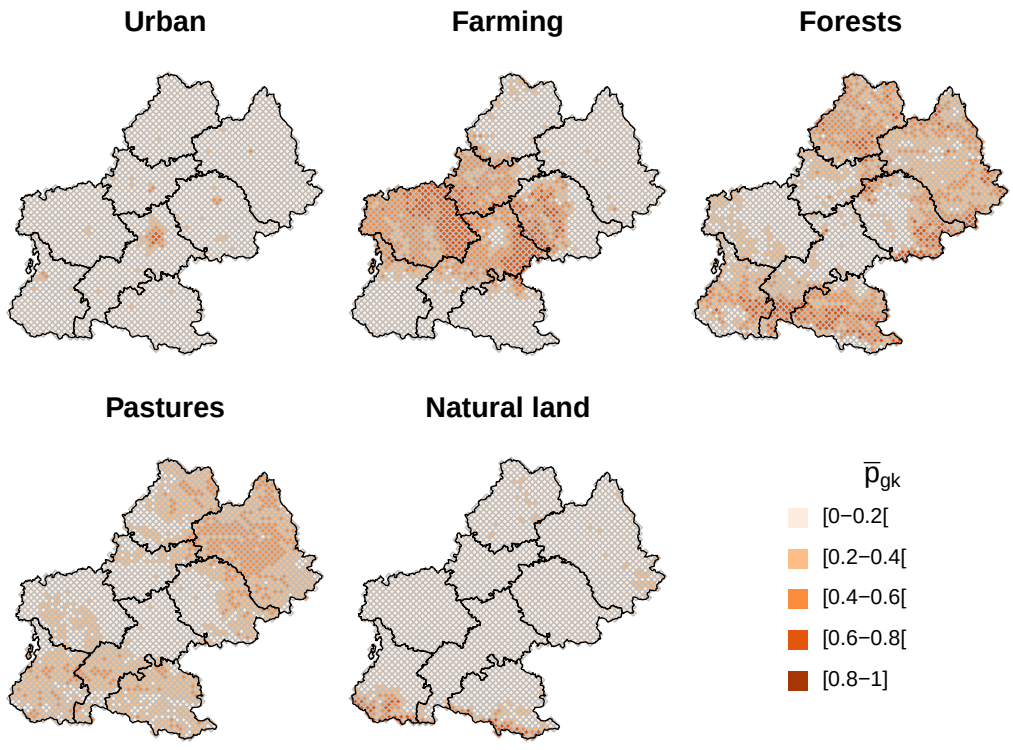

Figure 5: DGP probabilities $\bar{p}_{gk}$.

land uses with the lowest probabilities. Most of them are below 0.10 with 95.4% of the values less than 0.072 for p_{i1} and 90.0% of the values less than 0.084 for p_{i5}. Forests, farming and pastures have multimodal distributions with 23.7% of the values equal to 0.726 for farming, 26.0% of the values equal to 0.822 for forests and 21.9% of the values equal to 0.570 for pastures. It is not surprising to observe that only few points are concerned with high probabilities of urban and natural land uses while there exists a large number of points where the probability of forests and farming and to a lesser extent pastures are quite high.
The maps of Figure 5 show the spatial pattern of the aggregated land use probabilities $\bar{p}_{gk}$ at the areal level.

4. Impurity of the DGP probabilities: the Gini-Simpson index

For a population of individuals classified into K categories, impurity indices are used in many fields for example in ecology for measuring biodiversity, and in economics (Herfindahl index) to measure competition. In the land use classification problem, we want to measure how homogeneous or diverse land use is at a given point or in a given region. For a vector of probabilities $(p_1, \cdots, p_K)$, the Gini-Simpson impurity index (Simpson, 1949) is defined by $gs = \sum_{k=1}^{K} p_k(1-p_k) = 1 - \sum_{k=1}^{K} p_k^2$ where p_k is the probability of category k. The vector of probabilities is said to be pure if one probability is very high and all others are low, corresponding to a low gs index. It is said to be impure if all categories have similar probabilities, corresponding to a high gs index. Note that the related Gini-Simpson index $1 - gs$ is equal to the probability that two individuals taken at random from the data set of interest are of the same category. In statistics, the Gini-Simpson impurity index is also used for classification trees (e.g. Therneau et al., 2014) under the name of Gini impurity index (not to be confused with the Gini concentration index). It is always between 0 and 1 and equal 0 if one probability is equal to 1 and the others are 0. In the case of a uniform distribution between the classes, $p_k = 1/K$, $k = 1, \ldots, K$, the Gini index is equal to $(K-1)/K$.

At point level, we denote this index by $gs_i = 1 - \sum_{k=1}^{K} p_{ik}^2$ for point i.
At the areal level with squares G_g, we use $\overline{gs}_g = \frac{1}{\#G_g} \sum_{i \in G_g} gs_i$, the average of the point level gs_i for locations i inside square number g.
Because we are convinced that classification is going to be more difficult when there is diversity, i.e. impurity, we would like to relate the Gini-Simpson impurity index with the classification error and hence with the response error which is the main component of the classification error.

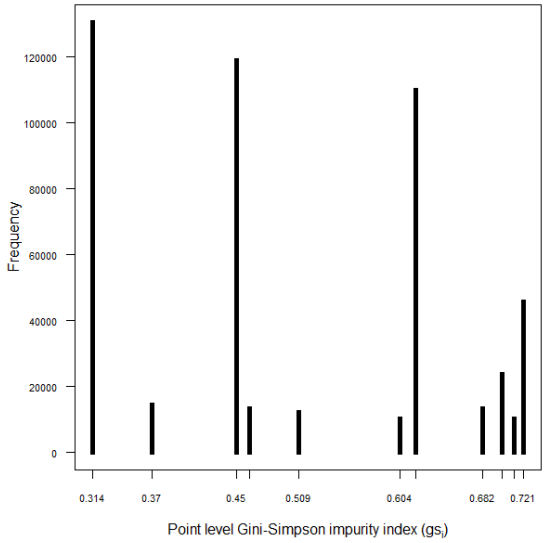

Figure 6: Bar chart of point level Gini-Simpson impurity index (gs_i).

We are first going to analyze the impurity of the vectors of probabilities generated by our model. As for the DGP probabilities, values of gs_i correspond to terminal nodes of the classification tree of the DGP (see Figure 2) and define a discrete variable with a finite number of values less than or equal to the number of terminal nodes, which gives 11 values in our case. Figure 6 presents the distribution of the gs_i across locations. As low values of the index

correspond to purity, we are mostly interested in the four smallest values of gs_i which constitute homogeneous groups in terms of land uses.

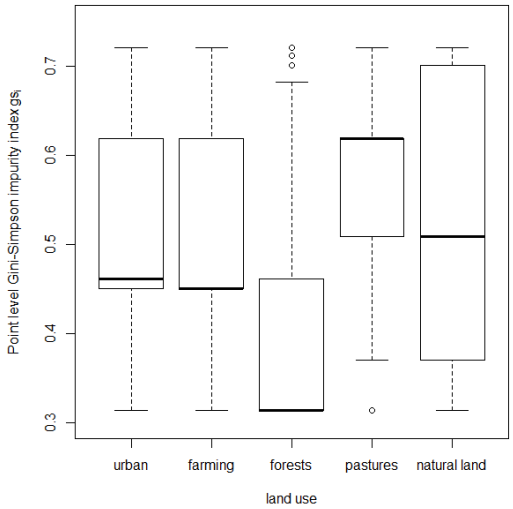

Figure 7: Boxplots of gs_i by land use.

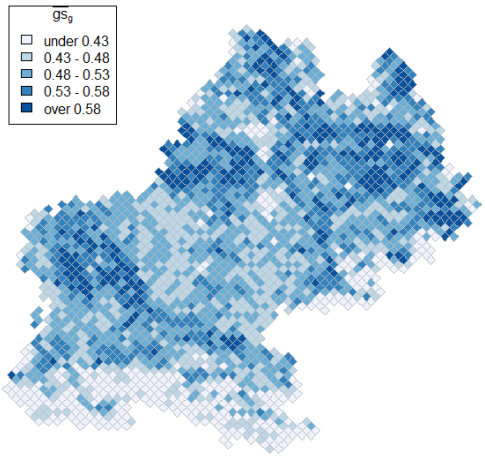

Figure 8: Map of $\overline{gs}_g$, the mean of gs_i at areal level.

Table 2: Characteristics of groups according to the Gini-Simpson index value (gs_i).

gs_i	**Frequency**	**Principal land uses**
0.314	130532	82.1% of forests
0.370	14376	77.4% of natural lands
0.450	119010	72.6% of farming
0.462	13158	72.0% of urban
0.509	12200	62.2% of pastures and 31.6% of natural lands
0.604	10167	56.0% of pastures and 27.4% of farming
0.619	109798	57.1% of pastures, 17.3% of forests and 13.3% of farming
0.682 to 0.721	92964	mix of all uses

4.1. Descriptive analysis

Using the bar chart of Figure 3, we make groups of gs_i values and Table 2 details the main land uses observed in each of these groups.

The first group defined by $gs_i = 0.314$ is the largest and the purest, it is composed by 82.1% of forests. The second group defined by $gs_i = 0.370$ is also very pure with 77.4% of natural lands. The third and fourth groups ($gs_i = 0.450$ and $gs_i = 0.462$ respectively) are quite pure with 72.6% of farming and 72.0% of urban use respectively, note that the third group is one of the largest. The other groups have at least two main land uses and are more and more impure.

Let us now examine the relationship between the distribution of the Gini-Simpson index and land use with Figure 7 which presents the distribution of gs_i by land use. We emphasize the notable behavior of the forests type which is mainly associated to low Gini-Simpson index

and hence low diversity. At the other extreme, the pastures type is very often encountered with other uses.

For mapping purposes, we use the cell levels averages $\overline{gs}_g$ of the Gini-Simpson's indices. Figure 8 compared to Figure 5 confirms that the mean of the Gini-Simpson impurity index is low (which means homogeneity or purity) in zones with very high proportion of a unique land use (forests or natural land) and high (which means heterogeneity or impurity) in zones with medium probabilities of several land uses.

4.2. Spatial analysis

In order to analyze the patterns of impurity, it is now natural to turn attention to the spatial distribution of the impurity index and to analyze its spatial autocorrelation using the Moran index (Cliff and Ord, 1981) and local indicators of spatial association (Anselin, 1995). For this purpose, we need to define a neighborhood matrix and we consider below the row-standardized 8-nearest neighbours matrix, which corresponds to a queen contingency matrix (Cliff and Ord, 1981). The Moran test ("free sampling model" version) is highly significant and the permutation test ("randomization model" version) is significant at the level 1%, thus we conclude that the Gini-Simpson impurity index presents a significant level of positive spatial autocorrelation. The Moran plot is presented in Figure 9.

The normalized Moran index is equal to 0.589. Figure 10 colors the map according to which quadrant of the Moran plot (high-high, low-low, high-low and low-high) the averaged Gini-Simpson index belongs and it shows that the Gini-Simpson index is highly clustered.

Figure 11 shows zones of significant positive local spatial association (LISA). Cold spots (zones with high positive LISA and low values of impurity index - i.e. purity - for segments and their neighbors, in green) are those with very high proportions of forests or natural land situated mainly in the South. Hot spots (zones with high positive LISA and high values of impurity index - i.e. impurity - for segments and their neighbors, in red) are zones where several land uses are observed with three principal patterns observed at several places: i) high proportion (40-60%) of pastures and medium proportion (20-40%) of forests in Ariege, most of Aveyron and some part of Lot, ii) medium proportions (20-40%) of forests and pastures in West of Aveyron and some part of Lot, and iii) medium proportion (20-40%) of farming, forests and pastures in Tarn-et-Garonne, South of Lot and Gers.

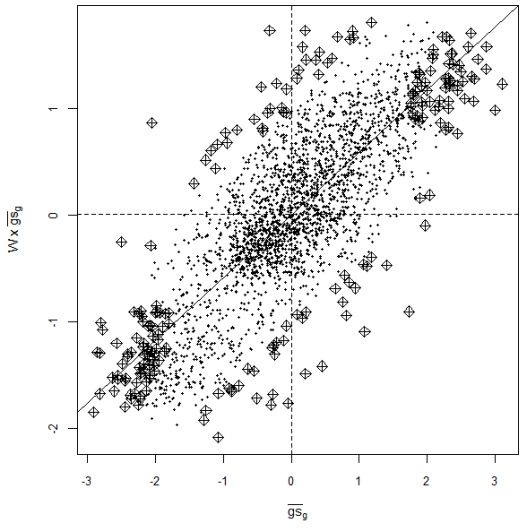

Figure 9: Normalized Moran scatterplot of the Gini-Simpson index $\overline{gs}_g$.

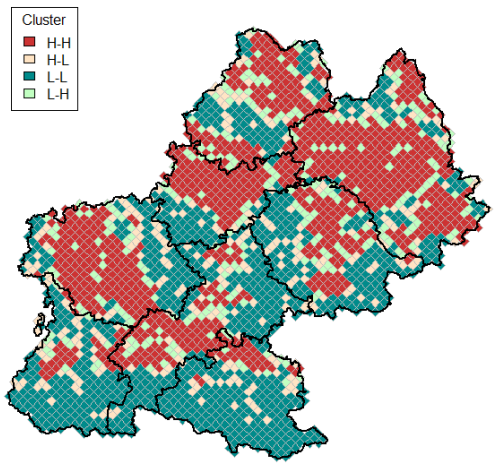

Figure 10: Clusters of the Gini-Simpson index $\overline{gs}_g$.

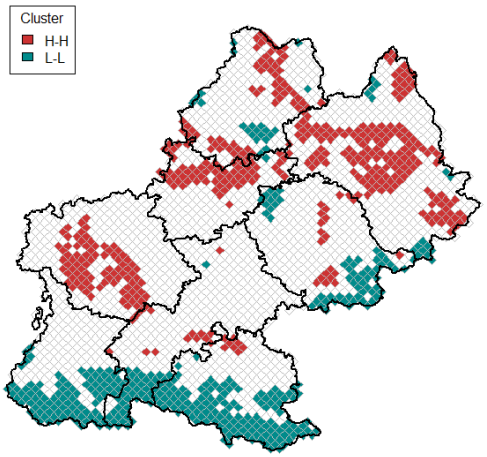

Figure 11: Segments with significant positive local spatial association (LISA) for the Gini-Simpson index $\overline{gs}_g$.

5. Analysis of the response error

5.1. *Analysis of the response error at point level*

The absolute error at point level between p_{ik} and d_{ik} is defined by $|d_{ik} - p_{ik}|$. As d_{ik} is a dummy variable, the absolute error equals $1 - p_{ik}$ if land use k is observed (i.e. if $d_{ik} = 1$) and p_{ik} otherwise (i.e. if $d_{ik} = 0$). Our analysis of this error includes descriptive statistics (see Table 3) and parallel boxplots of this error for the different values of the point level Gini-Simpson impurity index (see Figure 12).
For a given location i, the vector d_i has five coordinates, four equal to 0 and one equal to 1 hence we have more points in the first part of Table 3. We observe that the absolute response error quartiles are higher when $d_{ik} = 1$ than when $d_{ik} = 0$. This can be explained by the fact that many probabilities are quite small and, at locations where $d_{ik} = 1$, the value $1 - p_{ik}$ becomes quite large compared to p_{ik}. In particular, the conclusions for urban and natural land uses are the same because the probabilities of these two categories are very small (see Figure 4): the absolute response errors are the smallest when the land use is not observed ($d_{ik} = 0$) at least until the third quartile and these errors are the highest when the land use is observed ($d_{ik} = 1$). For the locations where the forests land use is observed, the minimum, first quartile and median values of the absolute response error are by far the smallest because there are quite many locations where the probability of forests is very large (see Table 3) compared to other land uses.

The relative error between p_{ik} and d_{ik} is defined by $|d_{ik} - p_{ik}|/p_{ik}$. As d_{ik} is a dummy variable, the relative error equals $1/p_{ik} - 1$ if the land use k is observed (i.e. if $d_{ik} = 1$) and 1 otherwise (i.e. if $d_{ik} = 0$). Descriptive statistics of this error at point level can be found in Table 3. Overall, the relative errors are rather variable with mean values close to 2 for farming, forests and pastures, and of the order of 10 for urban and natural land which are often associated with small probabilities. Note that the relative error for urban use reaches a maximum of 323, due to a point with a very low probability of urban but for which the multinomial draw gives urban use.

Figure 12 analyzes the link between the response error and the impurity by category of land use in detail. The rows correspond to the observed land use ($d_{ik} = 1$ for $k = 1, \ldots, 5$) and the columns to the response error ($|d_{ik} - p_{ik}|$ for $k = 1, \ldots, 5$). First note that most of the errors are constant because there is a unique value of the Gini-Simpson index which results in horizontal lines instead of boxplots.

Let us compare the graphs in a same column, the first one for example. In the first row of column one, we have an observed land use which is urban ($d_{i1} = 1$) and the error is on the urban category. We note that there is one boxplot showing low errors while the remaining ones display large errors. The low error boxplot corresponds to the value 0.462 of the Gini-Simpson index with points in a rather pure environment, where the main land use is urban and hence a value of p_{i1} close to 1 which explains the low error. On the contrary, for the other values of the Gini-Simpson index which correspond to points in a non purely urban environment, the error is large because

Table 3: Descriptive statistics of absolute and relative response error at point level.

		n	Minimum	Q1	Median	Mean	Q3	Maximum
Absolute response error								
urban	$\lvert d_{i1} - p_{i1}\rvert$ when $d_{i1} = 0$	464964	0.00	0.02	0.06	0.06	0.07	0.72
farming	$\lvert d_{i2} - p_{i2}\rvert$ when $d_{i2} = 0$	370122	0.00	0.02	0.07	0.16	0.13	0.73
forests	$\lvert d_{i3} - p_{i3}\rvert$ when $d_{i3} = 0$	347114	0.06	0.06	0.17	0.17	0.18	0.82
pastures	$\lvert d_{i4} - p_{i4}\rvert$ when $d_{i4} = 0$	372394	0.05	0.08	0.12	0.20	0.26	0.62
natural land	$\lvert d_{i5} - p_{i5}\rvert$ when $d_{i5} = 0$	454226	0.03	0.03	0.05	0.07	0.06	0.78
urban	$\lvert d_{i1} - p_{i1}\rvert$ when $d_{i1} = 1$	37241	0.28	0.28	0.93	0.73	0.94	1.00
farming	$\lvert d_{i2} - p_{i2}\rvert$ when $d_{i2} = 1$	132083	0.27	0.27	0.27	0.44	0.59	0.98
forests	$\lvert d_{i3} - p_{i3}\rvert$ when $d_{i3} = 1$	155091	0.18	0.18	0.18	0.38	0.81	0.94
pastures	$\lvert d_{i4} - p_{i4}\rvert$ when $d_{i4} = 1$	129811	0.38	0.43	0.43	0.58	0.75	0.95
natural land	$\lvert d_{i5} - p_{i5}\rvert$ when $d_{i5} = 1$	47979	0.22	0.60	0.69	0.69	0.94	0.97
Relative response error								
urban	$\lvert d_{i1} - p_{i1}\rvert / p_{i1}$ when $d_{i1} = 1$	37241	0.39	0.39	13.11	12.40	14.80	323.00
farming	$\lvert d_{i2} - p_{i2}\rvert / p_{i2}$ when $d_{i2} = 1$	132083	0.38	0.38	0.38	2.60	1.46	45.20
forests	$\lvert d_{i3} - p_{i3}\rvert / p_{i3}$ when $d_{i3} = 1$	155091	0.22	0.22	0.22	2.23	4.29	16.05
pastures	$\lvert d_{i4} - p_{i4}\rvert / p_{i4}$ when $d_{i4} = 1$	129811	0.60	0.75	0.75	2.89	2.96	18.12
natural land	$\lvert d_{i5} - p_{i5}\rvert / p_{i5}$ when $d_{i5} = 1$	47979	0.29	1.50	2.18	9.47	16.26	30.12

the probability p_{i1} is low. For the other rows of column one, the observed land use is no longer urban ($d_{i1} = 0$) and the errors corresponding to the 0.462 of the Gini-Simpson index are large because p_{i1} is large. On the contrary, for the other values of the Gini-Simpson index which correspond to points in a non purely urban environment, the error is low because p_{i1} is low and $d_{i1} = 0$. The columns urban and forests behave similarly whereas for pastures for example, we simultaneously see errors which are neither very low nor very large because the pastures are situated in more heterogeneous (impure) regions (see also Chakir et al., 2016b, for a comparison between the forests and pastures land use).

More generally, when a land use is observed, the error for this category is low at locations where this category is mainly observed and high at other locations, whereas when it is not observed, the opposite results are found.

Another plot for exploring the absolute response error is proposed by Haaf et al. (2014) and can be found in Chakir et al. (2016b) for our data set of interest. It is called the Cumulative Distribution Function of Error Tolerance (CDFET) and consists in the empirical cumulative distribution function of the response error for each land use. The percentage of points with absolute response error less than a specified threshold is plotted for each land use and we note that, for low values of the error threshold, the curves are very similar with small errors for approximately 35% of the points. These good predictions correspond to homogeneous zones with either low or high probabilities. This plot also confirms a different behavior of the urban and natural land uses compared with farming, forests and pastures. For these last land uses, the presence of medium probabilities causes a deterioration of the curve behavior.

In other words, the response error at point level shows different patterns across land uses and the analysis reveals that the larger errors correspond to heterogeneous zones.

5.2. *Analysis of the response error at areal level*

At areal level, we consider two different ways of aggregating the response errors. In the first case we consider the average point level absolute response error $\frac{1}{\#G_g} \sum_{i \in G_g} |d_{ik} - p_{ik}|$. In the second case we consider the difference between

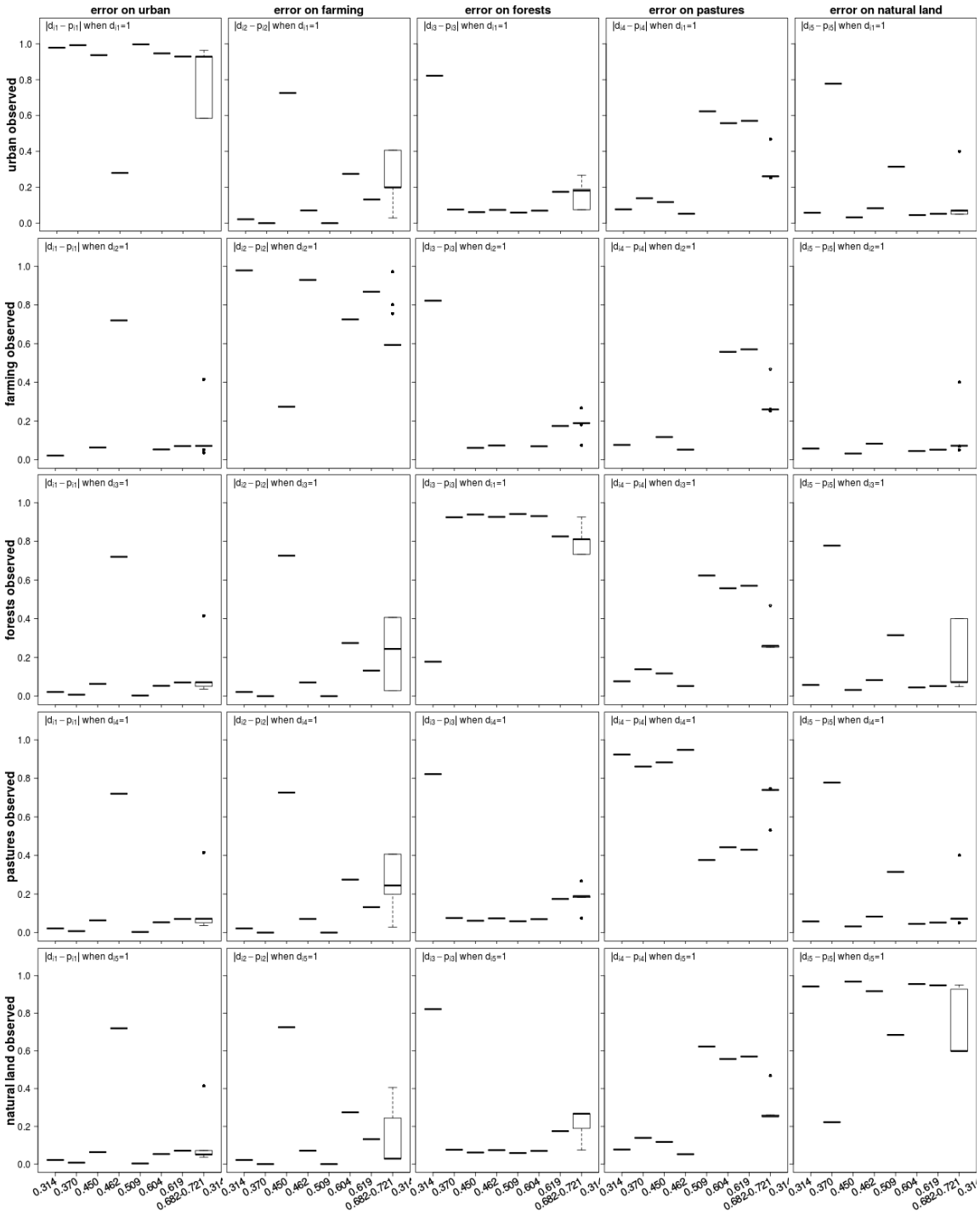

Figure 12: Absolute response error vs the Gini-Simpson impurity index gs_i by observed land use (rows) and by response error component (columns).

Table 4: Descriptive statistics of the two types of absolute response error at areal level.

		Minimum	Q1	Median	Mean	Q3	Maximum
Average point level absolute response error							
urban	$\frac{1}{\#G_g}\sum_{i\in G_g}\lvert d_{i1}-p_{i1}\rvert$	0.00	0.08	0.11	0.11	0.13	0.39
farming	$\frac{1}{\#G_g}\sum_{i\in G_g}\lvert d_{i2}-p_{i2}\rvert$	0.00	0.13	0.22	0.23	0.35	0.46
forests	$\frac{1}{\#G_g}\sum_{i\in G_g}\lvert d_{i3}-p_{i3}\rvert$	0.09	0.18	0.25	0.24	0.29	0.41
pastures	$\frac{1}{\#G_g}\sum_{i\in G_g}\lvert d_{i4}-p_{i4}\rvert$	0.08	0.24	0.29	0.30	0.35	0.49
natural land	$\frac{1}{\#G_g}\sum_{i\in G_g}\lvert d_{i5}-p_{i5}\rvert$	0.03	0.09	0.11	0.13	0.14	0.43
Areal level absolute response error							
urban	$\lvert\bar{d}_{g1}-\bar{p}_{g1}\rvert$	0.00	0.00	0.01	0.01	0.02	0.11
farming	$\lvert\bar{d}_{g2}-\bar{p}_{g2}\rvert$	0.00	0.01	0.01	0.02	0.03	0.11
forests	$\lvert\bar{d}_{g3}-\bar{p}_{g3}\rvert$	0.00	0.01	0.02	0.02	0.03	0.12
pastures	$\lvert\bar{d}_{g4}-\bar{p}_{g4}\rvert$	0.00	0.01	0.02	0.02	0.03	0.15
natural land	$\lvert\bar{d}_{g5}-\bar{p}_{g5}\rvert$	0.00	0.01	0.01	0.01	0.02	0.10

aggregated probabilities $|\bar{d}_{gk}-\bar{p}_{gk}|$ where $\bar{d}_{gk}=\frac{1}{\#G_g}\sum_{i\in G_g}d_{ik}$ and $\bar{p}_{gk}=\frac{1}{\#G_g}\sum_{i\in G_g}p_{ik}$ and call it the "areal level absolute response error".

In comparison to Table 3, Table 4 shows that the two types of areal level errors are lower than point level errors. Areal level absolute response error are very small for all categories with maximum values between 0.10 and 0.15. When averaging point level response errors, there is not so much difference between the land uses except for the urban and the natural land uses which are associated with smaller values.

The maps of Figure 13 plot the average point level absolute response error and are quite similar to those in Figure 5 where the aggregated observed land uses are plotted. This confirms the fact already noticed in Table 3 that the error is larger when d_{ik} equals one than when d_{ik} is zero because the probabilties are usually low. It also confirms that the disaggregated level leads to very poor results. Unlike the average point level absolute response error (see Figure 13), Figure 14 does not show any spatial pattern for the areal level absolute response errors. We can however notice that errors on farming and urban in a lesser extent are very low in the south of the region where forests and natural lands are very frequent.

On Figure 15, there is a positive trend for average point level response error for pastures and farming. The error for these land uses is lower when impurity is low which confirms the point level analysis. No relationship between the areal level response error and the Gini-Simpson impurity index can be found (see Figure 16).

In Chakir et al. (2016b), the CDFET for the areal level absolute response error is compared with the CDFET for the average point level response error. The comparison confirms once more that the areal level response error is very small and for all land uses.

In other words, the areal level absolute response error is much lower than the average point level response error. The impact of heterogeneity vanishes. These two points emphasize the interest of giving results at aggregated levels only. The reader can also refer to Chakir et al. (2016b) for a discussion on the choice of the aggregated level considering different grids.

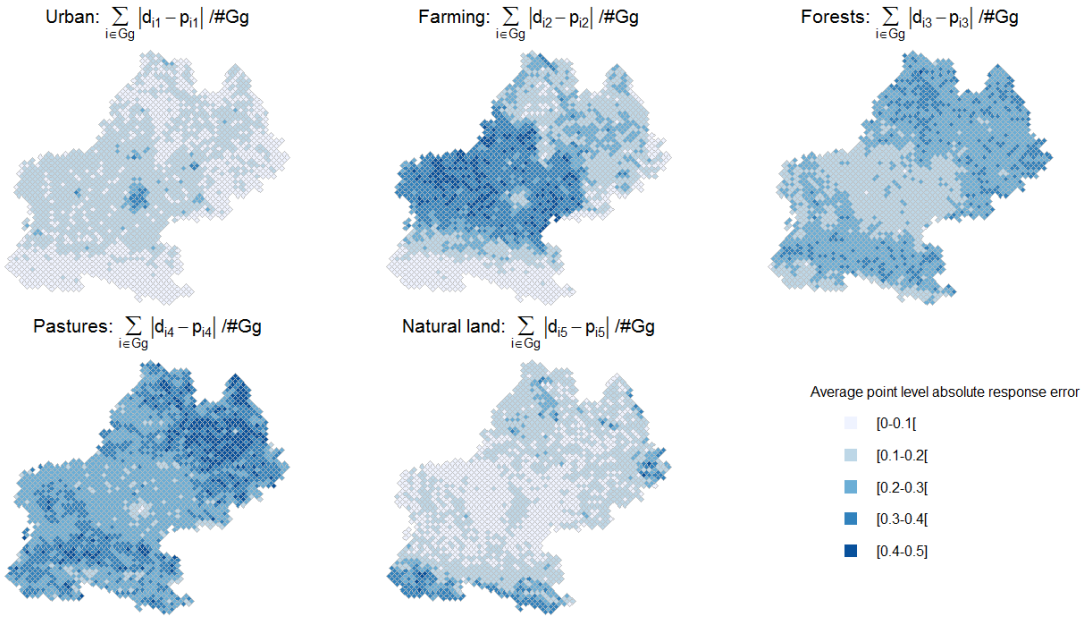

Figure 13: Average point level absolute response error.

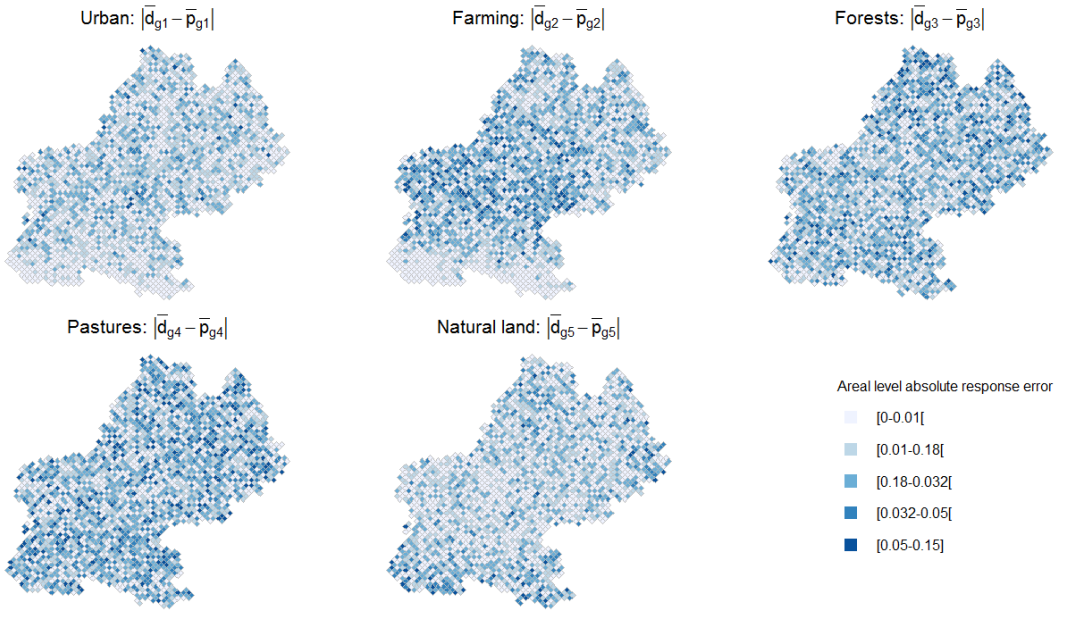

Figure 14: Areal level absolute response error ($|\bar{d}_{gk} - \bar{p}_{gk}|$).

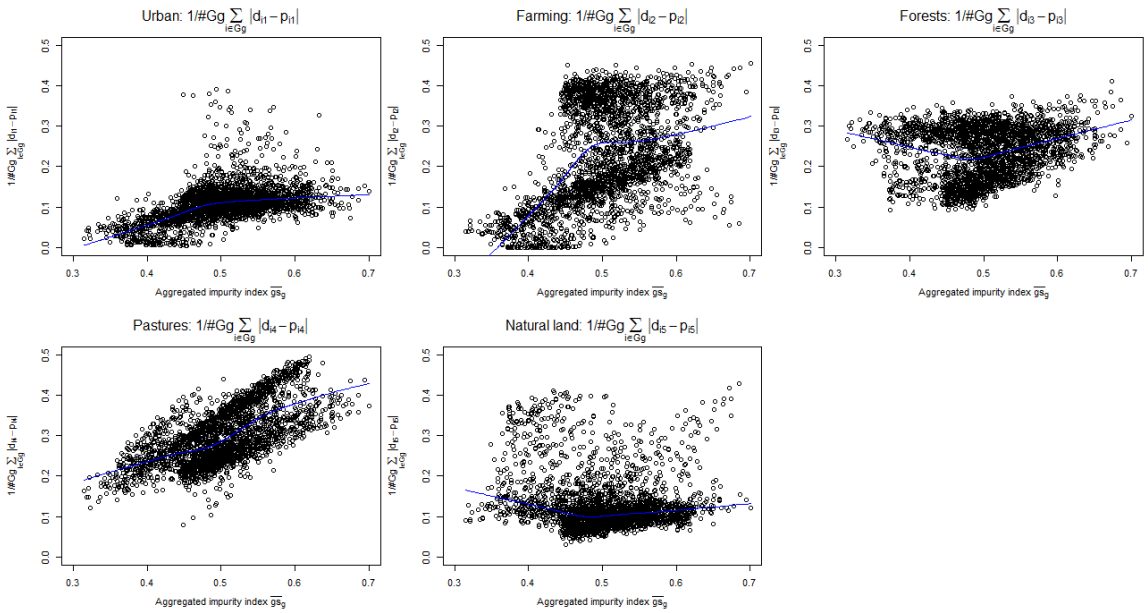

Figure 15: Average point level absolute response error ($\frac{1}{\#G_g}\sum_{i\in G_g}|d_{ik}-p_{ik}|$) versus aggregated Gini-Simpson impurity index $\overline{gs}_g$.

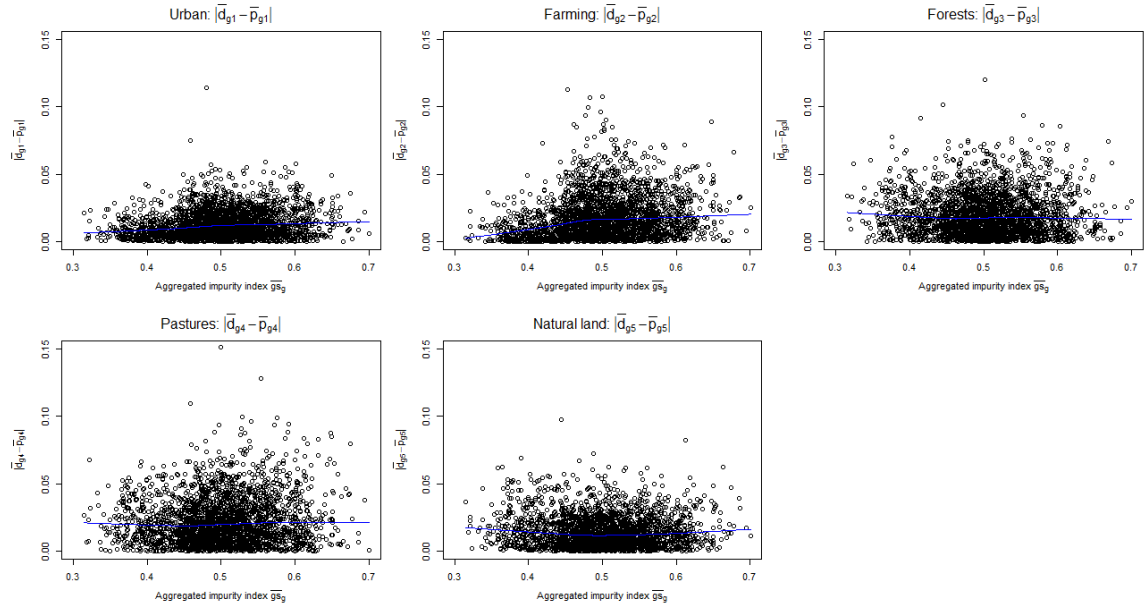

Figure 16: Areal level absolute response error ($|\bar{d}_{gk}-\bar{p}_{gk}|$) versus aggregated Gini-Simpson impurity index $\overline{gs}_g$.

6. Conclusion

In the present paper, we propose a synthetic data set generated by simulations with a model fitted to the Teruti-Lucas data and analyse in detail the response error when the land use is defined in 5 categories. Several conclusions can be derived from our study. The main one is that the response error at point level is larger in areas where the probabilities are quite similar (heterogeneous areas) and this fact can be illustrated using the Gini-Simpson measure. We also notice that the response error at point level is often larger for the land use which is observed than for the other land uses because the probabilities are rarely above 50%. Finally, aggregating the land uses is preferable in order to reduce the error. Note that in Chakir et al. (2016b), we even go further and advise the data analyst not to calculate predictions at the point level but aggregate directly the estimated probabilities so that the response error vanishes completely.

The covariates were only used in order to generate the data set but not for estimating a model as in Chakir et al. (2016b, 2017). However, it is possible to use the proposed data set for comparing different classification methods (in a data science challenge for instance) but also for comparing different sampling designs with the Teruti-Lucas systematic design for instance.

Acknowledgements

This work was partially financed by the Agence Nationale de la Recherche through the ModULand project (ANR-11-BSH1-005). We thank the Service de la Statistique et de la Prospective of the French Ministry of Agriculture for letting us use the Teruti-Lucas survey data in the framework of the ANR ModULand.

References

Anselin, L. (1995). Local indicators of spatial association-lisa. *Geographical Analysis*, 27:93–115.

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

Chakir, R., Laurent, T., Ruiz-Gazen, A., Thomas-Agnan, C., and Vignes, C. (2016a). Land use predictions on a regular grid at different scales and with easily accessible covariates. Working Paper.

Chakir, R., Laurent, T., Ruiz-Gazen, A., Thomas-Agnan, C., and Vignes, C. (2016b). Spatial scale in land use models: application to the teruti-lucas survey. *Spatial Statistics*, 18:246–262.

Chakir, R., Laurent, T., Ruiz-Gazen, A., Thomas-Agnan, C., and Vignes, C. (2017). Prédiction de l'usage des sols sur un zonage régulier à différentes résolutions et à partir de covariables facilement accessibles. *Revue Economique*, 68:435–469.

Cliff, A. D. and Ord, J. K. (1981). *Spatial processes: models & applications*, volume 44. Pion London.

Haaf, C. G., Michalek, J. J., Morrow, W. R., and Liu, Y. (2014). Sensitivity of vehicle market share predictions to discrete choice model specification. *Journal of Mechanical Design*, 136(12):121402.

Simpson, E. (1949). Measurement of diversity. *Nature*, 163:688.

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

Correspondence: christine.thomas@tse-fr.eu.