Prédiction de l'usage des sols sur un zonage régulier à différentes résolutions et à partir de covariables facilement accessibles. Application à l'enquête Teruti-Lucas

Raja Chakir (INRA-AgroParisTech), Thibault Laurent (TSE-R), Anne Ruiz-Gazen (TSE-R), Christine Thomas-Agnan (TSE-R) and Céline Vignes (TSE-R)

Séminaire INRA - MIAT - 1er avril 2016

Introduction

Objectives :

1. predict land use in the Midi-Pyrénées region of France (in 5 categories) using data easily accessible
2. at different spatial scales (points level and on regular grids).
3. Determine the different components of the prediction error.
4. Understand better the prediction error.

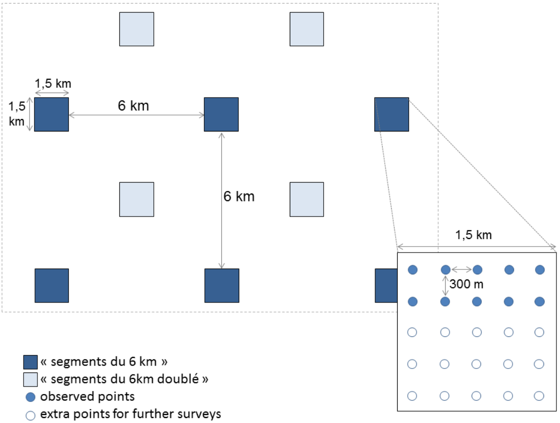
We focus on two quality criteria :

- The percentage of good prediction at the point level.
- The mean squared error of the estimated proportions (MSE), or the squared root of the MSE, or the Brier score (1/2 MSE), or the weighted Brier score, at the point and grid levels .

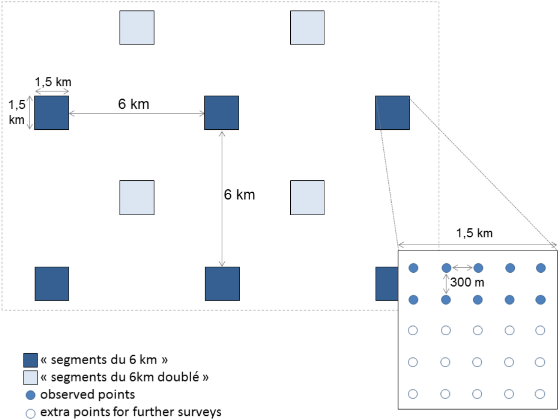
Data sources

Name	Geographical level	Source	Year
Land use (5 categories)	6km segment	Teruti-Lucas	2010
Soil constitution	UCS zones	BGSF (GISSOL)	1998
main surface			
base material			
evolution of soil texture			
presence of a waterproof layer			
Meteorology	grid 25x25km	Agri4cast	2010
annual minimum of daily temperature			
annual maximum of daily temperature			
annual mean of daily temperature			
annual sum of rain quantity			
mean speed of wind			
Land and empty meadow price	32 NRA	Agreste	2010
Socio-economic data		Insee	2010
population density (new)	grid 200x200m		
percentage of farmers	municipalities		
percentage of executives	municipalities		
metropolitan center	municipalities		
CLC2 (15 categories)	zones ($>$ 25 ha)	Corine Land Cover	2006
Altitude	grid (250m)	BDAlti de l'IGN	-

Teruti-Lucas : points and segments



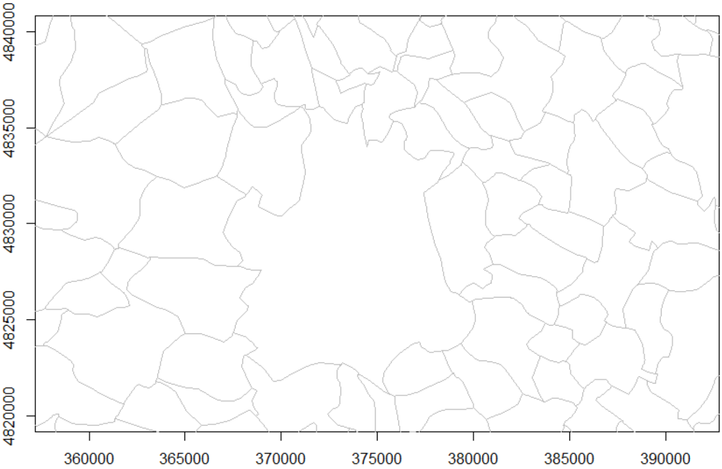
Teruti-Lucas : points and segments



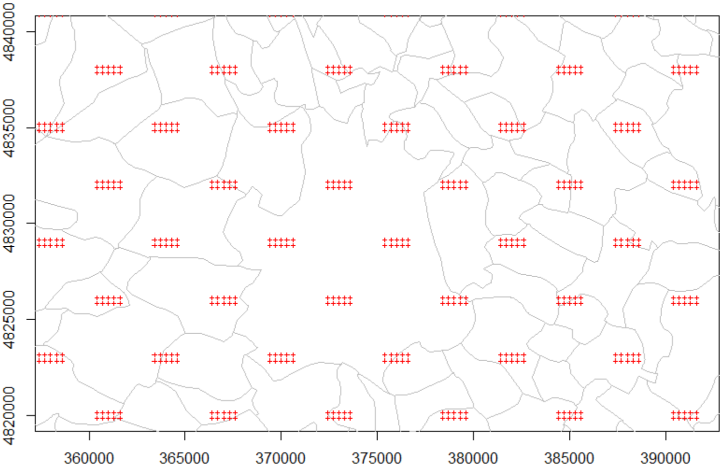
Teruti Lucas

- 10 points per "segment" (or less)
- Dark blue : learning sample
- Light blue : test sample

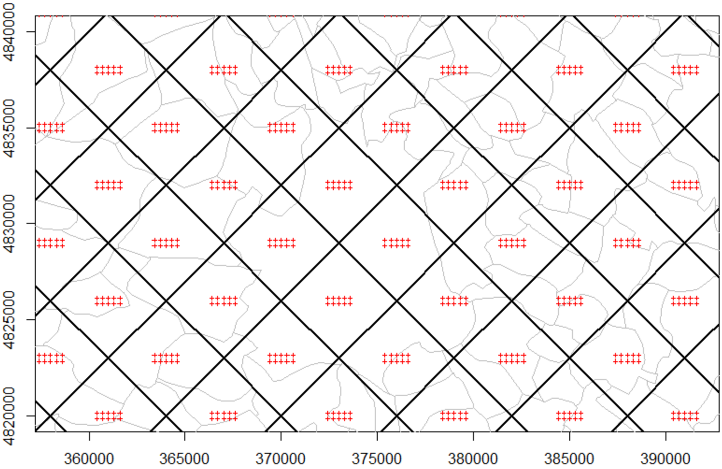
Municipalities



Teruti Lucas points



The grid (level A1)

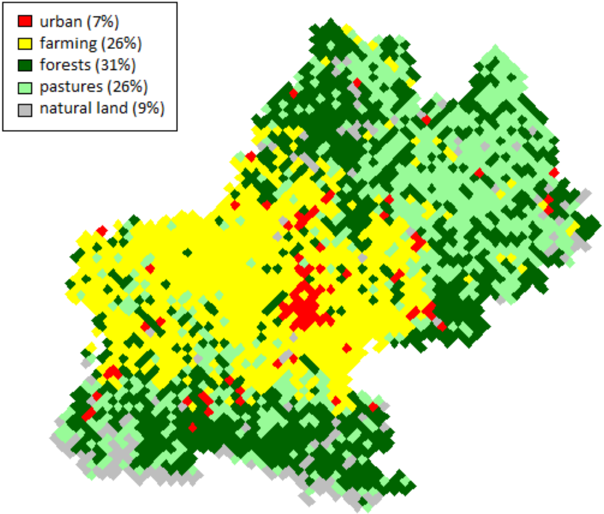


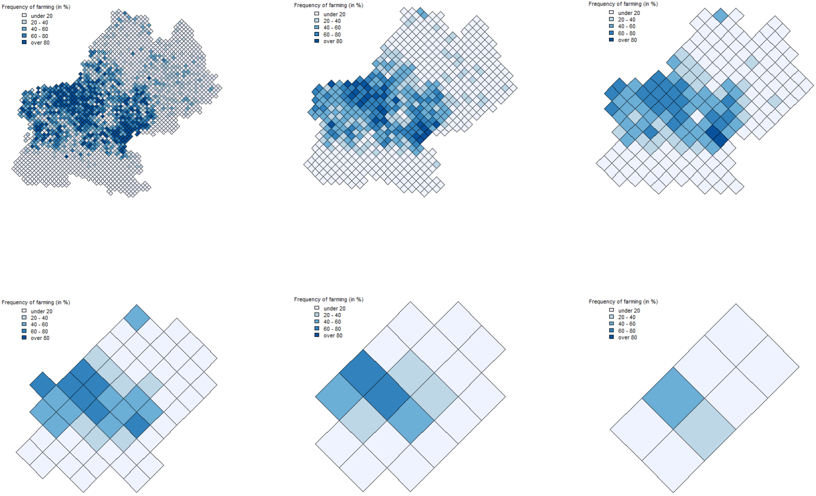
From points to a grid

Remark :

- Land use at the point level.
- Proportion of land use or main land use at a grid level.

Main land use





Aggregation levels from A1 to A6

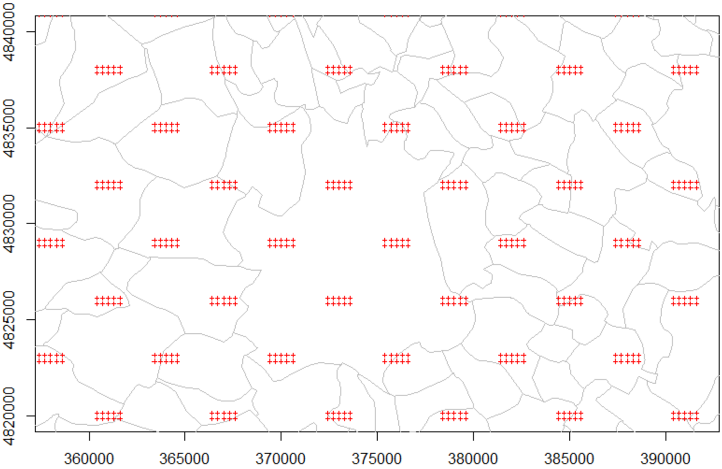
Summary of the spatial levels

TABLE: Characteristics of the grids

Grid	Number of aggregated "unit squares"	Approximate area	Number of points per square	Total number of squares
A_1	1	18 km^2	1 à 10	2 579 squares
A_2	4	72 km^2	1 à 40	689 squares
A_3	16	288 km^2	4 à 160	192 squares
A_4	64	1 152 km^2	10 à 640	59 squares
A_5	256	4 608 km^2	184 à 2 559	20 squares
A_6	1 024	18 432 km^2	184 à 6 605	8 squares

A_0 is the Teruti-Lucas points level and A_7 is the whole Midi-Pyrénées region.

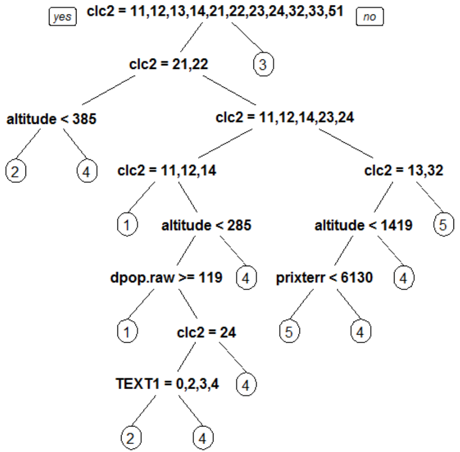
Teruti Lucas points

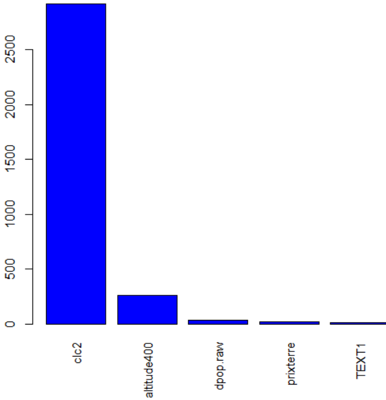


Prediction of the land use at Teruti-Lucas points

- There exist several methods for predicting a categorical variable with more than two categories.
- Multinomial logit model (MNL), discriminant analysis, classification tree,. . .
- We compared MNL and trees and get very similar results in terms of percentage of good prediction (number of points correctly predicted divided by the number of points).
- In this presentation, we focus on classification trees only.

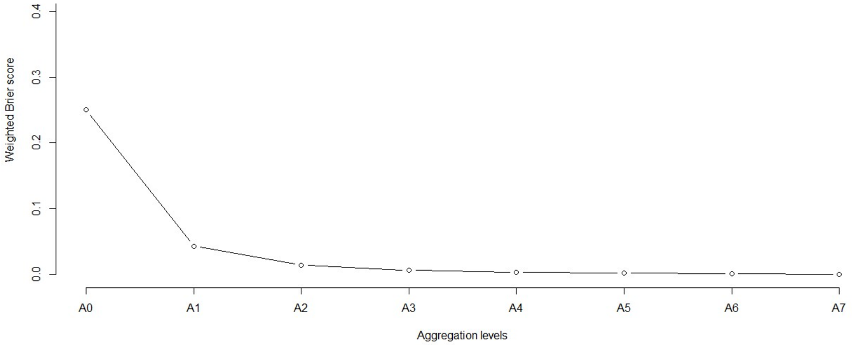
Classification tree and importance of variables



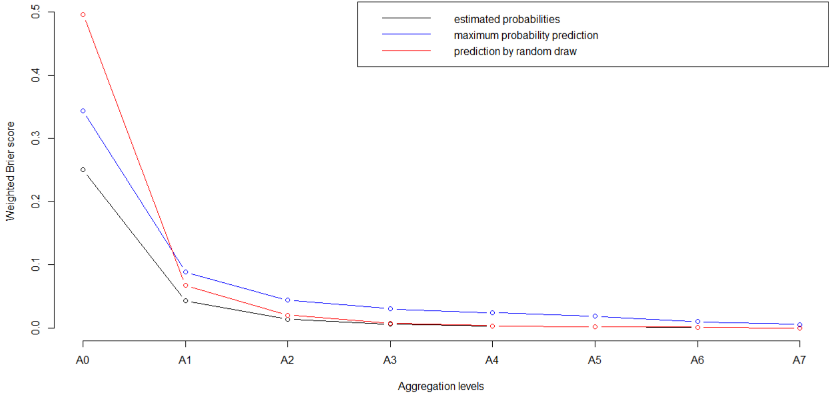


Classification tree

Results :
Percentage of correctly classified points : 65% with the maximum probability and 50% with a multinomial draw.



Comparison depending on the response prediction



Remark

- if $X_i \sim$ Multinomial$(1, p_{i1}, \ldots, p_{iK})$, $Y_i = (Y_{i1}, \ldots, Y_{iK})'$ with $Y_{iq(i)} = 1$ if j is such that $p_{iq(i)} = \max(p_{ij}, j = 1, \ldots, K)$ and X_i and Y_i independent, then

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{K} X_{ij} Y_{ij} = \frac{1}{n}\sum_{i=1}^{n} X_{iq(i)}$$

  and

$$E\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{K} X_{ij} Y_{ij}\right) = \frac{1}{n}\sum_{i=1}^{n} p_{iq(i)}$$

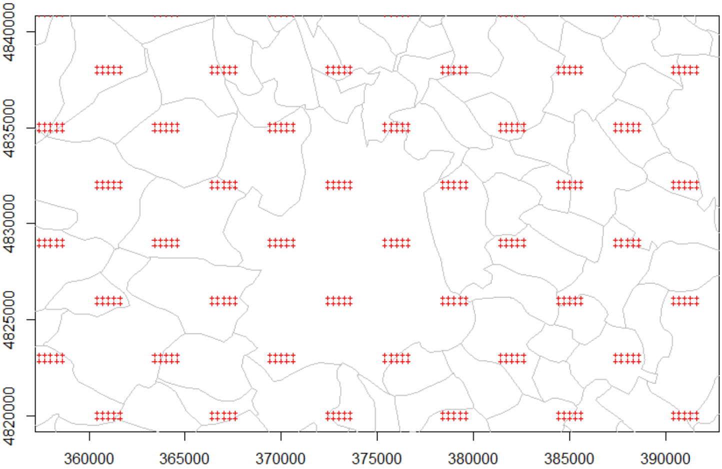
- if $X_i \sim$ Multinomial$(1, p_{i1}, \ldots, p_{iK})$, $Y_i \sim$ Multinomial$(1, p_{i1}, \ldots, p_{iK})$ and X_i and Y_i independent, then

$$E\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{K} X_{ij} Y_{ij}\right) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{K} p_{ij}^2$$

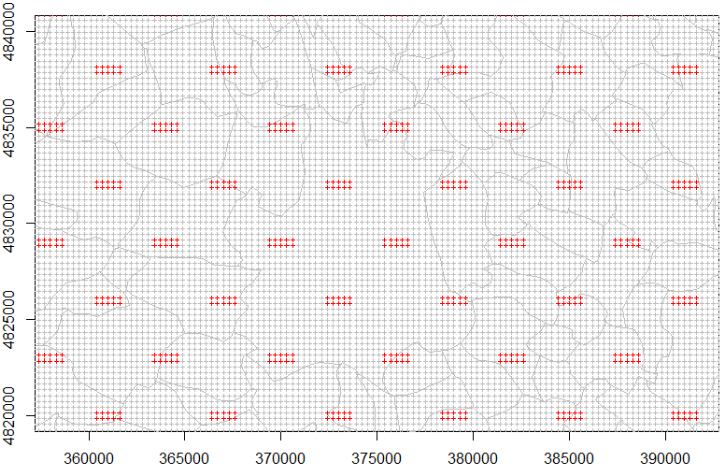
Remark

	well-classified rate	"pmax mean"
$K = 5$	65.12%	65.16%
$K = 4$	72.84%	73.19%
	well-classified rate	mean squares prob.
$K = 5$	50.01%	50.45%

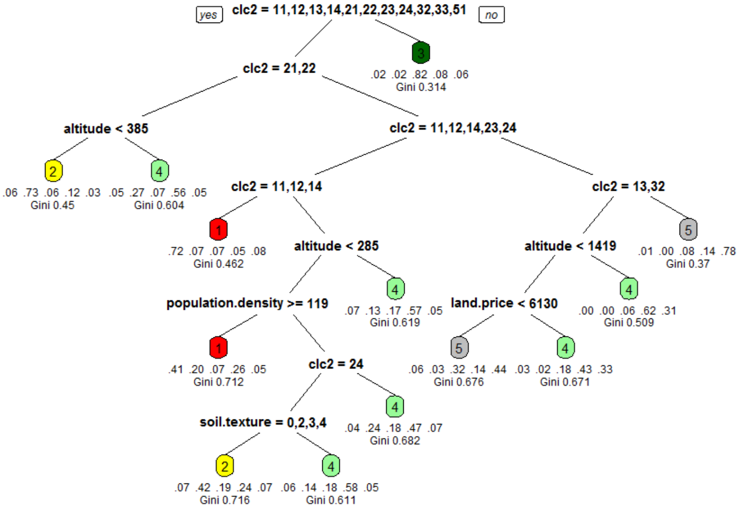
More points than Teruti-Lucas



More points than Teruti-Lucas



Classification tree chosen for the DGP



DGP

- Locations $i = 1, \ldots, 502205$,
- land uses $k = 1, \ldots, K$, $K = 5$,
- explanatory variables x_i,
- vector of probabilities $p_i = (p_{i1}, \ldots, p_{iK})$ at location i such that :

$$p_i = f(x_i).$$

DGP

- Locations $i = 1, \ldots, 502205$,
- land uses $k = 1, \ldots, K$, $K = 5$,
- explanatory variables x_i,
- vector of probabilities $p_i = (p_{i1}, \ldots, p_{iK})$ at location i such that :

$$p_i = f(x_i).$$

- The variable to explain dummy variable
  - "random draw" response denoted d^r_{ik} following a multinomial distribution with parameters 1 and p_i,
  - "maximum probability" response denoted d^m_{ik} ($d^m_{ik} = 1$ if p_{ik} is the maximum probability among the p_{ij} $j = 1, \ldots, K$ and 0 otherwise).

Prediction

- Locations $i = 1, \ldots, 25317$ (Teruti-Lucas),
- land uses $k = 1, \ldots, K$,
- explanatory variables x_i,
- variable to explain d^r_{ik},
- vector of probabilities estimates $\hat{p}_i = (\hat{p}_{i1}, \ldots, \hat{p}_{iK})$ at location i such that :

$$\hat{p}_i = \hat{f}(x_i).$$

Prediction

- Locations $i = 1, \ldots, 25317$ (Teruti-Lucas),
- land uses $k = 1, \ldots, K$,
- explanatory variables x_i,
- variable to explain d_{ik}^r,
- vector of probabilities estimates $\hat{p}_i = (\hat{p}_{i1}, \ldots, \hat{p}_{iK})$ at location i such that :

$$\hat{p}_i = \hat{f}(x_i).$$

- The prediction $\hat{d}_{ik}^m$ dummy variable at $i = 1, \ldots, 502205$ by
  - "random draw" predicted response denoted $\hat{d}_{ik}^r$ following a multinomial distribution with parameters 1 and $\hat{p}_i$,
  - "maximum probability" predicted response denoted $\hat{d}_{ik}^m$ ($\hat{d}_{ik}^m = 1$ if $\hat{p}_{ik}$ is the maximum probability among the $\hat{p}_{ij}$ $j = 1, \ldots, K$ and 0 otherwise).

The Sum of Squared Errors (SSE) between $\hat{d}_{ik}^m$ and d_{ik}^r defined by $SSE = \sum_{i=1}^{n} \left(\hat{d}_{ik}^m - d_{ik}^r\right)^2$ can be decomposed into :

$$\sum_{i=1}^{n} (\hat{d}_{ik}^m - d_{ik}^r)^2 = \sum_{i=1}^{n} (\hat{d}_{ik}^m - \hat{p}_{ik})^2 + \sum_{i=1}^{n} (\hat{p}_{ik} - p_{ik})^2 + \sum_{i=1}^{n} (p_{ik} - d_{ik}^r)^2 + R$$

where

$$\begin{aligned} R &= -2 \sum_{i=1}^{n} \Big[(\hat{d}_{ik}^m - \hat{p}_{ik})(\hat{p}_{ik} - p_{ik}) \\ &\quad - (\hat{d}_{ik}^m - \hat{p}_{ik})(d_{ik}^r - p_{ik}) - (\hat{p}_{ik} - p_{ik})(d_{ik}^r - p_{ik}) \Big]. \end{aligned}$$

Error decomposition

At the point level, "response error"

	urban	farming	forests	pastures	natural land
$\sum_i(\hat{d}^m_{ik} - d^r_{ik})^2$	33363.00	84407.00	71412.00	111038.00	41972.00
$\sum_i(\hat{d}^m_{ik} - \hat{p}_{ik})^2$	5983.91	26540.87	12118.45	37009.47	8168.27
$\sum_i(\hat{p}_{ik} - p_{ik})^2$	13.20	43.07	36.12	107.11	87.25
$\sum_i(d^r_{ik} - p_{ik})^2$	27324.69	57232.18	59282.26	73850.71	32923.67
R	41.20	590.87	-24.83	70.70	792.81

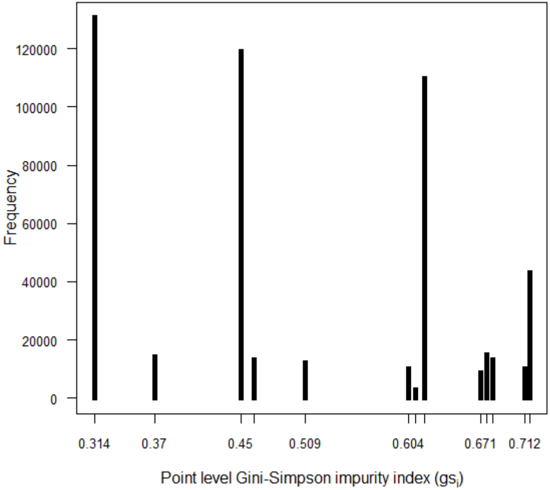
As a consequence, from now on, we forget the predicted probabilities and consider only the true probabilities.

We are now going to analyze the impurity of these probabilities and their spatial autocorrelation as we suspect that there is a relationship between the errors and the impurity. More preciselu, we measure how homogeneous or diverse is land use at a given point or in a given region, with the idea that classification is going to be more difficult when there is diversity.

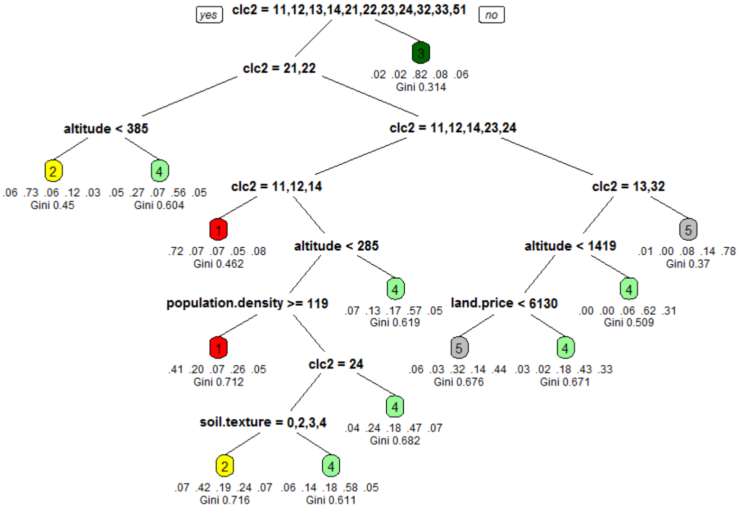
The impurity of probabilities p_{ik} is measured by the Gini-Simpson impurity index $gs_i = 1 - \sum_{k=1}^{K} p_{ik}^2$.

As values of gs_i correspond to terminal nodes of the tree of the DGP, gs_i is a discrete variable with 13 values.

Bar chart of point level Gini-Simpson impurity index



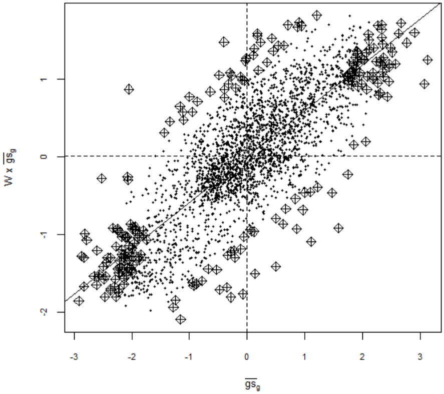
Classification tree chosen for the DGP

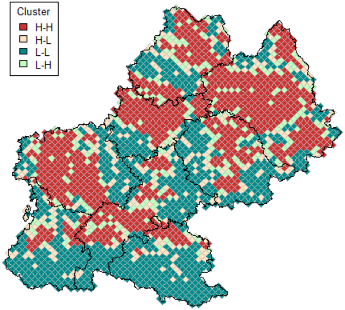


Characteristics of groups according to the gs_i value

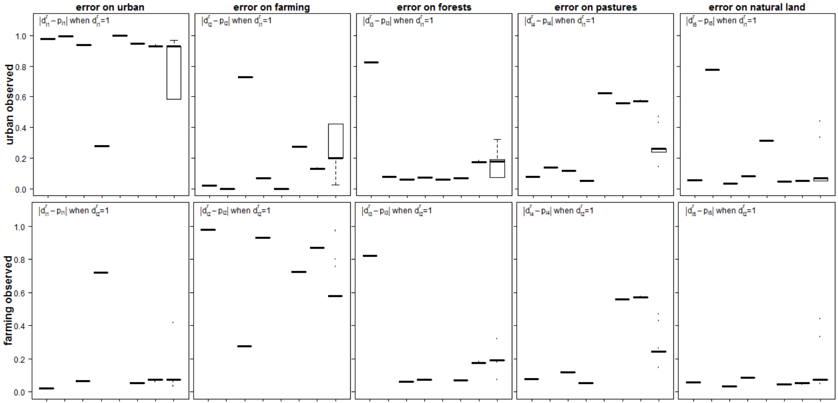
gs_i	frequency	principal land uses
0.314	130532	82.3% of forests
0.370	14376	77.4% of natural lands
0.450	119010	72.9% of farming
0.462	13158	71.1% of urban
0.509	12200	63.0% of pastures and 31.0% of natural land
0.604	10167	56.1% of pastures and 27.1% of farming
0.611 to 0.619	112647	57.1% of pastures, 17.4% of forests and 13.1% of farming
0.671 to 0.716	90115	mix of all uses

Moran scatterplot of the Gini-Simpson index



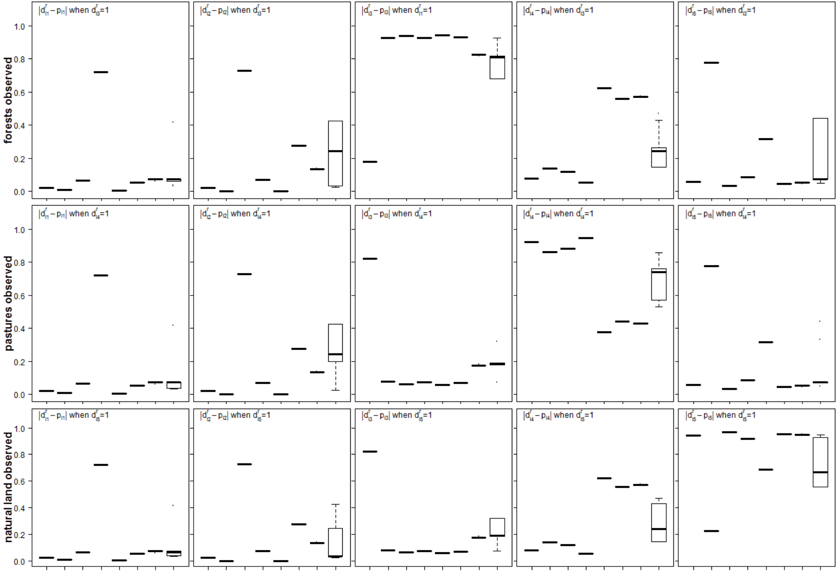


Absolute response error vs the Gini-Simpson impurity index



by observed land use (rows) and by response error component (columns)

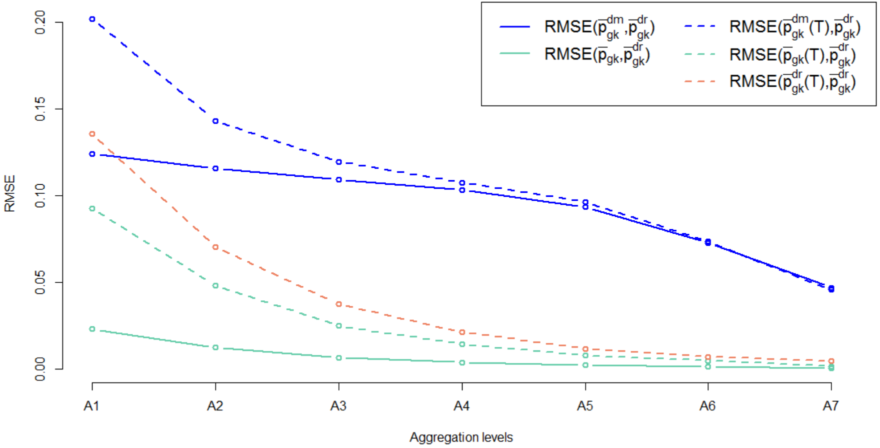
Absolute response error vs the Gini-Simpson impurity index



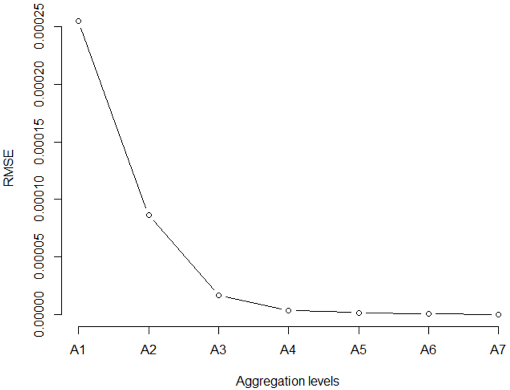
For each cell G_g, we define three aggregated probabilities :

- $\bar{p}_{gk}$ denotes the average of the probabilities p_{ik} derived from our initial model $p_i = f(x_i)$ for the points i that belong to the same cell G_g : $\bar{p}_{gk} = \frac{1}{\#G_g} \sum_{i \in G_g} p_{ik}$ where $\#G_g$ denotes the number of points in the cell G_g.
- $\bar{p}^{dr}_{gk} = \frac{1}{\#G_g} \sum_{i \in G_g} d^r_{ik}$, where we recall that d^r_{ik} is the prediction by multinomial random draw
- $\bar{p}^{dm}_{gk} = \frac{1}{\#G_g} \sum_{i \in G_g} d^m_{ik}$, where we recall that d^m_{ik} is the maximum probability prediction

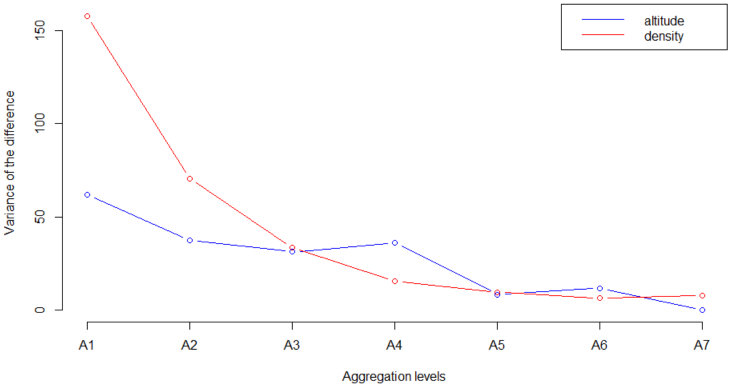
- We can make a part of the response error disappear by aggregating the probability estimates.
- We can measure another type of error called the sampling error. This error is due to the fact that we estimate the probabilities only at the Teruti-Lucas points while the explanatory variables are available at any point.


RMSE($\bar{p}^{dm}_{gk}$,$\bar{p}^{dr}_{gk}$)
RMSE($\bar{p}_{gk}$,$\bar{p}^{dr}_{gk}$)
RMSE($\bar{p}^{dm}_{gk}$(T),$\bar{p}^{dr}_{gk}$)
RMSE($\bar{p}_{gk}$(T),$\bar{p}^{dr}_{gk}$)
RMSE($\bar{p}^{dr}_{gk}$(T),$\bar{p}^{dr}_{gk}$)
RMSE
0.00
0.05
0.10
0.15
0.20
A1
A2
A3
A4
A5
A6
A7
Aggregation levels

RMSE for CLC - sampling error



Altitude and density



Variance of the difference of the means (Teruti-Lucas points vs. all points)

- Predict at aggregated levels by aggregating estimated probabilities.
- Use more points than Teruti-Lucas to estimate the probabilities at the point level.
- Work in progress : allocation methods.

This work was supported by the French Agence Nationale de la Recherche through the ModULand project (ANR-11-BSH1-005).

- Predict at aggregated levels by aggregating estimated probabilities.
- Use more points than Teruti-Lucas to estimate the probabilities at the point level.
- Work in progress : allocation methods.

This work was supported by the French Agence Nationale de la Recherche through the ModULand project (ANR-11-BSH1-005).
We thank

- the Service de la Statistique et de la Prospective from the Ministère de l'Agriculture in France for giving us access to the data Teruti-Lucas,
- the Unité INFOSOL from INRA for the data base BDGSF
- and the Joint Research Center-MARS from the European Commission for the meteorological data.

- Predict at aggregated levels by aggregating estimated probabilities.
- Use more points than Teruti-Lucas to estimate the probabilities at the point level.
- Work in progress : allocation methods.

This work was supported by the French Agence Nationale de la Recherche through the ModULand project (ANR-11-BSH1-005).
We thank

- the Service de la Statistique et de la Prospective from the Ministère de l'Agriculture in France for giving us access to the data Teruti-Lucas,
- the Unité INFOSOL from INRA for the data base BDGSF
- and the Joint Research Center-MARS from the European Commission for the meteorological data.

Thank you for your attention !