Land use predictions on a regular grid at different scales and with easily accessible covariates. Application to the Teruti-Lucas survey.

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

SSIAB - Rennes - May 2016

Introduction

Two papers in revision :

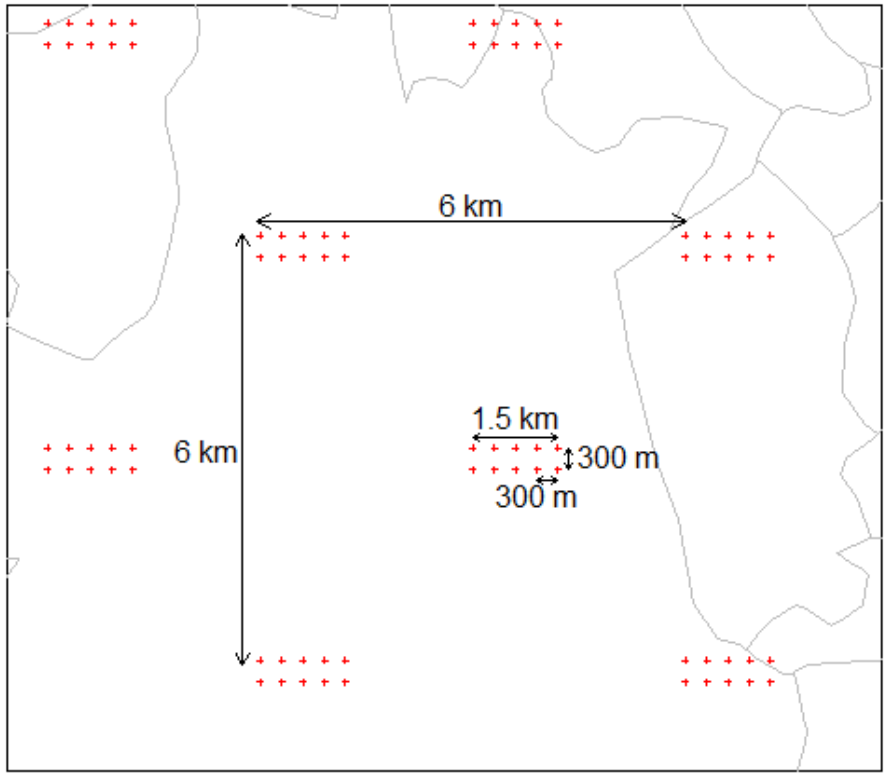
- "Prédiction de l'usage des sols sur un zonage régulier à différentes résolutions et à partir de covariables facilement accessibles"
- "Spatial scale in land use models : application to the Teruti-Lucas survey"

Introduction

The context : the Midi-Pyrénées region in France,

Introduction

the Teruti-Lucas survey in 2010 (not easily accessible),



Land use in 5 categories (urban, farming, forests, pastures, natural land) to predict.

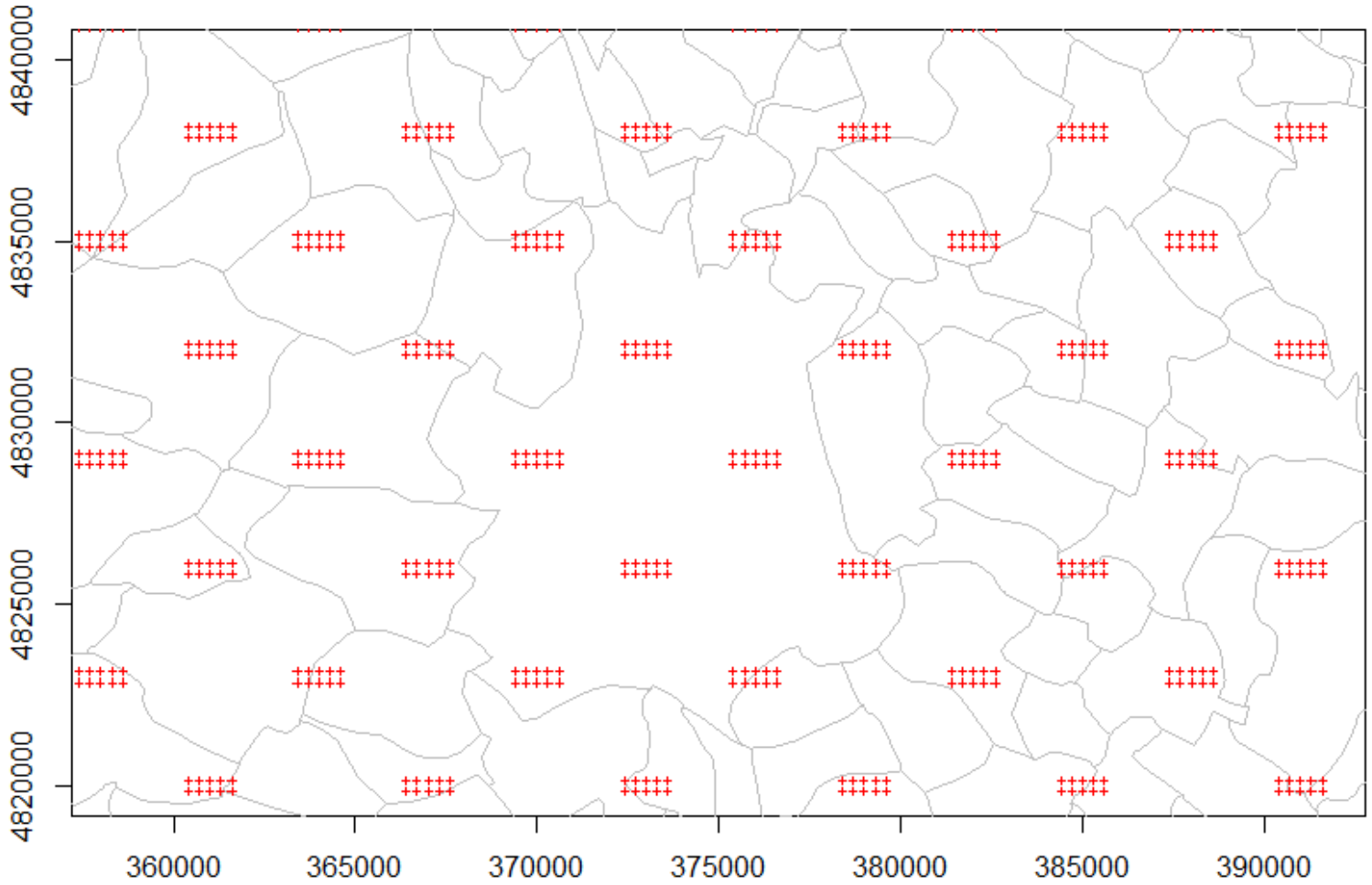
Introduction

Covariates easily accessible :

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	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	-

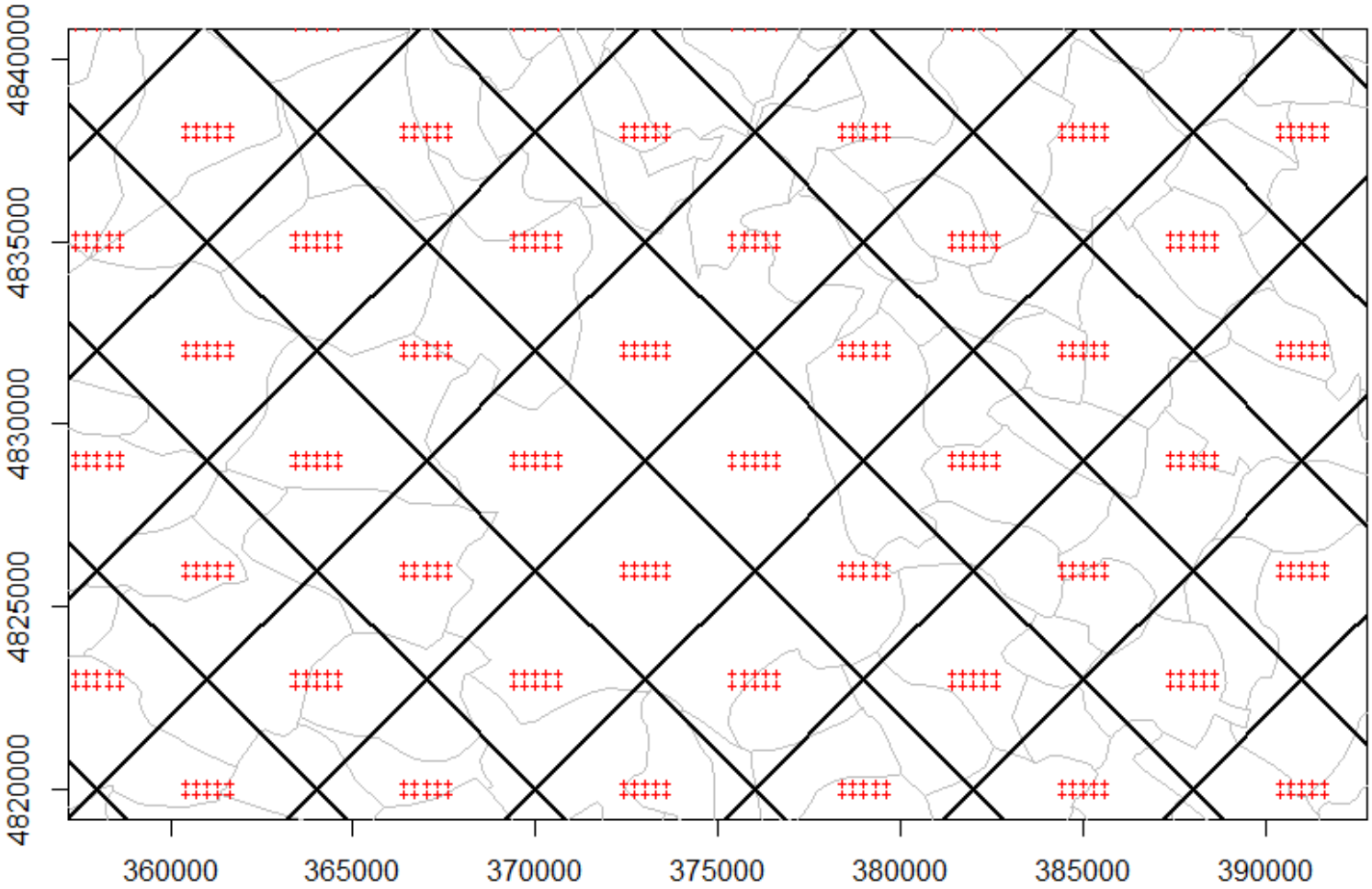
Introduction

Regular grids at different scales : Teruti-Lucas (TL) points



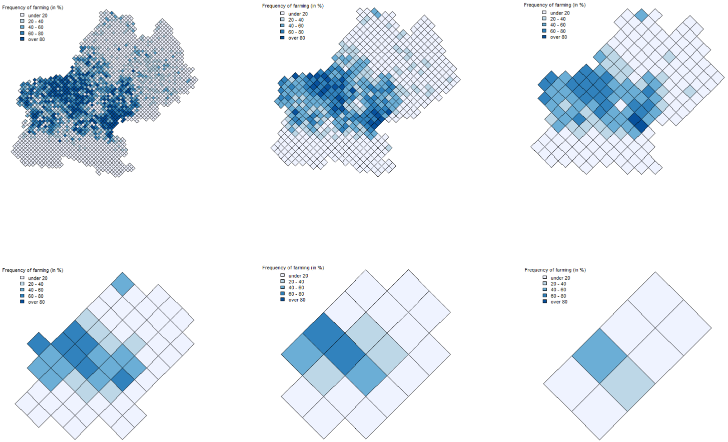
Introduction

Regular grids at different scales : grid A1



Introduction

Regular grids at different scales : from grid 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.

Introduction

1. predict TL 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. Investigate possibilities of improving predictions using a synthetic data set in order to validate the results.

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 RMSE, Brier score (1/2 MSE), weighted Brier score), at the point and grid levels .

1. Introduction
2. First study
3. Second study
   - Synthetic data set
   - Improving the predictions
4. Conclusion

Prediction of the land use at Teruti-Lucas points

- Locations $i = 1, \ldots, 25317$,
- land uses $k = 1, \ldots, K$, $K = 5$,
- explanatory variables $\mathbf{x}_i$,
- vector of probabilities $p_i = (p_{i1}, \ldots, p_{iK})$ at location i such that :

$$p_i = f(\mathbf{x}_i).$$

Prediction of the land use at Teruti-Lucas points

- Locations $i = 1, \ldots, 25317$,
- land uses $k = 1, \ldots, K$, $K = 5$,
- explanatory variables $\mathbf{x}_i$,
- vector of probabilities $p_i = (p_{i1}, \ldots, p_{iK})$ at location i such that :

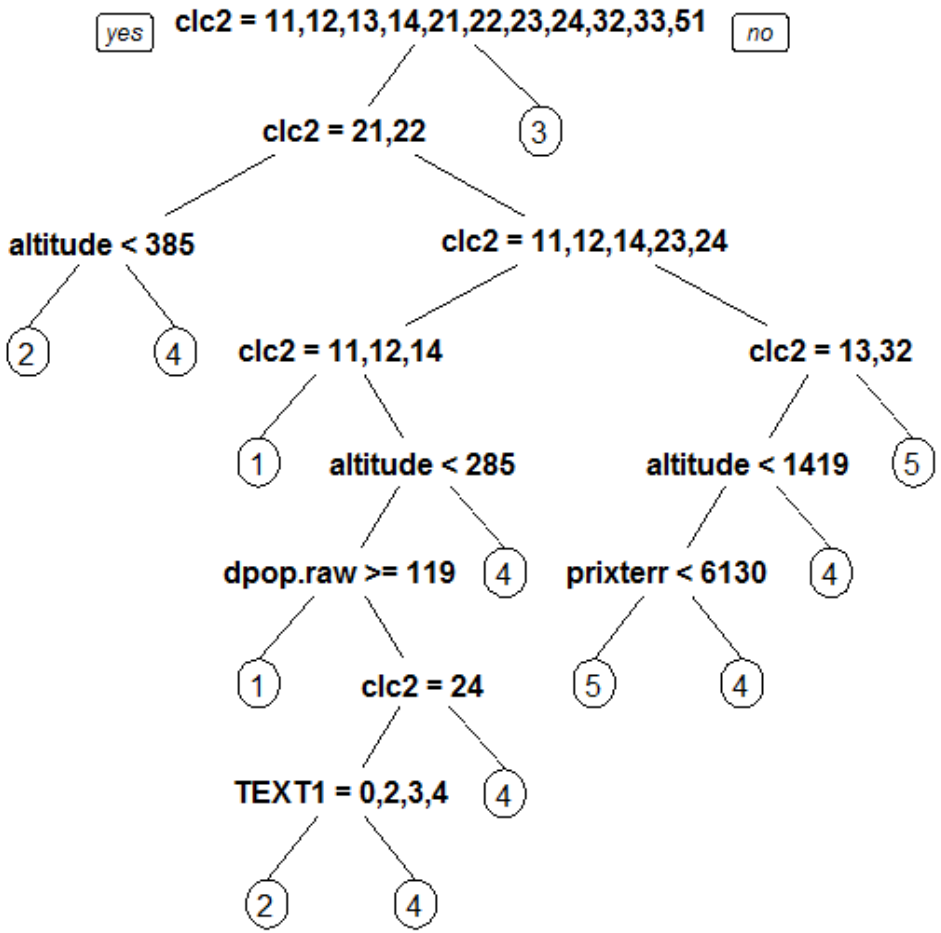
$$p_i = f(\mathbf{x}_i).$$

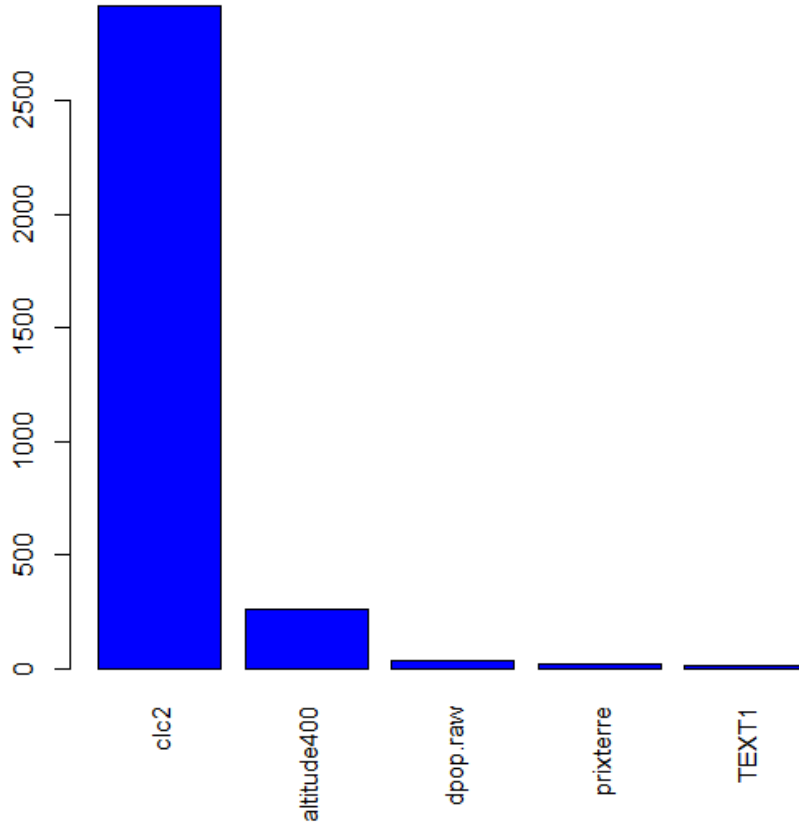
- The variable to explain is a vector with components denoted d_{ik}^r (dummy variables) following a multinomial distribution with parameters 1 and p_i,

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.
- We compared at the grids levels, different strategies (aggregate the estimated probabilities, predict the land use with the maximum probability, use a multinomial draw).

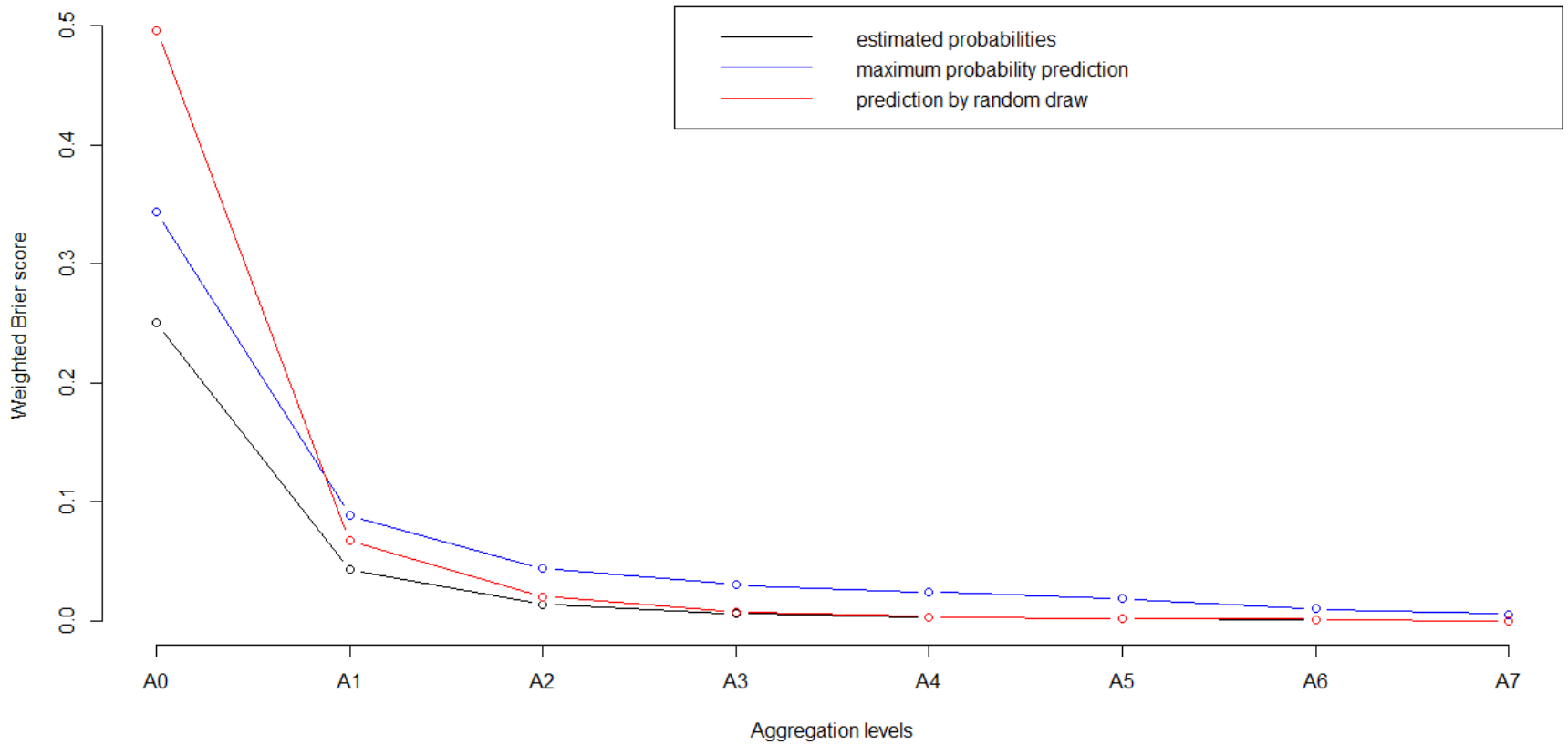
Classification tree and importance of variables





Results of the classification tree

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



Results

- We tried also some spatially lagged explanatory variables but, at the points level, not possible to improve the percentage of correctly classified points.
  The reason may be that the classification problem is difficult.
  If the probability vectors p_i are known, the proportion of points correctly classified is

$$\frac{1}{N}\sum_k \sum_i p_{ik}^2$$

  (1 - Bayes risk) and can be low if the p_{ik} are not close to 0 or 1.

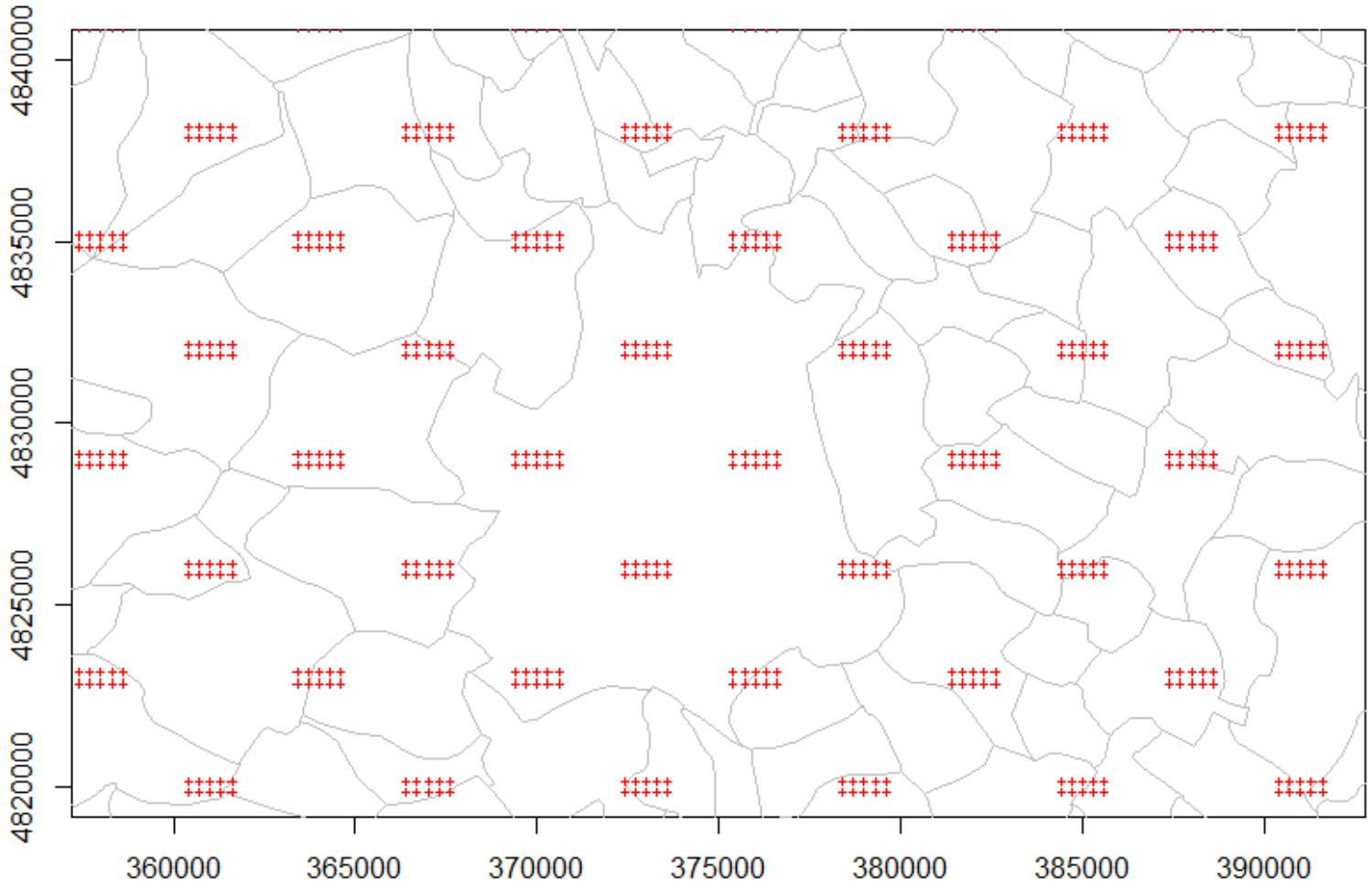
- At the grids levels, there is no need to predict the land use since it is better to aggregate estimated probabilities.

Questions

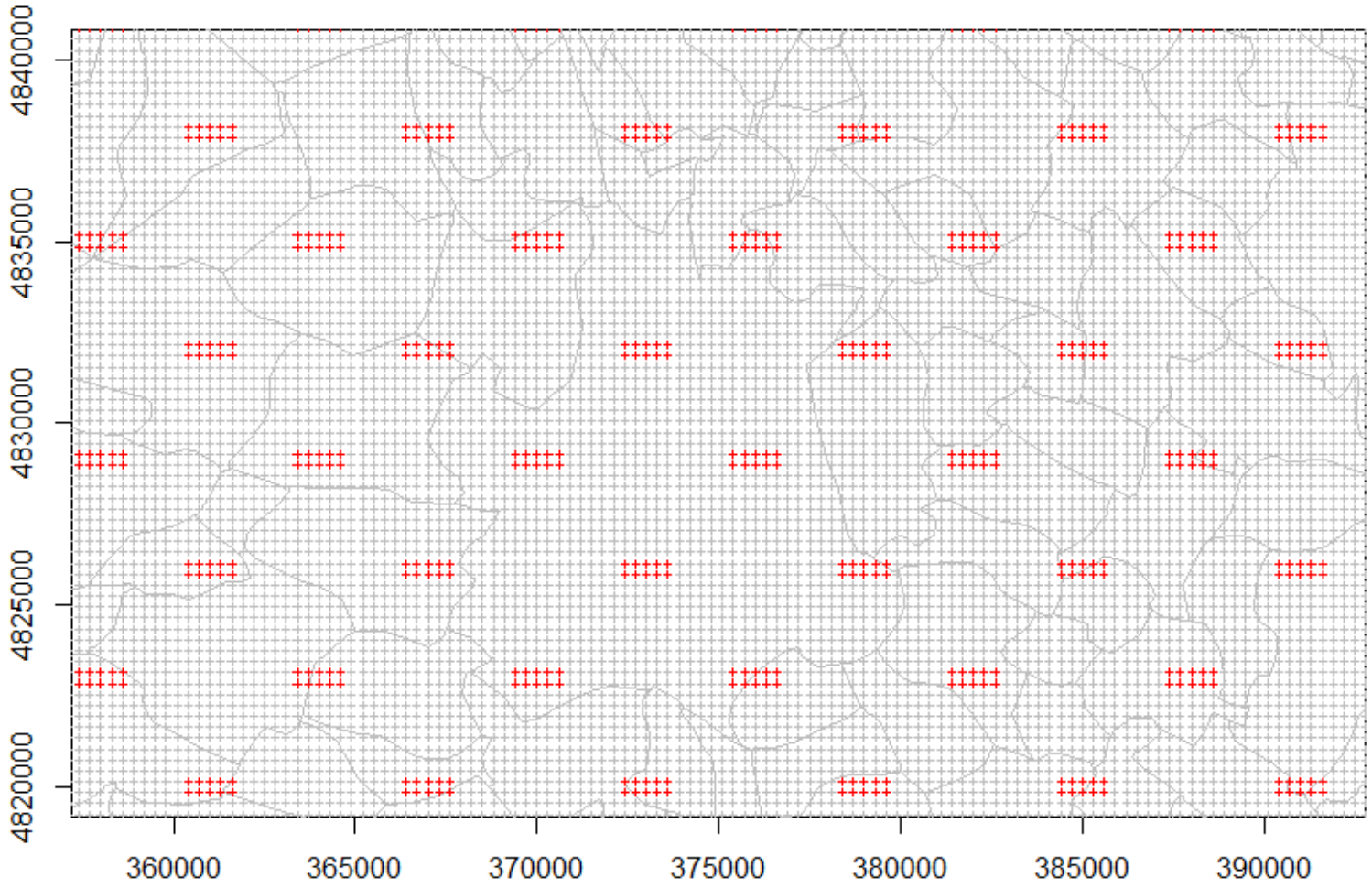
Can we expect an improvement of the results

- if we could observe more TL points to estimate the p_i ?
- if we predict at more points than TL using the fact that the explanatory variables are available elsewhere ?

More points than Teruti-Lucas (25 317)

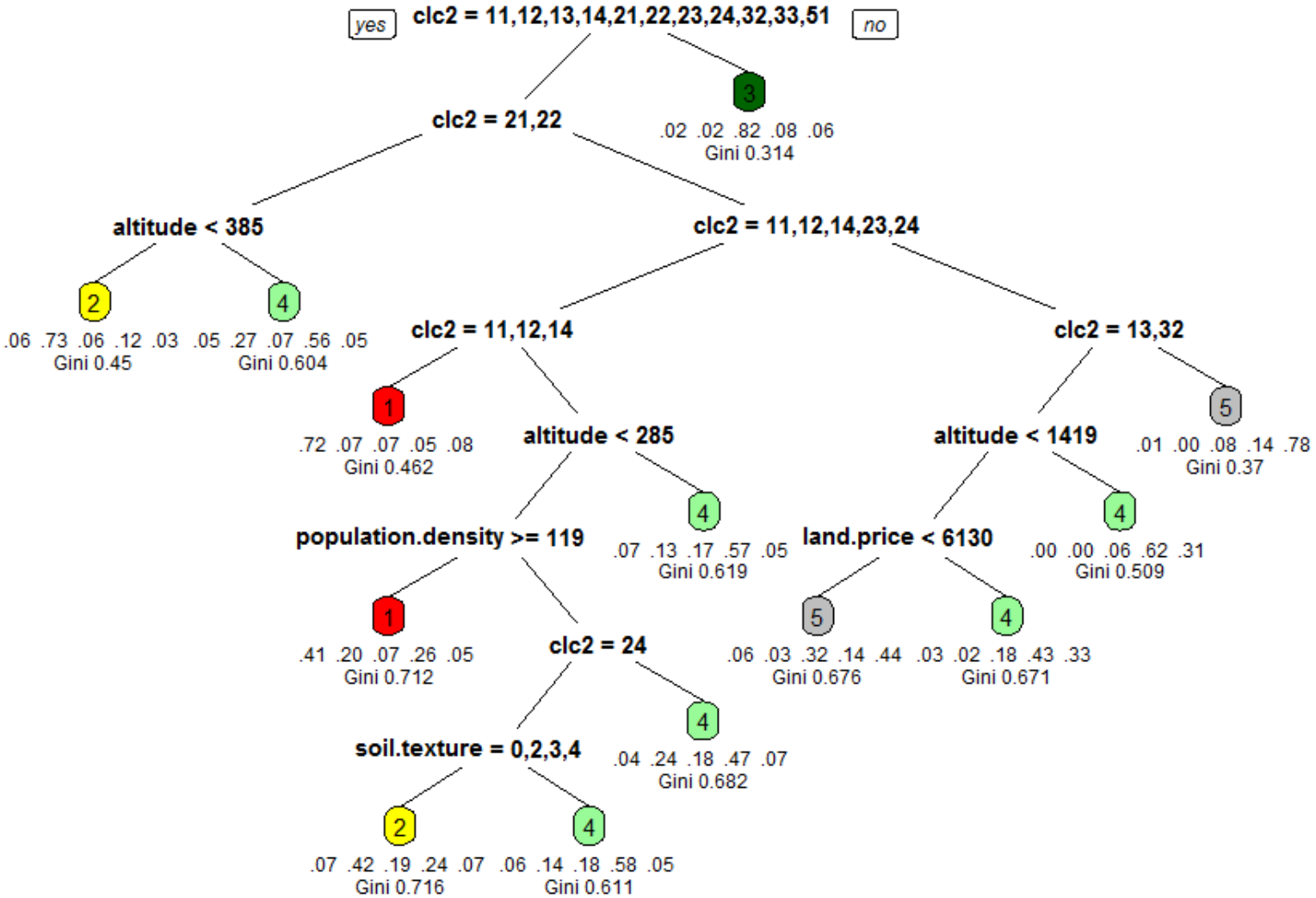


More points than Teruti-Lucas (502 205)



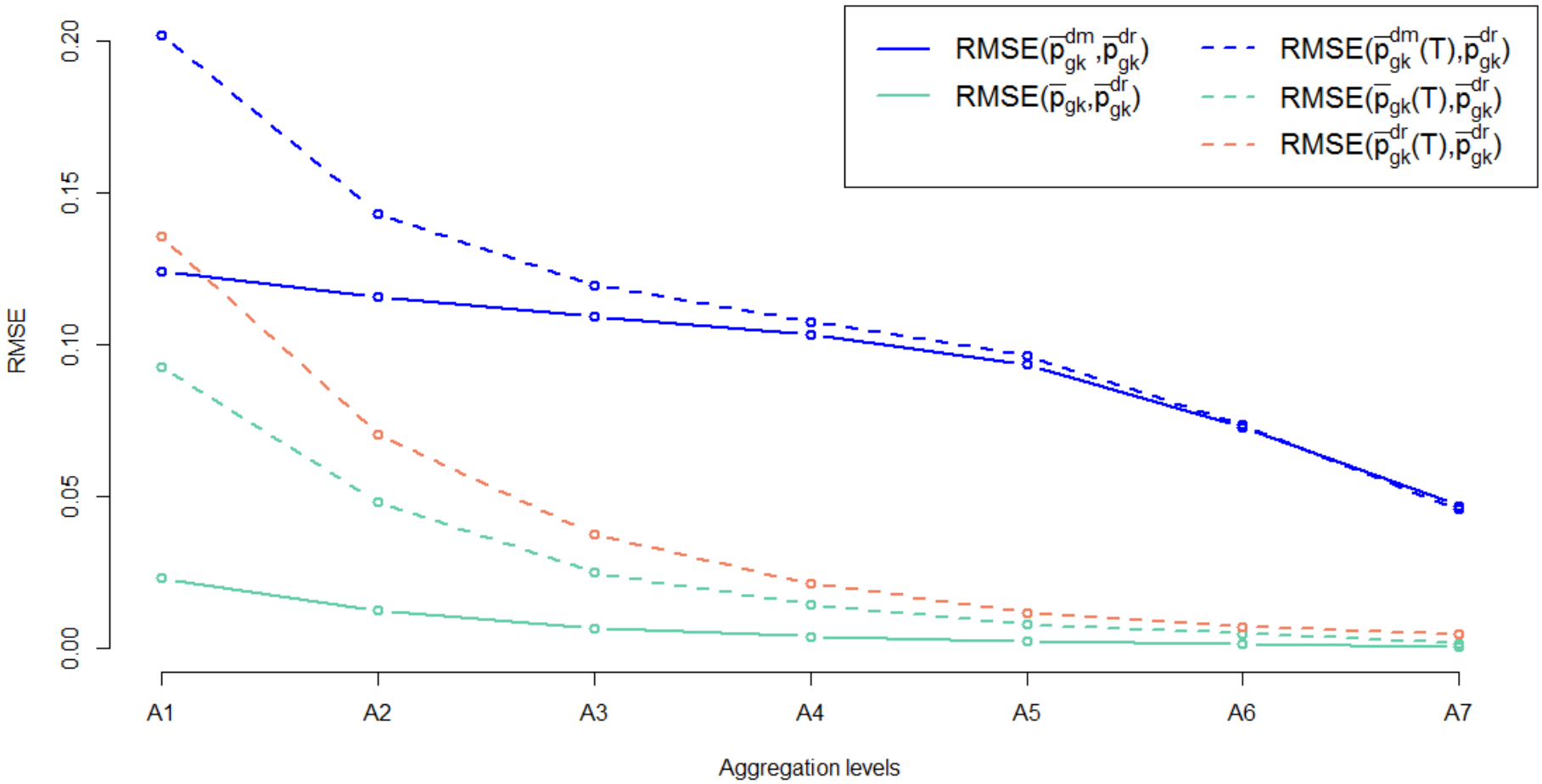
Data generation

Classification tree + multinomial draw

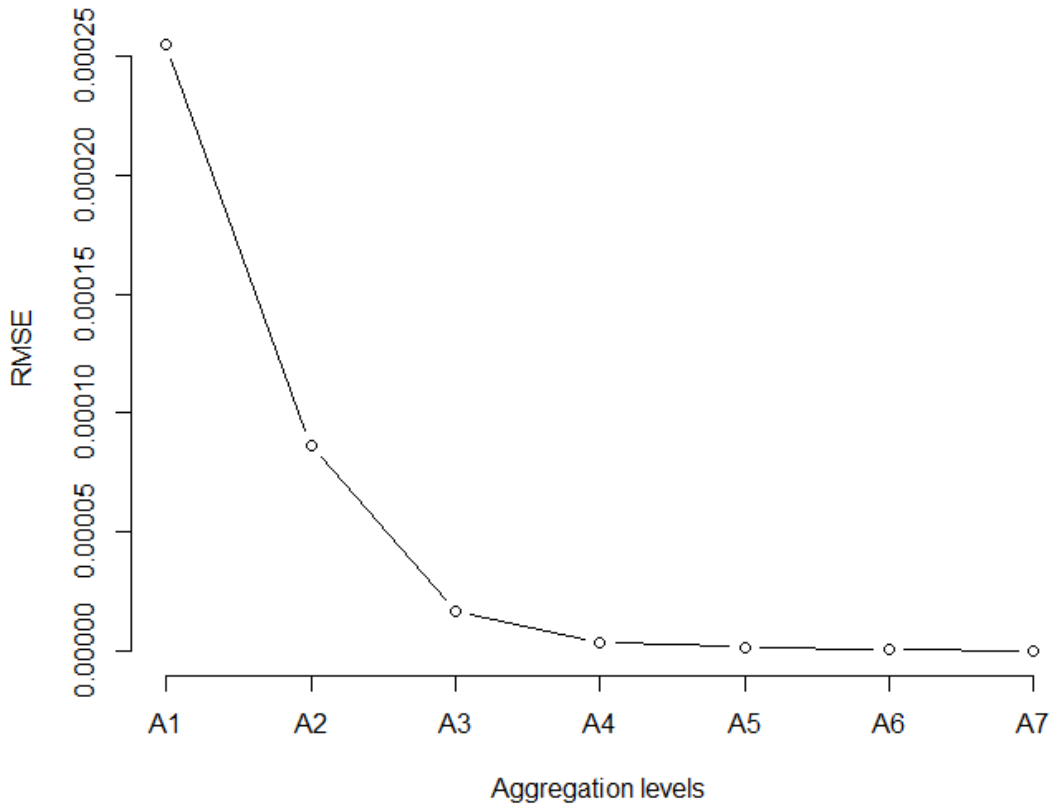


We consider the synthetic data as the truth (we forget about the TL survey).

- At the points level, we use the synthetic data at the TL locations only and at all locations. The two fitted models and the results were extremely similar. We conclude that (given our data set) there is no need to have more locations than TL to fit the model.
- So, we fit the model using only the TL locations but since the explanatory variables are available at any location, we estimate the probabilities at all locations and then aggregate the probabilities at the different grids levels.

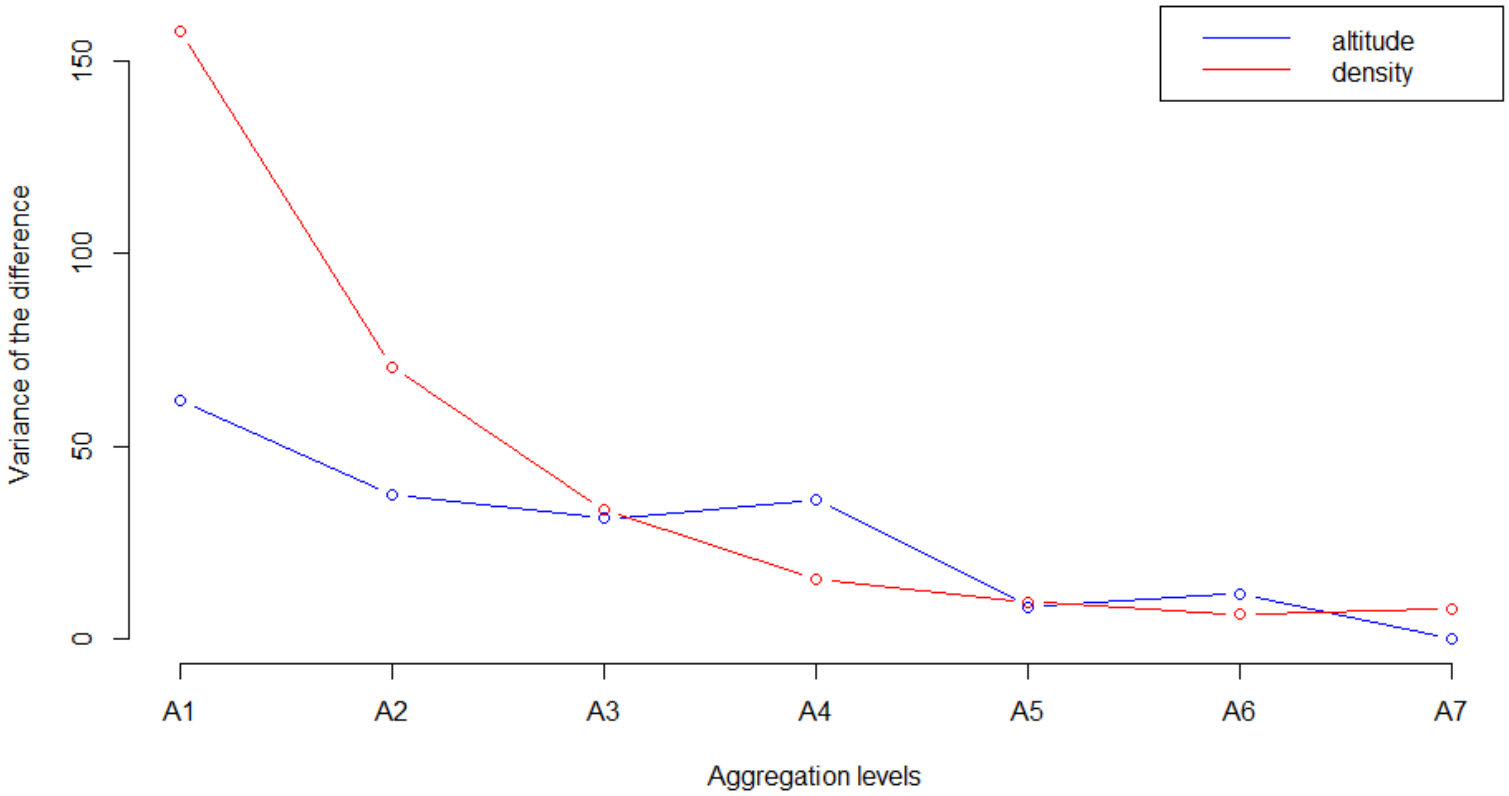

RMSE($\overline{p}_{gk}^{dm}$,$\overline{p}_{gk}^{dr}$)
RMSE($\overline{p}_{gk}$,$\overline{p}_{gk}^{dr}$)
RMSE($\overline{p}_{gk}^{dm}$(T),$\overline{p}_{gk}^{dr}$)
RMSE($\overline{p}_{gk}$(T),$\overline{p}_{gk}^{dr}$)
RMSE($\overline{p}_{gk}^{dr}$(T),$\overline{p}_{gk}^{dr}$)
RMSE
0.00
0.05
0.10
0.15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

CLC



Root mean squared of the difference between the proportions of land use (TL locations vs. all points)

Altitude and density



Variance of the difference of the means (TL locations vs. all points)

Conclusion

- Predict at aggregated levels by aggregating estimated probabilities.
- Use more points than Teruti-Lucas to estimate the probabilities at the point level.
- More in the papers (decomposition of the prediction error and spatial analysis of the error).
- Literature ?
- Work in progress : allocation methods and CODA models at the grids levels.

Acknowledgements

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.

Acknowledgements

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 !