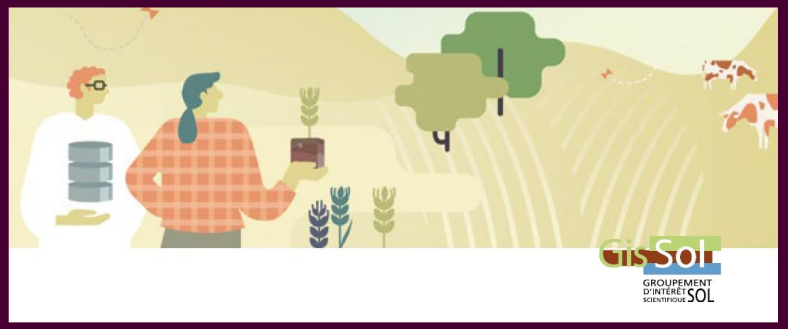
The French agricultural soil database: mine of information and headache for statisticians

N Saby, B Lemercier, F Munos, D Arrouays

Journée RESSTE

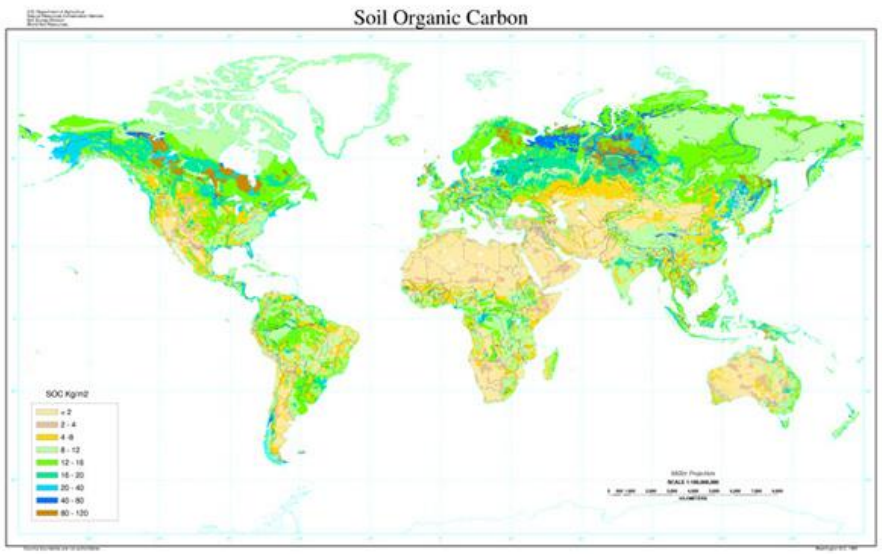


_01

Introduction

SOILS VARY IN SPACE AND TIME !

Soils vary in space & time !



USDA Soil organic maps

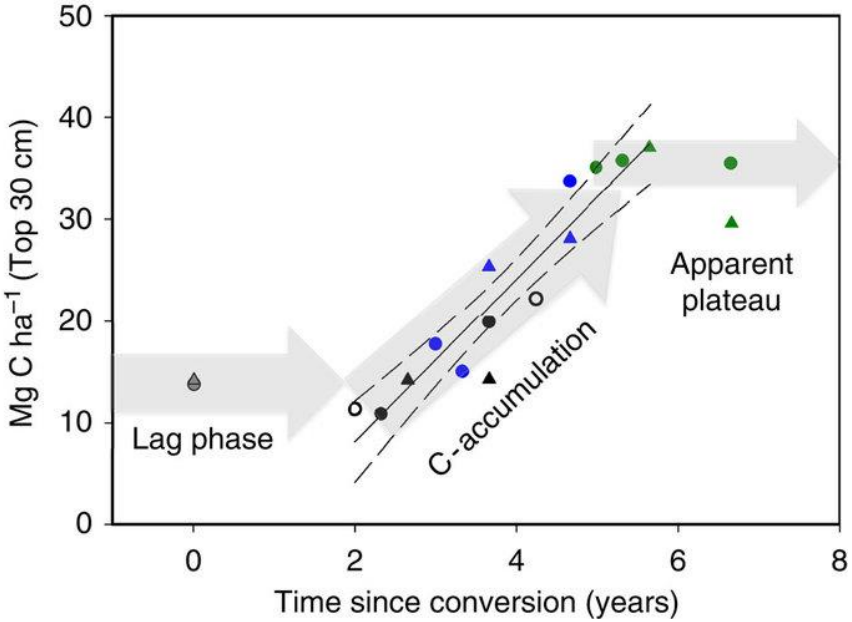


Figure 2: Soil carbon rapidly increases with conversion of row crop to intensive grazing.

Machmuller et al., 2015, *Nature Communications* **6**,

Context

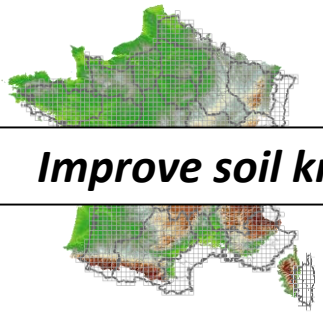
- High demand for soil monitoring, for example

Soil information system The « GIS Sol »

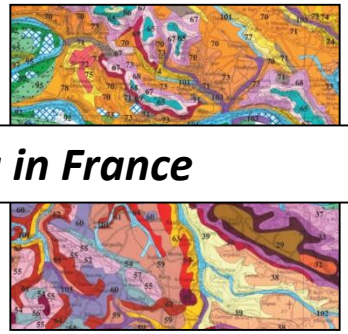
Four main soil survey and monitoring programmes

GIS Sol

RMQS

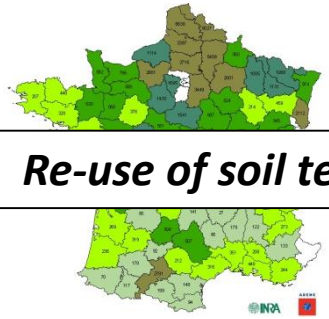


IGCS

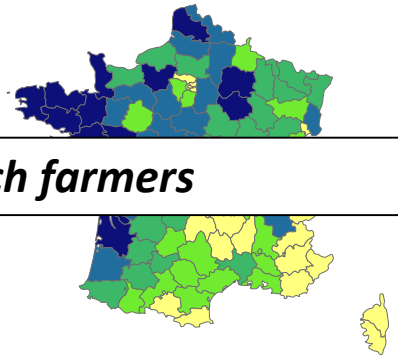


Improve soil knowledge and monitoring in France

BDETM



BDAT



Re-use of soil tests requested by French farmers

Outline

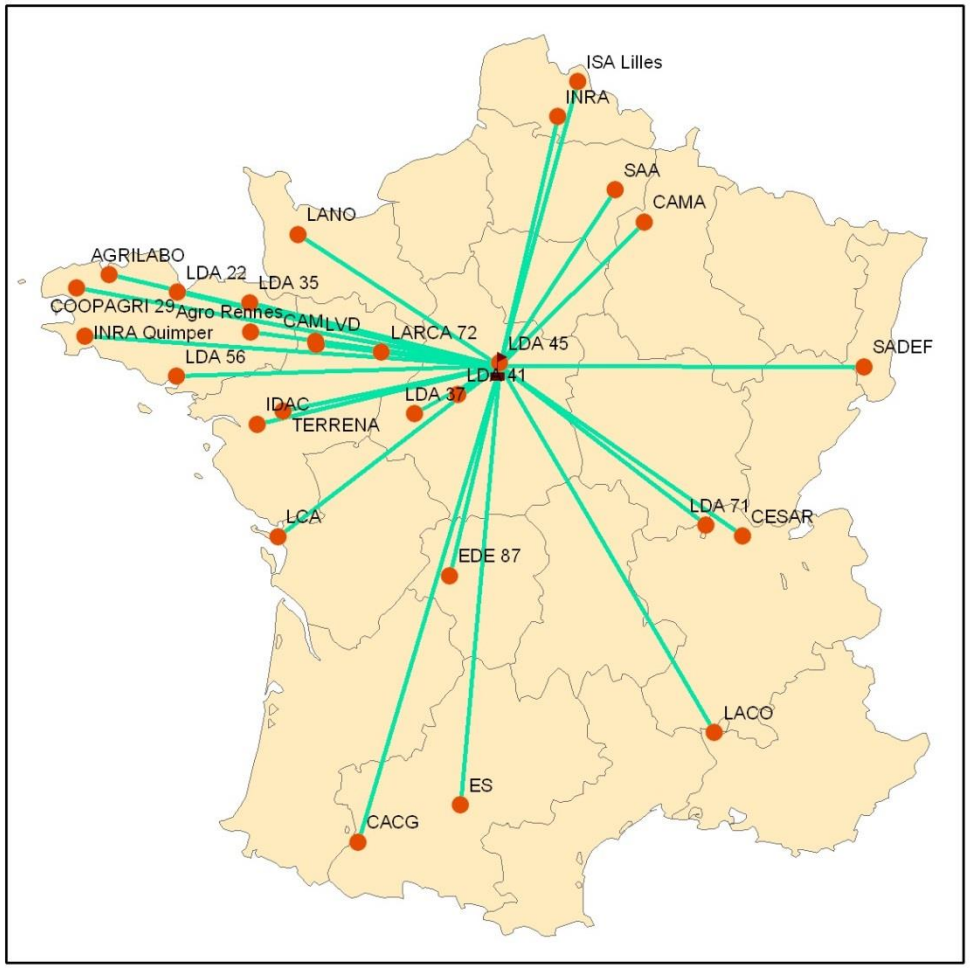
- The dataset
- The questions
- A few examples of what has been done

BDAT

The Soil Testing database

NATIONAL PROGRAM FOR SOIL MONITORING OF AGRICULTURAL SOIL IN THE FRAMEWORK OF THE GIS SOL

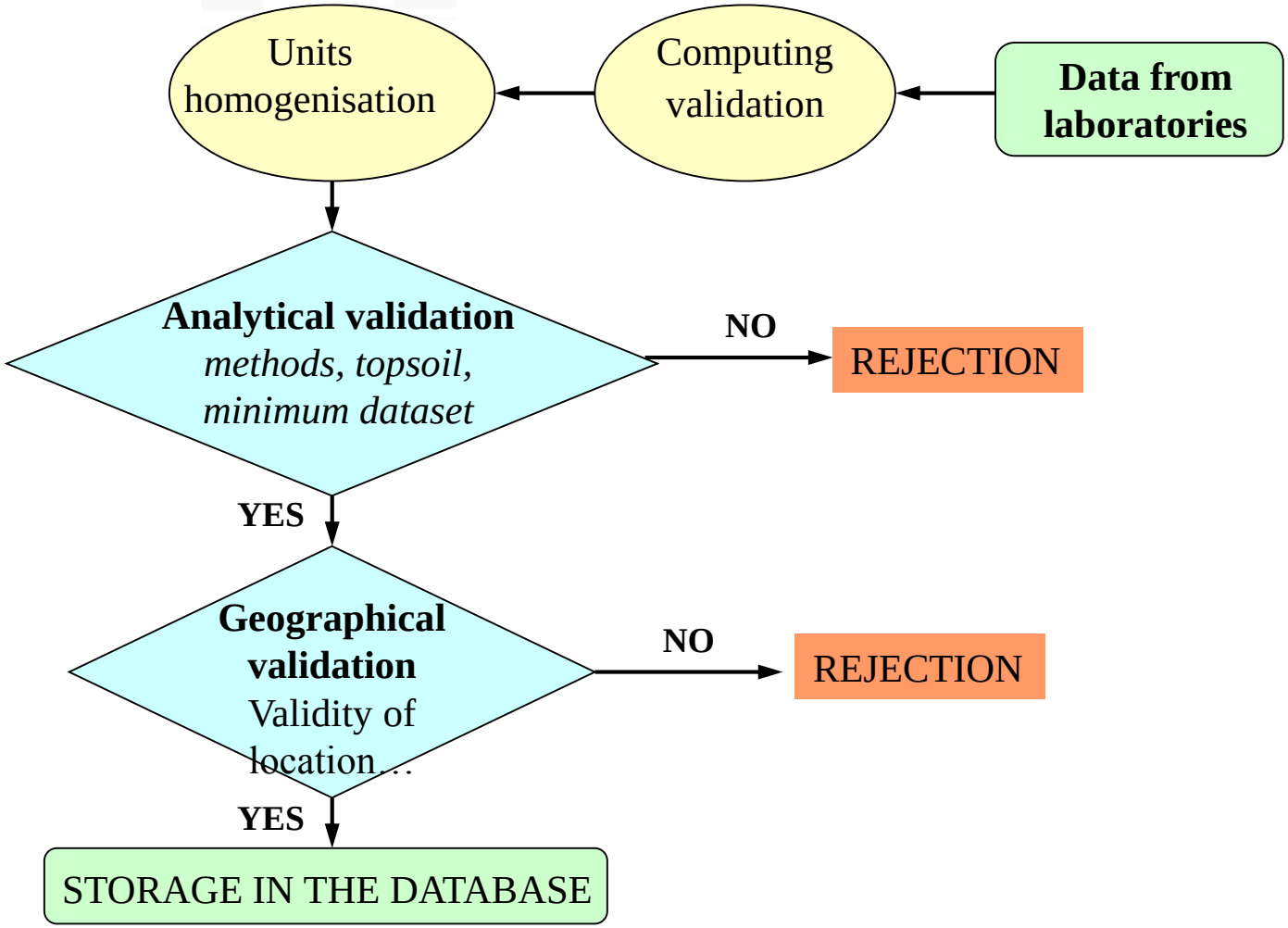
Soil testing for farmers



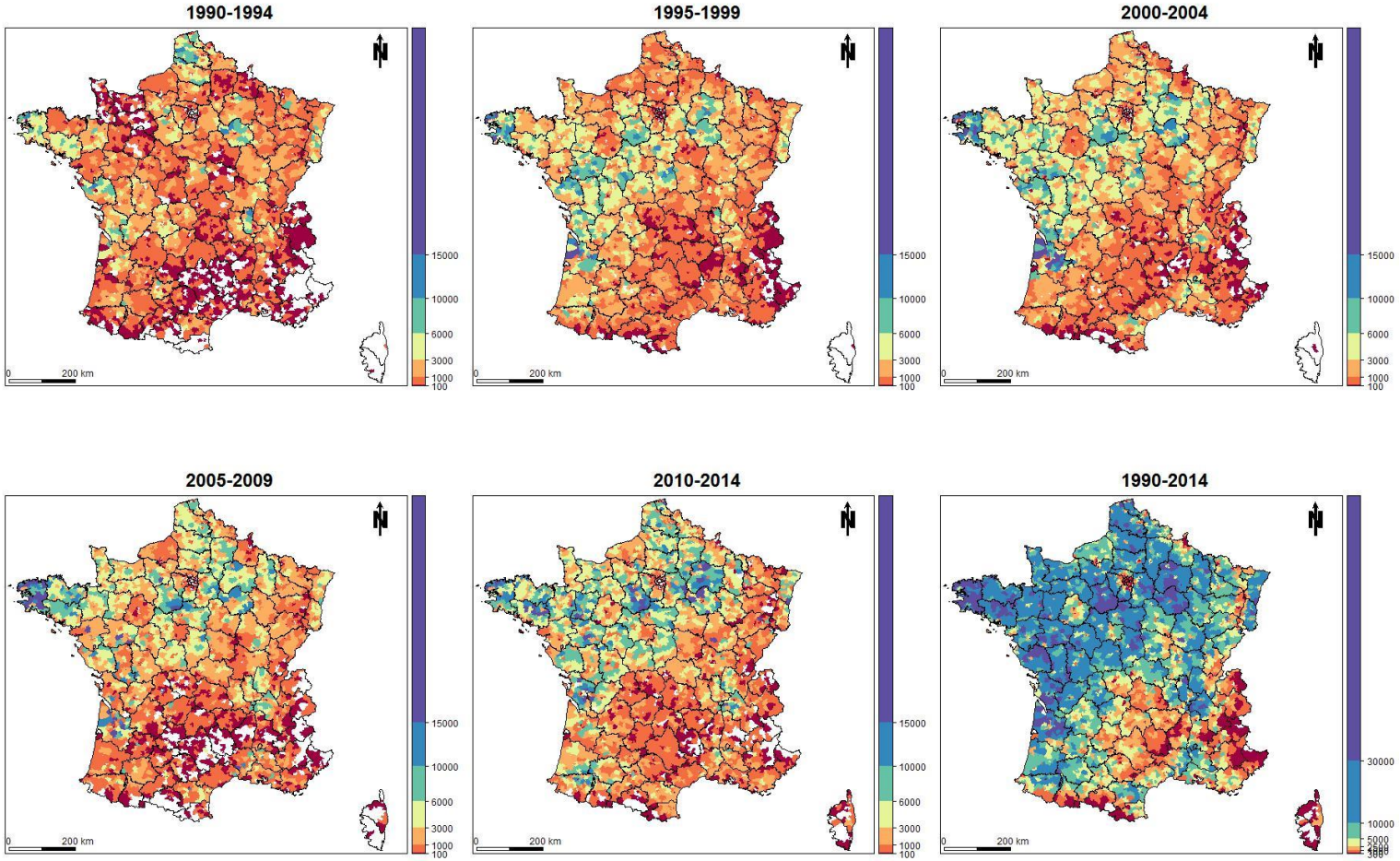
Characteristics of the data

- **Georeferencing:** imprecise – origin of the sample based on administrative area (municipality) for privacy reasons
- **Sampling:** no control of the overall sampling strategy (no detailed information on how, where and when).
- **Analytical procedures:** identical for the selected laboratories : standardisation
- **Available data in the database:**
  - Particle Size analysis
  - Organic C and N, pH, CEC,
  - Macronutrients (P K Ca Mg)
  - Micronutrients
  - Basic information on land use

Data assimilation procedure



In the BDAT data base:
Analytical determinations : 31,000,000
Samples: 2,608,880

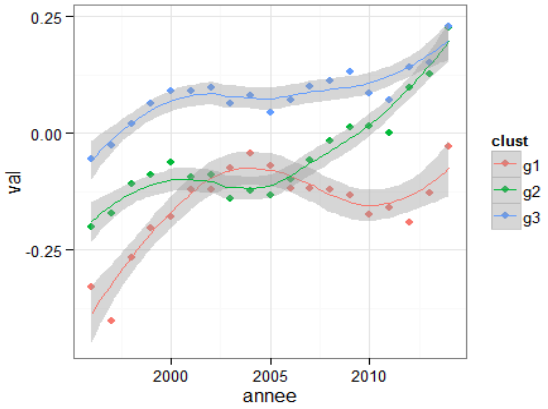


1

Scientific questions and statistical options

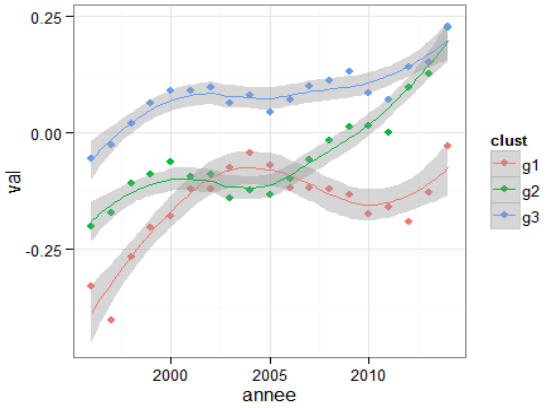
Scientific questions

- Is there any global or local **spatio temporal** trend in the properties of agricultural soils ?
- Is it possible to highlight the main drivers ?

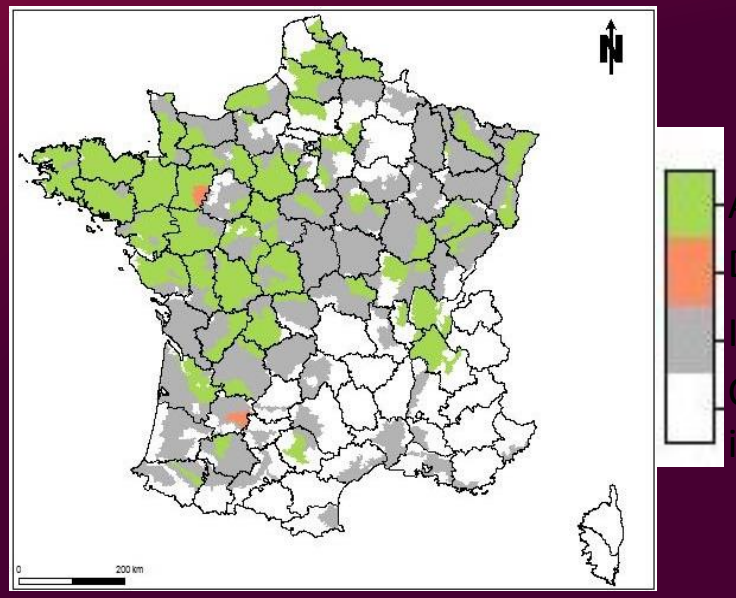


Some modelling challenges

1. Presence of outliers
2. Skewness of the marginal distributions
3. areal support of the data
4. Uncontrolled sampling strategy (sampling resolution in space and time)
5. a flexible spatio temporal model : eg not necessarily linear
6. Summarize the trend
7. Link to possible drivers (areal support)



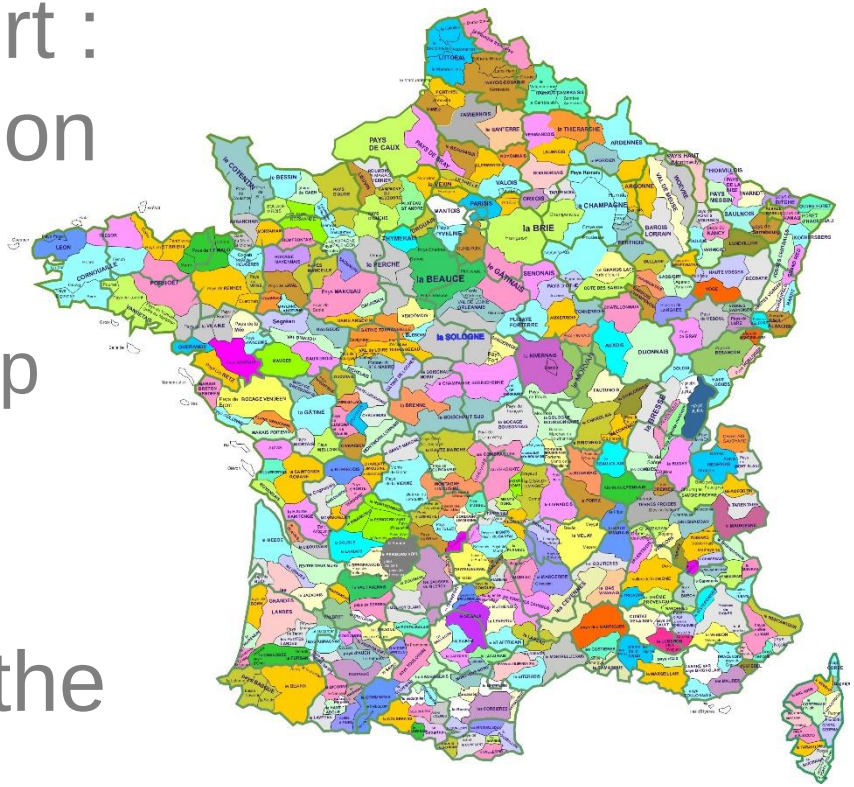
Trend



Run multiple non parametric statisical tests

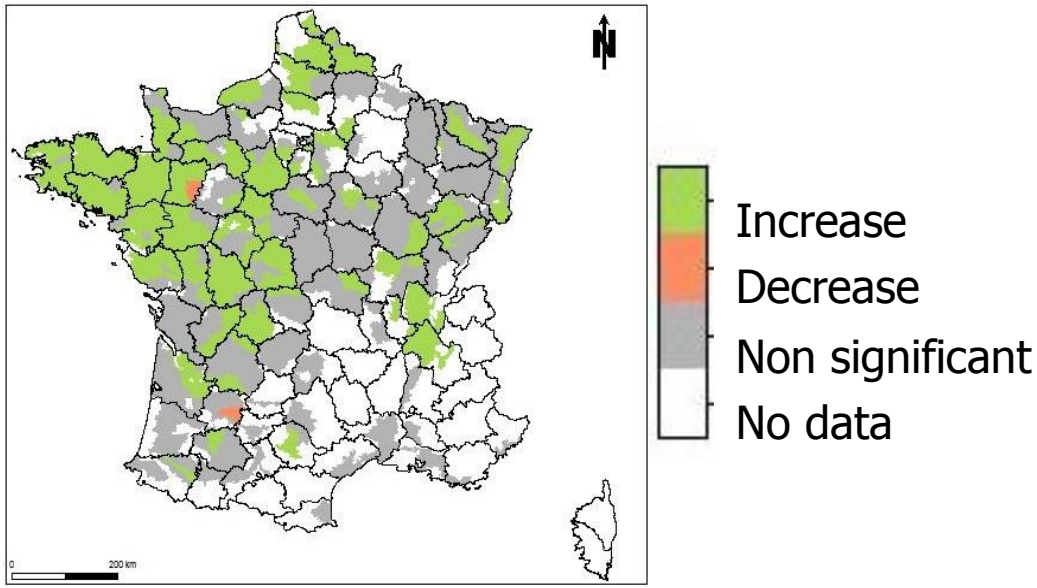
Statistical algorithm

- Two time periods
- Spatial decision support : Small Agricultural Region
- Statistical approach
  - MCMC Resampling step
  - Mann–Whitney tests
- Maps of the results of the inferences



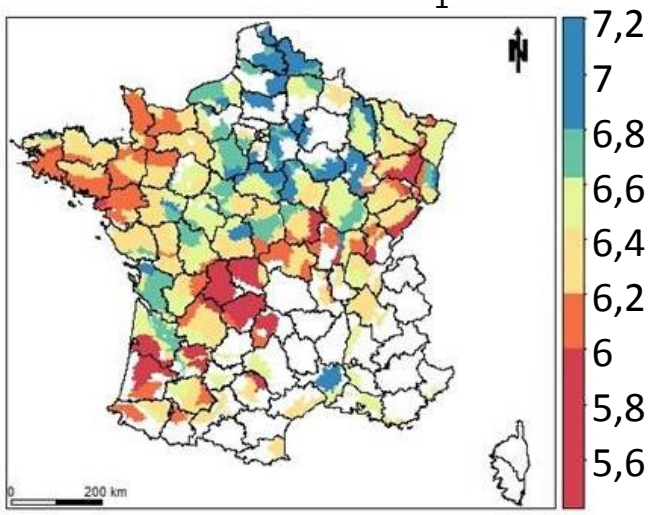
Trend in pH

Trend

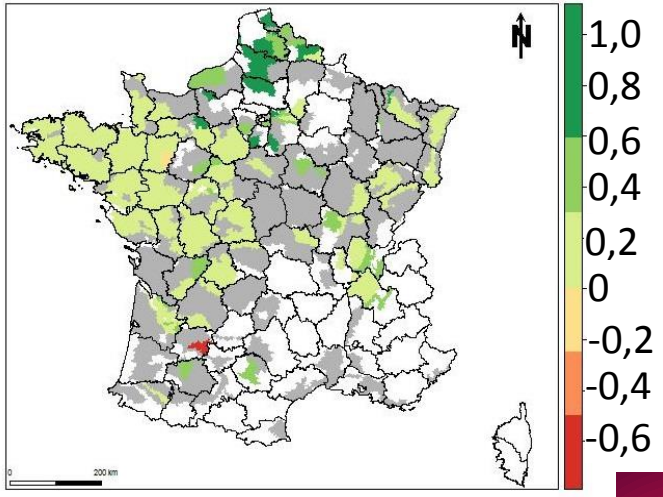


general increase in soil pH

Médiane à t_1



Médiane de la variation

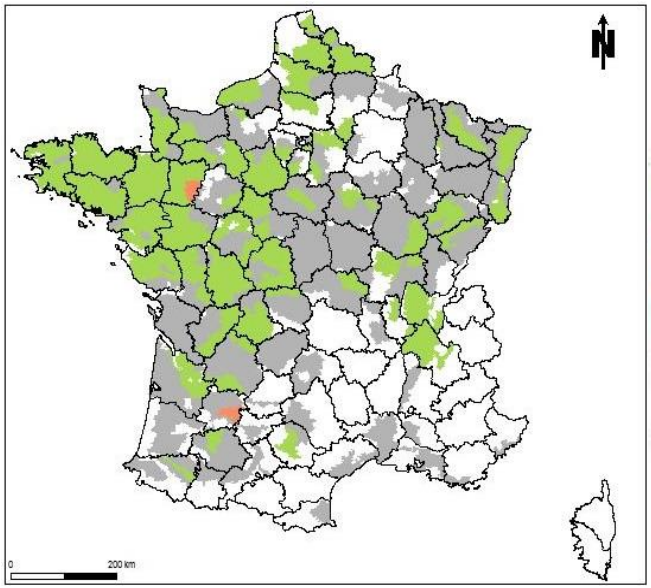


Saby NPA. et al., submitted, Soil Use Manag

Limitations

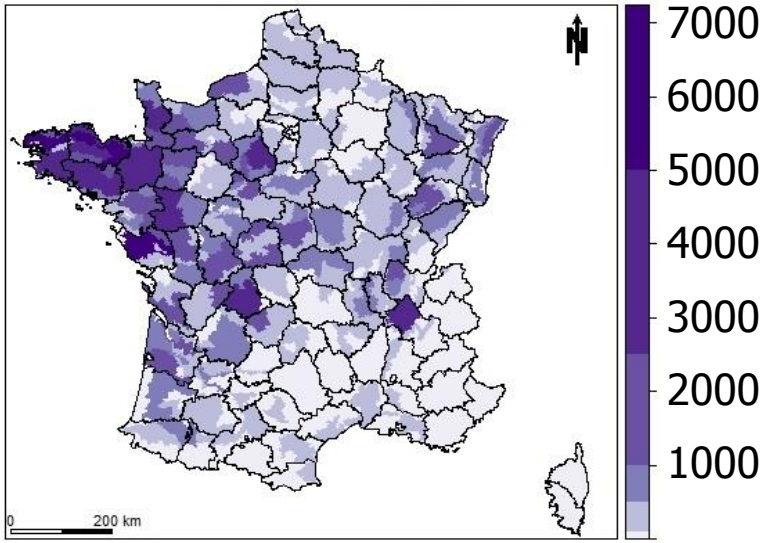
- Power of the test = f(n)

Changes



Increase
Decrease
Non significant
No data

Number of samples

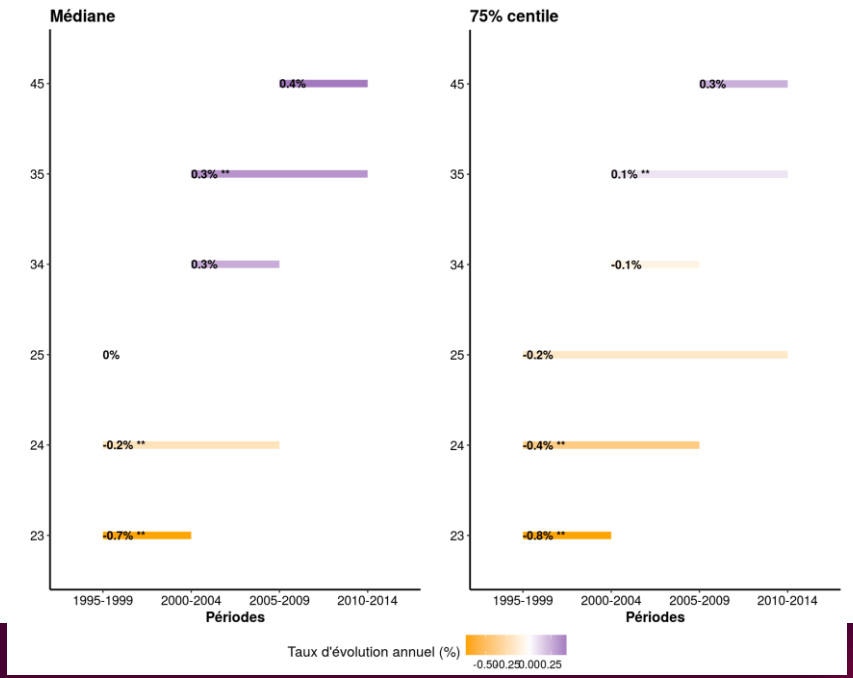


Saby NPA et al., submitted, Soil Use Manag

(Swiderski et al, 2008, in prep.)

Limitations

- Only pairwise comparison : Period 1 vs period 2
- Challenging when the number of pairwise comparisons increases
- No spatial correlation

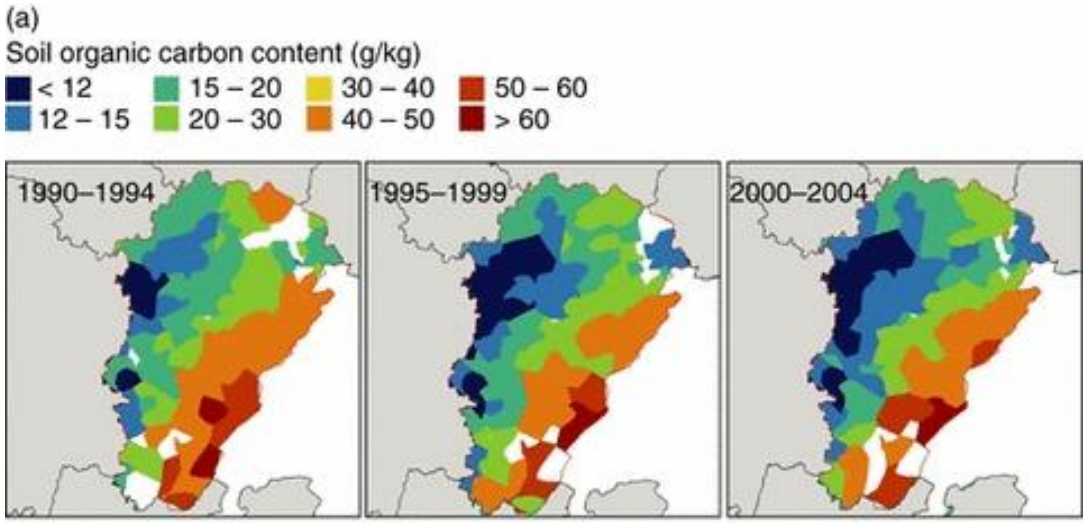


3

How to take into account spatial correlations

A geostatistical approach

- fit a linear model of coregionalisation between 2 dates
- Case study in Franche Comté



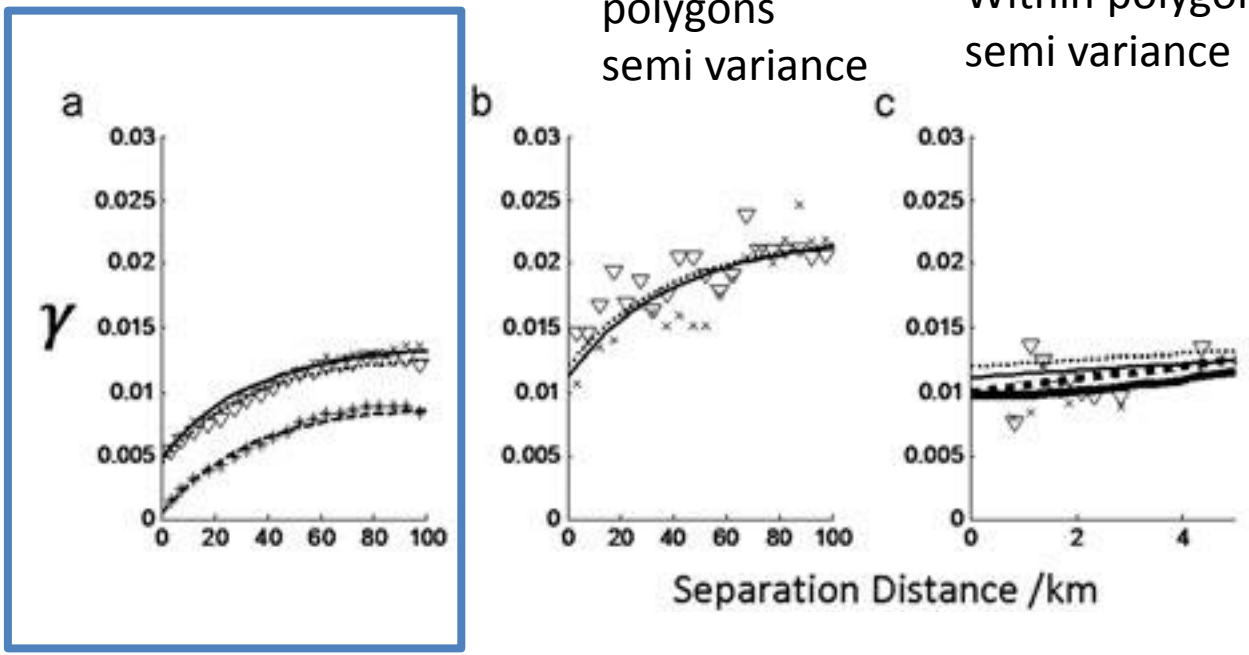
Orton et al., Environmetrics (2012)

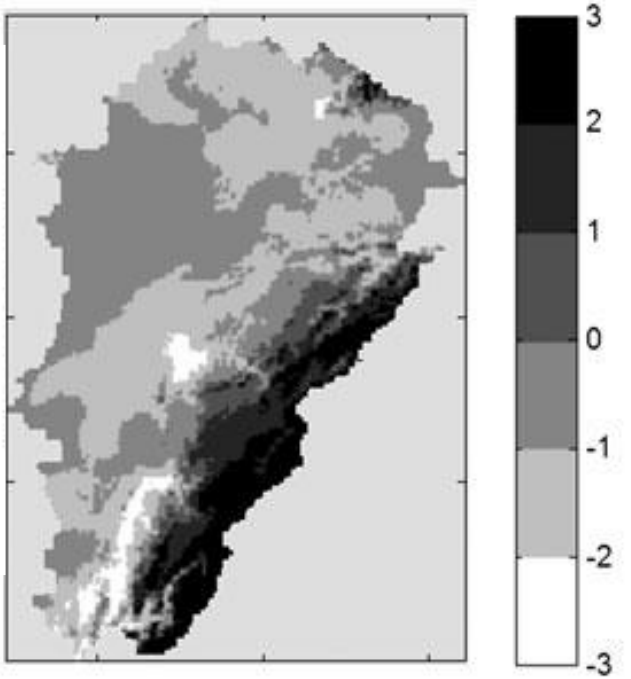
Geostatistics

- Area to point Universal co kriging approach

Between polygons semi variance

Within polygons semi variance





Orton et al., Environmetrics (2012)

Discussion

- Challenge to fit a LMCR for the whole territory (non stationarity of the variance)
- Challenge to use REML with huge matrix (n=36000 !) => need to use composite likelehood
- Only pairwise periods comparison : Period 1 vs period 2
- Skewed data

Bayesian hierarchical model

An INLA approach

- Accounting for the overdispersion : Student's *t* likelihood with 3 degrees of freedom
- Accounting for spatial autocorrelation: Besag's Intrinsic Conditionally Autoregressive process
- Spatially varying slope and intersept
- INLA

$$
\begin{aligned}
y_{ij} &\sim \mathrm{t}_3(\mu_{ij}, \sigma_y^2) \\
\mu_{ij} &= (\alpha_0 + \alpha_i) + (\beta_0 + \beta_i) j \\
\boldsymbol{\alpha}, &\sim \mathrm{N}(\mathbf{0}, \sigma_\alpha^2 \mathbf{Q}) \\
\boldsymbol{\beta} &\sim \mathrm{N}(\mathbf{0}, \sigma_\beta^2 \mathbf{Q})
\end{aligned}
$$

Munoz et al., in prep

- Maps the trend

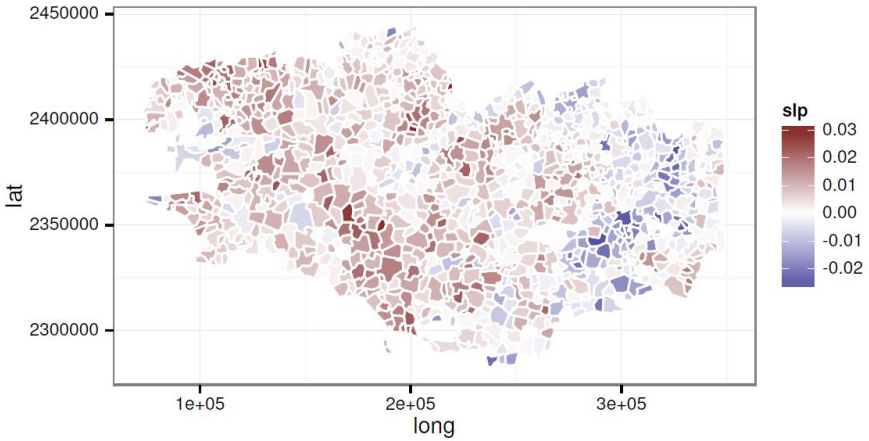
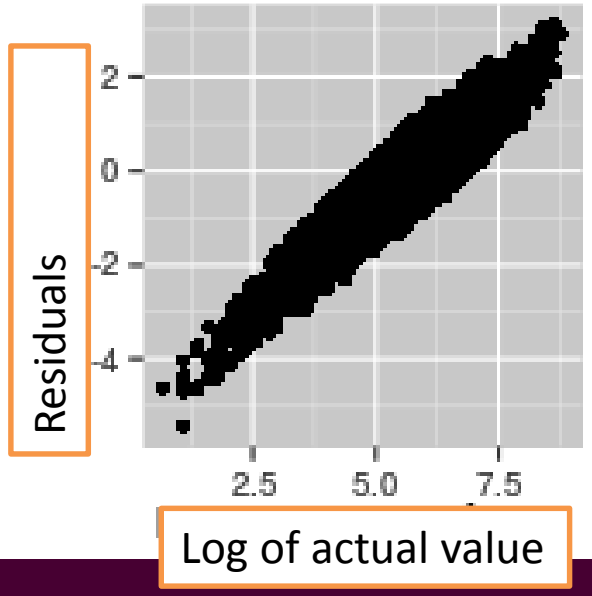


Figure 10: plot of chunk stm3-maps-slp

- The fit is not great

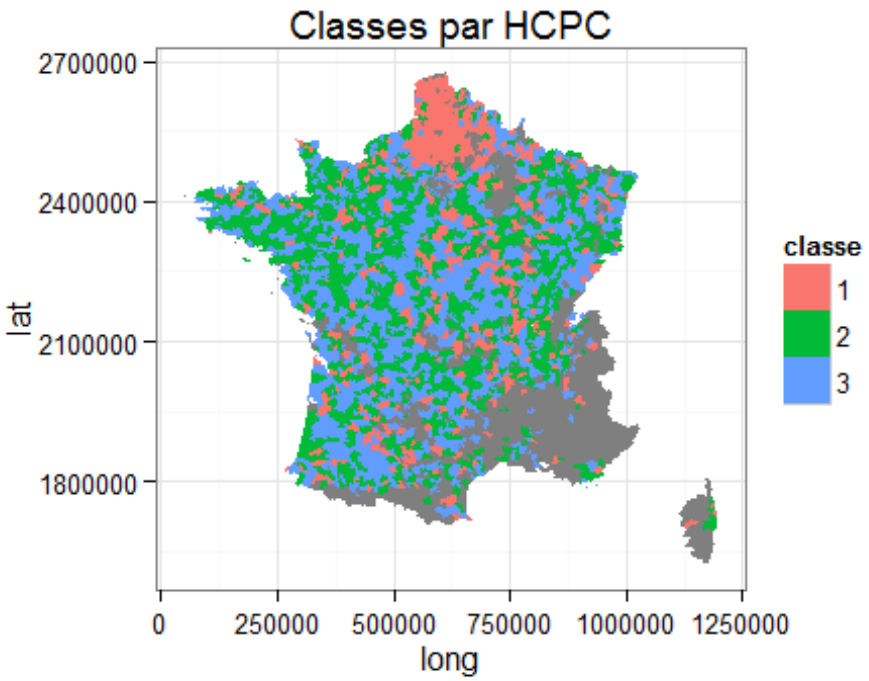


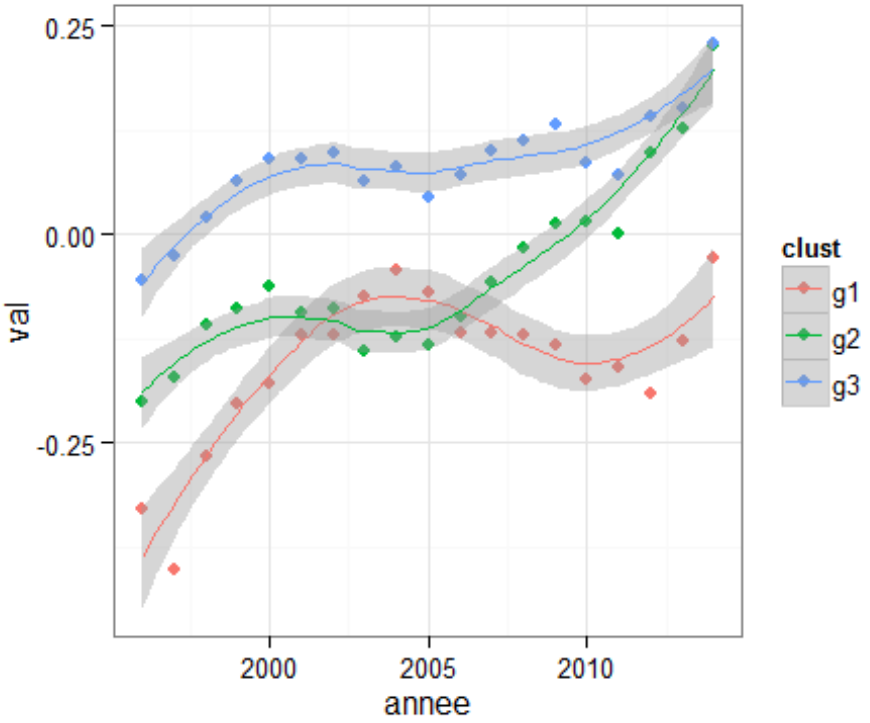
A last step ?

HOW TO SUMMARIZE THE TREND ?

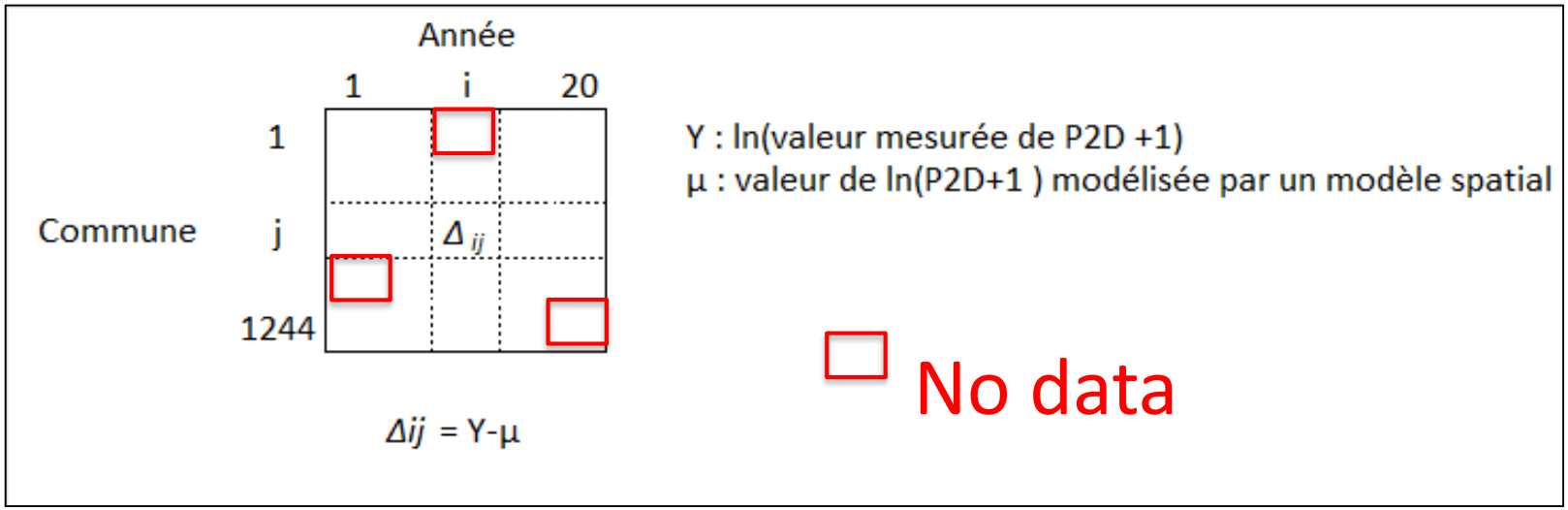
A last step :How to summarize the trend

- Hierarchical clustering on the Principal Components of the full spatio temporal matrix





- How to estimate the full data matrix



- Need a spatial clustering algorithm to make cluster more compact (Allard et al.)

Conclusions

Conclusions

- To address
  - areal support of the data: overdispersion of the residuals,
  - The fit of a flexible spatio temporal model
  - the possible bias linked to the uncontrolled sampling strategy
- And what about fitting a Multivariate model ?

Thank you !