Core data model & French insights

—

WP5 meeting, March 18th, Madrid

Benjamin Sakarovitch
INSEE

1. The core data model by Positium
    a. main steps
    b. pros...
    c. ... and cons
2. Investigation on a 5 months individual CDR dataset
    a. location of events or spatial interpolation
    b. home detection
    c. from SIM cards to overall population
    d. lessons learnt

The Core Data Model

from Positium report

The core data model by Positium

Goal : a **continuous description of the location** of every user that is aggregated according to the indicator to be produced

- detecting **anchor points**
  - place of residence
  - second home anchor points
  - work time anchor points
  - regular anchor points
  - + country of residence and usual environment
- **continuous description** of movements and locations
  - identification of stay sections vs movement sections
  - identification of trips and transit points

Pros...

- a comprehensive and refined model, that has proved useful to an advanced MPD company
- a core model that **sets the ground for different aggregation and indicators**
- a possible **common methodology** for different NSIs
- thus enabling mirror statistics (tourism)

... and cons

- need to define beforehand the smallest geographical unit and accuracy level, no explanation in Positium report for a smart spatial interpolation
- requires a **very extended access to data at an individual level** (more than CDR)
- has not been tested by any of us
- lacks the part on extending the results to the overall population

Spatial interpolation

from the telecommunication network cells to the official zoning in NUTS and LAU

Spatial interpolation

- **Voronoi tesselation** to approximate the best coverage areas of the antennae (strong approximation)
- From Voronoi cells to local administrative units (LAU)
  - NUTS 3
  - municipality
  - subcommunal divison, IRIS (every 2000 inhabitants)
- The **“area proportion method”**
  - computing the intersection between LAU and Voronoi cells
  - interpolation according to the proportion of the Voronoi cell in the LAU

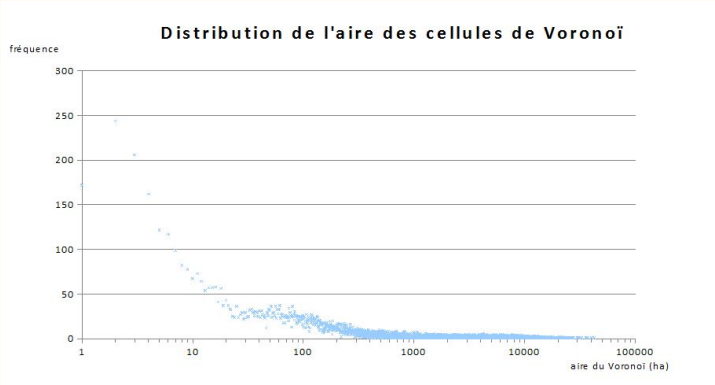
Estimating the cost of spatial interpolation

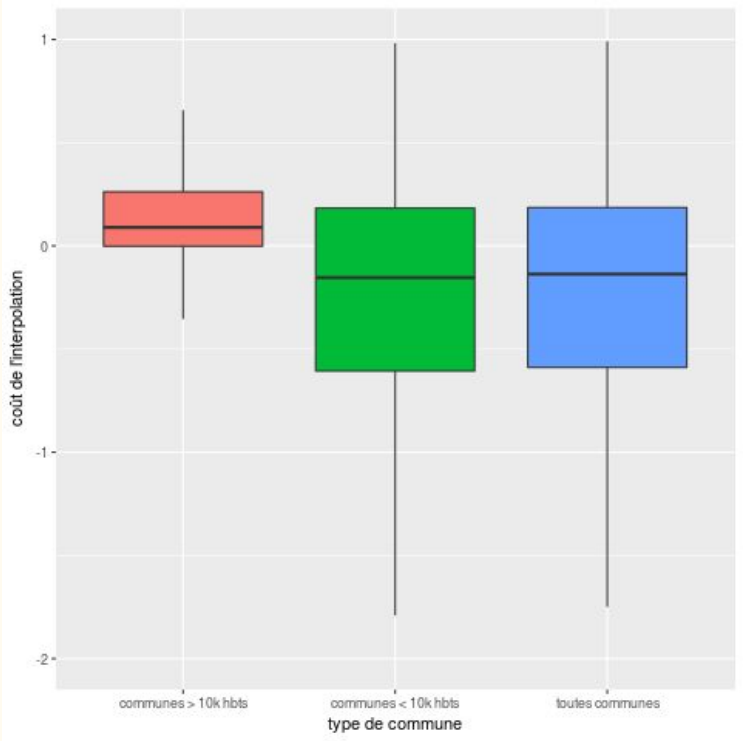
- Using only geocoded fiscal data
- estimation of the LAU (municipality) population with the “area proportion method” and the “real” population in each Voronoi cell
- comparing it to the official population of the LAU

⇒ we compute what is the ***cost* of this spatial interpolation** and **estimate the precision we lose** with such a method

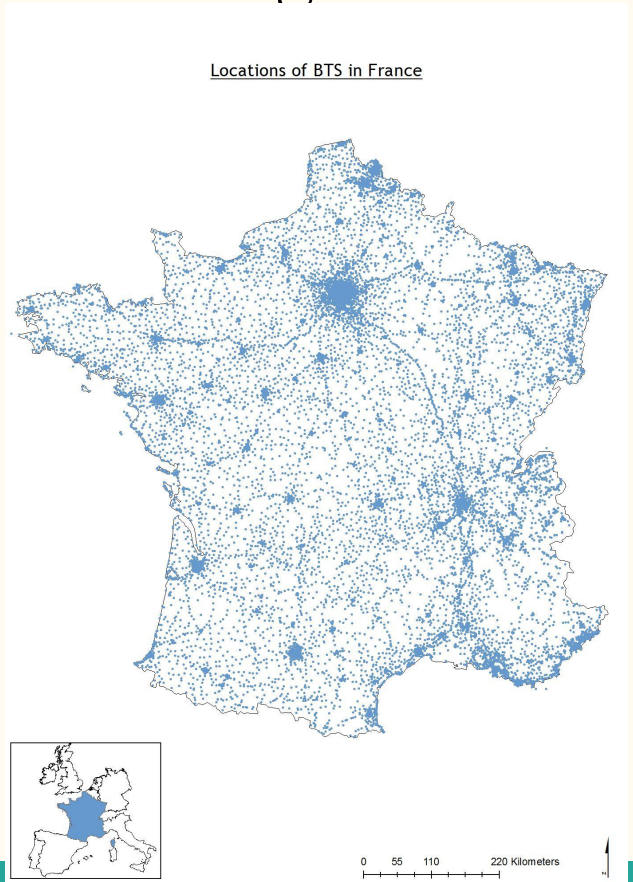
Estimating the cost of spatial interpolation

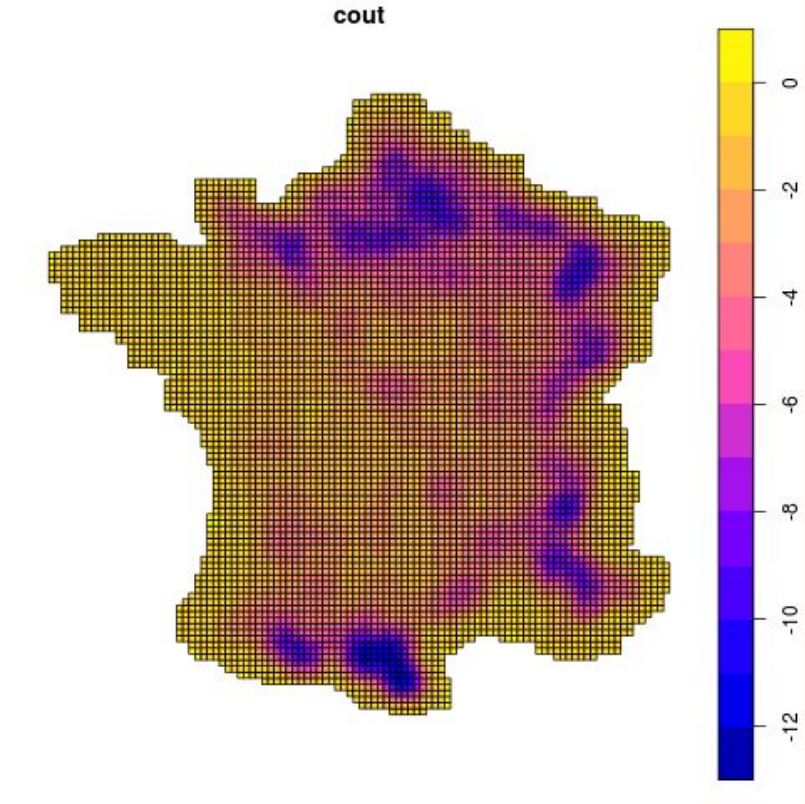
- this cost can be very important : **mean of over estimation by 10%**
- different for **small** and **large** municipality : **under** or **over** estimation
- very **dependent of the density of BTS**





Estimating the cost of spatial interpolation





Home Detection with CDR

a key step for computing indicators and many kind of studies

a first hint on validation

From home detection to population density

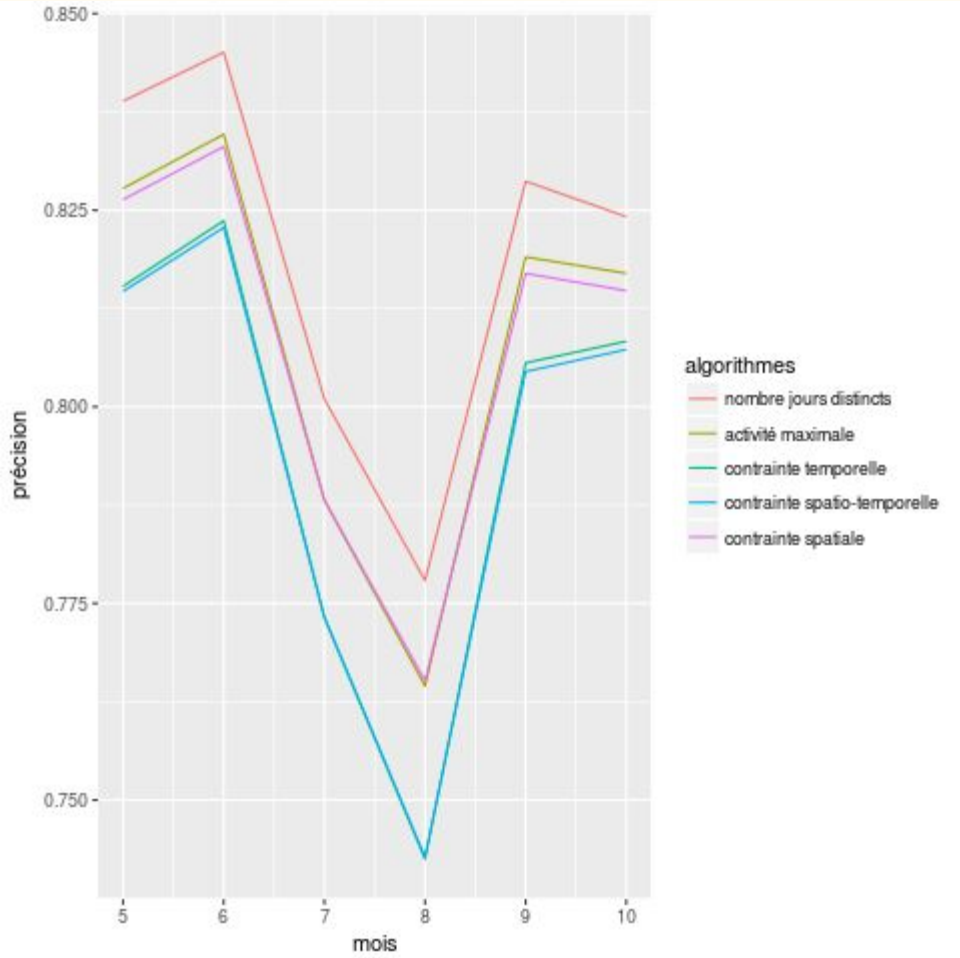
- home detection is a **key preliminary step**
- first idea : produce population density and **estimate its quality**

Several heuristics for defining “home”

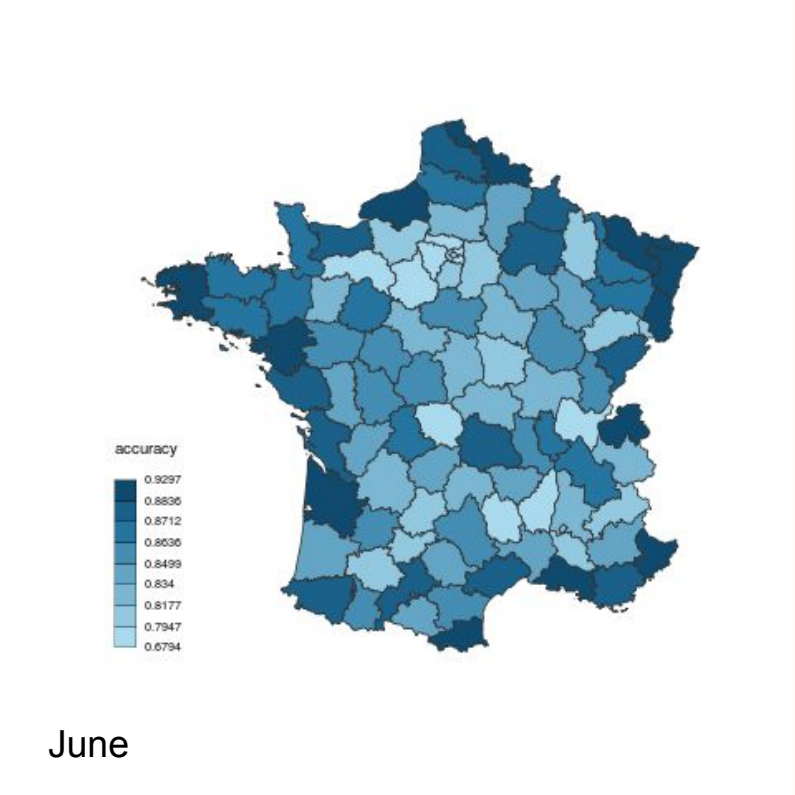
- **maximum activity**
- max activity with **time restraint** (at night)
- max activity with **space restraint** (within a certain radius)
- max activity with **time&space restraint**
- **distinct days** with max activity by night

Home detection precision

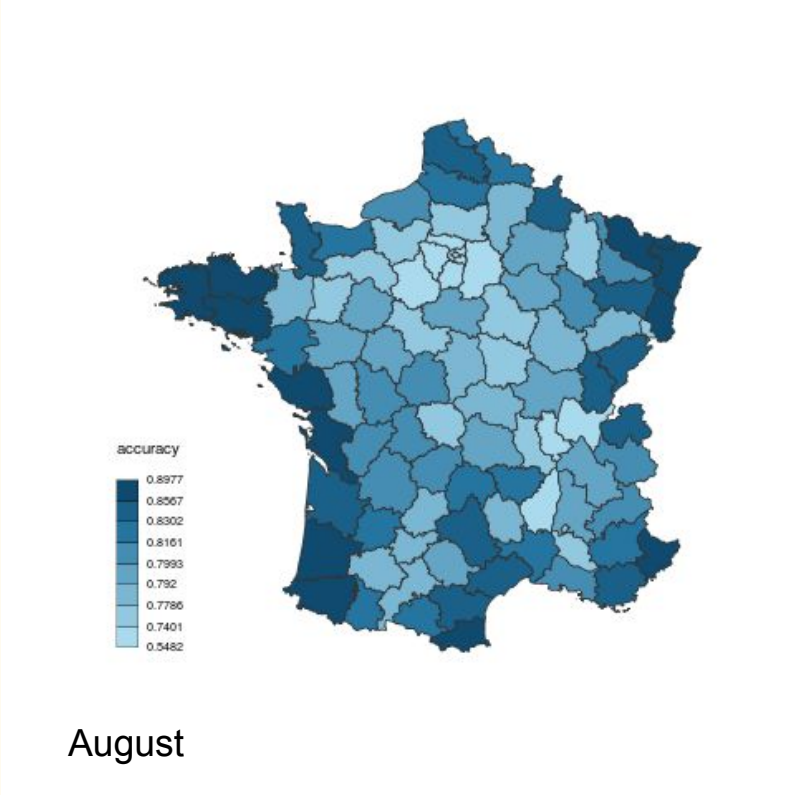
- thanks to CRM dataset possible estimation of the accuracy at a large scale (12 billions users)
- good accuracy for the **“distinct days” algorithm**
- clear time pattern due to summer hollidays
- limits :
  - at the NUTS 3 level (département)
  - CRM dataset records the client not the effective user
  - declared area of residence when contracting



NUTS 3 accuracy of Home Detection



June



August

Towards global population estimates

- from SIM cards to overall population : simple adjustments
- zooming in and out for validation

How to estimate population density ?

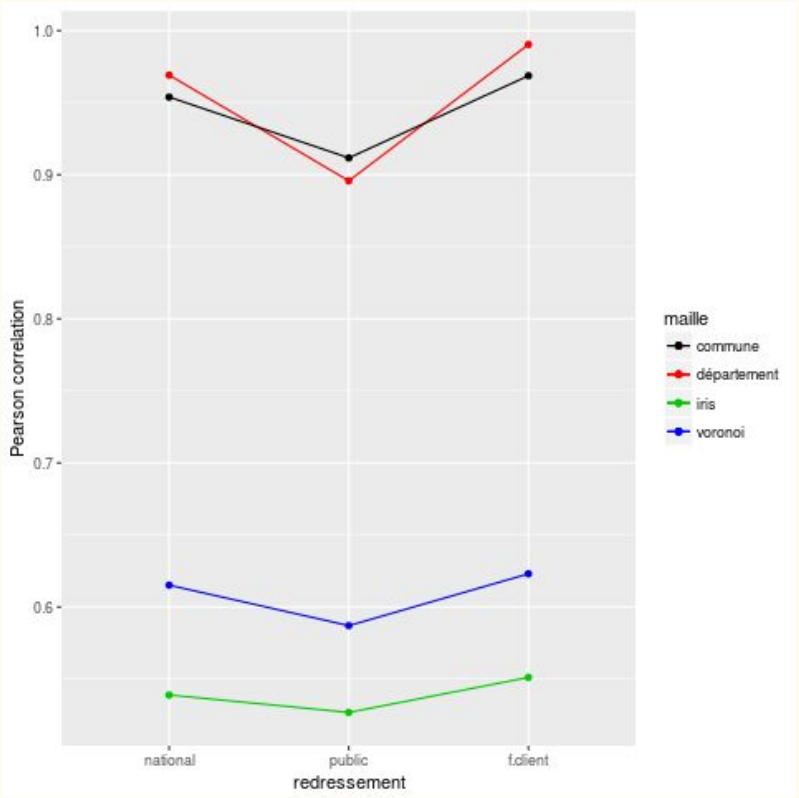
Need for extending the estimates to the global population

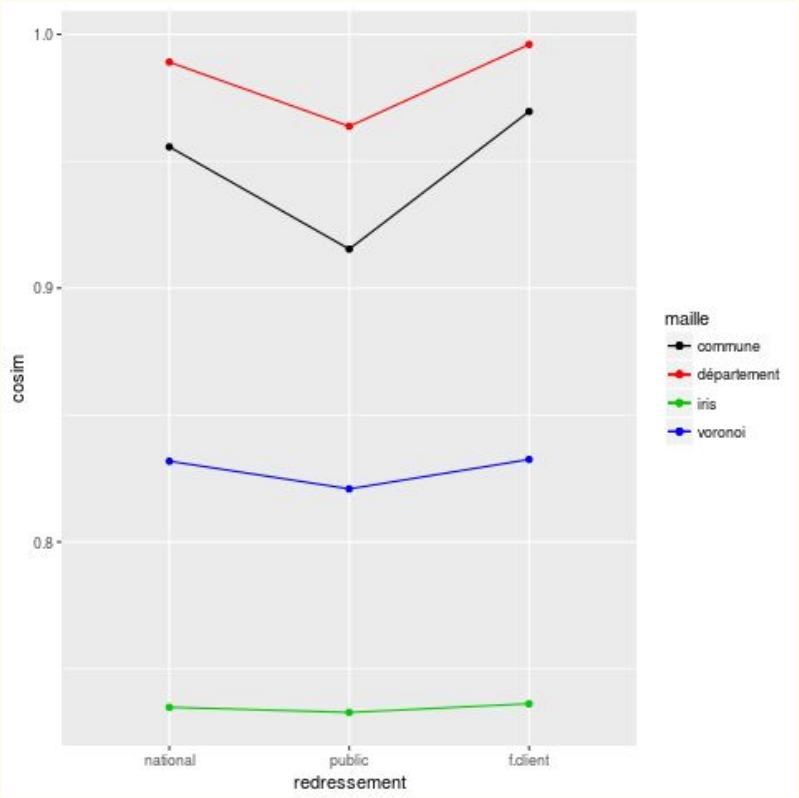
- **simplest solution** : multiplying by the ratio total population / number of users
- using **public information** : national market share and regional penetration rates
- using more of the MNO data (CRM dataset) to estimate **local market shares** and penetration rate (at the NUTS3 level)

What should the **geographical unit** be ?

- Voronoi cell
- NUTS 3 level (département)
- LAU2 (municipality)
- sub communal (IRIS)

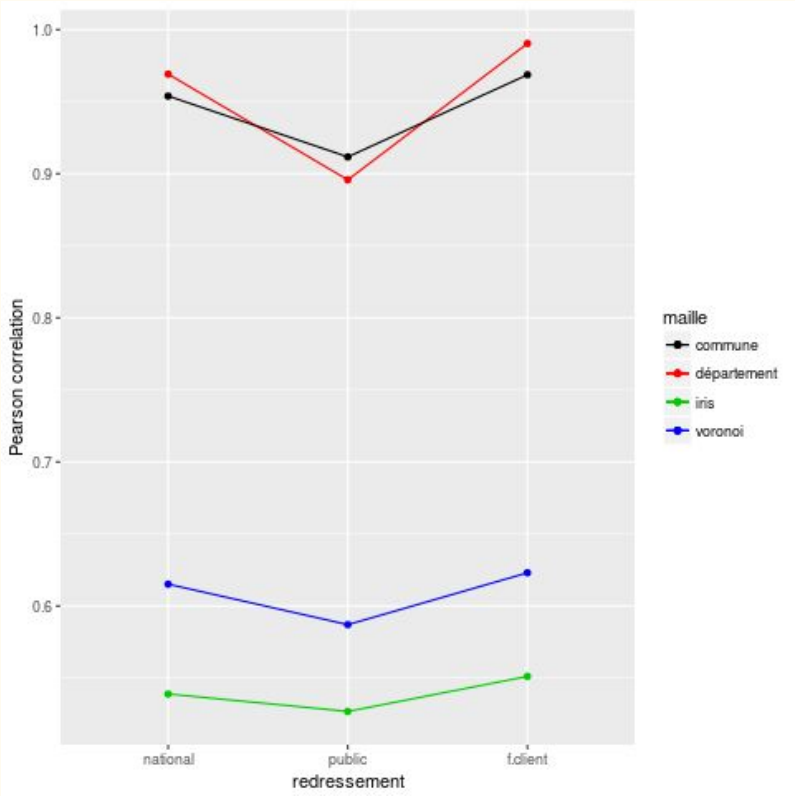
How to validate the estimates ?



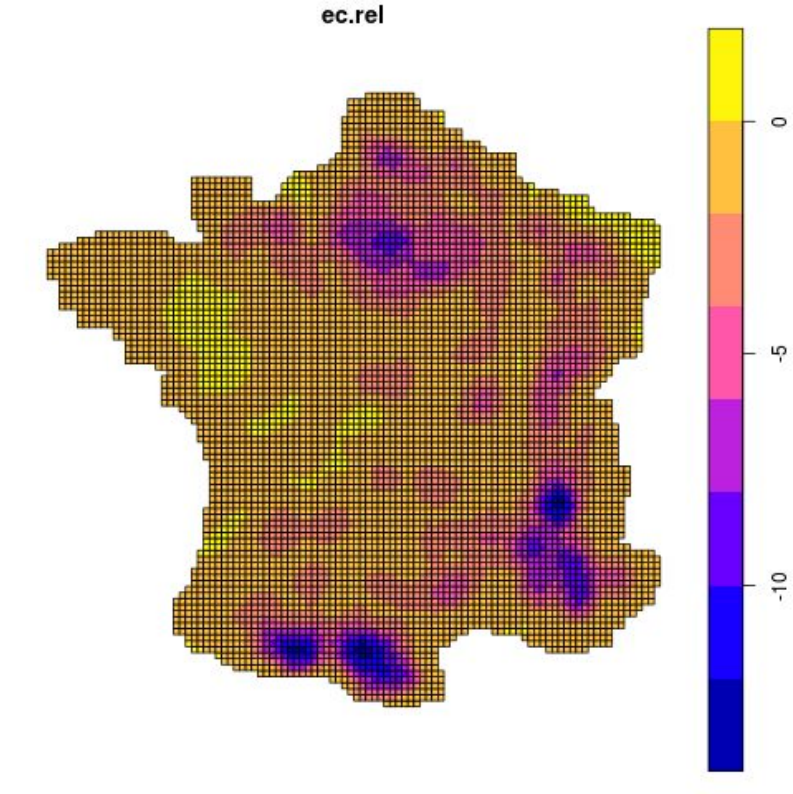


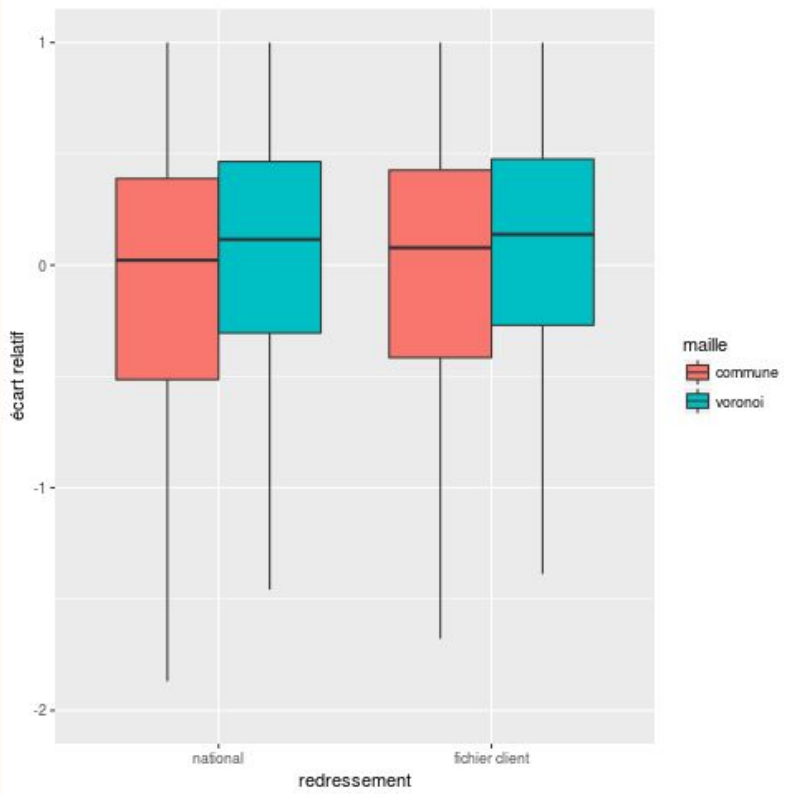
How to validate the estimates ?

- zooming out to municipality level gives fair results, denoising effect
- best performance adjusting with local market shares
- slightly different conclusion with Pearson correlation and cosine similarity



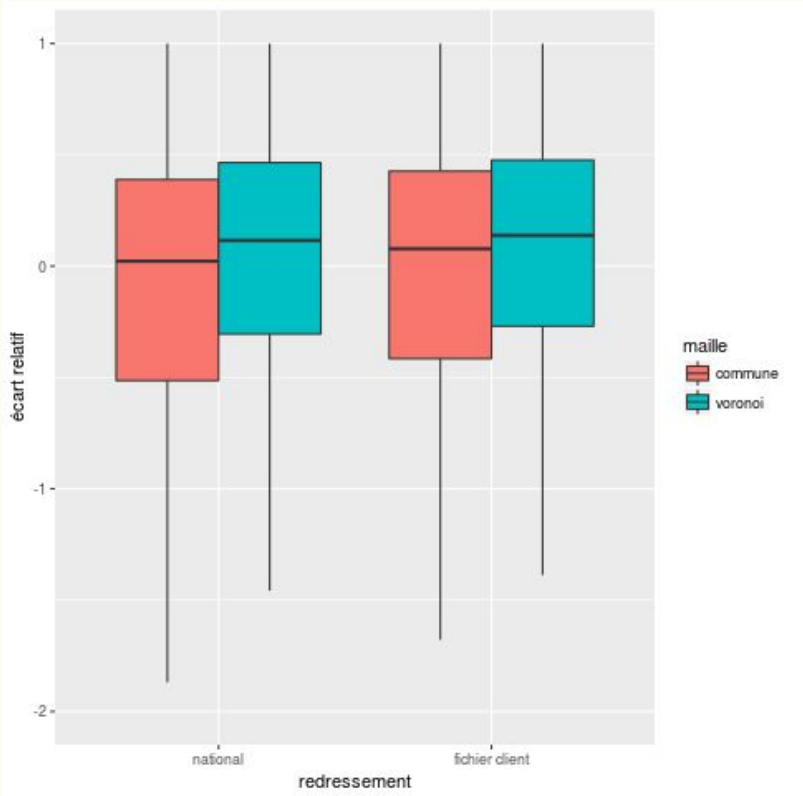
Relative difference in municipality pop estimates





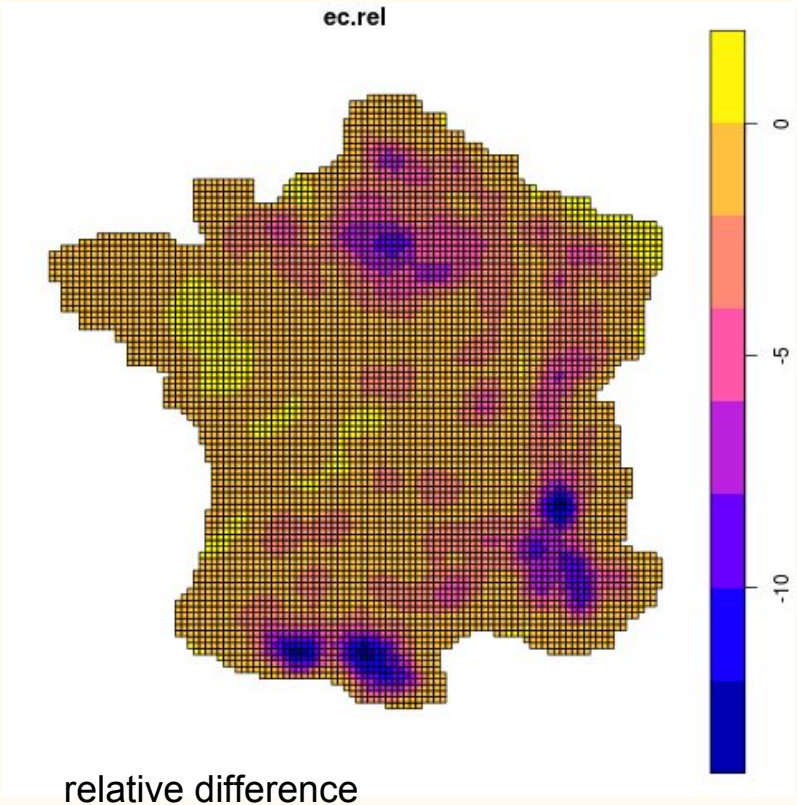
Relative difference in municipality pop estimates

- adjusting makes comparison possible at the municipality level
- double bias-variance trade-off with adjustment mode and geographical unit

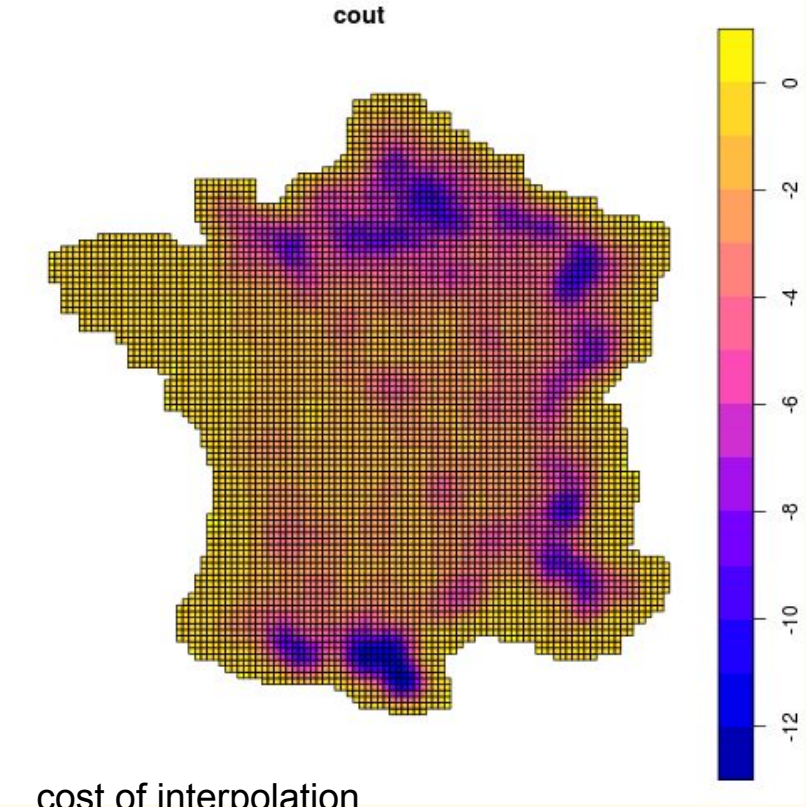


lessons learnt

Spatial interpolation is a crucial point

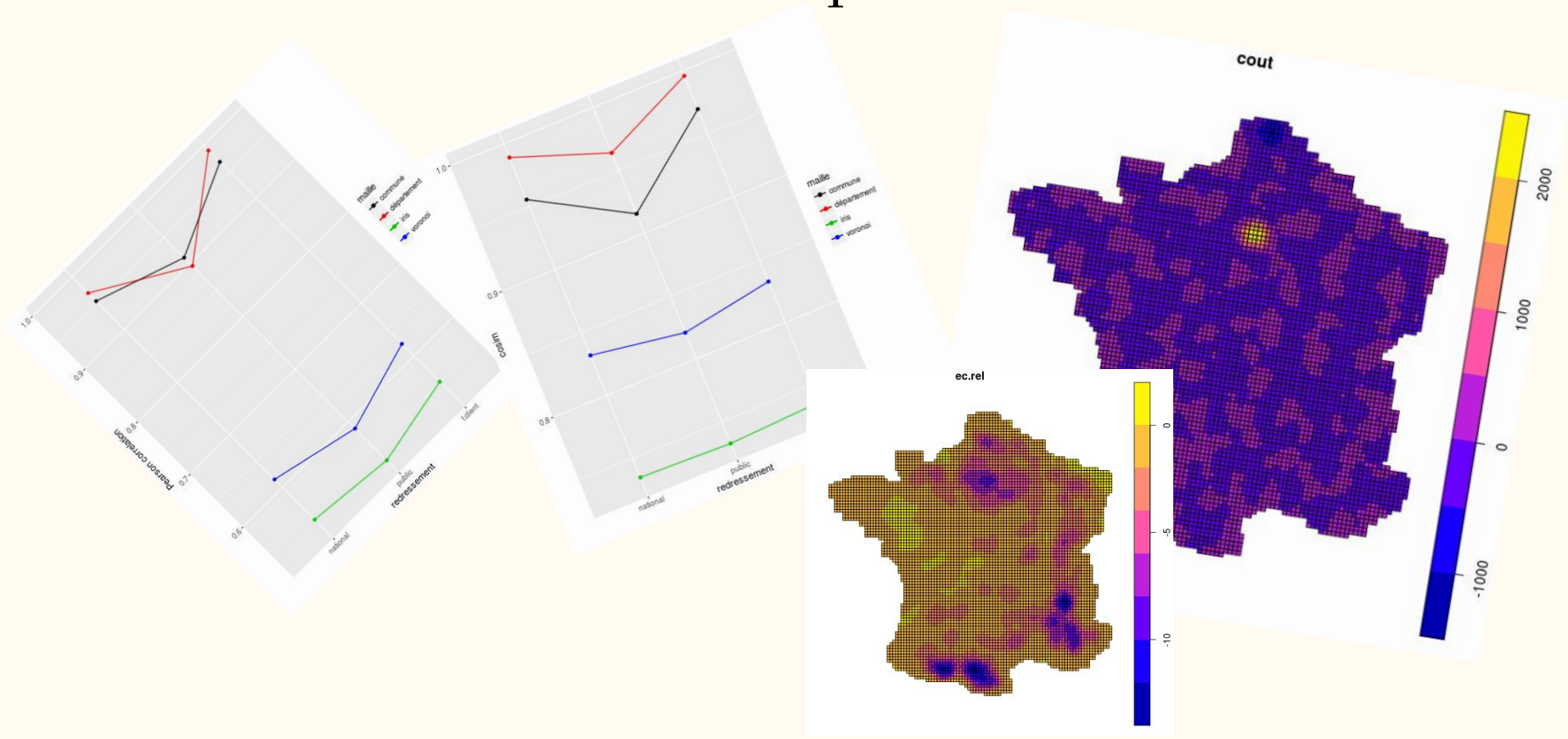


relative difference



cost of interpolation

Need for a clear validation procedure



thank you