# Spatial sampling for the French Master Sample

Thomas Merly-Alpa (INSEE)

ITACOSM 2019

SPE. 8 - Spatial sampling and applications - 06/06/2019

Common work with the Sampling Methodological Unit at Insee :

- Martin Chevalier
- Laurent Costa
- Thomas Deroyon
- Clément Guillo
- Nicolas Paliod – `nicolas.paliod@insee.fr`
- Pierre-Arnaud Pendoli – `pierre-arnaud.pendoli@insee.fr`
- Ludovic Vincent – `ludovic.vincent@insee.fr`

1. A new Master Sample in France
   - Context
   - Primary units
2. Spatial sampling
   - From Balanced sampling...
   - ...to Doubly Balanced Sampling
3. Spatial coordination with the LFS sample

# The Master Sample

- Each sampling design is a two-phase sampling :
  - 1st degree : a set of primary units is drawn ;
  - 2nd degree : a collection of households is surveyed ;
  - Master Sample : the first degree stays the same.

- The French National Statistical Institute (INSEE) manages a Master Sample :
  - It's a collection of geographical areas ;
  - These areas are covered by surveyers for social surveys ;
  - Each Master Sample lasts 10 years ; the next one starts in 2020.

# Why change ?

- Our current Master Sample is based on Population Census (very complex process in France) ;
- Transition to administrative data :
  - More and more tax data (on income, housing. . . ) are available ;
  - These data are processed by INSEE to achieve a good statistical quality.
- Exhaustion of the Master Sample :
  - We aim to not survey the same people twice in a row ;
  - The statistical properties of the last Master Sample are obsolete.

# Primary units

To draw our Master Sample, we need to build primary units.

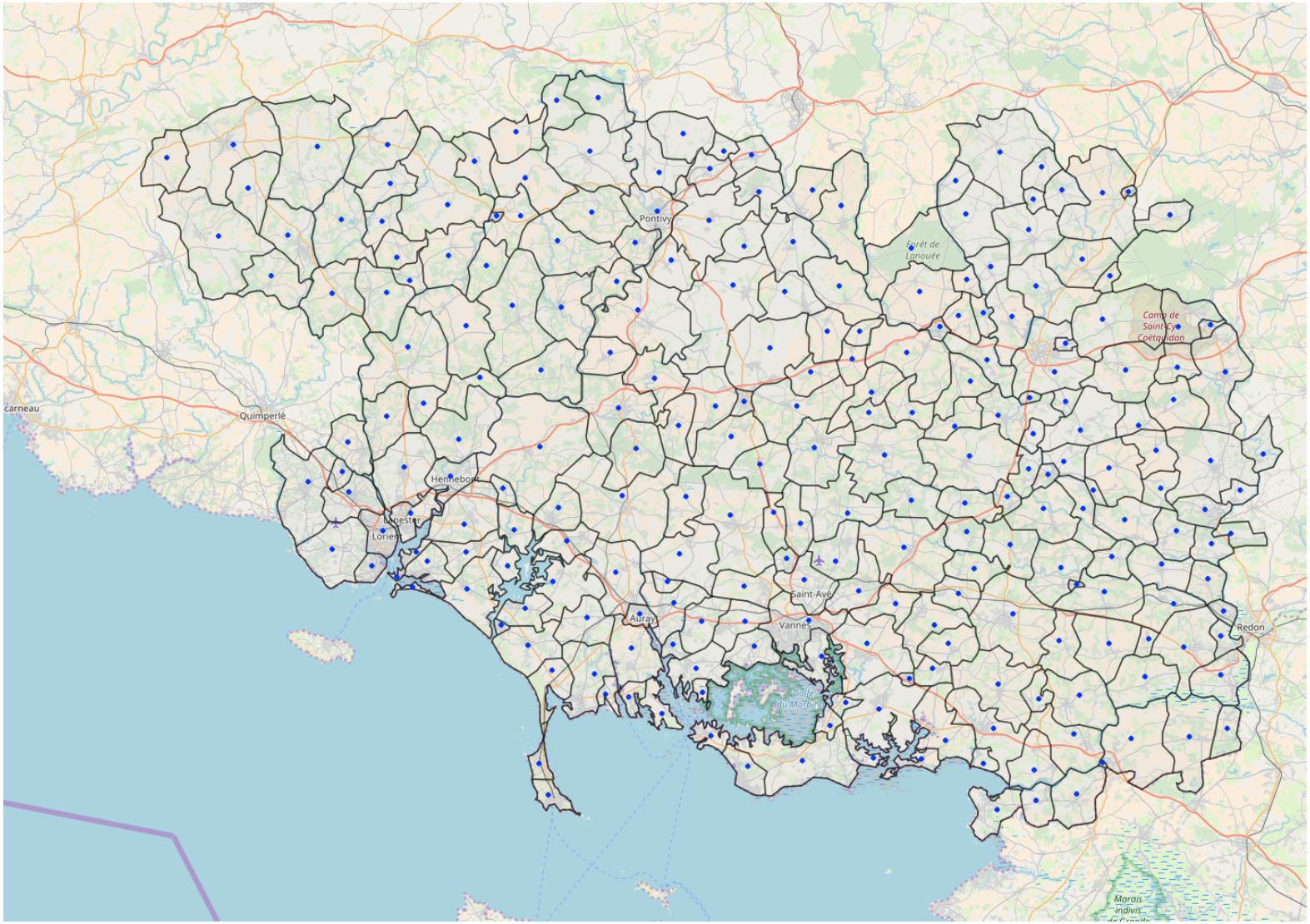
- Primary units are geographical areas :
  - based on administrative units : towns, municipalities...
  - other approach : purely geographical areas ; more flexible but hard to use in our processes

- Constraints :
  - a large enough number of households : min 2500 to reduce the burden of survey ;
  - as small as possible, in order to reduce the travel time of surveyers.

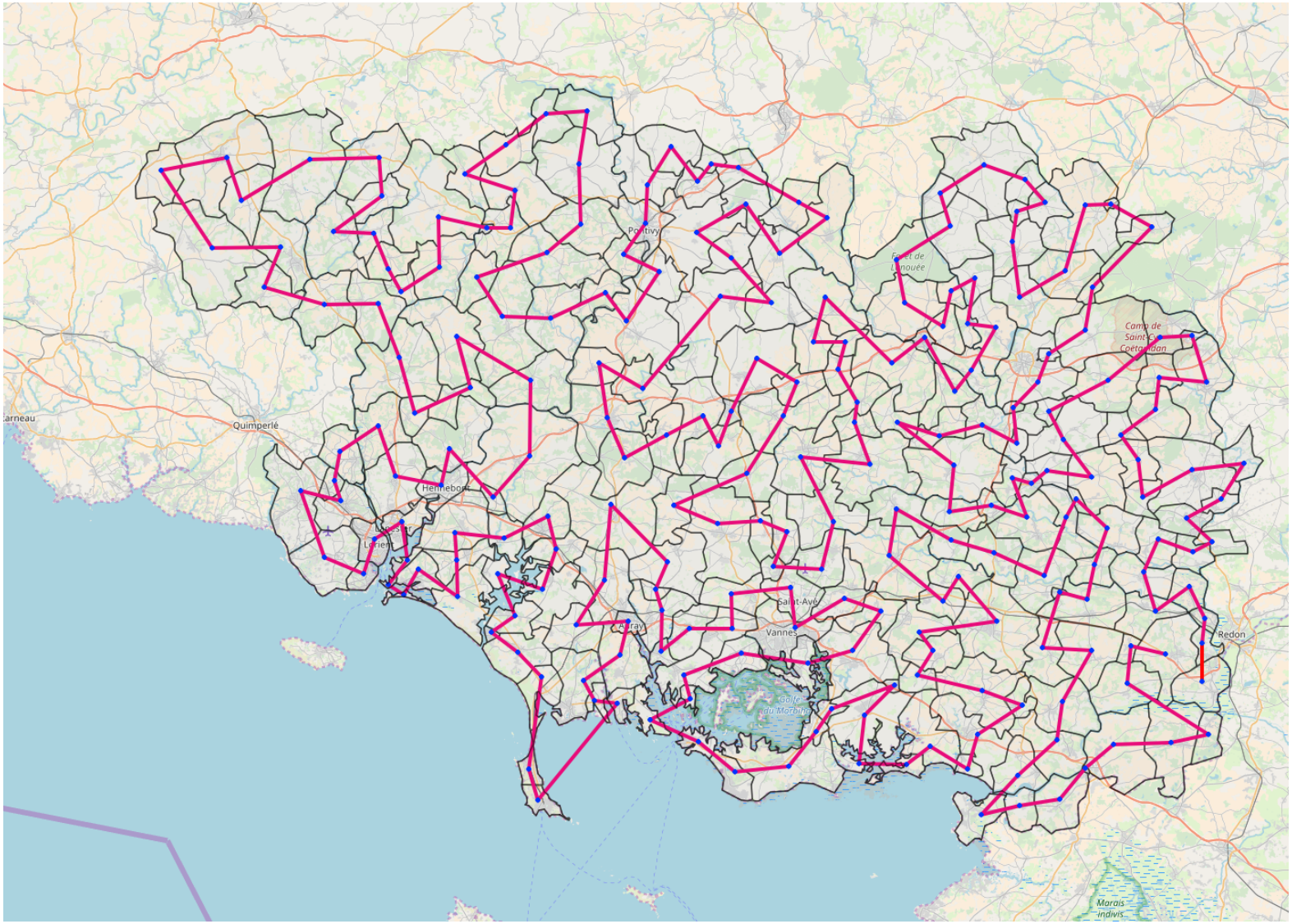
# Primary units

- How are the primary units built ?
- In each French NUTS3, we regroup cities to form primary units, with respect to the constraints
- Method used : the traveling salesman problem and its approximal solutions offer us a path around the area we can follow to build the primary units

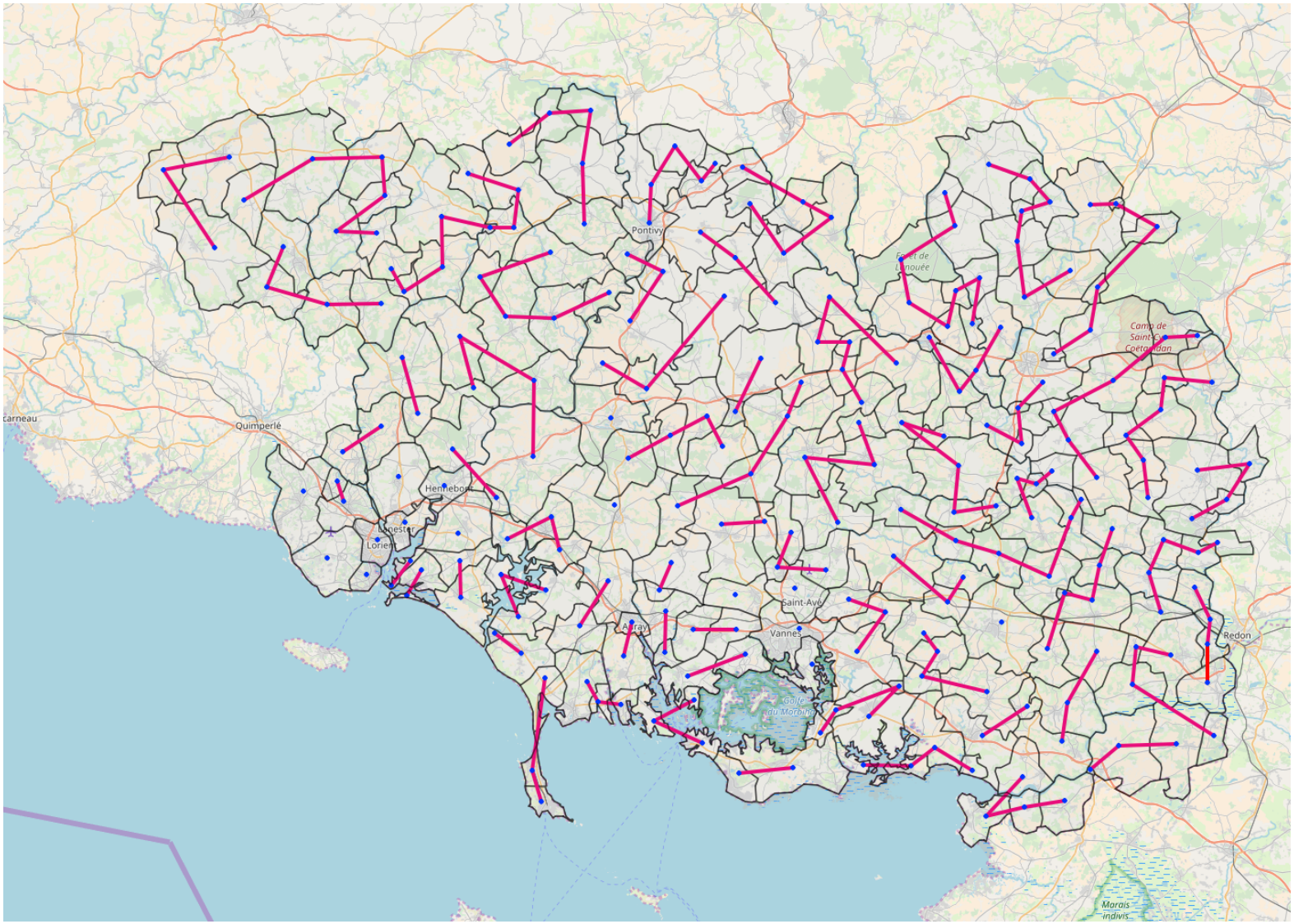
# Example in Brittany



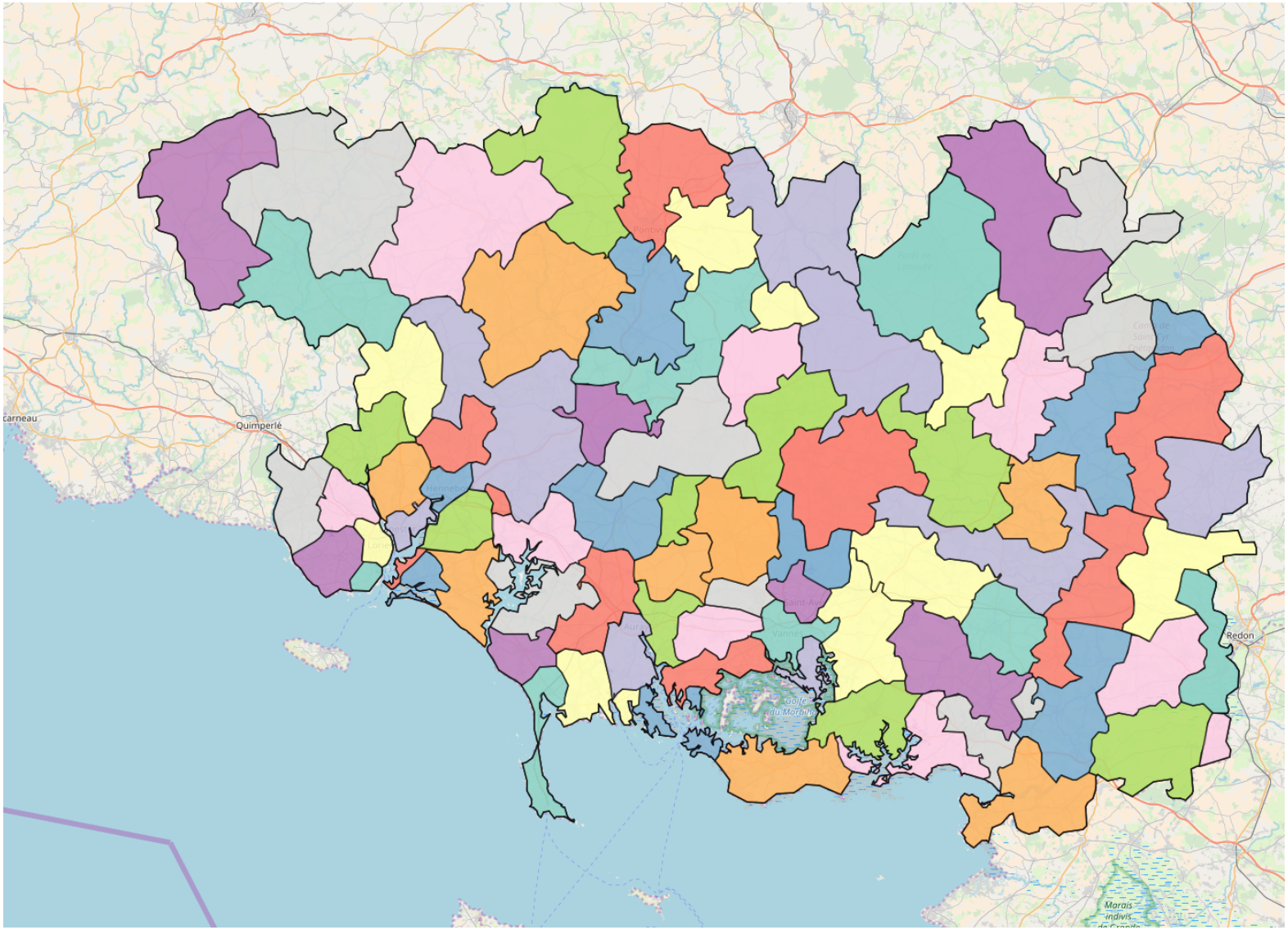
# Example in Brittany



# Example in Brittany

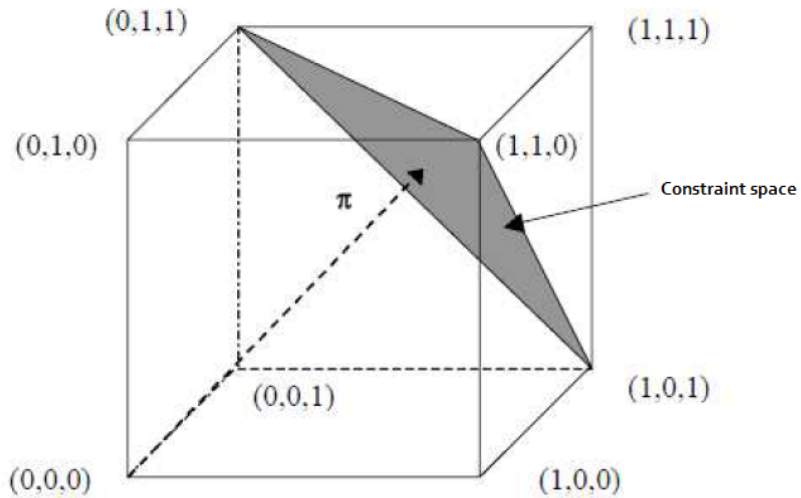


# Example in Brittany



# Balanced Sampling

- We want our Master Sample to be balanced : used for numerous surveys.
- Method used by INSEE : Cube (Deville & Tillé, 2004)




- Inequal probabilites based on the size of the primary unit.

# Balancing variables

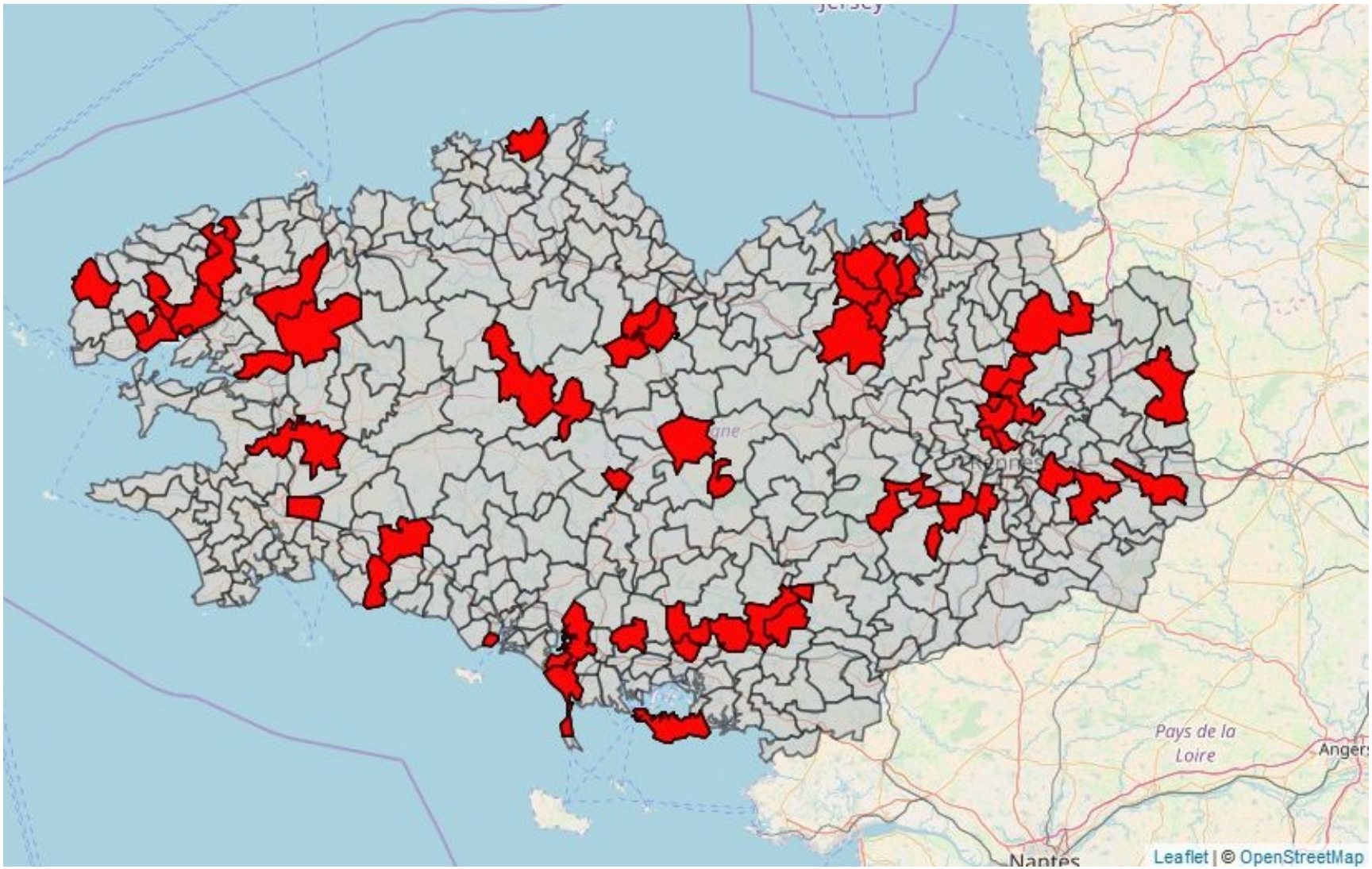
- Large number of socio-economical variables available
- Using too much variables affects the quality of the final sample

- Solution :
  - Choosing wisely the variables ;
  - Using data analysis methods (PCA. . . ) to reduce the dimensionality of the set of variables ;
  - Using a spatial method.

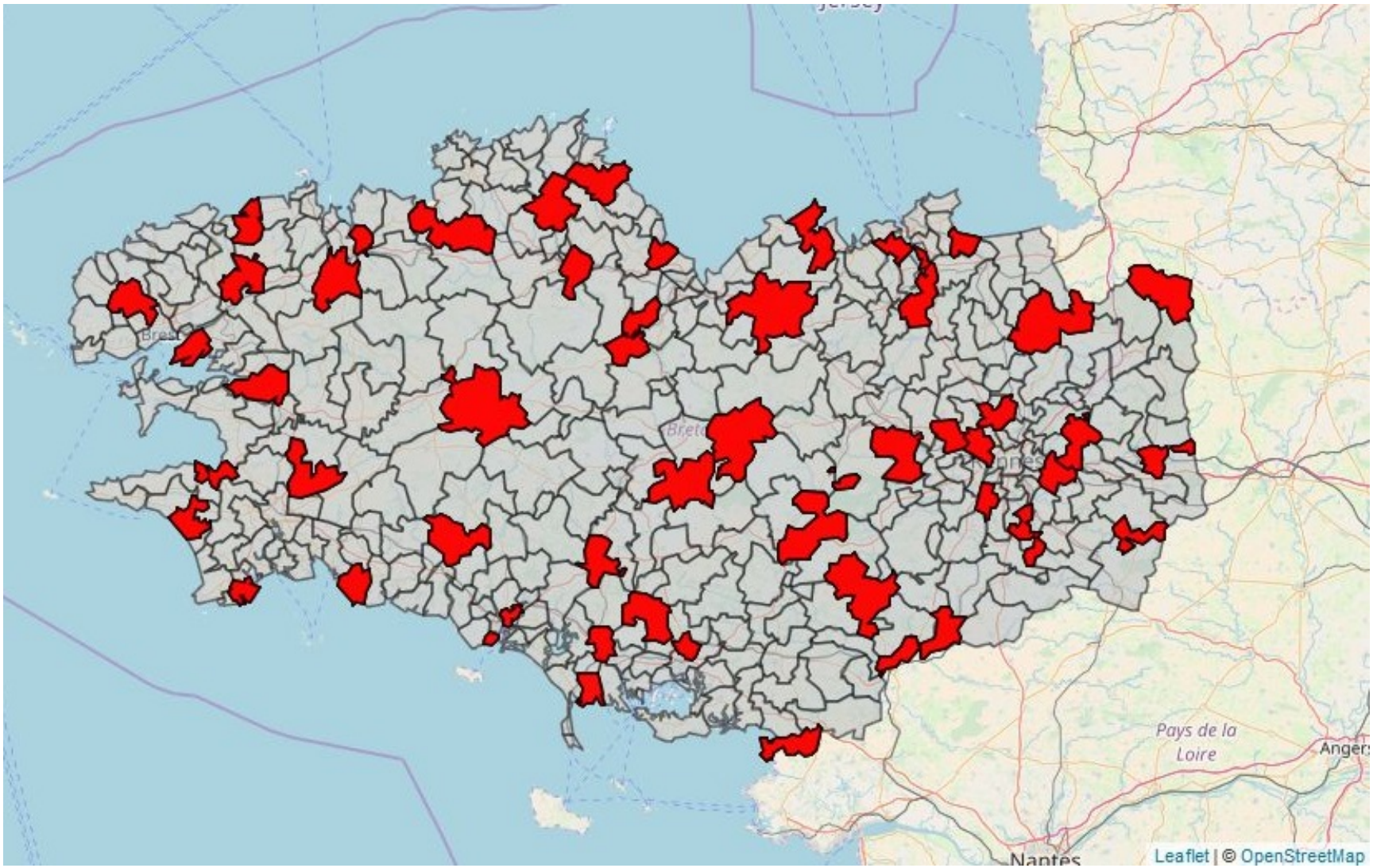
# Why spatial sampling ?

- Spatial sampling : spreading the sample geographically
- Taking into account spatial correlation : areas next to each other are similar
- A better coverage of the French territory
- Data collection easier to manage : no accumulation points

# Example of a balanced sample



# Example of a spatially balanced sample



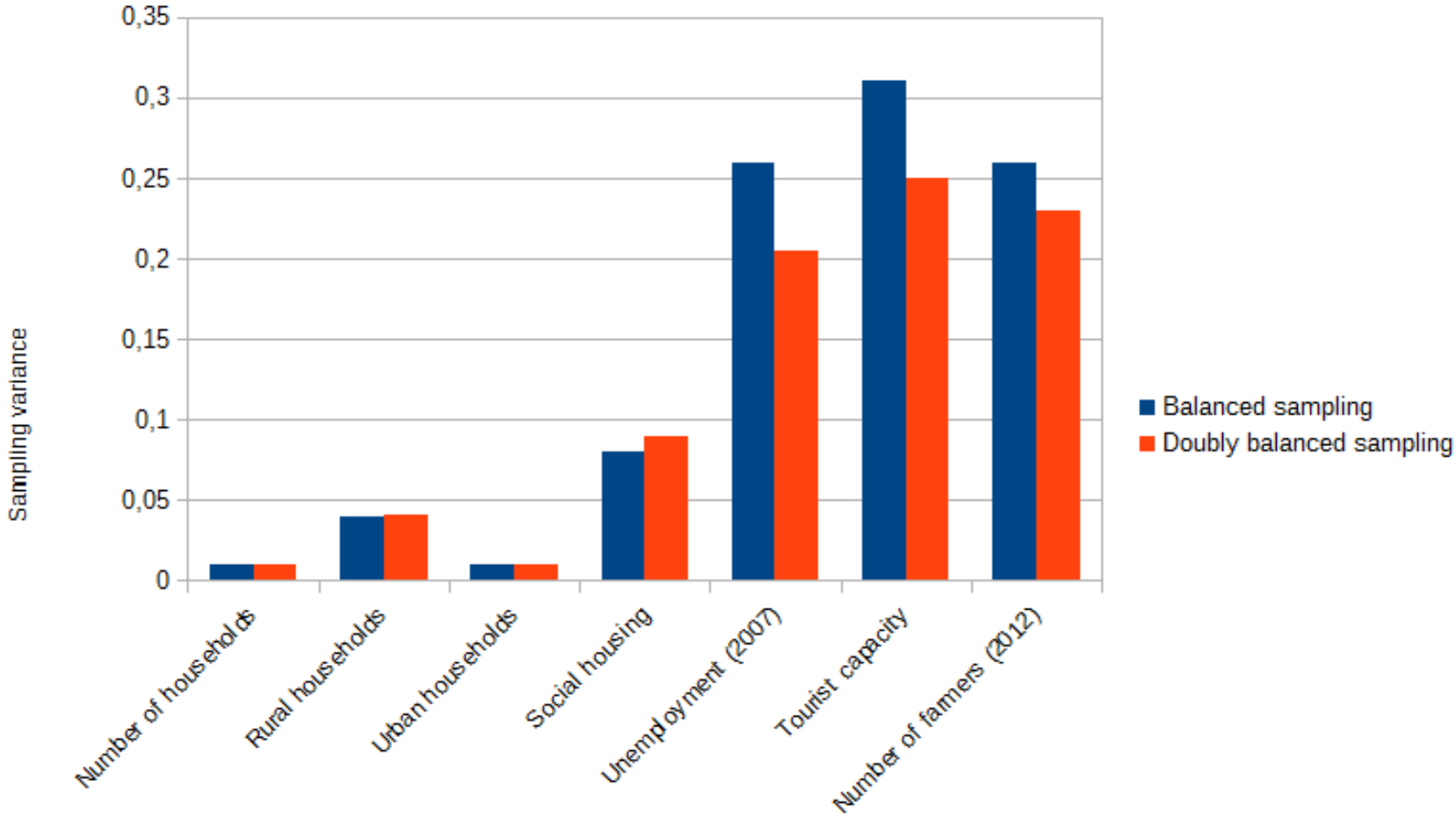
# Which method ?

- Lots of methods : pivotal, determinantal...
- We need to integrate balancing for key socio-economical variables
- Using doubly balanced sampling (Grafström & Tillé, 2013) allows us to combine :
  - balancing variables ;
  - geographical spreading.

# Doubly Balanced Sampling

- Two-step algorithm (as Cube) ;
- Let $p$ be the number of balancing variables ;
  - Local clusters of $p + 1$ units are built ;
  - On each cluster, a flight phase is done which leads to update the inclusion probabilities ;
  - The clusters are built once again ;
  - Landing phase : as soon as there is only $p$ units left to draw
- R package *BalancedSampling*

# Precision gains with a spatial design



# The Labour Force Survey

- The Labour Force Survey (LFS) measures unemployment and activity rates.
- Its sampling design is clustered : groups of 20 or so households are surveyed just after the reference date.
- The clusters are built in order to minimize the geographical dispersion.

# Sampling of LFS

- Balancing variables (from tax data) : Income, specific unemployment allowance.
- Spatial sampling : useful because unemployment is spatially (auto)correlated.
- Stratified sampling : dissemination of results at NUTS2 level.

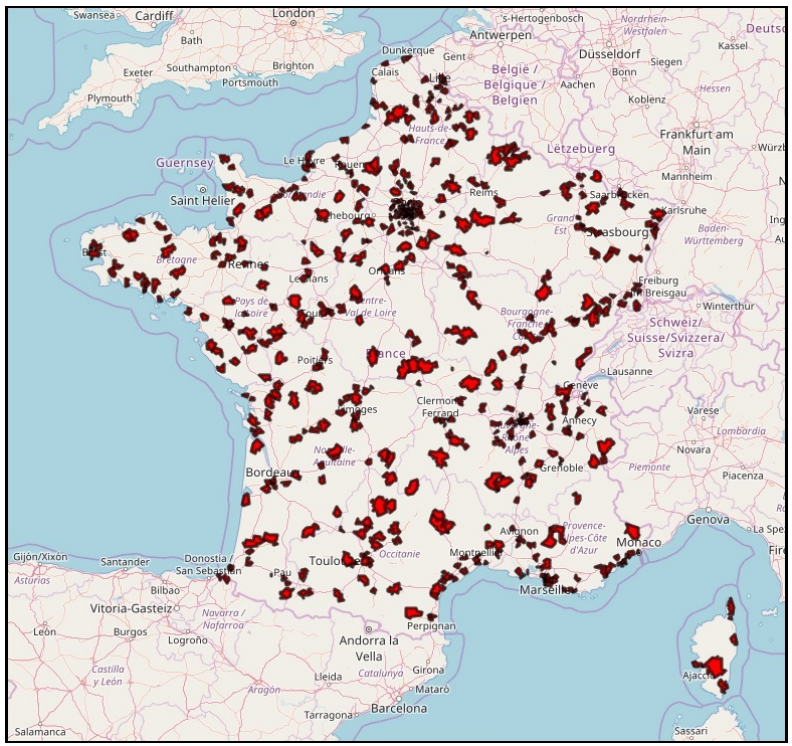
# Relation to the Master Sample

- LFS used to be drawn independantly from the Master Sample because of :
  - a large sample ;
  - a specific method.
- Issues : some household drawn for LFS are far from our surveyers
- Solution : coordination ?

# LFS sample in the Master Sample

Is it possible to draw the LFS sample within the Master Sample ?

- Exhaustion of the areas, which leads to burden of response.
- Clustering effect $\rightarrow$ important loss of precision



# Constitution of Coordination Units

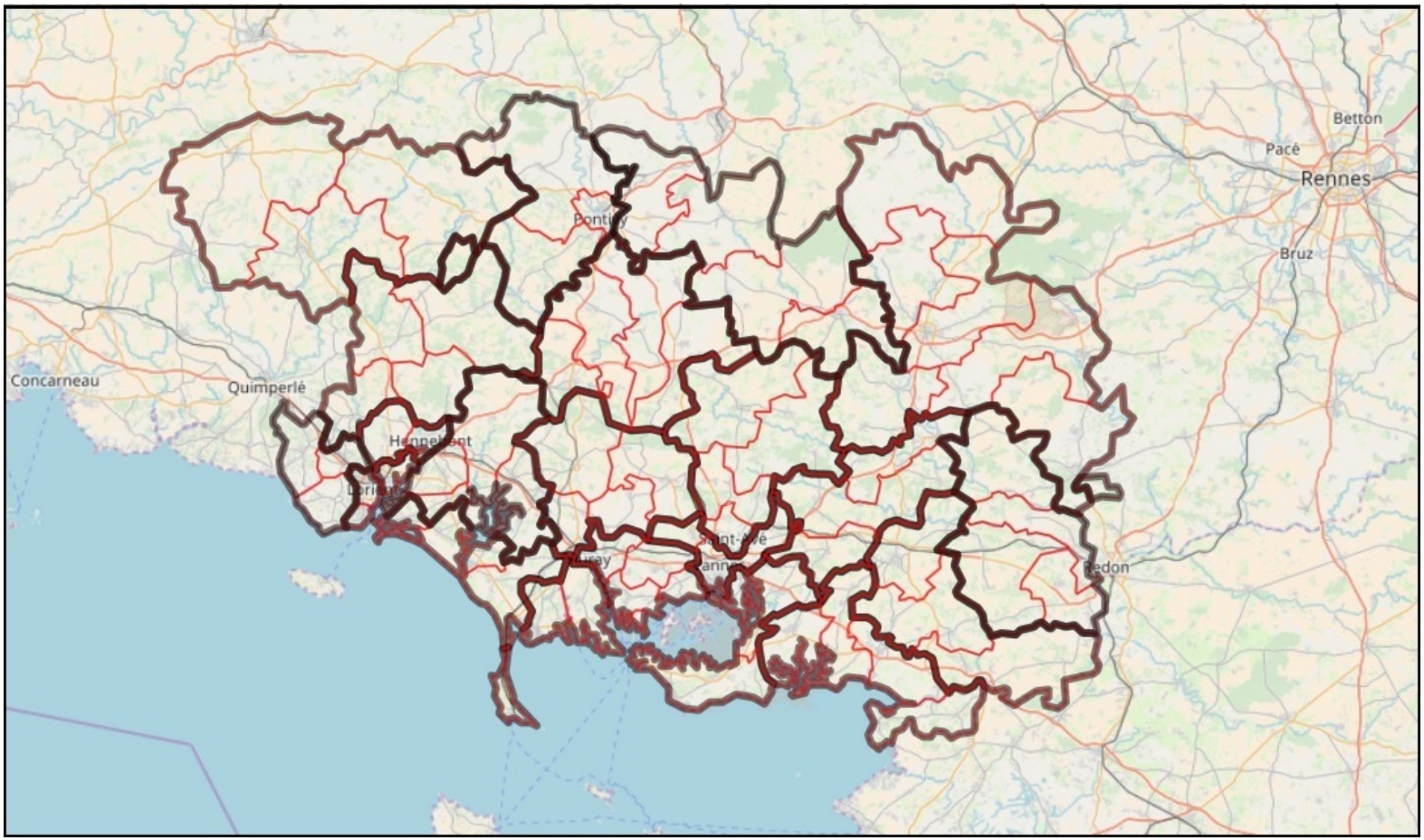
- We don't need to draw LFS within Master Sample ;
- We just want the two samples to be nearby.

- Solution : create new units, named Coordination Units, which are larger than Primary Units.

- How : aggregate Primary Units to form larger areas, which remain compact.

- The size of these Coordination Units has an impact on :
  - precision of the LFS ;
  - degree of coordination.

# Example in Brittany



# Sampling method

- Classical approach : make a 3-phase sample
  1. Draw a sample of Coordination Units ;
  2. Draw a sample of Primary Units and LFS clusters within these Coordination Units ;
  3. Draw households within the Master Sample.

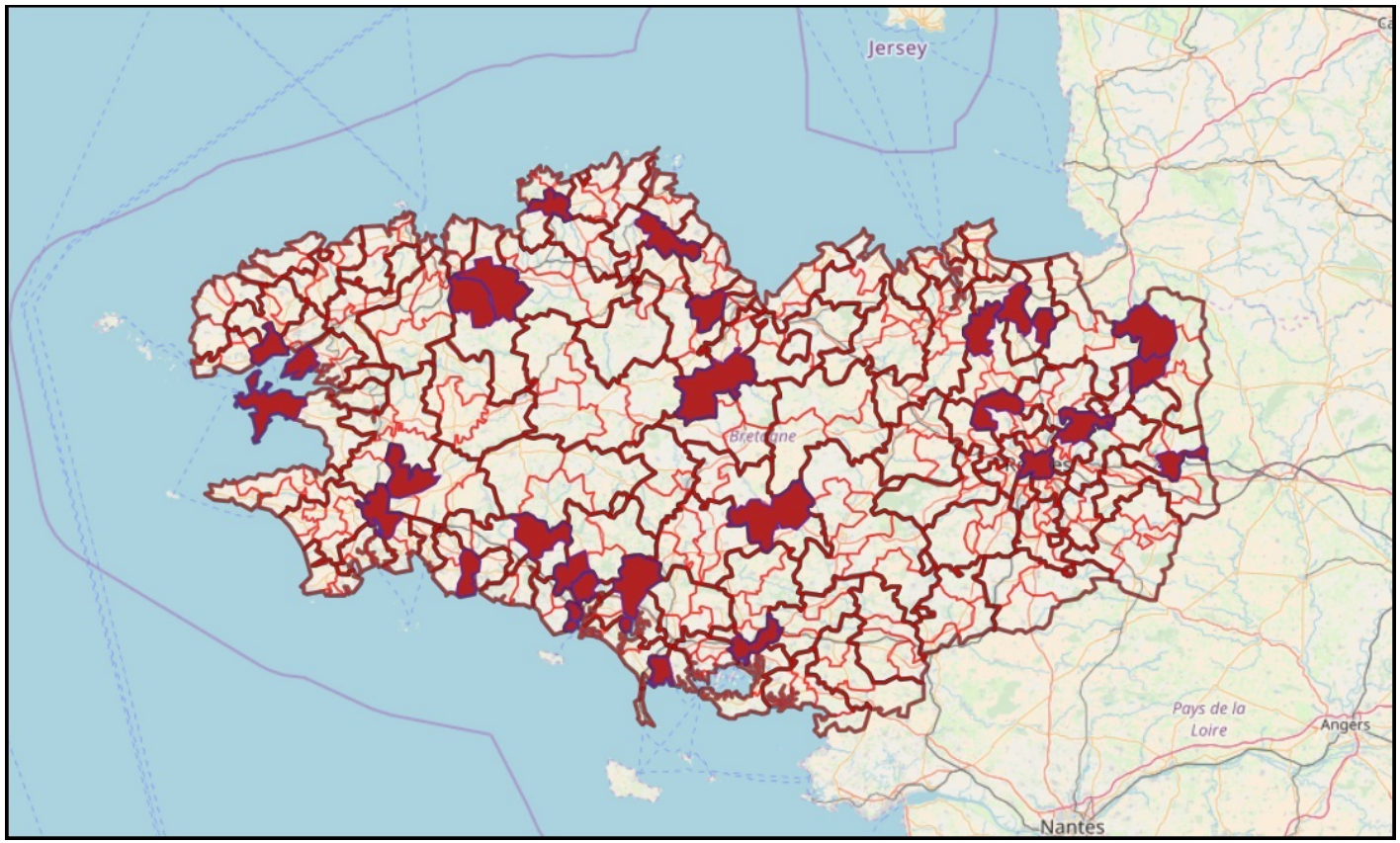
- Constraint of drawing exactly 1 Primary Unit within each Coordination Unit $\rightarrow$ loss of precision

# Indirect sampling

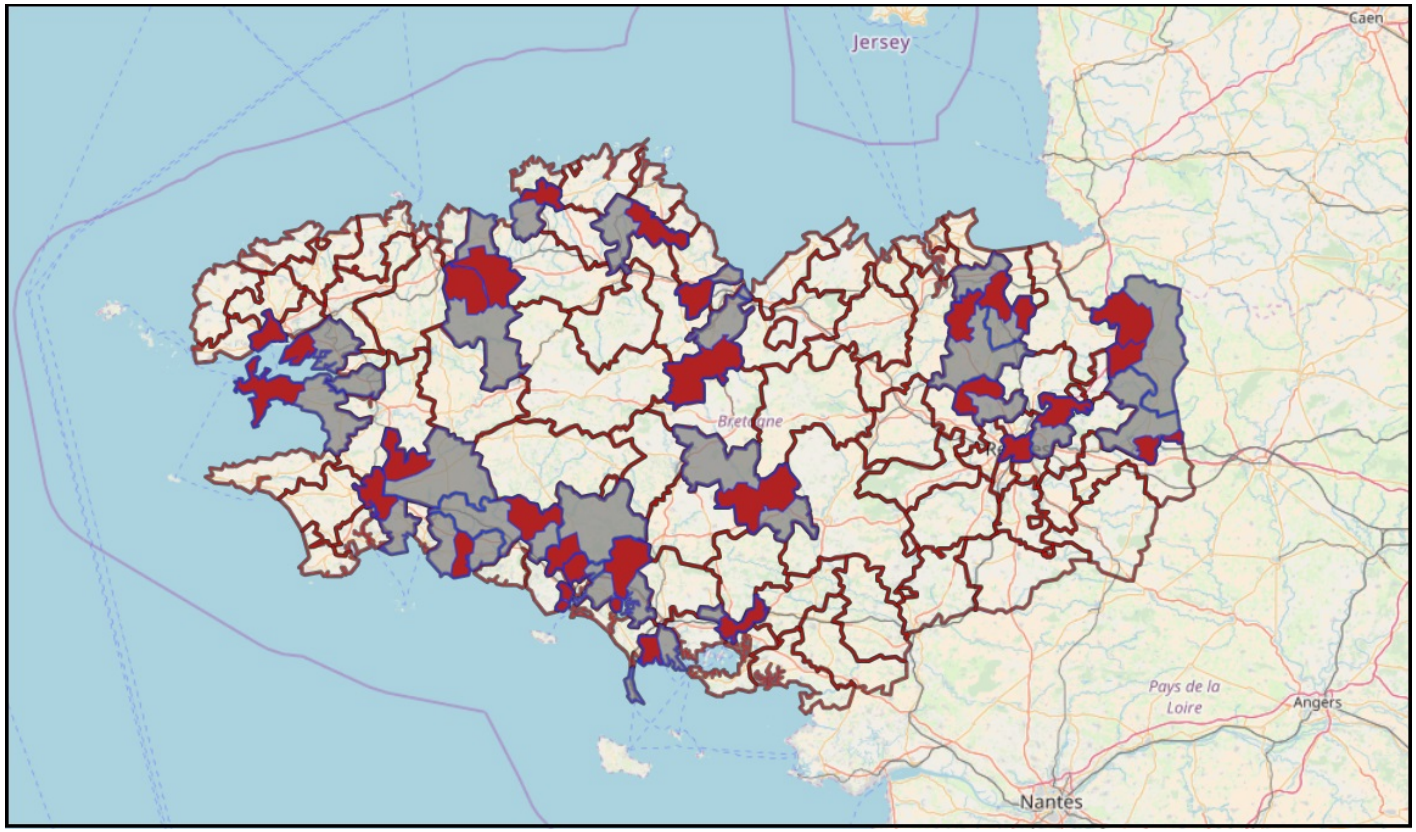
- New idea : start the sampling method by drawing the Primary Units, which indirectly select a set of Coordination Units

- Another 3-step sampling :
  - 1st step : Primary Unit drawing
  - 2nd step : Deducing associated Coordination Units
  - 3rd step : Drawing LFS clusters in these Coordination Units

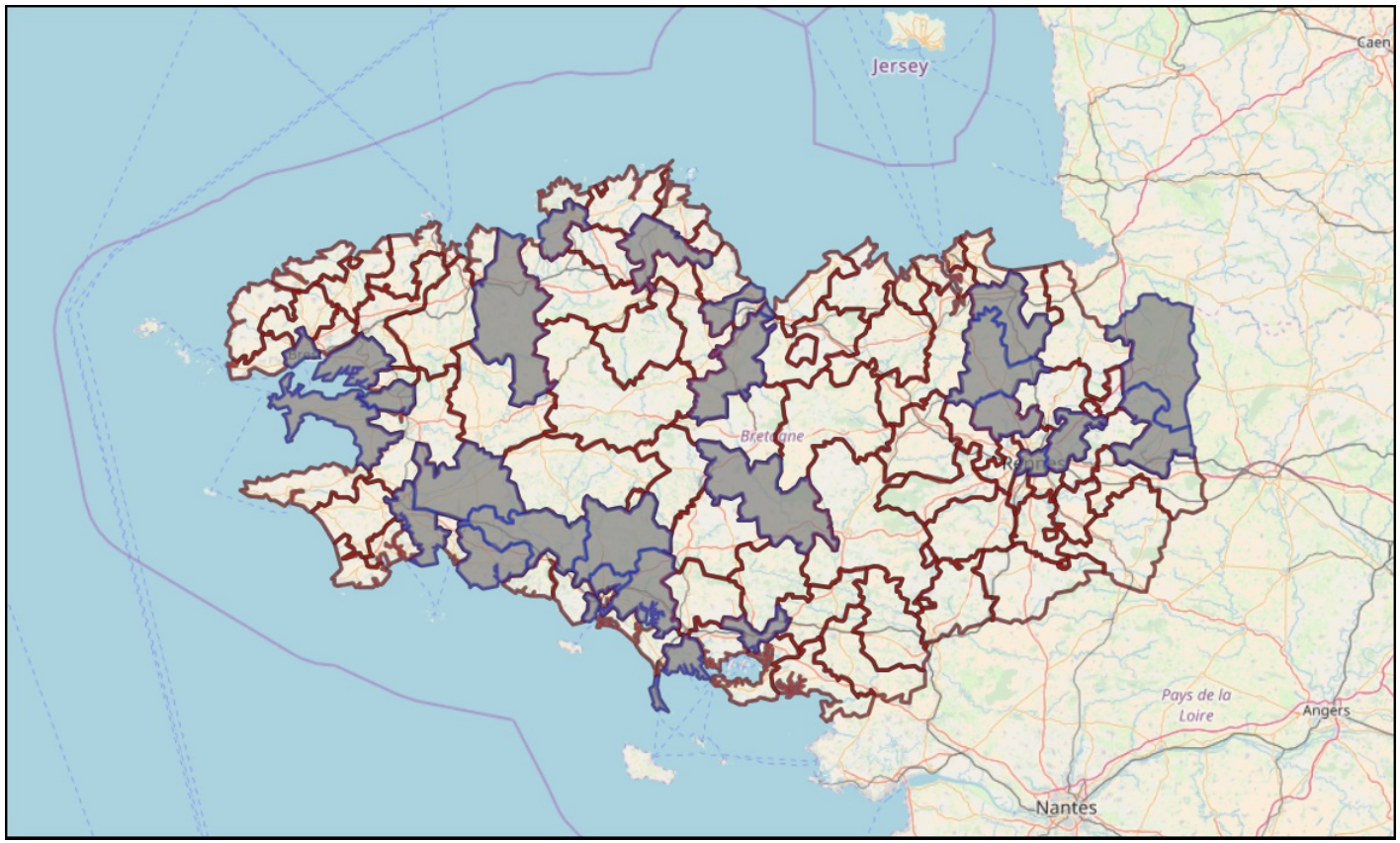
# 1st step : Primary Unit drawing



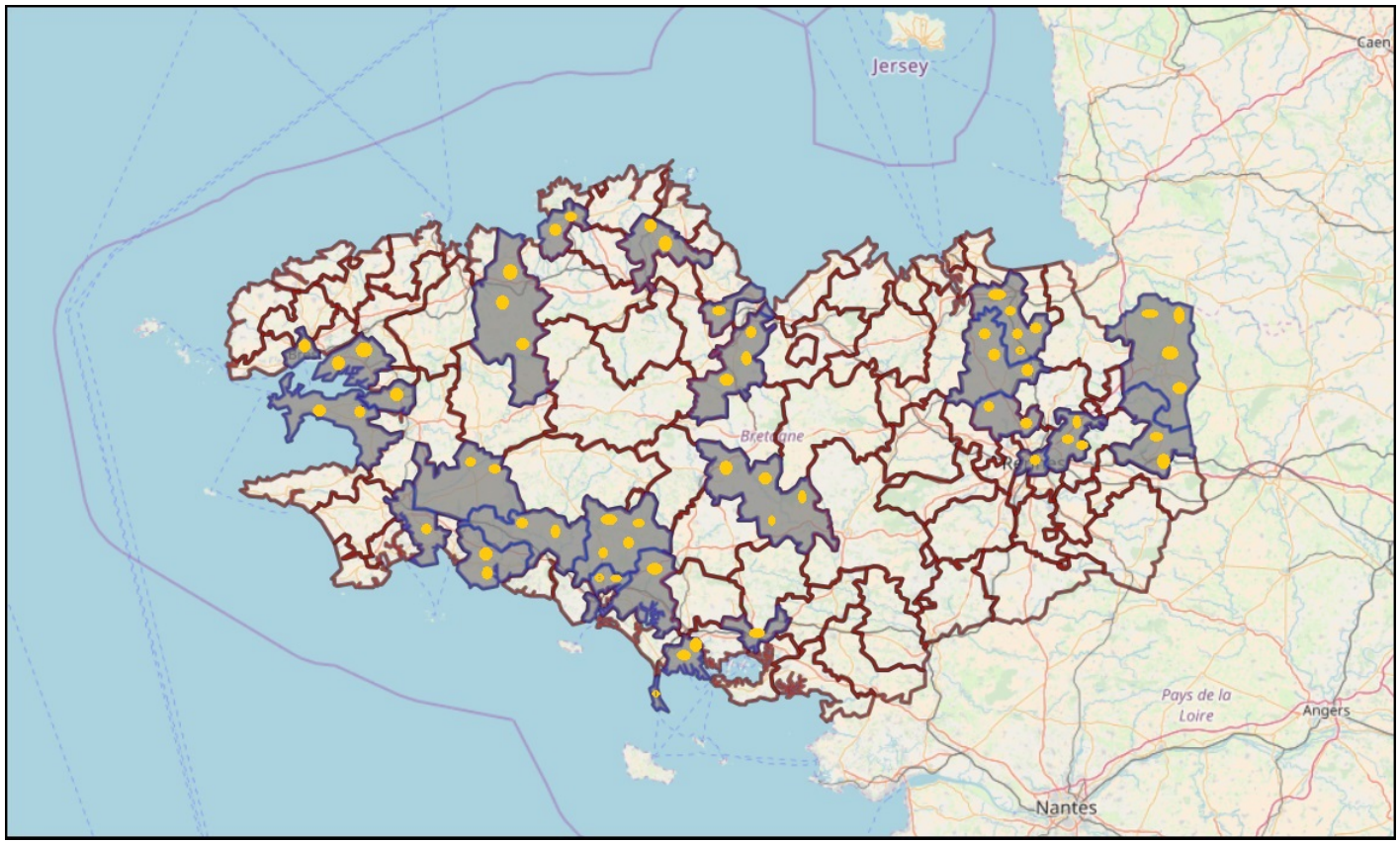
# 2nd step : Associated Coordination Units (1/2)



# 2nd step : Associated Coordination Units (2/2)



# 3rd step : Drawing LFS clusters



# Indirect sampling

- Indirect sampling :
  - 1st step : Primary Unit drawing
  - 2nd step : Deducing associated Coordination Units
  - 3rd step : Drawing LFS clusters in these Coordination Units
- This solution combine our criterion of :
  - precision (for the Master Sample) ;
  - coordination.
- New question : how to balance the sampling of LFS in this setup ?

# Indirect balancing

- Primary Units are balanced on socio-economical variables
- But in this setup, Primary Units sample decide where the LFS clusters can be drawn
- We need to include specific LFS variables in the sampling of the Master Sample to account for this
- How : using weight sharing methods in order to build "indirect balancing variables".

# Conclusion

Using spatial sampling methods :

- Leads to a better precision, even with the coordination with LFS
- Ease the process of data collection
- Local precision :
  - Increased interest on this matter ;
  - Spatial sampling : different impact on different geographical areas ?

Thank you for your attention ! Any question ?