Mining mobile phone data to recognize urban areas

Stéphanie Combes, Marie-Pierre de Bellefon (INSEE),
Maarten Vanhoof (Orange Labs)

Mining mobile phone data to recognize urban areas?

Understanding territory organization, for example in terms of employment, home location and mobility, is crucial for the implementation of policy measures.

In France, the National Statistics Office (INSEE) produces a zoning **(ZAU: Urban Area Zoning)** to identify the geographical extent of cities' influence at the national level: **each French municipality is assigned a type** from Major urban pole to rural municipality (9 classes)

This typology is **complex/long to produce** (several data sources to process, complex algorithm): first published in 2002, it was updated in 2010 (based on 2006-2008 data)

In this study, we aim at evaluating the **potential of mobile phone data** as a complementary source for generating this indicator between two official releases: **a collaboration between INSEE and Orange**

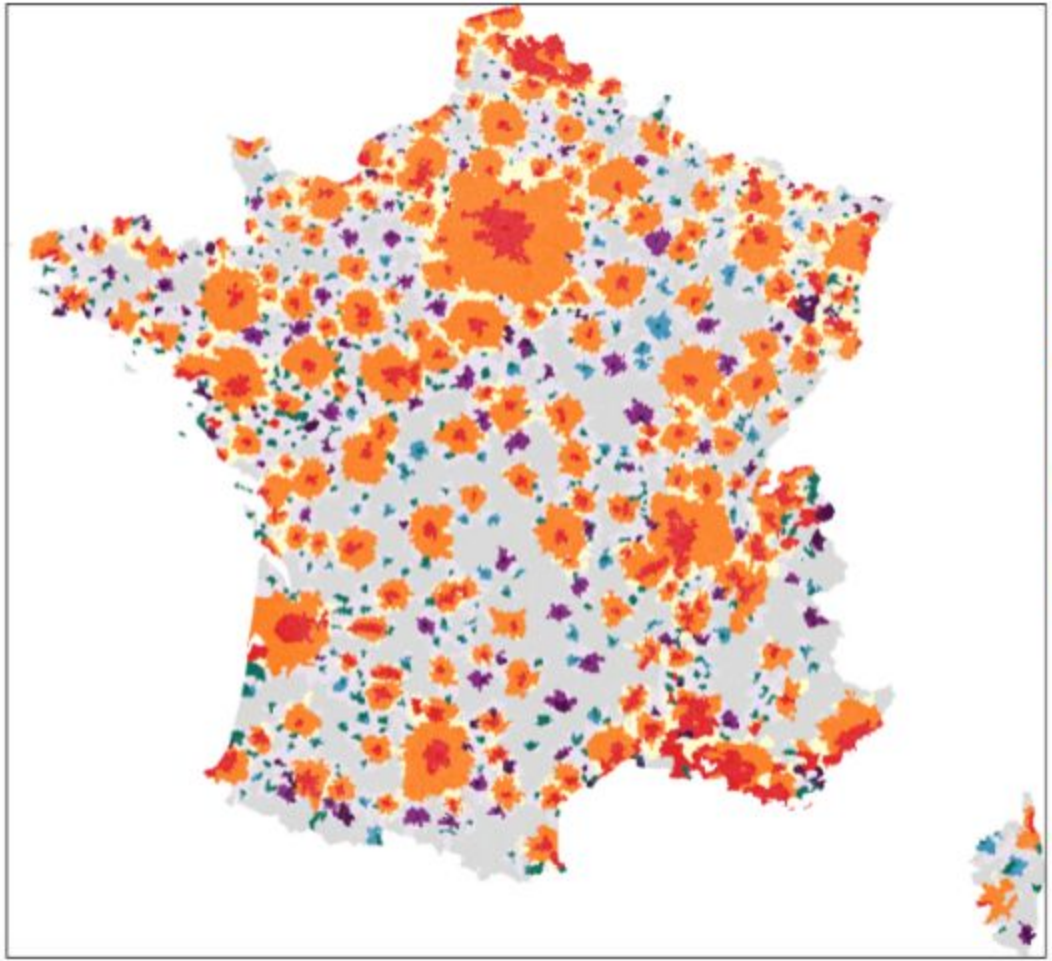
The official urban area typology: ZAU

Built in 2010 out of Census data 2008

- **Pole/urban units:** <200 meters between 2 constructions, > 2000 inhabitants, >1500 jobs
- **Surrounding:** 40 % inhabitant work in the pole or surrounding municipalities
- Urban area = pole +surrounding
- **Multi polarized municipality:** > 40 % inhabitants work in more than two urban areas
- **Isolated municipalities:** less than 40 % inhabitants work in urban areas

Major pole > 10 000 jobs

Small and medium pole = from 1500 to 10000 jobs



Urban Area Zoning 2010

The potential of mobile phone data for territory analysis?

Recently, mobile phone data have shown promising results for land use identification as they provide for **disaggregated, geo-localized and timely information** on cell phone uses of large shares of the population.

The timeliness of the data is interesting since the area covered by municipalities under the influence of major urban centers has increased by **39,5%** between 1999 and 2008 (Floch et Levy, 2011)

Mobile phone data may inform both about **local population density** (network geography) and **mobility** (intuition: people use their phones at different times when they are **at work or at home**, Toole et al 2012)

Data

Aggregated data used here are processed from the **CDR 2007 dataset** which is an anonymized GSM dataset collected by Orange for **billing and operational purposes**.

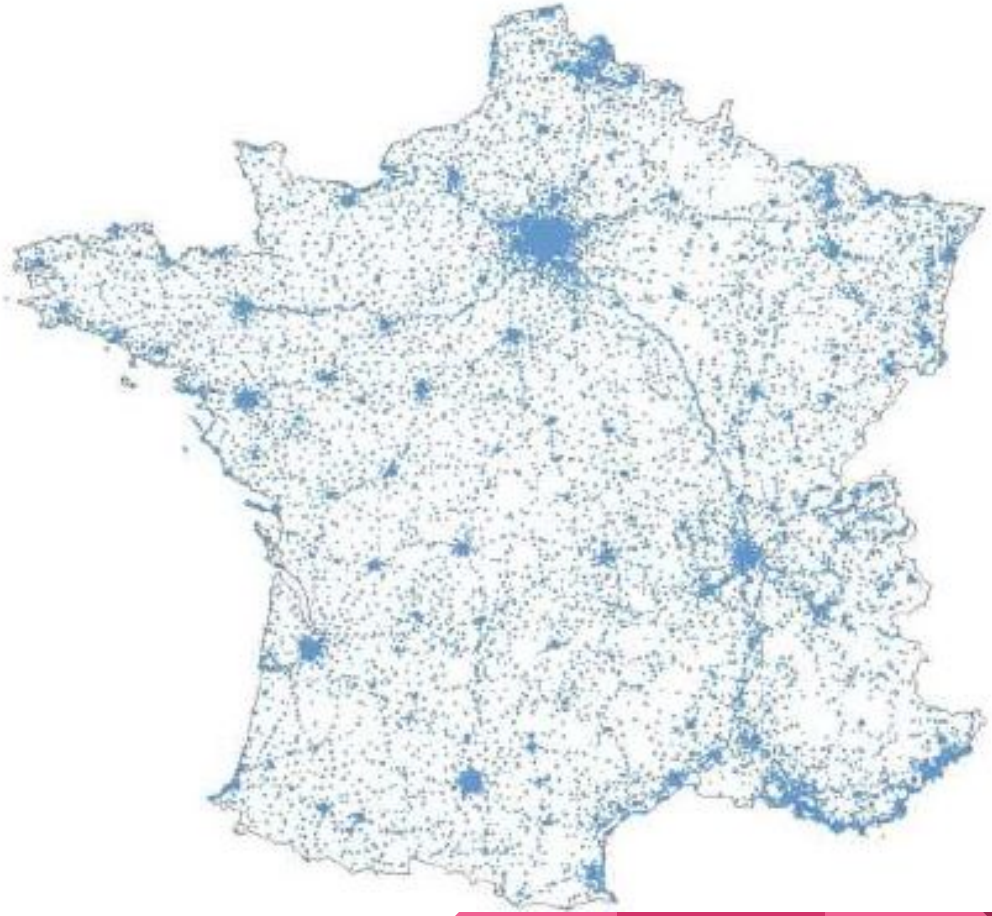
The original dataset consists of Call Detail Records (CDR) describing information about **each phone call and text message (SMS) sent or received by Orange users in a period of 6 months** (from 15/05/2007 to 15/10/2007).

In particular: the **timestamp**, the caller identifier, the callee identifier, the event type (incoming or outgoing call or text message), the duration of the call (in seconds or characters), an urban area identifier **and a tower identifier**.

timestamp	caller	callee	event	duration	area id	tower id
2007/10/01 23:45:00	HJ123423	R482G9342	VO	3656s	1548	53571
2007/13/01 12:10:04	TR234S3	43FG3423	SI	125c	32768	53571

Data

In this study, we exploit the **aggregates of mobile phone events (number of calls/SMS) registered at each tower for each hour** between May and October 2007



The urban area detection problem

Why aggregated? if individual data seem pretty relevant for rendering professional migrations and to a lesser extent local populations, exploiting them raise **legal and technical issues**.

Proposition:

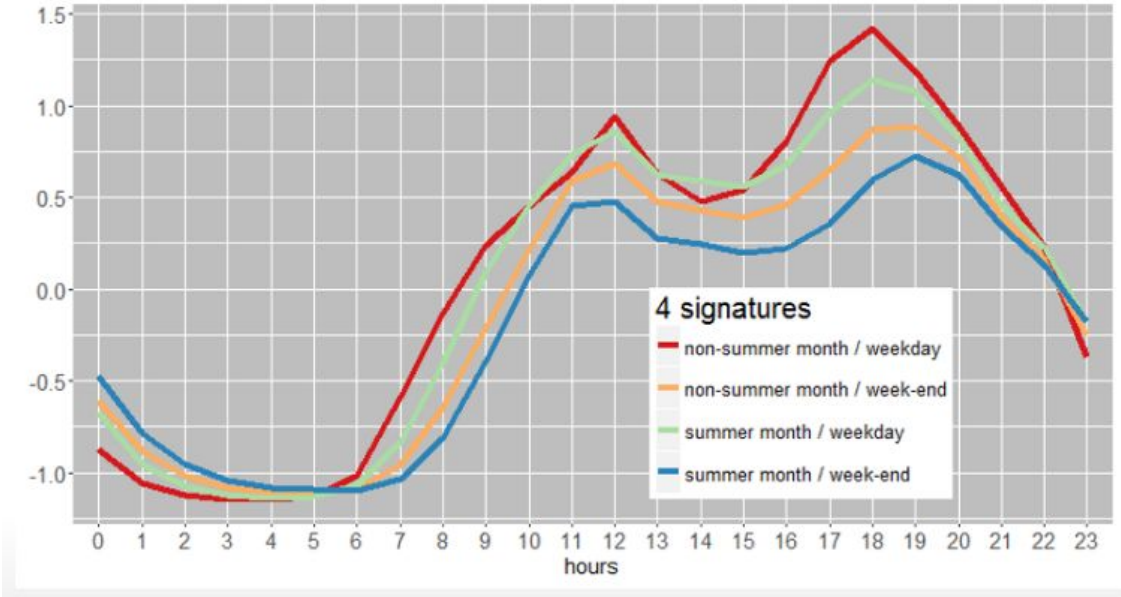
- **Step 1:** turn six months hourly time series of events' counts into **24 hours signature for each tower**.
- **Step 2: link signatures with area types** based on **2010 Zau**, **mapping of towers and municipalities** and **prediction methods**
- **Step 3: evaluate** the performances of step 2

Goal: if the procedure performs really good, it could be used for generating an update of the zoning based on a more recent mobile phone dataset.

Step 1: Features engineering (1/2)

Turning raw time series into signatures for each tower following 3 steps:

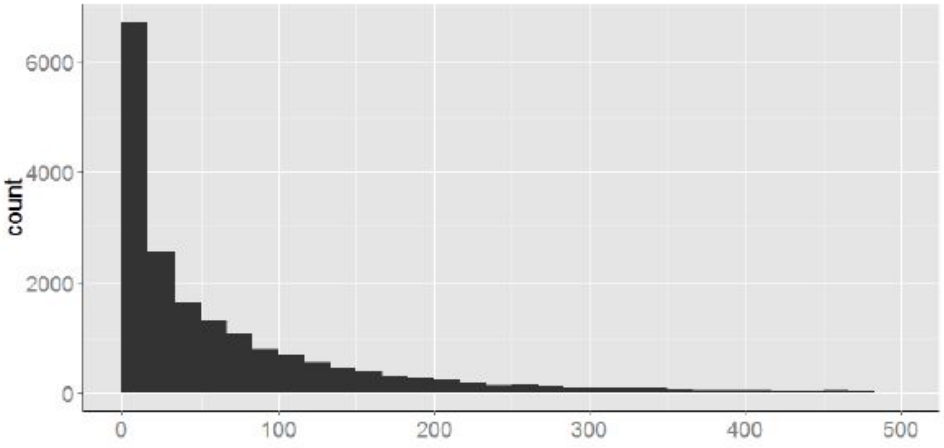
- absolute number of events differs between towers -> series are **standardized** for comparative purposes
- then **averaged** per hour of the day resulting in a typical day activity profile for each antenna
- **week day and week-end day were distinguished**, same for summer / non summer months, **resulting in 4 and not 1 signature per tower**

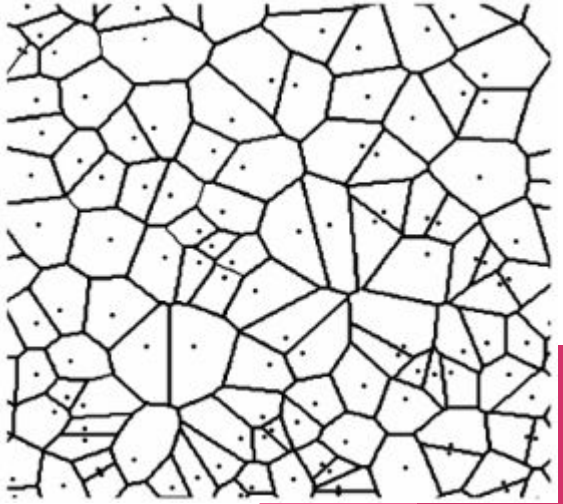


Step 1: Features engineering (2/2)

In addition to these characteristics of activity profiles (which can be regarded as 24x4 features for each tower), we also considered:

- Characteristics informing on **local density of the population**: **averages of events' counts**, **Voronoï cells' shapes**

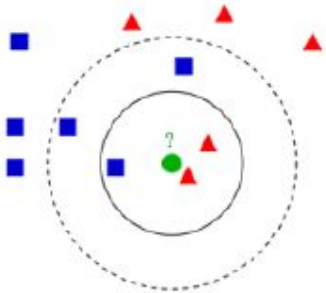




Step 2: prediction tools

Given the differences between official data and mobile phone data, **the same original/official algorithm could not be used for generating a typology with both data sources**.

We therefore recoursed **to machine learning techniques typically used for classification problem. Our goal: design a tool able to predict the area type of a tower given its characteristics (activity profile, voronoi cell shape…).**



The 2010 ZAU is used as a **reference** to estimate the link between characteristics and observed area types (**supervised classification**).

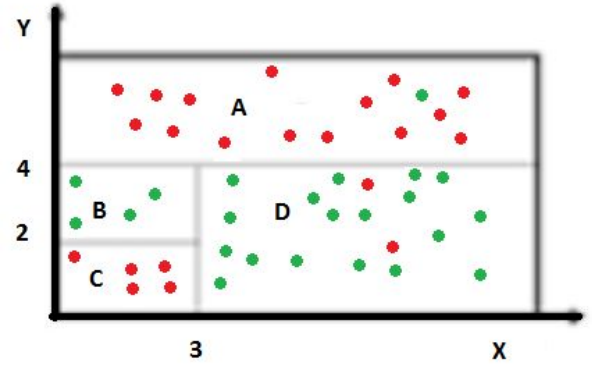
Once calibrated, the tool can be used to predict the area type of a tower given its characteristics measured **one year or more later for example**.

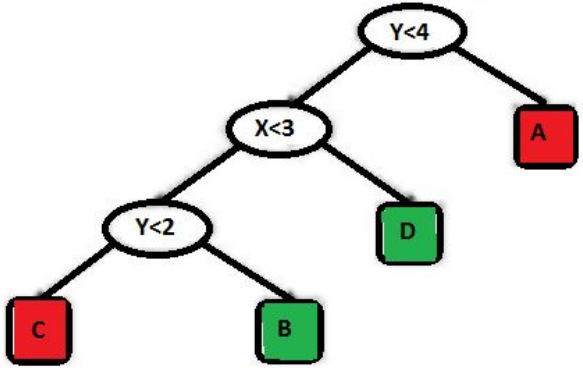
Step 2: prediction tools

Automatic prediction methods are relevant here because **we do *a priori* not know how to extract the useful information** included in those data.

Different methods were tested, but results are given for **random forests**: aggregation of classification trees with double randomisation for more diversity

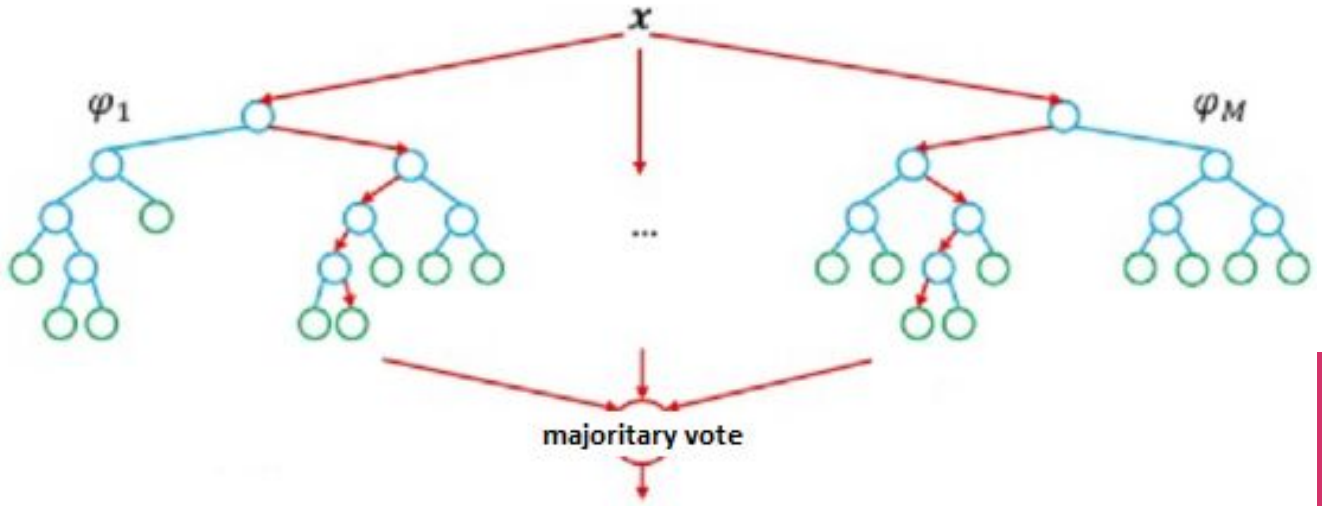
- a classification tree is built by **recursive partitioning** according to a splitting rule **maximising similarity** between observations and a stopping criteria
- the **predicted label** for a new observation is the **majoritary label** in the corresponding leaf.





Step 2: prediction tools

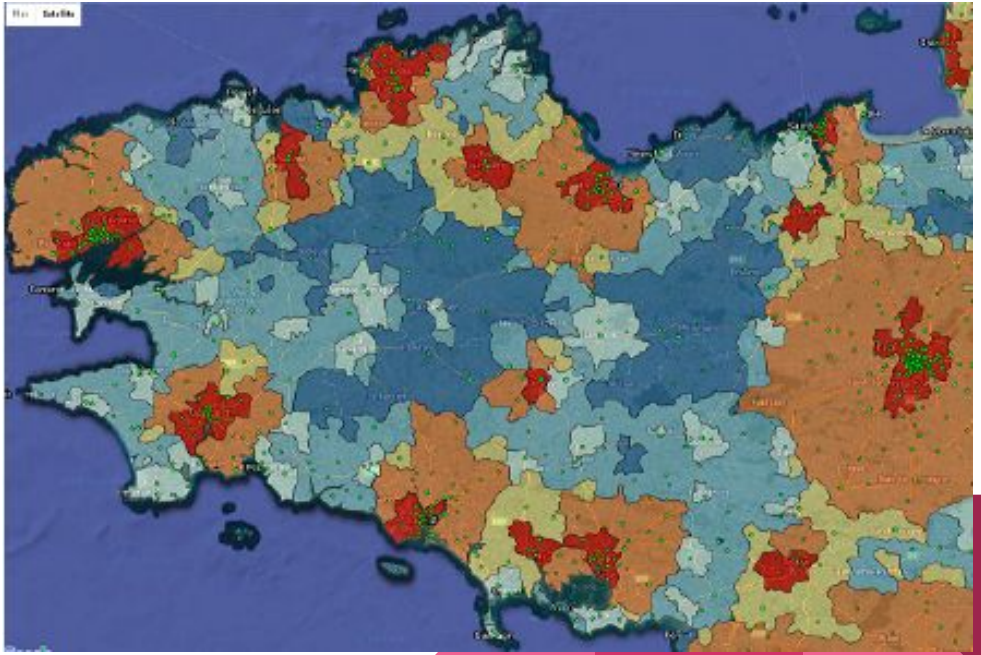
- a tree is a **weak classifier** (but easy to compute and interpret). Idea: aggregate multiple trees to reduce the variance.
- each tree is built on a **bootstrap sample** (bagging) for more diversity, the space of features is also reduced (randomised) in the random forest approach.



Step 2: prediction problem, about the labels

Predicting 9 different area types is not possible, especially because some area types are very badly represented (less than 3%), **some types were merged**

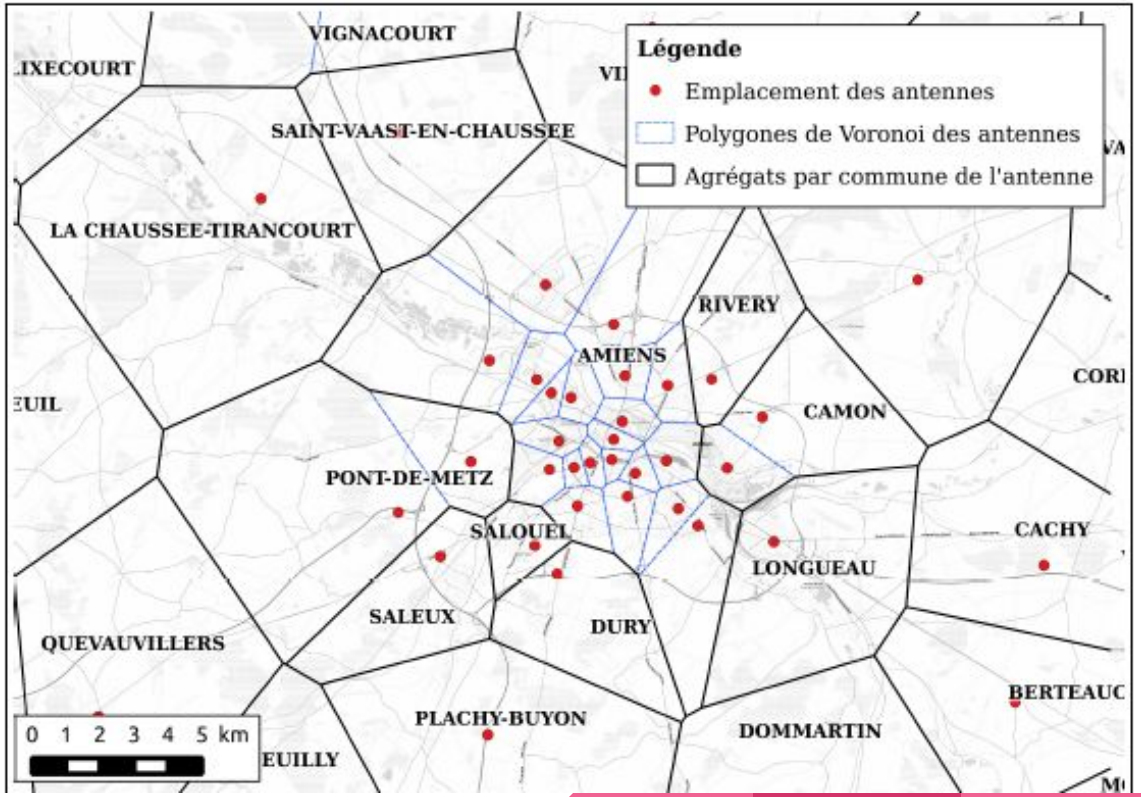
Code	Urban areas	Municipalities	Antennas
111	Major urban center	9%	54%
112	Surroundings of a major urban center	34%	18%
120	Multipolarized in a large urban area	11%	5%
211	Medium urban center	1%	3%
212	Surroundings of a medium urban center	2%	1%
221	Small urban center	2%	3%
222	Surroundings of a small urban center	2%	0.3%
300	Other multipolarized municipality	19%	6%
400	Isolated municipality outside influence	20%	10%



Step 2: prediction problem, about the predicted units

Features were built for towers but we aim at predicting the area type of a **municipality**.

Features were therefore averaged with respect to the **surface of the intersection** between the voronoï cell of a tower with the municipality contours.



Step 3: Evaluation of the performances

To assess the **quality of the prediction**, we need to compare the **predicted** area types using 2007 features with **actual** 2010 ZAU.

Global accuracy of the tool reveals its ability to predict all the classes, **marginal rates of classification** (class per class) must also be analysed.

Different criteria exist for measuring the global accuracy (beyond the global rate of correct classified units): **G-mean** is an evaluation criteria that favors models which detect all the classes (geometric mean of the classification rates per class):

$$G - mean = \left(\prod_{i=1}^{M} R_i \right)^{1/M}$$

Since classes may be semantically close (a municipality in a pole or in its close neighborhood may have similar characteristics), we define a **weighted G-mean** which penalises less a model when semantically close types are mistaken.

Results

We also computed the **Kappa statistic** (and variants accounting for approximations), which is a reference in remote sensing literature: values **between 0.41 and 0.60 are considered moderate**, values between **0.61 and 0.80 substantial.**

Results are **mitigated: excellent prediction of urban clusters** but more difficult disentanglement of suburban and rural areas.

Table 1: Global accuracy

Scenario	G-mean	w G-mean	Fuzzy G	Kappa	Fuzzy K	Accuracy	Accuracy 2
Orange	0.52	0.57	0.55	0.49	0.58	0.61	0.78

Table 2: Class detection rates

Scenario	Class1	Class2	Class3	Class4	Class5	Class6
Orange	0.82	0.44	0.39	0.53	0.46	0.59

The higher level of the weighted G-mean compared to the original G-mean underlines the **fuzziness of our classification problem**

Results

Still, a substantial part of the mitigated results can be explained by the recourse to general classification tools as opposed to the official algorithm **designed specifically for the construction of the ZAU**.

Indeed, running our classification algorithms on **official data close to the one used in the official ZAU** production provides **better but not outstanding performances.**

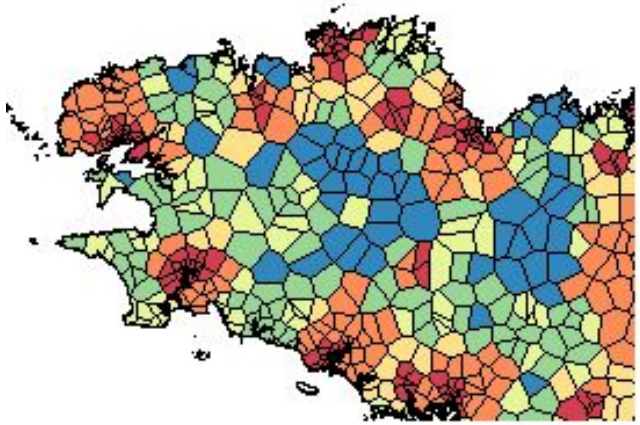
Table 1: Global accuracy

Scenario	G-mean	w G-mean	Fuzzy G	Kappa	Fuzzy K	Accuracy	Accuracy 2
Orange	0.52	0.57	0.55	0.49	0.58	0.61	0.78
INSEE	0.59	0.63	0.61	0.54	0.63	0.65	0.83
INSEE+CORINE	0.63	0.67	0.65	0.61	0.67	0.70	0.86

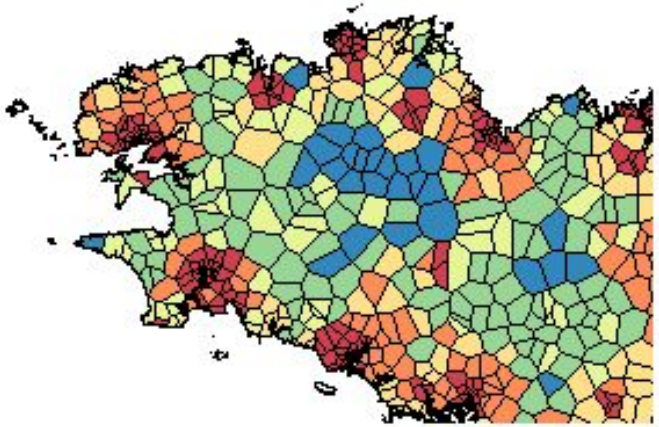
Table 2: Class detection rates

Scenario	Class1	Class2	Class3	Class4	Class5	Class6
Orange	0.82	0.44	0.39	0.53	0.46	0.59
INSEE	0.81	0.53	0.51	0.60	0.45	0.68
INSEE+CORINE	0.87	0.58	0.54	0.62	0.51	0.69

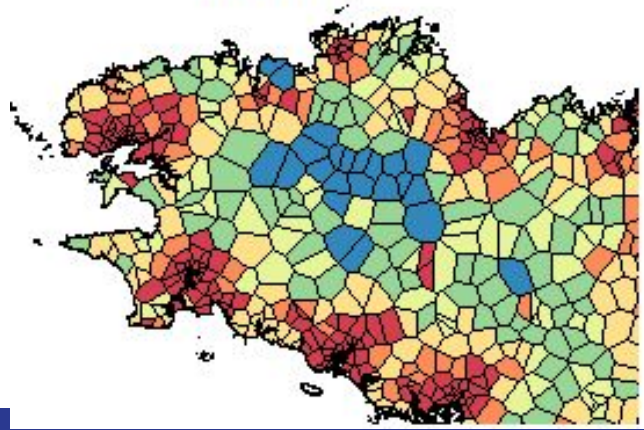
ZAU - official



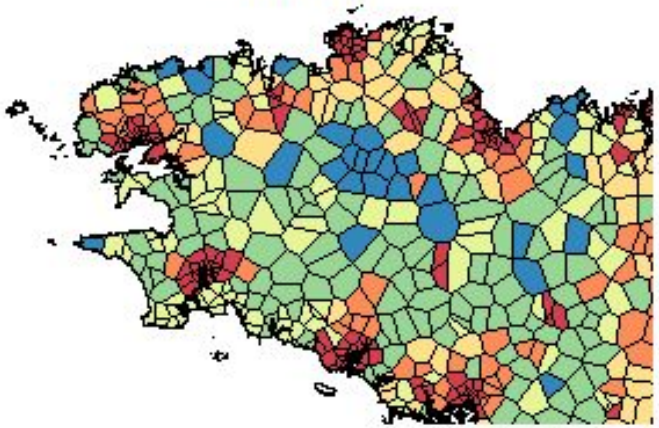
INSEE+CORINE



INSEE



Orange



Results

Given the difficulties to distinguish close classes, we recreated **3 classes**: poles and their surroundings, distinguishing major poles from others, and isolated areas.

With this new typology, results are way better but **still not good enough** to propose an official indicator based on this methodology.

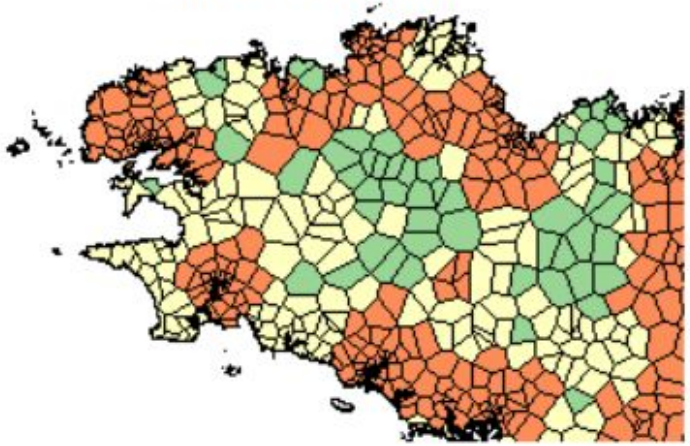
Table 3: Global accuracy

Scenario	G-mean	*w* G-mean	Fuzzy G	Kappa	Fuzzy K	Accuracy	Accuracy 2
Orange	0.71	0.78	0.75	0.58	0.72	0.81	0.94
INSEE	0.80	0.84	0.83	0.68	0.78	0.86	0.97
INSEE+CORINE	0.81	0.85	0.85	0.70	0.80	0.87	0.97

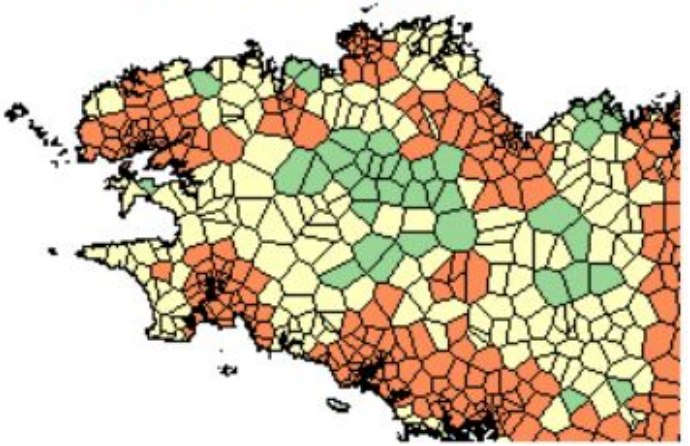
Table 4: Class detection rates

Scenario	Class1	Class2	Class3
Orange	0.87	0.59	0.69
INSEE	0.89	0.73	0.78
INSEE+CORINE	0.90	0.76	0.78

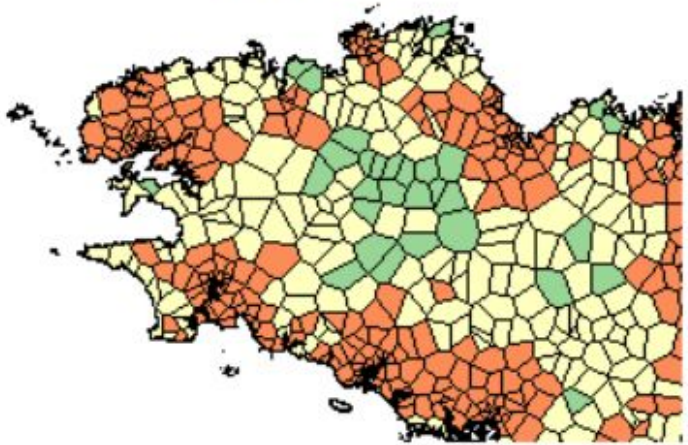
ZAU - official



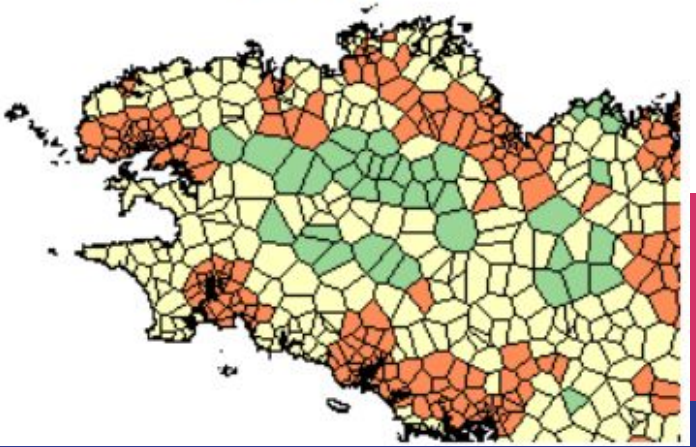
INSEE+CORINE



INSEE



Orange



Conclusion

We tried to answer the question: **can we infer area type from mobile phone patterns**? given that activity of mobile phone owners varies with the type of area: residential vs working...

Results were mitigated, the **classification problem is hard** (multiple classes, uneven distribution, fuzziness, question of scales, questionable reference: iterative algorithm, hard thresholds).

Consequently the lack of accuracy of the Orange scenario **is not exclusively due to the potential limits of mobile phone data** but limits their potential for this application.

Still, the study allowed to **compare a new source with a reliable reference. Mobile phone data as inputs allow an excellent detection of major urban poles (comparable to official data).**

These data **may be interesting to explore further**, probably for some **more local analysis in urban areas** where the mobile phone operator network is dense.

Annex - important features

Table 7: Most important features

Shape Area
Shape Leng
G v8 1 9
event
G v9 0 18
v8 1 12
G v9 0 17
v8 1 17
v8 1 16
G Shape Area

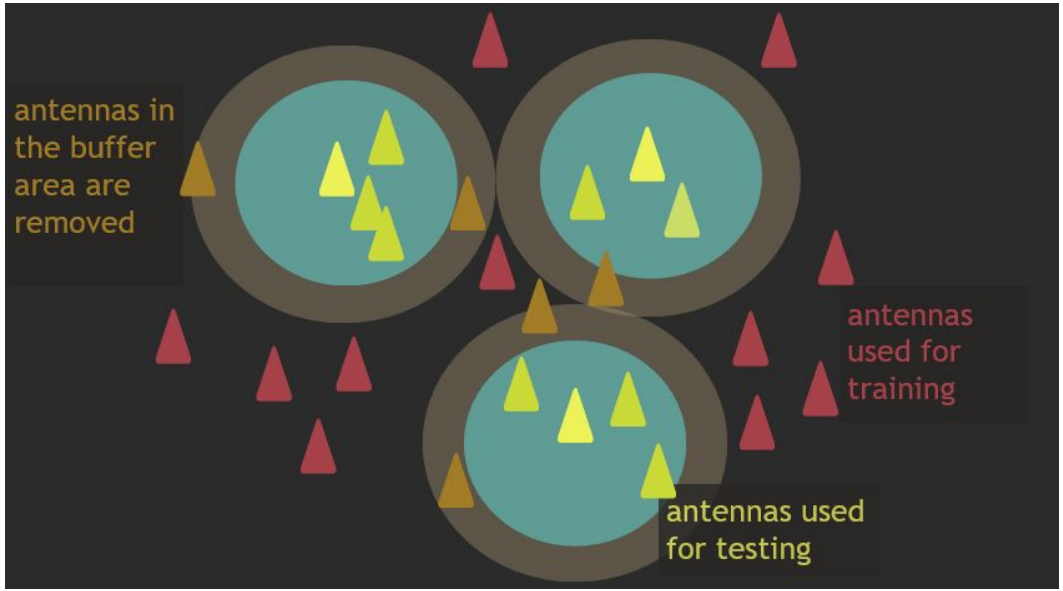
Annex - classic steps in ML

Main classic steps of the procedure :

- **Step 1** Separate the observations between a **training sample** and a **testing sample**
- **Step 2** Use the training sample to **estimate the link between a class** (urban area type) and an antenna depending on its characteristics for a chosen method (for example, random forests) that may need be calibrated during this step.
- **Step 3 Compare and evaluate the quality of different methods** on the testing (and therefore unused) sample: requires a **performance criteria** (founded on the correspondance between predicted urban area type and actual urban area type for each antenna/municipality)

Annex - subtleties

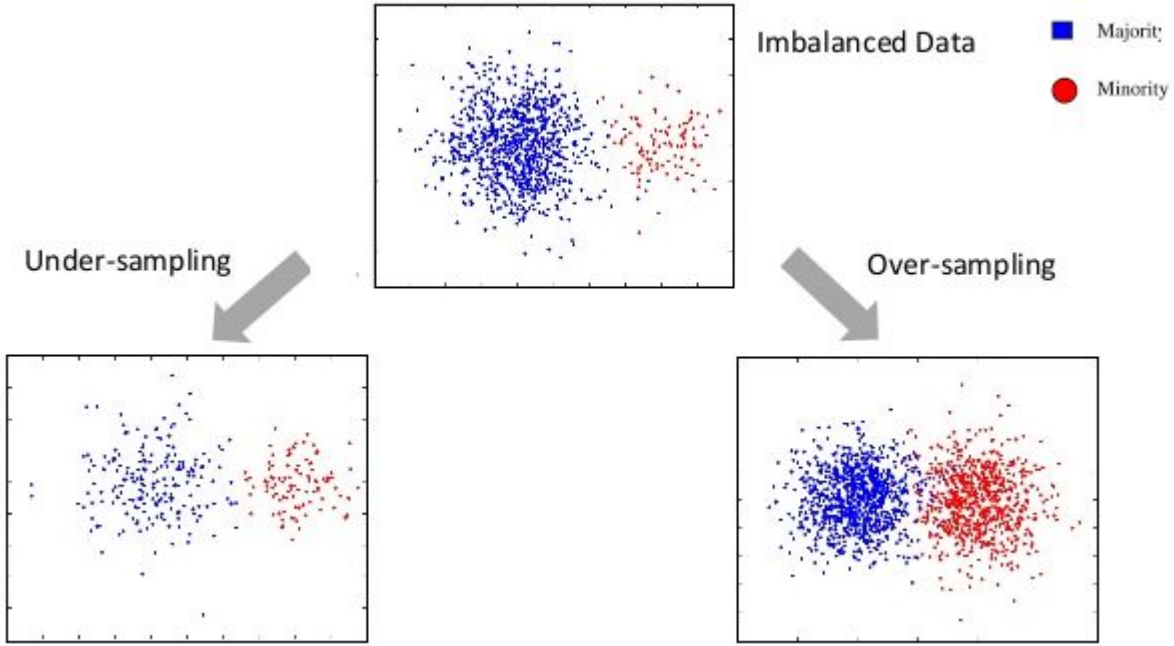
In this problem, we need to pay attention to **spatial decorrelation** of training and testing samples.



We also added to features the spatially lagged features to take into account for the **neighborhood** when predicting a label.

Annex - subtleties

Our data are characterized by uneven distribution (50% of towers in major urban centers), we can recourse to sampling methods when preprocessing



Annex - subtleties

To compute weighted G-mean we use a confusion matrix and a weight/cost matrix

Confusion matrix

		Predicted Class		
		Class 1	Class 2	Class 3
Actual Class	Class 1	2	1	1
	Class 2	1	2	1
	Class 3	1	2	3

weight matrix

	Class 1	Class 2	Class 3
Class 1	1	0.6	1
Class 2	0.6	1	1
Class 3	1	1	1

Annex - fuzziness of the problem