Étude théorique : stations de base du réseau téléphonique français

Progression du stage

Paul MÉHAUD, Brendan SÉVELLEC

České Vysoké Učení Technické v Praze

mai - août 2024

Déroulé de la présentation

Introduction

Contexte général

Objectifs

- Déterminer si les stations de base sont en zone urbaine ou rurale ;
- Chercher les stations de bases voisines les unes des autres pour aider à déterminer si les utilisateurs sont en mouvement.

Méthodes

- Approche par la théorie des graphes ;
- Approche par le machine learning.

Données

Données

Arcep

Autorité de régulation des communications électroniques, des postes et de la distribution de la presse.

Jeu de données

Le jeu de données `2023_T4_sites_Metropole.csv`[a] représente les stations de bases au trimestre 4 de 2023 avec leur position géographique (taille : $16,7\,\mathrm{Mo}$).

a. https://data.arcep.fr/mobile/sites/

A retenir :

- $108\,838$ sites ;
- 29 attributs.

A quoi ressemble notre base ?

code_op	nom_op	num_site	id_site_partage	id_station_anfr	x	y	latitude	longitude	nom_reg
20801	Orange	00000001A1	nan	0802290015	687035	6985761	49,97028	2,81944	Hauts-de-France
20801	Orange	00000001B1	nan	0642290151	422853	6249263	43,28861	-0,41389	Nouvelle-Aquitaine
20801	Orange	00000001B2	nan	0332290026	416932	6422196	44,84112	-0,58333	Nouvelle-Aquitaine
20801	Orange	00000001B3	nan	0472290005	511106	6349234	44,21666	0,63556	Nouvelle-Aquitaine
20801	Orange	00000001C1	nan	0512290147	836824	6889450	49,09028	4,87333	Grand Est

nom_dep	insee_dep	nom_com	insee_com	site_2g	site_3g	site_4g	site_5g	mes_4g_trim	site_ZB
Somme	80	Curlu	80231	1	1	1	0	0	0
Pyrénées-Atlantiques	64	Jurançon	64284	1	1	1	1	0	0
Gironde	33	Bordeaux	33063	1	1	1	1	0	0
Lot-et-Garonne	47	Agen	47001	1	1	1	0	0	0
Marne	51	Sainte-Menehould	51507	1	1	1	0	0	0

site_DCC	site_strategique	site_capa_240mbps	date_ouverturecommerciale_5g	site_5g_700_m_hz	site_5g_800_m_hz
0	0	0	nan	0	0
0	0	1	2020-12-14	0	0
0	0	1	2021-02-22	0	0
0	0	1	nan	0	0
0	0	1	nan	0	0

site_5g_1800_m_hz	site_5g_2100_m_hz	site_5g_3500_m_hz
0	0	0
0	1	0
0	0	1
0	0	0
0	0	0

Table 1 – Premières valeurs de la base

ČVUT
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

Description (1/2)

Ce qui nous intéresse

1. *longitude*, *latitude* : coordonnées de chaque site ;
2. *nom_op* : nom commercial de l'opérateur ;
3. *nom_reg*, *nom_dep* et *nom_com* : nom de la région, du département et de la commune d'implantation du site ;
4. *site_xg* : équipement du site en technologie xG ($x \in \{2, \dots, 5\}$) ;
5. *num_site* : identifiant du site issu du SI de l'opérateur.

Description (2/2)

Ce qu'il faut retenir

1. Répartition équitable du nombre de sites en fonction de l'opérateur ($\simeq 27\,000$) ;
2. $99,6\%$ des sites équipés en 4G ;
3. 6 stations en moyenne par commune.

La construction de cette base ne nous permet pas de faire de statistiques descriptives intéressantes.

Semaine du 03/06/24 au 06/06/24

Semaine du 03/06/24 au 06/06/24

Comparaison des résultats de 2 détection de villes

Affichage

Affichage des différences

Afin de visualiser les différences, nous pouvons tracer les 2 classifications en même temps sur une carte en adoptant ce code couleur :

- Rouge : station en ville dans les 2 classification ;
- Bleu : station en ville dans exactement 1 classification ;
- Vert : station en campagne dans les 2 classifications.

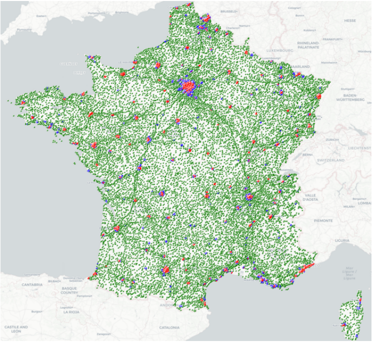


Figure 1 – Comparaison des résultats de détection de villes : DBScan vs HDBScan

Indicateurs

Indicateurs de base

Nous avons dévéloppé quatre indicateurs de base afin de caractériser la similarité entre deux classifications :

- a : Le pourcentage de station en ville dans les deux classifications ;
- b : Le pourcentage de station en ville dans la première classification et non la deuxième ;
- c : Le pourcentage de station en ville dans la deuxième classification et non la première ;
- d : Le pourcentage de station en campagne dans les deux classifications.

Résultats

Sur les classifications présentées dans la slide précédentes, voici les résultats obtenus :

- $a = 0,688$;
- $b = 0,028$;
- $c = 0,051$;
- $d = 0,232$.

Semaine du 03/06/24 au 06/06/24

Nouvelle manière de détecter les villes

ČVUT
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

Changement de cap

Problèmes rencontrés avec DBScan et HDBScan

- Méthodes complexes donc peu prévisibles ;
- Résultats non-satisfaisants quand le méthode est appliquée à seulement une partie de la France.
- DBScan est trop binaire, HDBScan est à densité variable donc ne détecte pas toutes les villes de la même façon (problème avec Paris notamment)
- Les probabilités de HDBScan ne caractérisent pas exactement la probabilité d'être en ville mais plutôt la certitude avec laquelle on peut rattacher un point à son cluster.

Réflexion sur une méthode plus simple

Nous souhaitions savoir quelle station se trouvait en ville, car on sait qu'en ville, les stations sont plus proches les unes des autres, donc les rayons de couverture plus courts. Cependant, il n'est pas nécessaire de detecter les villes pour cela, nous pouvons simplement regarder la distance moyenne aux k plus proches voisins.

Méthodologie

Nouvelle méthode

Au lieu d'utiliser une méthode de clustering, nous allons utiliser quelque chose de plus simple. On classifie chaque station selon la distance moyenne aux 3 plus proches voisins. Soit d cette distance, on regroupe les stations de la manière suivante :

- $d \in]0, 1]$: centre ville dense ;
- $d \in]1, 2]$: couronne périhurbaine ;
- $d \in]2, 5]$: campagne ;
- $d \in]5, 10]$: campagne profonde ;
- $d \in]10, \infty[$: trou paumé ou valeur aberrante.

Méthodologie : choix techniques

Calcul des plus proches voisins

Nous utilisons la bibliothèque *sklearn.neighbors.NearestNeighbors*[1].

Choix du nombre de voisins

Après plusieurs expérimentations, nous avons choisi de conserver 3 voisins dans notre calcul. En effet, un chiffre inférieur à 2 serait aberrant car ce ne serait pas une vraie moyenne (ici on cherche plus une mesure de la densité de stations). De plus, un chiffre supérieur à 4 prendrait en compte des stations trop éloignées, ce qui serait aberrant.

Métrique

Nous avons décidé de partir de la projection de Lambert 93 (qui est déjà présente dans nos données) pour calculer cette distance. L'avantage est le gain de temps (environ 30 fois plus rapide), sans perte de performance, par rapport à la conversion en km depuis les coordonnées Longitude, Latitude.

1. `https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.NearestNeighbors.html`

Mise à jour des critères

Angle

Voici les différents paliers que nous appliquons :

- $d \in]0,1]$: angle_min $= 40°$;
- $d \in]1,2]$: angle_min $= 30°$;
- $d \in]2,5]$: angle_min $= 25°$;
- $d \in]5,10]$: angle_min $= 15°$;
- $d \in]10,\infty[$: angle_min $= 10°$.

Distance

Voici les différents paliers que nous appliquons :

- $d \in]0,1]$: distance_max $= 2\,\mathrm{km}$;
- $d \in]1,2]$: distance_max $= 5\,\mathrm{km}$;
- $d \in]2,5]$: distance_max $= 10\,\mathrm{km}$;
- $d \in]5,10]$: distance_max $= 15\,\mathrm{km}$;
- $d \in]10,\infty[$: distance_max $= 15\,\mathrm{km}$.

From 10/06/24 to 14/06/24

ČVUT
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

From 10/06/24 to 14/06/24

Modification of criteria

New criteria and city classification

After some reflexions, we have agreed to change the way we classify the city-ness of each base station :

New city-ness classification

- $d \in]0, 1]$: city center;
- $d \in]1, 2]$: urban area;
- $d \in]2, 4]$: extra-urban area;
- $d \in]4, \infty[$: courntryside.

Angle

- $d \in]0, 1]$: angle_min $= 40°$;
- $d \in]1, 2]$: angle_min $= 30°$;
- $d \in]2, 4]$: angle_min $= 25°$;
- $d \in]4, \infty[$: angle_min $= 15°$.

Distance

- $d \in]0, 1]$: distance_max $= 2\,\text{km}$;
- $d \in]1, 2]$: distance_max $= 5\,\text{km}$;
- $d \in]2, 4]$: distance_max $= 10\,\text{km}$;
- $d \in]4, \infty[$: distance_max $= 15\,\text{km}$.

From 10/06/24 to 14/06/24

City detection results comparison

Reminder : the brute results

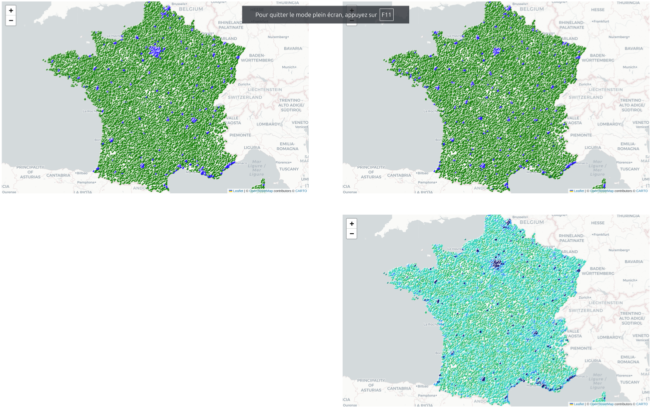


Figure 2 – City detection by respectively DBScan, HDBScan and 3-NN methods

Graphical methods comparison

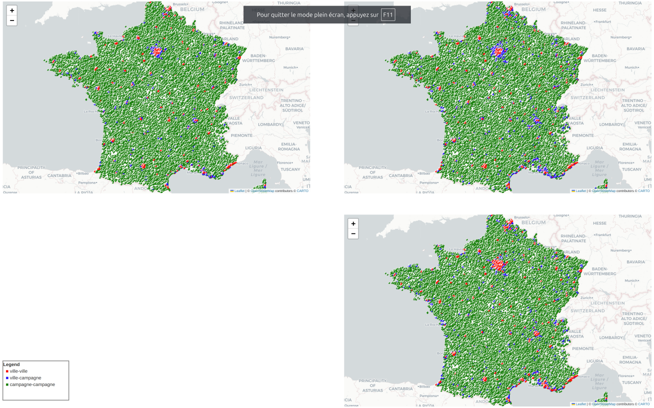


Figure 3 – DBScan vs HDBScan / HDBScan vs 3-NN / DBScan vs 3-NN

Numerical methods comparison

DBScan vs HDBScan

- $a = 0.688$
- $b = 0.028$
- $c = 0.052$
- $d = 0.232$

HDBScan vs 3-NN

- $a = 0.626$
- $b = 0.114$
- $c = 0.010$
- $d = 0.250$

DBScan vs 3-NN

- $a = 0.630$
- $b = 0.086$
- $c = 0.007$
- $d = 0.277$

From 10/06/24 to 14/06/24

Improvement of City detection

Reminder : functionnement of base 3-NN

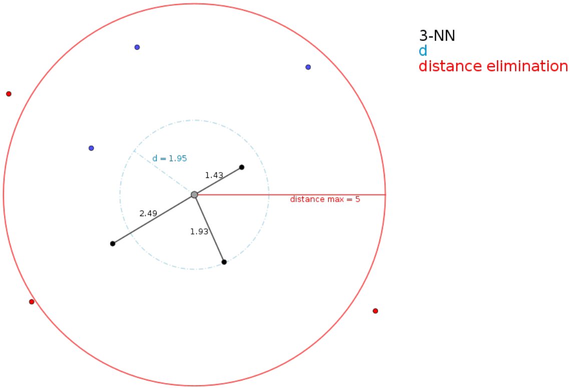


Figure 4 – How 3-NN method works

From 17/06/24 to 21/06/24

From 17/06/24 to 21/06/24

Road detection method - in progress

Methodology

We have seen that one classification with either DBScan or HDBScan isn't enough. So, why not combining them.

The base idea

We will combine the clusters of DBScan, HDBScan and OPTICS (another density base clustering method) to try to find roads.

One problem

We are now able to detect road/railways but the problem is that there is still little clusters that are useless.

New method illustration (1/2)

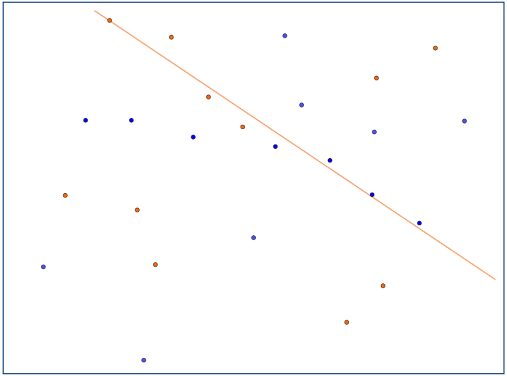


Figure 5 – One clustering method

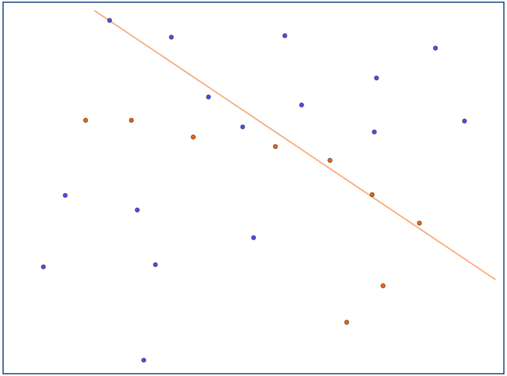


Figure 6 – Another clustering method

New method illustration (2/2)

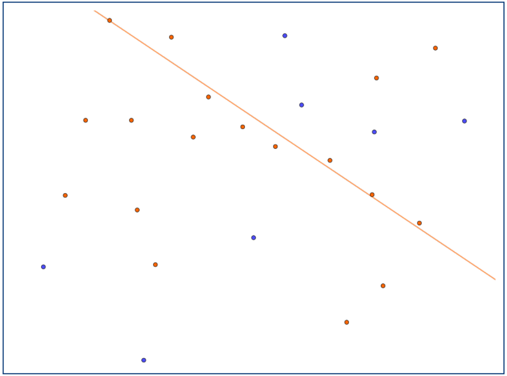


Figure 7 – Merge of the clusters

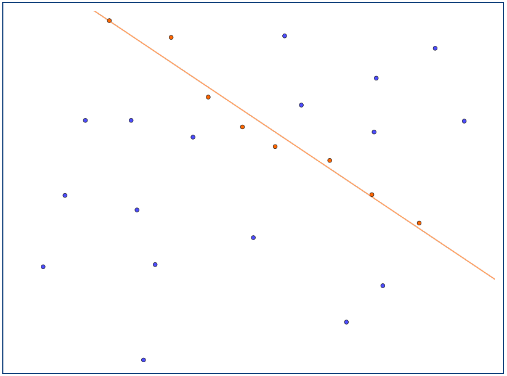


Figure 8 – Elimination of noise

DBScan clustering

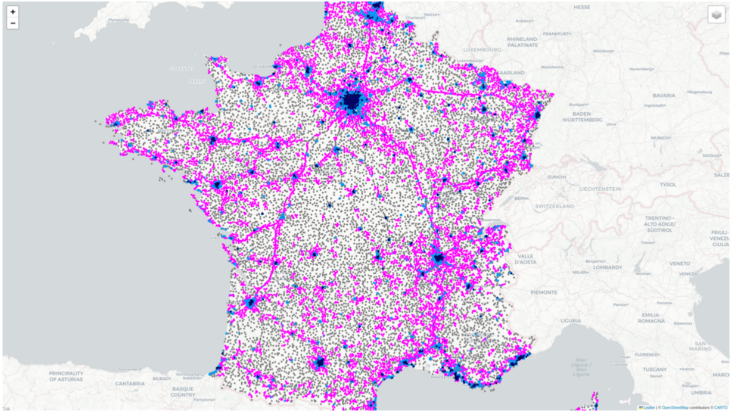


Figure 9 – DBScan clustering

HDBScan clustering

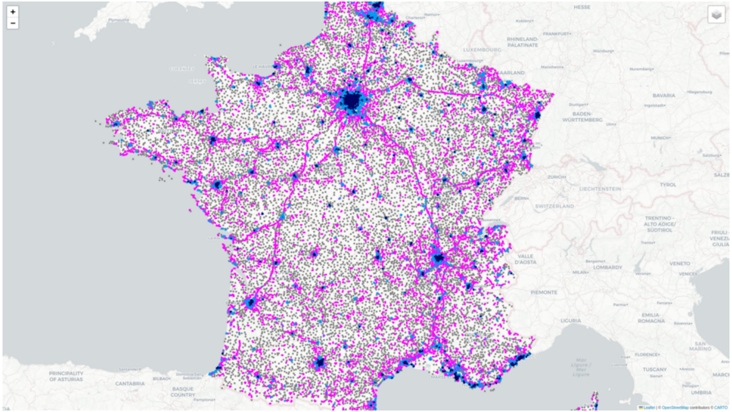


Figure 10 – HDBScan clustering

OPTICS clustering

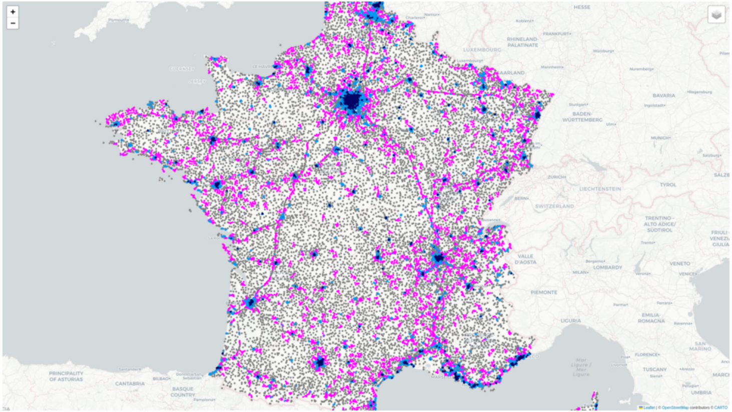


Figure 11 – OPTICS clustering

Result

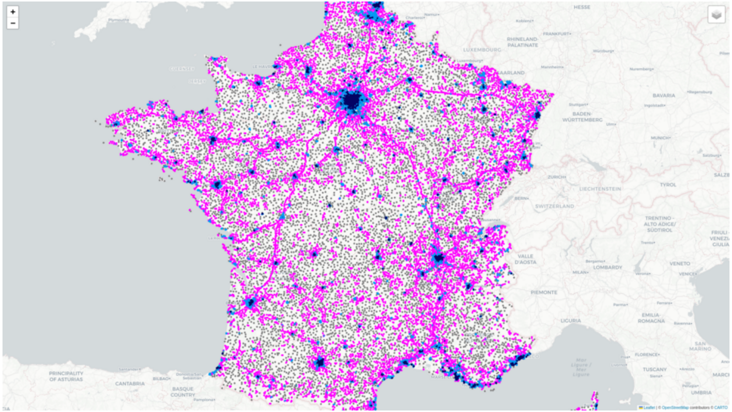


Figure 12 – Road detection

From 17/06/24 to 21/06/24

More details and problems

ČVUT
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

Detailed clusters detected

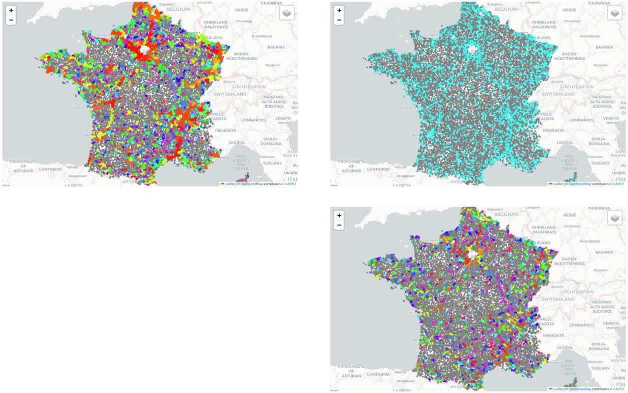


Figure 13 – Detailed clusters detected in coutryside by each method

Problems and ideas

Problems

- A lot of little citys detected : not only roads ;
- The areas around big citys are a mess ;
- HDBScan has only one cluster.

Ideas

- Refine the parameters of each method ;
- Use linear regressions to detect parts of roads and maybe help propagate them.

From 17/06/24 to 21/06/24

Altair AI Studio Software Review

ČVUT
ČESKÉ VYSOKÉ
UČENÍ TECHNICKÉ
V PRAZE

Altair AI Studio Software Review

Objective is to evaluate Altair AI Studio for their potential to enhance our research on classifying the terrain of mobile base stations. Available at: `https://altair.com/altair-ai-studio`

Altair AI Studio

This is a platform designed for data analysis and machine learning model building. Possible benefits for us:

- Supports clustering and classification algorithms.
- Support for various machine learning algorithms.
- Enables effective result visualization (graphs) for enhanced analysis.

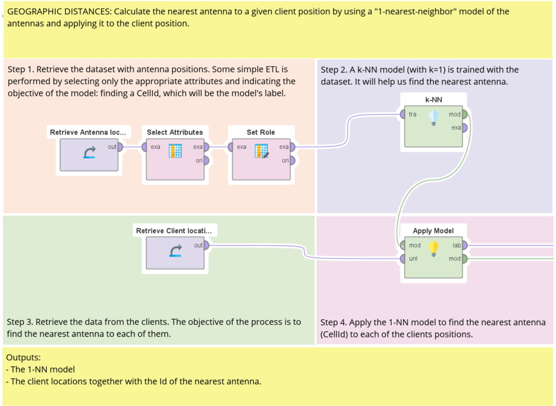
Key Features:

- Integration with various data sources.
- Interactive model creation and testing.
- Support for various machine learning algorithms.

Users:

- Data researchers
- Analysts
- Machine learning developers

Example of Usage in Altair AI Studio



Result:

id	prediction(CellId)	CoordinateX	CoordinateY
1	1251	611814	5636166
2	1305	595066	5627691
3	4187	599336	5588247
4	1305	587844	5601378
5	7131	608191	5634863
6	8757	592244	5659340
7	1305	589278	5589412
8	312	616828	5652313
9	1305	580590	5626004
10	1305	590716	5624514

Figure 14 – How Altair AI Software works

From 17/06/24 to 21/06/24

New Possible Approach for Terrain Classification

More accurate verification of the classification of base station locations

We want to validate and enhance the previous classification of base station locations by leveraging a new geographic dataset. We found and will try to use a comprehensive dataset from `data.enseignementsup-recherche.gouv.fr`, which includes precise geographic data points across France with pre-defined zone classifications.

Methodology for Base Station Classification

Previously, we worked with a dataset containing locations of all base stations. Now, we utilize an additional dataset with geographic points across France, each classified into zones (urban, suburban, rural). By comparing base station coordinates with these geographic points, we determine the zone of each base station. The closest geographic point's classification is assigned to the base station, providing a more accurate terrain classification.

Benefits:

- Cross-reference base station coordinates with geographic data points to accurately assign zone classifications.
- Provides an additional layer of verification and accuracy for our classification methods.

Classification of points in the new dataset

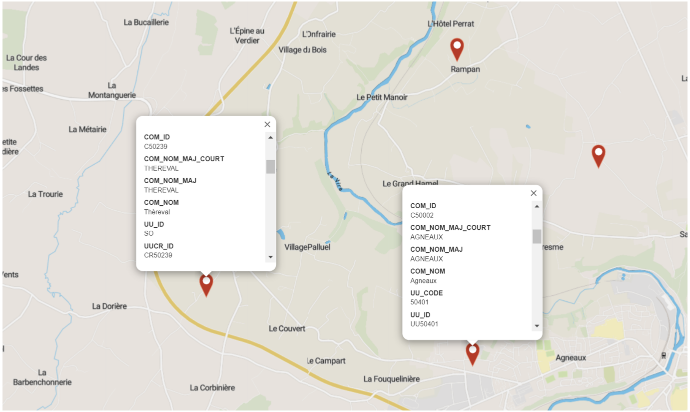


Figure 15 – Classification of points from the new dataset

Classification of points in the new dataset

Geographic identifiers:

- UU (Urban Units): Areas classified as urban units.
- CR (Rural Communes): Areas classified as rural communes.
- AU (Urban Areas): Larger urban areas or agglomerations.

Methodology and Challenges

This method helps classify base station locations by comparing them with geographic data points. However, we need to preprocess the dataset carefully and address the issue of having a limited number of points. To ensure high accuracy, we might need to augment the data or combine it with other datasets for better coverage and precision

From 17/06/24 to 21/06/24

Another new database

New possibilities on city/countryside classification

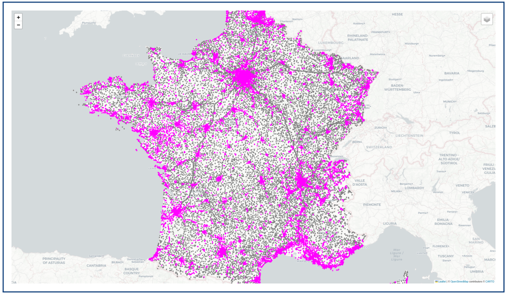


Figure 16 – Database we presented last week

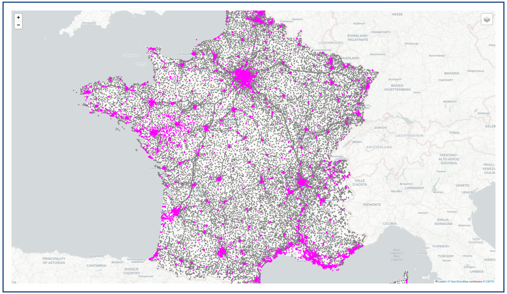


Figure 17 – A new database (cities with more than 5000 inhabitants)

What's this?

We found a new database on the population of each french commune in 2021 [2].

INSEE?

It is the National Institute of Statistics and Economics Studies. It collects, analyses and disseminates information on the French economy and society.

Why this one?

Firstly, this database is really recent (updated on 26/01/2024). Then, the other database we introduced was to permisive in the definition of what is a city (a commune with more than 2000 inhabitants). So we wanted something more flexible.

2. https://www.insee.fr/fr/statistiques/7739582

From 17/06/24 to 21/06/24

Back to road detection

HDBScan parameters adaptation

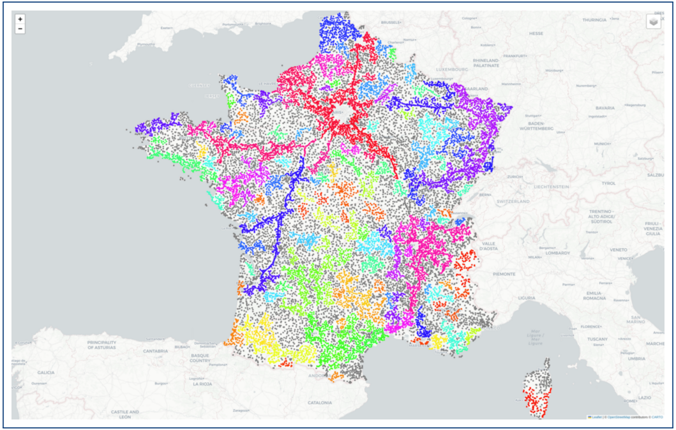


Figure 18 – Result of HDBScan with new parameters on countryside to detect roads

A bit of math theory (1/2)

Multiple linear regression

Let a dataset be composed of m vectors$\left(x^i\right)_{i\in\{1\ldots m\}}$ of $\mathbb{R}^n$, ie $\forall i \in \{1,\ldots,m\}$, $x^i = (x^i_1,\ldots,x^i_n)$.
Let $\left(y^i\right)_{i\in\{1\ldots m\}}$ be m associated values in $\mathbb{R}$.
Then, we can find a linear function $f : \mathbb{R}^n \to \mathbb{R}$ (ie $f(x_1,\ldots,x_n) = c_1x_1 + \cdots + c_nx_n$) that minimises the error : $\|\sum_{i=1}^{m}(f(x^i) - y^i)\|$
f is called a multiple linear regression associated to $\left(x^i\right)_{i\in\{1\ldots m\}}$ and $\left(y^i\right)_{i\in\{1\ldots m\}}$.

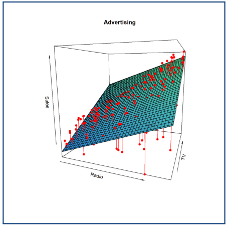


Figure 19 – Example of multiple linear regression with $n = 2$

A bit of math theory (2/2)

Polynomial regression

Let a dataset be composed of m values $(x^i)_{i\in\{1\ldots m\}}$ of $\mathbb{R}$.
Let $(y^i)_{i\in\{1\ldots m\}}$ be m associated values in $\mathbb{R}$.
We can transform the data in m vectors$(x^i)_{i\in\{1\ldots m\}}$ of $\mathbb{R}^{n+1}$ defined by $(x^i_0, x^i_1, x^i_2, \ldots, x^i_n) = (1, x^i, (x^i)^2, \ldots, (x^i)^n)$, and then apply the multiple linear regression seen earlier on this transomed data.
We therefore obtain a function

$$f : \begin{cases} \mathbb{R} \to \mathbb{R} \\ x \mapsto c_1 + c_2 x + \cdots + c_n x^n \end{cases}$$

That is a degree n polynome that approximates the relation between x and y.

Quantify the quality of the approximation

As for the simple linear regression, we obtain a R coefficient between 0 and 1 that indicates if the approximation is close to the original data (i.e. if the polynome we find at the end is close to the real appearance of the data).

Application of the method on each cluster of HDBScan

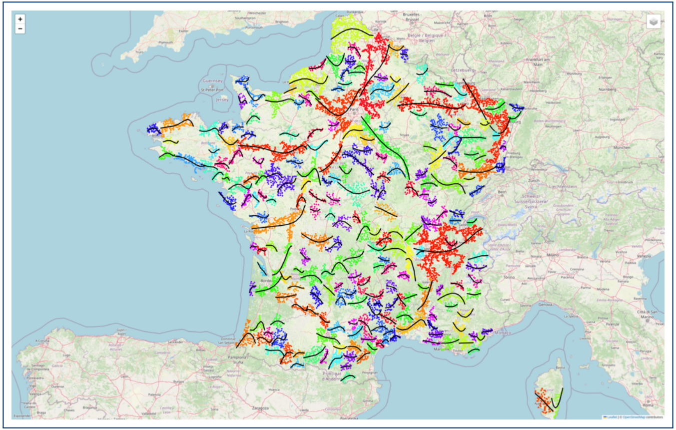


Figure 20 – Approximation of each cluster by a degree 20 polynome

Filtration of the cluster by keeping the ones that have a R coefficient above a value

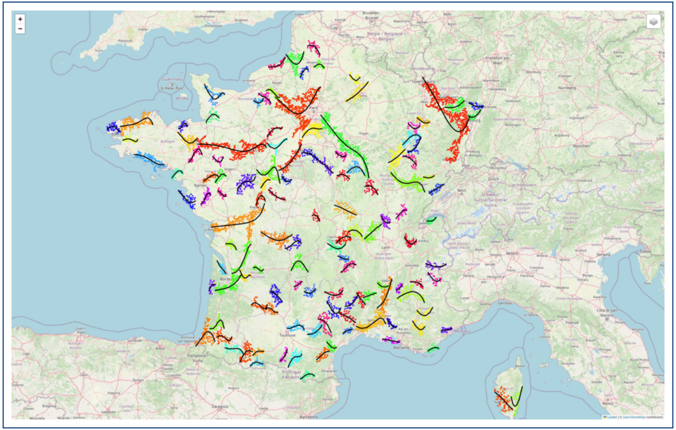


Figure 21 – Approximation of each cluster by a degree 20 polynome, keeping only the ones with a R coefficient above 0.2

From 24/06/24 to 28/06/24

From 24/06/24 to 28/06/24

New ideas on neighbours detection

Neighbours detection without Delaunay

Just to see, we asked ourselves: why not trying to find neighbours without the Delaunay triangulation? Here is our methodology:

Creating the potential-neighbours-graph

The base idea was to firstly create a graph. But how?
We will use the maximum distance to connect base stations according to the distance criteria, i.e. we connect a base station to surrounding stations that are in the radius of $max\_distance$.

And then we apply the angle and quadrant criteria like before.

Is it worth it?

It gives results that are more or less the same than before, but there are some imperfections. So, we could say that this is interesting but not really useful.

From 24/06/24 to 28/06/24

Here come...

...a newer database!

Because the other databases we found earlier were not really convincing (in fact they are just useful to compare results). We therefore did some other research and found some interesting things!

ANFR

The National FRequency Agency, a public administrative body, was set up by the Telecommunications Regulation Act of 26 July 1996 to plan, manage and monitor the use of public radio frequencies in France.

The database itself

In this database [3], you have a map of all base stations in France. The interest of this database is that we can have all the frequencies used in one site and the orientation of the antennas.
However, the problem is that we cannot download this database.

So, let's find a downloadable database.

3. `https://www.cartoradio.fr/#/cartographie/stations`

. . . a newer database! - bis

This is another database from ANFR. This time we have the frequencies used in the databases.
We linked those results in the first database by only keeping the highest frequency per base station.

From 24/06/24 to 28/06/24

Analysis of Frequency Data for Mobile Operators in France

France spectrum - mobile network frequency plan

We have detailed data on the frequency spectrums used by mobile operators in France. This data allows us to understand the specific frequencies at which each base station operates.

Frequency Bands:

- The primary frequency bands used for 4G LTE in France include 700 MHz, 800 MHz, 1800 MHz, 2100 MHz, and 2600 MHz. Each frequency band has its own characteristics in terms of range, penetration, and data capacity.

Layer (MHz)	3GPP band	FDD / TDD / SDL	Uplink Start (MHz)	Uplink End (MHz)	Downlink Start (MHz)	Downlink End (MHz)	Total (MHz)
700	n28	FDD	708.0	718.0	763.0	773.0	10.0
800	n20	FDD	852.0	862.0	811.0	821.0	10.0
900	n8	FDD	888.8	898.6	933.8	943.6	9.8
900	n8	FDD	888.8	897.5	933.8	942.5	8.7
900	n8	FDD	889.9	898.6	934.9	943.6	8.7
1800	n3	FDD	1710.0	1730.0	1805.0	1825.0	20.0
2100	n1	FDD	1960.0	1974.8	2155.0	2169.8	14.8
2600	n7	FDD	2515.0	2535.0	2635.0	2655.0	20.0
3500	n78	TDD	3710.0	3800.0	3710.0	3800.0	90.0

Figure 22 – Orange Mobile network frequency plan

Analysis of Frequency Data for Orange

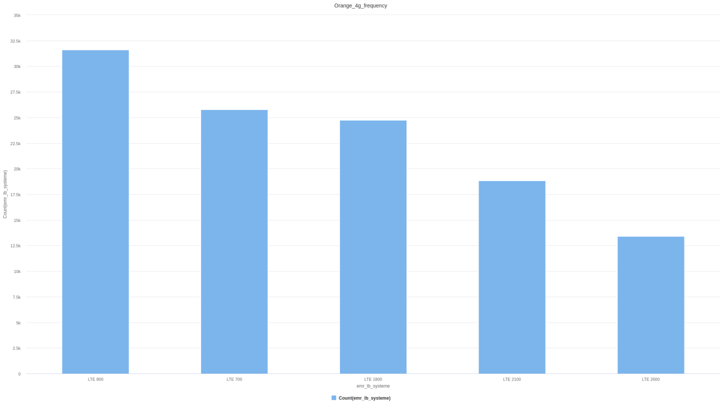


Figure 23 – Count of antennas operating at each frequency band

Analysis of Frequency Data

Characteristics of each band

- 700 MHz and 800 MHz: Long-range, good building penetration.
- 1800 MHz and 2100 MHz: Balanced range and capacity.
- 2600 MHz: Short-range, high capacity.

Methodology for Coverage Calculation

To estimate the coverage zones of mobile base stations using only frequency data, we can implement the following approach.
We can apply standard radio propagation models such as the Hata Model for urban areas and the Cost-231 for suburban and rural areas. These models allow us to estimate coverage zones based on the operating frequency of each base station. Despite lacking specific data on transmission power, antenna gain, path loss, and receiver sensitivity, these propagation models provide a generalized prediction of signal coverage areas by leveraging established empirical data and frequency characteristics. This methodology can enable us to approximate the coverage zones for each frequency band.

LTE RF Link Budgeting Model

- The link budget calculations are needed for the calculation of cellular coverage parameters.
- A link budget takes into consideration all the losses and gains involved in equipment (e.g., BTS), end-terminals (e.g., mobile devices), and communication medium (e.g., free space).
- The results obtained indicate the maximum allowable propagation loss (MAPL). The formulae used in the link budget calculations are expressed below:

Link budget calculations

$$EIRP_{Tx} = P_{Tx} + G_{Tx} - L_b \tag{1}$$

$$EIRP_{Tx} = NB_{noise} + Th_{noise} + SINR \tag{2}$$

$$MAPL = EIRP_{Tx} + R_{SENS} - IM - L_{cable} + G_{Rx} - M + G_{soft} \tag{3}$$

Note: We do not have exact data for our base stations, but we can make reasonable assumptions based on industry standards.

Reference Research Findings

We have referenced research that provides input parameters for radio propagation models and cell-range calculation results.

Input parameters for radio propagation model

- Carrier Frequencys (MHz): 700, 800, 1800, and 2100 MHz.
- Base Station Height (hb): Typically around 30 meters.
- Mobile Station Height (hm): Generally around 1.5 meters.

Frequency	**Urban (km)**	**Sub-urban (km)**	**Rural (km)**
700 MHz	1.25	2.51	4.31
800 MHz	1.11	2.12	3.81
1800 MHz	0.707	1.16	2.49
2100 MHz	0.625	0.99	2.24

Figure 24 – Cell-range calculation results

Frequency Reuse Mode Challenges

Challenges with Unknown Frequency Reuse Mode

- Each cell in a mobile network can be divided into sectors, each served by a different frequency.
- The reuse of frequencies in different cells is essential to maximize the efficient use of the available spectrum.
- Frequency reuse mode determines how frequencies are allocated across cells and sectors to minimize interference.
- We lack detailed information on the specific frequency reuse mode utilized for each sector of our base stations.
- Assumptions about uniform frequency reuse may not reflect the actual, varied configurations used in practice.

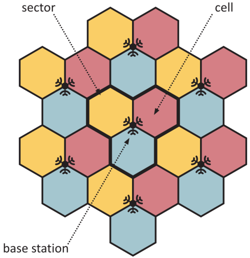


Figure 25 – Sectorization