Wednesday July 19th 2023,
Spatial Statistics 2023
Boulder, Colorado

What geostatistical model for uncertainly geolocated Energy Performance Certificates (EPC)?

Marc Grossouvre[1], supervised by Didier Rullière[2] and Jonathan Villot[3]

[1]U.R.B.S. SAS, marcgrossouvre@urbs.fr
[2]Mines Saint-Etienne - LIMOS - Univ Clermont Auvergne
[3]Mines Saint-Etienne - U.R.B.S. SAS

What geostatistical model for uncertainly geolocated Energy Performance Certificates (EPC)?

1. Uncertain label, uncertain geolocation

2. Integrate these uncertainties into a geostatistical model

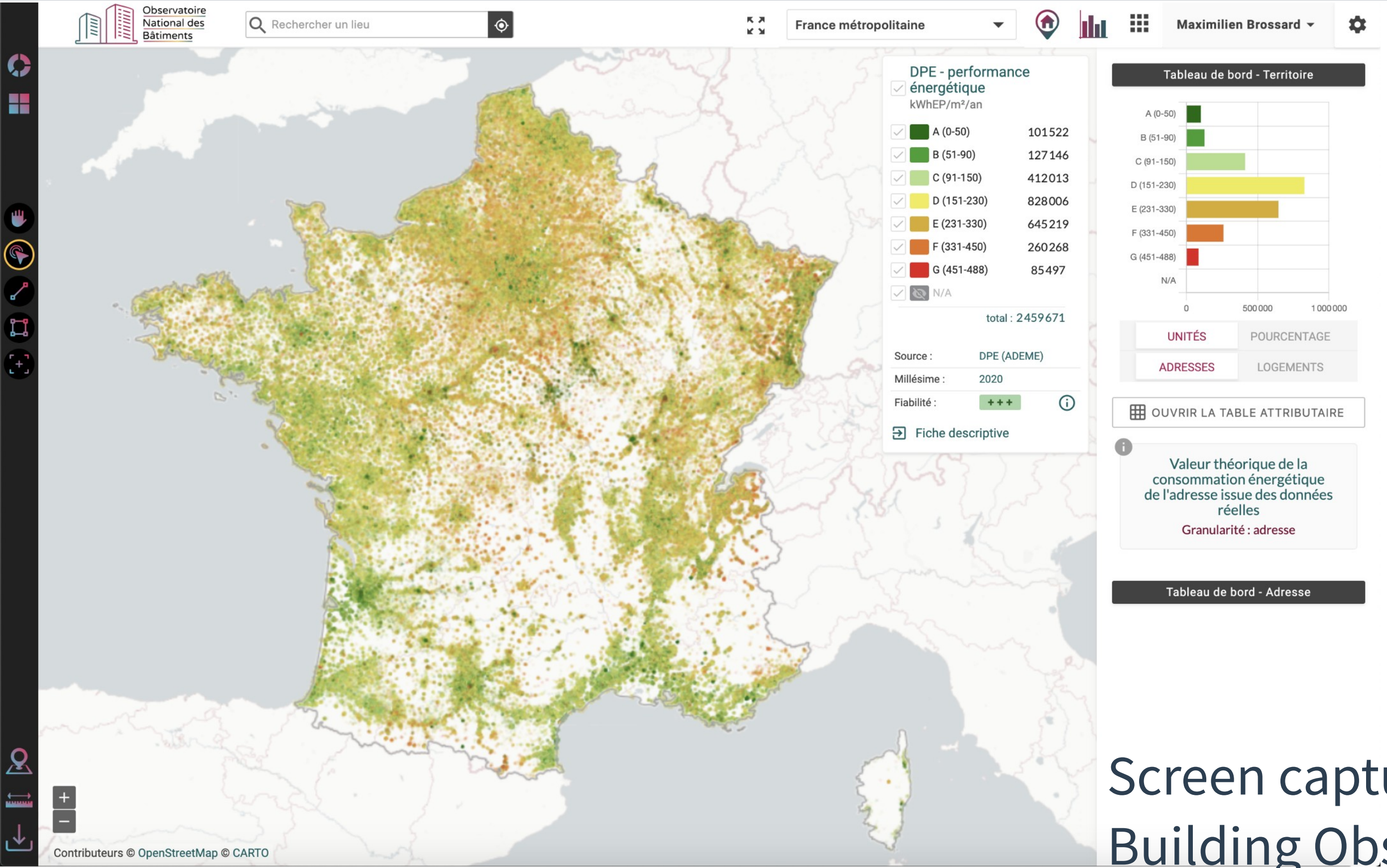
3. Data processing and model implementation in R

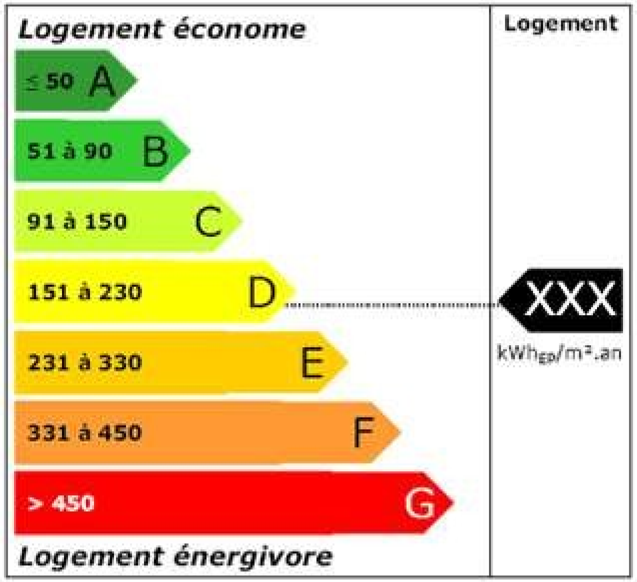
What geostatistical model for uncertainly geolocated Energy Performance Certificates (EPC)?

ADEME
AGENCE DE LA TRANSITION ÉCOLOGIQUE

1. Uncertain label, uncertain geolocation
2. Integrate these uncertainties into a geostatistical model
3. Data processing and model implementation in R

~2 millions EPCs are collected each year



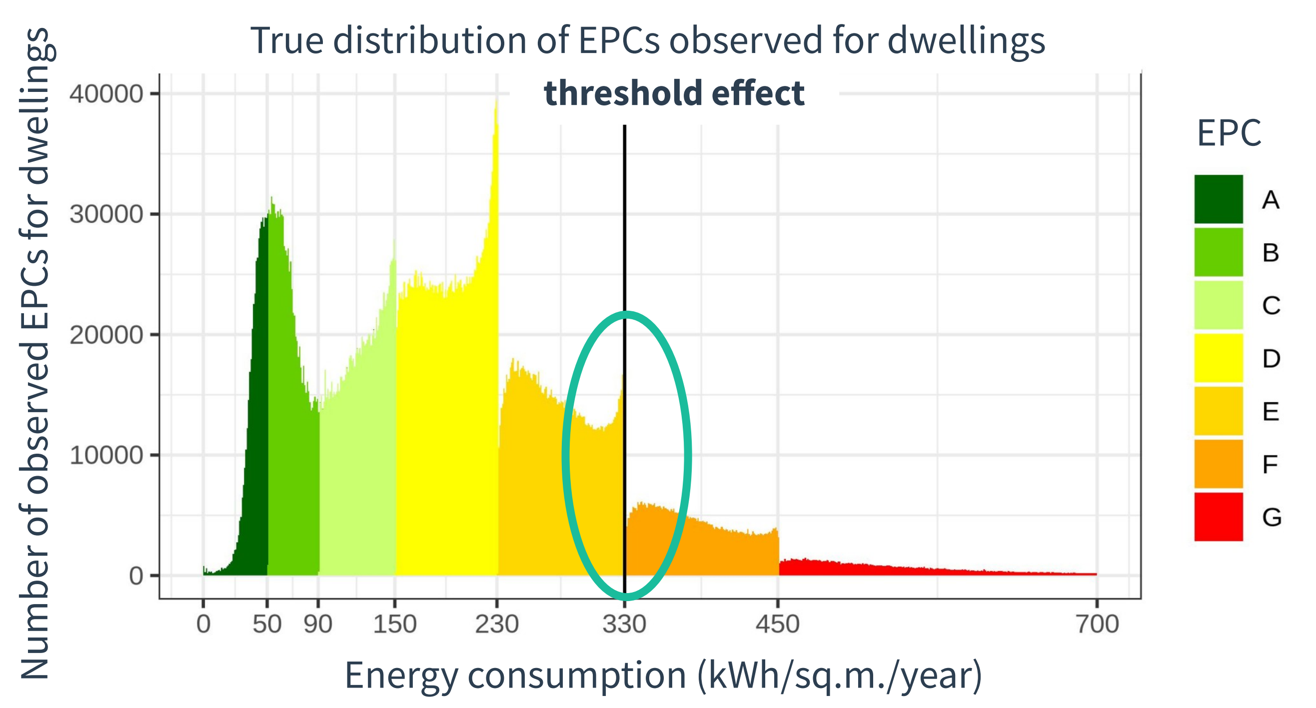


French EPC vignette (until 2021)

Screen capture of National Building Observatory (ONB)

Uncertain label:
human error, threshold, missing values

A threshold is a regulatory value bounding the domain associated with an EPC label.
The threshold effect encompasses phenomena that occur around this value.



Moreover...
EPC changes over time,
EPC of a dwelling ≠ EPC of a building

Uncertain geolocation: address, building, land plot

street_name
28 & 30 Rue Gabriel Vicaire \nRue Prosper Convert

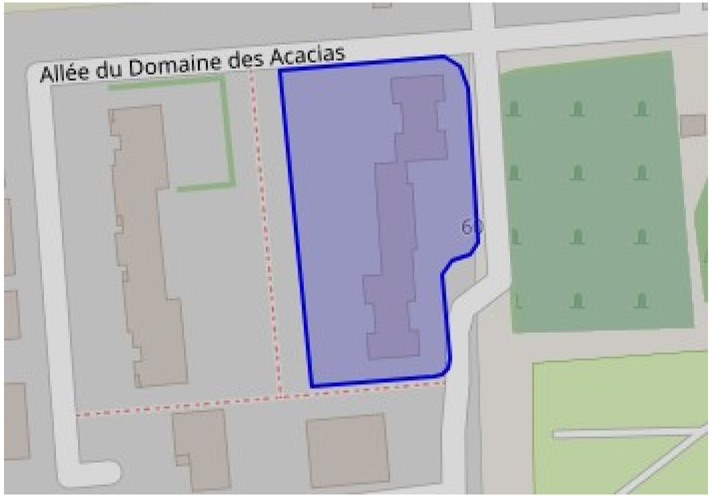
Understand: 28 & 30 rue Gabriel Vicaire,
at the corner of rue Prosper Convert

An address can be associated with a land plot but it is difficult to differentiate addresses on the same land plot

14, 22, 30, 40, 52 impasse des Acacias

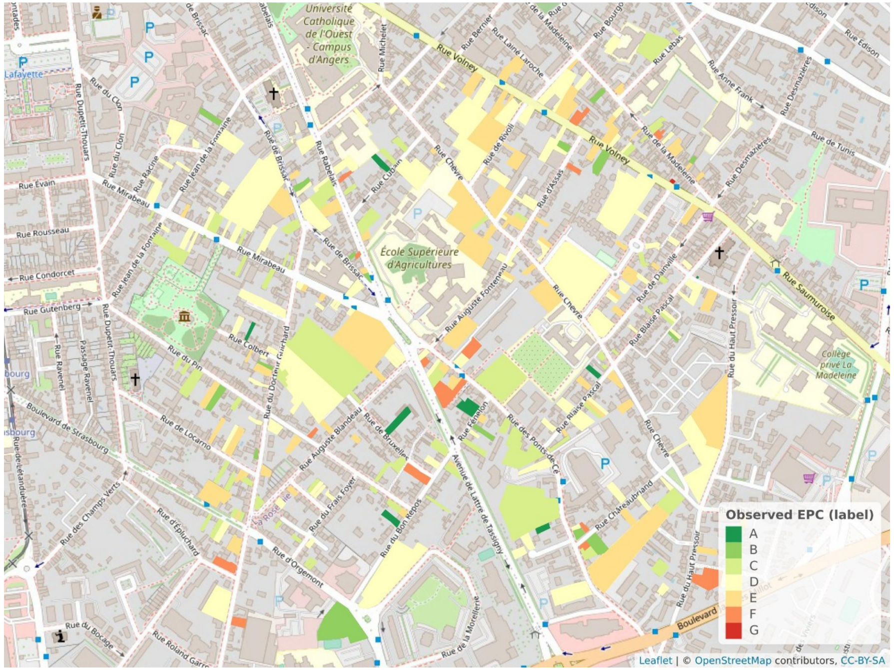
- 5 row houses,
- 1 land plot,

- Observations : 3 labels E, 2 labels D.
- We do not know where are the observed dwellings on the land plot.



A lot of missing values

An urban neighbourhood



- About 15% of addresses have at least 1 observed dwelling
- What can we say about unobserved addresses?
- Can we detect "energy sieves" (labels F and G)?

What geostatistical model for uncertainly geolocated Energy Performance Certificates (EPC)?

1. Uncertain label, uncertain geolocation

2. Integrate these uncertainties into a geostatistical model

3. Data processing and model implementation in R

2. Integrate these uncertainties into a geostatistical model

Issue: How to predict EPCs at the address level without physical inspection?

Hypothesis: EPCs can be modelled and predicted as geospatial data rather than relying on thermal engineering.

Scientific challenge: We need to handle uncertainty of both the positioning of observations and their values.

2. Integrate these uncertainties into a geostatistical model

Issue: How to predict EPCs at the address level without physical inspection?

Hypothesis: EPCs can be modelled and predicted as geospatial data rather than relying on thermal engineering.

Scientific challenge: We need to handle uncertainty of both the positioning of observations and their values.

Observations are <u>mixture distributions</u> of energy consumptions

Each dwelling has a random energy consumption

AND

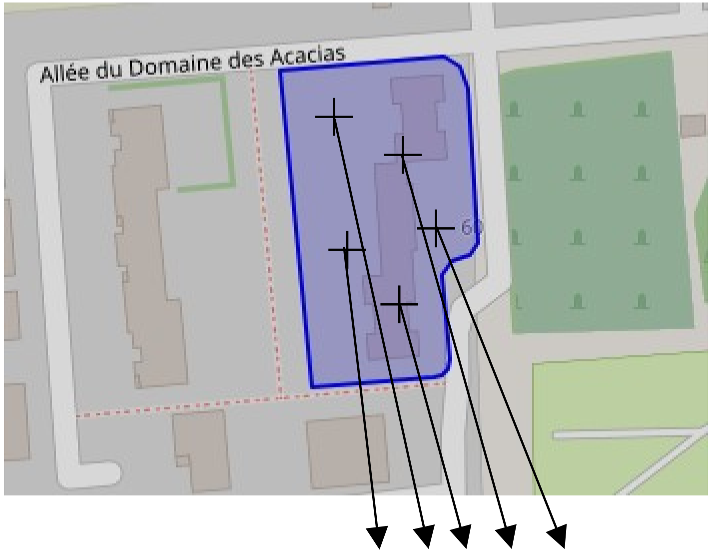
Each observed EPC is associated with a random dwelling among the dwellings on the land plot.

THEREFORE

Each observed EPC is

a <u>random</u> value associated with a <u>random</u> position.

What can we do with that?



Observations E, E, E, D, D
at random positions

Modelling EPCs as a mixture distribution.

Let:

x a point on the territory
$Y(x)$ the energy consumption per square meter of living space at x
g a land plot
X_g a random position on g

Then $Y(g) = Y(X_g)$ is the energy consumption per square meter of living space associated with the land plot g.

Observed consumptions make an observation vector

$$\underline{\mathbf{Y}} = (Y(g_1), ..., Y(g_n))^T.$$

Covariance between 2 mixture distributions

Denote:
$k(x, x') = \mathrm{Cov}\left[Y(x), Y(x')\right]$
$\mu(x) = \mathbb{E}\left[Y(x)\right]$.

$$\mathrm{Cov}\left[Y(g), Y(g')\right] = \mathbb{E}\left[k(X_g, X_{g'})\right] + \mathrm{Cov}\left[\mu(X_g), \mu(X_{g'})\right]$$

$$\mathrm{Var}\left[Y(g)\right] = \mathbb{E}\left[k(X_g, X_g)\right] + \mathrm{Var}\left[\mu(X_g)\right]$$

Notations:
x point
g land plot
$Y(x)$ random energy consumption at x
X_g random position on g
$Y(X_g) = Y(g)$ random energy consumption at g
$\underline{Y}$ observations

Covariance between 2 mixture distributions

Denote:

$k(x, x') = \mathrm{Cov}\left[Y(x), Y(x')\right]$

$\mu(x) = \mathbb{E}\left[Y(x)\right].$

$$\mathrm{Cov}\left[Y(g), Y(g')\right] = \boxed{\mathbb{E}\left[k(X_g, X_{g'})\right]} + \mathrm{Cov}\left[\mu(X_g), \mu(X_{g'})\right]$$

$$\mathrm{Var}\left[Y(g)\right] = \boxed{\mathbb{E}\left[k(X_g, X_g)\right]} + \mathrm{Var}\left[\mu(X_g)\right]$$

Notations:

x point

g land plot

$Y(x)$ random energy consumption at x

X_g random position on g

$Y(X_g) = Y(g)$ random energy consumption at g

$\underline{Y}$ observations

The best linear unbiased predictor (BLUP): Kriging of mixture distributions

For an unobserved plot g, a predictor is: $\hat{Y}(g) = \sum_{i=1}^{n} \alpha_i Y(g_i) = \boldsymbol{\alpha}^T \underline{\mathbf{Y}}$.

Denote:
$\mathbf{K}$ the covariance matrix of $\underline{\mathbf{Y}}$
$\mathbf{h}_g$ the covariance vector between $\underline{\mathbf{Y}}$ et $Y(g)$.

If $\mathbb{E}\left[\underline{\mathbf{Y}}\right] = 0$, the predictor is: $\boldsymbol{\alpha}^* = \mathbf{K}^{-1}\mathbf{h}_g$

Also possible if $\mathbb{E}\left[\underline{\mathbf{Y}}\right] \neq 0$.
The prediction error can be estimated too.

Grossouvre M., Rullière D., Villot J., *Spatial interpolation using mixture distributions: A Best Linear Unbiased Predictor*, 2023, preprint https://hal.science/hal-03276127/

What geostatistical model for uncertainly geolocated Energy Performance Certificates (EPC)?

1. Uncertain label, uncertain geolocation

2. Integrate these uncertainties into a geostatistical model

3. Data processing and model implementation in R

Implementing mixture Kriging raises technical questions (1/2)

$$\mathrm{Cov}\left[Y(g), Y(g')\right] = \mathbb{E}\left[k(X_g, X_{g'})\right] + \mathrm{Cov}\left[\mu(X_g), \mu(X_{g'})\right]$$

If $X_g, X_{g'}$ are discrete uniform:

$$\mathrm{Cov}\left[Y(g), Y(g')\right] = \frac{1}{[g]\,[g']} \sum_{(x,x')\in g\times g'} k(x, x')$$

$\Rightarrow$ Number of point to point covariances to compute $\propto$ (density of points)^2.
+ The covariance kernel has an exponential term which is costly.

$\Rightarrow$ Covariance kernels are implemented in `C++`, `RcppArmadillo`.

Implementing mixture Kriging raises technical questions (1/2)

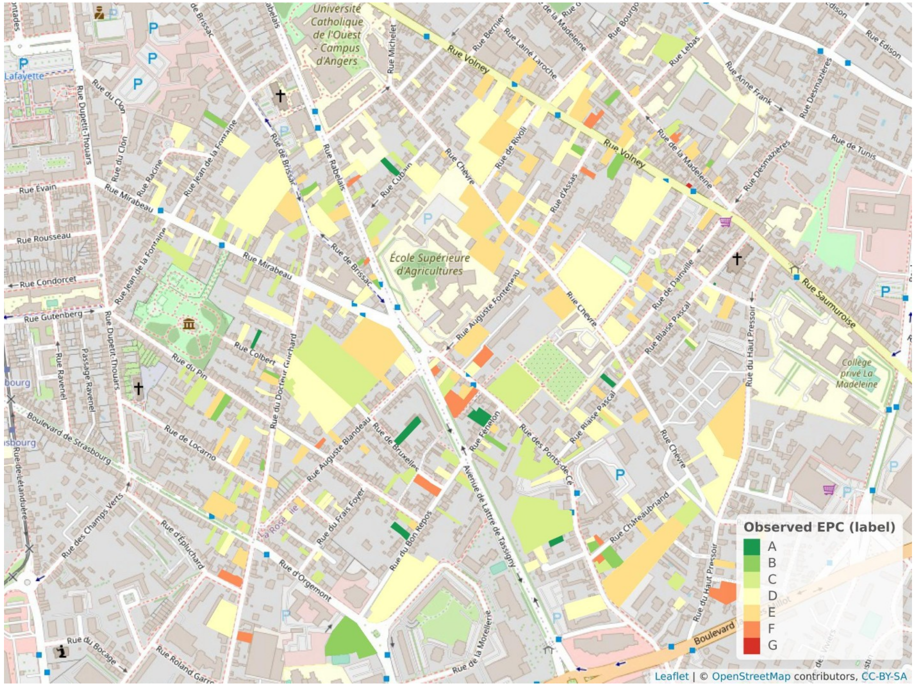
The observations' covariance matrix must be inverted.

$\Longrightarrow$ Difficult to learn from a large number of observations

We are trying to combine a family of sub-models giving the same predictions as a large model using compact support kernels (under research).

Results of a first test on an urban area

Observations

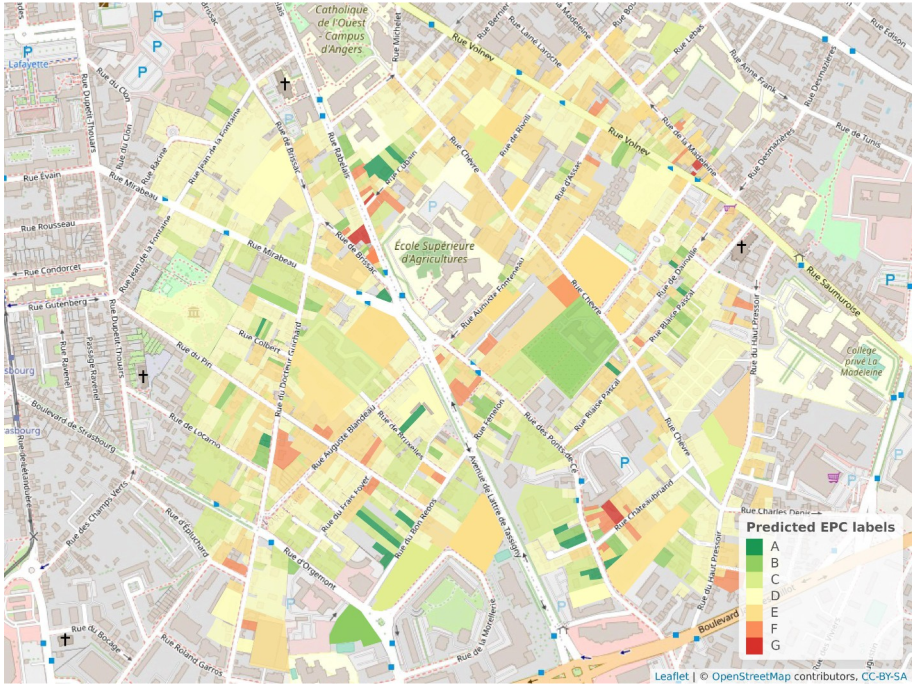


Confusion matrix

True values	Predicted values						
	A	B	C	D	E	F	G
A	2	1	1	2	2	0	0
B	1	3	3	9	2	2	0
C	1	3	3	26	15	4	0
D	3	5	5	80	33	5	1
E	4	2	2	36	36	5	1
F	0	3	3	4	5	3	0
G	0	0	0	1	1	0	0

Results of a first test on an urban area

Predictions



Confusion matrix

True values	Predicted values						
	A	B	C	D	E	F	G
A	2	1	1	2	2	0	0
B	1	3	3	9	2	2	0
C	1	3	3	26	15	4	0
D	3	5	5	80	33	5	1
E	4	2	2	36	36	5	1
F	0	3	3	4	5	3	0
G	0	0	0	1	1	0	0

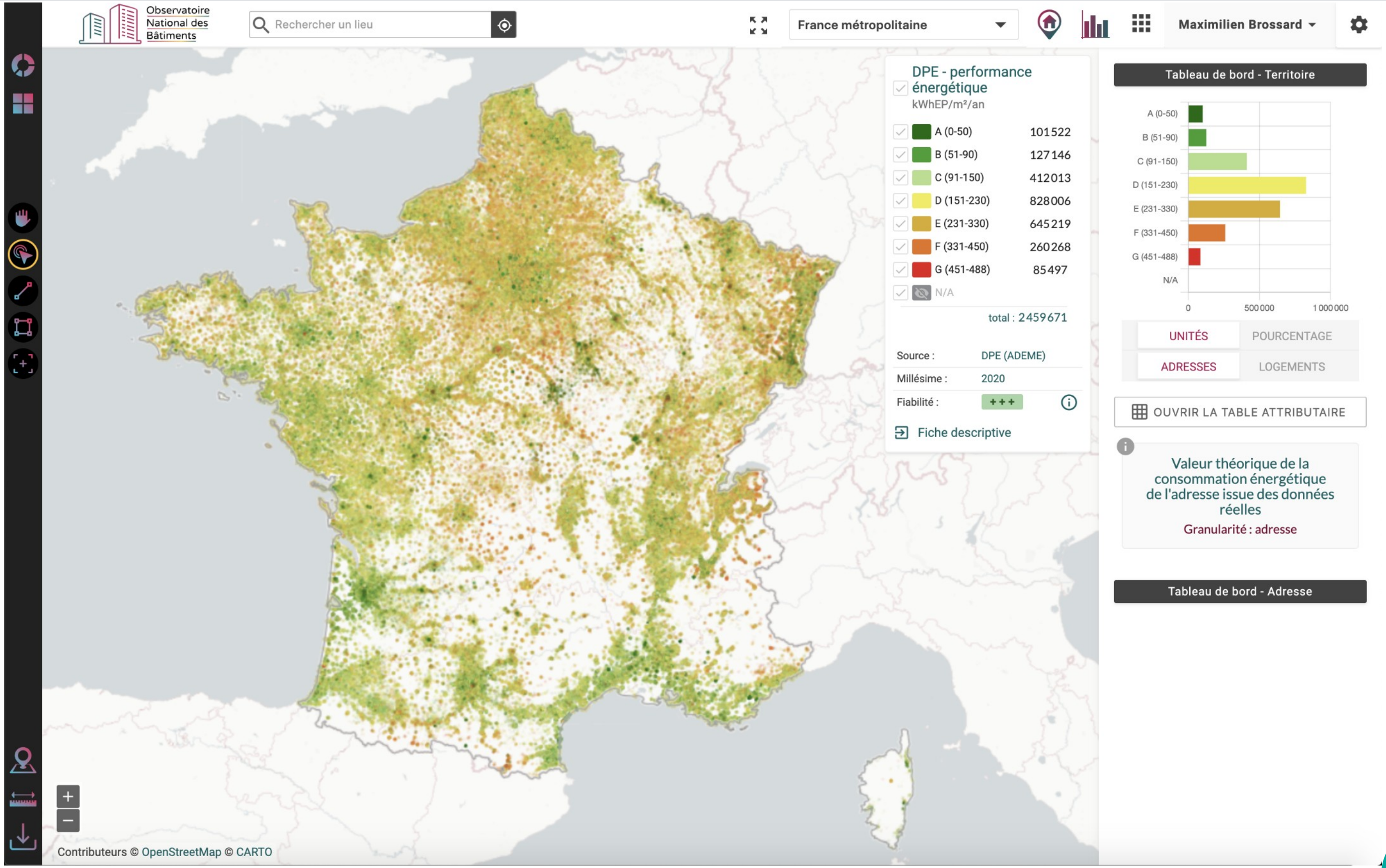
Conclusion
EPC can be regarded as a geolocated data

- It is possible to construct a geostatistical model instead of attempting to estimate physical variables.
- The Mixture Kriging model accommodates both the uncertainties of observed locations and values.
- With 3 variables (latitude, longitude, age), we achieve as good results as KNN with 40 variables.

Thanks for listening. Merci de votre attention.

Access all open data that have been discussed in this presentation with the Observatoire National des Bâtiments

www.imope.fr/onb.html



Annexes

Peut-on “lisser” les effets seuil ? Légitimité et validation ?

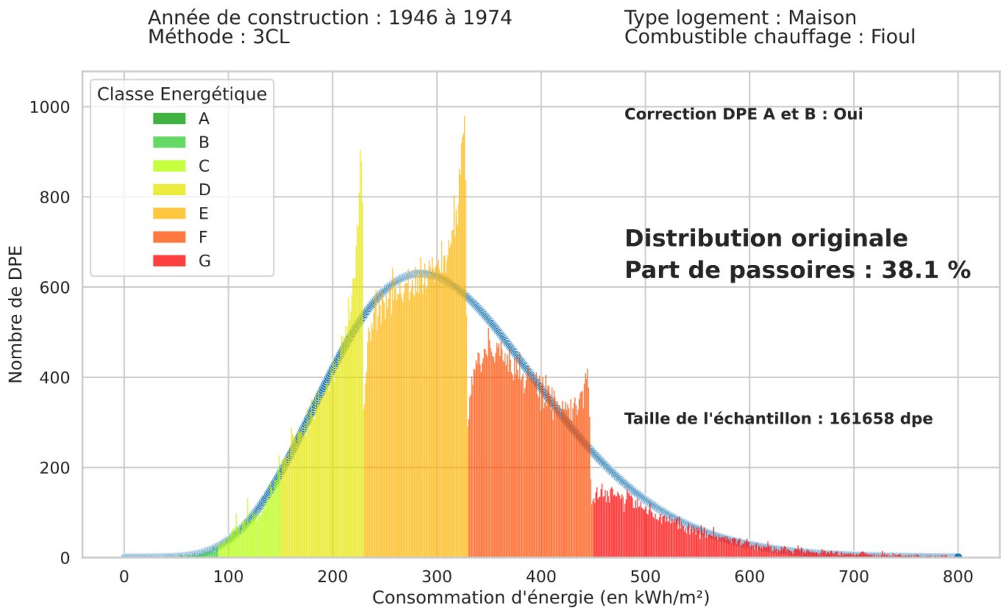
Géolocalisation, qu’en dit la BAN ?

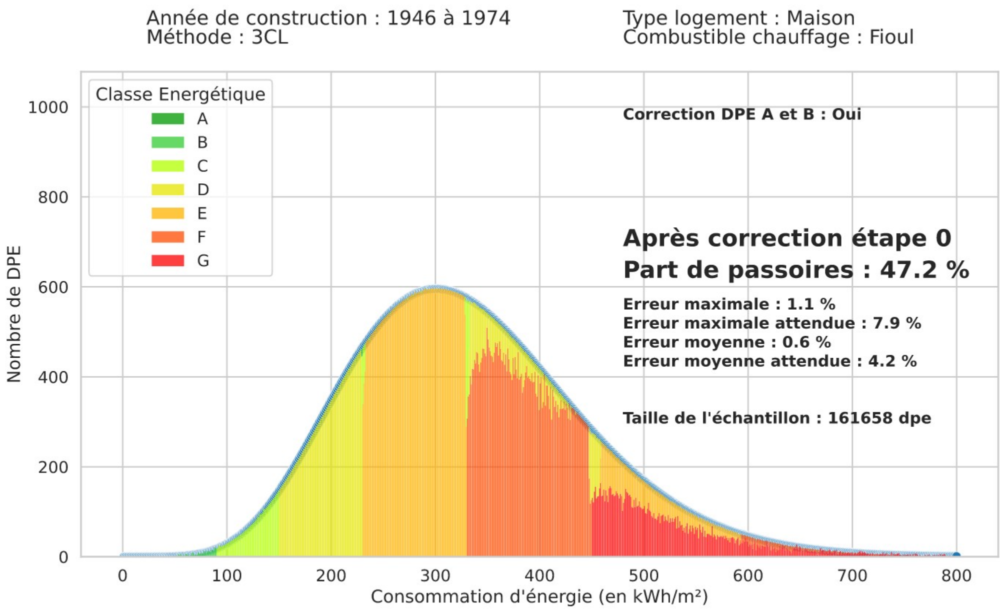
Le process de traitement des données.

Peut-on “lisser” les effets seuil ? Légitimité et validation ?

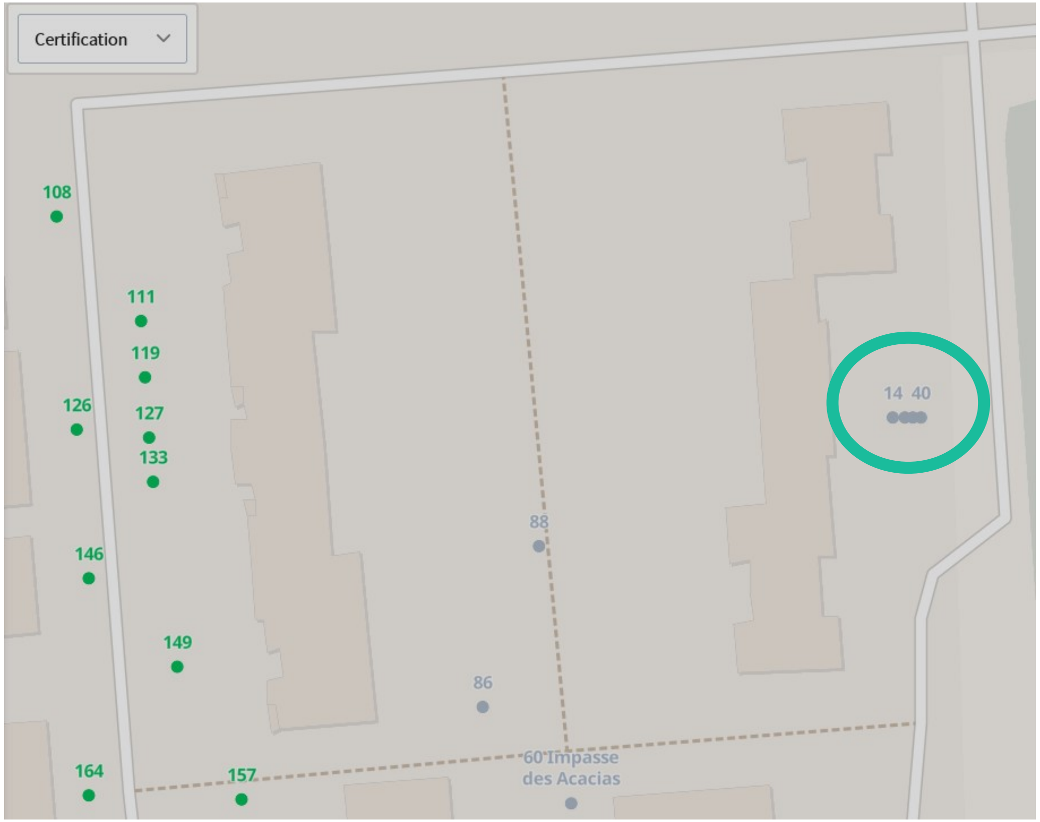
Source : Alternatives énergétiques, Yassine Abdelouadoud

Que mesure-t-on ?





Géolocalisation,
qu'en dit la Base d'Adresses Nationale (BAN) ?



Il faut apparier les DPE aux fichiers fonciers pour les fiabiliser

Les fichiers fonciers (Ministère des finances) permettent de relier le DPE observé à une parcelle.

Le DPE renseigne sur l'un des logements de l'un des bâtiments de la parcelle.