Prédire l'efficacité énergétique des bâtiments

Apprendre de données granulaires

Marc Grossouvre[1] sous la direction de Didier Rullière[2] et la supervision de Jonathan Villot[3]

[1]Doctorant CIFRE, 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

mardi 29 mars 2022, Institut Henri Fayol

MINES Saint-Étienne

U R B S Urban Retrofit Business Services

Le logiciel IMOPE, intégrer les données des logements

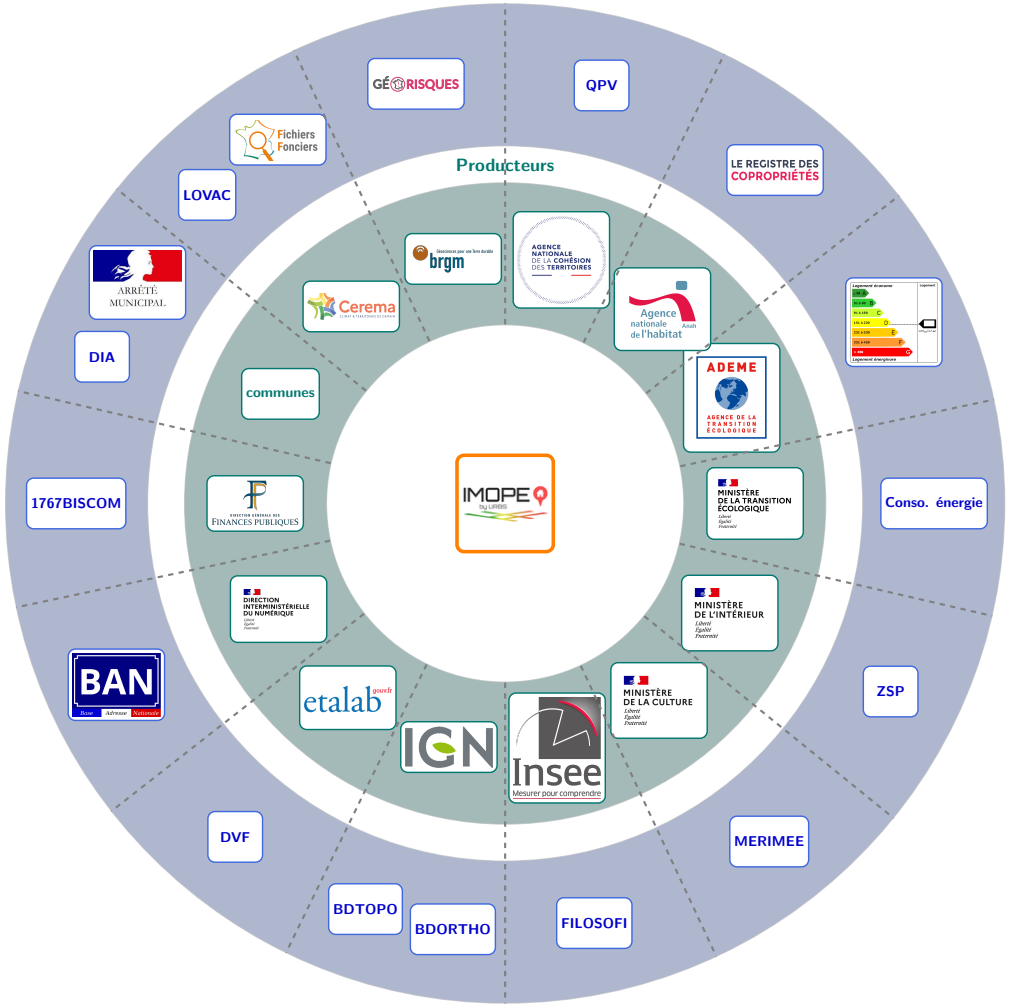
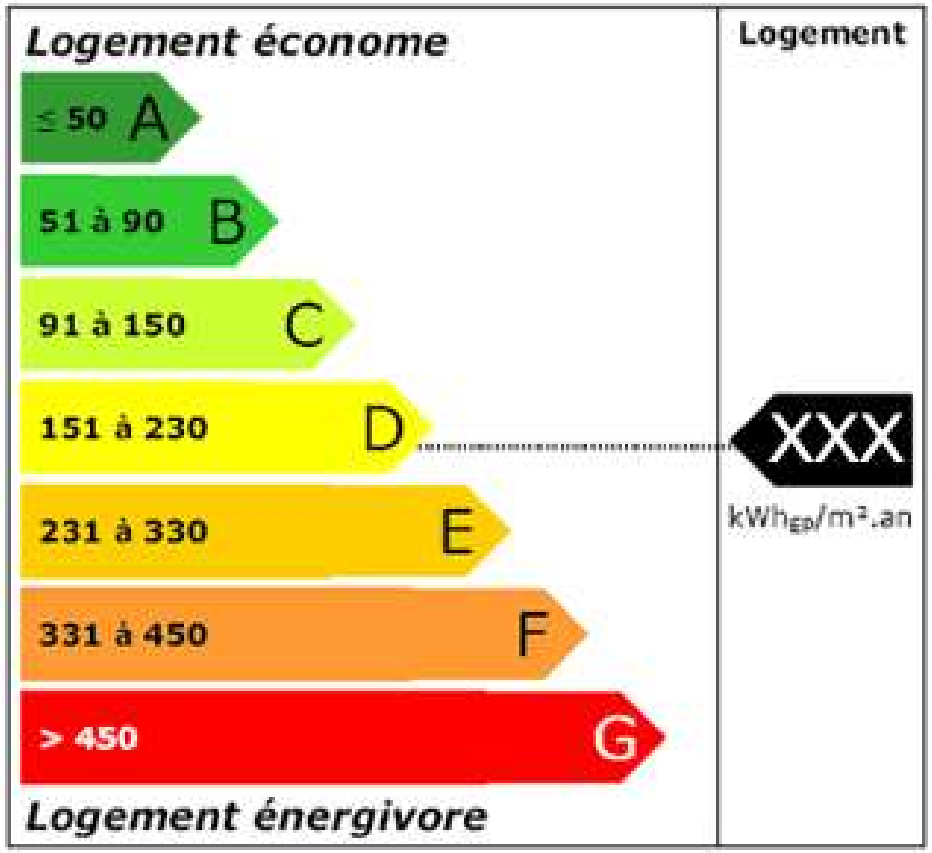


Figure 1: Les bases de données agrégées dans le logiciel IMOPE et leurs producteurs. Source U.R.B.S.

Le DPE, diagnostic de performance énergétique



Le logiciel IMOPE, visualiser les données à l'adresse

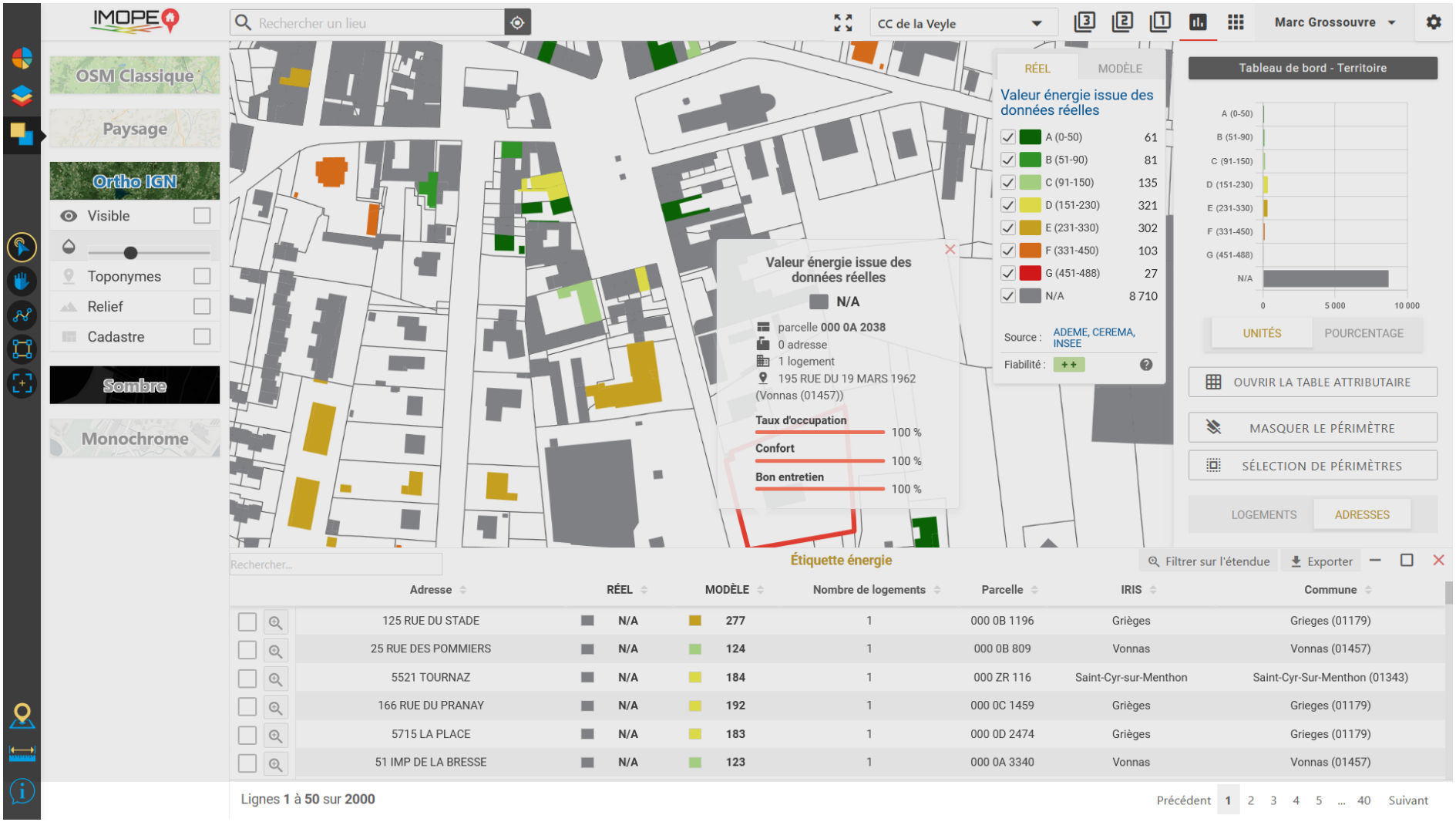


Figure 2: Impression d'écran du logiciel IMOPE.

Le problème des données incomplètes

On peut faire des statistiques globales avec des données incomplètes : "Combien a-t-on de passoires énergétiques en France ?"

Pour répondre à cette question, un échantillon peut suffire.

Mais on ne peut pas filtrer sur des données incomplètes : "Identifier toutes les passoires énergétiques pour proposer aux propriétaires une aide à la rénovation."

Répondre à cette question nécessite de compléter les données.

Des données granulaires pour prédire la performance énergétiques

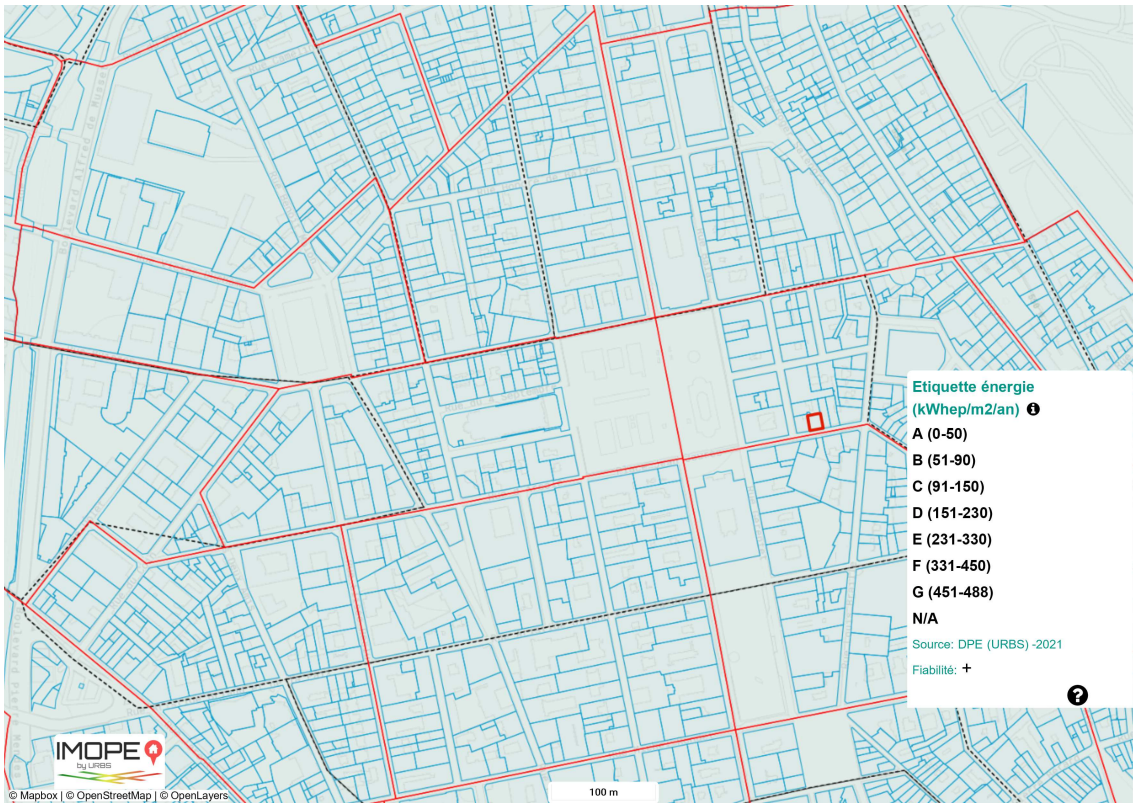
- Diagnostics de performance énergétique (DPE) : à l'adresse non géolocalisé
- Année de construction : partie de logement, avec adresse, géolocalisé sur une parcelle
- Revenus du foyer : au logement anonymisé géolocalisé à l'iris

En savoir plus...

Comment utiliser toutes les données disponibles dans un modèle de machine learning ?

Regarder les données comme des distributions de mélange

- On observe le DPE de mètres carrés d'habitation aléatoires parmi les mètres carrés d'habitation de l'adresse.
- On observe le revenu d'un logement aléatoire parmi les logements de l'iris.



Formalisme des distributions de mélange

A chaque **point** x d'un **territoire** χ, est associé une variable de sortie $Y(x)$. Les points de χ sont groupés en **grains**. On définit sur les points d'un grain g une variable de position aléatoire X_g. La variable de sortie associée au grain g est $Y(X_g)$.
Vocabulaire : on utilise indistinctement **distribution de mélange** ou **mixture**.

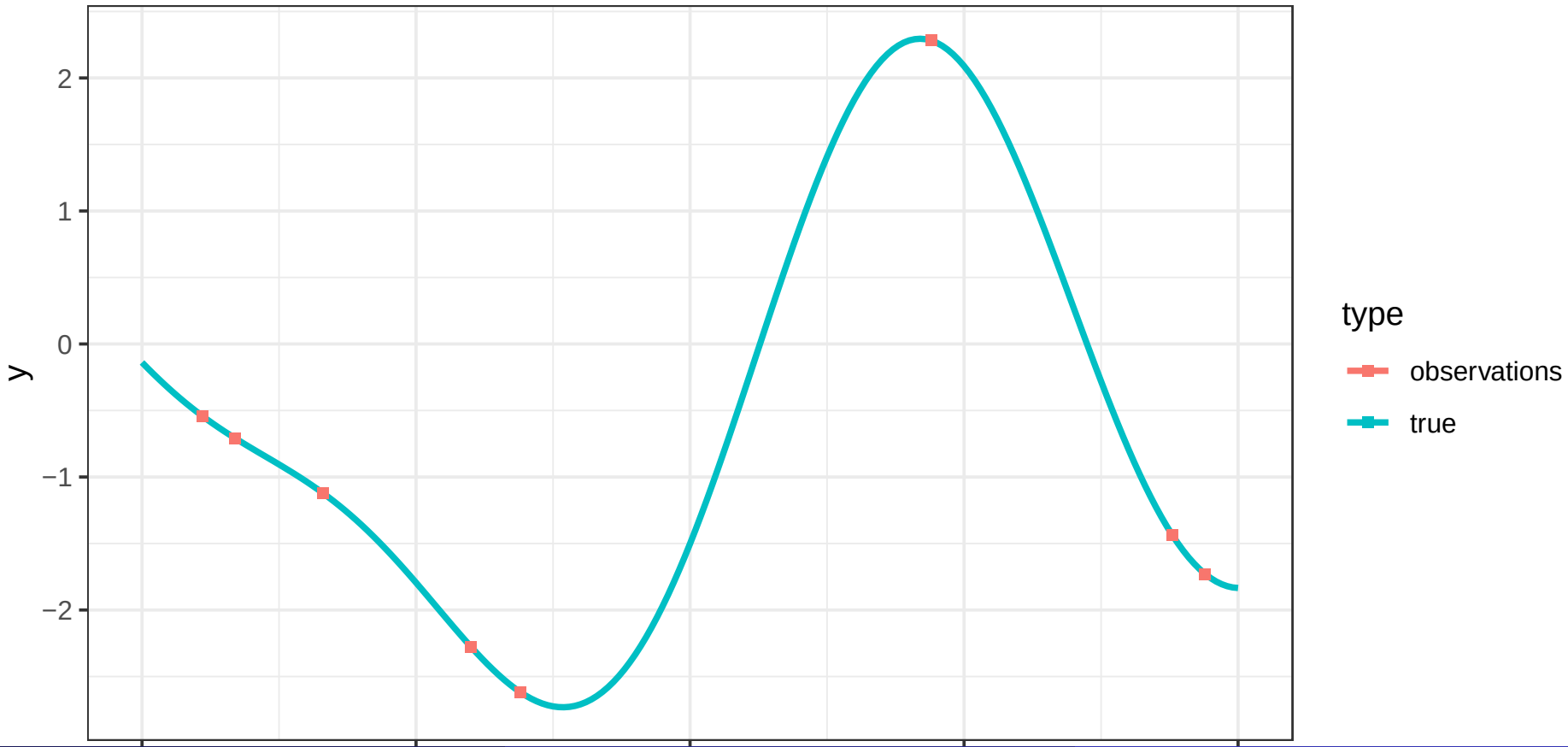
Méthodologie d'apprentissage

Un champ aléatoire gaussien est simulé.

Le noyau de covariance est $k_{\sigma,\theta}(x, x') = \sigma \exp(-\frac{(x-x')^2}{2\theta^2})$.

Observations : 6 vraies valeurs du champ.

À l'aide de ces 6 valeurs seulement, on entraîne un modèle pour tenter de retrouver les paramètres du champ réel.



Le krigeage simple

Soit un champ aléatoire d'espérance nulle Y défini sur tous les points d'un territoire χ. On connaît une fonction appelée un noyau de covariance k tel que $k(x, x') = \mathrm{Cov}\left[Y(x), Y(x')\right]$. On observe Y en n points de χ :

$$\underline{\mathrm{Y}} = (Y^1, ..., Y^n)^\top = (Y(x_1), ..., Y(x_n))^\top \ .$$

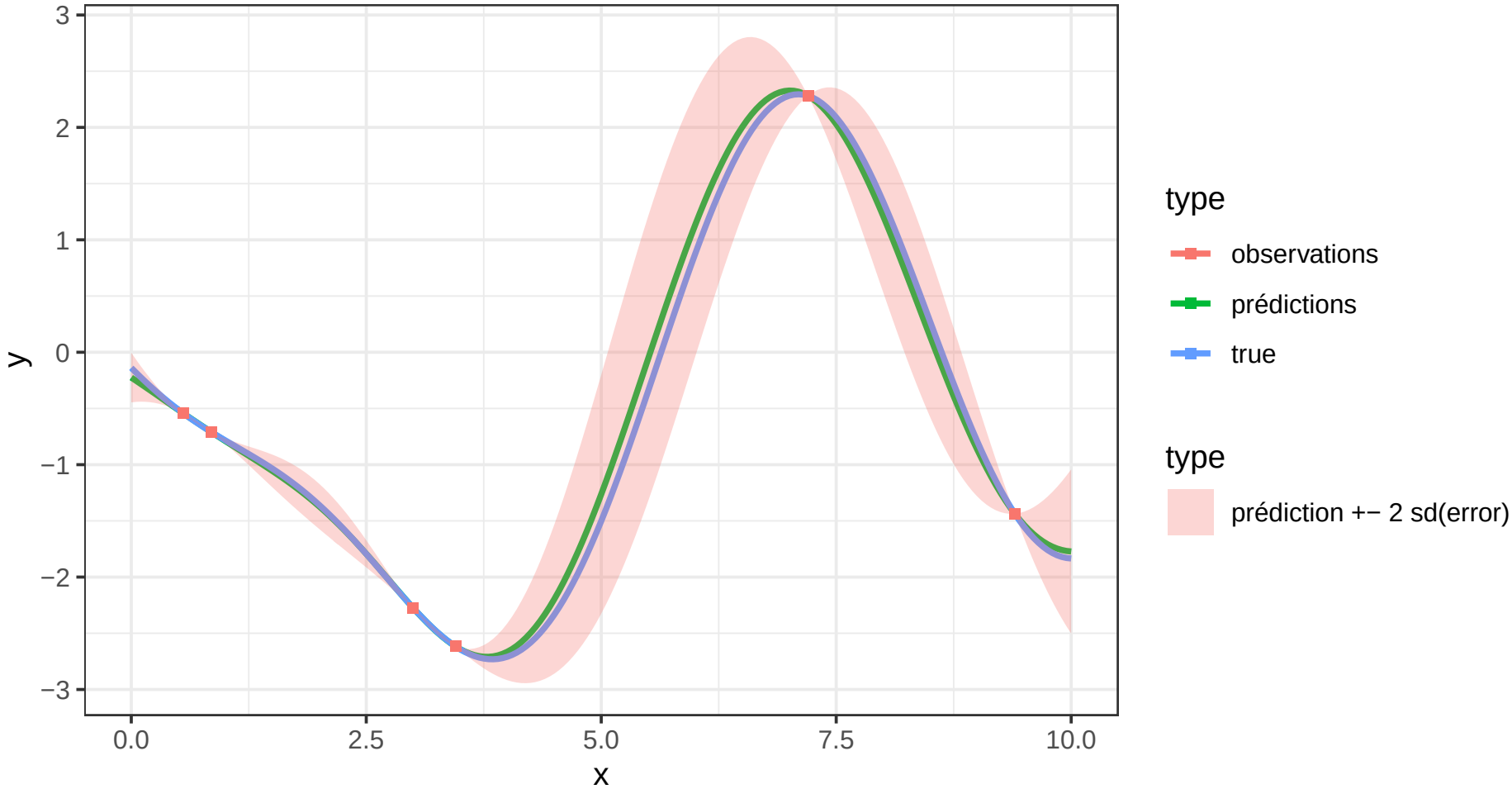
Pour $x \in \chi$, on veut prédire $Y(x)$ comme une combinaison linéaire des observations.
Soit $\mathbf{\Sigma}$ la matrice de covariance des observations : $\mathbf{\Sigma}_{i,j} = k(x_i, x_j)$.
Soit K_x le vecteur tel que $\mathrm{K}_{x,i} = k(x, x_i)$.
Alors le meilleur (qui minimise le rmse) prédicteur linéaire sans biais est :

$$\hat{Y}(x) = \mathbf{\Sigma}^{-1}\mathrm{K}_x\underline{\mathrm{Y}} \ .$$

De plus, on a une estimation de l'erreur. Et il existe une formule similaire dans le cas où le champ n'est pas d'espérance nulle (krigeage ordinaire).

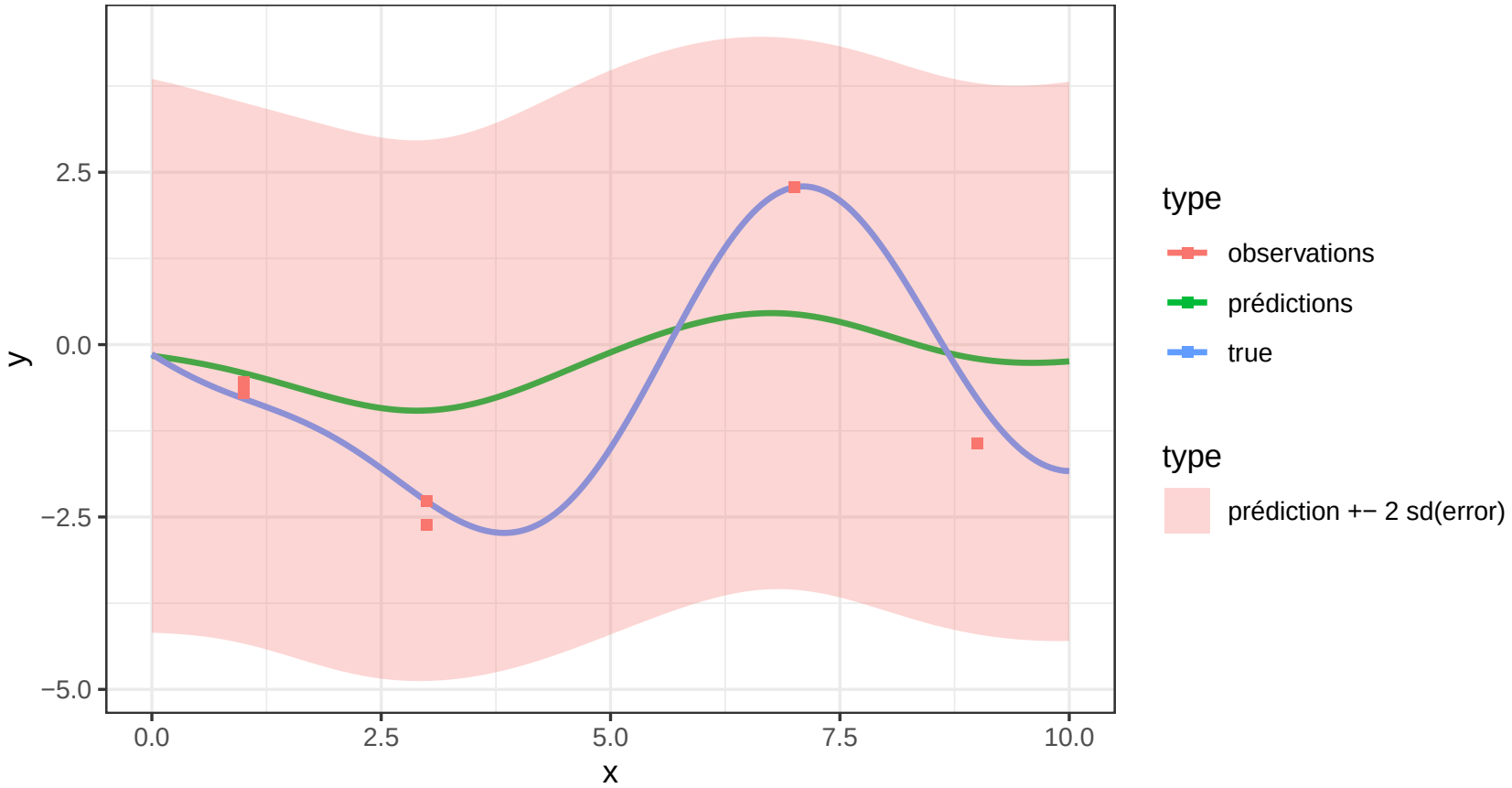
Observations non bruitées

Le krigeage est interpolant. C'est merveilleux



Observations avec bruit sur x

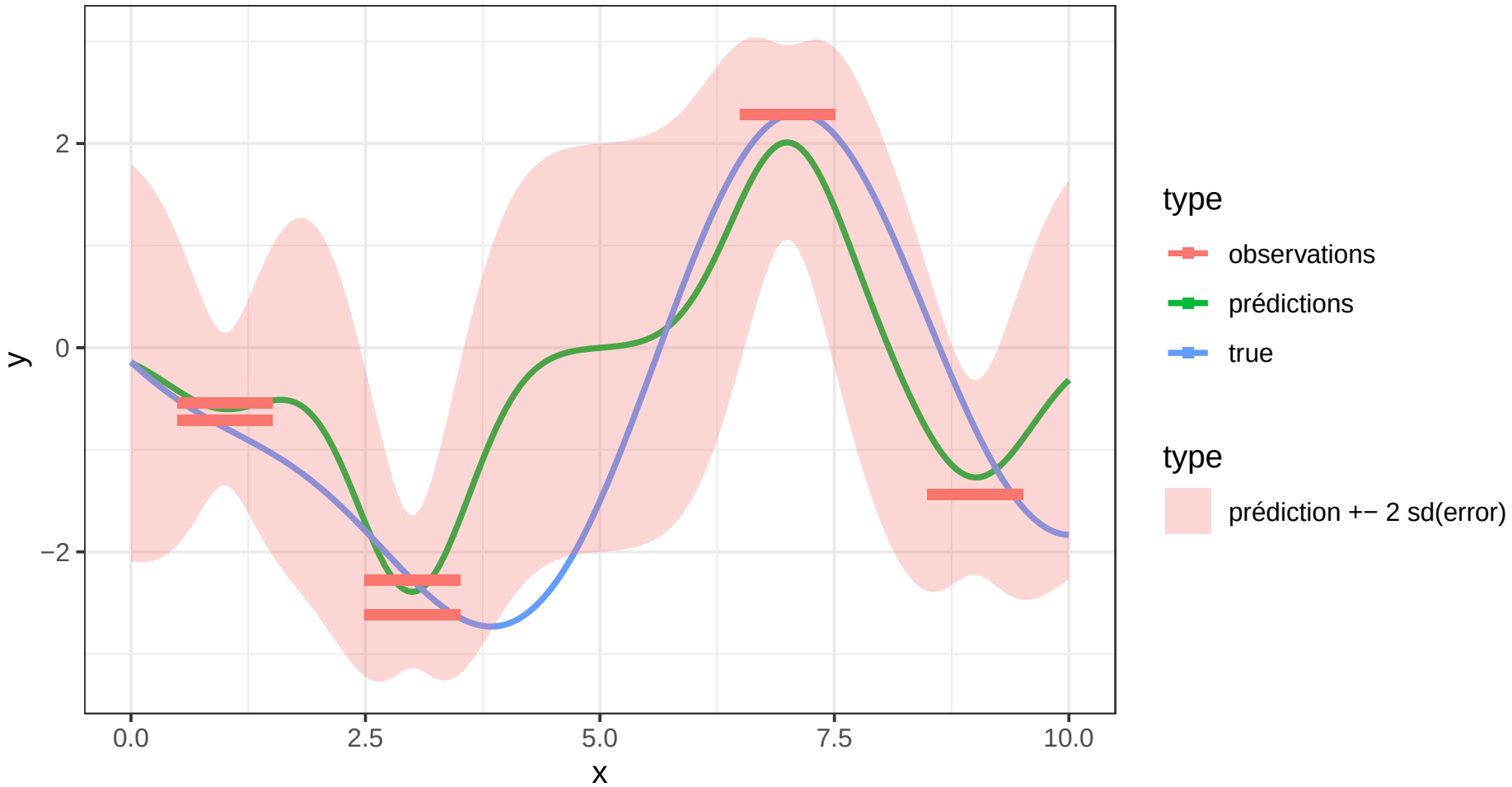
x subit un bruit ϵ : "arrondir x à l'unité". Les valeurs de sortie observées sont $Y(x+\epsilon)$. Pour prendre en compte ce bruit dans le krigeage, les logiciels usuels ajoutent au modèle un bruit sur Y : on observe $Y(x)+\epsilon'$ (effet pépite).



Au cours de l'optimisation, le modèle surestime le bruit.

Observations granulaires

On observe $Y(g)$ sur un grain. Inutile d'ajouter un effet pépite.



Le krigeage de mixtures n'est pas interpolant mais le bruit est controlé.

Mélange à moitié plein, à moitié vide

Bonnes nouvelles sur le krigeage de mixtures (voir les détails dans Grossouvre and Rullière [2021]) :

- il modélise l'incertitude que l'on a sur les variables d'entrée ;
- les prédictions ont d'avantage d'amplitude ;
- il est le meilleur prédicteur linéaire non biaisé ;
- on peut estimer la variance de l'erreur de prédiction.

Mauvaise nouvelle :

- les mélanges de variables gaussiennes ne sont pas gaussiens : l'interprétation du krigeage en termes gaussiens est perdue ;
- on a perdu l'interpolation.

Perpsectives

- Est-que les données INSEE peuvent améliorer la prédiction du DPE ?
- Les distributions de mélange modélisent bien les bruits sur les variables d'entrée, les données anonymisées...
- Peut-on élargir l'approche développée sur le krigeage des mixtures à l'intégration des mixtures dans d'autres algorithmes de machine learning ?

Pour aller plus loin

Marc Grossouvre and Didier Rullière. Mixture Kriging on granular data, July 2021. URL https://hal.archives-ouvertes.fr/hal-03276127.
Voir aussi le site www.imope.urbs.fr.

Merci à Didier Rullière et Jonathan Villot pour leur temps et leur bienveillance.
Merci à Benoît Génot (U.R.B.S.) pour le schéma présenté en Figure 1.

Merci de votre attention !

Modélisation des données