

**100% ACTUAIRES &
100% DATA SCIENCE**

INSTITUT DES
ACTUAIRES



29 / NOV / 2019

Hôtel Marriott Rive Gauche
Paris 14ème

Zonier avec Open Data et Machine Learning

Montassar BEN LAIBA, Nabil RACHDI – ADDACTIS France

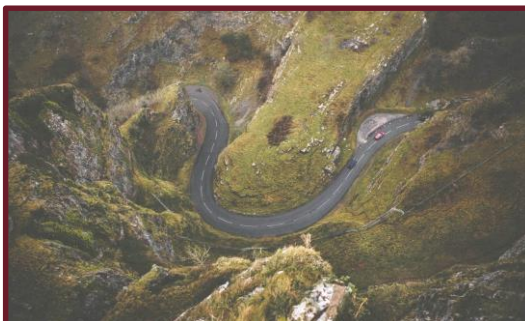
Olivier VERMASSEN, Joanna CHARDON – AXA France

SOMMAIRE

- 01 Contexte
- 02 Geo-Smoothing
- 03 Méthodologie
- 04 Données Externes et Modélisation
- 05 Résultats
- 06 Conclusion & Perspectives

Contexte

- Le zonier est l'une des variables les plus explicatives lors de la modélisation de la prime pure automobile. C'est pourquoi il s'agit d'un **enjeu majeur** pour de nombreux acteurs du marché.
- La méthodologie de la construction du zonier progresse depuis des années. La robustesse, la performance ainsi que la stabilité de ces modèles sont des enjeux considérables pour les acteurs du marché.
- L'apport des données externes permettra sans aucun doute d'apporter encore plus de **performance** et de **stabilité** dans les modèles.





Le Zonier Automobile

POURQUOI UN ZONIER ?



MIEUX MESURER LA **SINISTRALITÉ**



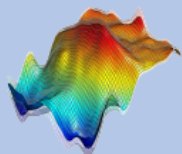
ESSOR DE LA **CONCURRENCE** ET PROGRESSION DU DIRECT



DES ASSURÉS EXIGEANTS ET INFORMÉS DES **TARIFS**

Segmenter plus précisément les facteurs de risques est donc **indispensable**, et cette tendance inclut l'étude du **risque géographique**.

Geo-smoothing - methodology



True function we want to obtain



Methodology

- Offset Emblem predictions
- Smooth the residuals by longitude/latitude
- Output are corrective relativities per ZIP code



Requirements

- Intuitive method
- Fast and automated
- Statistically sound method

OPTIONAL

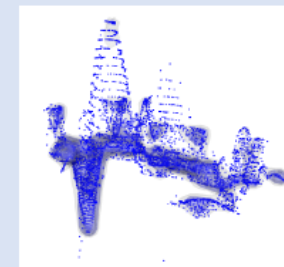
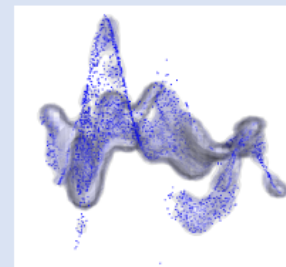


Geo-smoothing



Factor

Several ways to bin relativities into a factor



Results

- We obtained good results ...
- Yet no method consistently outperforms the other
- Necessary to experiment with different methods
- Easy since everything is fast + open-source

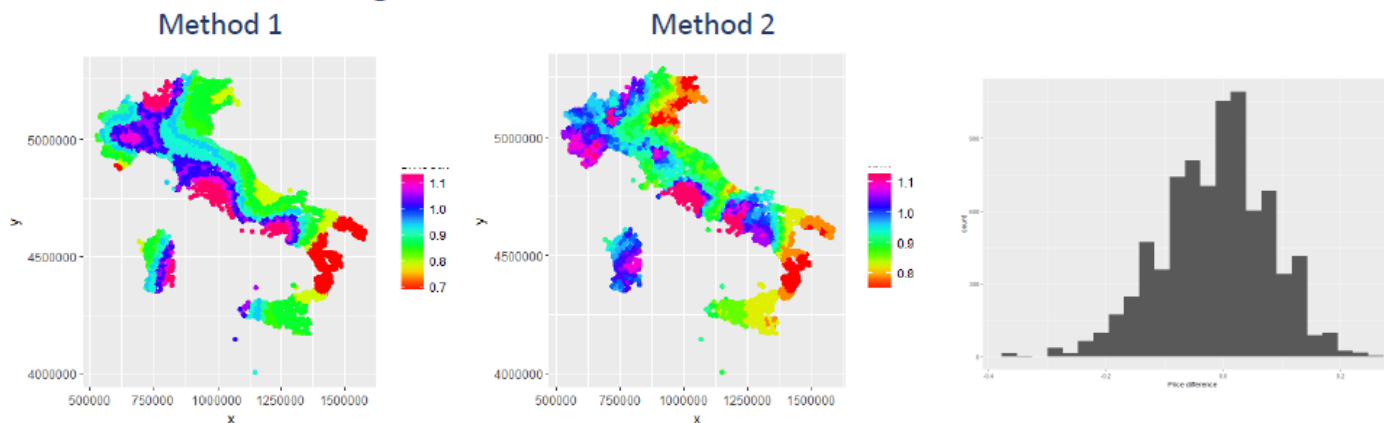


Deployment

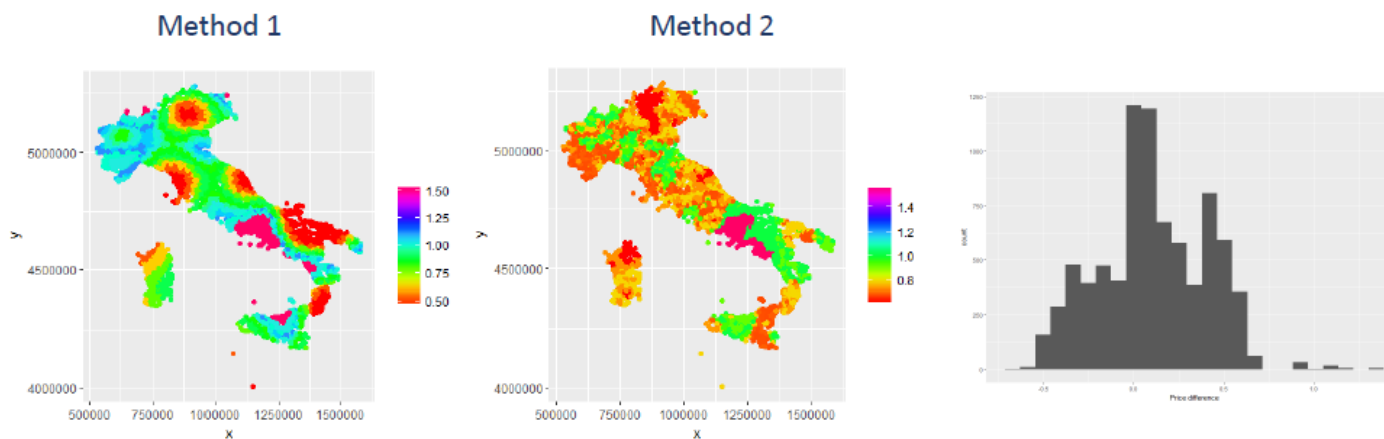
- Statistical fit shows classification performance on ZIP codes observed within the portfolio
- Also necessary to look at classification of ZIP codes not observed in order to prevent UW leakage

Comparison of two methods

Large dataset



Small dataset



Geo-smoothing Tool



Certains portefeuilles comportent des spécificités qui peuvent limiter les méthodes classiques de construction d'un Zonier



MANQUE D'EXPOSITION DANS DE NOMBREUSES REGIONS



MANQUE DE PRÉCISION



FORTE DÉPENDANCE AU PORTEFEUILLE D'ASSURÉS

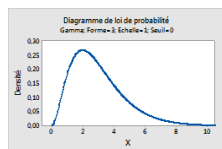


PAS D'INFORMATION GÉOGRAPHIQUES EN INTERNE

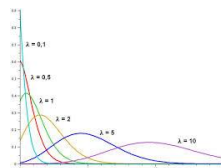
Extension de la méthodologie classique



GLM
DONNEES INTERNES
sans risques externes



Coût moyen



Fréquence

LISSAGE
GEOSPATIAL
et
CREDIBILITE
des résidus

Crédibilité

L'impact de cette fonction croît à mesure que l'exposition de la zone en question augmente. Elle permet d'affecter une influence plus importante aux zones fortement exposées.

& Distance

L'impact de cette fonction décroît à mesure que la distance entre deux zones augmente. Elle permet d'affecter une influence plus importante aux zones voisines.

+

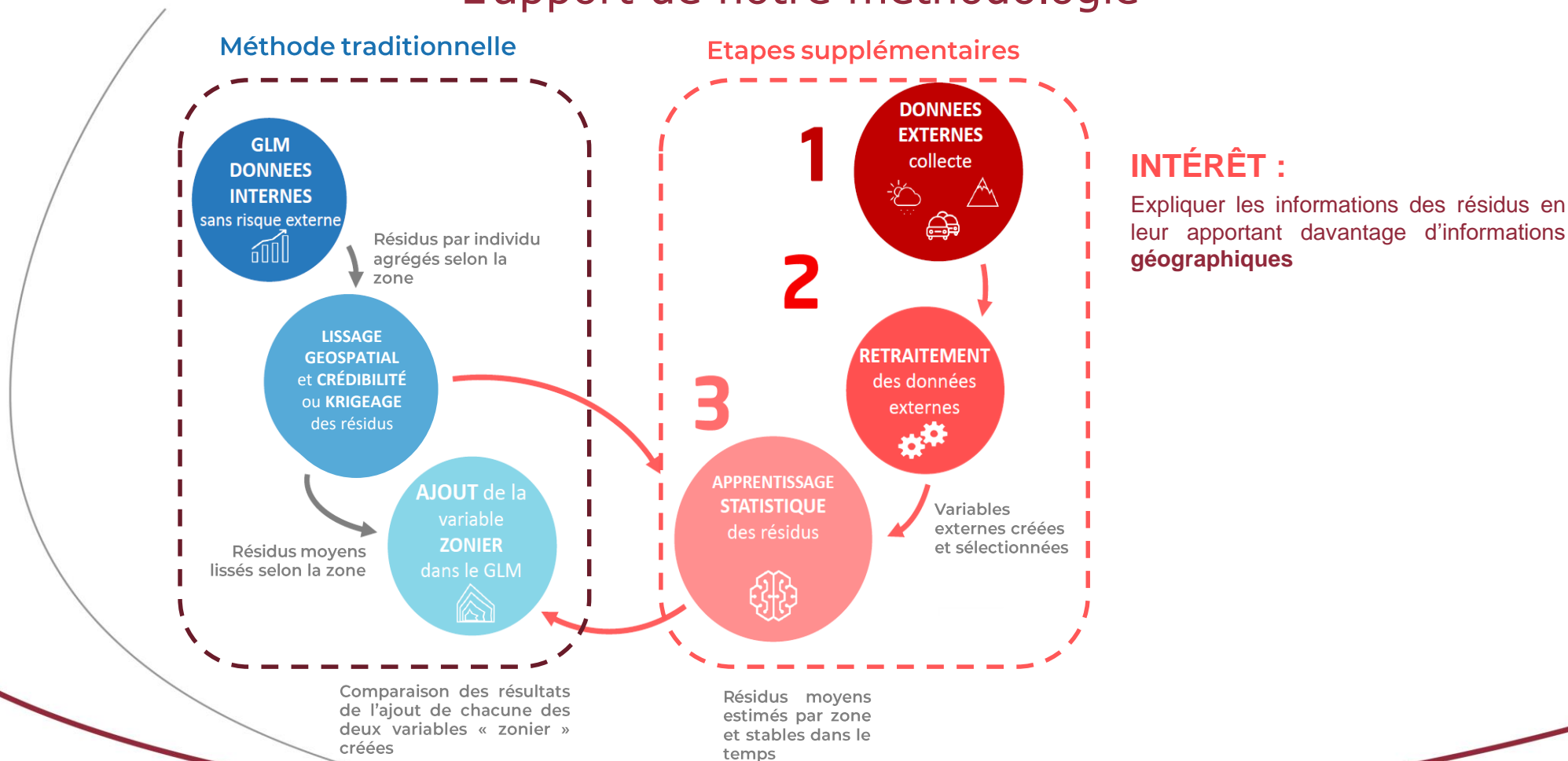
AJOUT de la
variable ZONIER
dans le GLM



Contrôler que la variable zonier n'est **pas corrélée** avec les autres variables incluses dans le modèle.

Vérifier la **significativité** du zonier à l'aide d'une analyse de type I et III
Test de significativité (Khi², AIC).

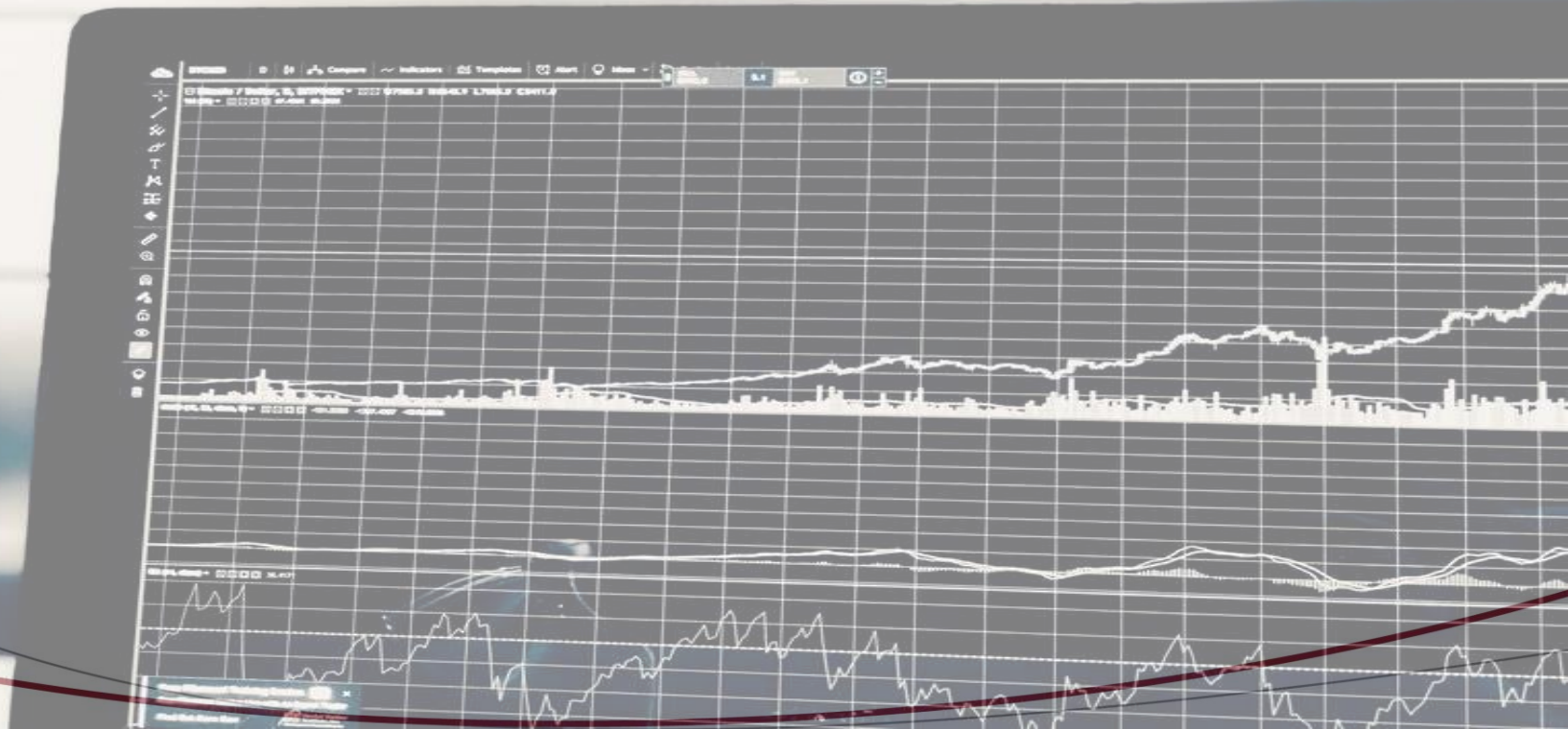
L'apport de notre méthodologie



INTÉRÊT :

Expliquer les informations des résidus en leur apportant davantage d'informations géographiques

Données Externes & Modélisation



COLLECTE D'INFORMATIONS GEOGRAPHIQUES EXTERNES

Addactis travaille depuis de nombreuses années sur l'alimentation d'une base de data externes.



Bases de données
publiques



Apis pour les
données du réseau
routier



Informations
météorologiques et
climatiques



flotte auto



activité éco



sécurité routière



API – réseau routier



météo



criminalité



Carte routière de la ville de Granville (50) obtenue via l'API Osmnx sous Python

**DONNEES
EXTERNES
collecte**



Retraitement des données externes

Nettoyage de la base après analyse de la pertinence, de la fiabilité et du rafraichissement des différentes sources.

Création de nouvelles variables pertinentes, stables et adaptées à chaque situation à partir des données brutes collectés (exploitables sur de nombreuses données brutes).

Analyse des corrélations des variables et traitement des données manquantes.

RETRAITEMENT
des données
externes



Calculées



Ajouter des variables calculées

Retraitées



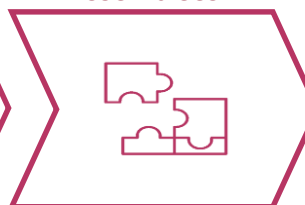
Analyse des données extrêmes, atypiques, manquantes

Analysées



One-one analysis, Statistiques descriptives

Assemblées



Croiser des variables, regrouper, assembler par corrélation

Lissées & Complétées

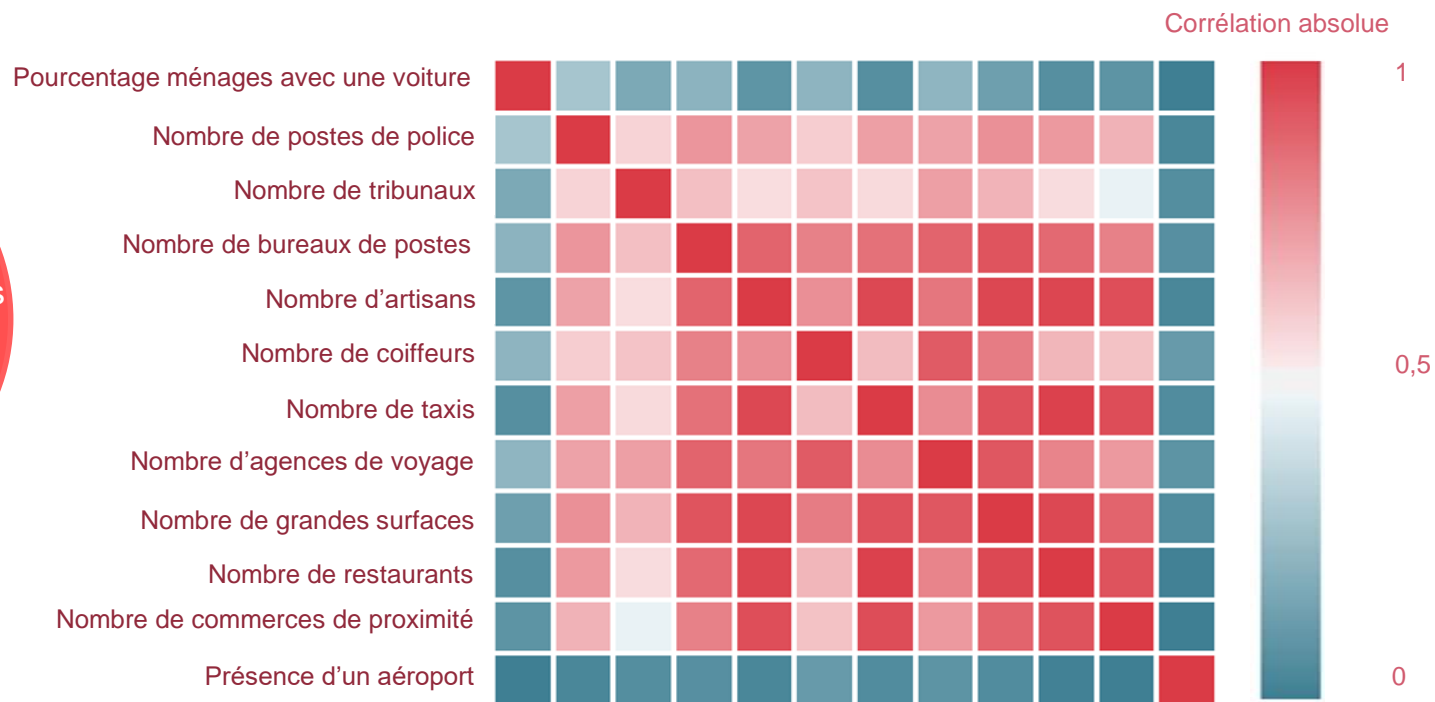


Lissage par crédibilité et/ou krigeage

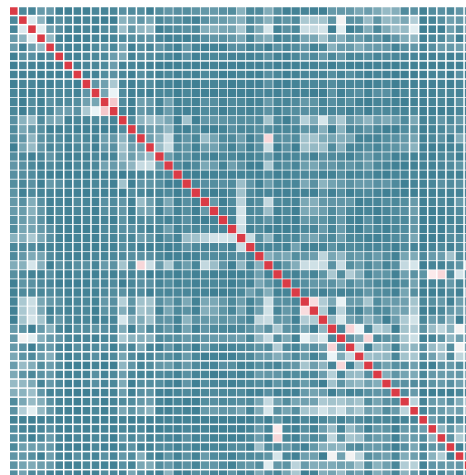
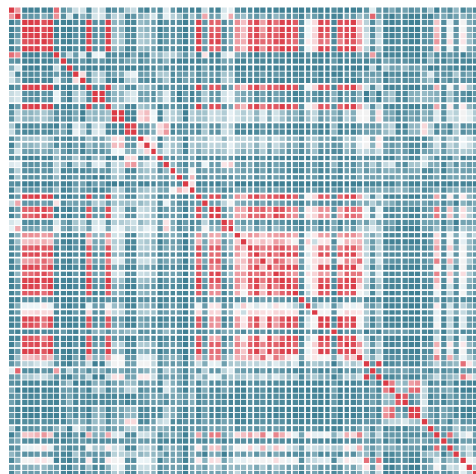
Etude des corrélations

Matrice de **corrélation** des variables avant sélection des variables, présentant des corrélations parfois fortes.

RETRAITEMENT
des données externes



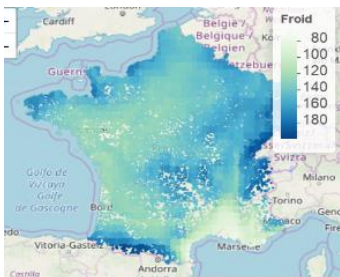
Sélection des variables avec du Machine Learning



De la « Raw Data » à la « Smart Data »



Météo



Criminalité



Mesures



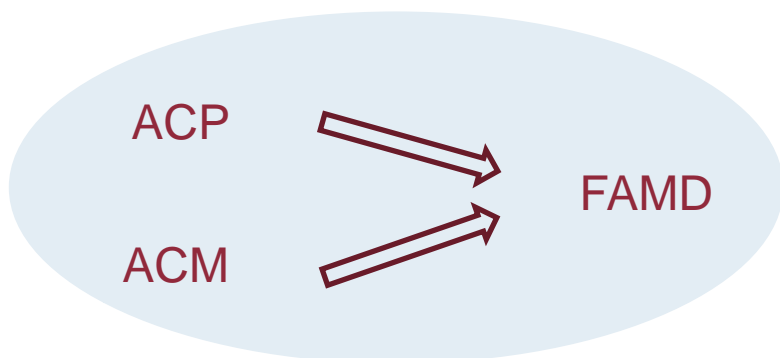
arrêté CAT-NAT



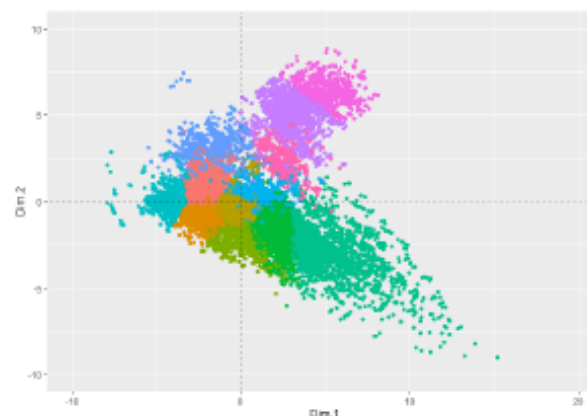
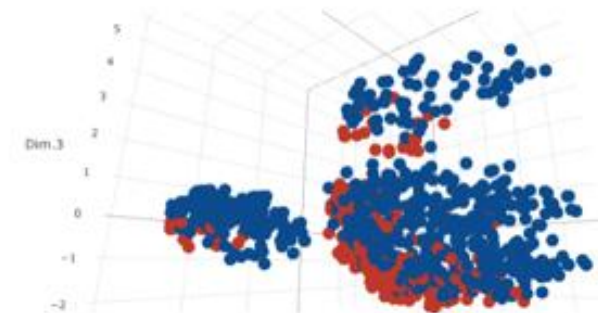
Eyrieux	Inondation	7221	Sair
Eyrieux	Inondation	7237	Sair
Eyrieux	Inondation	7248	Sair
Eyrieux	Inondation	7252	Sair
Eyrieux	Inondation	7256	Sair
Eyrieux	Inondation	7261	Sair
Eyrieux	Inondation	7269	Sair
Eyrieux	Inondation	7274	Sair
Eyrieux	Inondation	7278	Sair
Eyrieux	Inondation	7295	Sair
Eyrieux	Inondation	7303	Sair
Eyrieux	Inondation	7349	LaV
e Chassezac	Inondation	7017	Les
Chassezac	Inondation	7028	Beau
Chassezac	Inondation	7031	Berr
Chassezac	Inondation	7050	Char

Comment choisir et retraiter les données ?

Sélection de variables :



Analyse **simultanée** des données
quantitatives et qualitatives

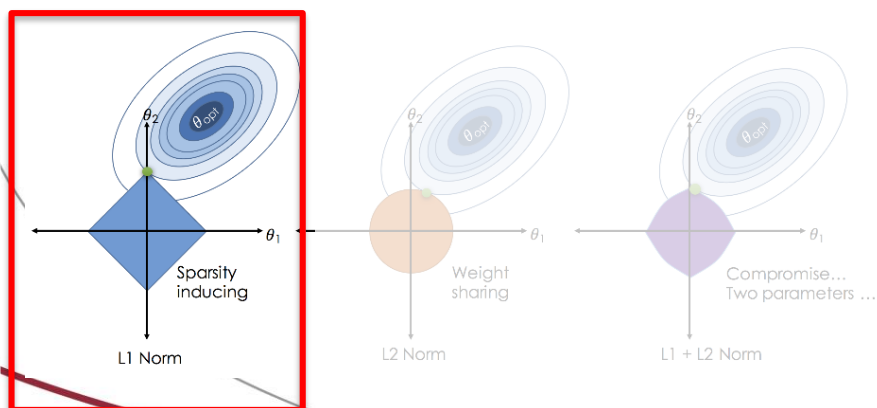


Comment choisir et retraiter les données ?

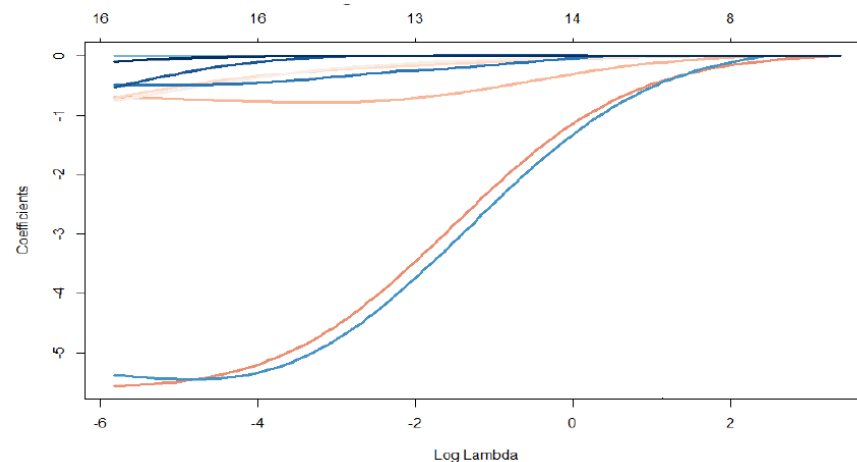
Sélection de variables :

$$\hat{\beta} = \operatorname{argmin} \left\{ \sum_{i=1}^n \left[y_i - \left(\beta_0 + \sum_{j=1}^k \beta_j x_{ji} \right) \right]^2 + \lambda \sum_{j=1}^k |\beta_j| \right\}$$

Méthode **LASSO** (pénalisation en norme L1)

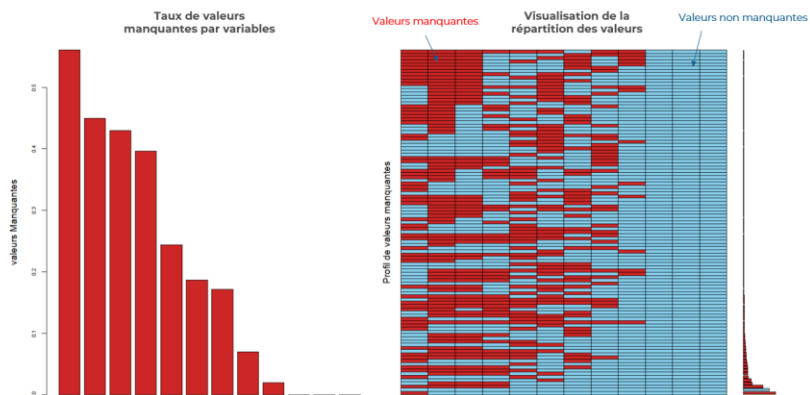


Trajectoires de régularisation



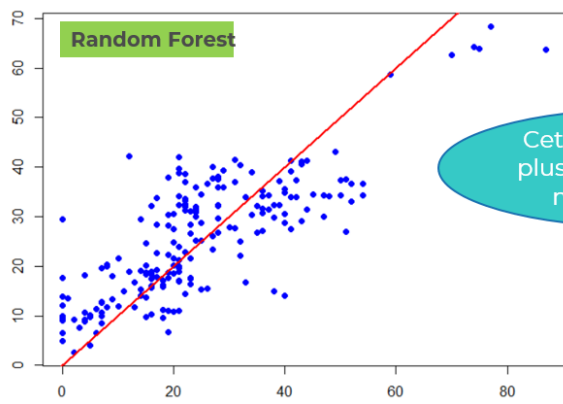
Comment choisir et retraiter les données ?

DATA WRANGLING (Retraitement, structuration, enrichissement)

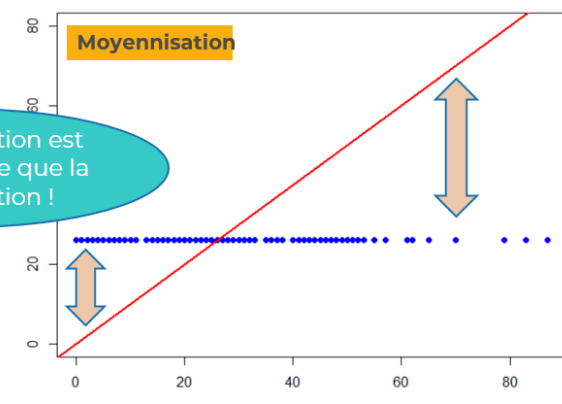


Diagnostic des données manquantes

Complétion des données manquantes (au-delà de la moyenne !!)



Cette complétion est plus pertinente que la moyennisation !



Comment choisir et retraiter les données ?

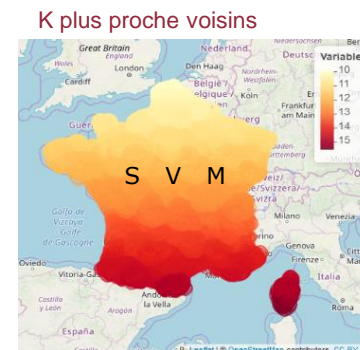
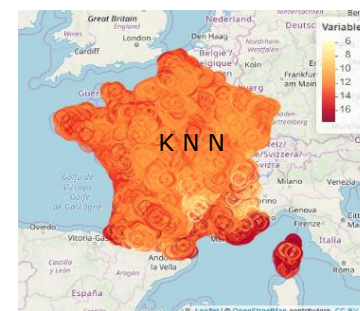
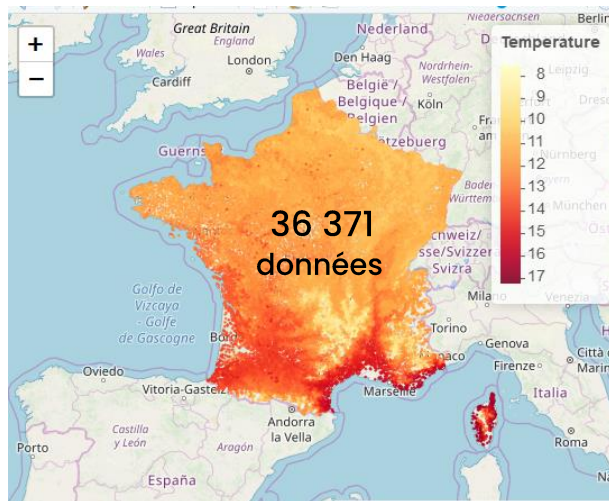
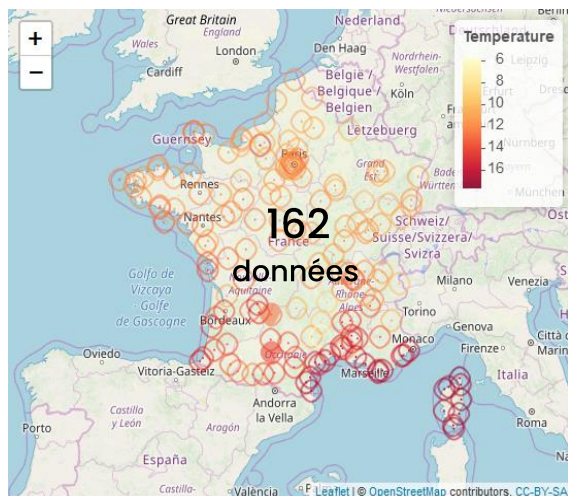
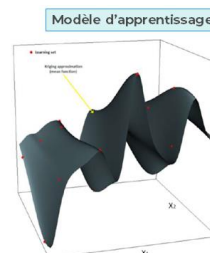
DATA WRANGLING (Retraitement, structuration, enrichissement)

Enrichissement par Kriging :

Interpolation par processus Gaussien conditionné

$$Y(x^*)|Y_n \sim \mathcal{N}(m(x^*), \tilde{\sigma}^2(x^*))$$

$$m(x^*) = \mathbb{E}[Y(x^*)|Y_n] = \mu(x^*) + k(x^*)^T K^{-1}(Y_n - \mu)$$



Support vector machine

exemple avec des données météo

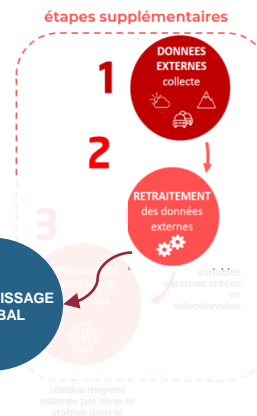
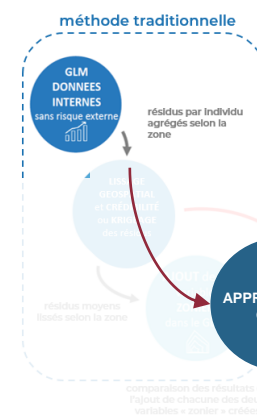
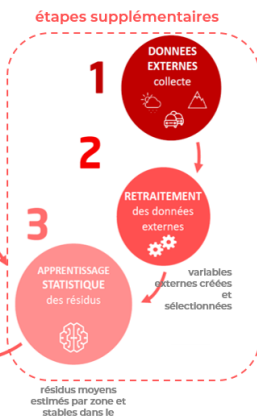
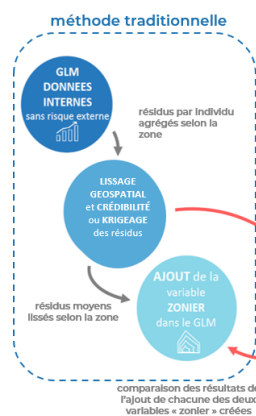
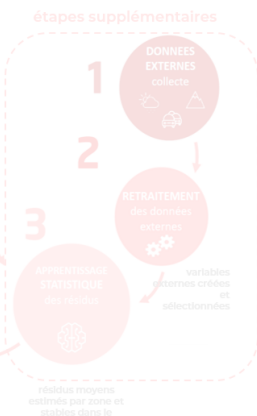
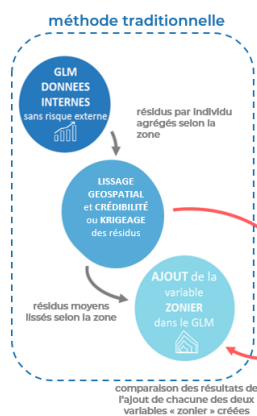
Stratégie de modélisation

Il existe plusieurs stratégies de modélisation pour la construction et l'intégration d'un zonier

GLM
+
Lissage des résidus

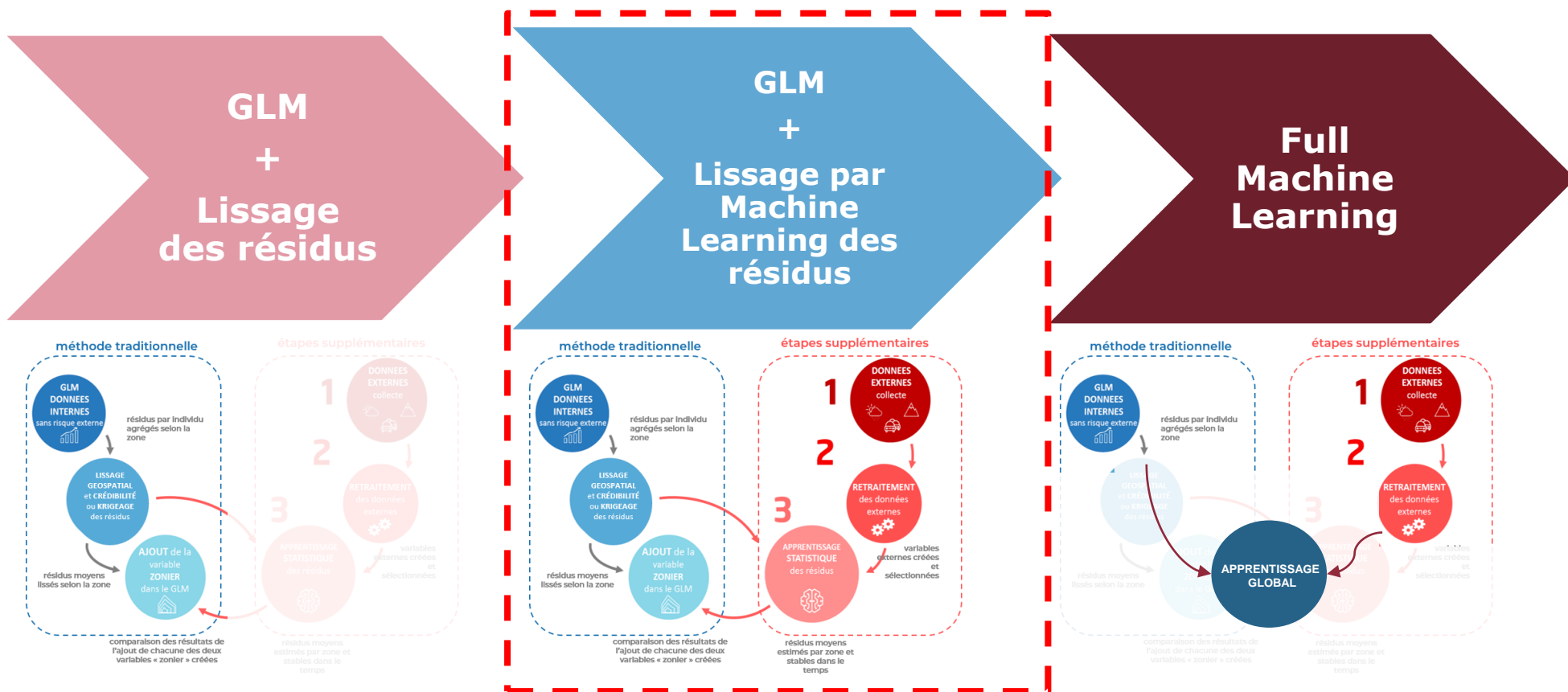
GLM
+
Lissage par Machine Learning des résidus

Full
Machine
Learning



Stratégie de modélisation

Il existe plusieurs stratégies de modélisation pour la construction et l'intégration d'un zonier

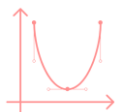


Apprentissage statistique des résidus

Modèles envisageables :



RANDOM
FOREST



GRADIENT
BOOSTING






RÉSEAUX DE
NEURONES

APPRENTISSAGE
STATISTIQUE
des résidus

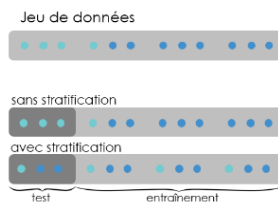


Objectifs du Machine Learning

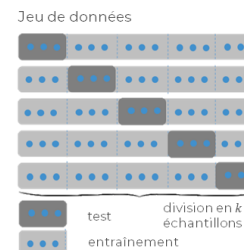
-  Réexpliquer le risque géographique uniquement à l'aide de données externes
-  Comprendre les raisons des niveaux de danger
-  Etre stable dans le temps et capable d'absorber les chocs

Valider la robustesse de chaque modèle

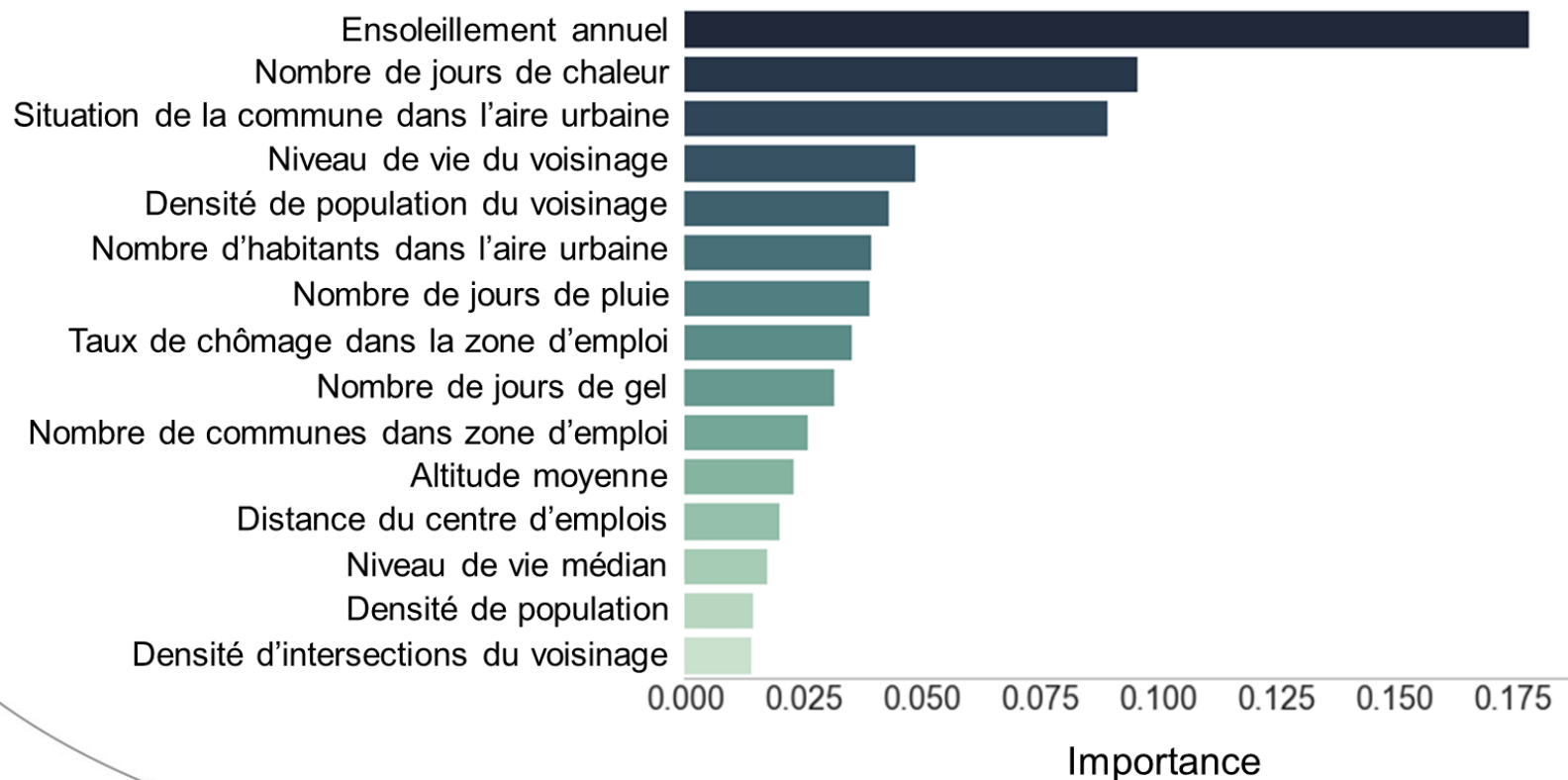
Stratification



Validation croisée

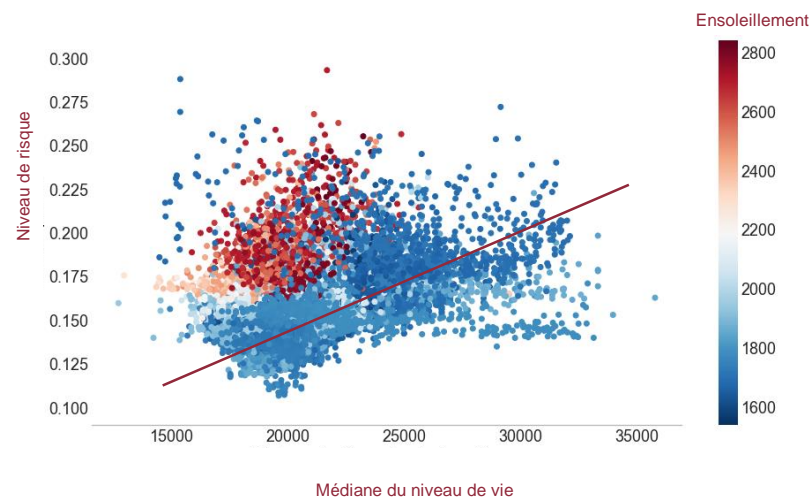
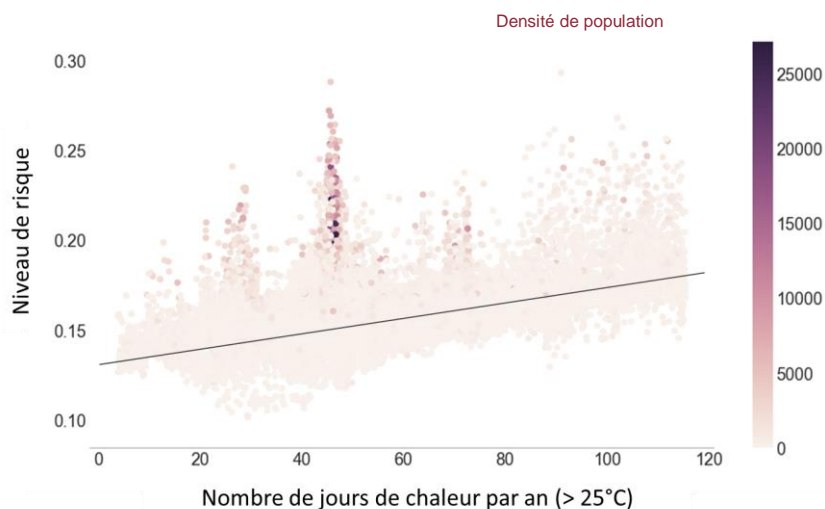


DOMMAGE – L'IMPORTANCE DES FACTEURS CLIMATIQUES ET DES DYNAMIQUES DE PEUPLEMENT

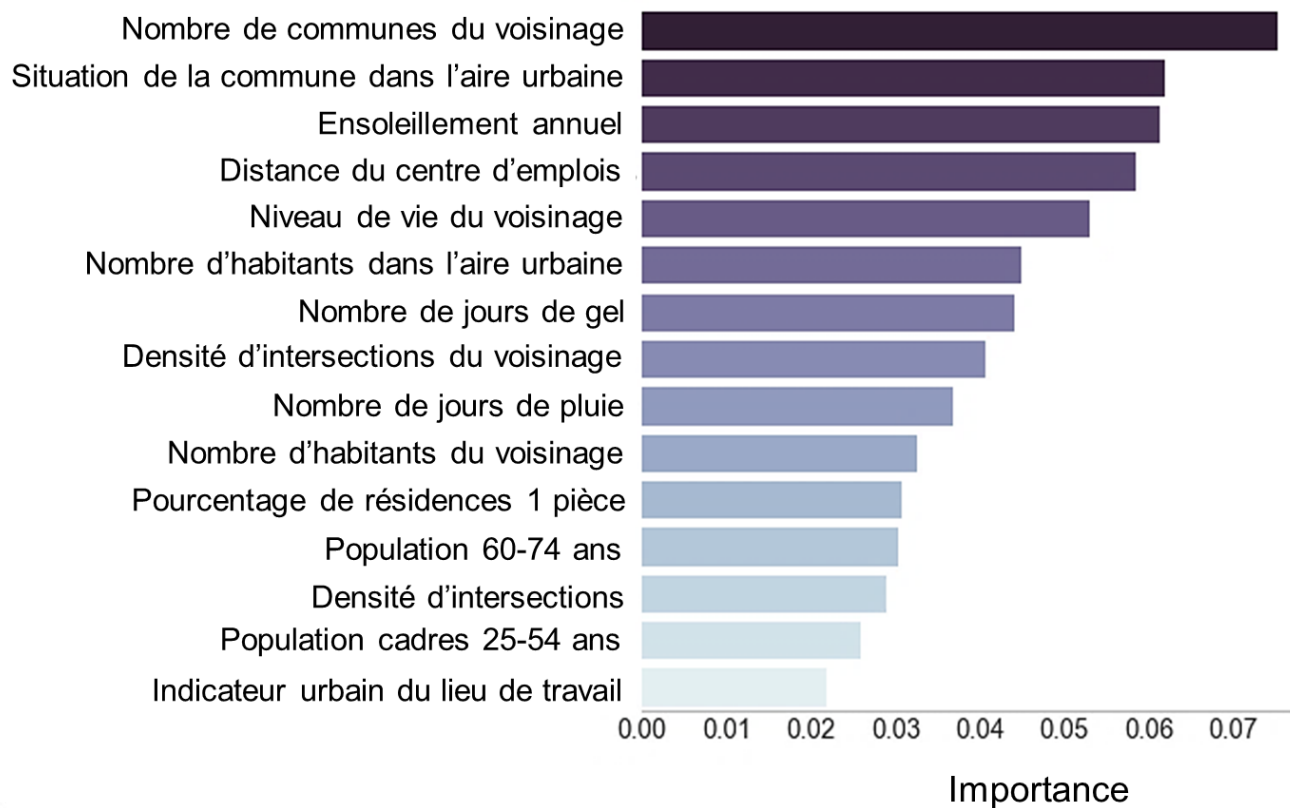


IMPORTANCE DES FACTEURS EXTERNES – DOMMAGE

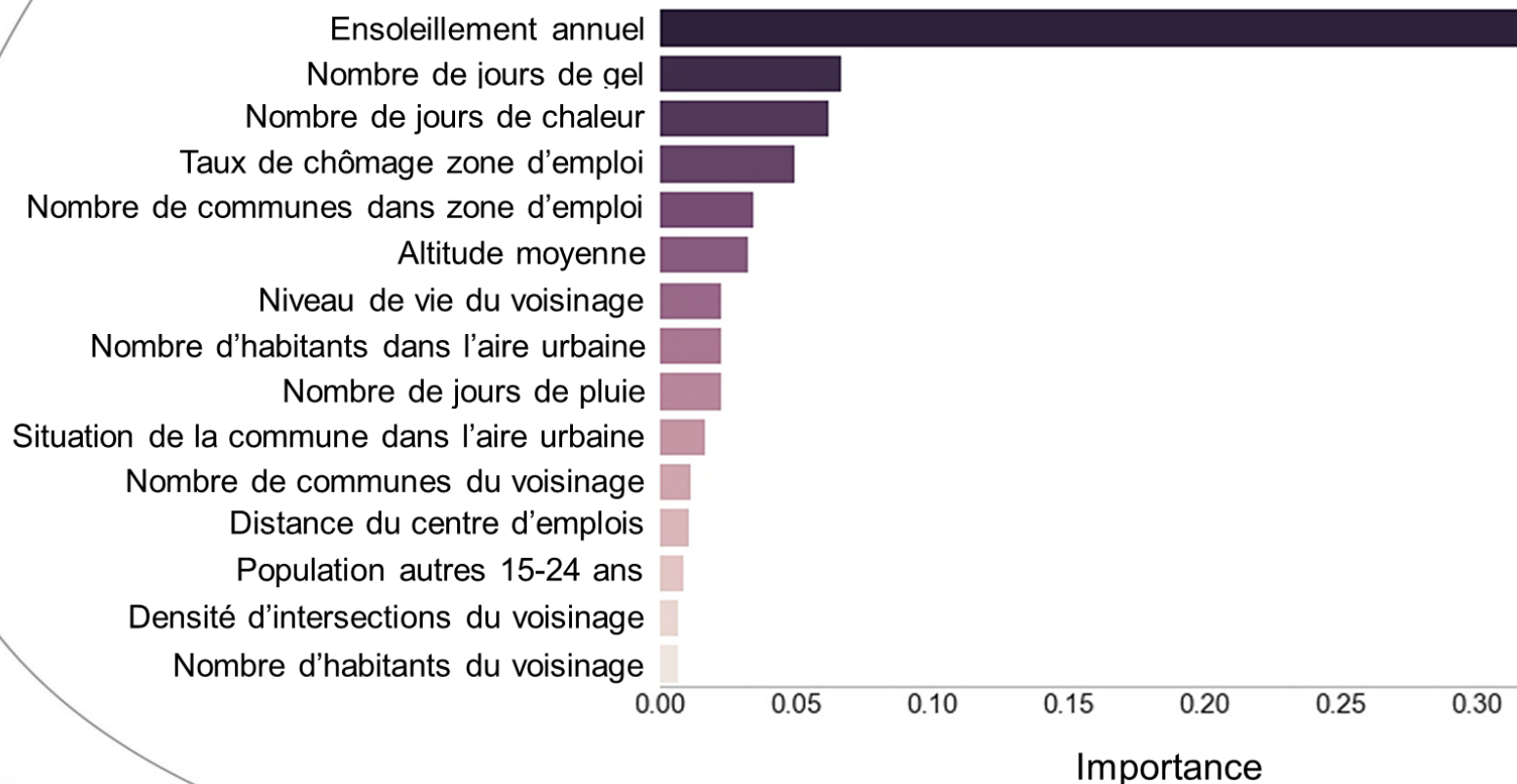
Pouvoir explicatif significatif des variables météorologiques



VOL – L'IMPORTANCE DES FACTEURS SOCIO- ECONOMIQUES ET DES DONNÉES DU VOISINAGE



BRIS DE GLACE – LES FACTEURS METEOROLOGIQUES CONTRIBUENT MASSIVEMENT A L'APPRENTISSAGE



Choix du modèle

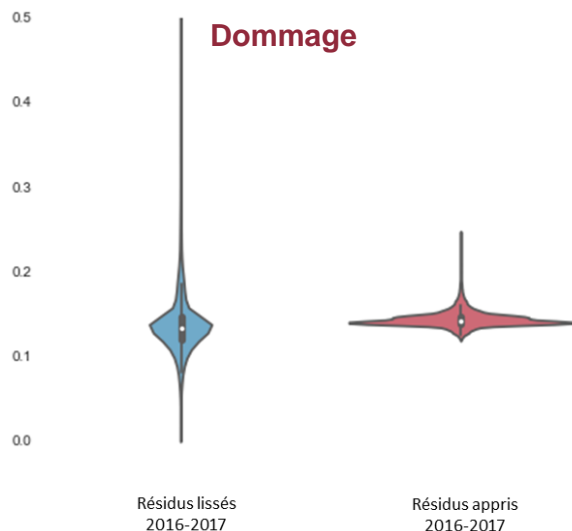
Modèle retenu pour chaque garantie : **GRADIENT TREE BOOSTING**

- **RÉSEAUX DE NEURONES** : Inefficaces pour l'importance des variables
- **RANDOM FOREST** : Moins fidèle à la réalité de l'échantillon d'apprentissage

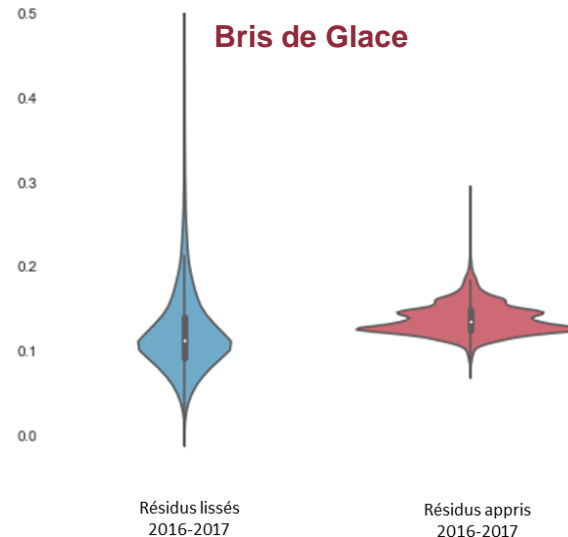
Validation



Répartition des valeurs des résidus lissés et prédits pour la garantie Dommage



Répartition des valeurs des résidus lissés et prédits pour la garantie Bris de Glace





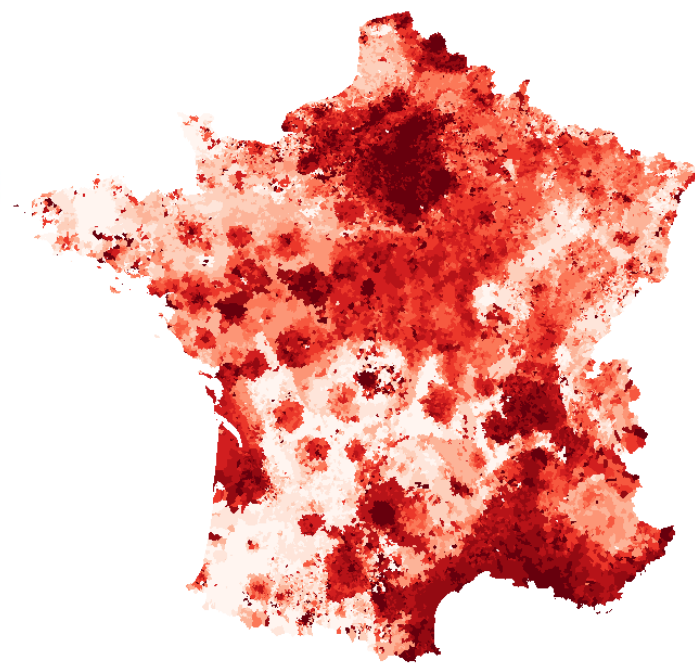
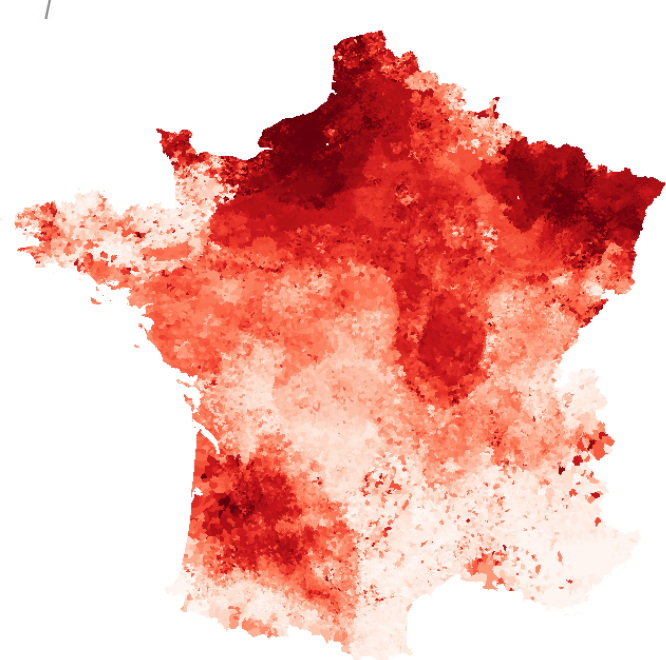
Les Résultats

RESULTATS CARTOGRAPHIQUES – IMPACT INÉGAL DES FACTEURS REGIONAUX SELON LES GARANTIES

GARANTIE
BRIS DE GLACE

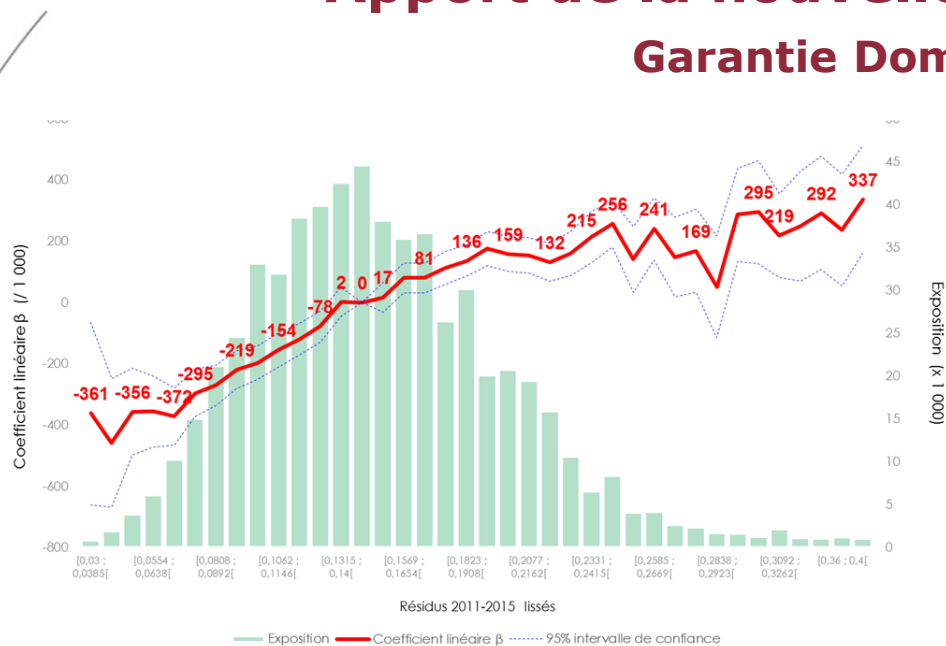
GARANTIE
DOMMAGE

GARANTIE
VOL

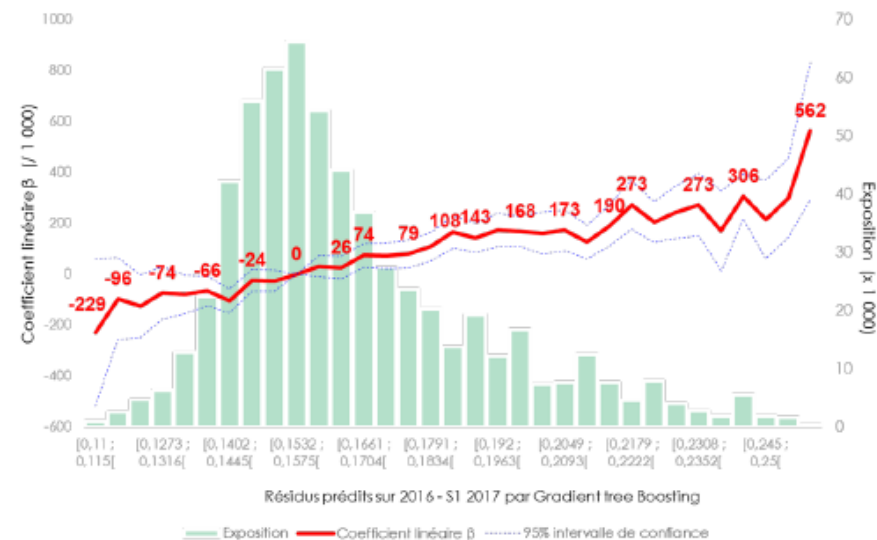


Éclairer les risques, tracer l'avenir

Apport de la nouvelle méthodologie Garantie Dommage



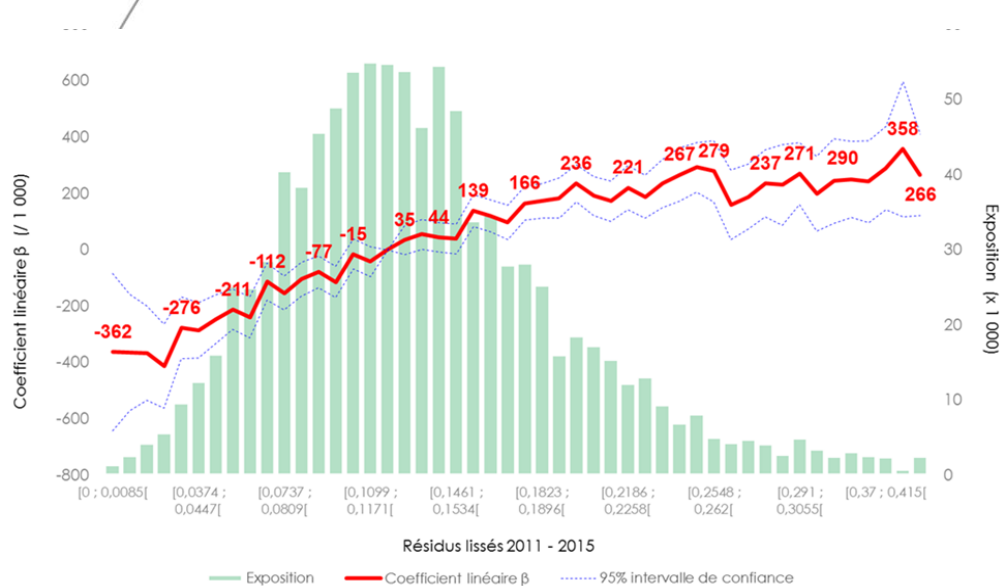
Valeurs des coefficients estimés (rouge) pour chaque segment de valeurs des résidus lissés de la méthode traditionnelle associé à son exposition (vert)



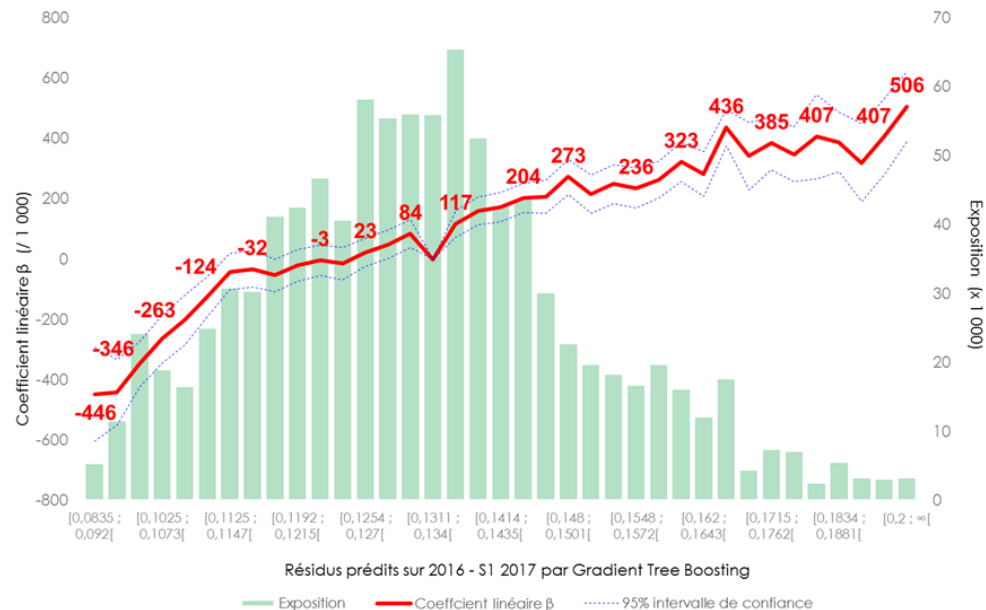
Valeurs des coefficients estimés (rouge) pour chaque segment de valeurs de prédiction des résidus moyens lissés par le Gradient Tree Boosting associé à son exposition (vert)

Gain en amplitude des coefficients β
et une meilleure stabilité aux valeurs extrêmes

Apport de la nouvelle méthodologie Garantie Bris de Glace



Valeurs des coefficients estimés (rouge) pour chaque segment de valeurs des résidus lissés de la méthode traditionnelle associé à son exposition (vert)



Valeurs des coefficients estimés (rouge) pour chaque segment de valeurs de prédiction des résidus moyens lissés par le Gradient Tree Boosting associé à son exposition (vert)

Gain en amplitude des coefficients β
et une meilleure stabilité aux valeurs
extrêmes

Conclusion & Perspectives





IMPORTANCE DE LA COLLECTE ET DU RETRAITEMENT
DES DONNÉES



LA QUALITÉ DES DONNÉES CONDITIONNE LA PERFORMANCE
DE TOUS LES MODÈLES



LA VISÉE OPÉRATIONNELLE FAÇONNE LE
TRAVAIL TECHNIQUE

Et le Véhiculier ?



MÉTHODOLOGIE



MERCI DE VOTRE ATTENTION