100% ACTUAIRES & 100% DATA SCIENCE

INSTITUT DES ACTUAIRES

16 Novembre 2018

Hôtel Marriott Rive Gauche
Paris 14ème

100% ACTUAIRES & 100% DATA SCIENCE

INSTITUT DES ACTUAIRES

Modélisation du risque géographique en assurance habitation

Jennifer PARIENTE BOUNAN – Prix Scor 2017 du jeune actuaire

Cadre et objectifs du mémoire

- Un marché concurrentiel avec une nécessité d'excellence technique liée à l'offre internet et à la législation française
- Nécessité d'améliorer constamment la segmentation afin de fournir des tarifs adaptés aux différents profils de risque

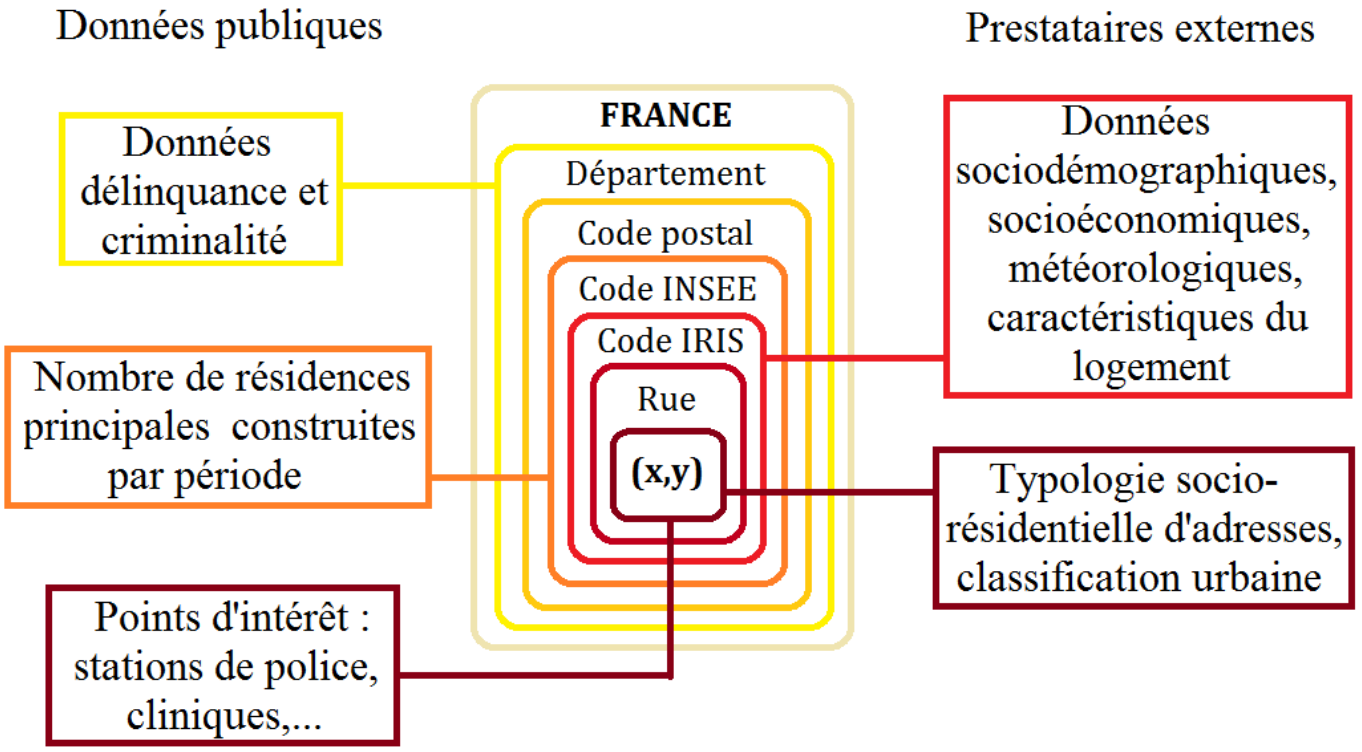
Problématiques principales

- Tester l'intégration de nouvelles données géographiques à des mailles fines
- Construire un nouveau processus du traitement du signal géographique à la fois stable et précis

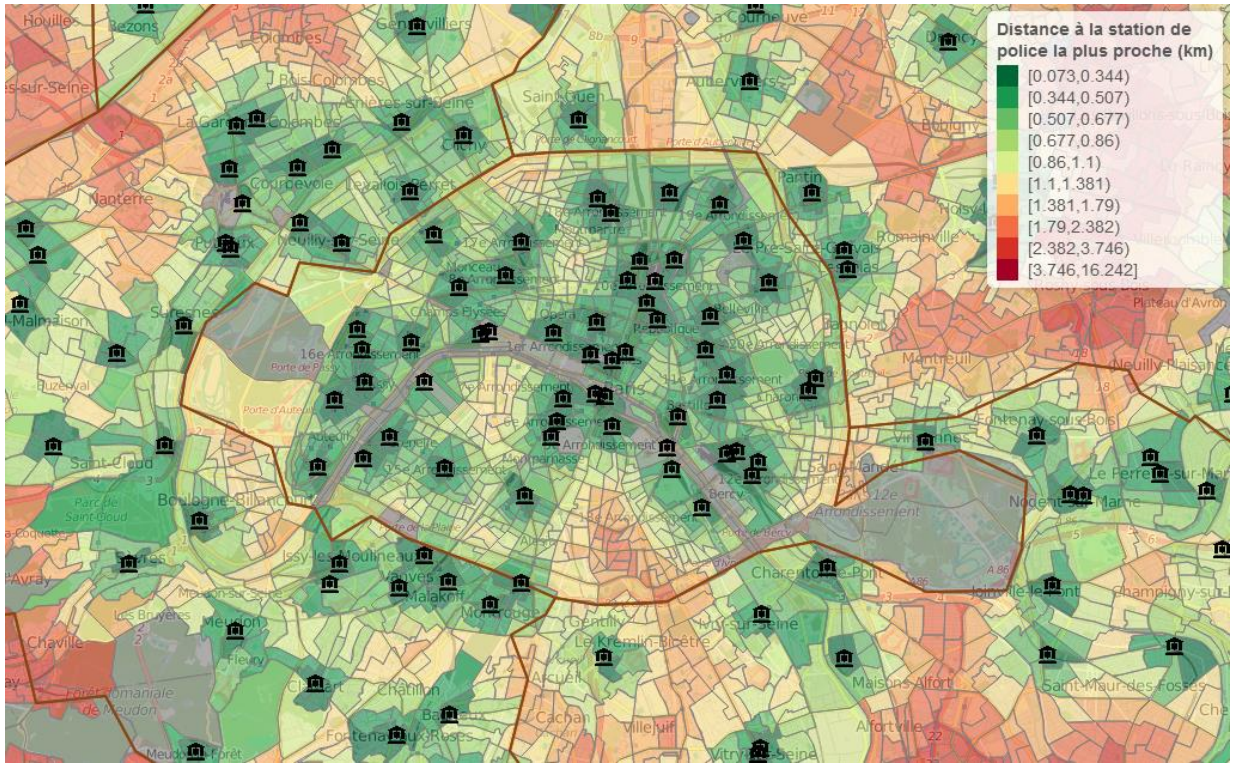
Sommaire

Sommaire

Des données provenant de différentes sources et concernant plusieurs mailles géographiques.



De nouvelles variables créées afin d'exploiter au mieux l'information



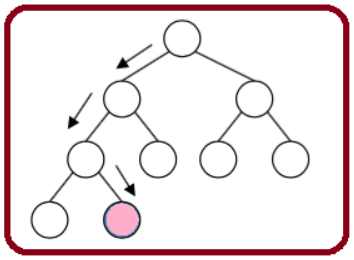
Visualisation de la distance à la station de police la plus proche pour Paris et ses alentours (par IRIS).

Sommaire

Sommaire

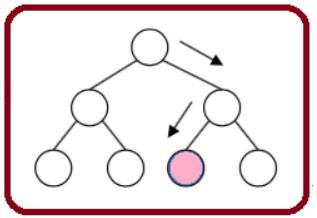
L'algorithme utilisé : le Gradient Tree Boosting pour traiter les données

➡ Plus de 600 variables traitées par Gradient Tree Boosting



$$\hat{f} = \hat{f}_{Arbre1} + \hat{f}_{Arbre2} + \dots$$

Gradient



Gradient

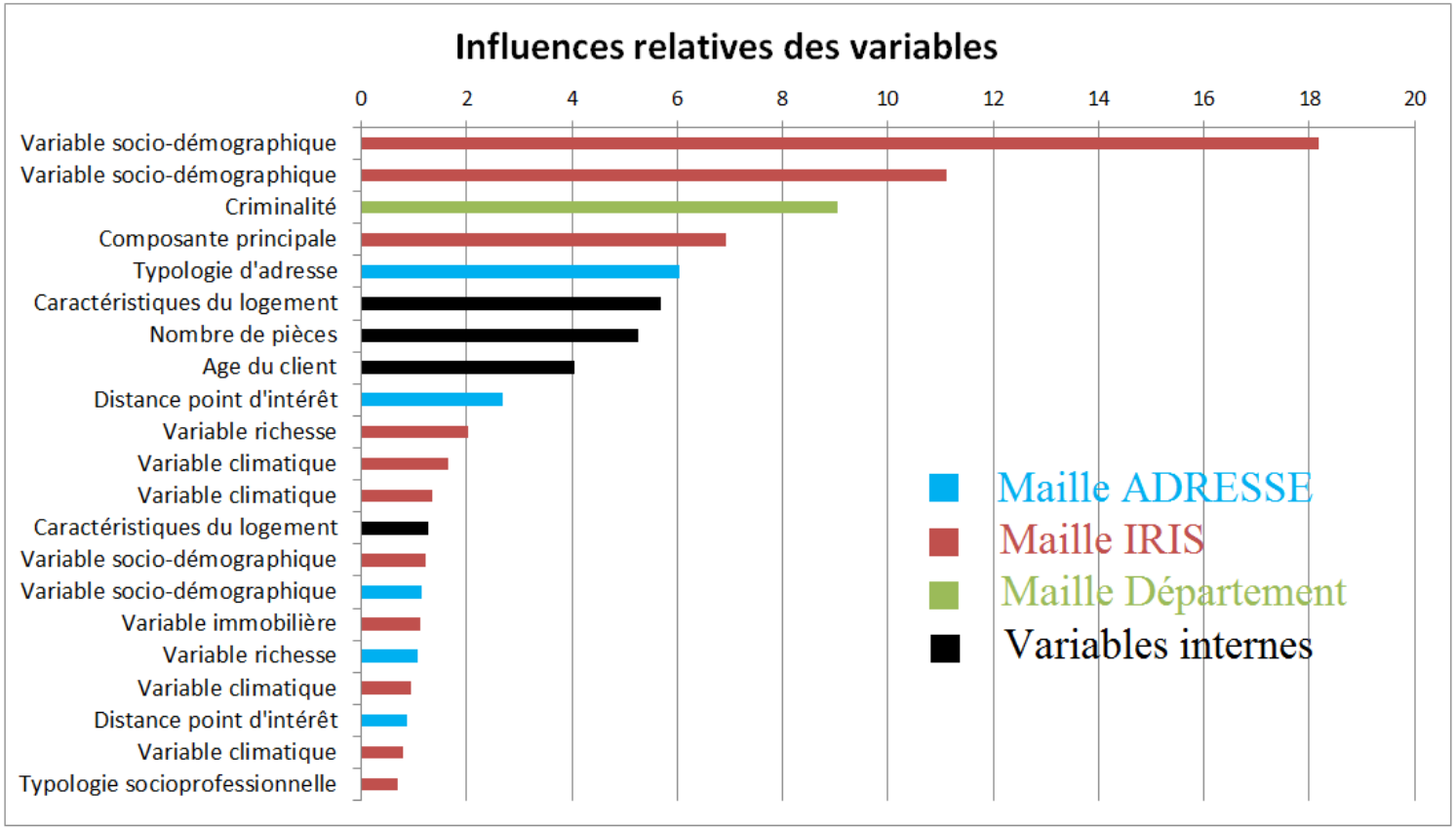
...

Prédiction finale

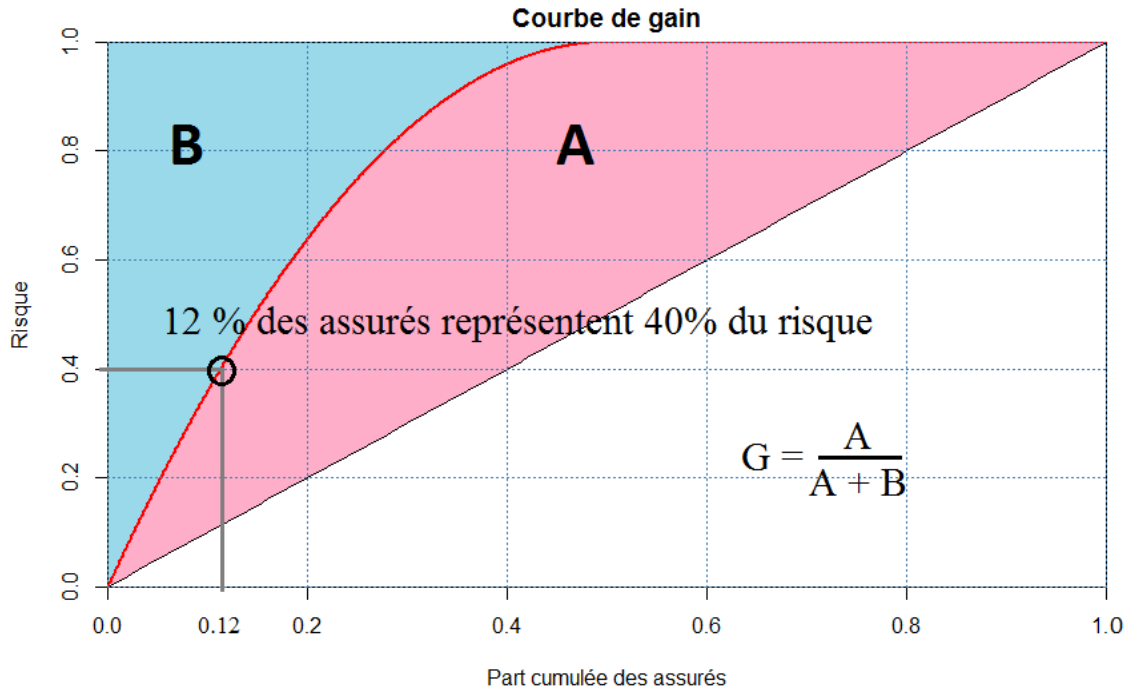
Avantages

- Sélection automatique des variables
- Bonne gestion des valeurs manquantes
- Modèle non paramétrique qui introduit des effets non linéaires

L'importance relative des variables



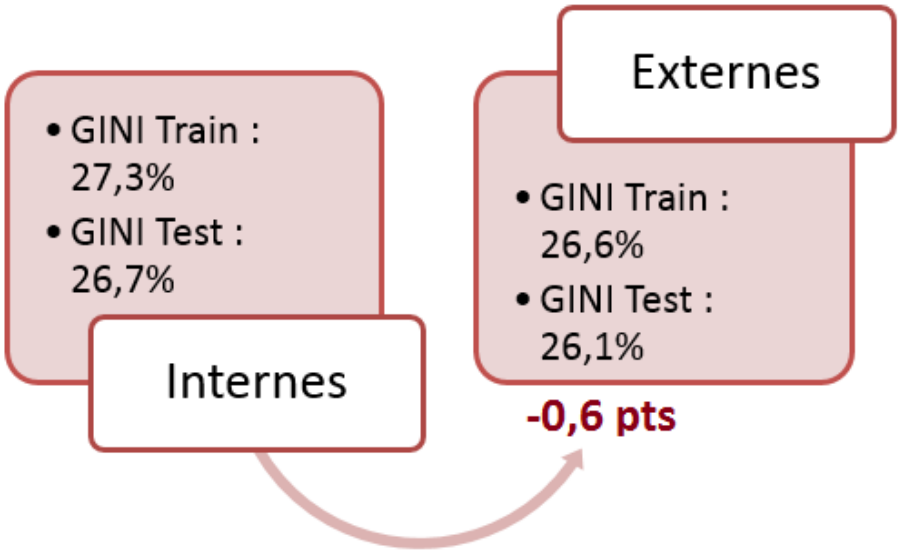
L'indicateur de performance utilisé : l'indice de GINI



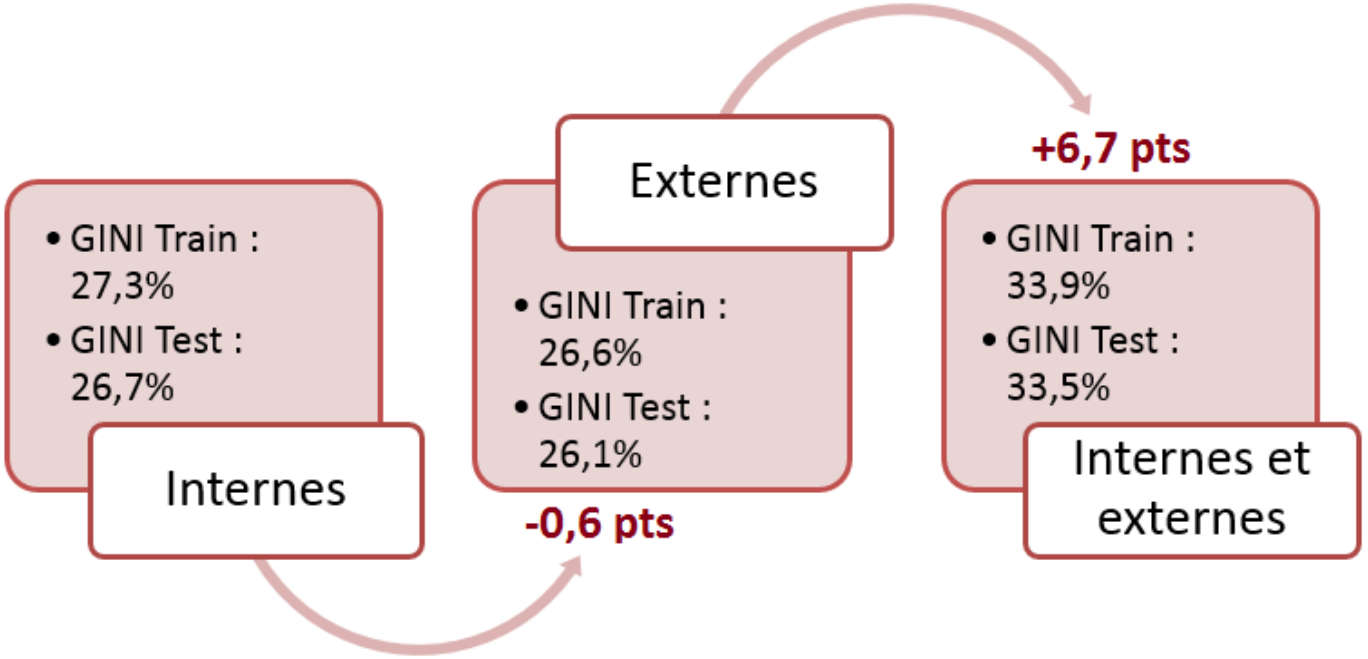
➡ Pour identifier les individus les plus risqués

Sommaire

Simplification du questionnaire client

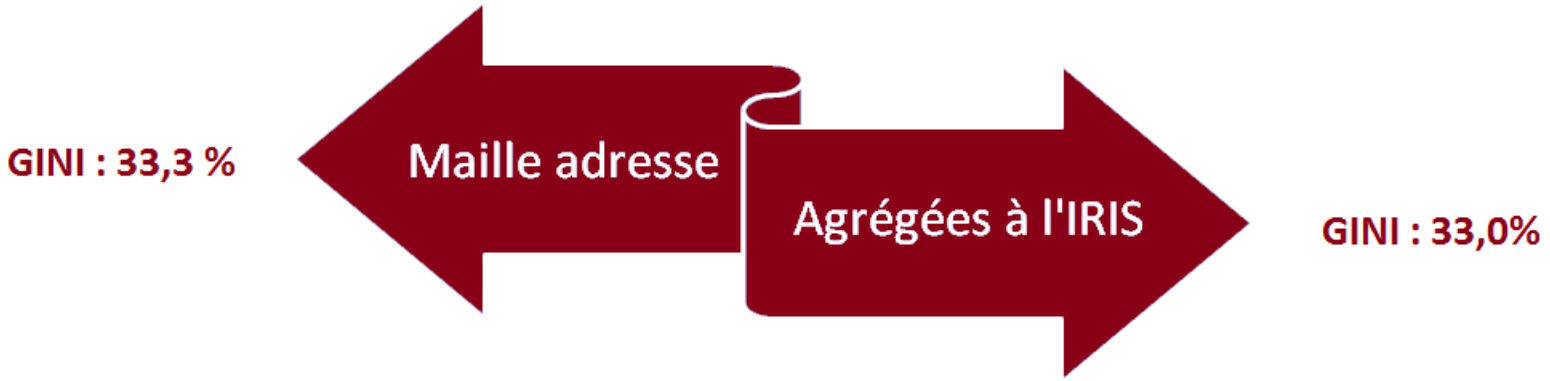


Simplification du questionnaire client et gain d'information



Données internes et externes sont des sources d'information **complémentaires**.

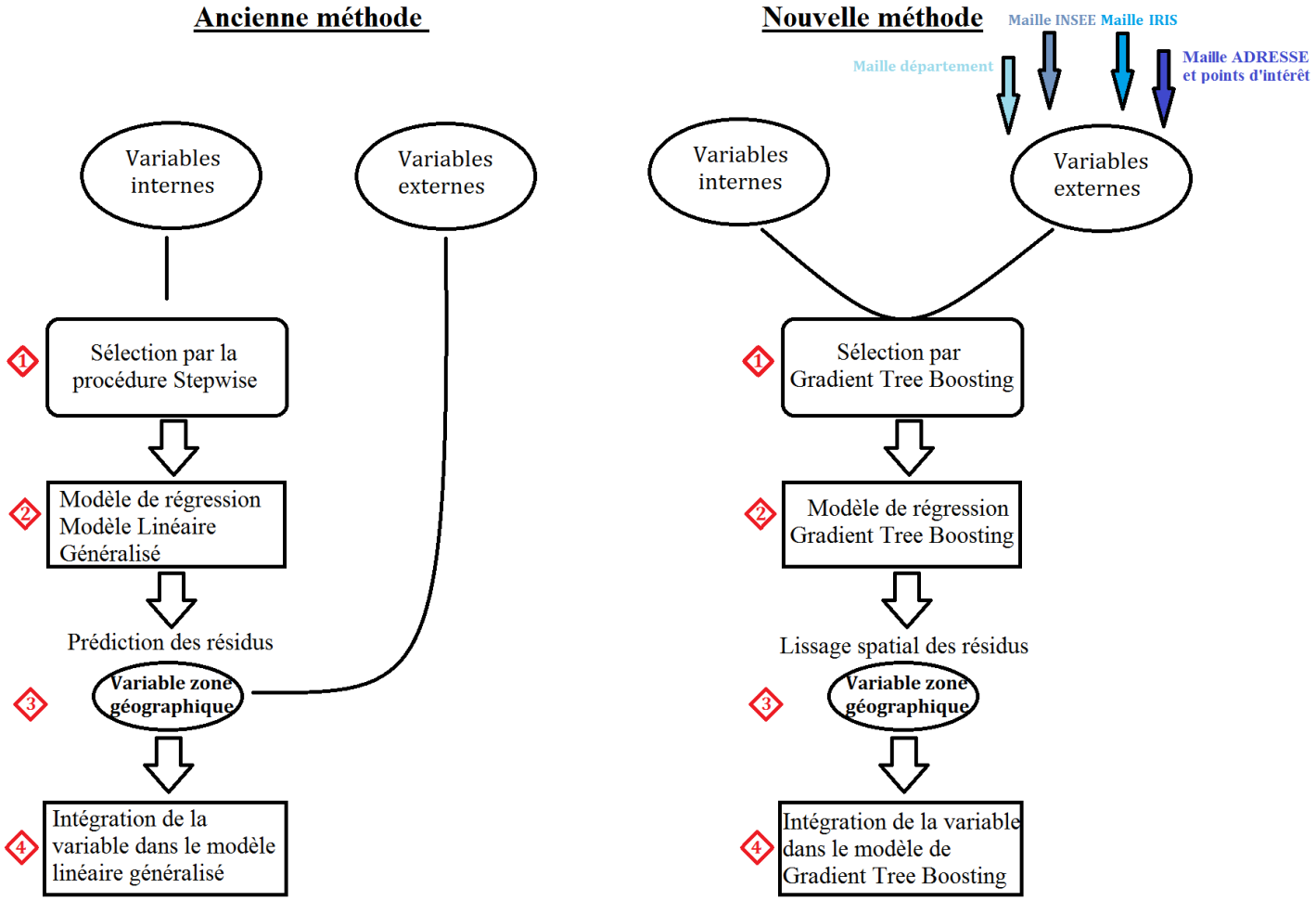
→ Gain faible d'utiliser certaines données du prestataire à la maille ADRESSE.



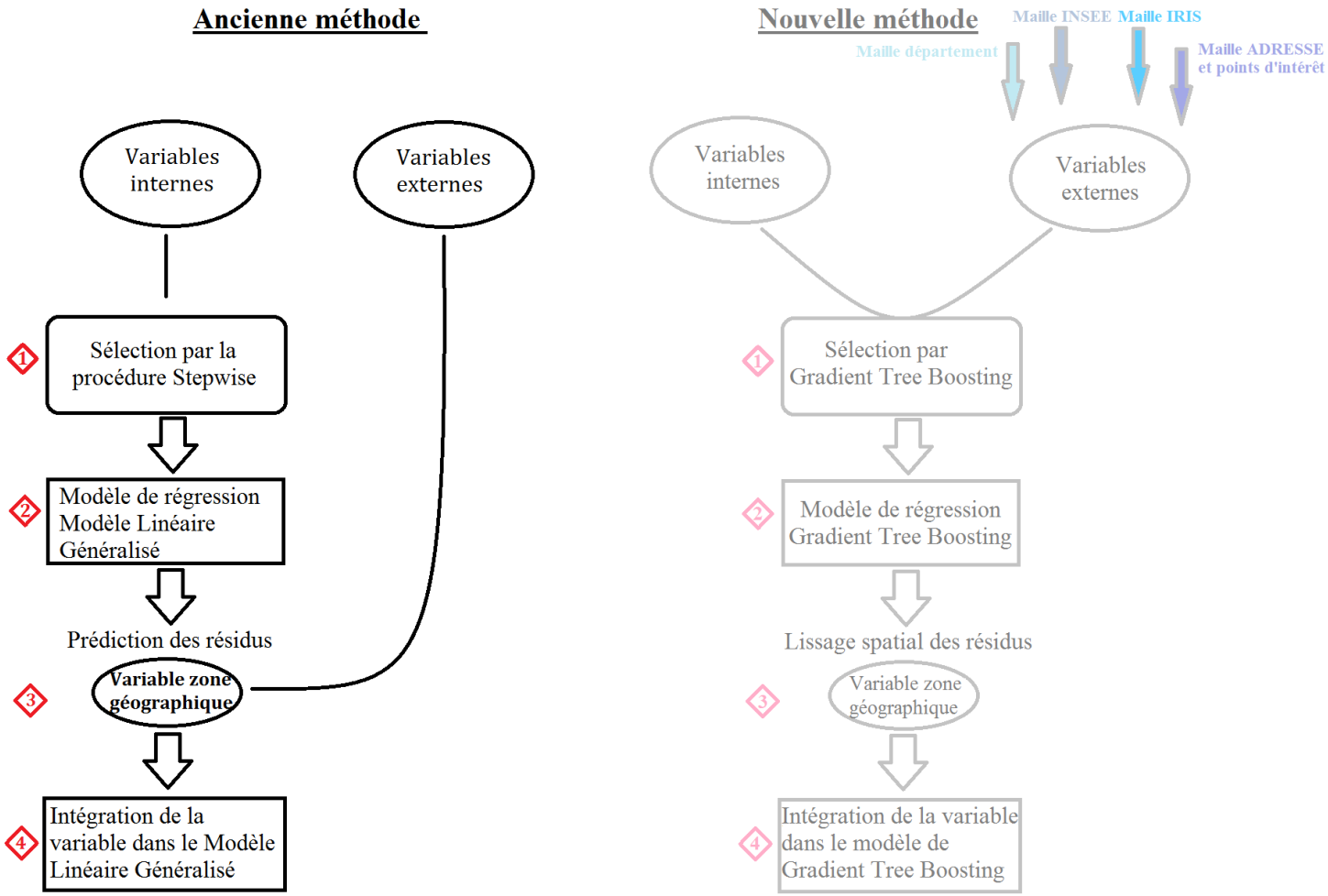
→ Substitution de variables internes peu fiables comme les montants assurés par des variables externes

Sommaire

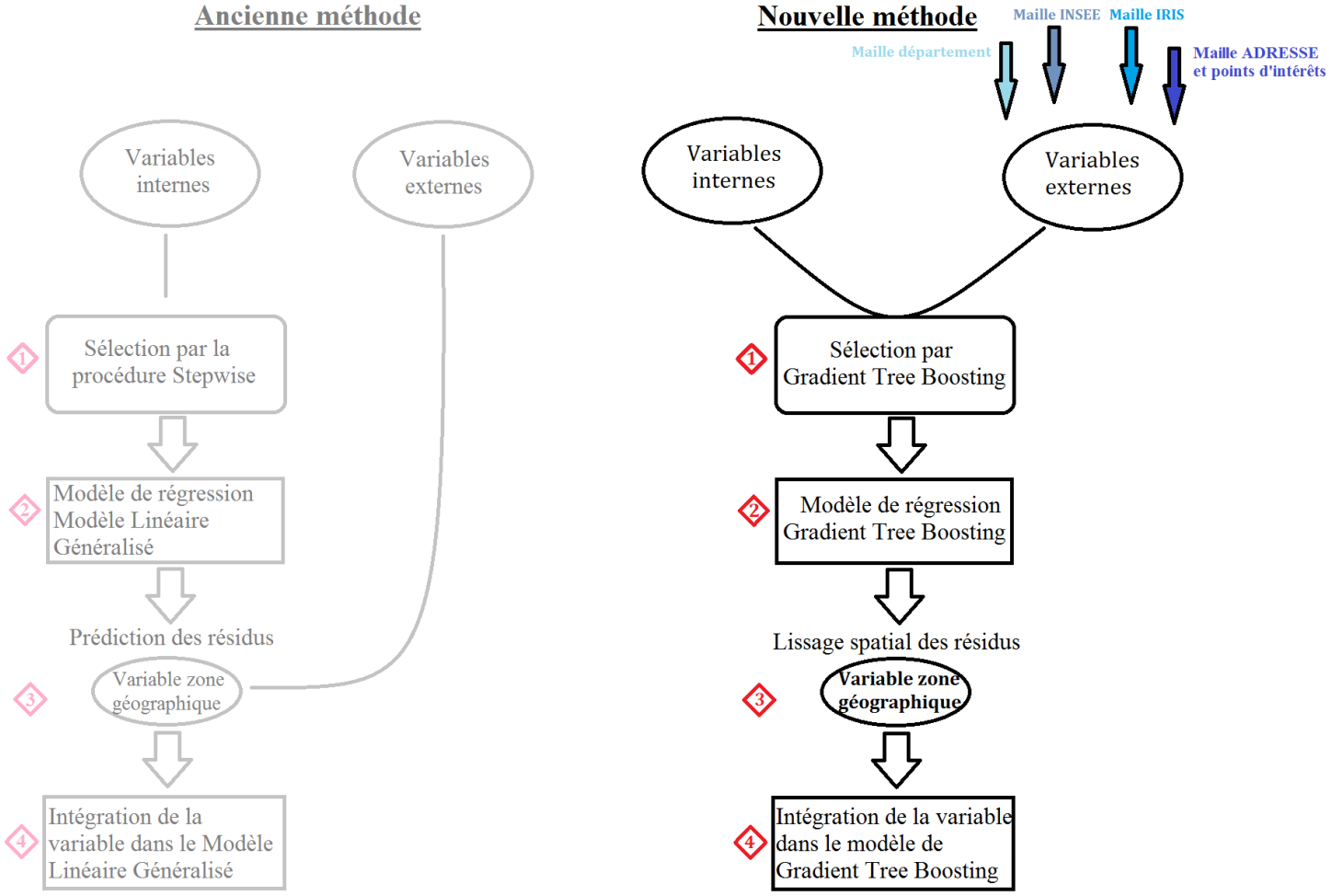
Sommaire

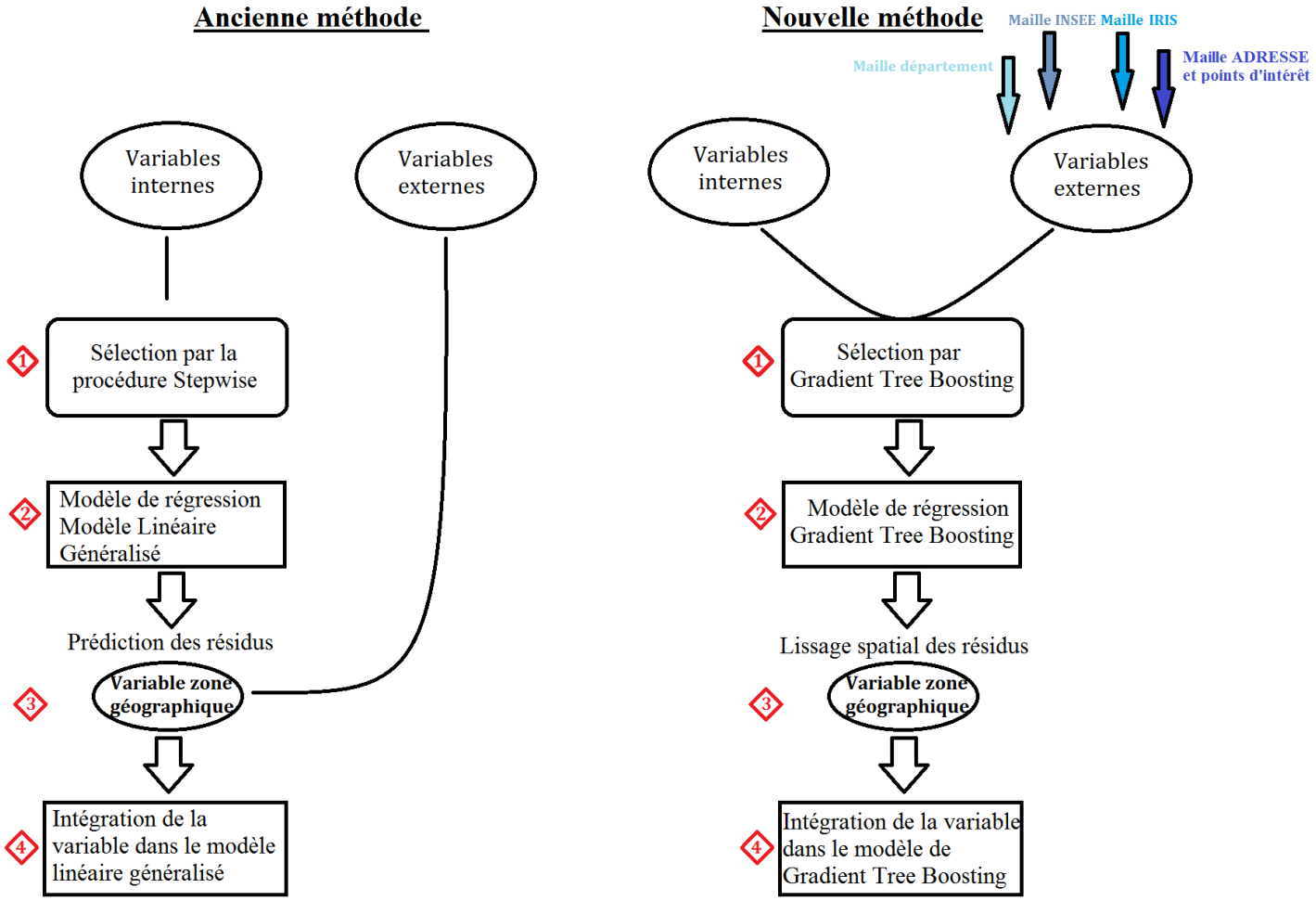

Ancienne méthode
Variables internes
Variables externes
1
Sélection par la procédure Stepwise
2
Modèle de régression Modèle Linéaire Généralisé
Prédiction des résidus
3
Variable zone géographique
4
Intégration de la variable dans le modèle linéaire généralisé
Nouvelle méthode
Maille INSEE
Maille IRIS
Maille département
Maille ADRESSE et points d'intérêt
Variables internes
Variables externes
1
Sélection par Gradient Tree Boosting
2
Modèle de régression Gradient Tree Boosting
Lissage spatial des résidus
3
Variable zone géographique
4
Intégration de la variable dans le modèle de Gradient Tree Boosting

3. Création d'un micro-zonier à l'aide de méthodes alternatives



3. Création d'un micro-zonier à l'aide de méthodes alternatives



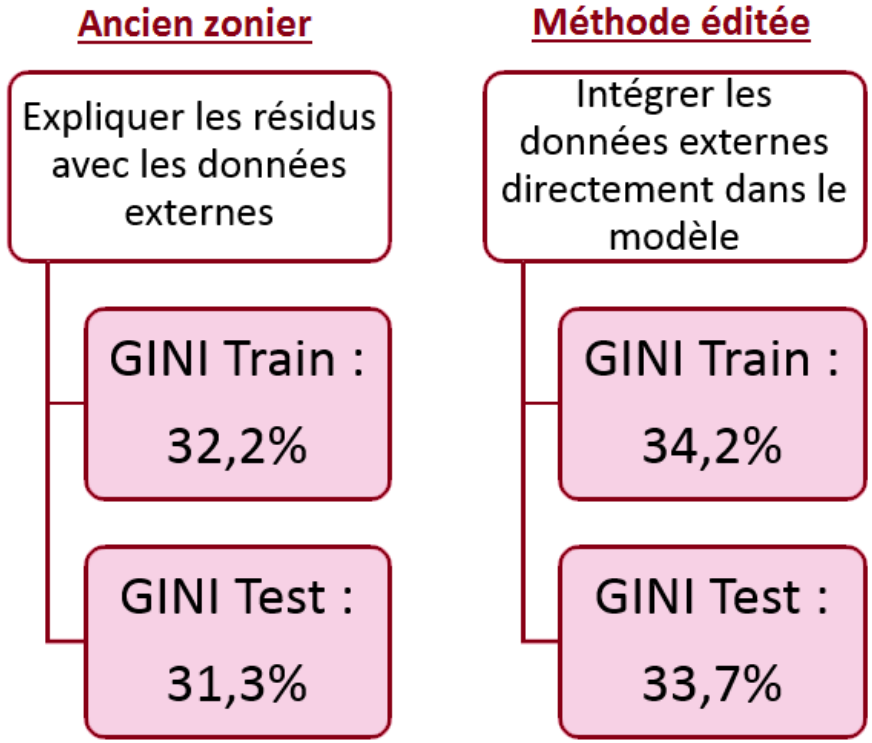

Ancienne méthode
Variables internes
Variables externes
1
Sélection par la procédure Stepwise
2
Modèle de régression Modèle Linéaire Généralisé
Prédiction des résidus
3
Variable zone géographique
4
Intégration de la variable dans le modèle linéaire généralisé
Nouvelle méthode
Maille INSEE
Maille IRIS
Maille département
Maille ADRESSE et points d'intérêt
Variables internes
Variables externes
1
Sélection par Gradient Tree Boosting
2
Modèle de régression Gradient Tree Boosting
Lissage spatial des résidus
3
Variable zone géographique
4
Intégration de la variable dans le modèle de Gradient Tree Boosting

Sommaire

Intégration directe des variables externes dans le modèle

Comparaison avec l'ancienne méthode

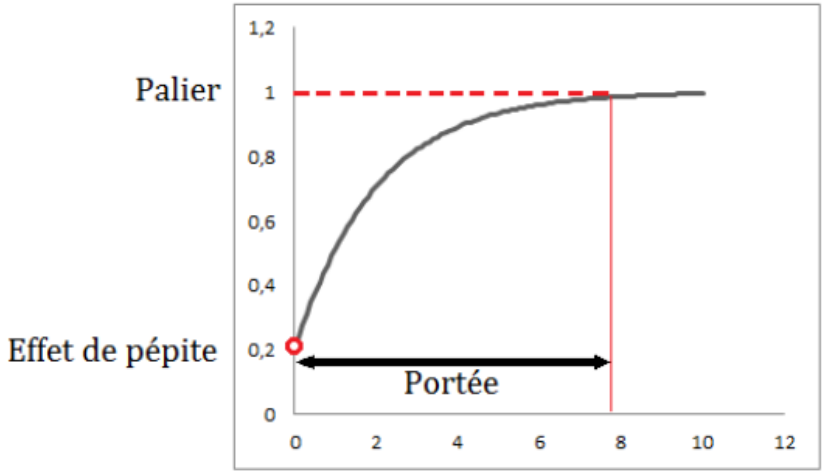
Profondeur d'historique : 1 an

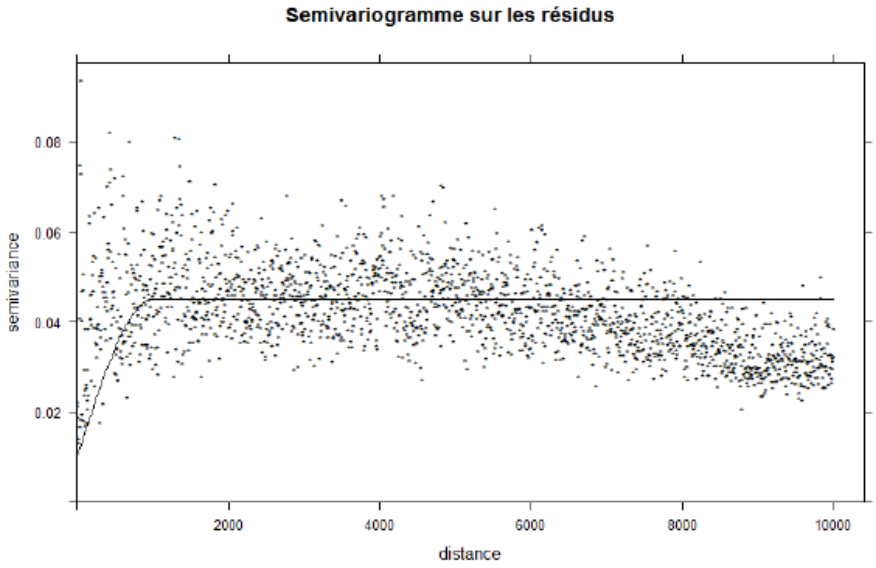


La méthode éditée est d'ores et déjà plus performante mais le signal géographique a-t-il été entièrement capturé ?

Sommaire

Le semivariogramme est un outil géostatistique qui permet d'étudier **l'auto-corrélation spatiale** d'une variable aléatoire.





Semivariogrammes théorique et empirique

Sommaire

Les résidus du modèle sont lissés par des méthodes d'interpolation spatiale dont l'idée repose sur la formule suivante :

Interpolation spatiale

$$\hat{r}(s) = \sum_{i=1}^{n} w_i(s) r(s_i), \qquad avec \qquad \sum_{i=1}^{n} w_i(s) = 1$$

Pondération inverse à la distance

$$w_i(s) = \frac{\frac{1}{d^{\beta}(s,s_i)}}{\sum_{j=0}^{n} \frac{1}{d^{\beta}(s,s_j)}}$$

Krigeage

w_i sont déterminés à l'aide du semivariogramme

Les poids w_i sont obtenus grâce au calcul matriciel suivant

$$W = A^{-1}B,$$

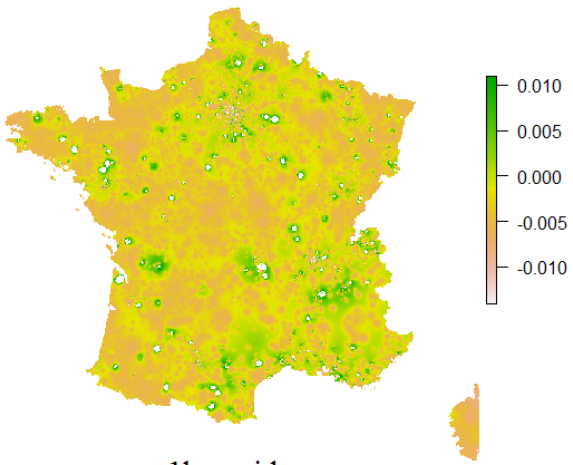
avec

$$W = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \\ \lambda \end{pmatrix}, \quad A = \begin{pmatrix} \gamma(h_{11}) & \cdots & \gamma(h_{1n}) & 1 \\ \gamma(h_{21}) & \cdots & \gamma(h_{2n}) & 1 \\ \cdots & \cdots & \cdots & \cdots \\ \gamma(h_{n1}) & \cdots & \gamma(h_{nn}) & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix} \text{ et } B = \begin{pmatrix} \gamma(h_1) \\ \gamma(h_2) \\ \vdots \\ \gamma(h_n) \\ 1 \end{pmatrix}$$

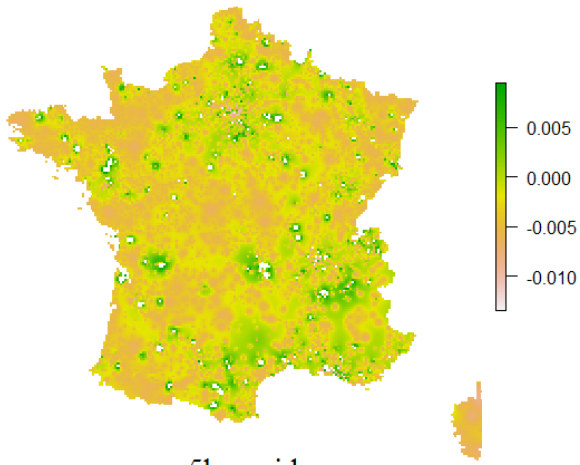
où les $\gamma(h_{ij})$ sont les valeurs du semivariogramme correspondant à la distance h_{ij} entre les points x_i et x_j, et les $\gamma(h_i)$ proviennent du modèle ajusté sur le semivariogramme.

Trouver les meilleurs paramètres pour réaliser le lissage spatial :

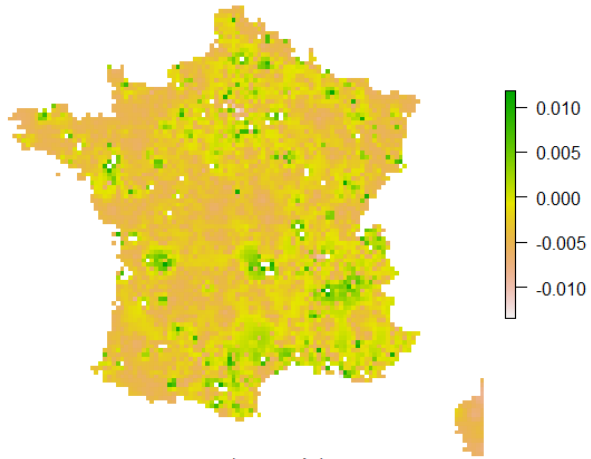
➡ La sensibilité au maillage.



1km grid



5km grid



10km grid

Résidus prédits par l'interpolation spatiale sur chaque carreau. Une taille de carreau de 10km limite le sur-apprentissage.

➡ La sensibilité de la maille d'observation des résidus

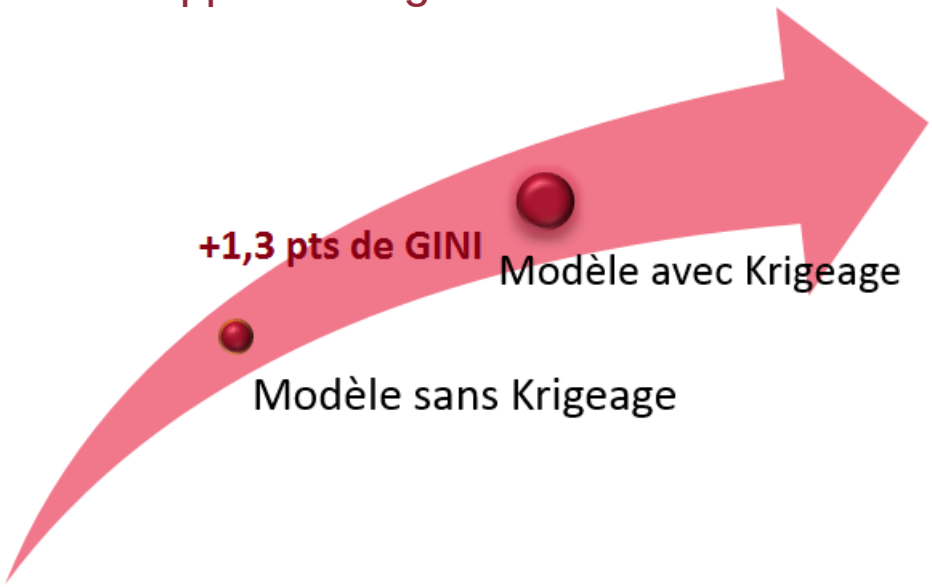
La méthode géostatistique du **Krigeage** :

- **Sophistication** de la méthode de la pondération par l'inverse de la distance.
- Plus efficace car prend en compte la répartition spatiale entre tous les points par l'intermédiaire du **semivariogramme**.
    - Meilleure segmentation du risque
    - Réduction du sur-apprentissage

Profondeur d'historique : 5 ans.

En utilisant les paramètres les plus adaptés pour l'étude :

- Amélioration supplémentaire de la segmentation du risque
- Limitation du sur-apprentissage



Conclusion

- L'intégration directe des données externes dans le modèle pour tirer profit du pouvoir prédictif de la géolocalisation de l'adresse du contrat.
- L'étape supplémentaire du lissage spatial des résidus est requise.

Extensions possibles

- Des carreaux de tailles variables selon la répartition géographique des observations
- Coupler l'étude avec une analyse temporelle pour fournir des actions préventives
- Sur d'autres garanties et d'autres lignes métier comme l'auto

Pour plus de détails

- **Tomislav HENGL**

A Practical Guide to Geostatistical Mapping

- **Sophie BAILLARGEON**

Le krigeage : revue de la théorie et application à l'interpolation spatiale de données de précipitations

Principe de la descente de gradient :

Algorithme d'optimisation ayant pour but de minimiser une fonction différentiable $f(\cdot)$ en procédant par améliorations successives.

Formule d'itération de l'algorithme du gradient

$$x_{k+1} = x_k - \alpha_{k+1} \nabla f(x_k).$$

$\Rightarrow$ Pour le Gradient Tree Boosting : la fonction $f(\cdot)$ correspond à l'espérance de la fonction de perte $L(y, \cdot)$.

Formule d'itération pour le Gradient Boosting

$$F_{k+1}(X) = F_k(X) - \alpha_{k+1} \nabla \mathbb{E}[L(y, F_k(X))]$$

Le modèle additif du Gradient Tree Boosting découle alors :

$$F_{approx}(X) = \underbrace{\underbrace{F_0(X) - \alpha_1 T_1(X)}_{F_1(X)} - \alpha_2 T_2(X)}_{F_2(X)} - \ldots - \alpha_M T_M(X)$$

$$F_M(X)$$

Pour limiter le sur-apprentissage, deux hyper-paramètres :

- **Le shrinkage** : Retarder la vitesse d'apprentissage de l'algorithme. La formule devient :

$$F_{approx}(X) = F_0(X) - \nu\alpha_1 T_1(X) - \nu\alpha_2 T_2(X) - ... - \nu\alpha_M T_M(X)$$

- **Le bagging** : Utiliser à chaque étape un sous-échantillon différent pour réaliser l'arbre.