Université Côte d'Azur

Master M1 Ingénierie Mathématique

Atelier sur l'analyse de données spatialisées via les outils géostatistiques

Rapport de TP Krigeage

soutenu le 22/04/2024

par

Lucie Agnello

Encadrant entreprise : Bertrand Iooss

Chapitre 1

Introduction

Ce rapport de TP présente l'analyse de données spatialisées en utilisant les outils géostatistiques de l'entreprise EDF. L'objectif principal est de réaliser une cartographie des précipitations en Suisse en utilisant la méthode du krigeage, qui est une technique d'interpolation géostatistique couramment utilisée dans les sciences de la terre et de l'environnement.

Contexte

La géostatistique est une branche de la statistique appliquée qui se concentre sur l'analyse et l'interprétation de phénomènes spatialisés. Elle est particulièrement utile pour modéliser des variables continues dans l'espace, comme les précipitations, les concentrations de polluants ou les caractéristiques du sol. Le krigeage, développé par le géologue sud-africain Danie Krige, est une méthode d'interpolation qui permet de prédire les valeurs d'une variable à des emplacements non échantillonnés en utilisant les valeurs observées aux points de mesure environnants. C'est une technique géostatistique de modélisation spatiale permettant, à partir de données dispersées, d'obtenir une représentation homogène des informations étudiées. Dans l'analyse de la pollution atmosphérique, on ne dispose que d'un certain nombre de stations de mesures qui fournissent les données. Le krigeage permet donc, à l'aide des mesures de concentrations obtenues en stations, d'estimer les concentrations hors station. Il devient ainsi possible de créer une carte étendant les relevés à tout l'espace. D'autres techniques géostatistiques permettent de faire ce travail, mais le krigeage a l'avantage de prendre en compte les distances entre les données c'est-à-dire entre les stations de mesure, les distances entre les données et la cible (le point pour lequel on veut estimer la mesure) et la structure spatiale (grâce à l'analyse variographique).

Les méthodes statistiques classiques telles que la régression linéaire se basent sur une hypothèse fondamentale : l'indépendance des variables. Or, lorsqu'une variable est spatialement autocorrélée, cette hypothèse n'est plus vérifiée. Ainsi, le krigeage se base sur cette nouvelle hypothèse : l'autocorrélation spatiale des données. Concrètement, cela signifie que deux données rapprochées dans l'espace tendent à posséder des caractéristiques similaires.

Objectifs

Les principaux objectifs de ce TP seront la **Visualisation des données** en essayant de comprendre la distribution spatiale des stations de mesure de la pluie en Suisse en utilisant les graphes et les statistiques de R,l'**Analyse variographique** en étudiant la structure spatiale des données de précipitations à l'aide de variogrammes et déterminer si les données présentent une anisotropie, l'**Ajustement d'un modèle de variogramme** en ajuster des modèles de variogramme aux données pour capturer la variabilité spatiale et définir les paramètres nécessaires pour le krigeage, **Utiliser le krigeage** pour interpoler les valeurs de précipitations sur une grille régulière couvrant la Suisse et évaluer la précision des prédictions. Enfin la **Validation du modèle** en comparant les prédictions obtenues par krigeage aux observations supplémentaires pour évaluer la performance du modèle et explorer différentes méthodes pour sélectionner de manière optimale les stations de mesure afin de maximiser l'efficacité et la représentativité des mesures.

Chapitre 2

Visualisation des données

trois bases de données présentes dans le package geoR :

1. sic.100 : Il s'agit d'un échantillon de 100 observations qui serviront à effectuer les interpolations. Cette base de données est utilisée comme base d'apprentissage pour développer un modèle de prédiction des précipitations. Elle contient des données enregistrées par 100 stations météorologiques sur la quantité de pluie.
2. sic.367 : Cette base de données contient des observations qui ne sont pas incluses dans l'échantillon de 100. Elle est utilisée pour comparer les estimations fournies par le modèle de prédiction avec les observations réelles. Comme pour sic.100, il s'agit également d'une base d'apprentissage, mais elle contient 367 stations météorologiques.
3. sic.all : Cette base de données comprend l'ensemble des observations disponibles. Elle contient les données de toutes les stations météorologiques, sans aucune exclusion ou sous-échantillonnage. Elle peut être utilisée pour différentes analyses et comparaisons globales.

Nous examinons les graphiques produits à l'aide des fonctions points() et plot.geodata().

Les points verts représentent les 100 stations d'apprentissage, tandis que les points rouges représentent les 367 stations supplémentaires.

Nous pouvons remarquer ici que les points présentent une certaine organisation, comme une tendance générale ou des clusters. Nous pouvons observer des regroupements de points dans certaines zones (Nord-Ouest) et une dispersion particulière non homogène. Les observations présentent donc des schémas ou des structures spatiales identifiables. Il existe une structuration spatiale des données.

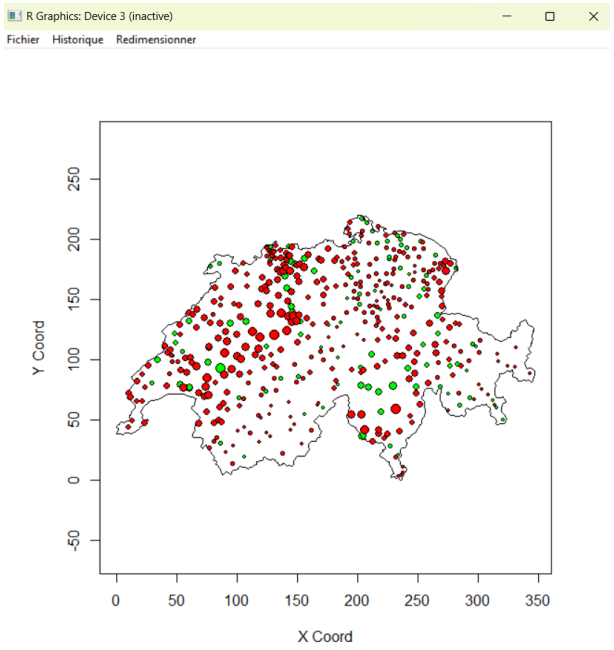


Figure 2.1 – Distribution spatiale des stations de mesure de la pluie en Suisse

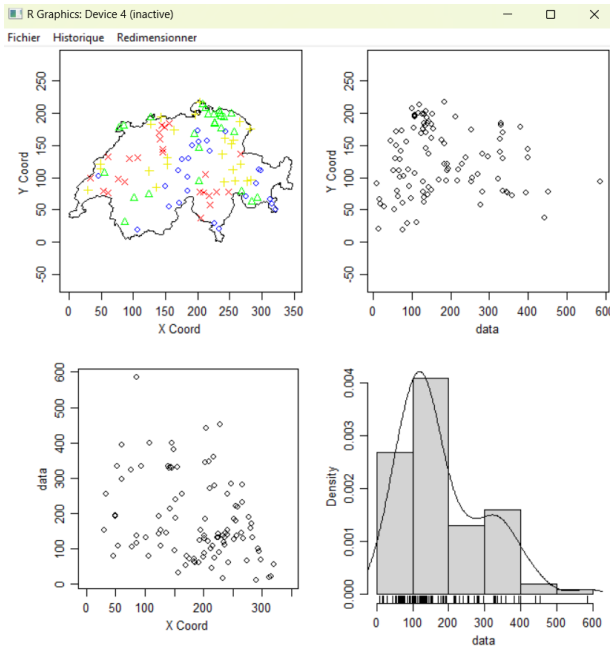


Figure 2.2 – Statistique des données spatiales

Chapitre 3

Analyse variographique

L'analyse variographique permet d'explorer et de comprendre la structure spatiale des données en calculant et en visualisant les variogrammes. Un variogramme montre la variation de la similarité entre les valeurs mesurées en fonction de la distance qui les sépare.

3.1 Calcul des variogrammes empiriques

Le variogramme empirique est calculé en fonction de la distance h entre les points de données. Lorsque h est nul, le variogramme est égal à 0. Lorsque h est petit, le variogramme est faible, ce qui indique une forte corrélation entre les points proches. À mesure que h augmente, le variogramme augmente également jusqu'à atteindre un plateau, indiquant l'absence de corrélation spatiale au-delà d'une certaine distance.Si h tend vers l'infini, la corrélation tend vers 0. On a donc le variogramme qui tend vers la variance.

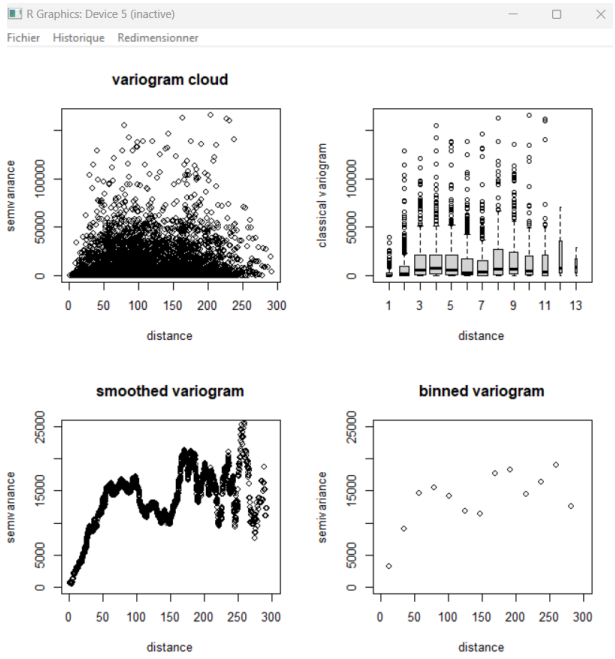


Figure 3.1 – Variogrammes

1. Variogram Cloud :

   Le variogram cloud montre la nuée variographique, illustrant la relation entre les distances séparant les points et leurs différences de valeurs. Plus la distance augmente, plus la valeur du variogramme tend à augmenter, indiquant une décroissance de la corrélation spatiale.

2. Clouds for Binned Variogram :

   Le binned variogram présente des boxplots des valeurs de variogrammes pour différentes classes de distance. Les lignes médianes des boxplots montrent une tendance à l'augmentation avec la distance, reflétant la décroissance de la corrélation.

3. Smoothed Variogram :

   Le variogramme d'une estimation non parametrique de la densite en utilisant lissage de la gaussienne montre la variation moyenne du variogramme en fonction de la distance, avec un lissage appliqué pour réduire les fluctuations aléatoires. Le variogramme tend vers un plateau indiquant la portée au-delà de laquelle les valeurs ne sont plus corrélées. On peut voir que la distance de séparation augmente le variogramme augmente.Elle augmente jusqu'à un certain seuil et puis cela devient constant les valeurs ne sont alors plus siognificative

4. Variogramme experimental :

Le binned variogram discretise les distances et calcule la moyenne des différences pour chaque classe de distance. Il aide à identifier la portée et les structures de corrélation spatiale dans les données. En effet, nous avons $bin = classepetitspoints = Z(X + h) - Z(h)$

3.2 Variogramme directionnel

Ces graphiques permettent d'observer les variations de la structure spatiale dans différentes directions. Une isotropie se manifesterait par des variogrammes similaires dans toutes les directions, tandis qu'une anisotropie serait indiquée par des variations distinctes dans les directions analysées. Par exemple, une plus grande portée dans une direction suggère une plus grande corrélation spatiale dans cette direction.

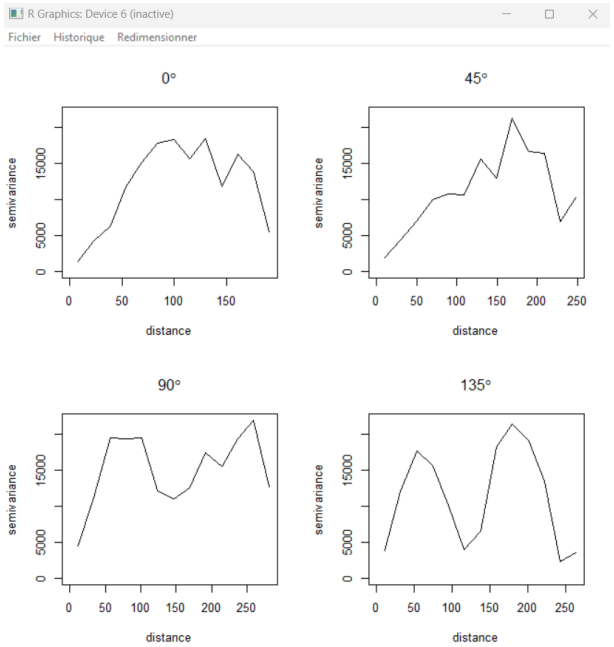


Figure 3.2 – Variogramme directionnel

On remarque ici que la pluie serait plus important en longitude qu'en largitude.

La direction horizontale, c'est-à-dire vario0°, évalue la distance horizontalement pour chaque point. La vario90° étudie les distances verticalement.

Le phenomene est plus corrélée dans la direction 45° que 135° l'angle de portée la plus forte est celle de vario45° car la vitesse de la portée et du début du variogramme est la plsu importante Alors que pour le vario135°, l'analyse n'est pas très stable. On passe d'un creu à une bande de valeurs forte avec ensuite aucune corrélation et enfin de nouveau une bande de valeurs très

fortes.

Les variogrammes montrent donc une structuration spatiale de la quantité de pluie. La tendance générale des variogrammes à augmenter avec la distance jusqu'à atteindre un plateau suggère que les valeurs de pluie sont spatialement autocorrélées. La présence de structures différentes dans les variogrammes directionnels peut indiquer une anisotropie, suggérant que la dépendance spatiale varie selon la direction.

Chapitre 4

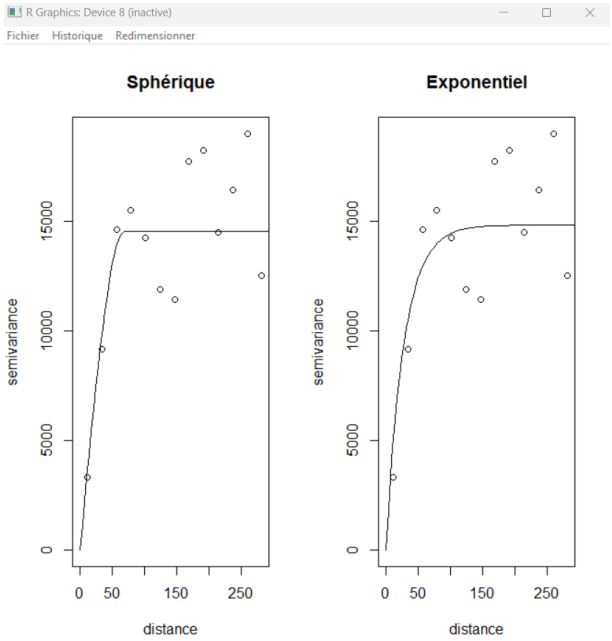
Ajustement d'un modèle de variogramme

L'ajustement d'un modèle de variogramme est essentiel pour la modélisation des données spatiales. Les modèles paramétriques courants incluent les modèles gaussien, exponentiel et sphérique. Ces modèles permettent d'estimer la dépendance spatiale à partir des variogrammes empiriques et de faire des prédictions plus précises avec des techniques comme le krigeage. Soit a la portée, si a est faible alors plus la pente est ardue. Vice-versa, si a est élevé alors la pente est faible. Lorsque la distance est plus grande ou égale à a, il y a aucune corrélation et si la distance est strictement plus petite, il y a corrélation.

Pour ajuster un modèle, nous devons d'abord calculer le variogramme empirique des données. Ensuite, on ajuste deux modèles de variogrammes : le modèle sphérique et le modèle exponentiel. Les paramètres initiaux choisis sont la semi-variance (15000) et la portée (100).Pour vérifier si un meilleur ajustement peut être obtenu, on teste d'autres valeurs de paramètres manuellement.

Les modèles sphérique et exponentiel ajustés initialement montrent une bonne adéquation avec les données empiriques. Le modèle sphérique semble particulièrement bien ajusté pour les premières distances, tandis que le modèle exponentiel montre une bonne adéquation globale.

La méthode utilisée pour ajuster les modèles est une descente de gradient qui optimise la somme des résidus au carré entre la courbe ajustée et les points du variogramme empirique. La pondération de certains points permet de lisser les données, en mettant plus de poids sur



les points qui sont plus fiables.

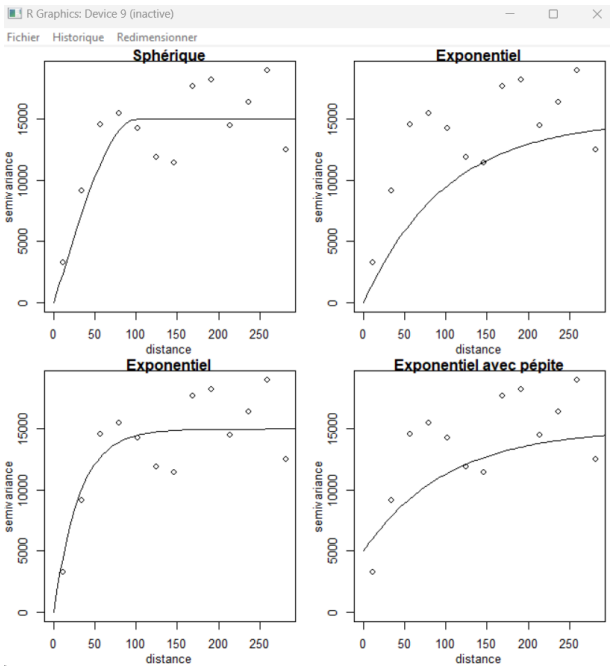
Les paramètres utilisés sont :

phi (a) : la portée, indiquant la distance à laquelle la corrélation devient négligeable.
sigma2 : la semi-variance au plateau.
tausq : l'erreur de mesure ou l'effet de pépite, qui relâche l'hypothèse de non fluctuation à distance zéro.

Le modèle sphérique tend à mieux s'ajuster aux premiers points du variogramme empirique, ce qui suggère qu'il est plus adapté pour capturer les corrélations à courte distance.
En ajustant manuellement les paramètres, nous pouvons observer que :

Une modification de la portée dans le modèle exponentiel montre que la portée (30) réduit la pente, suggérant une corrélation plus rapide. L'ajout d'un effet de pépite (nugget effect) dans le modèle exponentiel montre des variations dues à des erreurs de mesure mais permet d'affiner la précision du modèle.

Pour une approche alternative, nous pouvons utiliser likfit pour ajuster les modèles via une méthode de maximum de vraisemblance. Cette méthode peut parfois offrir des ajustements plus précis en utilisant des techniques de maximum de vraisemblance pour estimer les paramètres du modèle.

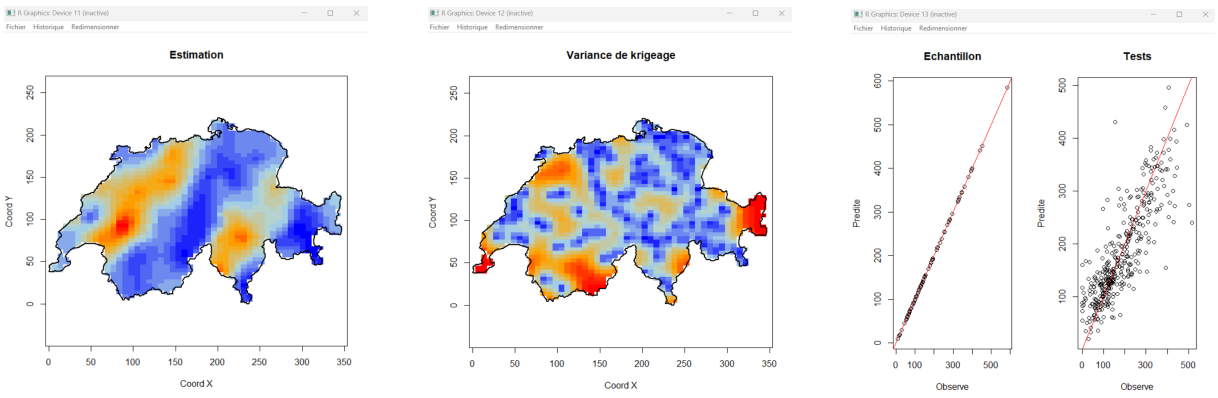

R Graphics: Device 9 (inactive)
Fichier Historique Redimensionner
Sphérique
Exponentiel
Exponentiel
Exponentiel avec pépite
semivariance
distance
0
5000
10000
15000
0 50 100 150 200 250

Chapitre 5

Krigeage

Le krigeage est une méthode d'interpolation spatiale permettant de prédire les valeurs d'une variable d'intérêt à des emplacements non observés en utilisant des modèles de variogramme. Nous utilisons ici le modèle sphérique sans effet de pépite, avec une variance de 15000 et une portée de 50 pour réaliser nos prédictions. Les étpes de cette méthode sont le calcul du variogramme empirique et ajustement du modèle sphérique, la création de la grille de prédiction et configuration du krigeage, l'exécution du krigeage et la validation du modèle.

5.1 Prédiction et Variance du Krigeage



krige.conv : Cette fonction réalise le krigeage en utilisant les données d'observation (sic.100) et les coordonnées des points de prédiction (pred.grid). Le modèle de variogramme utilisé est le modèle sphérique avec une variance de 15000 et une portée de 50.
La fonction image est utilisée pour visualiser les résultats du krigeage. Le premier appel à image affiche les valeurs prédites, tandis que le second affiche la variance de krigeage. Les noms des

attributs de l'objet kc peuvent être obtenus en utilisant names(kc). L'attribut krige.var représente la variance de krigeage, qui est une mesure de l'erreur de prédiction ou de la dispersion possible des prédictions.
L'estimation montre une direction préférentielle de 45 degrés, et la variance de krigeage est plus forte là où il n'y a pas d'observation, car il y a moins d'informations disponibles pour faire des prédictions précises.

La carte de prédiction obtenue montre une distribution des valeurs prédites de pluviométrie sur l'ensemble du territoire suisse. Nous observons que les valeurs sont cohérentes avec les observations d'entraînement.
La carte de la variance de krigeage montre que l'erreur de prédiction est plus élevée dans les zones où il y a peu ou pas d'observations. Cela est attendu car la précision de la prédiction diminue lorsque la distance par rapport aux points d'observation augmente.

5.2 Validation du modèle

Les graphes de comparaison entre les valeurs observées et prédites pour les échantillons d'apprentissage et de test montrent que les prédictions sont plutôt bonnes, mais il existe une certaine dispersion, particulièrement pour l'échantillon de test. Ceci indique que le modèle fonctionne bien en moyenne, mais peut avoir des difficultés avec les valeurs extrêmes.
L'erreur quadratique moyenne calculée pour l'échantillon de test est de 3175.393, indiquant une certaine variabilité entre les valeurs observées et prédites.
Le coefficient de prédictivité Q^2 est une mesure de la qualité du modèle. Un Q^2 proche de 1 indique un modèle performant. Le Q^2 obtenu montre que le modèle explique une part significative de la variance des observations, mais peut être amélioré.

L'intégration d'informations variographiques plus riches, comme des effets anisotropes ou des effets de pépite, pourrait améliorer la précision des prédictions. Des modèles mixtes ou des approches intégrant des variables supplémentaires comme l'altitude pourraient également fournir des informations précieuses pour affiner les prédictions. De plus, on pourrait ajouter un effet de tendance pour permettre d'améliorer les prédictions.
En conclusion, le modèle de krigeage utilisé montre une bonne performance globale, mais des ajustements et des intégrations supplémentaires pourraient encore améliorer sa précision,

notamment pour les valeurs extrêmes et dans les zones avec peu d'observations.

Chapitre 6

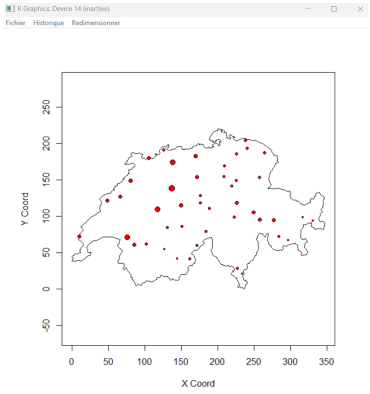
Planification d'expériences

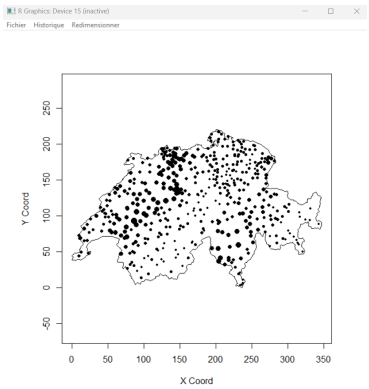
Face à une crise financière majeure, la Suisse doit réduire le nombre de ses stations pluviométriques de 467 à 45. L'objectif est de sélectionner ces 45 stations de manière optimale pour minimiser l'erreur de test dans les prédictions de krigeage. Voici trois méthodes pour y parvenir :

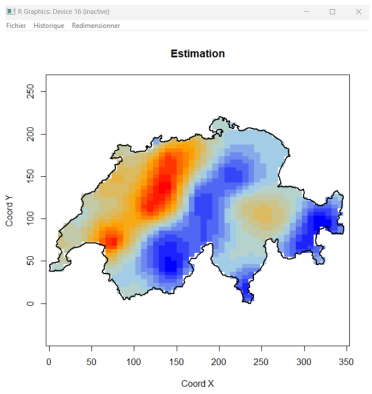
6.1 Méthode a) Sélection aléatoire avec maximisation de la distance minimale

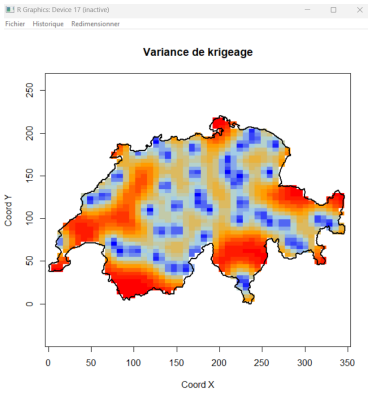
Cette méthode utilise une approche Monte Carlo pour sélectionner aléatoirement des ensembles de 45 stations et choisir celui avec la meilleure répartition spatiale. La fonction mindist() est utilisée pour calculer la distance minimale entre les stations sélectionnées.

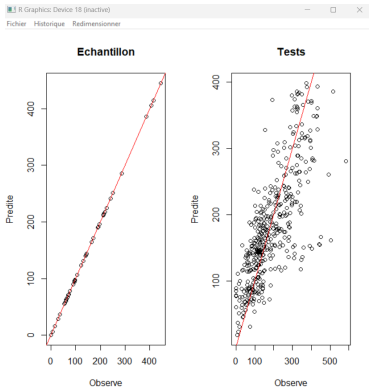
En utilisant cette méthode, nous avons obtenu une erreur quadratique moyenne (MSE) de 6672.168 et un coefficient de détermination croisé (Q2) de 0.457. Bien que cette méthode ait généré une distribution spatiale des stations relativement équilibrée, les performances de prédiction ne sont pas optimales.







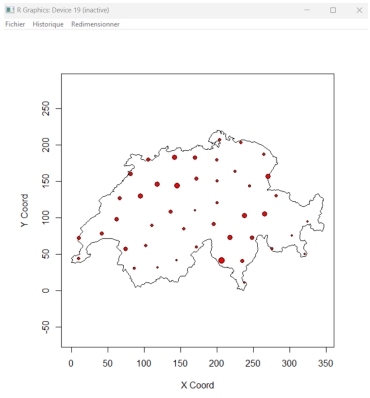


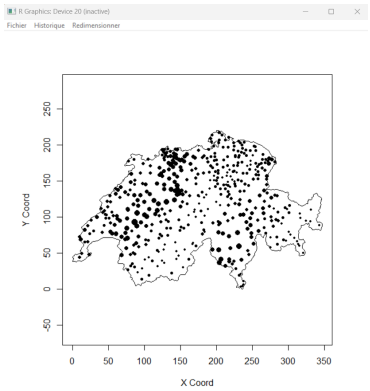


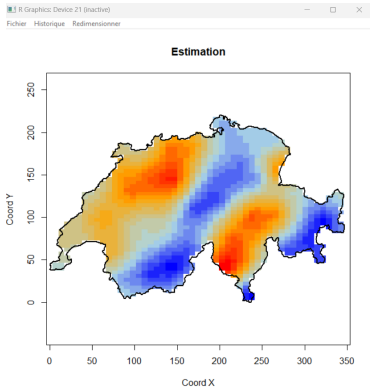
6.2 Méthode b) Algorithme de suppression des points

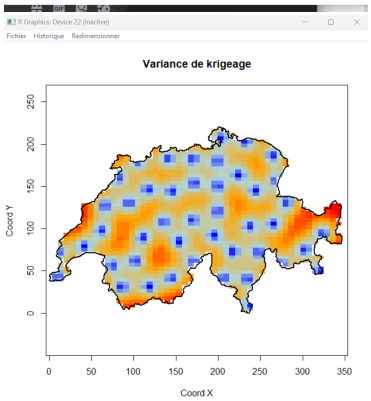
Cette méthode utilise la fonction wspDesign() pour sélectionner des stations en excluant celles dans un cercle de rayon prédéfini autour de chaque station sélectionnée. Le rayon est ajusté pour obtenir exactement 45 stations.

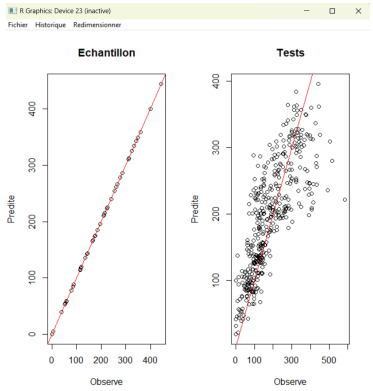
La sélection des 45 stations avec cet algorithme a produit des résultats légèrement meilleurs, avec un MSE de 4811.87 et un Q2 de 0.617. Cependant, le choix du rayon initial pour exclure les stations peut influencer considérablement les résultats et nécessite une optimisation plus poussée.







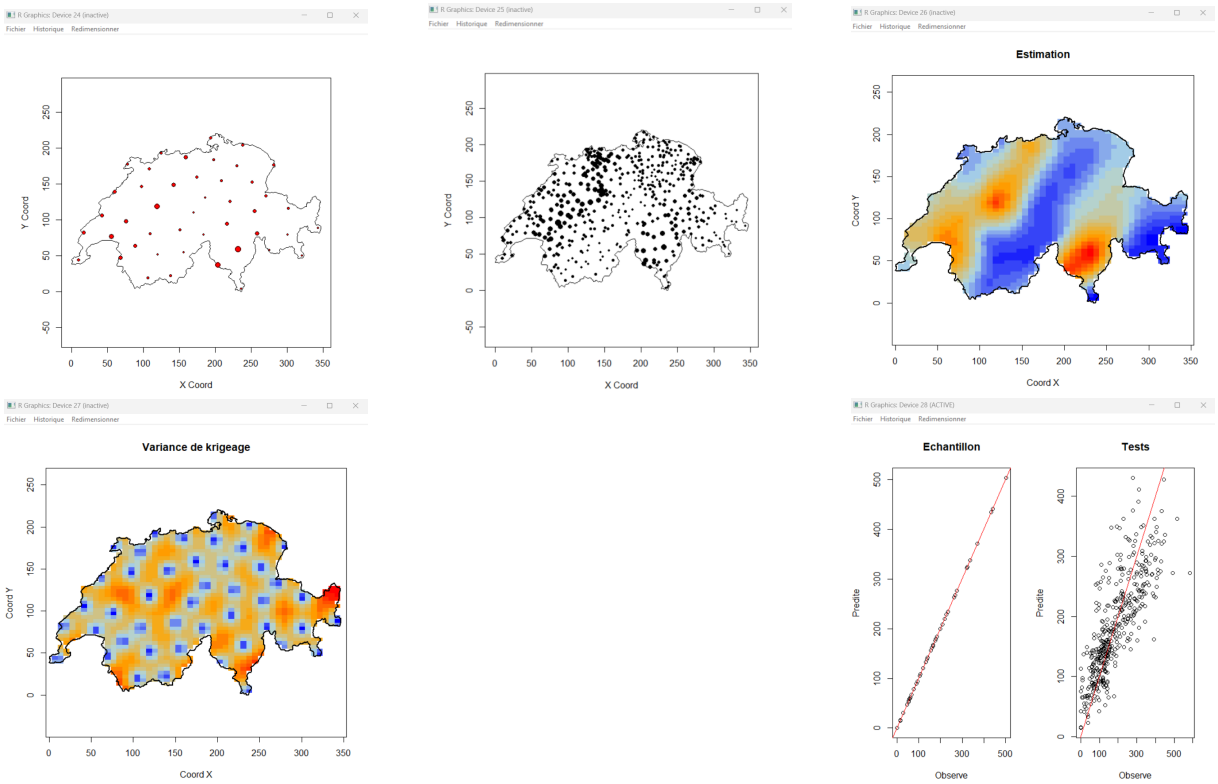




6.3 Méthode c) Algorithme "coffee-house design"

Cettte methode repose sur la méthode des k plus proche voisins. Son nom n'est pas laissé au hasard car en effet il illustre bien la logique de cette méthode : un client choisit sa table le deuxieme sinstalle le plus loin possible. En principe, on tire au hasard la premiere station et on choisit la station la plus lointaine et on construit l'algorithme récursivement. Cette méthode sélectionne donc séquentiellement des stations en maximisant la distance à la station sélectionnée la plus proche.

Cette méthode a donné les meilleurs résultats avec un MSE de 4303.575 et un Q2 de 0.653. En sélectionnant séquentiellement les stations les plus éloignées les unes des autres, cette approche a permis d'obtenir une distribution spatiale optimale des stations pour la prédiction par krigeage.



Pour conclure, nous pouvons identifié plusieurs aspects problématiques :

Prise en compte des frontières et de l'altitude : Les distances entre les stations ne prennent pas en compte les frontières ni les variations d'altitude, ce qui peut affecter la précision des prédictions, notamment dans les régions montagneuses. Ces facteurs peuvent avoir un impact significatif sur les schémas de précipitations et devraient être pris en compte pour améliorer la précision des prédictions.

La taille de l'échantillon de 45 stations pourrait ne pas être suffisante pour construire un variogramme précis, en particulier dans les régions où les précipitations varient considérablement.

La sélection des stations n'a pas pris en compte **la distribution des quantités de pluie**, ce qui peut entraîner des biais dans les prédictions, en particulier dans les zones où les précipitations sont très variables.

Ainsi, malgré les performances relativement bonnes de la méthode c), des améliorations peuvent être apportées en tenant compte de ces aspects lors de la sélection des stations pluviométriques.