MASTER EN SCIENCES MATHÉMATIQUES À FINALITÉ SPÉCIALISÉE EN PERSPECTIVES PROFESSIONNELLES DES MATHÉMATIQUES APPLIQUÉES

ANALYSE DE DONNÉES SPATIALES

DELVOSAL, Marine

Award date:
2022

Awarding institution:
Universite de Namur

UNIVERSITE DE NAMUR

Faculté des Sciences

ANALYSE DE DONNÉES SPATIALES

Promoteur : Germain VAN BEVER

Mémoire présenté pour l'obtention du grade académique
de master en sciences mathématiques à finalité spécialisée en Project Engineering

Marine DELVOSAL

Août 2022

Remerciements

Mon premier remerciement est pour Monsieur Germain VAN BEVER. D'abord, en qualité de promoteur, je le remercie pour son approche pédagogique, ses précieux conseils, sa structuration d'esprit et le recul qu'il a inspirés tant sur la matière que sur la vision stratégique dans la rédaction de mon mémoire. Ensuite, je le remercie pour ses qualités humaines, pour son soutien, ses encouragements, ses précieux conseils et pour sa patience qu'il m'a témoignés tout au long du projet.

Mon deuxième remerciement est pour l'ensemble du département de mathématique. Ma vocation a toujours été les mathématiques et ce rêve est devenu réalité grâce au soutien du département, grâce à la bienveillance du personnel de l'UNamur, grâce à une majorité de professeurs passionnés et conscients de leurs rôles dans la destinée des étudiants et enfin grâce aux assistants qui assurent harmonieusement le lien entre les jeunes universitaires et les professeurs.

Mon troisième remerciement est pour mes amis. Si le travail de ma compétence intellectuelle est l'oeuvre du département, l'évolution de mon intelligence émotionnelle est une des conséquences bénéfiques des relations sociales entretenues d'une part avec les compagnons d'université, comme Sarah, Chloé, Céline, Gladys, Lara, Laura, Gaëtan, Célia, Margaux et Bertrand et d'autre part, mes compagnons d'humanité comme Manon, Sophie, Nicolas et Tony. Une mention spéciale à Sarah qui m'a poussé, soutenu et aidé à chaque moment de ma vie estudiantine et lors de mon mémoire.

Mon quatrième remerciement est un remerciement du cœur. Quel que soit le style choisi ou la prose utilisée, je n'arriverai pas à exprimer l'importance de leurs rôles dans la rédaction de ce mémoire, dans ma réussite universitaire ou dans la création de mon identité individuelle. Je vais, donc, juste les citer : mes grands-parents, René et Claudine ; mes parents, Ives et Cathy ; ma sœur, Romane.

Résumé

Dans le cadre des variables spatiales, la caractéristique de localisation est cruciale. La première approche à réaliser avec ce type de données est l'étude de la relation entre la valeur et la localisation de la variable spatiale. Afin d'y parvenir nous formalisons les positions de la variable spatiale par une matrice de poids ou de voisinage et nous utilisons l'indice de Moran dans le but de détecter une possible autocorrélation spatiale. Les variables spatiales se décomposent en trois grandes classifications. La seconde approche est l'utilisation des méthodes propre à chaque catégorie. Dans un espace géographique définit, nous pouvons étudier la répartition des localisations et établir la présence d'une interaction entre les positions. Une autre démarche, réalisée grâce à la méthode du Kriging, est d'interpoler la valeur de la variable spatiale sur un nouvel emplacement géographique. La dernière méthode est l'utilisation de modèles de régression, adaptés dans le cadre spatiale, dans le but d'étudier l'influence de certains facteurs sur la variable spatiale.

Mots-clefs : données ponctuelles, données continues, données surfaciques, Ripley, Moran, Kriging, variogramme, interaction spatiale, maximum de vraisemblance, Lagrange

Abstract

In the context of spatial variables, the location feature is crucial. The first approach to be adopted with this type of data is to study the relationship between the value and the location of the spatial variable. In order to get this, we formalize the positions of the spatial variable by a weight or neighbourhood matrix and use the Moran index in order to detect a possible spatial autocorrelation. The spatial variables can be broken down into three broad classifications. The second approach is the use of methods specific to each category. In a defined geographical space, we can study the distribution of locations and establish the presence of an interaction between the positions. Another approach, using the Kriging method, is to interpolate the value of the spatial variable to a new geographical location. The last method is the use of regression models, adapted in the spatial environment, in order to study the influence of certain factors on the spatial variable.

Keywords : point pattern data, continuous data, surface area data, Ripley, Moran, Kriging, variogram, spatial interaction, maximum likelihood, Lagrange

Table des matières

Introduction

Dans le cadre de ce mémoire, nous nous concentrons sur les données dites spatiales où le support de ces dernières est une zone géographique. Les données spatiales prennent une place importante dans le monde actuel. Ces données apportent une compréhension supplémentaire à un problème posé. Par exemple, en météorologie, la dimension spatiale est primordiale afin de comprendre certains phénomènes naturels. Entre autres les entreprises, comme Amazon, utilisent ce type de données afin d'établir un suivi de leur livraison.

L'approche descriptive la plus courante est une carte géographique où nous observons la répartition ou l'amplitude des données sur celle-ci. Le système de coordonnées que nous utilisons est le format numérique décimal de la longitude et de la latitude. Dans le but de repérer une zone géographique, nous utilisons les fichiers de type *shapefile* qui se trouvent sur le site *GADM* [18] et qui détiennent les différentes niveaux de délimitation d'un pays. Sur base du livre de référence [11], les limitations géographiques d'un fichier *shapefile* sont appréhendées en *R*. Précisons que la version du logiciel est 4.1.3.

Nous nous concentrons ici sur l'approche inférentielle. L'objectif est donc d'introduire les méthodes qui permettent d'analyser les variables spatiales. En particulier, il existe plusieurs types de variables spatiales : ponctuelles, continues et surfaciques. Les méthodes que nous utilisons varient en fonction du type des variables spatiales que nous considérons. En effet, certains outils sont plus adéquats que d'autres. De plus, nous explicitons certains phénomènes qui apparaissent lors de l'utilisation de variables spatiales comme l'hétérogénéité et l'autocorrélation spatiale.

Pour chaque type de données, nous explicitons de manière théorique le choix des approches choisies ainsi que leurs formalismes en *R*. En particulier, nous appliquons les méthodes dédiées aux données surfaciques sur un certain jeux de données lié à la criminalité. L'analyse est fondée sur base des modèles de régression spatiale.

La structure de ce mémoire est composée de quatre chapitres. Le premier chapitre introduit les caractéristiques des données spatiales. Ensuite, les trois derniers chapitres sont consacrés à l'analyse de chacun des types de données spatiales.

Un rappel sur la régression classique qui est effectuée en annexe A. Par la suite, nous explicitons les adaptations de ce modèle dans le cadre spatial. En effet, l'article [8] met en évidence que l'application des modèles de régression linéaire classique qui n'est pas toujours envisageable car celle-ci fournira alors des estimateurs biaisés des paramètres.

Chapitre 1

Variables spatiales

Dans le cadre de la régression linéaire classique, nous sommes face à des variables de type quantitatives pour les régresseurs. Dans le cadre de ce mémoire, toutefois nous nous intéressons à un type de variables qui se nomme "spatiales". Nous parcourons dans ce chapitre les particularités et les propriétés de ce genre de variables en se basant sur [21]. Afin de comprendre la différence entre les variables quantitatives et spatiales, il est primordial de les définir toutes les deux grâce aux sources [37] et [21].

Définition 1.1
Une variable quantitative X est à valeurs dans un sous-ensemble (non-dénombrable) de $\mathbb{R}$.

Définition 1.2
Une variable spatiale $S = (X, P)$ est définie par deux composantes qui sont la valeur de l'observation X et sa localisation P.

Nous observons ici que l'information qui diffère entre les deux types de variables est la localisation. Nous pourrons constater par la suite que cette information est cruciale. En effet, elle joue un rôle important lorsque nous analysons ce type de données. Notons que les variables spatiales peuvent être classé en trois catégories différentes par la source [36] :

— données ponctuelles détaillées dans le chapitre 2 ;

— données continues détaillées dans le chapitre 3 ;

— données surfaciques détaillées dans le chapitre 4.

Le premier type de variables spatiales, que nous nommons données ponctuelles, étudie la répartition des localisations P de la variable dans l'espace. Un exemple de données ponctuelles serait les lieux d'apparition des incendies lors des périodes de sécheresse. Le but de l'analyse spatiale avec ce type de variables est de comparer la distribution des variables avec une autre distribution où les variables seraient dispersées de manière aléatoire. La comparaison se réalise en termes de distance.

Le deuxième type de variables spatiales est les variables continues, elles sont aussi appelées variables géostatistiques. Nous nous concentrons sur la valeur X de la variable. En effet, nous observons la valeur X de la variable spatiale en fonction de la localisation P dans une zone géographique considérée. Notons que la valeur de la variable est conçue de manière continue, c'est-à-dire, qu'elle varie de manière graduelle dans l'espace. Par exemple, si nous mesurons la pluviométrie dans toutes les provinces de la Belgique, nous sommes face à une variable continue. L'aspiration avec ce type de données permet de

pouvoir déterminer la valeur de la variable spatiale S pour une nouvelle localisation.

Le dernier type de variable spatiale que nous explorons est les données surfaciques. La localisation P et la valeur X de la variable S sont toutes deux prises en considération. En effet, nous nous focalisons sur les relations entre les différentes localisations de P se situant dans une zone territoriale définie. Nous déterminons l'existence ou non d'une variation de la valeur X lorsque les localisations de P sont voisines. Un exemple est l'influence de l'âge des habitants sur le nombre de décès dans une commune. L'analyse avec ce type de données se réalise en trois étapes :

— déterminer l'agencement des localisations de P ;
— définir l'attraction entre les valeurs de X ;
— établir si cette influence est révélatrice.

Nous observons que les trois catégories ne sont pas strictes. En effet, une variable spatiale peut faire partie de plusieurs catégories en fonction du contexte dans lequel nous nous situons. Par exemple, prenons le nombre de nouvelles constructions en Belgique. Ici, la maison peut être une donnée ponctuelle mais aussi une donnée surfacique si nous analysons l'influence du revenu moyen.

Lors de l'utilisation de données spatiales, nous sommes confrontés à deux phénomènes qui sont l'hétérogénéité spatiale et l'autocorrélation spatiale. L'hétérogénéité spatiale se traduit comme le comportement ainsi que la variation de la variable spatiale dans l'espace. Dans un premier temps, nous définissons les relations entre les localisations de la variable spatiale. Ensuite, nous passons à l'autocorrélation spatiale et aux outils qui permettent de la mesurer.

1.1 Les relations spatiales

Pour un espace considéré, nous définissons les relations entre les localisations d'une variable spatiale grâce à la source [23]. La notion de distance entre ces localisations influence la valeur X de la variable spatiale. C'est pourquoi nous définissons les *relations spatiales*.

Lors de l'étude d'une zone géographique, nous scindons cet espace en différentes zones où nous définissons le centroïde de chaque zone comme nous le faisons avec les provinces de la Belgique à la Figure 1.1. L'ensemble des centroïdes correspond à la composante localisation P d'une variable spatiale tel que P contient 11 observations. De plus, les coordonnées géographiques de chaque centroïde sont reprises à la Table 1.1. Ensuite, nous définissons les relations de voisinage par, soit la matrice de distance, soit la matrice de voisinage. Ces matrices sont carrées de taille n où n est le nombre d'observations de la variable spatiale S. Notons que la notion de matrice de voisinage est très utilisée dans le cadre des données surfaciques tandis que la matrice de distance est plus utilisée dans le cadre des données ponctuelles et continues. La matrice de distance mesure la dissimilarité entre les différentes zones. Considérons deux observations de la variable spatiale S, (x_i, p_i) et (x_j, p_j), nous pouvons définir la distance qui sépare les localisations p_i et p_j de différentes manières. Nous introduisons la distance euclidienne entre p_i et p_j comme

$$d_{ij} = d_{ji} = \sqrt{(m_i - m_j)^2 + (q_i - q_j)^2}.$$

où m et q sont les coordonnées géographiques, c'est-à-dire, la longitude et la latitude de la localisation p.

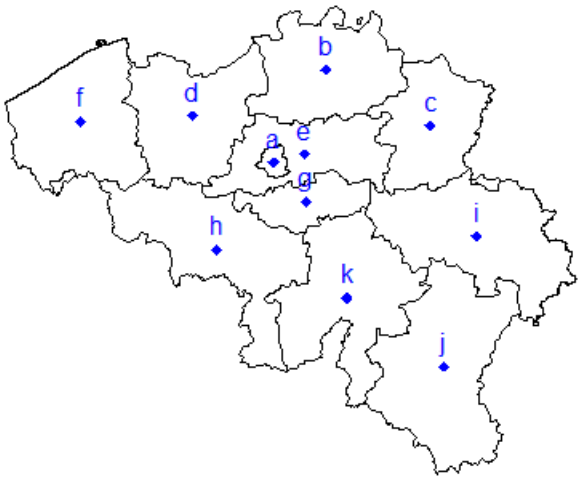

FIGURE 1.1 – Carte des centroïdes des provinces de la Belgique.

TABLE 1.1 – Les coordonnées des centroïdes de la Belgique

	m	q
a	4.37	50.84
b	4.72	51.23
c	5.43	50.99
d	3.82	51.04
e	4.57	50.87
f	3.06	51.01
g	4.59	50.66
h	3.98	50.46
i	5.74	50.52
j	5.52	49.96
k	4.86	50.25

La matrice de distance est construite à partir des coordonnées des centroïdes reprises à la Table 1.1 tel que pour $i, j = 1, \ldots, n$, nous avons

$$W(i,j) = \begin{cases} d_{ij} & \text{si } i \neq j, \\ 0 & \text{sinon.} \end{cases}$$

La matrice de voisinage, aussi appelée matrice de poids, détermine la proximité entre les différentes zones géographiques. Cette matrice dépend du *graphe de voisinage* qui met en évidence les relations entre les localisations. De nombreuses méthodes sont utilisées afin de définir ce graphe comme la triangularisation de Delauney ou la méthode des k plus proches voisins. Les termes W_{ij} de la matrice, où $i \neq j$, auront une valeur toute aussi importante que l'ampleur de l'effet de la localisation p_j sur la localisation p_i. Un exemple courant sont les matrices de contiguïté. Ce genre de matrice détermine la présence ou non d'une relation entre deux localisations. Notons qu'une localisation ne peut être en relation avec elle-même alors nous insérons la convention

$$W(i,i) = 0 \quad \forall i = 1, \ldots, n.$$

Afin de connaître le nombre de localisations qui sont contiguës à une localisation p_i en particulier, nous prenons simplement la somme des éléments de la ligne i de la matrice comme suit

$$W(i,:) = \sum_{j=1}^{n} W(i,j).$$

Ainsi pour obtenir le nombre total de liens entre toutes les localisations, nous sommons simplement toutes les localisations $W(i,:)$ où i est l'indice d'une localisation en particulier. Il ne faut pas oublié de diviser par deux le résultat car une symétrie intervient. Si une localisation p_i a un lien avec une localisation p_j alors l'inverse est vrai. Cette somme peut alors s'écrire

$$W = \frac{1}{2} \sum_{i=1}^{n} W(i,:).$$

En appliquant la matrice de poids W à une variable X, nous formalisons la caractéristique de la localisation sur les valeurs de la variable X. Cette relation précise l'influence que peut avoir une localisation sur une autre. On dit alors que la variable WX est spatialement décalée. Dans le cadre de notre exemple, nous prenons la matrice de contiguïté d'ordre 1 qui est de la forme

$$W(i,j) = \begin{cases} 1 & \text{si les localisations } p_i \text{ et } p_j \text{ sont liées,} \\ 0 & \text{sinon.} \end{cases}$$

Par la triangularisation de Delaunay grâce à la fonction *tri2nb* en R, le graphe de voisinage de la Figure 1.2 aboutit à la matrice de voisinage présente à la Table 1.2.

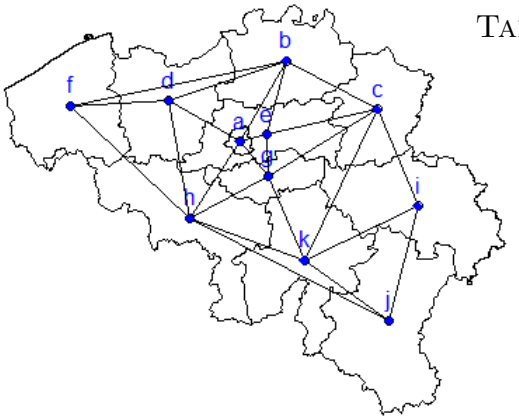

FIGURE 1.2 – Carte des provinces de la Belgique où les relations sont obtenues par la triangularisation de Delaunay.

TABLE 1.2 – La matrice de voisinage de la Belgique.

	a	b	c	d	e	f	g	h	i	j	k
a	0	1	0	1	1	0	1	1	0	0	0
b	1	0	1	1	1	1	0	0	0	0	0
c	0	1	0	0	1	0	1	0	1	0	1
d	1	1	0	0	0	1	0	1	0	0	0
e	1	1	1	0	0	0	1	0	0	0	0
f	0	1	0	1	0	0	0	1	0	0	0
g	1	0	1	0	1	0	0	1	0	0	1
h	1	0	0	1	0	1	1	0	0	1	1
i	0	0	1	0	0	0	0	0	0	1	1
j	0	0	0	0	0	0	0	1	1	0	1
k	0	0	1	0	0	0	1	1	1	1	0

1.2 Autocorrélation spatiale

Il est courant que les valeurs X de la variable spatiale soient influencées par la localisation P dans un espace géographique considéré. C'est pourquoi nous définissons l'autocorrélation spatiale comme la corrélation d'une variable spatiale avec elle-même qui peut être négative ou positive. Lorsque nous faisons face à une autocorrélation nulle, la répartition des valeurs de la variable S est aléatoire dans l'espace. Lorsque l'autocorrélation spatiale est positive, la répartition des localisations P est concentrée et les valeurs X associées sont similaires. Tandis que l'autocorrélation spatiale négative induit une dispersion des localisations. Par conséquent, si nous considérons une observation (x, p) de la variable S, les valeurs associées aux localisations voisines de la localisation p sont différentes de la valeur x. Nous pouvons faire une première analyse de manière globale en regardant si les valeurs de la variable sont corrélées de manière positive ou négative sur l'ensemble de l'espace considéré. Ensuite, nous pouvons réaliser une analyse locale en observant si les valeurs de la variable sont corrélées de manière positive ou négative dans une certaine zone de l'espace considéré.

1.2.1 Dépendance spatiale globale

Nous introduisons différents indices d'autocorrélation afin de répondre au problème de test suivant,

$$\begin{cases} \mathcal{H}_0 : \text{la répartition des valeurs } X \text{ de la variable spatiale est aléatoire,} \\ \mathcal{H}_1 : \text{les valeurs } X \text{ de la variable spatiale sont spatialement corrélés sur tout l'espace.} \end{cases}$$

Lors du rejet de l'hypothèse nulle, il est intéressant d'évaluer si l'autocorrélation est positive ou négative. Lors de l'acceptation de l'hypothèse nulle, la répartition des valeurs de X est aléatoire. En particulier, nous supposons que la distribution de X est normale. Les indices d'autocorrélation spatiale ont pour but de calculer la corrélation entre les valeurs X qui sont spatialement proches. Considérons X les valeurs de la variable spatiale S et WX la variable spatialement décalée des valeurs X, l'autocorrélation spatiale peut donc s'écrire comme

$$\text{Corr}(X, WX) = \frac{Cov(X, WX)}{\sqrt{\text{Var}(X)\text{Var}(WX)}}.$$

Par conséquent, nous observons que le choix de la matrice de poids aura un impact sur l'indice d'autocorrélation.

Définition 1.3
L'indice de Moran I est considéré comme le rapport entre la covariance des valeurs de localisations voisines et la variance de la valeur de la variable spatiale. De plus, il utilise la différence entre la valeur observée et la moyenne de X.

$$I = \frac{n}{\sum_{i=1}^{n} \sum_{\substack{j=1 \\ i \neq j}}^{n} W(i,j)} \frac{\sum_{i=1}^{n} \sum_{\substack{j=1 \\ i \neq j}}^{n} W(i,j)(x_i - \bar{X})(x_j - \bar{X})}{\sum_{i=1}^{n} (x_i - \bar{X})^2},$$

où n est le nombre d'observations de la variable P et $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} x_i$.

Cet indice est compris entre -1 et 1. Les valeurs positives indiquent une autocorrélation spatiale positive (respectivement négatives pour l'autocorrélation spatiale négative). Notons que l'utilisation de la statistique de Moran est favorisée grâce à sa robustesse. La loi

de la statistique de test sous $\mathcal{H}_0$ est

$$T_I = \frac{I - \mathbb{E}(I)}{\sqrt{\text{Var}(I)}} \sim \mathcal{N}(0,\, 1).$$

Définition 1.4
L'indice de Geary C est considéré comme le rapport entre la variance des valeurs de localisations voisines et la variance de la variable spatiale. De plus, il utilise la différence entre les valeurs de localisations voisines.

$$C = \frac{n-1}{2\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n} W(i,j)} \frac{\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n} W(i,j)(x_i - x_j)^2}{\sum_{i=1}^{n}(x_i - \bar{X})^2}$$

où n est le nombre d'observations de la variable P et $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Cet indice est compris entre 0 et l'infini. Lorsque l'indice est inférieur à 1, l'autocorrélation spatiale est positive. Nous pouvons aussi en déduire la loi de la statistique de test sous $\mathcal{H}_0$,

$$T_C = \frac{C - \mathbb{E}(C)}{\sqrt{\text{Var}(C)}} \sim \mathcal{N}(0,\, 1).$$

.

1.2.2 Dépendance spatiale locale

Jusqu'à présent, nous avons supposé la stationnarité spatiale ce qui correspond au fait que l'autocorrélation spatiale varie seulement selon la distance qui sépare les deux localisations et non des régions dans lesquelles elles se trouvent. Or en pratique, le livre de référence [32] indique que cela est rarement le cas. C'est pourquoi, sur base des sources [29] et [17], nous mettons en place les indices d'autocorrélation spatiale locale qui permettent de déterminer la dépendance entre les valeurs reprises dans une certaine zone géographique. Par conséquent, nous fixons une certaine localisation p_i et nous étudions la ressemblance ou la dissemblance des valeurs de X des localisations voisines de p_i. Pour chaque localisation, l'indice détermine l'intensité de l'agglomération des valeurs similaires qui se trouvent autour de la localisation considérée. Ainsi, le problème de test se réécrit de la manière suivante pour chaque localisation,

$$\begin{cases} \mathcal{H}_0 : \text{les valeurs de } X \text{ sont réparties de manière aléatoire autour d'une localisation choisie,} \\ \mathcal{H}_1 : \text{il existe une autocorrélation spatiale locale ente les valeurs de la variable spatiale.} \end{cases}$$

Définition 1.5
L'indice de Getis G_i quantifie la concentration de localisations voisines par rapport à la localisation particulière p_i,

$$G_i = \frac{\sum_{\substack{j=1\\ i\neq j}}^{n} W(i,j)x_j}{\sum_{\substack{j=1\\ i\neq j}} W(i,j)},$$

où n est le nombre d'observations de la variable P.

Lorsque la valeur de l'indice est positif, cela signifie qu'il existe un regroupement de localisations autour de la localisation choisie p_i dont les valeurs associées sont élevées. Tandis que si l'indice est négatif alors les valeurs sont faibles autour de p_i. Nous pouvons en déduire la loi de la statistique de test sous $\mathcal{H}_0$,

$$T_{G_i} = \frac{G_i - \mathbb{E}(G_i)}{\sqrt{\text{Var}(G_i)}} \sim \mathcal{N}(0,\, 1).$$

Indice de Moran local

Définition 1.6
L'indice de Moran I_i détermine si les valeurs dans une zone géographique particulière i sont similaires ou non.

$$I_i = (x_i - \bar{X}) \sum_{\substack{j=1 \\ i \neq j}}^{n} W(i,j)(x_j - \bar{X}),$$

où n est le nombre d'observations de la variable P et $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Lorsque la valeur de l'indice est positive, cela signifie qu'il existe un regroupement de valeurs similaires autour de la localisation p_i. Il est possible que les valeurs soient élevées ou faibles. Tandis que si l'indice est négatif alors les valeurs sont très différentes autour de la localisation p_i. Nous pouvons en déduire la loi de la statistique de test sous $\mathcal{H}_0$,

$$T_{I_i} = \frac{I_i - \mathbb{E}(I_i)}{\sqrt{\text{Var}(I_i)}} \sim \mathcal{N}(0,\, 1).$$

1.2.3 Diagramme de Moran

Il est possible de réaliser un diagramme de Moran afin de pouvoir analyser de manière locale et globale l'autocorrélation. Nous analysons un nuage de points dont l'abscisse donne les valeurs de X dont la variable est standardisée et l'ordonnée les valeurs de la variable WX qui sont les valeurs voisines de X. Lorsque l'autocorrélation spatiale est présente, la pente d'une régression linéaire devient non nulle. Le signe de la pente correspond à une autocorrélation positive ou négative.

Ensuite, nous analysons les différentes catégories d'autocorrélation spatiale grâce aux quatre quadrants du diagramme représenté sur la Figure 1.3. Notons que les quadrants sont délimités par X=WX=0. Dans le cadre de notre exemple, X mesure le taux de chômage en 2017 sur l'ensemble des communes de la Belgique. La matrice de poids est obtenue par la méthode basée sur les 4 plus proches voisins grâce à la fonction *knn2nb* en *R*. En raison de la pente de régression, nous sommes face à une autocorrélation positive. Lorsque nous nous déplaçons de la gauche vers la droite sur l'axe des abscisses, la valeur de la variable X qui était faible devient élevée. Les quadrants *high-high* et *low-low* représentent les valeurs de la variable spatiale possédant une autocorrélation spatiale positive avec une valeur élevée de X et réciproquement faible. Les quadrants *high-low* et *low-high* représentent les valeurs de la variable spatiale possédant une autocorrélation spatiale négative avec une valeur élevée de X et réciproquement faible. De plus, les points marqués par un losange représentent les points possédant une autocorrélation plus marquante que

les points marqués par un rond.

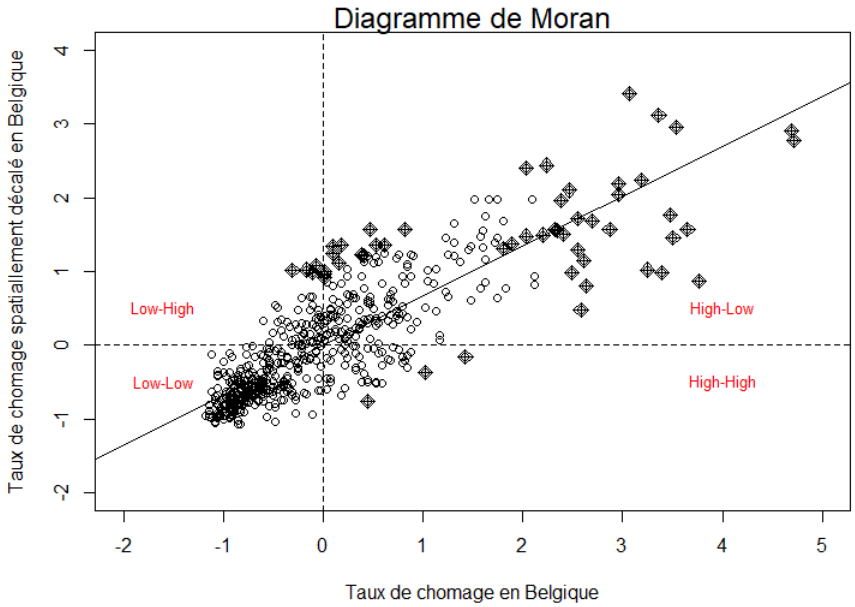

FIGURE 1.3 – Diagramme de Moran du taux de chômage en 2017 dans l'ensemble des communes de la Belgique

Chapitre 2

Données ponctuelles

Dans le chapitre précédent, nous avions défini une variable spatiale X comme un couple de deux composantes (X, P). A présent, nous étudions les données spatiales ponctuelles. Notons qu'ici, les données spatiales ponctuelles ont seulement la composante P. Passons maintenant aux méthodes statistiques utilisées pour analyser ce type de données qui se nomment les méthodes de configuration de points.

Définition 2.1
Une configuration de n points, $\mathcal{C} = \{x_1, \ldots, x_n\}$, se définit comme un ensemble de points où $n(\mathcal{C})$ est le nombre de points de cette configuration.

La distribution d'une configuration de points peut être de trois types : aléatoire, agrégée ou régulière. En effet, il est possible d'avoir une interaction entre les points. De plus, cette interaction peut pousser les données à s'écarter (régulière) ou à se rassembler (agrégée).

Un processus ponctuel est une collection de points dans un espace donné qui se nomme fenêtre (*window*). Notons X le processus ponctuel. Nous travaillons avec des processus qui génèrent une unique réalisation. En effet, si nous observons la répartition des universités en Belgique, nous ne pouvons pas avoir deux endroits qui situent l'Unamur et nous avons donc bien une seule répartition.

Définition 2.2
Un processus X est défini si on connaît pour toute région B la loi de la variable aléatoire fournissant $n(\mathcal{C} \cap B)$ points.

Habituellement, nous travaillons avec des processus dit localement définis lorsque $n(\mathcal{C} \cap B) < \infty$, pour toute région B. Si nous souhaitons tenir compte des caractéristiques des points de la configuration, nous travaillerons avec les processus ponctuels marqués (*marked point pattern*). Ces caractéristiques se nomment *marques du point* et elles sont de deux types : qualitatives et quantitatives. Prenons comme exemple, les restaurants dont le type de cuisine : asiatique, française, italienne,... est une caractéristique qualitative. Tandis que le nombre de couverts, la superficie du restaurant,... est une caractéristique quantitative.

2.1 Moments d'ordre 1 et 2

Les processus doivent tenir compte de la répartition des données. Afin de réaliser une comparaison entre la distribution considérée et une distribution aléatoire, nous utilisons les indicateurs d'ordre 1 et 2.

Définition 2.3
L'intensité d'ordre 1 *peut être définie comme une constante* λ *ou une variable* $\lambda(x)$*, avec* x *qui représente la localisation, en fonction du processus avec lequel nous travaillons,*

$$\mathbb{E}[n(X \cap B)] = \mu(B) = \int_B \lambda(x)dx.$$

Définition 2.4
L'intensité d'ordre 2*,* λ_2*, se définit comme le produit de densité pour deux régions disjointes, avec* y *et* x *qui représentent les localisations des régions* A *et* B *respectivement ,*

$$\mathbb{E}[n(X \cap A)n(X \cap B)] = \int_A \int_B \lambda_2(x, y)dxdy.$$

En mettant en commun les deux définitions de l'intensité d'ordre 1 et 2, nous obtenons la fonction de corrélation de paire de points :

$$g_2(x, y) = \frac{\lambda_2(x, y)}{\lambda(x)\lambda(y)}.$$

2.2 Les processus de Poisson

2.2.1 Le processus de Poisson homogène

Nous nous intéressons aux processus ponctuels complètement aléatoires (*Complete Spatial randomness - CRS*)[14].

Méthode 2.2.1 (Processus CSR)
Le processus de Poisson homogène est défini par

$$P(n(X \cap B) = k) = e^{-\lambda|B|}\frac{\lambda^k|B|^k}{k!}$$

où la distribution des points est celle de Poisson. Notons que l'homogénéité est respectée par la condition $\mathbb{E}[n(X \cap B)] = \lambda|B|$*. De plus, la condition d'indépendance l'est aussi vu que nous avons* m *variables aléatoires indépendantes* $n(X \cap B_1), \ldots, n(X \cap B_m)$*.*

La condition d'homogénéité correspond au fait que les points peuvent être placés à n'importe quelle position. Il n'y a donc pas de prédominance pour une localisation particulière. De plus, celle-ci permet d'établir que la quantité de points attendus dans la région B doit être proportionnelle à la quantité de celle-ci. L'indépendance démontre que la répartition des points d'une région n'influence pas celle d'une autre région.

L'intensité d'ordre 1, λ, se définit comme le nombre de points par unité de surface. Lorsque l'homogénéité est respectée, cela signifie que λ est constante mais cela ne sera pas toujours le cas lors de l'utilisation d'autres processus comme celui de Poisson inhomogène. De plus, l'intensité d'ordre 2 se formule comme $\lambda_2(x, y) = \lambda^2$, ainsi $g_2(x, y) = 1$.

Pour résumer, nous notifions que ce processus possède deux propriétés importantes. Tout d'abord, un processus CRS est *stationnaire* ce qui signifie qu'il est invariant sous translation et qu'il est homogène avec une intensité constante. Cette propriété est primordiale lors de l'utilisation de certains outils comme la fonction de Ripley que nous verrons

en détails dans la section 2.4.1. En plus d'être stationnaire, le processus est isotrope ce qui indique qu'il est invariant sous translation. En particulier, le processus de Poisson homogène est adapté pour des données réparties de manière aléatoire. Ces données n'ont donc pas d'interaction entre elles. Le processus de Poisson homogène peut être produit dans R grâce à la fonction $rpoispp(\kappa)$ où κ représente l'intensité du processus.

2.2.2 Le processus de Poisson inhomogène

Le processus de Poisson inhomogène se définit comme une adaptation du processus homogène. L'intensité n'est plus constante mais variable. Par conséquent, l'intensité se définit grâce à une fonction. De plus, la propriété d'indépendance est conservée. Les points sont i.i.d. avec une densité de probabilité $f(x) = \frac{\lambda(x)}{\int_B f(u)du}$. Encore une fois, la fonction utilisée dans R est *rpoispp* où nous fournissons l'expression de la fonction d'intensité. Ainsi, nous pouvons différencier deux types de configuration de points aléatoires, soit homogène par un processus de Poisson homogène, soit inhomogène par un processus de Poisson inhomogène comme nous le voyons à la Figure 2.1.

2.3 Les processus de Matérn

Les configurations de points considérées comme régulières ou agrégées sont couramment exploitées grâce au processus de Matérn expliqué dans le livre de référence [6]. Lors de la répartition régulière, la relation d'interaction entre les points a pour objectif de repousser les points afin de les rendre plus régulièrement espacés que lors d'une simple distribution aléatoire. A l'inverse d'une distribution agrégée où la relation d'interaction entre les points permet de rapprocher les points entre eux. Nous présentons les trois processus de Matérn qui vont tenir compte de la relation d'interaction et de dépendance entre les données.

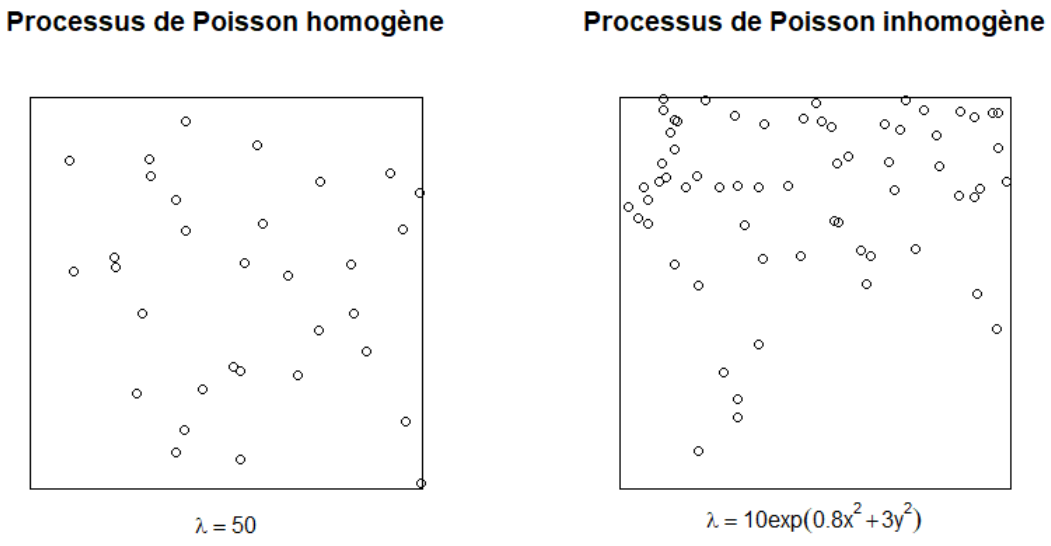

FIGURE 2.1 – Processus de Poisson homogène et inhomogène avec $\lambda = 50$ et $\lambda = 10e^{(0.8x^2+3y^2)}$ avec x et y qui représente les coordonnées.

2.3.1 Le processus de Matérn 1 et 2

En opposition du processus de Poisson homogène où nous avons considéré que les points étaient indépendants entre eux, ici nous considérons l'hypothèse de dépendance. Nous introduisons le processus de Matérn 1 et 2.

Le processus de Matérn 1 s'exécute en deux étapes :

— Étape 1 : grâce au processus de Poisson homogène, nous générons une configuration de points.

— Étape 2 : la suppression des couples de points qui ont une distance entre eux inférieure à r. Notons que r est une distance arbitrairement fixée.

Sur la Figure 2.2, les étapes du processus sont représentées de la gauche vers la droite. La fonction $rMaternI(\kappa, r)$, où κ est l'intensité du processus de Poisson homogène et r est la distance, est utilisée en R afin de générer un tel processus.

Le processus de Matérn 2 est semblable au processus de Matérn 1 mais cette fois-ci, une notion temporelle intervient. Lorsque nous générons notre série de points lors de la première étape, chaque point est lié à une variable t_i qui est le temps d'arrivée du point. L'ensemble des t_i sont indépendants et uniformément distribués sur l'intervalle $[0, 1]$. Lors de la deuxième étape, la distance entre deux points se calcule avec un point considéré et un des points qui le précède temporellement. La fonction *rMaternII* en R prend les mêmes arguments que *rMaternI*.

2.3.2 Le processus de Matérn agrégé

Lors d'une répartition agrégée, la relation d'interaction entre les points va provoquer des agglomérations de points car celle-ci va les rassembler. Nous utilisons dans ce cas, le processus de Matérn cluster qui se réalise en trois étapes :

— Étape 1 : grâce au processus de Poisson homogène, nous générons un certain nombre de points appelé *parents* avec une certaine intensité constante κ ;

— Étape 2 : pour chaque *parents*, nous générons μ points appelés *enfants*. Notons que les enfants sont des points indépendants et uniformément distribués dans un cercle de rayon r centré au parent considéré ;

— Étape 3 : nous conservons les points de type *enfant* et nous supprimons les *parents*.

Sur la Figure 2.3, nous observons les différentes étapes, de la gauche vers la droite, du processus de Matérn agrégé. La fonction que nous utiliserons en R afin d'établir un processus de ce type est $rMatClust(\kappa, r, \mu)$ où les points ont tendance à se rassembler.

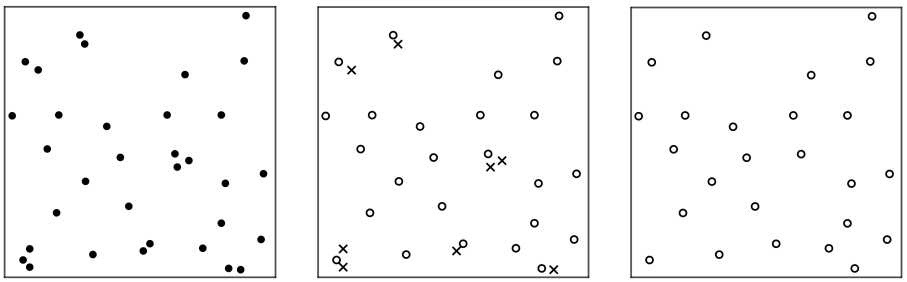

FIGURE 2.2 – Illustration du processus de Matérn 1.

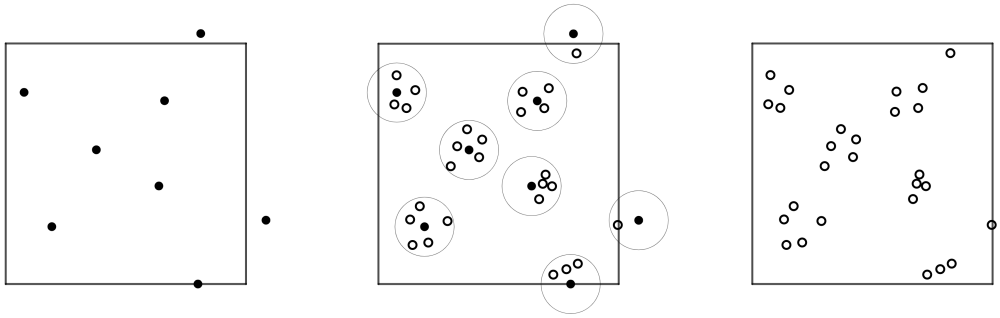

FIGURE 2.3 – Illustration du processus de Matérn agrégé.

2.4 Caractérisation d'une configuration de points

En fonction du contexte dans lequel nous nous situons, il est facile de repérer si la distribution de points est aléatoire, agrégée ou régulière. Si nous prenons un positionnement de vignes, il est cohérent que les données soient distribuées de manière régulière et que les données aient une relation d'interaction entre elles. Toutefois, il est possible que cela ne soit pas toujours si évident visuellement. Prenons un exemple d'un processus de Poisson inhomogène, même si les données sont indépendantes, le processus de Poisson inhomogène de la Figure 2.1 peut laisser supposer que la configuration soit agrégée. Afin de pouvoir comparer nos distributions à des distributions complètement aléatoires et ainsi déterminer le type de configuration, nous introduisons les différentes fonctions de Ripley.

2.4.1 La fonction K de Ripley

Une des fonctions les plus utilisées pour étudier les configurations de points est la fonction K de Ripley. Elle permet d'évaluer le nombre moyen de voisins des points du processus. Le point considéré et ses points voisins sont séparés d'une distance maximale de r. Si nous considérons un point quelconque, le nombre de points espérés, $\lambda\pi r^2$, se formule comme l'intensité homogène sur un cercle d'aire πr^2. Afin de comparer l'ensemble des points, nous réalisons une standardisation par l'intensité $\frac{n}{|W|}$, où $|W|$ est l'aire de la fenêtre d'observation.

Définition 2.5
L'estimateur de la fonction K de Ripley s'écrit sous la forme suivante :

$$\hat{K}(r) = \frac{|W|}{n(n-1)} \sum_i \sum_{j \neq i} \mathbb{1}\{||x_i - x_j|| \leq r\} c(x_i, x_j; r), \tag{2.1}$$

où

- *n est le nombre total de points sur l'espace d'observation*
- $\mathbb{1}\{||x_i - x_j|| \leq r\} = \begin{cases} 1 \text{ si la distance qui sépare les points } i \text{ et } j \text{ est inférieure ou égale à } r, \\ 0 \text{ sinon.} \end{cases}$
- *$c(x_i, x_j; r)$ est la correction d'effets de bord*[1]
- *W est la fenêtre d'observation*

1. Lorsque nous considérons un point proche de la limitation de la fenêtre d'observation et que nous observons les points qui l'entourent d'une distance maximale r, nous constatons que certains de ses voisins sont hors du domaine. Par conséquent, nous risquons une sous estimation de son voisinage et appliquons donc une pratique rectificatrice qui se nomme *correction d'effets de bord*

Concrètement, calculer $\hat{K}(r)$ revient à faire varier la distance pour chaque point. Ensuite, nous calculons le nombre de voisins présents dans le cercle de rayon r centré au point considéré et en apportant une correction aux effets de bord si besoin. Lorsque nous avons calculé le nombre moyen de points, nous comparons l'écart entre cette valeur et le nombre moyen de points pour une distribution de Poisson homogène $K_{poisson}(r)$ qui vaut πr^2. Ainsi, nous pouvons déterminer le type de distribution des configurations de points.

— $\hat{K}(r) \approx K_{poisson}(r)$: le processus est aléatoire et la condition indique alors que la distribution est proche d'une distribution d'un processus *CRS* ;

— $\hat{K}(r) > K_{poisson}(r)$: le processus est agrégé car la condition s'interprète comme le fait qu'il y a plus de πr^2 points dans le cercle de rayon r donc les points s'amassent ;

— $\hat{K}(r) < K_{poisson}(r)$: le processus est régulier car la condition s'interprète comme le fait qu'il y a moins de πr^2 points dans le cercle de rayon r donc les points s'écartent.

Ainsi, lorsque nous générons un processus de Poisson homogène, de Matérn 2 et de Matern agrégée, nous pouvons voir à la Figure 2.4 que leur distribution sont respectivement aléatoire, régulière et agrégée. La fonction de Ripley est obtenue par la fonction *Kest* en *R*.

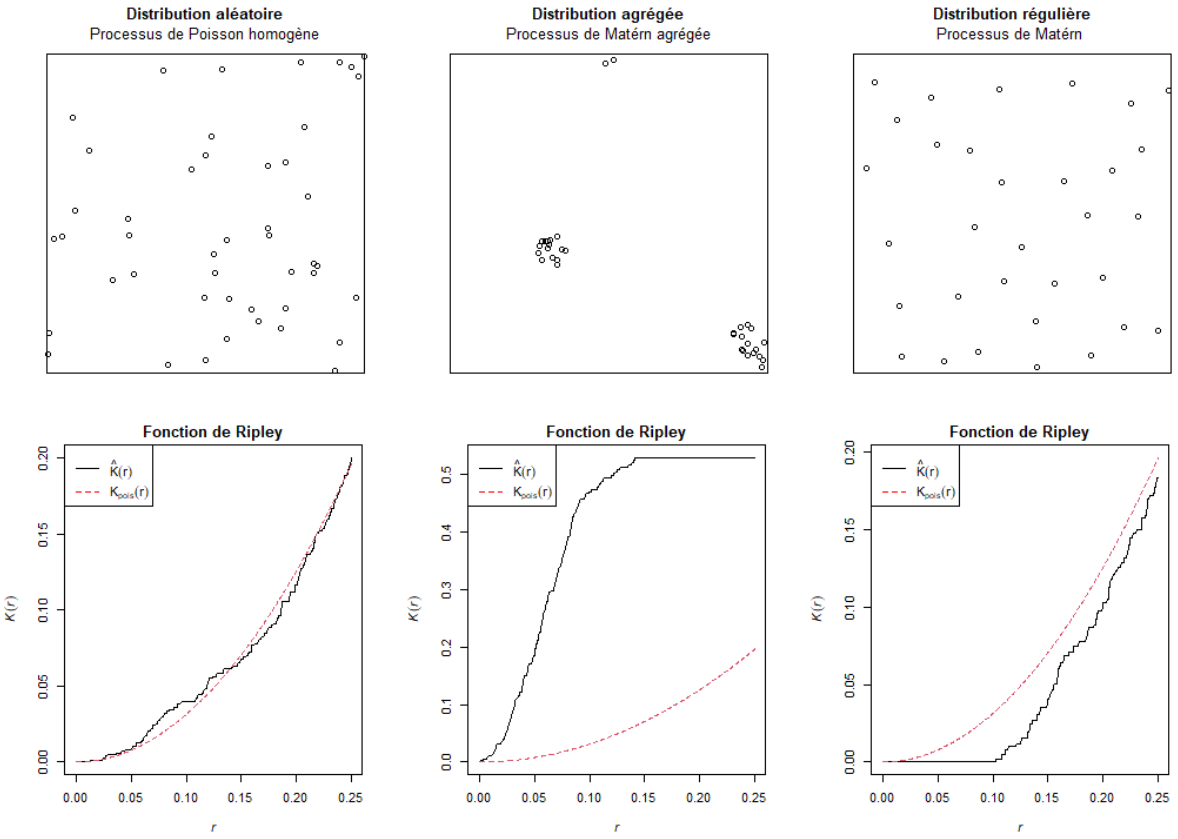

FIGURE 2.4 – Distributions de points aléatoire, agrégée et régulière avec leur fonction de Ripley respective.

2.4.2 La fonction L de Ripley

Nous introduisons une première transformation de la fonction de Ripley K.

Définition 2.6
L'estimateur de la fonction L de Ripley s'écrit sous la forme suivante :

$$\hat{L}(r) = \sqrt{\frac{\hat{K}(r)}{\pi}}.$$

Lors d'un processus de Poisson homogène $K_{poisson}(r) = \pi r^2$, la fonction devient $\hat{L}_{poisson}(r) = r$ ce qui permet de rendre la comparaison entre les distributions encore plus lisibles. De plus, la racine carrée permet de stabiliser la variance de l'estimateur $\hat{L}(r)$.

2.4.3 La fonction K_{inhom} de Ripley

Nous introduisons une autre transformation de la fonction de Ripley K. En effet, plusieurs raisons ont motivé ce choix. Tout d'abord, l'hypothèse de stationnarité ne permet pas de prendre en compte des processus plus complexes comme le processus de Poisson inhomogène. En effet, nous aurons une forte variation de l'intensité. Lorsque nous observons le résultat fourni par la fonction de Ripley K à la Figure 2.5, nous observons que $\hat{K}(r) > K_{poisson}(r)$ ce qui laisserait sous entendre que les points soient agrégés alors qu'ils sont indépendants. Ensuite, la corrélation entre les points n'est pas toujours de nature intrinsèque comme nous l'avons vu jusqu'à présent avec la distribution agrégée et régulière. De fait, des facteurs extrinsèques peuvent intervenir. La régularisation intervient dans différents contextes comme l'écologie avec le type de sol ou l'importance de certaines ressources. Un facteur qui influence l'agrégation pourrait être la biologie avec la reproduction ou la contamination.

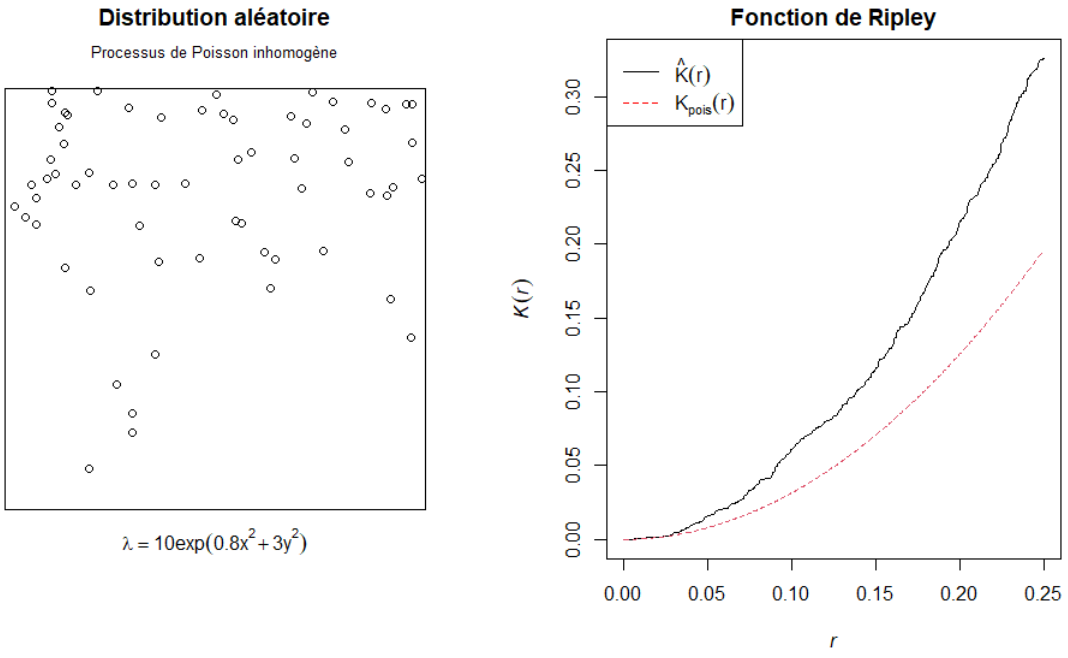

FIGURE 2.5 – Distribution de points avec sa fonction de Ripley associée.

Enfin, même si la dépendance des points induit une corrélation. L'opposé n'est pas vrai, le manque de corrélation n'implique pas que les données soient indépendantes. Prenons le processus cellulaire de Baddeley-Silverman dont l'espace d'observation est divisé en cellules de même taille. Dans chaque cellule, un certain nombre de points (0, 1 ou 10) sont placés de manière indépendantes et uniformes. Une certaine dépendance entre les points est présente au vu de la construction de l'espace. Nous observons à la Figure 2.6 que $\hat{K}(r) \approx K_{poisson}(r)$ ce qui signifie que le processus serait aléatoire alors qu'il ne l'est pas.

L'objectif ici est d'établir une version de la fonction de Ripley K lorsque l'hypothèse d'homogénéité n'est pas respectée. Notons que l'intensité ne va plus se définir comme une constante mais comme une variable dépendante du point considéré. Nous apportons une notion de poids w_i sur chaque point x_i de la configuration de points par la formulation suivante

$$w_i = \frac{1}{\lambda(x_i)}.$$

Lorsque nous considérons une paire de points x_i et x_j où $i \neq j$, le poids qui lui est affecté est

$$w_i w_j = w_{ij} = \frac{1}{\lambda(x_i)\lambda(x_j)}.$$

Reprenons la formule caractérisant le nombre de points attendus pour l'ensemble des points de la fenêtre d'observation où nous faisons intervenir la notion de poids,

$$\begin{aligned} \sum_i w_i \lambda(x_i)\pi r^2 &\Leftrightarrow \sum_i \frac{1}{\lambda(x_i)}\lambda(x_i)\pi r^2 \\ &\Leftrightarrow \pi r^2. \end{aligned}$$

Nous observons que nous pouvons interpréter nos résultats de la même manière que dans le cas homogène comme $K_{inhom,poisson}(r) = K_{homo,poisson}(r) = \pi r^2$.

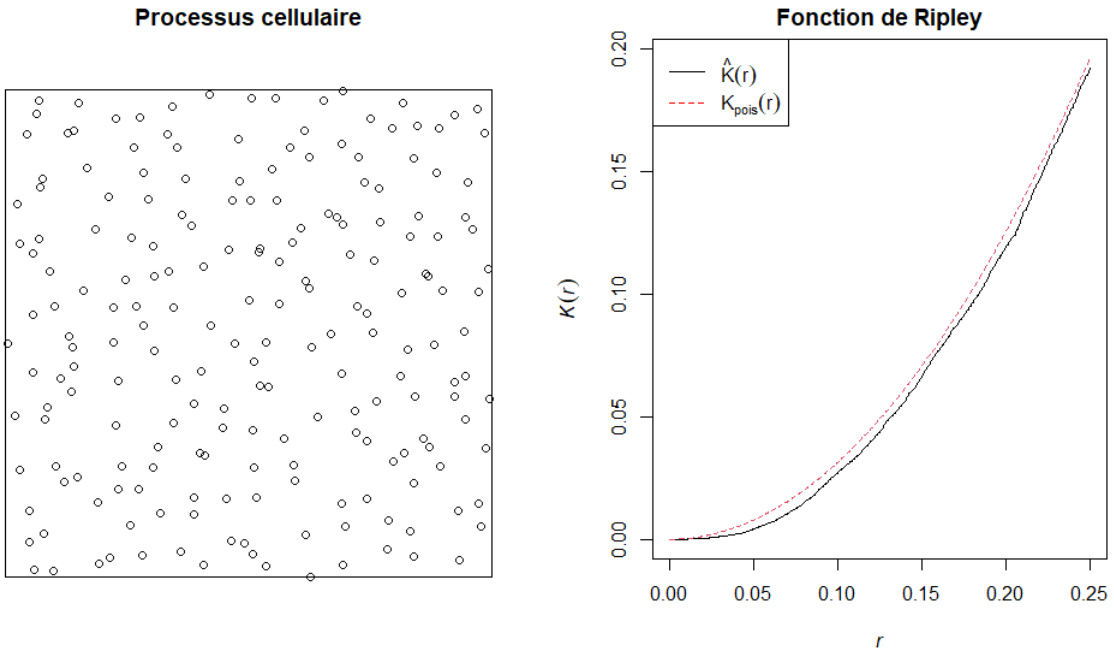

FIGURE 2.6 – Un processus cellulaire et sa fonction de Ripley associée.

Définition 2.7
L'estimateur de la fonction K_{inhom} de Ripley s'écrit sous la forme suivante :

$$\hat{K}_{inhom}(r) = \frac{1}{D}\sum_i \sum_{j\neq i} \frac{\mathbb{1}\{||x_i - x_j|| \leq r\}}{\hat{\lambda}(x_i)\hat{\lambda}(x_j)} c(x_i, x_j; r), \tag{2.2}$$

où

- *n est le nombre total de points sur l'espace d'observation ;*
- $\mathbb{1}\{||x_i - x_j|| \leq r\} = \begin{cases} 1 \text{ si la distance qui sépare les points } i \text{ et } j \text{ est inférieure ou égale à } r, \\ 0 \text{ sinon.} \end{cases}$
- *$c(x_i, x_j; r)$ est la correction d'effets de bord ;*
- *$\hat{\lambda}(x_i)$ est l'estimation de l'intensité au point x_i ;*
- *$D = \frac{1}{|W|}\sum_i \frac{1}{\hat{\lambda}(x_i)}$ est l'estimation globale de l'intensité sans erreur.*

Dans l'expression de l'estimateur $\hat{K}_{inhom}(r)$, nous avons l'estimateur de la fonction de densité au point x_i. Afin d'obtenir cet estimateur, nous utilisons la méthode du noyau (*kernel estimation*). Il existe un grand nombre de fonction du noyau mais nous utilisons la fonction de correction uniforme et la fonction de correction de Diggle,

$$\begin{cases} \hat{\lambda}^{(u)}(x) = \frac{1}{e(x)} \sum_{i=1}^n \kappa(x - x_i) \\ \hat{\lambda}^{(d)}(x) = \sum_{i=1}^n \frac{1}{e(x_i)} \kappa(x - x_i), \end{cases} \tag{2.3}$$

où x est la localisation dans la fenêtre d'observation W et $e(x)$ est la correction d'effets de bords. Notons qu'un choix courant de fonction du noyau $\kappa(x)$ est la densité d'une fonction Gaussienne tel que $\kappa(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$. Toutefois, un biais $\kappa(0)$ apparaît lorsque nous considérons les estimateurs $\hat{\lambda}^{(u)}$ et $\hat{\lambda}^{(d)}$ au point x_i. Par conséquent, si cet estimateur est biaisé alors $\hat{K}_{inhom}(r)$ le sera aussi. C'est la raison pour laquelle nous utilisons la méthode de validation croisée pour obtenir l'estimateur $\hat{\lambda}(x_i)$. Nous calculons l'estimateur avec l'ensemble des points x_j où $j \neq i$,

$$\begin{cases} \hat{\lambda}_i^{(u)}(x_i) = \frac{1}{e(x_i)} \sum_{i\neq j} \kappa(x_i - x_j), \\ \hat{\lambda}_i^{(d)}(x_i) = \sum_{i\neq j} \frac{1}{e(x_j)} \kappa(x_i - x_j). \end{cases} \tag{2.4}$$

Lorsque nous reprenons la distribution de la Figure 2.5 avec la fonction de Ripley inhomogène, *Kinhom* en R, nous observons sur la Figure 2.7 que les deux courbes sont approchées ce qui signifie que la distribution des points est bien aléatoire.

2.5 Processus multitypes

Nous avons introduit les processus ponctuels marqués où chaque caractéristique des données était représentée par un symbole. A présent, nous regroupons l'ensemble des caractéristiques de la configuration de points en utilisant des processus multitypes.

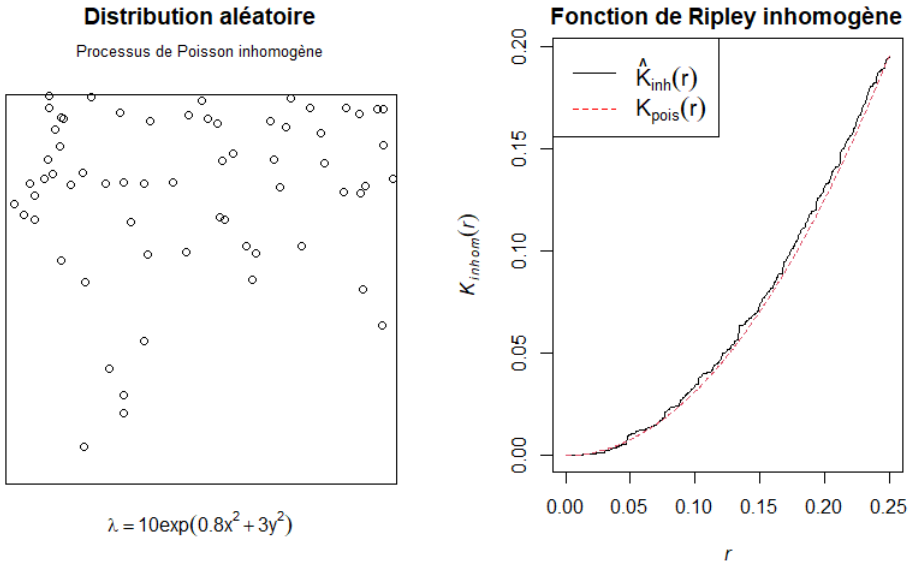

FIGURE 2.7 – Distribution aléatoire de points avec sa fonction de Ripley inhomogène respective.

Définition 2.8
Un processus ponctuel multitype Y est fondé à partir de sous processus-ponctuels $X^{(1)}, \ldots, X^{(M)}$ où $X^{(i)}$ est le processus ponctuel marqué de type i pour chaque $i = 1, \ldots, M$.

$X^{(\bullet)}$ est le chevauchement de l'ensemble des sous-processus ponctuels marqués. Par conséquent, $X^{(\bullet)}$ est un processus ponctuel non marqué qui détient uniquement les informations relatives à la localisation.

Par exemple, nous pouvons afficher une carte reprenant l'ensemble des restaurants d'une région classifié par le type de cuisine. Dans le cas multivarié, nous étudions la relation entre la localisation et le type de marque. Si la relation est dépendante, nous étudierons si les données ont une tendance à s'agglomérer ou à se repousser. Tout d'abord, nous redéfinissons la notion d'intensité présente dans la formulation de nombreux processus.

2.5.1 Fonction d'intensité

Afin de définir l'intensité du processus ponctuel multiple Y, nous calculons l'intensité pour chaque processus ponctuel marqué $X^{(i)}$ pour $i = 1, \ldots, M$. L'homogénéité du processus Y est respecté lorsque

$$\mathbb{E}[n(X^{(i)} \cap B)] = \lambda_i |B|,$$

pour chaque $i = 1, \ldots, M$ où λ_i est l'intensité lié au processus ponctuel $X^{(i)}$. Par conséquent, l'intensité du premier ordre du processus $X^{(\bullet)}$ se définit par $\lambda_\bullet = \sum_{i=1}^{M} \lambda_i$. Toutefois, comme nous l'avons déjà observé dans le cadre univarié, l'intensité peut être définie comme une variable lors d'utilisation de certains processus. La fonction d'intensité du processus $X^{(\bullet)}$ est donc la somme des fonctions d'intensités individuelles des sous-processus, $\lambda_\bullet = \sum_{i=1}^{M} \lambda_i(x) = \sum_{i=1}^{M} \lambda(x, i)$.

Définition 2.9
La fonction d'intensité $\lambda(x, i)$ avec la localisation x et la caractéristique i pour un processus ponctuel multiple se définit par

$$\mathbb{E}[n(X^{(i)} \cap B)] = \int_B \lambda(x, i)dx.$$

Tout d'abord, nous généralisons au cas multivarié les processus de Poisson homogène et inhomogène que nous avons détaillés dans le cas univarié.

2.5.2 Le processus multitype de Poisson homogène

Un processus est complètement aléatoire et indépendant (*Complete Spatial Randomness and Independance - CRSI*) si les deux premières ou les deux dernières caractéristiques sont respectées :

1. Le processus ponctuel non marqué est un *CRS* avec la densité $\lambda_\bullet$;
2. Les M catégories sont i.i.d., c'est-à-dire indépendamment des autres points, chaque point est attribué de manière aléatoire à une des M catégories avec une probabilité de $p_i = \dfrac{\lambda_i}{\lambda_\bullet}$;
3. Chaque sous-processus ponctuel marqué $X^{(i)}$ est un *CRS* avec la densité λ_i pour $i = 1, \ldots, M$;
4. Les points de chaque catégorie i pour $i = 1, \ldots, M$ sont indépendants.

Notons que chaque catégorie possède une distribution avec une certaine probabilité. Par conséquent, il est possible qu'une distribution de points d'une certaine catégorie ne soit pas proportionnelle par rapport aux autres catégories. Cependant, cela n'empêche pas le processus multitype Y d'être homogène. La fonction *rmpoisspp*(κ) où κ est le vecteur reprenant l'ensemble des intensités pour les sous-processus est utilisé pour générer un $CRSI$ en R.

2.5.3 Le processus multitype de Poisson inhomogène

La source [5] nous indique que lorsque nous nous plaçons dans le cadre d'une intensité variable cela produit l'inhomogénéité du processus multitype. Si les deux premières ou les deux dernières caractéristiques suivantes sont respectées alors le processus est dit multitype de Poisson inhomogène :

1. Le processus ponctuel non marqué est un processus de Poisson inhomogène avec la densité $\lambda_\bullet(x) = \sum_{i=1}^{M} \lambda_i(x)$;
2. Les M catégories sont indépendantes les unes des autres et chaque point est attribué à une des M catégories avec une probabilité de $p(i|x) = \dfrac{\lambda_i(x)}{\lambda_\bullet(x)}$;
3. Chaque sous-processus ponctuel marqué $X^{(i)}$ est un processus de Poisson inhomogène avec la densité $\lambda_i(x)$ pour $i = 1, \ldots, M$;
4. Les points de chaque catégorie i pour $i = 1, \ldots, M$ sont indépendants.

Nous observons que l'intensité pour chaque sous-processus marqué varie spatialement et n'est plus constante comme dans le cas homogène. En R, nous fournirons les expressions des fonctions de densité à la fonction *rmpoisspp*. Sur la Figure 2.8, nous générons par exemple un processus de Poisson homogène avec une intensité de 150 pour la catégorie A et de 50 pour la catégorie B, ce qui provoque un plus petit nombre de points pour la

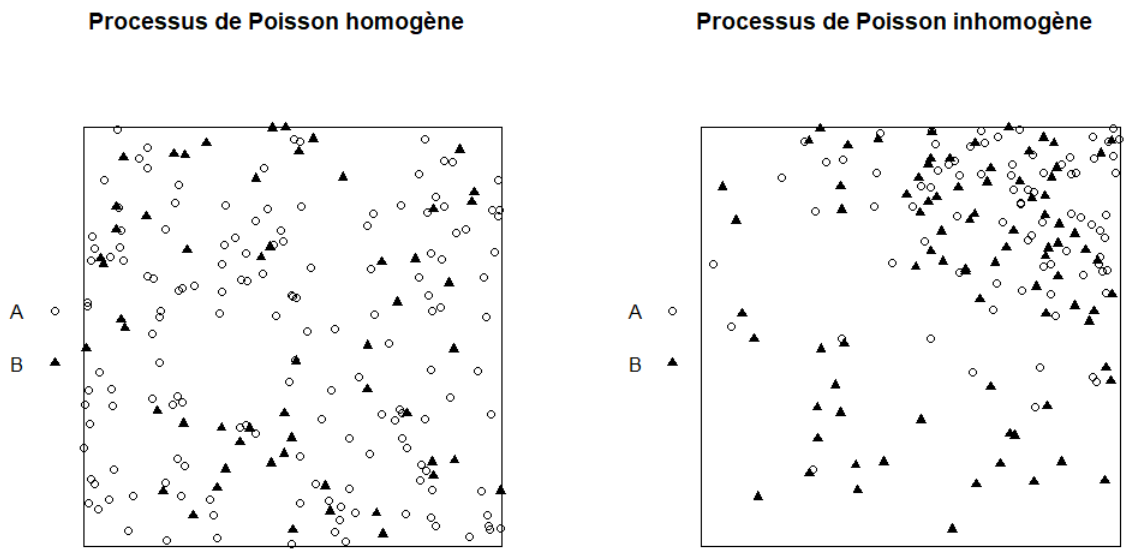

FIGURE 2.8 – Processus multiple de Poisson homogène et inhomogène.

catégorie B. De plus, nous générons un processus de Poisson inhomogène avec $\lambda = 250e^{x^2}$ pour la catégorie A et $\lambda = 250e^{1-x}$ pour la catégorie B.

2.6 Fonction intertype

A présent, nous nous intéressons aux interactions entre les sous-processus ponctuels marqués. L'objectif est d'explorer les configurations de points pour chaque marque du point. Nous généralisons la fonction K et K_{inh} de Ripley.

2.6.1 La fonction K intertype

Nous souhaitons étudier les catégories A et B d'un processus multitype. Nous étudions la configuration de points du type B qui se situe à une distance inférieure ou égale à r de la configuration de points A.

Définition 2.10
L'estimateur de la fonction intertype K s'écrit sous la forme suivante :

$$\hat{K}_{A,B}(r) = \frac{1}{\hat{\lambda}_A n_A} \sum_{i \in X^{(A)}} \sum_{j \in X^{(B)}} \mathbb{1}\{||x_i - x_j|| \leq r\} c(x_i, x_j; r), \tag{2.5}$$

où

- n_A *est le nombre total de points du sous-processus ponctuel* A,
- $\hat{\lambda}_A$ *est l'estimation de l'intensité du sous-processus ponctuel* A,
- $\mathbb{1}\{||x_i - x_j|| \leq r\} = \begin{cases} 1 \text{ si la distance qui sépare les points } i \text{ et } j \text{ est inférieure ou égale à } r, \\ 0 \text{ sinon.} \end{cases}$
- $c(x_i, x_j; r)$ *est la correction d'effets de bord.*

L'estimation de l'intensité du sous-processus A peut être définie par les fonctions de noyaux (2.3) définies dans le cas univarié. De plus, ces estimateurs sont non biaisés grâce à l'hypothèse d'homogénéité. Si nous supposons que $A = B$, $\hat{K}_{A,B}(r)$ devient $\hat{K}(r)$. Nous retrouvons l'estimateur de la fonction K de Kipley avec πr^2 qui représentait le nombre de points espéré sur un cercle d'aire πr^2 standardisé par l'intensité. Lorsque $A \neq B$, la valeur de référence reste πr^2 grâce à l'hypothèse d'indépendance entre les sous-processus. En effet, le nombre attendu de points de $X^{(B)}$ espacés d'une distance inférieure ou égale d'un point de $X^{(A)}$ est simplement le nombre attendu de points de $X^{(B)}$ repris dans un cercle de rayon r qui est $\pi r^2 \lambda_B$. Ainsi, nous pouvons déterminer le type d'interaction qu'il y a entre les processus :

— $\hat{K}_{A,B}(r) \approx \pi r^2$: les sous-processus n'ont pas de relation entre eux ;

— $\hat{K}_{A,B}(r) > \pi r^2$: les points du processus de $X^{(B)}$ sont attirés vers les points de $X^{(A)}$ dans un rayon r ;

— $\hat{K}_{A,B}(r) < \pi r^2$: les points du processus de $X^{(B)}$ sont repoussés autour des points de $X^{(A)}$.

La fonction intertype homogène est *Kcross* en *R*. Ainsi à la Figure 2.9, lorsque nous comparons les sous-processus du processus multitypes homogène que nous avons généré précédemment, nous constatons qu'ils n'ont pas de relation entre eux.

2.6.2 La fonction intertype inhomogène $K_{A,B}^{inhom}$

Comme nous l'avions mentionné lors du cas univarié, lorsque le processus n'est pas stationnaire, il est parfois difficile d'identifier comment les points sont caractérisés entre aléatoire et agrégé. La notion de poids intervient pour chaque point en prenant l'inverse de la fonction d'intensité et en considérant une paire de points. Nous introduisons donc la fonction suivante définie dans le cas multivarié.

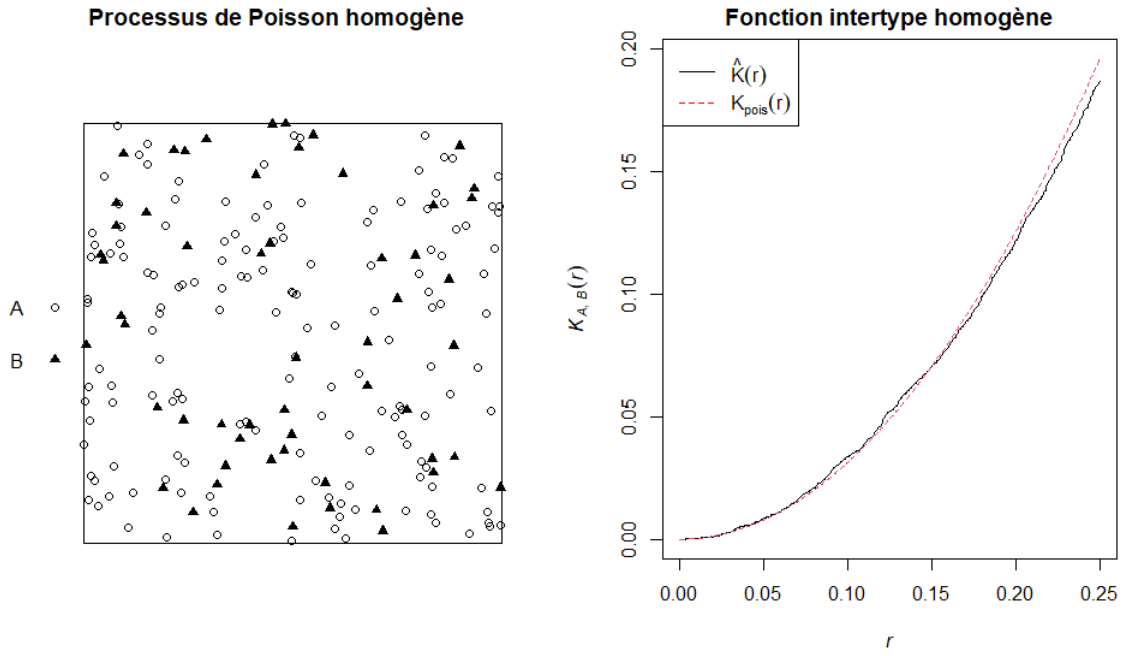

FIGURE 2.9 – Processus multiple de Poisson homogène et sa fonction intertype.

Définition 2.11
L'estimateur de la fonction intertype $K_{A,B}^{inhom}$ *s'écrit sous la forme suivante :*

$$\hat{K}_{A,B}^{inhom}(r) = \frac{1}{|W|} \sum_{i \in X^{(A)}} \sum_{j \in X^{(B)}} \frac{\mathbb{1}\{||x_i - x_j|| \leq r\}}{\hat{\lambda}_A(x_i)\hat{\lambda}_B(x_j)} c(x_i, x_j; r), \tag{2.6}$$

où

- $\mathbb{1}\{||x_i - x_j|| \leq r\} = \begin{cases} 1 \text{ si la distance qui sépare les points } i \text{ et } j \text{ est inférieure ou égale à } r, \\ 0 \text{ sinon.} \end{cases}$
- $c(x_i, x_j; r)$ *est la correction d'effets de bord ;*
- $\hat{\lambda}_A(x_i)$ *est l'estimation de l'intensité du sous-processus ponctuel* A *au point* x_i *;*
- $\hat{\lambda}_B(x_j)$ *est l'estimation de l'intensité du sous-processus ponctuel* B *au point* x_j *;*
- W *est la fenêtre d'observation.*

Les estimateurs de densité que nous utilisons pour les sous-processus ponctuels A et B sont les mêmes que ceux que nous avions utilisés dans le cas univarié défini en (2.4). La fonction intertype inhomogène est *Kross.inhom* en *R*. Ainsi à la Figure 2.10, lorsque nous comparons les sous-processus du processus multitype inhomogène que nous avons générés précédemment, nous constatons que les points du processus sont attirés vers les points du processus B. Notons toutefois que la distribution de chaque sous-processus est bel et bien aléatoire.

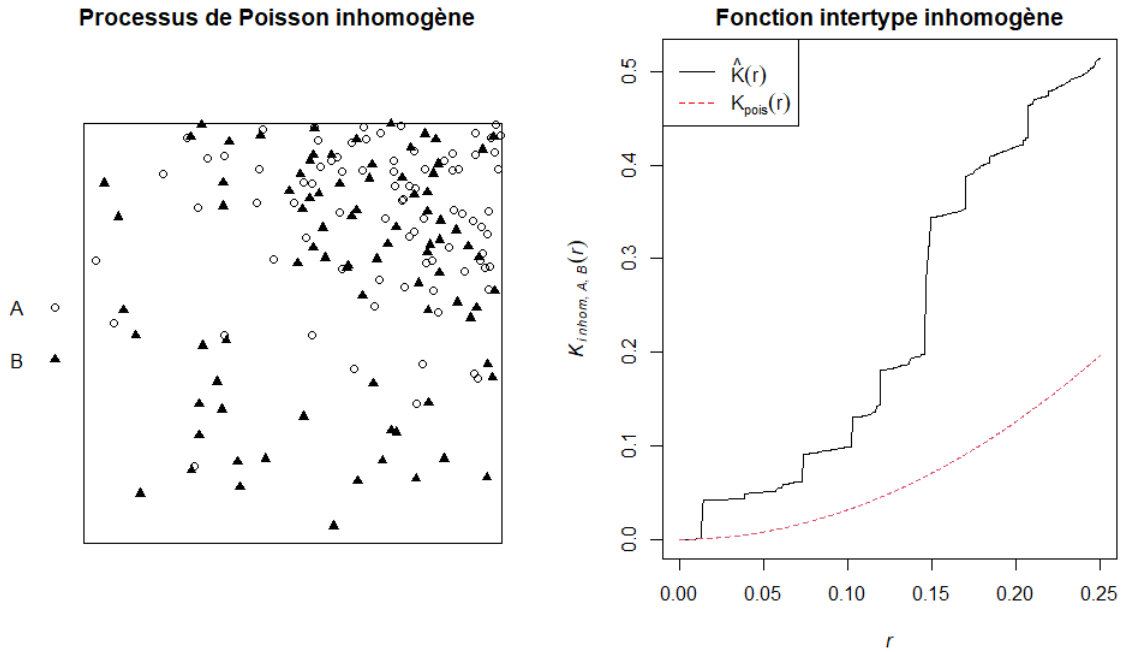

FIGURE 2.10 – Processus multiple de Poisson inhomogène et sa fonction intertype.

Chapitre 3

Données continues

Pour rappel, une variable spatiale se définit comme (X, P) où X est la valeur de la variable et P est la localisation. Lorsque nous travaillons avec des données continues, nous nous concentrons sur la valeur X de la variable vis à vis de sa localisation P.

L'objectif de ce chapitre est de déterminer la valeur de nouveaux emplacements géographique, ce que nous ferons grâce à la méthode du Kriging. Toutefois, afin de pouvoir utiliser cette méthode, une étape intermédiaire intervient qui est la modélisation des données grâce à l'ajustement du variogramme expérimentale. En particulier, ce modèle permet d'étudier les valeurs en fonction de leur position. De plus, nous découvrons les fonctions qu'il faut utiliser en R grâce à la source [20].

3.1 Covariogramme et variogramme

Comme dans le cadre des données ponctuelles, nous travaillons avec une unique réalisation. Si nous effectuons des relevés de température, il n'est pas possible qu'il fasse à la fois 15°C et 17°C à un même endroit. L'espace géographique D dans lequel nous travaillons se nomme *champ*.

Nous supposons que nous avons $(x_1, p_1), \ldots, (x_n, p_n)$ observations spatiales dans le domaine D. Nous introduisons la variable régionalisée $z(p_i)$ qui prend comme argument p_i et renvoie la valeur x_i de cette position pour $i = 1, \ldots, n$. Toutefois, afin de modéliser les données continues, nous utilisons un modèle probabiliste en faisant intervenir la notion de fonction aléatoire sur base des sources [7] et [9]. La fonction aléatoire est $Z(p)$ avec sa réalisation $z(p)$ qui est une variable variant en fonction de sa position p.

A présent, nous introduisons le concept de covariogramme et de variogramme grâce au livre de référence [10]. Ces outils permettent d'étudier le comportement des valeurs de la variable spatiale en fonction de la distance h qui sépare leurs localisations. Toutefois, nous avons une unique réalisation de la fonction aléatoire, nous ne pouvons donc pas caractériser l'espérance, la variance et la covariance de celle-ci. Par conséquent, le livre de référence[27] nous indique d'imposer une condition de stationnarité.

3.1.1 Covariogramme

Comme nous l'avons précisé précédemment. Nous imposons une condition de stationnarité sur la fonction aléatoire $Z(\cdot)$. Afin d'obtenir l'expression du covariogramme,

l'hypothèse de stationnarité d'ordre 2 est mise en place et celle-ci permet de s'assurer de l'existence de l'espérance, la variance et la covariance de la fonction aléatoire.

Définition 3.1
La stationnarité d'ordre 2 impose que l'espérance et la covariance de la fonction aléatoire, $Z(\cdot)$, soient invariantes sous translation, $\forall p \in D$:

- $\mathbb{E}[Z(p)] = \mu$ *tel que* $\mu(p+h) = \mu(p)$;
- $Var(Z(p)) = \mathbb{E}[(Z(p) - \mu)^2] = C(0) = \sigma^2$;
- $Cov[Z(p+h), Z(p)] = \mathbb{E}[(Z(p+h) - \mu)(Z(p) - \mu)]$.

Sous l'hypothèse de stationnarité, nous définissons la fonction de covariance, aussi appelée covariogramme comme

$$C(h) = \mathrm{Cov}[Z(p+h), Z(p)] = \mathbb{E}[(Z(p+h) - \mu)(Z(p) - \mu)].$$

La fonction $C(h)$ dépend du vecteur de translation h qui se définit comme une distance entre deux localisations p_i et p_j. De plus, comme nous l'indique la source [28], elle exprime la variation de la covariance lors de l'accroissement de la distance h. Nous cherchons à mesurer la similarité entre $Z(p+h)$ et $Z(p)$. Notons que la covariance et la variance sont égales lorsque la distance h est nulle : $C(0) = \mathbb{E}[(Z(p) - \mu)^2] = \sigma^2$. De plus, les propriétés de la fonction de covariance sont

- $C(-h) = C(h)$;
- $|C(h)| \leq C(0)$;
- $\mathrm{Var}[\sum_{i=1}^{n} \lambda_i z(p_i)] = \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j C(p_i - p_j) \geq 0$.

Le livre de référence [2] établit la condition d'un covariogramme défini positif afin d'assurer la positivité de la variance. A partir des observations $z(p_1), \ldots, z(p_n)$ de la fonction aléatoire, il est alors possible d'estimer le covariogramme associé.

Définition 3.2
L'estimateur de la fonction de covariance est défini comme

$$\hat{C}(h) = \frac{1}{n(h)} \sum_{i=1}^{n(h)} (z(p_i) - \mu)(z(p_i + h) - \mu),$$

où $n(h)$ est le nombre de paires de localisations (p_i, p_j) telle que la distance entre p_i et p_j est comprise entre $h - \delta h$ et $h + \delta h$.

Nous introduisons un certain intervalle pour h dans lequel la distance de la paire de localisation peut être toujours admise. Toutefois, le covariogramme que nous venons de définir possède certaines limites dues à la forme de stationnarité choisie. Pour rappel, sous stationnarité d'ordre 2, l'espérance et la variance sont constantes sur l'ensemble du domaine et la covariance ne dépend que du vecteur de translation h. Toutefois, dans le cadre des variables régionalisées, il est possible que la dispersion soit infinie ce qui impliquerait que la variance ne soit pas bornée comme nous l'avons imposé précédemment. C'est pour cette raison que nous privilégions la définition du variogramme avec une stationnarité intrinsèque.

3.1.2 Variogramme

Nous fixons une nouvelle condition de stationnarité pour la fonction aléatoire. Cette condition est la stationnarité intrinsèque. Nous privilégions cette stationnarité car celle-ci est plus faible et a donc des conditions d'utilisation moins strictes.

Définition 3.3
La stationnarité intrinsèque impose que l'espérance et la variance de l'accroissement $Z(p+h) - Z(p)$ *ne dépendent pas de la localisation* p, $\forall p \in D$:

- $\mathbb{E}[Z(p+h) - Z(p)] = 0$;
- $Var[Z(p+h) - Z(p)] = \mathbb{E}[(Z(p+h) - Z(p))^2] = 2\gamma(h)$.

Sous l'hypothèse de stationnarité, nous définissons la fonction de variance, appelée aussi variogramme, comme

$$\gamma(h) = \frac{1}{2}\mathbb{E}[(Z(p+h) - Z(p))^2]$$

et son but est de mesurer la discordance entre $Z(p+h)$ et $Z(p)$. Les propriétés du variogramme sont

- $\gamma(h) = \gamma(-h)$;
- $\gamma(0) = 0$;
- $\frac{\gamma(h)}{||h||^2} \to 0$ quand $||h|| \to \infty$.

Notons que sous l'hypothèse de stationnarité intrinsèque, si les coefficients de toutes les combinaisons linéaires respectent la condition $\sum_{i=1}^{n} \lambda_i = 0$, nous avons une propriété supplémentaire qui est

$$\begin{aligned} \text{Var}\left[\sum_{i=1}^{n} \lambda_i z(p_i)\right] &= -\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \gamma(p_i - p_j) \geq 0 \\ &\Leftrightarrow \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \gamma(p_i - p_j) \leq 0. \end{aligned}$$

Cela signifie, d'après la source [2], que le variogramme est conditionnellement défini négatif.

Ensuite, nous définissons la fonction d'autocorrélation $\rho(h)$ où nous établissons une relation entre le covariogramme et le variogramme. La fonction d'autocorrélation se définit comme le rapport $\frac{C(h)}{C(0)}$, dont la valeur est comprise entre -1 et 1, telle que

$$\gamma(h) = C(0) - C(h) \Leftrightarrow \gamma(h) = \sigma^2(1 - \rho(h)),$$

où la fonction $\gamma(h)$ qui intervient se définit comme le semi-variogramme de la fonction aléatoire.

Lors d'une autocorrélation spatiale positive, les valeurs associées à des localisations voisines sont semblables. Par conséquent, si nous augmentons de plus en plus la distance h entre les localisations, les variables régionales vont de plus en plus être différentes. Comme le covariogramme mesure la similarité des variables régionales entre les localisations, celui-ci diminue au fur et à mesure que la distance h augmente en opposition au variogramme qui lui va augmenter comme nous pouvons le voir à la Figure 3.1.

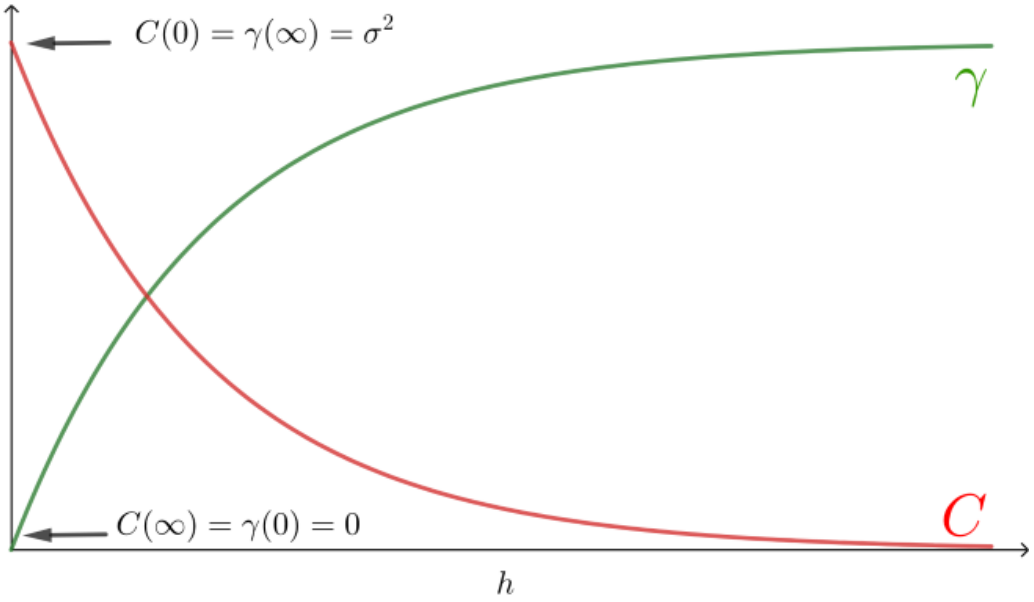

FIGURE 3.1 – Illustration du covariogramme et du variogramme.

A présent, nous étudions les spécificités que peut avoir un variogramme à la Figure 3.2. Nous prenons la forme la plus classique où nous observons que la fonction $\gamma(h)$ croit jusqu'à une certaine valeur qu'on nomme *palier*. La valeur h associée à la valeur du palier se nomme *portée*. A partir de la distance de la *portée*, les variables régionalisées ne sont plus corrélées. Ainsi, la fonction $\gamma(h)$ devient constante et comme la covariance est nulle, nous établissons que la valeur de $\gamma(h)$ après ce palier est de $C(0) = \sigma^2$. Nous avions établit théoriquement que $\gamma(0) = 0$ mais en pratique il se peut que cela ne soit pas le cas. Nous nommons cet écart, *effet pépite*. La cause de cet écart peut être due à une erreur lors de la mesure de la variable ou à une variation importante de la variable elle-même.

Comme pour le covariogramme, il est possible d'estimer le variogramme de la fonction aléatoire $Z(\cdot)$ à partir de n observations $z(p_1), \ldots, z(p_n)$.

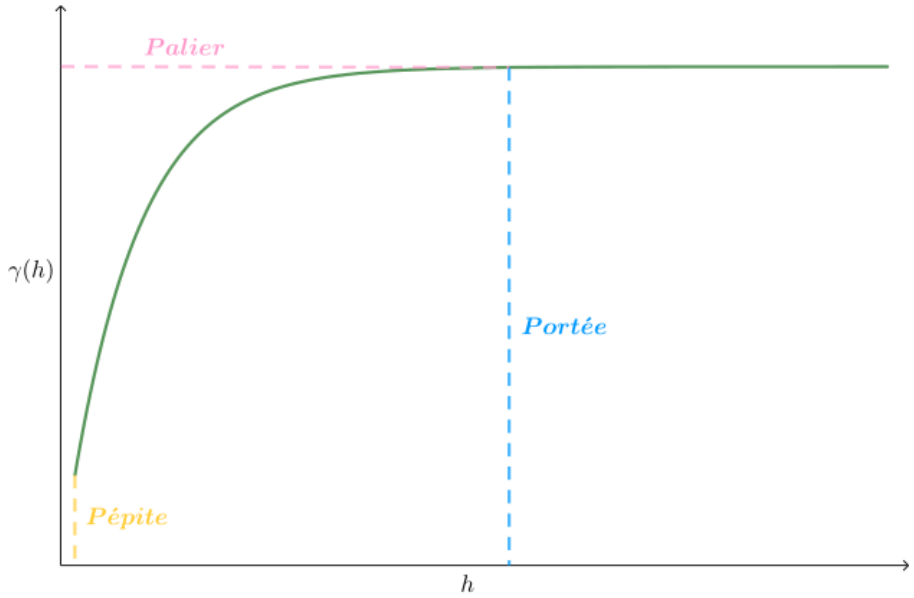

FIGURE 3.2 – Illustration du variogramme.

Définition 3.4
L'estimateur non biaisé du variogramme, nommé variogramme expérimental, est défini comme

$$\hat{\gamma}(h) = \frac{1}{2n(h)} \sum_{i=1}^{n(h)} (z(p_i + h) - z(p_i))^2,$$

où n(h) est le nombre de paires de localisation (p_i, p_j) telle que la distance entre p_i et p_j est comprise entre $h - \delta h$ et $h + \delta h$.

Il est possible que le variogramme expérimental ne respecte pas les propriétés du variogramme théorique que nous avons défini précédemment. C'est pour cette raison que nous introduisons divers modèles afin d'ajuster le variogramme expérimental. Toutefois, nous ne pouvons donc pas choisir n'importe quelle fonction. En effet, il est nécessaire que la fonction $\gamma(h)$ soit conditionnellement définie négative. Cette modélisation est cruciale car celle-ci est utilisée dans la méthode du Kriging.

3.1.3 Modèles

Les sources [7] et [21] présentent les modèles qui sont les plus utilisés avec les valeurs réelles a qui est la *portée*, c qui est le *palier* et c_0 qui est *l'effet pépite*. L'ajustement du variogramme expérimental se construit en R avec différentes étapes :

- Étape 1 : nous générons le variogramme expérimental lié aux observations de la variable régionalisée z grâce à la fonction $variogram(z, data)$ où $data$ est le dataframe contenant la variable.
- Étape 2 : nous choisissons le modèle du variogramme expérimental, $vgm(c, model, a, c_0)$, en considérant un modèle en particulier.
- Étape 3 : nous obtenons l'ajustement du variogramme expérimental en combinant les deux premières étapes par la fonction $fit.variogram(variogram, vgm)$ où $variogram$ est le variogramme obtenu à l'étape 1 et vgm est le modèle construit à l'étape 2.

Définition 3.5
Le modèle pépitique est caractérisé par

$$\gamma(h) = \begin{cases} 0 & pour\ h = 0, \\ c & pour\ h > 0, \end{cases}$$

et représente une indépendance entre les variables.

Définition 3.6
Le modèle sphérique est défini par

$$\gamma(h) = \begin{cases} 0 & pour\ h = 0, \\ c_0 + c[\frac{3}{2}(\frac{h}{a}) - \frac{1}{2}(\frac{h}{a})^3] & pour\ 0 < h \leq a, \\ c_0 + c & pour\ h > a. \end{cases}$$

Définition 3.7
Le modèle exponentiel est spécifié par

$$\gamma(h) = \begin{cases} 0 & pour\ h = 0, \\ c_0 + c[1 - e^{-\frac{h}{a}}] & pour\ h > 0. \end{cases}$$

Définition 3.8
Le modèle Gaussien prend la forme

$$\gamma(h) = \begin{cases} 0 & pour\ h = 0, \\ c_0 + c[1 - e^{-(\frac{h}{a})^2}] & pour\ h > 0. \end{cases}$$

Définition 3.9
Le modèle puissance est déterminé par

$$\gamma(h) = \begin{cases} 0 & pour\ h = 0, \\ c_0 + bh^p & pour\ h > 0, \end{cases}$$

où b est le facteur d'échelle et p est l'exposant.

Afin de pouvoir sélectionner le modèle, nous recourons à la validation croisée [20]. Ce processus "leave-one-out" consiste à enlever une observation $z(p_i)$ de notre ensemble $\{z(p_1), \ldots, z(p_n)\}$ et à calculer la prédiction de cette valeur sur base de notre nouveau sous-ensemble. Ce processus s'exécute pour toutes les valeurs. Nous comparons ainsi les valeurs réelles $z(p_i)$ et les valeurs prédites $\hat{z}(p_i)$ sur base de différents critères,

- biais moyen : $\frac{1}{n}\sum_{i=1}^{n}(\hat{z}(p_i) - z(p_i))$;
- MSE : $\frac{1}{n}\sum_{i=1}^{n}(\hat{z}(p_i) - z(p_i))^2$;
- Critère d'adéquation : $\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{z}(p_i) - z(p_i)}{\sqrt{\mathrm{Var}(\hat{z}(p_i) - z(p_i))}}\right)^2$;
- PRESS : $\sum_{i=1}^{n}(\hat{z}(p_i) - z(p_i))^2$.

Nous choisirons le modèle qui possède la plus petite valeur du critère sélectionné grâce à la fonction $krige.cv(z, data, fit.variogram)$.

3.2 Kriging ordinaire

Afin de réaliser une interpolation de la fonction $Z(\cdot)$ sur une nouvelle position $p_0 \in D$ à partir des observations $\{z(p_1), \ldots, z(p_n)\}$, nous introduisons la méthode du Kriging ordinaire obtenue par $krige(z, data, fit.variogram, grid)$ en R. L'argument $grid$ est la grille sur laquelle nous réalisons l'interpolation. Nous supposons la stationnarité intrinsèque de la fonction aléatoire $Z(\cdot)$ avec le variogramme $\gamma(h)$ associé. La moyenne μ est inconnue. Nous imposons une première contrainte de linéarité sur la prévision, celle-ci doit se formuler comme une combinaison linéaire des observations, $\hat{z}(p_0) = \sum_{i=1}^{n}\lambda_i z(p_i)$.

Ensuite, nous imposons les deux contraintes suivantes :

- contrainte de non biais : $\mathbb{E}[\hat{z}(p_0) - z(p_0)] = 0$ impose que la prédiction de la nouvelle observation est non biaisée ;
- contrainte d'optimalité : l'erreur quadratique moyenne de la prédiction, $\mathrm{Var}[\hat{z}(p_0) - z(p_0)]$, est minimisée avec un choix particulier de $\lambda_1, \ldots, \lambda_n$ et μ.

L'estimateur de Kriging est un estimateur exact qui fournit la valeur exacte de la valeur observée, $\hat{z}(p_i) = z(p_i) \forall p_i \in D$.

Nous déduisons la contrainte de minimisation par la contrainte de non biais,

$$\begin{aligned}\mathbb{E}[\hat{z}(p_0) - z(p_0)] = 0 &\Leftrightarrow \mathbb{E}\left[\sum_{i=1}^{n} \lambda_i z(p_i) - z(p_0)\right] = 0 \\ &\Leftrightarrow \sum_{i=1}^{n} \lambda_i \mathbb{E}\left[z(p_i)\right] - \mathbb{E}[z(p_0)] = 0 \\ &\Leftrightarrow \sum_{i=1}^{n} \lambda_i \mu - \mu = 0 \\ &\Leftrightarrow \mu\left(\sum_{i=1}^{n} \lambda_i - 1\right) = 0 \Leftrightarrow \sum_{i=1}^{n} \lambda_i = 1.\end{aligned}$$

Définition 3.10
Les équations de Kriging ordinaires,

$$\begin{cases} \sum_{j=1}^{n} \lambda_j \gamma(p_i - p_j) + \mu = \gamma(p_0 - p_i), \\ \sum_{i=1}^{n} \lambda_i = 1, \end{cases}$$

sont obtenues par la méthode de Lagrange en minimisant $Var[\hat{z}(p_0) - z(p_0)]$ *sous la contrainte* $\sum_{i=1}^{n} \lambda_i = 1$.

Démonstration
Nous utilisons la méthode de Lagrange , détaillée dans la source [12], pour déterminer les coefficients $\lambda_1, \ldots, \lambda_n, \mu$ qui minimisent la fonction $\text{Var}[\hat{z}(p_0) - z(p_0)] = \mathbb{E}[(\hat{z}(p_0) - z(p_0))^2]$ sous la contrainte $\sum_{i=1}^{n} \lambda_i = 1$. Le Lagrangien s'écrit donc de la manière suivante,

$$\mathcal{L}(\lambda_1, \ldots, \lambda_n, \mu) = \mathbb{E}[(\hat{z}(p_0) - z(p_0))^2] - 2\mu(\sum_{i=1}^{n} \lambda_i - 1).$$

Tout d'abord, nous réécrivons l'argument du terme $\mathbb{E}[(\hat{z}(p_0) - z(p_0))^2]$.

$$\begin{aligned}(\hat{z}(p_0) - z(p_0))^2 &= \left(\sum_{i=1}^{n} \lambda_i z(p_i) - z(p_0)\right)^2 \\ &= \left(\sum_{i=1}^{n} \lambda_i z(p_i)\right)^2 - 2z(p_0) \sum_{i=1}^{n} \lambda_i z(p_i) + z(p_0)^2 \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j z(p_i) z(p_j) - 2z(p_0) \sum_{i=1}^{n} \lambda_i z(p_i) + z(p_0)^2 \\ &= \underbrace{\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j z(p_i) z(p_j) - \sum_{i=1}^{n} \lambda_i z(p_i)^2}_{\text{terme 1}} + \underbrace{\sum_{i=1}^{n} \lambda_i z(p_i)^2 - 2z(p_0) \sum_{i=1}^{n} \lambda_i z(p_i) + z(p_0)^2}_{\text{terme 2}}.\end{aligned}$$

Le premier terme peut se reformuler comme

$$\begin{aligned}\text{terme 1} &= \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j z(p_i) z(p_j) - \sum_{i=1}^{n} \lambda_i z(p_i)^2 \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j z(p_i) z(p_j) - \underbrace{\sum_{j=1}^{n} \lambda_j}_{=1} \sum_{i=1}^{n} \lambda_i z(p_i)^2\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j z(p_i)z(p_j) - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j z(p_i)^2 - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j z(p_j)^2 \\
&= -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j [z(p_i)^2 + z(p_j)^2 - 2z(p_i)z(p_j)] \\
&= -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j [z(p_i) - z(p_j)]^2.
\end{aligned}
$$

Le second terme quant à lui peut se reformuler comme

$$
\begin{aligned}
\text{terme } 2 &= \sum_{i=1}^{n} \lambda_i z(p_i)^2 - 2z(p_0)\sum_{i=1}^{n} \lambda_i z(p_i) + z(p_0)^2 \\
&= \sum_{i=1}^{n} \lambda_i z(p_i)^2 - 2z(p_0)\sum_{i=1}^{n} \lambda_i z(p_i) + \underbrace{\sum_{i=1}^{n} \lambda_i}_{=1} z(p_0)^2 \\
&= \sum_{i=1}^{n} \lambda_i [z(p_i)^2 + z(p_0)^2 - 2z(p_i)z(p_0)] \\
&= \sum_{i=1}^{n} \lambda_i [z(p_0) - z(p_i)]^2.
\end{aligned}
$$

Le variance du Kriging s'écrit alors de la manière suivante

$$
\begin{aligned}
\mathrm{Var}[\hat{z}(p_0) - z(p_0)] &= \mathbb{E}[(\hat{z}(p_0) - z(p_0))^2] \\
&= \mathbb{E}\left[-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j (z(p_i) - z(p_j))^2 + \sum_{i=1}^{n} \lambda_i (z(p_0) - z(p_i))^2\right] \\
&= -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j \mathbb{E}[(z(p_i) - z(p_j))^2] + \sum_{i=1}^{n} \lambda_i \mathbb{E}[(z(p_0) - z(p_i))^2] \\
&= -\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j \gamma(p_i - p_j) + 2\sum_{i=1}^{n} \lambda_i \gamma(p_0 - p_i).
\end{aligned}
$$

L'expression du Lagrangien devient alors

$$
\mathcal{L}(\lambda_1, \ldots, \lambda_n, \mu) = -\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j \gamma(p_i - p_j) + 2\sum_{i=1}^{n} \lambda_i \gamma(p_0 - p_i) - 2\mu\left(\sum_{i=1}^{n} \lambda_i - 1\right).
$$

A présent, nous obtenons les équations du Kriging en annulant le gradient du Lagrangien par rapport aux paramètres $\lambda_1, \ldots, \lambda_n, \mu$

$$
\begin{aligned}
\frac{\partial}{\partial \lambda_i}\mathcal{L}(\lambda_1, \ldots, \lambda_n, \mu) = 0 &\Leftrightarrow -\sum_{j=1}^{n} \lambda_j \gamma(p_i - p_j) + \gamma(p_0 - p_i) - \mu = 0 \\
&\Leftrightarrow \sum_{j=1}^{n} \lambda_j \gamma(p_i - p_j) + \mu = \gamma(p_0 - p_i); \\
\frac{\partial}{\partial \mu}\mathcal{L}(\lambda_1, \ldots, \lambda_n, \mu) = 0 &\Leftrightarrow \sum_{i=1}^{n} \lambda_i - 1 = 0 \\
&\Leftrightarrow \sum_{i=1}^{n} \lambda_i = 1.
\end{aligned}
$$

□

3.3 Géostatistique multivariée

3.3.1 Covariogramme et variogramme croisés

Grâce aux livres de référence [39] et [1] nous nous plaçons à présent dans le cadre multivarié avec r fonctions aléatoires $Z_1(p), \ldots, Z_r(p)$ avec $p \in D$. Notons que chaque variable $Z_j(\cdot)$ possède son ensemble de points r_i pour $j = 1, \ldots, r$ et $i = 1, \ldots, n$. Il est possible que les ensembles de points soient différents ou égaux. Notons que les fonctions en R que nous avons introduites sont toujours exploitables dans le cas mutlivarié comme l'indique [31].

Pour les r fonctions aléatoires, nous définissons le covariogramme croisé sous l'hypothèse de stationnarité d'ordre 2 ainsi que le variogramme croisé sous l'hypothèse intrinsèque de stationnarité afin de pouvoir introduire le co-kriging.

Définition 3.11
La stationnarité d'ordre 2 conjointe impose que l'espérance et la covariance d'une paire de variables sont invariantes sous translation tel que $\forall p \in D$

- $\mathbb{E}[Z_i(p)] = \mu_i$ *pour* $i = 1, \ldots, r$;
- $Cov[Z_i(p), Z_j(p+h)] = \mathbb{E}[(Z_i(p) - \mu_i)(Z_j(p+h) - \mu_j)]$.

Le *covariogramme croisé* de la paire de variables $Z_i(p)$ et $Z_j(p)$ où $p \in D$ se définit comme

$$C_{ij}(h) = \mathrm{Cov}[Z_i(p), Z_j(p+h)] = \mathbb{E}[(Z_i(p) - \mu_i)(Z_j(p+h) - \mu_j)]$$

et mesure la similarité entre $Z_i(p)$ et $Z_j(p)$ en fonction de la distance h. Nous pouvons établir que $C_{ij}(h) = C_{ji}(-h)$ pour $i \neq j$. En effet,

$$\begin{aligned} C_{ji}(-h) &= \mathbb{E}[(Z_j(p) - \mu_j)(Z_i(p-h) - \mu_i)] \\ &= \mathbb{E}[(Z_j(p+h) - \mu_j)(Z_i(p-h+h) - \mu_i)] \\ &= C_{ij}(h). \end{aligned}$$

De plus, nous avons les deux propriétés suivantes : $C_{ij}(h) \neq C_{ji}(h)$ et $C_{ij}(h) \neq C_{ij}(-h)$. De plus, nous établissons que $\forall p \in D$, $C_{ij}(0) = \mathrm{Cov}[Z_i(p), Z_j(p)] = \sigma_{ij}$. La propriété d'être définie positive est toujours valable pour la fonction de covariance croisée,

$$\sum_{i=1}^{r}\sum_{j=1}^{r}\sum_{a=1}^{n}\sum_{b=1}^{n} \lambda_a^i C_{ij}(p_a - p_b)\lambda_b^j \geq 0.$$

La fonction de corrélation $P(h)$ peut être généralisée au cas multivarié et prend la forme

$$C(h) = \Sigma P(h),$$

où

- $C(h) = [C_{ij}(h)]_{i,j=1,\ldots,r}$ est la matrice $r \times r$ des fonctions de covariances ;
- $\Sigma = [\sigma_{ij}]_{i,j=1,\ldots,r}$ est la matrice $r \times r$ variance-covariance.

Définition 3.12
L'estimateur de la fonction de covariance croisée est définie de manière suivante,

$$\hat{C}_{ij}(h) = \frac{1}{r(h)} \sum_{k=1}^{r(h)} (z_i(p_k) - \mu_i)(z_j(p_k + h) - \mu_j),$$

où r(h) est le nombre de paires de localisations (p_k, p_l) telle que la distance entre p_k et p_l est comprise entre $h - \delta h$ et $h + \delta h$.

Définition 3.13
La stationnarité intrinsèque conjointe impose que l'espérance et la variance de l'accroissement $Z_i(p + h) - Z_i(p)$ ne dépendent pas de la localisation $p \in D$:

- $\mathbb{E}[Z_i(p + h) - Z_i(p)] = 0$ *pour* $i = 1, \ldots, r$;
- $\mathbb{E}[(Z_i(p + h) - Z_i(p))(Z_j(p + h) - Z_j(p))] = 2\gamma_{ij}(h)$ *pour* $i = 1, \ldots, r$.

Le variogramme croisé se définit comme

$$\gamma_{ij}(h) = \frac{1}{2}\mathbb{E}[(Z_i(p + h) - Z_i(p))(Z_j(p + h) - Z_j(p))]$$

et mesure la dissimilarité entre $Z_i(p)$ et $Z_j(p)$ en fonction de la distance h. Nous établissons que $\gamma_{ij}(0) = 0$. Nous pouvons établir que $\gamma_{ji}(h) = \gamma_{ij}(h) = \gamma_{ij}(-h)$. En effet,

$$\begin{aligned}
\gamma_{ji}(h) &= \mathbb{E}[(Z_j(p + h) - Z_j(p))(Z_i(p + h) - Z_i(p))] \\
&= \mathbb{E}[(Z_i(p + h) - Z_i(p))(Z_j(p + h) - Z_j(p))] \\
&= \gamma_{ij}(h) \\
&= \mathbb{E}[(Z_i(p + p - h) - Z_i(p - h))(Z_j(p + h - h) - Z_j(p - h))] \\
&= \mathbb{E}[(-1)(Z_i(p) - Z_i(p - h))(-1)(Z_j(p) - Z_j(p - h))] \\
&= \mathbb{E}[(Z_i(p - h) - Z_i(p))(Z_j(p - h) - Z_j(p))] \\
&= \gamma_{ij}(-h).
\end{aligned}$$

Définition 3.14
L'estimateur du variogramme croisé est défini de manière suivante,

$$\hat{\gamma}_{ij}(h) = \frac{1}{2r(h)} \sum_{k=1}^{p(h)} (z_i(p_k + h) - z_i(p_k))(z_j(p_k + h) - z_j(p_k)),$$

où r(h) est le nombre de paires de localisation (p_k, p_l) telle que la distance entre p_k et p_l est comprise entre $h - \delta h$ et $h + \delta h$.

3.3.2 Modèles

Comme dans le cadre univarié, nous cherchons à ajuster l'estimateur du variogramme croisé grâce à l'utilisation d'un modèle. Plusieurs modèles sont introduits, dans les livres de référence [1] et [4], pour le variogramme croisé. La sélection de celui-ci se fera encore une fois grâce à la validation croisée.

En particulier, nous étudions *le modèle de corrégionalisation linéaire*. Nous devons tenir compte de l'ensemble des fonctions aléatoires $Z_i(\cdot)$ pour $i = 1, \ldots, r$ et ne pas réaliser une

analyse individuelle pour chaque fonction. Nous supposons avoir la stationnarité d'ordre 2 conjointe pour les r fonctions aléatoires $Z_1(\cdot), \ldots, Z_r(\cdot)$. La fonction de covariance croisée est réécrite comme une combinaison linéaire des covariances univariées

$$C(h) = \sum_{u=0}^{L} C^u(h) = \sum_{u=0}^{L} B^u \rho^u(h),$$

où

- L est le nombre de champs spatiales indépendants ;
- $\rho^u(h)$ est la fonction de corrélation univariée telle que $\rho^u(0) = 1$;
- La matrice de corrégionalisation $B^u = [b^u_{ij}]_{i,j=1,\ldots,r}$ est une matrice de variance-covariance.

Ainsi, l'expression du variogramme croisé est réécrite de la même façon.

Définition 3.15
Le modèle de corrégionalisation linéaire s'écrit

$$\gamma(h) = \sum_{u=0}^{L} B^u \gamma^u(h),$$

où

- *L est le nombre de champs spatiales indépendants ;*
- *$\gamma^u(h)$ est la fonction du variogramme univariée ;*
- *La matrice de corrégionalisation $B^u = [b^u_{ij}]_{i,j=1,\ldots,r}$ avec $h \in D$ est une matrice de variance-covariance.*

3.3.3 Co-Kriging

Pour terminer, nous introduisons le co-Kriging ordinaire. Le but est de déterminer la valeur de $z_*(p_0)$ à partir des observations $\{z_*(p_1), \ldots, z_*(p_n)\}$ et des autres fonctions aléatoires $Z_i(\cdot)$ pour $i \neq *$. La contrainte de linéarité sur l'estimateur du co-Kriging se présente comme $\hat{z}_*(p_0) = \sum_{i=1}^{r} \sum_{a=1}^{n_i} \lambda^i_a z_i(p_a)$. Les r contraintes de minimisation deviennent

$$\sum_{a=1}^{n_i} \lambda^i_a = \begin{cases} 1 \text{ si } i = *, \\ 0 \text{ sinon.} \end{cases}$$

Quant à la contrainte de non biais, elle se reformule comme $\mathbb{E}[\hat{z}_*(p_0) - z_*(p_0)] = 0$.

Un développement similaire à celui présent dans la section consacrée au univariée est réalisé afin de formuler la variance du co-Kriging comme

$$\begin{aligned} \mathrm{Var}[\hat{z}_*(p_0) - z_*(p_0)] &= \mathbb{E}[(\hat{z}_*(p_0) - z_*(p_0))^2] \\ &= -\sum_{i=1}^{r}\sum_{j=1}^{r}\sum_{a=1}^{n_i}\sum_{b=1}^{n_j} \lambda^i_a \lambda^j_b \gamma_{ij}(p_a - p_b) + 2\sum_{i=1}^{r}\sum_{a=1}^{n_i} \lambda^i_a \gamma_{*i}(p_a - p_0). \end{aligned}$$

Définition 3.16
Les équations de Co-Kriging ordinaires sont

$$\begin{cases} \sum_{j=1}^{r} \sum_{b=1}^{n_j} \lambda^j_b \gamma_{ij}(p_a - p_b) + \mu_i = \gamma_{*i}(p_a - p_0) & \text{pour } i = 1, \ldots, r \text{ et } \alpha = 1, \ldots, n_i, \\ \sum_{a=1}^{n_i} \lambda^i_a = \delta_{*i} & \text{pour } i = 1, \ldots, r. \end{cases}$$

Elles sont obtenues par la méthode de Lagrange en minimisant $Var[\hat{z}_*(p_0) - z_*(p_0)]$ sous les r contraintes de minimisation.

Chapitre 4

Données surfaciques

Pour rappel, lorsque nous travaillons avec des données surfaciques, nous prenons en considération la valeur et la localisation de la variable spatiale. L'analyse se réalise en trois étapes : caractériser les relations entre les variables spatiales, mesurer l'autocorrélation spatiale et estimer le modèle de régression spatiale.

Les deux premières étapes ont été vues dans le chapitre 1. Nous caractérisons les relations entre les variables spatiales grâce au choix d'une certaine matrice de poids W. Afin de vérifier la dépendance spatiale, nous utilisons le test de Moran I. Lors du rejet de l'hypothèse nulle, nous affirmons la présence d'une autocorrélation spatiale. Ensuite il s'agit d'introduire des modèles de régression tenant compte de la composante spatiale. En effet, les modèles classiques de régression détaillés dans l'annexe A ne sont pas toujours fiables avec ce type de données. Pour terminer ce chapitre, nous étudierons les facteurs endogènes qui influencent la criminalité en Belgique et nous justifierons le choix des modèles choisis.

4.1 Modèles de régression

Afin d'introduire les modèles spatiaux, nous partons du modèle de régression linéaire multiple classique.

Définition 4.1
Le modèle de régression linéaire multiple, sous forme matricielle, est donné par

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,k-1} \\ 1 & x_{2,1} & \dots & x_{2,k-1} \\ 1 & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,k-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \Leftrightarrow Y = X\beta + \epsilon$$

où

- Y = *vecteur aléatoire observable composé de n lignes,*
- X = *matrice de constantes connues composé de n lignes et k colonnes tel que* $rg(X) = k$,
- β = *vecteur de paramètres inconnus composée de k colonnes,*
- ϵ = *vecteur aléatoire inobservable composé de n lignes tel que* $\epsilon \overset{iid}{\sim} \mathcal{N}_n(0_{n\times 1}, \sigma^2 I_n)$.

En admettant que le modèle de régression classique soit appliqué aux données spatiales, le coefficient de régression résultant est alors baisé ce qui est causé par le fait que la caractéristique d'autocorrélation des données spatiales n'est pas prise en compte dans le

modèle. L'utilisation de ce modèle n'est alors pas envisageable car celui-ci fournit alors des estimateurs biaisés des paramètres.

Dans le cadre spatiale, il existe trois types d'interactions spatiales possibles sur le modèle de régression : interaction de la variable endogène Y, interaction sur la variable exogène X et interaction sur le terme d'erreur ϵ. Dans un premier temps, nous étudions les modèles spatiaux qui possèdent l'une de ces interactions.

En particulier, lorsque nous avons une interaction spatiale sur la variable à expliqué Y, nous introduisons le modèle *autorégressif spatial*. Les valeurs de Y sont donc influencées par les valeurs voisines de Y, $WY = \sum_{j=1}^{n} w_{ij} y_j$.

Définition 4.2 *Spatial AutoRegression, SAR*
Le modèle SAR est donné par

$$\begin{aligned} Y &= \rho WY + X\beta + \epsilon \\ \Leftrightarrow (I_n - \rho W)Y &= X\beta + \epsilon \end{aligned}$$

où

- *W est une matrice de poids standardisée, c'est-à-dire que $\sum_{j=1}^{n} w_{ij} = 1$;*
- *WY est la variable dépendante décalée par la matrice de poids W ;*
- *ρ est le paramètre évaluant l'interaction entre les valeurs de la variable Y,*
- *$\epsilon \overset{iid}{\sim} \mathcal{N}_n(0_{n\times 1}, \sigma^2 I_n)$.*

Comme l'autocorrélation n'est pas sur le terme d'erreur, nous conservons son hypothèse de normalité. De plus, nous verrons par la suite que la matrice $(I_n - \rho W)$ est non singulière. Le modèle peut donc prendre la forme

$$Y = (I_n - \rho W)^{-1} X\beta + (I_n - \rho W)^{-1}\epsilon.$$

Ainsi, l'espérance et la variance de Y sont

$$\begin{aligned} \mathbb{E}(Y) &= (I_n - \rho W)^{-1} X\beta + (I_n - \rho W)^{-1}\mathbb{E}(\epsilon) \\ &= (I_n - \rho W)^{-1} X\beta \end{aligned}$$

et

$$\begin{aligned} \mathrm{Var}(Y) &= \mathbb{E}[(Y - \mathbb{E}(Y))^2] \\ &= \mathbb{E}[((I_n - \rho W)^{-1} X\beta + (I_n - \rho W)^{-1}\epsilon - (I_n - \rho W)^{-1} X\beta)^2] \\ &= \mathbb{E}[((I_n - \rho W)^{-1}\epsilon)^2] \\ &= \mathbb{E}(\epsilon\epsilon^T)(I_n - \rho W)^{-1}(I_n - \rho W^T)^{-1} \\ &= \sigma^2 I_n (I_n - \rho W)^{-1}(I_n - \rho W^T)^{-1}. \end{aligned}$$

Nous établissons que $Y \sim \mathcal{N}_n((I_n - \rho W)^{-1} X\beta, \sigma^2 I_n (I_n - \rho W)^{-1}(I_n - \rho W^T)^{-1})$.

Comme dans le cadre de la régression classique, nous voulons trouver un estimateur pour les paramètres β et σ^2 du modèle. Toutefois, nous avons un paramètre supplémentaire à estimer qui est ρ.

Une première intuition serait d'utiliser la méthode des moindres carrés afin d'obtenir un estimateur $\hat{\beta}$ qui serait $BLUE$. Néanmoins, nous ne pouvons pas utiliser cette méthode en raison de la corrélation spatiale sur la variable endogène. En effet, le terme d'erreur et la variable endogène décalée sont corrélés ce qui induit un terme d'erreur non sphérique

$$\begin{aligned}\text{Cov}(WY, \epsilon) &= \mathbb{E}(WY\epsilon^T) \\ &= \mathbb{E}[W(I_n - \rho W)^{-1}X\beta\epsilon^T + W(I_n - \rho W)^{-1}\epsilon\epsilon^T] \\ &= W(I_n - \rho W)^{-1}X\beta\mathbb{E}(\epsilon^T) + W(I_n - \rho W)^{-1}\mathbb{E}(\epsilon\epsilon^T) \\ &= \sigma^2 I_n W(I_n - \rho W)^{-1} \neq 0\end{aligned}$$

Par conséquent, nous utilisons la méthode du maximum de vraisemblance afin d'obtenir les estimateurs des paramètres. Alors, l'estimateur ne possède pas la propriété $BLUE$ que nous avions lors de l'utilisation des moindres carrés. Toutefois, sur base du cours de Statistiques [38], nous pouvons établir certaines propriétés de la qualité de l'estimateur sous les hypothèses de régularité. Ainsi, l'estimateur du maximum de vraisemblance est convergent, invariant, asymptotiquement efficace et asymptotiquement distribué selon une normale. Nous verrons par la suite que ces propriétés sont primordiales lors de la création des tests de spécifications dans la section 4.1.1. Par ailleurs, d'après l'article [25], afin de pouvoir calculer l'estimateur des paramètres du modèle spatial, nous devons imposer quelques hypothèses supplémentaires

— $\epsilon \overset{iid}{\sim} \mathcal{N}_n(0_{n\times 1},\, \sigma^2 I_n)$,
— la matrice X est de rang plein et non singulière,
— les éléments diagonaux de W sont nuls,
— la matrice $(I_n - \rho W)$ est non singulière pour toute valeur de $\rho \in [-1, 1]$,
— les colonnes et les lignes des matrices W et $(I_n - \rho W)^{-1}$ sont bornées.

Notons que la fonction de vraisemblance générale prend la forme

$$L(e) = (2\pi\sigma^2)^{-\frac{n}{2}} e^{\frac{1}{2\sigma^2}e^T e} \left|\frac{\partial e}{\partial Y}\right|$$

Avec la formulation du terme d'erreur $\epsilon = (I_n - \rho W)Y - W\beta$, nous maximisons la log-vraisemblance

$$\begin{aligned}\log\, L(\beta, \rho, \sigma^2) &= -\frac{n}{2}ln(2\pi) - \frac{n}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}[\epsilon^T\epsilon] + ln|I_n - \rho W| \\ &= -\frac{n}{2}ln(2\pi) - \frac{n}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}\Big([(I_n - \rho W)Y - W\beta]^T\,[(I_n - \rho W)Y - W\beta]\Big) \\ &\quad + ln|I_n - \rho W|\end{aligned}$$

et calculons les extrema de la fonction

$$\begin{cases} \frac{\partial \log\, L(\beta,\sigma^2,\rho)}{\partial\beta} = 0 \Rightarrow \hat{\beta} = (X^TX)^{-1}X^T(I_n - \rho W)Y \\ \frac{\partial \log\, L(\beta,\sigma^2,\rho)}{\partial\sigma^2} = 0 \Rightarrow \hat{\sigma}^2 = \frac{1}{n}\Big([(I_n - \rho W)Y - W\hat{\beta}]^T\left[(I_n - \rho W)Y - W\hat{\beta}\right]\Big) \end{cases}$$

en supposant que ρ soit connu. Ensuite, nous remplaçons les estimateurs des paramètres β et σ^2 dans l'expression de la fonction de vraisemblance afin que celle-ci ne dépende plus que du paramètre ρ,

$$\log\, L(\rho) = -\frac{n}{2}[ln(2\pi) + 1] + ln|I_n - \rho W| - \frac{n}{2}ln(\hat{\sigma}^2).$$

Par conséquent, l'estimateur de ρ peut être déterminé par une méthode numérique avec un certain critère de convergence (Newton).

Lors d'une interaction spatiale sur la variable exogène X, le modèle à *interactions exogènes* est mis en place. Soit la valeur y_i de Y considérée, celle-ci est influencée par la valeur de X prise dans la $i$ème zone géographique ainsi les autres valeurs de X prises dans les régions voisines.

Définition 4.3 *Spatial Lag X, SLX*
Le modèle SLX est donné par

$$Y = X\beta + WX\theta + \epsilon,$$

où WX est la variable explicative décalée par la matrice de poids W et θ est le paramètre évaluant l'autocorrélation entre les valeurs de X et Y.

Dans le cas de ce modèle, nous observons que le terme d'erreur est sphérique. Nous pouvons donc utiliser la méthode des moindres carrés afin d'obtenir l'estimateur $\hat{\beta}$ qui sera $BLUE$.

Lors d'une corrélation spatiale sur le terme d'erreur ϵ, nous utilisons le modèle *d'erreurs spatialement autocorrélées*. Les valeurs de Y sont influencées par le terme d'erreur et les valeurs voisines de celui-ci.

Définition 4.4 *Spatial Error Model, SEM*
Le modèle SEM est donné par

$$Y = X\beta + \epsilon,\ \epsilon = \lambda W\epsilon + u,$$

où

- *W est une matrice de poids standardisée, c'est-à-dire que $\sum_{j=1}^{n} w_{ij} = 1$,*
- *λ est le paramètre évaluant l'autocorrélation entre les résidus,*
- *$u \overset{iid}{\sim} \mathcal{N}_n(0_{n\times 1},\, \sigma^2 I_n)$.*

Comme pour le modèle SAR, l'interaction spatiale sur le terme d'erreur induit un terme d'erreur non sphérique. C'est pour cette raison que nous utilisons la méthode du maximum de vraisemblance afin d'obtenir les estimateurs du modèle. De plus, une méthode numérique est utilisée pour l'estimateur de λ.

Précédemment, la notion spatiale intervenait soit sur la variable dépendante, indépendante soit le terme d'erreur. Toutefois, il est possible d'avoir un modèle qui détient plus d'une interaction spatiale comme nous le voyons à la definition 4.5.

Définition 4.5
Le modèle de Manski est donné par

$$Y = \rho WY + X\beta + WX\theta + \epsilon,\ \epsilon = \lambda W\epsilon + u,$$

le modèle SDM, Spatial Durbin Model, est donné par

$$Y = \rho WY + X\beta + WX\theta + \epsilon,$$

le modèle SDEM, Spatial Durbin Error Model, est donné par

$$Y = X\beta + WX\theta + \epsilon,\ \epsilon = \lambda W\epsilon + u,$$

Le modèle SAC, Spatial Autoregressive Consused, est donné par

$$Y = \rho WY + X\beta + \epsilon,\ \epsilon = \lambda W\epsilon + u,$$

où

- W *est une matrice de poids standardisée, c'est-à-dire que* $\sum_{j=1}^{n} w_{ij} = 1$,
- WY *est la variable dépendante décalée par la matrice de poids* W,
- WX *est la variable indépendante décalée par la matrice de poids* W,
- ρ *est le paramètre évaluant l'interaction entre les valeurs de la variable* Y,
- θ *est le paramètre évaluant l'autocorrélation entre les valeurs de* X *et* Y,
- λ *est le paramètre évaluant l'autocorrélation entre les résidus,*
- $u \overset{iid}{\sim} \mathcal{N}_n(0_{n\times 1},\, \sigma^2 I_n)$.

Afin de pouvoir utiliser ces modèles en R, nous utilisons les fonctions reprises à la Table 4.1 qui proviennent du package *spatialreg*. De manière générale, les arguments d'une fonction de régression spatiale sont : *modele* qui correspond à la relation entre la variable endogène et exogène ; *data* qui est le dataset comprenant les valeurs des variables et W la matrice de poids. Lorsque nous travaillons avec plus d'une interaction spatiale, nous le spécifions en R grâce à l'argument "type".

TABLE 4.1 – Modèles de régression spatiale et leurs fonctions en R associées

Modèle	SAR	SLX	SEM	Manski	SDM	SDEM	SAC
Fonction en R	lagsarlm	lmSLX	errorsarlm	sacsarlm	lagsarlm	errorsarlm	sacsarlm

Comme nous l'avons vu dans le cadre des modèles avec une interaction spatiale, nous utilisons la méthode du maximum de vraisemblance pour déterminer les paramètres du modèle. En particulier, les paramètres liés à l'interaction spatiale, comme λ, ρ et θ, sont obtenus de manière numérique. La procédure est identique dans le cas de plus d'une interaction spatiale [33]. Par ailleurs, le package *spatialreg* utilise cette procédure afin d'obtenir nos estimateurs.

4.1.1 Tests de spécification

En pratique, nous devons sélectionner le modèle le plus représentatif de nos données. Une première intuition est d'utiliser le test de Moran qui détecte la présence de l'autocorrélation spatiale. Toutefois, ce test ne précise pas le type de l'interaction spatiale présente sur nos données. C'est pour cette raison que, sur base de la source [26], nous introduisons les *tests de spécification* . Lors du calcul du maximum de vraisemblance pour le modèle SAR, nous avions imposé les hypothèses de régularité et nous avions obtenu des propriétés intéressantes pour l'estimateur du maximum de vraisemblance. Sur base de ces propriétés, nous pouvons établir le test du ratio de vraisemblance et du multiplicateur de Lagrange.

L'objectif est de tester la présence ou non de contrainte sur le modèle. Sur base du livre de référence [19], nous reprenons les formes théoriques de chaque test avec le problème de test suivant,

$$\begin{cases} \mathcal{H}_0 : \text{le modèle est restreint sous } g \text{ contraintes,} \\ \mathcal{H}_1 : \text{le modèle n'est pas restreint.} \end{cases}$$

Nous supposons que nous calculons l'estimateur du maximum de vraisemblance des paramètres ξ du modèle que nous notons $\hat{\xi}$. De plus, les contraintes sur les paramètres du modèle s'écrivent $f(\xi)$. Commençons par le test du rapport de vraisemblance. Le ratio de vraisemblance se formule par

$$r = \frac{L(\hat{\xi}_c)}{L(\hat{\xi}_{nc})}$$

où $\hat{\xi}_{nc}$ représente l'estimateur du maximum de vraisemblance de ξ lorsque le modèle n'est pas contraint et réciproquement pour $\hat{\xi}_c$ avec le modèle contraint. Sous les hypothèses de régularité et l'hypothèse nulle, nous établissons la statistique de test

$$LR = -2ln(r) \sim \chi^2_g.$$

Nous observons que nous avons dû calculer l'estimateur du maximum de vraisemblance pour le modèle contraint et non contraint. Une alternative est le test du multiplicateur de Lagrange qui calcule uniquement l'estimateur lié au modèle contraint. L'aspiration est de maximiser la fonction du log vraisemblance associée au modèle contraint. Le Lagrangien s'écrit donc de la forme suivante

$$ln\mathcal{L}(\xi) = lnL(\xi) + \lambda^T(f(\xi) - g),$$

où λ est le vecteur reprenant les multiplicateurs de Lagrange. A présent, nous annulons le gradiant du Lagrangien par rapport aux paramètres ξ et λ :

$$\begin{aligned} \frac{\partial}{\partial\lambda} ln\mathcal{L} = 0 \Leftrightarrow f(\xi) - g = 0 \\ \frac{\partial}{\partial\xi} ln\mathcal{L} = 0 \Leftrightarrow \frac{\partial}{\partial\hat{\xi}_c} lnL(\hat{\xi}_c) = -C^T\lambda. \end{aligned}$$

où $C = \left[\frac{\partial f(\hat{\xi})}{\partial\hat{\xi}^T}\right]$. Sous les hypothèses de régularité et l'hypothèse nulle, la statistique de test est

$$LM = [\frac{\partial}{\partial\hat{\xi}} lnL(\hat{\xi}_c)]^T [I(\hat{\xi}_c)]^{-1} \frac{\partial}{\partial\hat{\xi}_c} lnL(\hat{\xi}_c) \sim \chi^2_g,$$

où $I(\hat{\xi}_c)$ est la matrice d'information de Fisher et $\frac{\partial}{\partial\hat{\xi}_c} lnL(\hat{\xi}_c)$ se nomme le vecteur score.

Afin de choisir parmi l'ensemble des modèles, une première approche est celle dite ascendante. Nous supposons que le modèle contraint est le modèle de régression linéaire. Dans un premier temps, nous utilisons le test du multiplicateur de Lagrange afin de savoir si l'interaction spatiale s'effectue sur le terme d'erreur (LM_λ) ou sur la variable endogène (LM_ρ). Ainsi, le modèle non contraint prend la forme soit du modèle *SEM* soit du modèle *SAR*. L'hypothèse nulle est donc soit $\mathcal{H}_0 : \lambda = 0$ ou soit $\mathcal{H}_0 : \rho = 0$. Comme nous fixons une seule contrainte, les deux tests suivent une loi de chi carré de degré 1.

Toutefois, une version plus robuste de ces tests est mise en place lors de la présence d'une deuxième autocorrélation spatiale. En effet, le livre de référence [3] motive ce choix en raison du fait que la statistique de test ne suit plus une loi chi carré. Considérons la double contrainte $\mathcal{H}_0 : \lambda = 0 = \rho$ où le modèle contraint est toujours le modèle de régression linéaire. Le modèle non contraint n'est pas spécifié. Comme nous avons deux contraintes, la statistique de ce test, que nous nommons $SARMA$, suit une loi de chi carré avec deux degrés de liberté. De plus, nous calculons l'alternative plus robuste des tests LM_λ et LM_ρ présentée précédemment. Considérons la présence de l'interaction spatiale ρ, l'hypothèse nulle est $\mathcal{H}_0 : \lambda = 0$ avec comme modèle non contraint, le modèle *SEM*. La présence d'une unique contrainte implique que la statistique de test RLM_λ suit une loi de chi carré avec un degré de liberté. Le raisonnement est similaire pour la statistique de test RLM_ρ avec la présence de l'interaction spatiale λ, l'hypothèse nulle est $\mathcal{H}_0 : \rho = 0$, le modèle non contraint est le modèle *SAR*. Afin de réaliser ces tests en *R*, nous utilisons la fonction *lm.LMTests* qui prend en argument le modèle de régression linéaire et une matrice de poids.

Cependant, les deux approches les plus conseillées d'utilisation sont l'arbre de décision dit "mixte" et "descendante" présentées dans le livre de référence [16]. Le test du maximum de vraisemblance permet de comparer un modèle contraint et un modèle non contraint. Dans l'approche "mixte", nous effectuons une comparaison entre le modèle *SAR* et *SDM* lors du test $\mathcal{H}_0 : \theta = 0$. De plus, le test $\mathcal{H}_0 : \theta = -\rho\beta$ compare le modèle *SEM* et *SDM*. En *R*, nous utilisons la fonction *LR.Sarlm* qui prend en argument le modèle contraint et le modèle non contraint. L'ensemble des modèles que nous avons défini de manière théorique n'est pas repris dans le reste de l'analyse. En effet, cette approche permet d'adopter les modèles les plus courants d'utilisation en pratique qui sont les modèles *SAR*, *SDM*, *SLX*, *SEM* et le modèle linéaire. La loi pour l'ensemble des statistiques de test utilisées dans cette approche est une loi chi carré avec un degré de liberté sauf pour le test $SARMA$ qui a deux degrés de liberté. Les étapes de l'approche mixte sont :

1. Étape 1 : test $SARMA$
   (a) $\mathcal{H}_0 : \lambda = \rho = 0$;
   (b) modèle contraint : modèle linéaire ;
   (c) modèle non contraint : *SLX* ;
   (d) Si rejet de l'hypothèse nulle, test LR_θ :
      i. $\mathcal{H}_0 : \theta = 0$;
      ii. modèle contraint : modèle linéaire ;
      iii. modèle non contraint : *SDM* ;
      iv. si rejet de l'hypothèse nulle, test LR_ρ :
         A. $\mathcal{H}_0 : \rho = 0$;
         B. modèle contraint : *SLX* ;

C. modèle non contraint : *SDM*.

2. Étape 2 : test RLM_ρ
   (a) $\mathcal{H}_0 : \rho = 0$ avec $\lambda \neq 0$;
   (b) modèle contraint : modèle linéaire ;
   (c) modèle non contraint : *SDM* ;
   (d) Si rejet de l'hypothèse nulle, test LR_θ :
      i. $\mathcal{H}_0 : \theta = 0$;
      ii. modèle contraint : *SAR* ;
      iii. modèle non contraint : *SDM* ;
3. Étape 3 : test RLM_λ
   (a) $\mathcal{H}_0 : \lambda = 0$ avec $\rho \neq 0$;
   (b) modèle contraint : modèle linéaire ;
   (c) modèle non contraint : *SDM* ;
   (d) Si rejet de l'hypothèse nulle, test $LR_{\theta+\rho\beta}$:
      i. $\mathcal{H}_0 : \theta = -\rho\beta$;
      ii. modèle contraint : *SEM* ;
      iii. modèle non contraint : *SDM*.

Les étapes de l'approche descendante sont :

1. Étape 1 : test LR_λ
   (a) $\mathcal{H}_0 : \lambda = 0$;
   (b) modèle contraint : modèle linéaire ;
   (c) modèle non contraint : *SEM* ;
   (d) Si rejet de l'hypothèse nulle, test $LR_{\theta+\rho\beta}$:
      i. $\mathcal{H}_0 : \theta + \rho\beta = 0$;
      ii. modèle contraint : *SEM* ;
      iii. modèle non contraint : *SDM*.
2. Étape 2 : test LR_ρ
   (a) $\mathcal{H}_0 : \rho = 0$;
   (b) modèle contraint : modèle linéaire ;
   (c) modèle non contraint : *SAR* ;
   (d) Si rejet de l'hypothèse nulle, test LR_θ :
      i. $\mathcal{H}_0 : \theta = 0$;
      ii. modèle contraint : *SAR* ;
      iii. modèle non contraint : *SDM*.
3. Étape 3 : test LR_θ
   (a) $\mathcal{H}_0 : \theta = 0$;
   (b) modèle contraint : modèle linéaire ;
   (c) modèle non contraint : *SLX* ;
   (d) Si rejet de l'hypothèse nulle, test LR_ρ :
      i. $\mathcal{H}_0 : \rho = 0$;
      ii. modèle contraint : *SLX* ;
      iii. modèle non contraint : *SDM*.

Afin de réaliser une première comparaison entre les modèles, les facteurs suivants sont utilisés

— AIC (*Akaike information criterion*) : $-2ln(L) + 2k$;

— BIC (*Bayes information criterion*) : $-2ln(L) + k\ ln(n)$;

où

— k est le nombre de paramètres du modèle ;

— n est le nombre d'observations ;

— L est la fonction de vraisemblance du modèle.

Le modèle comprenant le plus petit AIC et BIC est privilégié. Si nous devons choisir entre le facteur AIC et BIC, nous favorisons le facteur AIC en raison de sa robustesse. Nous pouvons nous questionner sur la qualité de celui-ci. Les fonctions correspondantes en R sont *AIC(model)* et *BIC(model)*. Dans le cadre de la régression linéaire classique, nous avions introduit le coefficient de détermination R^2 ajusté. Toutefois, celui-ci n'est pas exploité pour les modèles ayant une interaction spatiale sur ρ ou λ en raison de l'autocorrélation spatiale. Nous introduisons donc le coefficient *Pseudo* $R^2 = 1 - \frac{lnL}{lnL_0}$ où L est la fonction de vraisemblance du modèle et L_0 est la fonction de vraisemblance pour un modèle nul. Un modèle nul indique qu'aucun régresseur explique la variable dépendante et s'exprime donc uniquement par le terme constant. Plus le modèle est qualitatif alors plus ce coefficient sera proche de 1.

4.2 Analyse de la criminalité

Dans cette section, nous analysons les relevés de certains délits pour l'ensemble des communes de la Belgique en 2015. Nous avons relevé les délits de coups et blessures, d'assassinats, de cyberharcelment et de personnes disparues. L'ensemble de ces données ont été obtenues sur le site de la police fédérale en Belgique [30].

Certains facteurs comme des facteurs sociaux, économiques, démographiques,.. ont tendance à influencer la criminalité comme nous l'indiquent les sources [15], [41] et [35]. C'est pourquoi, les facteurs exogènes que nous avons récoltés sont le groupe d'âge, le sexe, le nombre de divorce (X_1), le nombre de chômeurs (X_9), le nombre d'employés (X_{10}), le revenu moyen (X_{11}), la densité de population (X_8), le nombre de familles monoparentale (X_{13}) et le nombre de personnes possédant un diplôme supérieur (X_{12}) pour l'ensemble des communes de la Belgique en 2015. En particulier, nous avons le nombre de femmes et hommes de moins de 18 ans (X_2 et X_3), le nombre de femmes et hommes entre 18 et 64 ans (X_4 et X_5) et le nombres de femmes et hommes de plus de 65 ans (X_6 et X_7). L'ensemble de ces données a été obtenu sur le site Statbel[34] et Iweps[22]. Toutefois, il est important de notifier que les deux derniers facteurs ont été obtenus à partir des indicateurs statistiques du census 2011. En effet, la Belgique créa plusieurs facteurs géographiques de la population en 2011. Par conséquence, les facteurs n'ont pas été mis à jour depuis cette date et fournissent une idée approximative du nombre de familles monoparentale et du nombre de personnes ayant un diplôme supérieur.

Afin de mieux comprendre les variables endogènes, nous prenons les mesures suivantes : la médiane, les quartiles, le minimum et le maximum à la Table 4.2.

TABLE 4.2 – Descriptif des variables endogènes

Y	Minimum	Médiane	Maximum	25%	75%
Meutres	0	1	113	0	2
Blessures	0	49	4918	26	106
Cyberharcelement	0	6	479	2	14
Personnes disparues	0	4	456	2	10

Nous observons que le minimum pour l'ensemble des variables est de zéro ce qui correspond au fait que ce type de délits ne s'est pas produit dans certaines communes. Nous réalisons l'analyse pour la variable liée au nombre de coups et blessures. L'analyse est identique pour le reste des variables. Nous observons que le nombre de coups et blessures est

— inférieur ou égale à 49 dans 50% des communes,

— inférieur ou égale à 26 dans 25% des communes,

— inférieur ou égale à 106 dans 75% des communes.

Pour chaque variable endogène, nous testons la présence d'autocorrélation spatiale grâce au test de Moran. Ainsi, nous pouvons en déduire la présence ou non d'une autocorrélation spatiale et choisir entre le modèle linéaire classique ou parmi un des modèles spatiales définis précédemment. Il est évident que le choix de la matrice de poids va influencer notre analyse. C'est pourquoi, nous en choisissons six différentes, dont la matrice de contiguïté, des 2,5,10 plus proches voisins, l'inverse et la standardisée sur les lignes. Notons que lors de l'ensemble de nos tests d'hypothèses, nous imposons que le seuil d'acceptation soit de 0.05

Afin d'évaluer la performance du modèle choisis, nous étudions le facteur pseudo R^2, le facteur AIC et le facteur BIC.

4.2.1 Assassinats

Dans un premier temps, nous utilisons l'approche mixte afin de déterminer le choix du modèle à la Table 4.3. Nous observons que la p-valeur associée au test de Moran est presque toujours inférieure au seuil d'acceptation ce qui induit le rejet de l'hypothèse nulle. Le test détecte de l'autocorrélation spatiale pour l'ensemble des matrices de poids sauf pour celles des deux voisins les plus proches. Ensuite, nous observons que l'hypothèse nulle est toujours acceptée pour les tests de $SARMA$, RLM_ρ et RLM_λ ce qui conduit au choix du modèle contraint qui est le ***modèle de régression linéaire***.

TABLE 4.3 – Les p-valeurs des tests de spécifications pour l'approche mixte

Tests \ Matrices	2 voisins	5 voisins	10 voisins	contiguë	inverse	standardisée
Moran	0.006291	0.00423	9.569×10^{-6}	3.132×10^{-5}	3.586×10^{-5}	4.14×10^{-15}
$SARMA$	0.9043	0.8266	0.1794	0.1158	0.3414	0.1215
RLM_ρ	0.6539	0.7514	0.09579	0.1622	0.1466	0.1537
RLM_λ	0.8365	0.7354	0.1464	0.2889	0.4195	0.318

Ensuite, nous utilisons l'approche descendante à la Table 4.4. Nous observons que les résultats obtenus pour l'ensemble des tests de spécifications sont équivalents pour l'ensemble des matrices de poids. Lors de la première et de la deuxième étape de l'approche descendante, nous observons que les hypothèses nulles des tests LR_λ et LR_ρ sont acceptées ce qui induit l'utilisation du modèle contraint qui est ***le modèle linéaire***. Toutefois lors de la troisième étape, l'hypothèse nulle du test LR_θ est rejetée et ensuite celle du test LR_ρ avec $\theta \neq 0$ est majoritairement acceptée ce qui indique l'utilisation du modèle ***SLX***.

TABLE 4.4 – Les p-valeurs des tests de spécifications pour l'approche descendante

Tests \ Matrices	2 voisins	5 voisins	10 voisins	contiguë	inverse	standardisée
Moran	0.006291	0.00423	9.569×10^{-6}	3.132×10^{-5}	3.586×10^{-5}	4.14×10^{-15}
LR_λ	0.9878	0.5726	0.3747	0.1507	0.8146	0.1676
$LR_{\theta+\rho\beta}$	x	x	x	x	x	x
LR_ρ	0.6984	0.6082	0.2454	0.09693	0.2142	0.09578
$LR_{\rho\theta}$	x	x	x	x	x	x
LR_θ	0.0002051	0.00665	1.49×10^{-5}	0.03391	4.068×10^{-13}	0.0413
$LR_{\theta\rho}$	0.9772	0.1739	0.02029	0.9383	0.3663	0.999

Deux modèles sont ressortis lors de notre analyse : le modèle linéaire et le modèle *SLX*. Afin de choisir le modèle qui convient le mieux, nous utilisons les critères de *AIC* et *BIC* ainsi que le coefficient pseudo R^2 pour déterminer la qualité du modèle. Nous comparons les différents facteurs pour l'ensemble des matrices de poids pour le modèle *SLX* à la Table 4.5. De plus, voici les différentes mesures pour le modèle linéaire :

— *AIC* : 2409.094 ;

— *BIC* : 2474.566 ;

— R^2 ajusté : 0.9104 .

TABLE 4.5 – Critères qualitatifs pour le modèle *SLX*.

Facteurs \ Matrices	2 voisins	5 voisins	10 voisins	contiguë	inverse	standardisée
AIC	2396.164	2406.143	2389.213	2411.385	2347.254	2412.063
BIC	2518.377	2528.356	2511.426	2533.598	2473.832	2534.276
R^2 ajusté	0.9143	0.9128	0.9153	0.912	0.9213	0.9119

Nous observons que le modèle *SLX* avec la matrice inverse possède les plus petits facteurs *AIC* et *BIC* ainsi que le R^2 ajusté le plus proche de 1. A présent que nous avons choisi notre modèle et sa matrice de poids associée, nous utilisons la fonction *summary* afin d'obtenir les informations de notre modèle qui sont reprises à la Table 4.6. Comme dans le cadre de la régression linéaire classique, nous observons la p-valeur associée au test de Fisher afin d'accepter ou de refuser l'hypothèse linéaire générale. La p-valeur étant de 2.2×10^{-16}, l'hypothèse nulle est rejetée et nous concluons qu'au moins un des régresseurs explique la variable dépendante. Par conséquent, nous effectuons un test individuel sur chaque régresseur afin de déterminer si celui-ci permet d'expliquer la variable Y. De plus, comme nous avons une interaction spatiale sur la variable des régresseurs X, notons que nous avons une estimation du paramètre pour la variable décalée spatialement WX.

TABLE 4.6 – Estimations des paramètres du modèle *SLX*.

Régresseurs	$\hat{\beta}$	p-valeur	$\hat{\theta}$	p-valeur
terme constant	-3.008654	0.004497	0.1162421	0.957697
X_1	-0.0004167357	0.743776	0.0007669059	0.615232
X_2	-0.002628011	0.053854	-0.004028357	0.201913
X_3	0.004213696	0.001333	0.004293584	0.130602
X_4	-0.001796957	0.000188	0.005275826	1.82×10^{-9}
X_5	0.001329153	0.007260	-0.006110624	8.10×10^{-10}
X_6	6.120818×10^{-6}	0.992779	0.000294174	0.820598
X_7	-0.0002991035	0.741862	-0.0007279978	0.677845
X_8	0.0001252251	0.234985	-0.0004718634	7.30×10^{-7}
X_9	0.0009613887	6.68×10^{-6}	0.001877683	3.39×10^{-7}
X_{10}	4.795746×10^{-5}	0.766698	0.0007234126	0.004159
X_{11}	0.0001638546	0.004781	-1.692526×10^{-5}	0.880835
X_{12}	1.773554×10^{-5}	0.794848	-0.0001800435	0.189628
X_{13}	-0.0003868126	0.293536	-0.003671481	6.47×10^{-8}

Ainsi, les régresseurs et régresseurs décalés qui expliquent le nombre d'assassinats en Belgique en 2015 sont

— le nombre d'hommes de moins de 18 ans ;

— le nombre de femmes et d'hommes entre 18 et 64 ans ainsi que sa variante spatialement décalée ;

— la variante spatialement décalée de la densité de population ;

— le nombre de chômeurs ainsi que sa variante spatialement décalée ;

— la variante spatialement décalée du nombre d'employés ;

— le revenu moyen ;

— la variante spatialement décalée du nombre de familles monoparentale.

4.2.2 Coups et blessures

Tout d'abord, nous observons à la Table 4.7 que la p-valeur associée au test de Moran est inférieure au seuil, sauf pour la matrice des deux voisins les plus proches, ce qui met en évidence une autocorrélation spatiale. La première approche que nous utilisons est l'approche mixte. Nous remarquons ici que les hypothèses nulles associées aux tests de $SARMA$, RLM_ρ et RLM_λ sont acceptées pour les matrices avec les voisins les plus proches ce qui induit l'utilisation du **modèle linéaire**.

Toutefois pour la matrice contiguë et standardisée, l'hypothèse nulle du test de $SARMA$ est rejetée ainsi que celle du test LR_θ. De plus, comme l'hypothèse nulle du test LR_ρ est acceptée, le modèle ***SLX*** est alors une possibilité. De plus, l'hypothèse nulle du test RLM_λ est rejetée pour ces matrices ainsi que celle du test $LR_{\theta+\rho\beta}$ ce qui entraîne l'utilisation du modèle ***SDM***.

Pour terminer l'approche mixte, nous observons que l'hypothèse nulle associée au test RLM_ρ pour la matrice inverse est rejetée ainsi que celle associée au test LR_θ ce qui induit l'utilisation du modèle ***SDM***.

TABLE 4.7 – Les p-valeurs des tests de spécifications pour l'approche mixte

Tests \ Matrices	2 voisins	5 voisins	10 voisins	contiguë	inverse	standardisée
Moran	0.01461	0.002955	0.000168	0.000324	5.135×10^{-12}	0.0004291
$SARMA$	0.4043	0.1633	0.06537	0.002034	0.109	0.002315
LR_θ	x	x	x	2.852×10^{-7}	x	1.53×10^{-7}
LR_ρ	x	x	x	0.9504	x	0.8676
RLM_ρ	0.2565	0.9531	0.3255	0.06703	0.0496	0.06467
LR_θ	x	x	x	x	$< 2.2\times10^{-16}$	x
RLM_λ	0.623	0.06013	0.05725	0.005915	0.7012	0.006958
$LR_{\theta+\rho\beta}$	x	x	x	2.791×10^{-6}	x	1.343×10^{-6}

Ensuite, lors de l'approche descendante reprise à la Table 4.8, nous observons que l'hypothèse nulle associée au test LR_λ est rejetée ainsi que celle du test $LR_{\theta+\rho\beta}$ pour l'ensemble des matrices sauf pour la matrice avec les deux voisins les plus proches et la matrice inverse. Nous pourrons alors utiliser le modèle ***SDM***.

En ce qui concerne l'ensemble des matrices des plus proches voisins, l'hypothèse nulle du test LR_ρ est acceptée, ce qui met en évidence le ***modèle linéaire***. Tandis que les hypothèses nulles du test LR_ρ et LR_θ sont rejetées pour les autres matrices de poids, le modèle à privilégier est le modèle ***SDM***.

Pour terminer l'approche descendante, nous observons que la p-valeur du test LR_θ est inférieure au seuil ce qui mène à l'utilisation du test LR_ρ. Comme l'hypothèse nulle de ce dernier est acceptée, nous pouvons utiliser le modèle ***SLX***.

TABLE 4.8 – Les p-valeurs des tests de spécifications pour l'approche descendante

Tests \ Matrices	2 voisins	5 voisins	10 voisins	contiguë	inverse	standardisée
Moran	0.01461	0.002955	0.000168	0.000324	5.135×10^{-12}	0.0004291
LR_λ	0.4117	0.04394	0.02644	0.005762	0.2511	0.006774
$LR_{\theta+\rho\beta}$	x	3.187×10^{-6}	1.476×10^{-11}	2.791×10^{-6}	x	1.343×10^{-6}
LR_ρ	0.2136	0.7635	0.1745	0.03258	0.03844	0.03205
$LR_{\rho\theta}$	x	x	x	8.265×10^{-7}	$<2.2\times10^{-16}$	4.478×10^{-7}
LR_θ	4.89×10^{-12}	6.377×10^{-7}	4.305×10^{-12}	1.302×10^{-7}	$<2.2\times10^{-16}$	6.966×10^{-8}
$LR_{\theta\rho}$	0.5431	0.8826	0.1486	0.9504	0.8891	0.8676

Trois modèles sont ressortis lors de notre analyse : le modèle linéaire, le modèle *SLX* et le modèle *SDM*. Afin de choisir le modèle qui convient le mieux, nous utilisons les critères de *AIC* et *BIC* ainsi que le coefficient pseudo R^2 pour déterminer la qualité du modèle. Nous comparons les différents facteurs pour l'ensemble des matrices de poids pour le modèle *SLX* et *SDM* à la Table 4.9. De plus, voici les différentes mesures pour le modèle linéaire :

— *AIC* : 5946.965 ;

— *BIC* : 6069.178 ;

— R^2 ajusté : 0.984 .

TABLE 4.9 – Critères qualitatifs pour le modèle *SLX* et *SDM*

Matrices / SLX	2 voisins	5 voisins	10 voisins	contiguë	inverse	standardisée
AIC	5924.404	5952.411	5924.11	5948.492	5815.905	5946.965
BIC	6046.617	6074.624	6046.323	6070.705	5942.483	6069.178
R^2 ajusté	0.9858	0.9851	0.9858	0.9852	0.9883	0.9853
Matrices / SDM	2 voisins	5 voisins	10 voisins	contiguë	inverse	standardisée
AIC	5926.034	5954.389	5924.024	5950.489	5817.886	5948.937
BIC	6052.612	6080.967	6050.602	6077.066	5948.828	6075.515
pseudo R^2	0.98646	0.98578	0.98651	0.98588	0.9888	0.98592

Nous observons que le modèle *SLX* avec la matrice inverse possède les plus petits facteurs *AIC* et *BIC* ainsi que le R^2 ajusté le plus proche de 1.A présent que nous avons choisi notre modèle et sa matrice de poids associée, nous utilisons la fonction *summary* afin d'obtenir les informations de notre modèle qui sont reprises à la Table 4.10. La p-valeur associée au test de Fisher est inférieure au seuil d'acceptation, ce qui signifie qu'au moins un des régresseurs explique la variable Y. Encore une fois, l'interaction spatiale est présente sur la variable X ce qui entraîne l'estimation du paramètre β et θ.

TABLE 4.10 – Critères qualitatifs pour le modèle *SLX*.

Régresseurs	$\hat{\beta}$	p-valeur	$\hat{\theta}$	p-valeur
terme constant	-6.726832	0.74733	-33.03938	0.44623
X_1	0.194205	6.25×10^{-14}	0.02174012	0.47155
X_2	-0.1290637	2.09×10^{-6}	-0.08684754	0.16448
X_3	0.05977756	0.02113	0.1377539	0.01441
X_4	-0.05540339	8.00×10^{-9}	0.09550328	3.48×10^{-8}
X_5	0.07265697	3.76×10^{-13}	-0.1333901	1.46×10^{-11}
X_6	0.07915179	5.75×10^{-9}	0.07662106	0.00295
X_7	-0.1131837	6.00×10^{-10}	-0.101605	0.00351
X_8	-0.002633253	0.20697	-0.0169123	$< 2\times10^{-16}$
X_9	0.04634989	$< 2\times10^{-16}$	0.03438747	2.26×10^{-6}
X_{10}	0.005300893	0.09786	0.01637481	0.00106
X_{11}	-0.0001746749	0.87876	0.001983574	0.37478
X_{12}	-0.000239449	0.85918	2.249582×10^{-5}	0.99339
X_{13}	0.0007943822	0.91315	-0.07704194	1.05×10^{-8}

Ainsi, les régresseurs et les régresseurs décalés qui expliquent le nombre d'agressions et de coups et blessures en Belgique en 2015 sont

— le nombre de divorces ;

— le nombre de femmes de moins de 18 ans ;

— le nombre de femmes et d'hommes entre 18 et 64 ans ainsi que sa variante spatialement décalée ;

— le nombre de femmes et d'hommes de plus de 65 ans ainsi que sa variante spatialement décalée ;

— la variante spatialement décalée de la densité de population ;

— le nombre de chômeurs ainsi que sa variante spatialement décalée ;

— la variante spatialement décalée du nombre d'employés ;

— la variante spatialement décalée du nombre de familles monoparentale.

4.2.3 Cyberharclement

Nous observons à la Table 4.11 que l'hypothèse nulle du test de Moran est rejetée ce qui signifie la présence d'une autocorrélation spatiale. De plus, le rejet des hypothèses nulles des tests de $SARMA$, LR_θ et LR_ρ indique l'utilisation du modèle ***SDM***. Ensuite, l'hypothèse nulle du test RLM_ρ est toujours acceptée, sauf pour la matrice des 10 voisins les plus proches, ce qui indique l'utilisation du ***modèle linéaire***. Pour la matrice des 10 voisins les plus proches, nous utilisons le modèle ***SDM*** comme l'hypothèse nulle est rejetée pour le test LR_θ. Afin de conclure l'utilisation de la méthode mixte, nous observons que les tests d'hypothèses RLM_λ et $LR_{\theta+\rho\beta}$ mènent au modèle ***SDM***.

TABLE 4.11 – Les p-valeurs des tests de spécifications pour l'approche mixte

Tests \ Matrices	2 voisins	5 voisins	10 voisins	contiguë	inverse	standardisée
Moran	4.759×10^{-6}	3.523×10^{-10}	1.482×10^{-13}	2.605×10^{-12}	$< 2.2\times10^{-16}$	2.514×10^{-12}
$SARMA$	1.506×10^{-5}	2.464×10^{-9}	1.234×10^{-10}	6.008×10^{-11}	4.552×10^{-15}	3.378×10^{-11}
LR_θ	1.931×10^{-5}	1.222×10^{-9}	3.749×10^{-9}	6.257×10^{-9}	4.216×10^{-12}	4.474×10^{-9}
LR_ρ	0.0001128	5.612×10^{-7}	4.166×10^{-5}	3.821×10^{-8}	6.161×10^{-9}	2.625×10^{-8}
RLM_ρ	0.1385	0.1673	0.04851	0.4173	0.07046	0.4018
LR_θ	x	x	4.205×10^{-10}	x	x	x
RLM_λ	1.784×10^{-5}	3.335×10^{-8}	4.628×10^{-10}	3.982×10^{-10}	2.509×10^{-13}	2.235×10^{-10}
$LR_{\theta+\rho\beta}$	0.02942	0.0005329	0.001678	0.005217	0.004179	0.005627

Passons à l'approche descendante qui est reprise à la Table 4.12. L'étape 1 et 3 de l'approche descendante met en évidence l'utilisation du modèle ***SDM***. En effet les hypothèses nulles du test LR_λ et $LR_{\theta+\rho\beta}$ ont été rejetées ainsi que celles du test LR_θ et LR_ρ. De plus, nous tirons les mêmes conclusions quant au choix du modèle avec le rejet des hypothèses nulles des test LR_ρ et LR_θ. Toutefois, nous observons que pour la matrice des deux voisins les plus proches, l'hypothèse nulle du test LR_ρ est acceptée ce qui signifie qu'un ***modèle linéaire*** est adapté pour cette matrice.

TABLE 4.12 – Les p-valeurs des tests de spécifications pour l'approche descendante

Tests \ Matrices	2 voisins	5 voisins	10 voisins	contiguë	inverse	standardisée
Moran	4.759×10^{-6}	3.523×10^{-10}	1.482×10^{-13}	2.605×10^{-12}	$< 2.2\times10^{-16}$	2.514×10^{-12}
LR_λ	1.8×10^{-6}	3.666×10^{-9}	2.739×10^{-9}	9.249×10^{-10}	1.369×10^{-13}	5.428×10^{-10}
$LR_{\theta+\rho\beta}$	0.02942	0.0005329	0.001678	0.005217	0.004179	0.005627
LR_ρ	0.05487	0.003204	0.01142	0.009257	0.0003141	0.009171
$LR_{\rho\theta}$	x	1.942×10^{-8}	2.293×10^{-8}	4.463×10^{-8}	4.205×10^{-10}	3.217×10^{-8}
LR_θ	0.002339	1.395×10^{-5}	1.574×10^{-6}	0.0004249	2.027×10^{-6}	0.0004126
$LR_{\theta\rho}$	0.0001128	5.612×10^{-7}	4.166×10^{-5}	3.821×10^{-8}	6.161×10^{-9}	2.625×10^{-8}

Deux modèles sont ressortis lors de notre analyse : le modèle linéaire et le modèle *SDM*. Afin de choisir le modèle qui convient le mieux, nous utilisons les critères de *AIC* et *BIC* ainsi que le coefficient pseudo R^2 pour déterminer la qualité du modèle. Nous

comparons les différents facteurs pour l'ensemble des matrices de poids pour le modèle *SDM* à la Table 4.13. De plus, voici les différentes mesures pour le modèle linéaire :

— *AIC* : 4872.855 ;

— *BIC* : 4938.326 ;

— R^2 ajusté : 0.7258 .

TABLE 4.13 – Critères qualitatifs pour le modèle *SDM*.

Matrices / Facteurs	2 voisins	5 voisins	10 voisins	contiguë	inverse	standardisée
AIC	4853.867	4829.761	4832.45	4833.687	4816.234	4832.877
BIC	4980.445	4956.339	4959.028	4960.264	4947.177	4959.454
pseudo R^2	0.75281	0.76286	0.76176	0.76125	0.76911	0.76158

Nous observons que le modèle *SDM* avec la matrice inverse possède les plus petits facteurs *AIC* ainsi que le pseudo R^2 le plus proche de 1. Le coefficient *BIC* est plus petit pour le modèle mais nous privilégions le modèle qui explique le plus la variable Y. De plus, le facteur *AIC* est plus robuste que le facteur *BIC*. A présent que nous avons choisi notre modèle et sa matrice de poids associée, nous utilisons la fonction *summary* afin d'obtenir les informations de notre modèle qui sont reprises à la Table 4.14. L'interaction spatiale est présente sur la variable X et Y, nous avons donc trois estimateurs pour les paramètres du modèle : $\hat{\beta}$, $\hat{\theta}$ et $\hat{\rho}$. L'estimateur de l'interaction sur la variable endogène est $\hat{\rho} = 0.3598068$. De plus, grâce aux p-valeurs associées à chaque estimateur, nous déterminons quelle variable exogène explique la variable dépendante.

TABLE 4.14 – Estimations des paramètres du modèle *SDM*.

Régresseurs	$\hat{\beta}$	p-valeur	$\hat{\theta}$	p-valeur
terme constant	8.183552	0.335548	11.83535	0.503347
X_1	-0.001838201	0.857928	0.007722962	0.530106
X_2	0.0157803	0.150025	-0.04360158	0.086121
X_3	0.01027564	0.328994	0.01455782	0.524210
X_4	0.00286536	0.456867	-0.005378552	0.439085
X_5	0.01164937	0.003382	-0.007600941	0.334497
X_6	0.001611239	0.767527	-0.01976907	0.059049
X_7	0.02240483	0.002188	0.00354483	0.802724
X_8	5.075122×10^{-5}	0.952316	-0.0004859155	0.522506
X_9	-0.01005789	3.726×10^{-9}	0.008949392	0.002268
X_{10}	-0.00982485	5.018×10^{-14}	0.008300426	4.440×10^{-5}
X_{11}	-0.0008232334	0.077323	-0.0002079289	0.819498
X_{12}	-0.0008271026	0.132224	0.001776456	0.107854
X_{13}	-0.007594647	0.010480	0.01551965	0.004045

Ainsi, les régresseurs et régresseurs décalés qui expliquent le nombre de cas de cyberharcelement en Belgique en 2015 sont

— le nombre d'hommes entre 18 et 64 ans et plus de 65 ans ;

— le nombre de chômeurs ainsi que sa variante spatialement décalée ;

— le nombre d'employés ainsi que sa variante spatialement décalée ;

— la variante spatialement décalée du nombre de familles monoparentale.

Nous observons qu'il y a moins de variables exogènes qui expliquent la variable Y par rapport aux analyses faites précédemment pour le recensement d'assassinats et d'agressions. Afin de pouvoir améliorer notre coefficient pseudo R^2, il serait astucieux de considérer des variables exogènes supplémentaires.

4.2.4 Personnes disparues

Dans un premier temps, nous utilisons l'approche mixte afin de déterminer le choix du modèle à la Table 4.15. Nous observons que la p-valeur associée au test de Moran est toujours inférieure au seuil d'acceptation ce qui induit le rejet de l'hypothèse nulle. Le test détecte donc de l'autocorrélation spatiale pour l'ensemble des matrices de poids. Ensuite, nous observons que l'hypothèse nulle est toujours acceptée pour les tests de $SARMA$, RLM_ρ et RLM_λ, sauf pour la matrice de poids inverse, ce qui conduit au choix du modèle contraint qui est le ***modèle de régression linéaire***. Pour la matrice de poids inverse, nous observons que l'hypothèse nulle associée au test $SARMA$ est rejetée ainsi que celle du test LR_θ et le test LR_ρ met en évidence le choix du modèle contraint qui est le modèle ***SLX***.

Table 4.15 – Les p-valeurs des tests de spécifications pour l'approche mixte

Tests \ Matrices	2 voisins	5 voisins	10 voisins	contiguë	inverse	standardisée
Moran	4.491×10^{-5}	9.797×10^{-6}	1.964×10^{-7}	2.901×10^{-6}	$< 2.2\times10^{-16}$	3.192×10^{-6}
$SARMA$	0.1799	0.8092	0.602	0.155	0.03171	0.147
LR_θ	x	x	x	x	1.782×10^{-8}	x
LR_ρ	x	x	x	x	0.3268	x
RLM_ρ	0.1425	0.6682	0.6048	0.5461	0.1242	0.5179
RLM_λ	0.9636	0.5165	0.3152	0.1663	0.4167	0.1673

Ensuite, nous utilisons l'approche descendante à la Table 4.16. Pour l'ensemble des matrices, sauf pour la matrice inverse, les tests LR_λ et LR_ρ mènent au choix du modèle contraint, c'est-à-dire, le ***modèle linéaire***. De plus, pour toutes les matrices de poids, l'hypothèse nulle du test LR_θ est rejetée tandis que celle du test LR_ρ est acceptée. Par conséquent, un autre choix de modèle intervient : le modèle ***SLX***. En particulier, pour la matrice de poids inverse, l'hypothèse nulle du test LR_ρ est rejetée en contrario avec celle du test LR_ρ ce qui mène à l'utilisation du modèle ***SDM***.

Table 4.16 – Les p-valeurs des tests de spécifications pour l'approche descendante

Tests \ Matrices	2 voisins	5 voisins	10 voisins	contiguë	inverse	standardisée
Moran	4.491×10^{-5}	9.797×10^{-6}	1.964×10^{-7}	2.901×10^{-6}	$< 2.2\times10^{-16}$	3.192×10^{-6}
LR_λ	0.1726	0.609	0.3609	0.0762	0.01897	0.07283
$LR_{\theta+\rho\beta}$	x	x	x	x	7.506×10^{-8}	x
LR_ρ	0.0506	0.96	0.9393	0.1891	0.02353	0.1751
$LR_{\rho\theta}$	x	x	x	x	6.445×10^{-8}	x
LR_θ	9.018×10^{-12}	0.0002542	0.002678	0.0469	1.172×10^{-8}	0.04171
$LR_{\theta\rho}$	0.8139	0.6002	0.6955	0.1879	0.3268	0.1889

Deux modèles sont ressortis lors de notre analyse : le modèle linéaire et le modèle *SLX*. De plus, le modèle *SDM* a été mentionné pour la matrice de poids inverse. Afin de choisir le modèle qui convient le mieux, nous utilisons les critères de *AIC* et *BIC* ainsi que le coefficient pseudo R^2 pour déterminer la qualité du modèle. Nous comparons les différents facteurs pour l'ensemble des matrices de poids pour le modèle *SLX* à la Table 4.17. De plus, voici les différentes mesures pour le modèle linéaire :

— *AIC* : 4904.705 ;

— *BIC* : 4970.176 ;

— R^2 ajusté : 0.8082 ;

et pour le modèle *SDM* avec la matrice de poids inverse :

— *AIC* : 4868.096 ;

— *BIC* : 4999.039 ;

— pseudo R^2 : 0.8328.

TABLE 4.17 – Critères qualitatifs pour le modèle *SLX*.

Matrices / Facteurs	2 voisins	5 voisins	10 voisins	contiguë	inverse	standardisée
AIC	4850.239	4892.36	4899.024	4908.117	4867.058	4907.708
BIC	4972.453	5014.573	5021.237	5030.33	4993.636	5029.921
R^2 ajusté	0.8291	0.8162	0.8141	0.8112	0.8243	0.8113

Nous observons que le modèle *SLX* avec la matrice des deux voisins les plus proches possède les plus petits facteurs *AIC* et *BIC* mais pas le R^2 ajusté le plus proche de 1. Le modèle possédant le meilleure pseudo R^2 est le modèle *SDM* avec la matrice inverse. Toutefois, nous faisons le choix de sélectionner le modèle *SLX* avec la matrice de poids des deux voisins les plus proches car le coefficient R^2 ajusté est proche de celui du modèle *SDM*. Les informations du modèle *SLX* avec la matrice des deux voisins les plus proches sont reprises dans la Table 4.18. Notons que nous avons la présence de l'interaction spatiale sur le terme exogène ce qui implique l'estimation du paramètre $\hat{\theta}$. De plus, la p-valeur associée au test de Fisher est inférieure au seuil d'acceptation ce qui indique qu'au moins un des régresseurs explique la variable Y. Nous observons que nous n'avons pas de terme constant décalé car la matrice choisie n'est pas standardisée.

TABLE 4.18 – Estimations des paramètres du modèle *SDM*.

Régresseurs	$\hat{\beta}$	p-valeur	$\hat{\theta}$	p-valeur
terme constant	-0.8906748	0.919451	x	x
X_1	0.02924763	0.000274	0.07223985	1.25×10^{-9}
X_2	0.009407002	0.425290	-0.01188358	0.512195
X_3	-0.01465431	0.189977	0.01046216	0.542628
X_4	0.003913614	0.318555	-0.00304687	0.567206
X_5	-0.009192363	0.025208	0.001072362	0.853114
X_6	0.003960046	0.523316	0.01885909	0.023415
X_7	-0.01423745	0.078103	-0.02409497	0.024125
X_8	-0.002555707	0.002778	0.0007672573	0.391421
X_9	0.008475689	2.67×10^{-6}	0.002385425	0.330952
X_{10}	0.005075194	0.000200	0.001328115	0.465244
X_{11}	-0.0008628123	0.078422	0.0008207915	0.175005
X_{12}	0.001612733	0.005875	7.979227×10^{-5}	0.923957
X_{13}	-0.007814322	0.013065	-0.006976482	0.118347

Ainsi, les régresseurs et régresseurs décalés qui expliquent le nombre de personnes disparues en Belgique en 2015 sont

— le nombre de divorces ainsi que sa variante spatialement décalée ;

— le nombre d'hommes entre 18 et 64 ans

— la variante spatialement décalée du nombres de femmes et d'hommes de plus de 65 ans ;

— la densité de population ;

— la variante spatialement décalée de la densité de population ;

— le nombres de chômeurs et d'employés ;

— le nombre de familles monoparentale ;

— le nombre de personnes possédant un diplôme supérieur.

Conclusions et perspectives

Lors de l'étude de la relation entre la valeur et la localisation de la variable spatiale, nous avons introduit plusieurs matrices de poids. Cette notion de matrices a été très utilisée afin d'étudier l'autocorrélation spatiale et afin d'émettre de nouveaux modèles de régression dans le cadre du chapitre 4. Nous avons pu nous rendre compte de l'impact d'un choix particulier de matrices de poids. En effet, ce choix peut être déterminant pour le rejet de l'hypothèse nulle de nos problèmes de tests. Dans le cadre de ce mémoire, nous avons exploité les matrices de poids les plus répandues lors de l'analyse spatiale. Toutefois, il serait intéressant d'approfondir nos analyses avec une plus grande variété de matrices.

Nous avons parcouru dans le cadre du second chapitre les méthodes liées aux donnés ponctuelles où nous nous concentrions uniquement sur la composante localisation. Pour rappel, l'objectif a été de déterminer la distribution de la configuration de points. L'outil mit en place a été les fonctions de Ripley dans le cadre homogène et inhomogène. Nous devions observer si l'estimateur de la fonction de Ripley associé à notre configuration de points se trouvait au-dessus ou en-dessous de l'estimateur de la fonction de Ripley pour une distribution de Poisson homogène dans le but d'établir la présence d'une interaction ou d'une répulsion entre les localisations. Toutefois, une amélioration que nous pouvons énoncer est l'intervention des intervalles d'acceptation. Ces intervalles peuvent nous permettre d'établir avec précision le positionnement de l'estimateur de Ripley par rapport à la courbe de référence. Si notre estimateur se trouve en dehors de cet intervalle, nous concluons que la distribution de points n'est pas aléatoire.

Pour les données continues, nous avons défini les modèles classiques d'ajustement pour les variogrammes expérimentaux univariés et multivariés. Nous sélectionnons le modèle le plus adapté en utilisant une validation croisée avec plusieurs critères de minimisation de la différence entre la valeur exacte et la valeur approximée de la variable spatiale. Lors de leurs implémentions en R, nous devons imposer manuellement les valeurs correspondantes au palier, à la portée et à l'effet pépite. Toutefois, il serait plus optimal d'optimiser ces choix afin d'augmenter l'ajustement du modèle. De plus, nous avons introduit la méthode du Kriging ordinaire. Il serait intéressant d'approfondir ce concept avec d'autres formes de Kriging comme le Kriging universel et de réaliser une comparaison des résultats d'interpolation obtenus.

Pour terminer, nous avons travaillé avec les données surfaciques. Une plénitude de modèles ont été étudiés. Toutefois, il en existe d'autres découlant des modèles généraux que nous avons introduit théoriquement. De plus, afin de sélectionner le modèle adéquat, nous avons utilisé les approches "mixte" et "descendante" qui sont les arbres de décisions les plus répandus dans la régression spatiale et les tests de spécifications basées sur les multiplicateurs de Lagrange et du rapport de vraisemblance. Toutefois, il aurait été intéressant de tester d'autres approches afin d'envisager d'autres modèles plus complexes comme le modèle de Manski. De plus, nous avons étudié les modèles de régression spatiales mais il

aurait été intéressant d'introduire la notion temporelle à ces modèles, sur base des livres de référence [13] et [40], afin d'étudier l'influence des variables explicatives à travers le temps et l'espace.

De plus, après plusieurs recherches dans le domaine de la criminologie, nous avons introduit certains régresseurs pouvant expliquer les variables de recensement de certains délits criminels. Ces données ont été récoltées sur les sites [34] et [22]. Toutefois, la récolte des données a été parfois un obstacle quant au manque de certaines catégories de données. Même si nous avons obtenu de bons coefficients pour la qualité des modèles, nous pourrions apporter une amélioration en considérant plus de facteurs exogènes liés à la criminalité. En parallèle la récolte des données criminelles sur le site [30] a aussi été limitée. En effet, le plus haut degré de précision que nous pouvions avoir pour la localisation est celui par commune. Il aurait été intéressant d'avoir accès à des données plus précises avec la longitude et la latitude exacte de certains recensements de crimes afin de pouvoir utiliser les méthodes vues dans le cadre du chapitre 2 et chapitre 3. Ainsi, nous aurions pu compléter l'analyse de la criminalité faite à la section 4.2 avec l'étude de la distribution de la localisation des crimes et l'interpolation de la valeur d'une certaine variable spatiale criminelle à un nouvel endroit.

Bibliographie

[1] Allard D., *Géostatistique multivariée*, diapositives de cours théoriques, année académique 2015-2016

[2] Allard D., Statistiques spatiales : introduction à la géostatistique, Université de Montpellier, support de cours, année académique 2012-2013

[3] Anslin L., *Spatial Regression :10. Specification Tests (2)*, Université de Chicago, diapositives de cours théoriques, année académique 2017-2018

[4] Arnaud M., *L'analyse krigeante pour le classement d'observations spatiales et multivariées*, 2001

[5] Baddeley A., *Analysing spatial point patterns in R*, 2010

[6] Baddeley A., *Spatial point patterns : methodolody and applications with R*, 2015

[7] Baillargeon S., *Le krigeage : revue de la theorie et application a l'interpolation spatiale de données de précipitations*, Universtié de Laval, mémoire, 2005

[8] Beale, C.M. et al., *Regression analysis of spatial data*, Ecology letters, 2010.

[9] Boualla N., *Cours de Géostatistique*, Université d'Oran, année académique 2018-2019

[10] Bradaï A., *Géostatistique appliquée*, Université Hassiba Benbouali de Chlef, notes de cours, année académique 2015-2016

[11] Casajus N., *Analyses spatiales sous R*, 2013

[12] Casanova JJ., *La méthode du Lagrangien*, 2015

[13] Cisse P.O., *Étude de modèles spatiaux et spatio-temporels*, 2019

[14] Colorado School of Mines, `http://inside.mines.edu/~jdzimmer/tutorials/Section2.html`, 19 juillet 2022

[15] Cours de droit.net, `https://cours-de-droit.net`, 5 Août 2022

[16] Elhorst P., *Applied Spatial Econometrics : Raising the Bar*, 2010

[17] Esri,`https://pro.arcgis.com/fr/pro-app/latest/tool-reference/spatial-statistics`, 12 Août 2022

[18] GADM, urlhttps ://gadm.org/, 5 Août 2022

[19] Greene W., *Econometric Analysis*, 2017

[20] Hennequi M., *Spatialisation des données de modélisation par Krigeage*, 2010

[21] Insee-Eurostat, *Manuel d'analyse spatiale*, 2018

[22] Iweps, `https://www.iweps.be/`, 5 Août 2022

[23] Jayet H., *Économétrie et données spatiales - Une introduction à la pratique*, 2001

[24] Kiriliouk A., *Régression linéaire et non linéaire [SMATM111]*, notes de cours théoriques, année académique 2020-2021.

[25] Lee LF., *Asymptotic Distributions of Quasi-Maximum Likelihood Estimators for Spatial Autoregressive Models*, Econometrica, 2004

[26] Le Gallo J., *Econometrie spatiale (Autocorrélation spatiale)*, 2002

[27] Matheron G., *La théorie des variables régionalisés et ses applications*, Les Cahiers du Centre de Morphologue Mathématique de Fontainebleau, 1970

[28] Méneroux Y., *Introduction à la géostatistique*, support de cours, année académique 2018-2019

[29] Oliveau S., *Autocorrélation spatiale : leçons du changement d'échelle*, 2010

[30] Police fédérale, `https://www.stat.policefederale.be/`, 29 juillet 2022

[31] Rahma A., *RPubs*, `https://rpubs.com/r_anisa/kriging-interpolation`, 2 Août 2022

[32] Saas Y., *Modèles statistiques spatialement explicites de données de comptage*, 2012

[33] SAS Institute Inc., *SAS/ETS 14.2 User's Guide TheSPATIALREG Procedure*, 2016

[34] Statbel, `https://bestat.statbel.fgov.be/bestat/`, 5 Août 2022

[35] Tarling, R., *Statistical applications in Criminology*, J. Royal. Stat. Society, Series D., 1986.

[36] Thomas-Agnan C., *Analyse statistique des données spatiales I*, diapositives de cours théoriques, année académique 2012-2013

[37] Van Bever G., *Analyse mutlivariée et introduction aux logiciels statistiques [SMATM102]*, notes de cours théoriques, année académique 2020-2021

[38] Van Bever G., *Statistiques [SMATN211]*, notes de cours théoriques, année académique 2017-2018

[39] Wackernagel H., *Cours de géostatistique multivariable*, 1993

[40] Wilke C. , *Spatio-Temporal Statistics with R*, 2019

[41] Weisburd, D. and Britt, C. (2014), *Statistics in criminal justice*, Springer.

Annexe A

Modèles de régression classique

Ce chapitre introduit les modèles de régression, vu dans le cadre du cours de *Régression linéaire et non linéaire* [24], et en particulier les modèles de régression linéaire. Dans celui-ci, nous verrons en détail les propriétés des modèles de régression linéaire simple et multiple ainsi que l'obtention de leurs paramètres. Notons que les figures ont été réalisées grâce au logiciel R.

A.1 Modèle de régression linéaire simple

A.1.1 Modèle normal

L'objectif de la régression est d'expliquer une variable aléatoire Y, appelée variable de réponse ou variable dépendante ou encore variable endogène, grâce à une variable X, appelée variable explicative ou variable indépendante ou encore variable exogène. Plus précisément, nous voulons réaliser une interpolation c'est-à-dire réaliser une prédiction de la variable Y en fonction de la variable X toutefois l'extrapolation de Y n'est pas possible.

Dans le meilleurs des cas, il est préférable d'obtenir une relation linéaire exacte à partir l'observation X afin d'expliquer la variable Y. Toutefois nous n'avons pas toujours la possibilité d'obtenir une telle relation, c'est pourquoi l'étude du modèle de régression linéaire normal est mis en pratique.

Définition A.1
Le modèle de régression linéaire simple normal est donné par

$$Y = \beta_0 + \beta_1 X + \epsilon$$

où

- X *est la variable explicative,*
- Y *est la variable à expliquer et*
- $\epsilon \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ *est un terme d'erreur.*

Le modèle est dit simple car la relation linéaire permettant de définir la variable à expliquer est fondée à partir d'une seule variable explicative. De plus, le modèle est dit linéaire et cela vient du fait que la variable aléatoire Y s'écrit comme une combinaison linéaire de la variable X avec des paramètres réels β_0 et β_1. La caractéristique de norma-

lité est appliquée au terme d'erreur du modèle.

De manière à réaliser la création du modèle, celui-ci est fondé sur un nombre fini n d'observations $(x_1, y_1), \ldots, (x_n, y_n)$ avec $\epsilon_1, \ldots, \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. Le modèle de régression linéaire simple normal pour chaque couple (x_i, y_i) où $i = 1, \ldots, n$ est alors donné par

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i. \tag{A.1}$$

En outre, nous pouvons notifier que la variable x_i est mesurée sans erreur et de plus qu'il ne s'agit pas d'une variable aléatoire. Cependant, la variable y_i est une variable aléatoire où le terme de l'erreur ϵ_i contient l'information de y_i qui n'est pas expliquée par x_i.

Il est important de percevoir l'impact de l'hypothèse de normalité du terme d'erreur sur l'espérance, la variance et la covariance de la variable de réponse Y. Tout d'abord, sachant que $\mathbb{E}(\epsilon_i) = 0$ et que l'espérance est une application linéaire, nous observons que l'espérance de la variable aléatoire Y ne dépend pas du terme d'erreur,

$$\mathbb{E}(y_i) = \beta_0 + \beta_1 x_i. \tag{A.2}$$

De plus, comme $\mathrm{Var}(\epsilon_i) = \sigma^2$ et par propriété de la variance, la variance de Y vaut elle aussi σ^2,

$$\mathrm{Var}(y_i) = \mathrm{Var}(\epsilon_i) = \sigma^2. \tag{A.3}$$

En utilisant la définition de la covariance, ainsi que le résultat (A.2) et que l'espérance du terme d'erreur est nulle, la covariance de la variable aléatoire Y est alors égale à la covariance de la variable aléatoire ϵ.

$$\begin{aligned}
Cov(y_i, y_j) &= \mathbb{E}[(y_i - \mathbb{E}(y_i))(y_j - \mathbb{E}(y_j))] \\
&= \mathbb{E}[(\beta_0 + \beta_1 x_i + \epsilon_i - \beta_0 - \beta_1 x_i)(\beta_0 + \beta_1 x_j + \epsilon_j - \beta_0 - \beta_1 x_j)] \\
&= \mathbb{E}(\epsilon_i \epsilon_j) \\
&= \mathbb{E}[(\mathbb{E}(\epsilon_i) - \epsilon_i)(\mathbb{E}(\epsilon_j) - \epsilon_j)] \\
&= Cov(\epsilon_i, \epsilon_j) \\
&= 0.
\end{aligned}$$

Ainsi, comme $\epsilon_1, \ldots, \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ alors $y_1, \ldots, y_n \overset{iid}{\sim} \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$.

A présent, nous présentons deux méthodes afin d'estimer les paramètres β_0 et β_1 du modèle de régression linéaire simple. La première méthode permet d'attribuer une propriété très importante aux estimateurs des paramètres β_0 et β_1. La deuxième méthode fournira une estimation supplémentaire pour le paramètre σ^2 du modèle.

Méthode des moindres carrés

Le but du critère des moindres carrés est de minimiser la somme des résidus, c'est-à-dire, la somme des écarts entre la valeur exacte de y_i et sa valeur approximée, $\hat{y}_i = \beta_0 + \beta_1 x_i$, par le modèle. L'objectif est donc de trouver les coefficients β_0 et β_1 qui minimisent

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

Pour ce faire, nous calculons les extrema de $S(\beta_0, \beta_1)$ et étant donné que la matrice hessienne de $S(\beta_0, \beta_1)$ est strictement définie positive alors les extrema sont des minimums.

$$\begin{cases} \frac{\partial S(\beta_0,\beta_1)}{\partial \beta_0} = 0 \Rightarrow \sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i \\ \frac{\partial S(\beta_0,\beta_1)}{\partial \beta_1} = 0 \Rightarrow \sum_{i=1}^{n} x_i y_i = \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 \end{cases}$$

Définition A.2
Les estimateurs des moindres carrés de β_0 et β_1 sont

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \quad et \quad \hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \frac{\sum_{i=1}^{n}(x_i - \bar{X})y_i}{\sum_{i=1}^{n}(x_i - \bar{X})^2},$$

où

- $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$ *et*
- $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

De façon à ce que les estimateurs des moindres carrés soient bien définis, nous imposons que S_{XX} soit non nulle. Cette condition n'amène pas de restriction importante dans la définition du modèle car si S_{XX} est nulle cela signifie que les valeurs sont toutes égales ce qui n'est pas le cas en pratique.

Comme nous l'avons mentionné précédemment, la méthode des moindres permet de mettre en évidence une propriété des estimateurs de certains paramètres du modèle grâce au théorème suivant.

Théorème A.1 (Gauss-Markov)
Pour le modèle linéaire simple

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \ i = 1, \ldots, n,$$

sous les hypothèses

- $\mathbb{E}(\epsilon) = 0_n$ *et*
- $Var(\epsilon) = \sigma^2 I_n$

pour $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$, les estimateurs des moindres carrés $\hat{\beta}_0$ et $\hat{\beta}_1$ sont, dans la classe des estimateurs linéaires non biaisés de β_0 et β_1, ceux qui ont la plus petite variance. Par ailleurs, ces estimateurs sont connus sous le nom de BLUE, Best Linear Unbiased Estimator.

Démonstration

Par la définition (A.2), l'estimateur des moindres carrés de β_1 est défini par $\hat{\beta}_1 = \sum_{i=1}^{n} c_i y_i$ où $c_i = \frac{(x_i - \bar{X})}{(n-1)S_{XX}}$. Ensuite, par propriété de la variance et par la relation (A.3), nous observons que $\mathrm{Var}(\hat{\beta}_1) = \sigma^2 \sum_{i=1}^{n} c_i^2$. L'objectif de cette démonstration est de montrer que cet estimateur est *BLUE*. Pour ce faire, nous supposons l'existence de $\hat{\beta}_1^* = \sum_{i=1}^{n} k_i y_i$ qui est un estimateur *BLUE* de β_1. Sachant que cet estimateur est non biasié

$$\mathbb{E}(\hat{\beta}_1^*) = \mathbb{E}(\sum_{i=1}^{n} k_i y_i) = \sum_{i=1}^{n} k_i \mathbb{E}(y_i) = \sum_{i=1}^{n} k_i(\beta_0 + \beta_1 x_i) = \beta_0 \sum_{i=1}^{n} k_i + \beta_1 \sum_{i=1}^{n} k_i x_i = \beta_1$$

alors nous devons imposer que

$$\sum_{i=1}^{n} k_i = 0 \text{ et } \sum_{i=1}^{n} k_i x_i = 1. \tag{A.4}$$

De plus, par propriété de la variance et par la relation (A.3), nous observons que $\text{Var}(\hat{\beta}_1^*) = \sigma^2 \sum_{i=1}^{n} k_i^2$.

Par la suite, nous fixons $d_i = k_i - c_i$ et nous pouvons alors réécrire la variance de l'estimateur $\hat{\beta}_1^*$ de la manière suivante

$$\text{Var}(\hat{\beta}_1^*) = \sigma^2 \sum_{i=1}^{n} k_i^2 = \sigma^2 \sum_{i=1}^{n} (c_i + d_i)^2 = \sigma^2 (\sum_{i=1}^{n} c_i^2 + \sum_{i=1}^{n} d_i^2 + 2 \sum_{i=1}^{n} c_i d_i).$$

Toutefois, par les conditions imposées en (A.4), le terme $\sum_{i=1}^{n} c_i d_i$ est nul. Par conséquent, la variance de l'estimateur $\hat{\beta}_1^*$ s'écrit de la manière suivante

$$\text{Var}(\hat{\beta}_1^*) = \sigma^2 (\sum_{i=1}^{n} c_i^2 + \sum_{i=1}^{n} d_i^2) = \text{Var}(\hat{\beta}_1) + \sigma^2 \sum_{i=1}^{n} d_i^2.$$

En outre, nous observons que $\text{Var}(\hat{\beta}_1^*) \geq \text{Var}(\hat{\beta}_1)$. Par définition de l'estimateur $\hat{\beta}_1^*$ de β_1, les variances doivent être par conséquent égales ce qui est le cas si et seulement si $\sum_{i=1}^{n} d_i^2 = 0$. Ainsi, nous observons alors l'égalité suivante $k_i = c_i$, $\forall i = 1, \ldots, n$. Pour conclure, l'estimateur des moindres carrés de β_1 est *BLUE*. Le raisonnement est similaire pour l'estimateur des moindres carrés de β_0.

□

Méthode du maximum de vraisemblance

A présent, nous réalisons une nouvelle estimation des paramètres β_0 et β_1 grâce à la méthode du maximum de vraisemblance. L'objectif de cette méthode est de trouver la valeur la plus probable pour chaque paramètre du modèle permettant d'engendrer les valeurs exactes y_i. Grâce à l'hypothèse $y_i \overset{iid}{\sim} \mathcal{N}(\beta_0 + \beta_1 x_i,\, \sigma^2)$, la fonction de vraisemblance des paramètres $(\beta_0, \beta_1, \sigma^2)$ et des variable aléatoires y_i pour $i = 1, \ldots, n$ se présente comme un produit des fonctions de densité des variables aléatoires. De plus, comme la dérivée seconde de la fonction de vraisemblance est négative nous maximisons la log-vraisemblance,

$$\begin{aligned} \log\, L(\beta_0, \beta_1, \sigma^2; y_1, \ldots, y_n) &= \log\, \left(\prod_{i=1}^{n} (2\pi\sigma^2)^{-\frac{1}{2}} \exp[-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2] \right) \\ &= -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \end{aligned}$$

et calculons les extrema de la fonction :

$$\begin{cases} \frac{\partial \log\, L(\beta_0, \beta_1, \sigma^2; y_1, \ldots, y_n)}{\partial \beta_0} = 0 \Rightarrow \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \\ \frac{\partial \log\, L(\beta_0, \beta_1, \sigma^2; y_1, \ldots, y_n)}{\partial \beta_1} = 0 \Rightarrow \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \\ \frac{\partial \log\, L(\beta_0, \beta_1, \sigma^2; y_1, \ldots, y_n)}{\partial \sigma^2} = 0 \Rightarrow -n\sigma^2 + \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = 0. \end{cases}$$

La dérivée seconde de la fonction de vraisemblance est négative ce qui rend les extrema trouvés des maximum locaux pour la fonction de vraisemblance.

Définition A.3
Les estimateurs du maximum de vraisemblance de β_0, β_1 et σ^2 sont

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}, \quad \hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} \quad et \quad \widetilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

Grâce à l'hypothèse de normalité des erreurs, les deux méthodes nous fournissent les mêmes résultats. Nous pouvons alors établir que les estimateurs de maximum de vraisemblance pour β_0 et β_1 sont $BLUE$.

Toutefois, il n'est pas possible de tirer les mêmes conclusions pour l'estimateur du maximum de vraisemblance de σ^2 qu'avec les estimateurs des paramètres du modèle de régression linéaire simple β_0 et β_1. En effet, l'estimateur de σ^2, défini précédemment, est un estimateur biaisé de σ^2 car $\mathbb{E}(\widetilde{\sigma}^2) = \left(\frac{n-2}{n}\right)\sigma^2$. C'est pourquoi, nous définissons un nouveau estimateur afin d'avoir un estimateur non biaisé de σ^2.

Définition A.4
L'estimateur non-biaisé de σ^2 est donné par

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = \frac{SSR}{n-2} = MSE,$$

où

- *SSR = Sum of Squared Residuals et*
- *MSE = Mean Squared Error.*

Maintenant que nous avons trouvé un estimateur non-biaisé pour chaque paramètre du modèle, nous définissons leurs expressions.

Définition A.5
Les estimateurs de β_0, β_1 et σ^2 sont

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}, \quad \hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} \quad et \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}.$$

Les estimateurs $\hat{\beta}_0$, $\hat{\beta}_1$ et $\hat{\sigma}^2$ sont les estimateurs optimaux de β_0, β_1 et σ^2. En effet, ces estimateurs font partie de la classe des estimateurs non-biaisés de β_0, β_1 et σ^2 et sont ceux qui détiennent la plus petite variance.

Maintenant que les estimateurs des paramètres du modèle β_0 et β_1 ont été choisis avec soin, nous appliquons les concepts définis précédemment à un exemple. L'objectif est de réaliser la droite de régression estimée de la variable Y qui est le résultat obtenu à un examen sur 100 grâce à X qui est le nombre d'heures passées à étudier le cours associé.

Dans le cadre de cet exemple, les estimateurs des moindres carrés de β_0 et β_1 sont

- $\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \frac{55}{36.75} = 1.50$

— $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$=40

TABLE A.1 – Résultats obtenus à l'examen et le temps d'étude associés

i	1	2	3	4	5	6	7	8
x_i	20	16	34	23	27	32	18	22
y_i	64	61	84	70	88	92	72	77

et donc la droite de régression estimée est $\hat{Y} = 40 + 1.5X$. La droite de régression est représentée sur la figure (1.1).

Grâce au modèle de régression linéaire simple défini pour la table (A.1), nous pouvons alors prédire la valeur de Y en fonction de la valeur que nous donnons à X. Ainsi, si un étudiant passe 30 heures à étudier son cours alors la note qu'il peut espérer est de 85 sur 100.

A.1.2 Intervalles de confiance

Maintenant que les estimateurs des paramètres β_0, β_1 et σ^2 ont été fixés, nous pouvons construire leurs intervalles de confiance afin de voir les valeurs que peut prendre chaque paramètre avec une certaine probabilité.

Pour cela, certaines propriétés des estimateurs des paramètres dans le cadre du modèle normal sont nécessaires afin de réaliser la construction des intervalles.

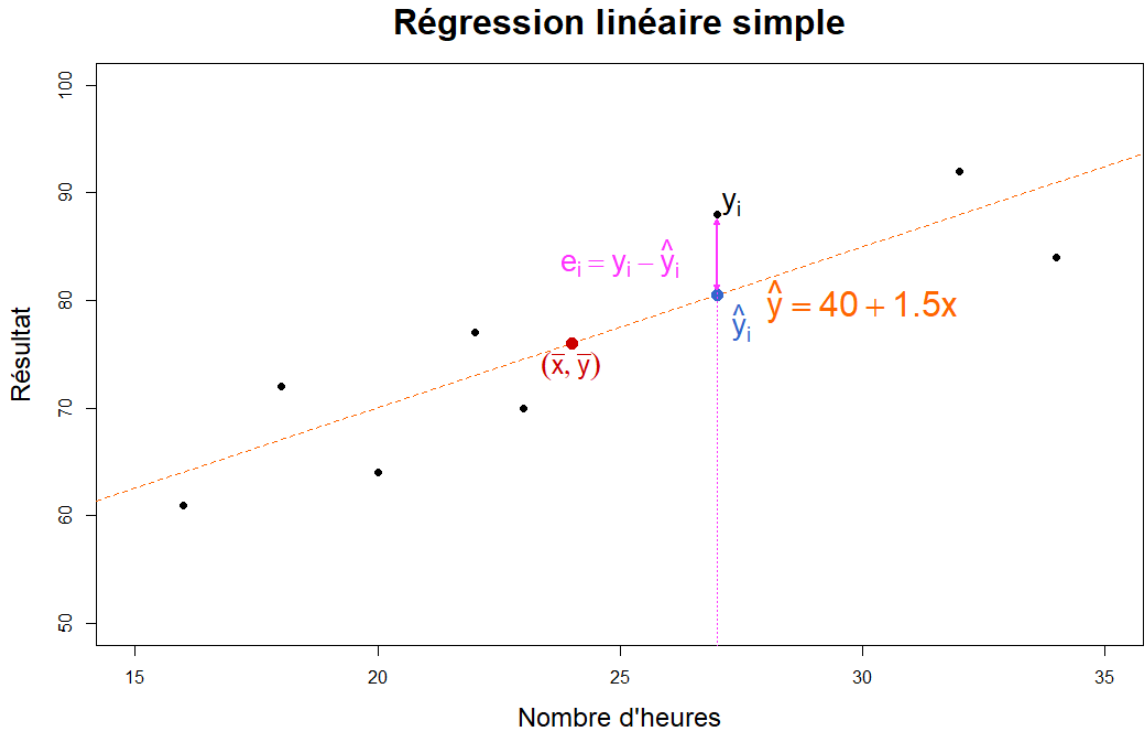

FIGURE 1.1 – La droite de régression linéaire simple de la table (A.1)

Proposition A.1
Dans le modèle linéaire simple normal, les estimateurs sont donnés par

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}, \quad \hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} \quad et \quad \hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

Ces estimateurs possèdent les propriétés suivantes

— $(\hat{\beta}_0, \hat{\beta}_1) \perp\!\!\!\perp \hat{\sigma}^2$,

— $(\hat{\beta}_0, \hat{\beta}_1)^T \sim \mathcal{N}_2 \left(\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \Sigma \right), \ \Sigma = \frac{\sigma^2}{(n-1)S_{XX}} \begin{pmatrix} \frac{1}{n} \sum_{i=1}^{n} x_i^2 & -\bar{X} \\ -\bar{X} & 1 \end{pmatrix}$ *et*

— $\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$.

Intervalle de confiance pour β_1

Les propriétes précédentes, énoncées dans la proposition (A.1), ont deux conséquences :

— $\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{(n-1)S_{XX}}\right)$ donc $Z_1 = \dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{(n-1)S_{XX}}}} \sim \mathcal{N}(0, 1)$ et

— $Z_1 \perp\!\!\!\perp U$ où $U = \frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$.

Dès lors par définition d'une loi de Student,

$$T_1 = \frac{Z_1}{\sqrt{\frac{U}{(n-2)}}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{(n-1)S_{XX}}}} \sim t_{n-2}$$

et T_1 est une fonction pivotable.

Nous pouvons alors en déduire que

$$\mathbb{P}\left[-t_{n-2,1-\frac{\alpha}{2}} \leq T_1 \leq t_{n-2,1-\frac{\alpha}{2}} \right] = 1 - \alpha$$

et donc d'obtenir l'intervalle de confiance au niveau de confiance α pour β_1 :

$$I_\alpha(\beta_1) = \hat{\beta}_1 \pm t_{n-2,1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\sigma}^2}{(n-1)S_{XX}}}. \tag{A.5}$$

Intervalle de confiance pour β_0

La construction de l'intervalle de confiance pour β_0 est similaire à celle de β_1. Par conséquent, l'intervalle de confiance au niveau de confiance α pour β_0 est

$$I_\alpha(\beta_0) = \hat{\beta}_0 \pm t_{n-2,1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\sigma}^2}{n(n-1)S_{XX}} \sum_{i=1}^{n} x_i^2}. \tag{A.6}$$

Intervalle de confiance pour σ^2

La construction de l'intervalle de confiance pour σ^2 est similaire à celle réalisée précédemment pour les paramètres β_0 et β_1. De sorte que, l'intervalle de confiance au niveau de confiance α pour σ^2 est

$$I_\alpha(\sigma^2) = \left[\frac{(n-2)\hat{\sigma}^2}{\chi^2_{n-2,1-\frac{\alpha}{2}}}, \frac{(n-2)\hat{\sigma}^2}{\chi^2_{n-2,\frac{\alpha}{2}}} \right]. \tag{A.7}$$

A.1.3 Tests d'hypothèses

A présent que les intervalles de confiances ont été obtenus dans la sous-section (A.1.2), nous pouvons alors établir le lien avec les tests d'hypothèses.

Test d'hypothèses sur β_0

Soit l'intervalle de confiance $I_\alpha(\beta_0)$, défini en (A.6), pour β_0 au niveau de confiance α alors le problème de test est

$$\begin{cases} \mathcal{H}_0 : \beta_0 = \beta_0^0 \\ \mathcal{H}_1 : \beta_0 \neq \beta_0^0,\ \beta_0 < \beta_0^0,\ \beta_0 > \beta_0^0. \end{cases}$$

Par construction de l'intervalle de confiance de β_0, sous $\mathcal{H}_0$, la statistique du test

$$T_0 = \frac{\hat{\beta}_0 - \beta_0^0}{\sqrt{\frac{\hat{\sigma}^2}{n(n-1)S_{XX}} \sum_{i=1}^{n} x_i^2}} \sim t_{n-2}.$$

Si bien que l'hypothèse nulle $\mathcal{H}_0$ est rejetée au niveau α et donc que $\mathcal{H}_1$ est acceptée dans les cas suivants repris dans la Table A.2.

TABLE A.2 – Règle de rejet pour le test d'hypothèse sur β_0

$\mathcal{H}_1$	Condition
$\beta_0 > \beta_0^0$	$t_0 \geq t_{n-2,1-\alpha}$ ou $\mathbb{P}(T_0 \geq t_0 \mid \mathcal{H}_0) \leq \alpha$
$\beta_0 < \beta_0^0$	$t_0 \leq -t_{n-2,1-\alpha}$ ou $\mathbb{P}(T_0 \leq t_0 \mid \mathcal{H}_0) \leq \alpha$
$\beta_0 \neq \beta_0^0$	$\lvert t_0 \rvert \geq t_{n-2,1-\frac{\alpha}{2}}$ ou $\begin{cases} \mathbb{P}(T_0 \leq t_0 \mid \mathcal{H}_0) \leq \frac{\alpha}{2} \\ \text{ou } \mathbb{P}(T_0 \geq t_0 \mid \mathcal{H}_0) \leq \frac{\alpha}{2} \end{cases}$

Test d'hypothèses sur β_1

Le test d'hypothèses pour β_1 est construit de manière similaire que celui de β_0. Soit l'intervalle de confiance $I_\alpha(\beta_1)$, défini en (A.5), pour β_1 au niveau de confiance α alors le problème de test est

$$\begin{cases} \mathcal{H}_0 : \beta_1 = \beta_1^0 \\ \mathcal{H}_1 : \beta_1 \neq \beta_1^0,\ \beta_1 < \beta_1^0,\ \beta_1 > \beta_1^0. \end{cases}$$

Par construction de l'intervalle de confiance de β_1, sous $\mathcal{H}_0$, la statistique du test

$$T_1 = \frac{\hat{\beta}_1 - \beta_1^0}{\sqrt{\frac{\hat{\sigma}^2}{(n-1)S_{XX}}}} \sim t_{n-2}.$$

Si bien que l'hypothèse nulle $\mathcal{H}_0$ est rejetée au niveau α et donc que $\mathcal{H}_1$ est acceptée dans les cas suivants.

TABLE A.3 – Règle de rejet pour le test d'hypothèse sur β_1

$\mathcal{H}_1$	Condition
$\beta_1 > \beta_1^0$	$t_1 \geq t_{n-2,1-\alpha}$ ou $\mathbb{P}(T_1 \geq t_1 \mid \mathcal{H}_0) \leq \alpha$
$\beta_1 < \beta_1^0$	$t_1 \leq -t_{n-2,1-\alpha}$ ou $\mathbb{P}(T_1 \leq t_1 \mid \mathcal{H}_0) \leq \alpha$
$\beta_1 \neq \beta_1^0$	$\lvert t_1 \rvert \geq t_{n-2,1-\frac{\alpha}{2}}$ ou $\begin{cases} \mathbb{P}(T_1 \leq t_1 \mid \mathcal{H}_0) \leq \frac{\alpha}{2} \\ \text{ou } \mathbb{P}(T_1 \geq t_1 \mid \mathcal{H}_0) \leq \frac{\alpha}{2} \end{cases}$

En outre, il est intéressant de voir s'il existe ou non une relation de régression entre la variable de réponse Y et la variable explicative X.

Lorsque $\mathcal{H}_0 : \beta_1 = 0$ n'est pas rejetée, c'est-à-dire que la pente de la droite de régression est nulle, alors nous avons deux cas de figure représentés sur la figure (1.2) :

— Dans le cas (a), la variable explicative X ne détient pas le rôle espéré, c'est-à-dire que la variable Y n'est pas expliquée par la variable X. En effet, la meilleure estimation de Y est donnée par $\hat{Y} = \bar{Y}$ pour tout X.

— Dans le cas (b), un autre problème intervient puisque la relation entre Y et X n'est pas linéaire ce qui n'est pas possible par définition du modèle.

Lorsque $\mathcal{H}_0 : \beta_1 = 0$ est rejetée, c'est-à-dire que la pente de la droite de régression est non nulle, alors nous avons aussi deux cas de figure représentés sur la figure (1.3) :

— Dans le cas (a), la variable explicative X détient le rôle espéré, c'est-à-dire que la variable Y est expliquée par la variable X et donc que le modèle linéaire simple est approprié. En effet, pour chaque X l'estimation de Y semble de bonne qualité.

— Dans le cas (b), la relation entre Y et X parait être linéaire. Toutefois, le modèle linéaire simple ne sera pas le plus adéquat.

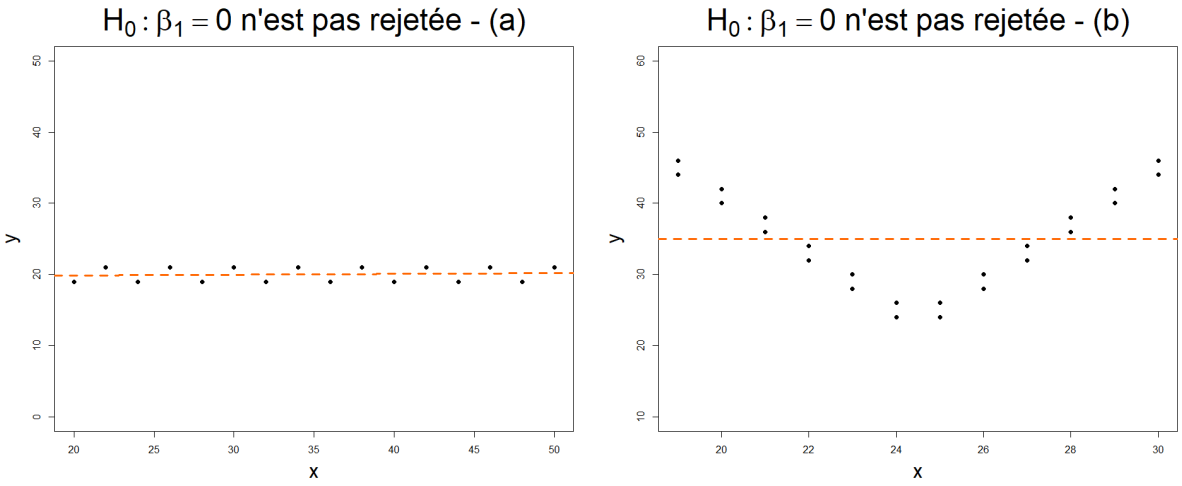

FIGURE 1.2 – Illustration lorsque l'hypothèse nulle n'est pas rejetée

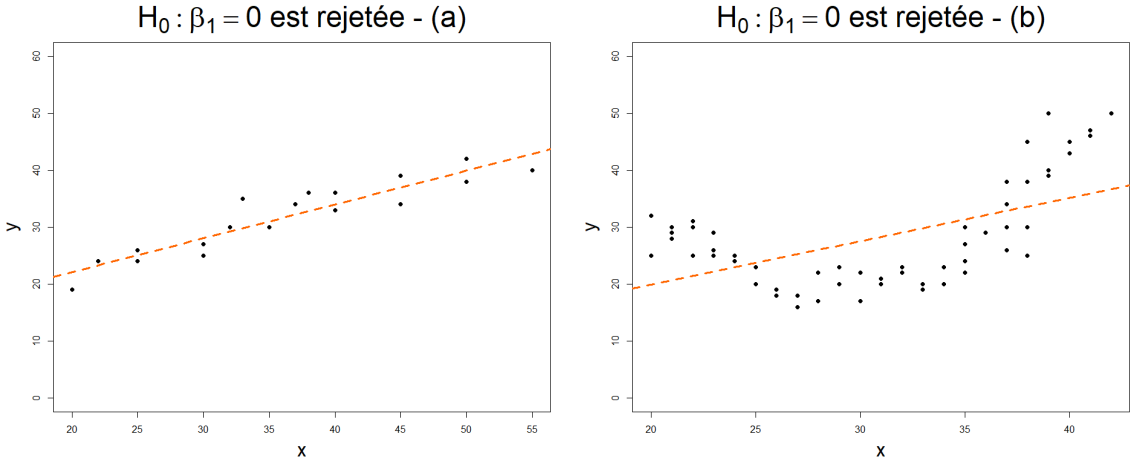

FIGURE 1.3 – Illustration lorsque l'hypothèse nulle est rejetée

Afin de juger la qualité du modèle de régression linéaire simple, nous mettons en place le coefficient de détermination. Ce coefficient permet de mesurer le pourcentage de la variation totale expliquée par le modèle.

Définition A.6
Le coefficient de détermination simple R^2 est défini de la façon suivante

$$R^2 = \frac{SSE}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(y_i - \bar{Y})^2} \quad et \quad 0 \leq R^2 \leq 1,$$

où

- *SSE = Sum of Squares Explained = variation expliquée due à la régression et*
- *SST = Sum of Squares Total = variation totale avant la régression.*

Lorsque le coefficient est nul, cela signifie que la variable X ne permet pas d'expliquer la variable Y et par conséquent, $\hat{y}_i = \bar{Y}$ pour $i = 1, \ldots, n$. Alors que si celui-ci vaut 1,

la variable X permet d'expliquer parfaitement la variable Y de sorte que $\hat{y}_i = y_i$ pour $i = 1, \ldots, n$.

A.2 Modèle de régression linéaire multiple

A.2.1 Modèle normal

Dans cette partie, l'objectif est d'expliquer une variable aléatoire Y non pas grâce à une variable X mais grâce à plusieurs variables X. Tout comme pour le modèle de régression linéaire simple, nous mettons en place un modèle afin d'obtenir la relation recherchée. Toute fois dans le cadre du modèle simple, la relation recherchée est une droite alors qu'ici le relation est obtenue sous forme d'un hyperplan.

Définition A.7
Le modèle de régression linéaire multiple est donné par

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_{k-1} + \epsilon,$$

où

- $X_1, \ldots, X_{k-1}$ *sont les variables explicatives (régresseurs),*
- Y *est la variable à expliquer et*
- $\epsilon \sim \mathcal{N}_n(0_{n\times 1},\, \sigma^2 I_n)$ *est un terme d'erreur.*

Semblablement au modèle de régression linéaire simple, la création du modèle est fondée sur un nombre fini n d'observations $y_1, \ldots, y_n$ sur lesquelles les $k-1$ variables explicatives ont été mesurées. Le modèle de régression linéaire multiple pour chaque vecteur $(x_{i,1}, \ldots, x_{i,k-1}, y_i)$ où $i = 1, \ldots, n$ est alors donné par

$$y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_k x_{i,k-1} + \epsilon_i, \tag{A.8}$$

ce qui donne lieu à la réécriture de la définition du modèle.

Définition A.8
Le modèle de régression linéaire multiple, sous forme matricielle, est donné par

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \ldots & x_{1,k-1} \\ 1 & x_{2,1} & \ldots & x_{2,k-1} \\ 1 & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \ldots & x_{n,k-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \Leftrightarrow Y = X\beta + \epsilon$$

où

- Y *= vecteur aléatoire observable composé de n lignes,*
- X *= matrice de constantes connues composé de n lignes et k colonnes tel que* $rg(X) = k$*,*
- β *= vecteur de paramètres inconnus composé de k colonnes et*
- ϵ *= vecteur aléatoire inobservable composé de n lignes tel que* $\epsilon \overset{iid}{\sim} \mathcal{N}_n(0_{n\times 1},\, \sigma^2 I_n)$.

Tout comme pour le modèle simple, nous avons dû poser certaines hypothèses sur les paramètres du modèle $\beta_0, \ldots, \beta_{k-1}$ afin de réaliser la prédiction de la variable Y. Dans le cas du modèle de régression linéaire simple, les hypothèses portent sur le terme d'erreur du modèle tandis que dans le cas du modèle de régression linéaire multiple, les hypothèses portent aussi sur la matrice X contenant les variables explicatives. En effet, l'hypothèse sur la matrice X nous assure d'avoir un modèle linéaire de rang plein et donc de bien définir le modèle. Si celle-ci est respectée cela signifie que $k \leq n$ et qu'il y a alors au moins autant d'équations que d'inconnues. De plus, par la normalité du terme d'erreur avec $\mathbb{E}(\epsilon) = 0_{n\times 1}$ et $\text{Var}(\epsilon) = \sigma^2 I_n$, nous mettons en évidence que $\mathbb{E}(Y) = X\beta$ et $\text{Var}(Y) = \sigma^2 I_n$. De plus, $Y \overset{iid}{\sim} \mathcal{N}_n(X\beta,\, \sigma^2 I_n)$.

Nous observons que si les colonnes de X ne sont pas linéairement indépendantes, les paramètres du modèle ne seront pas identifiables. Dans ce cas, il y a de la (multi)colinéarité parfaite entre le régresseurs.

Au vue de l'hypothèse $Y \overset{iid}{\sim} \mathcal{N}_n(X\beta,\, \sigma^2 I_n)$, nous utilisons ici qu'une des deux méthodes décrites précédemment afin d'obtenir l'estimateur de β vu que celles-ci établissent le même résultat. Le critère des moindres carrés est donc utilisé dans le but d'estimer les paramètres $\beta_0, \ldots, \beta_{k-1}$ du modèle de régression linéaire multiple. L'objectif est donc de trouver le coefficient β qui minimise

$$\begin{aligned} S(\beta) &= \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{k-1} \beta_j x_{ij})^2 \\ \Leftrightarrow S(\beta) &= \epsilon^T \epsilon = Y^T Y - 2Y^T X\beta + \beta^T X^T X\beta. \end{aligned}$$

Pour ce faire, nous calculons l'extremum de $S(\beta)$ et étant donné que la matrice hessienne de $S(\beta)$ est strictement définie positive alors l'extremum est un minimum.

$$\frac{\partial S(\beta)}{\partial \beta} = \text{O} \Rightarrow -2X^T Y + 2X^T X\beta = \text{O} \Leftrightarrow X^T X\hat{\beta} = X^T Y.$$

Définition A.9
L'estimateur des moindres carrés de $\beta = (\beta_0, \ldots, \beta_{k-1})$ est

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

Comme le rang de la matrice X est égale à k alors la matrice $X^T X$ est non singulière ce qui permet de bien définir l'estimateur des moindres carrés de β.

Encore une fois, nous utilisons le théorème de Gauss-Markov afin d'obtenir la propriété $BLUE$ pour l'estimateur de β.

Théorème A.2 (Gauss-Markov)
Pour le modèle de régression linéaire multiple défini en (A.8), l'estimateur des moindres carrés $\hat{\beta}$, qui fait partie de la classe des estimateurs linéaires non-biaisés de β, est celui qui a la plus petite variance. De plus, l'estimateur des moindres carrés $\hat{\beta}$ est $BLUE$. Par conséquence, la surface de régression estimée $\hat{Y} = X\hat{\beta}$ est un estimateur $BLUE$ de $\mathbb{E}(Y) = X\beta$.

Par ailleurs, sur base du résultat du théorème (A.2) et de la définition des estimateurs des moindres carrés (A.9), nous pouvons mettre en évidence certaines propriétés.

Proposition A.2
Sous les hypothèses du modèle de régression linéaire multiple, l'estimateur des moindres carrés $\hat{\beta}$ satisfait les propriétés suivantes :

- *$\hat{\beta}$ est non-biaisé, c'est-à-dire que $\mathbb{E}(\hat{\beta}) = \beta$ et*
- *$Var(\hat{\beta}) = \sigma^2(X^TX)^{-1}$.*

Ensuite, tout comme dans le cadre du modèle de régression linéaire simple, la valeur de σ^2 n'est pas connue c'est pourquoi nous définissons l'estimateur suivant afin d'avoir une estimation de la variance des erreurs dans le cadre d'un modèle linéaire multiple.

Définition A.10
Sous les hypothèses du modèle de régression linéaire multiple, l'estimateur non-biaisé de σ^2 dans le modèle de régression linéaire multiple est donné par

$$\hat{\sigma}^2 = \frac{(Y - X\hat{\beta})^T(Y - X\hat{\beta})}{n-k} = \frac{1}{n-k}\|Y - X\hat{\beta}\|^2.$$

Reformulons l'ensemble des estimateurs, non-biasés, des paramètres du modèle de régression linéaire multiple.

Définition A.11
Les estimateur de β et de σ^2 sont

$$\hat{\beta} = (X^TX)^{-1}X^TY \quad et \quad \hat{\sigma}^2 = \frac{1}{n-k}\|Y - X\hat{\beta}\|^2.$$

Les estimateurs $\hat{\beta}$ et $\hat{\sigma}^2$ sont les estimateurs optimaux de β et de σ^2, c'est-à-dire, ces estimateurs sont, dans la classe des estimateurs non-biaisées de β et de σ^2, ceux qui ont la plus petite variance.

Etant donné que l'estimateur des paramètres du modèle $\beta = (\beta_0, \ldots, \beta_{k-1})$ a été défini dans la définition (A.9), nous appliquons les concepts vus précédemment sur un exemple afin d'avoir une idée visuelle du comportement du modèle dans le cas multiple. Dans cet exemple, la variable de réponse Y, est le maximum journalier de la concentration en ozone. Les variables explicatives qui lui sont attribuées sont X_1 qui est la température à 12h, X_2 qui est le vent à 12h et X_3 qui est la nébulosité à 12h. Son modèle de régression linéaire multiple est

$$O3 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon.$$

Toutefois, tous les régresseurs ne sont pas toujours pertinents lors de l'élaboration d'un modèle. En conséquence, la concentration en ozone peut être expliquée uniquement par la température et le vent à 12h ce qui donne l'équation d'hyperplan suivante

$$\hat{Y} = -14 + 5X_1 + 2X_2$$

et est représentée à la figure (1.4).

A.2.2 Intervalles de confiance

Précédemment, dans le cadre du modèle simple, nous avons admis que la variable explicative X ne détient pas toujours le rôle espéré, c'est-à-dire que la variable Y n'est pas expliquée par la variable X. Dans le cadre du modèle linéaire, il y a $k-1$ variables explicatives potentiels. Par conséquent, le problème de test associé à l'hypothèse linéaire générale est mis en place afin de déterminer l'existence d'une relation de régression entre la variable à expliquer Y et les variables explicatives $X_1, \ldots, X_{k-1}$.

Définition A.12
Dans le modèle $Y = X\beta + \epsilon$, où $X \in \mathbb{R}^{n\times k}$ et $\epsilon \sim \mathcal{N}_n(0_{n\times 1}, \sigma^2 I_n)$, l'hypothèse linéaire générale est de la forme suivante

$$\mathcal{H}_0 : A\beta = C$$

où $A \in \mathbb{R}^{q\times k}$ est de rang q et $C \in \mathbb{R}^q$.

Par exemple, si

$$A = \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 0 & 1 \end{pmatrix} \text{ et } C = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

alors l'hypothèse $\mathcal{H}_0 : A\beta = C$ correspond à $\mathcal{H}_0 : \beta_1 = \ldots = \beta_{k-1} = 0$ et si l'hypothèse nulle est acceptée cela signifie que les variables explicatives $X_1, \ldots, X_{k-1}$ ne permettent pas d'expliquer la variable de réponse Y. Dans le cas contraire, si l'hypothèse nulle est rejetée alors cela signifie qu'au moins un β_i pour $i = 1, \ldots, k-1$ est non nul et qu'il y a donc au moins une parmi $k-1$ variables endogènes qui permet d'expliquer Y.

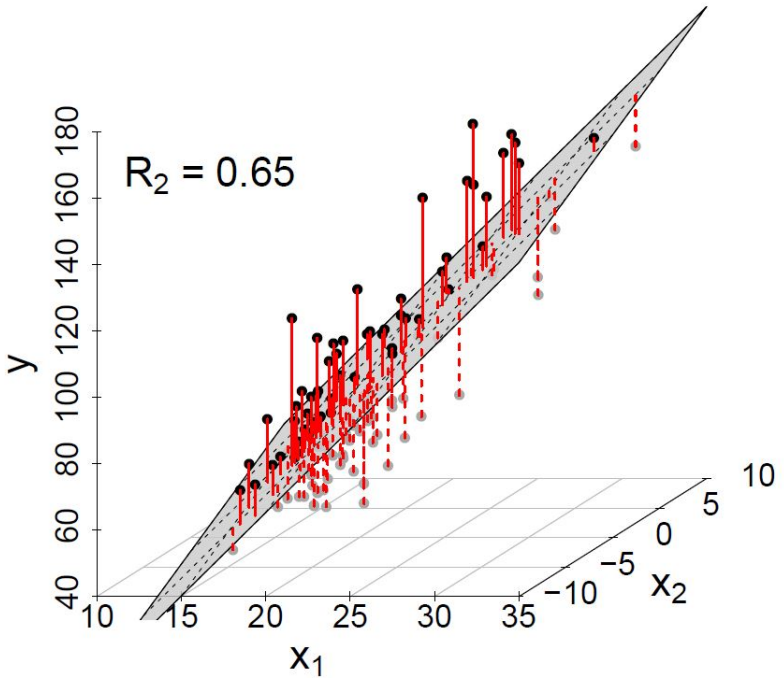

FIGURE 1.4 – Régression linéaire multiple de O3 tirée de [24].

C'est ainsi que l'élaboration d'une statistique de test basée sur le *Sum of Squared Residuals*, qui correspond à la variation inexpliquée après la régression, suivant une loi de Fisher est mise en place.

Définition A.13
Dans le modèle linéaire multiple normal $Y = X\beta + \epsilon$, *avec* $Y, \epsilon \in \mathbb{R}^n$, $\beta \in \mathbb{R}^k$ *et* $X \in \mathbb{R}^{n\times k}$. *Sous les hypothèses du modèle de régression multiple normal et sous* $\mathcal{H}_0 : A\beta = C$ *pour* $A \in \mathbb{R}^{q\times k}$, $C \in \mathbb{R}^q$, *la statistique de test*

$$F := \frac{(SSR_{\mathcal{H}_0} - SSR)/q}{SSR/(n-k)}$$

suit une loi de Fisher avec $(q, n-k)$ *degrés de liberté, où*

- SSR=$(Y - X\hat{\beta})^T(Y - X\hat{\beta})$,
- $SSR_{\mathcal{H}_0}$=$(Y - X\hat{\beta}_{\mathcal{H}_0})^T(Y - X\hat{\beta}_{\mathcal{H}_0})$ *et*
- $\hat{\beta}_{\mathcal{H}_0} = \hat{\beta} - (X^TX)^{-1}A^T[A(X^TX)^{-1}A^T]^{-1}(A\hat{\beta} - C)$.

Par ailleurs, l'hypothèse nulle est rejetée au niveau α *si* $\mathbb{P}(F \geq f|\mathcal{H}_0) \leq \alpha$.

Lorsque le test F conduit à l'acceptation de l'hypothèse $\mathcal{H}_0$, les estimateurs non-biaisés de σ^2 sont $\dfrac{SSR}{n-k}$ et $\dfrac{SSR_{\mathcal{H}_0} - SSR}{q}$. Toutefois, dans le cas contraire, ces estimateurs deviennent biaisés.

En outre, il est intéressant de se questionner sur la cause du rejet de l'hypothèse nulle $\mathcal{H}_0$. Une des façon d'y parvenir est de considérer les test individuels

$$\mathcal{H}_0 : a_j^T\beta = c_j, \text{ pour } j = 1, \ldots, q$$

où a_j^T est la j-ième ligne de la matrice A.

Afin de construire l'intervalle de confiance de β, nous devons mettre en évidence certaines propriétés des estimateurs des paramètres.

Proposition A.3
Dans le modèle linéaire multiple normal, les estimateurs sont donnés par

$$\hat{\beta} = (X^TX)^{-1}X^TY \quad et \quad \hat{\sigma}^2 = \frac{1}{n-k}\|Y - X\hat{\beta}\|^2.$$

Ces estimateurs possèdent les propriétés suivantes

- $\hat{\beta} \perp\!\!\!\perp \hat{\sigma}^2$,
- $\hat{\beta} \sim \mathcal{N}_k\left(\beta, \sigma^2(X^TX)^{-1}\right)$,
- $\frac{(n-k)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-k}$ *et*
- $Q := \frac{1}{\sigma^2}(\hat{\beta} - \beta)^TX^TX(\hat{\beta} - \beta) \sim \chi^2_k$.

Intervalle de confiance pour $a^T\beta$

Les propriétes précédentes, énoncées dans la proposition (A.3), ont plusieurs conséquences :

— $a^T\hat{\beta} \sim \mathcal{N}\left(a^T\beta, \sigma^2 a^T(X^TX)^{-1}a\right)$ donc $Z = \dfrac{a^T\hat{\beta} - a^T\beta}{\sqrt{\sigma^2 a^T(X^TX)^{-1}a}} \sim \mathcal{N}(0,\,1)$ et

— $Z \perp\!\!\!\perp U$ où $U = \frac{(n-k)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-k}$.

Dès lors par définition d'une loi de Student,

$$T = \frac{Z}{\sqrt{\frac{U}{(n-k)}}} = \frac{a^T\hat{\beta} - a^T\beta}{\sqrt{\hat{\sigma}^2 a^T(X^TX)^{-1}a}} \sim t_{n-k}$$

et T est une fonction pivotable pour $a^T\beta$.

Nous pouvons alors en déduire que

$$\mathbb{P}\left[-t_{n-k,1-\frac{\alpha}{2}} \leq T \leq t_{n-k,1-\frac{\alpha}{2}} \right] = 1 - \alpha$$

et donc d'obtenir l'intervalle de confiance au niveau de confiance α pour $a^T\beta$:

$$I_\alpha(a^T\beta) = a^T\hat{\beta} \pm t_{n-k,1-\frac{\alpha}{2}} \sqrt{\hat{\sigma}^2 a^T(X^TX)^{-1}a}. \tag{A.9}$$

Intervalle de confiance pour β_i avec $i = 0, \ldots, k-1$

Afin d'obtenir l'intervalle de confiance au niveau de confiance α pour β_i, il suffit de prendre $a^T = (0, \ldots, 0, 1, 0, \ldots, 0)$ où la composante non nulle se situe à la $i+1$ ème ligne :

$$I_\alpha(\beta_i) = \hat{\beta}_i \pm t_{n-k,1-\frac{\alpha}{2}} \sqrt{\hat{\sigma}^2 (X^TX)^{-1}_{i+1,i+1}}. \tag{A.10}$$

A.2.3 Tests d'hypothèses

Maintenant que les intervalles de confiance ont été obtenus dans la sous-section (A.2.2), nous pouvons établir le lien avec les tests d'hypothèses.

Test d'hypothèses sur $a^T\beta$

Soit l'intervalle de confiance $I_\alpha(a^T\beta)$, défini en (A.9), pour $a^T\beta$ au niveau de confiance α alors le problème de test est

$$\begin{cases} \mathcal{H}_0 : a^T\beta = c \\ \mathcal{H}_1 : a^T\beta \neq c,\ a^T\beta < c,\ a^T\beta > c. \end{cases}$$

Par construction de l'intervalle de confiance de $a^T\beta$, sous $\mathcal{H}_0$, la statistique du test

$$T = \frac{a^T\hat{\beta} - c}{\sqrt{\hat{\sigma}^2 a^T(X^TX)^{-1}a}} \sim t_{n-k}.$$

Si bien que l'hypothèse nulle $\mathcal{H}_0$ est rejetée au niveau α et donc que $\mathcal{H}_1$ est acceptée dans les cas suivants :

TABLE A.4 – Règle de rejet pour le test d'hypothèse sur $a^T\beta$

$\mathcal{H}_1$	Condition
$a^T\beta > c$	$t \geq t_{n-k,1-\alpha}$ ou $\mathbb{P}(T \geq t \vert \mathcal{H}_0) \leq \alpha$
$a^T\beta < c$	$t \leq -t_{n-k,1-\alpha}$ ou $\mathbb{P}(T \leq t \vert \mathcal{H}_0) \leq \alpha$
$a^T\beta \neq c$	$\vert t \vert \geq t_{n-k,1-\frac{\alpha}{2}}$ ou $\begin{cases} \mathbb{P}(T \leq t \vert \mathcal{H}_0) \leq \frac{\alpha}{2} \\ \text{ou } \mathbb{P}(T \geq t \vert \mathcal{H}_0) \leq \frac{\alpha}{2} \end{cases}$

Test d'hypothèses sur β_i

Soit l'intervalle de confiance $I_\alpha(\beta_i)$, défini en (A.10), pour β_i au niveau de confiance α alors le problème de test est

$$\begin{cases} \mathcal{H}_0 : \beta_i = 0 \\ \mathcal{H}_1 : \beta_i \neq 0,\ \beta_i < 0,\ \beta_i > 0. \end{cases}$$

Par construction de l'intervalle de confiance de β_i, sous $\mathcal{H}_0$, la statistique du test

$$T_i = \frac{\hat{\beta}_i}{\sqrt{\hat{\sigma}^2 (X^T X)^{-1}_{i+1,i+1}}} \sim t_{n-k}.$$

Si bien que l'hypothèse nulle $\mathcal{H}_0$ est rejetée au niveau α et donc que $\mathcal{H}_1$ est acceptée dans les cas suivants :

TABLE A.5 – Règle de rejet pour le test d'hypothèse sur β_i

$\mathcal{H}_1$	Condition
$\beta_i > 0$	$t_i \geq t_{n-k,1-\alpha}$ ou $\mathbb{P}(T_i \geq t_i \vert \mathcal{H}_0) \leq \alpha$
$\beta_i < 0$	$t_i \leq -t_{n-k,1-\alpha}$ ou $\mathbb{P}(T_i \leq t_i \vert \mathcal{H}_0) \leq \alpha$
$\beta_i \neq 0$	$\vert t_i \vert \geq t_{n-k,1-\frac{\alpha}{2}}$ ou $\begin{cases} \mathbb{P}(T_i \leq t_i \vert \mathcal{H}_0) \leq \frac{\alpha}{2} \\ \text{ou } \mathbb{P}(T_i \geq_i \vert \mathcal{H}_0) \leq \frac{\alpha}{2} \end{cases}$

Afin de juger la qualité du modèle de régression linéaire multiple, nous mettons en place le coefficient de détermination multiple. Ce coefficient est une généralisation du coefficient de détermination simple.

Définition A.14
Le coefficient de détermination multiple $R^2_{Y,\hat{Y}}$ est défini comme le coefficient de corrélation entre Y et $\hat{Y}$ au carré,

$$R^2_{Y,\hat{Y}} = \frac{\left(\sum_{i=1}^{n}(Y_i - \bar{Y})(\hat{Y}_i - \bar{\hat{Y}})\right)^2}{\left(\sum_{i=1}^{n}(Y_i - \bar{Y})\right)^2 \left(\sum_{i=1}^{n}(\hat{Y}_i - \bar{\hat{Y}})\right)^2}.$$

Toutefois, le coefficient de détermination multiple, $R^2_{Y,\hat{Y}}$, n'est pas toujours le plus efficace. En effet, le coefficient d'un modèle plein est toujours plus grand que celui basé sur un modèle réduit, c'est-à-dire sur un modèle qui tient compte que de certains régresseurs. Toutefois, dans le cas multiple, nous privilégierons un modèle plus simple si certaines variables explicatives n'influencent pas la variable à expliquer Y et par ailleurs si elles changent très peu le coefficient $R^2_{Y,\hat{Y}}$. C'est pourquoi nous introduisons un nouveau coefficient dans le cas d'un modèle multiple.

Définition A.15
Le coefficient de détermination multiple ajusté R^2_{aju} est

$$R^2_{aju} = 1 - \left(\frac{n-1}{n-k}\right)(1 - R^2_{Y,\hat{Y}}).$$

Dans ce chapitre, le modèle de régression linéaire multiple ainsi que ses propriétés ont été vues en détails. Nous les appliquons sur un jeu de données, grâce au logiciel R, afin de pouvoir analyser le résultat qui en découle. Le jeu de données utilisé est un jeu de données disponible sur R et se nomme $LifeCycleSavings$ qui correspond à des données sur le taux d'épargne entre 1960 et 1970. L'objectif est donc d'analysé la dépendance entre la variable explicative, *sr*, qui est le ratio d'épargne avec les régresseurs suivants :

— X_1 : *pop15* est le pourcentage de la population qui est de moins de 15 ans,
— X_2 : *pop75* est le pourcentage de la population qui est de plus de 75 ans,
— X_3 : *dpi* est le revenu réel disponible par habitant et
— X_4 : *ddpi* est le pourcentage de croissance de *dpi*.

Afin de pouvoir tirer une relation valide entre la variable explicative et les régresseurs, le jeu de données doit vérifier plusieurs hypothèses. En effet, les résidus doivent être indépendants et doivent suivre une loi normal de moyenne 0 et d'écart-type constant. Pour ce faire, nous utilisons plusieurs tests et leur hypothèse nulle respective est acceptée lorsque la p-valeur du test est supérieur au seuil $\alpha = 0.05$.

En premier lieu, le test de Durbin-Watson qui donne une solution au problème de test suivant

$$\begin{cases} \mathcal{H}_0 : \text{il n'y a pas d'autocorrélation entre les résidus} \\ \mathcal{H}_1 : \text{il y a une autocorrélation entre les résidus ,} \end{cases}$$

montre que les résidus sont indépendants car la p-valeur associée au test vaut 0.774. Ensuite, le test de Shapiro, qui donne une solution au problème de test suivant

$$\begin{cases} \mathcal{H}_0 : \text{les données sont normalement distribuées} \\ \mathcal{H}_1 : \text{les données ne sont pas normalement distribuées} , \end{cases}$$

est appliqué aux résidus de la régression et met en évidence que les résidus suivent une loi normale de moyenne 0 car la p-valeur associée au test vaut 0.8524. Pour finir, le test de Breush-Pagan qui donne une solution au problème de test suivant

$$\begin{cases} \mathcal{H}_0 : \text{les résidus possèdent la même variance} \\ \mathcal{H}_1 : \text{les résidus ne possèdent pas la même variance}, \end{cases}$$

signale que les résidus sont distribués de manière homogène car la p-valeur associée au test vaut 0.13153.

Maintenant que les hypothèses ont été vérifiées, nous utilisons la fonction *summary*, dont la sortie est représentée à la figure (1.5), afin d'accéder au résultat de la régression. En conséquence, l'équation de l'hyperplan de régression est

$$\hat{Y} = 28.57 - 0.46X_1 - 1.69X_2 - 0.0003X_3 + 0.41X_4.$$

En outre, la p-valeur du test de Fisher est 0.0007904 ce qui mène au rejet de *l'hypothèse linéaire générale* et donc signifie qu'au moins un des régresseurs explique la variable de réponse. Dans ce cas, nous devons regarder la p-valeur des tests individuels réalisés pour chaque régresseur afin de voir si le paramètre β_i pour $i = 1, \ldots, 4$ associé au régresseur est nul. Notamment, seulement les deux régresseurs *pop15* et *ddpi* possèdent une p-valeur inférieure au seuil α ce qui signifie qu'ils permettent d'expliquer la variable de réponse.

De plus, de sorte à déterminer la qualité du modèle tout en tenant compte du nombre de régresseurs utilisés, nous analysons la valeur du coefficient de détermination multiple ajusté. Toutefois, seulement 28% de la variation totale est expliquée par le modèle.

De façon à obtenir les intervalles de confiance des paramètres du modèle, nous utilisons la fonction *confint*. Par exemple, l'intervalle de confiance de 95% de l'estimation du paramètre du modèle associé à *pop15* est $[-0.752517542, -0.169868752]$.

```
Residuals:
    Min      1Q  Median      3Q     Max
-8.2422 -2.6857 -0.2488  2.4280  9.7509

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.5660865  7.3545161   3.884 0.000334 ***
pop15       -0.4611931  0.1446422  -3.189 0.002603 **
pop75       -1.6914977  1.0835989  -1.561 0.125530
dpi         -0.0003369  0.0009311  -0.362 0.719173
ddpi         0.4096949  0.1961971   2.088 0.042471 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.803 on 45 degrees of freedom
Multiple R-squared:  0.3385,    Adjusted R-squared:  0.2797
F-statistic: 5.756 on 4 and 45 DF,  p-value: 0.0007904
```

FIGURE 1.5 – Résultat de la fonction summary dans le cadre de notre exemple