2. Codifier la structure de voisinage

MARIE-PIERRE DE BELLEFON, VINCENT LOONIS, RONAN LE GLEUT
Insee

Résumé

Après avoir choisi l'échelle d'agrégation des données et effectué une première analyse descriptive grâce aux outils cartographiques, la deuxième étape d'une analyse spatiale est la définition du voisinage d'un objet. La définition du voisinage est indispensable pour mesurer la force des relations spatiales entre les objets, c'est-à-dire la façon dont les voisins s'influencent les uns les autres. Elle permet de calculer des indices d'autocorrélation spatiale, de mettre en œuvre les techniques d'économétrie spatiale, d'étudier la distribution spatiale des observations, mais aussi d'effectuer un échantillonnage spatial ou de partitionner un graphe.

L'enjeu de ce chapitre est de réussir à définir des relations de voisinage cohérentes avec les véritables interactions spatiales entre les objets. Ce chapitre présente plusieurs notions de voisinage, fondées sur la contiguïté ou sur les distances entre observations. La question du poids accordé à chaque voisin est aussi abordée. La mise en œuvre pratique s'appuie sur les packages R *spdep*, *tripack*, *spsurvey* et *tsp*.

R La lecture préalable du chapitre 1 : "Analyse spatiale descriptive" est recommandée.

2.1 Définir les voisins

2.1.1 Caractéristiques des relations entre objets spatiaux

Considérons une surface $\mathfrak{R}$. Cette surface peut être divisée en n zones mutuellement exclusives. Deux zones adjacentes sont séparées par une frontière commune. Les frontières peuvent naître de discontinuités spatiales (frontières administratives ou environnementales). Elles peuvent également être issues des polygones de Voronoï calculés à partir des points d'intérêt (voir chapitre 1 : "Analyse spatiale descriptive").

Encadré 2.1.1 — Définition mathématique des relations spatiales . Les relations spatiales $\mathscr{B}$ sont un sous-ensemble du produit cartésien $\mathbb{R}^2 \times \mathbb{R}^2 = \left\{(i,j) : i \in \mathbb{R}^2, j \in \mathbb{R}^2\right\}$ des couples (i,j) d'objets spatiaux, c'est-à-dire l'ensemble des couples (i,j) tels que i et j soient tous deux des objets spatiaux identifiés par leurs coordonnées géographiques, et que (i,j) soit différent de (j,i).
Un objet spatial ne peut pas être relié à lui-même : $(i,i) \notin \mathscr{B}$. De plus si $(i,j) \subseteq \mathscr{B}$ et $(j,i) \subseteq \mathscr{B}$ pour tout couple d'objets spatiaux, les relations spatiales sont dites *symétriques* (TIEFELSDORF 1998).

Les relations spatiales sont multidirectionnelles et multilatérales. Elles se distinguent en cela des relations temporelles qui n'autorisent que des relations séquentielles le long de l'axe passé-présent-futur.

La figure 2.1 illustre la démarche de codification des relations spatiales. Cette démarche permet de transcrire de manière systématique la complexité de l'espace géographique en un ensemble fini de données analysables par un ordinateur.
Tout d'abord, la zone d'étude est subdivisée en aires mutuellement exclusives. Chaque aire contient un point de référence (souvent son centroïde). Ensuite, les relations spatiales peuvent être spécifiées par un graphe de voisinage reliant les aires considérées comme voisines, ou par une matrice contenant les coordonnées géographiques des points de référence. La troisième étape consiste à coder le graphe dans une matrice de voisinage, ou à transformer les coordonnées géographiques en une matrice de distances.
La matrice de voisinage mesure la similarité entre les observations. Une valeur supérieure strictement à zéro indique que les observations sont considérées comme voisines. Par exemple, dans le cas de la matrice binaire présentée en figure 2.1 :

$$w_{ij} = \begin{cases} 1 & si \quad i \quad et \quad j \quad sont \quad reliés \quad dans \quad l'espace \\ 0 & sinon \end{cases} \tag{2.1}$$

Inversement, la matrice de distances mesure une dissimilarité entre zones. Plus d_{ij} est élevé, plus les zones diffèrent. Avec, si l'on utilise une distance euclidienne : $d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$, α et β étant les coordonnées géographiques des observations.

La matrice de voisinage est utilisée dans l'étude des données spatiales surfaciques, tandis que la matrice de distances sert plutôt à la géostatistique (voir chapitre 5 : "Géostatistique"). On peut cependant passer de l'une à l'autre en définissant une distance minimale au-delà de laquelle les observations ne sont plus voisines.

La structure de la dépendance spatiale peut ne pas être géographique. Toute relation duale pertinente permet de définir un graphe de voisinage. Citons par exemple :

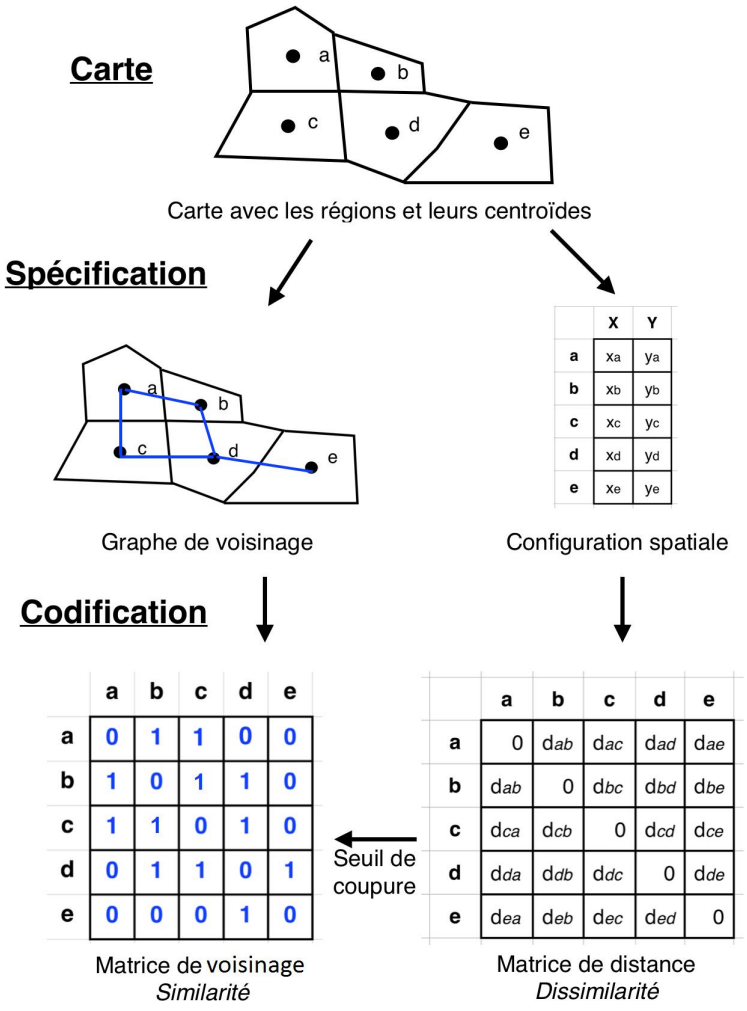


FIGURE 2.1 – Codification des relations spatiales
Source : TIEFELSDORF *1998*

— **au niveau des individus** : les liens d'amitié, la fréquence des communications, les citations dans les articles de recherche scientifique ;
— **au niveau des entreprises** : les liaisons siège-filiale, les similitudes en termes de marchés ;
— **au niveau international** : les alliances stratégiques, les flux commerciaux, l'appartenance commune à une organisation, les échanges culturels, les flux migratoires.

Les sections suivantes détaillent différentes spécifications de voisinage.

L'objet "liste de voisins" en R

Le package *spdep* permet de définir les relations de voisinage entre objets spatiaux. Dans R, la classe d'un objet définit l'ensemble de ses propriétés et la façon dont le statisticien peut l'utiliser. Les relations de voisinage sont enregistrées dans un objet de classe `nb`.
Soit n observations spatiales et *voisins_nb* l'objet spatial contenant les relations de voisinage associées. *voisins_nb* est une liste de longueur n. Chaque élément $[i]$ de la liste contient un vecteur avec l'index des voisins de l'élément d'index i. Si $[i]$ n'a pas de voisins, la liste contient uniquement 0. La liste contient également un vecteur de caractères associés à chaque zone de voisinage, ainsi qu'une valeur logique indiquant si la relation est symétrique ou pas (voir figure 2.2). Les informations principales sur l'objet *voisins_nb* peuvent être obtenues grâce à la fonction :

```
summary(voisins_nb)
```

La documentation du package *spdep* donne de plus amples informations (BIVAND et al. 2013b).

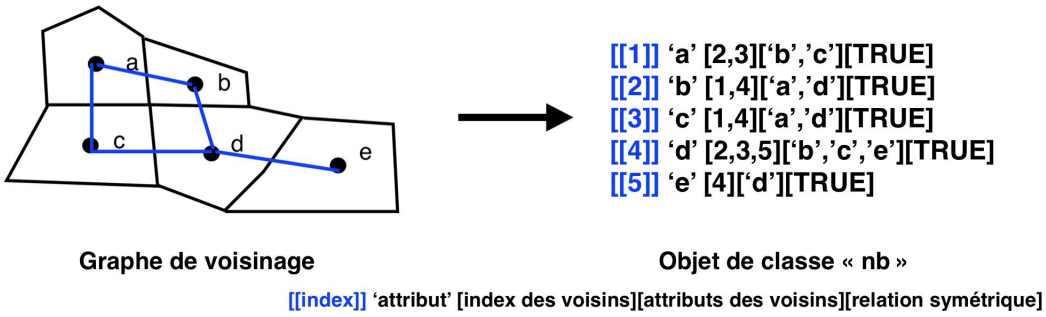


FIGURE 2.2 – La liste de voisins dans *spdep*

2.1.2 Définir les voisins en s'appuyant sur la distance

Dès lors qu'on dispose d'un ensemble de points répartis sur le territoire, on peut calculer les distances entre eux. Ces points peuvent être des lieux particuliers où l'information a été observée, ou l'ensemble des points représentatifs de chaque zone, par exemple leur centroïde. Dans ce cas, l'hypothèse sous-jacente est que la répartition de la valeur de la variable d'intérêt au sein de chaque zone est suffisamment homogène pour que l'approximation de l'attribuer à un unique point ne soit pas trop grossière.

Les graphes de voisinage matérialisent les liens entre les différentes entités. On les définit de façon à ce qu'ils représentent le plus fidèlement possible la structure spatiale sous-jacente. Il existe de nombreux graphes de voisinage différents. Nous présenterons ici les graphes fondés sur des notions géométriques et sur les voisins les plus proches.

Graphes de voisinage fondés sur des notions géométriques

La **triangulation de Delaunay** est une méthode géométrique qui relie les points sous forme de triangles tels que l'angle minimal de l'ensemble des triangles soit maximisé (cette triangulation

cherche à éviter les triangles "allongés"), voir figure 2.3 et 2.5a. La triangulation de Delaunay possède d'intéressantes propriétés géométriques et mathématiques. On peut cependant affiner la notion de voisinage.

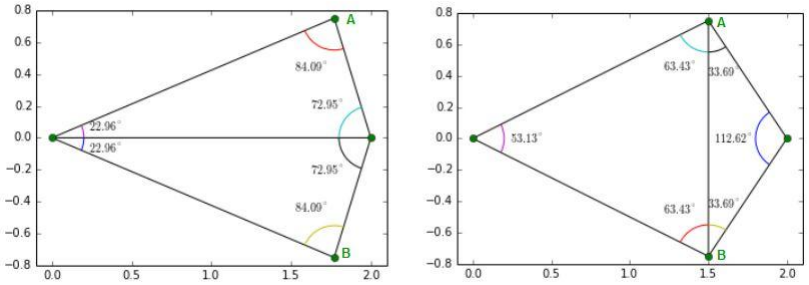


FIGURE 2.3 – Triangulation de Delaunay associée à différentes positions des points A et B
Source : *Gustavo [CC BY-SA 3.0 (https ://creativecommons.org/licenses/by-sa/3.0)], de Wikimedia Commons*

Le **graphe de la sphère d'influence** relie deux points si leurs "cercles du voisin le plus proche" se coupent. Le "cercle du voisin le plus proche" d'un point P est le plus grand cercle centré en P et qui ne contient pas d'autres points que P (voir figure 2.4 et 2.5b). Les graphes de la sphère d'influence ne sont pas nécessairement connectés, c'est-à-dire que tous les points de l'ensemble d'étude ne sont pas forcément reliés entre eux.

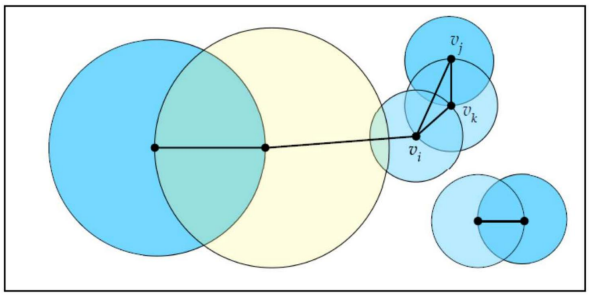


FIGURE 2.4 – Le graphe de la sphère d'influence d'un ensemble de points
Source : TOUSSAINT *2014*

Le **graphe de Gabriel** relie deux points p_i et p_j si et seulement si tous les autres points sont en dehors du cercle de diamètre $[p_i, p_j]$. Le graphe de Gabriel élimine certaine des liaisons du graphe de Delaunay, voir figure 2.5c.

Le **graphe des voisins relatifs** considère que deux points p_i et p_j sont voisins si

$$d(p_i, p_j) \leq max\,[d(p_i, p_k), d(p_j, p_k)] \quad \forall k = 1, ..., n \quad k \neq i, j \tag{2.2}$$

avec $d(p_i, p_j)$ la distance entre p_i et p_j. Le graphe des voisins relatifs impose moins de connexions que la triangulation de Delaunay ou le graphe de la sphère d'influence, voir figure 2.5d. TOUSSAINT 1980 juge qu'il s'adapte mieux aux données en imposant le moins de liaisons.

Les graphes de voisinage présentés ici sont tous des sous-graphes de la triangulation de Delaunay (voir figure 2.5). Ils ont l'avantage de ne laisser aucune unité sans voisins. En revanche, ils ne sont implémentés en R qu'avec la distance euclidienne, alors que d'autres types de distance, comme la distance du grand cercle, peuvent être plus adaptées à certaines études.

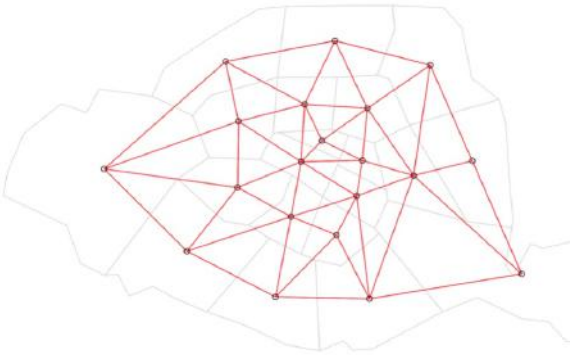
Application avec R

```
library(rgdal) #Pour importer les fichiers MIF/MID
library(maptools) #Pour importer les fichiers Shapefile
library(tripack) #Pour calculer les voisins basés sur la distance
library(spdep)

#Importation du fichier spatial
arr75 <- readOGR("~/ArmF.TAB", "ArmF")

#Voisins fondés sur la notion de graphe
#Le fichier en entrée est une matrice de coordonnées géographiques
#ou un objet de type SpatialPoints
coords <- coordinates(arr75)
IDs <- row.names(as(arr75,"data.frame"))

#Triangulation de Delaunay
Sy4_nb <- tri2nb(coords, row.names=IDs)
plot(arr75, border='lightgray')
plot(Sy4_nb,coordinates(arr75),add=TRUE,col='red')

#Graphe de la sphère d'influence
Sy5_nb <- graph2nb(soi.graph(Sy4_nb,coords),row.names=IDs)
plot(arr75, border='lightgray')
plot(Sy5_nb,coordinates(arr75),add=TRUE,col='red')

#Graphe de Gabriel
Sy6_nb <- graph2nb(gabrielneigh(coords), row.names=IDs)
plot(arr75, border='lightgray')
plot(Sy6_nb,coordinates(arr75),add=TRUE,col='red')

#Graphe des voisins relatifs
Sy7_nb <- graph2nb(relativeneigh(coords), row.names=IDs)
plot(arr75, border='lightgray')
plot(Sy7_nb,coordinates(arr75),add=TRUE,col='red')
```

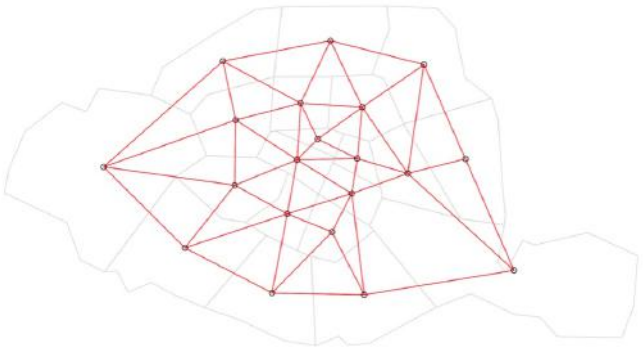
Graphes de voisinage fondés sur les voisins les plus proches

Une deuxième méthode consiste à sélectionner comme voisins les k points les plus proches (figure 2.6). Cette méthode a l'avantage de ne laisser aucun point sans voisin, ce qui n'est pas nécessaire pour conduire une analyse spatiale, mais reflète en général mieux la réalité (il est rare qu'une zone géographique soit complètement isolée). En revanche il est parfois difficile d'identifier la valeur k qui reflète les vraies relations spatiales sous-jacentes. Les graphes fondés sur les k voisins les plus proches ne sont pas nécessairement symétriques.

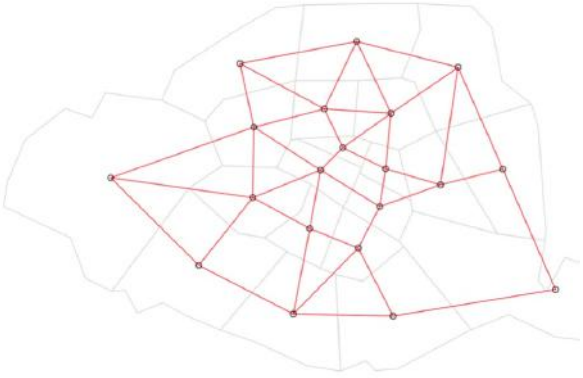
On peut également ne conserver que les points situés à une certaine distance. La fonction



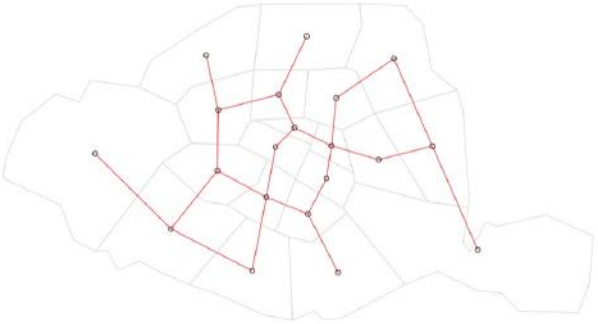
(a) Triangulation de Delaunay



(b) Graphe de la sphère d'influence



(c) Graphe de Gabriel



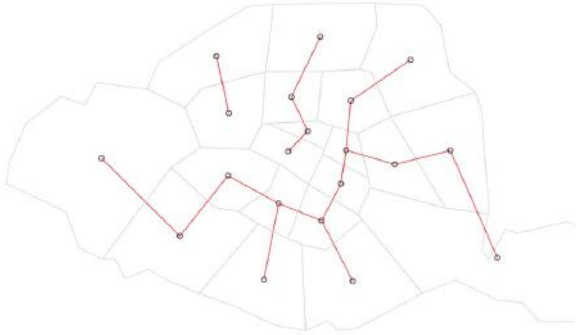
(d) Graphe des voisins relatifs

FIGURE 2.5 – Quatre graphes de voisinage des arrondissements parisiens fondés sur des notions géométriques

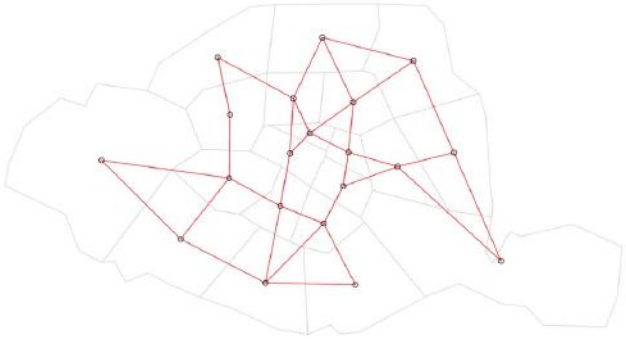
`nbdists` de R permet de calculer le vecteur des distances entre les voisins. On peut ainsi obtenir la distance minimale d_{min} au-delà de laquelle tous les points ont au moins un voisin, puis utiliser la fonction `dnearneighb` pour retenir comme voisins les seuls points situés entre les distances 0 et d_{min}. Cette méthode "de la distance minimale" n'est pas adaptée aux données irrégulièrement espacées car la distance minimale nécessaire pour qu'un point relativement isolé ait au moins un voisin est beaucoup plus élevée que la distance du plus proche voisin d'un point situé dans une zone dense. Il y aura donc de grandes disparités dans le nombre de voisins (BIVAND et al. 2013b), voir figure 2.6d.

Application avec R - Source : BIVAND *et al. 2013b*

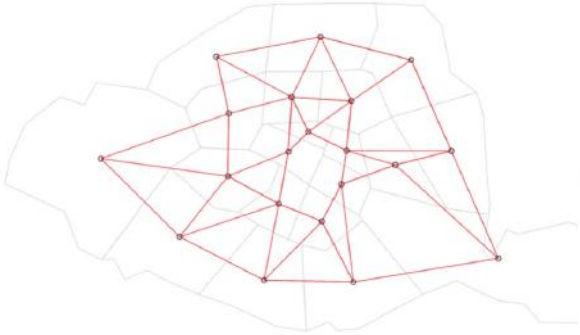
```
#Graphes fondés sur les plus proches voisins
Sy8_nb <- knn2nb(knearneigh(coords,k=1),row.names=IDs)
Sy9_nb <- knn2nb(knearneigh(coords,k=2),row.names=IDs)
Sy10_nb <- knn2nb(knearneigh(coords,k=3),row.names=IDs)

plot(arr75, border='lightgray')
plot(Sy8_nb,coordinates(arr75),add=TRUE,col='red')

#Etude de la distance moyenne du voisin le plus proche
dsts <- unlist(nbdists(Sy8_nb,coords))
summary(dsts)
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    820    1188    1678    1707    2016    3412
max_1nn <- max(dsts)

#Calcul et représentation des voisins à la distance minimale
Sy11_nb <- dnearneigh(coords, d1=0, d2=max_1nn, row.names=IDs)
plot(arr75, border='lightgray')
plot(Sy11_nb,coordinates(arr75),add=TRUE,col='red')
```



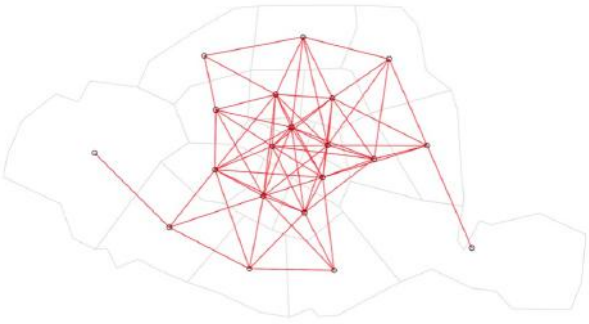
(a) Voisin le plus proche



(b) Deux plus proches voisins



(c) Trois plus proches voisins



(d) Voisins à la distance minimale

FIGURE 2.6 – Quatre graphes fondés sur les plus proches voisins des arrondissements parisiens

2.1.3 Définir les voisins en s'appuyant sur la contiguïté

Lorsque les données surfaciques consistent en une partition de l'ensemble du territoire, la notion de "distance entre les observations" peut devenir assez ambiguë. L'exemple 2.1 illustre les limites de l'utilisation de la distance entre centroïdes pour définir la notion de voisinage.

■ **Exemple 2.1 — Ambiguïté de la notion de distance entre centroïdes.** Soient R_1, R_2, R_3 trois zones distinctes. On peut considérer que comme R_2 et R_3 sont séparées dans l'espace, mais toutes les deux adjacentes à R_1, elles sont toutes les deux plus proches de R_1 que l'une de l'autre. Cependant, les centroïdes de ces zones sont équidistants entre eux (voir figure 2.7). Résumer la proximité entre zones par la distance entre centroïdes conduit à perdre une partie de la richesse des relations spatiales.

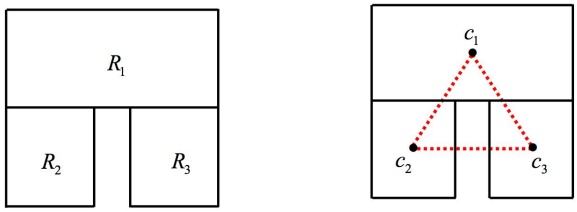


FIGURE 2.7 – Gauche : trois zones - Droite : distances entre centroïdes
Source : SMITH *2016*

■

Cette sous-section introduit différentes notions de contiguïté et présente la façon dont le package R *spdep* permet de créer une liste de voisins.

Au sens de la contiguïté **Rook**, les voisins possèdent au moins un segment de frontière commune. Cela correspond aux déplacements de la Tour du jeu d'échecs. Pour que deux zones soient voisines au sens de la contiguïté **Queen**, il suffit qu'elles partagent un point de frontière commune. Cela correspond aux déplacements de la Reine du jeu d'échecs. La figure 2.8 illustre ces notions dans le cas d'une grille régulière de points. Quand les polygones ont une forme et une surface irrégulières, les différences entre voisinage Rook et Queen deviennent plus difficiles à appréhender. Notons également qu'une zone très étendue entourée de plus petites zones aura un nombre de voisins beaucoup plus important que ses voisines.
Le voisinage au sens de la contiguïté est souvent utilisé pour étudier des données démographiques et sociales, pour lesquelles être d'un côté ou de l'autre d'une frontière administrative peut avoir plus d'importance qu'être situé à une certaine distance l'une de l'autre.

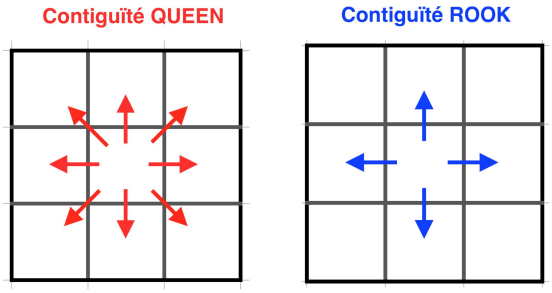


FIGURE 2.8 – Définition de la contiguïté Queen et Rook

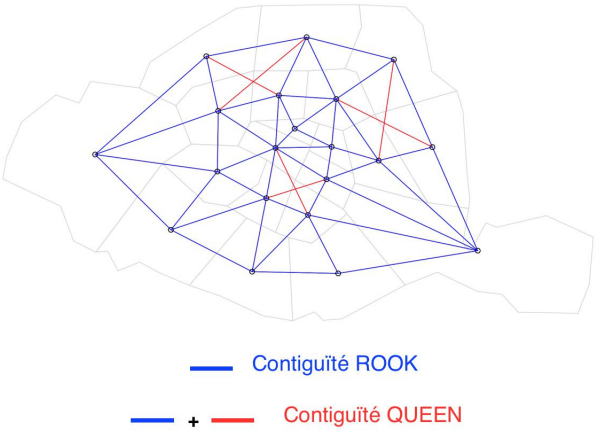


FIGURE 2.9 – Contiguïté Queen et Rook des arrondissements parisiens

Application avec R

Construction des graphes de voisinage Queen et Rook pour les arrondissements parisiens (figure 2.9)

```
#Le fichier en entrée est un fichier SpatialPolygons
#Extraction de la liste des voisins au sens QUEEN (par défaut)
arr75.nb<-  poly2nb(arr75)
```

```
#Extraction de la liste des voisins au sens ROOK
arr75.nb.ROOK <- poly2nb(arr75, queen=FALSE)

#Représentation visuelle des voisins :
plot(arr75, border='lightgray')
plot(arr75.nb, coordinates(arr75),add=TRUE,col='red')
plot(arr75.nb.ROOK, coordinates(arr75),add=TRUE,col='blue')
```

2.1.4 Définir les voisins en s'appuyant sur l'optimisation d'une trajectoire

Autour du voyageur de commerce

Certaines méthodes comme celle de l'échantillonnage spatial (voir chapitre 10 : "Échantillonnage spatial") nécessitent un tri préalable des données. Quand ces dernières sont caractérisées par deux variables (à savoir leurs coordonnées géographiques dans le plan), le choix de la méthode de tri est un problème théorique complexe.

Une solution consiste à faire passer un *chemin* par l'ensemble des points puis à les trier selon leur ordre d'apparition quand on parcourt le chemin. Les voisins d'un point donné sont alors les points situés juste avant ou juste après sur le chemin.

Parmi l'ensemble des chemins possibles, certains ont des caractéristiques plus adaptées à l'objectif recherché, comme par exemple celui de réduire la variance d'échantillonnage. C'est le cas du *plus court chemin*. Il minimise la somme des distances entre deux points consécutifs. Ce chemin qui ne fixe pas de contraintes particulières sur le point de départ ou d'arrivée est connu dans la littérature de la théorie des graphes comme *chemin de Hamilton* (figure 2.11b) associé à un graphe dont les arêtes sont pondérées. Un cas particulier célèbre de *plus court chemin* est celui dit du voyageur de commerce. Il représente le chemin que doit suivre un voyageur de commerce pour visiter l'ensemble de ses clients, tout en minimisant la distance parcourue et en rentrant chez lui le soir. Un tel chemin correspond à un cycle hamiltonien (figure 2.11c).

La recherche d'un plus court chemin est un problème classique d'optimisation dans le cadre de la théorie des graphes. Il est apparu notamment pour la résolution par Euler du problème des sept ponts de Königsberg [1]. Il intervient dans les questions relatives aux graphes eulériens ou hamiltoniens [2]. Il n'existe pas aujourd'hui d'algorithme en temps polynomial permettant de trouver le plus court chemin. Quand le nombre de points est grand, la recherche du chemin optimal passe par des heuristiques [3] conduisant à un optimum local. Elles sont disponibles dans le package *TSP* de R (HAHSLER et al. 2017).

Quand la distance est euclidienne et le nombre de points raisonnable, de l'ordre de quelques centaines, une solution exacte peut être trouvée grâce au programme `concorde` (APPLEGATE et al. 2006). Ce programme peut être appelé directement depuis R et le package *TSP*.

Enfin, la recherche d'un chemin hamiltonien à partir d'une matrice de distances est équivalente à celle d'un cycle hamiltonien pour peu que l'on rajoute une ligne et une colonne formées de 0 à la matrice originelle (GARFINKEL 1985). Le package *TSP* prévoit explicitement ce cas avec la fonction `insert-dummy`.

1. La question étudiée par Euler était : dans la ville de Königsberg peut-on faire une promenade en parcourant chacun des 7 ponts une fois et une seule ?

2. Un graphe eulérien est un graphe que l'on peut parcourir en partant d'un sommet quelconque et en empruntant exactement une fois chaque arête pour revenir au sommet de départ. Il correspond à un dessin qu'on peut tracer sans lever le crayon. Un graphe hamiltonien est un graphe que l'on peut parcourir en passant par tous les sommets une fois et une seule. Un graphe hamiltonien n'est pas nécessairement eulérien car dans un cycle hamiltonien, on peut très bien négliger de passer par certaines arêtes.

3. une heuristique est une méthode de calcul qui fournit rapidement (en temps polynomial) une solution réalisable, pas nécessairement optimale.

D'autres méthodes

La méthode *general randomized tessellation stratified* (GRTS , STEVENS JR et al. 2004) est populaire en échantillonnage spatial, puisqu'elle permet d'obtenir un échantillon spatialement réparti pour une population finie d'individus (unités distinctes et identifiables de dimension 0 d'une population discrète, par exemple des arbres d'une forêt), une population linéaire (unités continues de dimension 1, e.g. des rivières) ou une population de surfaces (unités continues de dimension 2, par exemple des forêts). Elle s'appuie sur un chemin construit à partir d'une classe de fonctions appelée *quadrant-recursive* (MARK 1990), permettant d'assurer que certaines relations de proximité de l'espace à deux dimensions soient toujours préservées dans l'espace à une dimension.

L'idée de la méthode est de projeter les coordonnées sur un carré unitaire, puis de découper ce carré en quatre cellules, chacune d'entre elles étant à nouveau divisée en quatre sous-cellules, etc. À chaque cellule on attribue une valeur résultant de l'ordre dans lequel le découpage a été effectué, ce qui permet finalement de placer les unités sur le chemin parcourant l'espace à deux dimensions.

La figure 2.10 montre les premières étapes du découpage, qui peut être mis en œuvre avec le package `spsurvey` de R (KINCAID et al. 2016). La méthode GRTS conduit cependant à créer de *grands sauts* (figures 2.11d) dans les chemins, ce qui peut affecter la précision des estimations.

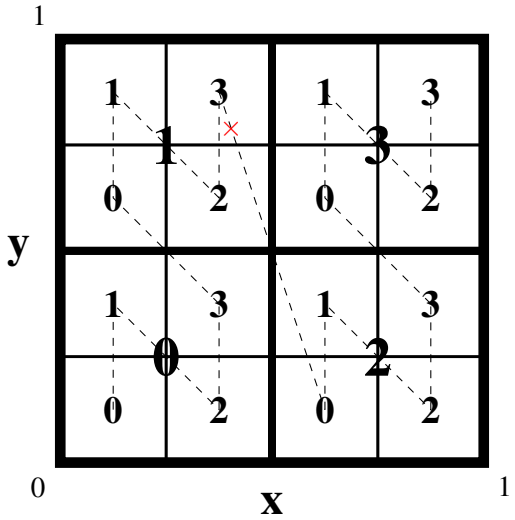
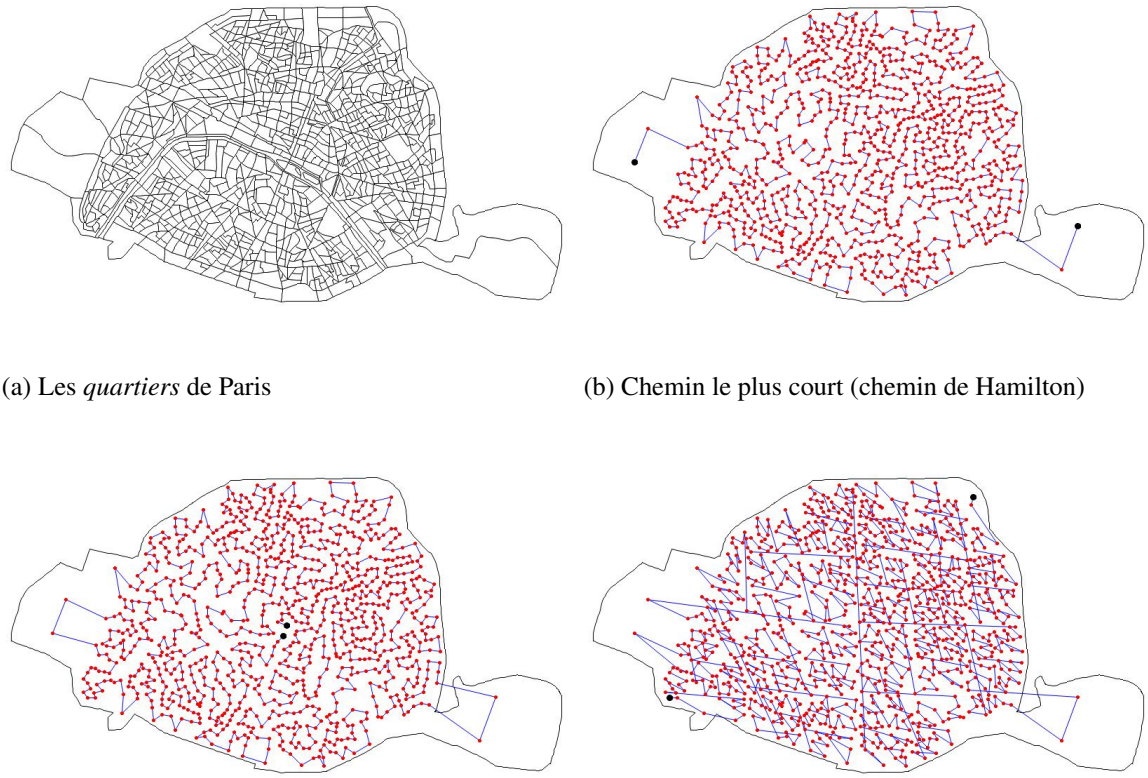


FIGURE 2.10 – Construction d'un chemin avec la méthode GRTS
Note : À l'unité dont la position est une croix rouge est associée la valeur "13" permettant ensuite de la positionner sur le chemin.

Application avec R - Source : Recherche d'un plus court chemin

```
library(TSP)
library(miscTools)

#Il faut télécharger l'utilitaire concorde à cette adresse :

http://www.tsp.gatech.edu/concorde/downloads/downloads.htm
#et l'appeler depuis R

Sys.setenv(PATH=paste(Sys.getenv("PATH"),"z:/cygwin/App/Runtime/Cygwin/bin"
    ,sep=";"))
concorde_path("Z:/concorde/")
```

```
#Les données en entrée sont une matrice de distances
test <-as.matrix(read.csv("U:/paris.csv",header=FALSE,sep="\t"))

#les erreurs d'arrondis peuvent conduire à ce que la matrice ne soit
    parfaitement symétrique.

tsp <-(symMatrix(test[upper.tri(test, TRUE)],nrow =nrow(test), byrow=TRUE))
#on crée un objet lisible par TSP
tsp<-TSP(tsp)
#On applique la méthode concorde à cet objet.
tour<-solve_TSP(tsp, method = "concorde")
```



(a) Les *quartiers* de Paris

(b) Chemin le plus court (chemin de Hamilton)

(c) Cycle hamiltonien, chemin du voyageur de commerce

(d) Construction d'un chemin avec la méthode GRTS

FIGURE 2.11 – Recherche de chemins passant par tous les *quartiers* de Paris

2.2 Accorder des poids aux voisins

2.2.1 Passer d'une liste de voisins à une matrice de poids

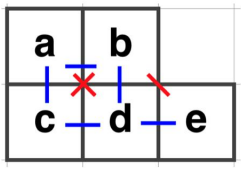
Une fois le graphe de voisinage défini et codifié sous forme d'une liste de voisins, on transforme la liaison entre les points i et j en l'élément w_{ij} de la matrice de poids $\mathbf{W}$. La matrice de poids $\mathbf{W}$

est "l'expression formelle de la dépendance spatiale entre observations" (ANSELIN et al. 1988).

Définition de la matrice de poids

— Le plus couramment, la matrice de poids est une matrice de contiguïté binaire (voir figure 2.12) :

$$w_{ij} = \begin{cases} 1 & si \quad i \quad et \quad j \quad sont \quad reliés \quad dans \quad l'espace \\ 0 & sinon. \end{cases} \tag{2.3}$$



Contiguïté ROOK

	a	b	c	d	e	Somme des poids des voisins
a	0	1	1	0	0	2
b	1	0	0	1	0	2
c	1	0	0	1	0	2
d	0	1	1	0	1	3
e	0	0	0	1	0	1

Contiguïté QUEEN

	a	b	c	d	e	Somme des poids des voisins
a	0	1	1	1	0	3
b	1	0	1	1	1	4
c	1	1	0	1	0	3
d	1	1	1	0	1	4
e	0	1	0	1	0	2

FIGURE 2.12 – Matrice de poids binaire

— Les matrices de poids peuvent également tenir compte de la distance entre les zones géographiques, les relations devenant plus faibles avec la distance : 1 si $d < d_0$, $\frac{1}{d^\alpha}$, ou $e^{-\alpha d}$ avec α un paramètre estimé ou défini *a priori*. Utiliser une distance maximale au-delà de laquelle $w_{ij} = 0$ permet de limiter le nombre d'éléments non nuls. Comme décrit en 2.1.2, lorsque les tailles des zones sont hétérogènes, on risque alors une grande variabilité du nombre de voisins.

— Enfin, certaines matrices tiennent compte de la force des relations entre les zones. Le poids peut par exemple être défini par $\frac{b_{ij}^\alpha}{d_{ij}^\beta}$ avec b_{ij} une mesure de la force des relations entre les zones i et j (qui n'est pas forcément symétrique), telle que le pourcentage de frontières communes, la population totale, la richesse et d_{ij} la distance entre les zones.

Certains travaux économétriques cherchent à endogénéiser les matrices de poids, mais elles sont considérées comme exogènes dans la plupart des applications d'économétrie spatiale (ANSELIN 2013). En général, les poids de voisinage ne doivent donc pas être des fonctions du phénomène qu'on cherche à expliquer.

L'objet "liste de poids" en R

La fonction `nb2listw` du package `spdep` permet de transformer un objet "liste de voisins" en un objet "liste de poids". Il est important de noter que l'objet "liste de poids", qui correspond à la matrice de poids décrite précédemment n'est pas une matrice $n \times n$ telle qu'on la représente théoriquement. Il s'agit d'une liste contenant le style de normalisation, puis pour chaque observation : son attribut, la liste des numéros d'observation de ses voisins, la liste des attributs de ses voisins et la liste des

poids de ses voisins. On parle souvent de *sparse matrix* ou *matrices creuses*.
Lorsqu'une zone n'a pas de voisins, l'option `zero.policy=TRUE` permet de générer tout de même une liste de poids, qui prend la valeur 'zéro' pour les observations sans voisins. (Si l'option est `FALSE`, un message d'erreur est généré).

Application avec R

```
#Matrice basée sur la contiguïté
#La fonction nb2listw convertit tout objet de type "nb" en une #liste de
    poids
arr75.lw <- nb2listw(arr75.nb)

#Matrice fondée sur la distance
#La fonction mat2listw convertit une matrice en une liste de poids
library(fields) #pour calculer la distance entre deux points
coords <- coordinates(arr75)
distance <- rdist(coords,coords)
diag(distance) <- 0
distance[distance >=100000] <- 0
#le poids décroit comme le carré de la distance, dans un rayon de 100km
dist <- 1.e12 %/% (distance*distance)
dist[dist >=1.e15] <- 0
dist.w <- mat2listw(dist,row.names=NULL)
```

Normalisation de la matrice de poids

La somme des poids des voisins d'une zone est appelée son *degré de liaison*. Si on ne normalise pas la matrice de poids (*schéma de codage "B"*), ce degré de liaison dépend du nombre de ses voisins, ce qui crée une hétérogénéité entre les zones. À la suite de TIEFELSDORF 1998, on peut distinguer quatre types de normalisation :

— Normalisation en ligne (*schéma de codage "W"*) : pour une zone, le poids accordé à chaque voisin est divisé par la somme des poids de ses voisins : $\sum_{j=1}^{n} w_{ij} = 1$. Cette standardisation facilite l'interprétation de la matrice de poids, puisque $\sum_{j=1}^{n} w_{ij}x_j$ représente la moyenne de la variable x sur tous les voisins de l'observation i. Chaque poids w_{ij} peut être interprété comme la fraction de l'influence spatiale subie par l'observation i imputable à j. En revanche, cette normalisation implique une certaine compétition entre les voisins : moins une zone a de voisins, plus ceux-ci ont un poids important. De plus, quand les poids sont inversement proportionnels à la distance entre les zones, normaliser en ligne rend difficile leur interprétation.

— Normalisation globale (*schéma de codage "C"*) : les poids sont standardisés de sorte que la somme de tous les poids soit égale au nombre total d'entités : tous les poids sont multipliés par $\frac{n}{\sum_{j=1}^{n}\sum_{i=1}^{n} w_{ij}}$.

— Normalisation uniforme (*schéma de codage "U"*) : les poids sont standardisés de sorte que la somme de tous les poids soit égale à 1 : $\sum_{j=1}^{n}\sum_{i=1}^{n} w_{ij} = 1$.

— Normalisation par stabilisation de la variance (*schéma de codage "S"*) : soit $\mathbf{q}$ le vecteur défini par : $\mathbf{q} = (\sqrt{\sum_{j=1}^{n} w_{1j}^2}, \sqrt{\sum_{j=1}^{n} w_{2j}^2}, ..., \sqrt{\sum_{j=1}^{n} w_{nj}^2})^T$.

Soit la matrice $\mathbf{S}^* = [diag(\mathbf{q})]^{-1}\mathbf{W}$.[4] À partir de $\mathbf{S}^*$, on calcule $Q = \sum_{j=1}^{n}\sum_{i=1}^{n} s_{ij}^*$ puis on en déduit la matrice de poids normalisée : $\mathbf{S} = \frac{n}{Q}\mathbf{S}^*$.

La normalisation par stabilisation de la variance a été introduite par Tiefelsdorf afin de réduire l'hétérogénéité dans les poids liée aux différences de taille et de nombre de voisins entre les zones. En effet, la normalisation en ligne donne plus de poids aux observations situées en bordure de la zone d'étude, avec un faible nombre de voisins. Au contraire, avec une normalisation globale ou uniforme, les observations situées au centre de la zone d'étude, avec un grand nombre de voisins, sont soumises à plus d'influences extérieures que les zones frontalières. Cette hétérogénéité peut avoir un impact significatif sur les résultats des tests d'autocorrélation spatiale.

Les poids de la matrice normalisée suivant le schéma "S" varient moins que ceux de la matrice normalisée suivant le schéma "W". La somme des poids des lignes varie plus pour le style "S" que pour le style "W", mais moins que pour les styles "B", "C" et "U" (BIVAND et al. 2013b).

Que le schéma de codage soit en ligne, global ou par stabilisation de la variance, la somme totale des éléments de la matrice vaut toujours n, ce qui permet aux statistiques d'autocorrélation spatiale utilisant la matrice d'être comparables entre elles.

Application avec R

```
#L'option style permet de définir le type de normalisation
arr75.lw <- nb2listw(arr75.nb,zero.policy=TRUE, style="W")
names(arr75.lw)
## [1] "style"      "neighbours" "weights"
summary(unlist(arr75.lw$weights))
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.1250  0.1667  0.1833  0.1961  0.2500  0.3333
```

2.2.2 Importance du choix de la matrice de poids

Lorsqu'on cherche à tester l'importance des relations économiques ou sociales entre certaines variables, la localisation géographique des observations est un paramètre clé. D'une part, les observations situées dans la même zone géographique sont soumises aux mêmes paramètres extérieurs (climat, pollution, etc.) ; d'autre part les observations voisines s'influencent mutuellement. Les modèles d'économétrie spatiale prennent en compte ces différentes interactions. Ces modèles utilisent la spécification du voisinage, par l'intermédiaire de la matrice de poids $\mathbf{W}$. Au sein de la communauté scientifique, les avis diffèrent sur l'influence de la définition de la matrice de poids sur les résultats.

BHATTACHARJEE et al. 2005 remarquent que : "Le choix des poids est souvent arbitraire [...] et le résultat des études varie considérablement en fonction de la définition des poids spatiaux". Une mauvaise spécification de $\mathbf{W}$ conduirait à de fausses conclusions. Ceci dit, puisque différentes méthodes de construction de la matrice de poids sont envisageables, "[...] il est possible que l'une des méthodes mène à des résultats pertinents, mais le risque d'une mauvaise spécification pèsera toujours sur le modèle choisi" (GETIS et al. 2004).

4. $diag(\mathbf{q})$ est une matrice diagonale avec les composantes de $\mathbf{q}$ sur sa diagonale principale

L'objectif est que les poids w_{ij} reflètent le plus fidèlement possible les interactions entre observations. Les hypothèses sous-jacentes peuvent s'appuyer sur des modèles économiques ou sociologiques. Par exemple, des poids nuls au-delà d'une certaine distance seront justifiés par le fait que l'influence d'une zone d'emploi sur son environnement est contrainte par la mobilité des individus, elle-même limitée par leur temps de trajet. HARRIS et al. 2011 soulignent cependant que le concept de 'distance' est lui-même flou. La distance est souvent définie par une distance géométrique entre deux points représentatifs des zones d'étude. Mais la distance peut également être un temps de transport entre deux régions (temps minimal, ou temps en empruntant la route la moins onéreuse), ou encore être proportionnelle aux échanges entre les zones. Pour HARRIS et al. 2011, "la conséquence de l'utilisation de mesures liées à la contiguïté ou à la distance pour pondérer les observations des régions voisines est qu'on impose une structure d'interaction spatiale dont on est incapable de tester la fiabilité, et qui est potentiellement mal spécifiée."

HARRIS et al. 2011 présentent quelques approches alternatives de construction de la matrice de poids. Ces méthodes ont pour objectif de diminuer au maximum les hypothèses *ad hoc* dans la spécification des matrices. Cependant aucune méthode n'arrive à s'en défaire totalement.

Tous les chercheurs ne sont pas aussi pessimistes : LESAGE et al. 2010 considèrent que la croyance selon laquelle la définition de la matrice de poids a une influence cruciale sur les résultats est due à des erreurs d'interprétation des coefficients des modèles d'économétrie spatiale, ou à des erreurs dans la spécification des modèles. Cette croyance serait, selon eux : "le plus gros mythe de l'économétrie spatiale". Ils soutiennent que si l'on s'intéresse à l'effet moyen des variables explicatives sur les variables dépendantes, les différences de spécification de la matrice de poids n'ont pas d'influence significative sur les résultats. LESAGE et al. 2010 reconnaissent cependant qu'il reste encore beaucoup à faire pour mieux caractériser la notion d'équivalence entre matrices.

Références - Chapitre 2

ANSELIN, Luc (2013). *Spatial econometrics : methods and models*. T. 4. Springer Science & Business Media.

ANSELIN, Luc et Daniel A GRIFFITH (1988). « Do spatial effects really matter in regression analysis ? » *Papers in Regional Science* 65.1, p. 11–34.

APPLEGATE, David et al. (2006). *Concorde TSP solver*.

BHATTACHARJEE, Arnab et Chris JENSEN-BUTLER (2005). « Estimation of spatial weights matrix in a spatial error model, with an application to diffusion in housing demand ». CRIEFF Discussion Papers.

BIVAND, Roger S, Edzer PEBESMA et Virgilio GOMEZ-RUBIO (2013b). « Spatial Neighbors ». *Applied Spatial Data Analysis with R*. Springer, p. 83–125.

GARFINKEL, R.S. (1985). « Motivation and modelling (chapter 2) ». *E. L. Lawler, J. K. Lenstra, A.H.G. Rinnooy Kan, D. B. Shmoys (eds.) The traveling salesman problem - A guided tour of combinatorial optimization,* Wiley & Sons.

GETIS, A et J ALDSTADT (2004). « On the specification of the spatial weights matrix ». *Geographical Analysis* 35.

HAHSLER, Michael et Kurt HORNIK (2017). *TSP : Traveling Salesperson Problem (TSP)*. R package version 1.1-5. URL : `https://CRAN.R-project.org/package=TSP`.

HARRIS, Richard, John MOFFAT et Victoria KRAVTSOVA (2011). « In search of 'W' ». *Spatial Economic Analysis* 6.3, p. 249–270.

KINCAID, Thomas M. et Anthony R. OLSEN (2016). *spsurvey : Spatial Survey Design and Analysis*. R package version 3.3.

LESAGE, James P et R Kelley PACE (2010). « The biggest myth in spatial econometrics ». *Available at SSRN 1725503*.

MARK, David M (1990). « Neighbor-based properties of some orderings of two-dimensional space ». *Geographical Analysis* 22.2, p. 145–157.

SMITH, Tony E. (2016). *Notebook on Spatial Data Analysis*. http ://www.seas.upenn.edu/ ese502/notebook.

STEVENS JR, Don L et Anthony R OLSEN (2004). « Spatially balanced sampling of natural resources ». *Journal of the American Statistical Association* 99.465, p. 262–278.

TIEFELSDORF, Michael (1998). « Modelling spatial processes : The identification and analysis of spatial relationships in regression residuals by means of Moran's I (Germany) ». Thèse de doct. Université Wilfrid Laurier.

TOUSSAINT, Godfried T (1980). « The relative neighbourhood graph of a finite planar set ». *Pattern recognition* 12.4, p. 261–268.

— (2014). « The sphere of influence graph : Theory and applications ». *International Journal of Information Theory and Computer Science* 14.2.