NAZIOARTEKO ESTATISTIKA MINTEGIA

SEMINARIO INTERNACIONAL DE ESTADISTICA

Analyse Statistique Des Données Spatiales

Christine Thomas-Agnan

Eustat
EUSKAL ESTATISTIKA ERAKUNDEA
INSTITUTO VASCO DE ESTADÍSTICA

54

Datu Espazialen Analisi Estatistikoa

Analyse Statistique des Données Spatiales

Análisis Estadísticos de Datos Espaciales

Christine Thomas-Agnan

Professeur de mathématiques, Université Toulouse 1

E-mail: christine.thomas@tse-fr.eu

AURKEZPENA

Nazioarteko Estatistika Mintegia antolatzean, hainbat helburu bete nahi ditu EUSTAT-Euskal Estatistika Erakundeak:

- Unibertsitatearekiko eta, batez ere, Estatistika-Sailekiko lankidetza bultzatzea.
- Funtzionarioen, irakasleen, ikasleen eta estatistikaren alorrean interesatuta egon daitezkeen guztien lanbide-hobekuntza erraztea.
- Estatistika alorrean mundu mailan abangoardian dauden irakasle eta ikertzaile ospetsuak Euskadira ekartzea, horrek eragin ona izango baitu, zuzeneko harremanei eta esperientziak ezagutzeari dagokienez.

Jarduera osagarri gisa, eta interesatuta egon litezkeen ahalik eta pertsona eta erakunde gehienetara iristearren, ikastaro horietako txostenak argitaratzea erabaki dugu, beti ere txostengilearen jatorrizko hizkuntza errespetatuz; horrela, gai horri buruzko ezagutza gure herrian zabaltzen laguntzeko.

Vitoria-Gasteiz, 2012ko Azaroa

JAVIER FORCADA SAINZ
EUSTATeko Zuzendari Nagusia

PRESENTATION

L´Institut Basque de Statistique se propose d´atteindre plusieurs objectifs par la promotion des Séminaires Internationaux de la Statistique:

- Encourager la collaboration avec l´Université et spécialement avec les départements de la Statistique.
- Faciliter le recyclage professionnel des fonctionnaires, professeurs, élèves, et tous ceux qui pourraient etre intéressés par la statistique.
- Inviter en Euskadi des professeurs mondialement renommés et des rechercheurs de premier ordre en matière de Statistique avec tout ce que cela pourrait entraîner commen avantage dans les rapports et l´échange d´expériences.

En outre, il a été décidé de publier les exposés de ces rencontres afin d´atteindre le plus grand nombre de personnes et d´institutions intéresées, et pour contribuer ainsi à développer dans notre pays les connaissances sur cette matière. Dans chaque cas la langue d´origine du conférencier sera respectée.

Vitoria-Gasteiz, Novembre 2012

JAVIER FORCADA SAINZ
Directeur General d´ EUSTAT

PRESENTACIÓN

Al promover los Seminarios Internacionales de Estadística, el EUSTAT-Instituto Vasco de Estadística- pretende cubrir varios objetivos:

- Fomentar la colaboración con la Universidad y en especial con los Departamentos de Estadística.
- Facilitar el reciclaje profesional de funcionarios, profesores, alumnos y cuantos puedan estar interesados en el campo estadístico.
- Traer a Euskadi a ilustres profesores e investigadores de vanguardia en materia estadística, a nivel mundial, con el consiguiente efecto positivo en cuanto a la relación directa y conocimiento de experiencias.

Como actuación complementaria y para llegar al mayor número posible de personas e Instituciones interesadas, se ha decidido publicar las ponencias de estos cursos, respetando en todo caso la lengua original del ponente, para contribuir así a acrecentar el conocimiento sobre esta materia en nuestro País.

Vitoria-Gasteiz, Noviembre 2012

JAVIER FORCADA SAINZ
Director General de EUSTAT

BIOGRAFI OHARRAK

CHRISTINE THOMAS-AGNAN TSEn (Toulouse School of Economics, Tolouse I unibertsitatekoa) irakasle eta ikertzailea da. Paris 7n matematikan lizentziaduna, matematikaren masterra (D.E.A) eskuratu zuen M. Marle-ren zuzendaritzapean (geometria sinplektika mekanika kuantikora aplikatua). CNRS-n (Centre National de la Recherche Scientifique –France) lanean aritua da.

Ikerketa-gai nagusiak: Ekonometria espaziala, prozesu puntual espazialak, kobariantzen funtzioen zenbatespen ez-parametrikoa, quantil eta espektil baldintzazkoak, ugalketa-guneen teoriaren aplikazioak estatistikari eta probabilitateei.
"GeoXp: An R Package for Exploratory Spatial Data Analysis" lanaren egilekide da.

NOTES BIOGRAPHIQUES

CHRISTINE THOMAS-AGNAN est professeur-chercheur à TSE (Toulouse Schoof of Economics) de l'Université de Toulouse I. Titulaire d'une Maîtrise de Mathématiques à Paris 7, Master de Mathématiques (D.E.A.) sous la direction de M. Marle (Géométrie symplectique appliquée à la mécanique quantique). Elle a travaillé au CNRS (Centre National de la Recherche Scientifique –France)

Ses principaux thèmes de recherche sont: Économétrie spatiale, Processus ponctuels spatiaux, Estimation nonparamétrique de fonctions de covariances, Quantiles et expectiles conditionnels, Applications à la théorie des noyaux reproduisants à la statistique et aux probabilités.
Elle est co-auteur de "GeoXp: An R Package for Exploratory Spatial Data Analysis"

NOTAS BIOGRÁFICAS

CHRISTINE THOMAS-AGNAN es profesora-investigadora en TSE (Toulouse Schoof of Economics de la Universidad Tolouse I). Licenciada en Matemáticas en Paris 7, Master de Matemáticas (D.E.A.) bajo la dirección de M. Marle (Geometría simpléctica aplicada a la mecánica cuántica). Ha trabajado en la CNRS (Centre National de la Recherche Scientifique –France)

Sus temas principales de investigación son: Econometría espacial, Procesos puntuales espaciales, Estimación no paramétrica de funciones de covarianzas, Quantiles y expectiles condicionales, Aplicaciones de la teoría de núcleos reproductivos a la estadística y las probabilidades.
Es coautora de "GeoXp: An R Package for Exploratory Spatial Data Analysis"

Table des matières

Chapitre 1

Introduction : nécessité de la prise en compte de la dimension spatiale

Les données spatiales ou géoréférencées sont des données comportant une dimension spatiale, c'est à dire pour lesquelles une information géographique est attachée à chaque unité statistique. L'information géographique est en général la position de l'unité sur une carte ou dans un référentiel spatio-temporel et peut par exemple prendre la forme de latitude et longitude ou de coordonnées UTM. Un traitement statistique de telles données qui ignore cet aspect ou l'intègre de façon inadéquate peut resulter en une perte d'information, des erreurs de spécifications, des estimations non convergentes et non efficaces. En effet, il ne suffit pas de juxtaposer l'analyse géographique à l'analyse statistique, il faut les faire interagir. Les systèmes d'information géographiques sont des outils sophistiqués qui permettent de faire de la cartographie professionnelle mais ils n'intègrent en général que des méthodes statistiques élémentaires (histogrammes, camemberts). Les outils propres à la statistique spatiale que nous allons exposer font intervenir la position spatiale à part entière dans leur définition.

Les domaines scientifiques privilégiés d'application de ces méthodes sont l'économie, la géographie, la sociologie, l'épidémiologie, la géologie, la météorologie. On trouve également des applications dans le secteur industriel avec l'industrie pétrolière et dans le tertiaire avec le géomarketing. Voici quelques exemples. En prospection pétrolière, il est utile de prédire la quantité de pétrole potentielle en un lieu donné en fonction de prélèvements effectués en certains points répartis sur une zone pour optimiser l'emplacement des forages. En économie urbaine, l'ajustement de modèles hédoniques qui expliquent le prix d'une transaction en fonction des caractéristiques du bien immobilier mais aussi des caractéristiques socio-économiques ou autres de

leur lieu d'implantation permet de mieux comprendre ce qui influence le marché immobilier. En aménagement du territoire, on peut vouloir étudier la répartition spatiale des établissements scolaires et chercher à augmenter l'efficacité du système scolaire en choisissant au mieux le lieu d'implantation de nouveaux établissements. En ce qui concerne l'environnement, la production de cartes de prédiction de niveaux de pollution utilise les outils de la géostatistique.

La distinction entre statistique spatiale et économétrie spatiale provient du fait que traditionnellement, les techniques de statistique spatiale se sont développées d'abord en géostatistique (au départ la statistique pour géologues) et concernent des données de nature différente de celles étudiées en économie comme nous le verrons en détail dans le chapitre suivant. Néanmoins on fait souvent référence à la statistique spatiale pour désigner l'ensemble de ces méthodes. Du point de vue historique, la géostatistique est née de l'industrie minière. L'ingénieur africain D.G. Krige s'est rendu célébre pour les estimations de gisements d'or (1951) et a donné son nom à la méthode de Krigeage. Ce domaine a ensuite été développé par l'école française de Fontainebleau avec G. Matheron et ses collaborateurs.
Quelques manuels de référence dans ce domaine sont : N. Cressie (1993), J. LeSage et K. Pace (2009) pour l'économétrie spatiale, Diggle (2003) et Illian et al. (2009) pour les semis de points et R. Bivand et al. (2008) pour l'implémentation en R.
L'outil de modélisation des données géoréférencées est le champ aléatoire. Lorsqu'une caractéristique $X(s, \omega)$ d'une unité statistique est mesurée en la position s pour la réalisation ω, on notera X_s la variable aléatoire associée, où l'indice s varie dans une partie $\mathcal{D}$ de $\mathbb{R}^d$. La dimension d varie de 1 à 3 dans les applications courantes.

1.1 Statistique spatiale et séries temporelles

Lorsque le champ aléatoire est indexé par un espace de dimension $d = 1$ on utilise, plutôt que le terme de champ, le terme de processus ou de série temporelle (le cas le plus fréquent étant celui où la variable aléatoire est indexée par le temps). L'étude des séries temporelles est un domaine en soi de la statistique et il est clair qu'il ne s'agit en aucun cas d'un cas particulier de la statistique spatiale (les ressemblances et différences seront signalées dans le texte). Certains des outils exposés dans ce manuel concernent le cas général de la dimension d supérieure à 1 mais c'est quand même le cas $d = 2$ qui reste l'objectif principal.
Le parallèle avec les séries temporelles est cependant intéressant. En effet, ce qui distingue les séries temporelles d'autres modèles statistiques est la prise en compte de la dépendance entre l'observation faite en un temps t et celle

faite en des temps voisins. Beaucoup de modèles supposent l'indépendance entre les observations (mathématiquement parlant entre les variables aléatoires associées) faites sur les diverses unités statistiques. Dans le cas d'observations temporelles, cette indépendance n'est pas une hypothèse réaliste car, dans beaucoup de phénomènes, ce qui se passe aujourd'hui est nécéssairement influencé par ce qui s'est passé hier et dans une moindre mesure par un passé lointain. Par ailleurs, dans le cas des séries temporelles, l'hypothèse de répartitions marginales identiques est aussi remise en question dans la mesure où le phénomène peut présenter une évolution en moyenne résultant en une non stationarité. De la même façon, les champs aléatoires spatiaux peuvent présenter à la fois

- une autocorrélation spatiale : les variables X_s et X_t étant d'autant plus corrélées que la distance entre s et t est petite.
- une hétérogénéité spatiale : la répartition marginale de X_s varie avec s.

Mais à la différence des séries temporelles, les notions de passé et de futur n'ont pas leur pendant en spatial et il n'y a pas d'ordre naturel dans $\mathbb{R}^d$.

1.2 Bénéfices de la prise en compte de la dimension spatiale

Quels sont les avantages d'une modélisation adaptée aux données spatiales ? Que perd-on si on ne la fait pas ? Dans le contexte d'un modèle de régression, on verra plus loin dans le document que si le processus de génération des données suit un modèle spatial alors que le statisticien utilise un modèle ordinaire en faisant abstraction des effets spatiaux, cela peut résulter, selon le type de processus spatial concerné, en la présence de biais dans les coefficients de la régression, d'une absence de convergence car ce biais n'est pas nécéssairement asymptotiquement nul, d'une inefficacité des estimations. L'absence d'impacts indirects (effet du changement d'une variable en un lieu donné sur les autres lieux) dans le modèle ordinaire peut masquer de réels effets de débordement. De plus cela entraîne aussi d'importants biais de prédiction.

Pour illustrer le biais d'estimation des coefficients, nous prenons ici comme exemple le découpage administratif de la région Midi-Pyrénées en 283 pseudo-cantons. On considèrera qu'une unité spatiale est voisine d'une autre si les unités spatiales partagent une frontière commune. On observe sur ces 283 unités une variable X simulée selon une loi $\mathcal{N}(\mu = 40, \sigma = 10)$. Considérons le modèle LAG qui sera présenté dans le chapitre 7

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

où ϵ est un bruit blanc spatial et WY désigne le vecteur des moyennes de

la variable Y dans le voisinage de chaque unité spatiale. La figure suivante représente une exemple de simulation de Y à partir de :

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

en prenant $\rho = 0.95$, $\beta = 50$ et ϵ simulée selon une loi $\mathcal{N}(\mu = 40, \sigma = 10)$. Nous avons également représenté les liens de voisinage entre les unités spatiales.

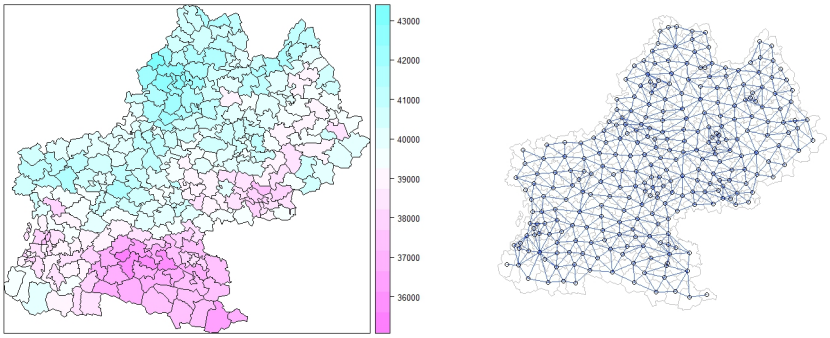


FIGURE 1.1 – A gauche la variable Y simulée avec $\rho = 0.95$. A droite les liens de voisinage entre unités spatiales.

Pour différentes valeurs de ρ, nous avons calculé le biais d'estimation du coefficient β donné, comme on le verra dans le chapitre 7, par :

$$(X'X)^{-1}X'(I - \rho W)^{-1}X - 1,$$

et la différence entre la variance de $\hat{\beta}$ estimée dans le modèle LAG et dans le modèle linéaire non spatial ordinaire donnée par

$$(X'X)^{-1}X'((I - \rho W)'(I - \rho W))^{-1}X - 1,$$

Ces quantités sont représentées respectivement en bleue et en rouge dans le graphique suivant en fonction de ρ.
Enfin, pour juger de l'hétéroscédasticité présente dans ce modèle spatial, nous avons représenté pour différentes valeurs de ρ, la distribution des éléments de la partie triangulaire supérieure de la matrice de variance de Y dans ce modèle, donné à facteur d'échelle près par :

$$((I - \rho W)'(I - \rho W))^{-1}$$

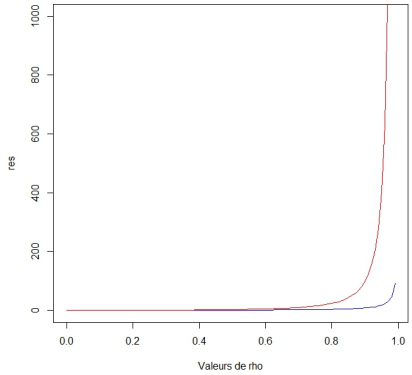


FIGURE 1.2 – Biais dans le modèle LAG en fonction de ρ. Différence entre variances estimées dans les modèles LAG et OLS

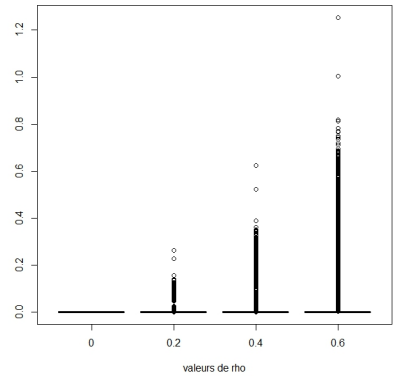


FIGURE 1.3 – Hétéroscédasticité

1.3 Rudiments de cartographie

Les données spatiales attachées à une position sur notre globe terrestre sont généralement représentées dans un plan. Avant de réaliser leur étude statistique, il est nécéssaire de les importer dans un logiciel d'analyse statistique. Pour importer et analyser une telle base de données, il y a cependant un minimum de connaissance cartographique à avoir. Tout d'abord, la surface réelle de la terre dite géoïde est de forme patatoïde ; on l'approxime par un ellipsoïde et il y a plusieurs approximations possibles (par exemple ellipsoïde de Clarke). Pour dessiner une carte il faut un système de coordonnées : des axes et une origine. De plus, comme la terre n'est pas plate, il faut choisir un système de projection cartographique. Cette projection est une correspondance entre les coordonnées planimétriques X et Y d'un point, mesurées sur une grille régulière, et sa latitude ϕ et longitude λ. La latitude est une mesure de l'angle ϕ par rapport à l'équateur, la longitude est une mesure de l'angle λ par rapport au méridien de référence. Il existe différentes unités pour mesurer ces angles : degrés-minutes-secondes, degrès-décimaux, radians, grades. Au besoin, l'altitude du point est mesurée au dessus du géoïde ou du niveau local zéro des mers. La projection est la méthode de réduction de la distorsion due à la rotondité de la terre appliquée sur une surface plate. On distingue plusieurs sortes de projections : conique, cylindrique, azimutale. Les projections les plus courantes sont : la projection de Mercator, la projection Lambert et la projection de Mercator Universelle. Un datum géodésique est la donnée d'un ellipsoïde et d'un système de projection : citons par exemple pour l'Europe le datum ED50, système européen unifié. Des logiciels de conversion permettent de passer d'un système à l'autre.

Pour illustrer les difficultés rencontrées, prenons un exemple. Dans le système WGS 84 (*World Geodetic System* 1984, système géodésique mondial, révision de 1984), les coordonnées longitude/latitude de la ville de Vitoria Gasteiz sont (2°41'0"W, 42°51'00"N) en degrés-minutes-secondes et (-2.683333°, 42.85°) en degrés-décimaux. Le CRS (système de coordonnées de référence) précise le système de projection (`proj=`) ainsi que l'ellipsoïde considéré (`ellps=`). Dans notre exemple, le CRS s'écrit :

`CRS("+proj=longlat ellps=WGS84")`.

Il est absolument essentiel de connaître le CRS d'un fichier de données spatiales. En effet, lorsqu'on travaille avec plusieurs sources de données, il est rare que les unités spatiales soient exprimées dans un même CRS. Par exemple, dans la Figure (1.3), les contours de la province d'Alava sont exprimés dans le `CRS("+proj=lcc +ellps=WGS84")` (projection *Lambert Conformal Conic*) alors que les coordonnées de la ville de Vitoria Gasteiz sont exprimées dans le `CRS("+proj=merc ellps=WGS84")` (projection *Mercator*).

Les SIG (Système d'Information Géographique) sont souvent munis d'outils

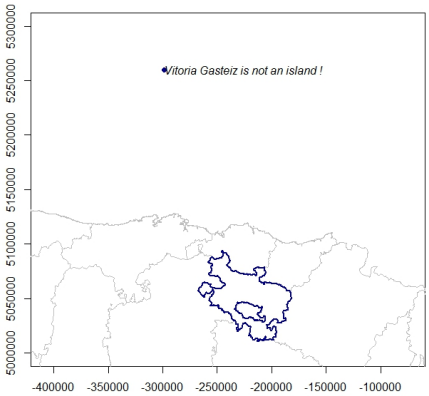


FIGURE 1.4 – Exemple de problème rencontré lorsque les CRS ne correspondent pas.

permettant la conversion d'unités spatiales d'un CRS vers un autre CRS. Dans la Figure (1.3), nous avons représenté la province d'Alava dans deux CRS différents :

`CRS("+proj=longlat ellps=WGS84")` et

`CRS("+proj=utm ellps=WGS84")` (projection *Universal Transverse Mercator*).

L'avantage avec la projection UTM est que les coordonnées sont exprimées en mètres et le calcul de distance entre deux points est ainsi facilité. Le package **rgdal** dans le logiciel R permet d'effectuer ces transformations.

1.4 Exemple de lecture d'un jeu de données spatiales avec R

Les fichiers de données spatiales sont :

- de type **vectoriel**, comme les fichiers Shapefile (avec une extension .shp) ou MapInfo (extension .MIF, .MID). L'unité spatiale de référence peut être assimilée à un point, un polygone ou un vecteur. Les unités spatiales peuvent être associées à des attributs. Dans le cas de données statistiques, ces attributs seront des variables (quantitatives ou qualitatives) observées sur les unités spatiales.
- de type **raster**, comme les fichiers au formats .bmp, .jpeg, .tiff, .asc. Dans ce cas, l'unité spatiale de référence est le pixel. En parlant de pixel, on pourra également utiliser le terme de cellule ou carreau. Une cellule peut être associée à une (ou plusieurs) valeur(s) visualisée par

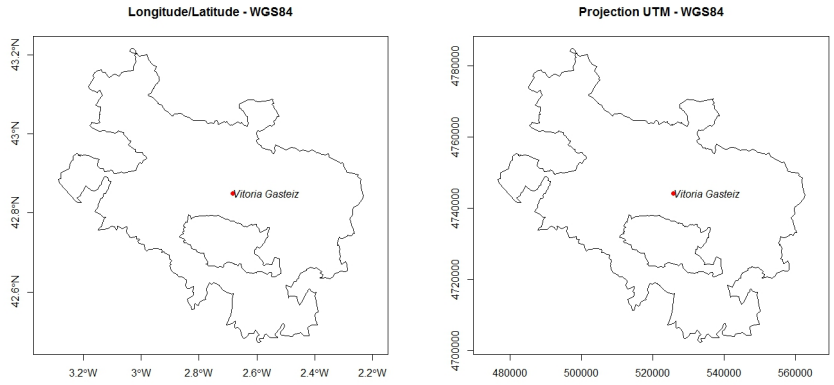


FIGURE 1.5 – Représentation de la région de Vitoria dans deux CRS différents.

la couleur.

a. Format vectoriel

Format ESRI shapefile

C'est le format de référence d'import/export pour des données géographiques (ESRI=*Environmental Systems Research Institute*). Un ESRI shapefile est formé de :

- un fichier principal (.shp) qui contient toute l'information liée à la géométrie des objets décrits qui peuvent être : des points, des lignes ou des polygones ;
- un fichier (.shx) qui stocke l'index de la géométrie ;
- un fichier dBASE (.dbf) pour les données attributaires (ou données statistiques) ;
- des fichiers facultatifs comme un fichier sur les datums/projections (.prj).

Dans le code R ci-dessous, le chargement de la librairie **sp** permet d'utiliser les classes d'objet `Spatial` et la librairie **maptools** permet l'importation de fichiers spatiaux. La commande `CRS("+init=epsg:4326")` de la fonction `readShapeSpatial()` permet d'indiquer le système géodésique utilisé. Il s'agit ici d'une simplification du `CRS("+proj=longlat +ellps=WGS84)` que nous avons vu précédemment. Le fichier correspond au découpage administratif des régions européennes utilisé dans le cadre de la législation européenne et représentées ici par le centroïde de la région.

```
> library("sp")
> library("maptools")
```

```
> xx <- readShapeSpatial("NUTS_LB_2010.shp", CRS("+init=epsg:4326"))
> class(xx)
[1] "SpatialPointsDataFrame"
attr(,"package")
[1] "sp"
```

`xx` est un objet de classe `SpatialPointsDataFrame`. Cela signifie que :

- les unités spatiales sont des points appartenant à la classe d'object `SpatialPoints`.
- chaque point est associé à des attributs inclus dans un objet `data.frame`.

Pour représenter l'objet `xx` :

```
> plot(xx,axes=TRUE)
> title("NUTS-2010 region centroids")
```

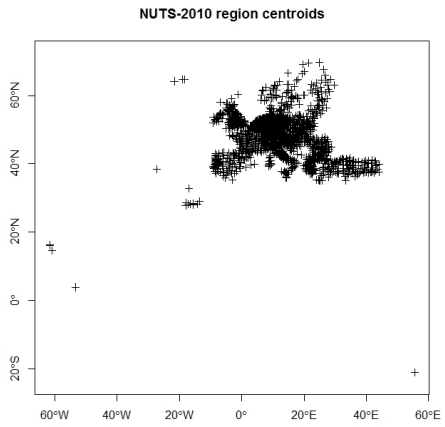


FIGURE 1.6 – Représentation des centroïdes des zones NUTS (*Nomenclature of territorial units for statistics*).

Pour savoir le nombre d'unités spatiales et le nombre de variables observées :

```
> dim(xx)
[1] 1921    4
```

Pour afficher les attributs des premières unités spatiales :

```
> head(xx@data)
  NUTS_ID      LAT      LON STAT_LEVL_
0   EL111 41.11184 26.11046          3
1   EL112 41.16937 24.82273          3
2   EL113 41.10678 25.50045          3
```

```
3   EL114 41.28699 24.18505          3
4   EL115 40.82641 24.33590          3
5   EL121 40.49527 22.26628          3
> plot(xx,axes=TRUE)
> title("NUTS-2010 region centroids")
```

Format MapInfo

Pour importer un fichier MapInfo, on utilisera la fonction `readOGR()` du package **rgdal** (qui peut aussi être utilisé pour importer un fichier Shapefile) de la façon suivante :

```
> xy <- readOGR("departements_region.mif","departements_region")
OGR data source with driver: MapInfo File
Source: "departements_region.mif", layer: "departements_region"
with 98 features and 7 fields
Feature type: wkbPolygon with 2 dimensions

> class(xy)
[1] "SpatialPolygonsDataFrame"
attr(,"package")
[1] "sp"
```

Ici la classe `SpatialPolygonsDataFrame` indique qu'une unité spatiale est représentée par un polygone. Il s'agit ici du découpage administratif de la France en départements. Parmi les attributs disponibles pour cet objet, on dispose de la population française en nombre d'habitants. Il est possible de représenter ainsi une carte choroplèthe en utilisant le code suivant :

```
> plotclr <- c("#EFF3FF", "#BDD7E7", "#6BAED6", "#3182BD", "#08519C")

> breaks<-quantile(xy@data$PSDC,c(0,0.2,0.4,0.6,0.8,1))

> plot(xy,col=plotclr[findInterval(xy@data$PSDC, breaks,
all.inside=TRUE)], border='grey')

> legend("topleft", legend = c("[29972,230296.0[","[29972,351983.8[",
[351983.8,554093.4[", "[554093.4,966320.0[","[966320.0,2554449.0]"),
title = "Nombre d'habitants",fill=plotclr,cex=0.7)
```

b. Format raster

Nous allons ici importer un fichier .asc à l'aide de la fonction `readAsciiGrid()` du package **maptools**.

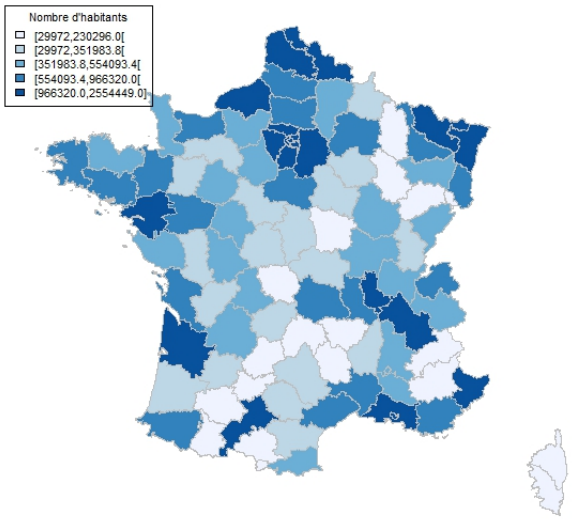


FIGURE 1.7 – Carte choroplèthe de la taille de la population dans les départements français.

```
> gr <- readAsciiGrid("pvgis_g13year00.asc")

> proj4string(gr)=CRS("+proj=longlat +ellps=WGS84")

> class(gr)
[1] "SpatialGridDataFrame"
attr(,"package")
[1] "sp"
```

L'avantage d'un objet de classe `SpatialGridDataFrame` est que sa structure n'est pas complexe. En effet, il suffit de connaître le nombre de cellules ($n_{row} \times n_{col}$), la taille d'une cellule (exprimées dans un CRS donné) et enfin les coordonnées de la cellule de référence (ou l'origine). Les valeurs associées aux cellules peuvent ensuite être stockées dans un vecteur de taille $n_{row} \times n_{col}$ sachant que le premier élément du vecteur correspond à la valeur observée à l'origine. Ici, l'image représente le temps d'ensoleillement annuel moyen observé en Europe. Pour représenter l'image :

```
> spplot(gr,axes=TRUE)
```

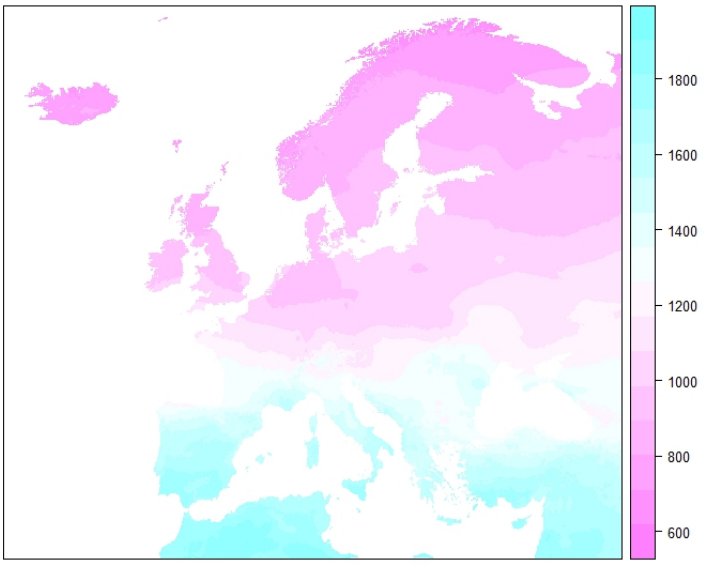


FIGURE 1.8 – Représentation d'un fichier de type raster représentant la durée d'ensoleillement annuel moyen.

1.5 Etapes d'une analyse spatiale

Comme dans toutes les études statistiques, l'analyse d'un jeu de données spatial commence par une étude exploratoire. Le but de cette étude, en sus des objectifs classiques tels que repérer les valeurs manquantes et atypiques, établir un premier résumé unidimensionnel de chaque variable, est ici d'explorer l'existence de tendances et d'autocorrélations spatiales. Si celles-ci sont mises en évidence, le reste de l'analyse s'attachera à corriger de l'hétérogénéité d'une part et de l'autocorrélation de l'autre de façon que l'analyse des impacts des facteurs explicatifs ainsi que les prédictions éventuelles soient les plus efficaces possible.

Chapitre 2

Divers types de données spatiales

On distingue trois grands types de données géoréférencées : les données ponctuelles ou de type géostatistique, les données surfaciques ou de type économétrie spatiale et les données de type semis de points. Ils différent d'abord par la nature de l'unité géographique attachée à chaque unité statistique, soit un lieu précis soit une surface, mais aussi par la qualité aléatoire ou non des positions spatiales. Avant de décrire ces types plus précisément, notons qu'il existe d'autres types moins répandus comme par exemple les données bilocalisées ou données de flux où chaque caractéritique se rapporte à un couple de sites. Notons qu'il existe aussi des données spatiales de type image pour lesquelles une ou plusieurs caractéristiques sont attachées à des pixels. Celles-ci peuvent justifier de traitements adaptés aux deux premiers types ci-dessous mais également à des traitements spécifiques au traitement d'image que nous n'aborderons pas dans ce document.

2.1 Données de type géostatistique ou ponctuelles

Les données de type géostatistique sont tout d'abord telles que la position observée n'est pas modélisée comme aléatoire car elle est choisie par le statisticien. Par exemple, un jeu de données météorologiques va être observé sur une collection de stations météo, des données de pollution atmosphérique sur une collection de lieux où l'on a implanté des appareils de mesure. Par ailleurs, l'unité géographique associée à la donnée est ici ponctuelle : on peut repérer la latitude et longitude des stations météo ou des appareils de mesure. Plus formellement, pour le champ aléatoire servant à modéliser notre phénomène, l'espace des indices sera un domaine $\mathcal{D}$ de $\mathbb{R}^d$ contenant un rectangle de volume strictement positif et l'indice s varie donc continument dans cet espace. Par contre, dans la pratique, les observations du champ sont faites en un nombre fini de points **déterministes** s_i de $\mathcal{D}$. Ceux-ci peuvent

dans certains modèles constituer une grille régulière mais ce n'est pas le cas en général.

2.2 Données de type économétrie spatiale ou surfaciques

De même que pour les données de type géostatistique, pour les données de type économétrie spatiale ou surfaciques, la position observée n'est pas modélisée comme aléatoire. Par contre, l'unité géographique associée à la donnée est ici de nature surfacique. Le territoire observé est partitionné en zones sur lesquelles le phénomène est observé. C'est le cas pour la majeure partie des données économiques qui sont mesurées sur des découpages administratifs du territoire comme par exemple le taux de chomage ou le revenu moyen par foyer fiscal d'une commune ou d'un département. L'indice s du champ aléatoire varie alors dans un nombre fini de localisations qui sont généralement les centroides des zones ou leurs représentants administratifs.

2.3 Données de type semis de points

Dans ce dernier cas, la localisation de la donnée est modélisée comme **aléatoire** car elle n'est pas choisie par le statisticien mais par le phénomène. Par exemple, supposons que l'on observe l'évolution d'une forêt et que l'on enregistre la localisation des arbres. Nous sommes alors en présence d'un semis de points et il y a une variable aléatoire bidimensionnelle pour chaque observation qui est la localisation de l'arbre exprimée par ses coordonnées dans un repère. Supposons que de plus on enregistre aussi le diamètre et le nombre de leurs feuilles de chaque arbre. On a alors un processus ponctuel marqué : il y a trois variables aléatoires pour chaque observation qui sont la localisation d'une part et le diamètre et le nombre de feuilles d'autre part. Ces deux dernières sont les marques aléatoires associées à cette localisation. On utilise la théorie des processus ponctuels pour modéliser les répartitions aléatoires de points. Ces points sont généralement inclus dans $\mathbb{R}^d$ avec d entier ≥ 1 mais nous considérerons plus simplement le cas le plus courant où $d = 2$. Les domaines classiques d'application de ces modèles sont la géologie, l'écologie, l'étude des forêts. Donnons quelques exemples : la disposition de certaines espèces végétales dans une forêt, les emplacements des épicentres de secousses sismiques enregistrées, la localisation de trésors archéologiques retrouvés sur un site, les adresses de patients affectés d'une certaine maladie dans une région, la répartition de cellules dans un tissu biologique, ...

Chapitre 3

Spécificité des données spatiales : hétérogénéité et autocorrélation

3.1 Considérations de modélisation

Que ce soit pour des données ponctuelles ou pour des données surfaciques, nous allons utiliser une même notation pour simplifier. On parlerons d'un champ X_s observé en des localisations $s_1, \cdots, s_n$. Lorsque les données sont ponctuelles, X_s désignera la variable aléatoire de la caractéristique au point s et lorsque les données sont surfaciques, X_s désignera la variable aléatoire de la caractéristique dans l'unité spatiale dont le représentant est s. Lorsqu'on utilise un modèle mathématique de champ aléatoire pour un phénomène observé spatialement, la loi du champ X_s est caractérisée par

- les lois marginales de X_s pour chaque localisation s
- les lois conjointes de vecteurs $X_{s_1}, \cdots, X_{s_n}$ pour un ensemble fini de localisations $s_1, \cdots, s_n$

On imagine donc que, pour un lieu s donné, il existe un univers de réalisations possibles de la caractéristique X_s mais dans la réalité on observe généralement une seule réalisation de X_s et ce pour un nombre fini de sites s. Par exemple si la donnée est un ensemble de niveaux de pluie mesurés en des stations météo à un instant donné, pour chaque station s, on dispose d'une réalisation de la variable "volume de pluie en s". On a une pluralité de données due à une pluralité de lieux mais non à une pluralité de réalisations sauf si on est dans le cas d'observations répétées. Dans ce dernier cas, il s'agit en général d'observations répétées au cours du temps. Cette dimension temporelle bien sûr pourrait induire la nécéssité d'utiliser un champ spatio-temporel et c'est un domaine de recherche très actif de nos jours, mais nous avons choisi de ne pas le développer ici. Par contre, pour un lieu donné s, si l'on est prêt à considérer une homogénéité temporelle du phénomène ainsi qu'une absence

de dépendance temporelle, on peut considérer que l'on dispose de plusieurs observations i.i.d. d'une même variable aléatoire X_s. Le fait que l'on dispose le plus souvent que d'une seule observation (coupe transversale) pourrait décourager le statisticien débutant de faire la moindre inférence. La solution est de puiser des forces dans la continuité spatiale du phénomène et dans la corrélation entre lieux voisins pour rendre cette inférence possible.
On suppose que le champ X_s admet un moment d'ordre un fini : $\mathbb{E}(X_s) < \infty$. On décompose alors le champ aléatoire en deux parties de la façon suivante

$$X_s = \mathbb{E}(X_s) + (X_s - \mathbb{E}(X_s))$$

Le terme déterministe $\mathbb{E}(X_s)$ s'appelle la tendance et modélise les variations à grande échelle du phénomène décrit par ce champ. Il représente la valeur moyenne du champ (valeur théorique que l'on pourrait estimer par exemple si l'on disposait de plusieurs réalisations temporelles de X_s). Le terme aléatoire $(X_s - \mathbb{E}(X_s))$ s'appelle la fluctuation et modélise les variations du champ à petite échelle. Notons que la fluctuation a une moyenne nulle par construction. Une décomposition similaire existe en séries temporelles. Dans la pratique cependant, cette décomposition en deux termes pour un phénomène observé n'est bien sûr pas unique, et c'est le choix du modélisateur d'affecter certains aspects à la partie aléatoire ou à la partie déterministe. Une coupe transversale ne permet pas de différencier, en présence d'un agrégat de résidus élevés, entre une hétérogénéité avec une forte tendance dans le voisinage de l'agrégat, et une autocorrélation spatiale positive. Pour comprendre ce découpage, il est bon de penser à une montagne : le détail de la variation de l'élévation mesuré avec précision constitue le champ ; on peut penser à l'allure de la montagne vue d'avion telle qu'elle se découpe sur l'horizon comme à une tendance ; la différence entre l'élévation précise et cette tendance représente alors les accidents de terrain visibles de près.
Dans le cas des modèles de régression où il y aura à la fois une variable dépendante Y_s et des variables explicatives X_s, nous raisonnerons conditionnellement à l'observation des variables explicatives.

3.2 Hétérogénéité spatiale

L'hétérogénéité des données spatiales se traduit par le fait que la répartition marginale du champ aléatoire X_s varie avec la localisation s. On dit qu'**il y a une tendance** lorsque $\mathbb{E}(X_s)$ est non constante dans l'espace : on dit aussi que la moyenne est non stationnaire. Si l'on mesure par exemple la quantité de précipitations sur des stations météo, il parait naturel de penser que le nombre moyen de centimètres cubes de pluie par semaine à Toulouse est différent de celui observé à Brest. Il s'agit là d'une différence sur la moyenne, mais on peut aussi imaginer qu'il y a une plus forte variabilité

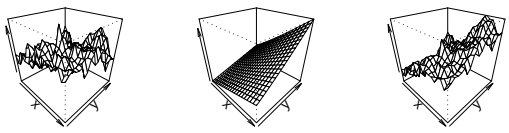


FIGURE 3.1 – Tendance et Fluctuation.

d'une semaine à l'autre à Brest qu'à Toulouse ou que les valeurs extrêmes sont très différentes. L'ensemble de la distribution des centimètres cubes de pluie par semaine a des raisons d'être spécifique du lieu.
L'hétérogénéité spatiale sera prise en compte par l'usage de variables explicatives pour modéliser la tendance. Certaines de ces variables peuvent être spatiales de nature comme, par exemple, la distance à certains lieux d'intérêt pour le problème. Mais notons cependant qu'il n'est pas suffisant de prendre en compte ces variables dans la moyenne pour évacuer totalement la structure spatiale du problème qui peut rester présente à l'ordre deux.

3.3 Autocorrélation spatiale

Une citation célèbre de Tobler (1979) est *Everything is related to everything else but closer things more so.*
Si la tendance est spécifique au moment d'ordre un d'un champ, l'autocorrélation concerne le moment d'ordre deux que l'on supposera exister dans ce paragraphe : on dit alors que le champ est du **second ordre**.
Pour les données spatiales, une corrélation peut se produire entre X_s et X_t du fait de leur proximité géographique. De façon qualitative, on parle d'autocorrélation spatiale positive pour une variable lorsqu'il y a regroupement géographique de valeurs similaires de la variable. De même, on parle d'autocorrélation spatiale négative pour une variable lorsqu'il y a regroupement géographique de valeurs dissemblables de la variable. Enfin, on parle

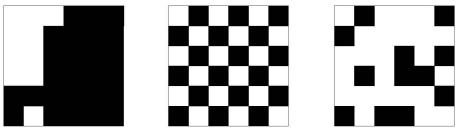


FIGURE 3.2 – Types d'autocorrélation.

d'absence d'autocorrélation pour une variable lorsqu'il n'y a pas de relation entre la proximité géographique et le degré de ressemblance des valeurs de la variable. Prenons pour illustrer cette notion l'exemple d'un champ dichotomique à valeurs 0 ou 1 représentées respectivement par les couleurs blanche et noire et constant sur les carrés d'une grille régulière. Comme on le voit sur la Figure 3.2, si le champ ne présente pas d'autocorrélation (à droite), une représentation graphique du champ montre des carrés blancs et noirs répartis au hasard. Si le champ présente une autocorrélation spatiale positive (à gauche), on verra des amas de carrés blancs et des amas de carrés noirs. Si le champ présente une autocorrélation spatiale négative (au centre), les carrés blancs auront souvent des voisins noirs et inversement.
Cette notion présentée ci-dessus de manière intuitive va se traduire par des propriétés du champ aléatoire portant sur l'ordre 2, c'est à dire sur la structure de covariance. La structure de covariance d'un champ du second ordre est définie par la fonction d'autocovariance

$$R(s,t) = \mathbb{C}ov(X_s, X_t).$$

Pour modéliser un tel champ, une des hypothèses simplificatrices que l'on est souvent amené à faire sur sa structure de covariance est celle de la stationnarité. La **stationnarité stricte** ou forte d'un champ suppose que la loi du vecteur $X_{s_1}, \ldots, X_{s_k}$ est invariante par translation quel que soit le nombre de points k et quelles que soient leurs positions $s_1, \ldots s_k$ i.e. $X_{s_1}, \ldots, X_{s_k}$ a même loi que $X_{s_1+h}, \ldots, X_{s_k+h}$ quel que soit $h \in \mathbb{R}^d$.
Une notion plus faible porte sur les deux premiers moments du champ. Un champ aléatoire X_s à valeurs réelles du second ordre est dit **stationnaire au second ordre** ou **au sens faible** s'il existe un vecteur $\mu \in \mathbb{R}$ et une fonction $R : \mathbb{R}^d \mapsto \mathbb{R}$ dite fonction d'autocovariance tels que

$$\mathbb{E}(X_s) = \mu \tag{3.1}$$
$$\mathbb{C}ov(X_s, X_{s+h}) = R(h) \tag{3.2}$$

Notons que, dans ce cas, la fonction d'autocovariance est une fonction d'une variable au lieu de deux. Il est clair que la stationnarité forte implique la stationnarité faible. Dans le cas gaussien, ces deux notions sont équivalentes puisque les moments d'ordre un et deux déterminent la distribution. Par la

suite le terme de stationnarité (sans précision) sera synonyme de stationnarité faible.
Les fonctions d'autocovariance peuvent être caractérisées par la propriété mathématique suivante. Une fonction $R(s,t)$ de $\mathbb{R}^2$ à valeurs dans $\mathbb{R}$ est une fonction d'autocovariance d'un champ aléatoire réel du second ordre si et seulement si elle est de type positif c'est à dire que quels que soit l'entier k, quels que soient les k sites $s_1, \ldots, s_k$ et les réels $a_1, \ldots, a_k$, on a

$$\sum_{i=1}^{k} \sum_{j=1}^{k} a_i a_j R(s_i, s_j) \geq 0.$$

Une fonction $R(s)$ de $\mathbb{R}$ à valeurs dans $\mathbb{R}$ est une fonction d'autocovariance d'un champ aléatoire réel stationnaire du second ordre si et seulement si elle est de type positif ce qui signifie dans ce cas que la fonction de deux variables $(s,t) \mapsto R(s-t)$ est de type positif. Notons que le vocabulaire "de type positif" est le même mais qu'il s'applique dans un cas à une fonction de deux variables et dans l'autre à une fonction d'une variable.

3.4 Notion d'homogénéité et d'interaction spatiale pour les semis de points

Dans le cas des semis de points, sans aborder les notions mathématiques précises, que nous verrons après avoir introduit le modèle, essayons de définir les notions d'homogénéité et d'interaction pour un processus non marqué.
La notion d'homogénéité est une notion d'ordre un : il s'agit de savoir si le nombre moyen de points par unité de surface est constant au travers du domaine. L'outil nécéssaire à son étude est l'intensité du processus.
La notion d'interaction est une notion d'ordre deux : il s'agit de savoir si le nombre (aléatoire) de points $N(A)$ dans une partie de l'espace A est dépendant ou indépendant (de façon probabiliste) du nombre de points $N(B)$ dans une autre partie B disjointe de A. Les phénomènes qui présentent de l'attraction ou de la répulsion entre les points comportent une dépendance entre $N(A)$ et $N(B)$. Par exemple, les positions d'animaux sur un territoire présentent de la répulsion en raison de la compétition pour la nourriture. Les positions de personnes atteintes d'une maladie épidémique vont au contraire montrer de l'attraction en raison de la contagion.

Chapitre 4

Outils statistiques pour données spatiales

Nous introduisons dans ce chapitre les outils spécifiques nécéssaires à l'étude des données spatiales. Le variogramme introduit dans le premier paragraphe est plutôt un outil de géostatistique pour la modélisation de la structure de covariance alors que les matrices voisinage et indices de Moran sont des outils pour les données de type surfacique. Le package "spdep" de R par R. Bivand permet de mettre en oeuvre les outils qui sont orientés vers les données surfaciques. Pour les données ponctuelles, on utilisera plutôt les packages "gstat", "geoR" et "geoRglm". Enfin le package "SpatStat" permet de modéliser les semis de points.

4.1 Variogramme pour variable ponctuelle continue

4.1.1 Variogramme théorique

La stationnarité est souvent une hypothèse trop forte dans les applications et une façon de l'affaiblir est de considérer la **stationnarité intrinsèque**. On n'exige pas l'existence d'un moment d'ordre un pour le champ lui-même mais seulement pour les accroissements du champ et l'on demande que

$$\mathbb{E}(X_{s+h} - X_s) = 0 \tag{4.1}$$

$$\mathbb{V}ar(X_{s+h} - X_s) = 2\gamma(h) = \mathbb{E}(X_{s+h} - X_s)^2 \tag{4.2}$$

La fonction γ s'appelle alors le semi-variogramme et 2γ le variogramme. Dans le cas où le champ est stationnaire (donc nécessairement intrinsèquement stationnaire), il existe la relation suivante entre variogramme et fonction d'autocovariance

$$\begin{aligned}\mathbb{V}\mathrm{a}r(X_{s+h} - X_s) &= \mathbb{V}\mathrm{a}r(X_{s+h}) + \mathbb{V}\mathrm{a}r(X_s) - 2\mathbb{C}\mathrm{o}v(X_s, X_{s+h}) \\ &= 2\sigma^2 - 2R(h) \\ &= 2\gamma(h)\end{aligned}$$

Le variogramme est donc un outil de description de la structure de covariance : on peut le définir pour une série temporelle mais il est peu utilisé dans ce contexte. Nous allons décrire à présent plusieurs aspects importants d'un variogramme, comme son comportement au voisinage de l'origine et à l'infini qui nous renseignent sur les propriétés du champ.
Remarquons d'abord que $\gamma(0) = 0$. On dit qu'un champ est continu en moyenne quadratique si

$$\lim_{h \longrightarrow 0} \gamma(h) = 0.$$

Cette condition équivaut à la continuité de la fonction d'autocovariance dans le cas stationnaire (dans le cas non stationnaire, cela équivaut à la continuité de la fonction de deux variables $R(s,t) = \mathbb{C}\mathrm{o}v(X_s, X_{s+h})$ sur la diagonale). Si par contre

$$\lim_{h \longrightarrow 0} \gamma(h) = c_0 \neq 0$$

alors c_0 est appellé **effet de pépite** ("nugget effect" en anglais) et témoigne d'un champ plus irrégulier. Les Figures 4.1 et 4.2 illustrent ces deux types de comportement pour un modèle de variogramme dit exponentiel défini par

$$\gamma_0(h) = \exp(-1.5 \parallel h \parallel). \tag{4.3}$$

Le graphique de gauche représente le variogramme et celui de droite montre une réalisation d'un champ gaussien stationnaire centré de variogramme donné par (4.3).
Lorsque le variogramme est borné, on appelle **seuil** ("sill" en anglais) la valeur de son asymptote et **portée** ("range" en anglais) dans la direction r la plus petite valeur de $\|r\|$ telle que $\gamma(r(1+\epsilon)) = R(0)$ quel que soit $\epsilon > 0$. La Figure 4.3 montre graphiquement ces deux paramètres dans le cas d'un variogramme sphérique (les divers modèles de variogrammes classiques seront définis dans le paragraphe 4.1.2) : la portée vaut 10 et le seuil vaut 1.
Un champ intrinsèquement stationnaire est **isotrope** si son variogramme $\gamma(h)$ ne dépend que de la norme du vecteur h et non de sa direction. Dans ce cas la fonction

$$\|h\| \longmapsto \mathbb{E}(X_{s+h} - X_s)^2 = \gamma_0(\parallel h \parallel)$$

est appellée variogramme omnidirectionnel isotrope. On parle d'**anisotropie** lorsque l'hypothèse d'isotropie n'est pas vérifiée. On peut alors représenter une fonction variogramme univariée pour chaque direction appellée **variogramme directionnel**. Si les lignes de niveau du variogramme sont des

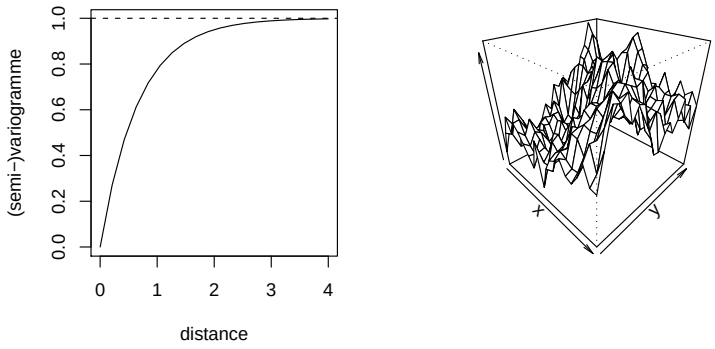


FIGURE 4.1 – Champ sans effet de pépite.

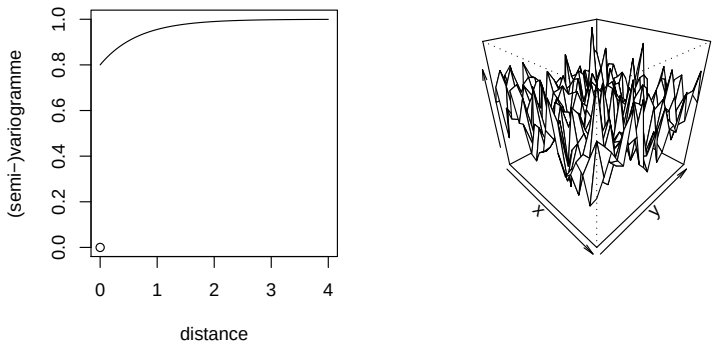


FIGURE 4.2 – Champ avec effet de pépite.

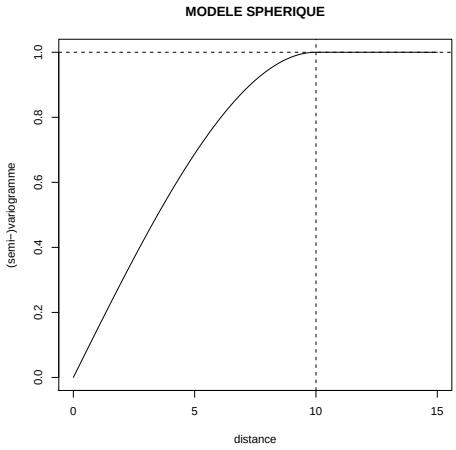


FIGURE 4.3 – Portée et Seuil d'un variogramme.

ellipses, on dit qu'il y a anisotropie géométrique. On peut alors se ramener à une configuration d'isotropie par une rotation composée par une affinité (A). Alors $\gamma(h) = \gamma_0(\| Ah \|)$.
Notons que, de façon similaire aux fonctions d'autocovariance, les fonctions variogrammes sont caractérisées par la propriété mathématique suivante. Une fonction $\gamma(t)$ de $\mathbb{R}$ à valeurs dans $\mathbb{R}$ est le variogramme d'un champ aléatoire réel intrinsèquement stationnaire si et seulement si elle est conditionnellement défini négative d'ordre un c'est à dire que quels que soit l'entier k, quels que soient les k sites $s_1, \ldots, s_k$ et les réels $a_1, \ldots, a_k$, on a

$$\sum_{i=1}^{k} \sum_{j=1}^{k} a_i a_j \gamma(s_i, s_j) \geq 0,$$

dès que les réels $a_1, \ldots, a_k$ satisfont la condition $\sum_{i=1}^{k} a_i = 0$. On dit qu'il s'agit d'une variogramme **valide**.

4.1.2 Estimation d'un variogramme

On appelle variogramme empirique un estimateur du variogramme introduit par Matheron (1962)

$$2\hat{\gamma}(h) = \frac{1}{\#N(h)} \sum_{(i,j) \in N(h)} (X_{s_i} - X_{s_j})^2,$$

où $h \in \mathbb{R}^2$, $N(h) = \{(i,j) : s_i - s_j = h\}$ et $\#A$ désigne le cardinal de l'ensemble A.
Dans le cas isotrope, on a $\gamma(h) = \gamma_0(\| h \|)$ et l'on appelle alors la fonction γ_0 le variogramme omnidirectionnel.
En pratique :

- il faut introduire une tolérance en distance ϵ_h et une tolérance angulaire ϵ_θ sinon les ensembles $N(h)$ sont souvent vides pour un "design" (disposition de points) irrégulier : $(i,j) \in N(h)$ si la valeur absolue de la différence entre $\| s_i - s_j \|$ et $\| h \|$ est inférieure à ϵ_h et si la valeur absolue de la différence entre l'angle de $s_i - s_j$ et celui de h est inférieure à ϵ_θ.
- cet estimateur n'est fiable que pour les h inférieurs au demi-diamètre de la région et tels que $N(h)$ contienne au moins 30 paires.
- le variogramme empirique n'est pas conditionnellement défini négatif.
- le variogramme empirique n'est pas robuste : en effet, pour un champ gaussien, la variable $\frac{(X_{s+h} - X_s)^2}{2\gamma(h)}$ a une loi de χ^2 à un degré de liberté et donc une forte asymétrie. En réduisant cette asymétrie, Cressie et Hawkins (1980) proposent une transformation en racine carrée de cette variable qui rend l'estimateur moins sensible aux points aberrants.

Finalement, un retour à l'échelle d'origine et une correction de biais les conduit à l'estimateur suivant :

$$2\tilde{\gamma}(h) = \frac{\{\frac{1}{\#N(h)} \sum_{(i,j)\in N(h)} (X_{s_i} - X_{s_j})^{1/2}\}^4}{0.457 + \frac{0.494}{\#N(h)}}$$

Nous avons mentionné que le variogramme empirique n'est pas un variogramme valide. Hors il est nécessaire d'avoir un estimateur conditionnellement défini négatif du variogramme pour que les variances de prédiction estimées soient positives. Pour cela, on ajuste au variogramme empirique un modèle théorique.
Après avoir choisi une famille paramétrique de variogrammes valides $\gamma(.;\theta)$, où $\theta \in \Theta$ est un vecteur de paramètres, on ajuste les valeurs du variogramme empirique $\hat{\gamma}(h_k)$ à cette famille par moindre carrés ordinaires :

$$\min_{\theta\in\Theta} \sum_{k=1}^{K} (\hat{\gamma}(h_k) - \gamma(h_k;\theta))^2.$$

Si l'on veut tenir compte de la variabilité des $\hat{\gamma}(h_k)$ ou même de leur structure de covariance, on peut aussi utiliser des moindres carrés pondérés ou même généralisés. En pratique, le choix de la famille se fait souvent en examinant visuellement la forme de la courbe empirique.
Les modèles fréquemment utilisés pour un variogramme isotropes sont

- le modèle exponentiel

$$\gamma(h) = \sigma^2(1 - \exp(-\frac{h}{\lambda})),$$

- le modèle sphérique

$$\gamma(h) = \sigma^2(\frac{3h}{2\alpha} - \frac{h^3}{2\alpha^3}),$$

- le modèle gaussien

$$\gamma(h) = \sigma^2(1 - \exp(-\frac{h^2}{\lambda^2})),$$

- le modèle de Matern

$$\gamma(h) = \sigma^2(1 - \frac{1}{2^{s-1}}\Gamma(s))\frac{h}{\lambda}^s K_s(\frac{h}{\lambda}).$$

4.2 Matrices de voisinage pour variables surfaciques

La matrice de voisinage est la version spatiale de l'opérateur retard en séries temporelles. Notons que le vocabulaire peut varier selon les auteurs et

on lui donne aussi parfois le nom de matrice de poids. Dans certains cas, elle se nomme matrice de contiguité (ce qui est en fait un cas particulier décrit plus loin). Elle constitue un outil de modélisation et non un paramètre du champ. Pour n sites géographiques, une matrice de poids W est de taille $n \times n$ et son élément w_{ij} indique l'intensité de la proximité de la zone i par rapport à la zone j (elle spécifie la topologie du domaine mais attention la proximité peut aussi avoir un sens autre que géographique comme on le verra plus loin). On impose en général que la diagonale soit nulle $w_{ii} = 0$.
Une matrice de voisinage W n'est pas nécéssairement symétrique. On peut symétriser une matrice W en la remplaçant par $(W + W')/2$.
Une matrice de poids est dite normalisée lorsqu'on impose la contrainte $\sum_{j=1}^{n} w_{ij} = 1$. Cette contrainte permet de rendre les paramètres spatiaux comparables entre divers modèles comme on le verra par la suite. On peut normaliser une matrice en divisant chaque ligne par sa somme.
Il faut faire attention au fait suivant : si on normalise une matrice symétrique, elle perd en général sa symétrie. De même si on symétrise une matrice normalisée, elle perd en général la normalisation. Seules les matrices doublement stochastiques peuvent être à la fois normalisées et symétriques. Une propriété plus faible que la symétrie qui est celle d'être semblable à une matrice symétrique va jouer un rôle plus tard. Si on normalise une matrice symétrique, elle reste semblable à une matrice symétrique.
On distingue plusieurs sortes de matrices de voisinage. Pour définir ces matrices, nous allons prendre un exemple : le tableau ci-dessous représente la position respective de neuf sites.

1	2	3
4	0	5
6	7	8

4.2.1 Matrices de contiguité

Une matrice de contiguité ne contient que des 0 et des 1. Dans le cas d'une grille régulière, on distingue les cas suivants, nommés d'après le vocabulaire des échecs :

- la matrice "rook"'consiste à poser $w_{ij} = 1$ si les sites i et j ont au moins une frontière commune ; dans notre exemple, 0 est voisin de 2, 7, 4, 5.

Ecrivons cette matrice et sa version normalisée

$$W_{rook} = \begin{pmatrix} 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{pmatrix} W^*_{rook} = \begin{pmatrix} 0 & 0 & \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 & \frac{1}{3} & 0 & 0 \\ \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 & 0 & 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \\ \frac{1}{3} & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{pmatrix}$$

- la matrice “bishop” consiste à poser $w_{ij} = 1$ si les sites i et j ont au moins un sommet commun ; dans notre exemple, 0 est voisin de 1, 3, 6, 8.

- la matrice “queen” consiste à poser $w_{ij} = 1$ si les sites i et j ont au moins une frontière ou un angle commun ; dans notre exemple, 0 est voisin de 1, 2 , 3, 4 , 5, 6, 7, 8.

Dans le cas de positions irrégulières, deux zones sont contigues si elles ont une frontière en commun. Ces matrices sont automatiquement symétriques.

4.2.2 Matrices basées sur la distance entre centroïdes

Notons $\mathbb{I}(A)$ la fonction indicatrice de l'évènement A et $d(s_i, s_j)$ une mesure de distance entre les sites s_i et s_j. Cette distance peut désigner tout simplement la distance euclidienne (distance géographique à vol d'oiseau) mais peut être aussi un temps de trajet entre les deux sites, ou encore de la forme $d(s_i, s_j) =\mid x_i - x_j \mid$, où x_i désigne une caractéristique socio-économique pertinente. Voici quelques façons couramment utilisées pour définir une matrice de voisinage à partie d'une distance :

- $w_{ij} = \mathbb{I}(d(s_i, s_j) \leq S)$, où S est un seuil fixé.
- $w_{ij} = \frac{C}{d(s_i,s_j)^\alpha}$, où C et α sont des constantes fixées.
- $w_{ij} = \exp(-\alpha d(s_i, s_j))$, où α est une constante fixée.

Notons que ces matrices sont automatiquement symétriques.

4.2.3 Matrices basées sur les plus proches voisins

Etant donné une notion de distance et un entier k, pour chaque site s_i, on ordonne les autres sites en fonction de leur distance à s_i et l'on détermine ainsi les k plus proches voisins de s_i. La matrice contient alors sur la ligne i des 1 pour les positions des k plus proches voisins et des 0 sinon. Ces matrices ne sont en général pas symétriques.

4.2.4 Matrices basées sur triangulation de Delaunay

On appelle triangulation de Delaunay l'unique triangulation telle que le cercle circonscrit à trois sommets quelconques ne contienne aucun autre

sommet. Cette triangulation permet de construire une matrice de la façon suivante : deux sites sont voisins si le segment les joignant est une arête de la triangulation. Ces matrices présentent cependant des liaisons pour les sites en bordure avec des voisins très éloignés.
Notons qu'on peut combiner le principe des plus proches voisins (ou de la contiguité) et celui de la distance en une même matrice. Ceci permet de combiner les avantages des deux approches dans le cas de positions très hétérogènes des centroides de zones dans l'espace.

4.2.5 Variable spatialement décalée

Si $\mathbf{X}$ désigne le vecteur colonne des valeurs X_{s_i} du champ aux points d'observation, on appelle **variable spatialement décalée** associée à $\mathbf{X}$ la variable $W\mathbf{X}$. Si W est normalisée, l'élément i du vecteur $(W\mathbf{X})$ est une moyenne pondérée des valeurs du champ dans les zones voisines de la position i. Dans le cas d'une variable de comptage, il peut être plus intéressant de ne pas normaliser la matrice de voisinage binaire de façon que $(W\mathbf{X})$ représente la somme (et non la moyenne) des valeurs voisines.
Notons que si la matrice W est normalisée et si le vecteur $\mathbf{X}$ est centré, le vecteur spatialement décalé $W\mathbf{X}$ est également centré.

4.3 Indice de Moran pour variable surfacique continue

Pour une matrice de voisinage W vérifiant $w_{ii} = 0$ et une variable $X_{s_i} = X_i, i = 1, \dots n$, l'indice de Moran est défini par :

$$I = \frac{\frac{\sum_{i,j} w_{ij}(X_i - \bar{X})(X_j - \bar{X})}{\sum_{i,j} w_{ij}}}{\frac{\sum_i (X_i - \bar{X})^2}{n}}$$

C'est le rapport d'une sorte de covariance entre unités contigües à la variance du champ : il est donc comparable à un coefficient d'autocorrélation. Cet indice est indépendant des unités dans lesquelles X est exprimé. Si l'on symétrise la matrice W, (i.e. $W \longrightarrow (W + W')/2$), I est inchangé.
Si X est une variable centrée, les valeurs de X de même signe et géographiquement proches contribuent positivement à $\mathbf{I}$. Les valeurs positives et fortes de $\mathbf{I}$ indiquent une autocorrélation spatiale positive, les valeurs négatives et fortes de $\mathbf{I}$ une autocorrélation spatiale négative et les valeurs proches de 0 une absence d'autocorrélation.
Attention : le $\mathbf{I}$ de Moran dépend du choix de la matrice W, et peut être affecté par le niveau d'agrégation (effet d'échelle) ainsi que par la forme des unités spatiales.

Il n'est pas possible d'interpréter un indice de Moran brut et nous verrons plus loin comment le normaliser et l'utiliser pour un test d'autocorrélation. De façon purement descriptive, nous pouvons cependant faire un **diagramme de Moran** et en tirer des conclusions qualitatives sur l'autocorrélation. Pour une matrice de voisinage W, le diagramme de Moran d'un champ X_s consiste en un diagramme de dispersion de la variable $\mathbf{X}$ contre la variable spatialement décalée $W\mathbf{X}$. On montre alors que la pente de la droite de régression linéaire simple de $W\mathbf{X}$ contre $\mathbf{X}$ est égale à l'indice de Moran. Grace au signe de la pente, on lit sur le graphique la tendance générale de l'autocorrélation, une pente positive correspondant à une autocorrélation positive et inversement.
Notons qu'il existe d'autres indices similaires comme celui de Geary et de Getis. Le coefficient **C de Geary** est défini par :

$$C = \frac{n-1}{2\sum_{i,j} w_{ij}} \frac{\sum_{i,j} w_{ij}(X_{s_i} - X_{s_j})^2}{\sum_i (X_{s_i} - \bar{X})^2}$$

Cet indice ressemble à la statistique de Durbin Watson en séries temporelles. Les valeurs faibles de $\mathbf{C}$ indiquent une autocorrélation spatiale positive et les valeurs fortes de $\mathbf{C}$ une autocorrélation spatiale négative. Cet indice est indépendant des unités dans lesquelles le champ X est exprimé.
Pour comparaison, on rappelle que la statistique de Durbin-Watson pour une série temporelle centrée est donnée par

$$DW = \frac{\sum_{t=2}^{n}(X_t - X_{t-1})^2}{\sum_{t=1}^{n} X_t^2}.$$

Il existe une version locale de l' indice de Moran qui mesure une version **locale** de la notion d'autocorrélation. L'indice de Moran local associé au site i se calcule simplement par

$$I_i = \sum_{j=1}^{n} w_{ij}(X_i - \bar{X})(X_j - \bar{X}) = (X_i - \bar{X}) \sum_{j=1}^{n} w_{ij}(X_j - \bar{X}).$$

La somme des indices de Moran locaux redonne l'indice de Moran si bien que I_i peut être considéré comme la contribution du point i à l'indice global. Il constitue une mesure d'influence du point i sur l'autocorrélation globale.

4.4 Statistique "join counts" pour variable surfacique qualitative

Ces statistiques sont souvent introduites dans le cas dichotomique. Si X_i a deux modalités 0 et 1 avec : $P(X_i = 1) = p$, on introduit les statistiques

suivantes appelées “join counts”

$$BB = \frac{1}{2}\sum_{i,j} w_{ij} X_i X_j$$

$$BW = \frac{1}{2}\sum_{i,j} w_{ij} (X_i - X_j)^2$$

Il est facile de comprendre par exemple que si W est une matrice binaire, la statistique BB compte le nombre de couples de sites voisins pour lesquels $X = 1$. Donc si BB prend une grande valeur, cela va plaider pour une autocorrélation spatiale positive. Inversement, BW compte le nombre de couples de sites avec une valeur différente de X. Nous verrons plus loin comment utiliser ces statistiques pour évaluer l'autocorrélation spatiale d'une variable surfacique binaire.

4.5 Processus ponctuels

La théorie des processus ponctuels est un cadre mathématique adapté à la modélisation de répartitions aléatoires de points. Le package “spatstat” de R (A. Badeley et R. Turner) permet la modélisation et la simulation de tels processus. Citons également les packages “splancs” et “VR” de R. Nous allons brièvement évoquer la définition mathématique d'un tel processus. Etant donné un sous-ensemble E de $\mathbb{R}^2$, un processus ponctuel X est une variable aléatoire à valeurs dans l'espace N_{lf} des sous-ensembles x localement finis de E, c'est à dire tels que le nombre de points de x contenus dans tout borné de E est fini. Ces sous-ensembles ou “configurations” sont considérés comme des suites non ordonnées de points et notés $\{x_1, \cdots, x_n\}$. Il faut bien sûr munir N_{lf} d'une tribu $\mathcal{N}$ pour définir proprement le processus mais nous ne rentrerons pas dans ces détails. Pour un borélien B de $\mathbb{R}^2$, on notera $N(B)$ le nombre de points d'une configuration appartenant à B : pour tout B, $N(B)$ est une variable aléatoire. La loi d'un processus ponctuel est définie par les probabilités $\mathbb{P}(X \in Y)$, pour tout $Y \in \mathcal{N}$: cette famille contient en particulier la famille des probabilités fini-dimensionnelles $\mathbb{P}(N(B_1) = n_1, \ldots, N(B_k) = n_k)$. Il est à noter qu'un processus ponctuel est caractérisé de manière unique par la famille des probabilités d'évitement $\mathbb{P}(N(B) = 0)$, lorsque B parcourt les boréliens.
Nous adopterons ici une approche plus commode pour les applications consistant à définir une densité jointe $f\big((x_1, \cdots, x_n), n\big)$ pour les variables N, nombre de points, et $X_1, \ldots, X_n$, localisations des N points (Cressie, 1993, p.622). On a alors

$$\sum_{n=0}^{\infty} \int_{E^n} f\big((s_1, \cdots, s_n), n\big) ds_1 \cdots ds_n = 1.$$

On dit qu'un processus ponctuel est **stationnaire** lorsque sa loi (et par conséquent toutes ses caractéristiques) est invariante par translation. On dit qu'il est **isotrope** lorsque toutes ses caractéristiques sont invariantes par rotation. Basée sur l'observation d'une seule réalisation sur une fenêtre bornée, l'étude statistique d'un processus ponctuel porte sur l'estimation de ses caractéristiques, le diagnostic d'homogénéité, d'interaction et enfin la modélisation en fonction de caractéristiques explicatives.

4.5.1 Un exemple : le processus de Poisson homogène

Le processus de Poisson homogène est le modèle de base en théorie des processus ponctuels car il formalise le concept de points répartis au hasard. Il est défini par les deux conditions suivantes pour un domaine Ω de $\mathbb{R}^2$:

1. il existe un réel $\lambda > 0$ tel que pour tout borélien A de $\mathbb{R}^2$, $N(A)$ suit une loi de Poisson de moyenne $\lambda \mid A \mid$, où $\mid A \mid$ désigne l'aire de A.
2. sachant que $N(A) = n$, les n points du processus qui sont dans A forment un échantillon de la loi uniforme sur A.

Ces deux conditions impliquent la condition (3) suivante : pour deux boréliens A et B, les variables aléatoires $N(A)$ et $N(B)$ sont indépendantes. Le processus de Poisson homogène est stationnaire et isotrope.
On démontre que les probabilités fini-dimensionnelles de ce processus sont données par

$$\mathbb{P}(N(B_1) = n_1, \ldots, N(B_k) = n_k) = \frac{\lambda^{n_1+\ldots+n_k} \mid B_1 \mid^{n_1} \ldots \mid B_k \mid^{n_k}}{n_1! \ldots n_k!} \exp(-\sum_{i=1}^{k} \lambda \mid B_k \mid).$$

Conditionnellement au nombre total de points $N = N(\Omega)$, les positions sont indépendantes et identiquement distribuées selon une loi uniforme sur Ω. Mais il ne faut pas confondre ce modèle avec celui de points uniformément répartis sur Ω pour lequel le nombre de points n'est pas aléatoire (ce processus porte le nom de processus ponctuel binomial car le nombre de points contenus dans un borélien A de Ω suit alors une loi binomiale).

4.5.2 Le processus de Poisson inhomogène

Le processus de Poisson homogène ayant une intensité constante ne peut servir à modéliser des phénomènes présentant une forte hétérogénéité spatiale. Etant donné une mesure d'intensité Λ, on peut définir le processus de Poisson X de mesure d'intensité Λ par les deux conditions suivantes

- (i) le nombre de points $N(A)$ de X dans tout borélien A de $\mathbb{R}^2$, suit une loi de Poisson de moyenne $\Lambda(A)$,
- (ii) les nombres de points de X dans k boréliens $A_1, \ldots, A_k$ disjoints de $\mathbb{R}^2$ sont k variables aléatoires indépendantes.

Ainsi défini, ce processus n'est pas stationnaire sauf si l'intensité est constante. Conditionnellement à $N = n$, les n points $X_1, \ldots, X_n$ sont alors i.i.d.. Lorsque la mesure d'intensité est absolument continue par rapport à la mesure de Lebesgue ($\Lambda(A) = \int_A \lambda(x)dx$), il existe une relation directe entre l'intensité du processus ponctuel $\lambda(.)$ et la densité d-dimensionnelle $f(.)$ de toute localisation X_i conditionnellement à N :

$$\forall s \in E,\ f(s) = \frac{\lambda(s)}{\int_E \lambda(s)\nu(ds)}.$$

4.5.3 Caractéristique d'ordre un : l'intensité

L'intensité est l'analogue pour le processus ponctuel de l'espérance pour une variable aléatoire. On commence par définir la mesure d'intensité comme une mesure sur les boréliens B de $\mathbb{R}^2$ vérifiant

$$\Lambda(B) = \mathbb{E}(N(B)),$$

de façon que $\Lambda(B)$ représente le nombre moyen de points du processus dans B. Si le processus est stationnaire, cette mesure est proportionnelle à la mesure de Lebesgue et le facteur de proportionalité, λ, appellé intensité, représente le nombre moyen de points du processus par unité de surface. Plus généralement, si Λ est absolument continue par rapport à la mesure de Lebesgue, il existe une fonction λ localement intégrable définie sur E telle que pour tout borélien B,

$$\Lambda(B) = \int_B \lambda(x)dx.$$

Cette fonction λ porte le nom de fonction d'intensité du processus ponctuel. Comme on l'a vu ci dessus, si le processus est stationnaire, la fonction d' intensité est constante. Inversement, si le fonction d'intensité est constante, le processus est dit stationnaire au premier ordre ou homogène (sinon, il est dit inhomogène). Dans le cas du processus de Poisson homogène, la fonction d'intensité est constante égale au paramètre λ de la définition du paragraphe 4.5.1.

4.5.4 Estimation de l'intensité

Dans le cas d'un processus homogène d'intensité λ, un estimateur sans biais de l'intensité est donné par $\hat{\lambda} = \frac{N}{|W|}$, où W est la fenêtre d'observation et $N = N(W)$ le nombre de points observés dans cette fenêtre. Il coincide en fait avec l'estimateur du maximum de vraisemblance dans le cas où le processus est un Poisson homogène.
Dans le cas inhomogène, on peut utiliser un estimateur non paramétrique, introduit par Diggle (1985) donné par

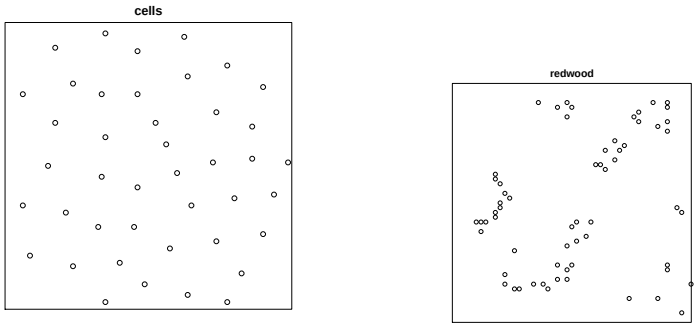


FIGURE 4.4 – Processus régulier (à gauche) et agrégé (à droite)

$$\hat{\lambda}_h(s) = \frac{\sum_{i=1}^{N} h^{-d} K\big(\frac{s-X_i}{h}\big)}{\int_E h^{-d} K\big(\frac{s-u}{h}\big) du} \tag{4.4}$$

où le dénominateur est un terme de correction au bord nécessaire lorsque le domaine d'observation est limité et où K est une fonction noyau. Cet estimateur est, de même qu'un estimateur non paramétrique de densité, peu sensible au choix du noyau K. Le choix de la largeur de bande ou fenêtre h permettant de minimiser l'erreur quadratique moyenne intégrée

$$EQMI(h) = \mathbb{E}\Big\{ \int_E \big(\hat{\lambda}_h(s) - \lambda(s)\big)^2 ds \Big\}$$

se fait selon des méthodes similaires au cas de l'estimation de densité.

4.5.5 Caractéristiques d'ordre deux : Fonctions F, G, J, K

Du fait de la propriété (ii) (voir paragraphe 4.5.2), le processus de Poisson implique une absence d'interaction entre les évènements. Les caractéristiques du second ordre vont permettre de mettre en évidence deux autres types de comportement. On distingue d'une part les processus pour lesquels les évènements ont tendance à s'attirer (agrégation) et ceux pour lesquels les évènements ont tendance à se repousser (régularité). On voit la différence entre ces deux comportements sur la figure suivante.
Nous allons d'abord introduire un certain nombre de fonctions associées à un processus ponctuel basées sur les distances entre points.

Distance d'un point courant au plus proche voisin

Soit x un point de E qui ne figure pas nécéssairement dans une configuration du PP X. Pour un processus ponctuel X homogène, on définit

$$F_x(r) = \mathbb{P}(d(x, \{x_1, \cdots, x_n\} \setminus \{x\}) \leq r).$$

Notons qu'en raison de l'homogénéité F_x ne dépends pas de x, c'est pourquoi nous le noterons plus simplement F. F est la fonction de répartition de la distance au plus proche voisin et peut aussi s'interpréter comme la mesure de "l'espace vide" (c'est pourquoi on l'appelle "empty space function" en anglais) dans le sens suivant : $1 - F(r)$ est la probabilité qu'une boule de centre 0 (ou un quelconque point de E fixé) ne contienne aucun point de X. Pour estimer F, on utilise en général une grille fine de points définie sur E qui permet d'approximer les distances au plus proche voisin.
Sous l'hypothèse CSR d'homogénéité spatiale sur $\mathbb{R}^2$, la fonction F a la forme analytique suivante pour $x > 0$

$$F(x) = 1 - \exp(-\pi\lambda x^2).$$

On en déduit la méthode suivante pour évaluer qualitativement l'hypothèse CSR par des simulations. On simule M réalisations d'un processus de Poisson homogène dans E et on calcule la fonction $\hat{F}_k(r)$ pour chaque simulation k. On détermine ensuite l'enveloppe supérieure F_U et inférieure F_L par

$$F_U(r) = \max_{k=1}^{M} \hat{F}_k(r), F_L(r) = \min_{k=1}^{M} \hat{F}_k(r).$$

Si la fonction $\hat{F}(r)$ de notre réalisation se trouve dans l'enveloppe, on en déduit que le modèle de Poisson homogène est compatible avec les données. Pour le jeu de données cells (positions de cellules) de spatstat, on voit que la fonction $\hat{F}$ en noir sur la figure 4.5.5 sort de l'enveloppe (en pointillés).

Distance d'un point du PP au plus proche voisin

Si cette fois, on s'intéresse à la distance entre un point du PP et son plus proche voisin, on définit la fonction de répartition de ces distances G par

$$G(r) = \mathbb{P}(d(x, \{x_1, \cdots, x_n\} \setminus \{x\}) \leq r \mid x \in X).$$

Un estimateur classique de G est donné par la fonction de répartition empirique définie par

$$\hat{G}(r) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(d(x_i, x_{j(i)}) \leq r),$$

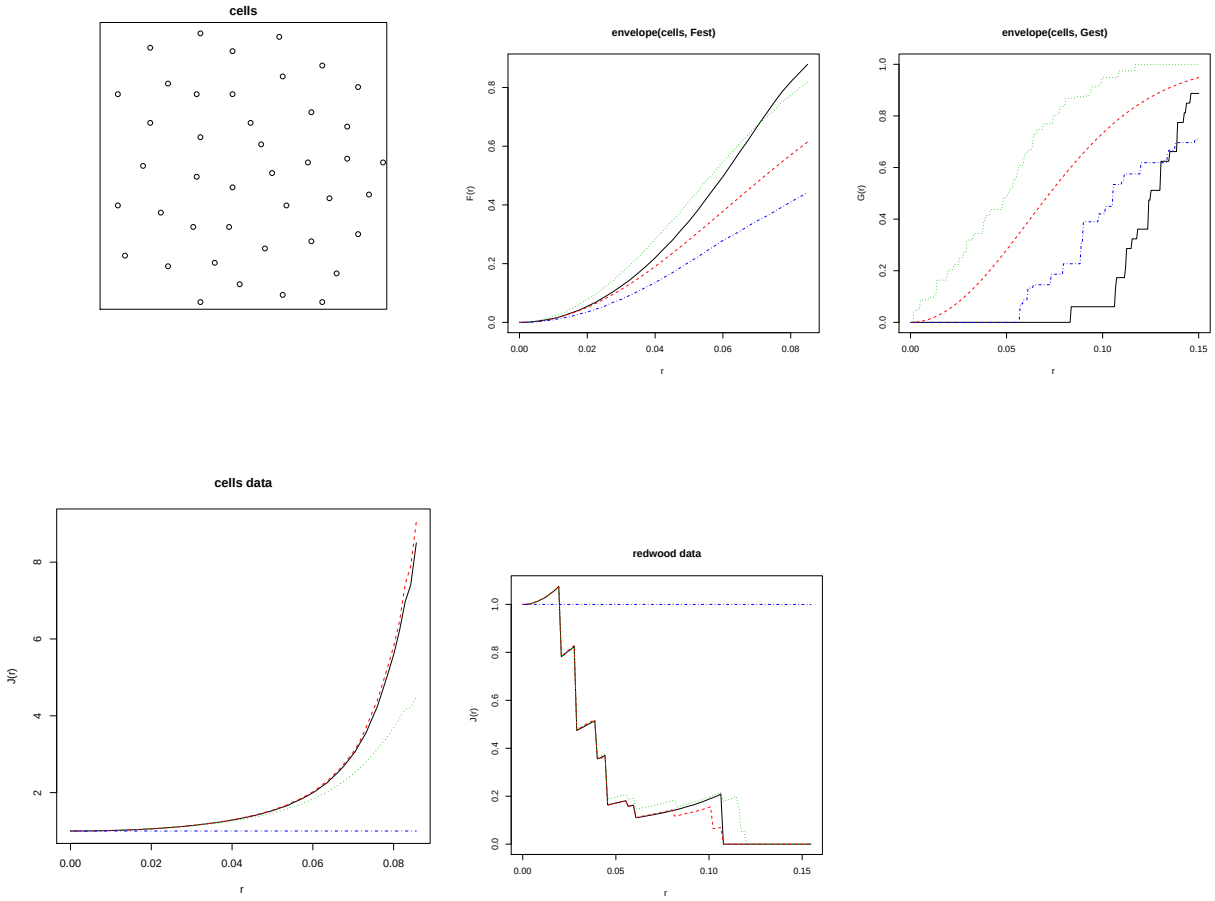


FIGURE 4.5 – Processus régulier à gauche et agrégé à droite

où $x_{j(i)}$ est le point de X le plus proche de x_i. Le même principe d'enveloppes peut être appliqué et l'on voit que sur les données cells, l'estimateur de la fonction G sort également de l'enveloppe sous l'hypothèse CSR.
A partir de F et G, on peut définir la fonction J par

$$J(r) = \frac{1 - G(r)}{1 - F(r)}.$$

$J = 1$ correspond au cas d'un processus poissonnien. $J > 1$ indique une tendance à la régularité et $J < 1$ à l'agrégation. La figure 4.5 montre la différence de comportement de J entre les données cells et les données redwood.

Fonction de corrélation des paires, fonction K de Ripley

De même que l'on a introduit la mesure d'intensité pour le moment d'ordre 1, le rôle du moment d'ordre 2 est joué par la mesure de moment factoriel d'ordre 2, donnée pour tous boréliens B_1 et B_2 de $\mathbb{R}^2$ par

$$\alpha_2(B_1 \times B_2) = \mathbb{E}(N(B_1)N(B_2)) - \Lambda(B_1 \cap B_2).$$

Lorsque cette mesure est absolument continue par rapport à la mesure de Lebesgue, on note ρ_2 sa densité, appellée densité d'intensité d'ordre 2. Pour un PP stationnaire, la fonction $\rho_2(x, y)$ ne dépends que de $x - y$. Si de plus le PP est isotrope, elle ne dépends que de $\parallel x - y \parallel$.
A partir de ρ_2, on définit la **fonction de corrélation des paires** g par

$$g(x, y) = \frac{\rho_2(x, y)}{\lambda(x)\lambda(y)}.$$

C'est cette fonction qui conduit à une autre méthode de comparaison avec un PP de Poisson. En effet, il est facile de voir que pour un PP de Poisson, on a $g(x, y) = 1$. Si $g(x, y) > 1$, cela indique que pour ce PP, il est plus probable d'observer un couple de points en x et y que pour un PP de Poisson ayant la même intensité. Si le PP est stationnaire et isotrope, g est une fonction de $r = \parallel x - y \parallel$; $g(r) > 1$ indique une tendance à l'agrégation pour des points à distance r, et inversement, $g(r) < 1$ indique une tendance à la répulsion pour des points à distance r.
Une façon alternative de caractériser les propriétés du second ordre est au travers de la fonction K de Ripley et de la fonction L qui lui est associée. Pour un PP stationnaire, introduisons la mesure κ, appellée mesure des moments réduits d'ordre deux, pour un borélien B par

$$\kappa(B) = \frac{1}{\lambda^2} \int_B \rho_2(x)dx.$$

Si de plus le PP est isotrope, en prenant pour B une boule $B(0, r)$ de centre l'origine et de rayon r, la fonction K de Ripley est définie par

$$K(r) = \kappa(B(0, r)).$$

$K(r)$ peut aussi s'interpréter comme le nombre moyen de points du PP dans une boule centrée en un des points du PP, hormis le centre lui-même. Pour un PP de Poisson homogène, $K(r) = \pi r^2$ et ceci engendre une autre méthode de comparaison avec un modèle de Poisson. Pour faciliter la comparaison et aussi pour réduire la variance, il est d'usage de renormaliser la fonction K en définissant la fonction L par

$$L(r) = (\frac{K(r)}{\pi})^{1/2}.$$

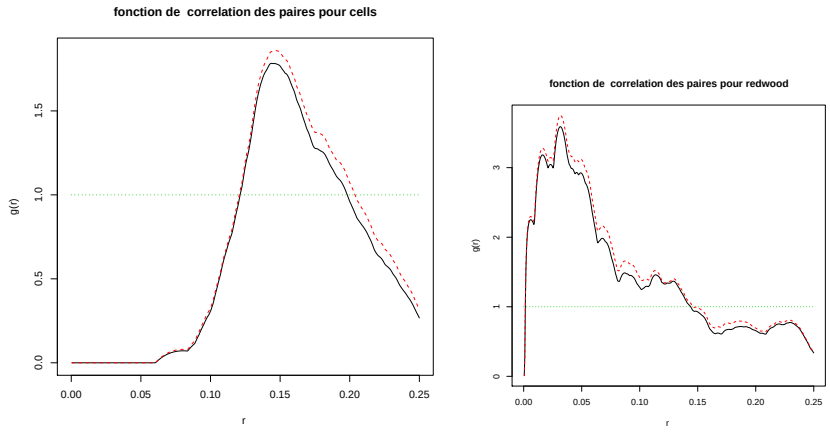


FIGURE 4.6 – Processus régulier à gauche et agrégé à droite

Pour le PP de Poisson homogène, la fonction L est donc égale à r. Lorsque $L(r) - r > 0$, cela indique un phénomène d'agrégation pour des distances inférieures ou égales à r, et lorsque $L(r) - r > 0$, cela indique un phénomène de régularité pour des distances inférieures ou égales à r.
Pour un PP stationnaire et isotrope, les relations suivantes existent entre g, ρ_2 et K :

$$g(r) = \frac{\rho_2(r)}{\lambda^2} = \frac{K'(r)}{2\pi r}, K(r) = \frac{2\pi}{\lambda^2} \int_0^r u\rho_2(u)du. \tag{4.5}$$

Pour estimer ces diverses caractéristiques du second ordre, on peut commencer par estimer ρ_2 par un estimateur à noyau de la densité incluant une correction de bord (diverses corrections existent). On peut alors en déduire un estimateur de la fonction de corrélation des paires en divisant par $\hat{\lambda}(x)\hat{\lambda}(y)$, où $\hat{\lambda}$ est par exemple l'estimateur de Diggle de l'intensité (voir 4.4).
On peut estimer directement la fonction K par

$$\hat{K}(r) = \sum_{x \in X, y \in W_{\ominus r}} \frac{1(x - y \in B(0,r))}{\hat{\lambda}(x)\hat{\lambda}(y)},$$

où $W_{\ominus r}$ désigne l'ensemble des points de la fenêtre W tels que la boule centré en ce point et de rayon r soit entièrement incluse dans W. D'autres formules existent mais consistent essentiellement à faire d'autres corrections de bord. Cet estimateur se calcule dans spatstat avec la fonction Kest et l'option correction="border". Notons que les relations 4.5 permettent aussi de déduire un estimateur de g à partir d'un estimateur de K.
La figure 4.6 montre un estimateur de la fonction de corrélation des paires pour les données cells et redwood.
La figure 4.7 présente des estimateurs des fonctions de Ripley pour les données cells et redwood et l'on voit bien à nouveau la différence de comportement entre processus régulier et agrégé.

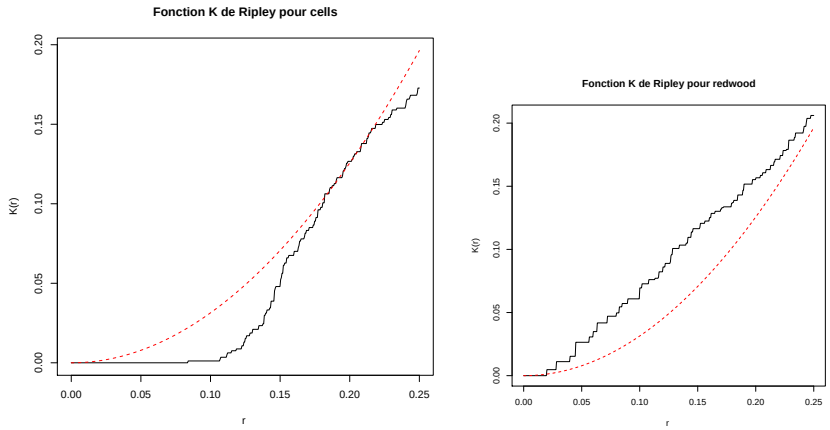


FIGURE 4.7 – Processus régulier à gauche et agrégé à droite

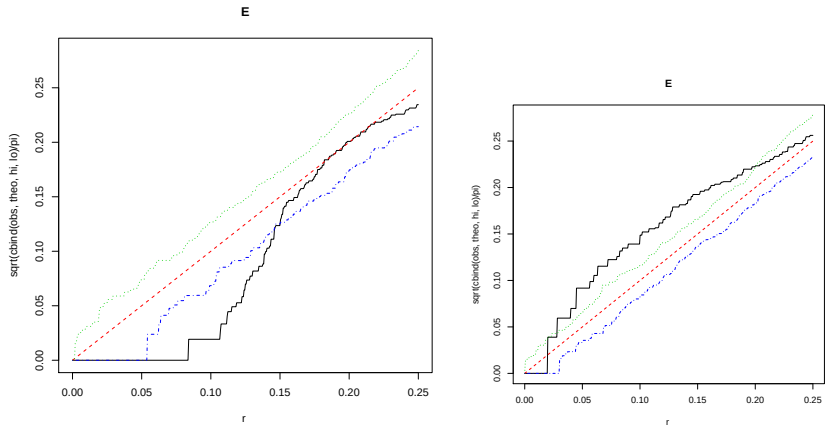


FIGURE 4.8 – Processus régulier à gauche et agrégé à droite

Enfin la figure 4.8 présente des enveloppes de la fonction L pour les données cells et redwood.

Chapitre 5

Méthodes exploratoires pour données spatiales

L'analyse exploratoire est préalable à toute modélisation statistique : c'est la phase de mise en place des bases de données, de leur nettoyage et du premier contact avec les variables. Il faut recenser et traiter les valeurs manquantes et les valeurs aberrantes, et produire les premiers diagnostics descriptifs uni et multidimensionnels. Dans le cas de données géoréférencées, aux techniques habituelles sur lesquelles nous ne reviendrons pas ici, s'ajoutent des méthodes spécifiques qui font l'objet de ce chapitre. Les Systèmes d'Information Géographique [1] ("SIG"), permettent de gérer et de cartographier des données géoréférencées mais ils n'intègrent pas ou peu d'outils statistiques sophistiqués, en particulier les outils spécifiques aux données spatiales. Nous utiliserons le terme "variable spatiale" pour désigner un ensemble de n observations d'un champ aléatoire en n sites ou n zones. Nous présentons dans ce chapitre les fondements de cette analyse. Un module de R dénommé "GeoXp" permet de mettre ces techniques en pratique (Laurent et al. 2012). Il permet une exploration interactive avec un dialogue entre graphique statistique et carte géographique.

5.1 Analyse exploratoire des matrices de voisinage

Avant de faire le choix d'utiliser une matrice de voisinage particulière, il est bon d'en faire une exploration. Par exemple, on peut représenter graphiquement les liens non nuls par des segments sur la carte et produire quelques caractéristiques de la distribution du nombre de voisins et de la distance au plus proche voisin. On peut ainsi comparer ces caractéristiques pour quelques choix différents de matrices.
Pour une matrice de voisinage W et une variable X données, le graphique

1. Exemple de SIG libre : Quantum GIS, `http://www.qgis.org/`

des voisinages consiste en un simple diagramme de dispersion où l'on porte pour tout site i, en abscisse la valeur X_i de la variable X au site i et en ordonnée les valeurs X_j de la variable X aux sites j voisins de i au sens de W, c'est-à-dire tels que $w_{ij} \neq 0$. Du point de vue de l'exploration de la matrice de voisinage, ce diagramme permet d'explorer la matrice dans le sens suivant :

1. il permet de visualiser qui est voisin de qui,
2. il permet d'aprécier visuellement la taille des voisinages lorsque la matrice est définie par un nombre de plus proches voisins : la "largeur" de la bande autour de la diagonale sur le nuage illustre l'étendue des voisinages,
3. il permet d'aprécier visuellement le nombre de voisins lorsque la matrice est définie par une distance seuil.

5.2 Analyse exploratoire d'une tendance directionnelle

Une variable présente une tendance dans une direction donnée, par exemple Sud-Est/Nord-Ouest, si celle-ci présente une moyenne non constante dans cette direction. Supposons dans un premier temps que la direction est connue et pour simplifier qu'il s'agit de la direction Nord-Sud ou Est-Ouest. Pour mettre en évidence cette tendance et la décrire, c'est-à-dire préciser comment varie la moyenne (croissante, décroissante, en forme de U, etc.), on superpose une grille régulière à la carte faite de petits rectangles, pour un nombre choisi de lignes et de colonnes. On calcule dans chaque rectangle les moyennes et médianes de toutes les unités dont le centroïde se situe dans ce rectangle, et on fait de même sur chaque ligne et colonne. On met ensuite la carte en regard avec à droite les moyennes et/ou médianes par ligne et en dessous les moyennes et/ou médianes par colonne (ainsi que du nuage des moyennes et/ou médianes par rectangle). La variation sur le graphique de droite des moyennes et/ou médianes (que l'on peut interpoler pour une meilleure lisibilité) met alors en relief une tendance Nord-Sud si les moyennes et/ou médianes ne sont pas constantes et respectivement la variation sur le graphique du dessous une tendance Est-Ouest si les moyennes et/ou médianes ne sont pas constantes. Si maintenant la direction est connue mais n'est ni Nord-Sud, ni Est-Ouest, on peut alors introduire un angle de rotation de la carte permettant de se ramener à la situation précédente. Finalement, dans le cas plus réaliste où l'on ne connait pas d'avance une direction de tendance, il faut alors utiliser un autre graphique exploratoire pour déterminer une telle direction. Pour un couple de sites i et j sur la carte, on définit l'angle θ_{ij} entre l'axe des abscisses et le vecteur d'origine i et d'extrémité j. On réalise ensuite un diagramme de dispersion dans lequel on associe à

l'angle θ_{ij} la valeur absolue $\mid X_i - X_j \mid$ de la différence entre les valeurs de la variable en ces deux sites. Si la variable présente une tendance directionnelle dans la direction θ, les différences $\mid X_i - X_j \mid$ pour les couples (i, j) tels que θ_{ij} est voisin de θ vont être plus importantes que dans les autres directions. Notons que ce graphique permet de détecter des tendances bien marquées.

5.3 Analyse exploratoire de l'autocorrélation spatiale

5.3.1 Le diagramme de Moran

Il s'agit d'un outil permettant d'explorer l'autocorrélation spatiale d'une variable surfacique continue. Un diagramme de Moran est un nuage de points présentant une variable d'interêt X en abscisse et la variable spatialement décalée WX en ordonnée. La variable X est centrée en abscisse et par conséquent la variable spatialement décalée WX en ordonnée est également centrée lorsque W est normalisée. Un point du quadrant $x \geq 0, y \geq 0$ correspond à un site où la variable X est supérieure à sa moyenne et où la variable WX également, témoignant d'une autocorrélation **locale** positive. Un point du quadrant $x \leq 0, y \geq 0$ correspond à un site où la variable X est inférieure à sa moyenne et où la variable WX est par contre supérieure à sa moyenne, témoignant d'une autocorrélation **locale** négative. Les deux autres quadrants s'interprètent de même. Une non linéarité du nuage indique plusieurs régimes d'association spatiale.

5.3.2 Le nuage de variogramme

Il s'agit d'un outil permettant d'explorer l'autocorrélation spatiale d'une variable ponctuelle continue. Pour une variable donnée possédant un variogramme isotrope, le "nuage de variogramme" est une représentation du demi-carré de la différence entre les valeurs de la variable mesurée en deux sites distants de h en fonction de la distance h pour tous les couples de sites. Si l'on revient à la formule (4.2) définissant le variogramme, on voit aisément qu'un lissage de ce nuage de points estime la fonction $\gamma(h)$. Ce lissage peut être superposé au nuage de points permettant ainsi d'analyser les diverses caractéristiques du variogramme (portée, seuil, effet de pépite).

5.4 Analyse exploratoire des points atypiques spatiaux

En statistique spatiale, il y deux sortes de points atypiques : les atypiques au sens ordinaire que nous nommerons ici "globaux" par opposition aux atypiques locaux que nous allons définir.

Au sens ordinaire, un point est dit **atypique global** pour la variable X si sa valeur pour X est extrême par rapport à l'ensemble de la distribution de X. Il y a bien sûr un degré de liberté dans la façon dont on définit "extrême".
Etant donnée une structure de voisinage sur l'ensemble des sites, un point est dit **atypique local** pour la variable X si sa valeur pour X est extrême par rapport à l'ensemble de la sous-distribution des X sur les sites voisins du site concerné.
Un aberrant global est en général un aberrant local (sauf dans le cas de groupes d'atypiques), mais un aberrant local peut très bien ne pas être un aberrant global.
Divers graphiques exploratoires permettent de détecter les atypiques locaux. On peut utiliser le nuage de variogramme, le diagramme des voisins, le diagramme de Moran, etc.
Avec le diagramme des voisins, les points éloignés de la diagonale correspondent à des couples de sites voisins dont les valeurs différent. Un atypique local aura donc sur sa verticale des points éloignés de la diagonale.
Avec le diagramme de Moran, on peut repérer certains atypiques locaux, ceux qui contribuent au I de Moran global avec un I local significatif.
Avec le nuage de variogramme, on procède ainsi. Tout d'abord, il est préférable pour cet objectif d'utiliser la version robuste du variogramme obtenue en remplaçant le carré de la différence par la racine carrée de la différence : en effet, pour un champ gaussien isotrope X, la loi de $\frac{(X_{s+h}-X_s)^2}{2\gamma(h)}$ est un χ^2 à un degré de liberté donc une loi asymétrique, alors que la loi de $\frac{(X_{s+h}-X_s)^{1/2}}{2\gamma(h)}$ est presque symétrique. Décider si une valeur élevée est un point atypique est plus facile sur une loi symétrique car une loi asymétrique peut produire des valeurs élevées qui ne sont pas atypiques. On peut repérer, sur le nuage de variogramme, des couples de sites atypiques en ce sens que la différence entre les valeurs du champ entre ces sites est grande comparée aux différences entre couples de sites distants de la même distance. En pratique, ce sont surtout les atypiques globaux qui ressortent.

Chapitre 6

Tests d'autocorrélation et d'homogénéité spatiale

On s'intéresse dans ce chapitre à la question de savoir si les données nécéssitent un traitement spécifique aux données spatiales. En effet une variable observée spatialement peut très bien dans l'absolu ne pas présenter d'hétérogénéité ni d'autocorrélation et dans ce cas elle peut être étudiée avec des techniques usuelles. Dans le cas de données de type ponctuel ou de type surfacique, il s'agira de répondre à la question : une variable observée présente-t-elle de l'autocorrélation spatiale et comment construire un test. Dans le cas de données de type semis de points, il s'agira de tester l'homogénéité spatiale du phénomène.

6.1 Test de Moran pour variable surfacique continue

Il s'agit de tester l'hypothèse d'absence d'autocorrélation spatiale pour une variable surfacique continue X. L'hypothèse nulle est H_0 : "absence d'autocorrélation spatiale" et l'alternative est H_1 : "présence d'autocorrélation spatiale". Cette spécification est trop vague pour construire un test et il est nécéssaire de faire des hypothèses plus précises pour H_0 de façon à avoir une statistique de test de distribution connue. Il existe deux modèles classiques pour cela.
Dans le modèle dit "free sampling", on suppose que sous H_0, $X_1, \cdots, X_n$ sont indépendantes et identiquement distribuées de loi $\mathcal{N}(0, \sigma^2)$. Ceci conduit au test, dit "test gaussien", qui teste en réalité si l'échantillon observé est représentatif de la distribution d'un vecteur gaussien de composantes i.i.d. La statistique de test est l'indice de Moran I associé à une matrice de voisinage W choisie (ce test dépend donc de ce choix). La loi de I sous H_0 ne peut pas être exprimée analytiquement et on utilise donc la loi asymptotique de I sous H_0. Pour cela, on a besoin de normaliser d'abord l'indice en lui

enlevant sa moyenne et en le divisant par son écart-type. Le calcul de cette moyenne et cet écart-type du I de Moran utilise le **Théorème de Pitman et Koopmans**. On obtient

$$\mathbb{E}(I) = -\frac{1}{n-1},$$

et

$$\mathbb{E}(I^2) = \frac{n^2 S_1 - nS_2 + 3S_0^2}{(n^2-1)S_0^2},$$

où les quantités dépendent de la matrice de voisinage W

$$S_0 = \sum_{i \neq j} w_{ij}, S_1 = \frac{1}{2}\sum_{i \neq j}(w_{ij} + w_{ji})^2, S_2 = \sum_{i \neq j}(w_{i+} + w_{+i})^2,$$

avec

$$w_{i+} = \sum_j w_{ij}, \quad w_{+j} = \sum_j w_{ji}.$$

On utilise alors la loi asymptotique $\mathcal{N}(0,1)$ de l'indice normalisé pour calculer une p-valeur associée.
Dans le modèle dit "non free sampling" ou modèle de randomisation, on suppose que conditionnellement aux observations $X_i = x_i$, en l'absence d'autocorrélation spatiale les $n!$ permutations des réalisations $x_1, \cdots, x_n$ sont équiprobables. Ceci conduit à l'aide de la statistique de Moran au test, dit "test de permutation", qui teste si l'échantillon observé est représentatif d'une allocation aléatoire uniforme des valeurs $x_1, \cdots, x_n$ aux n sites de la carte. On peut également calculer les moments de I sous cette hypothèse nulle et la moyenne est la même que pour le modèle "free sampling" mais la formule de la variance est plus compliquée.
Le choix entre "free sampling" et "non free sampling" peut être guidé par le contexte mais notons que si X suit une loi F inconnue de variance finie, on a toujours la même espérance pour l'indice de Moran et le moment d'ordre deux vérifie $\mathbb{E}(I^2) = \mathbb{E}(\mathbb{E}_R(I^2))$, où $\mathbb{E}_R$ désigne l'espérance sous l'hypothèse de randomisation.
Il existe également un test de Monte Carlo basé sur l'indice de Moran qui ne nécéssite pas le choix d'un modèle. En pratique, on tire au hasard T permutations des sites et pour chaque permutation on realloue les valeurs de la variable sur les sites permutés. On calcule les indices de Moran pour chacune de T permutations, leur minimum I_{min} et maximum I_{max}. On compare alors la valeur observée de l'indice de Moran avec l'intervalle $[I_{min}, I_{max}]$. On rejette H_0 si l'indice de Moran n'est pas dans cet intervalle. Le "pseudo-niveau de signification" empirique du test est égal à $(L+1)/(T+1)$ où L est le nombre de fois parmi les T permutations que l'indice de Moran recalculé dépasse la valeur observée sur l'échantillon. (le $+1$ vient du fait qu'on compte l'observation initiale ainsi que les T permutations).

6.2 Test de Moran pour variable surfacique qualitative

De même que pour les variables surfaciques continues, il y a deux modèles différents selon l'hypothèse nulle. Si X est qualitative avec k modalités, le modèle "free sampling" suppose un tirage aléatoire avec remise dans une population ayant k groupes de proportions $p_1, \cdots, p_k$ connues : les X_i sont alors indépendantes de loi multinomiale. En pratique, $p_1, \cdots, p_k$ doivent être estimées par les fréquences empiriques. Pour le modèle "non free sampling", on suppose un tirage aléatoire sans remise dans une population ayant k groupes d'effectifs connus $n_1, \cdots, n_k$: la loi du n-uplet $(X_1, \cdots, X_n)$ est la loi hypergéométrique conditionnelle aux effectifs de groupe observés.
Les statistiques utilisées pour construire le test sont les "join counts" et leurs moments sous l'hypothèse nulle sont connus dans le cas de variables dichotomiques.
Dans ce cas pour le modèle "free sampling", les X_i sont i.i.d. Bernouilli $\mathcal{B}(1,p)$. Les deux premiers moments sont

$$\mathbb{E}(BB) = \frac{1}{2} S_0 p^2$$

$$4\mathbb{V}ar(BB) = p^2(1-p)[S_1(1-p) + S_2 p]$$

$$\mathbb{E}(BW) = S_0 p(1-p)$$

$$4\mathbb{V}ar(BW) = [4S_1 p(1-p) + S_2 p(1-p)(1-4p(1-p))].$$

Pour le modèle "non free sampling", il y a $n_B = \sum_i X_i$ valeurs 1 et $n - n_B$ valeurs 0, et l'on fait un tirage sans remise.
Avec la notation $n^{(b)} = n(n-1)\cdots(n-b+1)$, on peut écrire les deux premiers moments et la variance asymptotique :

$$\mathbb{E}(BB) = \frac{S_0}{2} \frac{n_B^{(2)}}{n^{(2)}}$$

$$\begin{aligned} 4\mathbb{V}ar(BB) &= [S_1(\frac{n_B^{(2)}}{n^{(2)}} - 2\frac{n_B^{(3)}}{n^{(3)}} + \frac{n_B^{(4)}}{n^{(4)}}) \\ &+ S_2(\frac{n_B^{(3)}}{n^{(3)}} - \frac{n_B^{(4)}}{n^{(4)}}) + \frac{S_0^2 n_B^{(4)}}{n^{(4)}} - (\frac{S_0 n_B^{(2)}}{n^{(2)}})^2] \end{aligned}$$

$$4as\mathbb{V}ar(BB) = p^2(1-p)[S_1(1-p) + S_2 p - 4\frac{S_0^2 p}{n}]$$

6.3 Test d'autocorrélation pour variable ponctuelle continue

On peut tester l'absence d'autocorrélation spatiale d'une variable ponctuelle continue à l'aide du variogramme empirique avec une approche par simulations. De même que pour le test de Monte Carlo d'un indice de Moran, cela consiste à faire des permutations aléatoires des valeurs de la variable sur les sites et à recalculer le variogramme empirique sur chaque permutation. Si le variogramme empirique tombe dans 95 pour cent de l'étendue de ces variogrammes empiriques, alors on ne peut pas rejeter l'absence d'autocorrélation spatiale et on peut penser que la forme observée de la courbe, même si elle n'est pas plate, a pu être un effet du hasard.

6.4 Test d'autocorrélation des résidus d'un modèle de régression linéaire ordinaire pour variable surfacique continue

Il est intéressant de tester l'autocorrélation spatiale d'une variable mais dans une démarche de modélisation, on est fréquemment amené à tester l'autocorrélation spatiale de résidus dans un modèle linéaire ordinaire. En effet celui-ci servira de modèle de base et si une autocorrélation apparaît dans ses résidus, on s'orientera alors vers un modèle spatial. Il n'est cependant pas possible d'utiliser le même test de Moran que précédemment pour le cas d'une variable surfacique continue car même en l'absence d'autocorrélation spatiale des erreurs ϵ_i du modèle linéaire, les résidus estimés ne sont pas indépendants. On utilise comme statistique de test l'indice de Moran généralisé qui n'est autre que l'indice de Moran ordinaire appliqué aux résidus du modèle linéaire mais il faut ajuster les calculs de moments. Dans le cas $D = I_n$, on montre que sous l'hypothèse d'absence d'autocorrélation spatiale

$$\mathbb{E}(I) = -\frac{trA}{n-k},$$

où k est le nombre de colonnes de X et $A = (X'X)^{-1}X'WX$.

6.5 Tests d'homogénéité spatiale pour semis de points

On dit qu'un processus ponctuel vérifie l'hypothèse d'homogénéité spatiale (hypothèse CSR pour "complete spatial randomness") si c'est un processus de Poisson homogène. Cette hypothèse implique donc à la fois l'homogénéité de la répartition des points en moyenne mais aussi l'indépendance entre les observations dans des zones disjointes (une propriété d'ordre 1 et

une propriété d'ordre 2). Tester l'hypthèse CSR est la première étape dans la modélisation d'un processus ponctuel dans le sens où si le processus est Poisson homogène, il sera entièrement caractérisé par le réel λ du paragraphe 4.5.1. Si cela n'est pas le cas, c'est alors que le travail de modélisation peut commencer. Il existe de nombreux tests de CSR mais nous allons seulement développer deux approches.

6.5.1 Test basé sur les quadrats

Ce test très ancien consiste à diviser la fenêtre d'observation en m quadrats, c'est à dire en cellules rectangulaires ou carrées d'égale surface et à dénombrer les points du processus dans chaque cellule, notés $n_k, k = 1, \ldots, m$. Soit $\bar{n} = \frac{n}{m}$ le nombre moyen de points par cellule. Considérons alors la quantité suivante

$$I = \sum_{k=1}^{m} \frac{(n_k - \bar{n})^2}{(m-1)\bar{n}}.$$

I peut d'abord être interprété comme le rapport entre la variance empirique des effectifs n_k et leur moyenne (coefficient de variation). Les cellules étant de même surface, sous l'hypothèse CSR, les effectifs sont équidistribuées de loi de Poisson et comme la moyenne d'une loi de Poisson est égale à sa variance, I n'est autre que le ratio de deux estimateurs de la variance. Par ailleurs, conditionnellement au nombre total de points, $(m-1)I$ n'est autre que le χ^2 de Pearson d'ajustement de la série des effectifs des quadrats. Sous l'hypothèse CSR, la loi de $(m-1)I$ peut être approximée asymptotiquement par une loi de χ^2 à $m-1$ degrés de liberté. Lorsque cet indice est significativement grand et que l'homogénéité est respectée, il denote une tendance à l'agrégation, c'est à dire une dépendance entre les points de type attraction. Inversement, lorsque cet indice est significativement petit et que l'homogénéité est respectée, il traduit une tendance à la régularité, c'est à dire une dépendance entre les points de type répulsion.

6.5.2 Diagnostic basé sur des simulations

Une autre approche pour évaluer l'hypothèse CSR consiste à simuler M réalisations d'un processus de Poisson homogène et à calculer des caractéristiques du processus (fonctions F,G, K ou L) pour chaque simulation. On trace ensuite les enveloppes de ces courbes sur l'ensemble des simulations et on évalue si la caractéristique observée sur l'échantillon entre ou non dans ces enveloppes.

Chapitre 7

Modèles de régression spatiale

Le contexte général des modèles de régression spatiale pour variables surfaciques est le suivant. On dispose d'une variable dépendante dont les mesures en n sites donnent un vecteur aléatoire Y (quantitatif, univarié). Les sites sont représentés par leur centroide s_i. On dispose également d'une variable indépendante dont les mesures en n sites donnent un vecteur aléatoire X (quantitatif , multivarié de dimension p), observé sur les mêmes zones. En général on suppose de plus que X et Y ont une distribution gaussienne. On verra que la modélisation de la tendance ne présente pas de spécificité technique dans les modèles spatiaux alors que celle de l'autocorrélation en présente. C'est cette structuration de l'autocorrélation prenant en compte le fait que celle-ci découle de la proximité relative des points dans un certain espace qui fait la force des modèles spatiaux.

7.1 Un catalogue de modèles de régression spatiale

On peut faire entrer la plupart de ces modèles dans le cadre suivant :

$$Y = \mu + \epsilon$$

avec $\mu = \mathbb{E}(Y \mid X)$ (d'où $\mathbb{E}(\epsilon) = 0$ et $X \perp Y$), $\mathbb{V}ar(Y) = V$.
Les données spatiales présentent souvent une hétéroscédasticité, c'est pourquoi dans un premier temps le modèle de base non-spatial WLS (pour "weighted least squares") qui nous servira d'étalon est

$$Y = X\beta + \epsilon \tag{7.1}$$

avec $\mathbb{E}(\epsilon) = 0, \mathbb{V}ar(\epsilon) = \sigma^2 D$, où D est une matrice diagonale, $D = I_n$ correspondant au modèle OLS.

La présence de D correspond à l' hétéroscédasticité. Par exemple, si T_i (resp : τ_i) est le taux de chomage observé (resp : théorique) dans la zone i et P_i est la population de la zone. Alors $var(T_i) = \frac{\tau_i(1-\tau_i)}{P_i}$ donc même si le taux de chomage est constant, il faut prendre des poids sur la diagonale de D proportionnels à $\frac{1}{P_i}$. Plus généralement, si la variable à expliquer est une proportion ou une moyenne (par exemple un taux de chomage, un montant de dépenses mensuel par ménage), il est naturel de penser qu'un ratio avec un dénominateur plus grand est moins sujet à variabilité que si ce dénominateur est faible. Plus précisément, on peut supposer que la variance est inversement proportionnelle au dénominateur, ce qui se justifie par le raisonnement suivant. S'il s'agit d'une moyenne empirique, prenons par exemple le montant de dépenses mensuel par ménage, on sait que la variance d'une moyenne empirique est de la forme $\frac{\sigma^2}{n}$, où σ^2 est la variance de la variable sous-jacente, ici le montant de dépense d'un ménage, et n est la taille de la sous population sur laquelle cette moyenne est calculée, ici le nombre de ménages de la zone. Si l'on suppose que la variance σ^2 est homogène sur l'ensemble des zones, la différence de variance entre zones s'explique par la différence du nombre de ménages et l'on peut donc prendre la pondération inversement proportionnelle au nombre de ménages de la zone. S'il s'agit à présent d'une proportion, le raisonnement est le même (une proportion étant une moyenne de variables de Bernoulli) en supposant le paramètre de la Bernoulli homogène sur les zones, et dans le cas du taux de chomage, on peut donc prendre la pondération inversement proportionnelle à la population.
Rappellons également les formules usuelles relatives à l'estimation par maximum de vraisemblance du modèle WLS :

$$\begin{aligned}
\hat{\beta} &= (X'D^{-1}X)^{-1}X'D^{-1}Y \\
\mathbb{V}ar(\hat{\beta}) &= \sigma^2(X'D^{-1}X)^{-1} \\
\mathbb{V}ar(\hat{\epsilon}) &= \sigma^2 PDP', P = I_n - (X'D^{-1}X)^{-1}X'D^{-1} \\
\hat{\sigma}^2 &= \frac{(Y - X\hat{\beta})'D^{-1}(Y - X\hat{\beta})^2}{n-p}
\end{aligned}$$

Comme on l'a déja vu, un des problèmes de ce type de données est que l'on dispose en général d'une seule réalisation, c'est à dire de l'observation du couple (X, Y) en n sites. Sans autre restriction sur ce modèle, on a n observations pour estimer $n + \frac{n(n+1)}{2}$ paramètres d'où la nécéssité de réduire le nombre de paramètres.
Une première restriction consiste à exprimer la tendance μ comme une fonction des coordonnées géographiques ou de régresseurs (avec éventuellement des régresseurs spatialement décalés), ou encore une combinaison des deux. Les autres restrictions vont porter sur la modélisation de la structure de covariance V.

Etant donnée une matrice de voisinage W normalisée et une variable Z, la variable spatialement décalée WZ présente automatiquement une autocorrélation spatiale avec Z. La famille des modèles spatiaux simultanés consiste à introduire une telle variable dans le modèle non spatial OLS ou WLS à divers endroits de l'équation (7.1). On obtient ainsi les modèles suivants :

- introduire WX en explicative dans le modèle WLS conduit au modèle SLX, (en anglais "spatially lagged-X model")
- introduire WY dans le membre de droite du modèle WLS conduit au modèle LAG (pour "lagged-Y model")
- introduire WX dans le modèle LAG conduit au modèle SDM (pour "Spatial Durbin")
- utiliser le modèle LAG pour le terme d'erreur conduit au modèle SEM (pour "Spatial Error model")
- combiner les modèles LAG et SEM models conduit au modèle général SAC
- introduire $W\epsilon$ dans le modèle WLS conduit au modèle MA (pour "moving average")
- combiner les modèles LAG et MA conduit au modèle SARMA.

Nous nous concentrerons dans la suite sur les modèles de base LAG, SDM et SEM. Mais dans un premier temps écrivons le descriptif de chacun des modèles ci dessus en s'efforçant de comparer les différentes modélisations de la tendance μ et de la variance V dans chacun d'eux.

7.1.1 Le modèle SLX

Une première façon simple d'introduire de l'interaction entre unités spatiales est d'introduire une variable spatialement décalée parmi les explicatives :

$$Y = X\beta + WZ\delta + \epsilon,$$

où comme précedemment ϵ est centré de matrice de variance-covariance diagonale $\mathbb{V}ar(\epsilon) = \sigma^2 D$, la diagonale de D contenant la pondération. L'observation Y pour une unité spatiale donnée est donc ainsi expliquée par la valeur de X pour cette unité et par la moyenne des valeurs de Z pour les unités voisines. Par exemple, la production d'une région peut être expliquée par la disponibilité du travail et par le montant du capital public dans les zones voisines. L'ajustement de ce modèle peut se faire par moindres carrés ordinaires (OLS). Attention : si W est normalisée, il ne faut pas que la constante apparaisse à la fois dans X et dans Z sous peine de non identifiabilité. Z peut être égale ou différente de X. On obtient pour μ et V :

$$\mu = X\beta + WZ\delta$$

et

$$V = \sigma^2 D.$$

7.1.2 Le modèle LAG

Le modèle LAG consiste à prendre en compte pour expliquer la valeur de Y sur une unité spatiale donnée non seulement les explicatives X mais aussi la moyenne de Y dans les zones voisines ce qui conduit à

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

où ϵ est un bruit blanc spatial, WY est la variable endogène spatialement décalée, $(I - \rho W)Y$ est la variable endogène spatialement filtrée. Le paramètre ρ est lié à l'autocorrélation spatiale présente dans Y.
Si la matrice $(I - \rho W)$ est non singulière, le modèle prends la forme réduite suivante

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

On obtient alors aisément pour μ et V :

$$\mu = (I - \rho W)^{-1} X\beta$$

$$\mathbb{V}ar(Y) = \sigma^2\{(I - \rho W')(I - \rho W)\}^{-1}.$$

Notons que cette formule de variance implique la présence d'hétéroscédasticité même si les erreurs ϵ sont homoscédastiques.

7.1.3 Le modèle SDM

En ajoutant au modèle LAG une variable explicative spatialement décalée on obtient le modèle SDM

$$Y = \rho WY + X\beta + WZ\delta + \epsilon,$$

où ϵ est un bruit blanc spatial. Sa forme réduite s'écrit

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

Les expressions suivantes en découlent pour μ et V :

$$\mu = (I - \rho W)^{-1}(X\beta + WZ\delta)$$

et

$$\mathbb{V}ar(Y) = \sigma^2\{(I - \rho W')(I - \rho W)\}^{-1}.$$

Notons que lorsque le paramètre ρ est nul, le modèle SDM devient un modèle SLX.

7.1.4 Le modèle SEM

Le modèle SEM introduit l' autocorrélation spatiale dans le processus des erreurs

$$Y = X\beta + \epsilon \tag{7.2}$$
$$\epsilon = \lambda W\epsilon + U, \tag{7.3}$$

où U est un bruit blanc spatial. Le paramètre λ est lié à l'intensité de l'autocorrélation spatiale présente dans les erreurs résiduelles.
On peut écrire ce modèle de façon équivalente :

$$(I - \lambda W)Y = (I - \lambda W)X\beta + U.$$

Si la matrice $(I - \lambda W)$ est non singulière, ce modèle admet la forme réduite suivante

$$Y = X\beta + (I - \lambda W)^{-1}U$$

On en déduit aisément l'expression de μ et V :

$$\mu = X\beta$$

$$\mathbb{V}ar(Y) = \sigma^2\{(I - \lambda W')(I - \lambda W)\}^{-1}.$$

Comme pour le modèle LAG, cette variance implique une hétéroscédasticité automatique même si les erreurs ϵ sont homoscédastiques.

7.1.5 Le modèle SAC

En combinant les modèles LAG et SEM, on obtient le modèle SAC

$$Y = \rho W_1 Y + X\beta + \epsilon$$
$$\epsilon = \lambda W_2\epsilon + U,$$

Si la matrice $(I - \lambda W)$ est non singulière, ce modèle admet la forme réduite suivante

$$Y = (I - \rho W_1)^{-1}X\beta + (I - \rho W_1)^{-1}(I - \lambda W_2)^{-1}U$$

On en déduit aisément l'expression de μ et V :

$$\mu = (I - \rho W_1)^{-1}X\beta$$

et

$$V = [(I - \lambda W_1')(I - \rho W_2')(I - \lambda W_2)(I - \rho W_1)]^{-1}$$

7.1.6 Le modèle SARMA

On peut comme en séries temporelles construire un modèle moyenne mobiles en faisant intervenir à droite de l'équation de régression les erreurs spatialement décalées, c'est à dire $W\epsilon$. Le modèle MA s'écrit

$Y_i = \mu + \lambda \sum_{i=1}^{n} w_{ij}\epsilon_j + \epsilon_i$ où ϵ est un bruit blanc spatial, $\mathbb{E}(\epsilon) = 0, \mathbb{V}ar(\epsilon) = \sigma^2 D$ (D matrice diagonale). Alors on a $V = \sigma^2(I_n + \lambda W)D(I_n + \lambda W)'$.

En combinant ce modèle avec un modèle LAG, on obtient

$$Y = \rho W_1 Y + X\beta + \epsilon$$
$$\epsilon = (I_n + \lambda W_2)u,$$

où ϵ est un bruit blanc spatial.

On obtient alors

$$\mu = (I_n - \rho W_1)^{-1} X\beta$$

et

$$V = \sigma^2(I_n - \rho W_1)^{-1}(I_n + \lambda W_2)D(I_n + \lambda W_2)'(I_n - \rho W_1)^{-1'}.$$

7.2 Maximum de vraisemblance dans les modèles SAR

Nous allons développer la méthode du maximum de vraisemblance pour l'estimation des coefficients dans les modèles de la famille SAR. Notons cependant qu'il existe d'autres méthodes telles la méthode 2SLS (moindres carrés en deux étapes) ou la méthode GMM (méthode des moments généralisés que nous ne verrons pas dans ce cours).

7.2.1 Conditions sur les coefficients

Dans la famille des modèles simultanés autorégressifs SAR, on a vu que la condition de non singularité de la matrice filtre $I - \rho W$ est omniprésente. Cette condition va impliquer des contraintes sur les coefficients, ρ dans le modèle LAG et λ dans le modèle SEM. Soient ω_{min} et ω_{max} respectivement les valeurs propres plus faible et plus grande de la matrice de voisinage W (celles-ci peuvent être complexes si W n'est pas semblable à une matrice symétrique).

Si W est symétrique, les conditions

$$\frac{1}{\omega_{min}} < \rho < \frac{1}{\omega_{max}},$$

sont suffisantes pour la non-singularité de $I - \rho W$. Notons que comme (trace$(W) = 0$, on a que $\omega_{min} < 0$ et $\omega_{max} > 0$). Si W est normalisée, alors $\omega_{max} = 1$ et $\rho \in [0, 1[$ est une condition suffisante pour la non-singularité de $I - \rho W$.

7.2.2 Maximum de vraisemblance dans le modèle LAG

Rappelons que pour une matrice de voisinage normalisée donnée, et pour une variable Y donnée, le vecteur WY appellé variable spatialement décalée associée à Y, représente la moyenne des observations sur les unités spatiales voisines au sens de W. Il est donc naturel de penser que la valeur de Y peut dépendre de celle de ses voisins WY. Si l'on centre Y pour éliminer la constante, on peut imaginer le modèle suivant

$$Y = \rho WY + \epsilon,$$

où le paramètre ρ mesure l'influence moyenne des voisins sur une unité spatiale ou encore l'intensité de l'interaction entre Y et ses voisins. ϵ contient la variabilité de Y non expliquée par le voisinage et sera modélisé ici par une variable de coordonnées i.i.d. Pour modéliser la moyenne de Y, on peut naturellement envisager aussi de rajouter à ce modèle des variables explicatives

$$Y = \rho WY + X\beta + \epsilon$$

WY est la variable endogène décalée et $(I - \rho W)Y$ la variable endogène filtrée. Notons que si la matrice $(I - \rho W)$ est non singulière, ce modèle admet l'écriture équivalente suivante

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

On a donc l'expression suivante pour la moyenne et la variance

$$(I - \rho W)^{-1} X\beta,$$

$$\mathbb{V}ar(Y) = \sigma^2\{(I - \rho W')(I - \rho W)\}^{-1}.$$

Notons que cette variance implique une hétéroscédasticité même dans le cas où les erreurs sont homoscédastiques.
Il y a dans ce modèle des contraintes sur le paramètre ρ qui sont

$$\frac{1}{\lambda_{min}} < \rho < \frac{1}{\lambda_{max}},$$

où λ_{min} et λ_{max} représentent la plus petite et la plus grande valeur propre de la matrice de voisinage W.
On montre aisément que les estimateurs des moindres carrés ordinaires sont biaisés dans ce modèle et c'est pourquoi on doit recourir au maximum de vraisemblance. Sous l'hypothèse de normalité des erreurs $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$, la vraisemblance dans ce modèle s'écrit

$$L(y \mid \rho, \sigma^2) = \frac{1}{2\pi\sigma^n} \det(I - \rho W) \exp\{\frac{1}{2\sigma^2}(y - X\beta)'(I - \rho W)'(I - \rho W)(y - X\beta)\},$$

d'où la log-vraisemblance

$$\log L(y \mid \rho, \sigma^2) = -\log(2\pi) - n\log(\sigma) + \log(det((I-\rho W)) - \frac{1}{2\sigma^2}(y-X\beta)'(I-\rho W)'(I-\rho W)(y-X\beta).$$

Si l'on dérive par rapport à σ, β et ρ, on peut obtenir l' expression suivante de $\hat{\sigma}$ et $\hat{\beta}$ en fonction de $\hat{\rho}$

$$\hat{\sigma}^2(\rho) = \frac{1}{n}(y - X\hat{\beta}(\rho))'(I - \rho W)'(I - \rho W)(y - X\hat{\beta}(\rho)),$$

et

$$\hat{\beta}(\rho) = (X'X)^{-1}X'(I - \rho W)Y.$$

Lorsqu'on reporte ces expressions dans le log-vraisemblance, on obtient ce qui s'appelle la **log-vraisemblance concentrée** qu'il reste à minimiser par rapport à ρ et qui vaut à constante près

$$\log L(y \mid \rho) = \log(det((I - \rho W)) - \frac{n}{2}\log(y - \rho Wy)'(y - \rho Wy).$$

Cette vraisemblance concentrée doit être optimisée numériquement et le problème principal est celui de l'évaluation du terme en log déterminant $\log(det((I - \rho W))$ qui peut être couteux lorsque le nombre de sites devient grand : il faut alors recourir à des approximations de ce terme.
calcul du biais du beta chapeau OLS dans ce modèle (asymp sb au voisinage de zero)

7.2.3 Maximum de vraisemblance dans le modèle SEM

Considérons à présent un modèle SEM

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

où U est une variable de coordonnées i.i.d. Le paramètre λ mesure l'intensité de l'autocorrélation spatiale entre les résidus.
Notons que si la matrice $(I - \lambda W)$ est non singulière, ce modèle admet les écritures équivalentes suivantes

$$Y = X\beta + (I - \lambda W)^{-1}U$$

ou encore

$$(I - \lambda W)Y = (I - \lambda W)X\beta + U.$$

On a donc l'expression suivante pour la variance

$$\mathbb{V}ar(Y) = \sigma^2\{(I - \lambda W')(I - \lambda W)\}^{-1}.$$

Notons que cette variance implique une hétéroscédasticité (les éléments de la diagonale ne sont pas constants) même dans le cas où les erreurs U sont homoscédastiques.
Il y a dans ce modèle des contraintes sur le paramètre λ qui sont

$$\frac{1}{\lambda_{min}} < \lambda < \frac{1}{\lambda_{max}},$$

où λ_{min} et λ_{max} représentent la plus petite et la plus grande valeur propre de la matrice de voisinage W.
Si l'on pose $A = I - \lambda W$, on a alors $Y = X\beta + A^{-1}\epsilon$ et $\epsilon = A(Y - X\beta)$.
Sous l'hypothèse de normalité des erreurs $U \sim \mathcal{N}(0, \sigma^2 I)$, la vraisemblance de Y s'écrit alors :

$$\begin{aligned} f_Y(y) &= f_\epsilon(\epsilon) \mid det(\frac{\partial \epsilon}{\partial Y}) \mid \\ &= f_\epsilon(\epsilon) det(A) \\ &= \frac{1}{(\sigma\sqrt{2\pi})^n} \exp(-\frac{\parallel \epsilon \parallel^2}{\sigma^2}) \mid det(\frac{\partial \epsilon}{\partial Y}) \mid \end{aligned}$$

7.3 Interprétation des coefficients

Dans un modèle linéaire ordinaire $Y = X\beta + \epsilon$, les dérivées des coordonnées de Y par rapport à celles de X sont données par $\frac{\partial y_i}{\partial x_{ik}} = \beta_k$, pour tout i et k et $\frac{\partial y_i}{\partial x_{jk}} = 0$, pour tout k et $j \neq i$.
β_k s' interprète classiquement comme l'accroissement de $\mathbb{E}(Y)$ quand la k-ème variable explicative augmente d'une unité toutes choses égales par ailleurs. Le modèle SEM se comporte exactement de la même façon.
Par contre, dans le modèle LAG, ce n'est plus le cas et un changement de la variable explicative dans une unité spatiale peut se répercuter sur les Y de toutes les autres unités. L'écriture de LAG par composante est $y_i = \sum_{t=1}^{p} S_t(W)_{it} x_t + \tilde{\epsilon}_i$, où p est le nombre de variables explicatives, x_j est la j-ème colonne de la matrice X et $\tilde{\epsilon} = (I - \rho W)^{-1}\epsilon$.
Alors, les dérivées partielles de $\mathbb{E}(y_i)$ par rapport à x_{jt} sont

$$\frac{\partial \mathbb{E}(y_i)}{\partial x_{jt}} = S_t(W)_{ij}.$$

On remarque d'abord que la dérivée croisée de la i-ème composante $\mathbb{E}(y_i)$ par rapport à x_{jt} pour $j \neq i$ n'est plus nulle mais égale à $S_t(W)_{ii}$.
On en déduit qu'un changement sur l'une des variables explicatives pour l'individu i va affecter non seulement y_i mais aussi tous les y_j.

De plus, l'effet sur $\mathbb{E}(y_i)$ de l'accroissement d'une unité de la j-ème composante de la variable explicative x_{jt} n'est plus constant sur les i. On définit alors trois mesures résumant ces effets.
L'**impact direct moyen** $\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \mathbb{E}(y_i)}{\partial x_{it}}$ mesure l'effet moyen sur chaque composante de $\mathbb{E}(Y)$ de l'accroissement d'une unité de x_{it} pour l'individu i et la variable t.
L'**impact indirect moyen** ou "spillover" $\frac{1}{n}\sum_{i\neq j}\frac{\partial \mathbb{E}(y_i)}{\partial x_{jt}}$ mesure l'effet moyen sur chaque composante de $\mathbb{E}(Y)$ de l'accroissement d'une unité de x_{jt} pour tous les individus $j \neq i$ et la variable t.
L'**impact moyen total** , égal à $\frac{1}{n}\sum_{i,j}\frac{\partial \mathbb{E}(y_i)}{\partial x_{jt}}$, est la somme de l'impact direct moyen et de l'impact indirect moyen et mesure l'effet moyen de l'accroissement de x_t d'une unité pour tous les individus.

7.4 Le modèle conditionnel autorégressif CAR

Ce modèle, issu de la littérature géostatistique, est aussi utilisé pour des variables surfaciques. Contrairement aux autres modèles décrits précedemment dits simultanés, ce modèle est défini par une contrainte de type markovien sur la loi conditionnelle de Y_i sachant la valeur de Y pour les autres sites

$$Y_i \mid Y_1, \cdots, Y_{i-1}, Y_{i+1}, \cdots, Y_n \sim \mathcal{N}(\mu_i + \sum_{j=1}^{n} c_{ij}(Y_j - \mu_j), \tau_i^2),$$

où

- $C = (c_{ij})$ et $D = diag(\tau_1^2, \cdots, \tau_n^2)$ doivent satisfaire les deux conditions $D^{-1}C$ symétrique et $D^{-1}(I - C)$ définie positive.
- μ s'exprime par une combinaison linéaire d'explicatives $\mu = X\beta$

De façon équivalente dans le cas gaussien on peut écrire

$$Y \sim \mathcal{N}(X\beta, \tau^2(I - C)^{-1}D)$$

Pour le modèle CAR à un paramètre CAR(1) $C = \rho W$ avec W matrice de voisinage, la variance s'écrit alors $V = \tau^2(I_n - \rho W)^{-1}D$.
En faisant une hypothèse gaussienne, on peut écrire le modèle SAR

$$Y \sim \mathcal{N}((I - \rho W)^{-1}X\beta, \sigma^2\{(I - \rho W')(I - \rho W)\}^{-1})$$

et le modèle CAR

$$Y \sim \mathcal{N}(X\beta, \tau^2(I - C)^{-1})$$

d'où la même structure de covariance en posant $C = \rho(W + W') - \rho^2 WW'$ et $\sigma = \tau$ mais des moyennes modélisées de façon différente.
L'estimation des paramètres de ce modèle se fait par maximum de vraisemblance. Ecrivons la spécification CAR model composante par composante

$$Y_i \mid Y_1, \cdots, Y_{i-1}, Y_{i+1}, \cdots, Y_n \sim \mathcal{N}(\mu_i + \sum_{j=1}^{n} c_{ij}(Y_j - \mu_j), \tau_i^2)$$

En supposant les variances conditionnelles connues à un facteur près $\tau_i^2 = \tau^2\phi_i$ et en posant $\Phi = diag(\phi_1, \cdots, \phi_n)$, alors $D = \tau^2\Phi$.
On introduit la transformation $\tilde{Y} = \Phi^{-1/2}Y$, $\tilde{X} = \Phi^{-1/2}X$, et $\tilde{C} = \Phi^{-1/2}C\Phi^{1/2}$, de sorte que le modèle se simplifie en $\tilde{Y} \sim \mathcal{N}(\tilde{X}\beta, \tau^2(I - \tilde{C})^{-1})$
Les estimateurs du maximum de vraisemblance de (β, τ, ρ) peuvent alors se calculer en maximisant la vraisemblance des données transformées

$$LL = -\frac{n}{2}\log(2\pi\tau^2) - \log\det(I - \rho W)^{-1/2} - \frac{1}{2\tau^2}(\tilde{Y} - \tilde{X}\beta)'(I - \tilde{C})(\tilde{Y} - \tilde{X}\beta)$$

Première étape : pour le modèle CAR(1), $C = \rho W$ et $\tilde{C} = \rho\Phi^{-1/2}W\Phi^{1/2}$.
A ρ fixé, la maximisation de LL par rapport à (β, τ) est explicite

$$\hat{\beta}(\rho) = (\tilde{X}'(I - \tilde{C})\tilde{X})^{-1}\tilde{X}'(I - \tilde{C})\tilde{Y} \tag{7.4}$$
$$\hat{\tau}^2(\rho) = (\tilde{Y} - \tilde{X}\beta)'(I - \tilde{C})(\tilde{Y} - \tilde{X}\beta)/n \tag{7.5}$$

Deuxième étape : on substitue ces valeurs dans LL pour obtenir la log-vraisemblance dite concentrée

$$\begin{aligned} -LL(\rho) &= \frac{n}{2}(\log(2\pi) + 1) + \log\det(I - \rho W)^{-1/2} + \\ &\quad \frac{n}{2}\log\{\tilde{Y}(I - \tilde{C})\{I - \tilde{X}(\tilde{X}'(I - \tilde{C})\tilde{X})^{-1}\tilde{X}'(I - \tilde{C})\}\tilde{Y}/n\} \end{aligned}$$

La maximisation de $LL(\rho)$ est faite de façon numérique. Comme pour les modèles simultanés, la difficulté réside dans l'évaluation du terme $\log\det(I - \rho W)^{-1/2}$. Les conditions sur D et C impliquent des restrictions sur ρ qui sont les mêmes que pour le modèle LAG. Une fois $\hat{\rho}$ estimé, on obtient facilement $\hat{\beta}$ et $\hat{\tau}^2$ en insérant (7.4) et (7.5).

7.5 Modélisation géostatistique

Dans la littérature géostatistique, l'approche classique consiste à modéliser tendance et fluctuation en deux étapes. On commence par ajuster une tendance par exemple en ajustant des polynômes des coordonnées spatiales ou d'autres variables explicatives. On retranche ensuite celle-ci pour obtenir une fluctuation estimée. On ajuste ensuite un variogramme (comme on l'a vu au paragraphe 4.1) à la fluctuation. L'approche "model-based geostatistics" développée par Diggle et Ribeiro (2007), permet d'étendre ces modèles aux cas où la distribution de Y n'est plus gaussienne, et propose une estimation par maximum de vraisemblance en une seule étape.

7.6 Approximation du terme en log-déterminant

Il existe plusieurs méthodes :

– une approximation par troncature de série (Martin, 1992)

$$\log\det(I-\rho W) = \mathrm{t}r(\log\det(I-\rho W)) = -\sum_{i=1}^{\infty}\rho^i\frac{\mathrm{t}r(W^i)}{i}$$

– Pace and Barry (1997, 1999) font une approximation par Monte Carlo en utilisant la décomposition de Cholevsky pour profiter de la lacunarité de W
– Pace and LeSage (2004) utilisent une approximation de Chebyshev
– Cressie, Perrin and Thomas-Agnan (2005) proposent une approche par simulation dans le contexte des modèles CAR.

7.7 Les méthodes MWR et GWR

La régression par fenêtre glissante MWR et la régression geographiquement pondérée GWR sont des méthodes d'estimation locales. L'idée est de choisir une fenêtre centrée sur le point d'intérêt et d'utiliser seulement les observations qui sont dans cette fenêtre pour estimer la régression au point d'intérêt. En se sens il s'agit de méthodes localement linéaires où le local se mesure dans l'espace géographique et non dans l'espace des régresseurs.
Dans l'esprit des méthodes non paramétriques, GWR utilise une pondération avec une fonction noyau qui a pour effet de faire décroitre l'influence d'un voisin donné en fonction de sa distance au point d'intérêt. On peut aussi utiliser une fenêtre adapatative dans la fonction noyau pour éliminer les effets de la densité locale de points. C'est le cas par exemple pour la pondération suivante $w_{ij} = (1-\frac{d_{ij}^2}{d^2})^2$ si j est l'un des voisins de i et 0 sinon où d est la distance de i à ses k plus proches voisins (k est choisi par validation croisée) ce qui assure que chaque fenêtre contient le même nombre d'observations.
Notons que MWR et GWR **ne modélisent que l'hétérogénéité spatiale** et non l'autocorrélation.

7.8 Tests de spécification, comparaison de modèles

7.8.1 Autocorrélation des résidus

Outre le test de Moran des résidus d'une régression OLS ou WLS, il existe d'autres tests comme les tests du multiplicateur de Lagrange donnés par les formules

$$LM(err) = \{e'We/\sigma^2\}^2/tr(W'W+W^2)$$

dans le cas du modèle SEM en alternative et

$$LM(lag) = \{e'WY/\sigma^2\}^2/\{(WXb)'MWXb/\sigma^2 + tr(W'W+W^2)\},$$

dans le cas du modèle LAG en alternative, où e désignent les résidus et W une matrice de voisinage. Sous l'hypothèse d'absence d'autocorrélation, $LM(err)$ et $LM(lag)$ suivent asymptotiquement une loi de $\chi^2(1)$. On peut aussi tester l'hypothèse $H_0 : \lambda = \rho = 0$ dans le modèle complet par le test du multiplicateur de Lagrange et la statistique obtenue SARMA converge alors vers un $\chi^2(2)$.
IL existe des versions dites robustes de ces tests. La statistique $RLM(lag)$ permet de tester le modèle SEM en hypothèse nulle contre le modèle complet SAC en alternative. De même, la statistique $RLM(err)$ permet de tester le modèle LAG en hypothèse nulle contre le modèle complet SAC en alternative.
Voyons comment utiliser les tests de Lagrange pour orienter le choix de modèle. Si par exemple, les deux tests LM(err) et LM(lag) sont significatifs, mais lorsqu'on regarde les versions robustes, c'est le test RLM(lag) qui est le plus significatif : on se dirige alors vers un modèle LAG.

7.8.2 Tests sur les coefficients

Notons qu'il existe dans ces modèles une expression de la matrice de variance-covariance asymptotique des estimateurs des coefficients. On va s'intéresser aux tests d' hypothèses de la forme $H_0 : g(\theta) = 0$, où θ est le vecteur des paramètres. Par exemple $\theta = (\rho, \beta, \sigma^2)$ dans le modèle LAG et si l'on s'intéresse à l'hypothèse $H_0 : \rho = 0$, cela revient à tester le modèle OLS contre le modèle LAG.
Les trois types de tests classiques peuvent être utilisés pour cela : le test de Wald Wa (ou asymptotic t-test), le test du rapport de vraisemblance LR et le test du score ou de Lagrange LM. Si g est une contrainte de dimension q, les trois statistiques de test correspondantes suivent asymptotiquement une loi de $\chi^2(q)$ et les trois tests sont asymptotiquement équivalents. Notons que Wa et LR nécéssitent l'estimation du modèle non contraint alors que LM n'est fonction que de l'estimation du modèle contraint (souvent OLS).

7.8.3 Stratégies de choix de modèle

Une fois un modèle spatial LAG adopté, le test du multiplicateur de Lagrange $LM(err)^*$ permet de tester s'il est nécéssaire d'introduire également une autocorrélation spatiale des erreurs. De même, une fois un modèle spatial SEM adopté, le test du multiplicateur de Lagrange $LM(lag)^*$ permet de tester s'il est nécéssaire d'introduire également une variable endogène décalée. Notons qu'on ne peut utiliser un test du rapport de vraisemblance que si les modèles sont emboités.
Pour comparer des modèles non emboités, on peut aussi minimiser les critères usuels d'Akaiké et de Schwartz qui s'expriment en fonction de la log-vraisemblance

$$AIC = -2\log(L) + 2k, BIC = -2\log(L) + \log(nk),$$

où k est le nombre de paramètres.
Notons bien qu'il n'est pas légitime de faire un test de Moran pour tester l'autocorrélation spatiale des résidus d'une régression spatiale, mais on peut à titre descriptif faire un Moran plot de ces résidus. Pour finir, notons qu'une matrice de voisinage mal spécifiée peut engendrer de l'autocorrélation spatiale dans les résidus sans que pour autant le type de modèle soit à remettre en cause.

7.9 Prédiction dans les modèles spatiaux

7.9.1 Dans les modèles de la famille SAR

Dans un modèle non spatial ajusté par WLS, on calcule la prédiction de la variable Y avec la formule $\hat{y} = x\hat{\beta}$, que x soit un des points observés dans l'échantillon ou pas et cette prédiction correspond à la meilleure prédiction linéaire sans biais (BLUP).
Dans un modèle spatial, à cause de la présence d'autocorrélation spatiale, la meilleure prédiction linéaire sans biais ne se calcule plus ainsi et doit prendre en compte les autocorrélations (en particulier elle nécéssite le calcul des matrices de poids croisées correspondant à l'ensemble des points constitué des points de l'échantillon et des points où l'on veut prédire). Il faut donc se garder d'appliquer la formule $\hat{y} = x\hat{\beta}$.
Pour un point de l'échantillon, Bivand (2002) utilise la formule suivante

$$\hat{Y} = X\hat{\beta} + \hat{\rho}WY,$$

formule qui n'est qu'une approximation du BLUP.

7.9.2 Dans les modèles géostatistiques : le Krigeage

Dans le modèle géostatistique classique, le modèle $Y = \mu + \epsilon$ devient le modèle "signal plus bruit" suivant. Le signal est un champ X_s qui est l'objet d'intérêt sur lequel on veut faire de l'inférence et il présente lui-même une tendance $\mu = m(s) = \mathbb{E}(X_s)$ et une structure d'autocovariance donnée par la fonction $\sigma(s,t) = \mathbb{C}ov(X_s, X_t)$. Il est observé avec un bruit additif ϵ en un nombre fini de localisations $s_i, i = 1, \cdots, n$, d'où

$$Y_{s_i} = X_{s_i} + \epsilon_i,$$

où ϵ_i sont des réalisation d'un bruit i.i.d. de moyenne nulle et de variance σ^2.
L'objectif est de

- estimer la tendance $m(s) = \mathbb{E}(X_s)$,

– prédire les valeurs de X_s en une localisation s qui n'est pas nécéssairement parmi celles observées ou plus généralement prédire $\int_A w(s)X_s ds$.
– calculer des erreurs de prédiction

La méthode de Krigeage consiste à utiliser comme prédicteur celui qui possède la propriété d'optimalité suivante : il doit minimiser l'erreur quadratique de prédiction parmi les prédicteurs linéaires sans biais. On l'appelle le BLUP pour "Best Linear Unbiased Predictor". On suppose dans un premier temps que la structure de covariance σ est connue. En pratique il faut l'estimer au préalable. L'optimalité se traduit par les conditions suivantes. Ce prédicteur de X_s est une combinaison linéaire Y^* of $Y_i = Y_{s_i}$ satisfaisant

– Y^* est sans biais $\mathbb{E}(Y^*) = \mathbb{E}(X_s)$
– Y^* a la plus petite erreur de prédiction $\min \mathbb{E}(Y^* - X_s)^2$ parmi les prédicteurs linéaires sans biais.

Introduisons les notations suivantes

– Σ est la matrice de terme général $\sigma(s_i, s_j)$,
– Σ_s est le vecteur $\sigma(s_i, s)$
– Y est le vecteur $(Y_1, \cdots, Y_n)'$.

On utilise le résultat suivant : Σ est inversible dès que les localisations s_i sont distinctes et que la fonction d'autocovariance est strictement défini positive.

Krigeage simple - Cas du modèle sans bruit

Considérons d'abord le cas où il n'y a pas de bruit $\epsilon = 0$. De plus supposons d'abord que la tendance est constante et connue $m(s) = \mu$ et on parle alors de Krigeage simple ; on peut alors supposer la moyenne nulle. On montre alors que le BLUP est donné par

$$Y_s^* = \sum_{i=1}^n \lambda_i(s) Y_i$$

avec

$$\lambda^*(s) = \Sigma^{-1}\Sigma_s$$

L'erreur de prédiction est alors égale à

$$\mathbb{E}(Y_s^* - X_s)^2 = \sigma(s,s) - \lambda^*(s)'\Sigma_s.$$

En effet

$$\begin{aligned} \mathbb{E}\big(Y_s^* - X_s\big)^2 &= Var\big(Y_s^* - X_s\big) = Var\big(\sum_{i=1}^n \lambda_i(s)Y_i - X_s\big) \qquad (7.6) \\ &= \sum_{i=1}^n \sum_{j=1}^n \lambda_i(s)\lambda_j(s)\sigma(s_i, s_j) - 2\sum_{i=1}^n \lambda_i(s)\sigma(s_i, s) + \sigma(s,s). \end{aligned}$$

$$\frac{\partial}{\partial \lambda_i(s)} \mathbb{E}\big(Y_s^* - X_s\big)^2 = 2\sum_{j=1}^n \lambda_j(s)\sigma(s_i, s_j) - 2\sigma(s_i, s) = 0,$$

puisque

$$\frac{\partial^2}{\partial\lambda_i(s)\partial\lambda_j(s)}\mathbb{E}\big(Y_s^* - X_s\big)^2 = 2\sigma(s_i, s_j),$$

et donc la matrice hessienne est égale à 2Σ et donc définie positive ce qui assure que $\mathbb{E}\big(Z_s^* - Y_s\big)^2$ est convexe. Donc

$$\sum_{j=1}^{n} \lambda_j^*(s)\sigma(s_i, s_j) = \sigma(s_i, s) \qquad i = 1, \ldots, n$$

ou en notations matricielle

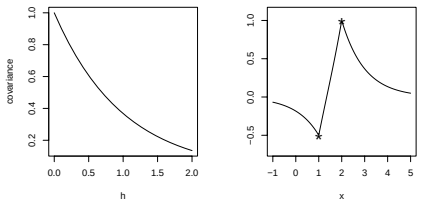
$$\Sigma\lambda^*(s) = \Sigma_s. \tag{7.7}$$

$$\lambda^*(s) = \Sigma^{-1}\Sigma_s.$$

L'erreur de prédiction se calcule alors par

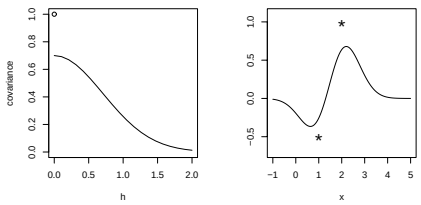
$$\begin{aligned}
\mathbb{E}\big(Y_s^* - X_s\big)^2 &= \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i^*(s)\lambda_j^*(s)\sigma(s_i, s_j) - 2\sum_{i=1}^{n} \lambda_i^*(s)\sigma(s_i, s) + \sigma(s, s) \\
&= \lambda^*(s)'\Sigma\lambda^*(s) - 2\lambda^*(s)\Sigma_s + \sigma(s, s) \\
&= \lambda^*(s)'\Sigma_s - 2\lambda^*(s)'\Sigma_s + \sigma(s, s) \\
&= \sigma(s, s) - \lambda^*(s)'\Sigma_s.
\end{aligned}$$

Ce prédicteur interpole les valeurs du champ dans le sens suivant $Y_i^* = Y_i$. Cela semble logique au vu du fait que le bruit est nul. Notons que ce prédicteur est une moyenne pondérée des valeurs observées. Le vecteur Σ_s a pour effet que le point s_i contribue d'autant plus à la prédiction de X_s que s_i est proche de s (proche impliquant plus corrélé). La matrice Σ^{-1} a pour effet que les points isolés sont pondérés plus fortement que les points agrégés en groupes. On remarque que $\lambda^*(s) = \Sigma^{-1}\Sigma_s$ doit être calculé pour chaque localisation s pour laquelle on souhaite une prédiction mais dans la formule de prédiction $Y_s^* = \lambda^*(s)'Y = \Sigma_s'\Sigma^{-1}Z$, le produit $\Sigma^{-1}Z$ est commun à toutes les prédictions et n'a besoin d'être calculé qu'une fois. La figure suivante montre un exemple de krigeage en dimension un avec un variogramme exponentiel et un modèle non bruité.



Krigeage simple - Cas du modèle bruité

Des argumens similaires aux précédents montrent que la solution est données par les mêmes formules mais en remplaçant $\sigma(s_i, s_j)$ par $\sigma(s_i, s_j) + \sigma^2\delta_0(s_i - s_j)$, et en laissant $\sigma(s_i, s)$ inchangé. Si σ^2 est non nul, ce prédicteur n'interpole plus les valeurs observées ce qui est intuitivement logique. La figure suivante montre un exemple de krigeage en dimension un avec un variogramme gausien et un modèle bruité.



Krigeage ordinaire - Cas du modèle non bruité

On parle de Krigeage ordinaire loesque la tendance $m(s) = \mu$ est toujours constante mais inconnue. Le champ X_s est supposé intrinsèquement stationnaire de vartiogramme γ.
Le prédicteur BLUP est alors donné par

$$Y_s^* = \sum_{i=1}^{n} \lambda_i(s) Y_i$$

avec les coefficients $\lambda^*(s)$ solution du système linéaire

$$\begin{aligned} \sum_j \lambda_j(s)\sigma(s_i - s_j) + \mu &= \sigma(s, s_i) \\ \sum_i \lambda_i(s) &= 1 \end{aligned} \tag{7.8}$$

pour $i = 1, \cdots n$ et $l = 0, \cdots L$.
L'erreur de prédiction est égale à

$$E(Y_s^* - X_s)^2 = \sigma(s, s) - \sum_i \lambda_i(s)\sigma(s_i, s) - \mu.$$

On peut aussi écrire ce système en fonction du variogramme (au lieu de la covariance)

$$\begin{aligned} \sum_j \lambda_j(s)\gamma(s_i - s_j) + \mu &= \gamma(s, s_i) \\ \sum_i \lambda_i(s) &= 1 \end{aligned} \tag{7.9}$$

et l'erreur de prédiction est alors

$$E(Y_s^* - X_s)^2 = \sum_i \lambda_i(s)\gamma(s_i, 0) + \mu.$$

En effet

$$\begin{aligned} & Var\big(Y_s^* - X_s\big) = Var\big(\sum_{i=1}^{n} \lambda_i(s) Y_i - X_s\big) \\ = & \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i(s)\lambda_j(s)\sigma(s_i, s_j) - 2\sum_{i=1}^{n} \lambda_i(s)\sigma(s_i, s) + \sigma(s, s) \\ = & \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i(s)\lambda_j(s)\Big(\frac{\sigma(s_i, s_i) + \sigma(s_j, s_j) - \gamma(s_i, s_j)}{2}\Big) \\ & -2\sum_{i=1}^{n} \lambda_i(s)\Big(\frac{\sigma(s_i, s_i) + \sigma(s, s) - \gamma(s_i, s)}{2}\Big) + \sigma(s, s) \end{aligned}$$

$$
\begin{aligned}
& Var\big(Y_s^* - X_s\big) \\
= & \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i(s)\lambda_j(s)\left[\frac{\sigma(s_i,s_i)+\sigma(s_j,s_j)}{2} - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i(s)\lambda_j(s)\gamma(s_i,s_j)\right] \\
- & \sum_{i=1}^{n}\lambda_i(s)\sigma(s_i,s_i) + \sum_{i=1}^{n}\lambda_i(s)\gamma(s_i,s) + \Big(1 - \sum_{i=1}^{n}\lambda_i(s)\Big)\sigma(s,s) \\
= & -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i(s)\lambda_j(s)\gamma(s_i,s_j) + \sum_{i=1}^{n}\lambda_i(s)\gamma(s_i,s).
\end{aligned}
$$

Ce modèle peut être généralisé au cas du Krigeage dit universel lorsque la tendance $m(s)$ n'est plus constante mais s'écrit comme combinaison linéaire de polynômes des variables spatiales à coefficients inconnus et lorsque la structure de covariance est intrinsèquement stationnaire d'ordre r.

7.10 Modèles de régression pour semis de points

La première étape de la modélisation d'un semis de points est le test de CSR. En effet, si celui-ci n'est pas significatif, on pourra adopter un modèle Poissonnien facile à manipuler alors que dans le cas contraire, il faudra se tourner vers des modèles avec interaction. Dans ce dernier cas, le choix d'un modèle spécifique se fait en utilisant quelques outils exploratoires comme la fonction de corrélation des paires. La modélisation de l'intensité prends toujours la forme suivante

$$\lambda(x) = \exp(\sum_{j=1}^{k}\theta_k Z_k(x)),$$

où Z_k sont des facteurs explicatifs et θ_k les paramètres correspondants.
Il reste ensuite à estimer les paramètres du modèle : ceux de l'intensité et ceux de la structure d'interaction. La méthode du maximum de vraisemblance est difficile à appliquer sauf dans quelques modèles particuliers en raison de la présence de la constante de normalisation difficile à évaluer. Une approche possible, similaire à la classique méthode des moments est de choisir une caractéristique du semis comme par exemple la fonction K de Ripley et de faire des moindres carrés entre la fonction théorique K qui dépends des paramètres et la fonction empirique correspondante pour déterminer les meilleurs paramètres. Une autre approche consiste à approximer la constante de normalisation par des méthodes de Monte Carlo. Enfin, une autre possibilité est d'utiliser le pseudo-maximum de vraisemblance qui consiste à remplacer la vraisemblance par le produit des densités conditionnelles.

La dernière étape consiste enfin en la validation du modèle et cela se fait généralement par la méthode des enveloppes basée sur des simulations. Cette méthode consiste d'abord à simuler un grand nombre M de réalisations du modèle ajusté et à déterminer si la fonction K de Ripley (version inhomogène) estimée tombe dans l'enveloppe des fonctions K associées aux M simulations du modèle supposé. Dans la pratique on utilise souvent $M = 19$ ou 99. Si pour une valeur de la distance r, dans l'étendue des valeurs observées dans la fenêtre, la courbe K observée sort de l'enveloppe, le modèle est rejeté. Le pseudo niveau de signification empirique associé à ce test est de $\frac{1}{M+1}$ ce qui conduit à 0.05 pour $M = 10$ et à 0.01 pour $M = 99$.

Remerciements. Ce document, qui est la trame d'un ouvrage, a été effectué en collaboration avec Thibault Laurent et Anne Ruiz-Gazen de l'université Toulouse 1 Capitole, Toulouse School of Economics.

Bibliographie

[1] R. Bivand, E.J. Pebesma, and V. Gómez-Rubio (2008) Applied Spatial Data Analysis with R, Series : Use R, Springer.

[2] R. Bivand (2002), Spatial econometrics functions in R : Classes and methods, Journal of geographical systems, 4(4), 405-421.

[3] N. Cressie (1993), Statistics for spatial data, Wiley, New-York.

[4] N. Cressie, Perrin and Thomas-Agnan (2005), Likelihood based estimation for gaussian MRF's, Statistical Methodology, Elsevier.

[5] N. Cressie, DM. Hawkins, (1980), Robust estimation of the variogram, Mathematical Geology, Springer.

[6] N. Cressie and C.K. Wilke (2011) Statistics for spatio-temporal data, Wiley, New-York.

[7] P.J. Diggle (1985), A kernel based method for smoothing point process data, Applies Statistics, 34, 138-147.

[8] P.J. Diggle (2003), Statistical Analysis of spatial point patterns, Oxford University Press.

[9] P.J. Diggle and P.J. Ribeiro (2007), Model based geostatistics, Springer Series in Statistics.

[10] J. Illian, A. Penttinen, H. Stoyan and D. Stoyan (2009), Statistical analysis and modelling of spatial point patterns, Wiley.

[11] T. Laurent , A. Ruiz-Gazen et C. Thomas-Agnan, (2012), GeoXp : An R Package for Exploratory Spatial Data Analysis, Journal Statistical Software, 47 (2).
Vol. 47, Issue 2, Apr 2012

[12] J. P. LeSage, R. K. Pace (2009), Introduction to spatial econometrics, CRC Press.

[13] Matheron (1962), Traité de Géostatistique appliquée, Technip, Paris.

[14] Martin (1992), Approximations to the determinant term in gaussian maximum likelihood estimation of some spatial models, Communications in Statistics, Theory and methods, 22(1), 189-205.

[15] K. Pace and R. Barry (1999) A Monte Carlo estimator of the log determinant of large sparse matrices, Linear Algebra and its applications.

[16] K. Pace and J. LeSage (2004), Chebyshev approximation of log determinants of spatial weight matrices, Computational Statistics and Data Analysis, 45(2), 179-196, Elsevier.