# Méthodes et outils pour l'analyse de données de mobilité

# **MEMO-F-403** : Preparatory work for the master thesis

Youri Hubaut

Année académique : 2016 - 2017

29 mai 2017

# Table des matières

# Chapitre 1

# Introduction

## 1.1 Plan

Dans ce travail, nous nous attarderons sur les méthodes et outils pour l'analyse de données liées à la mobilité. Nous commencerons par établir une définition de ce que nous entendons pour la notion de mobilité. Nous continuerons en donnant un bref aperçu des problématiques liées à ce genre de sujet ainsi que le contexte académique. Ensuite, nous aborderons une partie, plus théorique, en proposant un état de l'art qui tentera d'aborder de nombreux aspects liés à ce domaine : le temps, l'espace, leur union, les pratiques et études sur le sujet, les jeux de données disponibles ainsi que les bibliothèques développées afin de répondre à ces problématiques. Enfin, nous terminerons par tenter de définir le travail prévu pour l'année prochaine et le mémoire.

## 1.2 Mobilité

> *Caractère de ce qui peut être déplacé ou de ce qui se déplace par rapport à un lieu, à une position.*

Cette définition fournie par l'Académie française révèle un aspect capital, celui du mouvement. Parfois connotée péjorativement à l'instar de son usage dans les expressions : "personnes à mobilité réduite" ou "mobilité professionnelle". Il dénote néanmoins la notion de changement, d'évolution, qui est une conséquence du temps qui s'écoule inexorablement. Dans le cadre de ce travail, nous restreindrons sa polysémie à un usage quasi-unique, celle de la "mobilité spatiale". Ce terme peut alors tant désigner la circulation de personnes, de biens que d'idées.

De tout temps, les hommes se sont mis à voyager afin d'assouvir leur soif de connaissances ou de conquêtes. Et les progrès techniques, issus des révolutions technologiques, ont permis une réelle explosion dans la communication et les distances pouvant être accomplies. Conquérir le ciel ou la lune n'est plus une rêve et on cherche dès lors à améliorer les moyens existants en vue de perfectionner leur efficacité. Tous ces déplacements sont motivés par la réalisation d'activités (travail, études, affaires personnelles, loisirs), qui impliquent le plus souvent des contacts avec d'autres personnes dans l'optique de transformer nos sociétés.

Les études de mobilité sont aujourd'hui motivées tant par leur côté prévisionnel que dans leur analyse révélatrice des pratiques de nos sociétés où la division du travail s'amplifie et où l'interdépendance envers les uns et les autres ne fait que croître. Elles

permettent une meilleure compréhension des attitudes humaines. L'espace public et son parcours peuvent être conditionnés par l'évolution qu'on souhaite lui faire prendre. Code de la route, passages piétons sonores, aménagement de la voirie pour les handicapés sont tant de mesures qui formalisent nos valeurs ; comprendre cet aspect de nos vies permet avant tout de mieux appréhender le monde dans lequel nous évoluons et de mieux percevoir ce que nous désirons en faire.

## 1.3 Problématique

Les premières pas dans ce domaine sont souvent associées aux travaux d'Hägerstrand de 1968 [1] avec le concept de *diffusionisme* où, pour la première fois, des données géographiques et temporelles sont interprétées sous un même œil. Il a initié le mouvement de l'étude des rythmes journaliers des déplacements humains et de leur impact sur la vie en ville [2, 3]. Plusieurs approches pour l'analyse de l'activité humaine ont été proposées [4], toutes dans le but de mieux comprendre la qualité de vie et de mieux organiser les services publiques ou l'accessibilité. On peut mesurer les déplacements aux travers de plusieurs indicateurs ; on pense notamment aux questionnaires, compteurs de véhicules ou vidéo-contrôle. Avec la démocratisation des téléphones portables et des applications connectées, toutes ces données se sont multipliées et ne demandent qu'à être analysées [5].

Les dynamiques des populations et leur répartition sont des éléments capitaux pour la bonne compréhension de nombreux phénomènes. Les interactions des espèces, leur stabilité et leur diversité ont permis le développement d'une diversité biologique extraordinaire. Il est capital d'être en mesure de pouvoir analyser et prévoir les déplacements afin de prendre les mesures adéquates face à des maladies infectieuses par exemple [6]. Le futur des villes ou les prévisions routières sont également des problématiques importantes en vue de mieux gérer nos sociétés. Or notre compréhension des lois fondamentales régissant les déplacements humains demeure limitée malgré l'apparition de nombreuses études sur la régularité et l'aléatoire de nos trajets [7].

## 1.4 Contexte

Ce travail s'inscrit dans le cadre du mémoire de master, aboutissement des cinq années d'études à l'Université Libre de Bruxelles. Il vise à mettre en évidence les capacités de l'étudiant à employer les connaissances et méthodes acquises durant son enseignement. L'étudiant est évalué sur la qualité de sa présentation, de sa bibliographie, de son rapport et des contacts entretenus avec son superviseur.

Indépendamment, il existe à Bruxelles le consortium *Brussels MOBIlity-Advanced Indicators Dashboard (MOBI-AID)* [8], fusion du *Machine Learning Group*, ULB [9] et du *MOBI Research Group*, VUB [10], et qui vise à concevoir et construire un système de surveillance des performances au moyen d'un tableau de bord des indicateurs de la mobilité. Ceci permettrait de mieux comprendre les dynamiques de la région de Bruxelles-Capitale, de soutenir prises de décisions pour les autorités locales ainsi que de permettre à Bruxelles d'être reconnue comme modèle de Smart City. De même, un étudiant, actuellement en troisième année de bachelier, M. Romain à implémenter un outil permettant de visualiser l'utilisation des vélos en agglomération bruxelloise [11].

# Nomenclature

La liste suivante décrit certains symboles qui seront employés par la suite :

| | |
|---|---|
| $Y$ | Processus (le plus souvent inconnu) |
| $Z$ | Observations |
| $X$ | Majuscule, vecteur aléatoire |
| $x$ | Minuscule, un élément d'un vecteur aléatoire |
| $\epsilon$ | Erreur (généralement une gaussienne centrée) |
| $d$ | Nombre de dimensions |
| $n$ | Nombre d'observations |
| $\cdot_i$ | Variable indicée correspond à l'élément $i$ du vecteur aléatoire |
| $\cdot_{i:j}$ | Toutes les variables dont l'indice est compris entre $i$ et $j$ |
| $\cdot_{-i}$ | Variable indicée par $-i$ correspond à l'ensemble des éléments de la variable aléatoire excepté l'élément $i$ |
| $\hat{\cdot}$ | Variable surmontée d'un accent circonflexe, estimateur pour cette variable |
| $D$ | Domaine, sous ensemble fini ou non, spatial si indicé par $s$ et temporel par $t$ |
| $A$ | Aire |
| $s$ | Une position |
| $t$ | Un temps |
| $h$ | Une distance entre deux points |
| $\mathbb{E}(.)$ ou $\mu$ | Espérance globale de la variable |
| $\mu(.)$ | Espérance locale de la variable |
| $\mathbb{P}(.)$ | Probabilité associée à la loi |
| $var(.)$ | Variance de la variable |
| $cov(.,.)$ | Covariance entre les deux variables |
| $C_Y$ | Matrice de covariance, obtenue en faisant la covariance entre chacune des valeurs du processus $Y$ |
| $C_Y(h)$ | Covariance pour le processus $Y$ pour la distance $h$ |
| $C_Y(.,.)$ | Composante $(i,j)$ de la matrice de covariance pour le processus $Y$ |

$C_Y^{(t/s)}$ Matrice de covariance par rapport au temps/à l'espace pour $Y$

$\gamma_Y(A, B)$ Semi-variogramme, moitié de la variance de $A - B$ pour le processus $Y$

$MSE(\hat{\theta})$ Erreur quadratique moyenne, définie par $\mathbb{E}[(\hat{\theta} - \theta)^2]$

$MSPE(\hat{\theta})$ Erreur quadratique moyenne pour la prédiction

$W$ Bruit blanc

$S$ Ensemble des endroits

$T_i$ Ensemble des observations par rapport au temps pour un endroit $i$

$N(.)$ Voisins de l'élément

$\lambda(.)$ Fonction d'intensité

# Chapitre 2

# État de l'Art

## 2.1 Introduction

L'étude des données spatio-temporelles a connu un réel essor ces dernières années. En effet, avec les volumes croissants de données urbanistiques recensées et rendues disponibles, des nouvelles opportunités se sont créées pour l'analyse de données avec pour idée d'améliorer la vie des citoyens au travers de prises de décisions politiques basées sur des faits. Il s'agit donc essentiellement de méthodes de régression et de prédictions.

Cet état de l'art est inspiré par le livre *Statistics for spatio-temporal data*, écrit par Noel Cressie et Christopher K. Wikle [12]. Ceux-ci proposent une vision très étendue de tous les aspects statistiques qui rentrent en compte tant dans les processus de modélisation que dans les techniques employées dans ce vaste champ de l'informatique. Ainsi que de *Handbook of spatial Statistics* de Alan E. Gelfand, Peter J. Diggle, Montserrat Fuentes et Peter Guttorp [13] qui offrent un point de vue plus pratique aux travers des techniques et études sur le sujet.

## 2.2 Spatial processes

### 2.2.1 Introduction

Dans les problèmes liés à l'espace, les données sont rarement indépendantes entre-elles et, donc, certaines techniques standards qui font appels à l'indépendance des variables deviennent caduques dans ce domaine [14]. La dépendance spatiale se traduit par une covariation des propriétés, les caractéristiques et propriétés des lieux proches ont tendance à être plus corrélés tant positivement que négativement que des points fort éloignés. Ce phénomène d'auto-corrélation (non-indépendance des variables) viole des conditions d'application de techniques statistiques classiques [15].

Très naturellement, afin de mieux percevoir les effets locaux, on définit des notions de proximités (d'auto-corrélation spatiale) tant au niveau des distances (*Ripley's K/L-function*) [16] que des individus (*Moran's I global & Geary's C local*) [17].

Il y a trois branches majeures dans l'analyse spatiale statistique [13] :

— variation continue (*continuous spatial variation*)[2.2.2], axée sur la prédiction ;

— variation discrète (*discrete spatial variation*)[2.2.4], axée sur l'inférence et les interactions entre les données ;

— processus ponctuels (*spatial point processes*)[2.2.5], axée sur l'inférence et la positions des points par eux-mêmes ;

On attribue à Julian Besag des modèles et méthodes d'inférence pour analyser des données discrètes (sur des *lattices*) [18]. À Brian Ripley, des approches systématiques pour des processus à point spatial [19]. Les variations spatiales continues sont quant à elles plus anciennes et peuvent être associées aux travaux de Danie G. Krige [20] et Georges Matheron [21].

Le modèle géostatistique le plus classique (ou le plus général) décompose le processus aléatoire comme suit :

$$Z(s) = Y(s) + \epsilon(s) \tag{2.1}$$

avec $Y(s)$ le processus stochastique en fonction de la position $s$, $Z(s)$ l'observation liée et $\epsilon(s)$ l'erreur (de mesure, par exemple) associée.

### 2.2.2 Variation continue

Lorsqu'on aborde la notion d'espace, il est normal de penser tout d'abord à des variables continues a l'instar de la géométrie différentielle et son analyse vectorielle. Naturellement, plusieurs modèles de probabilité liés aux espaces (ayant $d$ dimensions) ont fait leur apparition et bon nombre considèrent une définition de processus stochastiques très génériques :

$$\{Y(s) : s \in D_s \subset \mathbb{R}^d\} \tag{2.2}$$

avec l'idée qu'à chaque position est associée une probabilité $\omega \in \Omega$. Ce processus est valide si la distribution jointe est bien définie et consistante (par permutation des éléments et marginalisation notamment) [22].

En outre, il est important de faire mention de deux définitions :

— On qualifie un processus stochastique de "strictement stationnaire" si les lois de probabilités sont invariantes par translation. Les probabilités jointes ne sont pas modifiées si on déplace toutes les positions par un vecteur $h$.

— Et de "stationnaire du second-ordre (ou faible)" si le processus satisfait ces deux propriétés :

$$\mathbb{E}(Y(s)) = \mathbb{E}(Y(s+h)) = \mu \tag{2.3}$$

et

$$cov(Y(s), Y(s+h)) = cov(Y(0), Y(h)) = C_Y(h) \tag{2.4}$$

où $C_Y(h)$ est la fonction de covariance. La matrice de covariance est souvent inconnue et l'exprimer purement en terme de distance permet de diminuer le nombre de dimensions afin de plus facilement pouvoir l'estimer.

Remarquons que la stationnarité du second ordre implique celle stricte pour les modèles gaussiens. D'autre part, sous l'hypothèse du second ordre, Mathéron a proposé la définition d'une fonction ($f : \mathbb{R}^+ \to \mathbb{R}$) dénommée semi-variogramme [23], plus connue sous sa forme forte comme le *semi-variogramme de Matérn* :

$$\gamma_Y(h) = \frac{1}{2} var(Y(s+h) - Y(s)) = C_Y(0) - C_Y(h) \tag{2.5}$$

Les variogrammes sont une classe de fonctions très importantes dans la théorie du *kriging*[2.2.3], elles sont définies par la manière dont les observations évoluent entre deux positions et sont donc par nature aléatoire ; les semi-variogrammes sont ces fonctions divisées par deux. Ils mesurent le degré de dépendance spatiale. Dans le cas stationnaire, les variogrammes ne dépendent que de la distance $h$ alors que les fonctions de covariance ne le deviennent qu'à partir du second ordre. De plus, l'estimation du variogramme n'est pas biaisée par la moyenne, au contraire de la covariance. Toutefois, tout variogramme respectant la condition de matrice définie non-positive peut être employé [24].

Un variogramme peut être qualifié d'"isotropique" s'il peut être écrit sous la forme d'une fonction de $\|h\|$ et lorsqu'on prend en considération un modèle muni d'un norme $L^2$ (distance euclidienne), les variogrammes peuvent montrer une discontinuité près de l'origine, on parle alors de *nugget effect* [12]. Sa qualité peut être mesurée par le biais des techniques liées au *Goodness of Fit* [25]. Lorsque le modèle n'est pas isotropique, on peut l'approcher par un isotropique associé à une transformation linéaire.

### 2.2.3 Kriging

Le *kriging* (ou krigeage) est une méthode d'estimation linéaire optimale non-biaisé qui minimise la variance ($BLUE$), sous les hypothèses de Gauss-Markov. Il s'agit d'une méthode des moindres carrés ordinaire ($OLS$). Il tient compte non seulement de la distance entre les données et le point d'estimation, mais également des distances entre les données deux-à-deux. L'idée fondamentale est de prédire la valeur de la fonction en un point comme une somme pondérée des valeurs connues et avoisinantes à celui-ci. Le *kriging* en lui-même est un nom générique pour une méthodologie afin de construire un estimateur ; en effet, plusieurs techniques existent en vue de satisfaire certains critères sous base de présomptions sur le modèle. Son principal avantage est qu'il permet d'obtenir l'erreur sur l'estimation [20].

$$\hat{Y}(s_0) = \sum_{i=1}^{n} l_i Z(s_i) + k = l'Z + k \tag{2.6}$$

où $s_0 \in D_s$ est la position de la prédiction et $s_i$ sont les observations (au nombre de $n$). Tout en minimisant la variance de la prédiction (*Mean Square Prediction Error*) :

$$MSPE(l,k) = var(Y(s_0) - l'Z - k) + \mathbb{E}(Y(s_0) - \hat{Y}(s_0))^2 \tag{2.7}$$

— Le *simple kriging* est sans doute le plus facile mathématiquement. Il fait l'hypothèse d'une moyenne constante, estimée par la moyenne de l'échantillon des données, et d'une covariance connue. Il peut être interpréter comme la tendance générale des données corrigée par la covariance locale des données.

$$\hat{Y}(s_0) = \mu_Y(s_0) + cov(Y(s_0), Z)' C_Z^{-1}(Z - \mu_Y) \tag{2.8}$$

Avec pour variance :

$$\sigma^2_{Y,simp.krig.}(s_0) = MSPE(\hat{l'}, \hat{k}) = C_Y(s_0, s_0) - cov(Y(s_0), Z)' C_Z^{-1} cov(Y(s_0), Z) \tag{2.9}$$

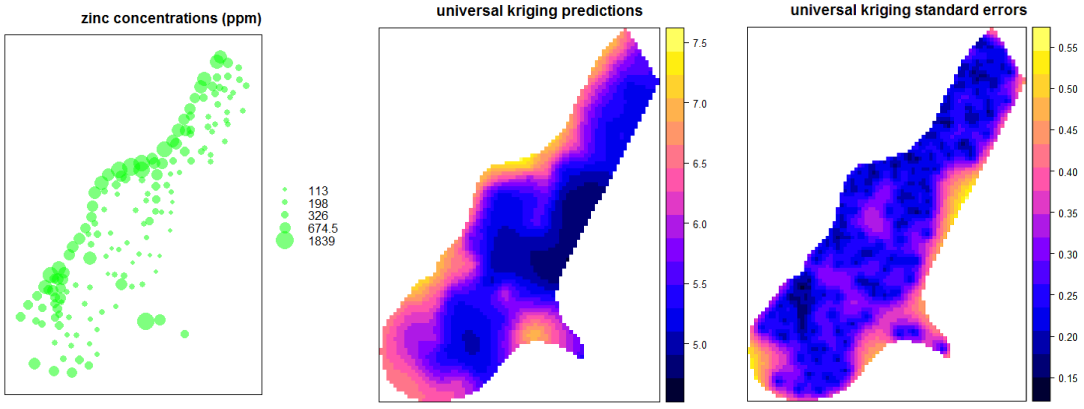
— L'*ordinary kriging* remplace la moyenne de l'échantillon par l'estimation des moindres carrés de celle-ci.

— L'*universal kriging* substitue la moyenne constante par un modèle de régression ; il est le plus employé mais introduit beaucoup de difficultés techniques mathématiquement et se décline en de nombreuses formes pour des cas spéciaux (valeurs binaires, valeurs discrètes, ...) [23].

Il existe des modèles connexes qui ne prennent pas en compte l'hypothèse de stationnarité (indépendance entre l'observation et le lieu considéré) suite aux résultats des tests de *Moran's I* ou de "Breusch-Pagan" [26]. Les deux autres grandes familles de modèles sont les *Spatial autoregressive model* (SAR) à l'instar des séries temporelles[2.3.2] et *Geographically Weighted Regression* (GWR régression locale sur un modèle) qui prennent en compte les phénomènes spatiaux [13].

Exemple de *kriging* en R, où on essaye de prédire la position de gisements de zinc le long de la Meuse en fonction de sites connus et de leur concentration :

```
library(sp)
data(meuse) # Dataset containing the positions of known zinc deposits

# Set the spatial coordinates -> it means (x, y)
coordinates(meuse) = ~x+y
# We load grid to make our predictions
data(meuse.grid)
gridded(meuse.grid) = ~x+y

# Plot our graph
bubble(meuse, "zinc",
       col=c("#00ff0080", "#00ff0080"),
       main = "zinc concentrations (ppm)")

# Variogram - Formula (2.5)
m <- vgm(.40, # Asymptotic value
         "Sph", # Model type, we can define ours
         954, # Range
         .06) # Nugget

# Universal kriging - Formula (2.6)
x <- krige(log(zinc)~x+y, # zinc prediction in function of the position (the s_i)
           meuse, # Training = meuse data
           meuse.grid, # Predicting = meuse.grid data
           model = m, # Variogram used
           block = c(40, 40)) # Size of the blocks
spplot(x["var1.pred"], main = "universal kriging predictions")
x$var1.se = sqrt(x$var1.var) # Formula (2.9)
spplot(x["var1.se"], main = "universal kriging standard errors")
```




Librairies R associées : sp (classes et méthodes pour les données spatiales) [27], gstat (variogramme et *kriging*) [28], spacetime (axée données spatio-temporelles) [29]

ou geoR (outils standards d'analyse géostatistque) [30].

### 2.2.4 *Lattice* & MRF

L'hypothèse d'un modèle continu n'est pas toujours nécessaire, il est parfois suffisant de travailler avec un sous-ensemble fini, discret, un treillis (*lattice graph*). Typiquement, une grille composée de petits rectangles. Ce procédé nécessite d'introduire une notion de distance au travers d'une matrice souvent notée $W$ et dont l'entrée $(i, j)$ correspond à la distance entre $i$ et $j$.

Une modélisation statistique sur base d'une collection finies de variables est souvent envisagée sous la forme des champs aléatoires de Markov (*Markov Random Field*). L'idée était de généraliser le concept des processus de Markov à d'autres dimensions tout en incluant la notion de dépendance [18]. Un MRF est un moyen simple de représenter des distributions conditionnelles. Ceci permet de s'intéresser à une variable à la fois tout en simplifiant la simulation. Le concept clef est qu'au lieu de s'intéresser à l'instant $t$ connaissant les $t-1, t-2, ...$, comme pour les processus de Markov (ce qui nécessite de définir une notion d'ordre), on regarde les positions voisines comme "historique" de la fonction [31].

$$Y(s_i)|Y_{-i} = Y(s_i)|Y(N(s_i)) \quad \forall i \in D_s \subset \mathbb{R}^d \tag{2.10}$$

avec $N(s_i)$ l'ensemble des voisins du point $s_i$ et $_{-i}$, tous les points exceptés $i$ ; on cherche à exprimer $Y(s_i)$ connaissant son voisinage $(.|Y(N(s_i)))$.

Les probabilités conditionnelles définissent notre MRF, seulement nous préférerions trouver les probabilités jointes associées afin de pouvoir appliquer de puissants théorèmes (Hammersley–Clifford [32] ou le résultat des Geman [33]) afin de ne plus être contraint à un ensemble de points fixés. Kaiser & Cressie donnent une procédure afin de calculer ces dites probabilités sous des hypothèses suffisamment généralistes [34].

Quelques modèles existent en rapport avec ces notions employant des distributions diverses. Citons les modèles dits "auto Poisson", "auto logistique", "auto binomial" ou "auto Beta" [35, 36]. Mais le plus classique est sans doute l'"auto Gaussien".

Le modèle CAR (*conditional autoregressive*) est défini par une distribution gaussienne :

$$\mathbb{P}(y(s_i)|y(N(s_i))) = (2\pi\tau_i^2)^{-1/2} exp[-(y(s_i) - \theta_i)^2/2\tau_i^2)] \tag{2.11}$$

où $\tau_i^2$ est la variance conditionnelle et $\theta_i$ est interprété comme une moyenne, fonction du voisinage du point, corrigée par la variance locale :

$$\theta_i(y(N(s_i))) = E(Y(s_i) \mid y(N(s_i))) = \mu(s_i) + \sum_{s_j \in N(s_i)} c_{ij}(y(s_j) - \mu(s_j)) \tag{2.12}$$

Delà, on peut définir la matrice $C$ de covariance avec les coefficients $c_{ij} = c_{ji}$ et $M$ diagonale avec les valeurs de $\tau_i^2$, on obtient alors une expression symétrique, définie positive qui permet d'obtenir la distribution jointe gaussienne [18] :

$$Y \sim Gau(\mu, (I - C)^{-1}M) \tag{2.13}$$

CAR est plus approprié aux situations où la dépendance est du premier ordre, où l'auto-corrélation locale est forte tandis que SAR[2.3.2] est adapté au second ordre et à une auto-corrélation plus globale. Cependant, l'interprétation de ces modèles n'est pas forcément intuitive. Les modèles SAR sont bien estimés par *maximum likelihood*

mais absolument pas par MCMC (lié à la difficulté d'introduire des effets aléatoires au contraire du modèle CAR) [37].

Librairies R associées : CARBayes (modèles CAR) [38], spdep (permet de définir des métriques et d'utiliser SAR) [39] ou spBayes (modèles bayésiens pour les problèmes spatio-temporels liés aux *lattice*) [40].

### 2.2.5 Spatial Point Process

Sous ce nom, on sous-entend un processus aléatoire dont les réalisations consistent en un ensemble fini ou dénombrable de points dans un espace. Un exemple serait la répartition des arbres dans une forêt. On tend à définir le processus $Z$ tel que $Z(A)$ dénote le nombre d'événements survenus dans une région $A \subset D_s \subset \mathbb{R}^d$ dite "Lebesgue mesurable" [41].

Intuitivement, la notion de Poisson semble bien correspondre à la description du problème, estimer le nombre d'évènements dans un intervalle. Seulement, on ne connaît pas le paramètre de la distribution associée à une région, qui serait une indication sur le nombre d'occurrences. Il faut donc créer une fonction "d'intensité" qui dépend de la position et qui représente le potentiel d'apparition en vue d'obtenir le nombre d'événements dans cet intervalle. On définit la fonction d'intensité du premier ordre comme suit (en relation avec le théorème de Campbell) [42] :

$$\lambda(s) = \lim_{|ds|\to 0} E(Z(ds)))/|ds| \quad \forall s \in D_s \sim E(Z(A)) = \int_A \lambda(s)ds \quad \forall A \subset D_s \tag{2.14}$$

Dans le contexte spatial, on parle souvent de processus de Cox qui est une généralisation de Poisson dans lequel la moyenne n'est pas constante mais varie avec l'espace ou le temps [43]. On peut étendre ce modèle en introduisant une notion analogue à la covariance, mais pour la fonction d'intensité, ainsi que l'équivalent des variogrammes (*K-functions*). Une autre famille très populaire est celle de Gibbs qui est équivalent à la notion de *Markov point process* et trouve son origine dans la physique. Le but est de modéliser un ensemble de points où certaines propriétés physiques (notamment d'attraction-répulsion) sont respectées. Mathématiquement, c'est loin d'être trivial, il faut créer une fonction de potentiel qui permet de minimiser les probabilités que deux points soient situées l'un à côté de l'autre. La simulation s'effectue par le biais d'un *MCMC* [44].

Librairies R associées : spatstat (incroyablement riche pour étudier les problèmes ponctuels) [45], lgcp (pour les processus de Cox) [46] ou PtProcess (analyse plus orientée sur l'aspect temporelle) [47].

## 2.3 Processus temporel

### 2.3.1 Introduction

Lorsqu'on se trouve face à des données récoltées, il n'est pas rare de définir et catégoriser les événements selon une notion d'ordre temporel. En effet, certaines observations sont fonction du temps et il peut être utile de voir si l'aspect temporel occupe une part importante dans l'analyse de ces données [48].

On définit les modèles temporaux de manière très générique :

$$\{Y(t) : t \in D_t\} \tag{2.15}$$

Traditionnellement, on décompose ces phénomènes en quatre éléments : les tendances (changements dans la moyenne à long terme), les variations saisonnières, les cycles (hebdomadaires ou pluriannuels) et les fluctuations irrégulières afin d'avoir une meilleure flexibilité au niveau de la modélisation. Il n'est pas rare non plus de considérer des phénomènes dits stationnaires où les premiers moments restent constants. Les définitions sont analogues aux processus spatiaux[2.2.2] mais le temps remplace la notion d'espace [49].

Pendant longtemps, on a employé des modèles purement déterministes liés aux équations aux différences ou différentielles, mais ceux-ci semblent mener, en pratique, à des résultats souvent inférieurs aux modèles stochastiques qui se basent sur les notions de bruit blanc ou de marche aléatoire (*random-walk*) [50].

### 2.3.2 Processus auto-régressif

Il n'est pas rare qu'une observation soit dépendante d'une ou plusieurs précédentes. On modélise ce genre de comportement par des processus ayant la propriété d'être auto-régressif. La taille de l'historique de la fonction définit son ordre.

$$AR(p) \equiv Y_t = \alpha_1 Y_{t-1} + ... + \alpha_p Y_{t-p} + \epsilon_t \tag{2.16}$$

avec $\epsilon_t$ est un processus de bruit blanc et $\alpha_k$ des constantes. Remarquons que le cas AR(1) correspond à la série géométrique (si $|a_k| < 1$). Seulement, il est souvent préférable de définir une valeur moyenne locale afin d'être en mesure de décrire l'erreur comme une combinaison de celle présente et passée. On a alors une fenêtre coulissante définissant une moyenne locale, le *moving average* [51] :

$$MA(q) \equiv Y_t = W_t + \beta_1 W_{t-1} + ... + +\beta_q \epsilon_{t-q} \tag{2.17}$$

avec $\epsilon_t$ des variables aléatoires (bruit blanc) et $\beta_k$ des constante. L'estimation de ces paramètres n'est pas aisée et on emploie souvent les équations de Yule-Walker et l'algorithme de Durbin-Levinson. Lorsqu'on combine ces deux notions, nous obtenons le modèle *ARMA* qui est à l'origine de bon nombre de variantes, notamment du *ARIMA* qui permet de gérer les cas non-stationnaires en effectuant récursivement la différence entre les données $t$ et $t + 1$ "$d$ fois" (terme d'"intégration" I(d)) [52]. Dans le cadre multivarié, ce modèle est étendu pour donner naissance à *VAR* et ses dérivés mais il peut parfois effectuer de l'overfitting et l'on préfèrera alors employé un *BVAR* [53].

Seulement, comment trouver les bons paramètres pour modéliser notre problème. Il faut d'abord se poser la question de la stationnarité par le biais du test de Dickey-Fuller. Ensuite, avons-nous affaire à un phénomène plutôt de type AR ou MA et quels sont leurs ordres ? AR donne des effets plus longs dans la durée parce que son auto-corrélation décline peu à peu tandis que MA en donne des brefs. On peut se baser sur le lag de l'auto-corrélation qui tend vers zéro au bout de $q$ étapes pour le MA(q). Le côté AR(p) est plus difficile, on se base sur la valeur à partir de laquelle elle devient négligeable au bout de $p$ étapes dans les *partial auto-correlation function*. Enfin, le terme d'intégration I(d) est évalué en regardant quand la tendance générale disparaît.

Bien sûr, on peut augmenter les paramètres petit à petit et étudier l'évolution du maximum de vraisemblance [50].

Lorsqu'on veut réaliser des prédictions, on peut employer les méthodes mentionnées ci-dessus, mais également, le *exponential smoothing* qui consiste simplement à appliquer une régression par des fonctions (exponentielles ou polynomiales), le fameux filtre de Kalman[2.4.6] ou encore des modèles bayésiens [51].

### 2.3.3 Représentation Spectrale

Très logiquement, il est normal de vouloir observer les phénomènes temporaux dans le domaine des fréquences. La notion d'auto-covariance est complétée par celle de périodogramme qui donne des indications sur les harmonies. Il ne reste plus qu'à calibrer les coefficients spectraux avec des techniques classiques (à l'instar de l'*exponential smoothing*). Un des grands avantages de cette représentation est qu'il devient facile de comparer deux séries temporelles afin de voir si elles sont corrélées. L'idée est de considérer une des séries comme l'entrée et l'autre comme la sortie pour trouver les propriétés du système linéaire qui pourrait y être associé [52].

Exemple de prédiction avec un modèle ARIMA, sur base du nombre de voyageurs prenant l'avion chaque mois :

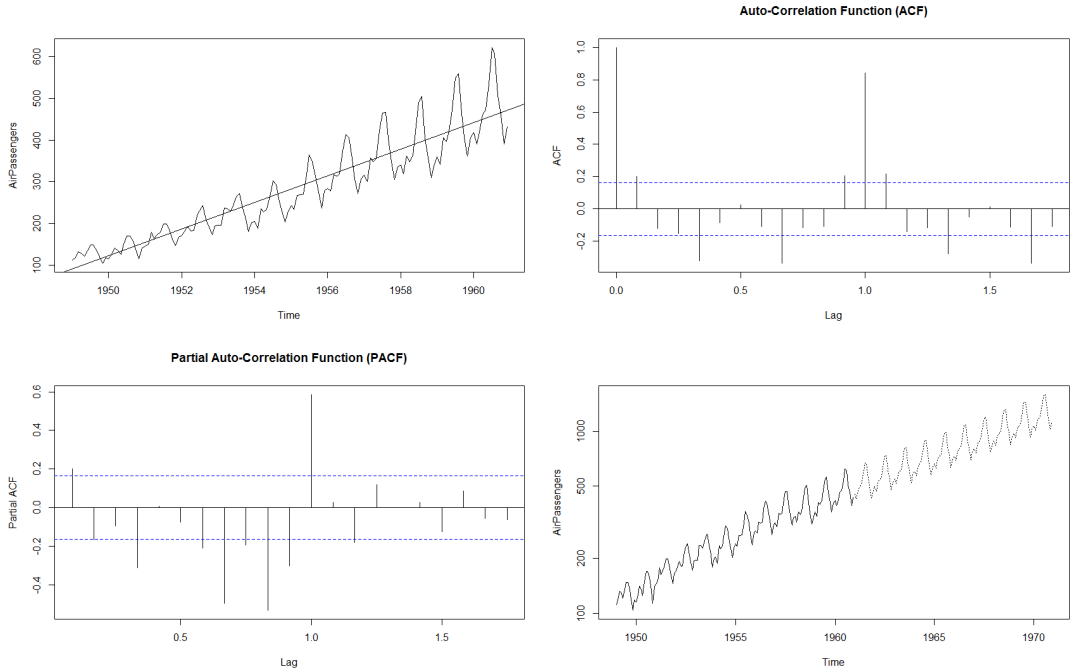
```
library(tseries)
data(AirPassengers) # Dataset number of passengers per year

plot(AirPassengers) # We plot our data
abline(reg = lm(AirPassengers ~ time(AirPassengers))) # We perform a linear regression

# Diff to remove the linear trend
adf.test(diff(log(AirPassengers)), alternative = "stationary", k = 0) # p-value = 0.01

acf(diff(log(AirPassengers)), main = "Auto-Correlation Function (ACF)")
pacf(diff(log(AirPassengers)), main = "Partial Auto-Correlation Function (PACF)")

fit <- arima(log(AirPassengers),
             c(0, 1, 1), # AR(0) - Formula (2.16), I(1), MA(1) - Formula (2.17)
             seasonal = list(order = c(0, 1, 1), period = 12)) # Yearly seasonal effects
pred <- predict(fit, n.ahead = 10 * 12) # 10 years later
ts.plot(AirPassengers,
        exp(1)^pred$pred, # Predictions
        log = "y", # Log scale for Y
        lty = c(1,3), # Plot solid, then dashed
        ylab = "AirPassengers")
```

Librairies R associées : timeSeries (modèles pour les séries temporelles financielles) [54], forecast (ensemble de modèles prédictifs) [55] ou AnomalyDetection (pour trouver des anomalies parmi les saisons) [56].

## 2.4 Processu spatio-temporel

### 2.4.1 Introduction

Les processus spatio-temporels combinent les notions liées aux sections précédentes afin de décrire un modèle sur de larges intervalles de temps et d'espace. Les objectifs sont nombreux : effectuer des prédictions dans le temps ou dans l'espace, inférer les comportements sur base des observations et tenter d'approcher ces processus par des modèles.




Illustrations des différentes notions pour la prédiction du nombre de passagers.

Un processus spatio-temporel est défini de manière très générale :

$$\{Y(s,t) : (s,t) \in D_s \times D_t \subset \mathbb{R}^d \times \mathbb{R}\} \tag{2.18}$$

avec l'idée sous-jacente que les valeurs sont dépendantes du voisinage et du passé.

Bien sûr, on peut envisager le problème sous la forme d'un système d'équations différentielles, mais ce modèle théorique est souvent inconnu et difficile à aborder [13].

### 2.4.2 Fonctions de covariance

Comme les solutions analytiques sont difficiles à appréhender, il vaut mieux décrire le processus par un modèle statistique :

$$Y(s,t) = \mu(s,t) + \beta(s) + \gamma(t) + \kappa(s,t) + \delta(s,t) \quad (s,t) \in D_s \times D_t \tag{2.19}$$

avec : $\mu(s,t)$ une moyenne déterministe, $\beta(s)$ la variation commune des endroits à travers le temps, $\gamma(t)$ du temps pour les positions, $\kappa(s,t)$ variation sur le modèle de $\mu$ et $\delta(s,t)$ le bruit (le *nugget effect*). De surcroît, on fait l'hypothèse de la description sur base des deux premiers moments [57].

A l'instar des processus spatiaux[2.2.2], les notions de covariance jouent un rôle clef. Une propriété importante est le fait que la fonction soit définie non-négative, ce qui permet d'appliquer le théorème de Bochner [58] et de parler de son spectre [59]. De même, il est intéressant d'avoir une certaine stationnarité tant spatiale que temporelle. Et on parle de fonction de covariance spatio-temporelle séparable lorsqu'on peut séparer les deux notions de covariance comme suit :

$$cov(Y(s,t), Y(x,r)) = C^{(s)}(s,x).C^{(t)}(t,r) \tag{2.20}$$

Cette hypothèse de séparabilité simplifie la construction de modèles, diminue le nombre paramètres et facilite les calculs (inversion de matrices) aux dépens des interactions spatio-temporelles (très présentes en physique) [60]. Cette séparabilité implique la notion de symétrie complète ($C(s,t) = C(s,-t) = C(-s,t) = C(-s,-t)$) et peut être testée [61]. Réciproquement, le cas inséparable a également été fortement étudié [62].

### 2.4.3 Kriging spatio-temporel

Les fonctions de covariance n'ont jamais qu'un rôle descriptif et, malgré leur souplesse, il est difficile de se rendre compte de l'étiologie du processus. Seulement, elles jouent un rôle dans la technique du *kriging* qui se base sur la définition de variogrammes. Or le *kriging* ne nécessite pas la notion de stationnarité pour exister [63]. On peut alors étendre le *kriging* aux données spatio-temporelles. En supposant que les données soient de la forme :

$$Z(s_i, t_{ij}) = Y(s_i, t_{ij}) + \epsilon(s_i, t_{ij}) \quad i \in S, j \in T_i \tag{2.21}$$

où $S$ représente l'ensemble des positions et $T_i$ les observations associé à l'endroit $i$ (par rapport au temps).

Le *simple kriging* revient alors à optimiser l'expression suivante :

$$\hat{Y}(s_0, t_0) \equiv \sum_{i=1}^{n} \sum_{j=1}^{T_i} l_{ij} Z(s_i, t_{ij}) + k \equiv l'Z + k \tag{2.22}$$

sous minimisation de l'expression :

$$E(Y(s_0, t_0) - l'Z - k)^2 \tag{2.23}$$

Cette technique nécessite que la moyenne soit connue. Si elle est supposée constante, elle donne naissance à l'*ordinary kriging* et si elle est combinaison linéaire, on parle de *universal kriging* [24].

Généralement, on préfère parler de *cokriging* qui est une extension multivariée de ce concept et qui prend en compte, non seulement l'information des mesures directes, mais également des autres composantes. Dans le but de fournir un meilleur estimateur, moins sujet à de grandes variations car étayées par les autres observations. Ainsi améliorer les prédictions des variables spatiales trop peu échantillonnées en exploitant davantage la corrélation spatiale attachée à d'autres variables plus facilement mesurables [64].

### 2.4.4 Séries temporelles

Il y a deux grandes familles de séries temporelles liées aux données spatio-temporelles, celle liée aux processus géostatistiques et celle des *lattice*. Dans les deux cas, la temps est vu de manière discrète et les processus spatiaux deviennent multivariés afin de compenser cette perte temporelle [65].

Pour les processus spatiaux continus, l'idée est d'exprimer le résultat à l'instant $t$ en un lieu comme une fonction (potentiellement non linéaire) de l'instant $t-1$ :

$$Y_t(s) = \mathcal{M}(s, Y_{t-1}(.)) + \epsilon_t(s) \text{ avec } \mathcal{M}(s, f(.)) \equiv \int_{\mathbb{R}^d} m(s,x) f(x) dx \tag{2.24}$$

avec $f(.) \in \mathbb{R}^d$ une fonction quelconque telle que l'intégrale existe et $m(s,x)$ qui contrôle l'influence de $Y_{t-1}$ sur l'instant actuel. Cette forme est liée au filtre de Kalman, offre beaucoup de flexibilité et possède de bonnes propriétés, notamment pour la réduction de dimensions ou la représentation de modèles spatiaux statiques [66].

Dans le cas des *lattice*, on voit le problème sous un autre angle. À un instant $t$, toutes les valeurs aux différentes positions représentent un élément de la série temporelle n-variée (une dépendance par rapport au temps étant plus naturelle). On emploie alors des modèles de type SAR ou CAR [67]. Ou, ce qui est plus classique pour les séries temporelles, la notion de modèle à vecteur autorégressif (VAR) qui est employé pour capturer les interdépendances linéaires parmi les multiples séries temporelles. La notion de VAR vise à simplifier celle de *Spatio-Temporal AutoRegressive Moving-Average* (STARMA) en limitant le nombre de paramètres [68].

### 2.4.5 Modèles hiérarchiques dynamiques spatio-temporels

Les processus spatio-temporels sont de nature à être modéliser par des dynamiques, on parle alors de *dynamical spatio-temporal models* ou DTSM. Or, ces processus sont difficile à appréhender et on tente alors de simplifier en reformulant le problème comme suit [69] :

| | |
|---|---|
| Data model : | $[data \vert process, parameters]$ |
| Process Model : | $[process \vert parameters]$ |
| Parameter Model : | $[parameters]$ |

$$\begin{aligned}[process, parameters|data] &\propto [data|process, parameters] \\ &\times [process|parameters] \\ &\times [parameters]\end{aligned}$$

Le modèle des données correspond à :

$$[\{Z(x,r) : (x,r) \in D_s \times D_t\} | \{Y(s,t) : (s,t) \in N(s) \times N(t)\}, \theta_D] \tag{2.25}$$

Les $Z(x,r)$ sont nos données récoltées, $Y(s,t)$ le processus inconnu à l'origine et $\theta_D$ les paramètres de ce modèle. En pratique, on fait l'hypothèse que l'observation n'est résultant que du processus et donc qu'on a indépendance conditionnelle dans les données.

Celui du processus :

$$[Y(s,t) | \{Y(w, t-\tau_1), ..., Y(w, t-\tau_p) : w \in N(s,p)\}, \theta_p] \tag{2.26}$$

$N(s,p)$ est le voisinage de $s$ à l'instant $\tau_p$ associé, $\theta_p$ les paramètres.

Et enfin, le modèle des paramètres :

$$[\theta_D, \theta_p | \theta_h] \tag{2.27}$$

$\theta_h$ est qualifié d'hyper-paramètre et peut également être décomposé en ajoutant un nouveau niveau à la hiérarchie.

Tout ceci est fort généraliste et change fortement en fonction des hypothèses mais permet de mettre au clair les différentes notions.

**Modèles des données**

Le plus simple des modèles consiste à créer une relation linéaire :

$$Z_t = a_t + H_t Y_t + \epsilon_t \tag{2.28}$$

où les erreurs sont toutes indépendantes et seulement déterminées par leur variance ($\epsilon_t \sim Gau(0, \sigma^2 I)$). Le $a_t$ et le $H_t \in \mathbb{R}^{n \times n}$ représentent le biais entre les $n$ données. Malgré sa simplicité, les paramètres peuvent être difficiles à estimer si les données sont trop peu nombreuses. La présence de la matrice, outre le fait d'offrir une relation linéaire entre le processus et les observations [70], permet de mieux gérer le cas où le nombre d'observations diffèrent du processus. Elle correspond souvent au phénomène d'incidence dans les graphes qui permet de décrire les liens entre les données. Enfin, elle permet de mettre en place les techniques de réductions de dimensions liés aux notions de spectre [13, 71].

Introduire de la non linéarité dans le modèle complexifie fortement la tâche en l'absence d'une bonne connaissance des paramètres. Et on peut alors envisager d'appliquer une transformation au processus [72]. On peut également enlever l'hypothèse de données gaussiennes mais il faut alors garder l'interdépendance conditionnelle, de nombreux modèles existent dans ce domaine, essentiellement des Poisson [73].

**Modèles du processus**

Les modèles spatio-temporels sont par nature dynamique, l'état actuel est déterminé par le passé récent du processus, et on peut espérer pouvoir appliquer l'approximation markovienne du premier ordre et ainsi être en mesure d'écrire :

$$Y_T = \mathcal{M}(Y_{t-1}, \epsilon_t; \theta_p) \approx Y_t - \mu_t = M(Y_{t-1} - \mu_{t-1}) + \epsilon_t \tag{2.29}$$

avec $\mathcal{M}$ une fonction quelconque, $\epsilon_t$ un bruit (souvent gaussien) et $M$ une matrice de transition. Cette matrice est souvent très difficile à estimer et on tente de trouver le nombre minimum de paramètres capables de capturer l'évolution du système et ses dynamiques [73].

On peut également employer des modèles auto-régressifs spatio-temporels (STAR), des équations différentielles ou des équations aux différences [13, 74].

Dans le cadre des modèles non linéaires, les problèmes liés au fléau des dimensions et d'une paramétrisation efficace sont d'autant plus exacerbés et les comportements chaotiques n'améliorent pas la situation. L'idée la plus logique consisterait donc à faire des approximations locales (développement de Taylor) et sont à mettre en relation avec les filtres étendus de Kalman [75]. Il existe bon nombre d'autres méthodes dont une basée sur des agents (automates cellulaires) afin d'étendre des choix locaux à tout le domaine [76].

### 2.4.6 Filtre de Kalman

L'un des modèles de processus linéaires et gaussiens les plus célèbres est le filtre de Kalman [77]. En effet, celui-ci est une méthode itérative qui offre des solutions théoriques pour les prédictions et la notion de filtre. Il permet également d'obtenir très rapidement une bonne idée de la valeur moyenne. On définit $Y_{t|T} \equiv \mathbb{E}[Y_t | Z_{1:T}]$ avec $Z_{1:T}$ qui représente toutes les observations dans l'intervalle de 1 à $T$ ordonné par rapport

au temps. Pour le filtre, si $T$ vaut $t-1$, il s'agit de la phase de prédiction du processus $Y$ en fonction des observations $Z$ et si $T$ égale $t$, celle de mise à jour du filtre. Des matrices de covariance d'erreur conditionnelle sont également définis comme suit (plus classiquement noté $P_{t|T}$) :

$$C_{t|T} \equiv \mathbb{E}[(Y_t - Y_{t|T})(Y_t - Y_{t|T})'|Z_{1:T}] \tag{2.30}$$

et le résultat théorique est celui-ci :

$$Y_t|Z_{1:T} \sim Gau(Y_{t|T}, C_{t|T}) \tag{2.31}$$

En fonction de ce paramètre $T$, on parle de distribution de prévision si $T = t-1$, de filtre si $T = t$ et de lisse (*smooth*) de manière générale. Cette dernière est surtout utile à des fins d'analyses rétrospectives mais nécessite de connaître les matrices de paramètres à chaque instant ainsi que les conditions initiales. Ceci permet également une définition récursive [75].

L'hypothèse de connaissance des paramètres de ces matrices est forte mais on peut les estimer si on considère qu'ils sont constants par rapport au temps. Il existe de nombreuses méthodes : méthode des moments, *max likelihood* par Newton-Raphson, *Expectation-Maximization* ou estimateur de Gibbs au travers des MCMC [12, 51].

Enfin, les modèles non linéaires ou non gaussiens peuvent utiliser le filtre étendu de Kalman ou les MCMC même si l'algorithme de Metropolis-Hastings peut devenir très difficile à mettre en place [78]. Des algorithmes de type *sampling importance resampling* (*SIR*) ou *Integrated Nested Laplace Approximation* (*INLA*) sont apparus plus récemment et sont plus rapides que les techniques de type MC pour les cas bayésiens [79].

Librairies R associées : SpatioTemporal (outils génériques mais complet pour la modélisation de problèmes spatio-temporels) [80], stem (estimation des paramètres des modèles) [81] ou spate (modèles à base d'équations différentielles) [82].

### 2.4.7 Visualisation

La visualisation de données spatio-temporelles révèle de nombreux problèmes [83]. En effet, les données sont, au minimum, tridimensionnelles et il n'est pas rare de vouloir observer plusieurs paramètres afin de déterminer les possibles interactions. Naturellement, représenter sous forme d'animation, de part la nature spéciale du temps, est tout indiqué. Bien sûr, toute représentation dépend du problème étudié mais une illustration puissante en météorologie sont les diagrammes de Hovmöller, où seule une direction est importante et qui permet facilement d'interpréter l'aspect temporel du processus [84].

En restant à deux dimensions, les séries temporelles condensent les valeurs d'une région pour un paramètre en fonction du temps. Inversement, les cartes spatiales représentent diverses observations à un moment donné. Bien sûr, des données peuvent manquer ou des intervalles de temps peuvent être plus larges et il faut alors recourir à des techniques d'interpolation [85].

La représentation des matrices de corrélation n'est pas dénuée d'intérêts. Tant les notions de *lag* temporel et spatial liés à l'auto-corrélation[2.2.2] [86] que les indicateurs locaux d'association spatiale (LISA) [87] (idée des *Moran's I*[2.2.1]) qui permettent de s'apercevoir des comportements anormaux du modèle sont des éléments-clefs à la bonne compréhension des données et du problème.

Un autre aspect important est liée à l'analyse spectrale. Ces outils ont beaucoup d'applications et permettent notamment de s'apercevoir des phénomènes de vagues

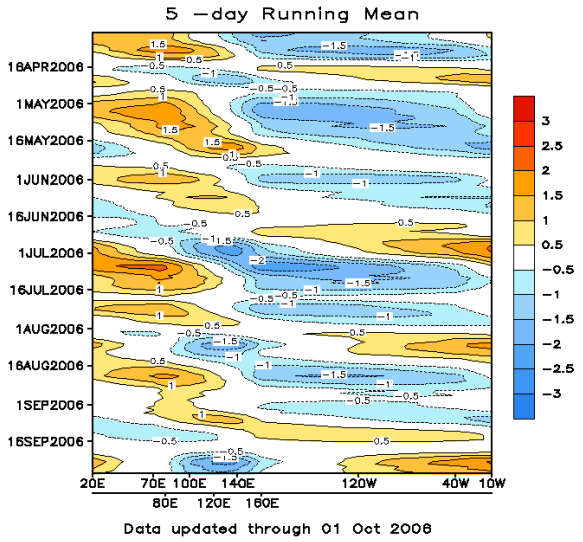



FIGURE 2.1 – Diagramme de Hovmöller - Émissions des radiations terrestres de longue portée - Source : *U.S. National Oceanic and Atmospheric Administration*

ou d'explorer les structures et motifs dans les données [83, 88]. La décomposition en valeurs propres est attachée aux fonctions orthogonales empiriques (EOF), qui, dans le cas discret, correspondent à l'analyse des composantes principales (PCA). Le cas continu est lié à toute la théorie de Karhunen-Loeve [89].

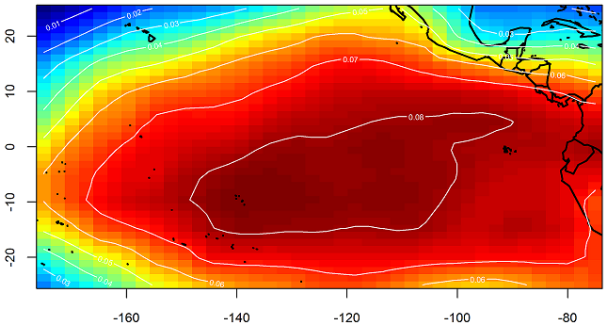



FIGURE 2.2 – Répresentation des EOF - Pression au niveau de la mer - Source : `http://menugget.blogspot.be/`

### 2.4.8 Processus ponctuel

Naturellement, les concepts liés aux phénomènes spatiaux[2.2.5] sont étendus pour inclure la notion de temporalité. Mais, on bénéficie de deux nouvelles propriétés, la possibilité de définir une dépendance par rapport au temps et celle de ne pas atteindre un équilibre. On définit alors la fonction d'intensité spatio-temporel $\lambda(s,t)$ qui représente le nombre d'événements attendus par région par unité de temps. On peut également réaliser une estimation non paramétrique par *kernel smoothing* mais en introduisant une dépendance entre les deux variables [90].

On parle de séparabilité du premier ordre si on est capable de réécrire pour toute région $A$ : $\lambda(s,t) \propto \lambda_A(s)\mu_A(t)$. Ceci permet de simplifier certains calculs. Il existe également une notion similaire au deuxième moment (et à la stationnarité) et qualifiée de fonction-$K$ :

$$K(u,v) = \lambda^{-1}E[N_0(u,v)] \simeq K_s(u)K_t(v)K(u,v) = \pi u^2 v + 2\pi\lambda^{-2}\int_0^v \int_0^u \gamma(u',v')u'du'dv' \tag{2.32}$$

Avec $N_0(u,v)$ le nombre d'événements dans un disque de rayon $u$ sur un intervalle de temps $[0,v)$ et $\gamma(u,v)$ la densité de covariance [91]. On peut dès lors employer les processus de Poisson ou de Cox avec le même phénomène que dans le cas spatial, une description naturelle des processus qui ne peuvent être décrits complètement par les variables considérées mais par des interactions stochastiques entre les points.

Enfin, de manière plus générale, on peut s'intéresser aux trajectoires que prennent les points. Historiquement, la première approche proposée a, sans doute, été la définition des mouvements Browniens au travers d'équations différentielles stochastiques [25]. Celle-ci demeure très populaire et a été complexifiée en vue d'inclure des notions de dérivées secondes afin de mieux répondre aux problèmes spatiaux. En outre, ces phénomènes ont été également envisagés par l'intermédiaire des VAR, d'équations aux différences ou décrits par des chaînes de Markov [13].

Librairies R associées : stpp (analyse et simulation de processus ponctuels) [92], spatstat (fort complet) [45] ou lgcp (pour les processus de Cox) [46].

## 2.5 Études antérieures

Les données spatio-temporelles liées à la mobilité sont très nombreuses. En effet, avec la multiplication des capteurs urbains, l'usage intensif du téléphone portable ou la géolocalisation des applications, il devient plus facile d'être en mesure de mieux déceler et comprendre les comportements humains et leurs déplacements. Bon nombre d'études ont été réalisées sur ce thème.

On peut évidemment s'intéresser à l'aspect principal de la mobilité en tant que déplacement d'une population. La notion capitale attachée est le principe de diffusion qui prévaut dans de nombreux modèles et qui trouve ses origines dans l'écologie. L'idée est que la majorité des gens se déplacement sur de courtes distance mais, avec le temps, ces distances deviennent de plus en plus longues et parfois, on a des voyages sur de fort longues distances [93].

Il n'est pas rare de modéliser ce genre de problème par une distribution de Weibull, de Pareto ou exponentielle [94]. Ou sous un problème de quantification des motifs tant dans leur changement que dans leur portée. Soit en marquant des individus et en étudiant leur comportement (techniques dites de *Mark-recapture/resighting*). Ou en exploitant un point de vue plus physique, celui de la dualité d'Euler-Lagrange, qui se focalise respectivement sur la population et son évolution sur un large nombre d'individus et sur la caractérisation des déplacements de certains spécimens [95, 96].

Lorsqu'on s'intéresse aux déplacements humains, des aspects plus pratiques rentrent en compte ; en effet, nous avons une certaine régularité temporelle et spatiale [7]. L'un des premiers modèles théoriques développés, outre celui brownien [97], fut celui du *Lévy flight* qui donne une distribution à des mouvements aléatoires et permet de représenter des transitions entre deux mouvements browniens [98, 99]. Et on peut alors étudier

la dispersion en terme de distances entre les différentes observations [93]. Une extension populaire est celle du *Random waypoint model* dans le cadre du *mobile ad hoc network* (*MANET*) qui semble être un meilleur modèle pour représenter les comportements humains [100]. Enfin, il existe également une approche basée sur des modèles de markov [101].

Seulement, toutes ces notions considèrent des mouvements libres où il n'existe aucun obstacle. De surcroît, les humains tendent à optimiser leur trajet. Dans cette optique, des modèles ont été créés [102]. On peut également mieux prendre en compte les relations sociales ainsi que les cycles journaliers [103].

La mobilité peut être observée par divers biais. On peut étudier les déplacements en tant que tels, soit au niveau des taxis qui sont indicateurs de l'activité économique et humaine [104]. On peut appliquer des télémétries et de la géolocalisation afin d'en déduire les statistiques liées [105]. Une approche plus sociale est également envisageable, on analyse les dynamiques géo-temporelles de l'activité des utilisateurs sur des plateformes de partages de positions telles que : *Twitter*, *Facebook Places*, *Gowalla* et *Foursquare* [106, 107, 108, 109]. Enfin, analyser les comportements sur base des données de téléphonie mobile est aussi très courant [110, 111, 112, 113].

## 2.6 Données

Depuis 2011, la Commission européenne préconise une démocratisation des données publiques en vue d'une meilleure réutilisation potentielle des services ou produits, de faire face aux challenges sociaux en nous aidant à découvrir des solutions innovantes, d'améliorer l'efficacité des communications dans les administrations et également permettre aux citoyens de prendre part à la vie de la société tout en bénéficiant d'une meilleure transparence pour le gouvernement [114].

Bon nombre de villes ont commencé à mettre en ligne divers ensembles de statistiques et jeux de données. La majorité des grandes villes et métropoles proposent des jeux fort fournis, on peut citer, pour les États-Unis : Chicago, New York, San Francisco, New Orleans, Seattle ou Atlanta et pour l'Europe : le site officiel de l'Union *Data Europa* [115] tente de collecter toutes les données existantes, cependant il est loin d'être complet et on peut alors avoir une vision rapide de nombreuses bases de données à partir de *Open Data Monitor* [116]. La ville de *Köln*, Allemagne semble être un référence en matière de données de transports (*TAPASCologne*). La Chine commence également à produire ses propres données mais ils sont difficiles d'accès pour les non-sinophones.

En Belgique, les données sont moindres malgré les plateformes de *Bruxelles Mobilité* [117] et *OpenData Bruxelles* [118] pour Bruxelles, *Gent stad Data* [119] pour Gand, *Opendata Antwerpen* [120] pour Anvers ainsi que *Data Gov* [121] et *Open Belgium* [122] pour la Belgique.

Dans les ensembles de données moins classiques. *Foursquare* fournit une interface indiquant les lieux où se sont rendus les gens [123] et des jeux peuvent être trouvés sur internet. On peut également penser à *Twitter* mais certaines restrictions sont appliquées. Pour certaines villes, il existe des données sur les déplacements des taxis et autres véhicules de transports. Enfin, certaines compagnies téléphoniques fournissent des détails de communication des GSM (généralement associé au mot *CDR*). Citons également un jeu de données peu banal, celui du recensement des billets de banque américains en fonction de la commune.

Un outil important dans la simulation du trafic et qui est employé par bon nombres

de chercheurs est *Simulation of Urban MObility* (SUMO) [124]. Il a remplacé un ancien simulateur appelé *VanetMobiSim*, extension du programme *CANU Mobility Simulation Environment*. *Geographic Resources Analysis Support System* (GRASS GIS) [125] est un grand classique dans la gestion et l'analyse de données géospatiales, il permet de modéliser des processus spatio-temporels, produire des images ou de visualiser des données. Dans le même ordre d'idées, il existe les programmes QGIS [126] et ArcGIS (propriétaire) [127] qui supportent une visualisation, une édition et une analyse des données spatiales.

## 2.7 R packages

De nombreux paquets pour le langage R existent sur le marché afin de répondre à bon nombres de questions liées aux données spatio-temporelles et à leur analyse. Voici une présentation non-exhaustive issue d'un article fort complet sur le sujet [128] :

- SpatioTemporal [80] qui fournit des utilitaires afin d'estimer, prédire et valider des modèles spatio-temporels. Conçu pour l'analyse de pollution atmosphérique et maladie respiratoire.
- sp [27] présente un ensemble de classes et méthodes pour les données spatiales.
- gstat [28] permet de modéliser des problèmes géostatistiques multivariés, de réaliser des prédictions et des simulations. Il supporte les variogrammes spatio-temporels et le *kriging*.
- spacetime [29] pour manipuler et explorer des données.
- RandomFields [129] permet d'estimer les hyper-paramètres des champs aléatoire Gaussiens sur base du maximum de vraisemblance ou des moindres carrés.
- stpp [92] afin d'analyser, simuler and montrer des motifs des points spatio-temporels.
- spatstat [45] est fort complète pour étudier les processus ponctuels et propose tous les outils possibles et imaginables dont vous pourriez avoir besoin.
- spBayes [40] pour les processus ponctuels univariés et multivariés sous les modèles bayésiens.
- spate [82] pour modéliser par le biais des équations aux dérivées partielles stochastiques.

## 2.8 Conclusion

Vous l'aurez compris, l'analyse de données spatio-temporel s est un incroyablement vaste domaine qui trouve ses origines dans les années 60. Malgré toutes ces années d'étude, il reste bon nombre de problèmes ouverts. Les enjeux liés à cette problématique sont immenses et offrent des aspects très diversifiés. La recherche de solutions à ces problèmes est encore activement sujet à études et est porteuse d'une diversité phénoménale tant dans la manière d'envisager et d'approcher ces notions que dans leur réalisation sur le terrain. Et ce, en dépit de l'avènement des ordinateurs et de leur augmentation en puissance de calcul qui a permis de développer de nouveaux modèles toujours plus complexes.

Ils demeurent de nombreux concepts clefs mal perçus même si les études sur les déplacements des populations, et des humains plus spécifiquement, se sont multipliés. Les applications sont innombrables, tant au niveau de la physique, de l'économie ou

de l'urbanisme. C'est un véritable bouleversement que propose l'analyse des données spatio-temporelles. Par ailleurs, notre monde devient de plus en plus connecté et les données affluent dans tous les sens et ne demandent qu'à être analysées. Nous pouvons espérer, par cette quantité, mieux comprendre les dynamiques sous-jacentes et les impacts sur notre vie au quotidien.

# Chapitre 3

# Prévision

## 3.1 Introduction

Dans cette partie, nous tenterons de définir des propositions de pistes de recherche pour le mémoire. Nous commencerons par donner un bref aperçu du contexte relatif à la mobilité des véhicules à l'échelle des grandes villes. Nous continuerons par fournir une description d'un ensemble de données liées à cette mobilité. Et, nous terminerons par donner un plan et des idées de solutions pour ce genre de problématique.

## 3.2 Contexte

Ces dernières années, il est apparu un nombre important de tentatives de solutions aux problèmes de réseaux routiers liés à la mobilité véhiculaire. Seulement, l'essentiel de ces solutions ont été éprouvés par le biais de simulateurs puisque l'expérimentation étant pratiquement impossible pour des raisons pratiques et les approches analytiques étant souvent trop complexes. La justesse de la simulation étant d'autant plus capitale qu'il n'existe pas de jeu de données capturant à la fois les dynamiques microscopiques et macroscopiques des trajets [105].

De fait, l'évaluation des performances des réseaux routiers est souvent biaisé par la manière dont se déplacent les véhicules, ce qui affecte le protocole utilisé et donc le déplacement de ces voitures. Cette situation peut mener à des mauvaises conclusions si le trafic est représenté incorrectement, et ce, indépendamment de la simulation [130].

On a alors cherché à améliorer les modèles représentant les déplacements des voitures, tout d'abord en employant des techniques stochastiques (*Random Waypoint*), puis en tenant en compte des propriétés topologiques [131]. Ensuite, on a voulu inclure le comportement des véhicules au niveau microscopique ou encore la signalisation [132]. Pour aboutir au simulateur le plus poussé sur le marché, SUMO [124].

SUMO est un simulateur, libre de sources et de droits, dédié à la simulation de trafic urbain. Il permet de modéliser tant les véhicules ou les transports publiques que les piétons. En outre, il propose nativement une pléthore d'outils afin de visualiser les trajets, d'importer des cartes et des comportements ou encore de calculer les taux d'émission en particule. Il propose également une interface riche où peut se greffer facilement de nombreux autres outils [124].

Nous l'avons dit, vérifier l'exactitude des simulateurs est une tâche peu aisée. Néanmoins, des études ont tenté de tester l'efficacité de ceux-ci en fournissant une carto-

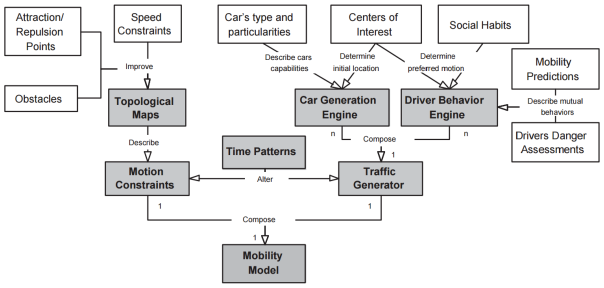



FIGURE 3.1 – Cadre de travail pour des modèles réalisates de simulation routière - Source : *Institut Eurécom* [132]

graphie entière de la région et en donnant le comportement macroscopique de la zone considérée [133] ou encore en comparant la simulation à des images prises d'un avion pour la ville de Porto, Portugal [134]. A l'inverse, à un niveau microscopique, des capteurs ont été placés sur les routes afin de mieux comprendre le comportements des véhicules à l'instar des études sur Bologna, Italie [135] ou sur le Luxembourg [136]. Le plus gros jeu de données est sans celui pour tout le canton de Zürich, Suisse qui correspond à l'ensemble des déplacements des véhicules pour une journée entière [137].

Cependant, les exemples précédents ont été générés en fournissant des données macroscopiques et microscopiques à un simulateur. Les trajets ont donc peut-être des comportements peu naturels et non représentatifs de l'activité réelle. Heureusement, il existe des exemples de déplacements réels tirés des réseaux de transport urbain ou de la circulation des taxis [138].

En vue d'obtenir une vision fort complète du trafic avec un haut niveau de granularité et ce sur une large région, une initiative appelée *TAPASCologne* [139] de l'*Institute of Transportation Systems at the German Aerospace Center* (ITS-DLR) vise à reproduire le plus fidèlement possible le trafic de véhicules dans la ville de Köln, Allegmane. Mais elle fait figure d'exception pour le moment.

## 3.3 Données

Un jeu de données qui a particulièrement retenu mon attention est lié à la ville de Aarhus, Danemark [140]. Celui-ci est une collections de données sur le trafic de véhicules, observés entre deux points pour une durée de plus de 6 mois (sur 449 points d'observations au total). Les données sont disponibles au format CSV et annoté au format des modèles d'information CityPulse. On peut également obtenir les informations en temps réel [141].

À chaque point d'observation est associé des méta-données qui donnent des informations sur le flux de données comme la position des senseurs, la distance entre-eux, le type de route, les précisions des capteurs, etc...

Chaque jeu de données est composée de plusieurs fichiers qui représentent les mesures effectuées entre deux capteurs. Dans ces fichiers, indexés par le temps par intervalle de 15 minutes, on retrouve le temps moyen, la vitesse moyenne et le nombre de véhicules ayant circulé entre ces points.

## 3.4 Prévisions

L'idée serait d'effectuer un travail analogue à celui de *TAPASCologne* mais pour la ville d'Arrhus, Danemark. Cela consisterait tout d'abord à trouver une cartographie de la ville qui contiendrait une idée sur les lieux d'intérêts ainsi que toutes les rues (cf. Open Street Map [142]) afin d'en extraire la topologie routière. Ce travail pourrait s'accompagner par une étude plus sociale sur la ville pour mieux comprendre les quartiers [143]. Ensuite, trouver des indications sur les problèmes de trafics dans cette ville et sur les demandes de déplacements des utilisateurs. Continuer par une simulation qui serait effectuée dans le logiciel SUMO en s'assurant que l'offre et la demande soit assurée (algorithme de Gawron [144]).

La première étape consisterait à tester la cohérence des données et à vérifier qu'elles soient bien complètes. Il faudrait voir l'évolution du nombre de véhicules en train d'effectuer un trajet, en partance et à destination ; ce qui donnerait des indications sur des signes de congestion excessive. Il y aurait alors sans doute du travail afin de corriger les comportements dans le but d'obtenir des résultats plus naturels (pics aux heures d'arrivées et de départs au travail, creux pendant la nuit, ...). Peut-être faudra-t-il penser à traiter des problèmes d'à-coups dans cette population et des informations inconsistantes sur les routes liées à des travaux. Tout cela s'effectuerait par le biais de diverses librairies écrites en R, on pense notamment à SpatioTemporal [80] et spastat [45] qui offrent de nombreux outils dédiés à cette problématique.

Enfin, on serait en mesure de tester une simulation, de s'assurer que des comportements logiques et cohérents soient bien rendus. Pour pouvoir lancer des analyses plus générales et à larges échelles. A l'instar d'études précédentes sur la ville de Köln nottament [145, 146]. Et finalement comparer les résultats du simulateur avec ceux observés. Notons qu'il s'agit d'un jeu de données qui n'a été le sujet d'aucune étude pour le moment.

# Bibliographie


[1] T. Hagerstrand *et al.*, "Innovation diffusion as a spatial process.," *Innovation diffusion as a spatial process.*, 1968. pages 3

[2] T. H. Newsome, W. A. Walcott, and P. D. Smith, "Urban activity spaces : Illustrations and application of a conceptual model for integrating the time and space dimensions," *Transportation*, vol. 25, no. 4, pp. 357–377, 1998. pages 3

[3] K. W. Axhausen, A. Zimmermann, S. Schönfelder, G. Rindsfüser, and T. Haupt, "Observing the rhythms of daily life : A six-week travel diary," *Transportation*, vol. 29, no. 2, pp. 95–124, 2002. pages 3

[4] H. J. Miller and J. Han, *Geographic data mining and knowledge discovery*. CRC Press, 2009. pages 3

[5] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W.-Y. Ma, "Understanding mobility based on gps data," in *Proceedings of the 10th international conference on Ubiquitous computing*, pp. 312–321, ACM, 2008. pages 3

[6] B. Grenfell, O. Bjørnstad, and J. Kappey, "Travelling waves and spatial hierarchies in measles epidemics," *Nature*, vol. 414, no. 6865, pp. 716–723, 2001. pages 3

[7] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008. pages 3, 20

[8] "Brussels mobi-aid : Brussels mobility advanced indicators dashboard." `http://mlg.ulb.ac.be/node/810`. Accessed : 2017-05-28. pages 3

[9] "Machine learning group." `http://mlg.ulb.ac.be/`. Accessed : 2017-05-28. pages 3

[10] "Mobi research group." `http://mobi.vub.ac.be/home/`. Accessed : 2017-05-28. pages 3

[11] "Villo dashboard." `https://github.com/Myxfall/MOBI-AID`. Accessed : 2017-05-28. pages 3

[12] N. Cressie and C. K. Wikle, *Statistics for spatio-temporal data*. John Wiley & Sons, 2015. pages 6, 8, 18

[13] A. E. Gelfand, P. J. Diggle, M. Fuentes, and P. Guttorp, *Handbook of spatial statistics*. CRC press, 2010. pages 6, 9, 14, 17, 20

[14] H. De Knegt, F. v. van Langevelde, M. Coughenour, A. Skidmore, W. De Boer, I. Heitkönig, N. Knox, R. Slotow, C. Van der Waal, and H. Prins, "Spatial autocorrelation and the scaling of species–environment relationships," *Ecology*, vol. 91, no. 8, pp. 2455–2465, 2010. pages 6

[15] P. Legendre, "Spatial autocorrelation : trouble or new paradigm ?," *Ecology*, vol. 74, no. 6, pp. 1659–1673, 1993. pages 6

[16] P. M. Dixon, "Ripley's k function," *Encyclopedia of environmetrics*, 2002. pages 6

[17] C. F Dormann, J. M McPherson, M. B Araújo, R. Bivand, J. Bolliger, G. Carl, R. G Davies, A. Hirzel, W. Jetz, W. Daniel Kissling, *et al.*, "Methods to account for spatial autocorrelation in the analysis of species distributional data : a review," *Ecography*, vol. 30, no. 5, pp. 609–628, 2007. pages 6

[18] J. Besag, "Spatial interaction and the statistical analysis of lattice systems," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 192–236, 1974. pages 7, 10

[19] B. D. Ripley, "Modelling spatial patterns," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 172–212, 1977. pages 7

[20] N. Cressie, "The origins of kriging," *Mathematical geology*, vol. 22, no. 3, pp. 239–252, 1990. pages 7, 8

[21] G. Matheron, "The intrinsic random functions and their applications," *Advances in applied probability*, vol. 5, no. 03, pp. 439–468, 1973. pages 7

[22] A. N. Kolmogorov and Y. A. Rozanov, "On strong mixing conditions for stationary gaussian processes," *Theory of Probability & Its Applications*, vol. 5, no. 2, pp. 204–208, 1960. pages 7

[23] G. Matheron, *The theory of regionalized variables and its applications*, vol. 5. École national supérieure des mines, 1971. pages 7, 9

[24] J. M. Ver Hoef and N. Cressie, "Multivariable spatial prediction," *Mathematical Geology*, vol. 25, no. 2, pp. 219–240, 1993. pages 8, 15

[25] P. J. Diggle, *Statistical analysis of spatial and spatio-temporal point patterns*. CRC Press, 2013. pages 8, 20

[26] T. S. Breusch and A. R. Pagan, "A simple test for heteroscedasticity and random coefficient variation," *Econometrica : Journal of the Econometric Society*, pp. 1287–1294, 1979. pages 9

[27] E. J. Pebesma and R. S. Bivand, "Classes and methods for spatial data in R," *R News*, vol. 5, pp. 9–13, November 2005. pages 9, 22

[28] E. J. Pebesma, "Multivariable geostatistics in s : the gstat package," *Computers & Geosciences*, vol. 30, no. 7, pp. 683–691, 2004. pages 9, 22

[29] E. Pebesma *et al.*, "spacetime : Spatio-temporal data in r," *Journal of Statistical Software*, vol. 51, no. 7, pp. 1–30, 2012. pages 9, 22

[30] P. Ribeiro Jr. and P. Diggle, "geoR : a package for geostatistical analysis," *R-NEWS*, vol. 1, no. 2, pp. 15–18, 2001. pages 10

[31] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm," *IEEE transactions on medical imaging*, vol. 20, no. 1, pp. 45–57, 2001. pages 10

[32] G. R. Grimmett, "A theorem about random fields," *Bulletin of the London Mathematical Society*, vol. 5, no. 1, pp. 81–84, 1973. pages 10

[33] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," *IEEE Transactions on pattern analysis and machine intelligence*, no. 6, pp. 721–741, 1984. pages 10

[34] M. S. Kaiser and N. Cressie, "The construction of multivariate distributions from markov random fields," *Journal of Multivariate Analysis*, vol. 73, no. 2, pp. 199–220, 2000. pages 10

[35] M. West, P. J. Harrison, and H. S. Migon, "Dynamic generalized linear models and bayesian forecasting," *Journal of the American Statistical Association*, vol. 80, no. 389, pp. 73–83, 1985. pages 10

[36] J. W. Lichstein, T. R. Simons, S. A. Shriner, and K. E. Franzreb, "Spatial autocorrelation and autoregressive models in ecology," *Ecological monographs*, vol. 72, no. 3, pp. 445–463, 2002. pages 10

[37] M. M. Wall, "A close look at the spatial structure implied by the car and sar models," *Journal of statistical planning and inference*, vol. 121, no. 2, pp. 311–324, 2004. pages 11

[38] D. Lee, "CARBayes : An R package for Bayesian spatial modeling with conditional autoregressive priors," *Journal of Statistical Software*, vol. 55, no. 13, pp. 1–24, 2013. pages 11

[39] R. Bivand and G. Piras, "Comparing implementations of estimation methods for spatial econometrics," *Journal of Statistical Software*, vol. 63, no. 18, pp. 1–36, 2015. pages 11

[40] A. O. Finley, S. Banerjee, and B. P. Carlin, "spbayes : an r package for univariate and multivariate hierarchical point-referenced spatial models," *Journal of Statistical Software*, vol. 19, no. 4, p. 1, 2007. pages 11, 22

[41] D. J. Daley and D. Vere-Jones, *An introduction to the theory of point processes : volume II : general theory and structure*. Springer Science & Business Media, 2007. pages 11

[42] A. Baddeley, I. Bárány, and R. Schneider, "Spatial point processes and their applications," *Stochastic Geometry : Lectures given at the CIME Summer School held in Martina Franca, Italy, September 13–18, 2004*, pp. 1–75, 2007. pages 11

[43] D. R. Cox, "Some statistical methods connected with series of events," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 129–164, 1955. pages 11

[44] J. Illian, A. Penttinen, H. Stoyan, and D. Stoyan, *Statistical analysis and modelling of spatial point patterns*, vol. 70. John Wiley & Sons, 2008. pages 11

[45] A. Baddeley, "Spatial point pattern analysis, model-fitting, simulation, tests," 2016. pages 11, 20, 22, 26

[46] B. M. Taylor, T. M. Davies, B. S. Rowlingson, and P. J. Diggle, "lgcp : An R package for inference with spatial and spatio-temporal log-Gaussian Cox processes," *Journal of Statistical Software*, vol. 52, no. 4, pp. 1–40, 2013. pages 11, 20

[47] D. Harte, "PtProcess : An R package for modelling marked point processes indexed by time," *Journal of Statistical Software*, vol. 35, no. 8, pp. 1–32, 2010. pages 11

[48] M. G. Kendall *et al.*, "The advanced theory of statistics," *The advanced theory of statistics.*, no. 2nd Ed, 1946. pages 11

[49] C. Chatfield, *Time-series forecasting*. CRC Press, 2000. pages 12

[50] C. Chatfield, *The analysis of time series : an introduction*. CRC press, 2016. pages 12, 13

[51] R. H. Shumway and D. S. Stoffer, *Time series analysis and its applications : with R examples*. Springer Science & Business Media, 2010. pages 12, 13, 18

[52] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis : forecasting and control*. John Wiley & Sons, 2015. pages 12, 13

[53] H. Lütkepohl, *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005. pages 12

[54] R. J. Hyndman and Y. Khandakar, "Environment for teaching "financial engineering and computational finance". managing financial time series objects.," *Journal of Statistical Software*, vol. 26, no. 3, pp. 1–22, 2008. pages 13

[55] R. J. Hyndman and Y. Khandakar, "Automatic time series forecasting : the forecast package for R," *Journal of Statistical Software*, vol. 26, no. 3, pp. 1–22, 2008. pages 13

[56] Twitter, Inc. and other contributors, "Anomalydetection r package," 2015. pages 13

[57] C. K. Wikle *et al.*, "\ bf hierarchical models in environmental science," *International Statistical Review*, vol. 71, no. 2, pp. 181–199, 2003. pages 14

[58] N. Cressie and H.-C. Huang, "Classes of nonseparable, spatio-temporal stationary covariance functions," *Journal of the American Statistical Association*, vol. 94, no. 448, pp. 1330–1339, 1999. pages 14

[59] M. L. Stein, *Interpolation of spatial data : some theory for kriging*. Springer Science & Business Media, 2012. pages 14

[60] M. G. Genton, "Separable approximations of space-time covariance matrices," *Environmetrics*, vol. 18, no. 7, pp. 681–695, 2007. pages 15

[61] M. Fuentes, "Testing for separability of spatial–temporal covariance functions," *Journal of statistical planning and inference*, vol. 136, no. 2, pp. 447–466, 2006. pages 15

[62] M. Fuentes, L. Chen, and J. M. Davis, "A class of nonseparable and nonstationary spatial temporal covariance functions," *Environmetrics*, vol. 19, no. 5, pp. 487–507, 2008. pages 15

[63] N. Cressie, "Kriging nonstationary data," *Journal of the American Statistical Association*, vol. 81, no. 395, pp. 625–634, 1986. pages 15

[64] H. Wackernagel, *Multivariate geostatistics : an introduction with applications*. Springer Science & Business Media, 2013. pages 15

[65] S. Rouhani and H. Wackernagel, "Multivariate geostatistical approach to space-time data analysis," *Water Resources Research*, vol. 26, no. 4, pp. 585–591, 1990. pages 15

[66] C. K. Wikle and N. Cressie, "A dimension-reduced approach to space-time kalman filtering," *Biometrika*, pp. 815–829, 1999. pages 16

[67] D. J. Allcroft and C. A. Glasbey, "A latent gaussian markov random-field model for spatiotemporal rainfall disaggregation," *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, vol. 52, no. 4, pp. 487–498, 2003. pages 16

[68] P. E. Pfeifer and S. J. Deutsch, "Identification and interpretation of first order space-time arma models," *Technometrics*, vol. 22, no. 3, pp. 397–408, 1980. pages 16

[69] C. K. Wikle, L. M. Berliner, and N. Cressie, "Hierarchical bayesian space-time models," *Environmental and Ecological Statistics*, vol. 5, no. 2, pp. 117–154, 1998. pages 16

[70] C. A. Gotway and L. J. Young, "Combining incompatible spatial data," *Journal of the American Statistical Association*, vol. 97, no. 458, pp. 632–648, 2002. pages 17

[71] C. K. Wikle, R. F. Milliff, D. Nychka, and L. M. Berliner, "Spatiotemporal hierarchical bayesian modeling tropical ocean surface winds," *Journal of the American Statistical Association*, vol. 96, no. 454, pp. 382–397, 2001. pages 17

[72] B. Sanso and L. Guenni, "Venezuelan rainfall data analysed by using a bayesian space–time model," *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, vol. 48, no. 3, pp. 345–362, 1999. pages 17

[73] S. Banerjee, B. P. Carlin, and A. E. Gelfand, *Hierarchical modeling and analysis for spatial data*. Crc Press, 2014. pages 17

[74] M. Cameletti, F. Lindgren, D. Simpson, and H. Rue, "Spatio-temporal modeling of particulate matter concentration through the spde approach," *AStA Advances in Statistical Analysis*, vol. 97, no. 2, pp. 109–131, 2013. pages 17

[75] M. S. Grewal, *Kalman filtering*. Springer, 2011. pages 17, 18

[76] C. K. Wikle and M. B. Hooten, "A general science-based framework for dynamical spatiotemporal models," *Test*, vol. 19, no. 3, pp. 417–451, 2010. pages 17

[77] R. E. Kalman *et al.*, "A new approach to linear filtering and prediction problems," *Journal of basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960. pages 17

[78] R. Van Der Merwe, A. Doucet, N. De Freitas, and E. Wan, "The unscented particle filter," in *Nips*, vol. 2000, pp. 584–590, Denver, CO, USA, 2000. pages 18

[79] H. Rue, S. Martino, and N. Chopin, "Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations," *Journal of the royal statistical society : Series b (statistical methodology)*, vol. 71, no. 2, pp. 319–392, 2009. pages 18

[80] J. P. Keller, C. Olives, K. Sun-Young, L. Sheppard, P. D. Sampson, A. A. Szpiro, A. P. Oron, J. Lindström, S. Vedal, and J. D. Kaufman, "A unified spatiotemporal modeling approach for predicting concentrations of multiple air pollutants in the multi-ethnic study of atherosclerosis and air pollution," *Environmental Health Perspectives (Online)*, vol. 123, no. 4, p. 301, 2015. pages 18, 22, 26

[81] M. Cameletti, *Stem : Spatio-temporal models in R*. BibSonomy, 2009. R package version 1.0. pages 18

[82] F. Sigrist, H. R. Künsch, W. A. Stahel, *et al.*, "spate : An r package for spatio-temporal modeling with a stochastic advection-diffusion process," *Journal of Statistical Software*, vol. 63, no. 14, pp. 1–23, 2015. pages 18, 22

[83] C.-h. Chen, W. K. Härdle, and A. Unwin, *Handbook of data visualization*. Springer Science & Business Media, 2007. pages 18, 19

[84] E. Hovmöller, "The trough-and-ridge diagram," *Tellus*, vol. 1, no. 2, pp. 62–66, 1949. pages 18

[85] M. P. Peterson, "Spatial visualization through cartographic animation : theory and practice," in *GIS/LIS*, pp. 619–628, 1994. pages 18

[86] N. Andrienko, G. Andrienko, and P. Gatalsky, "Exploratory spatio-temporal visualization : an analytical review," *Journal of Visual Languages & Computing*, vol. 14, no. 6, pp. 503–541, 2003. pages 18

[87] L. Anselin, "Local indicators of spatial association—lisa," *Geographical analysis*, vol. 27, no. 2, pp. 93–115, 1995. pages 18

[88] E. Renshaw and E. Ford, "The interpretation of process from pattern using two-dimensional spectral analysis : methods and problems of interpretation," *Applied Statistics*, pp. 51–63, 1983. pages 19

[89] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2002. pages 19

[90] P. Diggle, "A kernel method for smoothing point process data," *Applied statistics*, pp. 138–147, 1985. pages 19

[91] E. Gabriel and P. J. Diggle, "Second-order analysis of inhomogeneous spatio-temporal point process data," *Statistica Neerlandica*, vol. 63, no. 1, pp. 43–51, 2009. pages 20

[92] E. Gabriel, B. Rowlingson, P. Diggle, *et al.*, "stpp : an r package for plotting, simulating and analyzing spatio-temporal point patterns," *Journal of Statistical Software*, vol. 53, no. 2, pp. 1–29, 2013. pages 20, 22

[93] D. Brockmann, L. Hufnagel, and T. Geisel, "The scaling laws of human travel," *Nature*, vol. 439, no. 7075, pp. 462–465, 2006. pages 20, 21

[94] S. I. Higgins and D. M. Richardson, "Predicting plant migration rates in a changing world : the role of long-distance dispersal," *The American Naturalist*, vol. 153, no. 5, pp. 464–475, 1999. pages 20

[95] R. Nathan, G. Perry, J. T. Cronin, A. E. Strand, and M. L. Cain, "Methods for estimating long-distance dispersal," *Oikos*, vol. 103, no. 2, pp. 261–273, 2003. pages 20

[96] T. R. E. Southwood and P. A. Henderson, *Ecological methods*. John Wiley & Sons, 2009. pages 20

[97] A. Einstein, "Un the movement of small particles suspended in statiunary liquids required by the molecular-kinetic theory 0f heat," *Ann. d. Phys*, vol. 17, pp. 549–560, 1905. pages 20

[98] I. Rhee, M. Shin, S. Hong, K. Lee, S. J. Kim, and S. Chong, "On the levy-walk nature of human mobility," *IEEE/ACM transactions on networking (TON)*, vol. 19, no. 3, pp. 630–643, 2011. pages 20

[99] D. W. Sims, E. J. Southall, N. E. Humphries, G. C. Hays, C. J. Bradshaw, J. W. Pitchford, A. James, M. Z. Ahmed, A. S. Brierley, M. A. Hindell, *et al.*, "Scaling laws of marine predator search behaviour," *Nature*, vol. 451, no. 7182, pp. 1098–1102, 2008. pages 20

[100] T. Camp, J. Boleng, and V. Davies, "A survey of mobility models for ad hoc network research," *Wireless communications and mobile computing*, vol. 2, no. 5, pp. 483–502, 2002. pages 21

[101] C. Bettstetter, "Mobility modeling in wireless networks : categorization, smooth movement, and border effects," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 3, pp. 55–66, 2001. pages 21

[102] K. Lee, S. Hong, S. J. Kim, I. Rhee, and S. Chong, "Slaw : A new mobility model for human walks," in *INFOCOM 2009, IEEE*, pp. 855–863, IEEE, 2009. pages 21

[103] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility : user movement in location-based social networks," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1082–1090, ACM, 2011. pages 21

[104] N. Ferreira, J. Poco, H. T. Vo, J. Freire, and C. T. Silva, "Visual exploration of big spatio-temporal urban data : A study of new york city taxi trips," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2149–2158, 2013. pages 21

[105] S. Uppoor, O. Trullols-Cruces, M. Fiore, and J. M. Barcelo-Ordinas, "Generation and analysis of a large-scale urban vehicular mobility dataset," *IEEE Transactions on Mobile Computing*, vol. 13, no. 5, pp. 1061–1075, 2014. pages 21, 24

[106] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil, "An empirical study of geographic user activity patterns in foursquare.," *ICwSM*, vol. 11, pp. 70–573, 2011. pages 21

[107] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo, "A tale of many cities : universal patterns in human urban mobility," *PloS one*, vol. 7, no. 5, p. e37027, 2012. pages 21

[108] S. Hasan, X. Zhan, and S. V. Ukkusuri, "Understanding urban human activity and mobility patterns using large-scale location-based data from online social media," in *Proceedings of the 2nd ACM SIGKDD international workshop on urban computing*, p. 6, ACM, 2013. pages 21

[109] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui, "Exploring millions of footprints in location sharing services.," *ICWSM*, vol. 2011, pp. 81–88, 2011. pages 21

[110] R. Becker, R. Cáceres, K. Hanson, S. Isaacman, J. M. Loh, M. Martonosi, J. Rowland, S. Urbanek, A. Varshavsky, and C. Volinsky, "Human mobility characterization from cellular network data," *Communications of the ACM*, vol. 56, no. 1, pp. 74–82, 2013. pages 21

[111] Y.-A. De Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd : The privacy bounds of human mobility," *Scientific reports*, vol. 3, p. 1376, 2013. pages 21

[112] R. Ahas, A. Aasa, S. Silm, and M. Tiru, "Daily rhythms of suburban commuters' movements in the tallinn metropolitan area : Case study with mobile positioning data," *Transportation Research Part C : Emerging Technologies*, vol. 18, no. 1, pp. 45–54, 2010. pages 21

[113] P. Deville, C. Linard, S. Martin, M. Gilbert, F. R. Stevens, A. E. Gaughan, V. D. Blondel, and A. J. Tatem, "Dynamic population mapping using mobile phone data," *Proceedings of the National Academy of Sciences*, vol. 111, no. 45, pp. 15888–15893, 2014. pages 21

[114] Council of European Union, "Council regulation (EU) no 833/2011," 2011. `http://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:32011D0833`. pages 21

[115] "Data europa." `https://data.europa.eu/`. Accessed : 2017-05-03. pages 21

[116] "Open data monitor." `http://opendatamonitor.eu`. Accessed : 2017-05-03. pages 21

[117] "Bruxelles mobilité." `http://data-mobility.brussels/fr/`. Accessed : 2017-05-03. pages 21

[118] "Open data bruxelles." `http://opendatastore.brussels/dataset`. Accessed : 2017-05-03. pages 21

[119] "Open data portaal gent." `https://data.stad.gent/`. Accessed : 2017-05-03. pages 21

[120] "Opendata antwerpen." `http://opendata.antwerpen.be/`. Accessed : 2017-05-03. pages 21

[121] "The belgian open data initiative." `http://data.gov.be/fr`. Accessed : 2017-05-03. pages 21

[122] "Open belgium." `http://portal.openbelgium.be/dataset`. Accessed : 2017-05-03. pages 21

[123] "Foursquare api." `https://developer.foursquare.com/`. Accessed : 2017-05-04. pages 21

[124] D. Krajzewicz, J. Erdmann, M. Behrisch, and L. Bieker, "Recent development and applications of SUMO - Simulation of Urban MObility," *International Journal On Advances in Systems and Measurements*, vol. 5, pp. 128–138, December 2012. pages 22, 24

[125] GRASS Development Team, *Geographic Resources Analysis Support System (GRASS GIS) Software, Version 7.2.* Open Source Geospatial Foundation, 2017. pages 22

[126] QGIS Development Team, *QGIS Geographic Information System.* Open Source Geospatial Foundation, 2017. pages 22

[127] ESRI Development Team, *ESRI ArcGIS Desktop.* Redlands, CA : Environmental Systems Research Institute, 2017. pages 22

[128] E. Pebesma, R. Bivand, P. J. Ribeiro, *et al.*, "Software for spatial statistics," *Journal of Statistical Software*, vol. 63, no. 1, pp. 1–8, 2015. pages 22

[129] M. Schlather, A. Malinowski, P. J. Menck, M. Oesting, K. Strokorb, *et al.*, "Analysis, simulation and prediction of multivariate random fields with package randomfields," *Journal of statistical software*, vol. 63, no. 8, pp. 1–25, 2015. pages 22

[130] F. Bai, N. Sadagopan, and A. Helmy, "The important framework for analyzing the impact of mobility on performance of routing protocols for adhoc networks," *Ad Hoc Networks*, vol. 1, no. 4, pp. 383–403, 2003. pages 24

[131] A. K. Saha and D. B. Johnson, "Modeling mobility for vehicular ad-hoc networks," in *Proceedings of the 1st ACM international workshop on Vehicular ad hoc networks*, pp. 91–92, ACM, 2004. pages 24

[132] J. Harri, F. Filali, and C. Bonnet, "Mobility models for vehicular ad hoc networks : a survey and taxonomy," *IEEE Communications Surveys & Tutorials*, vol. 11, no. 4, 2009. pages 24, 25

[133] R. Baumann, F. Legendre, and P. Sommer, "Generic mobility simulation framework (gmsf)," in *Proceedings of the 1st ACM SIGMOBILE workshop on Mobility models*, pp. 49–56, ACM, 2008. pages 25

[134] M. Ferreira, H. Conceição, R. Fernandes, and O. K. Tonguz, "Stereoscopic aerial photography : an alternative to model-based urban mobility approaches," in *Proceedings of the sixth ACM international workshop on VehiculAr InterNETworking*, pp. 53–62, ACM, 2009. pages 25

[135] L. Bieker, D. Krajzewicz, A. Morra, C. Michelacci, and F. Cartolano, "Traffic simulation for all : a real world traffic scenario from the city of bologna," in *Modeling Mobility with Open Data*, pp. 47–60, Springer, 2015. pages 25

[136] Y. Pigné, G. Danoy, and P. Bouvry, "A vehicular mobility model based on real traffic counting data," in *International Workshop on Communication Technologies for Vehicles*, pp. 131–142, Springer, 2011. pages 25

[137] B. Raney, N. Cetin, A. Völlmy, M. Vrtic, K. Axhausen, and K. Nagel, "An agent-based microsimulation model of swiss travel : First results," *Networks and Spatial Economics*, vol. 3, no. 1, pp. 23–41, 2003. pages 25

[138] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun, and Y. Huang, "T-drive : driving directions based on taxi trajectories," in *Proceedings of the 18th SIGSPATIAL International conference on advances in geographic information systems*, pp. 99–108, ACM, 2010. pages 25

[139] "Tapas köln." `http://sumo.dlr.de/wiki/Data/Scenarios/TAPASCologne`. Accessed : 2017-05-03. pages 25

[140] "Aarhus dataset." `http://iot.ee.surrey.ac.uk:8080/datasets.html`. Accessed : 2017-05-03. pages 25

[141] "Aarhus dataset real time." `https://www.odaa.dk/dataset/realtids-trafikdata/resource/b3eeb0ff-c8a8-4824-99d6-e0a3747c8b0d`. Accessed : 2017-05-03. pages 25

[142] "Open street map." `https://www.openstreetmap.org/`. Accessed : 2017-05-03. pages 26

[143] "Open data aarhus." `https://www.odaa.dk/`. Accessed : 2017-05-03. pages 26

[144] C. Gawron, "An iterative algorithm to determine the dynamic user equilibrium in a traffic simulation model," *International Journal of Modern Physics C*, vol. 9, no. 03, pp. 393–407, 1998. pages 26

[145] S. Uppoor and M. Fiore, "Large-scale urban vehicular mobility for networking research," in *Vehicular Networking Conference (VNC), 2011 IEEE*, pp. 62–69, IEEE, 2011. pages 26

[146] S. Uppoor and M. Fiore, "Insights on metropolitan-scale vehicular mobility from a networking perspective," in *Proceedings of the 4th ACM international workshop on Hot topics in planet-scale measurement*, pp. 39–44, ACM, 2012. pages 26