UNIVERSITÉ CHEIKH ANTA DIOP DE DAKAR

FACULTÉ DES SCIENCES ET TECHNIQUES

ECOLE DOCTORALE DE MATHÉMATIQUES ET INFORMATIQUE

Année : 2014-2015 N°d'ordre : _0_/_0_/_6_/_7_/

THÈSE DE DOCTORAT UNIQUE

Mention : Mathématiques et Modélisation
Spécialité : Analyse, Statistique et Applications

Présentée par

LEBEDE NGARTERA

TITRE

Modélisation et prédiction de l'indice de la qualité de l'air dans Dakar

Soutenue publiquement le 10/02/2016 devant le jury composé de :

Président :	**M. Hamidou Dathe**	Professeur, (Sénégal)
Rapporteurs :	**M. Gabriel Biramé Ndiaye**	Maître de conférences, UCAD (Sénégal)
	M. Ali Souleymane Dabye	Professeur, UGB (Sénégal)
	M. Edouard Wagneur	Professeur, GERAD (Canada)
Examinateurs :	**M. Youssou Gningue**	Professeur, Laurentian University (Canada)
	M. Serigne Aliou Lô	Maître de conférences, UCAD (Sénégal)
	M. Idrissa Ly	Maître de conférences, UCAD (Sénégal)
Invité :	**Mme Aminata M. Diokhané**	Responsable du CGQA (Sénégal)
Directeur de thèse :	**Mme Salimata Gueye Diagne**	Maître de conférences, UCAD (Sénégal)

Table des matières

Table des figures

Liste des tableaux

Dédicace

Je dédie cet humble travail,

- *À Dieu le* ***Père****, le* ***Fils*** *et le* ***Saint Esprit****, pour son amour infini,*
- *À mes très chers parents* ***Lebede Telro Robert, Helène Maoubé****, pour leur amour et sacrifices,*
- *À mes adorables frères, sœurs pour leur patience et prières,*
- *À mes proches amis et toute ma grande famille, pour leurs soutiens et encouragements,*
- *À toutes les personnes qui me connaissent de prés ou de loin, seulement pour leur existence,*
- *À tous ceux qui œuvrent aujourd'hui pour la valorisation des modèles numériques à travers le monde.*

Remerciements

"Gloire à Dieu au plus Haut des Cieux et Paix sur la Terre aux Hommes qu'Il aime !" Merci infiniment au Dieu créateur, l'auteur de tout ce qui a été, qui est et qui sera. Gloire, Honneur et Louange à son Saint Nom en l'honneur de tous les hommes et femmes qui donnent toute leur Vie pour la formation, l'éducation de l'espèce humaine.

Je remercie principalement ***Madame Gueye Diagne Salimata*** *pour avoir accepté de diriger ma thèse. J'ai eu la chance d'apprécier ses qualités et sa rigueur scientifiques. Je lui suis très reconnaissant de la confiance qu'elle m'a toujours accordée en dirigeant mes travaux de thèse. Merci à toute l'équipe pédagogique de l'UCAD et de l'UGB, pour avoir assuré ma formation. Quelques lignes ne suffiront évidemment pas à exprimer l'étendue de ma gratitude à leur égard, inutile donc de se lancer dans d'interminables éloges.*

Je tiens à remercier tous les membres du jury qui ont bien voulu examiner ce travail:

M. Hamidou Dathe, Professeur à l'UCAD/FST ;

M.Gabriel Birame Ndiaye, Maître de conférences à l'UCAD/FST ;

M.Ali Souleymane Dabye, Professeur à l'UGB/Saint Louis ;

M. Edouard Wagneur, Professeur GERAD/Canada ;

M. Youssou Gningue, Professeur Université Laurentienne, Canada ;

M. Sérigne Aliou Lô, Maître de conférences à l'UCAD/FST ;

M.Idrassa Ly, Maître de conférences à l'UCAD/FST ;

Mme Aminata Mbow Diokhané, Responsable du CGQA ;

Mme Salimita Gueye Diagne, Maître de conférences à l'UCAD/FST.

Un très grand merci du fond du cœur à ***Madame la Directrice de l'Environnement et des Établissements Classés (DEEC)*** *qui a pris la décision de m'accepter pour faire un stage par la note service,* $N^0 02257/MEDD/DEEC/DAF.od$*. Merci à vous et à tous vos collaborateurs que je ne peux tous nommés pour la disponibilité et le sens de vie.*

Je tiens à remercier tout particulièrement et à témoigner toute ma reconnaissance au Responsable du CGQA et son équipe, pour l'expérience enrichissante et pleine d'intérêt qu'elles m'ont fait vivre durant le stage. Merci à vous pour les données pour le travail. Merci à l'ensemble du personnel du CGQA pour leur accueil sympathique et leur coopération professionnelle tout au long de ces trois mois.

Merci au **Dr BOUBACAR MBODJI***, Expert associé, Modélisation-Émission-Pollution au CGQA par qui j'ai connu le CGQA. Merci à tous et à toutes !*

Listes des abréviations et sigles

ADEME: Agence de l'Environnement et de la Maîtrise de l'Énergie
AIC: Akaike Information Criterion ;
AR: Auto-Regressive ;
MA: Moving Average ;
ARMA:Auto Regressive Moving Average ;
ARIMA: Auto Regressive Integrated Moving Average ;
BIC: Bayesian Information Criterion ;
NE-SO: Nord Est Sud Ouest ;
FND: Fonds Nordique de Développement ;
FST: Faculté des Sciences et Techniques ;
NILU: Institut Norvégien de Recherche sur l'Air ;
min: minimum ;
max: maximum ;
MCO : Moindre Carré Ordinaire ;
IJAM : International Journal of Applied Mathematics
IJAMAS : International Journal of Applied Mathematics and Statistics ;
iqa ou IQA: Indice de la Qualité de l'Air ;
CETUD: Conseil Exécutif des Transports Urbains de Dakar ;
DEEC: Direction de l'Environnement et des Établissements Classés ;
PAMU: Programme d'Amélioration de la Mobilité Urbaine ;
ASN: Association Sénégalaise de Normalisation ;
BTX: Benzène, Toluène et Xylènes ;
CGQA: Centre de Gestion de la Qualité de l'Air ;
UCAD: Université Cheikh Anta Diop de Dakar.
CO: Monoxyde de carbone ;
NO_2: Dioxyde d'azote ;
UGB: Université Gaston Berger de Saint.
O_3: Ozone ;
OMS: Organisation Mondiale de la Santé ;
$PM_{2,5}$: Particules en suspension de diamètre inférieur à 2,5 μm ;
PM_{10}: Particules en suspension de diamètre inférieur à 10 μm ;
SO_2: Dioxyde de soufre ;
ANOCI: Agence Nationale pour l'Organisation de la Conférence Islamique ;
APIX: Agence Chargée de la promotion de l'Investissement et des Travaux ;
IPCC: Intergovernmental Panel on Climate Change ;
PDU: Plan de Développement Urbain ;
EDP: Équations aux Dérivées partielles ;
GES: Gaz à Effet de Serre.

Une pensée

Le futur se construit dans le présent.
Le présent prend sa naissance à partir
des données du passé.
Prédire, c'est s'appuyer sur les données
du passé dans le présent pour se projeter dans
le futur en minimisant les risques de déviations
pour y arriver.

LEBEDE NGARTERA.

Résumé

La pollution atmosphérique issue du trafic automobile au cœur de la ville ainsi que par les activités industrielles est un problème aigu des grandes cités. Une fois émis dans l'atmosphère, les polluants subissent deux types de contraintes : d'une part ils réagissent chimiquement entre eux donnant naissance à de nouveaux polluants tels que l'ozone, et d'autre part ils sont transportés par les vents. L'indice de la qualité de l'air est calculé à partir de ces polluants. L'objet de cette thèse est de modéliser et prédire cet indice dans Dakar par les outils mathématiques afin d'aider les décideurs à prendre des mesures servant à mieux optimiser les pics de pollution dans son sein. Notre ambition n'est pas de réaliser un modèle complet mais d'avancer pas à pas sur chacun de ces aspects. Nous avons développé de petits codes de calcul qui pourront éventuellement ensuite être complétés et utilisés pour des études plus concrètes d'épisodes de pollution. Le premier chapitre est une généralité consacrée à la modélisation de la pollution atmosphérique. Le second chapitre présente la qualité de l'air dans Dakar, suivi du troisième chapitre qui fournit les outils mathématiques nécessaires pour la modélisation et prédiction. Les chapitres quatre et cinq présentent les modèles utilisés et le résultat de simulations numériques par nos deux publications. Nous insistons particulièrement sur les modèles de régression linéaire multiple et les modèles auto-régressifs à moyenne mobile. Le document de thèse se conclut par une comparaison des deux modèles suivie de quelques perspectives et suggestions.

Mots Clés: Modélisation, prédiction, qualité de l'air, Dakar, advection, diffusion, dispersion, différence finie, schéma up-wind, ARMA(2,1), BIC, AIC, régression.

Abstract

Air pollution resulting from car traffic in the heart of the city and by industrial activities is an acute problem in large cities. Once emitted into the atmosphere, pollutants undergo two types of constraints: first they react chemically with each other giving rise to new pollutants such as ozone, and secondly they are transported by winds. The air quality index is calculated based on these pollutants. The purpose of this thesis is to model and predict the index in Dakar using mathematical tools to help decision makers to take action for better control of pollution peaks. Our purpose is not to create a complete model, but to make incremental progress in several respects. We developed small computer codes which may optionally then be completed and used for more specific studies of pollution episodes. The first chapter is devoted to a general modeling of air pollution. The second chapter presents the air quality in Dakar, followed by the third chapter which provides the necessary mathematical tools for modeling and prediction. The fourth and fifth chapters present the models used and the results of numerical simulations described by our two publications. We rely particularly on multiple linear regression models and autoregressive moving average models. The thesis concludes with a comparison of the two models, followed by some perspectives and suggestions.

Keywords: Modeling, prediction, air quality, Dakar, advection, diffusion, dispersion, finite difference, up-wind scheme, ARMA(2,1), Bic, Aic, regression.

Introduction générale

La pollution atmosphérique est un véritable problème mondial de santé publique. Elle est engendrée dans la plupart des cas spécialement dans les zones urbaines à grande concentration telles que les grandes villes par des activités humaines liées au trafic routier et aux denses productions industrielles. Elle est la cause principale de la mauvaise qualité de l'air que nous respirons quotidiennement et source actuelle d'innombrables cas de décès. En effet, dans un communiqué du 25 mars 2015, l'organisation mondiale de la santé (OMS) indique que près de **7 millions de personnes sont décédées prématurément en 2012** une sur huit au niveau mondial du fait de l'exposition à la pollution de l'air. Ces chiffres représentent plus du double des estimations antérieures et confirment que la pollution de l'air est désormais le principal risque environnemental pour la santé dans le monde. Des millions de vies peuvent être sauvés en luttant contre la pollution de l'air. Mais comment stopper ce fléau à l'échelle mondiale spécialement dans nos zones urbaines ? Quelles stratégies de contrôle de rejets adopter ?

Les processus physico-chimique qui président au devenir des polluants dans l'atmosphère sont complexes. Afin d'éviter ou du moins **minimiser** les phénomènes de pollutions et particulièrement les épisodes les plus aigus, on peut que jouer que sur un seul paramètre: **nos émissions**. Mais cet objectif à atteindre est loin d'être simple. Actuellement, notre monde ne dispose pas assez de technologies nous permettant de répondre à nos besoins énergétiques sans rejet dans l'environnement. On peut simplement réduire plus ou moins certain de ces rejets. C'est le point focal des discussions de la conférence mondiale sur le climat **COP21** qui se tient cette année 2015 à Paris en France. Les mesures de réductions ne doivent pas être entreprises à la légère. Cela nécessite de gros investissements financiers, et les solutions appliquées peuvent s'avérer fort peu efficaces. La modélisation se révèle l'outil indispensable d'aide à la prise de décision en matière de contrôle de rejets.

Dans cette thèse on s'intéresse à la modélisation et prédiction de la qualité l'air dans la ville de Dakar. On présente dans le chapitre 1 la modélisation de la pollution atmosphérique dans son ensemble avec un état de l'art sur les modèles et différentes méthodes existants. Dans le chapitre 2, on parle de la qualité de l'air dans Dakar, la zone d'étude avec les stations qui ont fourni des mesures de son indice au cours de la période de 2010 à 2013. Le chapitre 3 fournit les outils d'optimisation dont on a besoin pour modéliser et prédire l'indice de la qualité de l'air dans Dakar. Ces outils sont utilisés dans les chapitres 4 et 5 pour modéliser et prédire l'indice de la qualité de l'air dans Dakar respectivement par les techniques de la régression linéaire multiple (moindre carré ordinaire) et par le modèle ARMA (Auto-Regressive Moving Average). Il faut noter que les chapitres 4 et 5 ont débouchés respectivement sur deux articles de recherches scientifiques dont le premier publié dans le journal IJAMAS et le deuxième dans IJAM. Il faut noter que les journaux IJAM et IJAMAS sont tous indexés par **MathSciNet** et **Zentralblatt MATH**. Le premier article porte sur la modélisation et prédiction de l'indice de la qualité de l'air dans Dakar et le deuxième sur l'analyse et l'optimisation de l'indice de la qualité de l'air par le modèle $ARMA(2,1)$ en vue de sa prédiction. On conclut la thèse par une comparaison de ces deux modèles de prédictions développés dans les deux articles et on finit par des perspectives.

Chapitre 1

La modélisation de la pollution atmosphérique

Sommaire

1.1 Introduction

La modélisation de la pollution atmosphérique a fait l'objet de nombreuses études. Plusieurs modèles et méthodes ont été proposés. Avant de nous lancer le vif du sujet que veut dire le mot modèle ? C'est la représentation d'un processus naturel simplifié servant à représenter et à étudier un système complexe. La modélisation de la pollution atmosphérique constitue l'ensemble des méthodes et outils qui permettent d'obtenir une information sur la qualité de l'air en dehors des points où sont réalisées les mesures.

Une des problématiques des études en pollution atmosphérique consiste à prouver de combien doivent être réduites les émissions de polluants pour que les concentrations ambiantes puissent être maintenues en dessous des valeurs limites acceptables pour la santé et le milieu naturel. C'est l'un des buts principaux de la modélisation de la pollution: pouvoir calculer les concentrations de polluants dans l'atmosphère à partir des émissions des différentes sources de polluants.

L'évolution de l'informatique a permis l'utilisation des approches qui demandent une puissance importante de calcul pour la simulation et la prédiction de la pollution atmosphérique à grande

échelle (locale, régionale voir mondiale)(Cihan et al 2006, Gitte et al 2005). La simulation sur ordinateur est un outil précieux qui permet de confronter le monde politique et économique aux risques ou aux bienfaits encourus par une augmentation ou une diminution des sources d'émissions sur la qualité de l'air.

Le but recherché est de fournir des informations sur la pollution de l'air aux différents acteurs: autorité publique, utilisateur simple, expert ou chercheur dans le domaine.

Dans ce chapitre de la thèse, on présentera dans un premier temps l'histoire de la modélisation de la pollution atmosphérique, dans un second temps on va s'approprier de quelques vocabulaires tels que atmosphère, pollution,... Le chapitre présente aussi le lien entre la pollution de l'air et le réchauffement climatique. On donne une classification avec quelques modèles allant d'échelle locale à régionale. Il se termine par une application simple de l'équation d'advection. Cependant comment a évolué la modélisation de la pollution atmosphérique, quelle est donc son histoire ?

Définitions et Historique

Avant de faire un état de l'art sur la modélisation de la pollution atmosphérique on définit d'abord les termes suivants:

(a) Modélisation : c'est la représentation d'un système par un autre, plus facile à appréhender. Il peut s'agir d'un système mathématique ou physique. Le modèle sera alors numérique ou analogique. La modélisation numérique consiste à construire un ensemble de fonctions mathématiques décrivant le phénomène. En modifiant les variables de départ, on peut ainsi prédire les modifications du système physique. La modélisation analogique consiste à construire un système physique qui reproduit plus ou moins un phénomène que l'on souhaite étudier. L'observation du comportement du modèle permet de tirer des enseignements sur le phénomène d'intérêt. Dans la thèse on modélise numériquement l'indice de la qualité de l'air dans Dakar.

(b) Qualité de l'air : C'est l'évaluation de l'état de l'air ambiant selon une échelle dépendant du taux de concentration des polluants. Elle est souvent mesurée par une combinaison de méthodes chimiques et électroniques. Des sondes sont reliées à un système informatique qui enregistre automatiquement une quantité de valeurs à intervalle réguliers et qui peuvent ensuite être visualisées facilement sous diverses formes. Un indice de la qualité de l'air est Un indicateur de qualité de l'air permettant de synthétiser différentes données en une valeur simple est l'Indice de la Qualité de l'Air (IQA ou iqa).

(c) Prédiction : c'est l'action de prédire, annoncer ce qui va arriver. c'est aussi l'action de prévoir (Prévision). C'est annoncer à l'avance un événement par calcul, par raisonnement, par induction. Dans cette thèse on fait la prédiction de l'indice de la qualité de l'air dans Dakar.

(d) Advection, Diffusion : En générale, c'est le transport d'une quantité telle que la chaleur, l'énergie interne, un élément chimique quelconque, les charges électriques par le mouvement (la vitesse) du milieu environnant. C'est un déplacement d'une masse d'air dans le sens horizontal, ou proche de l'horizontal. La **convection** est réservée aux mouvements provoqués par la poussée d'Archimède. La **diffusion ou conduction** est le transport relatif par rapport au milieu environnant en mouvement. On parle aussi dans ce cas de l'équation d'advection et de diffusion des polluants.

(f) Dispersion : C'est un phénomène de déconcentration de polluants instables dans un milieu en l'absence de confinement ou en raison de brèche dans celui-ci. Le vent est un excellent agent dispersant des polluants. En effet, plus le vent est fort plus les niveaux de pollutions seront bas. Il faut noter que la dispersion atmosphérique comporte 3 évènements importants
- Émission : rejet artificiel, libération "naturelle" (active), mise en suspension par l'écoulement (passive)
- Transport par le vent dans l'atmosphère
- Dépôt au sol et dans l'hydrosphère [1] : sec ou humide.

1. L'hydrosphère est un terme désignant l'ensemble des zones d'une planète où l'eau est présente. Elle concerne aussi bien l'eau sous forme liquide, que solide ou sous forme gazeuse.

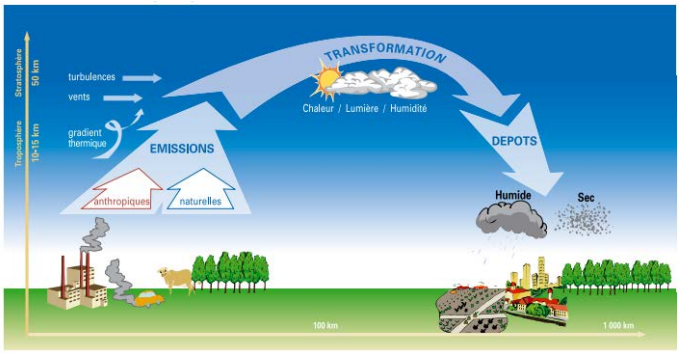


FIGURE 1.1 – Illustration de la dispersion atmosphérique. Source : ADEME

La modélisation de la pollution atmosphérique est la mise en œuvre d'équations physiques et chimiques pour récréer de manière numérique les phénomènes atmosphérique complexes. Cet outil de simulation numérique permet de prévoir, anticiper et analyser de façon objective les phénomènes atmosphériques sur toute ou une partie du territoire, en une période donnée.

L'histoire de la modélisation de la pollution photochimique est liée d'une manière étroite à celle de la modélisation numérique de l'atmosphère. Ces deux types de modélisations sont presque les mêmes avec les dernières générations de modèles couplés dit "on-line" chimie et météorologie sont manipulées simultanément avec rétroactions de l'une ou l'autre. Elles ont les repères chronologiques avec les premiers travaux théoriques de Bjerknes (1904) et Richardson (1922) puis les premières simulations dans les années 1950 de Charney et al. (1950) avec la puissance des moyens de calculs mais aussi certains objectifs. Cela se constate dans les prévisions à plus ou moins longue échéance, l' étude du changement climatique par la prescription des concentrations en gaz à effet de serre et en aérosols, problématique de l'assimilation.

Les modèles de boîtes destinés à mieux comprendre la formation du "smog" photochimique (modèle EKMA, Dodge (1977)) et leurs extensions au suivi de masses d'air lagrangiennes furent les tous premiers modèles.

La performance dans les calculs des modèles ont évoluées au cours des années 1950 à 1970. Les modèles de la qualité de l'air ont été étendus aux 3 dimensions géographiques avec une représentation en point de grille. Cela concerne les modèles eulériens et semi-lagrangiens qui traitent le transport, le dépôt, les émissions et la chimie gazeuse. Cette approche s'est basée au début sur la modélisation de la pollution photochimique urbaine à l'instar du premier CTM : Urban Airshed Model (UAM), conçu pour étudier la pollution à Los Angeles. On note au début une évolution des données météorologiques d'entrées en passant par des observations aux champs météorologiques numériques conçus ou analysés.

Autres que la pollution urbaine, les problèmes de pollution atmosphérique surgissent et transforment la formulation et la conception des modèles. On note la prise de conscience du caractère régional de la pollution photochimique ainsi que le phénomène des pluies acides (pollution transfrontière et transcontinentale). Cela a causé une l'extension horizontale des modèles urbains. Dans les années quatre vingt avec le phénomène du trou d'ozone stratosphérique a vu naitre les modèles bi puis tri-dimensionnels qui analyse la basse stratosphère en impliquant une extension verticale des modèles.

On note une variation dans la complexité des modèles et les résolutions des problèmes. Cela concerne les applications considérées et la puissance de calcul à la date considéré. Par exemple, actuellement, la résolution horizontale des CTM peut aller de quelques kilomètres pour les simulations régionales de la qualité de l'air à la centaine de kilomètres pour les problématiques plus globales.

A la fin des années 80, l'augmentation de la puissance de calcul et des bases de données sur les propriétés physiques et chimiques des particules et leurs interactions avec la phase gazeuse ont rendu possible la prise en compte des aérosols dans les modèles chimiques (Pilinis and Seinfeld (1988)).

Parallèlement au développement de ces modèles de chimie-transport, à partir des années 1990, des travaux portent sur les interactions et les rétroactions entre la dynamique turbulente et la chimie à des échelles sub-kilométriques par le biais de modèles de chimie on-line à méso- échelle (Yamartino et al. (1992)) et de modèles Large Eddy Simulation (LES) incluant des réactions chimiques : très peu à l'origine (une !) comme dans Schumann (1989) se complexifiant par la suite jusqu'à des mécanismes complets de photochimie troposphérique (Auger (2006)).

Notons que le développement poursuit son évolution jusqu'à présent.

1.2 L'atmosphère

Définie comme étant l'enveloppe gazeuse de la terre, l'atmosphère est l'environnement dans lequel la vie subsiste, beaucoup de transformations chimiques, en particulier d'origine photochimique surviennent à ce niveau. L'air qui est le fluide gazeux qui constitue l'atmosphère, est indispensable à la vie car il participe au processus de la respiration et à la photosynthèse des végétaux, cet environnement très sensible et qui subit le plus d'influence de la part de l'activité urbaine.

Composition chimique de l'atmosphère : Le tableau suivant nous donne les différentes concentrations des espèces constituant l'air et leur temps de résidence dans l'atmosphère.

Gaz	Concentration	Temps de résidence
Azote (N2)	78.084%	-
Oxygène (O2)	20.946%	-
Argon (Ar)	0.934%	-
Eau (H2O)	$[0.4..400] \times 102ppm$	10 jours
Dioxyde Carbone CO2	370 ppm	4 ans
Néon (Ne)	18.18 ppm	-
Hélium (He)	5.12 ppm	2 * 10 6 ans
Méthane (CH4)	1.75 ppm	10 ans
Krypton (Kr)	1.14 ppm	-
Hydrogène (H2)	0 .4 ppm	-
Xénon (Xe)	0.87 ppm	-

TABLE 1.1 – Composition chimique de l'air

Décomposition de la couche de l'atmosphère : L'atmosphère s'étend de la surface de la terre à plus d'une centaine de kilomètres. En fonction de l'altitude, l'atmosphère a des propriétés différentes, ce qui a permis de la "découper" en différentes épaisseurs successives (Figure 1.2). En ce qui concerne la météorologie liée à la pollution atmosphérique régionale, la zone d'intérêt sera la troposphère et plus particulièrement sa partie la plus basse, c'est à dire la couche limite. Un bon indicateur du type de couche atmosphérique où l'on se trouve est le comportement du gradient vertical de température : au sein de la troposphère, la température décroit régulièrement avec l'altitude (-6.5°C/km, en moyenne).

De nombreux ouvrages traitent de la couche limite atmosphérique. Nous citerons notamment ceux de Stull (1988) et de De Moor (1983). La troposphère est scindée en deux parties : une couche limite dans sa partie basse coiffé d'une couche limite libre (figure 1.2). La couche limite libre est la partie supérieure de la troposphère dans laquelle le vent est déterminé par de grands mouvements d'ensemble a l'échelle de la planète. Il résulte de l'équilibre entre les forces de pression et la force de Coriolis due a la rotation de la Terre. Le vent est appelé dans cette zone vent géostrophique.

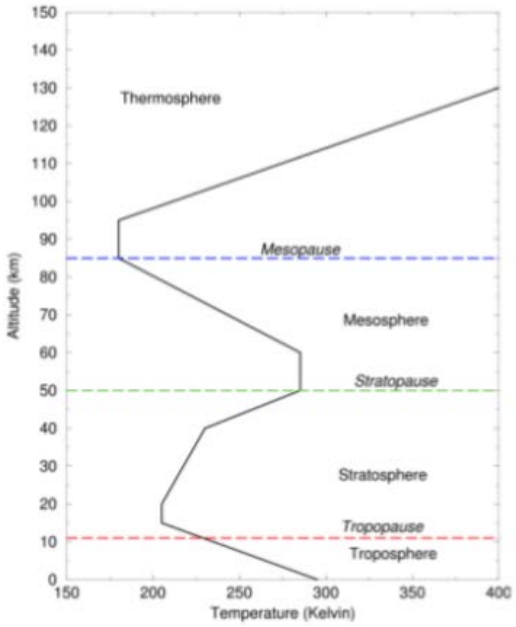


FIGURE 1.2 – Description schématique de la structure verticale de l'atmosphère

La couche limite atmosphérique : On s'intéresse à la couche limite atmosphérique, qui, dans de notre thèse, est le siège de la pollution, de la plupart des sources et puits ..., l'endroit où nous vivons et donc l'air que nous respirons. La couche limite atmosphérique (CLA) constitue

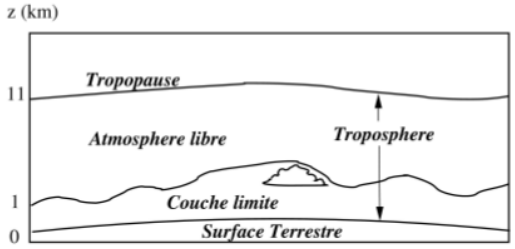


FIGURE 1.3 – Coupe verticale de la basse atmosphère

l'interface entre la surface terrestre et la troposphère libre (Figure 1.3). C'est la partie inférieure de l'atmosphère qui est sous l'influence directe des processus terrestres. Son extension verticale va de la surface à quelques centaines ou milliers de mètres d'altitude et dépend directement de tous les paramètres météorologiques (vent, température, humidité, insolation), mais aussi de la topographie et du type d'environnement (océan, continental rural ou urbain). La CLA est très mince (1 à $2km$) en comparaison avec le reste de la troposphère ($\approx 11km$) et avec toute l'atmosphère.

La notion d'épaisseur de la couche limite atmosphérique n'a pas de réalité physique instantanée ! Contrairement à d'autres variables bien plus palpables, comme le vent ou la température, l'épaisseur de la CLA est un bilan statistique. On peut mesurer instantanément une température, mais on ne peut que moyenner des grandeurs physiques pour en déduire une valeur moyenne de la hauteur de la couche limite.

Les processus principaux au sein de la couche limite sont des processus de transport de quantité de mouvement, de chaleur et d'humidité. Ces processus sont avant tout des flux, des processus turbulents.

Étudier la CLA revient avant toute chose à étudier la turbulence atmosphérique. Les processus

au sein de la CLA seront donc toujours décrits et étudiés en un terme moyen et un terme turbulent.

La couche limite urbaine : Elle caractérise l'évolution d'un ensemble de masses pour des espaces fortement construits. Le développement de la couche limite étant avant tout dirigé par les caractéristiques de la surface. Il apparait qu'entre la ville et la campagne, la basse atmosphère a un comportement aussi différent que peuvent l'être des milieux urbains et ruraux. La figure 1.4 schématise une ville entourée d'un milieu sub-urbain (la banlieue) ou rural. Sur cette figure sont regroupés tous les processus que l'on peut étudier dans la couche limite, qu'ils soient dynamiques ou chimiques. On ne peut pas parler de la couche limite sans parler des phénomènes suivants :

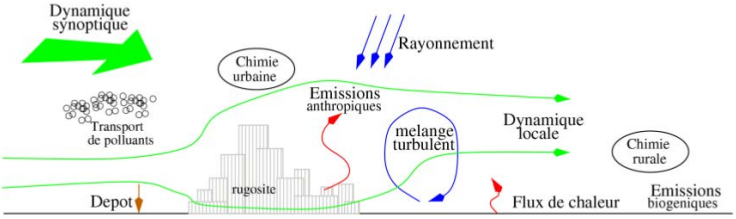


FIGURE 1.4 – Schéma d'un environnement urbain et de tous les processus inter-agissant dans la CLA.

(1) L'albédo et l'émissivité L'albédo ou l'effet réfléchi est une valeur physique qui permet de connaitre la quantité de lumière solaire incidente réfléchie par une surface. Du point de vue climatique, elle exprime la part de rayonnement solaire qui va être renvoyée par l'atmosphère et la surface terrestre vers l'espace et qui donc ne servira pas à réchauffer la planète. Elle est plus faible en ville qu'en milieu rural. De plus, la présence importante d'aérosols en milieu urbain tend à réduire le flux solaire incident de l'ordre de 10 à 20 % (en fonction des zones d' études). **L'émissivité** est la capacité d'une matière à émettre et à absorber du rayonnement. L'albédo concerne le rayonnement solaire tandis que L'émissivité est relative aux radiations émises par la terre. Le tableau 1.2 donne quelques valeurs d'albédo et d'émissivité, en fonction du type de milieu .

Type de sol	Herbes courtes (2cm)	Herbes longues (1cm)	Forêt	Ville
Albédo (A)	0,26	0,16	0,20	0,1-0,27
Emissivité (ϵ)	0,9-0,95	0,97	0,97	0,85-0,95

TABLE 1.2 – Exemples de valeurs d'albédo et d'émissivité en fonction du terrain.

(2) La température de surface : En générale la ville a une température plus élevée que la campagne. Cela s'explique principalement par les chauffages, le trafic automobile, les industries, ... Ce phénomène est accentué en saison sèche (Hiver). On définit alors un ilot de chaleur urbain . Il existe des relations empiriques établies pour tenter d'estimer cette différence thermique ΔT. On a notamment les relations (Bornstein (1987)) :

$$\Delta T = 15,27 - 13,87\Psi_s \tag{1.1}$$

$$\Delta T = 7,45 + 3,97\ln(\frac{H}{W}) \tag{1.2}$$

où Ψ_s désigne un rapport d'aspect des constructions et H la hauteur, W la largeur des constructions. En période diurne, une valeur "critique" de vitesse de vent est définie au delà de laquelle la formation de cet ilot ne peut se faire (Oke (1987)). Cette valeur est dépendante de la population (en millions d'habitants), telle que : $U_c = 3,4\ln(P) - 11,6$. Par exemple, considérant que la population de Dakar et de sa proche banlieue est de l'ordre de 3 millions (recensement de 2013), on obtient $U_c \approx 10,42ms^{-1}$. Cette valeur de vent étant une valeur

moyenne à $10m$. L'impact net est un flux de chaleur supplémentaire nommé flux anthropique, et qui vient s'ajouter au bilan radiatif que l'on fait sur un milieu.

(3) Le vent: Pendant le jour la norme du vent est plus faible en milieu urbain qu'en milieu rural car la présence de hautes constructions le ralentit. A cause de l'influence des forts gradients de température horizontaux, on a l'effet inverse la nuit.

Notons que par un vent moyen (au-dessus de $3m/s$), la couche limite atmosphérique (CLA) prend la forme d'un panache (figure 1.5), et l'influence de l'agglomération est alors ressentie par les villes se trouvant sous le vent provenant de cette agglomération.

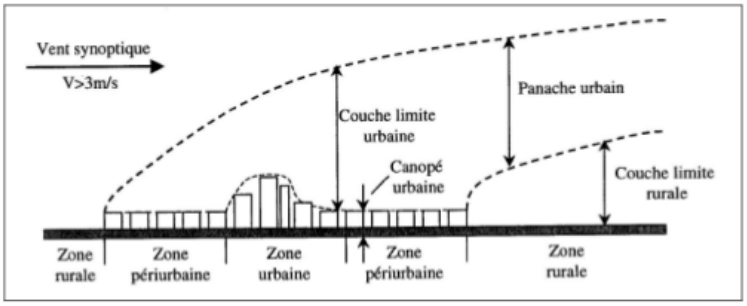


FIGURE 1.5 – Le "panache" urbain : structure verticale de l'atmosphère urbaine pour des vents supérieurs à 3m/s (Ringenbach (2004), d'après Mestayer et Anquetin (1995)).

(4) L'humidité: L'analyse de l'air montre que le milieu urbain est généralement plus sec qu'en milieu rural. Cela s'explique par le type de surface de la terre et à son revêtement. Cette différence ne se limite pas qu' à l'atmosphère : le sous-sol urbain est "creux" ce qui est rarement le cas de la terre. On observe donc des différences dans la façon dont l'humidité peut être stockée dans le sous-sol, moins longtemps en ville, et dont elle peut être restituée, beaucoup plus facilement et intensément en ville.

Brise de mer: Les différences de propriété de stockage de chaleur, de flux et d'assèchement du sol induisent des forts gradients de température horizontaux entre le milieu urbain et le milieu rural. La nuit, la ville est nettement plus chaude que la campagne environnante : on a un effet de brise similaire à la transition terre/mer : l'air froid rural va s' écouler vers la ville à basse altitude, puis repartir vers la campagne à des altitudes supérieures (Figure 1.6). Contrairement au cas précédent de couche interne, le phénomène de brise ne peut se produire que pour des vents très faibles : ce n'est qu'à cette condition que les effets thermiques peuvent dominer les forçages dynamiques : des études ont montrées que cela se produisait pour de grandes villes uniquement si $|U| < 3m/s$.

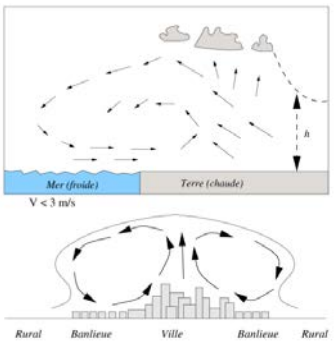


FIGURE 1.6 – Schéma de brise de mer et analogie avec la ville

1.3 pollution

Jonathan Raban cité par Claude Gagnière dans le Bouquin des citations (Robert Laffont, 2000) affirme que : "dans un pays sous-développé, ne buvez jamais d'eau. Mais, dans un pays industrialisé, ne respirez jamais l'air." D'après la Loi sur l'Air et l'Utilisation Rationnelle de l'Énergie du 30 décembre 1996(LAURE, loi numéro 96-1236), "constitue une pollution atmosphérique [...] l'introduction par l'homme, directement ou indirectement, dans l'atmosphère et les espaces clos, de substances ayant des conséquences préjudiciables de nature à mettre en danger la santé humaine, à nuire aux ressources biologiques et aux écosystèmes, à influer sur les changements climatiques, à détériorer les biens matériels, à provoquer des nuisances olfactives excessives" (`www.legifrance.gouv.fr`). On définit la pollution atmosphérique comme étant la présence indésirable d'impuretés ou l'élévation anormale de la proportion des certains constituants de l'atmosphère. Toutes les substances qualifiées de polluants atmosphériques ne sont pas étrangères à l'atmosphère et elles peuvent aussi avoir des conséquences positives pour la vie, comme l'ozone présent dans la stratosphère ou le dioxyde de carbone. Une substance présente dans l'atmosphère devient donc un polluant si sa concentration est modifiée de telle sorte que "des conséquences préjudiciables" apparaissent. Les polluants ne sont pas nécessairement différents des substances émises par des phénomènes naturels. Si les CFC (chlorofluorocarbones) sont produits et émis uniquement par l'homme, des activités humaines comme l'agriculture ou l'élevage conduisent à l'émission de méthane ou de poussières qui sont aussi naturellement émises par les bactéries des marais ou les déserts.

Remarquons que la définition donnée par la LAURE prend en compte les différentes échelles spatiotemporelles auxquelles la pollution atmosphérique peut survenir :

1. la macro-échelle, qui est celle des phénomènes globaux comme les changements climatiques,
2. la méso-échelle, celle des phénomènes qui concernent un continent ou une région comme les pluies acides qui nuisent aux ressources biologiques et aux écosystèmes,
3. la micro-échelle, qui est celle des phénomènes locaux comme la détérioration des biens matériels, par exemple les monuments, ou les nuisances olfactives excessives, notamment sous le vent de certains types d'usines.

La pollution est à l'origine de deux risques : le premier se situe à l'échelle locale ; ces composants chimiques (hydrocarbures non brûlés et oxydes d'azote) favorisent en effet la formation d'ozone sous l'effet du soleil. Le second se situe à l'échelle globale, l'émission croissante de ces polluants favorise les effets de serre et donc l'augmentation globale de température. Détaillons ces principaux composants chimiques polluants ainsi que leurs effets ; le premier de ces composants, le dioxyde de carbone CO_2 est dégagé lors de la combustion du charbon du gaz naturel et du pétrole destiné entre autres à la production d'énergie. Les gaz à effet de serre de type Chloro-fluoro-carbones se trouvent dans les aérosols, la combustion d'emballages plastiques, les réfrigérateurs et climatiseurs. Le monoxyde de carbone CO, le dioxyde d'azote N_2 et de plomb Pb_2 émis par les voitures à essence non catalysées sont massivement produits dans les embouteillages, les tunnels urbains et par temps froid. On ne saurait faire un panorama complet de ces substances en oubliant l'ozone O_3 formée par les polluants atmosphériques et dioxyde de soufre qui trouve sa source dans les diverses activités industrielles et qui contribue à la formation de pluies acides. Elle cause aussi des dégâts sur la santé humaine (irritation des yeux et de la gorge, dégradation de la capacité pulmonaire).

Pollution de l'air, une des principales causes de décès dans le monde L'organisation mondiale de la santé (OMS) indique dans son communiqué de presse du 25 mars 2015 [1] que près de 7 millions de personnes sont décédées prématurément en 2012, une sur huit au niveau mondial, du fait de l'exposition à la pollution de l'air (voir tableau 1.3). Ces chiffres représentent plus du double des estimations précédentes et confirment que la pollution de l'air est désormais le principal risque environnemental pour la santé dans le monde. Des millions de vies peuvent être sauvés en luttant contre la pollution de l'air.

Décès liés à la pollution extérieur et intérieure des habitations Il faut faire la distinction entre la mortalité induite par la pollution de l'air intérieur et extérieur. Souvent on pense que la

Région	Mortalité (en millions)	Mortalité pour 100 000 habitants
Pacifique ouest	2,868	102
Asie du Sud-est	2,3	124
Afrique	0,68	76
Europe	0,582	77
Est de la Méditerranée	0,414	50
Amérique centrale et du Sud	0,131	22
Amérique du Nord	0,096	25
Monde	7,071	100

TABLE 1.3 – 7 millions de morts par an dans le monde, près de 700 000 en Afriques

pollution provient de l'extérieur mais nos intérieurs sont également fortement pollués. Dans les pays à faible revenu, l'intérieur du logement subit la pollution d'équipements de chauffage et de cuisson rudimentaires.

Le tableau 1.4 suivant nous donne un lien entre la pollution de l'air à l'intérieur des habitations et de l'air à l'extérieur puis les maladies cardio-vasculaires ainsi que la pollution de l'air et le cancer. A cela vient s'ajouter le rôle de la pollution de l'air dans l'apparition de maladies respiratoires et notamment d'infections respiratoires aiguës et de bronchopneumopathies chroniques obstructives [1].

Décès dus à la pollution extérieure	
40 %	cardiopathies ischémiques
40%	accident vasculaire cérébral
11%	bronchopneumopathies chroniques obstructives (BPCO)
6%	cancer du poumon
3%	infections aiguës des voies respiratoires inférieures chez l'enfant
Décès dus à la pollution intérieure	
34%	accident vasculaire cérébral
26%	cardiopathies ischémiques
22%	bronchopneumopathies chroniques obstructives
12%	infections aiguës des voies respiratoires inférieures chez l'enfant
6%	cancer du poumon

TABLE 1.4 – Tableau de décès dus à la pollution intérieure et extérieur des habitations (OMS, 2014)

Pollution de l'air et réchauffement climatique

La température moyenne de la planète s'est élevée de 0.6℃ au XXème. Au cours du siècle à venir, elle devrait s'accroître d'au moins de 1.4℃, et jusqu'à 5.8℃ si on ne fait rien. Cette évolution, considérable, est d'une ampleur sans précédent depuis des dizaines de milliers d'années. Il est établi aujourd'hui avec certitude que ce phénomène tient à l'augmentation des émissions de gaz à effet de serre liées aux activités humaines, à commencer par le dioxyde de carbone (CO_2). Mais qu'est ce que l'effet de serre ?

L'effet de serre est un phénomène naturel. L'énergie solaire qui parvient au sol réchauffe la terre et se transforme en rayons infrarouges. De la même façon que les vitres d'une serre d'où le nom donné à ce mécanisme, les gaz présents dans l'atmosphère piègent une partie de ces rayons qui

tendent à la réchauffer. Ainsi, sans effet de serre, la température moyenne sur la Terre serait de -18℃ et peu d'eau serait sous forme liquide. Cet effet a donc une influence bénéfique puisqu'il permet à notre planète d'avoir une température moyenne de 15℃, et donc la vie sur terre. Depuis le début

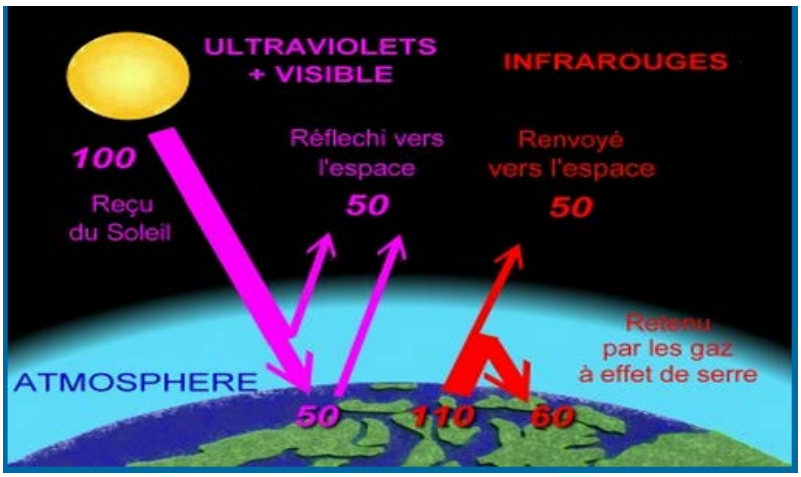


FIGURE 1.7 – Effet de serre. Source : Actions Vivres

de l'ère industrielle, l'homme a rejeté dans l'atmosphère des gaz (gaz carbonique, méthane, oxydes d'azote,...) qui augmentent artificiellement l'effet de serre. Si cet ajout à l'effet de serre naturel est faible (environ +1 %), il est amplifié par la vapeur d'eau et a ainsi contribué à l'augmentation de la température moyenne de notre planète d'environ 0.6℃ observée dans la seconde moitié du vingtième siècle.

Les dernières années ont donné quelques aperçus des risques que feraient courir le changement climatique aux continents: même s'il n'est généralement pas possible d'attribuer tel ou tel événement météorologique extrême (tempête, inondation, vague de chaleur,...) au dérèglement climatique, les faits observés matérialisent fidèlement les résultats du Groupe d'Experts Intergouvernemental sur l'Evolution du Climat (GIEC). Certains effets du dérèglement climatique sont d'ailleurs déjà visibles dans le monde: élévation plus de 0,9℃ en un siècle de la température moyenne annuelle et retrait des glaciers. A très long terme, des perturbations importantes pourront également intervenir dans les courants marins et les glaces polaires, avec des conséquences sur la répartition du réchauffement climatique selon les régions du globe.

1.4 Applications de la modélisation de la qualité de l'air

Les résultats de modélisation constituent un puissant moyen d'aide à la décision pour l'élaboration du déplacement urbain et pour la planification dans la gestion de l'aménagement urbain. Les différents types de la modélisation de la qualité de l'air ont un grand nombre d'applications pratiques:

1. la contribution à une meilleure **interprétation** des concentrations de polluants mesurées,
2. l'établissement d'une **cartographie** de la pollution,
3. l'**aide à la décision** dans le choix d'une **stratégie de contrôles des rejets** de polluants primaires,
4. le **suivi des masses d'air** contaminées en cas d'accidents industriels,
5. l'**évaluation de l'impact d'une source** industrielle ou de l'implantation d'une nouvelle source sur la qualité de l'air de la région,
6. en **mode prédictif**: la possibilité d'éviter les épisodes de pollution ou de prévenir les alertes.

En générale, on a pas dans la réalité toutes les mesures sur la qualité de l'air. Mais comment remédier au manque de mesures ? Les réseaux de surveillance de la qualité de l'air assurent,

comme leur nom l'indique, le suivi de la qualité de l'air dans leur région d'implantation, et sont chargés, entre autres missions, de l'information de la population. Or, il n'est bien entendu pas envisageable d'installer des stations de mesures partout. La modélisation, qui peut permettre d'obtenir la répartition spatiale et temporelle des différents polluants sur l'ensemble du domaine, s'avère alors très intéressante en complément de compagnes de mesures. On peut utiliser également la modélisation pour l'**optimisation** du réseau de mesures, comme outils d'aides à la décision dans les choix des lieux d'implantation des stations fixes.

1.5 Les étapes d'application d'un modèle de qualité de l'air

Pour la mise en place d'une stratégie de contrôle parfois très coûteuse, il est nécessaire d'abord de tenter d'évaluer son efficacité grâce à un modèle mathématique dit "de qualité de l'air". C'est un modèle qui calcule les variations de concentrations de différents polluants pour une région donnée, en simulant les processus physiques et chimiques de l'atmosphère. L'application d'un modèle de qualité de l'air doit se procéder en deux phases (figure1.8):

1. Validation du modèle,
2. application du modèle à l'évaluation d'une stratégie de contrôle.

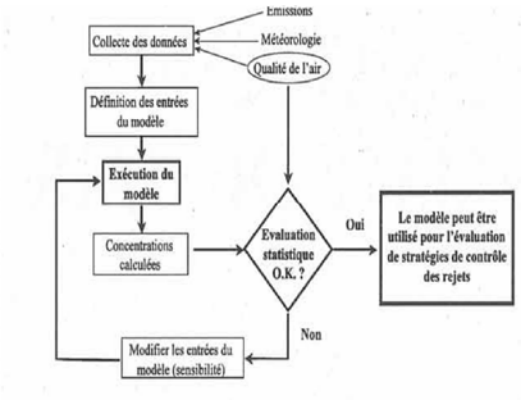


FIGURE 1.8 – Application optimale d'un modèle de qualité de l'air, Zanetti (1990).

La phase de la validation de nouveaux modèles nécessite l'existence d'une bases de données très bien documentée couvrant plusieurs épisodes de pollution. Mais cela constitue un véritable problème à Dakar et même voir en France. Pour une modélisation performante, il est très important d'apporter un grand soin à la collecte de données d'entrée de qualité. Puisque quelle que soit la qualité intrinsèque du modèle, si on a une grande erreur sur les mesures, il ne donnera pas de bons résultats. Mais c'est pas très fréquent d'avoir des champs de données d'entrée de grande qualité [2, 3]. Pour pallier à ce manque, on fait souvent des hypothèses. Le choix des données d'entrée est plus déterminant que le choix du modèle lui-même [4]. L'élaboration des hypothèses et le choix des données sont les étapes délicates dans la démarche de modélisation.

1.6 Les classes des modèles de la qualité de l'air

D'après Zanetti (1990), on divise les modèles de qualité de l'air en plusieurs catégories :

1. les modèles **physiques:** représentations à petite échelles, en laboratoire, de certains phénomènes tels que les vents de tunenel ;
2. les modèles **mathématiques:** des algorithmes numériques ou analytique décrivent les aspects physiques et chimiques du problème étudié.

Les modèles physiques fournissent dans la plus part des cas des données aux concepteurs de modèles mathématiques, tels les mécanismes chimiques élaborés à partir d'expériences en chambre de simulation. Dans cette thèse, on ne traitera pas les modèles physiques mais les modèles mathématiques sont au cœur de ce travail.

D'après Nicolas Moussiopoulos dans son livre titré << *Air Quality in Cities* >> [5], les modèles mathématiques de qualité de l'air peuvent être classés selon:

(α) l'échelle spatiale (globale, régionale à continentale,locale à régionale, locale),

(β) l'échelle temporelle (épisodes de pollution ou comportement à long terme). Selon le type de pollution étudiée, il conviendra de choisir l'échelle spatiale et temporelle adéquate pour la modélisation. Le tableau ci-dessous indique les différentes échelles spatiales correspondant aux types de pollution étudiés. L'échelle temporelle est liée en partie à l'échelle spatiale

	Échelle du phénomène de dispersion			
Type de pollution	Globale	Régionale à Continentale	Locale à Régionale	Locale
Changements Climatique	×			
Disparition de l'ozone stratosphérique	×	×		
Ozone troposphérique		×		
Acidification		×		
Smog d'été		×	×	
Smog d'hiver		×	×	
Qualité de l'air urbain			×	
Polluants industriels			×	×
accidents nucléaire		×	×	×
Accident chimique		×	×	×

TABLE 1.5 – Échelles spatiales des différents types de pollution atmosphérique,(extrait de Moussiopoulos et al.,1996).

choisie. Lorsque l'on travaille à l'échelle du globe, il convient de raisonner sur des échelles de temps assez longues (quelques semaines à quelques mois, voir quelques années). A l'échelle locale, inversement, les périodes de temps considérées seront très courtes (quelques heures à 1 ou 2 jours).

(γ) le traitement des équations de transport (lagrangien ou eulérien),

(δ) le traitement des différents processus (chimie, dépôts sec et humide). On appelle **dépôt** est un puits par lequel les polluants quittent l'atmosphère pour la surface du sol, des bâtiments ou de la végétation. Le dépôt humide a lieu en présence de précipitations qui **lessivent** [2] l'atmosphère en entraînant les polluants vers les surfaces. La grande variabilité des précipitations rend complexe la quantification du dépôt humide ; celui-ci dépend de plus de la solubilité de l'espèce dans la neige ou la pluie (qui varie en fonction de la température et du pH), de la taille des gouttes et de leur nombre (Finlayson-Pitts et Pitts 2000). Le dépôt sec (en l'absence de précipitations) dépend principalement du degré de turbulence de l'atmosphère, des caractéristiques de la surface et des propriétés chimiques des polluants en contact avec celle-ci. Son mécanisme peut être décomposé en trois étapes :

1. une étape aérodynamique qui est le transfert du gaz vers la couche laminaire c'est-à-dire celle qui est en contact avec la surface,

2. Le lessivage, ou dépôt humide, correspond à la perte de masse due aux nuages et à la pluie (dilution et entraînement dans les gouttes d'eau en suspension dans les nuages, et dans les gouttes de pluie). La capacité d'une espèce à être lessivée est modélisée par un coefficient de lessivage Λ. Pour plus d'information voir les références [6, 7, 8]

2. une étape de diffusion, à travers la couche laminaire et
3. une étape de piégeage du gaz par la surface. La vitesse de dépôt d'un gaz sur une surface est souvent estimée grâce à un modèle analogue à celui de résistance électrique, les résistances correspondant aux trois étapes du processus étant en série (Wesely 1989 ; Finlayson-Pitts et Pitts 2000).

(ϵ) le degré de complicité de l'approche.

Les modèles mathématiques sont divisés en deux grands types:
- les modèles **déterminismes:** basés sur la description mathématique des processus atmosphériques,
- les modèles **statistiques:** des relations semi-empiriques sont établies à partir d'un grand nombre d'observations.

1.6.1 Méthodes déterministes

Les méthodes déterministes expliquent ou à prévoient un phénomène de pollution à partir des mécanismes connus qui le régissent. Ces mécanismes sont traduits sous forme d'équation dans des modèles et permettent de simuler le fonctionnement du phénomène considéré (simulation numérique, modèle mécanique ou fonctionnel). Ces deux approches sont utilisées conjointement de plus en plus afin de prendre en compte à la fois des connaissances à priori des phénomènes de pollution et la réalité du terrain retranscrite par les appareils de mesures de la qualité de l'air.

Depuis plus d'une vingtaine d'années, les modèles mathématiques de qualité de l'air sont utilisés pour étudier le comportement des espèces traces dans l'atmosphère à différentes échelles. D'innombrables méthodes ont été utilisées au fil des années, depuis les techniques très simplistes telles le "**linear rollback**", qui supposait une relation de proportionnalité directe entre émissions et niveaux de pollution, jusqu'aux modèles eulériens de troisième génération, s'efforçant de décrire le plus fidèlement possible les phénomènes physico-chimiques ainsi que leurs interactions.

1.6.2 Méthodes statistiques

Elles visent à expliquer ou à prévoir un phénomène de pollution à partir d'observations enregistrés par des appareils de mesure de la qualité de l'air. Les méthodes statistiques permettent d'extraire de l'information contenue dans un ensemble d'observations du phénomène afin d'en décrire son fonctionnement (géostatistique, modèle d'apprentissage). Elles sont très nombreuses mais leur but est de construire des modèles. On peut citer quelques principaux modèles: le modèle linéaire (gaussien) de base, le modèle linéaire généralisé, les modèles non linéaires, les modèles mixtes, les modèles pour données répétées, les modèles pour séries chronologiques, l'analyse discriminante et la classification, les modèles par arbre binaire de régression et de classification. Dans cette thèse, on utilise le modèle linéaire de base en appliquant la régression linéaire simple et multiple puis les modèles pour les séries chronologique en appliquant l'outil des processus ARMA. Mais avant cela, on introduit les modèles gaussiens.

1.7 Les modèles gaussiens

On dénote deux types de modèles gaussiens. Pour le premier type on a le modèle de panache, ou gaussien stationnaire, modélise le panache émis par une source ponctuelle par une distribution gaussienne dans deux directions (horizontale perpendiculaire au vent, et verticale), et suppose une météorologie stationnaire. Le deuxième type est le modèle à bouffées qui modélise une émission instantanée par une bouffée gaussienne dans les trois directions (pour plus de détailles voir les documents [9, 10, 11]).

1.7.1 Modèles gaussiens et formulation eulérienne

Pour modéliser le comportement de la trace d'une espèce dans l'atmosphère, c'est- à-dire sa distribution spatiale et temporelle, on a deux approche. D'une part, **l'approche eulérienne**

consiste à décrire cette distribution dans un référentiel fixe, en fonction des caractéristiques du fluide porteur en un point donné. D'autre part **l'approche lagrangienne** décrit le comportement statistique d'un groupe de particules en déplacement en nous plaçant dans le référentiel du fluide qui se meut. Ces deux approches fournissent des formulations différentes, qui peuvent être reliées entre elles. Le modèle gaussien à bouffées peut être considéré comme un modèle lagrangien simplifié, dans la mesure où l'on "suit" le polluant émis sur sa trajectoire: la distribution statistique d'un grand nombre de particules est simplifiée et modélisée par une distribution gaussienne appelée bouffée. On présente ici la description eulérienne de la dispersion atmosphérique d'un polluant, et la façon dont la description gaussienne des sources ponctuelles s'en déduit. Pour plus de détails sur les calculs décrits ci-dessous, on se référera à Seinfeld et Pandis [12] dont provient l'inspiration de cette partie.

1.7.2 Approche eulérienne

L'approche eulérienne de la concentration c d'une espèce non réactive dans l'atmosphère est représentée par l'équation d'advection-diffusion:

$$\frac{\partial c}{\partial t} = -\nabla uc + \frac{\partial (K \frac{\partial c}{\partial z})}{\partial z} + \frac{\partial c(c_1, c_2, \ldots, c_n, T, J)}{\partial z} + u_d c + E \tag{1.3}$$

où c est la concentration du polluant, $u_d c$ = déposition sèche

$$\nabla uc = transport, \qquad \frac{\partial c(c_1, c_2, \ldots, c_n, T, J)}{\partial z} = Chimie,$$

$$\frac{\partial (K \frac{\partial c}{\partial z})}{\partial z} = diffusion\ turbulente, \qquad E = Emission,$$

K est la matrice de diffusion turbulente qui peut être utilisée pour fermer l' équation et représenter les termes du second ordre (fermeture du premier ordre appelée **théorie-K**). On suppose que la diffusion turbulente est très grande devant la diffusion moléculaire (négligeable). La matrice K est inconnue, et doit être estimée par des paramétrisations empiriques, où l'on suppose que les termes extra-diagonaux sont négligeables. Elle s'écrit alors

$$\begin{pmatrix} K_x & 0 & 0 \\ 0 & K_y & 0 \\ 0 & 0 & K_z \end{pmatrix}$$

Un tel modèle exige l'introduction de données météorologiques, des émissions, de la topographie, etc. De plus la résolution des équations peut entraîner des erreurs numériques qui peuvent se propager jusqu'aux résultats finaux. Ces incertitudes sont souvent le fruit d'un mauvais choix des conditions initiales et aux limites. Il faut encore signaler que les cellules d'un modèle eulérien peuvent se déformer. Ce qui permet de mieux suivre des phénomènes qui ne sont pas toujours constant dans l'espace.

L'équation de chimie-transport (1.3) (appelée aussi équation de diffusion atmosphérique ou équation de conservation de la masse) est utilisée pour représenter l'évolution des concentrations de polluants en fonction de l'espace et du temps.

On peut encore mieux écrire (1.3) dans l'espace l'équation de la manière suivante:

$$\frac{\partial c}{\partial t} + \underbrace{u_x \frac{\partial c}{\partial x} + u_y \frac{\partial c}{\partial y} + u_z \frac{\partial c}{\partial z}}_{Advection} = \underbrace{\frac{\partial (K_x \frac{\partial c}{\partial x})}{\partial x} + \frac{\partial (K_y \frac{\partial c}{\partial y})}{\partial y} + \frac{\partial (K_z \frac{\partial c}{\partial z})}{\partial z}}_{Diffusion}$$

$$+ \underbrace{R(c_1, c_2, \ldots, c_n)}_{Chimie} + \underbrace{E(x, y, z, t)}_{Sources} - \underbrace{S(x, y, z, t)}_{Puits}$$

On se place à présent dans le cas d'une seule source ponctuelle, de coordonnées $(0\,;\, y_s\,;\, z_s\,)$,

émettant une masse totale Q de façon instantanée à $t = 0$, et sans processus de pertes. On considère une situation météorologique constante et homogène, avec un vent moyen $u = (\overline{u}\,;\,0\,;\,0)$.

On considère enfin que les coefficients de diffusion K_x, K_y et K_z sont constants. L'équation (1.3) s'écrit alors

$$\frac{\partial c}{\partial t} + \overline{u}\frac{\partial c}{\partial x} = K_x\frac{\partial^2 c}{\partial x^2} + K_y\frac{\partial^2 c}{\partial y^2} + K_z\frac{\partial^2 c}{\partial z^2}. \tag{1.4}$$

avec, pour condition initiale

$$c(x, y, z, 0) = Q\sigma(x)\sigma(y - y_s)\sigma(z - z_s) \tag{1.5}$$

et pour conditions aux limites

$$c(x, y, z, t) = 0 \qquad x, y, z \longrightarrow \pm\infty \tag{1.6}$$

Pour résoudre cette équation, on écrit la concentration sous la forme

$$c(x, y, z, t) = QG_x(x, t) \times G_y(y, t) \times G_z(z, t) \qquad avec \tag{1.7}$$

$$G_\alpha(\alpha, 0) = \sigma(\alpha - \alpha_s) \qquad \alpha\epsilon\{x, y, z\}$$

Cette équation a pour solution analytique

$$G_\alpha(\alpha, t) = \frac{1}{2(\pi t K_\alpha)^{\frac{1}{2}}} \times \exp(\frac{-(\alpha - \alpha_s - V_\alpha t)^2}{4K_\alpha t}), \quad \alpha\epsilon\{x, y, z\} \tag{1.8}$$

avec $x_s = 0$, $u_x = \overline{u}$ et $u_y = u_z = 0$. On a donc finalement une expression de la concentration [13].

$$c(x, y, z, t) = \frac{Q}{8(\pi t)^{\frac{3}{2}}(K_x K_y K_z)^{\frac{1}{2}}} \times \exp(-\frac{(x - \overline{u}t)^2}{4K_x t})\exp(\frac{-(y - y_s)^2}{4K_y t})\exp(\frac{-(y - y_s)^2}{4k_z t}) \tag{1.9}$$

Modèle eulérien d'une source continue Soit une source ponctuelle, de coordonnées $(0, y_s, z_s)$, ayant un débit massique constant Q_s , et sans pertes processus de pertes. On suppose une météorologique constante et homogène, avec un vent moyen $V = (\bar{u}, 0, 0)$. On suppose que la turbulence K est constante et homogène. L'équation (1.3) est donc stationnaire et devient

$$\bar{u}\frac{\partial c}{\partial x} = k\left(\frac{\partial^2 c}{\partial x^2} + \frac{\partial^2 c}{\partial y^2} + \frac{\partial^2 c}{\partial z^2}\right) + Q_s\delta(x)\delta(y - y_s)\delta(z - z_s) \tag{1.10}$$

où les condition aux limites sont données par :

$$c(x, y, z, t) = 0 \;\; quand\; x, y, z \longrightarrow \pm\infty \tag{1.11}$$

On peut se résoudre cette équation en utilisant la variable $r^2 = x^2 + (y - y_s)^2 + (z - z_s)^2$. On a :

$$c(x, y, z) = \frac{Q_s}{4\pi \mid K \mid r} \times \exp\left(\frac{-\bar{u}(r - x)}{2k}\right) \tag{1.12}$$

on fait l'ajout ici d'une hypothèse supplémentaire, en supposant que le panache est étroit dans la direction perpendiculaire au vent y ($\ll$ slender plume approximation $\gg$). On a donc $x^2 \gg (y - y_s)^2 + (z - z_s)^2$, et la distance r s'écrit :

$$r \simeq x\left(\frac{1 + (y - y_s)^2 + (z - z_s)^2}{2x^2}\right) \tag{1.13}$$

L'approximation de panache étroit consiste à faire l'hypothèse que le vent (suivant la direction x) est largement grand par rapport à la turbulence. Par conséquent la diffusion turbulente

dans ce sens est négligeable. Donc on peut remplacer dans l'équation 1.12, r par x et $r-x$ par $\frac{(y-y_s)^2+(z-z_s)^2}{2x}$. Finalement, la procédure ci-dessus décrite se généralise aisément au cas de turbulence non homogène $K_x \neq K_y \neq K_z$. On a alors :

$$c(x,y,z,t) = \frac{Q_s}{4\pi(K_yK_z)^{1/2}x} \times \exp\left(-\frac{\bar{u}(y-y_s)^2}{4K_yx}\right)\exp\left(-\frac{\bar{u}(z-z_s)^2}{4K_zx}\right) \tag{1.14}$$

1.7.3 Représentation gaussienne de panache

Hormis le cadre eulérien, on suppose gaussienne la distribution des concentrations et on veut modéliser une source ponctuelle continue. On émet les hypothèses ci-après :

- Une émission ponctuelle continue (donc active pendant un temps assez long pour avoir un panache stabilisé entre la source et le point observé le plus lointain), de débit constant Q_s ,
- Des vents suffisamment importants pour que la diffusion turbulente dans la direction du vent soit négligeable en comparaison de l'advection (approximation de panache étroit),
- Conditions météorologiques uniformes et constantes (obtention d'un panache stable avant que la situation météorologique n'évolue).

Formule de panache gaussien

La première hypothèse montre que la concentration en un point ne varie pas au cours du temps ; en pratique, il s'agit donc de considérer des concentrations moyennes sur un temps suffisamment long (correspondant souvent au temps d'intégration d'un instrument de mesure). Du fait de la météorologie constante (la troisième hypothèse), il n'y a donc plus de dépendance explicite au temps (on parle parfois de **gaussien stationnaire**). De plus, la deuxième hypothèse permet de négliger la turbulence dans la direction x (direction du vent). Si l'on fait l'hypothèse que le panache est représenté par une distribution gaussienne dans les deux directions y et z, centré sur les coordonnées de la source, et d'écarts types (σ_y, σ_z), on peut alors écrire

$$c(y,z) = \frac{Q_s}{\bar{u}} G_y(y-y_s)G_z(z-z_s) \tag{1.15}$$

$$= \frac{Q_s}{\bar{u}}\left[\frac{1}{\sqrt{2\pi}\sigma_y}\exp\big(-\frac{(y-y_s)^2}{2\sigma_y^2}\big)\right]\left[\frac{1}{\sqrt{2\pi}\sigma_z}\exp\big(-\frac{(z-z_s)^2}{2\sigma_z^2}\big)\right] \tag{1.16}$$

avec Q_s le débit de la source (en unité de masse par seconde), et $\bar{u}$ la vitesse moyenne du vent.

Les fonctions G_α sont celles définies par l'équation 1.8, où la dépendance à t a été éliminée en supposant que $t = x/\bar{u}$ (hypothèse de météorologie stationnaire). L'équation 1.16 donne une expression analytique de la concentration émise par une source ponctuelle continue en tout point de l'espace (figure 1.9). L'équation 1.8 peut donc être reliée à la solution de l'équation eulérienne obtenue sous les mêmes hypothèses (équation 1.14), en supposant $t = x/\bar{u}$ et en reliant les écarts types gaussiens au coefficient de diffusion turbulente de l'équation 1.3 par la relation

$$\sigma_\alpha = \sqrt{2K_\alpha t}, \qquad \alpha \in \{x,y,z\}. \tag{1.17}$$

Notons que cette forme en σ_α proportionnelle $t^{1/2}$ ne s'applique que très loin de la source, une fois que la taille du panache couvre l'ensemble du spectre de taille des tourbillons de la turbulence. Proche de la source, σ_α croît plus rapidement. Les formules donnant les écarts types d'un panache atmosphérique en fonction de la distance à la source (ou du temps de trajet de la bouffée) sont, en pratique, déterminées de façon empirique sur des expériences de dispersion. Dans la plupart de ces formules, on a en général σ_α proportionnelle à $t^{1/2}$ lorsque t est suffisamment grand.

Il faut noter que la conservation de la masse

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \bar{u}c(y,z)dydz = Q \tag{1.18}$$

FIGURE 1.9 – Modèle gaussien de panache stationnaire : le panache émis par une cheminée est représenté par une distribution gaussienne dans deux directions. A gauche : photo de panache issu d'une cheminée (crédit : Yelva Roustan). A droite : exemple de sortie du modèle de panache gaussien de Polyphemus.

est assurée par

$$\int_{-\infty}^{\infty} G_\alpha(\alpha) d\alpha = 1. \tag{1.19}$$

Conditions aux limites

Un panache ne peut pas s'étendre dans tout l'espace. En effet, il existe des conditions aux limites au sol ($z = 0$), et éventuellement à la hauteur d'inversion ($z = z_i$). On appelle hauteur d'inversion, la hauteur à laquelle la température de l'atmosphère augmente avec l'altitude au lieu de diminuer, empêchant ainsi l'air chargé de polluants de s'élever davantage. C'est la hauteur maximale de mélange : elle est modélisée comme un **plafond**, de façon similaire au sol. Comme cette hauteur varie (de quelques dizaines de mètres la nuit, jusqu'à 1 ou 2 km), des phénomènes d'entraînement se produisent en pratique, mais ils sont négligés dans le cadre de l'échelle locale et de la météorologie stationnaire. Dans les modèles de Polyphemus, cette variation est décrite de façon binaire. Pendant la nuit, l'inversion est supposée être proche du sol, et la plupart des polluants sont au-dessus de cette inversion. Pendant le jour, la hauteur d'inversion est supposée élevée, c'est-à-dire de l'ordre de plusieurs centaines de mètres (et non quelques dizaines, comme de nuit), et donc au-dessus des sources : les réflexions sur la couche d'inversion sont donc modélisées. On note en réalité deux principaux cas d'inversion (figure 1.10):

1. De jour, le gradient vertical de température est en général négatif. Toutefois, une inversion peut se produire en hauteur lorsqu'une masse d'air chaud est transportée au-dessus d'une masse d'air plus froid. Dans ce cas, la hauteur d'inversion est élevée (plusieurs centaines de mètres). L'atmosphère est donc neutre ou instable entre le sol et la hauteur d'inversion, et stable au-delà : les échanges sont bloquées à cette hauteur.
2. De nuit, lorsque le sol se refroidit par rayonnement infrarouge, il devient plus froid que l'air ambiant : il y a donc inversion de température près du sol, sur quelques dizaines de mètres. Les polluants émis dans la journée précédente sont bloqués au-dessus, dans ce que l'on appelle la **couche résiduelle**. Sa hauteur est celle de la couche limite du jour précédent.

Notons que, lorsqu'il y a inversion, la hauteur de couche limite est la hauteur d'inversion puisqu'il n'y a pratiquement pas d'échanges entre la surface et la partie de l'atmosphère située au-dessus de la hauteur d'inversion.

Le panache n'est pas absorbé par le sol et la couche d'inversion, mais s'y réfléchit : si l'on n'a aucun processus de perte de type dépôt, et aucun échange avec la troposphère libre, les réflexions sont parfaites. La prise en compte du sol s'effectue en modélisant une source virtuelle **sous** le sol, à la distance $-z_s$. L'équation 1.16 est alors modifiée et devient :

$$c(x,y,z) = \frac{Q_s}{2\pi\sigma_y\sigma_z\bar{u}} \exp(-\frac{(y-y_s)^2}{2\sigma_y^2})[\exp(-\frac{(z-z_s)^2}{2\sigma_z^2}) + \exp(-\frac{(z+z_s)^2}{2\sigma_z^2})] \tag{1.20}$$

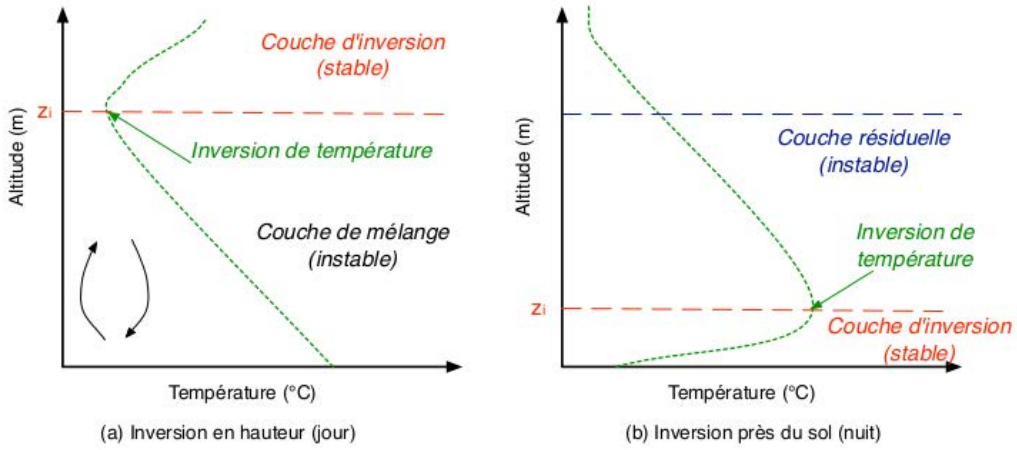


FIGURE 1.10 – Représentation de l'inversion de température dans les deux cas les plus fréquents : (a) inversion en hauteur dans la journée (z_i de l'ordre de 100 m) et (b) inversion proche du sol la nuit (z_i de l'ordre de 10 m).

En cas d'existence d'une couche d'inversion à la hauteur z_i, les réflexions sur celle-ci sont prises en compte de façon similaire. L'équation suivante nous donne la concentration 1.21

$$c(x,y,z) = \frac{Q_s}{2\pi\sigma_y\sigma_z\bar{u}} \exp(-\frac{(y-y_s)^2}{2\sigma_y^2}) \times \sum_{N=N_r}^{+N_r} [\exp(-\frac{(z-z_s+2Nz_i)^2}{2\sigma_z^2}) + \exp(-\frac{(z+z_s+2Nz_i)^2}{2\sigma_z^2})] \tag{1.21}$$

où N_r sont les réflexions sur la couche d'inversion sont considérées. La valeur $N_r = 5$ est présentée comme le maximum au-delà duquel les termes de la somme sont négligeables [14].

En pratique, N_r est souvent pris égal à 1, comme pour le sol. Les formulations précédentes présentent l'avantage de la simplicité, mais ne sont valides que pour $0 \leqslant z \leqslant z_i$. En effet, si l'on supposait que ces formules donnent la concentration dans tout l'espace, l'intégrale sur tout l'espace dans la direction z s'écrirait

$$\int_{-\infty}^{+\infty} [\exp(-\frac{(z-z_s)^2}{2\sigma_z^2}) + \exp(-\frac{(z+z_s)^2}{2\sigma_z^2})]dz = 2 \tag{1.22}$$

où

$$\begin{aligned}\int_{-\infty}^{+\infty} [\exp(-\frac{(z-z_s)^2}{2\sigma_z^2}) + \exp(-\frac{(z+z_s)^2}{2\sigma_z^2})]dz &= \int_{-z_i}^{+z_i} [\exp(-\frac{(z-z_s)^2}{2\sigma_z^2}) + \exp(-\frac{(z+z_s)^2}{2\sigma_z^2})]dz \\ &= 2\int_{0}^{+z_i} [\exp(-\frac{(z-z_s)^2}{2\sigma_z^2}) + \exp(-\frac{(z+z_s)^2}{2\sigma_z^2})]dz\end{aligned}$$

Les concentrations doivent être à zéro en dehors de la couche limite pour avoir la conservation de la masse . L'équation de conservation de la masse 1.19 peut se réécrire de la manière suivante:

$$\int_0^{z_i} \int_{-\infty}^{+\infty} c\bar{u}dydz = Q_s \tag{1.23}$$

Remarquons que lorsque le panache est suffisamment proche du sol ou de la couche d'inversion:

1. Il y a réflexion au sol si $\sigma_z > z_s$
2. Il y a réflexion sur la couche d'inversion si $z_s + \sigma_z > z_i$.

Enfin, lorsque le panache est suffisamment étendu sur la verticale ($\sigma_z > 1.5 z_i$), on considère que les différentes réflexions et le mélange turbulent l'ont rendu homogène sur la verticale. La concentration est alors donnée par la formule 1.24 suivante appelée formule de champ lointain :

$$c(x,y,z) = \frac{Q_s}{2\pi\sigma_y z_i \bar{u}} \exp(-\frac{(y-y_s)^2}{2\sigma_y^2}) \tag{1.24}$$

1.7.4 Modèle gaussien à bouffées

Formule de bouffée gaussienne

Pour un modèle instationnaire, le panache est représenté sous la forme d'une série de bouffées gaussiennes dans les trois directions. Les hypothèses présentées en partie 1.7.3 sont alors moins contraignantes : les hypothèses 1 et 2 n'ont plus lieu d'être. De plus, on suppose simplement que la météorologie est uniforme à l'intérieur d'une même bouffée (hypothèse 3), mais elle peut varier d'une bouffée à l'autre, et dans le temps. Dans le cas d'une seule bouffée émise, de quantité totale Q, La concentration de la bouffée supposée gaussienne dans les trois directions peut s'écrire

$$c(x,y,z) = \frac{Q}{(2\pi)^{2/3}\sigma_x\sigma_y\sigma_z} \exp(-\frac{(x-x_c)^2}{2\sigma_x^2}) \exp(-\frac{(y-y_c)^2}{2\sigma_y^2}) \exp(-\frac{(z-z_c)^2}{2\sigma_z^2}) \tag{1.25}$$

Les coordonnées du centre de la bouffée sont x_c dans la direction du vent, y_c et z_c dans les directions horizontale (perpendiculaire au vent) et verticale. Dans le cas d'un vent constant u, on a donc $x_c = ut_c$ avec t_c le temps écoulé depuis que la bouffée a été émise. Ce temps détermine l' **âge** de la bouffée. Les trois écarts types de la gaussienne sont σ_x et σ_y dans la direction horizontale (direction du vent et perpendiculaire au vent respectivement), et σ_z dans la direction verticale. De même que dans le cas du panache stationnaire, l'équation 1.25 est identique à la solution eulérienne 1.9 en supposant que les écarts types gaussiens et la diffusion eulérienne sont reliés par l'équation 1.17. En général, les formules de dispersion pour σ_y et σ_z sont les mêmes que dans le cas du modèle de panache. Pour les paramétrisations qui ne donnent pas de formulation spécifique de σ_x , celle de σ_y est utilisée. Les réflexions sur le sol et la couche d'inversion ainsi que le champ lointain sont pris en compte de manière similaire au modèle de panache (équations 1.21 et 1.24 respectivement).

Relation entre les formulations de panache et de bouffée

Pour une source continue avec une météorologie uniforme, la moyenne sur un temps suffisamment long du modèle à bouffées permet de retrouver la solution stationnaire du modèle de panache (figure 1.11). Dans le paragraphe (1.7.3), si la source continue modélisée, de débit Q_s

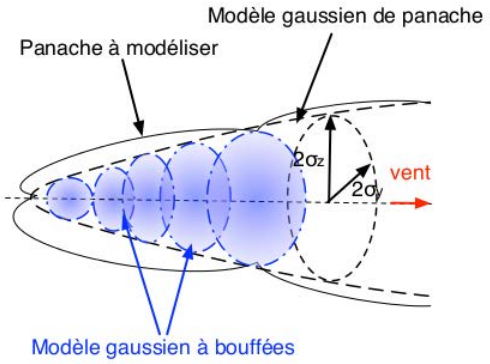


FIGURE 1.11 – Modèle gaussien non stationnaire à bouffées : le panache est discrétisé en une série de bouffées, gaussiennes dans les trois directions.

, est discrétisée en un certain nombre de bouffées, à un pas de temps Δt_p, alors chaque bouffée

contiendra la quantité $Q = \Delta t_p Q_s$. La concentration en un point est donnée par la somme de la contribution de toute les bouffées

$$c(x,y,z) = \sum_{i=1}^{N_p} \frac{Q_s \times \Delta t_p}{(2\pi)^{3/2}\sigma_x^i \sigma_y^i \sigma_z^i} \exp\left(-\frac{(x-x_c^i)^2}{2{\sigma_x^i}^2}\right) \exp\left(-\frac{(y-y_c^i)^2}{2{\sigma_y^i}^2}\right) \exp\left(-\frac{(z-z_c^i)^2}{2{\sigma_z^i}^2}\right) \quad (1.26)$$

Les coordonnées du centre de la bouffée i émise au temps $t_i = i\Delta t_p$ sont x_c^i, y_c^i, z_c^i. Les écarts types associés sont notés σ_x^i , σ_y^i, σ_z^i, et dépendent de la position de la bouffée par rapport à la source (ou de son âge). Ces coordonnées évoluent au cours du temps. Dans le cas d'un vent $\bar{u}$ constant et homogène dans la direction x, on a $x_c^i(t) = x_s + \bar{u}(t - t_i), y_c = y_s$ et $z_c = z_s$.

Le passage du cas discret (équation 1.26) au cas du panache continu s' écrit en intégrant sur des bouffées émises à des pas de temps infinitésimaux dt'

$$c(x,y,z) = \int_{-\infty}^{\infty} \frac{Q_s \times dt'}{(2\pi)^{3/2}\sigma_x\sigma_y\sigma_z} \exp\left(-\frac{(x-x_s-\bar{u}(t-t'))^2}{2{\sigma_x}^2}\right) \exp\left(-\frac{(y-y_c)^2}{2{\sigma_y}^2}\right) \exp\left(-\frac{(z-z_c)^2}{2{\sigma_z}^2}\right) \quad (1.27)$$

Revenons maintenant aux hypothèses faites dans la paragraphe 1.7.3: dans la première hypothèse on donne la concentration est faisant la moyenne sur un temps d'intégration T (puisque elle ne dépend pas du temps, et que la source émet en permanence), et dans la seconde hypothèse la dispersion turbulente dans la direction x est négligeable devant la distance ' parcourir dans cette direction:

$$\sigma_x \ll x - x_s. \quad (1.28)$$

De plus, le panache est supposé dans un état stationnaire, ce qui revient à dire que la source a commencé à émettre très longtemps avant la mesure (à $t = -\infty$). Les conditions météorologiques sont également stationnaires (la troisième hypothèse du paragraphe 1.7.3). Les valeurs de $\sigma_x, \sigma_y et \sigma_z$ en un point donné (x, y, z) dépendent du temps de transport de chaque bouffée $t - t'$. Cependant, étant données les conditions stationnaires, on considère que toutes les bouffées arrivent au point (x, y, z) avec le même temps de transport, et que les valeurs des écarts types en un point donné sont donc constantes au cours du temps. L'équation (1.27) réécrit alors comme suit :

$$c(x,y,z) = \frac{1}{T}\frac{Q_s}{2\pi\sigma_y\sigma_z} \exp\left(-\frac{(y-y_c)^2}{2\sigma_y^2}\right) \exp\left(-\frac{(z-z_c)^2}{2\sigma_z^2}\right) \int_0^T \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left(-\frac{(\bar{u}(t-t'))^2}{2\sigma_x^2}\right) dt'dt. \quad (1.29)$$

On pose t" $= t - t'$ on a après calcul

$$\int_0^T \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left(-\frac{(\bar{u}(t-t'))^2}{2\sigma_x^2}\right) dt'dt = \frac{T}{\bar{u}} \quad (1.30)$$

et on retrouve l'expression de la panache gaussien (équation 1.16)

$$c(y,z) = \frac{Q_s}{\bar{u}} G_y(y) G_z(z) \quad (1.31)$$

La compréhension des hypothèses et les limites des deux modèles résultent de cette preuve précédentes. Le modèle gaussien de panache fait des hypothèses plus contraignantes, notamment de météorologie constante et homogène sur tout le domaine, contrairement au modèle à bouffées. Cette hypothèse n'est plus valide hors du champ proche, et peut poser problème en cas de cisaillement de vent important sur la verticale (ce qui est également le cas pour les bouffées). De plus, le fait de négliger la turbulence horizontale en regard du transport (seconde hypothèse) n'est plus valide en cas de vent très faible.

Si l'on reste dans son domaine de validité, la formulation de panache gaussien est cependant plus appropriée, pour des sources émettant en continu, que la formulation non stationnaire. En effet, le modèle à bouffée nécessite de discrétiser le panache de façon suffisamment fine, et de faire une moyenne sur un temps assez long pour retrouver la solution stationnaire. Si l'on considère qu'une

bouffée a pour taille $2\sigma_x$ dans la direction x, deux bouffées successives émises à t_i et $t_{i+1} = t_i + \Delta t_p$ se recouvrent au temps t lorsque

$$\sigma_x(t - t_i) + \sigma_x(t - t_{i+1}) \geq \frac{\bar{u}}{2}\Delta t_p.$$

C'est pas important que pour bien modéliser un panache continu, de remplir cette condition à très faible distance de la source : il suffit qu'elle soit remplie aux points d'intérêt (par exemple, à un premier capteur de mesure). Si le vent est $\bar{u} = 5ms^{-1}$, et que $\sigma_x(t) = 1000\ m$ (taille atteinte au bout d'une heure environ dans une atmosphère neutre), on a $\Delta t_p = 800s$. Cette valeur minimale est adapté dans une application régionale, où les capteurs sont éloignés de la source. La valeur typique sera plutôt de l'ordre de $60s$ ou moins dans une application à échelle locale (quelques kilomètres).

1.7.5 Turbulence

À des échelles spatio-temporelles inférieures à la dizaine de mètres et à la minute, les mouvements de l'atmosphère ne peuvent plus être prévus de façon déterministes. L'écoulement atmosphérique à ces échelles est instable et son caractère chaotique fait qu'il ne peut être connu que par ses propriétés statistiques: il est turbulent. Les phénomènes non linéaires jouent un rôle prépondérant dans un écoulement turbulent, l'énergie étant transférée entre les structures de tailles très diverses présentes dans l'écoulement.

Bien qu'il soit difficile de caractériser la turbulence, ses nombreuses propriétés sont relativement bien connues: la turbulence est non stationnaire, dissipative, tri-dimensionnelle et rotationnelle, diffusive. La turbulence favorise donc le mélange : elle augmente les transferts de quantité de mouvement, de chaleur ou de masse de plusieurs ordres de grandeur. Elle permet donc les transferts de chaleur du sol vers l'atmosphère et le mélange des polluants.

1.7.6 Modèles stochastiques

Les séries chronologiques utilisent des modèles spécifiques: modèles AR (Auto-Regressive, ou auto-régressifs), MA (Moving Average, ou moyennes mobiles), ARMA, ARIMA (I pour Integrated). Dans cette partie de la thèse, on présente quelques outils de probabilité en guise de rappels pour annoncer le chapitre 5 sur les modèles ARMA.

Concepts de base en probabilité

Expérience aléatoire : c'est une expérience dont le résultat n'est pas connu avec certitude. Supposons que tous les résultats possibles de cette expérience soient connus. **L'espace d'échantillonnage** ou **l'univers de possibilités** Ω est cet ensemble de résultats possibles d'une expérience aléatoire.

Un **événement** E est un sous-ensemble de l'espace échantillonnage. Supposons que nous répétions l'expérience aléatoire un grand nombre de fois (n), et que l'événement E se produise m fois. La probabilité associée à l'événement E peut être approchée comme

$$P[E] \approx \frac{m}{n}.$$

Une définition empirique de la probabilité de E est :

$$P[E] = \lim_{n\to\infty} \frac{m}{n},$$

De manière plus formelle

Définition 1.7.1. *P est une mesure de probabilité si*
- $0 \leq P[E] \leq 1$, *pour tout* $E \subset \Omega$;
- $P[\emptyset] = 0$ *et* $P[\Omega] = 1$;
- $P[E_1 \cup E_2] = P[E_1] + P[E_2]$, *si* E_1 *et* E_2 *sont disjoints (i.e.* $E_1 \cap E_2 = \emptyset$*).*

Notons qu'un événement E_1 peut se produire tout en influença la probabilité d'un autre événement E_2. Par exemple, la probabilité qu'il pleuve demain dans Dakar (E_2) est plus élevée s'il pleut aujourd'hui (E_1) que s'il ne pleut pas. Si $P[E_1] > 0$, nous définissons la probabilité conditionnelle associée à l'événement E_2, étant donné E_1, comme:

$$P[E_2 \,|\, E_1] = \frac{P[E_1 \cap E_2]}{P[E_1]}.$$

Rappelons que la probabilité conditionnelle jouit des propriétés suivantes:
- $0 \leq P[E_2 \,/\, E_1] \leq 1$;
- $P[\emptyset \,/\, E_1] = 0$ et $P[\Omega \,/\, E_1\,] = 1$;
- $P[E_2 U E_3 \,/\, E_1) = P[E_2 \,/\, E_1] + P[E_3 \,/\, E_1]$, si E_2 et E_3 sont disjoints.

Deux **événements E_1 et E_2 sont indépendants** si

$$P[E_2 \,/\, E_1] = P[E_2].$$

De manière alternative, nous pouvons utiliser les définitions suivantes:

$$P[E_1 \,/\, E_2] = P[E_1]$$
$$P[E_1 \cap E_2] = P[E_1]P[E_2].$$

En général, nous postulons l'indépendance de deux événements pour se servir des définitions ci-dessus, plutôt que de déduire l'indépendance de deux événements à partir des définitions. On dit que n événements E_1, E_2,..., E_n sont indépendants si

$$P[E_1 \cap E_2 \cap \ldots E_n] = P[E_1]P[E_2] \ldots P[E_n].$$

Variable aléatoire x : c'est une fonction qui associe une valeur numérique $x(s)$ à chaque élément s de l'espace d'échantillonnage:

$$x : \Omega \to \mathcal{R}.$$

Il existe deux types principaux de variables aléatoires:
- variable aléatoire continue: valeurs réelles ;
- variable aléatoire discrète: valeurs entières ou nombre fini de valeurs.

La **fonction de répartition** associée à une variable aléatoire X est définie comme

$$F_X(b) = P[X \leq b] = P[s \in \Omega \,/\, X(s) \leq b],$$

Elle a comme propriétés:
- non décroissante ;
- $\lim_{b \to -\infty} F_X(b) = 0$ et $lim_{b \to \infty} F_X(b) = 1$;
- $P[a < X \leq b] = F_X(b) - F_X(a)$, vu que

$$\{s \in \Omega \,:\, X(s) \leq b\} = \{s \in \Omega \,:\, X(s) \leq a\} \cup \{s \in \Omega \,:\, a < X(s) \leq b\}.$$

Variables aléatoires discrètes : La fonction de masse associée à une variable aléatoire X est définie comme

$$P_X(k) = P[X = k] = P[s \in \Omega \,:\, X(s) = k].$$

Pour une variable aléatoire discrète, nous avons

$$F_X(b) = P[X \leq b] = \sum_{k \leq b} P[X = k] = \sum_{X \leq b} P_X(k).$$

aussi,

$$P[a < X < b] = F_X(b) - F_X(a) - P_X(b).$$

Variables aléatoires continues : Une **variable aléatoire X** est continue si sa fonction de répartition peut être représentée ainsi:

$$F_X(b) = \int_{-\infty}^{b} f_x(x)dx.$$

La fonction sous l'intégrale est appelée **fonction de densité** et satisfait les conditions suivantes:

$$f_x(x) \geq 0, \ \forall x;$$
$$\int_{-\infty}^{\infty} f_x(x)dx = 1$$

Si la fonction de densité est continue, alors elle est égale à la dérivée de la fonction de répartition:

$$f_X(x) = \frac{dF_X}{dx}(x).$$

La fonction de masse prend toujours la valeur 0:

$$P_X(x) = 0, \ \forall x.$$

Pour tout intervalle de la forme $< a, b >$, on a également

$$P[X \in < a, b >] = \int_a^b f(x)dx = F_X(b) - F_X(a).$$

Espérance mathématique :L'espérance mathématique (moyenne) associée à une variable aléatoire X est notée $E[X]$ et définie comme suit:

$$E[X] = \begin{cases} \sum_k kP_X(k) = \sum_k P[X = k] & \text{si } X \text{ est discrète,} \\ \int_{-\infty}^{\infty} xf_X(x)dx & \text{si } X \text{ est continue.} \end{cases}$$

On peut également considérer l'espérance d'une fonction $g(X)$. Si X est une variable aléatoire discrète, cela donne

$$E[g(X)] = \sum_k g(k)P_X(k).$$

Pour une variable continue, nous aurons

$$E[g(X)] = \int_{-\infty}^{+\infty} g(X)f_X(x)dx.$$

En particulier, nous avons, si $a \in \mathcal{R}$ et X, Y sont deux variables aléatoires,

$$E[aX] = aE[X],$$
$$E[X + Y] = E[X] + E[Y].$$

Variance : Elle est associée à une variable aléatoire X, dénotée $\sigma^2(X)$, est définie par la formule

$$\sigma^2(X) = E[X^2] - E[X]^2 = E[(X - E[X])^2].$$

Loi de probabilité c'est un modèle d'une expérience aléatoire. Elle est habituellement représentée par la fonction de répartition d'une variable aléatoire. Si cette dernière est discrète, la loi de probabilité est dite discrète. Une loi de probabilité discrète peut être représentée par sa fonction de masse. Si la variable aléatoire est continue, la loi de probabilité est dite continue, et peut être représentée par sa fonction de densité.

Loi de Bernouilli: l'espace d'échantillonnage se limite à deux éléments, notés par exemple par $\Omega = \{S, E\}$. On définit la variable aléatoire X comme suit:

$$X(S) = 1 \text{ et } X(E) = 0.$$

La fonction de masse est

$$P_X(1) = p \text{ et } P_X(0) = 1 - p,$$

où p est un paramètre. De manière équivalente, nous pouvons l'écrire comme

$$P_X(x) = p^x(1-p)^{1-x}.$$

Nous avons de plus

$$E[X] = p \text{ et } \sigma^2(X) = p(1-p).$$

Par exemple, le tirage d'une pièce de monnaie suit une loi de Bernouilli, avec p=1/2.

Loi uniforme Une variable aléatoire continue X (qui prend ses valeurs dans l'intervalle $[a, b]$) suit une **loi uniforme** (notée $U[a, b]$) si sa fonction de densité est:

$$f_X(x) = \frac{1}{b-a},\ \forall x \in [a, b].$$

Cette loi modélise l'expérience aléatoire consistant à choisir au hasard un point de $[a, b]$ (la probabilité de choisir un point dans un sous-intervalle est proportionnelle à la longueur de ce sous-intervalle).

Si X est une variable aléatoire continue (quelconque), nous avons la propriété suivante:

$$F_X(x) \sim U[0, 1].$$

Loi de Poisson : Une variable aléatoire X suivant une loi de Poisson est une variable aléatoire qui sert à compter le nombre d'apparitions d'un phénomène aléatoire durant un intervalle de temps de longueur t. Il pourrait s' agit par exemple du nombre d'appels reçus par un téléphoniste. X a alors pour fonction de masse

$$P_X(k) = P[X = k] = \frac{\theta t^k e^{-\theta t}}{k!},\ k = 0, 1, 2, \ldots,$$

où θ représente le taux moyen.

Loi exponentielle : soit une variable aléatoire X représentant le temps d'attente entre deux apparitions du phénomène aléatoire en supposant que le nombre d'apparitions durant un intervalle t suit une **loi de Poisson de paramètre** θ. La fonction de répartition vérifie alors

$$1 - F_X(x) = P[X > x] = e^{-\theta x},\ x \geq 0.$$

Il s'agit de la loi exponentielle de fonction de densité:

$$f_X(x) = \begin{cases} \theta e^{-\theta x} & \text{si } x > 0, \\ 0 & \text{sinon.} \end{cases}$$

L'espérance mathématique est:

$$E[X] = \frac{1}{\theta}.$$

C'est le taux moyen entre deux apparitions du phénomène aléatoire.

Modèles stochastiques : Un système stochastique est un système évoluant de manière probabiliste dans le temps. Les exemples sont nombreux, avec par exemple la température quotidienne ou un centre d'appels téléphoniques. Un modèle stochastique est une représentation mathématique d'un système stochastique. Nous verrons brièvement deux cas classiques de modèles stochastiques: les processus stochastiques et les files d'attente.

Processus stochastiques: Un **processus stochastique** est une suite de variables aléatoires évoluant dans le temps, que nous dénoterons $\{X_t\}$, $t \in T$. En général, T est un ensemble discret: $T = \{0, 1, 2, \ldots\}$. De plus, chaque variable aléatoire peut prendre une valeur parmi $M + 1$ états: $X_t \in \{0, 1, \ldots, M\}$.

Exemple 1.7.2. ***Précipitations quotidiennes***

$$X_t = \begin{cases} 1 & s'il\ y\ a\ des\ précipitations, \\ 0 & s'il\ n'y\ a\ pas\ de\ précipitations. \end{cases}$$

Chaînes de Markov : un processus stochastique est une chaîne de Markov [3] s'il possède la **propriété markovienne**:

$$P[X_{t+1} = j \,|\, X_0 = k_0, X_1 = k_1, \ldots, X_{t-1} = k_{t-1}, X_t = i] = P[X_{t+1} = j \,|\, X_t = i].$$

Cette propriété signifie que la probabilité d'un événement futur, étant donné des événements passés et un état au temps présent, ne dépend pas du passé, mais uniquement de l'état actuel. La probabilité de transition entre les états i et j est définie comme

$$p_{ij} = P[X_{t+1} = j \,|\, X_t = i].$$

Cette probabilité de transition est dite stationnaire si:

$$P[X_{t+1} = j \,|\, X_t = i] = P[X_1 = j \,|\, X_0 = i],\ t = 1, 2, \ldots$$

Puisqu'il s'agit de probabilité, nous devons en outre avoir

$$p_{ij} \geq 0,\ i, j \in \{0, 1, \ldots, M\};$$
$$\sum_{j=0}^{M} p_{ij} = 1 \geq 0,\ i \in \{0, 1, \ldots, M\}.$$

A partir des probabilités de transition, nous pouvons construire

- La matrice des transitions, ayant $M + 1$ rangées (les états présents) et $M + 1$ colonnes (les états futurs), chaque entrée (i, j) de la matrice correspondant à p_{ij}.
- Le graphe (ou diagramme) des transitions, ayant $M + 1$ sommets et tel qu'il y a un arc entre les états i et j si $p_{ij} > 0$.

Exemple 1.7.3. ***Précipitations:*** *supposons que la probabilité qu'il n'y ait pas de précipitations à Dakar demain, étant donné:*

- *qu'il n'y en a pas aujourd'hui est 0.8 ;*
- *qu'il y en a aujourd'hui : 0.6.*

Ces probabilités ne changent pas, même si nous tenons compte de ce qui se passe avant aujourd'hui. La propriété markovienne est satisfaite:

$$P[X_{t+1} = 0 \,|\, X_0 = k_0, X_1 = k_1, \ldots, X_{t-1} = k_{t-1}, X_t = 0] = P[X_{t+1} = 0 \,|\, X_t = 0]$$
$$P[X_{t+1} = 0 \,|\, X_0 = k_0, X_1 = k_1, \ldots, X_{t-1} = k_{t-1}, X_t = 1] = P[X_{t+1} = 0 \,|\, X_t = 1]$$

Nous avons donc une chaîne de Markov dont les probabilités de transition sont:

$$p_{00} = P[X_{t+1} = 0 \,|\, X_t = 0] = 0.8,$$
$$p_{10} = P[X_{t+1} = 0 \,|\, X_t = 1] = 0.6.$$

Grâce aux propriétés des probabilités de transition, nous pouvons déduire celles qui manquent :

$$p_{01} = P[X_{t+1} = 1 \,|\, X_t = 0] = 1 - 0.8 = 0.2.$$
$$p_{11} = P[X_{t+1} = 1 \,|\, X_t = 1] = 1 - 0.6 = 0.4.$$

Ceci donne la matrice de transition:

$$[P] = \begin{pmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{pmatrix}$$

Le graphe de transition est quant à lui représenté sur la Figure 1.12.

3. Les chaines de Markov avaient été introduites bien avant les travaux de Markov. En 1889, Galton a introduit des chaine de Markov pour étudier le problème de la disparition de noms de famille. En 1907,Ehrenfest a aussi introduit les chaines de Markov pour étudier la diffusion d'un gaz.

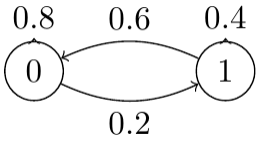


FIGURE 1.12 – Graphe de transition

1.8 Application

On considère l'équation suivante de la sous section 1.7.2 de l'approche eulérienne :

$$\frac{\partial c}{\partial t} + \underbrace{u_x\frac{\partial c}{\partial x} + u_y\frac{\partial c}{\partial y} + u_z\frac{\partial c}{\partial z}}_{Advection} = \underbrace{\frac{\partial(K_x\frac{\partial c}{\partial x})}{\partial x} + \frac{\partial(K_y\frac{\partial c}{\partial y})}{\partial y} + \frac{\partial(K_z\frac{\partial c}{\partial z})}{\partial z}}_{Diffusion}$$
$$+ \underbrace{R(c_1, c_2, \ldots, c_n)}_{Chimie} + \underbrace{E(x,y,z,t)}_{Sources} - \underbrace{S(x,y,z,t)}_{Puits} \tag{1.32}$$

Dans cette partie de la thèse, on s'intéresse au problème simple de transport linéaire. On compte en perspectif progressivement réfléchir et ajouter le terme de diffusion, de source puis des puits compte tenue de la complexité du problème (équations de Navier Stock). Cela ne sera pas traité dans la thèse. On traite ici le problème lié à l'advection.

Transport linéaire: advection Le modèle de la qualité de l'air est un problème de description de transport atmosphérique, un problème de diffusion, de réaction atmosphérique de polluants. Les variables inconnues sont des concentrations d'espèces chimiques dans l'air. Le but de l'étude et du développement de tels modèles c'est d'être capable de prédire comment les concentrations maximum changeront dans la réponse aux changements prescrits par la météorologie et dans la source de pollution. On place notre problème dans un contexte d'implantation d'usine qui émet des fumées toxiques (ie la source de pollution est stationnaire. Le problème peut être aussi étudié de manière un peu similaire dans le cas où la source polluante est mobile: pollution de proximité des routes). On suppose ces fumées toxiques émises en une fois par heure comme une partie du processus d'implantation ou non d'une usine. Il y a une belle maison que vous aimeriez acheter de nos jours. Mais comment habiter cette maison si celle-ci se trouve dans la direction de la fumée toxique ou bien à proximité d'une route à flux important de véhicules ? En d'autres termes, quelle concentration maximale de fumée prendra la direction de votre maison ? Il est nécessaire de connaître la densité de la fumée lorsque celle-ci atteint une maison, pour en estimer la nocivité. Supposons qu'il n'y a aucun changement chimique au cours de ce déplacement de l'air pollué: il y a deux processus essentiels à noter: **l'advection et la diffusion** . On traite dans cette sous-section de la thèse le processus de l'advection.

Définition 1.8.1. ***Advection:*** *L'advection est essentiellement l'effet du déplacement des vapeurs par le "soufflement du vent" dans une direction donnée sans disperser considérablement leur concentration. Un exemple de l'advection est celui du déplacement de nuage dans une direction donnée sans changement apparent de forme ou de dimension.*

Considérons la situation uni-dimensionnelle où il y a advection mais aucune diffusion. On suppose, qu'à l'instant $t = 0$, la densité des fumées a une distribution $c_0(x)$

Le profile de la concentration du polluant se déplace en fonction de la vélocité du vent u constante, donnant au profil en mouvement une concentration

$$c(x,t) = c_0(x - ut) \tag{1.33}$$

En différentiant partiellement (règle de chaîne) on obtient:

$$\frac{\partial c(x,t)}{\partial x} = \frac{\partial c_0(x-ut)}{\partial x} = (x-ut)'_x c'_0(x-ut) = c'_0(x-ut)$$

et

$$\frac{\partial c(x,t)}{\partial t} = \frac{\partial c_0(x-ut)}{\partial t} = (x-ut)'_t c'_0(x-ut) = -uc'_0(x-ut) \tag{1.34}$$

En supposant la vélocité du vent constante (en temps et en espace ie pour tout x réel et pour tout $t > 0$, $u(x,t) = u$, constante réelle), on aura:

$$\frac{\partial (uc)(x,t)}{\partial x} = u\frac{\partial c(x-ut)}{\partial x} = uc'_0(x-Ut) \tag{1.35}$$

En sommant (1.34) et (1.35) on obtient **l'équation de l'advection caractérisant le transport linéaire:**

$$\frac{\partial c(x,t)}{\partial t} + u\frac{\partial c(x,t)}{\partial x} = 0 \quad avec \quad c(x,0) = c_0(x) \tag{1.36}$$

On retrouve ici l'équation du transport en une dimension du problème d'advection dans l'équation générale 1.32. On regarde la forme avec la concentration du profil donnée par la formule (1.33), nous voyons que les fumées toxiques arrivent chez vous avec la même concentration que celle libérée au départ à l'usine ou par le véhicule: c'est une mauvaise nouvelle pour vous concernant votre maison. Heureusement tout n'est pas perdu car le processus de la diffusion peut éventuellement pallier à notre inquiétude. C'est la raison pour la quelle même sans la présence du vent, les odeurs fétides odorantes disparaissent habituellement après un temps. Par exemple, une maison récemment peinte cesse de sentir de peinture après un ou deux jours, sinon plutôt. Tenant compte de l'advection et de la diffusion, il y a au moins de possibilités que les vapeurs soient arrivées à votre maison, la diffusion a un grand effet pour que les fumées soient à peine notables.

Exemples

L'équation d'advection qui peut sembler très simple pose des difficultés numériques considérables, elle est toujours l'objet de recherches actuellement (il s'agit notamment de savoir calculer la solution sur des temps très grands). Par ailleurs, couplée à d'autres équations, elle pose des difficultés théoriques également.

On rencontre cette équation dans un grand nombre d'applications. Citons en quelques unes:

Exemple 1.8.2. ***circulation automobile*** *On étudie la circulation automobile sur une route. La variable $c(x,t)$ représente la quantité de voitures présentes entre les bornes x et $x+\Delta x$ à l'instant t et on appelle $F(x,t)$ le flux de voitures par minute qui passent à l'instant t devant la borne x. On suppose que chaque conducteur ajuste la vitesse de sa voiture en fonction uniquement de la vitesse de la voiture qui le précède. Alors la conservation de la quantité de voitures (il n'y a ni station-service, ni itinéraires de délestage) se traduit par :*

$$\frac{\partial c(x,t)}{\partial t} + \frac{\partial F(x,t)}{\partial x} = 0$$

Si toutes les voitures roulaient à vitesse constante donnée u, on aurait :

$$F(x,t) = uc(x,t)$$

et donc l'équation de transport linéaire (1.36).

Exemple 1.8.3. ***les équations cinétiques*** *La physique cinétique décrit les plasmas ou gaz dilués et fournit un ensemble important d'équations de transport dont les variables sont : le temps t, la position $x \in \Omega$ des particules avec Ω le domaine d'étude et leur vitesse $u \in U$ avec U l'ensemble des vitesses admissibles. L'exemple le plus classique est l'équation de scattering décrivant*

l'évolution d'une densité $f(x,t,u)$ de particules (neutrons, amibes ou bactéries en ce qui concerne les applications à la biologie)

$$\begin{cases} \dfrac{\partial f(x,t,u)}{\partial t} + u\dfrac{\partial f(x,t,u)}{\partial x} = K[f], \\ \\ f(x,t=0,u) = f_0(x,u) \quad \textit{donnée}, \end{cases}$$

où K est donnée par:

$$K[f] = \int_U k(u,u')f(x,t,u')du' - \int_U k(u',u)f(x,t,u')du',$$

où $k(u,u')$ représente une probabilité de "tourner" d'une vitesse u' à une vitesse u.
*Cette équation provient de la relation de **conservation** de la quantité des particules.*

Exemple 1.8.4. *Modèle démographie-Renouvellement cellulaire*

Ici $c(x,t)$ représente la densité d'individus d'âge $x > 0$ à l'instant t. Le taux de mortalité est noté $d(x)$ et les individus peuvent donner naissance à des nouveaux nés d'âge $x = 0$ avec un taux de fécondité $b(x)$:

$$\begin{cases} \dfrac{\partial c(x,t)}{\partial t} + \dfrac{\partial c(x,t)}{\partial x} + d(x)c(x,t) = 0, \\ \\ c(0,t) = \int_0^{+\infty} b(x')c(x',t)dx', \end{cases}$$

Dans le cas de la mitose cellulaire, il est naturel de choisir $d(x) = \chi_{\{x>x_0\}}$ [4] (disparition des cellules qui se divisent) et $b(x) = 2\chi_{\{x>x_0\}}$ (la mitose donne naissance à deux cellules identiques).

Le modèle

c désigne ici la concentration d'une espèce chimique ; c'est une fonction de position (x_1, x_2, x_3) et du temps t. Les espèces chimiques sont transportées par un vent de vitesse supposée connue $\overrightarrow{u} = \overrightarrow{u}(x_1, x_2, x_3, t)$. Les particules de l'espèce se diffusent aussi localement, elles ont tendance à se déplacer dans de régions de forte concentration vers de régions de faibles concentrations. Si la diffusion est ignorée alors l'équation du transport s'écrit:

$$\frac{\partial c}{\partial t} + \nabla(\overrightarrow{u}c) = 0 \tag{1.37}$$

cette équation est appelée dans certains contextes **l'équation de la continuité**.
Si on intègre (1.37) sur un domaine borné Δ de $\mathbb{R}^3$, on obtient:

$$\frac{d}{dt}\iiint_\Delta c(x_1,x_2,x_3,t)dx_1dx_2dx_3 = -\iint_{\partial\Delta} c\,\overrightarrow{u}\,\overrightarrow{n}\,ds \tag{1.38}$$

où $\partial\nabla$ est le bord de ∇ et $\overrightarrow{n}$ est la normale unitaire extérieure à $\partial\Delta$.

Cette équation exprime que le taux d'augmentation du chimique dans tout le domaine Δ est égal au flux du chimique à travers le bord. Inversement, si l'équation (1.38) est vraie sur tout Δ alors on peut retrouver la relation (1.37) en subdivisant Δ en plusieurs domaines ponctuels Δ_j. Si la diffusion n'est pas ignorée alors de l'équation (1.37), on obtient l'E.D.P complexe suivant:

$$\frac{\partial c}{\partial t} + \nabla(\overrightarrow{u}u) = \sum_{i,j=1}^{3} \frac{\partial}{\partial x_i}(K_{ij}\frac{\partial c}{\partial x_j}) \tag{1.39}$$

où K_{ij} est une matrice définie positive, appelée matrice de la diffusion. Dans l'un ou l'autre des cas, (1.37) ou (1.39) sont données avec la concentration c à l'instant initial $t = 0$.

$$c(x_1,x_2,x_3) = c_0(x_1,x_2,x_2) \tag{1.40}$$

4. $\chi_{(A)}$ est la fonction caractéristique de A

et notre tache est de calculer la concentration c à temps subséquents. Nous souhaiterons en particulier trouver les valeurs maximales de la concentration à un temps donné ; les règlementations des gouvernements pour contrôler la pollution prennent souvent la concentration maximale de pollution comme un facteur critique.

Équation d'advection

On se place ici dans un contexte sans diffusion et on suppose que la vitesse du vent est dans la direction horizontale seulement. Pour simplifier, on suppose que la direction du vent est orientée suivant la direction x. Alors $\overrightarrow{u} = (u, 0, 0)$ et l'équation du transport se réduit à

$$\frac{\partial c}{\partial t} + \frac{\partial (uc)}{\partial x} = 0 \tag{1.41}$$

Cette équation est appelée l'**équation de l'advection**. On suppose aussi que la concentration à l'état initial dépend de x seulement, ie

$$c(x, 0) = c_0(x), \quad -\infty < x < +\infty \tag{1.42}$$

La vitesse du vent $u = u(x)$ est fonction de x. Pour résoudre (1.41), (1.42), on réécrit (1.41) comme suit:

$$\frac{\partial c}{\partial t} + u\frac{\partial c}{\partial x} = f \qquad (f = -u_x c) \tag{1.43}$$

et on suppose que u(x) est continûment différentiable ($u_x = \frac{du}{dx}$). Soit l'équation différentielle

$$\begin{cases} \dfrac{dx}{dt} = u(x), t > 0 \\ \\ x(0) = x_0 \end{cases} \tag{1.44}$$

et on désigne sa solution $x(t)$ par $x(t, x_0)$. Géométriquement $x(t, x_0)$ détermine l'unique courbe γ_{x_0} passant par le point $(x_0, 0)$. On peut montrer que $x(t, x_0)$ est différentiable suivant x_0 et la dérivée est

$$z(t) \equiv \frac{\partial x(t, x_0)}{\partial x_0}$$

et vérifie

$$\frac{dz}{dt} = u_x(x(t, x_0))z, \quad z(0) = 1.$$

On examine maintenant la fonction $c(x(t, x_0), t)$ comme la fonction de la variable t. On a :

$$\frac{\partial c}{\partial t} = \frac{\partial c}{\partial t} + \frac{\partial c}{\partial x}\frac{dx}{dt}$$

en posant $\frac{dx}{dt} = u(x)$ on a :

$$\frac{\partial c}{\partial t} = \frac{\partial c}{\partial t} + \frac{\partial c}{\partial x}u$$

$$\Longrightarrow \frac{\partial c}{\partial t} = \frac{\partial c}{\partial t} + u\frac{\partial c}{\partial x} = f$$

comme $f = -u_x c \quad$ et $\quad u_x = u_x(x(t, x_0))$ on a

$$\frac{dc}{dt} = \frac{\partial c}{\partial x} + u\frac{\partial c}{\partial x} = f = -u_x(x(t, x_0)c$$

$$\Longleftrightarrow \frac{dc}{dt} = -u_x(x(t, x_0)c$$

$$\Longleftrightarrow \frac{dc}{c} = -u_x(x(t, x_0)dt$$

$$\begin{aligned}
&\Longleftrightarrow \frac{dc}{c} = -u_x(x(s,x_0)ds \quad (\text{ car la variable est muette})\\
&\Longleftrightarrow \int_0^t \frac{dc}{c} = \int_0^t -u_x(x(s,x_0)ds\\
&\Longrightarrow \Big[\log c\Big]_0^t = -\int_0^t u_x(x(s,x_0)ds\\
&\Longleftrightarrow \log c(x(x_0,t),t) - \log c(x(0,x_0,0) = -\int_0^t u_x(x(s,x_0)ds
\end{aligned}$$

On en déduit que:

$$c(x(x_0,t),t) = c_0(x_0)\exp\Big\{-\int_0^t u_x(x(s,x_0)ds\Big\} \tag{1.45}$$

On note que $\frac{d}{dt}$ est simplement la dérivation le long de la courbe γ_{x_0} de paramètre t. On voit que la solution de (1.41), (1.42) est donnée par la formule (1.45). Inversement on peut voir que si les courbes $(x(t,x_0),t)$ couvrent le demi-plan supérieur ($t \geqslant 0$), de (x,t), alors (1.45) est une solution de (1.41), (1.42). Les courbes (1.44) sont appelées les **caractéristiques** de (1.43) (pour tout f). La méthode décrite ci-dessus pour calculer la solution c est appelée la **méthode des caractéristiques**. Soit l'équation d'advection suivante (Le but de ce paragraphe est l'étude de l'existence et de l'unicité pour l'équation sans diffusion ie $K = 0$):

$$\begin{cases} \dfrac{\partial c}{\partial t}(x,t) + u\dfrac{\partial u}{\partial x}(x,t) = 0, & x \in \mathbb{R}, t > 0,\\ c(x,0) = c_0(x) & x \in \mathbb{R}. \end{cases} \tag{1.46}$$

On utilise la méthode des caractéristiques : on cherche une fonction $x(t)$ telle que c soit constante sur les courbes $(x(t),t)$. On suppose donc c solution de l'équation (1.46) et on pose $\phi(t) = c(x(t),t)$, on cherche $x(t)$ de sorte que ϕ soit constante. Par dérivation composée,

$$\begin{aligned}
\frac{d\phi(t)}{dt} &= \frac{\partial c}{\partial t}(x(t),t) + x'(t)\frac{\partial c}{\partial x}(x(t),t)\\
&= [u - x'(t)]\frac{\partial c}{\partial x}(x(t),t)
\end{aligned}$$

(car de l'équation (1.46) on tire $\frac{\partial c}{\partial t}(x,t) = -u\frac{\partial c}{\partial x}(x,t)$), qui est nulle si $x'(t) = u$. On en déduit les courbes, appelées caractéristiques de l'équation aux dérivées partielles (1.46):

$$x(t) = ut + constante \tag{1.47}$$

Réciproquement, on peut vérifier que:

$$\forall x \in \mathbb{R} \quad \forall t > 0 \quad c(x-t) = c_0(x-ut) \tag{1.48}$$

Cette expression permet d'obtenir un résultat d'existence et d'unicité suivant:

Theorem 1.8.5. *Si $c_0 \in \mathcal{C}^1$, alors le problème (1.46) admet une unique solution c de classe $\mathcal{C}^1(\mathbb{R}\times\mathbb{R}+)$, donnée par:*

$$\forall x \in \mathbb{R}, t > 0 \quad c(x,t) = c_0(x-ut). \tag{1.49}$$

Résolution numérique du problème: bien que la formule (1.45) est bonne pour la solution, elle n'est pas souvent utilisée pour les calculs de la modélisation de la qualité de l'air. En effet, on se donne un nombre fini d'états occupés par la qualité de l'air noté par $x_1, x_2, ..., x_N$ dans la réalité. On va évaluer la concentration c à chaque état x_i à la date $T_1, T_2, \ldots$ Pour calculer c à (x_j, T_1), la relation (1.45) exige la connaissance de x_0 tel que la caractéristique (courbe) à travers $(x_0, 0)$ passe par (x_j, T_1). Si on veut calculer c à l'instant T_2, on a besoin de connaitre à nouveau le point x_0 correspondant. Pour un temps assez grand T_N, on calcule le point x_0 correspondant. Ainsi on

développe une meilleur méthode calculatoire. Elle est basée sur les différences finies. On subdivise le x-espaces en des intervalles de longueur égale à Δx et l'axe du temps positif en des intervalles de longueur égale à Δt. On souhaite trouver une approximation de $c(j\Delta x, n\Delta t)$ par quelques valeurs c_j^n satisfaisant une certaine condition d'approximation. Pour simplifier, on considère dans un premier cas, le cas où u est indépendant de x, ainsi les équations (1.41) et (1.42) deviennent:

$$\frac{\partial c}{\partial t} + u\frac{\partial c}{\partial x} = 0 \qquad c(x,0) = c_0(x) \tag{1.50}$$

Pour simplifier la résolution numérique de la **valeur initiale du problème** (1.50), on remplace les dérivées par les différences finies. Cependant, avant de commencer on introduit un point **du treillis résolu** dans $x-t$ donné par $x = j\Delta x$, $t = n\Delta t$, $j = 0, \pm 1, \pm 2, ..., n = 0, 1, 2, 3, \ldots$ L'approximation à $c(j\Delta x, n\Delta t)$ est désignée par c_j^n, comme précisée-ci dessous. L'exposant "n" n'est pas une puissance.

La méthode des différences finies consiste à approximer les dérivées des équations de la physique au moyen de développement de Taylor.

$$\frac{\partial c(x,t)}{\partial x} = \lim_{\Delta x \longmapsto 0} \frac{c(x+\Delta x) - c(x,t)}{\Delta x}$$

Le développement de Taylor de $u(x+\Delta x, t)$ au voisinage de x donne:

$$c(x+\Delta x, t) = c(x,t) + \Delta x\frac{\partial c(x,t)}{\partial x} + \frac{\Delta x^2}{2!}\frac{\partial^2 c(x,t)}{\partial x^2} + \frac{\Delta x^3}{3!}\frac{\partial^3 c(x,t)}{\partial x^3} + ...$$

En tronquant la série au premier ordre on a:

$$\frac{c(x+\Delta x,t) - c(x,t)}{\Delta x} = \frac{\partial c(x,t)}{\partial x} + o(\Delta x)$$

En tenant compte de la notation indicielle précédente, on a les approximations suivantes:

$$\frac{\partial c}{\partial t} \approx \frac{c(j\Delta x, (n+1)\Delta t) - c(j\Delta x, n\Delta t)}{\Delta t} \approx \frac{(c_j^{n+1} - c_j^n)}{\Delta t}$$

et

$$\frac{\partial c}{\partial x} \approx \frac{c(j\Delta x, n\Delta t) - c((j-1)\Delta x, n\Delta t)}{\Delta x} \approx \frac{(c_j^n - c_{j-1}^n)}{\Delta x}$$

Alors l'équation (1.50) devient:

$$\frac{c_j^{n+1} - c_j^n}{\Delta t} + u\frac{c_j^n - c_{j-1}^n}{\Delta x} = 0$$

$\Longleftrightarrow$

$$c_j^{n+1} - c_j^n + u\frac{\Delta t}{\Delta x}(c_j^n - c_{j-1}^n) = 0$$

soit

$$c_j^{n+1} = c_j^n - u\frac{\Delta t}{\Delta x}(c_j^n - c_{j-1}^n) \tag{1.51}$$

De c_j au **temps** $n+1$ donné explicitement par le c_j au temps n, on fait référence à la formule (1.51) comme une formule explicite du schéma des différence finies. Le schéma des différences finies décrit ci-dessous est appelé **schéma décentré avant** ou **Schéma Upwind** ou encore **en avancé t, en arrière** x (peut-on deviner pourquoi ?). De (1.51), on remarque que si les valeurs de c sont explicitement connues dans la ligne $n = 0$ alors elles sont connues dans la ligne $n = 1$. Répétant ainsi ce processus, on peut alors déterminer tous les c_j^n pour la valeur initiale c_j^0 (voir figure 1.13) Deux questions surgissent: En premier, si Δx et Δt sont rendus suffisamment petit, est ce que c_j^n approche la valeur $c(j\Delta x, n\Delta t)$ aux points de la maille dans un certain sens ? La seconde, quand on garde Δx et Δt fixés, est ce que C_j^n reste t-il borner quand $n \longrightarrow +\infty$ quand j l'est uniformément (dans un certains sens) ? Si la réponse de la première question est oui on dit que le schéma des

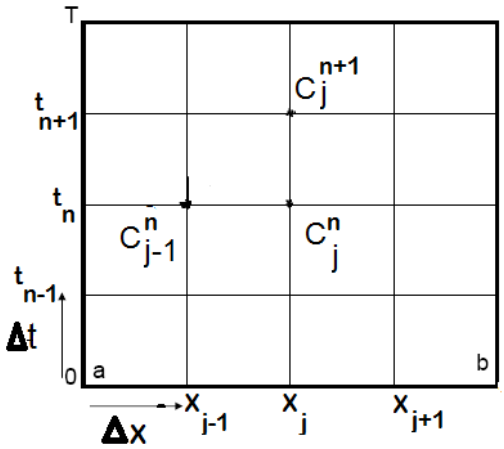


FIGURE 1.13 – Domaine de discrétisation

différences finies est convergente. Si la réponse de la seconde question est oui, on dit que le schéma est stable. On note que la seconde question s'intéresse vraiment au comportement de la solution au problème de discrétisation, la variable temps devient très grande: la solution est-elle bornée ou est ce qu'elle explose quand $n \longrightarrow +\infty$? Il apparait que la question de la convergence est plus pertinente. Cependant, on trouve que la stabilité est plus facile à vérifier. Heureusement il y a un théorème qui dit, en parlant rudement que si le problème (différentiel) de la valeur initiale est bien posé (il est intéressant de représenter la situation physique par exemple) et si on fait un travail raisonnable de remplacer les dérivées par les différences finies, alors la stabilité implique la convergence. La quantité

$$\sigma = \frac{u\Delta t}{\Delta x} \tag{1.52}$$

joue un rôle fondamental dans le schéma (1.51): si $0 < \sigma \leq 1$ alors le schéma numérique (1.51) converge. Cependant, si $\sigma > 1$ alors le schéma ne converge pas ie quand Δx et Δt deviennent très petits la solution de la différence finie développe de plus grandes oscillations.

En bref, le schéma est stable si $0 < \sigma \leq 1$ et instable si $\sigma > 1$.

La formule explicite du schéma des différences finies peut être étendue en général avec $u(x)$; elle est sous la forme

$$c_j^{n+1} = c_j^n - \frac{u_j \Delta t}{\Delta x}(c_j^n - c_{j-1}^n) - \frac{\Delta t}{\Delta x}(u_j - u_{j-1})c_j^n \tag{1.53}$$

où $u_j = u(j\Delta x)$.

NB: La concentration d'une espèce chimique est toujours une quantité non négative. Cela peut être vu en (1.45): si initialement $c_0 \geq 0$ alors $c \geq 0$ pour tout temps future. Cependant, ce résultat peut ne peut nécessairement en approximation avec la formule (1.51). Il se peut que même si $c_0 \geq 0$, c_i^n peut devenir négative pour quelques indices n, i. Bien sûr, comme Δt et Δx tendent vers 0, les valeurs négatives des quantités c_i^n doivent tendre vers 0 quand le schéma est convergent.

Upwind scheme

On pose:

$$c_t = \frac{\partial c_i}{\partial t} \quad , \quad u_x = u = constante \quad et \quad c_x = \frac{\partial c_i}{\partial x}, \quad on \quad a \quad :$$

$$c_t + uc_x = 0 \tag{1.54}$$

$$c(t,0) = c(t,1) \tag{1.55}$$

$$c(0,x) = g(x) \tag{1.56}$$

il vient:

$$g(0) = g(1) \tag{1.57}$$

voir même en valeur algébrique:

$$g'(0) = g'(1) \tag{1.58}$$

On cherche une fonction g qui vérifie cette condition. Dans notre cas de l'advection unidimensionnelle avec un champs de vent non divergent $u > 0$, une grille régulière de pas Δx et un pas de temps Δt. Si on note $\sigma = u\frac{\Delta t}{\Delta x}$ le nombre de Courant, l'évolution de la concentration du traceur pour la maille j est donnée dans le **Schéma de Godunov** par :

$$c_j^{n+1} - c_j^n = \sigma(c_{j-1}^n - c_j^n) \tag{1.59}$$

Quand l'on applique au rétro-traceur $c*$ avec le vent $-u$, le schéma amont s'écrit:

$$c_j^{*n} - c_j^{*n+1} = \sigma(c_{j+1}^{*n+1} - c_j^{*n+1}) \tag{1.60}$$

Résultat

Dans l'équation d'advection (1.54) le signe de la constante u détermine si le transport de l'espèce se fait de $x = 0$ à $x = 1$ ou à l'inverse. Avec cette méthode, l'information qu'on a besoin pour trouver le pas est assemblé et utilisé sur le schéma Upwind.

La discrétisation (n est le pas de temps, j pas d'espace) est:

$$\begin{cases} c_j^{n+1} & = & c_j^n + u\frac{\Delta t}{\Delta x}(c_j^n - c_{j-1}^n) & si \quad u\Delta t > 0 \\ c_j^{n+1} & = & c_j^n - u\frac{\Delta t}{\Delta x}(c_{j+1}^n - c_j^n) & si \quad u\Delta t < 0 \end{cases}$$

Et à cause du transport il y a deux discrétisations différentes. On pourrait imagine la méthode Downwind (différence arrière), mais il n'est pas numériquement stable et en réalité c'est général d'avoir de l'information d'avance dans le temps.

On note que la stabilité du schéma Upwind dépend de la condition du CFL (1.52):

$$\mid \sigma \mid = \mid \frac{u\Delta t}{\Delta x} \mid \leqslant 1$$

Sur une grille avec $N = 100$, nombre de points dans l'espace et $M = 700$, nombre de points en temps, prenons la fonction $g(x) = sin(\pi * x)$. On simule ici le transport $c_t + uc_x = 0$ avec le schéma Upwind.

Prenons $x \in [0, 1]$ et $t \in [0, T]$ avec la condition initiale $c(x, 0) = c_0(x)$,

$$u = 2,\ T = 1,\ I = [0, 1],$$

$$TT_{uw}(1) = TT(1) = M$$

et

$$XX_{uw}(1) = xx(1) = N,$$

$$\Delta t = \frac{t_{end} - t_0}{M} \quad et \quad \Delta x = \frac{1}{N}$$

On a en annexe A un bout de code matlab qui nous a donné cette simulation.

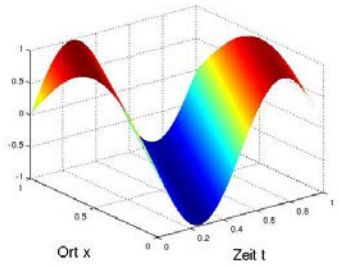


FIGURE 1.14 – simulation sous matlab de l'équation du transport linéaire (advection) avec T=1, $I = [0, 1]$, un vent de vitesse $u = 2$, $\Delta x = 0,01$, $\Delta t = 0,00142$, $\sigma = 0,284 \leq 1$: on a la stabilité du schéma

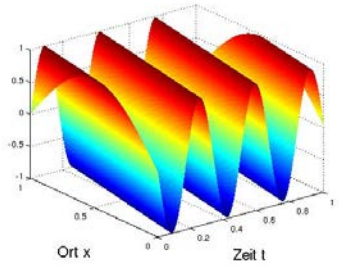


FIGURE 1.15 – simulation sous matlab de l'équation du transport linéaire(advection) avec $T = 1$, $I = [0, 1]$, un vent de Dakar avec une vitesse $u = 6$,$\Delta x = 0,01$,$\Delta t = 0,00142$, $\sigma = 0,852 \leq 1$: on a la stabilité du schéma

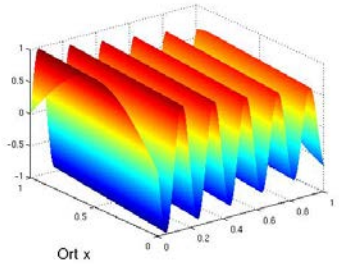


FIGURE 1.16 – simulation sous matlab de l'équation du transport linéaire(advection) avec T=3,$I = [0, 1]$, un vent $u = 10,4$ (voir 1.2), $TT(1) = 7000$, $XX(1) = 100$: on a toujours la stabilité du schéma.

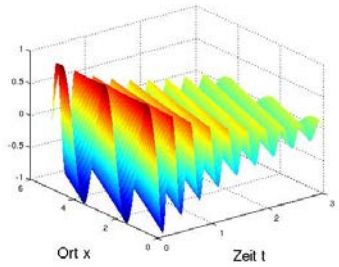


FIGURE 1.17 – simulation sous matlab de l'équation du transport linéaire(advection) avec T=3,$I = [0, 5]$, un vent $u = 6$, $TT(1) = 700$, $XX(1) = 100$, $\sigma > 1$: on a un schéma instable

1.9 Conclusion:

Dans ce chapitre on a un ensemble d'outils et d'analyses portant exclusivement sur la modélisation de la pollution atmosphérique, et sur l'évaluation des modèles. L'objectif est de fournir aux éventuels chercheurs qui vont s'intéresser dans ce domaine, un ensemble cohérent d'informations leur permettant une utilisation opérationnelle des modèles, souvent relativement simples, dédiés à la simulation des concentrations de polluants. Ce chapitre nous introduit dans la thèse avec de notions de bases qui vont être exploitées dans les prochains chapitres.

Chapitre 2

Qualité de l'air à Dakar

Sommaire

2.1 Introduction

Dans la plupart des grandes villes du monde, du fait d'intenses activités humaines liées au trafic automobile et à la production industrielle, la qualité de l'air se détériore. Cela cause des problèmes environnementaux et des risques graves de santé pour les habitants . C'est dans ce cadre que Dakar, une des grandes villes d'Afrique à forte densité humaine, développe des stratégies pour faire face à cette dégradation. Dans ce chapitre de la thèse, on présente Dakar avec ses cinq stations de surveillance des polluants. Le chapitre se termine par la genèse de données sur l'indice de la qualité de l'air en vue de faire sa modélisation et prédiction les chapitres 4 et 5.

2.2 Présentation

Située entre les 17°10 et 17°32 de longitude Ouest et les 14°53 et 14°35 de latitude Nord, localisée à l'extrême ouest du Sénégal, la région de Dakar est limitée à l'Est par la région de Thiès et par L'Océan Atlantique dans ses parties Nord, Ouest et Sud. Elle a une superficie de $550km^2$ soit 0,28% du territoire national. Depuis son érection en capitale de l'AOF, Dakar est une mégapole africaine de plus de 3 millions d'habitants, qui concentre 25% de la population du Sénégal et 80% des activités économiques [15, 16]. La région est restée une plaque tournante et sa position stratégique a renforcé son rayonnement au plan culturel et politique. A ces fonctions culturelles, économiques

et politiques viennent s'ajouter celles touristiques qui s'exercent dans un contexte géographique particulièrement favorable. Administrativement, la ville de Dakar est composée 19 Communes d'arrondissement. Sur le plan physique, Dakar est une presqu'île dont les caractéristiques ont

FIGURE 2.1 – Région de Dakar.

été décrites par Diaw et Mbow [15]. Au sud-est de la presqu'île s'élève le massif de Ndiass, qui correspond à un bas plateau avec une altitude maximum de 95 m. Il présente un relief de collines et de plateaux souvent cuirassés, couverts de litho-sols et de sols ferrugineux. Le long de la côte, les buttes de grès rouges sont bordées par des falaises. Dans la région de Rufisque-Bargny s'étendent des bas plateaux, dont la surface recoupe les calcaires et marnes éocènes. Sur ces terrains, des sols calcimorphes bruns alternent avec des sols vertiques gris-noirs. La majeure partie de la presqu'île est occupée par des dunes continentales fixées (Ogolien). Ces anciens cordons dunaires, orientés NE-SO portent des sols ferrugineux non lessivés. Sur ces dunes on note de fortes installations humaines (Cambérène, Yeumbeul, Malika). Au niveau des dépressions inter dunaires apparaissent des sols hydromorphes: ce sont les Niayes inondées par la nappe phréatique. Des dunes littorales vives ou semi-fixées s'étirent le long de la côte nord. Ces dunes récentes et actuelles, à sols minéraux bruts, ont isolé des lacs salés témoins de la dernière transgression. Ils sont bordés de cordons littoraux et de sols holomorphes. La presqu'île se termine à l'ouest par des reliefs volcaniques. Les buttes des Mamelles, culminant à 105 m, sont les restes d'un plateau édifié au début du Quaternaire. Des plateaux de laves basiques s'étendent autour et portent des sols vertiques. Les petits plateaux du cap Manuel à Dakar et de l'île de Gorée sont constitués de laves de la fin du Tertiaire. Tous ces reliefs volcaniques forment une côte rocheuse très échancrée [15]. Nous ne pouvons finir ce paragraphe sans dire un petit mot sur les vents qui jouent un rôle capital dans notre thèse. Il faut noter que le vent est un principal facteur météorologique de la dispersion des polluants. A Dakar, les vents dominants ont été majoritairement de nord avec une plus grande fréquence de vitesses élevées, supérieures à 6 m/s (21 km/h). En mars 2012, les vitesses les plus fréquentes sont comprises entre 7,56 à 14,7 km/h. Des vents plus forts (entre 4,7 km/h et 21,6 km/h) ont été observés pendant ce mois (figure 2.2). Nous allons dans notre étude nous intéresser à la qualité de l'air à la capitale Dakar chef lieux de la région de Dakar.

Le climat

Dakar bénéficie d'un climat de type canarien qui subit fortement l'influence des facteurs géographiques et atmosphériques. Par la présence d'une façade maritime ceinturant presque toute la

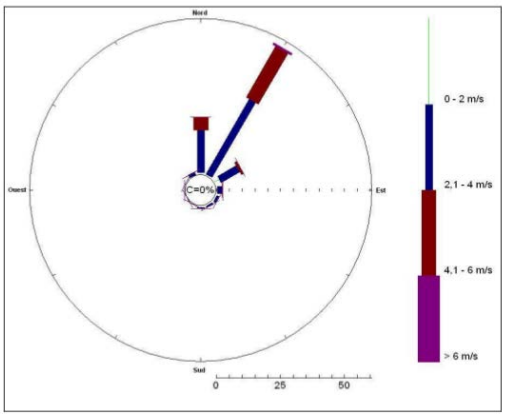


FIGURE 2.2 – Rose des vents à Dakar en mars 2012

ville, il est caractérisé, pendant une bonne partie de l'année, par un microclimat côtier marqué par l'influence de l'alizé maritime. D'où l'existence d'une fraîcheur et d'une humidité quasi permanente, relativement forte de l'ordre de 25%. Toutefois, l'harmattan, alizé continental saharien, se fait sentir faiblement en saison sèche et au fur et à mesure que l'on s'éloigne des côtes.

La température varie entre 17°et 22 °Cde Décembre à Avril et de 22°à 30 °Cde Mai à Novembre (saison des pluies).

Le régime des vents est marqué par l'influence prédominante de l'alizé. Ce dernier est issu de l'anticyclone des Açores.

La pluviométrie est caractérisée par la durée relativement courte de l'hivernage (3 à 4 mois de Juillet à Octobre) par rapport aux régions Sud du pays.

2.3 Centre de Gestion de la Qualité de l'Air

Le suivi de la qualité de l'air à Dakar est assuré par le Centre de Gestion de la Qualité de l'Air (CGQA) qui a été inauguré le 17 mars 2010. Le CGQA dispose d'un laboratoire de suivi et de contrôle des émissions atmosphériques. Il renseigne sur le niveau de pollution de l'air(suivi des immissions) dans Dakar. Il est une entité de la Direction de l'Environnement et des Etablissements Classés (DEEC) et dispose de cinq stations de mesure dans la ville de Dakar. Ces stations fixes sont complétées par un laboratoire mobile qui effectue des mesures dans des endroits ciblés.

Les missions du CGQA sont :

1. d'assurer la veille sur la pollution de l'air ambiant ;
2. d'évaluer les rejets de polluants à la source ;
3. de favoriser la mise en place d'un observatoire de la qualité de l'air ;
4. d'informer le public sur l'état de la qualité de l'air ;
5. de fournir des rapports sur la pollution de l'air pour une meilleure prise de décision.

L'équipe du CGQA est composée d'experts en modélisation, en informatique, en assurance et contrôle qualité et de techniciens instruments.

2.4 Les stations et leur localisation

Pour suivre et évaluer la qualité de l'air à Dakar, cinq stations de mesure sont installés à travers la ville. Il s'agit des stations de Médina, de Bel Air, de Yoff, de la Cathédrale (Bd République) et des HLM (voir figure 2.3) sont présentés ci-dessous. Dans les paragraphes suivants, nous présenterons les cinq stations avec leurs caractéristiques.

FIGURE 2.3 – les cinq stations de mesure de la qualité de l'air à Dakar

FIGURE 2.4 – Carte de la ville de Dakar avec les cinq stations. Source : Google maps modifié par C. Demay, 2011

Station de Médina Elle est placée dans l'enceinte de l'hôpital Abass Ndao. C'est une station de type suburbain situé en bordure de route au niveau d'une intersection munie de feux rouges. Elle est située entre le 14°41'14" et le 17°26'54" et mesure: le monoxyde de carbone (CO), les particules de poussière (PM_{10}), les oxydes d'azote (NO_X) avec le dioxyde d'azote (NO_2) et le monoxyde d'azote (NO).

Station de Bel Air Située entre le 14°40'50" et le 17°25'58" et se trouve dans la zone portuaire avec une grande partie des unités industrielles de la région de Dakar. C'est donc une station de type industriel. Elle mesure essentiellement le dioxyde de souffre (SO_2), les particules de poussière (PM_{10} et $PM_{2.5}$), les oxydes d'azote (NO_X) avec le dioxyde d'azote (NO_2) et le monoxyde d'azote (NO), et les BTX (Benzène, Toluène, Xylènes, ...).

Station de Yoff La station de Yoff est une station de type régional située en bordure de l'océan atlantique à 14°44'51" et 17°27'35". Elle évalue la pollution de fond. Elle mesure principalement les particules de poussières (PM_{10}), l'ozone (O_3), et les oxydes d'azote (NO_X) avec le dioxyde d'azote (NO_2) et le monoxyde d'azote (NO).

Station de HLM Elle est située à 14°42'37" N et 17°26'54" W et c'est une station de type périurbain située en proche banlieue. Elle est dotée des analyseurs qui mesurent le dioxyde de souffre (SO_2), les oxydes d'azote (NO_x) avec le dioxyde d'azote (NO_2) et le monoxyde d'azote (NO), l'ozone (O_3), et les particules de poussières (PM_{10}). Elle mesure aussi les paramètres météorologiques: température, humidité relative, radiation nette, pression, vitesse et direction du vent.

Station de la Cathédrale ou Boulevard de la République Elle est à 14°40'14" N et 17°26'11" W. Elle est de type urbain et est placée en plein centre ville, en bordure de route. Elle est munie des analyseurs qui mesurent le monoxyde de carbone (CO), le dioxyde de souffre ($SO2$), les oxydes d'azote (NOx) avec le dioxyde d'azote (NO_2) et le monoxyde d'azote (NO), l'ozone (O_3), les particules de poussières (PM_{10} et $PM_{2.5}$).

2.5 Les Polluants

Un polluant désigne un agent physique, chimique ou biologique qui provoque une gêne ou une nuisance dans le milieu liquide ou gazeux. Au sens large, le terme désigne des agents qui sont à l'origine d'une altération des qualités du milieu, même s'ils y sont présents à des niveaux inférieurs au seuil de nocivité. Du fait de leur nombre élevé dans l'atmosphère seuls quelques uns sont suivis par le CGQA. En effet, ils sont représentatifs des types de pollution (fixe, mobile ou surfacique) d'une part et d'autre part, leurs effets nuisibles pour l'environnement et/ou la santé ont été démontrés. Les polluants sont les indicateurs de pollution atmosphérique et font l'objet de réglementations. Dans ce qui suit on présentera les polluants atmosphériques suivis par les CGQA et leurs effets sur la santé et l'environnement.

2.5.1 Le SO_2, Dioxyde de soufre

Leurs émissions proviennent de la teneur en soufre des combustibles comme le gazole, le fuel, le charbon... Elles sont émises dans l'atmosphère par les cheminées des usines, les centrales thermiques. Ces émissions proviennent faiblement du secteur automobile Diesel.

2.5.2 Particulate Matter (PM)

Les PM, en anglais Particulate Matter, sont les particules en suspension dans l'air. Le CGQA mesure les PM_{10} (particules de diamètre inférieur $10\mu m$) et les $PM_{2,5}$ (particules de diamètre inférieur $2,5\mu m$). Les émetteurs de ces particules à Dakar sont le transport routier, les combustions

industrielles, l'incinération des déchets et certains phénomènes naturels. Les $PM_{2,5}$ progressent plus profondément dans l'appareil respiratoire. Selon leur taille (granulométrie), les particules pénètrent plus ou moins profondément dans l'arbre pulmonaire. Les particules les plus fines peuvent, à des concentrations relativement basses, irriter les voies respiratoires inférieures et altérer la fonction respiratoire dans son ensemble. Elles ont des propriétés mutagènes et cancérigènes. Les effets de salissure des bâtiments et des monuments sont les atteintes à l'environnement les plus observées.

2.5.3 Les NO_x, Oxydes d'azote

Les oxydes d'azote désignent principalement le monoxyde d'azote (NO) et le dioxyde d'azote (NO_2). Le NO se forme lors de réactions de combustion à haute température, par combinaison du diazote (N_2) et de l'oxygène atmosphérique (O_2). Il est ensuite oxydé en dioxyde d'azote (NO_2). Les sources principales sont les transports (routiers, maritime et fluvial), l'industrie, l'agriculture. Les NO_x sont émis aussi à l'intérieur des locaux où fonctionnent des appareils au gaz tels que gazinières, chauffe-eau, etc.

Les moteurs diesel en rejettent deux fois plus que les moteurs à essence catalyses. Mais la plupart des véhicules au Sénégal ne sont pas équipés avec des pots catalytiques pour réduire ces rejets. Le (NO) émis par les pots d'échappement est oxydé par l'oxygène et se transforme en dioxyde d'azote (NO_2).

Le NO_2 est un gaz irritant pour les bronches. Chez les asthmatiques, il augmente la fréquence et la gravité des crises. Chez l'enfant, il favorise les infections pulmonaires. Le NO_2 participe aux phénomènes des pluies acides, à la formation de l'ozone troposphérique, dont il est l'un des précurseurs, à l'atteinte de la couche d'ozone stratosphérique et à l'effet de serre.

2.5.4 O_3, l'ozone

C'est un gaz contenant trois atomes d'oxygène dans chaque molécule. Présent dans la haute atmosphère, il protège la Terre de la majorité du rayonnement ultraviolet du soleil. Dans la basse altitude, ce gaz est nuisible lorsque sa concentration augmente fortement. Cette réaction s'observe entre le dioxyde d'azote et les hydrocarbures. Elle est produite lors d'un fort ensoleillement, températures élevées, phénomène d'inversion de température, faible humidité et absence de vent. La nuisance de l' ozone s'observe au niveau de la muqueuse et peut entraîner une diminution de la fonction respiratoire, particulièrement chez les personnes sensibles (personnes ayant une insuffisance respiratoire, personnes âgées, ...). L'ozone a également un effet sur les végétaux (nécroses, altération de la croissance), entraînant notamment des pertes de production agricole. Il contribue en outre à l'effet de serre. C'est un polluant d'origine anthropique généré par les activités humaines.

2.5.5 Le CO, Monoxyde de carbone

Gaz inodore, incolore et inflammable, le monoxyde de carbone (CO) se forme lors de la combustion incomplète de matières organiques (gaz, charbon, fiouls, carburants, bois). son rejet est issu principalement du trafic automobile. Des taux importants de CO peuvent être rencontrés quand un moteur tourne au ralenti dans un espace clos. En cas de mauvais fonctionnement d'un appareil de chauffage domestique, des teneurs élevées en CO peuvent être relevées dans les habitations. Le CO se fixe à la place de l'oxygène sur l'hémoglobine du sang, conduisant à un manque d'oxygénation de l'organisme (cœur, cerveau, etc). Les premiers symptômes sont des maux de tête et des vertiges. Ces symptômes s'aggravent avec l'augmentation de la concentration de CO (nausée, vomissements...) et peuvent, en cas d'exposition prolongée, aller jusqu'au coma et à la mort. Le CO participe aux mécanismes de formation de l'ozone troposphérique. Dans l'atmosphère, il se transforme en dioxyde de carbone CO_2 et contribue à l'effet de serre.

2.5.6 Les BTX, Benzène, Toluène Xylène

Les BTX entrent dans la composition des carburants mais aussi de nombreux produits courants : peintures, encres, colles, solvant, etc. Les composés organiques volatils (COV) comprennent notamment Aldehydes, Cétones et Hydrocarbures Aromatiques Monocycliques (HAM) tels que les BTX. Parmi les COV, les BTX font, en raison de leur volatilité importante et du risque sanitaire qu'ils entraînent, l'objet de suivis importants pouvant aller jusqu'à la mesure en continu. Les risques sanitaires vont d'une certaine gène olfactive à des effets mutagènes et cancérigènes, en passant par des irritations diverses et une diminution de la capacité respiratoire. Les COV jouent un rôle majeur dans les mécanismes complexes de formation de l'ozone dans la basse atmosphère (troposphère). Ils interviennent également dans les processus conduisant à la formation des gaz à effet de serre et du "trou d'ozone".

Les composés organiques volatils sont libérés lors de l'évaporation des carburants (remplissage des réservoirs), ou par les gaz d'échappement.

Ils sont émis majoritairement par le trafic automobile, le reste des émissions provenant de processus industriels et éventuellement d'usage domestique de solvants.

Le plomb n'est pas suivi et n'est plus un indicateur de la pollution automobile, car il a été supprimé de l'essence depuis 2005.

2.6 La surveillance

La pollution atmosphérique est un problème de santé publique. Cela fait que l'air est sous surveillance. La surveillance consiste à mettre en œuvre les techniques et protocoles imposés par les autorités, à défaut ceux reconnus par la profession ou validés par les associations de qualité de l'air. Elle vise à mesurer, estimer, comparer et prévoir les niveaux de pollution à différents points de Dakar.

Les effets néfastes de la pollution atmosphérique ont été mis en évidence par des études épidémiologiques. Ils sont cohérents avec les études toxicologiques même si l'ensemble des phénomènes physiopathologiques n'est pas encore expliqué [17]. On distingue deux types d'effets:

1. les effets à court terme, de type cliniques, fonctionnels ou biologiques, qui se manifestent dans des délais brefs (quelques jours à quelques semaines) suite à des variations journalières du niveau des polluants atmosphériques ;
2. les effets à long terme, qui peuvent survenir après une exposition chronique (plusieurs mois ou années) et qui peuvent entraîner une surmortalité et une diminution de l'espérance de vie.

Dans le tableau B.1 de l' annexe, on a les principaux polluants atmosphériques surveillés par les AASQA (Associations Agréées Surveillance Qualité de l'Air), leurs principales sources et leurs effets néfastes sur la santé.

2.7 L'Indice de la Qualité de l'Air (IQA ou iqa)

Définition et mode de calcul

D'après le dictionnaire LAROUSSE (`http://www.larousse.fr/dictionnaires/francais/indice/42580`) un indice est un objet, un fait, un signe qui met sur la trace de quelque chose. Donc nous pouvons définir l'indice de la qualité de l'air comme un objet, un fait ou un signe qui nous met sur la trace de la qualité de l'air. Un indice de la qualité de l'air est un indicateur de qualité de l'air permettant de synthétiser différentes données en une valeur simple. D'après la Fédération ATMO France, il existe deux indices de qualité d'air différents selon la taille de l'agglomération:

1. l'indice ATMO: pour les agglomérations dont la population dépasse 100 000 habitants ;
2. l'indice iqa: pour les agglomérations de taille inférieure à 100 000 habitants.

L'indice ATMO est déterminé à partir des concentrations de quatre polluants: le dioxyde soufre (SO_2), le dioxyde d'azote (NO_2), l'ozone (O_3), le monoxyde de carbone (CO) et les particules

en suspension inférieur à 10 micromètres (PM_{10}). A chaque polluant correspond un sous-indice calculé à partir des concentrations mesurées. Ces sous-indices sont calculés à partir de la moyenne des maxima horaires pour le SO_2, NO_2, O_3 et CO et de la moyenne des moyennes horaires sur pour les PM_{10}. L'indice ATMO est le plus élevé des sous-indices par polluant et par station.

L'indice iqa est un indice ATMO simplifié, il peut être calculé à partir d'un, deux, trois, quatre ou cinq polluants. C'est un chiffre associé à un qualificatif (0-50 bon, 51 à 100 moyen, 101-200 mauvais et supérieur à 200 très mauvais). c'est ce qu'on appelle les 4 classes de l'indice. L'indice de la qualité de l'air (noté iqa ou IQA dans certains ouvrage) est évalué en comparant les niveaux de concentration de chaque polluant avec la valeur de référence de la norme sénégalaise [18]. Les valeurs limites d'immission de polluants ont été établies par l'Association Sénégalaise de Normalisation. Pour évaluer la qualité de l'air globale pour une station de surveillance particulière, un indice est calculé pour chaque polluant mesuré et le maximum est considéré comme l'indice de qualité de l'air pour cette station de surveillance, car il représente le mauvais des polluants mesurés. Cet indice est déterminé à partir des niveaux de pollution mesurés au cours de la journée par les cinq stations, caractéristiques de la pollution générale de l'agglomération de Dakar. Il intègre les principaux polluants atmosphériques, traceurs des activités de transport, urbaines et industrielles:

- Les poussières: PM_{10} et $PM_{2,5}$ (liées aux poussières désertiques, au trafic automobile, au chauffage et aux activités industrielles, mais aussi aux réactions chimiques dans l'atmosphère et aux transferts de pollution sur de grandes distances) ;
- L'ozone est le résultat de réactions chimiques, sous l'effet du rayonnement solaire, entre principalement les oxydes d'azote (NO_x) et les composés organiques volatils (COV : hydrocarbures, solvants, $\cdots$). Du fait du rôle joué par le soleil, les pics d'ozone surviennent principalement en saison sèche (l'été) : un fort ensoleillement et des températures élevées conduisent à des concentrations élevées d'ozone dans l'air ambiant ;
- Le dioxyde de soufre SO_2 (d'origine industrielle) ;
- Le dioxyde d'azoteNO_2 (lié aux transports, aux activités de combustion et de chauffage) ;
- le monoxyde de carbone (lié essentiellement au trafic routier, gaz d'échappement des véhicules).

Parmi les utilisateurs de ce système (iqa), l'Agence Américaine de Protection de l'Environnement (USEPA) a développé un iqa pour cinq principaux polluants, cités ci-dessus, réglementés par la loi sur la qualité de l'air. Pour chaque polluant, l'USEPA a déterminé des standards pour protéger contre les effets sanitaires.

Comment fonctionne l'iqa ?

D'après le CGQA [19], par exemple, une valeur de iqa de 50 représente une bonne qualité de l'air et un faible potentiel d'impact négatif sur la santé, alors qu'une valeur de iqa de 300 représente un air de très mauvaise qualité. La valeur de 100 correspond globalement au standard pour un polluant en-dessous duquel la santé des populations est préservée. Ainsi, des valeurs inférieures à 100 sont satisfaisantes. Quand les valeurs sont supérieures à 100, la qualité de l'air affecte d'abord la santé des populations sensibles, puis celle de tout le monde quand l'iqa devient plus élevé.

Au Sénégal, le Centre de Gestion de la Qualité de l'Air a adopté quatre classes de l'iqa et chaque classe correspond à un niveau d'impact sanitaire selon le groupe de population.

Les classes de l'IQA

Les classes de l'iqa sont symbolisées par les couleurs suivantes: le vert, le jaune, l'orange et le rouge.

Bon (vert) : rien à signaler. L'IQA est satisfaisant et la pollution de l'air pose très peu ou pas de risque sanitaire.

Jaune (moyen) : l'iqa est acceptable. Toutefois, pour certains polluants, il peut y avoir de légers risques sanitaires pour un nombre limité de personnes. Par exemple, les personnes qui ne sont pas d'habitude sensibles à l'ozone pourraient manifester quelques symptômes.

Mauvais (orange) : Certains groupes de personnes sont particulièrement sensibles aux effets nocifs de certains polluants. Ceci signifie qu'ils sont susceptibles d'être affectés pour les plus basses

valeurs que le grand public. C'est le cas pour les enfants et les adultes en activité à l'extérieur. Les personnes atteintes de maladies respiratoires sont soumises à un risque élevé en cas d'exposition à l'ozone, alors que les gens atteints de maladies cardiaques le sont en cas d'exposition au monoxyde de carbone. Avec des valeurs d'iqa entre 150 et 200, tout le monde peut commencer à sentir des effets sanitaires qui sont plus sérieux chez les gens des groupes sensibles.

Très mauvais (rouge) : Des valeurs d'IQA supérieures à 200 déclenchent une alerte sanitaire, car chacun peut ressentir de sérieux effets sur la santé. Avec des valeurs d'iqa supérieures à 300, toute la population est affectée. L'alerte générale doit être déclenchée et des mesures d'urgence doivent être prises.

2.8 Données

Genèse des données de l'iqa

Nous disposons d'une base de données historiques sur l'iqa pendant la période allant du 1^{er} janvier 2010 au 31 décembre 2013. Les données proviennent des cinq stations de mesures dans la ville de Dakar. Ces cinq stations sont équipées des analyseurs d'air. Chaque analyseur est doté d'une pompe lui permettant d'aspirer l'air ambiant et de mesurer les concentrations des polluants qui s'y trouvent. Les résultats sont transmis aux SAM-SK2 (Système d'acquisition de mesure de type SK2). Au niveau des SAM-SK2, par un système de connexion ADSL, les données arrivent chaque quart d'heure au laboratoire du CGQA. A partir du serveur XAIR qui abrite le logiciel Xair-Premium, elles sont converties en fichier excel puis texte et transmises au serveur Airquis. Ce dernier, qui exécute le logiciel du nom Airquis,$\cdots$, effectue le calcul de l'iqa. C'est ainsi que sont obtenues les données qui feront l'objet de notre étude. Nous allons présenter chaque station avec les données obtenues puis nous allons faire une étude sur l'iqa à Dakar au cours de la période de 2010 à 2013.

L'iqa à Dakar Période 2010 à 2013

Dans le tableau suivant, nous présentons l'indice de la qualité de l'air sur la période de 2010 à 2013. C'est un tableau de moyennes mensuelles de l'iqa.

Année	Jan	Fev	Mars	Avril	Mai	Juin	Juil	Août	Sept	Oct	Nov	Dec
2010	90	74	107	57	78	50	41	38	34	49	66	47
2011	80	107	68	62	51	43	49	39	37	65	60	127
2012	137	130	146	76	75	52	43	37	34	35	37	76
2013	107	100	57	59	59	45	36	31	32	26	50	93

Nous constatons que les périodes les plus polluées à Dakar se situent au début et à la fin de l'année. Pour l'année 2011, nous remarquons que les pics de pollution ont commencé à partir de décembre et se poursuivent jusqu'au mois de mars 2012. Nous sommes au dessus de 100 pendant cette période. C'est une période où l'iqa bascule de la classe "jaune" à la classe "orange" et de classe "orange" à la classe "rouge". Nous allons voir en profondeur ce qui passe pendant cette période. Les périodes concernées par les pics de pollution sont la fin 2011 et le début 2012 comme nous pouvons l'observer dans le tableau ci-après:

Annee	Jan	Fev	Mars	Avril	Mai	Juin	Juil	Aout	Sept	Oct	Nov	Dec
2011	80	107	68	62	51	43	49	39	37	65	60	127
2012	137	130	146	76	75	52	43	37	34	35	37	76

Voici un aperçu sommaire sur les données brutes de l'iqa pendant notre période d'étude:

1er	2	3	4	5	6	7	8	9	10	11	12	13	14		
Dec11	112	114	101	84	72	114	106	101	117	79	198	185	243	161	
Jan12	76	176	172	192	196	162	124	108	145	99	97	97	98	141	
Fev12	120	102	76	90	151	251	358	326	194	104	103	112	84	65	
Mar12	125	249	266	238	188	219	171	98	204	322	271	275	166	136	

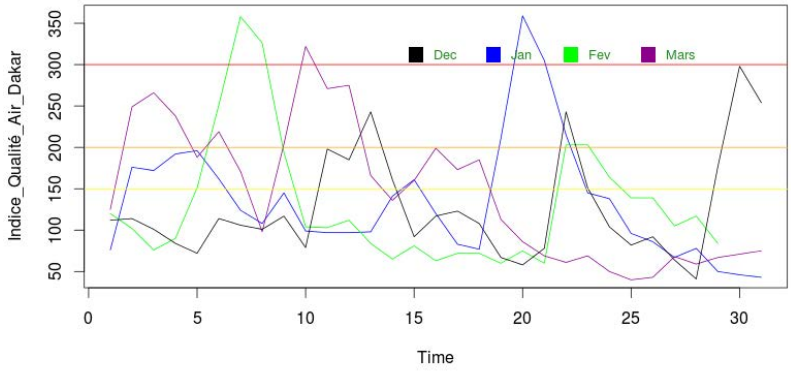


FIGURE 2.5 – Courbe des quatre mois les plus pollués de la période allant de 2010 à 2013: Décembre 2011, Janvier 2012, Février 2012 et Mars 2012.

Nous pouvons observer graphiquement comment évolue l'iqa pendant cette période par la Figure 2.5. Nous remarquons que les différentes classes de l'iqa sont franchies.

Par analyse des données sur l'iqa, nous constatons que le mois de Mars 2012 est le mois le plus pollué sur la période de notre étude. Au cours de ce mois la qualité de l'air est mauvaise à Dakar. Notre affirmation sur la mauvaise qualité de l'air à Dakar durant le mois de mars 2012 est conforme avec le bulletin mensuel du CGQA de mars 2012[20] Les PM_{10} et les $PM_{2,5}$ ont dépassé

IQA	Bel Air (Môle 10)	Bd. République	Médina (Hôp. Abass Ndao)	HLM	Yoff
Bon	3	0	3	1	-
Moyen	8	2	7	5	-
Mauvais	10	9	11	10	-
Très Mauvais	7	3	8	7	-

FIGURE 2.6 – État de la qualité de l'air de Dakar en mars 2012[20].

les valeurs limites fixées par la réglementation sénégalaise et les recommandations de l'OMS. La qualité de l'air a été mauvaise dans l'ensemble avec 43% des indices qui ont été mauvais, 23% moyens et seulement 7% bons [21].

Il faut noter que les données sur la température, la vitesse du vent, l'humidité, le point de rosé, les précipitations, le niveau de mer sont fournies par le site `http://www.wunderground.com/history/airport/GOOY/2013/1/1/CustomHistory.html?dayend=31&monthend=12&yearend=2013&req_city=NA&req_state=NA&req_statename=NA`.

2.9 Conclusion

Le chapitre nous a permis de faire un état de lieux sur la qualité de l'air dans Dakar au cours de la période 2010 à 2013. Cela permet d'affirmer que le mois de Mars 2012 est le mois le plus pollué de cette période. En perspective, il nécessaire d'analyser la situation météorologique de cette année et voir celle du mois de mars 2012 et puis trouver le polluant responsable de la dégradation de la qualité de l'air dans Dakar pendant cette période. Les données sur l'indice de la qualité de l'air dans Dakar étant disponibles, nous allons procéder à sa modélisation et prédiction. Mais avant cela, le chapitre 3 suivant nous présente quelques outils d'optimisation nécessaires pour ce travail.

Chapitre 3

Outils d'Optimisation

Sommaire

3.1 Introduction

L'objet de ce chapitre est de présenter les outils mathématiques de base qui seront utilisés dans les chapitres 4 et 5 dans notre démarche de modélisation et prédiction de l'indice de la qualité de l'air dans Dakar. Le chapitre s'oriente particulièrement vers l'optimisation sans contrainte étayée par les moindres carrés. Mais avant de poursuivre ce chemin, on donne quelques définitions :

Définitions

L'**optimisation** c'est l'ensemble des **techniques permettant de chercher les minima ou les maxima de fonctions ou de fonctionnelles**. Elle intervient dans pratiquement tous les processus de modélisation actuels. Qu'il s'agisse de problèmes directs (ajustement de données, contrôle optimal, résolution de systèmes linéaires par moindres carrées, etc .) ou inverses (identification de paramètres, contrôle des frontières libres etc.), il est rare qu'un problème d'optimisation plus ou moins complexe n'intervienne pas à un stade donnée de la modélisation et/ou de la simulation.

Un modèle tel que considéré dans ce chapitre, est une construction mathématique utilisée pour représenter certains aspects significatifs de problèmes du monde réel. Il y a beaucoup de types

différents de modèles mathématiques, mais nous nous focaliserons dans cette partie de la thèse sur les modèles d'optimisation. Mais qu'est ce qu'un problème d'optimisation ?

Un **problème d'optimisation** en mathématique (ou problème de programmation mathématique) consiste à trouver, parmi un ensemble donné, un élément (pour nous ici un vecteur de $\mathbb{R}^p$ mais ce peut être aussi un vecteur d'entiers, une fonction, ...) minimisant ou maximisant une fonction donnée de cet ensemble sur $\mathbb{R}$.

En Bref, optimiser revient à rechercher parmi un ensemble C de choix possibles le meilleur (s'il existe !). Si J est une application d'un ensemble E dans F . Considérons le problème (P) défini par:

$$(P)\begin{cases} minJ(x) \\ x \in C \subset E \end{cases}$$

Nous avons à pouvoir comparer 2 choix et donc avoir une structure d'ordre sur l'ensemble F. Généralement on prend toujours $F = \mathbb{R}$. Suivant les domaines d'applications :

- E s'appelle l'ensemble des stratégies, des états, des paramètres, l'espace des variables [1] ;
- C est l'ensemble des contraintes [2] ;
- J est la fonction coût, économique ou le critère, l'objectif [3].

Une fois le problème bien défini, il se pose deux questions. La première est de savoir si (P) admet une solution. Si la réponse est positive, il nous faut trouver la ou les solutions. Suivant la nature de l'ensemble E les réponses sont plus ou moins faciles. Si E est fini, l'existence de solution est évidente, mais le calcul est difficile si le nombre d' éléments est grand. Par contre si $E = \mathbb{R}^n$ ou est de dimension infinie (problème de la brachistochrone [4], le problème de transfert orbital [5]) la question de l'existence de solution est moins triviale, mais si les fonctions sont dérivables il est "plus" facile de calculer la solution.

Le problème d'optimisation peut aussi se présenter sous la forme [29] :

$$(P)\begin{cases} x^* \in C \\ J(x^*) = \inf\limits_{x \in C} J(x) \end{cases}$$

où $J : E \longrightarrow \mathbb{R}$ étant une application et $C \subset E$.

On dit que J est sous la contrainte $x \in C$.

Si $C = E$ on a un problème sans contrainte. En remplaçant J par $-J$ on transforme un problème de minimisation en un problème de maximisation.

Un point x^* est un **minimum local** de J sur C s'il existe un voisinage V de x^* tel que $J(x^*) \leqslant J(x)$ pour tout $x \in C \bigcap V$. Un point x^* est un **minimum local strict** de J sur C s'il existe un voisinage V de x^* tel que $J(x^*) < J(x)$ pour tout $x \in C \bigcap V$, $x = x^*$.

Une suite $(x_n)_{n\in\mathbb{N}}$ de points de C est une **suite minimisante** si $\lim\limits_{n\to+\infty} J(x_n) = \inf\limits_{x\in C} J(x)$.

Une telle suite existe toujours, par définition de inf, par contre on ne sait rien au sujet de sa convergence éventuelle.

La résolution du problème (P) nous impose des interrogations suivantes :

- L'existence de $\inf\limits_{x\in C} J(x)$: J est-il borné inférieurement ?
- L'infimum est-il atteint dans C ? Autrement dit existe t-il $x^* \in C$ vérifiant $J(x^*) = \min\limits_{x\in C} J(x)$?
- L'unicité de x^*: comment se présente la taille de l'ensemble des solutions ?
- Comment peut-on caractériser x^* ? ie peut-on trouver des conditions nécessaires pour caractériser un minimum ?
- Exhiber un algorithme d'optimisation pour déterminer les solutions de (P)

1. Les variables représentent les composantes du modèle qui peuvent être modifiées pour créer des configurations différentes.
2. Les contraintes représentent les limitations sur les variables.
3. Le terme "bjectif" vient du fait que l'objectif est d'optimiser cette fonction.
4. Le problème de la brachistochrone fut posé par Jean Bernoulli en 1696 et est considéré comme le problème fondateur du calcul des variations.
5. L'objectif de ce problème est de trouver une loi de commande du moteur qui réalise le transfert et qui minimise le temps de transfert. C'est un problème de contrôle optimal et l'inconnue est la commande

Répondre à ces interrogations nous pousse à voir

i la structure de E : espace vectoriel, muni d'une norme, d'un produit scalaire, de dimension finie ou infinie, ...

ii les propriétés de $C \bigcap E$: fermé, borné, convexe,...

iii les propriétés de $J : E \longrightarrow \mathbb{R}$: continuité, différentiabilité, convexité,...

Dans les sections qui suivent on va donner quelques éléments de réponse à ces questions. Avant de répondre à ces interrogations voici quelques problèmes types d'optimisation :

- calcul d'une trajectoire de rentrée dans l'atmosphère d'une navette
- déménageur de Piano,
- sac à dos, bibliothécaire, emploi du temps,
- voyageur de commerce, ...

On a dans ce qui suit un exemple appliqué à la modélisation et prédiction de l'indice de la qualité de l'air dans Dakar:

Exemple 3.1.1. *y désigne ici l'indice de la qualité de l'air dans la ville de Dakar. y est une variable aléatoire dépendant linéairement de n variables aléatoires $x_1, \ldots, x_n($*

Y: indice de la qualité de l'air (nombre sans unité) ; x_1 température maximale (en degré Celsius), x_2 point de rosée (en degré Celsius), x_3 humidité maximale (%) ; x_4: pression maximale au niveau de la mer (hPa), x_5 visibilité maximale (km), x_6 vitesse maximale du vent (Km/h), x_7 niveau de précipitation (mm), etc.), selon la relation :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n = \sum_{i=0}^{n} \beta_i x_i \tag{3.1}$$

On cherche, les coefficients $\beta_0, \beta_1, \ldots, \beta_n$ (coefficient de régression). Pour cela on effectue p observations (ou mesures) portant sur les variables x_i et y ; on note $Y = [y_1, y_2, \ldots, y_p]$ le vecteur des valeurs observées de y et $X = [x_{ij}], i = 1, \ldots, n, j = 1 \ldots p$, la matrice observée: x_{ij} est la $j - ième$ observation de la variable x_i. Le problème d'identification des paramètres β se formule alors de la manière suivante:

$$\underset{(\beta_0, \beta_1, \ldots, \beta_n \in) \mathbb{R}^n + 1}{min} \sum_{j=1}^{p} [y_j - (\beta_0 + \sum_{i=1}^{n} \beta_i x_{ij})]^2 \tag{3.2}$$

On revient sur le problème en détaille dans le prochain chapitre.

Classification

Considérons notre problème d'optimisation (P) suivant :

$$(P) \begin{cases} min f(x) \\ x \in C \subset E \end{cases}$$

Suivant la nature des ensembles C et E et de la fonction J nous avons différentes classes de problème d'optimisation. La figure 3.1 donne une classification des problèmes d'optimisation.

3.2 Rappels

Espace vectoriels normés

Dans tout ce qui suivra E désigne un espace vectoriel réel.

Définition 3.2.1. *Une distance sur E est une fonction $d : E \times E \longrightarrow \mathbb{R}$ vérifiant les conditions suivantes pour tous x, y, z dans E :*

(i) *$d(x, y) = 0$, si et seulement si $x = y$, (identité des indiscernables) ;*

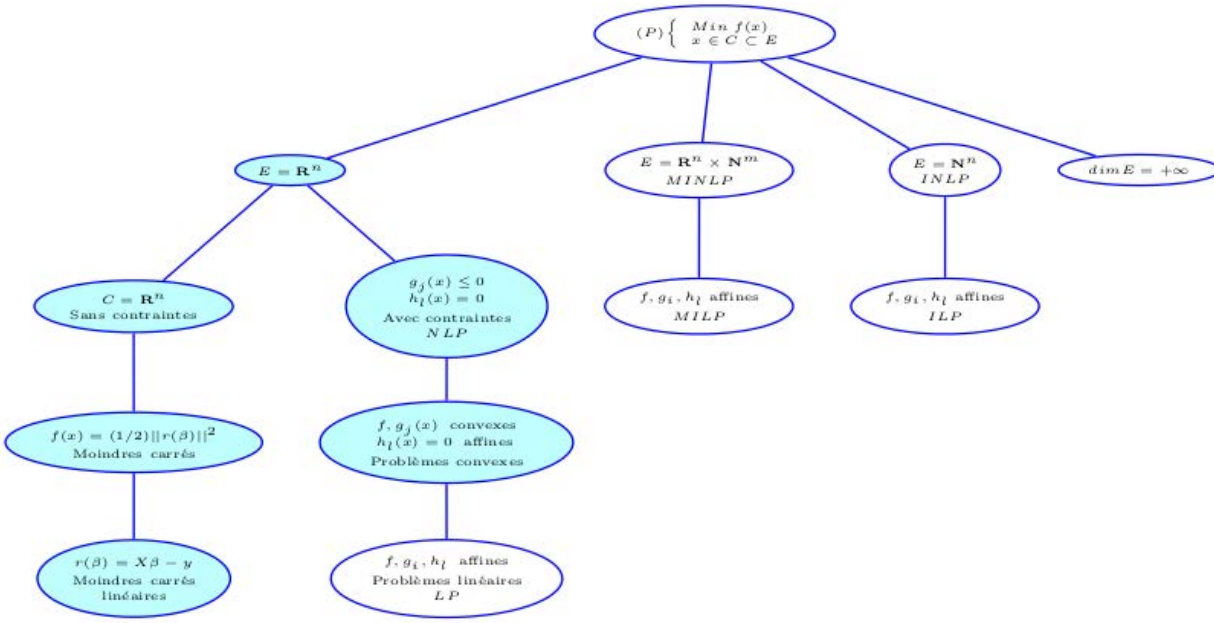


FIGURE 3.1 – Classification des problèmes d'optimisation.

(ii) $d(x,y) = d(y,x)$, *(symétrie) ;*

(iii) $d(x,z) \leq d(x,y) + d(y,z)$, *(inégalité triangulaire).*

Le couple (E,d) *est appelé un espace métrique.*

Une suite (x_n) dans un espace métrique est dite suite de Cauchy si pour tout $\epsilon > 0$ il existe un $N \in \mathbb{N}$ tel que $\forall n, m > N$ on a $d(x_n, x_m) < \epsilon$. On rappelle que dans un espace métrique toute suite convergente est de Cauchy.

Définition 3.2.2. *Un espace métrique* (E,d) *est complet, si toute suite de Cauchy de* E *a une limite dans* E*.*

Définition 3.2.3. *On appelle* ***norme*** *sur* E*, une application* $x \longmapsto \|x\|$ *de* E *dans* $\mathbb{R}+$ *vérifiant les propriétés suivantes:*

(1) $\forall x \in E : \|x\| = 0 \Leftrightarrow x = O_E$

(2) $\forall x \in E; \forall \lambda \in \mathbb{R} : \|\lambda x\| = |\lambda| \|x\|$

(3) $\forall (x,y) \in E \times E : \|x+y\| \leq \|x\| + \|y\|$

Exemple 3.2.4. *On définit sur* $E = \mathbb{R}^n$*, pour* $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ *les normes suivantes:*

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|, \quad \|x\|_2 = (\sum_{i=1}^{n} |x_i|^2)^{\frac{1}{2}}, \quad \|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$$

On a les inégalités classiques suivantes:

$$\forall x \in \mathbb{R}^n : \|x\|_\infty \leq \|x\|_2 \leq \|x\|_1 \ \text{ et } \|x\|_1 \leq \sqrt{n}\|x\|_2 \leq n\|x\|_\infty. \tag{3.3}$$

Exemple 3.2.5. *Soit* $E = \mathcal{C}([a,b], \mathbb{R})$*, l'espace vectoriel des fonctions réelles continues sur l'intervalle* $[a,b]$ $(\dim(E) = \infty)$*. Pour tout* $f \in E$ *on définit les normes suivantes:*

$$\|f\|_1 = \int_a^b |f(t)|dt, \quad \|f\|_2 = (\int_a^b |f(t)|^2 dt)^{\frac{1}{2}}, \quad \|f\|_\infty = \max_{[a,b]} |f(t)|.$$

On a les inégalités[6] *ci-dessous :*

$$\forall f \in E : \|f\|_1 \leq \sqrt{b-a}\|f\|_2 \leq (b-a)\|f\|_\infty \tag{3.4}$$

Un **espace vectoriel normé (e.v.n)** est un espace vectoriel muni d'une norme.
Étant donné un e.v.n $(E, \|.\|)$, on peut définir une distance sur E par $d(x,y) := \|x-y\|$.
La distance d est dite **induite par** la norme $\|\cdot\|$.
Un **espace de Banach** est un espace vectoriel normé complet.

Proposition 3.2.6. *Les applications* $E \times E \longrightarrow E$, $(x,y) \longmapsto x+y$; $\mathbb{R} \times E \longrightarrow E$, $(\lambda, x) \longmapsto \lambda x$; $E \longrightarrow \mathbb{R}_+$, $x \longmapsto \|x\|$ *sont continues*

Proposition 3.2.7. *Soient* $(E, \|.\|_E)$ *et* $(F, \|.\|_F)$ *des e.v.n.,* $\Phi : E \longrightarrow F$ *une application linéaire, alors*
Φ *est continue* $\Longleftrightarrow \exists c > 0, \forall x \in E : \|\Phi(x)\|_F \leq c\|x\|_E$

Proposition 3.2.8. *Soient* $(E, \|.\|_E)$ *et* $(F, \|.\|_F)$ *des e.v.n*[7]*. On note* $L(E,F)$ *l' e.v des applications linéaires continues de* E *vers* F*.*
Pour $f \in L(E,F)$ *on pose* $\|f\| = \sup_{\|x\|_E \leq 1} \|f(x)\|_F$*. Alors:*

(i) $f \longmapsto \|f\|$ *est une norme sur* $L(E,F)$ *;*

(ii) $\forall x \in E : \|f(x)\|_F \leq \|f\|\|x\|_E$ *;*

(iii) $\|f\| = \sup_{\|x\|_E \leq 1} \|f(x)\|_F = \sup_{\|x\|_E = 1} \|f(x)\|_F = \sup_{x \in E, x \neq 0_E} \frac{\|f(x)\|_F}{\|x\|_E}$

Application: Si $\dim E = n$ et $\dim F = m$ alors l'application linéaire $f : E \longrightarrow F$ est représentée par une matrice A de m lignes et n colonnes, on peut alors définir une norme de matricielle par :

$$\|A\| = \sup_{\|x\|_E = 1} \|Ax\|_F = \sup_{x \in E, x \neq 0_E} \frac{\|Ax\|_F}{\|x\|_E}. \tag{3.5}$$

Proposition 3.2.9. ***Espace dual***
Soit $(E, \|.\|_E)$ *un e.v.n. On note* $E' = L(E, \mathbb{R})$ *l'e.v. des formes linéaires continues sur* E*, alors : Une forme linéaire* $f \in E'$ *si et seulement si* $\ker f$ *est fermé dans* E*.* E' *est appelé le dual de* E*.*

Espace vectoriels normés de dimension finie

Proposition 3.2.10. *Soit* $(E, \|.\|)$ *un e.v.n. de dimension finie* n*, soit* $\{e_1, \ldots, e_n\}$ *une base de* E*. Alors l'application* $\Phi : (\mathbb{R}^n, \|.\|_\infty) \longrightarrow (E, .)$ *est une bijection continue.*

$$\alpha_1, \ldots, \alpha_n \longmapsto \sum_{i=1}^{n} \alpha_i e_i$$

De plus Φ^{-1} *est aussi continue,* Φ *est donc un isomorphisme topologique de* $\mathbb{R}^n$ *sur* E*. L'isomorphisme de* Φ *est à l'origine de nombreux corollaires suivants qui facilitent la manipulation des e.v. de dimension finie.*

Corollaire 3.2.11. *Sur un e.v.* E *de dimension finie toutes les normes sont équivalentes :*

$$\exists c_1, c_2 \in \mathbb{R}_+^2, \forall x \in E : c_1\|x\|' \leq \|x\| \leq c_2\|x\|' \tag{3.6}$$

Deux normes $\|.\|$ et $\|.\|'$ qui vérifient l'équation (3.6) sont dites équivalentes.

6. On a pas les inégalités inverses car pour $n \geq 2$, posons, pour $x \in [0, \frac{1}{n}]$, $f_n(x) = -2n^2x + 2n$ et pour $x \in [\frac{1}{n}, 1]$, $f_n(x) = 0$. Alors $\|f_n\|_1 = 1$ et $\|f_n\|_\infty = 2n$, l'on ne peut pas avoir $\|f_n\|_\infty \leq c\|f_n\|_1$
7. e.v.n = espace vectoriel normé.

Corollaire 3.2.12. *Soit $(E, \|.\|)$ un e.v. normé de dimension finie, alors E est* ***complet****.*

Corollaire 3.2.13. *Soient $(E, \|.\|_E)$ et $(F, \|.\|_F)$ des e.v.n., on suppose que E est de dimension finie. Alors toute application linéaire de E dans F est continue.*

Proposition 3.2.14. *Soit $(E, \|.\|)$ un e.v.n., F un sous-espace vectoriel fermé dans $(E, \|.\|)$ et W un sous-espace vectoriel de dimension finie. Alors $W + F$ est un sous-espace vectoriel fermé dans $(E, \|.\|)$.*

Corollaire 3.2.15. *Soit W un sous-espace vectoriel de dimension finie d'un e.v.n. $(E, \|.\|)$, alors W est fermé dans $(E, \|.\|)$.*

Espaces euclidiens. Espaces de Hilbert

[30]
Soit E un e.v. réel et $\phi : E \times E \longrightarrow R$ une **forme bilinéaire symétrique**, i.e.

1. $\forall (x, y) \in E^2 : \phi(x, y) = \phi(y, x)$;
2. $\forall (x_1, x_2, y) \in E^3, \forall (\lambda_1, \lambda_2) \in \mathbb{R}^2 : \phi(\lambda_1 x_1 + \lambda_2 x_2, y) = \lambda_1 \phi(x_1, y) + \lambda_2 \phi(x_2, y)$;
3. $\forall (x, y_1, y_2) \in E^3, \forall (\mu_1, \mu_2) \in \mathbb{R}^2 : \phi(x, \mu_1 y_1 + \mu_2 y_2) = \mu_1 \phi(x, y_1) + \mu_2 \phi(x, y_2)$

On dit que ϕ est **positive** ou **semi-définie positive** si : $\forall x \in E : \phi(x, x) \geq 0$.
On dit que ϕ est **non dégénérée** si : $\forall x \in E : \phi(x, x) = 0 \Longleftrightarrow x = O_E$.
Si ϕ est positive et non dégénérée, on dit que ϕ est **définie positive**.

Définition 3.2.16. *Une forme bilinéaire symétrique définie positive ϕ est appelée* ***produit scalaire*** *sur E. L'application $x \longmapsto \phi(x, x)$ définit alors une norme sur E. On note $\phi(x, y) = \langle x, y \rangle$ ou (x/y) et $\langle x, x \rangle = \|x\|^2$.*

Définition 3.2.17.
- *Un e.v. E, muni d'un produit scalaire $\langle ., . \rangle$ est un* ***espace préhilbertien.***
- *Un espace préhilbertien $(E, \langle ., . \rangle)$ est un* ***espace de Hilbert*** *s'il est complet pour la norme associée $\langle ., . \rangle$.*
- *Un e.v. E de dimension finie, muni d'un produit scalaire $\langle ., . \rangle$ est un* ***espace euclidien****.*

Exemple 3.2.18. *Soit $E = \mathbb{R}^n$, pour $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$ on définit :*

$$\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i \quad \langle x, x \rangle = \sum_{i=1}^{n} x_i^2 = \|x\|^2.$$

$(\mathbb{R}^n, \langle ., . \rangle)$ est un espace euclidien.

3.3 Théorème de la projection

Introduction

Le théorème de projection sur un convexe fermé est un outil fondamental de la théorie des espaces de Hilbert. Grâce à ce théorème nous déduirons la structure du dual d'un espace de Hilbert et nous serons en mesure de construire l'adjoint d'une application linéaire continue entre deux espaces de Hilbert.

Définition 3.3.1. ***Ensemble convexe***
Soit E un e.v. réel et $K \subset E$, K est dit convexe si et seulement si

$$\forall (x, y) \in K^2, \forall t \in [0, 1] : tx + (1 - t)y \in K. \tag{3.7}$$

Définition 3.3.2. ***Fonction convexe***
On dit qu'une fonction $J : K \subset E \longrightarrow \mathbb{R} \bigcup \{+\infty\}$ est convexe si K est convexe et si

$$\forall (x, y) \in K \times K, \forall t \in [0, 1], J(tx + (1 - t)y) \leq tJ(x) + (1 - t)J(y)$$

J est strictement convexe si :

$$\forall (x, y) \in K \times K (x \neq y), \forall t \in]0, 1[, J(tx + (1 - t)y) < tJ(x) + (1 - t)J(y)$$

Exemple 3.3.3. ***Fonctions convexes et strictement convexes***

1. *$J(x) = \|x\|^2$ est convexe.*
2. *Toute application affine ie de la forme*

$$J(x) = \langle b, x\rangle + \beta,$$

où b et x sont des éléments de E et $\beta \in \mathbb{R}$ est convexe mais pas strictement.

3. *Soit A une matrice carrée symétrique d'ordre n semi-définie positive et b un vecteur de $\mathbb{R}^n$. Alors J définie par*

$$J(x) = \frac{1}{2}\langle Ax, x\rangle_n - \langle b, x\rangle_n,$$

est convexe. Si de plus A est définie positive alors J est strictement convexe.
$\langle ., .\rangle_n$ désigne le produit scalaire de $\mathbb{R}^n$.

En générale si $\mathcal{A}$ est un opérateur linéaire de E (espace de Hilbert) dans E, auto-adjoint et monotone c'est à dire

$$\forall (x, y) \in E^2 \quad \langle \mathcal{A}(x) - \mathcal{A}(y), x - y\rangle \geq 0,$$

et $b \in E$, alors J définie par

$$J(x) = \frac{1}{2}\langle \mathcal{A}x, x\rangle - \langle b, x\rangle,$$

est convexe.

Définition 3.3.4. ***Domaine d'une fonction convexe***
Soit $J : E \longrightarrow \mathbb{R}\bigcup\{+\infty\}$ une fonction convexe. On appelle domaine de J l'ensemble

$$domJ = \{x \in E \ / \ J(x) < +\infty\}$$

domJ est convexe.
Lorsque le domaine de J est non vide J est dite ***propre.***

On rappelle que **l'épigraphe** de f est la partie de l'espace produit $E \times \mathbb{R}$ qui est au dessus de son graphe :

$$epif = \{(x, \alpha) \in E \times \mathbb{R} : f(x) \leq \alpha\}$$

Quant à **l'épigraphe stricte**, il est obtenu en prenant l'inégalité au sens strict et on note $epi_s f$

Définition 3.3.5. *On dit qu'une fonction $f : E \longrightarrow \mathbb{R}$ est convexe si son épigraphe (ou son épigraphe stricte) est convexe dans $E \times \mathbb{R}$. On dit que $f : E \longrightarrow \mathbb{R}$ est concave si $-f$ est convexe.*

Soit E un espace de Hilbert réel ou complexe, ie un espace vectoriel normé complet où la norme découle d'un produit scalaire $\langle ., .\rangle$. Rappelons que dans le cas complexe $\langle ., .\rangle$ vérifie :

1. $\langle ., .\rangle$ est sesquilinéaire, ie linéaire à gauche et semi-linéaire à droite.
2. $\langle ., .\rangle$ est hermitienne ie pour tout $x, y \in H$, $\langle x, y\rangle = \overline{\langle y, x\rangle}$.
3. $\langle ., .\rangle$ est définie positive.

Theorem 3.3.6. ***Projection sur un convexe fermé.***
Soient E un espace de Hilbert et K une partie convexe fermée non vide de E.
Pour tout $x \in E$, il existe un et un seul point $y \in K$ tel que $d(x, K) = \inf_{z \in K} \|x - z\| = \|x - y\|$ et on l'appelle projeté orthogonal de x sur K et on note $y = p_K(x)$. Il est caractérisé par :
$y = p_K(x) \Longleftrightarrow \forall z \in K, Re(\langle y - x, y - z\rangle) \leq 0$

Démonstration. La preuve du théorème va se faire dans les trois objectifs suivants:

Objectif 1 Existence du projeté orthogonal.

Posons $d = d(x, K) = \underset{z\in K}{inf}\|x - z\|$. Puisque $d^2 = \underset{z\in K}{inf}\|x - z\|^2$, par définition de la borne inférieure, on a :

$$\forall n \in \mathbb{N}^*, \exists y_n \in K \ / \ \|x - y_n\|^2 \leq d^2 + \frac{1}{n}.$$

Montrons que $(y_n)_{\in\mathbb{N}}$ est de Cauchy. L'identité du parallélogramme s'écrit :

$$\|z - z'\| = 2\|z\|^2 + 2\|z'\|^2 - \|z + z'\|$$

En prenant $z = y_n - x$ et $z' = y_p - x$ dans cette identité on a :

$$\|y_n - y_p\|^2 = 2\|y_n - x\|^2 + 2\|y_p - x\|^2 - \|y_n + y_p - 2x\|^2 \tag{3.8}$$

$$= 2\|y_n - x\|^2 + 2\|y_p - x\|^2 - 4\|\frac{y_n + y_p}{2} - x\|^2 \tag{3.9}$$

$$\leq (2d^2 + \frac{2}{n}) + (2d^2 + \frac{2}{p}) - 4d^2 \tag{3.10}$$

$$\leq \frac{2}{n} + \frac{2}{p} \tag{3.11}$$

Puisque K est convexe alors $\frac{y_n+y_p}{2} \in K$ et $\|\frac{y_n+y_p}{2} - x\|^2 \geq d$.
Et aussi comme $\lim\limits_{n\longrightarrow+\infty} \frac{1}{n} = 0$ et $\lim\limits_{p\longrightarrow+\infty} \frac{1}{p} = 0$ on a :

$$\forall \varepsilon > 0, \exists N \in \mathbb{N}, \forall n, p \geq N, \frac{1}{p} \leq \varepsilon$$

et on conclut que la suite $(y_n)_{n\in\mathbb{N}}$ est de Cauchy. Puisque E est complet, $(y_n)_{n\in\mathbb{N}}$ converge et on note sa limite $y \in K$ car est fermé. Par continué de la norme (proposition 3.2.6) on a alors: $\lim\limits_{n\longrightarrow\infty} \|y_n - x\|^2 = \|y - x\|^2$
puis:

$$\|y - x\|^2 \leq \lim_{n\longrightarrow\infty} d^2 + \frac{1}{n} = d^2$$

Finalement $\|y - x\| \leq d$ et la définition de d, $\|y - x\| \geq d$ et donc on a $\|y - x\| = d = d(x, K)$

Objectif 2 Unicité du projeté orthogonal.

Prouvons cette unicité par l'absurde. Supposons qu'il existe $y \neq y$ tels que $d = \|x - y\| = \|x - y'\|$. Comme K est convexe $\frac{y+y'}{2} \in K$ et par l'identité du parallélogramme on a :

$$\|\frac{y + y'}{2} - x\|^2 = \|\frac{1}{2}(y - x) + \frac{1}{2}(y' - x)\|^2 \tag{3.12}$$

$$= 2 \times \frac{1}{4}\|y - x\|^2 + 2 \times \frac{1}{4}\|y' - x\|^2 - \frac{1}{4}\|y - y'\|^2 \tag{3.13}$$

$$= \frac{1}{2}(\|y - x\|^2 + \|y' - x\|^2) - \frac{1}{4}\|y - y'\|^2 \tag{3.14}$$

$$= d^2 - \frac{1}{4}\|y - y'\|^2 \tag{3.15}$$

$$< d^2 \tag{3.16}$$

Absurde, par définition de d.

Objectif 3 Condition nécessaire et suffisante pour être le projeté orthogonal de x sur K.

Dans un premier temps supposons que y est le projeté orthogonal de x sur K et soit $z \in K$. Par convexité de K, $(1 - \lambda)y + \lambda z \in K$ pour $\lambda \in [0, 1]$ (le segment $[y, z]$ est contenu dans K. Par définition du projeté orthogonal on a :

$$\|x - ((1 - \lambda)y + \lambda z))\|^2 = \|x - y + \lambda(y - z)\|^2 \geq d^2 = \|x - y\|^2$$

En développant, il vient : $\|x-y\|^2 + 2\lambda Re(\langle x-y, y-z\rangle) + \lambda^2\|y-z\|^2 \geq \|x-y\|^2$ soit :

$$2\lambda Re(\langle x-y, y-z\rangle) + \lambda^2\|y-z\|^2 \geq 0$$

et pour $\lambda \neq 0, 2Re(\langle x-y, y-z\rangle) + \lambda\|y-z\|^2 \geq 0$
En faisant tendre λ vers 0 dans l'expression ci-après, on a bien:
$Re(\langle x-y, y-z\rangle) \geq 0 \Longleftrightarrow Re(\langle y-x, y-z\rangle) \leq 0$.
Inversement, supposons la condition ci-dessus satisfaite et montrons que y est alors le projeté orthogonal de x sur K. Il suffit de montrer que la condition ci-dessus implique que pour tout $z \in K$, on a : $\|z-y\| \geq \|y-x\|$. Or,

$$\|z-y\| = \|(z-y)-(x-y)\|^2 \tag{3.17}$$
$$= \|y-x\| + \|z-y\|^2 - 2Re(\langle y-x, y-z\rangle) \geq \|y-x\|^2 \tag{3.18}$$

car $-2Re(\langle y-x, y-z\rangle) \geq 0$ et $\|y-x\|+\|z-y\|^2-2Re(\langle y-x, y-z\rangle) \geq 0$ d'où $\|z-y\| \geq \|z-y\|$.
C.Q.F.D.

□

Corollaire 3.3.7. *Soit K un sous espace vectoriel fermé de E (en particulier K est naturellement convexe). Alors, $E = K \oplus K^{\perp}$ et p_K est une application linéaire continue.*

En application, voici l'important théorème de caractérisation du dual d'un espace hilbertien :

Theorem 3.3.8. ***(Théorème de représentation de Riesz-Fréchet)****[24, 30]*
Soit f une forme linéaire continue sur E. Alors, il existe un unique vecteur $a \in E$ tel que $f(x) = \langle x, a\rangle$ pour tout $x \in E$.

Démonstration. Supprimons d'abord le cas où $f = 0$, dans ce cas $a = 0$ convient. Sinon $f \neq 0$ implique que $\ker(f) \neq E$ et que $\ker(f)$ est un sous espace vectoriel fermé car f est continue. D'après le théorème précédent, on a alors la décomposition $E = \ker(f) \oplus \ker(f)^{\perp}$. Comme $\ker(f) \neq E$, on a $\ker(f)^{\perp} \neq \{0\}$. Soit h un élément h non nul dans $\ker(f)^{\perp}$. Pour $x \in E$, on a naturellement $x - \frac{f(x)}{f(h)}h \in \ker(f) \Longrightarrow \langle x - \frac{f(x)}{f(h)}h, h\rangle = 0$. En développant l'expression ci-dessus, on a donc :

$$\langle x, h\rangle = \langle \frac{f(x)}{f(h)}h, h\rangle = \frac{f(x)}{f(h)}\|h\|^2 \Longrightarrow f(x) = \langle x, \frac{\overline{f(h)}}{\|h\|^2}h\rangle$$

Le vecteur $a = \frac{\overline{f(h)}}{\|h\|^2}h$ répond à notre problème d'existence. Pour finir montrons l'unicité de a. Supposons qu'il existe un autre $a' \in E$ vérifiant $f(x) = \langle x, a'\rangle$. Alors on a :

$$\forall x \in E, \langle x, a'\rangle = \langle x, a\rangle \Longrightarrow \forall x \in E, \langle x, a'-a\rangle = 0$$

Ainsi $a - a' \in E^{\perp} = \{0\}$ et on a ainsi l'unicité de a. C.Q.F.D. □

Remarque 3.3.9. *Si E est hilbertien complexe, l'application $x \longmapsto \langle a, x\rangle$ n'est pas linéaire, mais semi-linéaire d'où il est important de placer l'élément a à droite dans l'écriture $\langle x, a\rangle$.*

3.4 Minimisation sans contrainte

Dans cette partie de la thèse nous allons étudier les problèmes d'optimisation évoquées dans la section 3.1 dans le cas où $E = \mathbb{R}^n$ muni du produit scalaire usuel et lorsqu'il n'y a pas de contraintes : on effectue la minimisation de la fonction J sur tout l'espace. Cela nous prépare le terrain d'optimisation pour notre article publié dans IJAMAS (International Journal of Applied Mathematics and Statistics) intitulé "Modeling and Prediction of Dakar Air Quality Index" (4.5).

Dans cet article nous avons la régression linéaire multiple avec la minimisation sans contrainte d'où l'intérêt de cette partie dans la thèse.

Le problème (P) se formule de la manière suivante

$$(P)\begin{cases} min J(x) \\ x\in\mathbb{R}^n \end{cases}$$

où J est une fonction de $\mathbb{R}^n$ dans $\mathbb{R}\bigcup\{+\infty\}$

3.4.1 Résultat d'existence et d'unicité

Avant d'étudier les propriétés de la solution (ou des solutions) de (P) il nous faut nous s'assurer de leur existence. Ensuite nous donnerons ensuite des résultats d'unicité.

Définition 3.4.1. *On dit que $J : E \longrightarrow \mathbb{R}$ est **coercive** si*

$$\lim_{\|x\|\longrightarrow+\infty} J(x) = +\infty$$

Exemple 3.4.2. ***Fonctions coercives***

(i) *$J(x) = \|x\|$ est coercive.*

(ii) *La fonction affine J définie par $J(x) = \langle\alpha, x\rangle + \beta$, $\alpha \in \mathbb{R}^n, \beta \in \mathbb{R}$ n'est pas coercive.*

(iii) *Soit A une matrice carrée d'ordre n symétrique, définie positive et b un vecteur de $\mathbb{R}^n$. Alors J définie par*

$$J(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle$$

est coercive.

Theorem 3.4.3. ***Existence***

Soit $J : \mathbb{R}^n \longrightarrow \mathbb{R}\bigcup\{+\infty\}$ propre, continue et coercive. Alors (P) admet au moins une solution.

Démonstration. Soit $d = \inf(P)$; $d < +\infty$ car J est propre. Soit $(x_p) \in \mathbb{R}^n$ une suite minimisante ie

$$\lim_{p\longrightarrow+\infty} J(x_p) = d$$

Montrons que (x_p) est bornée.

Supposons le contraire alors on pourrait extraire de cette suite une sous-suite (encore notée (x_p)) telle $\lim_{p\longrightarrow 0}\|x_p\| = +\infty$. Puisque J est coercive on aurait $\lim_{\|x\|\longrightarrow+\infty} J(x_p) = +\infty$, absurde car $\lim_{p\longrightarrow+\infty} J(x_p) = d < +\infty$ et donc (x_p) est bornée.

Puisque (x_p) est bornée, d'après Bolzano Weierstrass, on peut alors en extraire une sous-suite (encore notée (x_p)) qui converge $\bar{x} \in \mathbb{R}^n$. Par continuité de J, on a alors vers $d = \lim_{p\longrightarrow+\infty} J(x_p) = J(\bar{x})$.

En particulier $d > +\infty$ et $\bar{x}$ est une solution du problème (P). □

Remarque 3.4.4. *Ce résultat est encore vrai si on remplace $\mathbb{R}^n$ par un espace de Hilbert E. La continuité de la fonctionnelle J est alors remplacée par de la semi-continuité inférieure pour la topologie faible. On se ne limite au cas où l'espace de référence est de dimension finie.*

Il faut noter qu'on n'a pas forcément l'unicité. Ci-dessous on a un critère pour l'unicité.

Theorem 3.4.5. ***Unicité de la solution***

Si $J : \mathbb{R}^n \longrightarrow \mathbb{R}\bigcup\{+\infty\}$ est strictement convexe alors le problème (P) admet au plus une solution.

Démonstration. Supposons que J admette au moins un minimum m et soient $x_1 \neq x_2$ (dans $\mathbb{R}^n$) réalisant ce minimum : $J(x_1) = J(x_2) = m$. Par stricte convexité de la fonction J on a alors :

$$J(\frac{x_1 + x_2}{2}) < \frac{1}{2}(J(x_1) + J(x_2) = m$$

Absurde ar m désigne le minimum de J. Et par conséquent, $x_1 = x_2$

□

Nous admettons le théorème suivant qui donne un critère pour qu'une fonction soit strictement convexe et coercive :

Theorem 3.4.6. *Soit J une fonction de classe $\mathcal{C}^1$ de $\mathbb{R}^n$ dans $\mathbb{R}$. On suppose qu'il existe $\alpha > 0$ tel que*

$$\forall (x, y) \in \mathbb{R}^n \times \mathbb{R}^n, \langle \nabla J(x) - \nabla J(y), x - y \rangle \geq \alpha \|x - y\| \tag{3.19}$$

Alors J est strictement convexe et coercive ; en particulier le problème (P) admet une solution unique.

Définition 3.4.7. ***(Fonction elliptique)***
On dit que $J : E \longrightarrow \mathbb{R}$ est elliptique si la condition (3.20) suivante est vérifiée

$$\exists > 0 \quad / \quad \forall (x, y) \in E \times E, \langle \nabla J(x) - \nabla J(y), x - y \rangle \geq \alpha \|x - y\| \tag{3.20}$$

α est la constante d'ellipticité

Proposition 3.4.8. ***critère d'ellipticité***
Une fonction $J : E \longrightarrow R$ deux fois différentiable sur E est elliptique si et seulement si

$$\forall (x, y) \in E \times E, \quad \langle D^2 J(x)y, y \rangle \geq \alpha \|y\|^2$$

.

Démonstration. On applique la formule de Taylor à la fonction $\varphi : t \longrightarrow \varphi(t) = J(x + ty)$. □

Maintenant il nous faut des conditions pour pouvoir calculer la (ou les) solutions.

3.4.2 Conditions d'optimalité

Conditions nécessaires du premier ordre

Définition 3.4.9. *Soient E un e.v., F un e.v.n. et $J : E \longrightarrow F$ une fonction. On dit que J est directionnellement dérivable au sens de Dini en $x \in E$ dans la direction $h \in E$ si la limite $\lim\limits_{t \to 0^+} \frac{J(x+th)-J(x)}{t}$ existe dans F. On la note $J'_D(x; h)$*

Définition 3.4.10. ***G-différentiabilité***
Soient E et F deux e.v.n. et $J : E \longrightarrow F$ une fonction. On dit que J est ***Gâteaux différentiable*** *en $x \in E$ si*

1. *la dérivée directionnelle $J'_D(x; h)$ existe quel que soit $h \in E$,*
2. *l'application $D_G J(x) : h \in E \mapsto J'_D(x; h) \in F$ est linéaire continue.*

Les conditions que nous allons donner sont des conditions différentielles qui portent sur la dérivée de la fonction à minimiser. On va donc se restreindre au cas des fonctions Gâteaux-différentiables.

Theorem 3.4.11. ***(Condition nécessaire d'optimalité du premier ordre)***
Soit E un espace de Hilbert réel et $J : E \longrightarrow \mathbb{R}$ une fonctionnelle Gâteaux-différentiable sur E. Si x^* réalise un minimum (global ou local) de J sur E alors

$$\nabla J(x^*) = 0 \tag{3.21}$$

Démonstration. Si x^* réalise un minimum de J sur E alors

$$\forall x \in \mathcal{B}(x^*, \rho) J(x^*) \leq J(x),$$

où $\mathcal{B}(x^*, \rho)$ est une boule de rayon $\rho > 0$ centrée en x^*.

Soit $h \in E, h \neq 0$; choisir $t_h = \frac{\rho}{\|h\|} > 0$ tel que

$$\forall t \in]0, th[\quad x^* + th \in \mathcal{B}(x^*, \rho),$$

et donc

$$\forall t \in]0, th[\quad J(x^*) \leq J(x^* + th)$$

Or J est Gâteaux-différentiable en x^* , donc

$$\lim_{t \longrightarrow 0^+} \frac{J(x^* + t_h) - J(x^*)}{t} = \langle \nabla J(x^*), h \rangle$$

$$Donc \ \ \forall h \in E, \langle \nabla J(x^*), h \rangle \geq 0 \tag{3.22}$$

ie $\nabla J(x^*) = 0$ □

Définition 3.4.12.

Un point x^* de E vérifiant $\nabla J(x^*) = 0$ est appelé **point critique** ou **point stationnaire** .

La relation $\nabla J(x^*) = 0$ est appelée **équation d'Euler**.

Le théorème sur la condition nécessaire d'optimalité du premier ordre n'a pas de sens si la fonction J n'est pas différentiable.

Theorem 3.4.13. ***(CNS du premier ordre dans le cas convexe)*** *[27, 28]*

Soit $J : E \longrightarrow \mathbb{R}$ Gâteaux différentiable et convexe sur E. Un point x^* réalise un minimum global de J sur E si et seulement si $\nabla J(x^*) = 0$.

Démonstration. On a vu que la condition est toujours nécessaire. Montrons qu'elle est suffisante. Soit $x^* \in E$ tel que $\nabla J(x^*) = 0$. Comme J est convexe on peut appliquer la continuité des fonctions convexes et on obtient après calcul :

$$\forall x \in H, \ J(x) \geq J(x^*) + \langle \nabla J(x^*), x - x^* \rangle = J(x^*).$$

Et on a donc immédiatement le fait que x^* réalise un minimum de J sur E. □

Le cas où J est convexe est fréquent dans la pratique mais pas systématique. Nous allons donc donner maintenant des conditions suffisantes pour qu'un point critique réalise un minimum (ou un maximum). Ces conditions vont faire intervenir la dérivée seconde de J : ce sont des conditions du second ordre.

Conditions du deuxième ordre

Nous commençons par une condition nécessaire permettant de préciser encore les éventuels minima.

Theorem 3.4.14. ***(Condition nécessaire du second ordre)*** *[26]*

On suppose que x^* est un minimum (local) de J et que J est deux fois dérivable sur E. Alors on a :

1. $\nabla J(x^*) = 0$ *et*
2. $\forall x \in E, \quad \langle D^2 J(x^*)x, x \rangle \geq 0$

Démonstration. On a déjà le point 1. Montrons le point 2.

Soit $x \in E$. Appliquons la formule de Taylor à la fonction $\varphi : t \longrightarrow \varphi(t) = J(x^* + tx)$. Comme $\nabla J(x^*) = 0$ on obtient

$$0 \leq J(x^* + tx) - J(x^*) = \frac{t^2}{2}\langle D^2 J(x^*)x, x\rangle + o(t^2)$$

Après division par t^2 , on fait tendre t vers 0 et on a le résultat voulu. □

Remarque 3.4.15. *Dans le cas $E = \mathbb{R}^n$, 2. est équivalent à dire que la matrice Hessienne de J en x^*: $D^2J(x^*)$ est semi-définie positive.*

Rappelons qu'un critère pour que $D^2J(x^*)$ (qui est une matrice symétrique) soit semi-définie positive est que toutes ses valeurs propres soient positives ou nulles.

La réciproque du théorème précédent est fausse (il suffit de penser à la fonction $t \longrightarrow t^3$ pour s'en convaincre). Nous pouvons toutefois donner une réciproque sous forme de condition suffisante du second ordre plus forte (pour un résultat plus faible).

Theorem 3.4.16. ***(Condition suffisante du second ordre)*** *[26, 28]*

Soit J deux fois dérivable sur E vérifiant $\nabla J(x^*) = 0$ et

$$\exists \alpha > 0, \forall x \in E \langle D^2 J(x^*)x, x\rangle \geq \alpha \|x\|^2 \tag{3.23}$$

Alors la fonction J admet un minimum local strict en x^* .

Démonstration. Soit x dans E. On utilise la formule de Taylor appliquée a la fonction $\varphi : t \longmapsto \varphi(t) = J(x^* + tx)$. Nous avons

$$J(x * +tx) - J(x) = \frac{t^2}{2}\langle D^2(x^*)x, x\rangle + o(t^2) \geq \frac{t^2}{2}\alpha\|x\|^2 + o(t^2)$$

Ceci montre que x^* réalise un minimum local strict de J

□

La condition (3.23) est une condition d'ellipticité locale.

Remarque 3.4.17. *Si $E = \mathbb{R}^n$ la condition (3.23) revient à dire que la matrice Hessienne $D^2J(x^*)$ est définie positive, un choix possible pour α étant alors la plus petite valeur propre. C'est une condition de convexité (locale) stricte au voisinage de x^*.*

3.4.3 Application à la régression linéaire

Nous allons illustrer les résultats de la section précédente par l'exemple très important de la régression linéaire.

Considérons un nuage de n points de $\mathbb{R}^2$, $M_i = (t_i, x_i)$, $1 \leq i \leq n$. Ces données sont souvent le résultat de mesures et on cherche à décrire le comportement global de ce nuage. En général ces points ne sont pas alignés, mais on décide de chercher une droite les "approchant" au mieux ...

On utilise pour cela la méthode des moindres carrés : comme on n'a pas $x_i = at_i + b$ pour tout i, on cherche à minimiser le carré des différences. On veut donc trouver un couple de réels (a, b) solution de

$$\min_{(a,b)\in\mathbb{R}^2} \sum_{i=1}^{n} (x_i - at_i - b)^2$$

on prend :

$$J(a, b) = \sum_{i=1}^{n} (x_i - at_i - b)^2.$$

Calculons le gradient de J en un point quelconque (a, b) de $\mathbb{R}^2$.

$$\frac{\partial J(a,b)}{\partial x} = \sum_{i=1}^{n} 2(at_i + b - x_i)ti \qquad = 2a\sum_{i=1}^{n} t_i^2 + 2b\sum_{i=1}^{n} t_i - 2\sum_{i=1}^{n} t_i x_i$$

$$\frac{\partial J(a,b)}{\partial x} = \sum_{i=1}^{n} 2(at_i + b - x_i) \qquad = 2a\sum_{i=1}^{n} t_i - 2nb - 2\sum_{i=1}^{n} x_i$$

Posons $S_t = \sum_{i=1}^{n} t_i$, $S_x = \sum_{i=1}^{n} x_i$, $S_{xt} = \sum_{i=1}^{n} x_i t_i$ *et* $S_{t^2} = \sum_{i=1}^{n} t_i^2$

$$\nabla J(a,b) = 0 \Longleftrightarrow \begin{cases} S_{t^2}a + S_t b & = S_{xt} \\ S_t a + nb & = S_x \end{cases}$$

La résolution de ce système donne une solution unique si $(S_t)^2 - nS_t^2 \neq 0$. On obtient

$$a = \frac{S_x S_t - nS_{xt}}{(S_t)^2 - nS_t^2} \quad et \quad b = \frac{S_{xt}S_t - S_x S_{t^2}}{(S_t)^2 - nS_t^2}$$

Vérifions que le couple obtenu est bien un minimum. Calculons pour cela la matrice Hessienne de J en (a, b) :

$$D^2 J(a,b) = 2\begin{pmatrix} S_t^2 & S_t \\ S_t & n \end{pmatrix}$$

cette matrice est toujours (semi-définie) positive. Elle est définie positive si elle est inversible (car alors aucune valeur propre n'est nulle) c'est-à -dire lorsque que son déterminant $(S_t)^2 - nS_{t^2}$ est non nul. Dans ce cas, le couple obtenu est bien (l'unique) minimum strict de J. La droite ainsi obtenue est appelée **droite de régression**.

3.4.4 Algorithmes

Il existe de nombreux algorithmes classiques permettant de calculer (d'une manière approchée) la ou les solutions du problème (P) de départ. On peut citer entre autres l'**Algorithme du Gradient** (Méthode du Gradient), l'**Algorithme de Newston** (Méthode de Newston), l'**Algorithme du Gradient conjugué** (Méthode du Gradient conjugué), **Algorithme de Relaxation** (Méthode de Relaxation),... Toutefois, il faut noter que la plupart de ces algorithmes exploitent les conditions d'optimalité dont on a vu qu'elles permettaient (au mieux) de déterminer des minima locaux. La question de la détermination de minima globaux est difficile. Néanmoins, nous verrons en exemple dans le paragraphe (3.4.5) un **algorithme probabiliste** permettant de "déterminer" un minimum global.

Remarquons aussi que nous avons fait l'hypothèse de différentiabilité de la fonction J. Il existe des méthodes permettant de traiter le cas non différentiable (ou non régulier). Nous n'en parlerons pas ici. Le lecteur intéressé peut se référer à [22]. Nous commencerons par quelques définitions :

Définition 3.4.18. ***(Algorithme)*** *Un algorithme est défini par une application $\mathcal{A}$ de $\mathbb{R}^n$ dans $\mathbb{R}^n$ permettant la génération d'une suite d'élément de $\mathbb{R}^n$ par la formule:*

$$\begin{cases} x_0 \in \mathbb{R}^n \quad \textbf{\textit{donné}}, k = 0 \quad \textbf{\textit{Etape d'initialisation}} \\ x_{k+1} = \mathcal{A}(x_k), k = k+1 \quad \textbf{\textit{Itération k.}} \end{cases}$$

Écrire un algorithme revient que se donner une suite $(x_k)_{k\in\mathbb{N}}$ de $\mathbb{R}^n$ et étudier la convergence de l'algorithme c'est étudier la convergence de la suite $(x_k)_{k\in\mathbb{N}}$.

Définition 3.4.19. ***(Convergence d'un algorithme)*** *On dit que l'algorithme $\mathcal{A}$ converge si la suite $(x_k)_{k\in\mathcal{N}}$ engendrée par l'algorithme converge vers une limite x^* .*

Certes s'assurer de la convergence d'un algorithme par les hypothèses fixées est très capital mais la vitesse de convergence et la complexité sont aussi des facteurs à non négligeable lors de l'utilisation (ou de la génération) d'un algorithme ; on a en effet "intérêt" à ce que la méthode soit la plus rapide possible tout en restant précise et stable. Un critère de mesure de la vitesse (ou taux) de convergence est l' évolution de l'erreur commise ($e_k = \|x_k - x^*\|$).

Définition 3.4.20. ***Taux de convergence d' un algorithme***
Soit $(x_k)k \in \mathbb{N}$ une suite de limite x^* définie par la donnée d'un algorithme convergent $\mathcal{A}$. On dit que la convergence de $\mathcal{A}$ est

1. ***linéaire*** *si l'erreur $e_k = \|x_k - x^*\|$ décroit linéairement :*

$$\exists C \in [0,1[, \exists k_0, \forall k \geq k_0 \ e_{k+1} \leq Ce_k$$

2. ***super-linéaire*** *si l'erreur décroit de la manière suivante :*

$$e_{k+1} \leq \alpha_k e_k$$

où α_k est une suite positive convergente vers 0. Si α_k est une suite géométrique alors la convergence de l'algorithme est dite ***géométrique.***

3. ***d'ordre*** *p si l'erreur e_k décroit de la manière suivante :*

$$\exists C \geq 0, \exists k_0, \forall k \geq k_0 \quad e_{k+1} \leq C[e_k]^p$$

Si $p = 2$, la convergence de l'algorithme est dite ***quadratique****. Enfin, la convergence est dite* ***locale*** *si elle n'a lieu que pour des points de départ x_0 dans un voisinage de x^* . Dans le cas contraire la convergence est* ***globale****.*

Remarque 3.4.21. *La "classification" précédente des vitesses de convergence renvoie à la notion de comparaison des fonctions au voisinage de $+\infty$. En effet, si on suppose que l'erreur e_k ne s'annule pas, une convergence linéaire revient à dire que $\frac{e_{k+1}}{e_k} = O(1)$, alors qu'une convergence super-linéaire est équivalente à $\frac{e_{k+1}}{e_k} = o(1)$. De la même manière, un algorithme d'ordre $p \geq 2$ est tel que $\frac{e_{k+1}}{e_k} = o(e_k^{p-2})$. On a bien entendu intérêt à ce que la vitesse de convergence d'un algorithme soit la plus élevée possible (afin d'obtenir la solution avec un minimum d'itérations pour une précision donnée)*

3.4.5 Méthode probabiliste

On présente ici un algorithme stochastique ie un algorithme qui fait intervenir des variables aléatoires. La plupart des algorithmes cités dans le paragraphe (3.4.4) fournissent des minima locaux, sauf dans des cas très particuliers (cas convexe par exemple). L'algorithme du **recuit simulé** [22] permet d'obtenir les minimums globaux d'une fonction. On ne parlera que des algorithmes sur des ensembles finis. On suppose en effet qu'on a discrétisé l'ensemble des contraintes.

Dynamique de Métropolis

Soit E un espace fini. On considère une fonction V [8] de E dans $\mathbb{R}$ appelée fonction d'énergie ou potentiel que nous souhaitons minimiser. L'algorithme de Métropolis est un algorithme de recherche des minima de V. L'idée heuristique de cette méthode est la suivante : si à l' étape n l'itéré vaut $X_n = x$, on regarde la valeur de V pour un point y voisin de x choisi aléatoirement. Si $V(y) < V(x)$ on alors x n'est pas bon et on prend $X_{n+1} = y$. Dans le cas contraire, on prend $X_{n+1} = x$. Mais on veut éviter de rester piégé en un éventuel minimum local $X_n = x$. Donc on posera $Xn + 1 = y$ si $V(y) - V(x)$ est inférieur à une variable aléatoire positive simulée, et $X_{n+1} = x$ dans le cas contraire. L'algorithme se présente comme suit :

Dynamique de Metropolis

1. **Initialisation** $n = 0$: choix deX_0 dans E déterministe arbitraire.
2. **Itération** n : on observe $X_n = x$.

8. Nous changeons de notations et de terminologie : la fonction V n'est autre que la fonction coût J des chapitres précédents restreinte au sous-ensemble fini E.

On simule une variable aléatoire Y_{n+1} de loi $Q(x,.)$; puis on génère un nombre au hasard U_{n+1} (de loi uniforme sur $[0,1]$) indépendant de Y_{n+1} et on pose

$$X_{n+1} = \begin{cases} Y_{n+1} & si \quad V(Y_{n+1}) \leq -\tau \log U_{n+1} + V(x) \\ x & sinon \end{cases}$$

τ est un réel positif : c'est la température. Q est une matrice (de transition markovienne (1.7.6)) sur E, symétrique : c'est la règle de sélection des voisins. Cette matrice exprime en général une relation de voisinage : si chaque point x a $r+1$ voisins, la relation de voisinage étant symétrique, on peut prendre $Q(x,y) = \frac{1}{r}$ si x et y sont distincts et voisins. Pour plus de détails (en particulier sur les démonstrations de convergence) on peut se référer à [23].

3.5 Méthode des moindres carrés

3.5.1 Introduction

L'étude d'un phénomène peut, le plus souvent, être schématisé de la manière suivante :

on s'intéresse à une grandeur y, que nous appellerons par la suite réponse ou variable expliquée, qui dépend d'un certain nombre de variables $x_1, x_2, \ldots, x_n$ que nous appellerons facteurs ou variables explicatives. Ici dans cette étude, y représente ici l'indice de la qualité de l'air (nombre sans unité),

x_1 désigne la température maximale (en degré Celsius),

x_2 point de rosée (en degré Celsius), x_3 humidité maximale (%),

x_4 pression maximale au niveau de la mer (hPa),

x_5 visibilité maximale (km),

x_6 vitesse maximale du vent (Km/h),

x_7 niveau de précipitation (mm).

Ces variables sont prises dans la ville de Dakar.

3.5.2 Notion de modèle et de régression linéaire multiple

On cherche ici à mettre en évidence la liaison, relation fonctionnelle pouvant exister entre la variable expliquée y et les variables explicatives $x_1, x_2, \ldots, x_n$. On s'intéresse aux modèles dits linéaires, i.e. aux modèles du type :

$$y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_1 x_n = \sum_{j=1}^{n} \beta_j x_j \tag{3.24}$$

où les β_j sont des réels appelés coefficients du modèle.

L'idéal de l'estimation est d'avoir $\hat{y} = y$ mais le plus souvent $\hat{y} \approx y$ avec $\hat{y} \neq y$.

3.5.3 Critère des moindres carrés - formulation

Critère

On cherche donc un modèle qui nous permet d'obtenir un $\hat{y}$ le plus $\ll$ *proche* $\gg$ possible de y. Pour cela, on effectue m mesures ($m > n$) des variables $x_1, x_2, \ldots, x_n$ et de y. On cherche alors $\beta_1, \beta_2, \ldots, \beta_n$ tels que, pour $i = 1, \ldots m$:

$$\hat{y_i} = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_{i,1} x_{i,n} = \sum_{j=1}^{n} \beta_j x_{i,j}$$

soit le plus $\ll$ *proche* $\gg$ possible de y_i .

En utilisant les notations matricielles, le système :

$$\begin{cases} \hat{y_1} & = \beta_1 x_{1,1} + \beta_2 x_{1,2} + & \dots & +\beta_n x_{1,n} \\ \hat{y_2} & = \beta_1 x_{2,1} + \beta_2 x_{2,2} + & \dots & +\beta_n x_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ \hat{y_m} & = \beta_1 x_{m,1} + \beta_2 x_{m,2} + & \dots & +\beta_n x_{m,n} \end{cases}$$

s'écrit:

$$\underbrace{\begin{pmatrix} \hat{y_1} \\ \hat{y_2} \\ \vdots \\ \hat{y_m} \end{pmatrix}}_{\hat{y}} = \underbrace{\begin{pmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,n} \\ x_{2,1} & x_{2,2} & \dots & x_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m,1} & x_{m,2} & \dots & x_{m,n} \end{pmatrix}}_{X} \times \underbrace{\begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}}_{\beta}$$

On cherche ainsi $\beta = (\beta_1, \beta_2, \dots, \beta_n)^T$ tel que $X\beta$ soit le plus $\ll$ *proche* $\gg$ possible de y. On comprend alors que la notion de distance apparaît. On rappelle que la distance euclidienne usuelle est définie comme suit :

$$\forall x, y \in \mathbb{R}^m, d(x, y) = \sqrt{\|x - y\|^2}$$

où $\|.\|$ définit la norme euclidienne :

$$\forall x \in \mathbb{R}^m, \|x\|^2 = \langle x, x \rangle = x^T x = \sum_{i=1}^{m} x_i^2.$$

On souhaite que $d(\hat{y} = X\beta, y)$ soit minimale et cela se traduit par :

$$\min_{\beta \in \mathbb{R}^n} \|X\beta - y\|^2 \tag{3.25}$$

Forme standard

Définition 3.5.1. *On appelle forme standard d'un problème de moindres carrés la donnée de*

- *la matrice* $X = \begin{pmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,n} \\ x_{2,1} & x_{2,2} & \dots & x_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m,1} & x_{m,2} & \dots & x_{m,n} \\ \uparrow & \uparrow & & \uparrow \\ x_1 & x_2 & \dots & x_n \end{pmatrix} \begin{matrix} \longleftarrow \textit{mesure 1} \\ \longleftarrow \textit{mesure 2} \\ \\ \longleftarrow \textit{mesure 3} \end{matrix} \in \mathcal{M}_{m,n}$ *appelée matrice des données*
- *le vecteur réponse* $y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} \in \mathbb{R}^m$
- *l'expression du critère : on cherche* $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix} \in \mathbb{R}^n$ *réalisant :*

$$\min_{\beta \in \mathbb{R}^n} \|X\beta - y\|^2$$

Remarque 3.5.2. *Le critère peut encore se réécrire de la manière suivante :*

$$\beta = arg\min_{\alpha \in \mathbb{R}^n} \|X\alpha - y\|^2 \tag{3.26}$$

Remarque 3.5.3. *Il arrive parfois que la relation fonctionnelle entre la variable expliquée et les variables explicatives ne soit pas donnée sous forme linéaire, comme dans l'exemple suivant :*

$$y = f(x_1, x_2) = x_1^{\beta_1} x_2^{\beta_2}$$

Dans ce cas, on ≪ linéarise ≫ : $ln(y) = \beta_1 ln(x_1) + \beta_1 ln(x_1)$.
La nouvelle variable expliquée est $b' = ln(b)$ *et les nouvelles variables explicatives sont* $ln(x_1)$ *et* $ln(x_2)$.

3.5.4 Recherche d'une solution

On fait l'hypothèse que les variables explicatives sont linéairement indépendantes (i.e. $rang(X) = n$).

Solution géométrique

On cherche à exprimer un vecteur comme combinaison linéaire de m vecteurs indépendants. Cette combinaison linéaire appartient, par définition d'un espace-vectoriel, à l'espace vectoriel engendré par ces variables explicatives:

$$vect(x_1, x_2, \ldots, x_m) = Im(X) \text{ où } Im(X) \text{ désigne } l'ensemble \text{ image de } X.$$

C'est un sous espace vectoriel de $\mathbb{R}^m$ $(m > n)$.

La recherche consiste ici à exhiber donc $\hat{y} \in Im(X)$ tel que $\|\hat{y} - y\|^2$ soit minimal : c'est la définition de la projection orthogonale de y sur $Im(X)$.

Lemme 3.5.4. ***caractérisation de la projection orthogonale sur un sous-espace vectoriel.***
Soit $x \in \mathbb{R}^k$ *et* K *un sous espace vectoriel de* $\mathbb{R}^k$ *, on a :*
$\hat{x}$ *projeté orthogonal de* x *sur* $K \iff x - \hat{x}$ *est orthogonal à tout vecteur de* K

Il faut noter que dans notre cas le sous-espace vectoriel est $Im(X)$.

Remarque 3.5.5. *Tout vecteur de* $Im(X)$ *s'écrit* $X\beta$, $\beta \in \mathbb{R}^n$. β *est le représentant des coordonnées du vecteur dans la base* X.

Soit $X\beta \in Im(X)$, d'après la caractérisation de la projection orthogonale, on a :

$$\begin{aligned}
(X\beta | y - \hat{y}) &:= (X\beta | y - Xx) = 0 \ \forall \beta \in \mathbb{R}^n \\
&\iff \beta^T X^T (y - Xx) = 0 \ \forall \beta \in \mathbb{R}^n \\
&\iff X^T (y - Xx) = 0 \\
&\iff X^T y - X^T X x = 0 \\
&\iff X^T y = X^T X x
\end{aligned}$$

La solution du problème est donc la solution x du système $X^T y = X^T X x$ et cette solution est unique car c'est le projeté y sur un ensemble convexe.

Solution analytique

Notons $\varepsilon(x) = \|Xx - y\|^2$ la fonction d'erreur [9]. On peut vérifier si que la fonction $\varepsilon(x)$ est strictement convexe alors on a $\varepsilon(x) \iff \varepsilon'(x) = 0$ où $\varepsilon'(x)$ désigne la dérivée matricielle de $\varepsilon(x)$. Comme nous allons utiliser la dérivée matricielle, dans le paragraphe suivant est consacré à cette notion.

9. $\varepsilon(x) \Longrightarrow$ la dérivée de $\varepsilon'(x) = 0$

Dérivation matricielle

Soit une forme linéaire $f : \mathbb{R}^k \longrightarrow \mathbb{R}$ et $x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{pmatrix}$ un vecteur de $\mathbb{R}^k$.

Définition 3.5.6. *On appelle dérivée f en x et on note $\frac{\partial f}{\partial x}$ ou $\nabla f(x)$ ou encore $f'(x)$ le vecteur colonne des dérivée partielles de f par rapport aux x_i:*

$$\frac{\partial f}{\partial x} = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_k} \end{pmatrix}$$

Proposition 3.5.7. *Si f est dérivable en x, alors, quelque soit la direction $d \in \mathbb{R}^k$ on a:*

$$Df(x,d) = \langle f(x), d\rangle = f'(x)^T d$$

Remarque 3.5.8. *Cette égalité nous permet de calculer la dérivée d'une forme linéaire de la façon suivante :*

1. *on calcule $Df(x,d)$*
2. *on l'exprime sous la forme d'un produit scalaire en d : $(Q|d)$ ou on factorise d à droite : $Q^T d$*
3. *le facteur gauche Q (ou l'autre facteur que d du produit scalaire) est nécessairement $f'(x)$.*

Proposition 3.5.9. *Les dérivées directionnelles dans la direction des vecteurs de la base canonique $(e^i)_{i\in[1,k]}$ sont les dérivées partielles :*

$$\forall i \in [1,k], \frac{\partial f(x)}{\partial x_i} = Df(x, e^i)$$

Cette proposition donne un moyen simple de calculer les dérivées partielles.

Dérivation de formes linéaires

$$Soit \quad f : \left\{ \begin{array}{ccc} \mathbb{R}^k & \longrightarrow & \mathbb{R} \\ x = (x_1, \ldots, x_k) & \longmapsto & \alpha_1 x_1 \quad + \ldots + \alpha_k x_k = \langle x, \alpha\rangle = \langle \alpha, x\rangle = \alpha^T x = x^T \alpha \end{array} \right.$$

$\forall i \in [1,k], \frac{\partial f(x)}{\partial x_i} = \alpha_i$ donc $f'(x) = \nabla f(x) = \frac{\partial f}{\partial x} = \alpha$

Retenons que $\forall \alpha, x \in \mathbb{R}^k$ *on a* : $\frac{\partial(\alpha^T x)}{\partial x} = \frac{\partial(x^T \alpha)}{\partial x} = \alpha$

Remarque 3.5.10. *On retrouve ce résultat par les dérivées partielles. En effet, pour une direction d quelconque, on a:*

$$\begin{aligned} Df(x,d) &= \lim_{\epsilon \longrightarrow 0} \frac{f(x+\epsilon d) - f(x)}{\epsilon} \\ &= \lim_{\epsilon \longrightarrow 0} \frac{\langle x + \epsilon d, \alpha\rangle - \langle x, \alpha\rangle}{\epsilon} \\ &= \lim_{\epsilon \longrightarrow 0} \frac{\langle x, \alpha\rangle + \epsilon\langle d, \alpha\rangle - \langle x, \alpha\rangle}{\epsilon} \\ &= \lim_{\epsilon \longrightarrow 0} \langle d, \alpha\rangle \\ &= \langle d, \alpha\rangle = \langle \alpha, d\rangle = \alpha^T d \\ &\Longrightarrow f'(x) = \alpha \end{aligned}$$

Dérivation d'une forme quadratique

Définition 3.5.11. *Une forme quadratique est un polynôme homogène de degré 2 avec un nombre quelconque de variables :*

$$f : \mathbb{R}^k \longrightarrow \mathbb{R}$$
$$x = (x_1, \ldots, x_k) \longmapsto \sum_{i,j=1}^{k} a_{i,j} x_i y_j$$

En notant $A = (a_{i,j}) \in \mathcal{M}_{(k,k)}(\mathbb{R})$, f s'écrit :

$$f : \mathbb{R}^k \longrightarrow \mathbb{R}$$
$$x \longmapsto x^T A x = \langle x, Ax \rangle = \langle Ax, x \rangle$$

A l'aide des dérivées directionnelles on obtient la dérivée de la forme quadratique: $Df(x,d) = \langle (A + A^T)x, d \rangle$ et on a : $f'(x) = (A + A^T)x$.

En bref $\forall x \in \mathbb{R}^k, A \in \mathcal{M}_{(k,k)} : \frac{\partial(x^T A x)}{\partial x} = (A + A^T)x$

Remarque 3.5.12. *Si A est symétrique, $A = A^T$ alors on a:$\frac{\partial(x^T A x)}{\partial x} = (A + A^T)x = 2Ax$ En dimension 1 on retrouve le résultat bien connu $\frac{\partial(x^T a x)}{\partial x} = \frac{\partial(a x^2)}{\partial x} = 2ax$.*

Calcul de la solution

La fonction d'erreur

$$\begin{aligned}\varepsilon(x) &= \|Xx - y\|^2 \\ &= \|Xx\|^2 - 2\langle Xx, y \rangle + \|y\|^2 \\ &= x^T X^T X x - 2x^T X^T y + y^T y\end{aligned}$$

Calculons sa dérivée de $\varepsilon'(x)$

$$\begin{aligned}\varepsilon'(x) &= \frac{\partial(x^T X^T X x)}{\partial x} - 2\frac{\partial(x^T X^T X y)}{\partial x} + \frac{\partial y^T y}{\partial x} \\ &= 2X^T X x - 2X^T y + 0 \\ &= 2X^T X x - 2X^T y\end{aligned}$$

$$\begin{aligned}\varepsilon'(x) = 0 &\Longleftrightarrow 2X^T X x - 2X^T y = 0 \\ &\Longleftrightarrow X^T X x = X^T y\end{aligned}$$

3.5.5 Interprétation statistique

On suppose que y est la réalisation d'une variable aléatoire et que $y = X\beta + \varepsilon$ où ε est l'erreur, encore appelée variable aléatoire résiduelle, suivant une loi $\mathcal{N}(0, \sigma^2)$. $\hat{y} = X\beta$ est alors l'estimateur sans biais du maximum de vraisemblance de y (l'espérance mathématique $E(y) = \hat{y}$).

Theorem 3.5.13. *[31]*

Soit $X \in \mathcal{M}_{(m,n)}$ avec $m > n$ et $y \in \mathbb{R}^m$. Une condition nécessaire et suffisante pour que $\beta \in \mathbb{R}^n$ réalise le minimum de $E(\beta) = \|X\beta - y\|^2$ est que

$$X^T X \beta = X^T y \tag{3.27}$$

*Les équations (3.27) sont appelées **équations normales**. Ce système admet toujours au moins une solution. Si la matrice $X^T X$ est régulière, i.e. si $rang(X) = n$, alors la solution est unique.*

3.5.6 Inconvénients

La résolution d'un problème de moindres carrés par les équations normales possède deux inconvénients majeurs [25] :

D'une part, la perturbation due aux erreurs d'arrondi lorsque l'on passe par les équations normales peut être importante. En effet, si la matrice des données X est légèrement perturbée : $X' = X + \lambda X$, le passage aux équations normales va amplifier la perturbation : $(X + \lambda X)^T(X + \lambda X) = X^TX + \lambda X^TX + X^T\lambda X + \lambda X^T\lambda X \ldots$ alors qu'en passant par d'autres méthodes de résolution (par exemple factoriser X sous la forme OT où O est orthogonale et T triangulaire) la perturbation des données sera moindre. D'autre part, le calcul de X^TX peut faire intervenir des overflow ou underflow parasites comme dans l'exemple suivant : Soit la matrice des données suivante

$$X = \begin{pmatrix} 1 & 1 & 1 \\ \epsilon & 0 & 0 \\ 0 & \epsilon & 0 \\ 0 & 0 & \epsilon \end{pmatrix}$$

avec $\epsilon \neq 0$. On a bien $rang(A) = 3$.

$$X^TX = \begin{pmatrix} 1+\epsilon^2 & 1 & 1 \\ 1 & 1+\epsilon^2 & 1 \\ 1 & 1 & 1+\epsilon^2 \end{pmatrix}$$

Si ϵ est supérieur au plus petit flottant représentable alors que ϵ^2 lui est inférieur, soit :

$$\epsilon^2 < m < \epsilon$$

la matrice X^TX ne sera plus régulière !

Pour ces problèmes, on fait recours à des transformations orthogonales élémentaires, comme celles de Householder ou de Givens. Avec ces méthodes, on obtient une meilleure $\ll$ *stabilité* $\gg$ de notre système.

3.6 Conclusion

Ce chapitre offre les outils d'optimisation qu'on utilise dans les deux prochains chapitres qui suivent. Il met à notre disposition les techniques mathématiques pour modéliser et prévoir l'indice de la qualité de l'air dans Dakar par la régression linéaire multiple. C'est l'objet du chapitre 4 suivant.

Chapitre 4

Modélisation et prédiction de l'indice de la qualité de l'air dans Dakar

Sommaire

4.1 Introduction

Ce chapitre introduit la régression linéaire en passant de la régression linéaire simple à la régression linéaire multiple par les techniques des moindres carrés ordinaires (MCO). Il présente une démarche de modélisation et prédiction en utilisant la régression linéaire multiple. L'article publié dans le IJAMAS (International Journal of Applied Mathematics and Statistics) est une application de cette démarche et clôture le chapitre. La régression linéaire manipule les séries chronologiques. Dans ce qui suit, on commence par appréhender cette notion.

Série Chronologique

Une **série temporelle, ou série chronologique**, est une suite de valeurs numériques représentant l'évolution d'une quantité spécifique au cours du temps. C'est une suite d'observations numériques (mesures) indicées par le temps. Ces observations sont représentées par :

$$x_1, x_2, \ldots, x_n$$

où $1, 2, \ldots, n$ représentent les marques temporelles, x_i est la valeur de la mesure réalisée au temps i. De telles suites de variables aléatoires peuvent être exprimées mathématiquement afin d'en

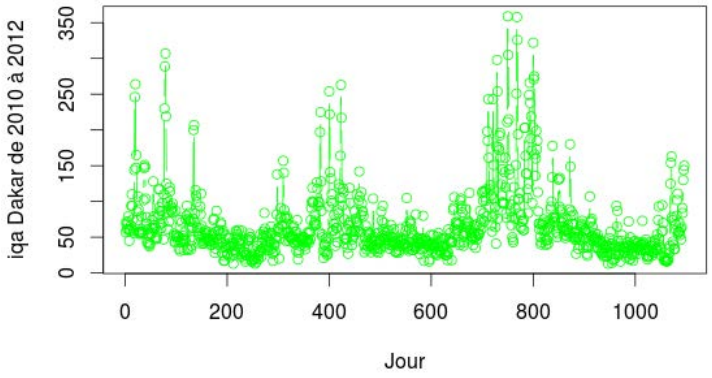


FIGURE 4.1 – La série indice de la qualité de l'air à Dakar de la période 2010 à 2012

analyser le comportement, généralement pour comprendre son évolution passée et pour en prévoir le comportement futur. Une telle transposition mathématique utilise le plus souvent des concepts de probabilités et de statistique. L'étude des séries temporelles a bénéficié de nombreux apports théoriques fondés, essentiellement, sur les progrès de leur modélisation.

Le but des séries temporelles est de décrire, d'expliquer ou de prévoir un phénomène. En générale des analyses (décompositions) suivantes peuvent être faites sur les séries chronologiques :

- **Analyse avec un modèle additif.** La série se décompose dans ce cas sous la forme (voir la figure 4.2):

$$x_n = t_n + s_n + \sigma_n$$

où

(i) t_n est la tendance. C'est la composante essentielle de la série. La tendance (trend en anglais) comme un comportement moyen de la série.

(ii) s_n est la série des coefficients de variations saisonnières. Cette série est périodique. Elle mesure les variations saisonnières de la série dues à divers comportements sociaux et aux variations climatiques.

(iii) σ_n est ce qui reste. C'est le hasard de la série qui a priori n'est pas analysable.

- **Analyse avec un modèle multiplicatif.** La séries se décompose comme de la manière suivante :

$$x_n = t_n s_n \sigma_n$$

où t_n , s_n et σ_n ont les mêmes définitions que dans le cas additif.

Mais cependant comment retenir la bonne décomposition ? Le choix se fait à l'aide des observations faites à partir de la représentation graphique de la série. Pour un modèle additif, l'amplitude des oscillations reste constante, alors que pour un modèle multiplicatif, l'amplitude des oscillations varie (l'amplitude est strictement croissante dans le temps). Ce sont ces remarques qui vont guider le choix du modèle. En effet, la représentation d'une série chronologique (x_n) faisant apparaître des oscillations, nous permettra de constater si ces dernières restent constantes ou varient.

Définition 4.1.1. ***Notion de bruit blanc****[46, 49].*

Un processus ϵ_t est qualifié de bruit blanc indépendant si :

(i) $E[\epsilon_t] = 0$,

(ii) $E[\epsilon_t^2] = \sigma^2$

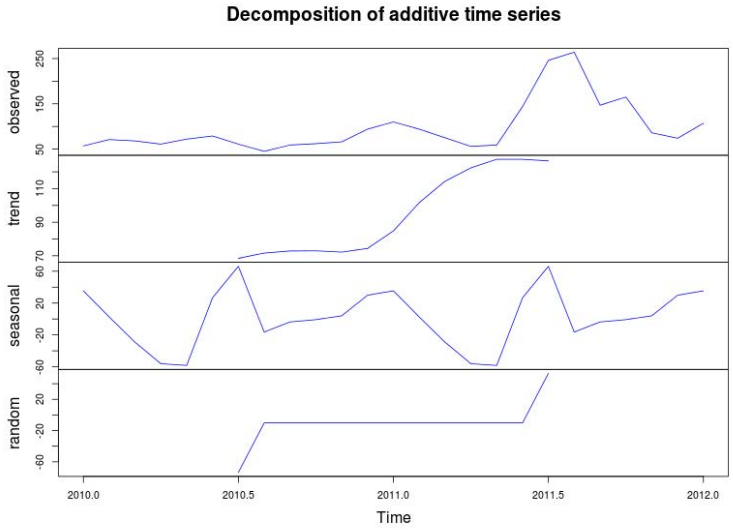


FIGURE 4.2 – Décomposition additive de la série indice de la qualité de l'air (iqa) à Dakar : période 2010 à 2012

(iii) *ϵ_t, et ϵ_τ, sont indépendants $\forall t \neq \tau$*

On remarque que la condition **(iii)** de la définition du bruit blanc indépendant, celle d'indépendance, implique la condition d'autocovariance nulle du bruit blanc, tandis que la réciproque n'est pas forcément vraie. On a la définition suivante qui est plus stricte que la première.

Définition 4.1.2. ***Bruit blanc gaussien***
Un processus ϵ_t est qualifié de bruit blanc gaussien si

(i) *ϵ_t* ***est un bruit blanc indépendant*** *,*

(ii) $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$

Remarque 4.1.3. *Il faut distinguer les deux définitions suivantes:*
- *Un processus stochastique x sur un espace E est un* ***bruit blanc fort*** *(BBF) si les variables $x(e), e \in E$ sont centrées, indépendantes et identiquement distribuées.*
- *Un processus stochastique x sur un espace E est un* ***bruit blanc faible*** *(BBf) si les variables $x(e), e \in E$ sont centrées, décorrélées, et de variances finies constantes ($Cov(x_e, x_t) = \sigma^2 \delta_{e=t}$)*

On parle d'**analyse de la variance (ANOVA)** si y quantitative, $x_1, \ldots, x_m$ qualitatives. Expliquer y revient à attribuer une valeur moyenne dans chaque classe définie à partir des valeurs de $x_1, \ldots, x_m$ (par exemple si x_i peut prendre k_i valeurs possibles, il existe $k_1 \times, \ldots \times k_p$ classes différentes). On peut alors essayer d'évaluer si chaque variable explicative a une influence ou non sur y.

On a une **analyse de covariance (ANCOVA)** dans le cas où y est quantitative, $x_1, \ldots, x_m$ qualitatives et quantitatives. Les valeurs différentes des variables explicatives qualitatives définissent des classes dans lesquelles on effectue la régression linéaire de y sur les variables explicatives quantitatives.

Généralement la meilleure approximation de y (pour le coût quadratique) par une fonction des x_i est donnée par l'espérance conditionnelle

$$J(x_1, \ldots, x_m) = E(y|x_1, \ldots, X_m),$$

qui est bien sûr inconnue en pratique. Lorsque y admet un moment d'ordre 2, l'espérance conditionnelle minimise l'erreur quadratique $E[(y - J(x_1, \ldots, x_m))^2]$. Si le vecteur $(y, x_1, \ldots, x_m)$

est gaussien alors l'espérance conditionnelle est une fonction affine. Dans ce cas, on peut donc se restreindre aux fonctions J linéaires en 1, $x_1, \ldots, x_m$,

$$J(x_1, \ldots, x_m) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_m,$$

ce qui justifie le terme régression linéaire.

Une régression est dite **régression de modèle I** lorsque les variables x sont contrôlées c'est à dire non aléatoires et de régression de modèle II dans le cas de variables aléatoires.

On parle d'un problème de **régression linéaire simple** si le problème n'implique qu'une seule variable prédictive, utilisée simplement au premier degré (et non pas sous la forme $x^2, x^3, \ldots$).

Un problème est dit de **régression multiple** lorsque l'estimation est fondée sur plusieurs variables prédictives.

4.2 La régression linéaire simple

On rappelle qu'un modèle linéaire simple est un modèle de régression linéaire avec une seule variable explicative. Ce modèle est souvent présenté dans certains manuels sous le titre d'ajustement affine.

On a donc deux variables aléatoires, une variable expliquée y, qui est un scalaire, une variable explicative x, également scalaire. On dispose de n réalisations de ces variables, $(x_i)_{1\leq i\leq n}$ et $(y_i)_{1\leq i\leq n}$, soit :

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

ou ϵ_i est le terme d'erreur ; chaque terme d'erreur lui-même est une réalisation d'une variable aléatoire ϵ_i. L'objectif est de chercher à **expliquer** les variations d'une variable quantitative y (par exemple, l'indice de la qualité de l'air dans Dakar) par une variable explicative x également quantitative (par exemple, la vitesse du vent). La forme matricielle du modèle est:

$$y = \beta x + \epsilon \tag{4.1}$$

avec $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$, $x = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$, $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$

Dans le cadre d'un modèle linéaire simple, on peut représenter graphiquement la relation entre x et$y$ à travers un nuage de points. L'estimation du modèle linéaire permet de tracer la droite de régression. Le paramètre β_0 représente l'ordonnée à l'origine et β_1 le coefficient directeur de la droite. On suppose que les bruits ϵ_i sont

1. **centrées**: $E(\epsilon_i) = 0$,
2. **non-corrélés** :$\forall i \neq j, cov(\epsilon_i, \epsilon_j) = 0$,
3. **de variances** égales **(homoscédastiques)** : $var(\epsilon_i) = \sigma^2 < \infty$.

On définit le **coefficient de corrélation linéaire** ou **coefficient de Pearson** par

$$\rho(x, y) = \frac{Cov(x, y)}{\sigma_x \sigma_y} \tag{4.2}$$

IL est compris entre -1 et 1 (on le montre avec l'inégalité de Cauchy-Schwarz). Le lien affine entre x et y est d'autant plus net que $\rho(x, y)$ est proche de 1 ou -1 (1 pour une relation croissante et -1 pour une relation décroissante).

4.3 Régression linéaire multiple

Il arrive souvent qu'on veuille expliquer la variation d'une variable dépendante par l'action de plusieurs variables explicatives. La technique de la régression linéaire multiple est un outil pour faire cette explication.

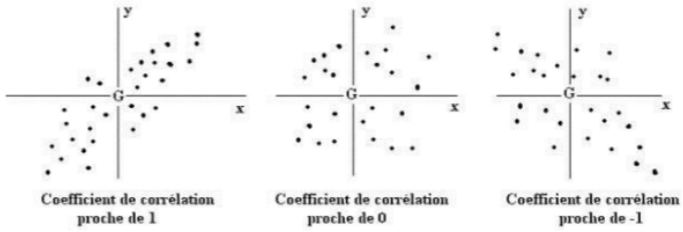


FIGURE 4.3 – Interprétation de différentes valeurs prises par le coefficient de corrélation.

Exemple 4.3.1. *L'indice de la qualité de l'air iqa est influencée par la température* (T) *et par la vitesse du vent* (V_x)*. Le vent est mesuré en degré (direction) et mètre par seconde (vitesse).*

Nous avons synthétisé ces 2 variables en créant une variable (V_x) *qui est la projection du vent sur l'axe est-ouest. Nous avons* $n = 50$ *observations. Au-delà de 2 variables explicatives, il est impossible de visualiser simplement les données. Le logiciel R nous donne l'image de la régression simple, en traçant les données (voir figure 4.4).*

L'issue de l'estimation avec le logiciel R nous donne le modèle $iqa = -74.70299 + 5.85599T - 0.03025V_x$

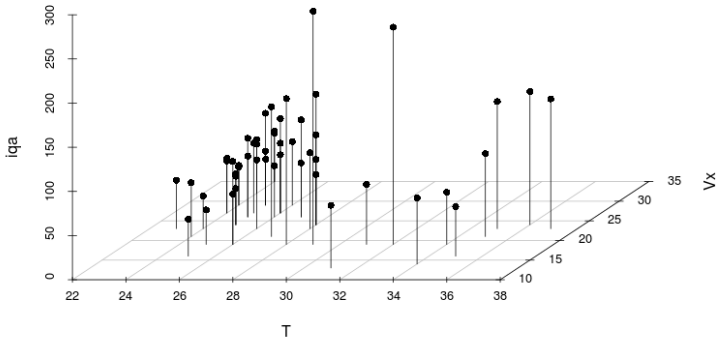


FIGURE 4.4 – Représentation brute des données : modèle d'explication de l'indice de la qualité de l'air dans Dakar (iqa) par la température sur les 50 premiers jours de l'année 2010 (T) et le vent (V_x).

Dans cette partie de la thèse on s'intéresse à un cas plus générale de modèle d'une variable y en fonction de plusieurs variables explicatives $x_1, \ldots, x_p$. Le modèle est une généralisation de la régression linéaire simple. On observe des réalisations indépendantes $\{y_i, x_{1,i}, \ldots, x_{p,i}\}_{i=1,\ldots,n}$, où

$$y_i = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_p x_{p,i} + \epsilon_i \quad i = 1, \ldots, n.$$

Comme dans la régression linéaire simple, les ϵ_i sont centrés, de même variance $\sigma^2 < \infty$ et non-corrélés. Le modèle s'écrit sous la forme matricielle suivante:

$$y = x\beta + \epsilon,$$

avec

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, x = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{p,1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1,n} & \dots & x_{p,n} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} \text{ et } \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

- y désigne le vecteur à expliquer de taille $n \times 1$,
- x la matrice explicative de taille n $\times(p+1)$,
- le vecteur d'erreurs ϵ de taille $n \times 1$.

4.3.1 Régression linéaire multiple et moindres carrés ordinaires (MCO)

De même manière que dans le cas de la régression simple, on cherche à estimer β et σ. L'estimateur des MCO $\hat{\beta}$ est défini comme l'unique minimiseur de la fonction

$$J(\beta) = \|y - x\beta\|^2, \quad \beta \in \mathbb{R}^{p+1}$$

On suppose que $p < n$ et que x est de rang $p+1$. Sous ces hypothèses, l'estimateur β est l'unique solution des conditions du premier ordre

$$\nabla J(\hat{\beta}) = -2x^T y + 2x^T x\hat{\beta} = 0 \Longleftrightarrow \hat{\beta} = (x^T x)^{-1} x^T y.$$

$x^T x$ est inversible car la matrice x est de plein rang.

Proposition 4.3.2. *[48] L'estimateur des MCO $\hat{\beta}$ est un estimateur sans biais de β de matrice de variance*

$$var(\hat{\beta}) = \sigma^2 (x^T x)^{-1}.$$

Theorem 4.3.3. ***Gauss-Markov*** *[49, 53].*
L'estimateur des moindres carrés $\hat{\beta}$ est optimal (au sens du coût quadratique) parmi les estimateurs sans biais linéaires en y .

L'optimalité au sens $\mathbb{L}^2$ ne nécessite pas la normalité du modèle. Un résultat plus fort est valable dans le cas Gaussien $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ où la variance de β atteint la borne de Cramer-Rao. L'estimateur des moindres carrés est donc optimal parmi tous les estimateurs sans biais de β dans ce cas.

La matrice $\Pi_x := x(x^T x)^{-1} x^T$utilisée dans la preuve du théorème de Gauss-Markov est la projection orthogonale sur l'image de x. On le montre simplement en vérifiant que Π_x est symétrique et vérifie $\Pi_x^2 = \Pi_x$ et $Im(\Pi_x) = Im(x)$. Ainsi, le **vecteurs des prévisions**

$$\hat{y} = x\hat{\beta} = x(x^T x)^{-1} x^T y \tag{4.3}$$

est la **projection orthogonale** de y sur $Im(x)$. C'est en quelque sorte la part de $y \in \mathbb{R}^n$ expliquée par les variables $\mathbf{1}, x_1, \dots, x_p$ (les colonnes de x). De même, le vecteur des résidus

$$\hat{\epsilon} = y - \hat{y} = (I - x(x^T x)^{-1} x^T)y = (I - x(x^T x)^{-1} x^T)(x\beta + \epsilon) = (I - x(x^T x)^{-1})\epsilon$$

est la projection orthogonale de y sur $Im(x)^{\perp}$ et par conséquent celle de ϵ puisque $y = x\beta + \epsilon$. Une conséquence immédiate est que les vecteurs $\hat{\beta}$ et $\hat{\epsilon}$ sont non-corrélés. En effet,

$$cov(\hat{\beta}, \hat{\epsilon}) = E[(\hat{\beta} - \beta)\hat{\epsilon}^T] = (x^T x)^{-1} x^T E[\epsilon\hat{\epsilon}^T](I - x(x^T x)^{-1} x^T) = 0$$

La norme du vecteur des résidus permet de construire un estimateur de σ^2 par

$$\hat{\sigma} := \frac{1}{n-p-1}\|y - \hat{y}\|^2 = \frac{1}{n-p-1}\|\hat{\epsilon}\|^2 \tag{4.4}$$

Proposition 4.3.4. *L'estimateur σ^2 est sans biais.*

Un résultat plus fort est valable dans le cas gaussien.

Proposition 4.3.5. *Dans le modèle gaussien $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$, les estimateurs $\hat{\beta}$ et $\hat{\sigma}^2$ sont indépendants et vérifient*

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (x^T x)^{-1}) \quad et \quad (n-p-1)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p-1)$$

Il est intéressant de remarquer que dans le cas Gaussien, $\hat{\beta}$ est l'estimateur du maximum de vraisemblance. En revanche, l'estimateur du maximum de vraisemblance de σ^2 est différent, donné par

$$\hat{\sigma}^2_{MV} = \frac{1}{n}\|y - \hat{y}\|^2 = \frac{n-p-1}{n}\hat{\sigma}^2.$$

4.3.2 Comportements asymptotiques des estimateurs

Dans ce paragraphe on s'intéresse au comportement des estimateurs quand n tend vers l'infini. Sous les hypothèses fortes de la régression, les lois de β et σ^2 sont connues ce qui permet de déduire facilement leur comportement asymptotique. La convergence de $\hat{\beta}$ dans $\mathbb{L}^2$ est soumise à la seule condition que $(x^T x)^{-1}$ tend vers 0. La convergence de σ , que ce soit dans $\mathbb{L}^2$ ou même presque sûrement, est vérifiée dans le cas Gaussien sans hypothèse supplémentaire sur x. On peut se demander si ces résultats restent valables sans la normalité des bruits. Un premier résultat immédiat est que sous les hypothèses faibles, β reste convergent dans $\mathbb{L}^2$ dès que $(x^T x)^{-1}$ tend vers 0. On peut également montrer que si les ϵ_i sont iid et $h_n := \max_{1\leq i,j\leq p+1} |\Pi_{x,ij}|$ tend vers 0, alors $\hat{\beta}$ est asymptotiquement Gaussien. Si de plus $\frac{1}{n}x^T x$ converge quand $n \longrightarrow \infty$ vers une matrice inversible M, alors

$$\sqrt{n}(\hat{\beta}, \beta) \xrightarrow{loi} \mathcal{N}(0, \sigma^2 M^{-1})$$

L'hypothèse $\frac{1}{n}x^T x \longrightarrow M$ est souvent vérifiée en pratique. Par exemple, elle est vérifiée presque sûrement si l'échantillon $x_{1i}, \ldots, x_{pi\,i=1,\ldots,n}$ est issu de réalisations indépendantes de variables aléatoires $x_1, \ldots, x_p$ de carré intégrable. Dans ce cas, $\frac{1}{n}x^T x$ converge presque sûrement vers la matrice des moments d'ordre deux, par la loi forte des grands nombres. Ce résultat est important car il confirme que même sous les hypothèses faibles, la plupart des tests du modèle linéaire restent valables asymptotiquement.

4.3.3 Analyse de la variance (coefficient de détermination)

Elle consiste à diviser la variance de y en une partie expliquée par les variables $x_1, \ldots, x_p$ et une partie résiduelle. Cela revient à remarquer que

1. $\hat{y} = x(x^T x)^{-1}x^T y = \Pi_x y$ est la projection orthogonale de y sur $Im(x)$.
2. $\bar{y}\mathbf{1}$ est la projection orthogonale de y sur l'espace engendré par le vecteur constant $\mathbf{1}$, noté $vec\{\mathbf{1}\}$.
3. $vec\{\mathbf{1}\}$ étant un sous un espace de $Im(x)$, $\bar{y}\mathbf{1}$ est également la projection orthogonale de $\hat{y}$ sur $vec\{\mathbf{1}\}$(on peut vérifier que $\hat{\bar{y}} = \bar{y}$)
4. $\hat{\epsilon} = y - \hat{y} = (I - x(x^T x)^{-1}x^T)y$ est la projection orthogonale de y sur $Im(x)^{\perp}$.

On obtient ainsi la décomposition de $y - \bar{y}\ \mathbf{1} = \hat{y} - \bar{y}\mathbf{1} + \hat{\epsilon}$, le théorème de Pythagore nous fournit:

$$\underbrace{\|y - \bar{y}\ \mathbf{1}\|^2}_{SCT} = \underbrace{\|\hat{y} - \bar{y}\mathbf{1}\|^2}_{SCE} + \underbrace{\|\epsilon\|^2}_{SCR}$$

Le coefficient de détermination R^2 qui donne un indicateur de la qualité de la modélisation est défini par

$$R^2 := \frac{SCE}{SCT} = \frac{\|\hat{y} - \bar{y}\ \mathbf{1}\|^2}{\|y - \bar{y}\ \mathbf{1}\|^2}$$

Dans le cas univarié, le coefficient de détermination est égal au carré du coefficient de corrélation de Pearson $\rho(x, y)$. Dans le cas multivarié, le R^2 correspond à la valeur maximale du carré du coefficient de Pearson entre y et une combinaison linéaire des variables explicatives [46, 47, 54]. :

$$R^2 = \sup_{\beta \in \mathbb{R}^{p+1}} \rho(y, x\beta)^2$$

4.3.4 Prédiction

On fait une observation d'un nouveau jeu de variables $x_{1,n+1}, \ldots, x_{1,n+1}$ et on cherche à prédire la valeur y_{n+1} correspondante. On note $x_{n+1} = (1, x_{1,n+1}, \ldots, x_{p,n+1})$. Sous l'hypothèse de normalité (qui est essentielle ici), la prédiction $\hat{y}_{n+1} = x_{n+1}\hat{\beta}$ suit une loi normale $\mathcal{N}(x_{n+1}\beta, \sigma^2 x_{n+1}(x^T x)^{-1} x_{n+1}^T)$ et est indépendante de $y_{n+1} = x_{n+1}\beta + \epsilon_{n+1}$ [50, 51, 52, 47]. . On montre alors que

$$\frac{\hat{y}_{n+1} - y_{n+1}}{\hat{\sigma}\sqrt{x_{n+1}(x^T x)^{-1} x_{n+1}^T}} \sim \mathcal{T}_{n-p-1}, \tag{4.5}$$

Cela donne un outil de construction d'un intervalle de prédiction, qui n'est valable que sous l'hypothèse de normalité.

4.3.5 Vérification des hypothèses

Une grande des résultats de la régression linéaires reposent sur les hypothèses de **normalité, homoscédasticité et non-corrélations des résidus**. C'est donc nécessaire de pouvoir vérifier la validité de ces hypothèses.

1. **Normalité** : pour vérifier si les bruits ϵ_i sont gaussiens, on effectue un test de normalité sur les résidus $\hat{\epsilon_i}$. En effet, la normalité de ϵ_i entraîne la normalité de $\hat{\epsilon_i}$. Plusieurs tests existent comme le test de Shapiro-Wilk (commande `shapiro.test` sour R) ou encore le test de Lilliefors (commande `lillie.test` du package `nortest`). Le diagramme quantile-quantile (ou qq-plot) permet également de vérifier graphiquement la normalité des résidus.
2. **Homoscédasticité (ou homogénéité)** : Le test de Breusch-Pagan (commande `bptest` du package `lmtest`) permet de tester si la variance des bruits est constante. On peut également utiliser le test de White (commande `white.test` du package `bstats`). Graphiquement, l'hétéroscédasticité du bruit se traduit par une répartition d'ampleurs inégales du nuage de points autour de la droite de régression.
3. **Non-corrélation** : Le test de Breusch-Godfrey (commande `bgtest` du package `lmtest`) permet de tester une corrélation à l'ordre 1 ou supérieur des bruits ϵ_i . Pour tester une corrélation à l'ordre 1, on peut également utiliser la statistique de Durbin-Watson (commande `dwtest` du package `lmtest`), définie par

   $$D = \frac{\sum_{i=2}^{n}(\hat{\epsilon}_i - \hat{\epsilon}_{i-1})^2}{\sum_{i=2}^{n}\hat{\epsilon}_i^2} \tag{4.6}$$

   On montre facilement que D est comprise entre 0 et 4 mais sa loi sous l'hypothèse nulle de non-corrélation n'est pas une loi usuelle. Une règle de décision couramment utilisée est de conclure qu'il n'y a pas de corrélation entre ϵ_i et ϵ_{i+1} si la statistique de Durbin-Watson est comprise entre 1 et 3.

On a utilisé les techniques de la régression linéaire multiple par les moindres carrés ordinaires (MCO) ci-dessus pour publier un article dans le Journal IJAMAS [32] pour prédire l'indice de la qualité de l'air à Dakar. Voici à présent la présentation de l'article en Français.

4.4 Modèle et prédiction de l'iqa

La qualité de l'air joue un rôle très important dans la vie actuelle de nos sociétés. Elle est intrinsèquement liée aux politiques stratégiques de régulation et d'amélioration de la santé humaine

et de l'environnement. Dans cette étude, on présente l'indice de la qualité de l'air dans Dakar pendant la période de 2010 à 2012 et on propose un modèle de sa prédiction par la technique de la régression multiple. Cette technique utilise les méthode des moindres carrés ordinaire.

On cherche à déterminer un modèle de prédiction de l'indice de la qualité de l'air à Dakar en fonction de la température, l'humidité, la vitesse du vent, le point de rosée, la pression au niveau des mer,la visibilité et les précipitations.

Dans cette partie, on utilise les modèles **statistiques** notamment la régression linéaire multiple (RML). Parlant de la régression, on cite en premier lieu les modèles classiques linéaires : la régression linéaire simple (on exprime la concentration d'un polluant en fonction d'une seule variable qui peut être le temps, un facteur météorologique, etc.) ou la régression linéaire multiple (la concentration de polluant est exprimée en fonction de plusieurs variables explicatives). La RLM a été utilisée dans plusieurs études, mais dans la plupart des cas à titre de comparaison. Goyal and al. (2006) [34] ont étudié la prévision des niveaux totaux de particules respirables (moyennes journalières) dans deux métropoles : Delhi et Hong Kong. Les auteurs ont développé un premier modèle basé sur une RLM, les variables explicatives étant des paramètres météorologiques : vitesse du vent, rayonnement solaire, humidité relative et température de surface. La RLM a montré que les contributions de certaines variables étaient significatives (vitesse du vent, humidité relative), mais le rayonnement solaire n'était pas une variable influente. Le modèle explique 58 % de la variance des concentrations journalières de particules, avec une erreur quadratique moyenne de 76 $\mu g.m^{-3}$, ce qui n'est pas très satisfaisant. Le deuxième modèle utilisé est de type ARIMA, et le troisième, une combinaison linéaire des deux premiers. La RLM a été utilisée à titre de comparaison dans d'autres travaux, comme ceux de Salini et Perez en 2008, de Slice and al. en 2006, ou bien Robles-Diaz and al en 2008 [35, 36, 37]. Généralement, les différents auteurs ont retenu la RLM comme une alternative intéressante en raison de sa simplicité de mise en œuvre, et menant à des résultats plutôt convenables, mais avec des performances plus faibles que celles des modèles plus élaborés. Une façon particulière d'appliquer la RLM est d'utiliser les composantes principales à la place des variables explicatives, lorsque ces dernières sont corrélées entre elles (ce qui arrive souvent), mais dans cette recherche on ne présentera pas d'exemple d'application pour la prédiction de l'indice de la qualité dans ce cas. Nous proposons un modèle de prévision par la technique de régression multiple. Afin de souligner l'impact significatif de chaque variable explicative sur la qualité de l'air, nous utilisons la statistique de student pour mettre en évidence l'incidence significative de chaque variable exogène sur la variable dépendante. En outre, nous prédisons le logarithme néperien des valeurs de l'indice de la qualité de l'air pour (31) jours de Janvier 2013 et comparons ces résultats avec les mesures réelles prises dans les stations au cours de la même période.

4.4.1 Modélisation

Cette recherche s'intéresse principalement à la qualité de l'air dans la ville de Dakar. Elle est appréciée à travers l'indice de la qualité de l'air que nous modélisons par la technique de régression linéaire multiple.

Régression linéaire multiple

La qualité de l'air est généralement influencée à des degrés différents par les variables que sont l'humidité, la vitesse du vent, les précipitations, le point de rosée, la visibilité, la pression au niveau de la mer et la température. L'objectif de cette recherche est d'identifier les variables qui influencent de manière déterminante l'indice de qualité de l'air dans la ville de Dakar pendant la période allant de 2010 à 2012. Les variables indiquant les moyennes journalières maximales (high) sont successivement présentées ci-dessous.

1. Y: indice de la qualité de l'air (nombre sans unité) ;
2. X_1: température maximale (en degré Celsius) ;
3. X_2: point de rosée (en degré Celsius) ;
4. X_3: humidité maximale (%) ;
5. X_4: pression maximale au niveau de la mer (hPa) ;

6. X_5: visibilité maximale (km) ;
7. X_6: vitesse maximale du vent (Km/h) ;
8. X_7: niveau de précipitation (mm).

Remarque 4.4.1. *La variable expliquée ou endogène est l'indice de la qualité de l'air ; iqa. Par contre toutes les autres variables sont des variables explicatives ou exogènes. Pour notre étude ces variables sont des données météorologiques. Nous avons obtenu leurs données historiques pendant la période allant de 2010 à 2012 grâce au site* `http://www.wunderground.com/history/airport/GOOY/2013/1/1/CustomHistory.html?dayend=31&monthend=12&yearend=2013&req_city=NA&req_state=NA&req_statename=NA`.

1. Y est la variable expliquée ou endogène.
2. Les variables explicatives ou exogènes sont $X_1, ..., X_7$.
3. Le modèle de régression devient [18] :

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_7 X_7 + \varepsilon \;\; ou \;\; simplement \;\; Y = \beta X + \varepsilon.$$

avec $\beta = (\beta_0, \beta_1, ..., \beta_7)$ sont les coefficients inconnus du modèle que l'on cherche à estimer et $\varepsilon = (\varepsilon_i)$ est le vecteur des erreurs indépendantes et identiquement distribuées (iid) suivant une loi $N(0, \sigma^2)$.

Le modèle de régression repose sur les hypothèses suivantes (sur les erreurs)ε_i :
- homoscédasticité: la variance σ^2 des erreurs est constante ;
- normalité: les erreurs sont distribuées selon une loi normale centrée ;
- indépendance: les résidus sont indépendants ou de manière équivalente non corrélées dans le cas gaussien.

Pour estimer de manière optimale les paramètres β_0,...,β_{p-1} du modèle de régression, nous allons utiliser la méthode des moindres carrés pour minimiser la quantité suivante:

$$\min_{\beta_0,...,\beta_{p-1}} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{i,1} + ... + \beta_{p-1} X_{i,p-1}))^2.$$

On s'intéresse à l'explication de la variable Y par les variables X_1, X_2, X_3, X_4, X_5, X_6, X_7.

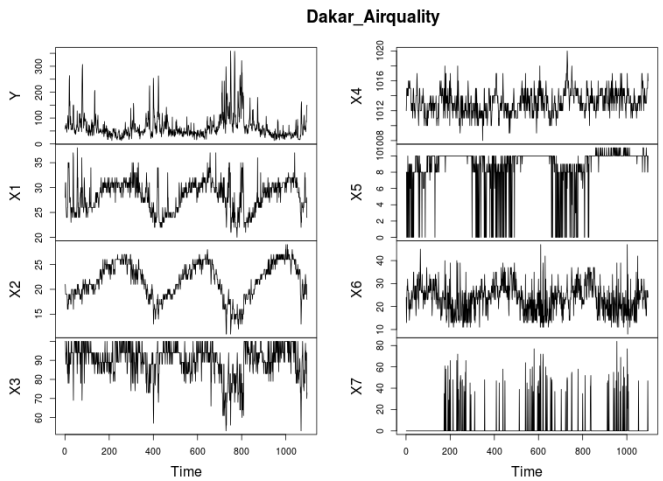


FIGURE 4.5 – Courbes des différentes séries temporelles.

Au regard de la base de données et de la figure (4.5), nous remarquons que les variables composant nos séries temporelles sont inégalement reparties et ne peuvent pas être représentées sur

une même échelle. Par exemple les valeurs prises par la variable X_4 sont très élevées par rapport aux valeurs des autres variables des différentes chroniques. La représentation graphique des séries laisse présager des mouvements accidentels, par conséquent l'issue envisageable est de passer chacune des chroniques au lissage. Pour ce faire nous faisons appel au filtre logarithmique. A cet effet, nous prenons le logarithme népérien de variable ne contenant pas de zéro.

Les valeurs des probabilités critiques nous renseignent qu'au seuil de 5%, il n'y a pas chaque jour, dans la ville de Dakar, de 2010 à 2012, interdépendance entre la qualité de l'air et la pluie, l'orage, la pluie et l'orage, la pluie et l'orage dans la ville de Dakar. Les résultats issus de ce test nous aident à retirer de l'estimation l'impact de tous ces événements.

Test de stationnarité

Afin de mettre en évidence des changements structurels, nous effectuerons deux tests de stationnarité à savoir, le test de Dickey et Fuller (DF) augmenté et le test de Phillips Perron PP.

DF	$logY$	$logX_1$	$logX_2$	$logX_3$	$logX_4$	X_5	$logX_6$	X_7
Statistique	-11.393	-10.194	-5.138	-12.932	-14.585	-19.818	-20.192	-28.906
P-value	0,000	0,0007	0,000	0,000	0,000	0,000	0,000	0,000

PP	$logY$	$logX_1$	$logX_2$	$logX_3$	$logX_4$	X_5	$logX_6$	X_7
Statistique	-10.900	-9.543	-3.927	-12.565	-14.708	-21.420	-21.613	-29.248
P-value	0,000	0,0007	0,0018	0,000	0,000	0,000	0,000	0,000

Nous constatons que pour le DF augmenté comme pour le PP, toutes les variables sont stationnaires.

4.4.2 Estimation des paramètres β_i

L'estimation des paramètres β_i du modèle se fait à partir des données observées dans un échantillon de taille n extrait de la population d'étude. Pour chaque sujet $i(i = 1, 2, \ldots, n)$ de l'échantillon, on observe le vecteur de dimension (p) de valeurs (Y_i, X_i), réalisation du vecteur de variables (Y_i, X_i), où $X_i = (X_{i1}, X_{i2}, ..., X_{i(p-1)})$ est le vecteur des valeurs des variables explicatives.

L'estimation du vecteur des paramètres de régression, noté $\beta = (\beta_0, \beta_1, \cdot, \beta_{p-1})$, se fait de la même manière par la méthode des moindres carrés comme pour la régression linéaire simple [38]. Dans le cadre d'une régression linéaire multiple à p (p=7 dans notre cas) variables explicatives, le critère des moindres carrés s'écrit :

$$\Phi(\beta) = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{i,1} + ... + \beta_{p-1} X_{i,p-1}))^2.$$

L'estimateur $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_{p-1})$ du vecteur $\beta = (\beta_0, \beta_1, \cdot, \beta_{p-1})$ est obtenu en minimisant le critère des moindres carrés $\Phi(\beta)$.

L'estimateur des moindres carrés $\hat{\beta}$ est défini comme suit :

$$\hat{\beta} = arg \min_{\beta \in \mathbb{R}^{p-1}} \sum_{i=1}^{n} (Y_i - \sum_{j=0}^{p-1} \beta_j X_{ij})^2 = arg \min_{\beta \in \mathbb{R}^{p-1}} \|Y - X\beta\|^2.$$

La matrice du plan d'expérience $X = [X_0|...|X_{p-1}]$ est formée de p vecteurs colonnes (la première colonne étant généralement constituée de 1). Le sous-espace engendré par les p vecteurs colonnes de X est appelé espace image, ou espace des solutions, et noté $\mathcal{M}(X)$ $(dim(\mathcal{M}(X)) = p)$ et tout vecteur de cet espace est de la forme $X\alpha$, où α est un vecteur de $\mathbb{R}^p$:

$$X\alpha = \alpha_0 X_0 + ... + \alpha_{p-1} X_{p-1}$$

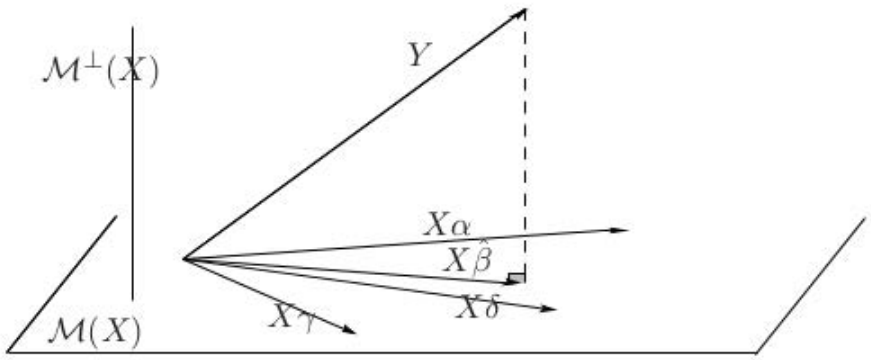


FIGURE 4.6 – Représentation de $\hat{X}\beta$ dans l'espace des variables.

Le vecteur Y est la somme d'un élément de $\mathcal{M}(X)$ et d'un bruit élément de $\mathbb{R}^n$, lequel n'a aucune raison d'appartenir à $\mathcal{M}(X)$. Minimiser $\|Y - X_\alpha\|^2$ revient à chercher un élément de $\mathcal{M}(X)$ qui soit le plus proche de Y au sens de la norme euclidienne classique. Cet unique élément est, par définition, le projeté orthogonal de Y sur $\mathcal{M}(X)$. Il sera noté $\hat{Y} = P_X Y$, où P_X est la matrice de projection orthogonale sur $\mathcal{M}(X)$. Il peut aussi s'écrire sous la forme $\hat{Y} = X\hat{\beta}$, où $\hat{\beta}$ est l'estimateur des MCO de β. L'espace orthogonal à $\mathcal{M}(X)$, noté $\mathcal{M}^{\perp}(X)$, est souvent appelé espace des résidus. En tant que supplémentaire orthogonal, il est de dimension

$n - p = dim(\mathcal{R}^n) - dim(\mathcal{M}(X))$.

Proposition 4.4.2. *L'estimateur $\hat{\beta}$ des Moindres Carrés Ordinaires a pour expression :*
$\hat{\beta} = (X^{'}X)^{-1}X^{'}Y$,
et la matrice P_X de projection orthogonale sur $(\mathcal{M})$ s'écrit :
$P_X = X(X^{'}X)^{-1}X^{'}$.

Le logiciel R nous fournit les valeurs des estimations $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_{p-1}$ qui sont des réalisations du vecteur aléatoire $\hat{\beta}$ (les valeurs de $\hat{\beta}$ sont lues sur la deuxième colonne Estimate Std du tableau ci-après) [39].

	Estimate Std	Error t	t value	$Pr(>\|t\|)$
$logX_1$	1.9078564	0.1684258	11.328	< 2e-16 ***
$logX_2$	-2.8269400	0.1287505	-21.957	< 2e-16 ***
$logX_3$	-0.1609299	0.1580903	-1.018	0.3089
$logX_4$	0.9963846	0.1275077	7.814	1.3e-14 ***
X_5	0.0082808	0.0041909	1.976	0.0484 *
$logX_6$	0.0239046	0.0535752	0.446	0.6555
X_7	-0.0016791	0.0009222	-1.821	0.0689.
Signif. codes:	0 '***' 0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' ' 1

Residual standard error: 0.3929 on 1089 degrees of freedom	—
F-statistic: 1.659e+04 on 7 and 1089 DF	$p - value :< 2.2e - 16$
Multiple R-squared: 0.9907	Adjusted R-squared: 0.9906

Nous avons une estimation avec un $R^2 = 0.9907$ et $R^2_{adj} = 0.9906$. Dans ce qui suivra nous allons faire une série de tests suivants pour justifier notre application de la méthode des moindres carrés ordinaires.

Test de Normalité des résidus

Pour l'étude de la normalité des résidus, faisons un test de normalité de Kolmogorov-Smirnov et puis un histogramme :
> ks.test(x,pnorm,mean(x),sd(x))
One-sample Kolmogorov-Smirnov test
data: x
D = 0.023, p-value = 0.6094
alternative hypothesis: two-sided

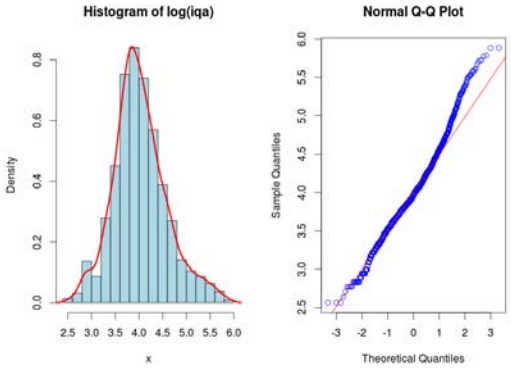


FIGURE 4.7 – Histogramme des résidus et $Q - Qplot$ de $log(iqa)$ [40]

Au regard de la P-value de ce test (p-value = 0.6094>0.05), de l'histogramme et de la courbe quantile des quantiles, on ne peut pas rejeter l'hypothèse nulle et on affirme qu'il y a la normalité des résidus.

Homoscédasticité:

Nous utilisons le test de Harrison-McCabe, qui a pour hypothèse nulle et les résidus suivent des lois normales de même variance. Le test effectué sous le logiciel R nous donne le résultat:
> hmctest(fit1)
Harrison-McCabe test
data: fit1
HMC = 0.486, p-value = 0.27

L'hypothèse nulle de ce test est : les résidus suivent des lois normales de même variance . Puisque la p-value =0.27> 0.05 alors on ne peut donc pas rejeter l'hypothèse nulle. On peut conclure que les résidus suivent des lois normales de même variance.

Test d'autocorrélation

Pour cela, nous effectuerons un test assez général ; c'est la statistique LM (Multiplicatur de Lagrange de Breusch et Godfrey) dont l'hypothèse nulle est l'absence d'autocorrélation [18, 41].

Lags (P)	Chi2	df	$prob > chi2$
1	322,697	1	0,0000

La P value nous permet de rejeter l'hypothèse nulle, c'est-à-dire, les résidus de différentes périodes sont autocorrélés. Pour palier à ce problème d'autocorrélation, nous ferons recours à la méthode itérative de Cochrane-Orcutt.

Résultat final par itération: la méthode de Cohrane-Orcutt

Après neuf (9) itérations avec le logiciel stata [41, 42], nous aboutissons au résultat ci-dessous : Prais-Winsten AR(1) regression – iterated estimates [39]

$logY$	coef	Std. Err.	t	$P > \|t\|$	[95% $int.\ confiance$]
$logX_1$	1,049889	0,1696522	6,19	0,000	0,717007 1,382771
$logX_2$	-1,979188	0,1664031	-11,89	0,000	-2,305695 -1,652681
$logX_3$	-0,4320999	0,1696388	-2,55	0, 011	-0,7649559 -0,099244
$logX_4$	1,200409	0,1237848	9,70	0,000	0,9575247 1,443292
X_5	0,0080417	0,036396	2,21	0,027	0,0009003 0,0151831
$logX_6$	0,0487099	0,0449114	1,08	0,278	-0,0394129 0,1368326
X_7	-0,0017805	0,0007329	-2,43	0,015	-0,0032187 -0,0003424

F(7, 1089) = 3962,77	Prob $> F = 0,0000$	—
$R^2\ normal = 0,9622$	$R^2 ajusté = 0,9620$	rho=0,5961765

Résultats de la Modélisation

La formule théorique de la régression multiple est

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_{p-1} X_{p-1} + \varepsilon \ \ ou \ \ soit \ \ Y = \beta X + \varepsilon.$$

Numériquement nous avons le résultat suivant :
$(logY_i) = 1,049889(logX_{1i}) - 1,979188(logX_{2i}) - 0,4320999(logX_{3i}) + 1,200409(logX_{4i}) + 0,0080417(X_{5i}) + 0,0487099(logX_{6i}) - 0,0017805(X_{7i})$.

Ce modèle prédit l'indice de la qualité de l'air (iqa) en fonction de la température, du point de rosée, de l'humidité, de la pression au niveau de la mer, de la visibilité, de la vitesse du vent et des précipitations. Les variables explicatives sont en majeur partie fortement significatives (p-value quasi-nulle), le coefficient de détermination est de 96,22%. Le modèle explique ainsi environ 96,20% de la variation de l'IQA. Cela signifie en d'autres termes que notre modèle a un pouvoir explicatif de 96,22%. Le modèle est très bon et réaliste car R^2 est assez proche de 1. C'est un très bon modèle d'ajustement des données. Plus R^2 est proche de 1, plus les données sont alignées sur la droite de régression.

4.4.3 Validation statistique du modèle

Présomption à la causalité

Dans notre modèle numérique, le coefficient de détermination, rapport entre la somme de carrée estimé sur la somme de carré total, montre que $96,22\%$ des nuages de points ajustent au mieux la variable expliquée. Autrement dit, on accorde 95% de crédit au fait que la qualité de l'air à Dakar trouve en grande partie $(96,20\%)$ son explication dans la température, l'humidité, la pression de rosé, la visibilité, le vent, la précipitation et la pression du niveau de mer.

Causalité globale

En effet, comme le coefficient de détermination donne une idée partielle sur la causalité [43], par conséquent, il faudrait vérifier est ce que cette présomption tapisse derrière ce coefficient par une statistique adéquate. C'est la statistique de Fisher. Elle se transforme de la manière suivante:

sous l'hypothèse nulle qu'aucune des variables n'explique de manière significative l'iqa, contre l'hypothèse alternative qu'au moins une des variables explique l'iqa. C'est un test bilatéral. Sous l'hypothèse nulle, on constate que la valeur de la probabilité critique est de $0,000$ et cela permet de réfuter l'hypothèse nulle. C'est-à-dire, qu'il existe au moins une des variables explicatives qui rend de manière significative compte de la qualité de l'air à Dakar. Afin, de ressortir l'impact significatif de chacune des variables explicatives sur la qualité de l'air, nous passerons aux tests de causalité individuelle captée par la statistique de Student.

Causalités individuelles : la statistique de Student

Afin de mettre en évidence l'incidence significative de chaque variable exogène sur la variable dépendante, Student a élaboré une statistique paramétrique pour le faire [44]. La valeur de cette statistique est donnée par le rapport en valeur absolue du coefficient estimé sur son écart type. Sous l'hypothèse nulle, le coefficient est différent de zéro. Soit β ce coefficient

$$\begin{cases} \beta = 0 & (Absence\ de\ causalité) \\ \beta \neq 0 & (Causalité) \end{cases}$$

La règle de décision se base soit sur la lecture du t, soit à l'une de valeur de la probabilité critique ou soit encore via la lecture de l'intervalle de confiance. Ainsi, l'hypothèse nulle est validée si et seulement si, la valeur de t est inférieure à 1,96 ; ou la valeur de la probabilité est inférieure à 5% (0,05) ou encore la valeur du coefficient estimé se trouve dans l'intervalle à 95% du seuil de confiance. En choisissant comme critère de décision la valeur de la probabilité critique, on constate que les variables température (temp), le point de rosée (ptrosé) et la pression au niveau de mer (slvpres) expliquent de manière significative l'indice de la qualité de l'air (iqa) à Dakar à 99%, tandis que la visibilité (visblt), l'humidité (hmdt) et les précipitations expliquent significativement et cela à 95% l'iqa au jour le jour de 2010 à 2012 à Dakar. Enfin, le vent (wind) explique de manière significative l'iqa à 72%.

4.4.4 Interprétation des résultats

Le résultat trouvé est globalement satisfaisant. L'interprétation des résultats nécessite de prime à bord certaines clarifications. Ainsi donc, une valeur très grande de l'iqa implique une mauvaise qualité de l'air. Cette remarque nous sera un puissant instrument d'aide lorsque seront interprétés les valeurs des coefficients estimés. Il faut le rappeler que le modèle tel que estimé est un modèle non stationnaire. Sa linéarisation par le filtre logarithme permet d'interpréter les résultats en termes d'élasticité. Enfin, nous allons pouvoir déceler laquelle des variables explicatives contribue le plus à l'explication de l'iqa. L'influence des différentes variables sur la qualité de l'air à Dakar, est captée par les effets marginaux [18].

$logY$	dy/dx	Z	$P > \|t\|$	[95% $int.\ confiance$]
$logX_1$	1,049889	6,19	0,000	0,717007 1,382771
$logX_2$	-1,979188	-11,89	0,000	-2,305695 -1,652681
$logX_3$	-0,4320999	-2,55	0, 011	-0,7649559 -0,099244
$logX_4$	1,200409	9,70	0,000	0,9575247 1,443292
$logX_5$	0,0487099	1,08	0,278	-0,0394129 0,1368326
X_6	0,0080417	2,21	0,027	0,0009003 0,0151831
X_7	-0,0017805	-2,43	0,015	-0,0032187 -0,0003424

A travers les effets marginaux, on note qu'au fur et à mesure que de la température augmente de 1%, l'indice de la qualité de l'air augmente de $1,05\%$, ce qui implique une détérioration de la qualité de l'air de 2%. Au final, la baisse de la température entraîne une amélioration de la qualité de l'air. Aussi une augmentation de la pression au niveau des mers (X_4) entraîne une mauvaise qualité de l'air. Une augmentation de 1% de la (X_4) engendre une augmentation de $1,2\%$ de l'iqa. Cependant, notons que les effets de la temp et de la slvpres sur l'iqa sont très prononcés sur l'iqa au regard de la p-value.

De même la visibilité et le vent expliquent négativement l'iqa, car une bonne visibilité (accroissement de 1%) conduit à une détérioration de la qualité de l'air (augmente de $0,008\%$). S'agissant particulièrement de la visibilité, mesurée en kilomètre, l'on peut retenir que plus l'on s'écarte de 1% en termes de distance, l'iqa augmente de $0,008\%$ et donc une mauvaise qualité de l'air. Un vent ayant une augmentation de la vitesse de 1% entre une augmentation de $0,05\%$ de l'iqa donc une détérioration de $0,05\%$. Leurs effets sont très peu prononcés sur l'iqa.

Par cette même analyse, on pourra conclure qu'un fort point de rosée améliore la qualité de l'air. Autrement dit, pour tout accroissement de point de rosée de 1% de degré Celsius, l'iqa

diminue sensiblement de 2%. Aussi, l'humidité contribue à améliorer la qualité de l'air. Bien que les précipitations n'expliquent significativement la qualité de l'air, elle contribue à l'améliorer.

Cette analyse ne nous permet pas de nous prononcer sur la proportion de la contribution de chaque variable à l'explication de l'iqa. Pour ce faire, nous recourrons à la bêta [41] régression afin de faire transparaître les proportionnalités.

	$logX_1$	$logX_2$	$logX_3$	$logX_4$	$logX_5$	X_6	X_7
β	0,3508	-0,8785	-0,0524	-0,0788	0,0003	0,0541	-0,0190

Le principe de comparaison est basé sur la valeur absolue des coefficients beta. Par conséquent, une valeur absolue de β très grande implique que la variable affectée au coefficient contribue le plus à l'explication du phénomène. D'après ce critère l'on peut déduire que le point de rosée contribue le plus à expliquer la qualité de l'air dans Dakar, suivie par la température, la pression du niveau de la mer, la visibilité, l'humidité, la précipitation et le vent.

4.4.5 Test du modèle

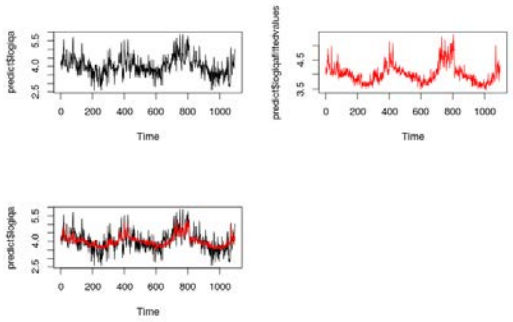
Dans cette section prédisons le logarithme néperien des valeurs de l'iqa pour les trente un (31) jours du mois de Janvier 2013 et comparons ces résultats avec les vraies mesures relevées au niveau des stations pendant cette même période.

$(logY_i) = 1,049889(logX_{1i}) - 1,979188(logX_{2i}) - 0,4320999(logX_{3i}) + 1,200409(logX_{4i}) + 0,0080417(X_{5i}) + 0,0487099(logX_{6i}) - 0,0017805(X_{7i})$.

	1erJan2013	2jan2013	3jan2013	4jan2013	5jan2013	6jan2013	7jan2013
$logY$ Predict	4.623918	4.351801	4.381915	4.324212	4.522027	4.485942	4.480862
$logY$ measured	4.624973	4.634729	4.454347	4.26268	4.836282	5.01728	4.394449
Error	0.001055	0.282928	0.072432	0.061532	0.314255	0.531338	0.086413
	8jan2013	9jan2013	10jan2013	11jan2013	12jan2013	13jan2013	14jan2013
$logY$ Predict	4.318746	4.310354	4.16586	4.176713	4.257765	4.069178	4.105231
$logY$ measured	5.594711	5.010635	4.75359	4.543295	4.110874	4.682131	4.394449
Error	1.275965	0.700281	0.58773	0.366582	0.146891	0.612953	0.289218
	15jan2013	6jan2013	17jan2013	18jan2013	19jan2013	20jan2013	21jan2013
$logY$ Predict	4.236575	4.546349	4.724899	4.306895	4.247306	4.026511	4.011349
$logY$ measured	4.430817	4.691348	5.062595	4.584967	4.094345	3.583519	4.343805
Error	0.194242	0.144999	0.337696	0.278072	0.152961	0.442992	0.332456
	22jan2013	23jan2013	24jan2013	25jan2013	26jan2013	27jan2013	28jan2013
$logY$ Predict	3.965532	3.933329	4.205845	4.116431	4.154939	4.167326	4.129025
$logY$ measured	4.248495	3.871201	4.430817	4.406719	4.905275	4.248495	4.060443
Error	0.282963	0.062128	0.224972	0.290288	0.750336	0.081169	0.068582
	29jan2013	30jan2013	31jan2013				
$logY$ Predict	4.64672	4.482949	4.456865				
$logY$ measured	4.158883	4.75359	5.26269				
Error	0.48784	0.270641	0.805825				

TABLE 4.1 – Prédiction de $log(iqa)$ et sa valeur mesurée en Janvier 2013

Les erreurs sont obtenues par la formule $Erreur = |logiqaPrédit - logiqa\ mesuré|$. Nous remarquons qu'on a une très bonne approximation car les erreurs sont très négligeables. La figure ci-dessous nous donne l'allure [45] des valeurs réelles et des valeurs ajustées du logiqa avec le logiciel R.



Cette étude nous a permis de déterminer un modèle de prédiction de l'indice de la qualité de l'air à Dakar. Cet modèle prédit l'IQA en fonction de la réponse aux changements prescrits par la météorologie et dans la source de pollution. Ce modèle peut servir à prédire l'indice de la qualité de l'air au moins durant la prochaine année et peut être réajusté au fur et à mesure de la disponibilité des données de l'année en cours. Il fournit un éventuel outil que le centre de surveillance de la qualité de l'air de Dakar pourrait exploiter dans une perspective de santé publique.

4.5 L'article en anglais

International Journal of Applied Mathematics and Statistics,
Int. J. Appl. Math. Stat.; Vol. 54; Issue No. 2; Year 2016,
ISSN 0973-1377 (Print), ISSN 0973-7545 (Online)

Modeling and Prediction of Dakar Air Quality Index.

Lebede Ngartera[1,2], Gueye Diagne Salimata[2] and Youssou Gningue[3]

[1]Department of Mathematics, Faculty of Exact and Applied Sciences
N'Djamena University, Chad
B.P:1739-N'djamena Chad
ngarteralebede12@gmail.com

[2]Department of Mathematics, Faculty of Science and Technology
Cheikh Anta Diop University, Dakar, Senegal
gueyesalli@yahoo.com

[3]Departement of Mathematics and Computer Sciences
Laurentian University, Sudbury,ON, Canada
ygningue@cs.laurentian.ca

ABSTRACT

Air quality plays a very important role in today's life in our societies. It is intrinsically linked to strategic policy regulation and improvement of human health and environment. In this article, we present the index of the air quality for Dakar during the period from 2010 to 2012 and propose a prediction model by the technique of multiple regressions.

Keywords: Air pollution, air quality, multiple regressions, optimization, least squares, modeling and prediction.

2010 Mathematics Subject Classification: 62J05, 62M10, 90C15, 91B76.

1 Introduction

A model is a representation of a simplified natural process used to represent and study a complex system. One of the problems of air pollution studies is to demonstrate to what extent pollutant emissions must be reduced in order for ambient concentrations to be maintained below value limits acceptable for health and the natural environment (Buchard V. and C., 2000). One of the major goals of modeling is to estimate the concentrations of pollutants in the atmosphere from Emissions of different sources of pollutants. A computer simulation is a valuable tool that makes it possible to confront the political and economic world to the risks or benefits incurred by an increase or a decrease of emission sources on air quality. Models of air quality can be divided into several categories (P., 1990; Corinne Schadkowski, 2002) :

1. **Physical** models: representations of certain phenomena (such as tunenel winds) on small scales, in the laboratory;

2. **Mathematical** models: analytical or numerical algorithms describe the physical and chemical aspects of the problem studied.

Physical models are often used to provide data to designers of mathematical models, such as chemical mechanisms developed from experiments in a simulation chamber.
Mathematical models can be classified in two major types:

- **Determinism** models: based on the mathematical description of atmospheric processes,
- **Statistical** models: semi-empirical relationships are established from a large number of observations.

In this research, we use statistical models including multiple linear regression (MLR). Speaking of regression, classical linear models are mentioned first: the simple linear regression (We express the concentration of a pollutant according to a single variable which may be the weather, a meteorological factor, etc.) or the Multiple linear regression (the pollutant concentration is expressed according to several explanatory variables). The MLR has been used in several studies, but in most cases it has been done so for comparison purposes. Goyal and al. (Goyal P., 2006) studied the forecasting of the total levels of breathable particles (daily averages) in two cities: Delhi and Hong Kong. The authors developed a first model based on MLR, the explanatory variables being meteorological parameters: wind speed, solar radiation, relative humidity and surface temperature. The MLR showed that the contributions of some variables were significant (wind speed, relative humidity), but solar radiation was not an influential variable. The model explains 58% of the variance of daily particle concentrations with a mean square error of 76 $\mu g.m^{-3}$, which is not very satisfactory. The second model used is of the ARIMA type, and the third, a linear combination of the first two. The MLR was used for the sake of comparison in other works, like those of Salini and Perez, Slini and al., or Robles-Diaz and al. (Perez P., 2008; Slini Th., 2006; Diaz-Roblès LA, 2008). Generally, different authors have retained the MLR as an interesting alternative because of its implementation simplicity, and leading to rather decent results but with lower performance than more sophisticated models. One particular way of applying the MLR is to use the principal components instead of the explanatory variables, when the latter are correlated with each other (which happens often), but in this research we will not present an example of application for the prediction of the quality index in this case. We propose a model of forecast by the multiple regression technique. To emphasize the significant impact of each explanatory variable on air quality, we use the student statistics to highlight the significant impact of each exogenous variable on the dependent variable. In addition, we predict the natural logarithm of the values of the index of air quality for thirty-one (31) days of January 2013 and compare these results with actual measurements taken in (weather) stations during the same period.

2 Modeling

This research is primarily concerned with the quality of the air in the city of Dakar. It is assessed through the index of the quality of air we model by the multiple linear regression technique.

2.1 Multiple Linear Regression

The quality of air is generally influenced to different degrees by variables such as humidity, wind speed, rainfall, dew point, visibility, the pressure at sea level and temperature. The objective of this research is to identify the variables that decisively influence the Index of air quality in the city of Dakar during the period from 2010 to 2012. The variables showing the maximum (high) daily averages are successively presented below.

1. Y: Index of air quality (number without units);
2. X_1: Maximum temperature (in Celsius degrees);
3. X_2: dew point (in Celsius degrees);
4. X_3: Maximum humidity (%);
5. X_4: maximum pressure at sea level (hPa);
6. X_5: maximum visibility (km);
7. X_6: maximum wind speed (Kph);
8. X_7: Level of precipitation (mm).

Remark 2.1. The dependent or endogenous variable is the air quality index; iqa. However, all other variables are explanatory or exogenous variables. For our study these variables are weather data. We obtained the historical data related to these for the period from 2010 to 2012 thanks to the site (http://www.wunderground.com/history/airport/GOOY)

1. Y is explained or endogenous variable.
2. The explanatory or exogenous variables are $X_1, ..., X_7$.
3. The regression model becomes (C. and S.Hadi, 2013) :

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_7 X_7 + \varepsilon \;\; or \;\; simply \;\; Y = \beta X + \varepsilon.$$

with $\beta = (\beta_0, \beta_1, ..., \beta_7)$ are the unknown coefficients of the model that seeks to estimate and $\varepsilon = (\varepsilon_i)$ is the vector of independent errors identically distributed (iid) according to a law $N(0, \sigma^2)$.
The regression model is based on the following assumptions (on errors) ε_i :

- Homoscedasticity: the variance σ^2 of errors is constant;
- Normality: the errors are distributed according to a centric (standard) normal distribution;
- Independence: the residues are independent or equivalently uncorrelated in the Gaussian case.

To optimally estimate the parameters β_0,...,β_{p-1} of the regression model, we will use the least square method to minimize the following quantity:

$$\min_{\beta_0,\ldots,\beta_{p-1}} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{i,1} + \ldots + \beta_{p-1} X_{i,p-1}))^2.$$

We are interested in the explanation of the variable Y by the variables X_1, X_2, X_3, X_4, X_5, X_6, X_7.

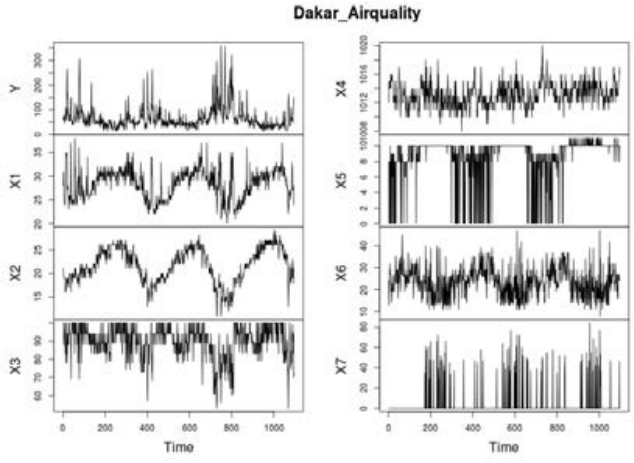


Figure 1: Curves of different time series.

In regard to the database and the following figure we notice that the variables making up our time series are unevenly distributed and cannot be represented on the same scale. For example, the values taken by the variable X_4 are very high compared to the values of the other variables of the different chronicles. The graphical representation of the series suggests accidental movements, therefore the possible outcome is to have each chronic undergo smoothing (shading). To do this we use the logarithmic filter. To this end, we take the natural logarithm of variables that do not contain zero.

The values of critical probabilities inform us that on the threshold of 5%; in the city of Dakar, from 2010 to 2012, there is no interdependence between the quality of air and rain, storn, rain and storm, rain and thunderstorm. The results of this test help us remove the impact of all such events from the assessment.

2.2 Test of stationarity

To highlight structural changes, we will make two stationary tests ie, the Dickey and Fuller (DF) enhanced test and the Phillips Perron (PP) test.

DF	$logY$	$logX_1$	$logX_2$	$logX_3$	$logX_4$	X_5	$logX_6$	X_7
Statistique	-11.393	-10.194	-5.138	-12.932	-14.585	-19.818	-20.192	-28.906
P-value	0,000	0,0007	0,000	0,000	0,000	0,000	0,000	0,000

PP	$logY$	$logX_1$	$logX_2$	$logX_3$	$logX_4$	X_5	$logX_6$	X_7
Statistique	-10.900	-9.543	-3.927	-12.565	-14.708	-21.420	-21.613	-29.248
P-value	0,000	0,0007	0,0018	0,000	0,000	0,000	0,000	0,000

We find that for the DF enhanced test as well as for the PP, all variables are stationary.

3 Estimation of parameters β_i

The estimation of the parameters β_i of the model is based on data observed in a sample of size n extracted from the study population. For each individual $i(i = 1, 2, \ldots, n)$ of the sample, we observe that the vector of dimension (p), and values (Y_i, X_i), carrying out the vector of variables (Y_i, X_i), wherein $X_i = (X_{i1}, X_{i2}, \ldots, X_{i(p-1)})$ is the vector of values of explanatory variables. The estimation of the regression parameter vector, noted $b = (b_0, b_1, \ldots, b_{p-1})$, is done in the same way by the least square method as for simple linear regression (Leisch, 2002). As part of a multiple linear regression to p (p = 7 in this case) explanatory variables, the criterion of least squares is:

$$\Phi(\beta) = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{i,1} + \ldots + \beta_{p-1} X_{i,p-1}))^2.$$

The estimator $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_{p-1})$ of the vector $b = (b_0, b_1, ..., b_{p-1})$ is obtained by minimizing the criterion of the least squares $\Phi(\beta)$.
The estimator of the least squares $\hat{\beta}$ is defined as follows:

$$\hat{\beta} = arg \min_{\beta \in \mathbb{R}^{p-1}} \sum_{i=1}^{n} (Y_i - \sum_{j=0}^{p-1} \beta_j X_{ij})^2 = arg \min_{\beta \in \mathbb{R}^{p-1}} \|Y - X\beta\|^2.$$

The experience plane array $X = [X_0|\ldots|X_{p-1}]$ is formed of p vectors columns (the first column being generally made up of I). The sub-space generated by the p vectors columns of X is called image space, or solution space and noted $\mathcal{M}(X)$ $(dim(\mathcal{M}(X)) = p)$ and any vector of this space is of the form $X\alpha$, where α is a vector of $\mathbb{R}^p$:

$$X\alpha = \alpha_0 X_0 + \ldots + \alpha_{p-1} X_{p-1}$$

The vector Y is the sum of an element of $\mathcal{M}(X)$ and a noise element $\mathbb{R}^n$, which has no

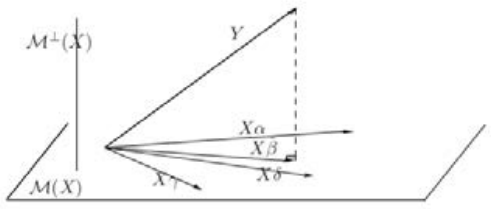


Figure 2: Representation $\hat{X}\beta$ in variable space.

reason to belong to $\mathcal{M}(X)$. Minimize $\|Y - X_\alpha\|^2$ like finding an element of $\mathcal{M}(X)$ which is closest to the direction Y classical Euclidean norm. This single element is, by definition, Y orthogonal projection of $\mathcal{M}(X)$. It will be noted $\hat{Y} = P_X Y$ where P_X is the orthogonal projection matrix $\mathcal{M}(X)$. It can also be written as $\hat{Y} = X\hat{\beta}$, where $\hat{\beta}$ is the MCO estimator of β. The space orthogonal to $\mathcal{M}(X)$, noted $\mathcal{M}^{\perp}(X)$, is often called residuals space. As an additional orthogonal, it has dimension $n - p = dim(\mathcal{R}^n) - dim(\mathcal{M}(X))$.

Proposition 3.1. *The $\hat{\beta}$ estimator Ordinary Least Square is expressed as:*
$\hat{\beta} = (X'X)^{-1}X'Y$,
and the orthogonal projection matrix $(\mathcal{M})$ is written:
$P_X = X(X'X)^{-1}X'$.

The R software provides us with estimates of values $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_{p-1}$ are realizations of the random vector β (β values which are read from the second column of the table Estimate Std below) (Cornillon and Matzner-Lober, 2010).

	Estimate Std	Error t	t value	$Pr(>\|t\|)$
$logX_1$	1.9078564	0.1684258	11.328	$< 2e-16 * **$
$logX_2$	-2.8269400	0.1287505	-21.957	$< 2e-16 * **$
$logX_3$	-0.1609299	0.1580903	-1.018	0.3089
$logX_4$	0.9963846	0.1275077	7.814	1.3e-14 ***
X_5	0.0082808	0.0041909	1.976	0.0484 *
$logX_6$	0.0239046	0.0535752	0.446	0.6555
X_7	-0.0016791	0.0009222	-1.821	0.0689.
Signif. codes:	0 *** 0.001 **	0.01 $*$	0.05 .	0.1 ' '1

Residual standard error: 0.3929 on 1089 degrees of freedom	—
F-statistic: 1.659e+04 on 7 and 1089 DF	$p - value :< 2.2e-16$
Multiple R-squared: 0.9907	Adjusted R-squared: 0.9906

We estimate with an $R^2 = 0.9907$ and $R^2_{adj} = 0.9906$. In what follows we will do a series of tests to justify our following application of the ordinary least square method.

Normality test residues

To study the normality of residuals, making a normality Kolmogorov-Smirnov and then a histogram:
ks.test(x,pnorm,mean(x),sd(x)) One-sample Kolmogorov-Smirnov test
data: x D = 0.023, p-value = 0.6094
alternative hypothesis: two-sided
Given the P-value of this test (p-value$= 0.6094 > 0.05$), histogram and quantile quantile curve (Davison, 2002), we cannot reject the null hypothesis and argues that there has normality of residuals.

Homoscedasticity:

We use the test Harrison-McCabe, whose null hypothesis and residues follow normal distributions with the same variance. The test carried out under the R software gives us the following result:
hmctest(fit1) data: fit1
Harrison-McCabe test HMC = 0.486, p-value = 0.27

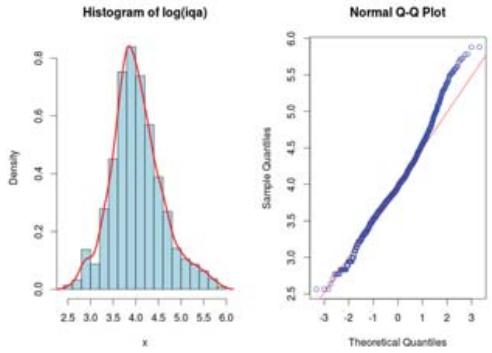


Figure 3: Residuals histogram and $Q - Qplot$ of $log(iqa)$

The null hypothesis of this test is: residues follow normal distributions with the same variance. Since the $p - value = 0.27 > 0.05$ we therefore cannot reject the null hypothesis. It can be concluded that the residues follow normal distributions with the same variance.

Autocorrelation test

For this, we will make a fairly general test; it is the LM statistic (Lagrange Multiplicatur Breusch and Godfrey) whose null hypothesis is the absence of autocorrelation (C. and S.Hadi, 2013; Cochrane and Orcutt, 1946).

Lags (P)	Chi2	df	$prob > chi2$
1	322,697	1	0,0000

The P value allows us to reject the null hypothesis, that is to say, the residues from different periods are autocorrelated. To overcome this problem autocorrelation, we will use the iterative method of Cochrane-Orcutt.

Final result per iteration: the method of Cohrane-Orcutt

After nine (09) iterations with the Stata software (Becketti, 2013), we arrive at result below:
Prais-Winsten AR(1) regression – iterated estimates (Cochrane and Orcutt, 1946)

$logY$	coef	Std. Err.	t	$P > \|t\|$	$[95\% \; intervalle \; de \; confi]$
$logX_1$	1,049889	0,1696522	6,19	0,000	0,717007 1,382771
$logX_2$	-1,979188	0,1664031	-11,89	0,000	-2,305695 -1,652681
$logX_3$	-0,4320999	0,1696388	-2,55	0, 011	-0,7649559 -0,099244
$logX_4$	1,200409	0,1237848	9,70	0,000	0,9575247 1,443292
X_5	0,0080417	0,036396	2,21	0,027	0,0009003 0,0151831
$logX_6$	0,0487099	0,0449114	1,08	0,278	-0,0394129 0,1368326
X_7	-0,0017805	0,0007329	-2,43	0,015	-0,0032187 -0,0003424

F(7, 1089) = 3962,77	Prob$> F = 0,0000$	—
$R^2 \; normal = 0,9622$	$R^2 ajusted = 0,9620$	rho=0,5961765

Results Modeling

The theoretical formula of multiple regression is

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_{p-1} X_{p-1} + \varepsilon \ \ or \ \ is \ \ Y = \beta X + \varepsilon.$$

Numerically we have the following result:
$(logY_i) = 1,049889(logX_{1i}) - 1,979188(logX_{2i}) - 0,4320999(logX_{3i}) + 1,200409(logX_{4i}) + 0,0080417(X_{5i}) + 0,0487099(logX_{6i}) - 0,0017805(X_{7i})$.
This model predicts the index of the air quality (iqa) as a function of the temperature, dew point, humidity, sea level press, visibility, wind speed and precipitation. The explanatory variables are highly significant in the major part (p-value close to zero), the determining factor is 96.22%. The model explains about 96.20% of the change in the iqa. This means in other words that our model has explanatory power of 96.22%. The model is very good and realistic because R^2 is fairly close to 1. This is a very good fit to the data model. More R^2 is to 1, the more data are aligned with the regression line.

4 Statistical Validation of the model

4.1 Presumption of Causality

In our numerical model, the coefficient of determination, ratio of the sum of squared estimated on the total sum of the square, shows that 96.22% of the point cloud best fit the dependent variable. In other words, 95% of credit granted the fact that air quality in Dakar is largely (96.20%) explained is by the temperature, humidity, pressure, dew, visibility, wind , precipitation and sea level press.

4.2 Global Causality

Indeed, as the coefficient of determination gives a partial idea of causality (Angrist, 1996), therefore, it should be verified that this presumption is lining behind this coefficient by adequate statistics. This is the Fisher statistic. It becomes as follows:
under the null hypothesis that none of the variables significantly explains iqa, against the alternative hypothesis that at least one of the variables explains iqa. It is a bilateral test. Under the null hypothesis, we find that the value of the critical probability is 0.000 and it helps disprove the null hypothesis. That is to say; that there is at least one of the variables that significantly accounted for the air quality in Dakar. In order to highlight the significant impact of each explanatory variable on the air quality, we will go to individual causality tests captured by the Student statistic.

4.3 Individual Causality: Statistics Student

To highlight the significant impact of each exogenous variable on the dependent variable, Student has developed a parametric statistics to do so (Granger, 1969). The value of this statistic is given by the ratio in absolute value of the estimated coefficient on its standard deviation. Under the null hypothesis, the coefficient is not zero.

$$\begin{cases} \beta = 0 & (No\ causal) \\ \beta \neq 0 & (Causality) \end{cases}$$

The decision is based on either the reading of t rule, either in terms of value of the critical probability or still through the reading of the confidence interval. Thus, the null hypothesis is validated if and only if the value of t is less than 1.96; or the probability value is less than 5% (0.05) or the value of the estimated coefficient is in the range of 95% confidence level. By choosing as a decision criterion the value of

the critical probability, we find that the temperature variables (temp), dew point (ptros) and pressure at sea level (slvpres) explain significantly the index of the air quality (iqa) in Dakar to 99%, while visibility (visblt), humidity (HMDT) and precipitation significantly and this explains the 95% iqa daily from 2010 to 2012 in Dakar. Finally, the wind explains significantly up to 72% the iqa.

5 Interpretation of results

The results are generally satisfactory. The interpretation of results is intended to premium onboard certain clarifications. Thus, a very high value of iqa implies poor air quality. This remark will be a powerful tool to help interpret the values of estimated coefficients. It must be remembered that the model is considered as a non-stationary model. Linearization by the logarithm filter enables interpreting the results in terms its linearization of elasticity. Finally, we can identify which variables contributes most to the explanation of the iqa. The influence of different variables on the air quality in Dakar, is captured by the marginal effects.

$logY$	dy/dx	Z	$P > \|t\|$	$[95\% \; intervalle \; de \; confi]$
$logX_1$	1,049889	6,19	0,000	0,717007 1,382771
$logX_2$	-1,979188	-11,89	0,000	-2,305695 -1,652681
$logX_3$	-0,4320999	-2,55	0, 011	-0,7649559 -0,099244
$logX_4$	1,200409	9,70	0,000	0,9575247 1,443292
$logX_5$	0,0487099	1,08	0,278	-0,0394129 0,1368326
X_6	0,0080417	2,21	0,027	0,0009003 0,0151831
X_7	-0,0017805	-2,43	0,015	-0,0032187 -0,0003424

Through the marginal effects, it is to remember that: as as the temperature increases by 1%, the index of the quality of the air increases to 1.049889% or 1.05%, which implies a deterioration of the air quality of 2%. Finally, the drop in temperature causes an improvement in air quality. Also an increase in the pressure at sea level (X_4) leads to poor air quality. An increase of 1% of the (X 4) generates an increase of 1.2% of iqa. However, it is to be noted that the effects of temp and slvpres on iqa are very pronounced on iqa under the p-value. Similarly, the visibility, and wind negatively explain the iqa because good visibility (increase of 1%) leads to a deterioration in air quality (increased by 0.008%). With specific regard to visibility, measured in kilometers, we can say that the more one moves away from 1% in terms of distance, the iqa increases of 0.008% and therefore poor air quality. A wind having a speed increase between a 1% increase 0.0487099% or 0.05% of iqa therefore a deterioration of 0.05%. Their effects are very pronounced on iqa. In this same analysis, we can conclude that a high dew point improves air quality. In other words, for every dew point increase 1% celsus degree the iqa decreases 1.979188% is substantially 2%. Also, the moisture helps to improve the air quality. Although rainfall is significantly explain the air quality, it helps to improve it. This analysis does not allow us to comment on the proportion the contribution of each variable to the explanation of the iqa. To do this, we will use the Beta regression (Becketti, 2013) in order to transpire proportionalities.

	$logX_1$	$logX_2$	$logX_3$	$logX_4$	$logX_5$	X_6	X_7
beta	0,3508323	-0,8784867	-0,0524051	-0,0788565	0,0003116	0, 0540706	-0,0189856

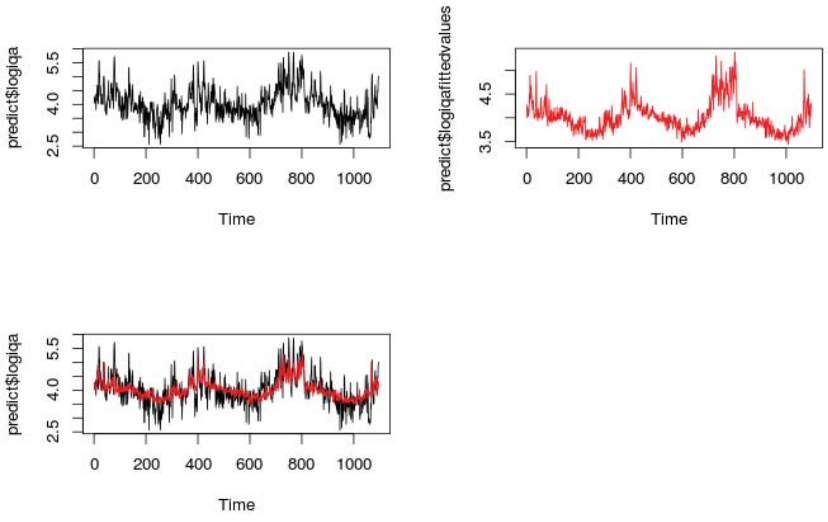
The comparison is based on the absolute value of beta coefficients. Therefore, an absolute value of high beta implies that the variable assigned to the factor contributing most to explaining the phe enon. On aps this criterion we can see that the dew point contributes most to explain the air quality in Dakar, followed by temperature, pressure sea level, visibility, humidity, precipitation and wind.

6 Test model (significance tests)

In this section predict the natural logarithm of the iqa values for (31) days of January 2013 and compare these results with real measurements taken at stations during the same period.
$(logY_i) = 1,049889(logX_{1i}) - 1,979188(logX_{2i}) - 0,4320999(logX_{3i}) + 1,200409(logX_{4i}) + 0,0080417(X_{5i}) + 0,0487099(logX_{6i}) - 0,0017805(X_{7i})$.

	1erJan2013	**2jan2013**	**3jan2013**	**4jan2013**	**5jan2013**	**6jan2013**	**7jan2013**
$logY$ Predict	4.623918	4.351801	4.381915	4.324212	4.522027	4.485942	4.480862
$logY$ measured	4.624973	4.634729	4.454347	4.26268	4.836282	5.01728	4.394449
Error	0.001055	0.282928	0.072432	0.061532	0.314255	0.531338	0.086413
	8jan2013	**9jan2013**	**10jan2013**	**11jan2013**	**12jan2013**	**13jan2013**	**14jan2013**
$logY$ Predict	4.318746	4.310354	4.16586	4.176713	4.257765	4.069178	4.105231
$logY$ measured	5.594711	5.010635	4.75359	4.543295	4.110874	4.682131	4.394449
Error	1.275965	0.700281	0.58773	0.366582	0.146891	0.612953	0.289218
	15jan2013	**16jan2013**	**17jan2013**	**18jan2013**	**19jan2013**	**20jan2013**	**21jan2013**
$logY$ Predict	4.236575	4.546349	4.724899	4.306895	4.247306	4.026511	4.011349
$logY$ measured	4.430817	4.691348	5.062595	4.584967	4.094345	3.583519	4.343805
Error	0.194242	0.144999	0.337696	0.278072	0.152961	0.442992	0.332456
	22jan2013	**23jan2013**	**24jan2013**	**25jan2013**	**26jan2013**	**27jan2013**	**28jan2013**
$logY$ Predict	3.965532	3.933329	4.205845	4.116431	4.154939	4.167326	4.129025
$logY$ measured	4.248495	3.871201	4.430817	4.406719	4.905275	4.248495	4.060443
Error	0.282963	0.062128	0.224972	0.290288	0.750336	0.081169	0.068582
	29jan2013	**30jan2013**	**31jan2013**				
$logY$ Predict	4.64672	4.482949	4.456865				
$logY$ measured	4.158883	4.75359	5.26269				
Error	0.48784	0.270641	0.805825				

Errors are obtained by the formula error$= |logiqaPredict - logiqameasueed|$. We note that we have a very good approximation since errors are very significant. The figure below gives us the shape of the actual values and the adequate adjusted logiqa values (F., 1973).



7 Conclusion and outlook

This study allowed us to determine a model for predicting for the Dakar air quality index. This model predicts the air quality index depending on the response to the changes required by the meteorology and the pollution source. This model can be used to predict the air quality index at least during the next year and can be adjusted as and extent of data availability of the current year. It provides a tool that the quality monitoring center air Dakar could exploit a public health perspective. Note that the consideration of a linear regression model necessarily diminish its sharpness and degree of prediction. The study conducted under this research could be improved by nonlinear regression model approach. We have introduced an approach by the PROCESS AUTO REGRESSIVE MOVING AVERAGE ARMA(2,1) method (L. Ngartera and Gningue, 2015). We plan the comparison with the approach proposed in this article that could be conducted to take advantage of the two forms.

Acknowledgment

We thank God the Almighty author of perfect knowledge to have attracted to men the techniques of modeling, simulation and prediction mathematical. A great thank to all those who work today for the valuation of numerical methods worldwide. We give thank to anonymous reviewer of journal for their suggestions to improve the publication of this paper. My thanks also go to professor Christopher Thron from the Texas A & M University Central Texas USA for his advice and guidance.

References

Angrist, J.D., I. G. R. D. 1996. Identification of causal effects using instrumental variables (with comments), *Journal of the American Statistical Association* **91**: 434–444.

Becketti, S. 2013. Introduction to time series using stata, *College Station, TX: Stata Press* .

Buchard V., Helfer P., M. P. and C., M. 2000. Etude de la pollution atmosphrique transfrontalire, *Springer* **110**.

C., S. and S.Hadi, A. 2013. Rgression analysis by exemple, *John Weley et Sons* .

Cochrane, D. and Orcutt, G. H. 1946. Application of least squares regression to relationships containing auto-correlated error terms, *Journal of the American Statistical Association* **44**: 32–61.

Corinne Schadkowski, Jean Claude Dechau, V. N. Y. F. 2002. Introduction la modlisation de la qualit de l'air, *Air Pur* pp. 5–8.

Cornillon, P.-A. and Matzner-Lober, E. 2010. Rgression avec r, *Springer* .

Davison, A. C. e. D. K. 2002. "an introduction to the bootstrap with applications in r. statistical computing and statistical graphics, *Newsletter* **13**: 6–11.

Diaz-Roblès LA, O. J. F. e. a. 2008. A hybrid arima and artificiel neural networks model to forcast particulate matter in urbain areas of temuco, chile, *Atmospheric Environment, 42: 8331-40.* .

F., A. 1973. Graphs in statistical analysis, *The American Statistician* **27, n1**: 17–21.

Goyal P., C. A. . J. N. 2006. Statistical models for the prediction of respirable suspended particulate matter in urban cities, *Atmospheric Evironment,40,2068-2077* **40**: 2068–2077.

Granger, C. 1969. Investigating causal relations by econometric models and cross-spectral methods, **37, No 3,July**: 424–438.

L. Ngartera, S. D. and Gningue, Y. 2015. Optimization and analysis of the index of air quality in dakar by the process auto regressive moving average arma(2,1), *International Journal of Applied Mathematics (IJAM)* **28, No 5**: 621–636.

Leisch, F. 2002. Sweave, part i: Mixing r and latex, *R News* **2 (3)**: 28–31.

P., Z. 1990. Air pollution modeling theories, computational methods and available software, *Van Nostrand Reinhold, New York,* .

Perez P., S. G. 2008. $Pm_{2.5}$ forecasting in a large city: Comparaison of three methods., *Atmospheric Environment* **42: 8219-24**.

Slini Th., Karatzas K, M. N. 2006. Pm_{10} forecasting for thessaloniki, greece, *Environmental Modelling & Software, 21:559-65.* .

Chapitre 5

Optimisation et analyse de l'indice de la qualité de l'air dans Dakar par le processus $ARMA(2,1)$.

Sommaire

5.1 Introduction

Dans ce chapitre on introduit le processus $ARMA$ dans sa généralité avec quelques outils nécessaires pour l'étude des processus stationnaires. On présente deux algorithmes pour la détermination du prédicateur. La démarche de modélisation et prédiction par les processus $ARMA$

est abordée. Le chapitre se termine par un article 5.6 qui retient le processus $ARMA(2,1)$ pour prédire l'indice de la qualité de l'air dans Dakar. Dans le section suivante, on commence par les processus stationnaires d'ordre deux.

Processus stochastique stationnaire

Définition 5.1.1. ***(Processus stochastique)***

On considère une grandeur donnée sur des dates de 1 *à* T *. On fait des observations* $x_1, \ldots, x_T$*, réalisations des variables aléatoires* $X_1, \ldots, X_T : (\Omega, \mathcal{A}, P) \longrightarrow \mathbb{R}, \quad \omega \in \Omega$ *est un état de la nature tel que* $x_t = X_t(\omega)$*.*

$(X_t)_{t\in\mathbb{Z}}$ *est appelé un* ***processus stochastique*** *et* $(x_t)_{t\in\mathbb{Z}}$ *et* ***trajectoire du processus*** $(X_t)_{t\in\mathbb{Z}}$ *.*

Si $E(X_t) = m_t$, on a une seule observation (x_t en l'occurrence) pour estimer m_t . En revanche si pour tout $t \in \mathbb{Z}, E(X_t) = m$, on peut estimer m par $\hat{m} = \frac{1}{T}\sum_{t=1}^{T} X_t$.

C'est donc nécessaire de supposer que la suite X_t a certaines propriétés de régularité.

Dans la suite de ce chapitre, on considère $(X_t)_{t\in\mathbb{Z}}$ et on suppose $X_t \in \mathcal{L}^2(\Omega, \mathcal{A}, P), \forall t \in \mathbb{Z}$.

Définition 5.1.2. ***Stationnarité stricte ou forte***

$(X_t)_{t\in Z}$ *est dit* ***processus stationnaire au sens strict*** *si :* $\forall n \in \mathbb{N}, \forall (t_1, \ldots, t_n), \forall h \in \mathbb{Z}$*, la loi de* $(X_{t_1}, \ldots, X_{t_n})$ *est identique à la loi de* $(X_{t_{1+h}}, \ldots, X_{t_{n+h}})$

Theorem 5.1.3. ***Théorème de Kolmogorov***

$(Xt)_{t\in\mathbb{Z}}$ *est dit* ***processus stationnaire au sens strict*** *si et seulement si la loi de* $(X_t)_{t\in\mathbb{Z}}$ *est identique à la loi de* $(Y_t)_{t\in\mathbb{Z}}$ *où* $Y_t = X_{t+h}$ *.*

Définition 5.1.4. ***Stationnarité faible***

$(X_t)_{t\in\mathbb{Z}}$ *est un* ***processus stationnaire du second ordre*** *(ou* ***processus faiblement stationnaire****) s'il remplit les 3 conditions suivantes:*

(i) $\forall t \in \mathbb{Z}, E(X_t) = m$

(ii) $\forall t\mathbb{Z}, var(X_t) = \sigma^2 = \gamma(0)$

(iii) $\forall t \in \mathbb{Z}, \forall h \in \mathbb{Z}, cov(X_t, X_{t+h}) = \gamma(h)$ *(ne dépend que de h)*

$\gamma(h)$ *est l'****auto-covariance*** *d'ordre* h *de* X_t *.*

Remarque 5.1.5.

- *Dans la suite de ce chapitre de la thèse, les processus stationnaires désignent les processus de la définition 5.1.4 ;*
- $(iii) \Longrightarrow (ii)$ *pour* $h = 0$ *et* $\gamma(0) = \sigma^2$;
- *Si un processus est stationnaire au sens strict alors il est faiblement stationnaire ;*
- *Si* $(X_t)_{t\in\mathbb{Z}}$ *est un processus gaussien alors il y a équivalence entre stationnarité faible et forte ;*
- $E\begin{pmatrix} X_{t_1} \\ \vdots \\ X_{t_n} \end{pmatrix} = \begin{pmatrix} m \\ \vdots \\ m \end{pmatrix} \quad var\begin{pmatrix} X_{t_1} \\ \vdots \\ X_{t_n} \end{pmatrix} = \begin{pmatrix} \gamma(0) & & \gamma(t_j - t_i) \\ & \ddots & \\ & & \gamma(0) \end{pmatrix}$

Exemple 5.1.6. ***Processus stationnaire***

(i) ***Bruit blanc faible (white noise)****:* $(\epsilon_t)_{t\in\mathbb{Z}}$ *est un bruit blanc faible si et seulement si:*

$$E(\epsilon_t) = 0, \forall t \in \mathbb{Z}$$
$$var(\epsilon_t) = \sigma^2, \forall t \in \mathbb{Z}$$
$$cov(\epsilon_t, \epsilon_\tau) = 0, \quad si \quad t \neq \tau$$

on note pour la suite de ce chapitre $\epsilon_t \rightsquigarrow \mathcal{BB}(0, \sigma^2)$

(ii) $(\epsilon_t)_{t\in\mathcal{Z}}$ *est un* ***bruit blanc fort*** *si et seulement si les* ϵ_t *sont i.i.d.,* $E(\epsilon_t) = 0$ *et* $var(\epsilon_t) = \sigma^2$*.*

(iii) ***Processus moyenne mobile*** *d'ordre 1, noté* $MA(1)$ *(moving average of order 1)*

Soient $\theta \in \mathbb{R}^*$, $\epsilon_t \rightsquigarrow \mathcal{BB}(0, \sigma^2)$*. Le processus* $(X_t)_{t\in\mathbb{Z}}$ *défini par:* $\forall t \in \mathbb{Z}, X_t = \epsilon_t - \theta\epsilon_{t-1}$ *est un processus stationnaire appelé moyenne mobile d'ordre 1 et on note* $X_t \rightsquigarrow MA(1)$*.*

Remarque 5.1.7. *Dans la pratique, on ne fait pas la différence entre* x_t *et* X_t *.* (x_t) *ou* (X_t) *désigne toujours le processus et* $x_1, \ldots, x_T$ *ou* $X_1, \ldots, X_T$ *la suite des observations.*

Exemple de processus non stationnaires

Nous présentons ici sous forme d'un exemple un processus non stationnaire.

Exemple 5.1.8. ***Processus non stationnaires***

(i) ***Marche aléatoire (randon walk)***

Soit $\epsilon_t \rightsquigarrow \mathcal{BB}(0,\sigma^2)$. $(X_t)_{t\in\mathbb{Z}}$ est une ***marche aléatoire sans dérive*** *si et seulement si*
- $X_t = X_{t-1} + \epsilon_t, \forall t \geq 0$
- $cov(\epsilon_t, X_{t-k}) = 0, \forall 0 < k \leq t$.

Même si on a la propriété $EX_t = EX_{t-1} \Longrightarrow EX_t = m, \forall t \in \mathbb{Z}$, $(X_t)_{t\in\mathbb{Z}}$ n'est pas stationnaire :

$$\left.\begin{array}{ll} X_t & = X_{t-1} + \epsilon_t \\ X_{t-1} & = X_{t-2} + \epsilon_{t-1} \\ & \vdots \\ X_1 & = X_0 + \epsilon_1 \end{array}\right\} \Longrightarrow X_t = X_0 + \sum_{k=1}^{t} \epsilon_k$$

D'où

$$\begin{aligned} var(X_t) &= var(X_0) + 2\sum_{k=1}^{t} cov(\epsilon_t, X_0) + var(\sum_{k=1}^{t} \epsilon_k) \\ &= var(X_0) + t\sigma^2 \end{aligned}$$

Et par conséquent le processus n'est pas stationnaire en variance.

(ii) *Processus stationnaire autour d'un trend déterministe.*

$X_t = a + bt + Y_t$ où $(Y_t)_{t\in\mathbb{Z}}$ est un processus stationnaire.

Par exemple si $Y_t = \epsilon_t \rightsquigarrow \mathcal{BB}(0,\sigma^2), EX_t = a + bt$, le processus n'est pas stationnaire en espérance.

Fonction d'auto-covariance

Définition 5.1.9. ***Fonction d'auto-covariance*** *L'auto-covariance d'un processus stationnaire $(X_t)_{t\in\mathbb{Z}}$ est définie par:*

$$\begin{array}{rll} \gamma : & \mathbb{Z} & \longrightarrow \mathbb{R} \\ & h & \longmapsto \gamma(h) = cov(X_t, *X_{t-h}) \end{array}$$

Proposition 5.1.10. *Soit γ la fonction d'auto-covariance d'un processus stationnaire $(X_t)_{t\in\mathbb{Z}}$ alors γ est une fonction paire de type positif ie $\forall n \in \mathbb{N}, \forall (t_1,\ldots,t_n), \forall (a_1,\ldots,a_n) \in \mathbb{R}$*

$$\sum_{1\leq i\leq j} a_i a_j \gamma(t_i - t_j) > 0$$

Démonstration. Nous faisons cette preuve en deux partie:

1. Parité de γ :

$$\begin{aligned} \gamma(h) &= cov(_t, X_{t-h}) = cov(X_{t-h}, X_{(t-h)+h}) = cov(X_{t-h}, X_t) \\ &= cov(X_t, X_t - h) = \gamma(-h) \;\; d'\text{où } la \text{ parité de } \gamma \end{aligned}$$

2. Positivité :

$$\begin{aligned} var(\sum a_i X_{t_i}) &= cov(\sum_i a_i X_{t_i}, \sum_j a_j X_{t_j}) \\ &= \sum_{i,j} a_i a_j cov(X_{t_i}, X_{t_j}) \\ &= \sum_{i,j} a_i a_j \gamma(t_i - t_j) \geq 0 \end{aligned}$$

□

Fonction d'auto-corrélation

Définition 5.1.11. ***Fonction d'auto-corrélation***
La fonction d'auto-corrélation d'un processus stationnaire $(X_t)_{t\in\mathbb{Z}}$ est définie par :

$$\forall h \in \mathbb{Z}, \rho(h) = \frac{\gamma(h)}{\gamma(0)} = corr(X_t, X_{t+h})$$

L'autocorrélation partielle d'ordre k désigne la corrélation entre X_t et X_{t-k} obtenue lorsque l'influence des variables X_{t-k-i}, avec $i < k$, a été retirée.

On appelle variogramme (resp. corrélogramme), le graphe d'une fonction d'autocovariance (resp. d'autocorrélation). De même, on définie un corrélogramme partielle comme étant le graphe de la fonction d'autocorrélation partielle.

Proposition 5.1.12. *$\rho : h \longmapsto \rho(h)$ est une fonction paire de type positif à valeur dans $]-1;1.[$*

Démonstration. La définition même de ρ garantie la preuve de la proposition (5.1.12). En effet:

$$corr(X_t, X_{t+h}) = \frac{cov(X_t, X_t + h)}{\sigma_{X_t}\sigma_{X_{t+h}}} = \frac{\gamma(h)}{\gamma(0)} = \rho(h)$$

Comme γ est paire de type positif alors ρ l'est aussi. □

Définition 5.1.13. ***Auto-corrélogramme théorique*** *L'auto-corrélogramme de $(X_t)_{t\in\mathbb{Z}}$ est le graphe de la fonction:*

$$\begin{cases} \mathbb{N} \longrightarrow]-1;1[\\ h \longmapsto \rho(h) \end{cases}.$$

Processus stationnaire transformé par une moyenne mobile infinie

Définition 5.1.14. ***Proposition***
On considère $(X_t)_{t\in\mathbb{Z}}$ un processus stationnaire et $(a_j)_{j\in\mathbb{Z}}$ une suite de réels tels que $\sum_j |a_j| < +\infty$. Alors $Y_t = \sum_{j\in\mathbb{Z}} a_j X_{t-j}$ est défini presque surement (p.s.) pour tout t.
On a les propriétés suivantes :

1. *$Y_t \in \mathcal{L}^2(\Omega, \mathcal{A}, P), \forall t \in \mathbb{Z}$*
2. *$(Y_t)_{t\in\mathbb{Z}}$ est un processus stationnaire tel que $EY_t = m_Y = \left(\sum_{j\in\mathbb{Z}} a_j\right) mX$*
   $\gamma_Y(h) = \sum_{j,k} a_j a_k \gamma(h+k-j) = \sum_{j,k} a_j a_k \gamma(h+j-k), \forall h \in \mathbb{Z}$
   On dit que $(Y_t)_{t\in\mathbb{Z}}$ est la ***transformée de*** *$(X_t)_{t\in\mathbb{Z}}$* ***par la moyenne mobile infinie associée aux*** *$(a_j)_{j\in\mathbb{Z}}$.*

Régression linéaire sur un nombre fini de retards

Définition 5.1.15.

Soit $(X_t)_{t\in\mathbb{Z}}$ un processus stationnaire.

(i) La **régression linéaire théorique** de X_t sur $X_{t-1}, \ldots, X_{t-p}$ est la projection orthogonale dans $\mathcal{L}^2(\Omega, \mathcal{A}, P)$ de X_t sur $H = Vect(X_{t-1}, \ldots, X_{t-p})$.
On note généralement $EL(X_t | X_{t-1}, \ldots, X_{t-p})$ la régression linéaire théorique de X_t sur $X_{t-1}, \ldots, X_{t-p}$.

(ii) La **régression affine théorique** de X_t sur $X_{t-1}, \ldots, X_{t-p}$ est la projection orthogonale dans $\mathcal{L}^2(\Omega, \mathcal{A}, P)$ de X_t sur $H^* = Vect(1, X_{t-1}, \ldots, X_{t-p})$.
On note généralement $EL(Xt | 1, X_{t-1}, \ldots, X_{t-p})$ la régression affine théorique de X_t sur $X_{t-1}, \ldots, X_{t-p}$.

Proposition 5.1.16. *(i) coïncide (ii) si et seulement si $EX_t = 0$.*

Remarque 5.1.17. *1. Si $EX_t = 0$, on calculera toujours la régression affine. On la note aussi souvent $EL(X_t|X_{t-1},\dots,X_{t-p})$.*

2. $Vect(X_t|X_{t-1},\dots,X_{t-p})$ et $Vect(Xt|X_{t-1},\dots,X_{t-p})$ sont des sous espace vectoriel de dimension finie de $\mathcal{L}^2$ donc fermés (d'après le corollaire 3.2.15).

3. Si $(X_t)_t$ est gaussien, alors $EL(X_t|.) = E(X_t|.)$.

On peut calculer la régression affine théorique (ii). En effet :

$$H = Vect(1, X_{t-1},\dots,X_{t-p}) \text{ et } X_t^* = p_H(X_t) \text{ est caractérisé par } X_t^* \in H \text{ et } X_t - X_t^* \perp H.$$

$$X_t^* \in H \Longleftrightarrow \exists a_0, a_1,\dots,a_p / X_t^* = a_0 + \sum_{j=1}^{p} a_j X_{t-j}$$

L'existence de X_t^* est donnée par le théorème de la projection (3.3.6).

$X_t - X_t^* \perp H$ après un petit effort à nous conduit à

$$\begin{cases} a_0 &= m_X(1 - \sum_{j=1}^{p} a_j) \\ E(X_{t-k}X_{t-j}) &= m_X^2 + \sum_{k=1}^{p} a_k[E(X_tX_{t-k}) - m_X^2] \quad \forall j = 1,\dots,p. \end{cases}$$

On a alors

$$\forall j = 1,\dots,p \quad E(X_tX_{t-j}) - m_x^2 = \sum_{k=1}^{p} a_k[E(X_{t-k}X_{t-j}) - m_X^2]$$

Soit

$$\forall j = 1,\dots,p \quad cov(X_t, X_{t-j}) = \sum_{k=1}^{p} a_k cov(X_{t-k+j}, X_{t-j})$$

Ce qui fournit

$$\gamma(j) = \sum_{k=1}^{p} a_k \gamma(k-j)$$

Et

$$a_0 = m_X(1 - \sum_{j=1}^{p} a_j)$$

Il vient donc

$$\begin{pmatrix} 1 & \gamma(1) & \dots & \gamma(p-1) \\ \gamma(1) & 1 & \dots & \gamma(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(p-1) & \gamma(p-2) & \dots & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} \gamma(1) \\ \vdots \\ \gamma(p) \end{pmatrix} \tag{5.1}$$

Puisque $\rho(h) = \frac{\gamma(h)}{\gamma(0)}$, divisons l'équation (5.1) par $\gamma(0)$ et on a:

$$\begin{pmatrix} 1 & \rho(1) & \dots & \rho(p-1) \\ \rho(1) & 1 & \dots & \rho(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \rho(p-1) & \rho(p-2) & \dots & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} \rho(1) \\ \vdots \\ \rho(p) \end{pmatrix} \tag{5.2}$$

Si les X_t sont indépendants alors la matrice du premier terme de l'équation (5.2) est inversible.

Définition 5.1.18. ***Propriété***

On appelle auto-corrélation partielle d'ordre p l'expression suivante:

$$\begin{aligned} r(p) &= Corr(X_t - EL(X_t|X_{t-1},\dots,X_{t-p+1}), X_{t-p} - EL(X_{t-p}|X_{t-1},\dots,X_{t-p+1})) \\ &= \frac{cov(X_t - EL(X_t|X_{t-1},\dots,X_{t-p+1}), X_{t-p} - EL(X_{t-p}|X_{t-1},\dots,X_{t-p+1}))}{\sqrt{var(X_t - EL(X_t|X_{t-1},\dots,X_{t-p+1}))var(X_{t-p} - EL(X_{t-p}|X_{t-1},\dots,X_{t-p+1}))}} \end{aligned}$$

On montre que $r(p) = a_p$ coefficient [66] de X_{t-p} dans $EL(X_t|X_{t-1},\ldots,X_{t-p})$.

$$EL(X_{t-p}|X_{t-1},\ldots,X_{t-p+1}) = \sum_{j=1}^{p} a_j X_{t-j} \tag{5.3}$$

Définition 5.1.19. ***Auto-corrélogramme partiel***
L'auto-corrélogramme partiel de $(X_t)_{t\in\mathbb{Z}}$ est le graphe de la fonction:

$$\begin{cases} \mathbb{N} \longrightarrow]-1;1[\\ p \longmapsto r(p) \end{cases}.$$

Régression sur un nombre infini de retards

Définition 5.1.20. *Soit $(X_t)_{t\in\mathbb{Z}}$ un processus stationnaire.*

(i) *La **régression linéaire théorique** de X_t sur $X_{t-1},\ldots,X_{t-p},\ldots$ est la projection orthogonale dans $\mathcal{L}^2(\Omega,\mathcal{A},P)$ de X_t sur $H = Vect(X_{t-1},\ldots,X_{t-p},\ldots)$.*

(ii) *La **régression affine théorique** de X_t sur $1, X_{t-1},\ldots,X_{t-p},\ldots$ est la projection orthogonale $\mathcal{L}^2(\Omega,\mathcal{A},P)$ de X_t sur $H^* = Vect(1, X_{t-1},\ldots,X_{t-p},\ldots)$.*

On note $\bar{\mathcal{L}}(\underline{X_{t-1}})$ l'espace $\bar{\mathcal{L}}(1, X_{t-1},\ldots,X_{t-k},\ldots)$ et :

$$\begin{aligned} EL(X_t|\underline{X_{t-1}}) &= EL(X_t|X_{t-1},\ldots,X_{t-k},\ldots) \\ &= EL(X_t|1,X_{t-1},\ldots,X_{t-k},\ldots) \end{aligned}$$

est la régression linéaire (ou affine) sur $\bar{\mathcal{L}}(\underline{X_{t-1}})$.

Proposition 5.1.21. *(i) et (ii) coincident si et seulement si $EX_t = 0, \forall t$.*

Remarque 5.1.22. $X_t^* = EL(X_t|X_{t-1},\ldots,X_{t-p})$

$$\begin{aligned} \|X_t - X_t^*\|_2 &= \min_{a_0,\ldots,a_p} \left\| X_t - \left(a_0 + \sum_{j=1}^{p} a_j X_{t-j}\right)\right\|_2 \\ &= \min_{Y\in H} \|X_t - Y\|_2 \end{aligned}$$

Proposition 5.1.23. $EL(X_t|\underline{X_{t-1}}) = \lim_{n\longrightarrow+\infty} EL(X_t|X_{t-1},\ldots,X_{t-n})$ *au sens de $\mathcal{L}^2$.*

Theorem 5.1.24. *(Admis)*
Soient $(X_t)_{t\in\mathbb{Z}}$ un processus stationnaire et $X_t = EL(X_t|\underline{X_{t-1}}))$ la régression affine de X_t sur $\bar{\mathcal{L}}(1, X_{t-1},\ldots,X_{t-k},\ldots)$ et $\epsilon_t = X_t - X_t^*$, alors

(i) *$(\epsilon_t)_{t\in\mathbb{Z}}$ est un bruit blanc.*

(ii) $Cov(\epsilon_t, \epsilon_{t-k}) = 0 \ \forall\ k > 0$

Définition 5.1.25. ***Processus des innovations***
En considérant les notations du théorème (5.1.24) ci-dessus on a [67, 68]:

(i) *$(\epsilon_t)_{t\in\mathbb{Z}}$ est appelé le **processus des innovations** de $(X_t)_{t\in\mathbb{Z}}$;*

(ii) *ϵ_t est l'**innovation** de X_t ;*

(iii) *X_t^* est appelé la **prévision optimale** de X_t à la date $t-1$.*

Theorem 5.1.26. ***De Wold***
Soient $(X_t)_{t\in\mathcal{Z}}$ un processus stationnaire et $(\epsilon_t)_{t\in\mathcal{Z}}$ le processus des innovations correspondant. Alors

$$\exists (a_k)_{k\in\mathbb{Z}} / \sum_{k=0}^{+\infty} |a_k| < +\infty \text{ et } X_t = m + \sum_{k=0}^{+\infty} a_k \epsilon_{t-k}.$$

L'opérateur retard et l'opérateur avance

Définition 5.1.27. ***L'opérateur backward et l'opérateur forward***

1. ***L'opérateur retard*** $\mathbb{L}$ *(lag) ou* $\mathbb{B}$ *(backward) est défini sur la classe des processus stationnaires comme étant :*

$$\mathbb{B} : (X_t)_{t\in\mathbb{Z}} \longmapsto (Y_t)_{t\in\mathbb{Z}} \ \ tel \ \ que \ \ Y_t = X_{t-1}$$

On note: $\mathbb{L}X_t = X_{t-1}$.

2. *De même,* ***l'opérateur avance*** $\mathbb{F}$ *(forward) correspond à* $\mathbb{F} : (X_t)_{t\in\mathbb{Z}} \longmapsto (Y_t)_{t\in\mathbb{Z}}$ *tel que* $Y_t = X_{t+1}$ *On note :* $\mathbb{F}X_t = X_{t+1}$.

Remarque 5.1.28. $\mathbb{L}\circ\mathbb{F} = \mathbb{F}\circ\mathbb{L} = I$ *(opérateur identité) et on notera par la suite* $\mathbb{F} = \mathbb{L}^{-1}$ *et* $\mathbb{L} = \mathbb{F}^{-1}$.

1. Il est possible de faire la composition de ces opérateurs : $\mathbb{L}^2 = \mathbb{L}\circ\mathbb{L}$, en générale on a:

$$\mathbb{L}^p = \underbrace{\mathbb{L}\circ\mathbb{L}\circ\ldots\circ\mathbb{L}}_{p\ fois} \ o\grave{u}\ p\in\mathbb{N}$$

avec la convention $\mathbb{L}^0 = I$. Notons que $\mathbb{L}^p(Xt) = X_{t-p}$.

2. Soit A le polynôme,

$$A(z) = a_0 + a_1 z + a_2 z^2 + ... + a_p z^p.$$

On notera $A(\mathbb{L})$ l'opérateur

$$A(\mathbb{L}) = a_0 I + a_1\mathbb{L} + a_2\mathbb{L}^2 + ... + a_p\mathbb{L}^p = \sum_{k=0}^{p} a_k\mathbb{L}^k.$$

Soit (X_t) une série temporelle. La série (Y_t) définie par $Y_t = A(\mathbb{L})X_t$ vérifie

$$Y_t = A(\mathbb{L})X_t = \sum_{k=0}^{p} a_p X_{t-k}$$

Par passage à la limite, on peut aussi définir des séries formelles

$$A(z) = \sum_{k=0}^{\infty} a_k z^k \ et \ A(\mathbb{L}) = \sum_{k=0}^{\infty} a_k\mathbb{L}^k$$

Proposition 5.1.29. *Pour toutes moyennes mobiles* A *et* B*, alors*

$$\left\{\begin{array}{rcl} A(\mathbb{L}) + B(\mathbb{L}) & = & (A+B)(\mathbb{L}) \\ \alpha\in\mathbb{R},\ \alpha A(\mathbb{L}) & = & (\alpha A)(\mathbb{L}) \\ A(\mathbb{L})\circ B(\mathbb{L}) & = & (AB)(\mathbb{L}) = B(\mathbb{L})\circ A(\mathbb{L}) \end{array}\right.$$

La moyenne mobile $C = AB = BA$ vérifie donc

$$(\sum_{k=0}^{\infty} a_k\mathbb{L}^k)\circ(\sum_{k=0}^{\infty} b_k\mathbb{L}^k) = (\sum_{i=0}^{\infty} c_i\mathbb{L}^i) \ o\grave{u}\ c_i = \sum_{k=0}^{i} a_k\mathbb{L}^{i-k}.$$

Les Moyennes mobiles

Définition 5.1.30. *On appelle moyenne mobile, un opérateur linéaire provenant d'une combinaison linéaire d'opérateurs retard*

$$M = \sum_{i=-m_1}^{m_2} \theta_i\mathbb{L}^{-i},\ avec\ m_1,\ m_2\in\mathbb{N}$$

et M peut encore s'écrire

$$M = \mathbb{L}^{m_1} \sum_{i=0}^{m_1+m_2} \theta_{i-m_1} \mathbb{L}^{-i} = \mathbb{L}^{m_1} \sum_{i=0}^{m_1+m_2} \theta_{i-m_1} F^i = \mathbb{L}^{m_1} \Theta(F),$$

où $\Theta(.)$ est un polynôme appelé polynôme caractéristique de M, de degré m_1+m_2, et m_1+m_2+1 sera appelé ordre de M(correspondant au nombre(théorique) de terme de M).

Définition 5.1.31. *Si $m_1 = m_2 = m$, la moyenne mobile est dite centrée. Si de plus, M est centrée, et que pour tout i, $\theta_i = \theta_{-i}$ alors la moyenne mobile est dite symétrique.*

Exemple 5.1.32. *La moyenne mobile $M_1(X_t) = \frac{(X_t+X_{t-1})}{2}$,*
soit $M_1 = \frac{(\mathbb{L}+I)}{2} = \frac{\mathbb{L}(I+\mathbb{F})}{2}$ est de degré 1, d'ordre 2 et n'est pas centrée(ni symétrique).

Exemple 5.1.33. *La moyenne mobile $M_2(X_t) = \frac{(X_{t+1}+2X_t+X_{t-1})}{4}$,*
soit $M_2 = \frac{(\mathbb{L}^{-1}+2I+\mathbb{L})}{4} = \frac{\mathbb{L}(I+2\mathbb{F}+\mathbb{F}^2)}{4}$ est de degré 2, d'ordre 3. Elle est centrée et symétrique.

Notons que les moyennes centrées symétriques, sont nécessairement d'ordre impair(pour être centrées). Pour m impair, on prendra les moyennes mobiles d'ordre $m = 2p+1$ définie par

$$M_m(X_t) = \frac{1}{m}(X_{t-p} + X_{t-p+1} + \ldots + X_{t-1} + X_t + X_{t+1} + \ldots + X_{t+p-1} + X_{t+p}).$$

Exemple 5.1.34. *La moyenne mobile $M_3(X_t) = \frac{1}{3}(X_{t-1} + X_t + X_{t-1})$ est d'ordre 3 et a pour coefficients $\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$.*

Exemple 5.1.35. *La moyenne mobile*

$$M_9(X_t) = \frac{1}{9}(X_{t-4} + X_{t-3} + X_{t-2} + \ldots + X_t + \ldots + X_{t+4})$$

est d'ordre 9 et a pour coefficients $\frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \ldots, \frac{1}{9}$.

En général le filtre $M_{2p+1}(X_t)$ s'écrit de la manière suivante:

$$M_{2p+1}(X_t) = \frac{1}{2p+1}(X_{t-p} + X_{t-p+1} + \ldots + X_{t-1} + X_t + X_{t+1} \ldots + X_{t+p-1} + X_{t+p})$$

Pour les moyennes mobiles centrées d'ordre pair ($p = 2k$), on a:

$$M_p(X_t) = M_{2k}(X_t) = \frac{1}{p}\Big[\sum_{i=-k+1}^{k-1} X_{t+i} + \frac{1}{2}X_{t-k} + \frac{1}{2}X_{t+k}\Big]$$

Remarque 5.1.36. *1. Les moyennes mobiles permettent de lisser directement la série sans hypothèse à priori sur le modèle sous-jacent. La méthode est donc valable quel que soit le modèle de décomposition. Pour cette raison, on peut classer ce type de lissage dans les méthodes non-paramétriques(par opposition aux méthodes paramétriques). C'est un outil simple a mettre en œuvre qui met en évidence l'allure de la tendance en supprimant la composante saisonnière et en atténuant le bruit.*

2. On peut aussi remplacer la moyenne par la médiane et on obtient alors un lissage par médiane mobile. Ce procédé a alors l'avantage d'être moins sensible aux valeurs aberrantes.

5.2 Processus $ARMA(p,q)$

5.2.1 $AR(p)$: processus auto-régressifs d'ordre p

Définition et représentation canonique minimale

Définition 5.2.1. ***Processus*** *$AR(p)$*
$(X_t)_{t\in\mathbb{Z}}$ est un processus $AR(p)$ si:

(i) (X_t) *est stationnaire ;*

(ii) (X_t) *vérifie une équation* $X_t = \mu + \varphi_1 X_{t-1} + \ldots + \varphi_p X_{t-p} + \epsilon_t$ *avec* $\varphi_p \neq 0$ *et* $\epsilon_t \rightsquigarrow \mathcal{BB}(0, \sigma^2)$
On note $\phi(L)X_t = \mu + \epsilon_t$ *où* $\phi(L) = 1 - (\varphi_1 L + \ldots + \phi_p L^p)$

Exemple 5.2.2. $X_t \rightsquigarrow AR(1)$ *ie* $(1 - \rho L)X_t = \mu + \epsilon_t$ *où* $\epsilon_t \rightsquigarrow \mathcal{BB}(0, \sigma^2)$ *et* $|\rho| < 1$

Remarque 5.2.3. *On a de solution non stationnaires (en espérance) de la même équation.*
Soient Y_t *tel que* $(1-\rho L)Y_t = 0 \Longrightarrow Y_t = \rho Y_{t-1} \Longrightarrow Y_t = \rho^t Y_0$ *et* (X_t) *un processus stationnaire. On définit* (Z_t) *par* $Z_t = X_t + Y_t$*. Alors on a :*

$$\begin{aligned}(1 - \rho L)Z_t &= (1 - \rho L)X_t + (1 - \rho L)Y_t = \epsilon_t + 0 \\ EZ_t &= EX_t + EY_t = m_X + \rho^t EY_0 \neq \text{ constante.}\end{aligned}$$

Et par conséquent (Z_t) *n'est pas un processus stationnaire.*

Proposition 5.2.4. *Si* $X_t \rightsquigarrow AR(p)$ *tel que* $\phi(L)X_t = \mu + \epsilon_t$ *alors*

$$EX_t = \frac{\mu}{\phi(1)} = \frac{\mu}{1 - (\varphi_1 + \ldots + \varphi_p)}$$

Démonstration.

$$\begin{aligned}X_t &= \mu + \varphi_1 X_{t-1} + \ldots + \varphi_p X_{t-p} + \epsilon_t \\ EX_t &= \mu + \varphi_1 EX_{t-1} + \ldots + \varphi_p EX_{t-p} + E\epsilon_t \\ m &= \mu + \varphi_1 m + \ldots + \varphi_p m \\ m &= \frac{\mu}{1 - (\varphi_1 + \ldots + \varphi_p)} = \frac{\mu}{\phi(1)}\end{aligned}$$

CQFD. □

Proposition 5.2.5. *Si* $X_t \rightsquigarrow AR(p)$ *est tel que* $\phi(L)X_t = \mu + \epsilon_t$ *et si l'on pose* $Y_t = X_t - m$ *(où* $m = EX_t$*), on alors*

$$\phi(L)Y_t = \epsilon \text{ et } EY_t = 0$$

Écriture d'un $MA(p)$ lorsque les racines de ϕ sont de module strictement supérieur à 1

On se donne les hypothèses $\phi(L)X_t = \mu + \epsilon_t$ avec $\phi(L) = 1 - (\varphi_1 L + \ldots + \varphi_p L^p)$ et $|z| \leq 1 \Longrightarrow \phi(z) \neq 0$.
On suppose que $\phi(z) = \prod_1^p (1 - \lambda_i z)$ où $|\lambda_i| = \frac{1}{|z_i|} < 1$.
Alors $\phi(L)$ est inversible et $\phi^{-1}(L) = \sum_0^\infty a_k L^k = A(L)$ tel que $\sum |a_k| < \infty$ et $a_0 = 1$.
On en déduit

$$\begin{aligned}X_t &= A(L)\mu + A(L)\epsilon_t \\ &= A(1)\mu + \left(\sum_0^\infty a_k L^k\right)\epsilon_t \\ &= m + \sum_0^\infty a_k L^k \epsilon_{t-k}\end{aligned}$$

du fait que $\phi(1)^{-1}\mu = m$

Proposition 5.2.6. *Sous les hypothèses ci-dessus,* $(X_t)_{t\in\mathbb{Z}}$ *possède une représentation* $MA(\infty)$ *ie:*

$$X_t = m + \sum_0^\infty a_k L^k \epsilon_{t-k}, \quad \text{où} \quad a_0 = 1, a_k \in \mathbb{R}, \sum_0^\infty |a_k| < +\infty$$

Proposition 5.2.7. *On rappelle que (5.1.20)* $\bar{\mathcal{L}}(\underline{X_t}) = \bar{\mathcal{L}}(1, X_t, X_{t-1}, \ldots, X_{t-p}, \ldots)$ *et* $\bar{\mathcal{L}}(\underline{\epsilon_t}) = \bar{\mathcal{L}}(1, \epsilon_t, \epsilon_{t-1}, \ldots, \epsilon_{t-p}, \ldots)$ *et sous les hypothèses précédentes on a :*
(i) $\bar{\mathcal{L}}(\underline{X_t}) = \bar{\mathcal{L}}(\underline{\epsilon_t})$
(ii) ϵ_t *est l'innovation de* X_t.

Définition 5.2.8. *Soit* $X_t \rightsquigarrow AR(p)$ *et* ϕ *un polynôme vérifiant les 2 conditions (i) et (ii) suivantes:*
(i) $\phi(L)X_t = \mu + \epsilon_t$
(ii) $|z| \leq 1 \Longrightarrow \phi(z) \neq 0$
Alors la représentation $\phi(L)X_t = \mu + \epsilon_t$ *est appelée* ***représentation canonique*** *de* $(X_t)_{t\in\mathbb{Z}}$.

Propriétés des $AR(p)$

Proposition 5.2.9. *Si* $(X_t)_{t\in\mathbb{Z}} \rightsquigarrow AR(p)$ *et si* $\phi(L)X_t = \mu + \epsilon_t$ *est sa représentation canonique alors on a :*

$$r(h) = \begin{cases} 0 & si\ h > 0 \\ \neq 0 & sinon \end{cases}$$

Proposition 5.2.10. *Si* $(X_t)_{t\in\mathbb{Z}} \rightsquigarrow AR(p)$, *alors :*

$$\gamma(h) = \begin{cases} 0 & si\ |h| > p \\ \neq 0 & si\ |h| = p \end{cases}$$

$$\rho(h) = \begin{cases} 0 & si\ |h| > p \\ \neq 0 & si\ |h| = p \end{cases}$$

Définition 5.2.11. *Un processus* $AR(p)$ *est dit causal lorsqu'il existe une suite de nombres* α_k *telle que* $k \in \mathbb{Z}\ \sum_{k\in\mathbb{Z}} \mid \alpha_k \mid < \infty$ *et*

$$X_t = \sum_{k=0}^{\infty} \alpha_k \varepsilon_{t-k}$$

Par cette définition, nous pouvons remarquer tout processus à moyenne mobile est causal.

5.2.2 $MA(q)$: processus moyenne mobile d'ordre q

Définition 5.2.12. $(X_t)_{t\in\mathbb{Z}} \rightsquigarrow MA(q)$ *s'il existe* $\epsilon_t \rightsquigarrow \mathcal{BB}(0, \sigma^2)$ *et* $\theta_1, \ldots, \theta_q$ *tel que*

$$X_t = m + \epsilon_t - \theta_1\epsilon_{t-1} - \ldots - \theta_q\epsilon_{t-q}$$

Proposition 5.2.13. $EX_t = m$

Remarque 5.2.14. **(i)** $(X_t)_{t\in\mathbb{Z}}$ *est nécessairement stationnaire.*
(ii) *On note* $X_t = m + \theta(L)\epsilon_t$ *où* $\theta(L) = 1 - \theta_1 L - \ldots - \theta_q L^q$
(iii) *Puisque* $X_t - m = \theta(L)\epsilon_t$, X_t *est centré,* $EX_t = 0$

Sous les hypothèses $\theta(L)$ est inversible et $\theta(L)^{-1} = \sum_{k=0}^{+\infty} a_k L^k$ avec $a_0 = 1$ et $\sum |a_k| < +\infty$. Il en découle que

$$X_t - m = \theta(L)\epsilon_t \Longleftrightarrow \theta(L)^{-1}(X_t - m) = \epsilon_t \Longleftrightarrow \theta(L)^{-1}X_t - \frac{m}{\theta(1)} = \epsilon_t$$

Soit encore

$$\sum_{k=0}^{+\infty} a_k X_{t-k} - \mu = \epsilon_t \quad o\grave{u} \quad \mu = \frac{m}{\theta(1)}$$

D'où on a la **représentation canonique** $AR(\infty)$:

$$X_t = \sum_{k=1}^{+\infty} a_k X_{t-k} + \frac{m}{\theta(1)} + \epsilon_t \tag{5.4}$$

Proposition 5.2.15. *Sous les hypothèses précédentes on a:*
(i) $\bar{\mathcal{L}}(\underline{X_t}) = \bar{\mathcal{L}}(\underline{\epsilon_t})$
(ii) ϵ_t *est l'innovation de* X_t

Propriétés des processus $MA(q)$

On suppose que la représentation étudiée est la représentation canonique

$$\begin{cases} X_t & = & m + \theta(L)\epsilon_t \\ toutes & les & racines\ de\ \theta\ sont\ de\ modules\ > 1 \\ \theta(L) & = & 1 - \theta_1 L - \ldots - \theta_q L^q \\ \epsilon_t & \rightsquigarrow & \mathcal{BB} \end{cases}$$

Proposition 5.2.16. *Sous les hypothèses précédentes on a :*

$$\gamma(h) = \begin{cases} 0 & si & |h| > q \\ -\theta_q \sigma_\epsilon^2 & \neq 0 & si |h| = q \\ \sigma_\epsilon^2 & (-\theta_h + \sum_{i=h+1}^{q} \theta_i \theta_{i-h}) & si\ 1 \leq |h| < q \\ \sigma_\epsilon^2 & (1 + \sum_{i=1}^{q} \theta_i^2) & si\ h = 0 \end{cases}$$

Remarque 5.2.17. *Il n' y a de résultat particulier pour les auto-corrélations partielles des* $MA(q)$.

Proposition 5.2.18. $\rho^i(h)$ *décroit exponentiellement avec* h.

	AR(p)	MA(q)
$\rho(h)$	décroît exponentiellement vers 0 avec h	0 si $\|h\| > q$ et non nul si $h = q$
$r(h)$	0 si $h > p$ et non nul si $h = p$	-
$\rho^i(h)$	0 si $h > p$ et non nul si $h = p$	décroît exponentiellement vers 0 avec h

TABLE 5.1 – Tableau récapitulatif des différentes situations du processus $AR(p)$ et $MA(q)$

où $\rho^i(h)$ est l'auto-corrélation inverse d'ordre h et est définie par $\rho_X^i(h) = \frac{\gamma_X^i(h)}{\gamma_X^i(0)}$ et $r(h)$ est le coefficient de X_{t-h} $EL(Xt|X_{t-1}, \ldots, X_{t-h}) = \mu + \sum_{i=1}^{p} \varphi_i X_{t-1}$.

Les auto-corrélations inverses d'un processus $MA(q)$ ont les mêmes propriétés que les auto-corrélations d'un $AR(q)$.

Proposition 5.2.19. *[55, 54, 38] Le processus stationnaire centré* (X_t) *est engendré par une modélisation minimale* $AR(p)$ *si et seulement si* $\rho(p) \neq 0$ *et* $\rho(h) = 0$ *pour tout* $h > p$.

Proposition 5.2.20. *[54, 56] Le processus stationnaire centré* (X_t) *est engendré par une modélisation minimale* $MA(q)$ *si et seulement si* $\rho(q) \neq 0$ *et* $\rho(h) = 0$ *pour tout* $|h| > q$.

5.2.3 $ARMA(p,q)$: $AR(p) + MA(q)$

Les modèles ARMA (AutoRegressive Moving Average), ou aussi modèle de Box-Jenkins, sont les principaux modèles de séries temporelles. Ils s'appuient principalement sur deux principes mis en évidence par Yule et Slutsky, le principe autorégressif et moyenne mobile. Leur application à l'analyse et à la prédiction des séries temporelles fut généralisée (Box et Jenkins en 1970) en combinant les deux principes AR et MA. Ils montrèrent que ce processus pouvait s'appliquer à de nombreux domaines et était facile à implémenter.

Définition et représentation canonique minimale

Définition 5.2.21. *Un processus stationnaire $(X_t)_t \in \mathbb{Z}$ admet une* ***représentation canonique minimale*** $ARMA(p,q)$ *[59] s'il vérifie une équation :*

$$\phi(L)X_t = \mu + \theta(L)\epsilon_t$$

avec

(1) $\epsilon_t \rightsquigarrow \mathbb{BB}(0,\sigma^2)$

(2) $\phi(L) = 1 - \varphi_1 L - \ldots - \varphi_p L^P$, *avec* $\varphi_p \neq 0$.

(3) $\theta(L) = 1 - \theta L - \ldots - \theta_q L^q$ *avec* $\theta_q \neq 0$

(4) φ *et* θ *ont toutes leurs racines de module strictement supérieur à 1 (représentation canonique).*

(5) φ *et* θ *n'ont pas de racines communes (représentation minimale).*

Remarque 5.2.22. *1. Il existe des solutions non stationnaires : soit (X_t) un processus stationnaire et (Y_t) déterministe tel que $\varphi(L)Y = 0$. On définit $Z_t = X_t + Y_t$ qui vérifie l'équation.*

2. Revenons sur la représentation canonique :

- *Si le processus (X_t) est stationnaire, alors les racines de φ sont de module distinct de 1. On pouvait supposer qu'on est le cas où θ a des racines de module 1 (c'est compatible avec la stationnarité).*
- *En considérant φ et θ avec des racines de module distinct de 1, on peut toujours se ramener à la représentation $\phi^*(L)X_t = \mu^* + \theta^*(L)\eta_t$ où φ^* et θ^* ont des racines de module > 1.*
- *Si ϕ et θ ont des racines de module strictement supérieur à 1 mais admettent une racine commune, alors*

$$\varphi(L) = (1-\lambda L)\varphi_0(L) \text{ et } \theta(L) = (1-\lambda L)\theta_0(L)$$

Et donc on a :

$$\varphi_0(L) = \frac{\mu}{1-\lambda} + \theta_0(L)\epsilon_t \Longrightarrow X_t \rightsquigarrow ARMA(p-1, q-1)$$

Proposition 5.2.23. **(i)** $EX_t = \frac{\mu}{\phi(1)}$

(ii) $\phi(L)(X_t - m) = \theta(L)\epsilon_t$

Remarque 5.2.24. *On peut se ramener par centrage au cas où $\mu = 0$.*

Démonstration. **(i)**

$$E(X_t - \varphi_1 X_{t-1} - \ldots - \varphi_p X_{t-p}) = E(\mu + \epsilon_t - \theta_1\epsilon_{t-1} - \ldots - \theta_q\epsilon_{t-q})$$

Puisque (X_t) est stationnaire il vient:

$$m(1 - \varphi_1 - \ldots - \varphi_p) = \mu + 0 \Longrightarrow m = \frac{\mu}{\phi(1)}$$

(ii)

$$\phi(L)X_t = \phi(1)m + \theta(L)\epsilon_t = \phi(L)m + \theta(L)\epsilon_t$$

Et donc $\phi(L)(X_t - m) = \theta(L)\epsilon_t$

□

Proposition 5.2.25. *On garde les mêmes hypothèses comme précédemment.*

(i) (X_t) *admet une représentation* $AR(\infty), \sum_{k=0}^{+\infty} a_k X_{t-k} = \mu + \epsilon_t$ *avec* $a_0 = 1$ *et* $\sum_k |a_k| < +\infty$.

(ii) (X_t) *admet une représentation* $MA(\infty)$, $X_t = m + \sum_{k=0}^{+\infty} b_k\epsilon_{t-k}$ *avec* $b_0 = 1$ *et* $\sum_k |b_k| < +\infty$.

(iii) $\bar{\mathcal{L}}(\underline{X_t}) = \bar{\mathcal{L}}(\underline{\epsilon_t})$

(iv) ϵ_t *est une innovation de* X_t

Remarque 5.2.26. *Il faut noter et retenir que:*

1. $AR(p) \equiv ARMA(p,0)$;
2. $MA(q) \equiv ARMA(0,q)$;
3. $$\begin{aligned} ARMA(p,q) &\equiv AR(\infty) \neq AR(P) \text{ si } P \text{ grand} \\ &\equiv MA(\infty) \neq MA(Q) \text{ si } Q \text{ grand}; \end{aligned}$$

   Généralement l'un des paramètres (p ou q) est petit alors que l'autre est grand. Avec l'approximation précédente on a alors moins de paramètres à estimer.
4. *En vertu du théorème de Wold (5.1.26), $X_t = m + B(L)\epsilon_t$, où (ϵ_t) est le processus des innovations, si de plus $X_t \rightsquigarrow ARMA(p,q)$ alors $B(L) = \frac{\theta(L)}{\phi(L)}$.*

Propriétés des processus $ARMA(p,q)$

Soit un processus $ARMA(p,q)$ tel que :

1. $\varphi(L)X_t = \theta(L)\epsilon_t$ *(éventuellement après centrage),*
2. $\phi(L) = 1 - \varphi_1 L - \ldots - \varphi_p L^p$,
3. $\theta(L) = 1 - \theta_1 L - \ldots - \theta_q L^q$.

C'est la représentation canonique minimale de $ARMA(p,q)$.

Proposition 5.2.27. ***Auto-covariance et auto-corrélation des*** $ARMA$

(i) *Pour $h > q$, les $\gamma(h)$ et les $\rho(h)$ vérifient les équations de récurrence d'ordre d'ordre p :*

$$\gamma(h) - \varphi_1\gamma(h-1) - \ldots - \varphi_p\gamma(h-p) = 0 \tag{5.5}$$
$$\rho(h) - \varphi_1\rho(h-1) - \ldots - \varphi_p\rho(h-p) = 0 \tag{5.6}$$

(ii) *Les fonctions γ et ρ décroissent vers 0 exponentiellement avec h, pour $h > q$.*

Démonstration. **(i)** $X_t = \varphi_1 X_{t-1} + \ldots + \varphi_p X_{t-p} - \theta_1\epsilon_{t-1} - \ldots - \theta_q\epsilon_{t-q}$ et par conséquent :

$$\begin{aligned} \gamma(h) &= E(X_t X_{t-h}) \\ &= \varphi_1 E(X_{t-1}X_{t-h}) + \ldots + \varphi_p E(X_{t-p}X_{t-h}) - \theta_1 \underbrace{E(\epsilon_{t-q}X_{t-h})}_{=0} - \ldots - \theta_q \underbrace{E(\epsilon_{t-q}X_{t-h})}_{=0} \\ &= \varphi_1\gamma(h-1) + \ldots + \varphi_p\gamma(h-p) \end{aligned}$$

Il s'en suit que :

$$\rho(h) = \varphi_1\rho(h-1) + \ldots + \varphi_p\rho(h-p)$$

(ii) les $\gamma(h)$ et$\rho(h)$ vérifie une équation de récurrence dont le polynôme caractéristique est $z^{p+1}\phi(\frac{1}{z})$.

CQFD. □

Les conditions initiales sont $\gamma(q)$, $\gamma(q-1)$, ... , $\gamma(q-p+1)$ et $\rho(q)$,$\rho(q-1)$, ..., $\rho(q-p+1)$.

Proposition 5.2.28. *[54] Soit le processus (X_t) stationnaire engendré par la modélisation $ARMA(p,q)$ minimale $\mathcal{A}(L)X_t = \mathcal{B}(L)\varepsilon_t$, où (ε_t) est un bruit blanc de variance $\sigma^2 > 0$. Alors, pour $\lambda \in \mathbb{T} = [-\pi, \pi]$, sa densité spectrale est donnée par $f_X(\lambda) = \frac{\sigma^2|\mathcal{B}(e^{-i\lambda})|^2}{2\pi|\mathcal{A}(e^{-i\lambda})|^2}$*

Le caractère stationnaire de (X_t) est implicitement relié au fait que le polynôme $\mathcal{A}$ ne s'annule pas sur le cercle unité, garantissant ainsi l'existence de $f_X(\lambda)$ sur tout le tore $\mathbb{T}$.

Les équations de Yule-Walker

Dans l'équation précédente pour $k = q+1, \ldots, q+p$ nous donne :

$$\begin{pmatrix} \rho(q) & \ldots & \rho(q+p-1) \\ \vdots & \ddots & \vdots \\ \rho(q+p-1) & \ldots & \rho(q) \end{pmatrix} \begin{pmatrix} \varphi_1 \\ \vdots \\ \varphi_p \end{pmatrix} = \begin{pmatrix} \rho(p+1) \\ \vdots \\ \rho(p+q) \end{pmatrix} \tag{5.7}$$

Quand ρ est connu ou estimé, on peut alors calculer les ϕ_j . Ou inversement, quand les φ_j sont connus, on calcule $\rho(q+1), \ldots, \rho(q+p)$ qui seront les conditions initiales pour le calcul de $\rho(h)$ tel $h > q$.

Pour finir ce paragraphe de la thèse, on introduit la transformée de Fourrier de la fonction d'auto-covariance. C'est une expression directe en fonction des paramètres dans les modèles $ARMA(p,q)$.

Définition 5.2.29. ***Densité spectrale.*** *S'il existe une fonction $f : [-\pi, \pi[\longrightarrow \mathbb{R}^+$ telle que : $\forall k \in \mathbb{Z}, \gamma(k) = \int_{-\pi}^{\pi} e^{ik\lambda} f(\lambda) d\lambda$, alors on dit que X_t admet une densité spectrale* [1]

Proposition 5.2.30. ***Existence et propriété de la densité spectrale.***

1. *les covariances $\gamma(k)$ vérifient $\sum |\gamma(k)^2| < \infty \iff f$ existe et est de carré intégrable sur $[-\pi, \pi[$.*
2. *Les covariance $\gamma(k)$ sont telles que $\sum |\gamma(k)| < \infty \Longrightarrow f$ existe et est de continue sur $[-\pi, \pi[$.*

Proposition 5.2.31. *Soit $X = (X_k)_{k\in\mathbb{Z}}$ un processus du second ordre stationnaire centré à temps discret. Si la densité spectrale f existe sur $[-\pi, \pi[$ alors f est paire et $f(\lambda) = \frac{1}{2\pi} \sum_{k\in\mathbb{Z}} \gamma(k) e^{-ik\lambda}$ pour $\lambda \in [-\pi, \pi[$.*

Theorem 5.2.32. ***Formule de Kolmogorov*** *Soit $X = (X_k)_{k\in\mathbb{Z}}$ un processus à temps discret stationnaire du second ordre tel que $F = \int_{-\pi}^{\pi} \log(f(\lambda)) d\lambda > -\infty$. Alors la partie régulière* [2] *de X non nulle.*

On peut montrer la relation $\|X_t - \hat{X}_t\|^2 = 2\pi \exp(F/2)$. La relation $F > -\infty$ est donc équivalente à dire que la prévision n'est pas exacte, donc que le processus n'est pas singulier.

5.2.4 Notion du processus ARIMA

Dans certains cas des $ARMA$, la série chronologique considérée peut, même après soustraction d'une tendance, apparaître comme variant au cours du temps de la façon suivante : l'espérance reste apparemment nulle, mais la variabilité s'accroît au cours du temps. La variance empirique des données calculée sur un intervalle augmente avec le temps. Une croissance de la variance peut être le signe que la série observée correspond à un cumul d'une série stationnaire ; l'accumulation des petites fluctuations de la série d'origine a une variabilité qui s'accroît au cours du temps. Le modèle le plus simple ayant cette propriété est la marche aléatoire. Partant d'un bruit blanc ϵ centré de variance 1, on définit le processus de marche aléatoire

$$X_t = \sum_{k=0}^{t}$$

Les variables X_t ne sont pas indépendantes et leur variance vaut t. Par contre, par définition, le processus des différences $X_t - X_{t-1}$ est un processus indépendant. Selon cet exemple, lorsque l'on

1. Il est montrer que la densité spectrale d'un bruit blanc faible stationnaire à variance finie existe et on peut le calculer.
2. Si l'espace vectoriel sur lequel est projeté X est indépendant de t alors le processus est dit singulier. Un exemple du processus singulier est le processus aléatoire constant. On appelle processus régulier un processus dans lequel aucun tirage aléatoire n'a lieu avant que le début de l'évolution.

observe une série dont la variance semble s'accroître, on peut calculer les différences discrètes de la série pour rechercher une série stationnaire qu'on pourra à son tour modéliser comme un $ARMA$. Cette opération de différentiation peut être opérée plusieurs fois de suite. Les processus qui donnent des processus ARMA après différentiations forment la classe des processus $ARIMA(p,d,q)$ (Auto-Regressive Integrated Moving Average), où d est le nombre de fois qu'il faut différencier le processus pour obtenir un processus $ARMA(p,q)$.

5.3 Algorithme de calcul du prédicteur

Soit un processus stationnaire du second ordre régulier X_t de fonction d'auto-covariance γ connue. Le calcul du prédicteur linéaire peut être approché par la projection sur des espaces de plus en plus grand engendrés par les n plus proches observations dans le passé. Il correspond à la résolution d'un système linéaire. Nous présentons une résolution itérative de ce calcul appelé **algorithme de Durbin-Levinson**.

5.3.1 Algorithme de Durbin-Levinson

On note $\hat{X}_t^n$ la projection de X_t sur l'espace noté L_t^n engendré par les variables $(X_{t-1},\ldots,X_{t-n})$. les coefficients $\phi_{n,i}$ sont les coordonnées de cette projection :

$$\hat{X}_t^n = \phi_{n,1}X_{t-1} + \ldots + \phi_{n,n}X_{t-n}$$

De plus, on note $v_n = \|X_t^n - X_n\|^2$, la variance de l'erreur de cette projection. Par définition, $v_0 = \gamma(0)$

Proposition 5.3.1. *Supposons qu'on ait calculé les coefficients $\phi_{n-1,i}$ et v_{n-1}. Les coefficients $\phi_{n,i}$ et v_n se calculent explicitement par les formules suivantes :*

$$\begin{aligned}
\phi_{n,n} &= \frac{1}{v_{n-1}}\left(\gamma(n) - \sum_{i=1}^{n-1}\phi_{n-1,i}\gamma(i)\right)\\
\phi_{n,i} &= \phi_{n-1,i} - \phi_{n,n}\phi_{n-1,n-i} \text{ pour } i = 1,\ldots,n-1\\
v_n &= (1-\phi_{n,n}^2)v_{n-1}
\end{aligned}$$

Remarque 5.3.2. *La proposition (5.3.1) fournie l'algorithme de Durbin-Levinson. Notons que:*

1. *l'algorithme calcule les coefficients d'auto-régression de la série sur elle-même, et propose une représentation de la série par un modèle $AR(n)$ avec une valeur n aussi grande que l'on veut et limité en pratique par la taille de l'échantillon disponible. Tout processus stationnaire du second ordre causal régulier peut être représenté par une moyenne mobile infinie ; le résultat précédent implique qu'il peut également être représenté par un modèle auto-régressif de taille infinie. Les deux représentations ne sont pas égales, mais possèdent seulement la même structure de covariance.*
2. *L'algorithme permet de calculer le prédicteur pour tout modèle stationnaire du second ordre régulier lorsque la fonction de covariance du modèle est connue. On peut par ailleurs estimer cette covariance à partir des données par l'estimateur empirique de la covariance. Nous disposons donc déjà d'une méthode de prévision adaptable à toutes les séries d'observations stationnaires et facile à programmer en pratique. Mais cette méthode prend bien en compte tous les modèles possibles (méthode non paramétrique) au prix d'une moindre efficacité statistique par rapport à des méthodes ou un choix de modèle paramétrique a été effectué. C'est pour cette raison que nous utilisons la classe des modèles $ARMA(p,q)$ et développons des méthodes spécifiques de maximum de vraisemblance adaptées à ces méthodes.*

L'algorithme suivant permet aussi le calcul du prédicteur. IL fondé sur la représentation en moyenne mobile infinie.

5.3.2 Algorithmique des innovations

Dans l'algorithme précédent, l'espace L_t^n des n variables passées est découpé en deux espaces vectoriels orthogonaux. Dans cette méthode, on le découpons en n espaces orthogonaux (orthogonalisation de Gram-Schmidt de la base formée par les X_i).de Gram-Schmidt de la base formée par les X_i) : on pose $e_{t-n} = X_{t-n}$, $e_{t-n+1} = X_{t-n+1} - P_n(X_{t-n+1})$, $e_{t-i} = X_{t-i} - P_{i+1}(X_{t-i})$ où P_i est la projection sur l'espace engendré par les variables $X_{t-i}, \ldots, X_{t-n}$. Les $(e_{t-i})_{i=1,\ldots,n}$ forment une base orthogonale de L_t^n . On note $\theta_{n,i}$ la i-ième coordonnée de X_t sur cette base :

$$\begin{aligned}\hat{X}_t^n &= \theta_{n,1}e_{t-1} + \ldots + \theta_{n,n}e_{t-n}\\ v_n &= \|\hat{X}_t^n - X_t\|^2.\end{aligned}$$

Proposition 5.3.3. *Les projections intermédiaires s'expriment par rapport à la famille de coefficients $\theta_{i,j}$ suivant la relation :*

$$P_{n-i+1}(X_{t-n+i}) = \hat{X}_{t-n+i}^i = \theta_{i,1}e_{t-n+i-1} + \ldots + \theta_{i,i}e_{t-n} = \sum_{j=0}^{i-1}\theta_{i,i-j}e_{t-n+j} \tag{5.8}$$

de plus $\|e_{t-n+i}\|^2 = v_i$

Démonstration. $P_{n-i+1}(X_{t-n+i})$ est le projeté de X_{t-n+i} sur les i variables les plus proches de son passé. Le processus X étant stationnaire au second ordre, les covariances sont les mêmes qu'entre X_t et les variables X_{t-1} à X_{t-i} et la base orthonormée est construite de la même façon, donc les coefficients de projection sont égaux. La stationnarité implique de même que $\|e_{t-n+i}\|^2 = \|X_{t-n+i} - \hat{X}_{t-n+i}^i\| = \|X_t - \hat{X}_t^i\| = v_i$. □

Proposition 5.3.4. ***Algorithme des innovations***

Supposons que pour $j < n$, on ait calculé v_j et les coefficients $\theta_{j,i}$ pour $i \leq j$. Les coefficients $\theta_{n,i}$ et v_n se calculent explicitement par les formules suivantes :

$$\begin{aligned}v_0 &= \gamma(0)\\ \theta_{n,n} &= \frac{\gamma(n)}{v_0}\\ \theta_{n,n-i} &= \frac{1}{v_i}\left(\gamma(n-i) - \sum_{j=0}^{i-1}\theta_{i,i-j}\theta_{n,n-j}v_j\right)\\ v_n &= \gamma(0) - \sum_{j=0}^{n-1}\theta_{n,n-j}^2 v_j.\end{aligned}$$

5.3.3 Prévision récursive

Considérons une série chronologique d'observations Xè$1, \ldots, X_n$. On cherche à prévoir la valeur prochaine de la série. On suppose que la série est générée par un modèle stationnaire de fonction d'auto-covariance connue. Dans ce qui suit on modifie l'algorithme des innovations pour le prédicteur de X_n en fonction des valeurs précédentes (ici la date de la valeur à prévoir augmente comme la longueur des données prises en compte n) et pour un processus non nécessairement stationnaire de fonction d'auto-covariance $\gamma(i,j)$. On rappelle que $v_j = \|X_j - X_j\|^2$.

Proposition 5.3.5. *Récursivement, le prédicteur linéaire de X_n est défini par*

$$\begin{aligned}\hat{X}1 &= 0\\ \hat{X}_n &= \sum_{j=1}^{n-1}\theta_{n,j}(X_{n-j} - \hat{X}_{n-j}).\end{aligned}$$

avec les coefficients calculés par

$$\begin{aligned} v_1 &= \gamma(1,1) \\ \theta_{n,n-k} &= \frac{1}{v_k}\left(\gamma(n,k) - \sum_{j=1}^{k-1} \theta_{k,k-j}\theta_{n,n-j}v_j\right) \\ v_n &= \gamma(n,n) - \sum_{j=1}^{n-1} \theta_{n,n-j}^2 v_j. \end{aligned}$$

Proposition 5.3.6.

$$\hat{X}_{n+1} = \sum_{j=1}^{n} \theta_{n,j}(X_{n+1-j} - \hat{X}_{n+1-j}) \qquad si\ n < m$$

$$\hat{X}_{n+1} = a_1 X_n + \ldots + a_p X_{n+1-p} + \sum_{j=1}^{q} \theta_{n,j}(X_{n+1-j} - \hat{X}_{n+1-j}) \quad si\ n \geq m$$

Et $\quad E(X_{n+1} - \hat{X}_{n-1})^2 = v_n\sigma^2.$

5.4 Modélisation et prévision par les ARMA

Il y a présomption de processus ARMA si les conditions suivantes sont satisfaites :

1. le processus est stationnaire à l'analyse visuelle :
   - pas de tendance,
   - pas de saisonnalité,
   - variance constante.
2. la fonction de corrélation empirique est:
   - à décroissance pas trop lente,
   - sans pics périodiques.

Les étapes de la prévision se récapitulent de la manière suivant :

(i) Calcul de la tendance $f(t)$ par la méthode de régression précédente.

(ii) Modélisation de la série résultante U_t par un processus $ARMA(p,q)$.

(iii) Calcul de la prévision et de son intervalle de confiance.

Le principal objectif de cette partie est la cherche d'un modèle ARMA stationnaire approchant la série $U_t = X_t - f(t)$. Cela revient à faire un bon choix des ordres p et q du modèle puis estimer les coefficients a_i et b_j , ensuite estimer les ϵ_t résidus correspondant et enfin vérifier qu'ils forment bien un bruit blanc.

5.4.1 Sélection des ordres (p,q)

D'une manière analogue au cas de la régression multiple, il n'existe pas de meilleurs choix évidents ; plus les ordres sont grands, meilleure est l'adéquation, mais le nombre de paramètres à estimer augmente et leur estimation statistique devient imprécise. On emploie donc une règle heuristique en utilisant les corrélogrammes empiriques calculés à partir des données. Les deux séries de coefficient sont utilisés pour la propriété suivante:

Proposition 5.4.1. *Pour un processus $AR(p)$, $\rho_k = 0$ dès que $k > p$. Pour un processus $MA(q), \rho_k = 0$ dès que $k > q$.*

L'estimation de ces coefficients d'auto-corrélation se fait en utilisant les coefficients d'auto-corrélation empirique. Pour estimer le coefficient d'auto-corrélation partielle, on effectue la régression des $(X_{i+k})_{i=1,\ldots,n-k}$ sur les vecteurs $(X_{i+k-1})_{i=1,\ldots,n-k}, (X_{i+k-1})_{i=1,\ldots,n-k}$, jusqu'à

$(X_i)_{i=1,\ldots,n-k}$ et on retient le coefficient de régression sur ce dernier vecteur. L'étude théorique de ces deux estimateurs permet de construire un test de significativité qui est calculé par les logiciels statistiques tel que R que nous utilisons spécifiquement dans cette partie de la thèse. On observe les valeurs de k à partir desquelles les corrélogrammes ne sont plus significatifs. Cela donne une borne raisonnable, généralement trop grande, pour le choix des p et q.

5.4.2 Identification des paramètres d'un ARMA(p,q) stationnaire.

Méthode du maximum de vraisemblance et contraste

Tout d'abord on modélise la série de données par un modèle $ARMA(p,q)$ gaussien. Nous appliquons la méthode générale du maximum de vraisemblance pour déterminer les paramètres de notre modèle, c'est-à-dire les coefficients des deux polynômes P et Q et la variance σ^2 du bruit. Les degrés des polynômes ont été choisis préalablement. La méthode consiste à écrire la densité de probabilité correspondant aux données $(X_1, \ldots, X_n)$. Cette densité de probabilité dépend des paramètres du modèle. Nous choisissons les paramètres qui rendent cette fonction la plus grande possible. La densité de probabilité exprimée en fonction des paramètres est appelée vraisemblance, d'où le nom de maximum de vraisemblance pour cette méthode. Dans le cas d'un processus gaussien centré la vraisemblance a la forme :

$$f(X) = \frac{1}{(2\pi)^{n/2}\sqrt{det(\sum)}} \exp\left(-\frac{{}^tX \sum^{-1} X}{2} \right) \tag{5.9}$$

où $\sum$ est la matrice de covariance du vecteur X ($\sum$ est aussi appelé la matrice de Toeplitz à l'ordre n). Il suffit donc de trouver parmi les processus $ARMA(p,q)$, celui dont la matrice de covariance maximise cette quantité pour $X = (X_1, \ldots, X_n)$. Cette recherche de maximum peut être très coûteuse en temps si le calcul n'est pas correctement préparé. Pour cela on diagonalise la matrice de covariance en utilisant l'algorithme des innovations :

Proposition 5.4.2. *Soit un processus gaussien $(Xi)_{i>0}$ de fonction d'auto-covariance $\gamma(i,j)$. La densité de probabilité du vecteur $(X_1, \ldots, X_n)$ peut s'écrire*

$$f(X_1, \ldots, X_n) = \frac{1}{(2\pi 2\sigma^2)^{n/2}\sqrt{v_0 \ldots v_{n-1}}} \exp\left(-\frac{2^n}{\sum_{i=1}} \frac{(X_i - \hat{X}_i)^2}{v_{i-1}} \right) \tag{5.10}$$

Les v_i et les $\hat{X}_i$ sont calculés récursivement grâce à la proposition (5.3.5)

Il ne nous reste qu'à appliquer cette méthode au cas particulier des processus $ARMA$ gaussiens. La vraisemblance s'écrit :

$$L_n(a,b,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}\sqrt{v_0 \ldots v_{n-1}}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} \frac{(X_i - \hat{X}_i)^2}{v_{i-1}} \right) \tag{5.11}$$

où a at b sont respectivement les coefficients des polynômes P et Q et σ^2 la variance du bruit. Les v_i et $\hat{X}_i$ sont calculés par l'algorithme de proposition (5.3.6)

$$\hat{X}_{i+1} = \sum_{j=1}^{i} \theta_{i,j}(X_{i+1-j} - \hat{X}_{i+1-j}) \qquad si\ i < m$$

$$\hat{X}_{i+1} = a_1 X_i + \ldots + a_p X_{i+1-p} + \sum_{j=1}^{q} \theta_{i,j}(X_{i+1-j} - \hat{X}_{i+1-j}) \quad si\ i \geq m$$

IL faut noter que les coefficients v_i et $\theta_{i,j}$ sont indépendants de σ^2. En dérivant par rapport à σ^2, on voit que le maximum est atteint pour $\hat{\sigma}^2 = \frac{1}{n}S(a^*, b^*)$ où $S(a^*, b^*)$ est la valeur minimale de

$$S(a,b) = \sum_{i=1}^{n} \frac{(X_i - \hat{X}_i)^2}{v_{i-1}}. \tag{5.12}$$

Pour obtenir le maximum de $L_n(a, b, \hat{\sigma}^2)$ par rapport aux coefficients a et b, il suffit de déterminer le minimum de la fonction :

$$l(a,b) = \log(\frac{1}{n}S(a,b)) + \frac{1}{n}\sum_{i=1}^{n}\log(v_{i-1}) \tag{5.13}$$

Remarque 5.4.3. *La méthode du maximum de vraisemblance revient à déterminer les valeurs de a et b qui minimisent cette quantité. Cette recherche opérationnelle est réalisée par un algorithme d'optimisation non linéaire à partir d'une valeur initiale des paramètres a et b. Cette première valeur proposée doit se trouver près du vrai maximum et correspondre à un modèle causal, car les algorithmes de projections utilisés exploitent cette hypothèse. Les logiciels d'optimisation utilisent une valeur de départ obtenu par l'estimateur de Yule-Walker. La théorie du maximum de vraisemblance donne des conditions pour que cet estimateur soit non seulement consistant mais efficace, c'est-à-dire avec une vitesse de convergence en $\sqrt{n}$ et une variance asymptotique minimale. Cette méthode d'estimation n'est justifiée que lorsque le processus $ARMA$ est gaussien, mais lorsque l'on cherche le prédicteur linéaire, un modèle gaussien et un modèle non gaussien de même fonction d'auto-covariance sont parfaitement équivalents, car le prédicteur ne dépend que de la fonction d'auto-covariance.*

Estimation d'un modèle ARMA

L'estimation des paramètres d'un modèle $ARMA(p, q)$ lorsque les ordres p et q sont supposés connus peut se réaliser par différentes méthodes dans le domaine temporel :

- Moindres Carrés Ordinaires (modèle sans composante $MA, q = 0$). Dans ce cas, on retrouve les équations de Yule Walker. En remplaçant les autocorrélations théoriques par leurs estimateurs, on peut retrouver les estimateurs des MCO des paramètres du modèle par la résolution des équations de Yule Walker.
- Maximum de Vraisemblance approché **(Box and Jenkins 1970)**
- Maximum de Vraisemblance exacte **(Newbold 1974, Harvey et Philips 1979, Harvey 1981)**.

Nous allons présenter ici brièvement la démarche de l'estimation par le maximum de vraisemblance. Cette maximisation est réalisée à l'aide d'algorithmes d'optimisation non linéaire **(Newton-Rahpson, méthode du simplex)** que nous n'exposerons pas dans le cadre de ce chapitre. Nous nous contenterons ici de montrer comment s'écrit le programme de maximisation de la vraisemblance permettant d'estimer les paramètres d'un modèle $ARMA(p, q)$.

Theorem 5.4.4. *Considérons un processus $ARMA(p, q)$ stationnaire $X = (X_k)_{k\in\mathbb{Z}}$ de densité spectrale $f_{(a,b)}$. Soit $I_n(\lambda)$ l'estimateur de la densité spectrale défini par :*

$$I_n(\lambda) = \frac{1}{2\pi}\sum_{k=1-n}^{n-1}\hat{\gamma}(k)e^{-ik\lambda} = \frac{1}{2\pi n}|\sum_{k=1}^{n}X_k e^{-ik\lambda}|^2. \tag{5.14}$$

*On considère le **contraste de Whittle** défini par :*

$$U_n(a,b) = \frac{1}{4\pi}\int_{-\pi}^{\pi}\log(f_{(a,b)}(\lambda)) + \frac{I_n(\lambda)}{f_{(a,b)}(\lambda)}d\lambda. \tag{5.15}$$

L'estimateur de (a, b) par minimum de contraste est la valeur de (a, b) qui minimise $U_n(a, b)$.

Remarque 5.4.5. *Cet estimateur converge à la même vitesse que l'estimateur du maximum de vraisemblance. De plus, ces propriétés de convergence sont établies même dans le cas où le processus étudié n'est pas gaussien. Il est bien adapté au cas des processus $ARMA$, car la densité spectrale de ces processus s'exprime directement par rapport aux coefficients des polynômes, ce qui n'est pas le cas pour la fonction d'aut-ocovariance.*

5.4.3 Significativité des paramètres

Losqu'on a déjà calculé les estimations des paramètres, on peut tester si tous les paramètres estimés sont significativement distincts de 0. Les estimateurs de paramètres étant asymptotiquement gaussiens et de variance estimable, on peut construire pour chaque paramètre, un test de l'hypothèse "ce paramètre est nul" et ne conserver dans le modèle que les paramètres rejetés par ce test de nullité. Si les paramètres de grand ordre a_p ou b_q sont acceptés comme nuls, cela veut dire qu'ils ne sont pas nécessaires dans le modèle et qu'il faut baisser l'ordre p ou q correspondant puis estimer de nouveau tous les paramètres.

5.4.4 Test d'adaptation

On vérifie que le modèle est adéquat, au sens où ses résidus forment bien un bruit blanc. Sinon il y a encore de la dépendance dans les résidus. On cherchera à utiliser un modèle d'ordre p ou q plus grand afin de prendre en compte cette dépendance. Pour tester l'adéquation à un processus $ARMA$, on peut utiliser un test dit de portmanteau (ce qui signifie "fourre-tout" en anglais). Après avoir calculé les résidus estimés $(\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n)$ obtenus à partir $(X_1, \ldots, X_n)$ et des coefficients estimés par une méthode de type maximum de vraisemblance ou minimum de contraste. On définit

$$\hat{T}_n(k) = n\sum_{j=1}^{k}(\hat{\rho}_\epsilon(k))^2 \quad o\text{ù} \quad \hat{\rho}_\epsilon(k) = \frac{\frac{1}{n}\sum_{i=p}^{n-k}\hat{\epsilon}_i\hat{\epsilon}_{i+k} - (\frac{1}{n}\sum_{i=p}^{n}\hat{\epsilon}_i)^2}{\frac{1}{n}\sum_{i=p}^{n}\hat{\epsilon}_i^2 - (\frac{1}{n}\sum_{i=p}^{n}\hat{\epsilon}_i)^2}.$$

On note que $\hat{\rho}_\epsilon(k)$ est la corrélation empirique des résidus, qui doit tendre vers 0 si le modèle est bien un $ARMA(p,q)$ dès que $k = 0$, suivant un Théorème de la Limite Centrale (d'où le n devant la statistique de test). On note aussi que les ϵ_i ne sont calculables que quand $i \geq p+1$). On doit choisir k suffisamment grand pour donner plus de pertinence au test. Ainsi, sous l'hypothèse que le processus est bien un processus $ARMA(p,q)$, dont le bruit admet un moment d'ordre 4 on peut alors montrer que:

$$\hat{T}_n(k) \underset{n\longrightarrow+\infty}{\xrightarrow{\mathcal{L}}} \chi^2(k-p-q). \tag{5.16}$$

5.4.5 Sélection du modèle: critère AIC et BIC

On suppose que la procédure précédente ait permis de construire un modèle valide pour les données dont on dispose. C'est possible qu'un autre modèle soit également valide, et s'adapte mieux aux données. Pour cela, plus on va choisir des ordres p et q importants, plus le modèle a des chances de présenter des résidus plus petits ; l'inconvénient est que les paramètres seront plus nombreux et moins correctement estimés. On introduit un critère de choix entre les modèles qui met de balancer l'adéquation du modèle avec le nombre de paramètres. Ainsi, on établira un compromis entre les deux inconvénients. Pour ce faire on peut utiliser comme critère de sélection le critère d'information d'Akaike AIC (anglais Akaike information criterion) ou le critère $AICC$ qui met en balance le nombre de paramètres du modèle $p+q$ avec la vraisemblance obtenue par les estimateurs du maximum de vraisemblance.

$$AICC(p,q) = -2\log L_n(\hat{a}, \hat{b}, \hat{\sigma}^2) + 2(p+q+1)\frac{n}{n-p-q-2} \tag{5.17}$$

On peut aussi utiliser le critère bayésien BIC. Dans le cas d'un processus $ARMA$ causal inversible celui-ci s'écrit:

$$BIC(p,q) = \log\big(\frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_j^2\big) + \frac{\log n}{n}(p+q), \tag{5.18}$$

avec $\hat{\epsilon}_j = \frac{\hat{P}(B)}{\hat{Q}(B)}X_j$ (On doit avoir Q inversible causal pour que les racines de Q soient en dehors du disque trigonométrique). Les estimateurs $\hat{P}_N(B)$ et $\hat{Q}_N(B)$ sont calculés en remplaçant leurs

coefficients a_i et b_j par les estimateurs obtenus par une des méthodes évoquées plus haut. On calcule ce critère pour tous $0 \leq p \leq p_{max}$ et $0 \leq q \leq q_{max}$ et on fait le choix :

$$(\hat{p}, \hat{q}) = \underset{0 \leq p \leq p_{max},\ 0 \leq q \leq q_{max}}{Argmin} BIC(p,q) \tag{5.19}$$

En pratique il est plus coûteux en calculs de déduire l'ordre p et q pour un processus $ARMA(p,q)$, car on doit maintenant minimiser une fonction à deux variables. Les critères AIC et BIC pour un processus $ARMA(p,q)$ de s'écrient:

$$AIC(p,q) = log(\hat{\sigma}_\varepsilon^2) + 2\frac{(p+q)}{T}\ ,\ BIC(p,q) = log(\hat{\sigma}_\varepsilon^2) + \frac{(p+q)}{T}log(T)$$

Pour minimiser ces fonctions, une méthode consiste à faire deux boucles itératives sur p et q pour tester tous les couples (p,q) jusqu'à certaines bornes $p < p_{max}$ et $q < q_{max}$. A l'intérieur de ces boucles, on calcule d'abord les estimateurs $\hat{\phi}$, $\hat{\theta}$ et utilisant par exemple les moindres carrés ou le maximum de vraisemblance, $\hat{\sigma}_\varepsilon^2$ on calcule les critères AIC et BIC pour ces différents ordres et on trouve le minimum de ces quantités. On a donc les valeurs $\hat{p}$ et $\hat{q}$ qui minimisent l'AIC ou le BIC. Ensuite, on calcule des estimateurs efficaces des paramètres du modèle $ARMA(\hat{p},\hat{q})$ utilisant la méthode du maximum de vraisemblance. Si plusieurs modèles sont concurrents, on choisit le couple (p,q) qui minimise la statistique

$$AIC(p,q) = log(\hat{\sigma}_\varepsilon^2) + 2\frac{(p+q)}{T} \quad ou \quad BIC(p,q) = log(\hat{\sigma}_\varepsilon^2) + \frac{(p+q)}{T}log(T)[55,\ 56,\ 61].$$

5.4.6 Prévision

Ayant réuni les estimations de la tendance de f, les paramètres du modèle $ARMA$ et les estimations du bruit ϵ_i, on propose d'estimes X_{n-1} par

$$\hat{X}_{n+1} = f(n+1) + \hat{a}_1(X_n - f(n)) - \ldots - \hat{a}_p(X_{n-p+1} - f(n-p+1)) + \hat{b}_1\hat{\epsilon_n} + \ldots + \hat{b}_q\hat{\epsilon}_{n-q+1}.$$

Le calcul de l'intervalle de confiance repose sur l'étude des covariances des différents estimateurs utilisés [62, 63, 64]. Ces estimateurs étant très dépendants, le calcul est compliqué, mais il est généralement effectué par le logiciel qui calcule les estimations des paramètres. On observe dans la formule la diminution du nombre de termes correspondant aux valeurs réellement observées. L'information apportée par le modèle ARMA diminue et la prévision se rapproche de la prévision déterministe définie par la fonction f . La prévision utilisant les modèles $ARMA$ n'a d'intérêt qu'à court terme.

Toute théorie ci- dessus appliquée au logiciel R nous plonge dans notre article publié dans IJAM (International Journal of Applied Mathematics)[57]. Dans ce qui suit on présente le résumé de l'article en français.

5.5 Optimisation et analyse de l'iqa

De nombreux phénomènes dépendant du temps gouvernent notre monde. Les séries temporelles ou chronologiques font partie de l'une des méthodes souvent utilisées pour les comprendre mathématiquement. On peut utiliser les séries temporelles pour prévoir les événements futurs. Dans cette étude, nous utilisons le concept de série chronologique pour analyser et modéliser l'indice de la qualité de l'air à Dakar en vue de faire sa prévision à court terme. Pour couronner cette approche, le processus $ARMA(2,1)$ choisi par le critère optimal de sélection AIC etBIC a été utilisé pour effectuer quelques simulations.

Notons que la prédiction est appliquée dans de multiples domaines tels que l'atmosphère, l'astrologie, l'économie, les sciences socio-politiques, le traitement du signal,... Rappelons qu'une série temporelle ou série chronologique est une suite formée d'observations au cours du temps. A partir de la connaissance des informations antérieures, on peut estimer le comportement d'un

système dans le futur. Si l'estimation de l'état futur du système est exacte, on parle d'une méthode de prédiction entièrement déterministe. En réalité, plusieurs facteurs rendent le calcul exact de l'état futur du système impossible. Cependant, il est possible de générer un modèle qui peut être utilisé pour calculer la probabilité du comportement futur entre deux limites spécifiées. Un tel modèle est appelé modèle stochastique ou processus stochastique. Une importante classe de modèles stochastiques est utilisée pour la description des séries chronologiques stationnaires appelée la classe des modèles stochastiques stationnaires. Ces modèles supposent que les propriétés de la série temporelle sont invariantes par la translation temporelle. Parmi ces modèles on peut citer les modèles Autoregressif Moving (AR), Moving Average (MA) et Autoregressif Moving Average (ARMA). Les processus utilisés pour la description des séries temporelles non stationnaires (en moyenne, en variance et autres) sont: ARIMA, SARIMA,... Pour la construction des modèles quelles que soient leurs classes, Box et Jenkins ont introduit une méthodologie utilisée pour l'obtention d'un modèle linéaire qui s'ajuste au mieux à une série temporelle. Cette méthodologie se décompose en trois étapes: l'identification du modèle, l'estimation des paramètres et la validation du modèle [58].

5.5.1 La recherche des modèles candidats

Les fonctions d'autocorrélation et d'autocorrélation partielle nous permettent de déterminer l'ordre d'un modèle autorégressif ou à moyenne mobile. Maintenant, cherchons l'ordre du modèle à partir des critères statistiques .Les fonctions d'autocorrélation et d'autocorrélation partielle nous permettent de déterminer l'ordre d'un modèle autorégressif ou à moyenne mobile. Maintenant, cherchons l'ordre du modèle à partir des critères statistiques . Pour simplifier supposons que la recherche se fasse parmi des processus ARMA non troués, cela nous permet de mettre de coté les modèles saisonniers. Le bon processus ARMA est celui d'ordre le couple inconnu (p^*, q^*) tel que $(p^*, q^*) < (p_{max}, q_{max})$. Autrement dit, on postule que les ordres vrais sont respectivement inférieurs à deux ordres p_{max} et q_{max} que l'on se fixe. Le vrai problème qui pourrait se poser est de choisir des ordres p_{max} et q_{max} trop faibles qui ne prenant pas en compte les ordres vrais. Généralement, on examine les corrélogrammes représentant les autocorrélations et autocorrélations partielles estimées pour fixer ces bornes maximales. Fixer les bornes maximales p_{max} et q_{max} revient à définir une famille de $(p_{max}+1) \times (q_{max}+1)$ filtres candidats voir le tableau 5.2.

q \ p	0	1	...	p_{max}
0		$AR(1)$	...	$AR(p_{max})$
1	$MA(1)$	$ARMA(1,1)$	...	$ARMA(p_{max},1)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
q_{max}	$MA(q_{max})$	$ARMA(q_{max},1)$	...	$ARMA(p_{max},q_{max})$

TABLE 5.2 – Critère BIC de la série *iqa* pour $(p,q) \in [1,...,5]^2$

La recherche d'un modèle optimal dans le sens d'un certain critère se fera donc au sein de cette famille. Si la procédure aboutit au choix d'un modèle appartenant à la dernière colonne ou à la dernière ligne alors il est prudence de refixer des valeurs pour p_{max} ou q_{max} supérieures à celles choisies initialement.

Soit l_i la log-vraisemblance du i^{e} modèle, T la taille de l'échantillon de travail et k_i le nombre de paramètres, le critère de sélection s'écrit en générale de la manière suivante :

$$c_i(T,k_i) = \frac{-2l_i}{T} + \frac{k_i g(T)}{T} \tag{5.20}$$

où $g(T)$ est une fonction positive du nombre paramètres à estimer et donc désigne l'ampleur de la pénalité. La complexité du problème se trouve dans le choix des pondérations entre vraisemblance et complexité, et par conséquent sur la spécification de la fonction $k_i g(T)$ de pénalité .

5.5.2 Estimation, tests de validation et prévisions des processus ARMA

La procédure de modélisation de Box et Jenkins (1976) comporte les étapes suivantes :

1. Stationnarisation et Dessaisonalisation
2. Identification
3. Estimation
4. Validation et Test
5. Prévisions

Tests de stationnarité de la série iqa

Pour vérifier la stationnarité il existe plusieurs méthodes: l'examen visuel de la série, les calculs de la moyenne et de la variance sur des sous ensembles de la serie et tests d'égalité, l'analyse visuelle de la décroissance de la fonction d'autocorrélation et les tests de racine unité (Dickey-Fuller, Phillips-Perron, Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test,. . .). Nous retiendrons pour cette recherche les tests de la racine unité.

	Statistique	p-value
Phillips-Perron Unit Root Test	-11.0836	0.01
Augmented Dickey-Fuller Test	-4.8911	0.01
KPSS	0.5337	0.03408

Nous observons que pour le Dickey-Fuller Augmenté, le Phillips-Perron et tout comme pour le KPSS test, la p-value est inférieure à 0.05 donc la variable log(iqa) est stationnaire.

Identification du modèle de l'iqa: $ARMA(2,1)$

Notons qu'un modèle AR(p) présente un corrélogramme simple caractérisé par une décroissance géométrique de ses termes et un corrélogramme partiel caractérisé par ses p premiers termes différents de 0. Un modèle MA(q) présente un corrélogramme simple défini par ses q premiers termes significativement différents de 0 et un corrélogramme partiel caractérisé par une décroissance géométrique des retards. Le modèle ARMA(p,q) présente un corrélogramme simple et partiel qui sont un mélange des deux corrélogrammes des processus AR et MA purs.

Dans certaines littératures l'analyse des séries temporelles unidimensionnelles comme celle de notre recherche sur l'indice de la qualité de l'air dans la ville de Dakar par la méthode de Box et Jenkins se résume à trois étapes : identification, estimation, validation. La phase initiale d'identification est souvent difficile dans la pratique. Il n'est pas du tout évident de trouver souvent le bon modèle adapté à la série chronologie considérée. C'est dans cette phase que se déroule le grand travail la recherche opérationnelle du bon modèle d'ajustement de la série.

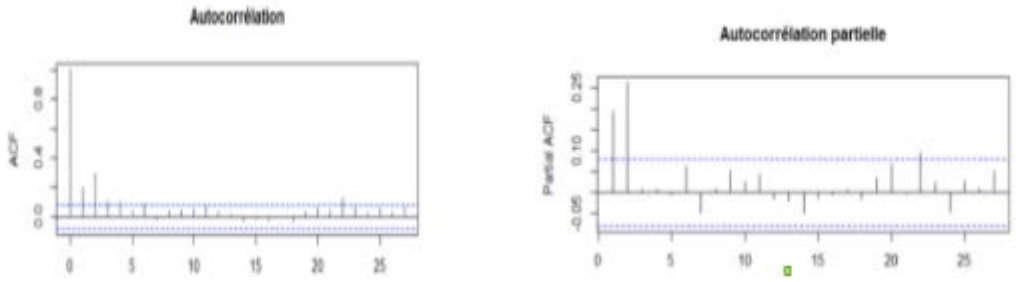


FIGURE 5.1 – ACF et PACF pour la série log(iqa).

En observant la figure5.1, on note remarque que l'autocorrélation et l'autocorrélation partielle décroissent tout doucement et prennent du temps avant d'être significativement nulles.

Pour cette raison on écarte la possible de retenir un modèle AR ou MA. Estimons les critères AIC et BIC pour un $ARMA(p,q)$ pour $p = 1,...,5$ et $q = 1,...,5$. Cherchons le couple (p,q) optimal (le minimum) pour notre modèle. Les tableaux 5.3 et 5.4 nous donnent ces valeurs.

p \ q	**1**	**2**	**3**	**4**	**5**
1	805.07	790.47	779.93	774.89	770.37
2	**762.64** 3e*candidat optimum*(2,1)	764.57	766.4	768.4	770.22
3	764.56	766.64	768.5	769.84	772.33
4	766.39	768.2	764.29	**762.25** 1er*candidat optimum*(4,4)	764.04
5	768.38	770.36	772.42	**762.52** 2e*candidat optimum*(4,5)	764.45

TABLE 5.3 – Critère AIC de la série iqa pour $(p,q) \in [1,...,5]^2$

La recherche opérationnelle par le critère AIC semble nous orienter vers l'opération (le modèle) $ARMA(4,4)$ suivit de $ARMA(5,4)$ puis du $ARMA(2,1)$. Mais le troisième choix optimal (minimal) qui aurait le plus bas AIC serait l'opération $ARMA(2,1)$. Nous avons à choisir d'estimer trois paramètres, huit ou bien encore d'en estimer neuf. Choisir un modèle à neuf ou à huit paramètres maximise la vraisemblance du modèle. Ce modèle pourrait nous donner un meilleur ajustement de notre échantillon, mais pourrait être complètement erroné pour faire des prédictions. Nous choisirons d'estimer donc trois paramètres au lieu de huit ou neuf. Pour vérifier si ce choix ajuste bien nos données, nous avons aussi la possibilité de comparer les deux estimateurs de la variance des résidus.

p \ q	**1**	**2**	**3**	**4**	**5**
1	825.0663	815.4707	809.9266	809.8786	810.3665
2	**787.6319** *optimum*(2,1)	800.9503	826.8575	810.1573	822.2392
3	794.5862	801.3645	808.2271	814.8847	822.1193
4	801.3752	808.201	811.2327	827.0412	819.5427
5	808.3771	814.8691	815.7096	819.4184	825.8988

TABLE 5.4 – Critère BIC de la série iqa pour $(p,q) \in [1,...,5]^2$

Le critère BIC du tableau 5.4 nous montre que le couple $(2,1)$ réalise le minimum. Nous retenons en conclusion pour notre ajustement l'opération $ARMA(2,1)$.

Estimation et simulation de l'iqa

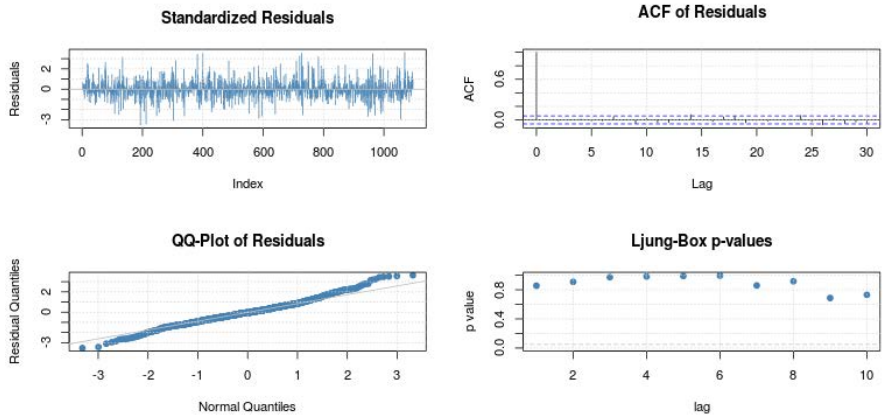
Dans cette partie nous faisons l'estimation et la simulation du logarithme néperien de l'iqa par un $ARMA(2,1)$ sous le logiciel R. Avec la librairie fArma de R, la simulation nous donne:
Titre: ARIMA Modelling
Call: armaFit(formula = arma(2, 1), data = log(iqa))
Moments: Skewness=0.2539 Kurtosis= 0.9756.
Skewness mesure l'asymétrie de la distribution du processus et Kurtosis mesure l'excès du Kurtosis : il est égal à $3 - \widehat{KU}$. L'estimation du Kurtosis et celle du Skewness sont respectivement données par :

$$\widehat{KU} = \frac{\frac{1}{T}\sum_{t=1}^{T}(\hat{u}_t - \bar{u})^4}{(\frac{1}{T}\sum_{t=1}^{T}(\hat{u}_t - \bar{u})^2)^2} \qquad et \qquad \widehat{SK} = \frac{\frac{1}{T}\sum_{t=1}^{T}(\hat{u}_t - \bar{u})^3}{(\frac{1}{T}\sum_{t=1}^{T}(\hat{u}_t - \bar{u})^2)^{3/2}}$$

Le logiciel R nous fournit les coefficients de l'estimation dans le tableau suivant:

```
Estimate  Std. Error  t value Pr(>|t|)
ar1         1.59923     0.03428    46.65   <2e-16 ***
```

```
ar2        -0.60384     0.03298   -18.31   <2e-16 ***
ma1        -0.93293     0.01862   -50.11   <2e-16 ***
intercept   4.04585     0.14121    28.65   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
sigma^2 estimated as: 0.9162
log likelihood:       -376.32
AIC Criterion:        762.64
```



Nous remarquons l'estimation du modèle paraît de bonne qualité, car la valeur de notre estimateur est très proche de la vraie valeur.

Le test QQ-PLOT de normalité est une méthode graphique : Le nuage de point est formé par (quantiles de N(0,1), quantiles empiriques réduits de $\hat{u}_t$), sous l'hypothèse H0 le nuage est rectiligne sur la droite y=x). Nous affirmons donc ici qu'il y a normalité (voir figure 5.5.2).

Tests d'adéquations

Avec le logiciel R, le Box.test nous donne le résultat suivant :
Box-Pierce test
data: predict$residuals
X-squared = 6383.219, df = 10, p-value = 0,6742
La p-value est grande $(0,6742 > 0,05)$ donc les résidus ne sont pas corrélés.

Et le test de Kolmogorov-Smirnov avec la fonction lillie.test() nous fournit :
Lilliefors (Kolmogorov-Smirnov) normality test
data: predict$residuals
D = 0.1125, p-value < 0.3221
On peut donc affirmer que les données des résidus suivent une loi de Gauss.

5.5.3 Prévision de l'iqa par $ARMA(2,1)$

Dans cette section prédisons le logarithme néperien des valeurs de l'iqa pour les trente un (31) jours du mois de Janvier 2013 et comparons ces résultats avec les vraies mesures relevées au niveau des stations pendant cette même période.

```
             1erJan2013 2jan2013  3jan2013 4jan2013 5jan2013 6jan2013 7jan2013  8jan2013
logiqa Prédit 4.7178     4.5378    4.4269   4.3582   4.3152   4.2880   4.2705    4.2589
Residuals     0.3409     0.4096    0.4388   0.4535   0.4620   0.4676   0.4716    0.4748

jan2013  10jan2013 11jan2013 12jan2013 13jan2013  14jan2013 15jan2013 16jan2013
4.2509    4.2451    4.2407    4.2372    4.2341     4.2314    4.2290     4.2266
0.4776    0.4800    0.4823    0.4844    0.4864     0.4883    0.4901     0.4919

17jan2013 18jan2013 19jan2013 20jan2013 21jan2013 22jan2013  23jan2013 24jan2013
4.2244    4.2222    4.2201    4.2180    4.2159    4.2139     4.2119    4.2099
0.4936    0.4953    0.4970    0.4986    0.5001    0.5016     0.5031    0.5045
```

25jan2013	26jan2013	27jan2013	28jan2013	29jan2013	30jan2013	31jan2013
4.2080	4.2061	4.2042	4.2023	4.2004	4.1986	4.1968
0.5059	0.5073	0.5086	0.5099	0.5111	0.5124	0.5135

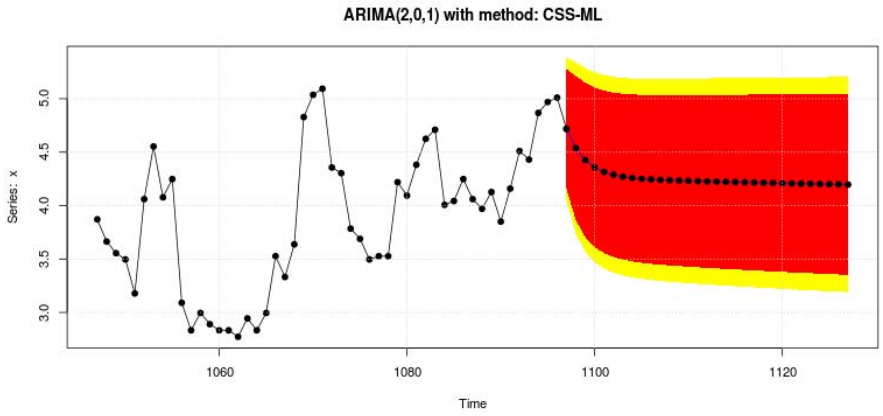


FIGURE 5.2 – Prévisions et intervalle de confiance à 90% en rouge et 95% en jaune.

Mesure de la qualité de prévision

Dans notre échantillon initial, $(X_1, ..., X_T)$ considérons seulement, $T_1 = [(1-\varepsilon)T]$ observations avec $\varepsilon > 0$. Les $L = T - [(1-\varepsilon)T]$ soit les $L = T - T_1$ seront à prévoir par le modèle. On peut alors considérer plusieurs critères :

– Mean Absolute Pourcentage Error

$$MAPE = \frac{1}{L}\sum_{i=1}^{L}\left|\frac{X_{T_1+i}-\hat{X}_{T_1+i}/T_1}{X_{T+i}}\right|$$

– Mean Square Error

$$MSE = \left(\sum_{i=1}^{L}\frac{(X_{T_1+i}-\hat{X}_{T_1+i}/T_1)^2}{L}\right)^{1/2}$$

Nous obtenons sous le logiciel R, MAPE= 4.687186 et MSE= 4.624215. En conclusion on peut affirmer ARMA(2,1) est un bon modèle pour notre ajustement et en tenant compte du critère d'information traité ci-haut, le couple (2,1) est un bon choix optimal du couple (p, q) qui minimise le critère.

5.5.4 Discussion

Dans la sous-section 5.5.3, on a prédit l'*iqa* sur les trente un (31) jours du mois de Janvier 2013. On a comparé les résultats avec les vraies mesures relevées au niveau des stations pendant cette même période dans la ville de Dakar. On remarque que plus le nombre de jours est grand plus l'erreur est grande. En d'autres termes l'erreur augmente avec le nombre de jours. Cela nous permet de dire que notre modèle $ARMA(2,1)$ permet de faire de prévisions à court terme et non de prévision à long terme.

5.6 L'article en anglais

International Journal of Applied Mathematics

Volume 28 No. 5 2015, 621-636
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
doi: http://dx.doi.org/10.12732/ijam.v28i5.13

OPTIMIZATION AND ANALYSIS OF THE INDEX OF AIR QUALITY IN DAKAR BY THE PROCESS AUTO REGRESSIVE MOVING AVERAGE ARMA(2,1)

Lebede Ngartera[1] §, Salimata Diagne[2], Youssou Gningue[3]

[1]Department of Mathematics
Faculty of Exact and Applied Sciences N'Djamena University, Chad
B.P: 1739, Ndjamena, CHAD

[2]Department of Mathematics
Faculty of Science and Technology
Cheikh Anta Diop University, Dakar, SENEGAL

[3]Department of Mathematics and Computer Sciences
Laurentian University
Sudbury, ON, CANADA

Abstract: Numerous time-dependent phenomena govern our world. The time series are part of one of the methods often used to understand mathematically. Time series can be used to predict future events. In this article, we use the concept of time series to analyze and model the Index of air quality in Dakar to make his short-term forecast. To top this approach, the auto-regressive moving average(2,1) selected by the optimal Akaike information criterion and Bayesian information criterion was used to perform some simulations.

AMS Subject Classification: 78M50, 65C20, 62P12, 37M10
Key Words: AIC and BIC, air quality, optimization, auto-regressive moving average, simulation and estimation

1. Introduction

The statistical prediction is applied in many fields such as the atmospheric

Received: August 29, 2015

§Correspondence author

science, astronomy, economics, socio-political science, signal processing, etc. Note that a time series or time series is a sequence formed of observations over time. From the knowledge of the previous information, we can estimate the behavior of a system in the future. If the estimate of the future state of the system is accurate, we speak of a method of entirely deterministic prediction. In fact, several factors make the exact calculation of the future state of the system impossible. However, it is possible to generate a model that can be used to calculate the probability of a range of future behaviors between two specified limits. Such a model is called a stochastic model and stochastic processes. An important class of stochastic models is used for the description detrending called class stationary stochastic models. These models assume that time series properties are invariant under the time translation. These models include the autoregression models (AR), Moving Average models (MA) and autoregressive moving average models (ARMA). The processes used for the description of non-stationary time series (average, variance and others) are: ARIMA, SARIMA, ... ARIMA and SARIMA models are extensions of ARMA class in order to include more realistic dynamics, in particular, respectively, non stationarity in mean and seasonal behaviours.

For the construction of models, whatever their class, Box and Jenkins have introduced a methodology for obtaining a linear model that best adjusts to a time series. This methodology consists of three steps: identification of the model, parameter estimation and validation of the model, cf Fiordaliso [1].

The article is structured as follows: Section 2 is devoted to the basic concepts rest of ARMA. Ssection 3 presents the optimal selection criterion. Section 4 discusses the document, the Box and Jenkins approach to retain ARMA(2,1) as our simulation model and we end with the conclusion and some perspectives.

2. Process ARMA(AR+MA)

The ARMA models (also known as Box Jenkins models), are the most common type of time series model. They are mainly based on two principles highlighted by Slutsky and Yule, the autoregressive and moving average principles. Their application to the analysis and prediction of time series was widespread Box and Jenkins in 1970. They showed that this process could be applied to many areas and was easy to implement.

Given a time series X_t, the ARMA model is a tool to understand and attempt to predict possibly future values in this series. The model consists of two parts: an Autoregressive (AR) part and Moving-Average (MA) part. The

model is generally denoted ARMA (p, q), where p is the order of the AR part and q the order of the party with MA $p \geqslant 0$; $q \geqslant 0$.

An autoregressive model and moving-average orders (p, q) (abbreviated ARMA (p, q)) is a discrete time process $(X_t, t \in \mathbb{N})$ satisfying, cf. Jonathan [5]:

$$X_t = \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}, \tag{2.1}$$

where the parameters φ_i and θ_i are constants, and the error terms ε_i are independent of the process.

- An autoregressive model AR(p) may be identified as an ARMA$(p, 0)$. In this case the series becomes :

  $$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t$$

  where $\varphi_1, \ldots, \varphi_p$ are the model parameters, c is a constant and ε_t white noise, cf Peter and al [6]. In some literature [5], the constant is often omitted, the process is then said to be centered. The first autoregressive processes were introduced by George Udny Yule. In his paper he uses the first autoregression model to model the time series of the number of sunspots than the Schuster periodogram method, cf. Jonathan and al [5].

- A moving average model MA(q) is an ARMA$(0, q)$. The series is then:

  $$X_t = \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

  where $\theta_1, \ldots, \theta_q$ are the parameters of the model $\varepsilon_t, \varepsilon_{t-1}, \ldots$ are again the error terms. Eugen Slutzky introduced this model for the first time in 1927 the moving average process in his article.

Note that the error terms of ε_i are generally assumed to be independent and identically distributed (iid) as a normal distribution with zero mean : $\varepsilon_t \sim N(0, \sigma^2)$ where σ^2 denotes the variance.

Condensed manner, using the delay operator, the ARMA may be written in condensed form as, for all $t \in \mathbb{Z}$,

$$\mathcal{A}(z) X_t = \mathcal{B}(z) \varepsilon_t \tag{2.2}$$

where we define for any $z \in \mathbb{C}$, polynomials $\mathcal{A}$ et $\mathcal{B}$ by, cf Jonathan et al [5]: $\mathcal{A}(z) = 1 - \theta_1 z - \ldots - \theta_p z^p$ and $\mathcal{B}(z) = 1 + \varphi_1 z + \ldots + \varphi_q z^q$.

In the case of AR(p), we have $\mathcal{B}(z) = 1$ whereas symmetrically, in the case of MA(q), we have $\mathcal{A}(z) = 1$. Modeling is referred to as "*minimal*" if $\theta_p \neq 0$, $\varphi_q \neq 0$ and if $\mathcal{A}$ and $\mathcal{B}$ have no common root. Without this, it is always possible to find a formulation ARMA(p', q') equivalent with $p' \leqslant p$ and $q' \leqslant q$ generating (X_t).

2.1. Autocorrelation and Correlograms Functions

The autocorrelation function of a process $(X_t, t \in \mathbb{Z})$ with average $E(X_t) = m$, denoted $\rho(k)$ or ρ_k, is defined by $\forall k \in \mathbb{Z}$:

$$\rho(k) = \rho_k = \frac{\gamma(k)}{\gamma(0)}, \tag{2.3}$$

where $\rho(k) \in [-1, 1]$, and $\gamma(k) = \gamma_k$ denotes the autocovariance function,

$$\forall \in \mathbb{Z}, \gamma(k) = \gamma_k = E[(X_t - m)(X_{t-k} - m)].$$

The partial autocorrelation [8] of order k denotes the correlation between X_t and X_{t-k} obtained when the influence of the variables X_{t-i} with $i < k$ was removed.

The graph of an autocovariance function (resp. Autocorrelation) is called a variogram (resp. correlogram). Similarly, we define a partial correlogram as the graph of the partial autocorrelation function (resp. correlogram).

An AR (p) has simple correlogram characterized by a geometric decrease in its terms and partial correlogram characterized by its first p terms different from 0.

	AR(p)	MA(q)
$\rho(h)$	decreases exponentially to 0 with h	0 if $\|h\| > q$ and non-zero if $h = q$
$r(h)$	0 if $h > p$ and non-zero if $h = p$	-
$\rho^i(h)$	0 if $h > p$ and non-zero if $h = p$	decreases exponentially to 0 with h

where $\rho^i(h)$ is the inverse auto-correlation of order h and is defined by $\rho^i_X(h) = \frac{\gamma^i_X(h)}{\gamma^i_X(0)}$ and $r(h)$.

Inverse autocorrelations [8] a MA(q) has the same properties as the autocorrelations of an AR(q).

Proposition 2.1. (cf. Jonathan et al. [5], Brockwell et al. [7] and Stocker [9]) The centered stationary process (X_t) is generated by minimal modeling AR(p) if and only if $\rho(p) \neq 0$ and $\rho(h) = 0$ for any $h > p$.

Proposition 2.2. (cf. Brockwell et al. [7]) The centered stationary process (X_t) is generated by a minimal MA(q) model if and only if $\rho(q) \neq 0$ and $\rho(h) = 0$ for any $|h| > q$.

There are other places in the rest of the paper where you need to make the same correction In practice, an ARMA process is often presumed under the following conditions:

1. the process is stationary in the visual analysis:
    - no trend,
    - no seasonality,
    - constant variance.
2. empirical correlation function is:
    - to decay too slow,
    - without periodic peaks.

Proposition 2.3. (cf. Brockwell et al. [7]) Either the process (X_t) generated by the stationary modeling minimal $ARMA(p, q)$ $\mathcal{A}(L)X_t = \mathcal{B}(L)\varepsilon_t$ where (ε_t) is a white noise variance $\sigma^2 > 0$. So, for $\lambda \in \mathbb{T} = [-\pi, \pi]$, its spectral density is given by $f_X(\lambda) = \frac{\sigma^2|\mathcal{B}(e^{-i\lambda})|^2}{2\pi|\mathcal{A}(e^{-i\lambda})|^2}$.

The stationary character of (X_t) is implicitly related to the fact that the polynomial $\mathcal{A}$ does not vanish on the unit circle, thus guaranteeing the existence of $f_X(\lambda)$ on over all the torus $\mathbb{T}$.

2.2. Estimation of ARMA Model

Parameter estimation of ARMA(p, q) where p and q commands are assumed to be known can be achieved by various methods in the time domain:

- Ordinary Least Squares (model without MA components, $q = 0$). In this case, there are the Yule Walker equations. Replacing theoretical

autocorrelations by their estimators, one can find the MCO estimators of the model parameters by solving the Yule Walker equations.

- Maximum Likelihood approach (Box and Jenkins 1970) [8].

- Exact Maximum Likelihood (Newbold 1974, Harvey and Philips 1979, Harvey 1981) [8].

We will present here briefly the approach of the estimate by maximum likelihood. This maximization is performed by using nonlinear optimization algorithms such as Newton-Raphson or the simplex method that we will not explain it in this chapter. Here it suffices to show how the writing of the likelihood maximization program to estimate the parameters of an ARMA (p, q).

2.3. Stationarity and Causality of the Process

Definition 2.4. We say that the process $(Y_t)_{t\in\mathcal{T}}$ ($\mathcal{T} = \mathbb{N}$ *or* $\mathbb{Z}$) is strictly stationary (or strongly stationary) if the law of $\{Y_{t_1}, \ldots, Y_{t_n}\}$ is the same as the law of $\{Y_{t_{1+\tau}}, \ldots, Y_{t_{n+\tau}}\}$ for all $(t_1, \ldots, t_n)$ with $t_i \in \mathbb{T}$ for $i = 1, \ldots, n$ and for any $\tau \in \mathcal{T}$ with $t_{i+\tau} \in \mathcal{T}$.

Thus, a random process is strictly stationary if all these statistical characteristics, that means all those moments are invariant for any change in the origin of time. But the stationary in strict sense is too restrictive, and this condition is relaxed by defining the stationary of second order.

Definition 2.5. A process $(Y_t)_{t\in\mathcal{T}}$, is called second-order stationary (or weakly stationary) if $(Y_t)_{t\in\mathcal{T}}$, is 2nd order and if the first two moments are time-invariant:

1. $E(Y_t) = m = constant \ \forall \ t \in \mathcal{T}$

2. $Var(Y_t) = \sigma^2 = \gamma(0) < \infty$

3. $Cov(Y_t, Y_{t-h}) = E(Y_t Y_{t-h}) - E(Y_t)E(Y_{t-h}) = \gamma(h) \ \forall \ t \in \mathcal{T}, \forall \ h \in \mathcal{T}$

In short, a process Y_t is called second-order stationary if its mean, its variance and its covariance are independent of time and if variance is non infinite. A such a process is without trend in mean and without trend in variance.

Example 2.6. The best known example of stationary process is white noise process (denoted BB or White Noise). A White Noise is a series of real random variable ε_t, $t \in \mathcal{T}$ such that: $E(\varepsilon_t) = 0 \ \forall \ t \in \mathcal{T}$,

$$\text{and} \ \ \gamma(h) = E(\varepsilon_t \varepsilon_{t-h}) = \begin{cases} \sigma^2 \ if \ h = 0 \\ 0 \quad h \neq 0 \ if \ not \end{cases}$$

Definition 2.7. An AR(p) process is called causal when there is a series of numbers α_k as $k \in \mathbb{Z}$,

$$\sum_{k \in \mathbb{Z}} \mid \alpha_k \mid < \infty$$

and

$$X_t = \sum_{k=0}^{\infty} \alpha_k \varepsilon_{tk}.$$

By this definition, we can see any moving average process is causal.

3. Selection Criterion

The autocorrelation functions and partial autocorrelation allow us to determine the order of an autoregressive or moving average model. Let us look for the model from the statistical criterion.

3.1. Candidate Models Search

For simplicity assume that research is done among not what is a sidestep ARMA process, it allows us to put aside seasonal patterns. What is a fine ARMA process is that of the unknown couple (p^*, q^*) such that $(p^*, q^*) < (p_{max}, q_{max})$. In other words, it is assumed that real orders are respectively less than two orders p_{max} and q_{max} that one focuses on. In practice, a problem arises when the orders p_{max} and q_{max} are chosen too small to find the best model. Generally we examine correlograms representing the autocorrelations and partial autocorrelations estimated in order to set these maximum limits. Setting maximum limits p_{max} and q_{max} gives rise to a family of $(p_{max} + 1) \times (q_{max} + 1)$ candidate models, as shown in Table 1.

The search for an optimal model in the sense of a certain criterion will therefore be in this family. If the procedure results in the choice of a model

q \ p	0	1	...	p_{max}
0		$AR(1)$	...	$AR(p_{max})$
1	$MA(1)$	$ARMA(1,1)$	...	$ARMA(p_{max},1)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
q_{max}	$MA(q_{max})$	$ARMA(q_{max},1)$	...	$ARMA(p_{max},q_{max})$

Table 1: Candidate models search

belonging to the last column or the last row then it is prudent to redefine the new values for $p_{\max}$ or $q_{\max}$ higher than those initially chosen.

Let l_i the log-likelihood of the ith pattern, T the size of the working sample and k_i the number of parameters, the selection criterion is written in general as follows:

$$c_i(T,k_i) = \frac{-2l_i}{T} + \frac{k_i g(T)}{T}, \tag{3.1}$$

where $g(T)$ is a positive function of many parameters to estimate and therefore denotes the magnitude of the penalty.

This minimization problem reflects the tradeoff between increasing likelihood and increasing complexity of the model.

3.2. AIC and BIC Criterion for Autoregressive Process

The Akaike Information Criterion (AIC) generates a function that estimates the quality of the fit. Recall that if the number of parameters increases, the variance $\hat{\sigma}_\varepsilon^2$ decreases. In order not to end up with an over-parameterization of the model, we add a factor that will make a compromise between the number of parameters and minimum variance. In the following, we will take a model $A(p)$ and considers $\hat{\sigma}_\varepsilon^2$ using maximum likelihood for many positive values of p . This method could also be used for a model $MA(q)$. The AIC consists of calculating

$$AIC(p) = log(\hat{\sigma}_\varepsilon^2) + 2\frac{p}{T}.$$

Using this criterion, we remark that if p is the obtained parameter of the minimization and p is the parameter of the real model, it has the following property: $P(\hat{p} \geqslant p) \longrightarrow 1$ when $T \longrightarrow +\infty$, [8]. The criterion therefore tends to select a larger number of parameters than the real model, which leads us to a small error term $\hat{\sigma}_\varepsilon^2$. If one wishes to have a better choice of order p, there is the Bayesian

information criterion (BIC) that uses a higher penalty. The BIC selects the parameter p which minimizes the following quantity $BIC(p) = log(\hat{\sigma}_\varepsilon^2) + \frac{p}{T} log(T)$.

3.3. AIC and BIC for ARMA Model

In estimation, it is a little more expensive to deduct the p and q order for a ARMA(p,q) process because to optimize the model it is necessary to minimize a function of two variables. The AIC and BIC criteria for a ARMA(p,q) process are written as:

$$AIC(p,q) = log(\hat{\sigma}_\varepsilon^2) + 2\frac{(p+q)}{T} \ , \ BIC(p,q) = log(\hat{\sigma}_\varepsilon^2) + \frac{(p+q)}{T} log(T).$$

To minimize these functions, one method is to make two iterative loops on p and q to test all pairs (p,q) p until some limits $< P$ and $q < Q$. Inside these loops, first compute the estimators $\hat{\Phi}$, $\hat{\theta}$ and using for example the least squares or maximum likelihood, $\hat{\sigma}_\varepsilon^2$ we calculate the AIC and BIC criteria for these various levels and there is the minimum of these quantities. It was therefore the values $\hat{p}$ and $\hat{q}$ that minimize the AIC or BIC. Then we calculate efficient estimators of the model parameters ARMA $(\hat{p}, \hat{q})$ using the maximum likelihood method. If several models are competing, we choose the pair (p,q) which minimizes statistics $AIC(p,q) = log(\hat{\sigma}_\varepsilon^2) + 2\frac{(p+q)}{T}$ or $BIC(p,q) = log(\hat{\sigma}_\varepsilon^2) + \frac{(p+q)}{T} log(T)$, cf. Peter et al. [6].

4. Estimation, Validation Testing and Forecasting ARMA

The modeling procedure of Box and Jenkins comprises the following steps, [8]:

1. Stationarity and seasonal adjustment
2. Identification
3. Estimate
4. Validation and Test
5. Forecast

4.1. Stationarity Tests of the iqa Series

To verify the stationarity there are several methods: visual examination in the series, the calculations of the mean and variance of the subsets of the series

and equality tests, visual analysis of the decrease in the function of autocorrelation and unit root tests (Dickey-Fuller, Phillips-Perron Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test ...). We use the unit root test in the current research.

	Statistic	p-value
Phillips-Perron Unit Root Test	-11.0836	0.01
Augmented Dickey-Fuller Test	-4.8911	0.01
KPSS	0.5337	0.03408

We observe that for the Dickey-Fuller Augmented, the Phillips-Perron and just like the KPSS test, the p-value less than 0.05 so the variable log(iqa) is stationary.

4.2. Model Identification of iqa: ARMA(2,1)

In some literature the analysis of univariate time series as that of our research on the Index of air quality in the city of Dakar in the Box and Jenkins boils down to three steps: identification, estimation, validation. The initial phase of identification is often difficult in practice. It is not clear at all to often find the right model to suit the timeline series considered. It is in this phase that takes place the great work operations research series good adjustment model.

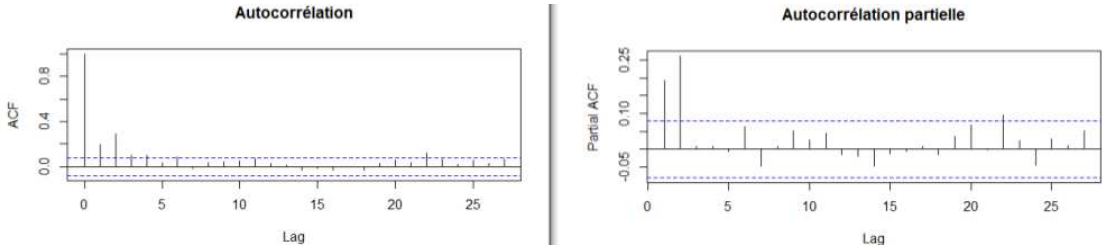


Figure 1: ACF and PACF for series log(iqa).

By observing Figure 1, there is seen that the autocorrelation and partial autocorrelation decay slowly and take time before they significantly nil.

For this reason we rule out possible to retain an AR or MA model. Estimates the AIC and BIC criteria for an $ARMA(p,q)$ $p = 1,\ldots,5$ and $q = 1,\ldots,5$. Looking couple (p,q) optimal (minimum) for our model. Tables 2 and 3 give us those values.

Operational research by the AIC seems to direct us towards the operation (the model) ARMA(4,4) followed by ARMA(5,4) and ARMA (2,1). But

p \ q	1	2	3	4	5
1	805.07	790.47	779.93	774.89	770.37
2	**762.64** *third candidate optimum(2,1)*	764.57	766.4	768.4	770.22
3	764.56	766.64	768.5	769.84	772.33
4	766.39	768.2	764.29	**762.25** *first candidate optimum(4,4)*	764.04
5	768.38	770.36	772.42	**762.52** *second candidate optimum(4,5)*	764.45

Table 2: Criterion AIC series iqa for $(p, q) \in [1, ..., 5]^2$

the third optimal choice (minimum) which would be the lowest AIC operation ARMA(2,1). We choose to estimate three parameters, eight or even nine to estimate. Selecting a model of nine or eight parameters maximizes the likelihood of the model. This model could give us a better fit in our sample, but could be completely wrong to make predictions. So we choose to estimate three parameters instead of eight or nine. To check if this choice fits well to our data, we also have the opportunity to compare the two variance estimators residues.

p \ q	1	2	3	4	5
1	825.0663	815.4707	809.9266	809.8786	810.3665
2	**787.6319** *optimum(2,1)*	800.9503	826.8575	810.1573	822.2392
3	794.5862	801.3645	808.2271	814.8847	822.1193
4	801.3752	808.201	811.2327	827.0412	819.5427
5	808.3771	814.8691	815.7096	819.4184	825.8988

Table 3: BIC Criterion for series $log(iqa)$ with $(p, q) \in [1, ..., 5]^2$

The BIC of the table 3 shows that the pair (2,1) gives us the minimum. We retain finally fit for our operation ARMA(2,1).

4.3. Estimation and Simulation of IQA

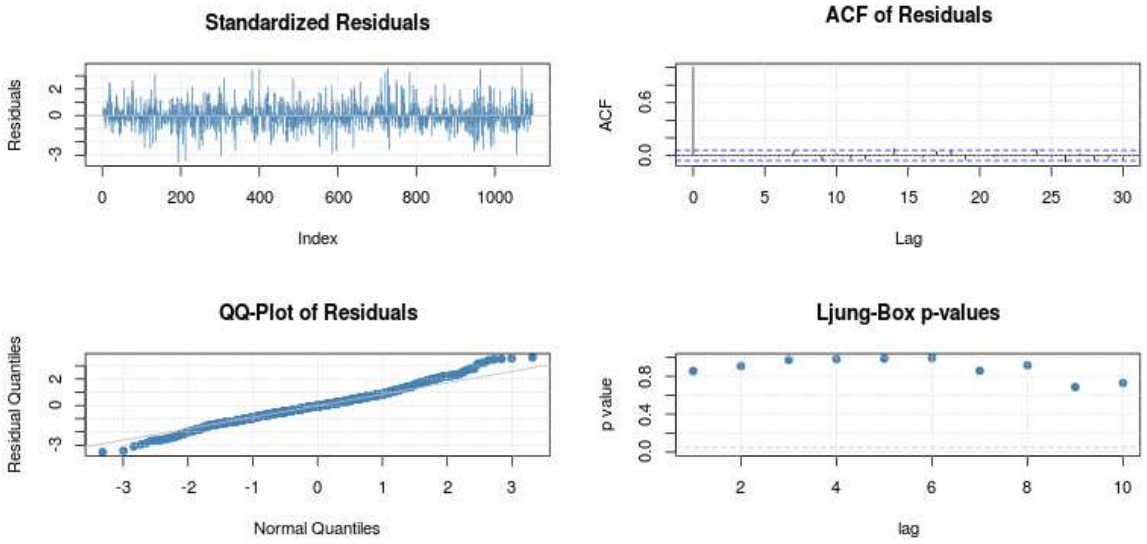
In this section we estimate and simulation of $log(IQA)$ by an ARMA(2,1) under the R software. With Farma library in R, the simulation gives us:
Title: ARIMA Modelling
Call: armaFit(formula = arma(2, 1), data = log(iqa))
Moments: Skewness=0.2539 Kurtosis= 0.9756.
Skewness measures the asymmetry of the distribution process and kurtosis measures the excess kurtosis: it is equal to $3 - \widehat{KU}$. The estimate of kurtosis and

skewness are respectively given by:

$$\widehat{KU} = \frac{\frac{1}{T}\sum_{t=1}^{T}(\hat{u}_t - \bar{u})^4}{(\frac{1}{T}\sum_{t=1}^{T}(\hat{u}_t - \bar{u})^2)^2} \quad \text{and} \quad \widehat{SK} = \frac{\frac{1}{T}\sum_{t=1}^{T}(\hat{u}_t - \bar{u})^3}{(\frac{1}{T}\sum_{t=1}^{T}(\hat{u}_t - \bar{u})^2)^{3/2}}.$$

The R software provides us the coefficients of the estimation in the following table:

```
           Estimate  Std. Error  t value Pr(>|t|)
ar1         1.59923     0.03428    46.65   <2e-16 ***
ar2        -0.60384     0.03298   -18.31   <2e-16 ***
ma1        -0.93293     0.01862   -50.11   <2e-16 ***
intercept   4.04585     0.14121    28.65   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
sigma^2 estimated as: 0.9162
log likelihood:       -376.32
AIC Criterion:        762.64
```



We note the model estimation seems good, because the value of our estimator is very close to the true value.

QQ-PLOT test of normality is a graphical method: The point cloud is formed by (quantile N(0.1), reduced empirical quantile of $\hat{u}_t$), assuming H0 the cloud is on the straight line y = x). We confirm that there is a normality.

4.3.1. Adequacies Tests

With R software, Box.test gives us the following result:
Box-Pierce test
Data: predict$residuals
X-squared = 6383.219, df = 10, p-value = 0,6742
The p-value is large ($0.6742 > 0.05$) so the residuals are not correlated.

And the Kolmogorov-Smirnov lillie test with the function () provides us:
Lilliefors (Kolmogorov-Smirnov) normality test
data: predict$residuals
D = 0.1125, p-value < 0.3221
We can therefore say the residues data follow a Gaussian distribution.

4.4. Forecast of iqa by ARMA(2,1)

In this section we predict log(IQA) for a 31-day period in January 2013, we compare these results with real measurements taken at stations during the same period.

	1erJan2013	2jan2013	3jan2013	4jan2013	5jan2013	6jan2013	7jan2013	8jan2013
logiqa Prdit	4.7178	4.5378	4.4269	4.3582	4.3152	4.2880	4.2705	4.2589
Residuals	0.3409	0.4096	0.4388	0.4535	0.4620	0.4676	0.4716	0.4748

9jan2013	10jan2013	11jan2013	12jan2013	13jan2013	14jan2013	15jan2013	16jan2013
4.2509	4.2451	4.2407	4.2372	4.2341	4.2314	4.2290	4.2266
0.4776	0.4800	0.4823	0.4844	0.4864	0.4883	0.4901	0.4919

17jan2013	18jan2013	19jan2013	20jan2013	21jan2013	22jan2013	23jan2013	24jan2013
4.2244	4.2222	4.2201	4.2180	4.2159	4.2139	4.2119	4.2099
0.4936	0.4953	0.4970	0.4986	0.5001	0.5016	0.5031	0.5045

25jan2013	26jan2013	27jan2013	28jan2013	29jan2013	30jan2013	31jan2013
4.2080	4.2061	4.2042	4.2023	4.2004	4.1986	4.1968
0.5059	0.5073	0.5086	0.5099	0.5111	0.5124	0.5135

4.5. Measuring Quality Prediction

In our initial sample, $(X_1, \ldots, X_T)$ only consider, $T_1 = [(1-\varepsilon)T]$ observations with $\varepsilon > 0$. The $L = T - [(1-\varepsilon)T]$ either are the $L = T - T_1$ will be predict by the model. We can then consider several criteria:

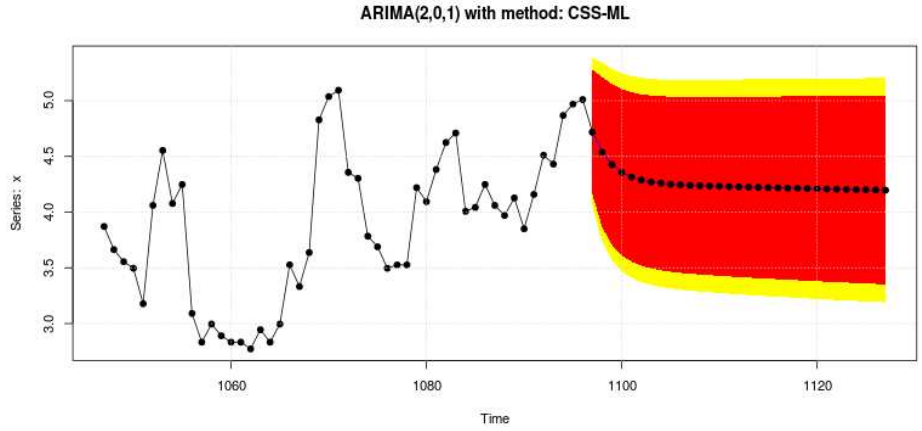


Figure 2: Forecasts and confidence interval to 90% (red) and 95%(yellow).

- Mean Absolute Percentage Error

$$MAPE = \frac{1}{L}\sum_{i=1}^{L}\left|\frac{X_{T_1+i}-\hat{X}_{T_1+i}/T_1}{X_{T+i}}\right|$$

- Mean Square Error

$$MSE = \left(\sum_{i=1}^{L}\frac{(X_{T_1+i}-\hat{X}_{T_1+i}/T_1)^2}{L}\right)^{1/2}$$

Here L is the number of fitted points. We get under the R software, MSE = 4.687186 and MAPE = 4.624215. In conclusion we can say ARMA(2,1) is a good model for our fit and taking into account the information criterion corrected above, the pair (2,1) is a good choice of optimal pair (p, q) that minimizes the criterion. Indeed MSE gives greater weight to larger deviations (which could result from outliers) and MAPE gives less overall weight to a large deviation if the time series value is large. MSE give us the averages of the squared deviations and MAPE averages the absolute percent errors. Although the concept of MAPE sounds very simple and convincing, it has major drawbacks in practical application [10].

It cannot be used if there are zero values (which sometimes happens for example in demand data) because there would be a division by zero. For forecasts

which are too low the percentage error cannot exceed 100%, but for forecasts which are too high there is no upper limit to the percentage error. When MAPE is used to compare the accuracy of prediction methods it is biased in that it will systematically select a method whose forecasts are too low. This little-known but serious issue can be overcome by using an accuracy measure based on the ratio of the predicted to actual value (called the Accuracy Ratio), this approach leads to superior statistical properties and leads to predictions which can be interpreted in terms of the geometric mean [10]. Mean square error (MSE) can also be utilized in the same fashion. Squaring the forecast errors eliminates the possibility of offsetting negative numbers, since none of the results can be negative.

5. Conclusion and Perspectives

This research leads us to choose the ARMA(2,1) to model and predict the index of the air quality in Dakar. The model can be used to make short-term predictions of IQA in Dakar by the air quality management center (CGQA). It provides a tool that the quality monitoring center air Dakar could use to benefit public health. An interesting avenue of research work would be to couple an urban sprawl model with a simulation model of emissions of pollutants from road transport. Indeed, there is evidence that as the city grows in extent, the more origin-destination trips made daily by motorists is growing, resulting in increased traffic. Such a model could be used to assess the impact of new areas of development plans and cities around Dakar on transport and the impact of that traffic on people and plants.

References

[1] A. Fiordaliso, *Systèmes Flous et Prévision de Séries Temporelles*, Hermès Science Publication, Paris (1999).

[2] Y. Aragon, Séies temporelles avec R. Méthodes et cas, Springer, Paris (2011).

[3] A. Schuster, *On the Periodicity of Sun-Spots*, Royal Society of London, London (1906).

[4] G. Colletaz, *Les Critères de sélection*, Notes de Cours Master 1, ESA, Novembre (2007).

[5] Jonathan D. Cryer and Kung-Sik Chan, *Time Series Analysis with Applications in R*, 2nd Ed., Springer (2008), 191-245.

[6] Peter J. Brockwell and Richard A. Davis, *Time Series: Theory and Methods*, Springer (2009).

[7] P.J. Brockwell and R.A. Davis, *Time Series: Theory and Methods*, 2nd Ed., Springer-Verlag, New York (1991).

[8] P.J. Brockwell and R.A. Davis, *Introduction to Time Series and Forecasting*, 2nd Ed., Springer Texts in Statistics, New York (2002), 83-421.

[9] T. Stocker, On the asymptotic bias of OLS in dynamic regression models with autocorrelated errors, *Statist. Papers*, **48** (2007), 81-93.

[10] C. Tofallis, A better measure of relative prediction accuracy for model selection and model estimation, *J. of the Operational Research Society*, **66**, No 8 (2015), 1352-1362.

5.7 Conclusion

Cette étude nous a permis de choisir le processus $ARMA(2,1)$ pour modéliser et prédire l'indice de la qualité de l'air à Dakar. C'est un modèle qui prédit l'*iqa* à court terme en fonction de données antérieurs recueillies sur la qualité de l'air à Dakar par le centre de gestion de la qualité de l'air (CGQA). Il peut éventuellement être un outil que le centre de surveillance de la qualité de l'air de Dakar pourrait exploiter dans une perspective de santé publique.

Avec la présentation de notre deuxième article, on se dirige maintenant vers la conclusion générale de la thèse dans la partie suivante.

Conclusion générale - perspectives

Cette thèse propose deux modèles de prédiction de l'indice de la qualité de l'air dans Dakar. Le premier modèle se repose sur la régression linéaire multiple par les moindres carrés ordinaires (MCO). Le deuxième modèle le prédit par le modèle $ARMA(2,1)$. On procède à la comparaison de ces deux modèles dans le tableau suivant:

Méthode	**Avantages**	**Inconvénients**
Modèle ARMA	Simplicité et rapidité dans la construction. Outil d'estimations par la variable elle- même. Permet de tracer des séries chronologiques stationnaires, des données manquantes.	Présume la stationnarité, paramètres figés sur la modélisation gaussienne. Nombre limité de variables explicatives. Pas d'explication des causes de l'évolution des variables dépendantes. Nombre minimal d'observations pour des prévisions acceptables. Exige d'avoir des données pour les variables explicatives sur l'horizon de prévision si des variables prévisionnelles sont présentes.
Modèle de régression linéaire multiple	Inclut des variables explicatives. Peut expliquer les causes de l'évolution de la variable dépendante. Permet de traiter des séries chronologiques stationnaire avec des données manquantes. Fournit non seulement une courbe ajustant les données, mais aussi des prédictions avec leurs incertitudes.	Construction plus complexe. Exige d'avoir des données pour les variables explicatives sur l'horizon de prédiction. Requiert un nombre d'observations minimale pour données de prévisions acceptables. Plus ce nombre est plus élevé, plus le modèle est complexe.

TABLE 5.5 – Comparaison des deux modèles.

Notons que la pollution atmosphérique est liée au niveau et type d'urbanisation dans une agglomération donnée. La notion de ville durable de plus en plus agitée dans les rencontres internationales renvoie à une ville qui maîtrise son environnement, en termes de déchets produits, de pollution atmosphérique afin de minimiser sa contribution aux changements climatiques à travers les gaz à effet de serre. De ce point de vue, Dakar offre l'exemple d'une agglomération à croissance démographique spectaculaire, avec une densité humaine vingt(20) fois supérieure à la moyenne nationale. Étouffant dans ses limites traditionnelles, Dakar s'étend et s'étire vers le seul espace que lui offre sa situation géographique: Le Nord Est de la presqu'île(Autoroute à péage, Keur Massar,

Guédiawaye, Sanghalkam, ...)

Des pistes intéressantes suivantes de travail de recherche peuvent s'ouvrir:

(1) Coupler un modèle d'étalement urbain avec un modèle de simulation des émissions de polluants issus des transports routiers. En effet, il est prouvé que plus la ville s'étire dans l'espace, plus les trajets origine- destination effectués chaque jour par les automobilistes s'allonge, entraînant une augmentation du trafic. Un tel modèle pourrait permettre d'évaluer l'impact des plans d'aménagement de nouveaux quartiers et villes autour de Dakar sur les transports, puis l'impact de ce trafic sur les populations et les végétaux(zone des niayes).

(2) Orienter le travail de modélisation vers les modèles non linéaires en matière d'analyse des données pour minimiser d'avantage les erreurs de prédiction.

(3) Introduire une meilleure modélisation locale des émissions, en couplant un modèle gaussien pour l' échelle locale à un modèle eulérien. Cela permet de quantifier la variabilité due à la mauvaise représentation de l'échelle locale autour de la source,

(4) Utiliser un traitement statistique de réduction d'échelle. Il s'agit de déterminer des relations entre les concentrations simulées et les observations aux stations dont l'échelle de représentativité est plus faible que l' échelle bien représenté par le modèle. Ces deux approches sont complémentaires. La première permet d'estimer les sources d'incertitude dans la représentation des phénomènes sous-maille et d'améliorer la modélisation localement autour des sources. La seconde améliore les performances des prévisions aux stations d'observations, mais ne permet pas une estimation spatiale de la variabilité.

(5) Développer des modèles mathématiques pour les sources polluantes mobiles pouvant aider les décideurs à prendre des mesures pour contrôler la pollution due au trafic causant dégâts à la santé et à l'environnement.

Quelques mesures s'imposent pour une bonne qualité de l'air à Dakar:

- Utilisation de pots catalytiques: le catalyseur est un corps chimique actif à base de métaux précieux. Entre 250 et 400°C, il brûle les reliquats de CO et de HC avant leur dispersion dans l'atmosphère et élimine les oxydes d'azote,
- Utilisation de filtre à particules: réservé aux motorisations diesel, ce filtre supprime quasi-totalement le rejet de composés polluants solides(fumées noires). Les constructeurs doivent équiper les véhicules diesel et son utilisation devrait se généraliser dans les prochaines années,
- contrôle technique: obligatoire pour les véhicules de plus de 4 ans et devant être renouvelé ensuite tous les 2 ans, ce contrôle objectif permet de vérifier le bon fonctionnement des points techniques. En cas de déficience, certaines fonctions majeures directement liées à la sécurité sont soumises à une obligation de réparation,
- élimination des **CARS RAPIDES** et **NDIAGA NDIAYE**, leur remplacement par de nouveaux moyens de transport moins polluants
- mener des actions d'éducation et de sensibilisation auprès des usagers, des associations de quartier ou de consommateurs et des jeunes à l'école,... Cela dans le but de la prise de conscience de la menace de la pollution afin d'améliorer la qualité de l'air.
- Création d'un programme municipal de maîtrise ou de réduction de l'effet de serre.
- Revoir la location et le nombre de station de surveillance de la qualité de l'air dans Dakar.
- Surveillance(d'avantage) des stations pour une bonne base de données sans grand nombre de données manquantes. Cela entraine spécialement pas de coupure régulière d'électricité.
- Pour un bon modèle de prédiction, associer à chaque station une mesure pour les données météorologiques.

D'une manière plus générale, ces études ont démontré que la modélisation de la qualité de l'air constitue un outil d'exploration efficace des moyens possibles à mettre en œuvre pour une amélioration de la qualité de l'air. Ainsi, les effets des normes d'émission, des aménagements du territoire et des changements de combustibles et/ou carburant par exemple, peuvent être déterminés à priori et donc avant même l'application réelle de ces mesures.

Enfin, il faut noter que la ville constitue le pouvoir le plus rapproché du plus grand nombre de consommateurs finaux d'énergie, individus, institutions ou entreprises. Elle est donc en mesure de les inciter, pour des raisons économiques ou environnementales, à adopter des mesures d'efficacité

énergétique, soit par des actions directes(exemple: appui à la réglementation sur l'efficacité énergétique des bâtiments ou organisation du réseau de transport urbain), soit par des actions promotionnelles(documentation, conseils techniques, financement des recherches et études, etc.)

Le développement durable des ville implique de nouvelles relations de partenariat entre les collectivités territoriales et l'État avec la participation de la société civile dans des stratégies conjointes à moyen et à long terme. Il conduit aussi à des changements dans le travail interdisciplinaire, à la révision de certaines dispositions législatives et réglementaires et à la décentralisation. Il requiert la révision des systèmes de mobilisation de ressources financières au niveau international (par exemple la **COP21**), national ou local et particulièrement pour élargir l'accès aux différentes formes d'énergie. Celle-ci joue un rôle primordial dans le développement des villes, l'habitat, le développement économique, le transport et la lutte contre la pauvreté. Les responsables des villes sont donc appelés à intervenir fortement au niveau de l'approvisionnement et de la maîtrise des consommations d'énergie.

Annexe A

Code matlab

```
% Solution and error of the transport equation
% u_t + a * u_x = 0
% x ∈ I = [a, b] and t ∈ [0, T] with the
% initialdata u(x, 0) = u_0(x) with different methods
% Euler forward for time cebtralforspace
% Upwind
% Author : Lebede Ngartera
% Date : september 12, 2013
% Institution : University Cheikh Anta Diop Dakar (Ucad),
% Thesis doctorat Applied
Analysis and Numerical Mathematics,
% Numerical Mathematics Airquality modeling
clear all; close all;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Parameters
%%%%%%%%%%%%%%%%%%%%%

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Initial data ('sin(pi * x)') ;
u0 = inline('sin(pi * x)');
% exact solution for computing theerror
u0T =' sin(pi * (x - a * T))';
% velocity
a = 2 ;
% Intervall
T = 1 ;
I = [0 1] ;
% Number of time steps
TT(1) =7000 ;
% Number of space steps
XX(1) = 100;
flag = menu('Please select a method',...
'Euler-forward for time / central for space',...
'Lax-Friedrichs',...
'Lax-Wendroff',...
'Upwind',...
```

```
'Errorplot');
disp('Please wait ...');
switch flag
case 1
errsteps = 1;
case 2
errsteps = 1;
case 3
errsteps = 1;
case 4
errsteps = 1;
case 5
% Discretization for first step of the error
TT(1) = 5;
XX(1) = 2;
% Number of refinement steps for error plot
errstepseuler = 5;
errsteps = 7;
end
% Euler forward for time / central for space
TTeuler(1) = TT(1);
XXeuler(1) = XX(1);
if flag==1
errstepseuler = 1;
end
if flag==1 | flag==5
for k = 1:errstepseuler
NT = TTeuler(k);
NX = XXeuler(k);
dt = T/NT;
dx = (I(2)-I(1))/NX;
lambda = dt/dx;
% Initial conditions
SOL = zeros(NX+2*NT+1,2);
SOLPLOT = zeros(NX+1,NT+1);
for j = 1:NX+2*NT+1
SOL(j,1) = u0(I(2)+(NT+1-j)*dx);
end
SOLPLOT(:,1) = SOL(NT+1:NT+NX+1,1);
% Finite Difference-Scheme: Euler-forward in
time / central in space
for n = 1:NT
for j = 1+n:NX+2*NT+1-n
SOL(j,2) = SOL(j,1) - 0.5*lambda*a*(SOL(j-1,1)-SOL(j+1,1));
end
SOL(:,1) = SOL(:,2);
SOLPLOT(:,n+1) = SOL(NT+1:NT+NX+1,1);
end
```

Annexe B

Santé et pollution atmosphérique

Polluants atmosphériques	Sources principales	Effets sur la santé
Dioxyde de soufre ou SO_2	Combustions de combustibles fossiles (fioul, charbon, lignite, gazole,...) contenant du soufre. La nature émet aussi des produits soufrés (volcans).	Irritation des muqueuses de la peau et des voies respiratoires supérieures (toux, gène respiratoire, troubles asthmatiques).
Métaux lourds plomb (Pb), mercure (Hg), arsenic (As), cadmium (Cd), nickel (Ni)	Proviennent de la combustion des charbons, pétroles, ordures ménagères mais aussi de certains procédés industriels (production du cristal, métallurgie, fabrication de batteries électriques). Plomb : principalement émis par le trafic automobile jusqu'à l'interdiction totale de l'essence plombée (01/01/2000).	S'accumulent dans l'organisme, effets toxiques à plus ou moins long terme, Affectent le système nerveux, les fonctions rénales hépatiques, respiratoires,...
Hydrocarbures aromatiques polycycliques (HAP) et composés organiques volatils (COV)	Combustions incomplètes, utilisation de solvants (peintures, colles) et de dégraissants, produits de nettoyage, remplissage de réservoirs automobiles, de citernes,...	Effets divers selon les polluants dont irritations et diminution de la capacité respiratoire, Considérés pour certains comme cancérogènes pour l'homme (benzène, benzo-(a)pyrène), Nuisances olfactives fréquentes.
Monoxyde de carbone ou CO	Combustions incomplètes (gaz, charbon, fioul ou bois), dues à des installations mal réglées (chauffage domestique) et provenant principalement des gaz d'échappement des véhicules.	Intoxications à fortes teneurs provoquant maux de tête et vertiges (voir le coma et la mort pour une exposition prolongée). Le CO se fixe à la place de l'oxygène sur l'hémoglobine du sang.
Ozone ou O_3	Polluant secondaire, produit dans l'atmosphère sous l'effet du rayonnement solaire par des réactions complexes entre certains polluants primaires (NO_x, CO et COV) et principal indicateur de l'intensité de la pollution photochimique.	Gaz irritant pour l'appareil respiratoire et les yeux, Associé à une augmentation de la mortalité au moment des épisodes de pollution (Étude ERPURS/ORS Ile-de-France).
Oxydes d'azote ou NO_x (NO et NO_2)	Toutes combustions à hautes températures de combustibles fossiles (charbon, fioul, essence,...). Le monoxyde d'azote (NO) rejeté par les pots d'échappement s'oxyde dans l'air et se transforme en dioxyde d'azote (NO_2) qui est à 90% un polluant "secondaire"	NO_2 : gaz irritant pour les bronches (augmente la fréquence et la gravité des crises chez les asthmatiques et favorise les infections pulmonaires infantiles), NO non toxique pour l'homme aux concentrations environnementales.
Particules ou poussières en suspension (PM)	Combustions industrielles ou domestiques, transport routier diesel, origine naturelle (volcanisme, érosion,...). Classées en fonction de leur taille : .PM_{10}: particules de diamètre inférieur à $10\mu m$ (retenues au niveau du nez et des voies aériennes supérieures) .$PM_{2,5}$: particules de diamètre inférieur à $2,5\mu m$ (pénètrent profondément dans l'appareil respiratoire jusqu'aux alvéoles pulmonaires)	Irritation et altération de la fonction respiratoire chez les personnes sensibles, Peuvent être combinées à des substances toxiques voire cancérigènes comme les métaux lourds et des hydrocarbures, Associées à une augmentation de la mortalité pour causes respiratoires ou cardiovasculaires (ERPURS/ORS Ile-de-France).

TABLE B.1 – Tableau récapitulatif des principaux polluants (Airparif).

Annexe C

Procédure itérative de Box-Jenksen

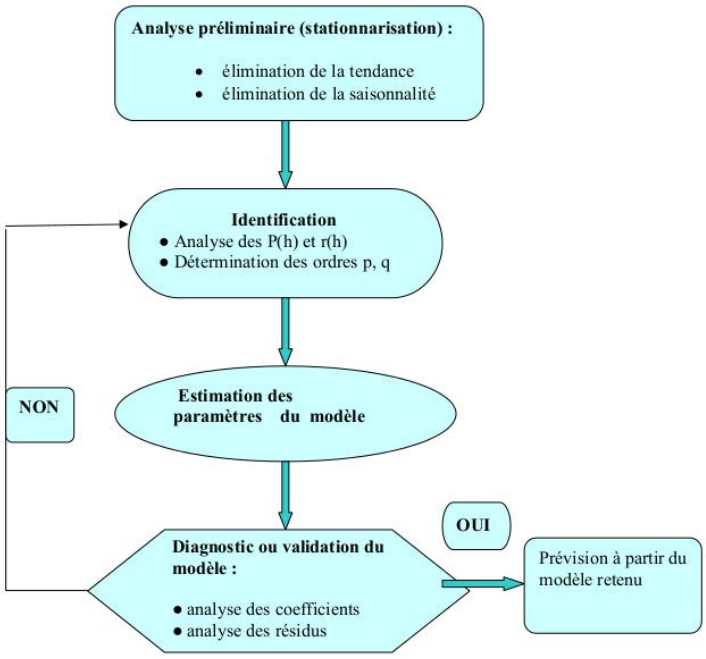


FIGURE C.1 – procédure itérative de Box-Jenkins (Bourbonnais, 1998)

PM_{10}	Norme sénégalaise ($\mu g/m^3$)		Maximum des concentrations moyennes sur 24 h ($\mu g/m^3$)				
	Valeur Limite	Nb de dépassements autorisés	Bel Air	Bd. République	Médina (Abass Ndao)	HLM	Yoff
Moyenne 24 h	260	1 fois / an	243	247	288	290	-
Nb de dépassements observés			0	0	2	2	-

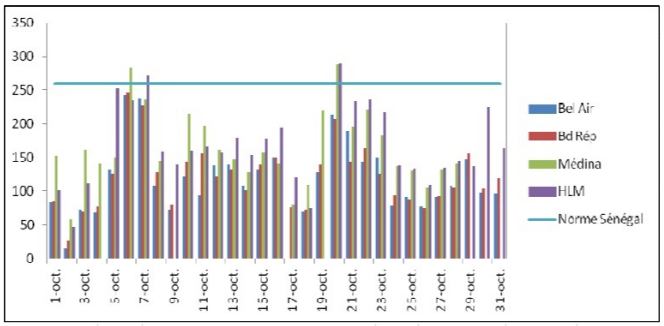


FIGURE C.2 – Concentrations moyennes journalières de PM_{10} à Dakar en octobre 2011: on note que la moyenne journalière des concentrations de PM_{10} a dépassé le seuil de 260 g/m^3, fixé par la norme NS-05-062 au cours de ce mois. Deux dépassements ont été observés à la Médina et aux HLM. [69]

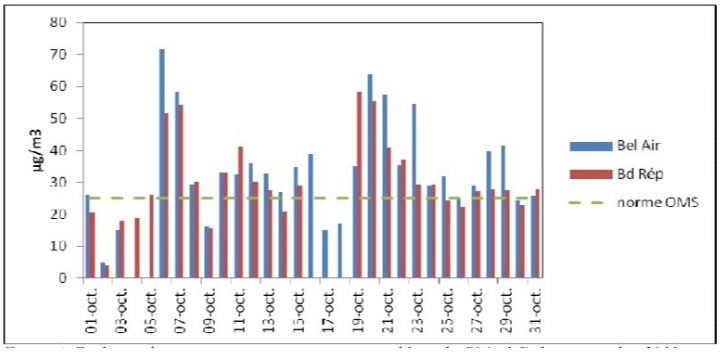


FIGURE C.3 – Evolution des concentrations moyennes journalières de $PM_{2,5}$ à Dakar en octobre 2011: Les concentrations moyennes journalières de $PM_{2,5}$ ont été élevées, car la valeur guide de l'OMS ($25\mu g/m^3$) a été dépassée 19 fois au Boulevard de la République et 23 fois à Bel Air, [69]

NO_2	Norme sénégalaise ($\mu g/m^3$)		Maximum des concentrations moyennes horaires ($\mu g/m^3$)				
	Valeur Limite	Dépassements autorisés	Bel Air	Bd. République	Médina (Abass Ndao)	HLM	Yoff
Moyenne 1 h	200	1 fois / mois	-	46	46	35	28

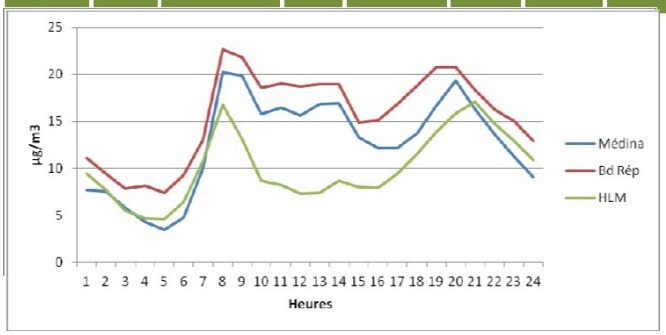


FIGURE C.4 – Évolution diurne des concentrations horaires de NO_2 en octobre 2011 à Dakar: Concernant le NO_2, les concentrations moyennes horaires n'ont pas dépassé la valeur limite de la norme NS-05-062 en octobre 2011. Les plus fortes concentrations ont été mesurées au Boulevard de la République (Cathédrale, 46 $\mu g/m_3$) et à Médina (Abass Ndao, 46 $\mu g/m_3$). Nous remarquons que l'évolution diurne montre deux maxima observés à 8h et à 20h, ce qui traduit l'influence des activités humaines, notamment le transport, sur la pollution au dioxyde d'azote. La baisse des concentrations en cours de journée pourrait être liée à la formation de l'ozone, suite à l'interaction du NO_2 avec le rayonnement solaire [69].

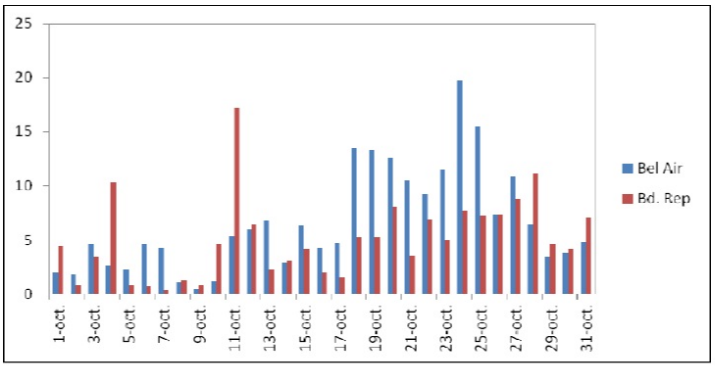


FIGURE C.5 – Evolution des concentrations moyennes journalières de SO_2 à Bel Air (Môle 10) et au Bd République en octobre 2011: Les concentrations moyennes journalières ont été faibles et n'ont jamais dépassé la limite de 125 $\mu g/m_3$ au cours de ce mois. Le maximum des moyennes journalières a été de 20 $\mu g/m_3$ à Bel Air, contre 11 $\mu g/m_3$ au Boulevard de la République.[69]

O_3	Norme sénégalaise (µg/m^3)		Maximum des concentrations moyennes sur 8 heures (µg/m^3)		
	Valeur Limite	Dépassements autorisés	Bd. République	Yoff	HLM
Moyenne 8 h	120	2 fois / mois	28	39	24

FIGURE C.6 – Concentrations moyennes horaires maximales d'ozone à Dakar en octobre 2011: l'ozone est mesuré dans trois sites, Boulevard de la République, HLM et Yoff. Les concentrations sont restées inférieures à la valeur fixée par la norme NS-05-062, pour une durée d'exposition de 8 heures (120 $\mu g/m_3$) [69].

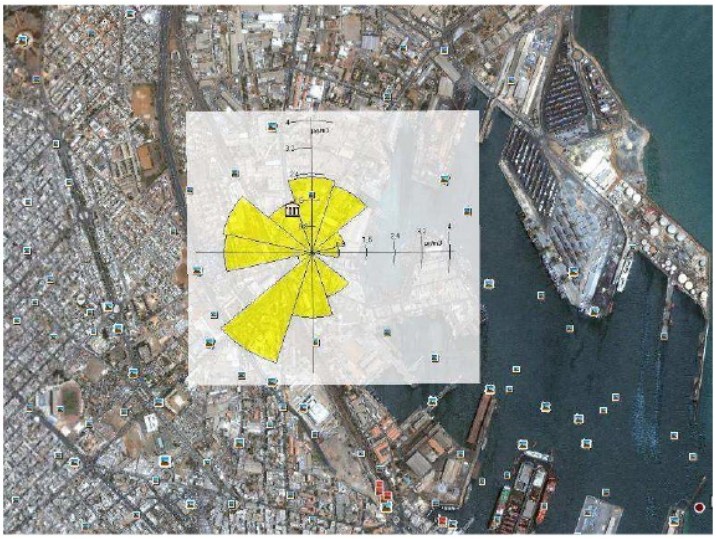


FIGURE C.7 – Identification des zones sources du SO_2 mesuré à Bel Air et au Bd de la République en septembre 2011 (image Google Earth): La rose de concentration de SO_2 indique que le SO_2 mesuré à Bel Air provient principalement de l'avenue Malick Sy, et secondairement de la gare routière des Pompiers, de la zone industrielle et du Boulevard de la Libération. C'est donc le trafic qui est à l'origine de l'essentiel du SO_2 mesuré sur le site du Port en ce mois.

Bibliographie

[1] Tarik J. Glenn T. Nada O. 7 millions de décès prématurés sont liés á la pollution de l'air chaque année. Technical report, OMS, http://www.who.int/mediacentre/news/releases/2014/air-pollution/fr/, 2014.

[2] Moussiopoulos N. SAHM P. KESSLER C. Numerical simulation of the photochimical smog in athens greece. a. case study. *Atmospheric Evironment*, 29(24):3619, 1995.

[3] Moussiopoulos N. BERGE E. BOHLER T. and all. Ambiant air quality dispersion and transport models. european environment agency. Technical Report 19/96, 1996.

[4] HANNAH S.R. MOOR G.E.and FERNAN M.E. Evaluation of photochimical grid models (uam iv, uam-v and rom-uam iv couple) using data from lake michigan ozone study (lmos). *Atmospheric Environment*, 30:3265–3281, 1996.

[5] Moussiopoulos Nicolas. *Air quality in Cities*. Springer Verlag Berlin Heidlberg News York, ISBN 3-540-00842-X, 2003.

[6] B. Sportisse and Du Bois L. Numerical and theoretical investigation of a simplified model for the parameterization of below-cloud scavenging by falling raindrops. Technical Report 36, Atmos. Env., 2002.

[7] Underwood B. Review of deposition velocity and washout coefficient. Technical report, AEA Technology, Harwell, 2001.

[8] C. Belot, Y. Caput and J. Guenot. Étude bibliographique du lavage par la pluie des radionucléides particulaires et gazeux émis en situation accidentelle. Technical report, IRSN et EDF., 1988.

[9] Moeng C. and J. Wyngaard. Statistics of conservative scalars in the convective boundary. *layer. J. Atmos. Sci.*, 41:3161–3169, 1984.

[10] Mallet V. *Estimation de l'incertitude et prévision d'ensemble avec un modèle de chimie transport. Application a la simulation numérique de la qualité de l'air.* PhD thesis, ENPC, 2005.

[11] L. Mejean P. Perkins, R. Soulhac and I. Rios. Modélisation de la dispersion des émissions atmosphériques d'un site industriel vers un guide de l'utilisateur. 1ère partie: état de l'art. Technical report, Rapport technique, Laboratoire de Mécanique des Fluides et d'Acoustique. UMR CNRS 5509-UCBL1- ECL., 2002.

[12] Sorensen J. Sensitivity of the derma long-range gaussian dispersion model to meteorological input and diffusion parameters. *Atmos. Env.*, 32(24):4195–4206, 1998.

[13] I. Korsakissok. *Changements d' échelle en modélisation de la qualité de l'air et estimation des incertitudes associées.* PhD thesis, Université Paris-Est, Namur, Belgium, Décembre 2009.

[14] Arya S. Air pollution meteorology and dispersion. *Oxford University press*, 1999.

[15] IAGU. Résumé du rapport géo ville région de dakar. Technical report, `http://www.iagu.org/PDF/resume_rapport_geodakar.pdf`, 2007.

[16] R. Tremblay and E. H. Mamadou Ndiaye. Les caractéristiques du secteur du transport urbain à dakar. *Perspective Afrique* *http://www.perspaf.org/index.php?id=95*, 3(1-3), 2008.

[17] afsse. Impact sanitaire de la pollution atmosphérique urbaine. Technical Report 2, Agence française de sécurité sanitaire environnementale, 94704 Maisons-Alfort Cedex - Tél. +33 (0)1 56 29 19 30, 2004.

[18] Association Sénégalaise de Normalisation (ASN). Normes de rejets. Technical Report NS05-062, `www.denv.gouv.sn/index.php?option=com_joomdoc&task=cat_view&gid=90&Itemid=59`, 2001.

[19] Centre de Gestion de la Qualité de l'Air (CGQA) Dakar. Comment fonction l'iqa. Technical report, `www.air-dakar.org/iqa/comment-fonctionne-liqa.html`, 2010.

[20] C.Honoré and all. Predictability of european air quality : Assessment of 3 years of operational forecasts and analyses by the prev'air system. *Journal of Geophysical Research*, 113, 2008.

[21] Mbaye D. and Aminata M.Diokhané. Suivi de la qualité de l'air à dakar. Technical Report 03, 2012.

[22] J.F. Bonnans-J.C. Gilbert C. Lemaréchal. *Optimisation numérique. Aspects théoriques et pratiques*, volume 27. Collection Mathématiques et Applications, 1998.

[23] M. Duflo. *Algorithmes stochastiques, Mathématiques et Applications 23*, volume 23. 1996.

[24] Georges Skandalis. *Topologie et analyse 3ème année*. 2004.

[25] Jorge Nocedal and Stephen J. Wright. Numerical optimization. *Springer*, second edition:2006.

[26] Jean-Baptiste Hiriart-Urruty. *Optimisation mathématique / Théorie des jeux. optimisation mathématique.* FR, France, ISBN 2-13-047981-2, 1996.

[27] R. Tyrrell Rockafellar. *Convex Analysis (Princeton Landmarks in Mathematics and Physics)*. ISBN 0-691-08069-0 (cloth), 1997.

[28] H. Cartan. *Cours de calcul différentiel*. 1997.

[29] F. Ecoto. *Initiation a la Recherche Opérationnelle*. 1986.

[30] H. Brezis. *Analyse fonctionnelle*. Collection Mathématiques Appliquées pour la Maîtrise, 1987.

[31] P.G. Ciarlet. *Introduction à l'analyse numérique matricielle et à l'optimisation*. 1988.

[32] Lebede Ngartera S. Diagne and Y. Gningue. Modeling and prediction of dakar air quality index. *IJAMAS (International Journal of Applied Mathematics and Statistics)*, 54(2):42–53, 2016.

[33] Valérie Nollet Yannick Flandrin Corinne Schadkowski, Jean Claude Dechau. Introduction à la modélisation de la qualité de l'air. *Air Pur*, pages 5–8, 2002.

[34] Goyal P. Chan A.T & Jaiswal N. Statistical models for the prediction of respirable suspended particulate matter in urban cities. *Atmospheric Evironment,40,2068-2077*, 40:2068–2077, 2006.

[35] Perez P. and Salini G. $pm_{2.5}$ forecasting in a large city: Comparaison of three methods. *Atmospheric Environment*, 42(8219-24), 2008.

[36] Moussiopoulos N. Slini Th., Karatzas K. Pm_{10} forecasting for thessaloniki, greece. *Environmental Modelling & Software, 21:559-65.*, 2006.

[37] Ortega JC. Fu et al. Diaz-Roblès LA. A hybrid arima and artificiel neural networks model to forcast particulate matter in urbain areas of temuco, chile. *Atmospheric Environment, 42: 8331-40.*, 2008.

[38] T. Stocker. On the asymptotic bias of ols in dynamic regression models with autocorrelated errors. *Statist. Papers*, 48:81–93, 2007.

[39] P.-A. Cornillon and E. Matzner-Lober. Régression avec r. *Springer*, 2010.

[40] Davison A. C. and Diego Kuonen. "an introduction to the bootstrap with applications in r." statistical computing and statistical graphics. *Newsletter*, 13:6–11, 2002.

[41] Bear J. Bachmat Y. Introduction to modeling of transport phenomena. *Dordrecht: Kluwer*, 1990.

[42] D. Cochrane and G. H. Orcutt. Application of least squares regression to relationships containing auto-correlated error terms. *Journal of the American Statistical Association*, 44:32–61, 1946.

[43] J.D. Imbens G.W.and Rubin D.B. Angrist. Identification of causal effects using instrumental variables (with comments). *Journal of the American Statistical Association*, 91:434–444, 1996.

[44] Granger C. W. J. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969.

[45] Anscombe F. Graphs in statistical analysis. *The American Statistician*, 27, n°1:17–21, 1973.

[46] Robert H. Shumway and David S. Stoffer. *Time Series Analysis and Its Applications. With R Examples*. Springer New York Dordrecht Heidelberg London, DOI 10.1007/978-1-4419-7865-3, 2011.

[47] Brockwell PJ and Davis RA. *Time series in statistics*. Springer Verlag, 1st ed. New York, 1987.

[48] Droesbeke JJ. Fichet B and Tassi P ed. *Séries chronologiques. Thorie et pratique des modèles ARIMA*. Paris: Economica, 1989.

[49] Gourieroux C. Monfort A. *Séries temporelles et modèles dynamiques*. 1st ed. Paris : Economica, 1990.

[50] Coutrot B. Droesbeke JJ. *Les méthodes de prévision*. 2nd ed. Paris : Presses universitaires de France (Que sais-je ?), 1990.

[51] Giraud R. and Chaix N. *Économétrie*. 2nd ed. Paris : Presses universitaires de France, 1994.

[52] Bresson G. and Pirotte A. *Économétrie des séries temporelles. Théorie et applications*. 1st ed. Paris : Presses universitaires de France, 1995.

[53] Peter J. Brockwell and Richard A. Davis. *Introduction to Time Series and Forecasting*. Springer, Second Edition, ISBN 0-387-95351-5 (alk. paper), 2002.

[54] Peter J. Brockwell and Richard A. Davis. *Time Series ; Theory and Methodes*. Springer Verlag News York, inc, ISBN 0-387-96406-1, 1987.

[55] Peter J. Brockwell and Richard A. Davis. *Time Series : Theory and Methods*. Springer, 2009.

[56] Peter J. Brockwell Richard and A. Davis. *Introduction to Time Series and Forecasting, Second Edition*. Springer, ISBN 0-387-95351-5, 2002.

[57] Lebede Ngartera S. Diagne and Y. Gningue. Optimization and analysis of the index of air quality in dakar by the process auto regressive moving average arma(2,1). *IJAM (International Journal of Applied Mathematics)*, 28(5):621–636, 2015.

[58] A. Fiordaliso. Systèmes flous et prévision de séries temporelles, paris. *Hermès Science*, 1999.

[59] Abadie and Meslier. Etude de l'utilisation des modèles arima pour la prévision très court terme de l'énergie journalière produite par Électricité de france. 13(1):37–54, 1979.

[60] K. P. Burnham and D. R. Anderson. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer-Verlag, ISBN 0-387-95364-7, 2002.

[61] K. P. Burnham and D. R. Anderson. *Multimodel inference: understanding AIC and BIC in Model Selection*. Sociological Methods and Research, 2004.

[62] C. M. Hurvich and C.-L. Tsai. Regression and time series model selection in small samples. *Biometrika 76*, 76:297–307, 1989.

[63] Speed T. P. Moussiopoulos and Yu B. Model selection and prediction: normal regression. *Annals of the Institute of Statistical Mathematics 1*, pages 35–36, 1993.

[64] M Dacunha-Castelle, D. et Duflo. *Probabilités et statistiques. Tome 1: Problèmes à temps fixe et Tome 2: Problèmes à temps mobile. Collection Mathématiques Appliquées pour la Maîtrise,*. Masson, 1983.

[65] D Azencott, R. et Dacunha-Castelle. *Séries d'observation irrégulières*. Masson, Paris, 1984.

[66] Amemiya T. *Advanced Econometrics. Cambridge*. MA: Harvard University Press., 1985.

[67] J.D. Hamilton. *Time series analysis.* Time series analysis., 1994.

[68] Gourieroux C. and Montfort A. *Séries temporelles et modèles dynamiques.* Economica, 1995.

[69] Mbaye D. and M.Diokhané. Suivi de la qualité de l'air à dakar. Technical Report 9,10/2011, CGQA, 13 octobre et 15 novembre 2011.