ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Modélisation temporelle de la concentration d'ozone.

Claudio Semadeni

Responsable du projet de semestre :
Prof. Stephan Morgenthaler

Assistant :
Thomas Gsponer

Juillet 2001

Table des matières

Introduction

L'ozone. Ce gaz particulier constitue d'une part une aubaine pour les être vivants de notre planète car sa présence dans la stratosphère nous protège des radiations UV mortelles émises par le soleil. Mais il constitue d'autre part un danger polluant lorsqu'il se trouve à la surface de la terre en trop haute concentration.

Dans le cadre de ce travail, nous allons nous intéresser à l'évolution du taux d'ozone à la surface de la terre et plus particulièrement aux mesures effectuées par le réseau suisse NABEL.

Le but de ce travail est d'effectuer des prédictions des taux d'ozone pour une station de mesure. Cela pourrait permettre de connaître à l'avance les risques de dépassement des taux limites imposés par l'Ordonnance sur la protection de l'air (OPair).

Après une brève présentation à propos de l'ozone et des réactions chimiques qui permettent sa création (chapitre 1) ainsi qu'une présentation du jeu de données (chapitre 2), nous effectuons une brève exploration des données (chapitre 3). Les chapitres 4 et 5 étant ensuite dédiés au traitement des valeurs manquantes dans le jeu de données et aux aspects de la prédiction.

Chapitre 1

L'ozone

Découvert en 1840 par le physicien allemand Schönbein, l'ozone est un gaz bleuté qui dégage une odeur âcre et irritante. L'ozone, ou trioxygène, est comme son nom l'indique formé de trois atomes d'oxygène (formule chimique : O_3). Les molécules d'ozone sont peu présentent dans l'atmosphère à la différence des autres constituants chimiques tels que le dioxygène ou l'azote. L'ozone est un gaz naturel qui est présent dans la haute atmosphère (*stratosphère*) et au niveau du sol (*troposphère*).
L'intéêt porté par la communauté scientifique à l'égard de l'ozone s'est accru au cours des dernières années à cause de l'activité humaine qui modifie de manière importante et dangereuse les taux naturels de ce gaz dans l'atmosphère.

1.1 Ozone stratosphérique

L'ozone stratosphérique se trouve essentiellement à une altitude comprise entre 15 et 20 km et joue un rôle capital pour les plantes et les êtres vivants, car il protège la surface terrestre des radiations UV qui seraient mortelles. C'est pour cette raison que certains utilisent parfois le terme de *bon ozone*.

Contrairement à l'oxygène, dont la formule chimique est O_2, l'ozone est très instable et réagit avec d'autres molécules présentes dans l'atmosphère. Sa durée de vie n'est donc que de quelques heures ou de quelques jours. Cependant, les polluants qui entrent dans la stratosphère provoquent un amincissement de la couche d'ozone, ce qui provoque une augmentation de l'intensité du rayonnement UV à la surface de la terre.

1.2 Ozone troposphérique

L'ozone troposphérique correspond à environ 10% de l'ozone contenu dans l'atmosphère, ce qui paraît peu. Toutefois, sa présence dans l'air proche du sol et le fait que les êtres humains le respire lui confère une grande importance. En effet, au delà d'une certaine concentration, il joue le rôle d'un polluant dangereux pour les êtres vivants ainsi que pour les plantes.
La présence de l'ozone troposphérique a été attribuée dans un premier temps à des transferts dynamiques d'ozone stratosphérique. En fait, seulement 10% de l'ozone troposphérique provient de la stratosphère alors que les 90% restants se forment à proximité du sol. Les oxydes d'azote, le monoxyde de carbone ainsi que le radical OH jouent un rôle important

dans le processus de formation.

Les oxydes d'azote NO et NO_2 qui résultent de la combinaison d'oxygène et d'azote sont des éléments très répendus dans la troposphère. Les réactions qui conduisent à ces oxydes sont soit d'origine naturelle (orages, incendies de forêts, etc.) soit induites par les activités humaines (combustion des hydrocarbures pour le transport ou le chauffage). La photodissociation du dioxyde d'azote en présence de rayonnement solaire de courte longueur d'onde constitue une source possible d'ozone :

$$NO_2 \rightarrow NO + O$$

et en présence d'autres oxydants on a :

$$O + O_2 \rightarrow \mathbf{O_3}$$

sinon NO_2 se reforme :

$$NO + O_3 \rightarrow NO_2 + O_2$$

D'autre part, le monoxyde de carbone ou gaz carbonique (formule chimique : CO_2), qui résulte en grande partie des activités humaines (combustion des hydrocarbures), contribue à la formation de l'ozone. Les processus de formation de l'ozone sont dans leur ensemble très compliqués, mais il est possible de schématiser la production d'ozone par l'équation de réaction suivante :

$$CO + 2O_2 \rightarrow CO_2 + \mathbf{O_3}$$

avec les oxydes d'azotes NO et NO_2 comme catalyseurs.
Les sites de forte activité industrielle et de circulation automobile intense sont donc propices à la formation de l'ozone car la production d'oxydes d'azote et d'oxydes de carbone y est importante ; des conditions météorologiques particulières comme les anticyclones pouvant favoriser sa création. D'autre part, notons que les concentrations de O_3 ont tendance à diminuer la nuit car le rayonnement solaire est indispensable à la formation de l'ozone, les concentrations de O_3 ont tendance à diminuer la nuit.

La pollution par l'ozone affecte surtout les grands centres urbains et leur périphérie, ce qui justifie la mise en place d'un dispositif de surveillance quotidienne de la composition chimique de l'atmosphère ainsi que des campagnes de mesures pour approfondir la connaissance de ces phénomènes.
D'une manière générale, la concentration d'ozone a tendance à augmenter dans la troposphère, mais pas seulement dans les zones urbanisées, car la déforestation tropicale a dans ce domaine les mêmes effets que la circulation automobile.

1.3 Ozone technique

En plus de l'ozone stratosphérique et troposphérique qui se forment de manière plus ou moins naturelle, il existe l'ozone dit *technique* qui est généré par l'homme de manière volontaire pour ses propriétés d'oxydant. En effet, l'ozone est un gaz oxydant très fort ; beaucoup plus fort que le chlore. Les principales applications de l'ozone *technique* sont :

- le traitement de l'eau, car c'est un excellent substitut du chlore ;

- le traitement des odeurs dans l'air ;
- le contrôle de qualité des matières plastiques (tests de vieillissement).

L'ozone suscite souvent des craintes lors de ses utilisations, car il est connu comme étant un polluant. Mais il faut mentionner que l'ozone est beaucoup moins dangereux que les produits qu'il remplace comme en particulier le chlore.

L'ozone est donc un gaz atmosphérique unique : il est fort bénéfique dans la haute atmosphère, mais constitue un danger polluant au niveau du sol ; on parle parfois du "bon" et du "mauvais" ozone. Ironiquement les activité anthropiques ont entraîné une réduction des quantités d'ozone dans la haute atmosphère et une augmentation de sa concentration au niveau du sol.

Chapitre 2

Présentation des données

2.1 Le réseau NABEL

Le jeu de données qui est utilisé dans ce travail provient du réseau national d'observation des polluants de l'air (NABEL). Le réseau NABEL effectue entre autre des mesures météorologiques (direction et vitesse du vent, température, etc.) et des mesures sur les gaz atmosphériques. Le réseau NABEL a été mis en route en 1979 ; il a été modernisé entre 1989 et 1991 pour passer de 8 à 16 stations qui sont réparties dans l'ensemble du pays (figure 2.1).

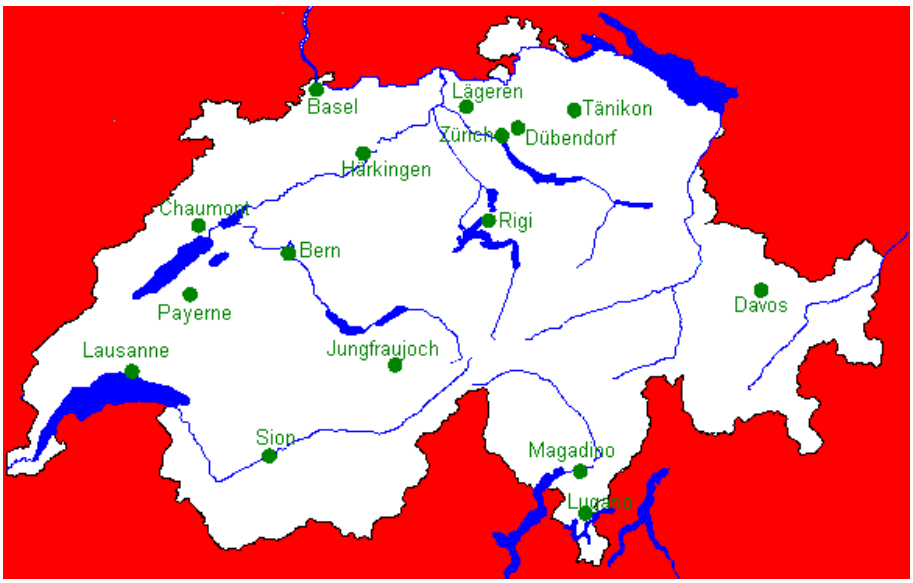

FIG. 2.1 – *Répartition actuelle des stations de mesure du réseau NABEL. Source : site internet de l'Office fédéral de l'environnement, des forêts et du paysage (OFEFP).*

Etant donné que la pollution de l'air varie fortement d'un endroit à l'autre, suivant l'emplacement géographique et les sources d'émission qui se trouvent à proximité, les stations sont situées dans des endroits typiques. Le tableau 2.1 présente la répartition des stations de

mesure du réseau NABEL. Une telle répartition permet de donner un aperçu des différents degrés de pollution allant de très faible à très élevé.

TAB. 2.1 – *Type d'emplacement des différentes stations du réseau NABEL*

Type d'emplacement	**Stations**
Centre-ville, route à grand trafic	Berne, Lausanne
Centre-ville, parc	Lugano, Zurich
Agglomération	Bâle-Binningen, Dübendorf, Magadino
Zone rurale, bordure d'autoroute	Härkingen, Sion
Zone rurale, altitude inférieure à 1000m	Lägeren, Payerne, Tänikon
Zone rurale, altitude supérieure à 1000m	Chaumont, Davos, Rigi
Haute montagne	Jungfraujoch

Les gaz polluants qui sont mesurés dans la plupart des stations de manière continue sont notamment : le gaz carbonique (CO_2), les oxydes d'azote (NO et NO_2), l'ozone (O_3) et l'anhydride sulfureux (SO_2).

2.2 Le jeu de données

Notre choix pour les mesures des taux d'ozone effectuées par la station lausannoise provient simplement du fait de la proximité de la station avec l'Ecole Polytechnique Fédérale de Lausanne (EPFL). La station lausannoise est située au centre ville, aux abords d'une route de transit (environ 30'000 véhicules par jour).

Les données dont nous disposons concernant les mesures effectuées par la station de Lausanne sont les taux quotidiens moyens d'ozone ($\mu g/m^3$) entre 1991 et 1999. Parmi les 3287 observations que le jeu de données devrait contenir, il manque 33 observations probablement à cause de pannes ou de révisions qui ont été effectuées sur les appareils de mesures. Notons que ces taux sont forcément positifs, la valeur la plus faible possible étant atteinte si l'air ne contient aucune molécule d'ozone. Il existe certainement une borne supérieure qui dépend de la pression pour le taux d'ozone , mais elle n'est très probablement pas atteinte dans le cas présent

Chapitre 3

Exploration des données

Etant donné que les observations représentent des mesures en fonction du temps qui sont espacées de manière régulière, nous allons utiliser l'approche des séries temporelles (ou séries chronologiques) pour les étudier. Les données constituent donc une suite de termes ordonnés par l'indice du temps $\{Y_t\}$ où Y_1 représente le taux moyen d'ozone du 1^{er} janvier 1991 et Y_{3287} représente le taux du 31 décembre 1999.

Comme 1992 et 1996 sont deux années bissextiles, nous avons choisi, pour des raisons de commodité, de transformer les données de la série initiale pour que chaque année comporte exactement 365 jours. Pour ce faire nous avons simplement supprimé de la série les deux termes provenant de 29 février et nous avons modifié les valeurs des taux des 28 février des deux années bissextiles en leurs assignant la moyenne des taux des 28 et 29 février.
Nous avons également envisagé la transformation suivante sur les taux des années bissextiles :

$$X_t = \frac{(366-t)Y_t + t \cdot Y_{t+1}}{366}, \quad 1 \leq t \leq 365$$

où Y_t représente les termes de la série initiale et X_t les termes de la nouvelle série. Cette transformation effectue la moyenne de deux observations consécutives en assignant des poids aux valeurs suivant la période de l'année d'où ces taux proviennent. La suite X_t contiendrait alors 365 termes. Mais nous avons renoncé à utiliser cette transformation car elle aurait beaucoup modifié les valeurs du milieu de l'année (mois de mai, juin, juillet et août) car les poids sont presque égaux.

Par la suite, nous allons utiliser la série dont les taux correspondant aux 29 févirer ont été retirés. La série comporte donc 3285 termes : $X_1, X_2, ..., X_{3285}$. Cette série est représentée dans la figure 3.1.
Selon l'Ordonnance sur la protection de l'air (OPair) qui règlemente les taux admissibles des gaz polluants en Suisse, 98% des moyennes semi-horaires d'un mois doivent être inférieures ou égales à $100\mu g/m^3$ et la moyenne horaire ne doit pas dépasser $120\mu g/m^3$ plus d'une fois par année. Ces deux limites sont également représentées dans la figure.

Avant d'aller plus avant dans l'exploration de cette série, nous introduisons quelques notions liées aux séries temporelles.

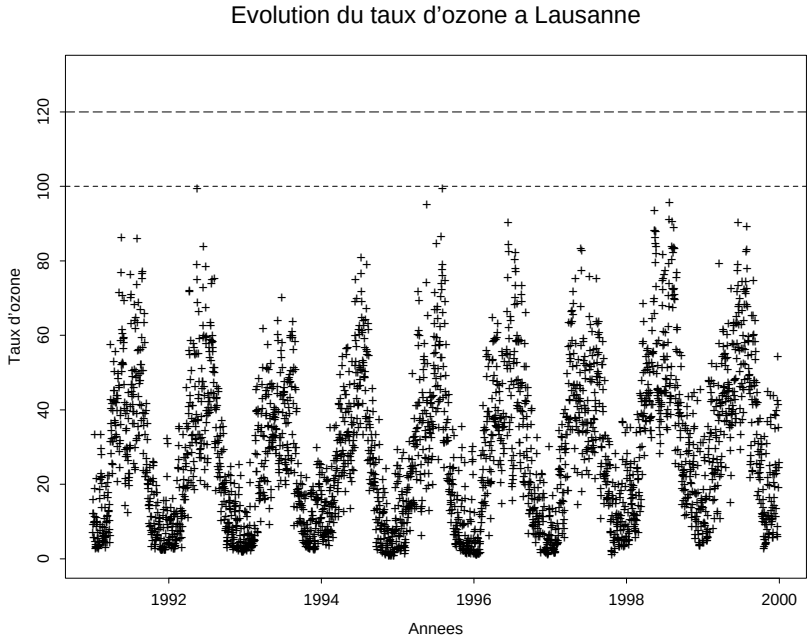

FIG. 3.1 – *Evolution de la moyenne quotidienne des taux d'ozone ($[\mu g/m^3]$) mesurés par la station lausannoise du réseau NABEL.*

3.1 Autocovariance, autocorrélation et stationnarité

Nous introduisons ici les notions d'autocovariance, d'autocorrélation et de stationnarité qui seront utiles pour la suite de ce travail.

Soit $\{X_t, t \in T\}$ une série temporelle où T est un sous-ensemble de $\mathbb{R}$, c'est-à-dire une suite d'observations d'une variable en fonction du temps effectuées à un rythme régulier, alors si $\{X_t,\ t \in T\}$ est tel que $\mathrm{Var}(X_t) < \infty \quad \forall t \in T$, on défini la *fonction d'autocovariance* $\gamma_X(\cdot,\cdot)$ de X_t par

$$\gamma_X(r,s) = \mathrm{Cov}(X_r, X_s)\ , \quad r,s \in T\ , \tag{3.1}$$

et la *fonction d'autocorrélation* est définie par

$$\rho_X(r,s) = \mathrm{Corr}(X_r, X_s)\ , \quad r,s \in T\ . \tag{3.2}$$

La série temporelle $\{X_t, t \in \mathbb{Z}\}$ est dite *stationnaire* ou plus souvent *faiblement stationaire* si :

$$\begin{array}{ll} \text{(i)} & \mathbb{E}({X_t}^2) < \infty\,, \quad \forall t \in \mathbb{Z} \\ \text{(ii)} & \mathbb{E}(X_t) = m\,, \quad \forall t \in \mathbb{Z} \\ \text{(iii)} & \gamma_X(r,s) = \gamma_X(r+t, s+t)\,, \quad \forall r,s,t \in \mathbb{Z} \end{array}$$

Remarquons que si $\{X_t, t \in \mathbb{Z}\}$ est stationnaire, alors $\gamma_X(r,s) = \gamma_x(r-s,0)$ pour tout $r,s \in \mathbb{Z}$. Sur la base des équation 3.1 et 3.2, il est alors commode de redéfinir les fonctions d'autocovariance et d'autocorrélation pour un processus stationnaire par les fonctions d'une

variable

$$\gamma_X(h) \equiv \gamma_X(h,0) = \mathrm{Cov}(X_{t+h}, X_t) \ , \quad \forall t, h \in \mathbb{Z} \ ,$$

et

$$\rho_X(h) \equiv \frac{\gamma_X(h)}{\gamma_X(0)} = \mathrm{Corr}(X_{t+h}, X_t) \ , \quad \forall t, h \in \mathbb{Z} \ .$$

La *fonction d'autocorrélation partielle* d'une série temporelle stationnaire est définie par

$$\alpha(1) = \mathrm{Corr}(X_2, X_1) = \rho(1) \ ,$$

et

$$\alpha(k) = \mathrm{Corr}(X_{k+1} - P_{\{1,X_2,\ldots,X_k\}}(X_{k+1}) \ , \ X_1 - P_{\{1,X_2,\ldots,X_k\}}(X_1)) \ , \quad k \geq 2 \ ,$$

où $P_{\{1,X_2,\ldots,X_k\}}X_{k+1}$ représente une projection. Se référer à [1] pour plus de détails. La valeur $\alpha(k)$ est appelée *autocorrélation partielle de différence k*.

La série présentée dans la figure 3.1 n'est donc manifestement pas stationnaire au sens de la définition ci-dessus.
Le caractère périodique de la série est très marqué. La période semble être d'une année et on peut remarquer que le taux quotidien moyen est faible en hiver et beaucoup plus élevé en été. Cela correspond parfaitement à ce que l'on pouvait s'attendre étant donné que le rayonnement solaire est nécessaire à la création de l'ozone (page 4).

Un moyen simple pour éliminer un effet périodique dans une série d'observations dont la période est connue est d'appliquer une moyenne mobile.

3.2 Moyenne mobile

Soit q un entier non négatif et n le nombre de terme du processus $\{X_t\}$, alors la *moyenne mobile bilatérale* (two-sided moving average) est le processus donné par :

$$W_t = \frac{1}{(2q+1)} \sum_{j=-q}^{q} X_{t+j} \ , \quad q+1 \leq t \leq n-q \tag{3.3}$$

Etant donné que la série n'est pas observée pour $t \leq 0$ et $t > n$, il n'est pas possible d'utiliser l'équation (3.3) pour $t \leq q$ ou $t > n-q$. Cependant, il est possible d'utiliser les familles de *moyennes mobiles unilatérales* (one-sided moving averages) du type :

$$W_t = \sum_{j=0}^{n-t} \alpha(1-\alpha)^j X_{t+j} \ , \quad 1 \leq t \leq q$$

et

$$W_t = \sum_{j=0}^{t-1} \alpha(1-\alpha)^j X_{t-j} \ , \quad n-q+1 \leq t \leq n$$

La construction de ces termes est très sensible à la valeur α. Il a été trouvé de manière empirique (voir Chatfield, 1974) que les valeurs de α comprises entre 0.1 et 0.3 donnent des lissages raisonnables.

Dans le but d'observer le comportement général de la série initiale, en supprimant l'aspect périodique de la série, nous nous contentons d'appliquer le lissage de la moyenne mobile bilatérale. Etant donné que nous avons remarqué une périodicité d'une année, nous utilisons le lissage de la moyenne mobile bilatérale avec $q = 182$ comme semi-largeur, c'est-à-dire une largeur totale de 365 jours. Le résultat est présenté dans la figure 3.2.

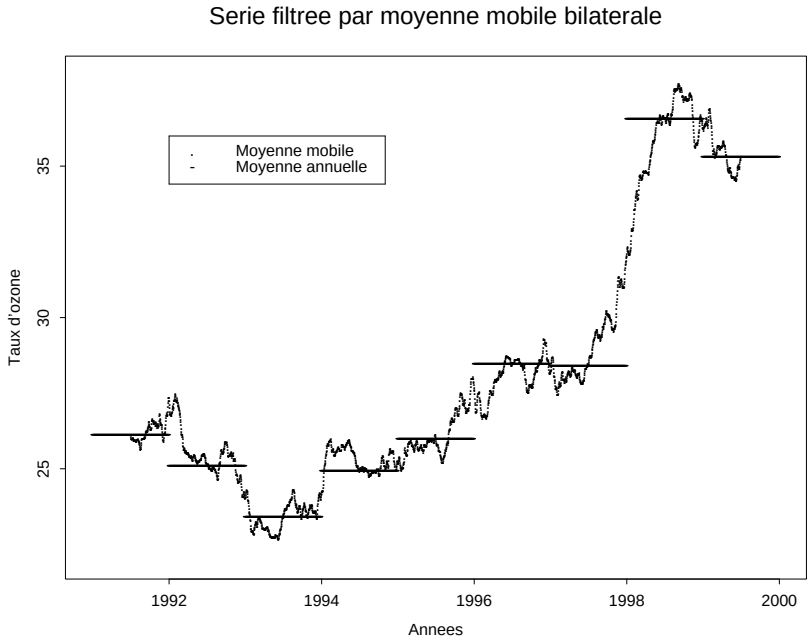

FIG. 3.2 – *Série initiale des taux d'ozone quotidiens ($[\mu g/m^3]$) filtrée par une moyenne mobile bilatérale ($q = 182$) et représentation des moyennes annuelles.*

La série filtrée ainsi obtenue n'est pas "plate" et l'évolution des moyennes annuelles le montre également. On peut donc raisonnablement penser qu'il existe une tendance à la hausse dans la série initiale. L'augmentation générale du taux d'ozone apparue depuis 1993 peut probablement s'expliquer par l'augmentation de l'ampleur du parc automobile suisse ; ce qui impliquerait une augmentation de l'émission de gaz carbonique qui favorise la création d'ozone (section 1.2).

Une approche classique consisterait à utiliser notamment les représentations des autocorrélations et des autocorrélations partielles ainsi que l'analyse harmonique. Ces approches seront traitées plus tard car il n'est pas possible d'utiliser ces outils sur la série initiale, étant donné qu'il manque des observations.

Dans ce qui suit nous allons donc tenter de construire les valeurs manquantes de manière à pouvoir disposer d'une série "sans trous".

Chapitre 4

Traitement des valeurs manquantes

Pour tenter de compléter la série intitiale dont il manque 33 valeurs, une première approche pourrait consister à reconstruire chaque valeur en utilisant une moyenne de valeurs qui l'entoure éventuellement avec des poids. Nous avons renoncé à utiliser cette méthode pour deux raisons. Premièrement, il peut arriver qu'il y ait plusieurs valeurs manquantes qui se suivent. Que faire dans ce cas? Deuxièmenent, si la valeur manquante devait se situer dans une période estivale, c'est-à-dire dans un pic, une approximation par des moyennes sous-estimerait très certainement la vraie valeur. Pour ces raisons, nous avons choisi d'utiliser la méthode qui est présentée dans la section suivante.

4.1 Tendance et saisonnalité

Le modèle classique de décomposition d'une série qui contient une tendance générale et un effet saisonnier est le suivant :

$$X_t = m_t + s_t + Z_t \,,$$

où m_t est une fonction qui varie lentement représentant la tendance générale, s_t est une fonction avec une période connue d qui représente la saisonnalité et Z_t est la composante stationnaire de bruit aléatoire.

Le but est d'estimer et d'extraire les composantes déterministes m_t et s_t dans l'espoir que le résidu, le terme Z_t, soit un processus stationnaire. Notons que dans ce cas, $E[Z_t] = 0$, $s_{t+d} = s_t$ et $\sum_{i=1}^{d} s_i = 0$. Pour des raisons de commodité dans les calculs et particulièrement dans les indices, nous indexons les données par l'année et le jour en effectuant la transformation suivante :

$$x_{j,k} = x_{k+365(j-1)} \,, \qquad j = 1, 2, \ldots, 9, \quad k = 1, 2, \ldots, 365$$

Le terme $x_{j,k}$ représente donc le taux de la année $j^{ème}$, c'est-à-dire $1990 + j$, et du $k^{ème}$ jour. Si l'évolution de la tendance est faible, il peut être raisonable de considérer qu'elle est constante durant une année. D'autre part, comme $\sum_{i=1}^{365} s_i = 0$, on a l'estimateur non-biaisé suivant pour la tendance annuelle m_j :

$$\hat{m}_j = \frac{1}{365} \sum_{k=1}^{365} x_{j,k} \,, \qquad j = 1, 2, \ldots, 9$$

alors que pour s_k on a :

$$\hat{s}_k = \frac{1}{9}\sum_{j=1}^{9}(x_{j,k} - \hat{m}_j)\,, \qquad k = 1, 2, \ldots, 365\,. \tag{4.1}$$

Notons que dans ce cas $\hat{s}_k$ satisfait forcément la condition $\sum_{k=1}^{365}\hat{s}_k = 0$. En effet,

$$\begin{aligned}\sum_{k=1}^{365}\hat{s}_k &= \sum_{k=1}^{365}\frac{1}{9}\sum_{j=1}^{9}(x_{j,k} - \hat{m}_j) \\ &= \sum_{k=1}^{365}\frac{1}{9}\sum_{j=1}^{9}(x_{j,k} - \frac{1}{365}\sum_{l=1}^{365}x_{j,l})\,,\end{aligned} \tag{4.2}$$

et en posant

$$\bar{x}_{j,\cdot} = \frac{1}{365}\sum_{l=1}^{365}x_{j,l}\,, \qquad j = 1, 2, \ldots, 9\,,$$

pour la moyenne des taux de l'année j, l'équation (4.2) devient :

$$\sum_{k=1}^{365}\hat{s}_k = \frac{1}{9}\sum_{j=1}^{9}\sum_{k=1}^{365}(x_{j,k} - \bar{x}_{j,\cdot}) = 0\,,$$

car pour chaque j la somme sur k est nul.

L'estimation du terme d'erreur du taux du $k^{ème}$ jour de la $j^{ème}$ année est donné par :

$$\hat{Z}_{j,k} = x_{j,k} - \hat{m}_j - \hat{s}_k\,, \qquad j = 1, 2, \ldots, 9, \quad k = 1, 2, \ldots, 365\,.$$

Lorsqu'il manque des valeurs, ce qui est le cas dans le jeu de données considéré, il suffit d'utiliser les estimateurs suivants :

$$\hat{m}_j = \frac{1}{n_j}\sum_{k\in K_j}x_{j,k}\,, \qquad j = 1, 2, \ldots, 9$$

où $K_j = \{k \in \mathbb{N} \mid x_{j,k} \text{ ne manque pas}\}$ et $n_j = \text{card}(K_j)$,

$$\hat{s}_k = \frac{1}{m_k}\sum_{j\in J_k}(x_{j,k} - \hat{m}_j)\,, \qquad k = 1, 2, \ldots, 365\,. \tag{4.3}$$

où $J_k = \{j \in \mathbb{N} \mid x_{j,k} \text{ ne manque pas}\}$ et $m_k = \text{card}(J_k)$, et enfin

$$\hat{Z}_{j,k} = x_{j,k} - \hat{m}_j - \hat{s}_k\,, \qquad \forall (j,k) \in I_{j,k}\,,$$

où $I_{j,k} = \{(j,k) \in \mathbb{N}^2 \mid x_{j,k} \text{ ne manque pas}\}$.
Les figures 4.1, 4.2 et 4.3 montrent la série sans la tendance annuelle $x_{j,k} - \hat{m}_j$, les estimations des composantes saisonnières $\hat{s}_k$, et enfin la série des observations sans tendance ni saisonnalité $\hat{Z}_{j,k} = x_{j,k} - \hat{m}_j - \hat{s}_k$. Les résidus ainsi obtenus (figure 4.3) ne sont mani-

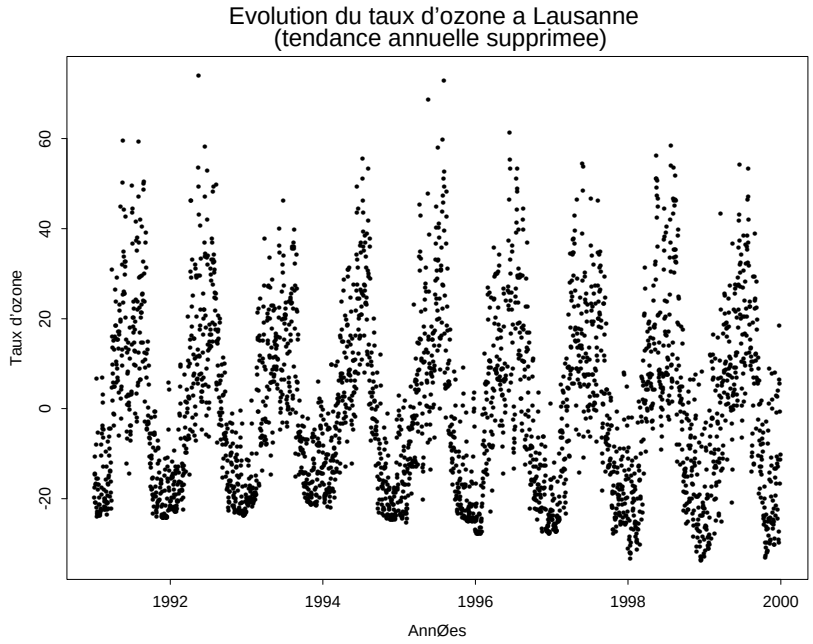

FIG. 4.1 – *Représentation des taux d'ozone ($[\mu g/m^3]$) de la série initiale dont les tendances annuelles ont été supprimées.*

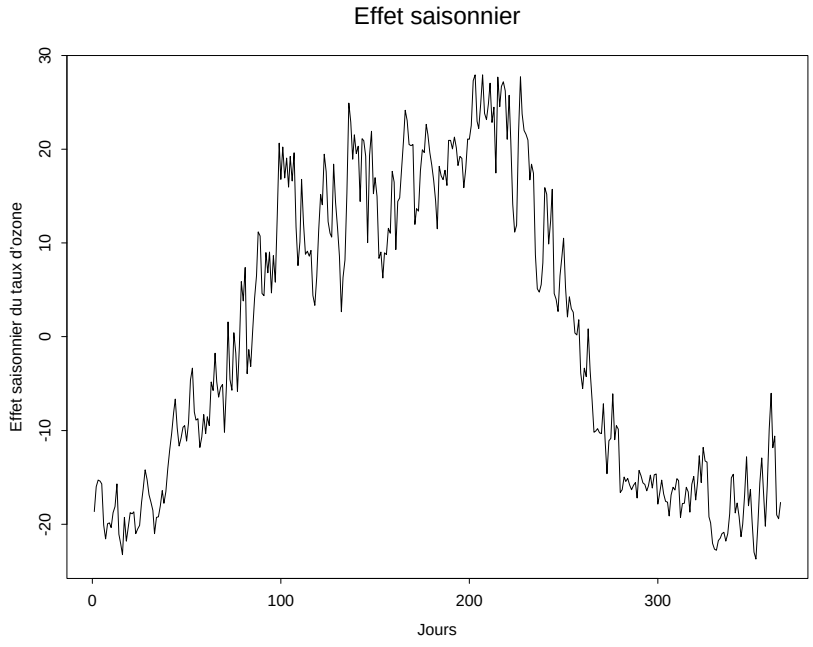

FIG. 4.2 – *Représentation des composantes saisonnières de la série initiale des taux d'ozone ($[\mu g/m^3]$) pour la période d'une année.*

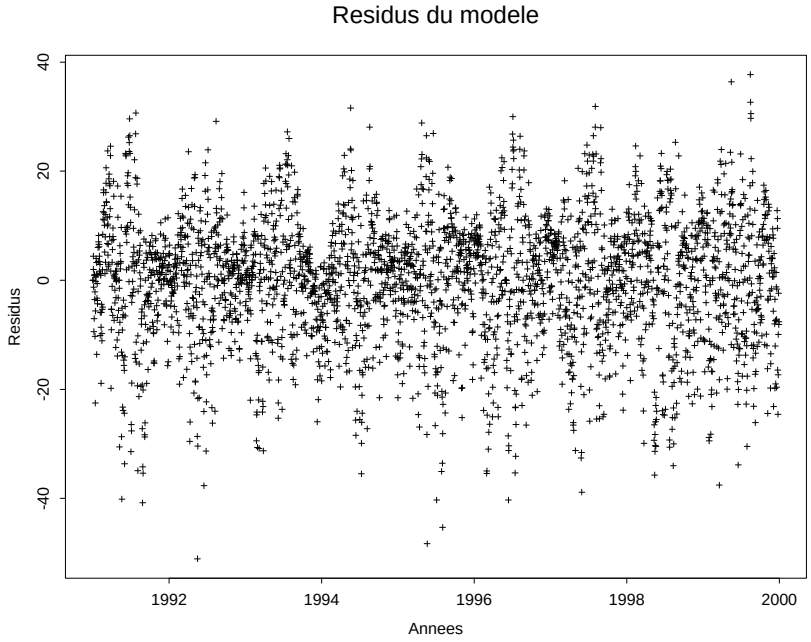

FIG. 4.3 – *Représentation des résidus du modèle de décomposition classique où la tendance et la saisonnalité ont été supprimées de la série initiale.*

festement pas stationnaires. Nous pensons que les reserrements des résidus autour de zéro proviennent des accumulations des taux lorsqu'ils sont faibles (hiver). En effectuant une transformation logarithmique sur les taux de la série initiale, il est possible d'éclater ces accumulations. Une telle transformation aura également comme effet de diminuer l'ampleur des résidus extrêmes. Les résultats obtenus par la transformation

$$Y_t = \log(X_t + c)\,, \qquad c \geq 0\,,$$

effectuée sur les données initiales avec $c = 14$ sont présentés dans les figures 4.4, 4.5 et 4.6. Les résidus ainsi obtenus sont certes plus acceptable, mais les resserrements observés précédemment sont encore présents.

Il peut alors être envisagé d'utiliser une autre transformation, comme par exemple la racine carrée, ou éventuellement de farbriquer une fonction de transformation continue en utilisant des morceaux de fonctions continues selon les besoins spécifiques et en fonction du jeu de donnés.

D'autre part, on peut se demander si la tendance m_t est significativement non-nulle. Pour ce faire, nous effectuons une ANOVA à deux voies avec le taux mesuré X_t comme variable réponse et la tendance annuelle m_t et la tendance saisonnière s_t comme facteurs. Le résultat de cette ANOVA est présenté ci-dessous, et il apparaît que les deux facteurs sont significatifs car les deux p-valeurs sont nulles, donc inférieures au taux de rejet de 5%.

```
Response: reponse.ozone
```

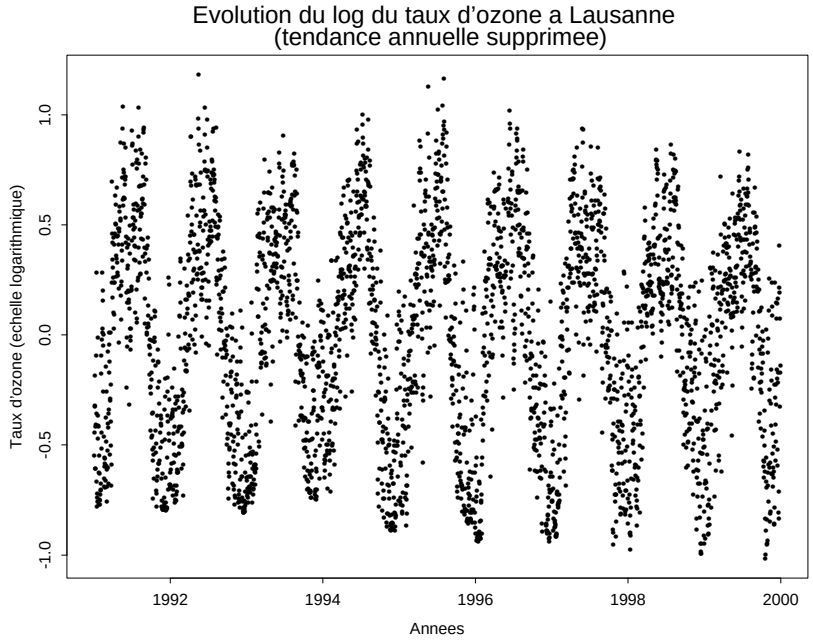

FIG. 4.4 – *Représentation de la série initiale après transformation logarithmique (*$\log(X_t+14)$*) des taux et suppression de la tendance annuelle.*

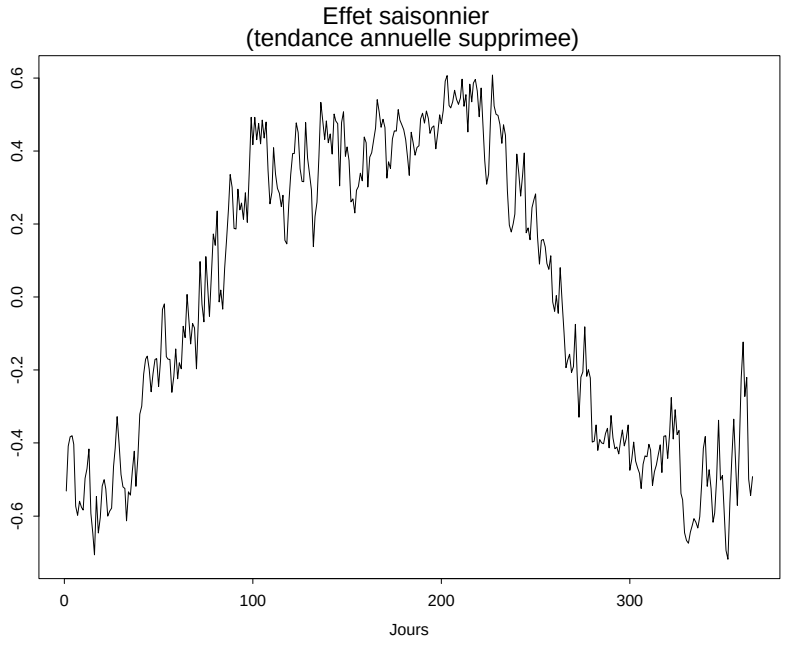

FIG. 4.5 – *Représentation des composantes saisonnières de la série initiale des taux d'ozone (*$[\mu g/m^3]$*) pour la période d'une année après transformation logarithmique des taux (*$\log(X_t+14)$*).*

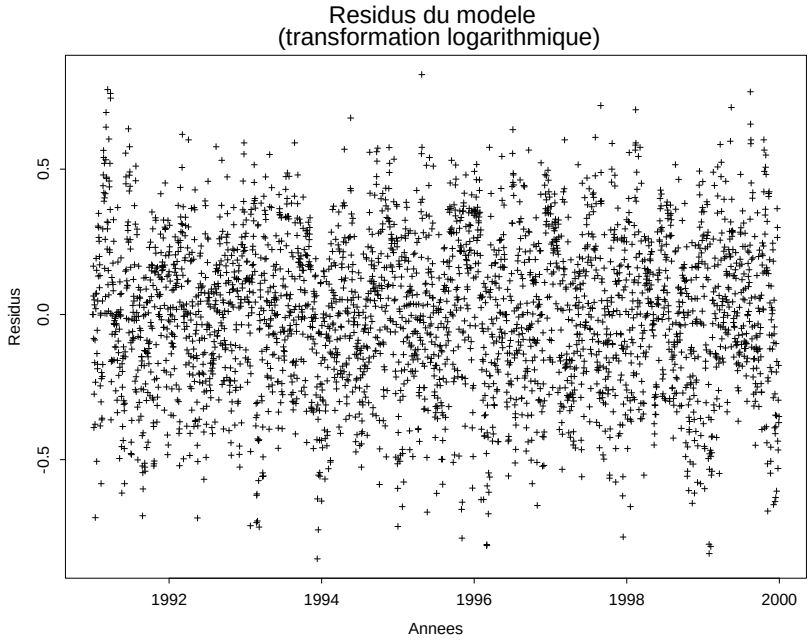

FIG. 4.6 – *Représentation des résidus du modèle de décomposition classique où la tendance et la saisonnalité ont été supprimées de la série initiale après transformation* $\log(X_t + 14)$.

```
Terms added sequentially (first to last)
                Df Sum of Sq Mean Sq  F Value Pr(F)
lignes.tend      1   2688088 2688088 20411.09     0
colonnes.sais    1    824078  824078  6257.36     0
Residuals     3283    432363     132
```

D'autre part, il semble que malgré la transformation logarithmique, les résidus ne sont pas totalement stationnaires. Cela pourrait provenir du fait que la série contient peut-être d'autres périodicités. Pour améliorer le modèle utilisé précédemment, on peut chercher une éventuelle périodicité inconnue en utilisant l'analyse harmonique ou plus précisément un périodogramme. Nous allons présenter cette approche dans la section suivante.

Etant donné que les méthodes que nous allons utiliser dans la suite ne tolèrent pas les valeurs manquantes, nous allons compléter la série initiale par les valeurs obtenues par le décomposition classique sans ajout de bruit. La figure 4.7 permet de comparer la série initiale avec la série construite de toutes les valeurs construite. Il apparaît que la série construite sous-estime les pics estivaux. Nous expliquons cela par la fait que l'estimation de la saisonnalité est obtenue par des moyennes. Il est donc normal que les valeurs les plus hautes ne puissent pas être prédites de cette manière.

4.2 Effets périodiques

Dans ce qui suit, nous allons utiliser la série complétée

En considérant une suite d'observations $x_1, x_2, \ldots, x_n$ effectuées au temps $1, 2, \ldots, n$, le

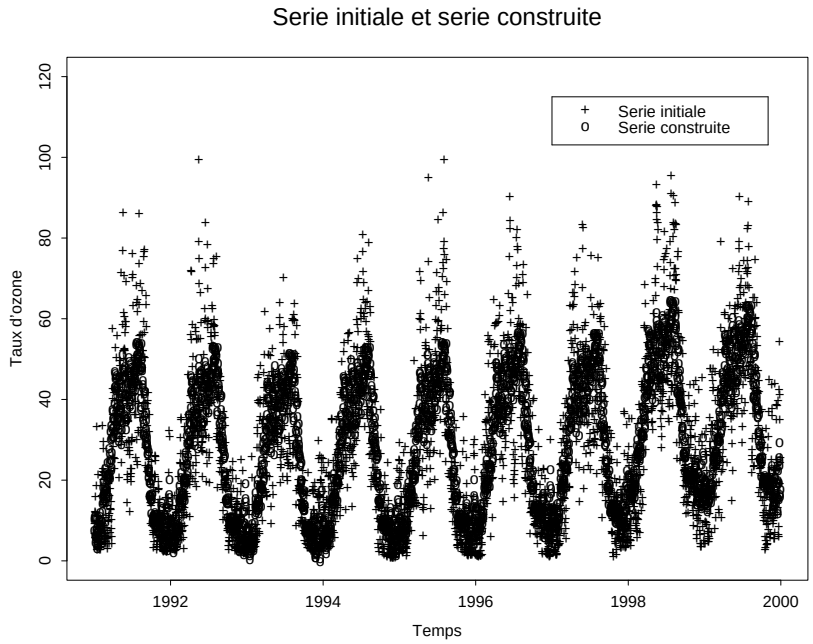

FIG. 4.7 – *Superposition de la série initiale et de la série construite par la méthode de décomposition classique.*

vecteur $\mathbf{x} := (x_1, x_2, \dots x_n)^T$ appartient à l'espace vectoriel $\mathbb{R}^n$ (ou $\mathbb{C}^n$ si les composantes sont complexes).

En supposant que les valeurs $x_1, x_2, \dots, x_n$ sont les valeurs d'une fonction de période n pour les valeurs $1, 2, \dots, n$, chaque x_t peut être exprimé comme une combinaison linéaire de fonctions harmoniques

$$x_t = \frac{1}{n^{1/2}} \sum_{-\pi < \omega_j \leq \pi} a_j e^{it\omega_j}\,, \qquad t = 1, 2, \dots, n\,, \tag{4.4}$$

où les $\omega_j = 2\pi j/n$ sont des multiples entiers de la fréqence fondamentale $2\pi/n$. Notons que les fonctions harmoniques en dehors de l'intervalle $(\pi, \pi]$ ne peuvent pas être distinguées sur la base d'observations effectuées en des temps entiers. L'ensemble des indices pour les ω_j est alors donné par $F_n = \{j \in \mathbb{Z} : -\pi < \omega_j \leq \pi\}$. On peut écrire (4.4) vectoriellement de la manière suivante :

$$\mathbf{x} = \sum_{j \in F_n} a_j \mathbf{e}_j\,,$$

où

$$\mathbf{e}_j = \frac{1}{n^{1/2}}(e^{i\omega_j}, e^{2i\omega_j}, \dots, e^{ni\omega_j})^T\,, \qquad j \in F_n\,.$$

La valeur $I(\omega_j)$ du périodogramme de $\mathbf{x}$ pour la fréquence $\omega_j = 2\pi j/n\,, j \in F_n$ est définie par les termes de la transformée de Fourier $\{a_j\}$ de $\mathbf{x}$:

$$I(\omega_j) := |a_j|^2 = \frac{1}{n}\,|\sum_{t=1}^{n} x_t e^{-it\omega_j}|^2\,, \qquad j \in F_n\,.$$

Les pics dans le périodogramme sont révélateurs de l'existence d'un effet périodique pour la fréquence associée.

Etant donné que la construction du périodogramme n'est pas possible s'il manque des valeurs dans la série, nous utilisons la série initiale complétée. La figure 4.8 montre le périodogramme de la série initiale complétée (section 4.1). Le tableau 4.1 montre les fré-

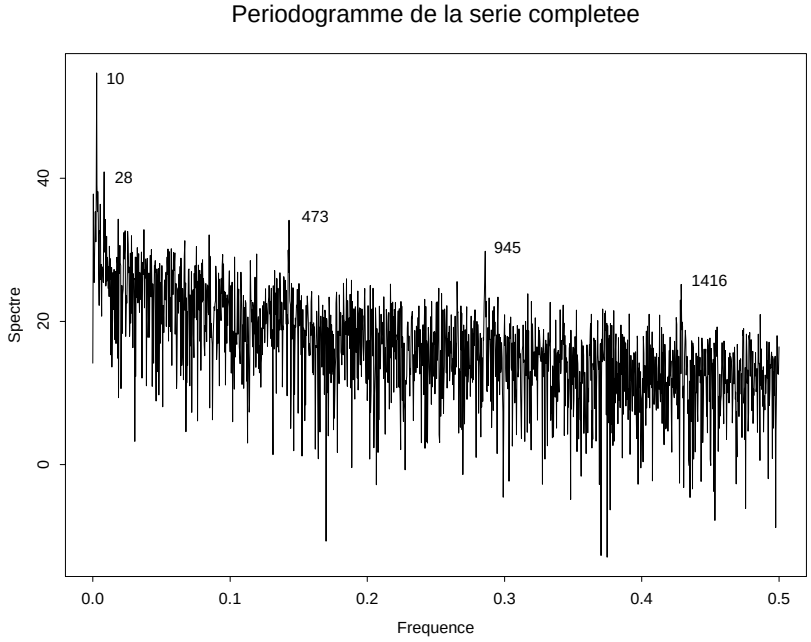

FIG. 4.8 – *Représentation du périodogramme de la série initiale complétée.*

quences et les périodes correspondantes pour les pics observés dans le périodogramme.

TAB. 4.1 – *Fréquences remarquables et périodes correspondantes pour la séries complétée.*

Valeurs	**Fréquences**	**Périodes [Jours]**
10	0.002727273	366.666672
28	0.008181818	122.222227
473	0.143030301	6.991526
945	0.286060601	3.495763
1416	0.428787887	2.332155

La périodicité annuelle est confirmée par le périodogramme. De plus, pour la construction d'un modèle complet, nous retiendrions la période de 122 jours, c'est-à-dire 4 mois, ainsi que celle de 7 jours.
La périodicité de 4 mois qui ressort du périodogramme nous étonne quelque peu, mais il est peut-être possible de l'expliquer en constatant que pour chaque période de quatre mois (janvier à avril, mai à août et septembre à décembre), les périodes les plus chaudes se trouvent souvent au milieu de la période. Le rayonnement solaire ayant une influence sur la création d'ozone, il nous semblerait alors possible que le taux d'ozone fluctue de manière périodique sur une période de 4 mois. L'explication pour la périodicité hebdomadaire nous semble quant à elle plus évidente.

En effet, de par le fait que la station lausannoise se situe au bord d'une route à grand trafic, et que le gaz carbonique contenus dans les échappements des voitures intervient dans le processus de création d'ozone (section 1.2), les concentrations d'ozone doivent alors suivre l'intensité du trafic routier au cours des semaines, avec des baisses lors des fins de semaines. Mais le but de la construction de ce modèle étant de remplacer les valeurs manquantes, nous nous contenterons du modèle proposé dans la section précédante.

L'ajustement d'un modèle qui contient également une saisonnalité sur une période de 7 jours rend la démarche plus complexe et pose notamment un probleme pour le traitement des années bissextiles car il y aurait un saut de jours entre le 28 février et le 1^{er} mars. Le but étant ici de reconstruire les valeurs manquantes, nous nous contenterons dans ce qui suit de la série complétée par le modèle de décomposition classique sur les données transformées par le logarithme.

Chapitre 5

Prédictions

Dans le but d'effectuer des predictions sur les taux d'ozone, nous allons utiliser les taux d'ozone du jeu de données des huit premières années (1991 à 1998) pour prédire les taux de l'année 1999. Cela nous permettra de comparer les valeurs prédites avec les taux mesurés. Pour effectuer ces predictions nous allons utiliser la série mensuelle obtenue en effectuant la moyenne des taux pour chaque mois (figure 5.1).

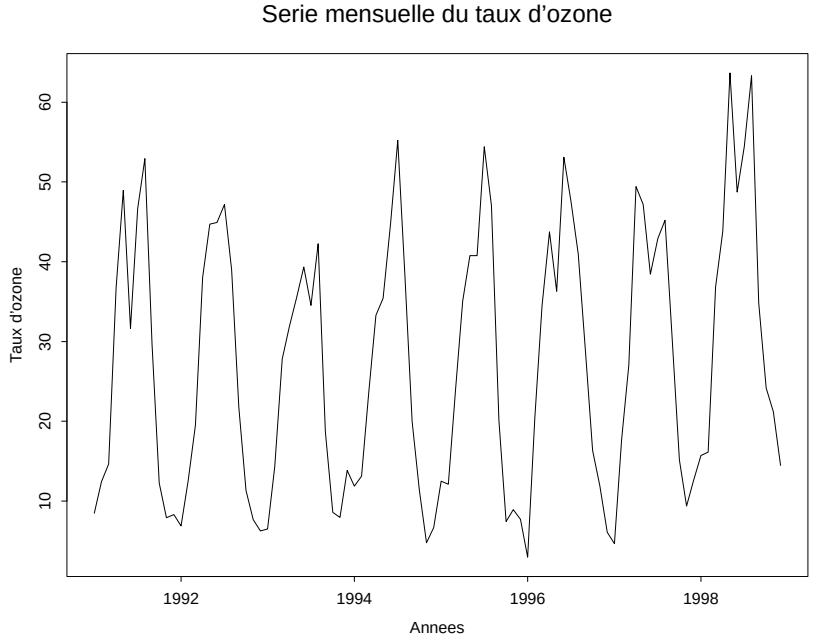

FIG. 5.1 – *Représentation de la série mensuelle du taux d'ozone mesuré par la station lausannoise du réseau NABEL.*

La série des taux mensuels $\{Y_t, t = 1, \ldots, 108\}$ est obtenue de la manière suivante. En posant

$$x_{i,j,k} = x_{k+n_1+\cdots+n_{j-1}+365(i-1)}\,, \qquad i = 1, \ldots, 8\,,\ j = 1, \ldots, 12\,,\ i = 1, \ldots, n_j\,,$$

pour le taux du k^{ème} jour du j^{ème} mois de la i^{ème} année $(1990 + i)$ où n_l indique le nombre

de jours du l^{ème} mois et

$$x_{i,j,\cdot} = \frac{1}{n_j} \sum_{k=1}^{n_j} x_{i,j,k} \,, \qquad i = 1, \ldots, 8 \,,\ j = 1, \ldots, 12 \,,\ k = 1, \ldots, n_j \,,$$

représente la moyenne des taux du j^{ème} mois de la i^{ème} année. La série mensuelle est alors donnée par

$$y_t := y_{j+12(i-1)} = x_{i,j,\cdot} \,. \tag{5.1}$$

Pour effectuer des prédictions sur les taux mensuels de l'année 1999, nous allons utiliser plusieurs approches ; elles sont décrites dans les sections suivantes. La première approche utilisée se base sur la décomosition classique utilisée au chapitre précédant.

5.1 Décomposition classique

Rappelons que la décomposition classique pour les huit premières années est donnée par

$$X_t = m_t + s_t + Z_t \,, \qquad t = 1, \ldots, 96 \,,$$

où m_t, la tendance annuelle, est considérée comme étant constante pour chaque année, s_t représente la saisonnalité et Z_t représente un bruit stationnaire. L'idée intuitive pour prédire les taux mensuels moyens de 1999 consiste à ajouter la tendance saisonnière à la tendance annuelle de l'année considérée. La tendance saisonnière utilisée est calculée à partir des huit premières années et la tendance annuelle est prédite par une regression se basant sur les tendances annuelles de 1991 à 1998. Il est clair que l'hypothèse sous-jacente est que la saisonnalité est égale à celle des années précédantes.

Le calcul de la saisonnalité se fait en utilisant l'équation (4.1) mais en sommant uniquement les termes pour $j = 1$ à 8. La figure 5.2 montre le résultat obtenu.
Une regression linéaire a été effectuée sur la série filtrée par la moyenne mobile afin d'obtenir la la valeur de la tendance annuelle de 1999. Notre choix s'est porté sur lasérie filtrée plutôt que sur la suite des moyennes annuelles car cette dernière ne contient que huit points alors que la série filtrée en contient 2921. Le résultat de la régression devrait ainsi être plus fiable. Le modèle utilisé comporte uniquement un terme constant est la première puissance de la variable explicative.

L'intervalle de confiance pour l'estimation de m_t est donné par

$$I_{1-\alpha}(m_t) = \left[m_t \ \pm \ \hat{\sigma} \cdot \mathrm{q}(1 - \frac{\alpha}{2}) \right]$$

où $\hat{\sigma}$ est l'écart-type des données et $\mathrm{q}(\cdot)$ représente le quantile de la loi normale. Le résultat ainsi obtenu est $I_{95\%}(m_t) = [31.13034 \pm 3.658185]$ alors que la valeur effective est 35.32332. La vraie valeur se trouve donc juste en dehors de l"intervalle de confiance à 95%. Les figure 5.3 et 5.4 montrent le résultat de la régression et de la prédiction pour 1999. Au vu de la disposition des points, on peut se demander s'il n'aurait pas été plus approprié d'effectuer une regression linéaire avec le terme quadratique. Mais en établissant une échelle orthonormée, il est aisé de constater qu'en fait la tendance est linéaire. Cela est confirmé par une regression linéaire avec le terme quadratique qui donne zéro pour le coefficient quadratique.

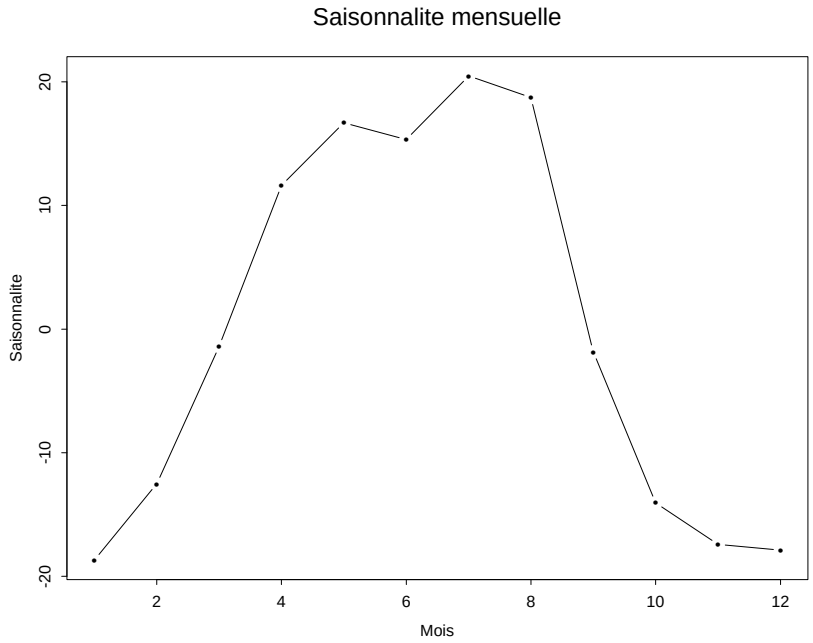

FIG. 5.2 – *Représentation de la saisonnalité calculée sur la base des taux d'ozone mensuels de 1991 à 1998.*

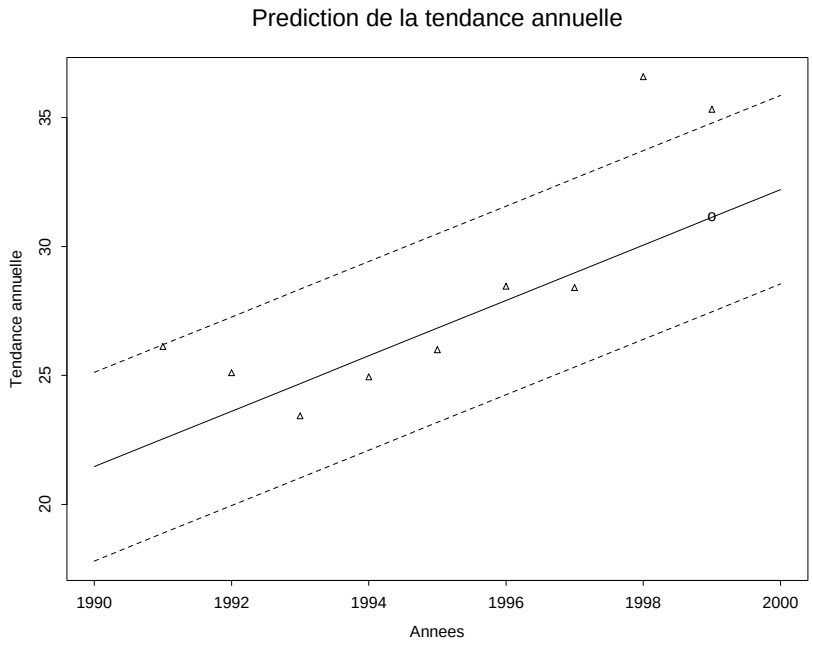

FIG. 5.3 – *Représentation de prediction de la tendance annuelle pour 1999 par une regression linéaire sur les tendances annuelles de 1991 à 1998.*

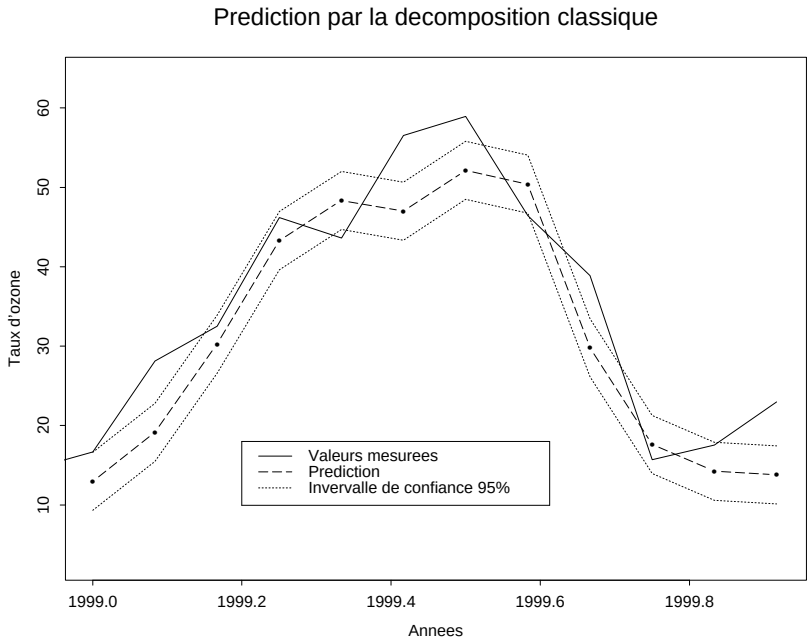

FIG. 5.4 – *Représentation de la prediction effectuée à l'aide du modèle de décomposition classique pour les taux mensuels d'ozone à Lausanne en 1999.*

5.2 Décomposition saisonnière

L'idée de la décomposition saisonnière consiste à extraire la périodicité de la série et d'ajuster un modèle ARIMA sur les résidus. Nous introduisons dans la section suivante quelques éléments théoriques concernant une classe très importante de séries temporelles $\{X_t\,,\, t = 0 \pm 1, \pm 2, \dots\}$ définies en termes d'équations linéaire des différences à coefficients constant. Il s'agit des processus ARMA (autoregressive moving average)

5.2.1 Processus ARMA

Un processus $\{Z_t\}$ est un *bruit blanc* (white noise) de moyenne 0 et de variance σ^2 noté $\{Z_t\} \sim \mathrm{WN}(0, \sigma^2)$ si et seulement si $\{Z_t\}$ a une moyenne égale à zéro et comme fonction de covariance

$$\gamma(h) = \begin{cases} \sigma^2 & \text{si } h = 0\,, \\ 0 & \text{si } h \neq 0\,. \end{cases}$$

Le processus $\{X_t\,,\, t = 0 \pm 1, \pm 2, \dots\}$ est un processus ARMA(p, q) si $\{X_t\}$ est stationnaire et si pour tout t

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}\,, \tag{5.2}$$

où $\{Z_t\} \sim \mathrm{WN}(0, \sigma^2)$. On dit que $\{X_t\}$ est un processus ARMA(p,q) de moyenne μ si $\{X_t - \mu\}$ est un processus ARMA(p,q).
L'équation (5.2) peut être écrite de manière symbolique plus compacte de la manière suivante

$$\phi(B) X_t = \theta(B) Z_t\,, \qquad t = \pm 1, \pm 2, \dots$$

où ϕ et θ sont les polynômes de degrés p et q

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \ ,$$

et

$$\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q \ ,$$

et B est l'opérateur de différence retrograde (backward shift operator) défini par

$$B^j X_t = X_{t-j} \ , \qquad j = 0, \pm 1, \pm 2, \ldots \ .$$

Une des raisons pour lesquelle la famille de processus ARMA joue un rôle important dans la modélisation des données de séries temporelles est la suivante: pour toute fonction d'autocovariance $\gamma(\cdot)$ telle que $\lim_{h\to\infty} \gamma(h) = 0$ et pour tout entier $k > 0$, il est possible de trouver un processus ARMA avec fonction d'autocovariance $\gamma_X(\cdot)$ tel que $\gamma_X(\cdot) = \gamma(h) \ , \quad h = 0, 1, \ldots, k \ .$

Notons encore qu'un processus ARMA(p,q) défini par les équations $\phi(B)X_t = \theta(B)Z_t$ est dit *causal* s'il existe une suite de constante $\{\Psi_j\}$ telle que $\sum_{j=0}^{\infty} |\Psi_j| < \infty$ et

$$X_t = \sum_{j=0}^{\infty} \Psi_j Z_{t-j} \ , \qquad t = 0, \pm 1, \pm 2, \ldots \ .$$

5.2.2 Processus ARIMA

Une généralisation de la classe des processus ARMA qui incorpore une grande quantité de séries non stationnaires est donnée par les processus ARIMA, c'est-à-dire les processus qui deviennent des processus ARMA après un nombre fini de différentiations.

Si d est un entier non négatif, alors $\{X_t\}$ est un processus ARIMA(p,d,q) si $Y_t := (1-B)^d X_t$ est un processus ARMA(p,q) causal.

Cette définition signifie que $\{X_t\}$ satisfait une équation de différences de la forme

$$\phi^*(B)X_t \equiv \phi(B)(1-B)^d X_t = \theta(B)Z_t \ , \qquad \{Z_t\} \sim \mathrm{WN}(0, \sigma^2)$$

où $\phi(z)$ et $\theta(z)$ sont des polynômes de degrés p et q respectivement et $\phi(z) \neq 0$ for $|z| \leq 1$. Le polynôme $\phi^*(z)$ a un zéro d'ordre d et $z = 1$. Le processus $\{X_t\}$ est stationnaire si et seulement si $d = 0$, et dans ce cas il s'agit d'un processus ARMA(p,q).

Les modèles ARIMA sont utiles pour représenter les données avec une tendance.

5.2.3 Prédictions

Comme cela a déjà été mentionné au début de cette section, l'idée de la décomposition saisonnière consiste à extraire la composante saisonnière de la série mensuelle puis d'ajuster un modèle ARIMA sur les résidus. La figure 5.5 montre le résultat de cette décomposition et la figure 5.6 montre le diagramme en boîte (boxplot).
Les représentations des fonctions d'autocorrélation et d'autocorrélation partielle du résidu est présenté dans la figure 5.7. La décroissance lente des coéfficients d'autocorrélation

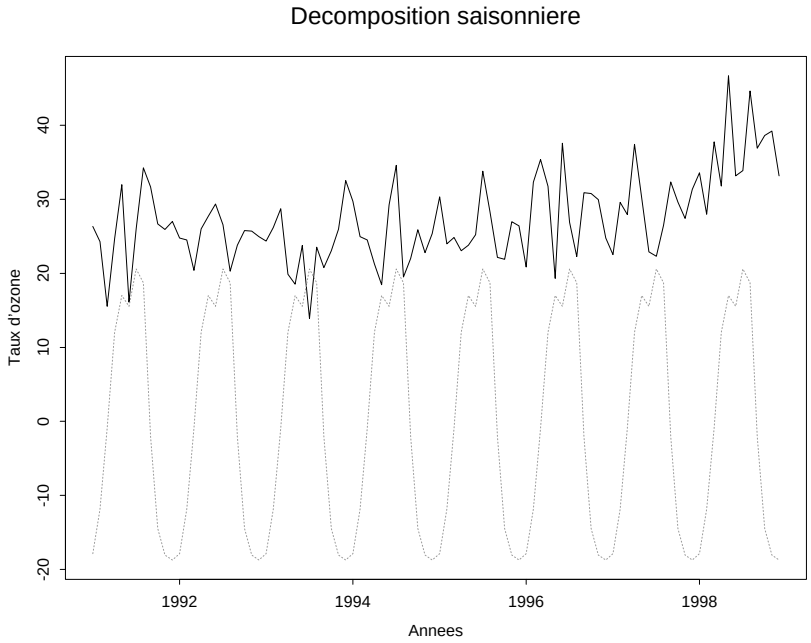

FIG. 5.5 – *Représentation de la décomposition saisonnière effectuée sur la série mensuelle des taux d'ozone ($[\mu g/m^3]$) mesurés par la station lausannoise du réseau NABEL.*

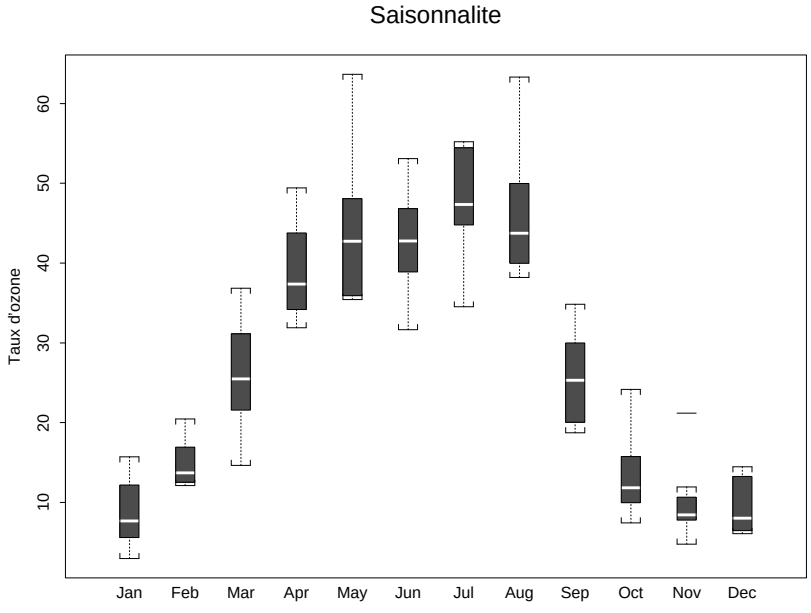

FIG. 5.6 – *Diagramme en boîte (boxplot) de la saisonnalité pour les taux d'ozone mensuels ($[\mu g/m^3]$) mesurés à Lausanne*

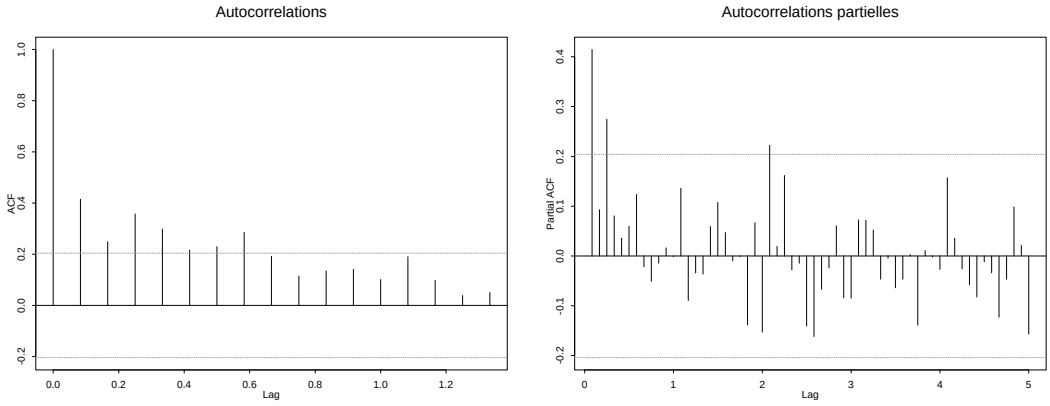

FIG. 5.7 – *Représentation des fonctions d'autocorrélation et d'autocorrélation partielle pour le résidus de la décomposition saisonnière de la série mensuelle des taux d'ozone mesurés par la stations lausannoise du réseau NABEL.*

indique la présence d'une tendance et la figure 5.5 suggère en effet une telle tendance. La figure 5.8 montre les fonctions d'autocorrélation et d'autocorrélation partielle obtenus en appliquant les différences d'ordre 1 ($Y_t = (1-B)X_t$), pour que les résidus soit stationnaires, sur les termes de la série, et suggère d'ajuster un modèle ARIMA(2,1,2). Le modèle ajusté

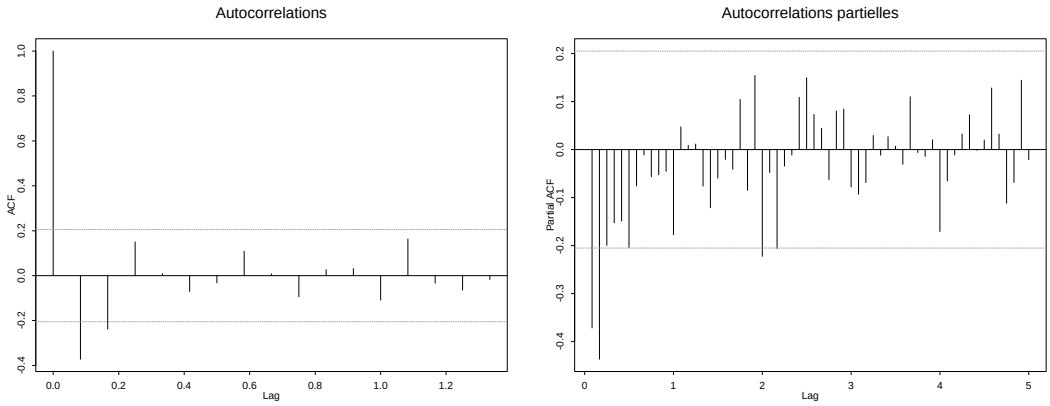

FIG. 5.8 – *Représentation des fonctions d'autocorrélation et d'autocorrélation partielle de la série des différences de premier ordre pour les résidus de la décomposition saisonnière de la série mensuelle des taux d'ozone mesurés par la stations lausannoise du réseau NABEL.*

est donné par

$$(1 - 0.002869026B - 0.230505853B^2)(1 - B)X_t = (1 + 1.7605145B - 0.7679783B^2)Z_t$$

et la prédiction pour les taux mensuels est donnée dans la figure 5.9.

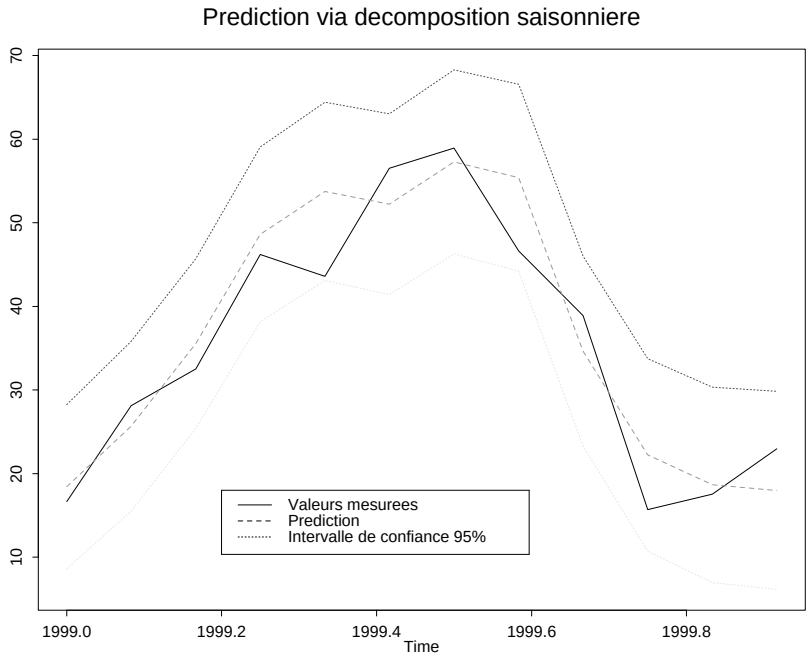

FIG. 5.9 – *Représentation de la prédiction pour les taux d'ozone mensuel de 1999 à Lausanne par la méthode de la décomposition saisonnière.*

5.3 Modèle SARIMA

La série mensuelle donnée par l'équation (5.1) est caractérisée par une forte corrélation entre les termes qui représente des taux du même mois. Cela est confirmé par la fonction de corrélation représentée dans la figure 5.10. Dans la section 5.1 nous avons utilisé la décomposition classique d'une série temporelle $X_t = m_t + s_t + Z_t$ où m_t est la composante de la tendance, s_t la composante saisonnière et Z_t la composante de bruit stationnaire. Mais dans bien des cas, il n'est pas raisonnable de supposer que la composante saisonnière est exactement la même de cycle en cycle. Les models SARIMA permettent de modéliser l'aspect aléatoire des composantes saisonnières d'année en année. Nous introduisons maintenant quelques éléments théoriques concernant des processus SARIMA.

5.3.1 Processus SARIMA

Si d et D sont des entiers non-négatifs, alors la série $\{X_t\}$ est un processus SARIMA(p, d, q) x $(P, D, Q)_s$ avec période s si le processus des différences $Y_t := (1-B)^d(1-B^s)^D X_t$ est un processus ARMA causal

$$\phi(B)\Phi(B^s)Y_t = \theta(B)\Theta(B^s)Z_t\,, \qquad \{Z_t\} \sim \mathrm{WN}(0, \sigma^2)\,,$$

où $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$, $\Phi(z) = 1 - \Phi_1 z - \cdots - \Phi_P z^P$, $\phi(z) = 1 + \phi_1 z + \cdots + \phi_q z^q$ et $\phi(z) = 1 + \phi_1 z + \cdots + \phi_Q z^Q$.
Notons que le processus $\{Y_t\}$ est causal si et seulement si $\phi(z) \neq 0$ et $\Phi(z) \neq 0$ pour $|z| \leq 1$.

Dans la pratique, D est plus grand que un et P et Q sont typiquement inférieurs à trois.

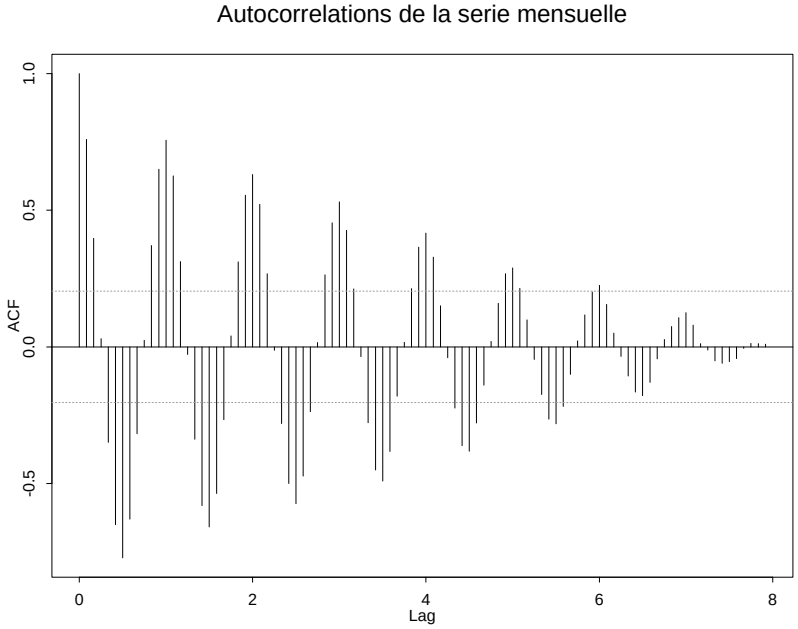

FIG. 5.10 – *Représentation des autocorrélations pour la série mensuelle.*

Pour identifier un modèle SARIMA, il faut dans un premier temps, trouver d et D et construire le processus des différences

$$Y_t = (1-B)^d(1-B^s)^D X_t$$

de telle sorte qu'il soit stationnaire. Il faut ensuite examiner les représentation des autocorrélations et des autocorrélations partielles de $\{Y_t\}$ dont les ordres sont des multiples de s de manière à identifier les ordres P et Q. Si $\hat{\rho}(\cdot)$ est l'estimation de la fonction d'autocorrélation de $\{Y_t\}$, alors P et Q doivent être choisis de telle sorte que $\hat{\rho}(ks),\ k = 1, 2, \ldots$ soient compatibles avec la fonction d'autocorrélation d'un processus ARMA(P,Q). Les ordre p et q sont alors sélectionnés pour que les estimations des autocorrélations $\hat{\rho}(1), \ldots, \hat{\rho}(s-1)$ correspondent aux fonctions d'autocorrélations d'un processus ARMA(p,q).

5.3.2 Prédictions

La fonction d'autocorrélation (figure 5.10) montre une périodicité de 12 pour la série mensuelle des taux d'ozone. Dans la notation du modèle SARIMA on a alors $s = 12$ et $D = 1$. Il faut alors construire la série des différences d'ordre 12 et observer les fonctions d'autocorrélations et d'autocorrélations partielles (figures 5.11) pour déterminer les autres paramètres du modèle SARIMA. Noter que la fonction d'autocorrélation décroît rapidement ce qui indique que la série des différences d'ordre 12 est stationnaire et on pose alors $d = 0$. La représentation de la fonction d'autocorrélation suggère les valeurs de $q = 0$ et $Q = 1$, alors que la représentation de la fonction d'autocorrélation partielle suggère les valeurs de $p = 0$ et $P = 1$. Le modèle à ajuster est donc SARIMA$(0,0,0)$ x $(1,1,1)_{12}$. Le modèle ajusté est donné par

$$(1 - 0.1613181B^{12})Y_t = (1 + 0.2903926B^{12})Z_t\ , \qquad \{Z_t\} \sim \text{WN}(0, 46.6937\ )$$

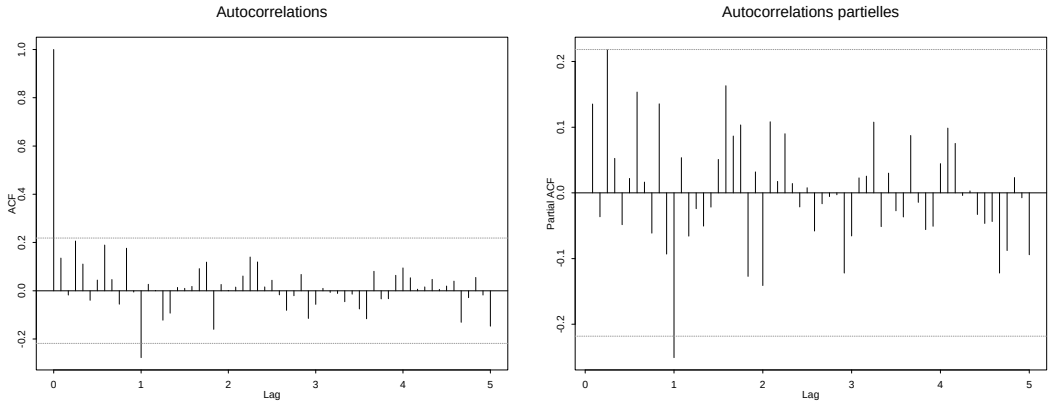

FIG. 5.11 – *Représentation des fonctions d'autocorrélation et d'autocorrélation partielle de la série des différences d'ordre 12 pour la série mensuelle des taux d'ozone mesurés par la station lausannoise du réseau NABEL.*

où

$$Y_t = (1 - B^{12})X_t \ ,$$

et le résultat de la prédiction est présenté dans la figure 5.12.

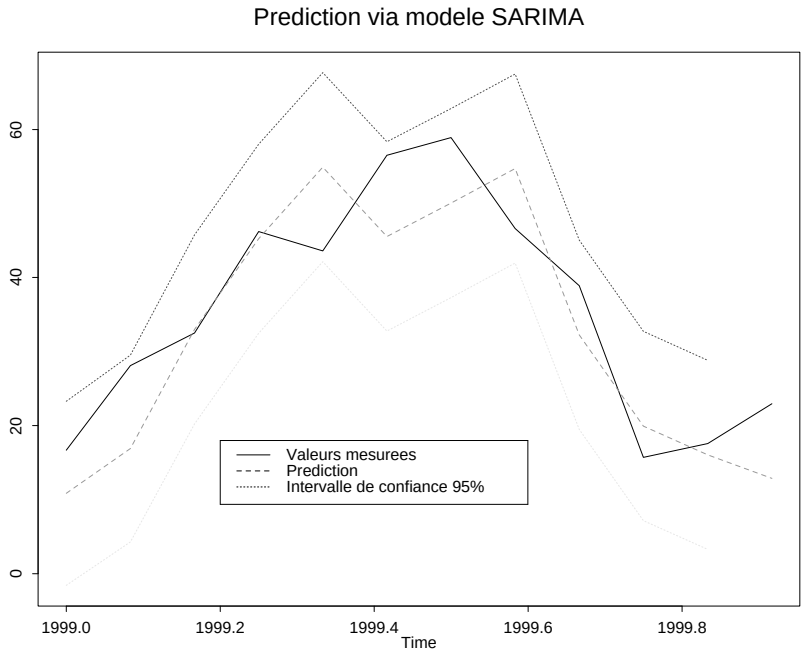

FIG. 5.12 – *Représentation de la prédiction des taux d'ozone mensuel de 1999 à Lausanne obtenue par le modèle* SARIMA $(0,0,0)$ x $(1,1,1)_{12}$.

Chapitre 6

Remarques

Les résultats présentés dans ce travail se distinguent en deux parties ; il y a la construction des valeurs manquantes et les prédictions des taux d'ozone mensuels.

Valeurs manquantes

Le modèle qui a été utilisé pour compléter la série initiale consiste en une approche simple et intuitive. Pour l'améliorer, il aurait été possible, comme nous l'avons déjà mentionné, de construire un modèle qui tient compte de la périodicité hebdomadaire, mais cela pose des problème notamment concernant le traitement des années bissextiles.
D'autre part, il aurait été possible d'ajouter aux deux termes déterministes m_t et s_t un terme de bruit stationnaire qui aurait les mêmes caractéristiques que les résidus du modèle. Il aurait également pu être envisagé de procéder par itérations, c'est-à-dire de construire les valeurs manquantes en utilisant les termes déterministes ainsi que le bruit stationnaire, de recalculer ensuite les termes de tendance et de saisonnalité, et sur la base de ceux-ci, reconstruire la valeur manquante, et ainsi de suite. On peut alors se demander si la méthode converge et si c'est le cas, vers quoi elle converge.

Prédictions

Parmi les trois modèles qui ont été utilisés pour effectuer des prédictions, c'est-à-dire par le modèle de la décomposition simple, par la décomposition saisonnière et par l'ajustement d'un modèle SARIMA, nous estimons que le premier de ceux-ci est trop simpliste. En effet, cette approche suppose que la saisonnalité est constante au cours des années et se répète exactement.
Le résultat obtenu pour cette prédiction (figure 5.4) n'est pas tellement bon, car les valeurs effectivement mesurées pour 1999 sortent à plusieurs reprises de la bande de confiance de la prédiction et celle-ci sous-estime globalement les taux effectifs. A propos de la bande de confiance, elle a été construite à l'aide de l'intervalle de confiance sur la prédiction de la tendance m_t, mais il aurait peut-être été préférable de construire également des intervalles de confiance pour les composantes de la saisonnalité s_t. La bande de confiance pour la prédiction serait alors certainement plus adaptée à la prédiction et sa largeur étant plus grande, elle aurait mieux recouvert les taux effectivement mesurés.

D'autre part, on peut également se demander s'il est raisonnable d'utiliser toutes les valeurs disponibles pour prédire la tendance de l'année à venir. En effet, on constate que l'estimation de la tendance pour 1999 sous-estime nettement la vraie tendance. Il aurait probablement été plus judicieux d'effectuer une regression sur les taux à partir de 1993, car la tendance entre 1991 et 1993 est décroissante et a pour conséquence d'applatir la droite de régression. Sans tenir compte des valeurs des deux premières années, la prédiction pour la tendance annuelle de 1999 aurait très certainement été meilleure. On peut alors, en toute généralité, se demander sur combien d'années il faut baser sa prédiction de la tendance annuelle. Nous pensons que le choix du nombre d'années dépend en fait de cas en cas.

La prédiction effectuée par la décomposition saisonnière, qui a consisté à retirer la composante périodique de la série et à ajuster un modèle ARIMA sur le résidu, nous semble nettement meilleure (figure 5.9). En effet, la bande de confiance recouvre totalement les taux effectivement mesurés et, à l'exception des mois de mai et d'octobre avec leur pic vers le bas, la prédiction est très proche de la réalité.

Le modèle SARIMA est le modèle le plus complet qui a été utilisé dans ce travail, et le résultat de la prédiction ainsi obtenue (5.12) est satisfaisant car la bande de confiance recouvre totalement les taux effectifs. Mais en comparant les prédiction obtenues par le modèle de décomposition saisonnière et par le modèle SARIMA, nous pensons que le résultat de la première de celle-ci est meilleure car elle reste plus proche des taux effectifs. Peut-être qu'en ajustant un modèle SARIMA plus compliqué, nous aurions obtenu de meilleurs résultats, mais nous pensons tout de même que le modèle ajusté dans la section 5.3 correspond très bien aux données malgré le fait qu'il est assez simple.
Pour des raisons analogues à celles présentées ci-dessus, on peut se demander sur combien d'années il faut baser ces prédictions compte tenu de l'éventuelle évolution de la tendance annuelle ou de la saisonnalité. Le fait d'utiliser des valeurs dont la tendance ne correspond plus à ce qui se passe actuellement pourrait avoir pour conséquence de fausser les résultats. Il peut en effet arriver que des facteurs extérieurs ponctuels, comme par exemple l'introduction d'un test anti-pollution qui diminuerait probablement les émissions de gaz carbonique et diminuerait ainsi la tendance annuelle des taux d'ozone, ou une reprise économique qui favoriserait l'augmentation du parc automobile et ainsi les émissions totales de CO_2. C'est peut-être un phénomène analogue qui s'est produit autour de 1993.

Parmi les trois prédictions proposées, nous pensons que la meilleure est obtenu par la méthode de décomposition saisonnière. Cela semble étonnant car le modèle SARIMA est le plus complet, mais il est peut-être plus sensible aux changements brusques de tendances que nous avons évoqués ci-dessus. On peut alors se demander, au vu de ce qui précède, quels seraient les résultats obtenus pour des prédiction basées uniquemenent sur les taux mesurés depuis 1993.

Remarques générales

Le boxplot (figure 5.6) qui représente les taux mensuels d'ozone indique nettement que la variance des taux est bien plus grande pour les taux dont la valeur est élevée. Il serait donc opportun de trouver une transformation continue à effectuer sur les données de manière à uniformiser la variance pour toutes les observations. Nous avons effectuer des essais

avec des transformations utilisant le logarithme ou la racine carrée, mais les resultats n'ont pas été très concluants. Peut-être faut-il alors construire à la main une transformation continue en utilisant des morceaux de fonctions continues.

D'autre part, nous pensons qu'il serait intéressant de d'effectuer des prédictions sur la série journalière car cela pourrait permettre de signaler des dangers de dépassement des valeurs limites fixées par l'Ordonnance sur la protection de l'air (OPair).

Conclusion

Dans ce travail, nous avons étudié les taux d'ozone mesurés par la station lausannoise du réseau NABEL en utilisant l'approche des séries temporelles.

Après exploration des données initiales, nous avons été en mesure de compéter le jeu de données en remplacant les valeurs manquantes par des predictions issues de la décomposition classique d'une série temporelle qui contient une tendance et une saisonnalité. Le resultat n'est pas totalement satisfaisant et nous avons alors proposé quelques possibilités d'amélioration de cette méthode.

Nous avons ensuite été en mesure de prédire les taux mensuels de la dernière année en utilisant les taux mensuels des huit premières années. Pour ce faire nous avons utilisé trois méthodes différentes. La première est basée sur la décomposition classique, et les résultats ainsi obtenus ne sont vraiment satisfaisants. Nous avons proposé deux moyens d'améliorer cette prédiction qui reste selon nous trop simpliste.
La seconde méthode a consisté à ajuster un modèle ARIMA sur les résidus de la série dont la saisonnalité a été retirée au préalable. Nous sommes très satisfaits de la prédiction ainsi obtenue.
La dernière méthode utilisée pour ces prédictions a consisté à ajuster un modèle SARIMA. Les résultats obtenus sont bons, mais notre préférence va malgré tout pour la seconde méthode utilisée.

Une des questions que ce travail nous a permis de soulever concerne le choix du nombre d'observations à utiliser dans l'ajustement des modèles ARIMA et SARIMA. Une étude future pourrait porter sur la détermination optimale du choix du nombre d'observations à considérer dans l'ajustement de ces modèles.

Remerciements

Je tiens à remercier le Professeur S. Morgenthaler et les membres de son groupe pour leur précieuse aide.
Et je tiens à remercier tout particulièrement Thomas Gsponer pour sa grande disponibilité et sa précieuse collaboration durant tout le semsetre.

Bibliographie

[1] P. J. Brockwell, R. A. Davis, *Time Series : Theory and Methods*, Springer Series in Statistics, 1987

[2] W. N. Venables, B. D. Ripley, *Modern Applied Statistics with S-SLUS*, Springer, Third Edition, 1999.

[3] S. Morgenthaler, *Introduction à la statistique*, Presses polytechniques et universitaires romandes, 1997.