UNIVERSITÉ DU QUÉBEC

MÉMOIRE PRÉSENTÉ À
L'UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

COMME EXIGENCE PARTIELLE
DE LA MAÎTRISE EN MATHÉMATIQUES ET INFORMATIQUE
APPLIQUÉES

PAR

KATLYN THIBODEAU

APPLICATION DE LA MÉTHODOLOGIE BOX-JENKINS AUX SÉRIES DU MINISTÈRE DE LA SANTÉ

AVRIL 2011

REMERCIEMENTS

Je remercie toutes les personnes importantes de ma vie qui ont contribué au bon déroulement de toutes mes années d'études universitaires dont résulte ce mémoire. D'abord, je tiens à remercier mon directeur de recherche, Mhamed Mesfioui, professeur au département de mathématiques et informatique de l'Université du Québec à Trois-Rivières et Latifa Elfassihi du Ministère de la Santé et des Services Sociaux. Ils ont cru en moi, m'ont proposé un sujet de recherche des plus intéressants et m'ont accompagnée tout au long du projet.

Je témoigne aussi beaucoup de reconnaissance à l'égard du département de mathématiques et informatique de l'Université du Québec à Trois-Rivières qui m'ont toujours encouragée et beaucoup aidée tout au long de mes études et représente pour moi une seconde famille. J'aimerais remercier spécialement Ismail Biskri et François Meunier d'avoir accepté d'évaluer mon mémoire et proposé des commentaires constructifs pour son perfectionnement.

Je remercie toutes les institutions qui m'ont soutenue financièrement lors de mes études : l'Université du Québec à Trois-Rivières et l'Institut des Sciences Mathématiques du Québec.

Finalement, je désire remercier sincèrement tous les membres de ma famille et en particulier ma sœur My-Linh, ma plus grande source d'inspiration dans la vie, et mon amoureux Peter.

RÉSUMÉ

Notre projet de recherche consiste en l'élaboration de modèles de prévision à l'aide des séries chronologiques basés sur la méthodologie Box-Jenkins aussi couramment appelée le modèle ARIMA.

Cette étude est le fruit d'une collaboration entre l'Université du Québec à Trois-Rivières (UQTR) et le Ministère de la Santé et des Services Sociaux (MSSS), plus précisément l'équipe de la direction de la surveillance de l'état de santé.

Les modèles développés servent à prédire les valeurs d'un indicateur d'intérêt d'un des programmes d'aide à la population que le MSSS met en action: le SIPPE (Services Intégrés en Petite Enfance et Périnatalité) et ce, pour les 5 prochaines années financières.

Les résultats obtenus n'ont pas permis d'établir que la méthodologie de Box-Jenkins était la méthode la plus appropriée afin de traiter les données du programme SIPPE. Cela s'explique par le fait que seulement 60 séries sur un total de 208 ont pu être ainsi modélisées. Le projet a toutefois permis de faire les prévisions pour chacune des séries, et à mener à de solides recommandations pour les prochaines prévisions ainsi qu'à la suggestion d'alternatives à explorer.

Mots clés : prévision, séries chronologiques, méthodologie Box-Jenkins, modèles ARIMA

ABSTRACT

Our research project consisted of the development of predictive models using chronological series based on Box-Jenkins methodology (also known as ARIMA models.)

This study is the fruit of collaboration between the Université du Québec à Trois-Rivières (UQTR) and the Ministère de la Santé et des Services Sociaux (MSSS) or, more precisely, its management team for monitoring public health.

The models developed are used to predict, for the next five financial years, the values of an indicator of interest for one of the public-assistance programs run by the MSSS: the SIPPE program.

The results obtained did not show that Box-Jenkins methodology was the most appropriate way to process the data from the SIPPE program. Indeed, only 60 out of 208 series were capable of being thus modelled. The project did however enable us to make forecasts for each of the series, and led to concrete recommendations for the next forecasts, as well as suggestions for other alternatives to explore. Considering the nature of the variable, one particularly interesting possibility is the development of full value chronological series models.

A censored version of the final report submitted to the MSSS on April 26, 2010 is attached to the dissertation. The report was censored in order to comply with the confidentiality requirements of the SIPPE program.

Keywords: forecasting, chronological series, Box-Jenkins methodology, ARIMA models.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

INTRODUCTION

Que ce soit sur le plan économique lorsqu'il est question des indicateurs financiers qui servent à la surveillance et à l'évaluation de la solidité, la stabilité et le rendement de l'économie [1] ou que ce soit sur le plan social à l'aide des différents indicateurs de santé servant à dresser le Portrait de santé du Québec et de ses régions [2], l'étude des séries chronologiques peut s'avérer un outil puissant de prévision pour nos différentes institutions gouvernementales.

En 2001, la Loi sur la santé publique [3] fut adoptée et c'est ainsi que la notion de surveillance de l'état de santé de la population a pris une place de premier choix au cœur des activités du Ministère de la Santé et des Services Sociaux (MSSS). Plus particulièrement, la **Direction de la surveillance de l'état de santé a pour mandat d'accomplir cette mission en réalisant les activités suivantes** [4]:

1. dresser un portrait global de l'état de santé de la population
2. observer les tendances et les variations temporelles et spatiales
3. détecter les problèmes en émergence
4. identifier les problèmes prioritaires
5. élaborer des scénarios prospectifs de l'état de santé de la population
6. suivre l'évolution, au sein de la population, de certains problèmes spécifiques de santé et de leurs déterminants.

Notre étude a pour but de contribuer à la réalisation de la cinquième activité, c'est-à-dire : «L'élaboration de scénarios prospectifs de l'état de santé de la population». Il s'agit de développer un outil d'aide à la décision pour un indicateur de santé contenu dans le programme d'aide SIPPE (Services Intégrés en Périnatalité et Petite Enfance). Cet outil consiste en l'utilisation de la méthodologie Box-Jenkins pour modéliser cent-quatre-vingts-huit séries de données rattachées au programme SIPPE. Les données proviennent des fichiers fermés et provisoires de naissances vivantes et mortinaissances du Registre

des événements démographiques du Québec. La population visée par le programme SIPPE est celle résidant dans les municipalités du Québec conventionnées.

Une introduction aux séries chronologiques et aux méthodes de prévision quantitatives courantes, telles que la régression linéaire simple, le lissage exponentiel simple, le lissage exponentiel double, la méthode de Winters et les modèles Box-Jenkins, composent notre premier chapitre.

Le chapitre 2 couvre en détails la méthodologie Box-Jenkins pour les modèles non saisonniers ainsi que les raisons qui ont conduit à ce choix de méthode pour le projet. Plus précisément, il s'agit respectivement des étapes de la démarche et des spécifications des modèles de moyennes mobiles, autorégressifs et mixtes ainsi que des résultats d'une brève exploration des données.

Le chapitre 3 fait la présentation de l'application des étapes d'identification du modèle et d'estimation des paramètres de la modélisation ARIMA aux séries à l'étude à l'aide d'un exemple. Il comporte aussi les modèles retenus de 3 séries chronologiques tirées des analyses produites pour le SIPPE. Les alternatives lorsqu'aucun modèle ARIMA n'est adéquat sont aussi brièvement présentées et appuyées d'un exemple pour chaque cas.

Un bilan du projet, des recommandations pour les prochaines prévisions et les perspectives à venir constituent la conclusion de ce mémoire.

CHAPITRE 1

MÉTHODES PRÉVISIONNELLES PORTANT SUR LES SÉRIES CHRONOLOGIQUES

Ce chapitre se divise en trois parties, la première expose les séries chronologiques de façon générale et décrit de façon succincte les séries chronologiques du programme SIPPE (Services Intégrés en Périnatalité et Petite Enfance). La deuxième partie présente quelques exemples de modèles simples de séries chronologiques. La troisième partie traite des méthodes de prévision quantitatives les plus utilisées. Ce chapitre se veut une introduction au projet de recherche qui fait l'objet de ce mémoire. Il présente les notions de base sur les séries chronologiques afin de mieux comprendre comment elles ont pu servir pour faire de la prévision pour le travail de collaboration qui a été effectué avec le Ministère de la Santé et des Services Sociaux du Québec.

1.1 Introduction aux séries chronologiques

1.1.1 Définition

Une série chronologique est représentée par une série de n observations numériques d'une variable X généralement mesurées à intervalles équidistants de temps. Notons toutefois qu'il est possible d'avoir des séries d'observations continues qui pourraient être représentées de façon continue à l'aide d'un graphe par exemple. Il est plus fréquent de travailler avec des séries d'observations qui ont été prises de manière systématique à des intervalles réguliers dans le temps. Ce type de séries chronologiques est donc caractérisé comme étant discret.

La série chronologique peut alors se noter comme suit :

x_t: observation au temps t

Série de n observations : $x_1, x_2, ..., x_n$

1.1.2 But

L'étude des séries chronologiques sert à faire de la prévision à court, moyen et long terme. Il existe des méthodes prévisionnelles quantitatives et qualitatives. Les méthodes quantitatives se subdivisent en deux catégories. Il y a les méthodes d'extrapolation qui produisent des prévisions sur le principe d'une corrélation de la variable étudiée avec le temps et les méthodes explicatives qui reposent sur les corrélations entre la variable étudiée et différentes variables explicatives. Les méthodes les plus utilisées de la première catégorie font l'objet de la section 1.3 et sont pour la plupart illustrées d'exemples tirés parfois des séries chronologiques du SIPPE et provenant d'autres fois de sources différentes.

1.1.3 Composantes

Les principales composantes décrivant une série chronologique sont : la tendance, le cycle, la saisonnalité et les fluctuations irrégulières [5] (p.4-5).

Tendance : cette composante porte sur les changements de croissance ou de décroissance tout au long de la série.

Cycle : cet aspect de la série fait référence à la présence d'une certaine récurrence et peut s'observer généralement sur des intervalles de plusieurs années.

Saisonnalité : il s'agit de la présence ou non d'un effet périodique qui se rapporte a une année (trimestres, semestres etc.)

Fluctuations irrégulières : cette caractéristique d'une série chronologique constitue la partie non expliquée par la tendance, le cycle ou la saisonnalité. Des événements rares qui peuvent difficilement être prédits sont souvent à l'origine de ces fluctuations.

1.1.4 Séries chronologiques du programme SIPPE

L'extrait suivant provient du rapport final remis au ministère et décrit bien les séries chronologiques étudiées dans le projet :

Les services intégrés en périnatalité et pour la petite enfance (SIPPE) visent à soutenir les familles vivant en contexte de vulnérabilité. La population ciblée par les SIPPE se divise en deux grands groupes. Le premier groupe est constitué des mères de moins de 20 ans et le second des mères de 20 ans et plus avec moins d'onze ans de scolarité. Les séries de données par réseau local de services (RLS) représentant le nombre de mères ayant accouché entre 1981 et 2009 en années financières (1^{er} mai au 30 avril) et associées au premier groupe, notées dans le cadre de ce rapport séries 1, sont généralement composées de plus petits nombres que les séries associées au deuxième groupe pour la même période de référence, notées dans la suite de ce rapport séries 2.

1.2 Modèles

Un modèle de série chronologique est en fait la spécification de sa loi de probabilité conjointe qui se traduit par sa fonction de répartition. L'élaboration d'un modèle probabiliste complet dans une analyse de séries chronologiques est rarement effectuée car ce processus impliquerait l'estimation de nombreux paramètres. On s'en tient généralement aux 2 premiers moments des lois conjointes : l'espérance $E[X_t]$ et la covariance $\text{cov}(X_t, X_{t-h})$ $t, h = 0, 1, ..., n$.

Rappelons rapidement les définitions de covariance et de corrélation. Soient X et Y deux variables aléatoires. La covariance s'exprime ainsi :

$$\text{cov}(X,Y) = E(XY) - E(X)E(Y).$$

La corrélation quant à elle, s'exprime de la façon suivante :

$$\text{cor}(X,Y) = \frac{\text{cov}(X,Y)}{\sigma_x \sigma_y}.$$

Des modèles simples de séries chronologiques souvent cités sont ceux avec une tendance linéaire, ceux avec une tendance polynomiale et ceux avec une tendance exponentielle. De plus, lorsqu'il y a présence d'une composante saisonnière on s'en remet généralement à la régression dynamique ou à l'utilisation de variables auxiliaires pour modéliser l'effet de saisonnalité. Ces différents modèles seront brièvement décrits dans ce qui suit. Le lecteur qui désire approfondir sur ces modèles, peut se rapporter aux références sur les séries chronologiques destinées aux étudiants gradués ou en voie de graduation suivantes : Bowerman & O'Connell (2006) et Brockwell & Davis (1996).

1.2.1 Exemples

Le modèle avec une tendance linéaire se compose d'une fonction déterministe notons-la m_t. Elle change dans le temps et est associée à une suite de variables aléatoires e_t de moyenne égale à 0. Ce modèle peut s'exprimer comme suit :

$$X_t = m_t + e_t.$$

Il est important de mentionner que les erreurs représentées par les variables aléatoires e_t ne sont pas obligatoirement indépendantes. Puisque le modèle est linéaire, l'équation représentant la composante m_t est de la forme :

$$m_t = \beta_0 + \beta_1 t.$$

Sur le même principe, le modèle avec tendance polynomiale est composé d'une fonction qui évolue dans le temps de façon polynomiale. Cette fonction s'exprime comme suit :

$$m_t = \beta_0 + \beta_1 t + \cdots + \beta_k t^k.$$

On utilise généralement la méthode des moindres carrés pour en estimer ses coefficients. Ce calcul s'effectue en minimisant l'entité suivante :

$$\sum_{t=1}^{n} (X_t - m_t)^2 = \sum_{t=1}^{n} \left(X_t - \beta_0 - \beta_1 t - \cdots - \beta_k t^k\right)^2.$$

Dans le même courant d'idées, une évolution exponentielle dans le temps caractérise le modèle avec tendance exponentielle. La tendance s'exprime donc par :

$$m_t = \beta_0 e^{\beta_1 t}.$$

1.2.2 Graphique

Le graphique suivant présente l'une des séries du projet SIPPE ayant une tendance exponentielle. Précisons que la plupart des séries affichaient une tendance exponentielle.

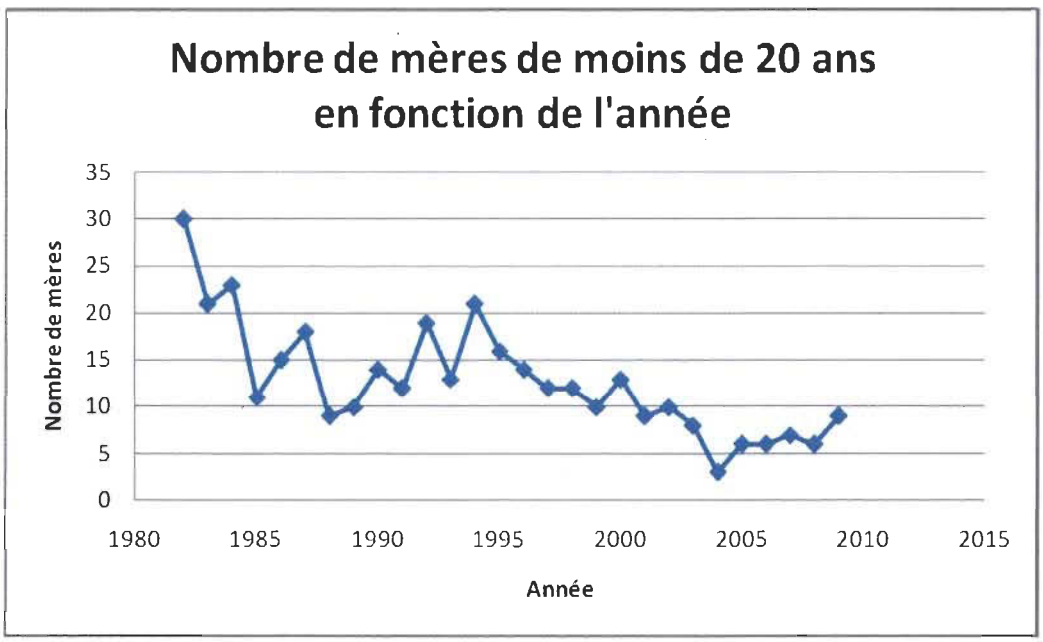


Figure 1 - Exemple de graphique d'une série affichant une tendance exponentielle

Pour estimer les paramètres de ce modèle il faut rendre la tendance linéaire en appliquant la transformation logarithmique et ensuite utiliser la méthode des moindres carrés sur la série transformée $m_t^{'}$. La transformation est traduite ici :

$$m_t^{'} = \log(m_t).$$

1.2.3 Estimation de la tendance

Pour rendre la série lisse, on peut appliquer un filtre qui fait partie des méthodes de lissage. Notons qu'à cet effet, un exemple de filtre de moyenne mobile finie pourrait être :

$$\hat{m}_t = W_t = \frac{1}{2g+1}\sum_{j=-g}^{g} X_{t-j} \quad \text{où } g+1<t\leq n-g,$$

et W_t correspond au filtre de moyenne mobile finie.

1.2.4 Composante saisonnière

Les séries chronologiques présentent souvent une composante saisonnière. Il se peut donc que l'amplitude soit stable dans le temps ou qu'elle varie en fonction du temps. Dans le second cas, il est préférable de transformer la série afin de stabiliser la composante saisonnière. Les transformations fréquemment utilisées à cette fin sont :

$$X_t^* = \log(X_t),$$

$$X_t^* = X_t^{\alpha} \text{ où } 0 < \alpha < 1 .$$

Par exemple, si l'on veut appliquer l'opération d'extraire la racine carrée des valeurs de la série initiale :

$$X_t^* = \sqrt{X_t}, \; X_t^* = X^{0.5}.$$

Une autre approche pour transformer les séries chronologiques avec composante saisonnière non stable dans le temps est d'appliquer la régression dynamique. La composante saisonnière s'exprime sous la forme suivante :

$$S_t = \beta_0 + \sum_{j=1}^{k}\left\{\beta_{1,j}\cos(\lambda_j t) + \beta_{2,j}\sin(\lambda_j t)\right\}$$

où les λ représentent les fréquences de Fourier et s'écrivent comme suit:

$$\lambda_j = 2\pi j / s$$

où $s = 12$. Les modèles peuvent s'exprimer par cycles de 4, 6 ou 12 mois par exemple. Un modèle par cycles de 4 mois aurait la forme suivante :

$$S_t = \beta_0 + \beta_1 \cos\left(2\pi \frac{3t}{s}\right) + \beta_2 \sin\left(2\pi \frac{3t}{s}\right).$$

Encore une fois, la méthode la plus utilisée pour estimer les coefficients pour l'estimation de la tendance est celle des moindres carrés. Le graphe suivant présente un exemple de série avec une composante saisonnière évidente.

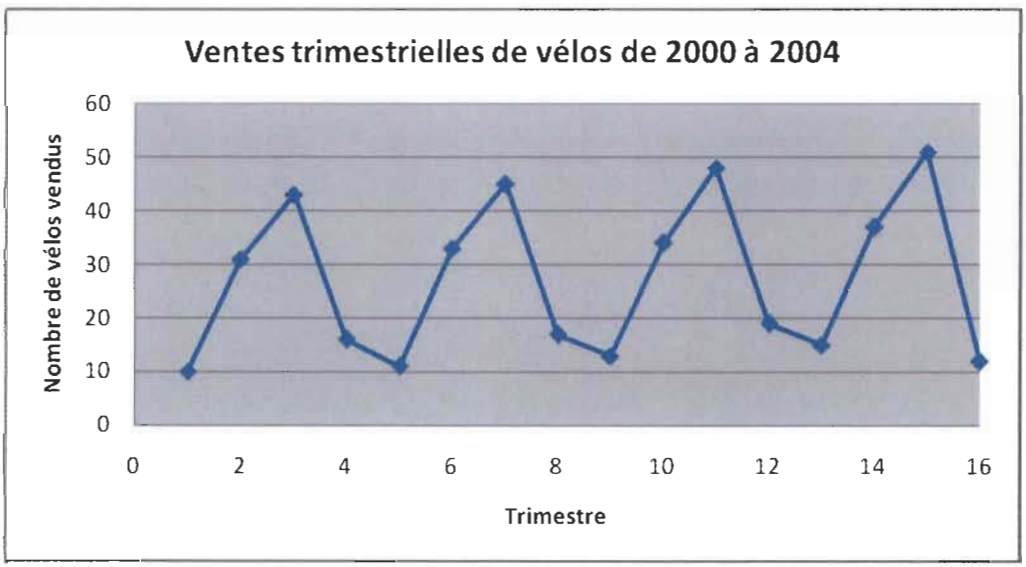


Figure 2 - Exemple d'une série affichant une composante saisonnière

Le modèle avec variables auxiliaires est une autre méthode très utilisée pour modéliser la saisonnalité. Le modèle est simple d'utilisation et s'exprime ainsi :

$$S_t = \beta_1 M_{1,t} + \cdots + \beta_{s-1} M_{s-1,t}$$

$$\text{où } M_{j,t} = \begin{cases} 1 & si\ t\ appartient\ à\ la\ saison\ j \\ 0 & \text{sinon} \end{cases}$$

$$\text{pour } j = 1, \ldots, s-1.$$

Ce modèle est souvent utilisé en pratique pour faire la modélisation de séries comportant une composante saisonnière.

1.3 Méthodes prévisionnelles courantes

1.3.1 Régression linéaire

La régression linéaire est l'une des méthodes les plus fréquemment utilisées pour faire de la prévision. Elle a pour but d'expliquer les fluctuations d'une variable dépendante notée Y en fonction d'une combinaison linéaire de variables $X_1, \ldots, X_k$ dites explicatives ou indépendantes. Le modèle général de la régression linéaire s'exprime avec la formule qui suit :

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$$

où les variables aléatoires ε_i avec $i = 1, \ldots n$ sont les termes d'erreur et les coefficients $\beta_0, \ldots, \beta_k$ sont les paramètres de la régression.

Précisons que la régression linéaire simple est le cas particulier où il n'y a qu'une variable explicative. Dans certains cas, il peut être intéressant d'utiliser la notation matricielle de la régression. Nous aborderons ici le cas général de la régression linéaire multiple.

1.3.1.1 Hypothèses fondamentales de la régression linéaire multiple

Les termes d'erreur sont indépendants et identiquement distribués selon une loi Normale $N(0,\sigma^2)$.

1.3.1.2 Méthode des moindres carrés

Les estimateurs $b_0,...,b_k$ des coefficients $\beta_0,...,\beta_k$ s'obtiennent en minimisant l'erreur quadratique moyenne relative aux résidus du modèle et s'écrit :

$$\sum_{i=1}^{n}\{y_i-(b_0+b_1x_{i1}+b_2x_{i2}+\cdots+b_kx_{ik})\}^2.$$

1.3.1.3 L'équation et la droite de régression

Une fois les estimateurs trouvés on peut écrire l'équation :

$$\hat{y}_i=b_0+b_1x_{i1}+b_2x_{i2}+\cdots+b_kx_{ik} \text{ avec } i=1,...n.$$

La droite de régression quant à elle s'écrit :

$$\bar{y}=b_0+b_1\bar{x}_1+b_2\bar{x}_2+\cdots+b_k\bar{x}_k.$$

On utilise l'équation de la régression pour prévoir les valeurs de la variable dépendante pour d'autres valeurs des variables indépendantes.

1.3.1.4 Résidus et variance associée aux termes d'erreur

Les résidus et l'estimation de la variance σ^2 associée aux erreurs ε_i sont définis tels que :

$$e_i=y_i-\hat{y}_i \text{ et } s^2=\frac{1}{n-k-1}\sum_{i=1}^{n}e_i^2$$

pour $i=1,...n.$

1.3.1.5 Étude de la significativité de la régression dans son ensemble

L'analyse de la variance est utile pour tester si la régression est significative dans son ensemble alors que les tests marginaux nous permettent de savoir si la contribution de chacune des variables explicatives est significative.

1.3.1.6 Analyse de la variance en régression linéaire multiple

La variance totale de la variable dépendante est la somme de la variation expliquée par la régression et la variation résiduelle. En effet,

$$SCT = SCR + SC_{res}$$

où SCT, SCR et SC_{res} désignent les sommes de carrés suivantes :

$$SCT = \sum_{i=1}^{n} (y_i - \bar{y})^2,$$

$$SCR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2,$$

$$SC_{res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

1.3.1.7 Carrés moyens

Les carrés moyens associées aux sommes SCR et SC_{res} sont obtenus en divisant les sommes de carrés par les nombres de degrés de liberté respectifs. Donc,

$$CMR = \frac{SCR}{k}$$

et $CM_{res} = \frac{SC_{res}}{n-k-1}$.

1.3.1.8 Coefficient de détermination

Le coefficient de détermination est une interprétation du pourcentage de variation expliquée par la régression. Il se calcule comme suit :

$$R^2 = \frac{SCR}{SCT}.$$

1.3.1.9 Test de signification global de la régression

On peut tester si la régression multiple à k variables indépendantes est significative dans son ensemble.

Posons $Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \varepsilon$ et confrontons les deux hypothèses suivantes :

$$H_0 : \beta_1 = \cdots = \beta_k = 0$$

$$H_1 : \text{il existe un } j = 1,...k \text{ tel que } \beta_j \neq 0.$$

Pour ce faire, nous utiliserons le rapport F du carré moyen avec le carré moyen résiduel. D'après le théorème de Cochran [6] (p.292), ce rapport devrait suivre une loi de Fisher sous H_0. Le rapport s'exprime :

$$F = \frac{CMR}{CM_{res}}.$$

Alors le test se résume à rejeter H_0 si $F > F_{\alpha,k,n-k-1}$. Si on rejette H_0, on conclut que la régression est significative dans son ensemble au seuil α. Sinon, la régression est considérée non significative.

Notons que la *p-value* associée à cette statistique est :

$$p - value = P(F > F_{obs} \mid H_0 : vraie).$$

1.3.1.10 Tests de contribution marginale

On peut s'intéresser à savoir si la contribution de chaque variable explicative est significative. Il s'agit de tester :

$$H_0 : \beta_j = 0 \text{ contre } H_1 : \beta_j \neq 0 \text{ pour } j = 1,...,k.$$

À cette fin, nous utilisons la statistique de test suivante :

$$t_j = \frac{b_j}{s(b_j)}.$$

Pour calculer $s(b_j)$, on s'en remet à un logiciel de statistique mais retenons que :

$$\text{var}(b) = \sigma^2 (X'X)^{-1} \text{ d'où } s^2(b) = s^2 (X'X)^{-1}$$

$$\text{alors que } s^2 = \frac{1}{n-k-1}\{Y'Y - b'X'Y'\}.$$

On rejette H_0 si $|t| > t_{\alpha/2,n-k-1}$ et on conclut que la variable explicative testée est significative au seuil α.

1.3.1.1 Intervalles de confiance de β_j

L'intervalle de confiance de β_j pour $j = 0,...,k$, est déterminé par :

$$IC(\beta_j) = \left[b_j - t_{\alpha/2,n-k-1} s(b_j), b_j + t_{\alpha/2,n-k-1} s(b_j) \right].$$

Il est bien de remarquer que l'expression $t_{\alpha/2,n-k-1} s(b_j)$ représente l'erreur d'estimation. On constate aussi que si $0 \in IC(\beta_j)$, la contribution marginale de la variable explicative X_j n'est pas significative.

1.3.1.12 Intervalles de confiance et prévision de la variable dépendante

Pour avoir une idée de la précision de l'estimation de $\hat{y}_h$, on construit un intervalle de confiance sur la moyenne $E(y_h)$. Au niveau de $1-\alpha$, l'intervalle est :

$$IC = \left[\hat{y}_h - s(\hat{y}_h)t_{\alpha/2,n-k-1}, \hat{y}_h + s(\hat{y}_h)t_{\alpha/2,n-k-1}\right],$$

$$\text{où } s(\hat{y}_h) = s\sqrt{X_h'(X'X)^{-1}X_h},$$

$$\text{et } s = \sqrt{\frac{Y'Y - b'X'Y}{n-k-1}}.$$

L'intervalle de prévision de y_h s'écrit :

$$IP = \left[\hat{y}_h - s(d_h)t_{\alpha/2,n-k-1}, \hat{y}_h + s(d_h)t_{\alpha/2,n-k-1}\right]$$

$$\text{où } s(d_h) = s\sqrt{1 + X_h'(X'X)^{-1}X_h}$$

$$\text{et } s = \sqrt{\frac{Y'Y - b'X'Y}{n-k-1}}.$$

1.3.1.13 Méthodes pour construire une équation de régression multiple

1-Introduction progressive

La méthode consiste à calculer les rapports des contributions marginales de chacune des variables explicatives et à sélectionner celle dont le rapport est maximal puis à l'introduire. Si aucune variable n'est retenue, la procédure s'arrête ici. Si une variable est retenue, on recalcule les rapports de contributions marginales des autres variables en tenant compte de l'introduction de cette première variable et on réapplique le procédé jusqu'à ce qu'il n'y ait plus de variables à introduire dans le modèle.

Le critère de sélection du rapport est :

$$\text{si } F_t > F_{\alpha,1,n-2}.$$

2-Élimination progressive

Cette méthode est similaire à l'introduction progressive à la différence qu'on élimine le plus petit rapport lorsque :

$$\text{si } F_i < F_{\alpha;1,n-k-1}.$$

3-Régression pas à pas

Cette procédure réunit les deux méthodes précédentes en introduisant et en retranchant les variables selon des critères d'entrée et de sortie déterminés au préalable.

Un exemple sera détaillé dans la section 3.7.1 portant sur les alternatives utilisées pour le traitement des séries problématiques du projet SIPPE.

Pour un approfondissement sur la théorie de la régression linéaire ainsi que l'illustration de nombreuses applications, il est mentionné dans *The Statistical analysis of Time Series*, par T. W. Anderson [7] de s'en remettre aux ouvrages des auteurs suivants : E. J. Williams (1959) [8], N.R Draper & H. Smith (1966) [9] et Franklin A. Graybill (1961) [10].

1.3.2 Méthodes de décomposition

Il n'y a pas de théorie statistique à proprement parler derrière les méthodes de décomposition, elles relèvent plutôt de l'intuition et de la pratique. Elles sont employées avec les séries qui montrent une tendance et une composante saisonnière. Le modèle additif est plus approprié dans le cas de séries dont la composante saisonnière est constante dans le temps alors que le modèle multiplicatif convient plus aux séries comportant une partie saisonnière qui décroit ou croit dans le temps. À titre d'exemple tiré de la vie courante, la méthode de décomposition développée par le *Bureau of the Census of the U.S department of Commerce* [11] sera brièvement présentée à la section 1.3.2.2.

1.3.2.1 Modèle additif

Une condition à l'application du modèle additif est que les paramètres décrivant la série ne changent pas au cours du temps. Le modèle additif résulte d'une addition de la tendance et de la partie saisonnière. Ce modèle s'écrit :

$$X_t = m_t + S_t + e_t \text{ avec } S_{t+s} = S_t .$$

Prenons l'exemple de la série de données suivantes :

Tableau 1 - Exemple d'un modèle additif

Ventes mensuelles d'une entreprise de 1967 à 1970			
1967	**1968**	**1969**	**1970**
153	228	187	170
189	283	201	243
221	255	292	178
215	238	220	248
302	164	233	202
223	128	172	163
201	108	119	139
173	87	81	120
121	74	65	96
106	95	76	95
86	145	74	53
87	200	111	94

Pour estimer la tendance m_t, on utilisera un filtre selon la période de la saisonnalité. Cela s'effectue à l'aide d'un filtre de moyenne mobile finie avec :

$$s = 2g$$

$$\text{et } \hat{m}_t = (0.5x_{t-g} + x_{t-g+1} + \cdots + x_{t+g-1} + 0.5x_{t+g})/s$$

où g+1<t ≤ n-g.

Pour estimer la partie saisonnière, il faut d'abord éliminer la tendance en calculant :

$$w_t = x_t - \hat{m}_t \text{ pour } t = g+1,...,n-g.$$

Ensuite, il faut calculer le résidu moyen $\overline{w}_j$ de tous les mois j qui consiste en la moyenne des $w_j, w_{j+s}, w_{j+2s},..., w_{j+ks}$ pour $k \leq (n-j)/s$.

Le résidu moyen s'écrit :

$$\overline{w}_j = \frac{1}{k+1}\sum_{t=0}^{k} w_{j+ks} \text{ où } k \leq (n-j)/s.$$

La composante saisonnière sera donc estimée par : $\hat{s}_t = \overline{w}_t - \frac{1}{s}\sum_{j=1}^{s} \overline{w}_j$, où $t = 1,...,s$.

On ajuste ensuite la tendance en prenant la série désaisonnalisée $d_t = x_t - \hat{s}_t$ avec la méthode des moindres carrés en l'appliquant aux points (t, d_t).

On en tire l'équation de prévision :

$$\hat{x}_t = \hat{d}_t + \hat{s}_t$$

où $\hat{d}_t$ s'exprime selon la nature de la tendance. Si la tendance est linéaire, $\hat{d}_t = b_0 + b_1 t$. Alors que si la tendance est quadratique, $\hat{d}_t = b_0 + b_1 t + b_2 t^2$.

Appliquons cette démarche à l'exemple des ventes mensuelles d'une entreprise de 1967 à 1970. D'abord, appliquons le filtre et trouvons les valeurs de m_t. Ensuite, il faut calculer le résidu moyen afin d'être en mesure d'estimer la composante saisonnière. La série dessaisonalisée nous permettra d'estimer la tendance et de sortir les prévisions à l'aide des paramètres estimés avec la méthode des moindres carrés. À quoi nous ajouterons l'effet saisonnier pour obtenir ainsi les résultats finaux. Les calculs précédemment mentionnés sont résumés dans le tableau de l'annexe A.

1.3.2.2 Modèle multiplicatif

Encore une fois, on doit vérifier que les paramètres décrivant la série ne changent pas au cours du temps avant d'appliquer le modèle. Le modèle se définit de la manière suivante:

$$X_t = m_t \times S_t \times e_t \text{ avec } S_{t+s} = S_t .$$

Le modèle tient son nom du fait qu'il se définit comme étant la multiplication de la fonction représentant sa tendance et du facteur représentant sa partie saisonnière. On mentionne aussi dans Bowerman (2006) [12] qu'à cela pourrait s'ajouter des facteurs qui représenteraient un effet cyclique ou une composante de fluctuations irrégulières s'il y a lieu. Pour l'appliquer, il suffit de suivre la même procédure que pour le modèle additif en l'adaptant pour tenir compte de l'effet multiplicatif de cette nouvelle décomposition.

1.3.2.3 Modèle de décomposition du *Bureau of the Census of the U.S department of Commerce*

Cette méthode constitue une extension de la méthode de décomposition multiplicative. La première version de la méthode fut développée par Julius Shiskin dans les années 50. Une particularité de cette méthode est qu'elle tient compte de faits inhérents à la situation du monde des affaires. Par exemple, elle tient compte qu'un mois ne contient pas toujours le même nombre de jours d'affaires. Il est intéressant de constater qu'on se sert de moyennes mobiles dans le procédé et qu'on identifie et remplace les valeurs extrêmes afin de limiter l'effet du hasard [5] (p.373). La procédure X-12 ARIMA permet d'utiliser une adaptation de la méthode à condition de détenir un minimum de 3 ans de données

historiques nécessaires à l'application. Pour plus d'informations sur le sujet, il est possible de consulter le site officiel du *U.S Census Bureau* [11].

1.3.3 Lissage exponentiel

Le lissage exponentiel ne repose sur aucune théorie statistique que ce soit mais relève plutôt de la pratique. Une fois de plus, on caractérise cette approche d'intuitive. On l'utilise avec les séries chronologiques sans composante saisonnière et dont la moyenne change lentement dans le temps. Le modèle accorde aussi plus d'importance aux observations les plus récentes. Cela permet de détecter tout changement dans les paramètres qui pourraient survenir [5] (p.379). Cela s'effectue avec un paramètre de lissage $\beta \in [0,1]$ retenu après une analyse des candidats potentiels lors de simulations. Pour les cas de lissage exponentiel double à un ou deux paramètres, les notations proposées de Bowerman (2006) ont été retenues.

1.3.3.1 Lissage exponentiel simple

Le lissage exponentiel simple convient aux séries qui n'ont pas de tendance. Le modèle prend la forme suivante :

$$\hat{m}_t = \beta x_t + (1-\beta)\hat{m}_{t-1} \text{ pour } t = 1,...,n.$$

La valeur initiale est estimée par la moyenne des données c'est-à-dire :

$$\hat{m}_0 = \frac{1}{n}\sum_{t=0}^{n} x_t .$$

Les intervalles de prévisions pour x_{n+1} et x_{n+2} sont :

$$IP = \left[\hat{m}_n - z_{\alpha/2}S, \hat{m}_n + z_{\alpha/2}S\right] \text{ et } IP = \left[\hat{m}_n - z_{\alpha/2}S\sqrt{1+\beta^2}, \hat{m}_n + z_{\alpha/2}S\sqrt{1+\beta^2}\right],$$

$$\text{avec } S^2 = \frac{1}{n-1}\sum_{t=1}^{n}(x_t - \hat{m}_{t-1})^2.$$

La valeur de $\hat{m}_n$ nous donne les prévisions pour x_{n+1} et x_{n+2}.

On remarque que les erreurs de prévision deviennent de plus en plus grandes en fonction du délai. Les contributions des données diminuent de façon exponentielle lorsqu'on remonte dans le temps en allant vers les données les plus anciennes.

Appliquons la méthode du lissage exponentiel simple aux données de l'exemple suivant afin de trouver les prévisions de x_{15} et x_{16} :

Tableau 2 - Série pour l'application du lissage exponentiel

Exemple fictif	
1	206
2	245
3	185
4	169
5	162
6	177
7	207
8	216
9	193
10	230
11	212
12	192
13	162
14	189

Les prévisions pour x_{15} et x_{16} sont les mêmes, ce sont les erreurs de prévision qui changent au fil du temps. Elle deviennent plus grandes.

Voici le tableau des résultats :

Tableau 3 - Résultats de l'application du lissage exponentiel

	Prévisions (mt)
1	199,05
2	212,835
3	204,4845
4	193,83915
5	184,287405
6	182,101184
7	189,570828
8	197,49958
9	196,149706
10	206,304794
11	208,013356
12	203,209349
13	190,846544
14	190,292581
15	**133,204807**
16	**133,204807**

Les prévisions de x_{15} et x_{16} sont donc toutes deux égales à 133,204807.

1.3.3.2 Lissage exponentiel double à un paramètre

Le lissage exponentiel double est utilisé dans le cas de séries chronologiques qui ont une tendance linéaire mais dont les paramètres ne changes pas dans le temps. Il est utilisé pour pondérer différemment les observations. Supposons que la série chronologique ait une tendance linéaire et s'écrive :

$$y_t = \beta_0 + \beta_1 t + \varepsilon.$$

1. Les paramètres de lissage peuvent alors s'exprimer comme :

$$S_T = \alpha y_T + (1-\alpha) S_{T-1} \text{ et } S_T^{[2]} = \alpha S_T + (1-\alpha) S_{T-1}^{[2]}$$

où α correspond au coefficient de lissage retenu.

2. Les estimations $b_0(T)$ et $b_1(T)$ de β_0 et β_1 se calculent :

$$b_0(T) = 2S_T - S_T^{[2]} - Tb_1(T)$$

$$\text{et } b_1(T) = \frac{\alpha}{1-\alpha}(S_T - S_T^{[2]}).$$

3. La prévision et son intervalle de confiance à $100(1-\alpha)\%$ s'expriment :

$$\hat{y}_{T+\tau} = \left(2 + \frac{\alpha\tau}{(1-\alpha)}\right)S_T - \left(1 + \frac{\alpha\tau}{(1-\alpha)}\right)S_T^{[2]} \text{ et}$$

4. $\left[\hat{y}_{T+\tau}(T) \pm z_{\alpha/2} d_t \Delta(T)\right]$,

$$\text{où } \Delta(T) = \frac{\sum_{t=1}^{T} |y_t - \hat{y}_t(t-1)|}{T},$$

$$d_t = 1.25\left[\frac{1 + \frac{\alpha}{(1+v)^3}\left[(1+4v+5v^2) + 2\alpha(1+3v)\tau + 2\alpha^2\tau^2\right]}{1 + \frac{\alpha}{(1+v)^3}\left[(1+4v+5v^2) + 2\alpha(1+3v) + 2\alpha^2\right]}\right]^{1/2}$$

$$\text{et } v = 1-\alpha.$$

1.3.3.3 Lissage exponentiel double à deux paramètres – Holt-Winters

Dans le cas particulier d'une série chronologique comportant une tendance linéaire [12] (p.467) f qui change avec un taux fixé dit β_1 et une composante saisonnière SN_t qui est additive et constante dans le temps, on écrit le modèle comme suit :

$$y = (\beta_0 + \beta_1 t) + SN_t + \varepsilon.$$

On note alors que la série à $T-1$ vaut β_0 et à T prend la valeur $\beta_0 + \beta_1 T$. Pour appliquer la méthode *Holt-Winters* on utilise l'opérateur l_{T-1} pour estimer le niveau de la série au temps $T-1$ et b_{T-1} estime le taux d'accroissement de la série en $T-1$. Supposons que

nous obtenons une nouvelle valeur de y_t et que sn_{T-L} est la plus récente estimation de la composante saisonnière où L correspond au nombre de saisons.

Nous pouvons donc exprimer l'estimation du niveau de la série à l'aide du coefficient de lissage α :

$$l_T = \alpha(y_t - sn_{T-L}) + (1-\alpha)(l_{T-1} + b_{T-1})$$

$$\text{avec } b_T = \gamma(l_T - l_{T-1}) + (1-\gamma)b_{T-1}$$

$$\text{et } sn_T = \delta(y_T - l_T) + (1-\delta)sn_{T-L}$$

où γ et δ sont les coefficients de lissage respectifs de b_T et sn_T.

Les prévisions sont données par l'équation suivante :

$$\hat{y}_{T+\tau}(T) = l_T + \tau b_T + sn_{T+\tau-L}.$$

L'intervalle de prévision avec un niveau de confiance de 95% qui lui est associé s'écrit :

$$\left[\hat{y}_{T+\tau} \pm z_{[0,025]} s\sqrt{c_\tau}\right]$$

$$\text{où lorsque } \tau = 1 : c_1 = 1,$$

$$2 \le \tau \le L : c_\tau = \left[1 + \sum_{j=1}^{\tau-1} \alpha^2(1+j\gamma)^2\right],$$

$$L \le \tau : c_\tau = 1 + \sum_{j=1}^{\tau-1}\left[\alpha(1+j\gamma) + d_{j,L}(1-\alpha)\delta\right]^2$$

$$\text{où } d_{j,L} = 1 \text{ si } j \text{ est un entier multiple de } L \text{ ou sinon égal à 0.}$$

L'écart-type s au temps T est défini de la manière suivante :

$$s = \sqrt{\frac{SSE}{T-3}} = \sqrt{\frac{\sum_{t=1}^{T}\left[y_t - \hat{y}_t(t-1)\right]^2}{T-3}} = \sqrt{\frac{\sum_{t=1}^{T}\left[y_t - (l_{t-1} + b_{t-1} + sn_{T-L})\right]^2}{T-3}}.$$

Conclusion

En résumé, le premier chapitre se voulait essentiellement un prélude sur la théorie des séries chronologiques. À cet effet, les caractéristiques et les principales méthodes de prévision quantitatives ont été abordées. Les quelques exemples et illustrations qui ont été présentés ont permis de se familiariser avec les séries chronologiques et faciliteront la compréhension de la méthodologie utilisée dans le projet. L'approche Box-Jenkins qui a été employée pour faire les prévisions du projet de collaboration avec le Ministère de la Santé et des Services Sociaux est présentée dans le chapitre suivant.

CHAPITRE 2

MÉTHODOLOGIE BOX-JENKINS - MODÈLES NON-SAISONNIERS

Ce chapitre traite de la modélisation des séries chronologiques à l'aide de l'approche Box-Jenkins. Il se divise en cinq parties. Il débute en résumant les étapes principales de cette approche qui seront ensuite plus détaillées et appuyées d'exemples dans le chapitre 3. En second lieu, la stationnarité est abordée car la méthodologie s'applique uniquement pour les séries stationnaires. En troisième partie, on traite des modèles classiques et des conditions d'application de ceux-ci. Ce traitement comprend l'analyse de la stationnarité, de l'inversibilité et des fonctions d'autocorrélation. Ces analyses nous conduisent à l'identification des candidats du modèle et à l'estimation des paramètres du modèle retenu et cela constitue la quatrième partie du chapitre. Finalement, la cinquième traite des aspects portant sur les prévisions rattachées au modèle désigné.

2.1 Étapes de la méthodologie Box-Jenkins

Dans la littérature sur le sujet, il est généralement admis que la méthodologie Box-Jenkins peut se résumer en 4 (parfois 5) étapes. Ici, la définition présentée dans Bowerman (2006) à été retenue [5] (p.436) :

1- Identification du modèle approprié

2- Estimation des paramètres

3- Vérification du diagnostic (test d'adéquation et amélioration du modèle s'il y a lieu)

4- Prévisions

Il faut d'abord appliquer les trois premières étapes et si les résultats sont concluants il est justifié de passer à la quatrième étape. Le cas échéant, il faut reprendre les trois premières étapes jusqu'à l'obtention de résultats satisfaisants. Il est possible qu'il n'y ait pas de modèle se prêtant à la situation. Dans le cadre du projet SIPPE, pour les cas où il n'y avait pas de modèle approprié des alternatives ont été proposées.

L'identification du modèle se fait après avoir vérifié que la série est stationnaire. Une fois cette condition assurée, on s'en remet aux diagrammes d'autocorrélation et d'autocorrélation partielle pour la suggestion de candidats de modèles ARIMA (*AutoRegressive Integrated Moving Average*).

Pour chacun des candidats retenus, on génère ensuite l'estimation des paramètres puis on compare les différents modèles.

Les modèles sont validés à l'aide de la statistique Ljung-Box puis comparés sur la base de l'erreur type (*Standard Error Estimate*) qui nous indique des erreurs de prévision moins grandes. Dans le projet SIPPE, un indicateur (RMRES2) a été créé à cet effet et il permettait de faire cette vérification. Il était calculé à partir des prévisions du modèle retenu en lui retirant les cinq dernières données. Il constitue la racine de la moyenne des erreurs de prévision mises au carré. Les erreurs de prévisions ont été obtenues en faisant la différence entre la donnée réelle et la prévision pour les cinq dernières années.

Une fois le meilleur modèle retenu, les prévisions s'y rattachant sont produites.

2.2 Stationnarité

2.2.1 Définition

Une série $\{X_t\}$ est considérée comme étant stationnaire lorsque ses propriétés statistiques (moyenne, variance) sont constantes dans le temps. Il existe plusieurs tests de stationnarité. Mentionnons entre autres le test de la racine unitaire. Il est possible de vérifier si la série est stationnaire ou pas en scrutant le graphique des valeurs de la série en fonction du temps ou en étudiant le graphique de la fonction d'autocorrélation lui étant associé. Cet examen sera détaillé ultérieurement.

2.2.2 Transformations

Dans le cas où la série n'est pas stationnaire, il existe des moyens pour la rendre stationnaire. Il faut parfois stabiliser la variance lorsqu'on observe une augmentation de la variance de la série dans le temps. Pour ce faire, il est possible d'utiliser la transformation logarithmique.

Elle se résume comme suit :

$$Y_t = \ln(X_t).$$

Sur le même principe, on peut extraire la racine carrée des données initiales de la série :

$$Y_t = \sqrt{X_t}.$$

Ensuite, il est possible de stabiliser la moyenne lorsque la série présente une tendance à l'aide d'une ou plusieurs différenciations avec l'opérateur Δ. Il est important d'appliquer la transformation logarithmique avant cette dernière si elle est nécessaire car le processus de différenciation implique des valeurs négatives qui ne sont pas compatibles avec le modèle logarithmique.

Soit $\{Y_t\}$ le processus :

$$Y_t = \Delta X_t = X_t - X_{t-1}.$$

2.3 Modèles classiques

Dans cette section, nous allons définir et étudier les caractéristiques des principaux processus de modèles ARIMA.

2.3.1 Moyennes mobiles

Soit $\{X_t\}$ une série chronologique stationnaire.

Commençons par traiter le modèle de moyenne mobile d'ordre q (noté $MA(q)$) défini par :

$$X_t = \mu + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}$$

où $\{a_t\}$ est un bruit blanc, c'est-à-dire une suite de variables aléatoires indépendantes et identiquement distribuées de moyenne 0 et de variance σ^2 où $\theta_1,...,\theta_q$ sont les paramètres tels que $\theta_q \neq 0$.

2.3.2 Autorégressifs

Le modèle autorégressif d'ordre p (noté $AR(p)$) est quant à lui défini par:

$$X_t - \mu = \phi_1(X_{t-1} - \mu) + \cdots + \phi_p(X_{t-p} - \mu) + a_t$$

où $\{a_t\}$ est le bruit blanc caractérisé de la même façon que pour le modèle de moyenne mobile à la différence que $\phi_1,...,\phi_p$ en sont les paramètres tels que $\phi_p \neq 0$.

2.3.3 Conditions d'inversibilité et de stationnarité

2.3.3.1 Condition d'inversibilité de $MA(q)$

Soit $\{X_t\}$ un processus moyenne mobile d'ordre q noté $MA(q)$.

Soit B l'opérateur de retard défini par :

$$BX_t = X_{t-1}.$$

Il est donc possible d'écrire le processus en fonction de B :

$$X_t = \mu + a_t - \theta_1 B a_t - \cdots - \theta_q B^q a_t \quad [1]$$

donc,

$$X_t - \mu = (1 - \theta_1 B a_t - \cdots - \theta_q B^q a_t) = \theta B(a_t)$$

où $\theta(z)$ désigne le polynôme : $\theta(z) = 1 - \theta_1 z - \cdots - \theta_q z^q$.

Pour être en mesure d'écrire a_t en fonction du passé, il nous faut inverser l'opérateur B dans l'équation [1].

2.3.3.2 Théorème d'inversibilité

Le processus $MA(q)$ est inversible si toutes les racines du polynôme $\theta(z)$ sont hors du disque unité ou autrement dit que leurs modules soient strictement plus grands que 1. Si tel est le cas alors,

$$(\theta(z))^{-1} = \sum_{k=0}^{\infty} \psi_k z^k.$$

En effet,

$$\theta(B) a_t = (X_t - \mu) \Rightarrow a_t = (\theta(B))^{-1} (X_t - \mu)$$

d'où

$$a_t = \left(\sum_{k=0}^{\infty} \psi_k B^k \right) (X_t - \mu) = \sum_{k=0}^{\infty} \psi_k (X_{t-k} - \mu).$$

2.3.4 Stationnarité des modèles autorégressifs

2.3.4.1 Stationnarité du modèle $AR(1)$

Étudions d'abord la stationnarité du cas particulier où $p=1$.

Soit $\{X_t\}$ le processus :

$$X_t-\mu=\phi(X_{t-1}-\mu)+a_t.$$

Posons :

$$\begin{aligned} Y_t &= \phi Y_{t-1}+a_t \\ &= \phi(\phi Y_{t-2}+a_{t-1})+a_t \\ &= \phi^2(Y_{t-2}+a_{t-1})+a_t \\ &\vdots \\ &= \phi^{n+1}Y_{t-n-1}+\sum_{k=0}^{n}\phi^k a_{t-k}. \end{aligned}$$

Nous constatons alors que si $|\phi|<1$, $\phi^{n+1}Y_{t-n-1}$ converge vers 0 . De ce constat nous en tirons que :

$$Y_t=\sum_{k=0}^{\infty}\phi^k a_{t-k}\,.$$

Cela nous permet d'affirmer que $\{Y_t\}$ est stationnaire si $|\phi|<1$.

Généralisons maintenant le résultat obtenu.

2.3.4.2 Stationnarité du modèle $AR(p)$

Soit $\{X_t\}$ un processus $AR(p)$, il s'ensuit que :

$$a_t = (1-\phi_1 B-\cdots-\phi_p B^p)(X_t-\mu) = \phi(B)\left(X_t-\mu\right)$$

où $\phi(z)$ désigne le polynôme caractéristique associé à $\{X_t\}$. Le polynôme est défini de la manière suivante :

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

Pour que $AR(p)$ soit stationnaire, l'opérateur $\phi(B)$ doit être inversible. Cette condition peut se résumer à l'aide du théorème suivant.

2.3.4.2 Théorème d'inversibilité pour $AR(p)$

Un processus $AR(p)$ est dit stationnaire si toutes les racines de son polynôme caractéristique sont hors du disque unitaire, ce qui implique que leurs modules sont strictement supérieurs à 1. Dans ce cas, on peut écrire le processus en fonction du bruit blanc $\{a_t\}$ à l'aide de l'équation suivante :

$$X_t = \mu+\sum_{k=0}^{\infty}\pi_k a_{t-k}\ .$$

Remarquons que pour le cas particulier du processus $AR(1)$ cela s'écrit :

$$a_t = (X_t-\mu)-\phi_1(X_{t-1}-\mu) = \phi(B)(X_t-\mu),$$

ici $\phi(z) = 1-\phi_1 z$. La racine du polynôme est donc tout simplement $1/\phi_1$, ce qui revient bel et bien à dire que le processus est stationnaire si $|1/\phi_1|>1$ ou dit autrement $|\phi_1|<1$.

2.3.5 Fonctions d'autocorrélation

2.3.5.1 Fonction de covariance

Considérons un processus stationnaire au sens large qui a pour conditions :

1. $\Gamma_X(t)$ indépendant de t.

2. $\gamma_X(t+h,t)$ indépendant de t pour tout h.

Il est important de mentionner que la moyenne est constante, que $\Gamma_X(t)=\mu$ et que la fonction covariance ne dépend que du délai h ainsi $\gamma_X(t+h,t)=\gamma(h)$.

Les propriétés de $\gamma(h)$ sont :

1. $\gamma(0)$ représente la variance du processus donc : $\gamma(0)=\text{var}(X_t)\geq 0$.

2. γ est une fonction paire c'est-à-dire : $\gamma(-h)=\gamma(h)\ \ \forall h$.

3. La variance du vecteur fonction $\tilde{X}_n=(X_1,...,X_n)^T$ s'exprime en termes de la fonction γ. En effet, pour tout $n\geq 1$ et $a_1,...,a_n\in\mathbb{R}$:

$$\text{var}(X_n)=\left[\text{cov}(X_i,X_j)\right]_{i,j=1}^{n}=\left[\gamma(i-j)\right]_{i,j=1}^{n}.$$

De ces propriétés découle :

$$\text{var}(X_n)=\begin{pmatrix}\gamma(0) & \gamma(-1) & \cdots & \gamma(-n+1)\\ \gamma(1) & \gamma(0) & \cdots & \gamma(-n+2)\\ \vdots & \vdots & \vdots & \vdots\\ \gamma(n-1) & \gamma(n-2) & \cdots & \gamma(0)\end{pmatrix}.$$

2.3.5.2 Définition et propriétés de la fonction d'autocorrélation

La fonction d'autocorrélation est définie par :

$$\rho(h) = cor(X_{t+h}, X_t) = \frac{\text{cov}(X_{t+h}, X_t)}{\sqrt{\text{var}(X_{t+h})\,\text{var}(X_t)}}.$$

Nous pouvons donc en conclure que :

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} \text{ où } h \in \mathbb{Z}.$$

Cette fonction représente l'autocorrélation de $\{X_t\}$ avec un délai h.

Les propriétés de $\rho(h)$ sont :

1. $\rho(0) = 1$

2. $|\rho(h)| \leq 1$

3. La fonction ρ est définie positive, c'est-à-dire que pour tout $n \geq 1$ et $a_1, ..., a_n \in \mathbb{R}$:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \rho(i-j) \geq 0.$$

2.3.5.3 Moyennes mobiles

2.3.5.3.1 Fonctions d'autocovariance et autocorrélation théoriques

Soit $\{X_t\}$ un processus $MA(q)$, alors :

$$E(X_t) = \mu,$$

$$\gamma(h) = \begin{cases} \sigma^2 \sum_{i=0}^{q-h} \theta_i \theta_{i+h}, \\ 0 \end{cases}$$

$$\rho(h) = \begin{cases} \dfrac{\sum_{i=0}^{q-h} \theta_i \theta_{i+h}}{\sum_{i=0}^{q} \theta_i^2} \\ 0 \end{cases}$$

avec $\theta_0 = -1$.

2.3.5.3.2 Fonctions d'autocovariance et d'autocorrélation empiriques

Soit $x_1, ..., x_n$ une série chronologique. La fonction d'autocovariance d'ordre k peut être estimée par :

$$\hat{\gamma}(k) = c_k = \sum_{t=1}^{n-k} (x_t - \bar{x})(x_{t+k} - \bar{x}).$$

Sur le même principe, l'estimation de la fonction d'autocorrélation est donnée par l'équation suivante :

$$\hat{\rho}(k) = r_k = \frac{\sum_{t=1}^{n-k} (x_t - \bar{x})(x_{t+k} - \bar{x})}{\sum_{t=1}^{n} (x_t - \bar{x})^2} = \frac{c_k}{c_0},$$

où

$$\bar{x} = \frac{1}{n} \sum_{t=1}^{n} x_t.$$

2.3.5.3.3 Corrélogramme de la fonction d'autocorrélation

Le corrélogramme est une représentation graphique traduisant r_k en fonction de k. Ce graphique peut servir à l'identification de l'ordre q d'un processus $MA(q)$. Par la définition du modèle de moyenne mobile, on sait que les autocorrélations théoriques $\rho(k)$ s'annulent pour $k > q$. Nous sommes donc en mesure de détecter l'ordre q en scrutant le corrélogramme et il s'agit de l'ordre q tel que les r_k sont petits pour $k > q$. Il ne reste plus qu'à modéliser la série à l'aide de cette information.

2.3.5.4 Autorégressifs

2.3.5.4.1 Fonction d'autocorrélation théorique

Soit $\{X_t\}$ un processus $AR(p)$, alors en écrivant:

$$X_t = \mu + \sum_{i=0}^{\infty} \pi_i a_{t-i},$$

il s'ensuit que :

$$\begin{aligned}
\gamma(h) &= \operatorname{cov}(X_t, X_{t+h}) \\
&= \operatorname{cov}\left(\mu + \sum_{i=0}^{\infty} \pi_i a_{t-i}, \mu + \sum_{j=0}^{\infty} \pi_j a_{t+h-j} \right) \\
&= \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \pi_i \pi_j \operatorname{cov}(a_{t-i}, a_{t+h-j}).
\end{aligned}$$

Le fait que les variables aléatoires $\{a_t\}$ soient indépendantes entraîne que :

$$\text{cov}(a_{t-i}, a_{t+h-j}) = \begin{cases} \sigma^2 & \text{si} \quad j = h+i \\ 0 & \text{si} \quad j \neq h+i \end{cases}.$$

Avec cette constatation on peut en déduire que :

$$\gamma(h) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \pi_i \pi_j \,\text{cov}(a_{t-i}, a_{t+h-j})$$
$$= \sigma^2 \sum_{i=0}^{\infty} \pi_i \pi_{i+h}.$$

Il en découle ainsi que :

$$\gamma(0) = \sigma^2 \sum_{i=0}^{\infty} \pi_i^2 ,$$

ce qui nous conduit au résultat suivant :

$$\rho(h) = \frac{\sum_{i=0}^{\infty} \pi_i \pi_{i+h}}{\sum_{i=0}^{\infty} \pi_i^2} \text{ pour tout } h .$$

Équation de Yule-Walker

Cette équation est intéressante car elle permet d'établir une relation de récurrence entre les autocorrélations $\rho(k)$ d'un processus $AR(p)$.

Elle s'exprime :

$$\rho(k) = \phi_1 \rho(k-1) + \phi_2 \rho(k-2) + \cdots + \phi_p \rho(k-p) \text{ pour } k > 0 .$$

Soit un processus $AR(p)$ défini par :

$$X_t - \mu = \phi_1(X_{t-1} - \mu) + \cdots + \phi_p(X_{t-p} - \mu) + a_t .$$

Écrivons ce processus de la manière suivante :

$$Y_t = X_t - \mu .$$

Il s'ensuit que :

$$Y_t = \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + a_t .$$

Considérons maintenant :

$$\begin{aligned}
\gamma(h) &= \operatorname{cov}(Y_t, Y_{t-k}) \\
&= \operatorname{cov}\left(\sum_{i=1}^{p} \phi_i Y_{t-i} + a_t, Y_{t-k} \right) \\
&= \sum_{i=1}^{p} \phi_i \operatorname{cov}(Y_{t-i}, Y_{t-k}) + \operatorname{cov}(a_t, Y_{t-k}) \\
&= \sum_{i=1}^{p} \phi_i \gamma(k-i).
\end{aligned}$$

Nous obtenons l'équation recherchée en divisant par $\gamma(0)$:

$$\rho(k) = \sum_{i=1}^{p} \phi_i \rho(k-i).$$

L'utilité de l'équation Yule-Walker réside dans le fait qu'il est ainsi plus facile de trouver la fonction d'autocorrélation $\rho(k)$ d'un processus $AR(p)$.

En effet, il s'agit de calculer les $\rho(k)$ avec l'équation suivante :

$$\rho(k) = A_1 \left(\frac{1}{r_1} \right)^k + \cdots + A_p \left(\frac{1}{r_p} \right)^k , \; k > 0$$

où $r_1, \ldots, r_p$ désignent les racines du polynôme $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$.

Les constantes $A_1,...,A_p$ sont déterminées par la condition initiale $\rho(0)=1$ et les $p-1$ premières équations de Yule-Walker.

2.3.5.4.2 Fonction d'autocorrélation partielle théorique et empirique

Pour l'identification des processus $AR(p)$, il nous faut introduire la notion d'autocorrélation partielle.

Soit $P_k, k \geq 1$, la matrice des corrélations telle que :

$$P_k = (\rho(i-j)), \qquad 1 \leq i, j \leq k.$$

Pour tout $k \geq 1$, définissons les nombres $\Phi_{k,1},...,\Phi_{k,k}$ comme solutions du système :

$$P_k \begin{pmatrix} \Phi_{k,1} \\ \vdots \\ \Phi_{k,k} \end{pmatrix} = \begin{pmatrix} \rho(1) \\ \vdots \\ \rho(k) \end{pmatrix}.$$

Le nombre $\Phi_{k,k}$ représente l'autocorrélation partielle d'ordre k.

Théorème

Pour tout processus $AR(p)$, on a :

$$\Phi_{k,k} = 0 \text{ si } k > p.$$

Estimation de la fonction d'autocorrélation partielle

$\Phi_{k,k}$ est estimée par $\hat{\Phi}_{k,k}$ qui est calculée à partir des r_i qui sont elles-mêmes les estimations de ρ_i.

2.3.5.4.3 Corrélogramme de la fonction d'autocorrélation partielle

La relation exprimée dans cette représentation graphique est celle de $\hat{\Phi}_{k,k}$ en fonction de k. Lorsque les valeurs de $\hat{\Phi}_{k,k}$ deviennent très faibles à partir d'un certain $k = p$, il est raisonnable de penser que p est l'ordre du modèle autorégressif. De cette façon, il est possible d'émettre des hypothèses sur l'identification du modèle.

Cas particuliers de $AR(1)$ **et** $AR(2)$

Lorsque les $\hat{\Phi}_{k,k}$ du corrélogramme de la fonction d'autocorrélation partielle deviennent très petits à partir de $k = 2$, il est raisonnable de penser que le modèle autorégressif d'ordre 1 modélise adéquatement la série.

Le modèle se résumera donc ainsi :

$$x_t = \mu - \phi_1(x_{t-1} - \mu) + a_t,$$

$$\hat{\Phi}_{1,1} = \rho(1),$$

$$\hat{\Phi}_{k,k} = 0 \text{ pour } k > 1.$$

Alors que dans le cas où les $\hat{\Phi}_{k,k}$ deviennent très petits à partir de $k = 3$, on recourt plutôt à un modèle autorégressif d'ordre 2 pour modéliser la série.

Le modèle se résumera donc ainsi :

$$x_t = \mu - \phi_1 x_{t-1} - \phi_2 x_{t-2} + a_t,$$

$$\hat{\Phi}_{1,1} = \rho(1),$$

$$\hat{\Phi}_{2,2} = \left[\rho(2) - \rho(1)^2\right] / \left[1 - \rho(1)^2\right],$$

$$\hat{\Phi}_{k,k} = 0 \text{ pour } k > 2.$$

2.2.3 Mixtes

2.2.3.1 $ARMA(p,q)$

2.2.3.1.1 Définition

Soit $\{a_t\}$ un bruit blanc de moyenne 0 et de variance σ^2.

Si $\{X_t\}$ suit un processus $ARMA(p,q)$ alors il existe des constantes $\mu, \phi_1, ..., \phi_p, \theta_1, ..., \theta_q$ telles que :

$$X_t - \mu - \sum_{i=1}^{p} \phi_i (X_{t-i} - \mu) = a_t - \sum_{j=1}^{q} \theta_j a_{t-j}.$$

Cela implique :

$$\phi(B)(X_t - \mu) = \theta(B) a_t,$$

où :

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \text{ et } \theta(z) = 1 - \theta_1 z - \cdots - \theta_q z^q.$$

2.2.3.1.2 Stationnarité et inversibilité

Si les racines du polynôme $\phi(z)$ sont en dehors du cercle unité, le processus est considéré stationnaire. Si les racines du polynôme $\theta(z)$ sont en dehors du cercle unité, le processus est donc inversible.

2.2.3.1.3 Autre écriture du processus $ARMA(p,q)$

Il est possible de décomposer un modèle $ARMA(p,q)$ en termes de processus MA et AR pures.

Cela s'écrit :

$$X_t - \mu = f(B)a_t \text{ et } a_t = g(B)(X_t - \mu)$$

où

$$f(B) = \frac{\theta(B)}{\phi(B)} \text{ et } g(B) = \frac{\phi(B)}{\theta(B)}.$$

2.2.3.2 Modèles intégrés $ARIMA(p,d,q)$

Dans la réalité, il est assez fréquent que la série ne soit pas stationnaire initialement. Il faut donc transformer la série pour la rendre stationnaire. La différence est souvent appliquée pour stabiliser la moyenne. Pour ce faire, on utilise l'opérateur différence ∇ autant de fois que nécessaire.

Cela s'exprime de la manière suivante :

$$\nabla X_t = X_t - X_{t-1}$$
$$\nabla^2 X_t = X_t - 2X_{t-1} + X_{t-2}$$
$$\vdots$$

2.4. Estimation des paramètres

2.4.1 Estimation des paramètres du modèle de la moyenne mobile

L'estimation des paramètres du modèle de moyenne mobile est un peu plus compliquée que celle du modèle autorégressif du au fait qu'il n'est pas possible de trouver des estimateurs explicites. Il faut donc s'en remettre à des procédures itératives afin de minimiser l'erreur quadratique résiduelle.

Procédure d'estimation itérative

1-Idée sous-jacente à la procédure

En remarquant que le paramètre θ est lié à l'autocorrélation d'ordre 1 par la relation suivante :

$$\rho(1) = \frac{\theta}{1+\theta^2} \text{ avec } |\theta| < 1,$$

et que par le fait même l'estimateur $\hat{\theta}$ est lié à l'autocorrélation empirique d'ordre 1 :

$$r_1 = \frac{\hat{\theta}}{1+\hat{\theta}^2} \quad [2].$$

Le principe de la méthode peut donc se résumer à l'estimation de μ par $\hat{\mu} = \bar{x}$ suivi de la détermination de $\hat{\theta}^*$ par la résolution de l'équation $[2]$.

2-Étapes de la procédure

1. D'abord, il faut choisir $a_0 = 0$ puis calculer $a_1,...,a_n$ de la manière suivante :`

$$\begin{aligned} a_1 &= x_1 - \bar{x}, \\ a_2 &= x_2 - \bar{x} - \hat{\theta}^* a_1, \\ &\vdots \\ a_n &= x_n - \bar{x} - \hat{\theta}^* a_{n-1}. \end{aligned}$$

2. Il faut ensuite calculer la somme des carrés résiduels :

$$\sum_{t=1}^{n} a_t^2 .$$

3. Il s'agit de répéter les étapes précédentes pour différentes valeurs de $(\hat{\mu},\hat{\theta})$ choisis au voisinage de $(\bar{x},\hat{\theta}^*)$ calculé au préalable.
4. Finalement, après avoir effectué les analyses, on détermine le couple $(\hat{\mu},\hat{\theta})$ qui minimise la somme des carrés résiduels.

2.4.2 Estimation des paramètres du modèle autorégressif

Pour faire l'estimation des paramètres du modèle autorégressif il suffit d'appliquer la méthode des moindres carrés pour minimiser :

$$\sum_{t=p+1}^{n} \left[x_t - \mu - \phi_1 (x_{t-1} - \mu) - \cdots - \phi_p (x_{t-p} - \mu) \right]^2$$

par rapport aux paramètres μ et $\phi_1,...,\phi_p$.

2.4.2.1 Cas particulier du processus $AR(1)$

Soit un processus $AR(1)$:

$$X_t - \mu = \phi(X_{t-1} - \mu) + a_t .$$

L'estimation des paramètres μ et ϕ par la méthode des moindres carrés consiste à minimiser par rapport à μ et ϕ la fonction suivante :

$$L(\mu,\phi) = \sum_{t=2}^{n} [x_t - \mu - \phi(x_{t-1} - \mu)]^2 .$$

Les estimations $\hat{\mu}$ et $\hat{\phi}$ de μ et ϕ sont des solutions des équations :

$$\frac{\partial}{L(\hat{\mu},\hat{\phi})} = 0 \text{ et } \frac{\partial L(\hat{\mu},\hat{\phi})}{\partial \hat{\phi}} = 0.$$

Après simplification, on obtient les deux équations :

$$\hat{\mu} = \frac{\overline{x}_2 - \hat{\phi}\overline{x}_1}{1-\hat{\phi}} \text{ et } \hat{\phi} = \frac{\sum_{t=1}^{n-1}(x_t - \hat{\mu})(x_{t-1} - \hat{\mu})}{\sum_{t=1}^{n-1}(x_t - \hat{\mu})^2}, \qquad [3]$$

où $\overline{x}_1$ et $\overline{x}_2$ représentent les moyennes respectives des $n-1$ premières et $n-1$ dernières observations. Notons qu'il n'est pas possible d'obtenir $\hat{\mu}$ et $\hat{\phi}$ à partir des équations [3]. Pour y parvenir, il faut utiliser les approximations suivantes :

$$\overline{x}_1 \approx \overline{x}_2 \approx \overline{x},$$

où $\overline{x}$ représente la moyenne des n observations : $x_1,...,x_n$.

À partir de ces approximations, on trouve donc :

$$\hat{\mu} = \overline{x} \text{ et } \hat{\phi} = \frac{\sum_{t=1}^{n-1}(x_t - \overline{x})(x_{t-1} - \overline{x})}{\sum_{t=1}^{n-1}(x_t - \overline{x})^2}.$$

En faisant l'approximation :

$$\sum_{t=1}^{n-1}(x_t - \overline{x})^2 \approx \sum_{t=1}^{n}(x_t - \overline{x})^2,$$

on obtient :

$$\hat{\phi} = \frac{\sum_{t=1}^{n-1}(x_t - \bar{x})(x_{t-1} - \bar{x})}{\sum_{t=1}^{n}(x_t - \bar{x})^2} = r_1,$$

où r_1 correspond à l'autocorrélation empirique d'ordre 1. Ceci suit la logique précédemment établie puisque pour un modèle $AR(1)$, l'autocorrélation théorique est $\rho(1) = \phi$.

2.4.2.1 Cas particulier du processus $AR(2)$

En appliquant des approximations similaires, on trouve que les estimateurs obtenus par la méthode des moindres carrés du modèle $AR(2)$ sont :

$$\hat{\mu} \approx \bar{x},$$
$$\hat{\phi}_1 \approx r_1(1 - r_2)/(1 - r_1^2),$$
$$\hat{\phi}_2 \approx (r_2 - r_1^2)/(1 - r_1^2),$$

où r_1 et r_2 désignent respectivement les autocorrélations empiriques d'ordres 1 et 2.

2.5 Prévisions

Rappel

2.5.1 Espérance conditionnelle

Soient X et Y deux variables aléatoires. L'espérance conditionnelle de Y étant donné X est l'espérance de Y par rapport à la densité conditionnelle de Y étant donné X et cette relation s'écrit :

$$E\{Y|X=x\} = \int_{-\infty}^{\infty} yf(y|x)dy.$$

Cela entraîne que lorsque $Y = X_{n+k}$ et $X = (X_1,...,X_n)$ on a :

$$\hat{X}_n(k) = \int_{-\infty}^{\infty} yf_{n+k}(y|x_1,...,x_n)dy,$$

où $\int_{-\infty}^{\infty} yf_{n+k}(y|x_1,...,x_n)$ représente la densité conditionnelle de X_{n+k} étant donné $X_1 = x_1,...,X_n = x_n$.

2.5.2 Propriétés de l'espérance conditionnelle

1. $E\{E\{Y|X=x\}\} = E\{Y\}$

2. $E\{a_1Y_1 + a_2Y_2|X\} = a_1E\{Y_1|X\} + a_2E\{Y_2|X\}$

3. $E\{g(X)Y|X\} = g(X)E\{Y|X\}$

4. Si X et Y sont indépendantes alors $E\{Y|X\} = E\{Y\}$

5. Pour toute fonction réelle g on a $E\left\{\left[Y - E\{Y|X\}\right]^2\right\} \leq E\left\{\left[Y - g(X)\right]^2\right\}$.

De la propriété 5, on tire la conclusion que l'erreur quadratique moyenne minimale est atteinte pour $g(X)=E\{Y|X\}$. Cela justifie le choix de $\hat{X}_n(k)=E\{X_{n+k}|X_1,...,X_n\}$ comme prévision de X_{n+k}.

2.5.3 Calcul des prévisions

2.5.3.1 Modèles autorégressifs

Cas particulier $AR(1)$

Soit $\{X_t\}$ un processus $AR(1)$ qui s'écrit :

$$X_t-\mu=\phi(X_{t-1}-\mu)+a_t.$$

En inversant le processus, on remarque :

$$X_t=\sum_{i=0}^{\infty}\pi_i a_{t-i}.$$

Cela met en évidence le fait que X_t dépend uniquement des valeurs de $a_t, a_{t-1}, a_{t-2},...$

Cela signifie que a_j est indépendante de X_i pour $j>i$.

Calculons maintenant la prévision $\hat{X}_n(k)$ de délai k associée à ce processus. Pour ce faire, nous calculons d'abord $\hat{X}_n(1)$ de la manière suivante :

$$\begin{aligned}\hat{X}_n(1)&=E\{X_{n+1}|X_1,...,X_n\}\\&=E\{\mu+\phi_1(X_n-\mu)+a_{n+1}|X_1,...,X_n\}\\&=\mu+\phi E\{(X_n-\mu)|X_1,...,X_n\}+E\{a_{n+1}|X_1,...,X_n\}\\&=\mu+\phi_1(X_n-\mu).\end{aligned}$$

D'où la prévision de $\hat{X}_n(1)$ est donnée par $\mu+\phi_1(X_n-\mu)$.

En se basant sur le même raisonnement, la prévision de $\hat{X}_n(k)$ est :

$$\begin{aligned}\hat{X}_n(k) &= E\left\{X_{n+k} \middle| X_1,...,X_n\right\} \\ &= E\left\{\mu + \phi_1(X_{n+k-1} - \mu) + a_{n+k} \middle| X_1,...,X_n\right\} \\ &= \mu + \phi E\left\{(X_{n+k-1} - \mu) \middle| X_1,...,X_n\right\} + E\left\{a_{n+k} \middle| X_1,...,X_n\right\} \\ &= \mu + \phi_1(\hat{X}_n(k-1) - \mu).\end{aligned}$$

D'où la prévision de $\hat{X}_n(k)$ est donnée par :

$$\mu + \phi_1(\hat{X}_n(k-1) - \mu).$$

Par itérations successives, on obtient le résultat :

$$\hat{X}_n(k) = \mu + \phi_1^k(X_n - \mu).$$

Cas général $AR(p)$

Soit $\{X_t\}$ un processus $AR(p)$ qui s'écrit :

$$X_t - \mu = \phi_1(X_{t-1} - \mu) + \cdots + \phi_p(X_{t-p} - \mu) + a_t.$$

En se basant sur les mêmes arguments que pour le cas particulier $AR(1)$ on obtient :

$$X_n(1) = \mu + \phi_1(X_n - \mu) + \cdots + \phi_p(X_{n-p} - \mu),$$

ainsi que

$$\hat{X}_n(k) = \mu + \phi_1(\hat{X}_n(k-1) - \mu) + \cdots + \phi_p(\hat{X}_n(k-p) - \mu).$$

À partir de ces deux équations, nous sommes en mesure de calculer toutes les prévisions en tenant compte du fait que $\hat{X}_n(0) = X_n$.

2.5.3.2 Modèles de moyennes mobiles

Cas particulier $MA(1)$

Soit $\{X_t\}$ un processus $MA(1)$ qui s'écrit :

$$X_t = \mu + a_t - \theta_1 a_{t-1}.$$

Il s'ensuit que :

$$\begin{aligned}\hat{X}_n(1) &= E\{X_{n+1} | X_1, ..., X_n\} \\ &= E\{\mu + a_{n+1} - \theta_1 a_n | X_1, ..., X_n\} \\ &= \mu - \theta_1 a_n.\end{aligned}$$

Par contre, $\hat{X}_n(k) = \mu$ pour $k \geq 2$.

Cas général $MA(q)$

Soit $\{X_t\}$ un processus $MA(q)$ qui s'écrit :

$$X_t = \mu + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}.$$

Il en découle que les prévisions se calculent comme suit :

$$\hat{X}_n(k) = \mu - \sum_{i=k}^{q} \theta_i a_{n+k-i}, \quad k \leq q$$

et

$$\hat{X}_n(k) = \mu, \quad k > q.$$

2.5.3.3 Erreur de prévision

L'erreur de prévision se caractérise par la différence entre la valeur réelle de la série et la valeur prédite.

On peut représenter ce concept par l'équation suivante :

$$e_n(k) = X_{n+k} - \hat{X}_n(k).$$

L'intervalle de prévision de X_{n+k} au niveau de 95% s'exprime comme suit :

$$\left[\hat{X}_n(k) - 2\sqrt{\text{var}(e_n(k))}, \hat{X}_n(k) + 2\sqrt{\text{var}(e_n(k))}\right].$$

Il s'agit donc maintenant de calculer $\text{var}(e_n(k))$ selon le contexte.

Exemple de calcul de $\text{var}(e_n(k))$ **pour le processus** $MA(1)$

Nous savons que :

$$\begin{aligned} e_n(1) &= X_{n+1} - \hat{X}_n(1) \\ &= \mu + a_{n+1} - \theta_1 a_n - \mu + \theta_1 a_n \\ &= a_{n+1}. \end{aligned}$$

Cela a pour conséquence que $\text{var}(e_n(1)) = \sigma^2$.

Pour $k \geq 2$ on trouve :

$$\begin{aligned} e_n(k) &= X_{n+k} - \hat{X}_n(k) \\ &= \mu + a_{n+k} - \theta_1 a_{n+k-1} - \mu \\ &= a_{n+k} - \theta_1 a_{n+k-1}. \end{aligned}$$

Finalement, on constate que $\text{var}(e_n(k)) = \sigma^2(1 + \theta_1^2)$.

Les intervalles de prévision de X_{n+1} et X_{n+k} au niveau de 95% sont respectivement :

$$\left[\hat{X}_n(1)-2\sigma, \hat{X}_n(1)+2\sigma\right]$$

et

$$\left[\hat{X}_n(k)-2\sigma\sqrt{1+\theta^2}, \hat{X}_n(k)+2\sigma\sqrt{1+\theta^2}\right].$$

Conclusion

Cette section clôt le second chapitre qui avait pour but de présenter la théorie qui sous-tend la méthodologie Box-Jenkins. L'approche se résume en quelques étapes à appliquer qui faisaient l'objet de la première section. Pour bien mener à terme l'analyse il faut vérifier quelques hypothèses qui sont traitées dans la deuxième section. En troisième lieu, les modèles classiques sont présentés. Les deux dernières sections abordent l'estimation des paramètres et les prévisions rattachées aux différents modèles. Le prochain chapitre traitera de la démarche d'analyse et des prévisions des séries chronologiques du programme SIPPE pour lesquelles l'approche Box-Jenkins a été principalement utilisée.

CHAPITRE 3

APPLICATION DE LA MÉTHODOLOGIE AUX SÉRIES DU PROGRAMME SIPPE

Ce chapitre présente les analyses de séries chronologiques effectuées dans le cadre du projet de collaboration avec le Ministère de la Santé et des Services Sociaux. Les séries chronologiques du projet sont celles du programme SIPPE et un certain nombre d'entre elles ont été modélisées à l'aide de la méthodologie Box-Jenkins. La première section introduit le projet de collaboration en précisant les spécificités des données. La deuxième section est constituée d'exemples de séries chronologiques du programme. La stationnarité et les transformations nécessaires pour rendre une série stationnaire y sont détaillées. La troisième section porte sur l'identification des modèles et propose plusieurs exemples souvent rencontrés lors des analyses. À la quatrième section, on s'attarde à la validation de l'adéquation du modèle et aux prévisions qui y sont rattachées. La cinquième section traite des prévisions et des équations des modèles retenus. La dernière section traite des méthodes alternatives qui ont été utilisées pour les cas où l'approche Box-Jenkins ne pouvait s'appliquer.

3.1 Présentation du projet de collaboration (extrait)

3.1.1 Exploration des données

Les analyses qui font l'objet de ce chapitre ont été réalisées lors d'un stage au sein des bureaux du Ministère de la Santé et des Services Sociaux durant l'année 2009 et 2010. Le projet de collaboration entre l'Université du Québec à Trois-Rivières et le Ministère a été supervisée par Mhamed Mesfioui et Latifa Elfassihi. Le Ministère de la Santé et des Services Sociaux du Québec voulait faire la prévision d'une variable pour les cinq prochaines années pour l'un de ses programmes : le SIPPE. Les services intégrés en périnatalité et pour la petite enfance (SIPPE) visent à soutenir les familles vivant en contexte de vulnérabilité. La population ciblée par les SIPPE se divise en deux grands sous-groupes de la population. Le premier groupe est constitué des mères de moins de 20

ans et le second des mères de 20 ans et plus avec moins d'onze ans de scolarité. Les séries de données sont réparties par réseau local de services (RLS). Les données historiques disponibles sont les nombres de mères ayant accouché entre 1981 et 2009 en années financières et associés au premier et au second groupe. Les deux types de groupe sont respectivement les séries 1 et séries 2. On remarque que les premières séries sont généralement composées de plus petits nombres que les séries associées au deuxième groupe pour la même période de référence.

Lors de l'exploration des données, une analyse multivariée a été effectuée en combinant les deux séries. Cette analyse nous a montré que les distributions des deux séries sont différentes. C'est pour cette raison que nous avons donc décidé de traiter les séries des deux groupes séparément et pour chacun des 95 RLS de la province. Dans ce cas, nous avons élaboré des modèles univariés. Donc, les prévisions ne dépendent que d'une variable explicative, ici le temps, et s'expliquent à l'aide des valeurs des années antérieures de la série [5] (p.11).

3.1.2 Source et traitement des données

Les données proviennent des fichiers fermés et provisoires de naissances vivantes et mortinaissances du Registre des événements démographiques du Québec. La population visée par le programme SIPPE est celle résidant dans les municipalités du Québec conventionnées. À cet effet, les mères qui résident dans les municipalités non conventionnées ont été exclues. En raison de perturbations dans le fichier de données, il a fallu procéder à l'estimation de données manquantes pour l'année 2006. Cette estimation a été réalisée à l'aide de calculs de proportions à l'aide des données de l'année précédente et de l'année suivante. Toutes les analyses ont été effectuées à l'aide du logiciel : SAS 9.2.

3.2 Étude de la stationnarité de quelques séries du projet

Soit $\{X_t\}$ une série chronologique du programme SIPPE où x_t représente le nombre de mères de 19 ans et moins en fonction de t qui représente le nombre d'années écoulées depuis 1982. Cette section présente plusieurs séries qui ont été analysées dans le projet. Pour chaque série, on retrouve les informations permettant d'étudier la stationnarité de la série. Il s'agit des valeurs de la série, du graphique des valeurs en fonction du temps, du corrélogramme de la fonction d'autocorrélation et des transformations appliquées pour rendre la série stationnaire dans le cas ou elle ne l'est pas. Il s'agit de reprendre les notions portant sur la stationnarité du chapitre précédent et de les mettre en application.

L'allure du graphique des valeurs de la série en fonction du temps peuvent nous donner un bon aperçu mais il est toujours préférable de poser un jugement après avoir examiné le corrélogramme d'autocorrélation. Le corrélogramme de la fonction d'autocorrélation d'une série stationnaire montre une atténuation rapide des valeurs de r_k au fil des décalages.

Pour l'étude de la stationnarité nous avons donc procédé à l'observation du graphique des valeurs de la série en fonction du temps et noté s'il était nécessaire d'appliquer des transformations afin de rendre la série stationnaire. Les observations relatives à chacune des séries étaient consignées dans un journal d'observations. Nous pouvons donc résumer l'étude de la stationnarité en trois points : l'allure du graphique de la série brute, l'amélioration du modèle par l'application de la transformation logarithmique et nécessité d'appliquer une ou plusieurs différenciations.

La façon de vérifier si la transformation logarithmique améliore le modèle est de comparer les valeurs du coefficient de détermination (r^2) associées aux modèles de régression linéaire de la série initiale et de la série transformée. Une amélioration considérable était une augmentation du r^2 de plus de 10%. Pour obtenir la série transformée, il s'agit d'appliquer l'opération du logarithme à chacune des valeurs de la série initiale. L'effet de la différenciation quant à lui s'observe en étudiant le comportement de la fonction d'autocorrélation sur le corrélogramme qui lui est associé.

Nous détaillerons la démarche précédemment expliquée dans l'étude de la stationnarité de quelques séries du projet dans la section qui suit.

3.2.1 Exemple 1

Tableau 4 -Valeurs de la série de l'exemple 1

Nb. de mères de moins de 20 ans en fonction de l'année	
1982	30
1983	21
1984	23
1985	11
1986	15
1987	18
1988	9
1989	10
1990	14
1991	12
1992	19
1993	13
1994	21
1995	16
1996	14
1997	12
1998	12
1999	10
2000	13
2001	9
2002	10
2003	8
2004	3
2005	6
2006	6
2007	7
2008	6
2009	9

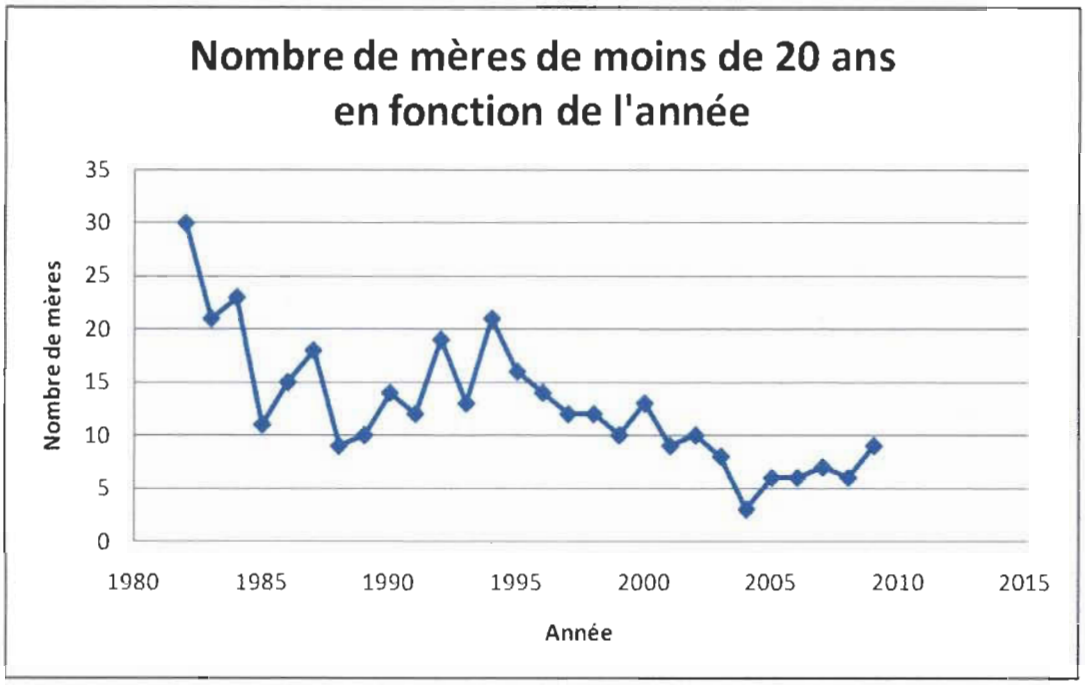


Figure 3 – Graphique des données brutes en fonction du temps de la série de l'exemple 1

L'allure du graphique nous amène à penser que la série n'est pas stationnaire car les propriétés ne semblent pas constantes dans le temps. Les hypothèses pourraient être qu'un modèle logarithmique serait adéquat pour modéliser les valeurs de la série et que la moyenne semble changer au fil du temps.

Le Système SAS 15:46 Tuesday, March 15, 2011

The ARIMA Procedure

Name of Variable = Y

Mean of Working Series	12.75
Standard Deviation	5.919852
Number of Observations	28

Autocorrelations

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	35.044643	1.00000	|********************|	0
1	18.462054	0.52682	|********** . |	0.188982
2	17.058036	0.48675	|********** . |	0.235665
3	9.305804	0.26554	|***** . |	0.269187
4	7.205357	0.20561	|**** . |	0.278385
5	5.319196	0.15178	|*** . |	0.283756
6	-1.031250	-.02943	| . * |	0.286641
7	-0.917411	-.02618	| . * |	0.286749

"." marks two standard errors

Figure 4 - Corrélogramme de la fonction d'autocorrélation de la série initiale de l'exemple 1

On remarque que le corrélogramme de la fonction d'autocorrélation de la série initiale semble indiquer que la série n'est pas stationnaire car les valeurs des r_k s'atténuent plutôt lentement dans le temps.

3.2.1.1 Étude des transformations dans le cas de l'exemple 1

Transformation logarithmique

Dans le cas de l'exemple 1, la valeur du r^2 du modèle de régression linéaire pour la série initiale est de 0,5572 alors que celle de la série transformée est de 0,5634. L'augmentation du r^2 du second modèle est seulement de 0,62%, ce qui est inférieur à 10%. La transformation n'améliore donc pas le modèle et celle-ci n'est pas retenue. Il est intéressant de noter que la transformation logarithmique améliore la plupart du temps le modèle des séries portant sur le sous-groupe cible des mères de 20 ans et plus avec moins de 11 ans de scolarité (type de série : 2).

Tableau 5 – Valeurs de la série transformée log de l'exemple 2

Valeurs de la série transformée log	
1982	2,340444115
1983	2,212187604
1984	2,243038049
1985	2,075546961
1986	2,033423755
1987	1,986771734
1988	1,892094603
1989	1,832508913
1990	1,913813852
1991	1,86332286
1992	1,819543936
1993	1,875061263
1994	1,748188027
1995	1,69019608
1996	1,62324929
1997	1,880813592
1998	1,662757832
1999	1,544068044
2000	1,477121255
2001	1,633468456
2002	1,447158031
2003	1,462397998
2004	1,633468456
2005	1,556302501
2006	1,643452676
2007	1,113943352
2008	1,414973348
2009	1,568201724

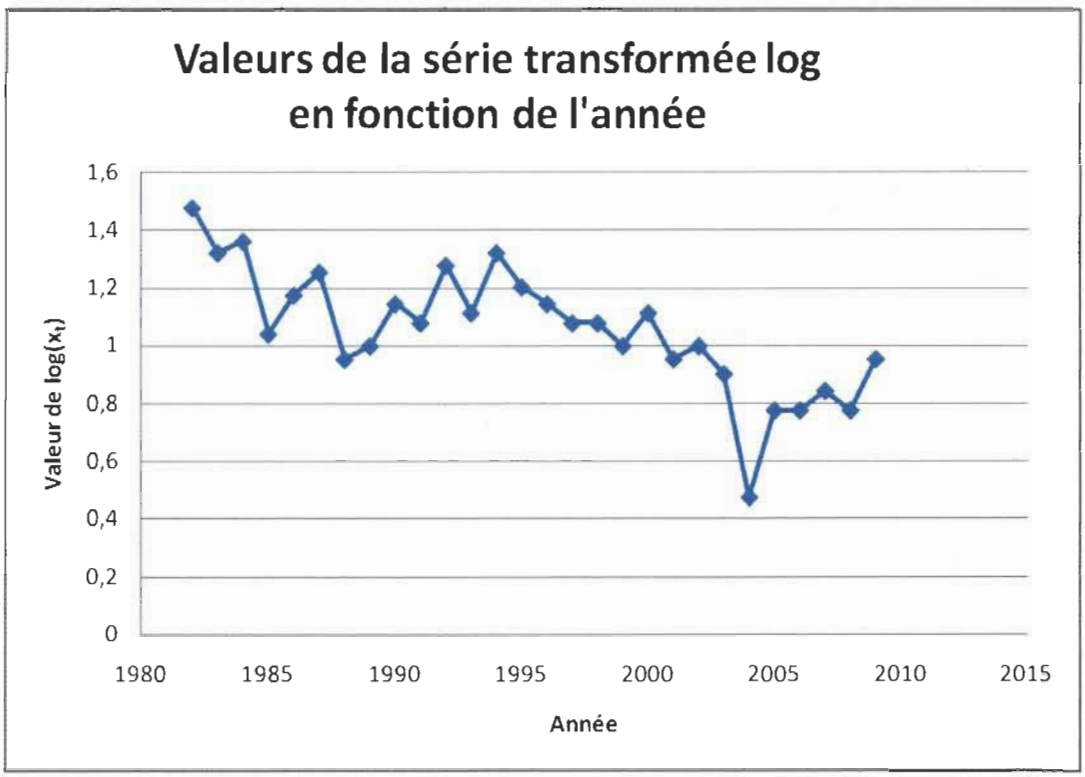
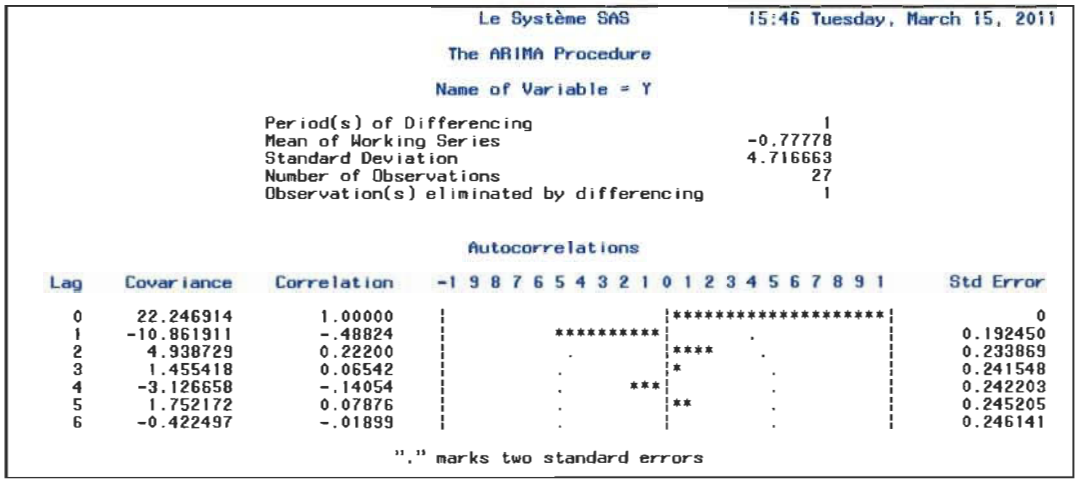


Figure 5 - Graphique de la série transformée log de l'exemple 1

Différenciation

Le corrélogramme de la fonction d'autocorrélation associé à la série différenciée d'ordre 1 sert à voir l'effet de l'application de la différenciation d'ordre 1.



Le Système SAS 15:46 Tuesday, March 15, 2011

The ARIMA Procedure

Name of Variable = Y

Period(s) of Differencing	1
Mean of Working Series	-0.77778
Standard Deviation	4.716663
Number of Observations	27
Observation(s) eliminated by differencing	1

Autocorrelations

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	22.246914	1.00000	||********************|	0
1	-10.861911	-.48824	|**********| . |	0.192450
2	4.938729	0.22200	| . |**** . |	0.233869
3	1.455418	0.06542	| . |* . |	0.241548
4	-3.126658	-.14054	| . ***| . |	0.242203
5	1.752172	0.07876	| . |** . |	0.245205
6	-0.422497	-.01899	| . | . |	0.246141

"." marks two standard errors

Figure 6 – Corrélogramme de la fonction d'aucorrélation de la différenciée d'ordre 1 de l'exemple 1

Dans le corrélogramme de la série différenciée, on constate une atténuation rapide à partir du décalage 1. À partir de ces faits, il est raisonnable de penser que la série différenciée d'ordre 1 est stationnaire.

3.2.2 Exemple 2

Tableau 6 – Valeurs de la série de l'exemple 2

Nb. de mères de 20 ans et plus avec moins de 11 ans de scolarité en fonction de l'année	
1982	219
1983	163
1984	175
1985	119
1986	108
1987	97
1988	78
1989	68
1990	82
1991	73
1992	66
1993	75
1994	56
1995	49
1996	42
1997	76
1998	46
1999	35
2000	30
2001	43
2002	28
2003	29
2004	43
2005	36
2006	44
2007	13
2008	26
2009	37

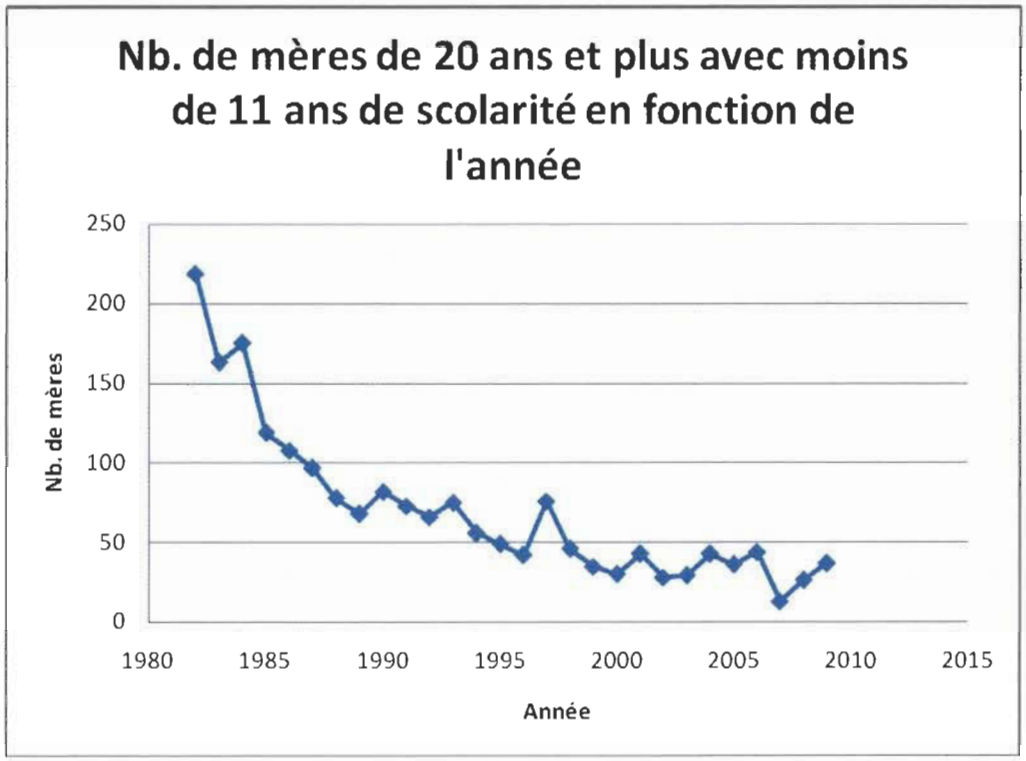


Figure 7 – Graphique des données brutes en fonction du temps de la série de l'exemple 2

Dans le cas présent, l'allure du graphique nous amène à penser que non seulement la série n'est pas stationnaire mais qu'un modèle logarithmique serait indiqué pour la modélisation de la série.

Le Système SAS 14:39 Wednesday, March 23, 2011

The ARIMA Procedure

Name of Variable = Y

Mean of Working Series	69.85714
Standard Deviation	48.02433
Number of Observations	28

Autocorrelations

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	2306.337	1.00000	|**********|	0
1	1673.030	0.72541	|**********. |	0.188982
2	1405.779	0.60953	|********** |	0.270742
3	939.534	0.40737	|********. |	0.315972
4	767.701	0.33287	|*******. |	0.334204
5	557.787	0.24185	|*****. |	0.345842
6	413.179	0.17915	|****. |	0.351830
7	334.740	0.14514	|***. |	0.355073

"." marks two standard errors

Figure 8 - Corrélogramme de la fonction d'autocorrélation de la série initiale de l'exemple 2

On remarque que le corrélogramme de la fonction d'autocorrélation de la série initiale de l'exemple 2 semble indiquer que la série n'est pas stationnaire pour la même raison que l'exemple 1. Les valeurs des r_k s'atténuent encore plus lentement dans le temps.

3.2.2.1 Étude des transformations dans le cas de l'exemple 2

Transformation logarithmique

Pour le cas de l'exemple 2, la valeur du r^2 du modèle de régression linéaire pour la série initiale est de 0,6905 alors que celle de la série transformée est de 0,8168. L'augmentation du r^2 du deuxième modèle est de 12,63%, ce qui est supérieur à 10%. Nous pouvons donc conclure que la transformation améliore le modèle.

Tableau 7 – Valeurs de la série transformée log de l'exemple 2

Valeurs de la série transformée log	
1982	2,340444115
1983	2,212187604
1984	2,243038049
1985	2,075546961
1986	2,033423755
1987	1,986771734
1988	1,892094603
1989	1,832508913
1990	1,913813852
1991	1,86332286
1992	1,819543936
1993	1,875061263
1994	1,748188027
1995	1,69019608
1996	1,62324929
1997	1,880813592
1998	1,662757832
1999	1,544068044
2000	1,477121255
2001	1,633468456
2002	1,447158031
2003	1,462397998
2004	1,633468456
2005	1,556302501
2006	1,643452676
2007	1,113943352
2008	1,414973348
2009	1,568201724

Puisque nous retenons la transformation logarithmique, faisons sortir de nouveau le diagramme de la fonction d'autocorrélation associé à la série transformée log.

Le Système SAS 14:39 Wednesday, March 23, 2011

The ARIMA Procedure

Name of Variable = Y

Mean of Working Series	1.756697
Standard Deviation	0.273545
Number of Observations	28

Autocorrelations

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	0.074827	1.00000	|********************|	0
1	0.055087	0.73619	|***************. |	0.188982
2	0.046052	0.61545	|************. |	0.272813
3	0.034978	0.46746	|*********. |	0.318563
4	0.033271	0.44464	|*********. |	0.342185
5	0.027118	0.36241	|*******. |	0.362232
6	0.020188	0.26979	|*****. |	0.374958
7	0.018968	0.25349	|*****. |	0.381828

"." marks two standard errors

Figure 9 – Corrélogramme de la fonction d'autocorrélation de la série transformée de l'exemple 2

Le diagramme de la fonction d'autocorrélation de la série transformée ne semble pas indiquer que la série est stationnaire. Il faut appliquer une différenciation et étudier le corrélogramme associé.

Le Système SAS 14:39 Wednesday, March 23, 2011

The ARIMA Procedure

Name of Variable = Y

Period(s) of Differencing	1
Mean of Working Series	-0.0286
Standard Deviation	0.161826
Number of Observations	27
Observation(s) eliminated by differencing	1

Autocorrelations

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	0.026188	1.00000	|********************|	0
1	-0.010274	-.39231	|********. |	0.192450
2	-0.0019104	-.07295	|. *. |	0.220085
3	-0.0029437	-.11241	|. **. |	0.220979
4	0.0042852	0.16363	|. |***. |	0.223087
5	0.0018279	0.06980	|. |*. |	0.227489
6	-0.0057056	-.21787	|. ****. |	0.228281

"." marks two standard errors

Figure 10 - – Corrélogramme de la fonction d'autocorrélation de la série transformée et différenciée de l'exemple 2

On remarque que les valeurs de r_k s'atténuent rapidement après le premier décalage. Il est important d'appliquer les transformations dans cet ordre s'il y a lieu car on ne peut pas faire une différenciation avant une transformation logarithmique. Cela repose sur le fait que la série différenciée peut comporter des valeurs négatives et le logarithme d'un nombre négatif n'est pas défini.

3.2.3 Exemple 3

Tableau 8 – Valeurs de la série de l'exemple 3

Nb. de mères de moins de 20 ans en fonction de l'année	
1982	17
1983	24
1984	17
1985	14
1986	14
1987	13
1988	9
1989	9
1990	11
1991	21
1992	15
1993	15
1994	17
1995	12
1996	20
1997	10
1998	8
1999	9
2000	6
2001	8
2002	15
2003	12
2004	9
2005	5
2006	8
2007	1
2008	5
2009	9

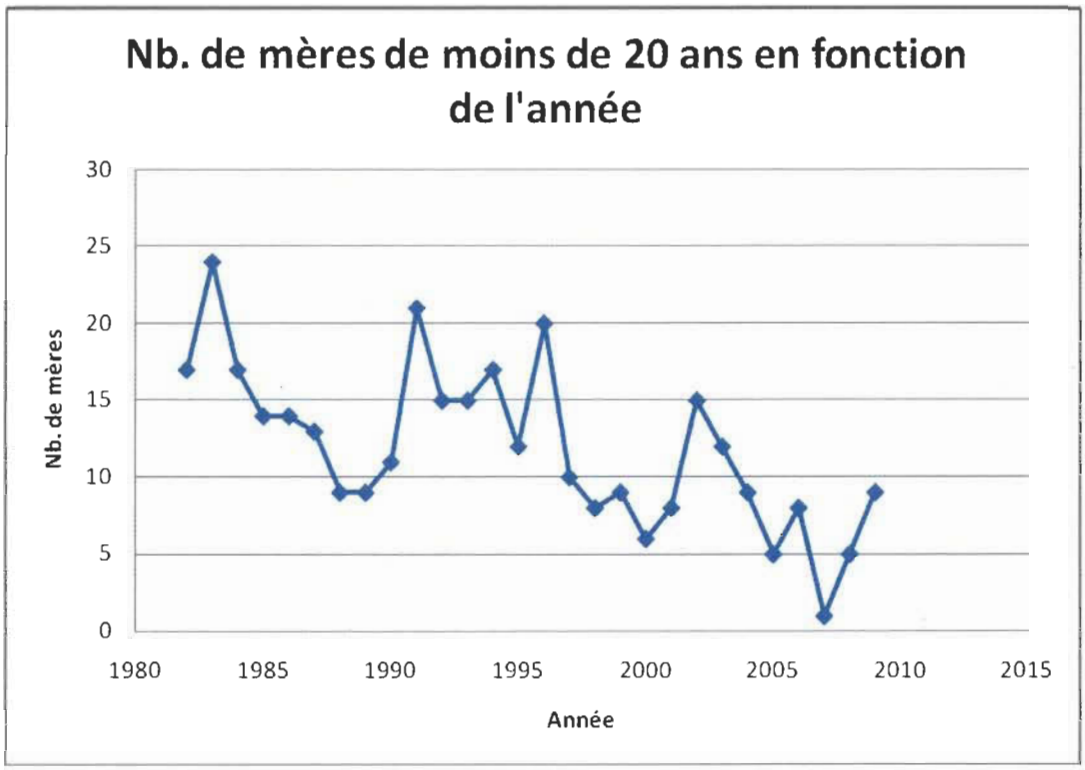


Figure 11 - Graphique des données brutes en fonction du temps de la série de l'exemple 3

Le Système SAS 14:39 Wednesday, March 23, 2011

The ARIMA Procedure

Name of Variable = Y

Mean of Working Series	11.89286
Standard Deviation	5.184725
Number of Observations	28

Autocorrelations

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	26.881378	1.00000	|********************|	0
1	14.694743	0.54665	| . |********** |	0.188982
2	8.532343	0.31741	| . |****** . |	0.238871
3	4.483464	0.16679	| . |*** . |	0.253486
4	0.240707	0.00895	| . | . |	0.257376
5	-0.240570	-.00895	| . | . |	0.257387
6	-0.590470	-.02197	| . | . |	0.257398
7	2.206314	0.08208	| . |** . |	0.257465

"." marks two standard errors

Figure 12 – Diagramme de la fonction d'autocorrélation de la série brute de l'exemple 3

Le diagramme de la fonction d'autocorrélation montre que les valeurs de r_k s'atténuent rapidement à partir du premier décalage. En effet, il suffit de regarder les valeurs de la fonction d'autocorrélation pour les quatre premiers décalages : $r_1 = 0.54665, r_2 = 0.31741, \ r_3 = 0.16679$ et $r_4 = 0.00895$. Puisque la série initiale est stationnaire, aucune transformation n'est requise et la série peut être directement modélisée avec la méthodologie Box-Jenkins.

Pour les cas où l'étude de la stationnarité n'était pas évident, les candidats potentiels de modèles étaient étudiés pour les deux possibilités (avec et sans différenciation) et le meilleur modèle était retenu. Nous traitons les étapes de ce procédé dans les sections suivantes : les étapes d'identification et de validation.

3.3 Identification des modèles

Une fois la stationnarité assurée, il est possible d'entreprendre la première étape de la méthodologie Box-Jenkins : l'identification du modèle approprié. Dans un premier temps, nous verrons un exemple détaillé de sélection d'un modèle. Ensuite nous verrons plusieurs exemples de sélection de modèles ARIMA différents. Finalement, un résumé de l'ensemble des modèles ARIMA retenus pour les séries du projet SIPPE ainsi qu'une discussion portant sur l'étape d'identification seront présentés.

3.3.1 Rappel sur l'identification des candidats à partir des corrélogrammes

Tout comme il est mentionné dans le deuxième chapitre, il est possible de proposer des candidats de modèles ARIMA en effectuant l'étude des diagrammes de la fonction d'autocorrélation et de la fonction d'autocorrélation partielle. D'une part, les candidats potentiels pour un modèle autorégressif se manifestent par une atténuation lente de la courbe du corrélogramme de la fonction d'autocorrélation partielle et rapide à partir d'un certain décalage p de la courbe du corrélogramme d'autocorrélation. D'autre part, les candidats potentiels pour un modèle de moyenne mobile se détectent par une atténuation lente de la courbe du corrélogramme de la fonction d'autocorrélation et par un pic à un certain décalage q dans le diagramme d'autocorrélation partielle. Finalement, les candidats de modèles mixtes peuvent se repérer par la présence de pics à la fois à partir d'un décalage p dans le corrélogramme de la fonction d'autocorrélation et d'un décalage q dans le corrélogramme de la fonction d'autocorrélation partielle. Généralement, on retient les trois candidats les plus plausibles et on les analyse ensuite selon certains critères durant les phases de l'estimation des paramètres et de la validation de l'adéquation du modèle. Dans la phase d'identification du modèle, nous nous contenterons de présenter trois exemples de sélection de candidats et les caractéristiques des modèles retenus servant à la comparaison des cas seront détaillées dans les sections suivantes : l'estimation des paramètres et la vérification de l'adéquation du modèle.

3.3.2 Sélection de candidats pour quatre séries du programme SIPPE

3.3.2.1 Exemple 4

L'étude de la stationnarité a été effectuée et il a été décidé que la série était stationnaire dès le départ et c'est pour cette raison que nous utilisons les données de la série brute.

Tableau 9 – Valeurs de la série brute de l'exemple 4

Nb. de mères de moins de 20 ans en fonction de l'année	
1982	18
1983	13
1984	13
1985	8
1986	6
1987	9
1988	7
1989	5
1990	7
1991	5
1992	5
1993	13
1994	9
1995	8
1996	6
1997	9
1998	6
1999	6
2000	6
2001	4
2002	1
2003	5
2004	5
2005	5
2006	8
2007	4
2008	3
2009	4

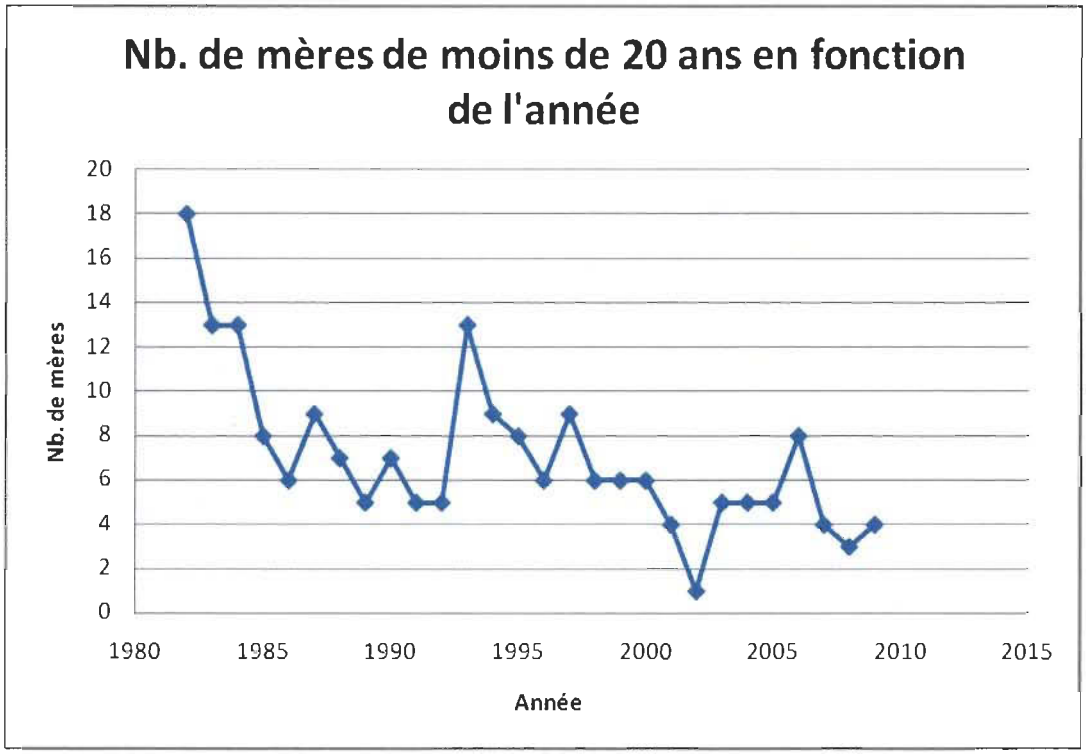
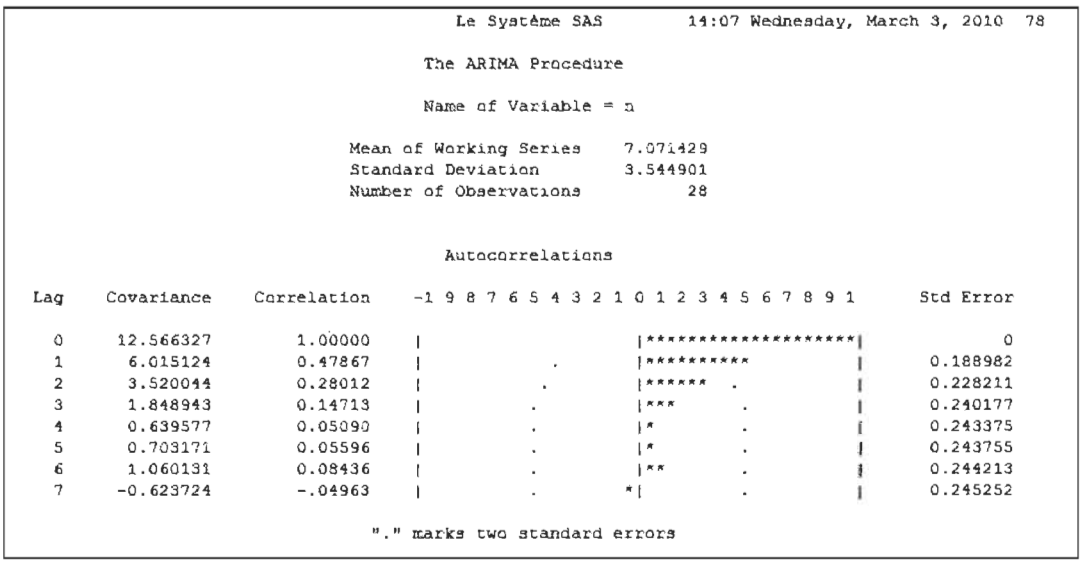


Figure 13 – Graphique des données brutes en fonction du temps de la série de l'exemple 4

3.3.2.1.1 Corrélogrammes des fonctions d'autocorrélation

Fonction d'autocorrélation



Le Système SAS 14:07 Wednesday, March 3, 2010 78

The ARIMA Procedure

Name of Variable = n

Mean of Working Series	7.071429
Standard Deviation	3.544901
Number of Observations	28

Autocorrelations

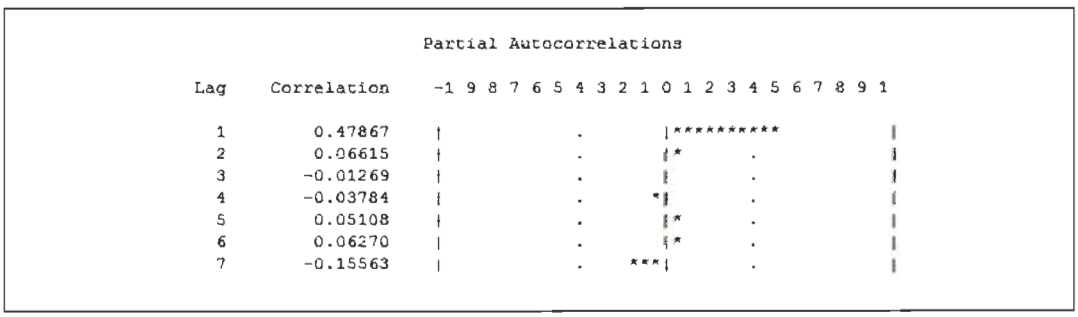
Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	12.566327	1.00000	|********************|	0
1	6.015124	0.47867	|********** . |	0.188982
2	3.520044	0.28012	|****** . |	0.228211
3	1.848943	0.14713	|*** . |	0.240177
4	0.639577	0.05090	|* . |	0.243375
5	0.703171	0.05596	|* . |	0.243755
6	1.060131	0.08436	|** . |	0.244213
7	-0.623724	-.04963	*| . |	0.245252

"." marks two standard errors

Figure 14 – Corrélogramme de la fonction d'autocorrélation associé à l'exemple 4

On remarque un petit pic au décalage 1. On peut donc soupçonner qu'un candidat autorégressif d'ordre 1 serait adéquat.

Fonction d'autocorrélation partielle



Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	0.47867	|********** |
2	0.06615	|* . |
3	-0.01269	| . |
4	-0.03784	*| . |
5	0.05108	|* . |
6	0.06270	|* . |
7	-0.15563	***| . |

Figure 15 - Corrélogramme de la fonction d'autocorrélation partielle associé à l'exemple 4

On remarque aussi un petit pic au décalage 1. On peut donc soupçonner qu'un candidat de moyenne mobile d'ordre 1 serait approprié.

Dans le cas présent, les trois candidats identifiés sont : un modèle autorégressif d'ordre $p=1$, un modèle de moyenne mobile d'ordre $q=1$ et un modèle mixte d'ordre

$(p,q)=(1,1)$. Ces modèles peuvent s'écrire respectivement comme étant des modèles ARIMA d'ordre : $(p,d,q)=(1,0,0)$, $(p,d,q)=(0,0,1)$ et $(p,d,q)=(1,0,1)$ où d représente l'ordre de différenciation ici égal à 0 car la série n'a pas été différenciée. Dans la prochaine section, on comparera les trois modèles en regardant les facteurs suivants : le fait que les paramètres du modèle soient significatifs, l'estimation de l'erreur type (Standard Error Estimate), la valeur de l'indicateur RMRES2 et les probabilités associées au calcul de la statistique de Ljung-Box (*Autocorrelation of Check residuals*).

3.3.2.2 Exemple 5

L'étude de la stationnarité a été effectuée et il a été décidé que la série était stationnaire après avoir effectué une différenciation d'ordre 1.

Tableau 10 - Valeurs de la série différenciée de l'exemple 5

Valeurs de la série différenciée d'ordre 1 de l'exemple 5	
1	1
2	-5
3	1
4	6
5	-9
6	-1
7	-1
8	2
9	1
10	1
11	-1
12	-3
13	5
14	-8
15	7
16	-2
17	2
18	-3
19	3
20	-4
21	-2
22	-3
23	4
24	-3
25	0
26	1
27	2

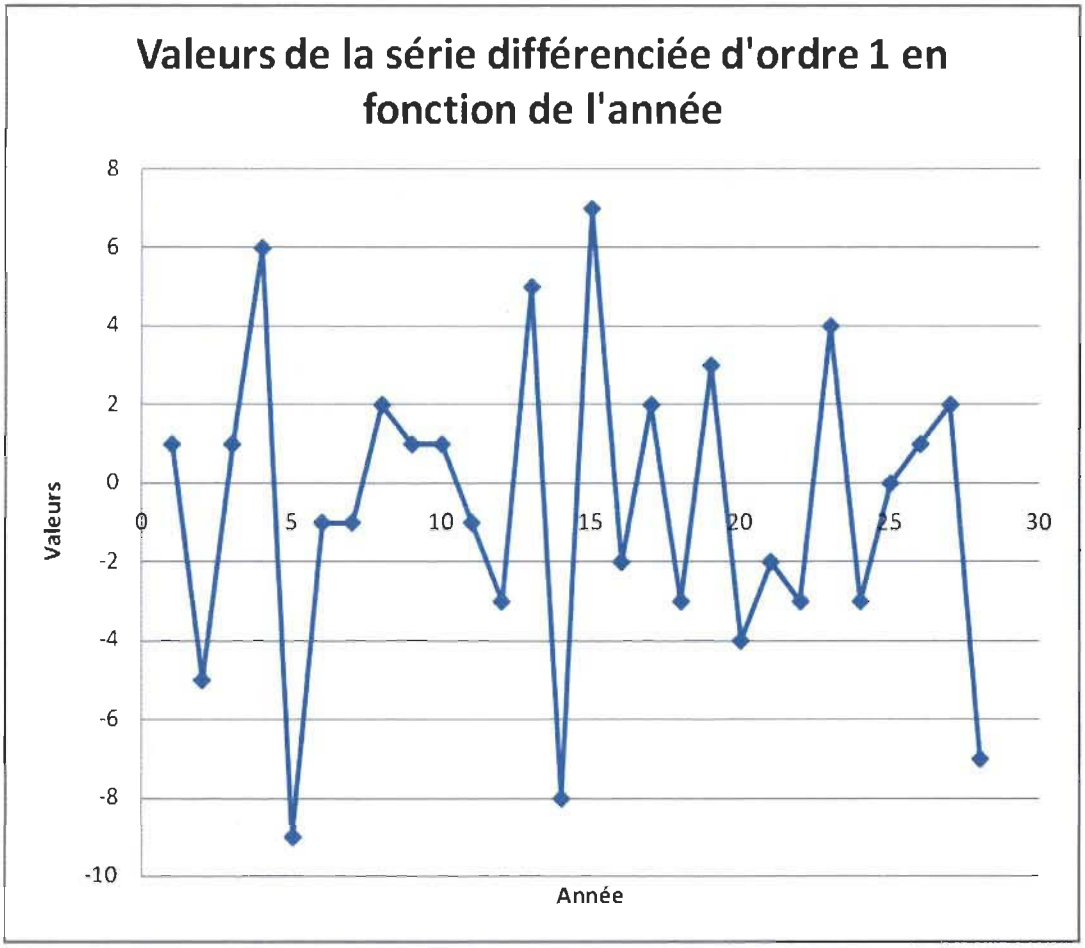


Figure 16 - Graphique de la série différenciée de l'exemple 5

3.3.2.2.1 Corrélogrammes des fonctions d'autocorrélation

Fonction d'autocorrélation

Le Système SAS 13:37 Wednesday, March 3, 2010 79

The ARIMA Procedure

Name of Variable = n

Period(s) of Differencing	1
Mean of Working Series	-0.33333
Standard Deviation	3.761402
Number of Observations	27
Observation(s) eliminated by differencing	1

Autocorrelations

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	14.148148	1.00000	|********************|	0
1	-7.893004	-.55788	*********** . |	0.192450
2	2.559671	0.18092	. |**** . |	0.245135
3	-2.530864	-.17888	. **** | . |	0.250032
4	2.983539	0.21088	. |**** . |	0.254728
5	-3.292181	-.23269	. ***** | . |	0.261113
6	1.679012	0.11867	. |** . |	0.268684

"." marks two standard errors

Figure 17 - Corrélogramme de la fonction d'autocorrélation associé à l'exemple 5

On remarque un petit pic au décalage 1. On peut encore une fois suspecter qu'un candidat autorégressif d'ordre 1 serait adéquat.

Fonction d'autocorrélation partielle

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	-0.55788	*********** . |
2	-0.18920	. **** | . |
3	-0.24756	. ***** | . |
4	0.01549	. | . |
5	-0.15166	. *** | . |
6	-0.13795	. *** | . |

Figure 18 - Corrélogramme de la fonction d'autocorrélation partielle associé à l'exemple 5

On remarque la présence d'un petit pic au décalage 1. On retient donc aussi un candidat de moyenne mobile d'ordre 1.

Dans le cas de l'exemple 5, les trois candidats retenus sont une fois de plus : un modèle autorégressif d'ordre $p=1$, un modèle de moyenne mobile d'ordre $q=1$ et un modèle mixte d'ordre $(p,q)=(1,1)$. Ces modèles peuvent s'écrire respectivement comme étant des modèles ARIMA d'ordre : $(p,d,q)=(1,1,0)$, $(p,d,q)=(0,1,1)$ et $(p,d,q)=(1,1,1)$ où d représente l'ordre de différenciation ici égal à 1 car la série a dû être différenciée pour s'assurer la stationnarité. On comparera encore une fois les trois modèles choisis en regardant les facteurs suivants : le fait que les paramètres du modèle soient significatifs, l'estimation de l'erreur type (Standard Error Estimate) et la valeur de l'indicateur RMRES2 et les probabilités associées au calcul de la statistique de Ljung-Box (*Autocorrelation of Check residuals*).

3.3.2.3 Exemple 6

L'étude de la stationnarité a été effectuée et il a été décidé que la série était stationnaire dès le départ.

Tableau 11 - Valeurs de la série brute de l'exemple 6

Nb. de mères de moins de 20 ans en fonction de l'année	
1982	27
1983	24
1984	15
1985	16
1986	25
1987	18
1988	17
1989	19
1990	21
1991	26
1992	23
1993	23
1994	9
1995	11
1996	17
1997	11
1998	17
1999	11
2000	16
2001	13
2002	8
2003	14
2004	5
2005	11
2006	15
2007	9
2008	4
2009	17

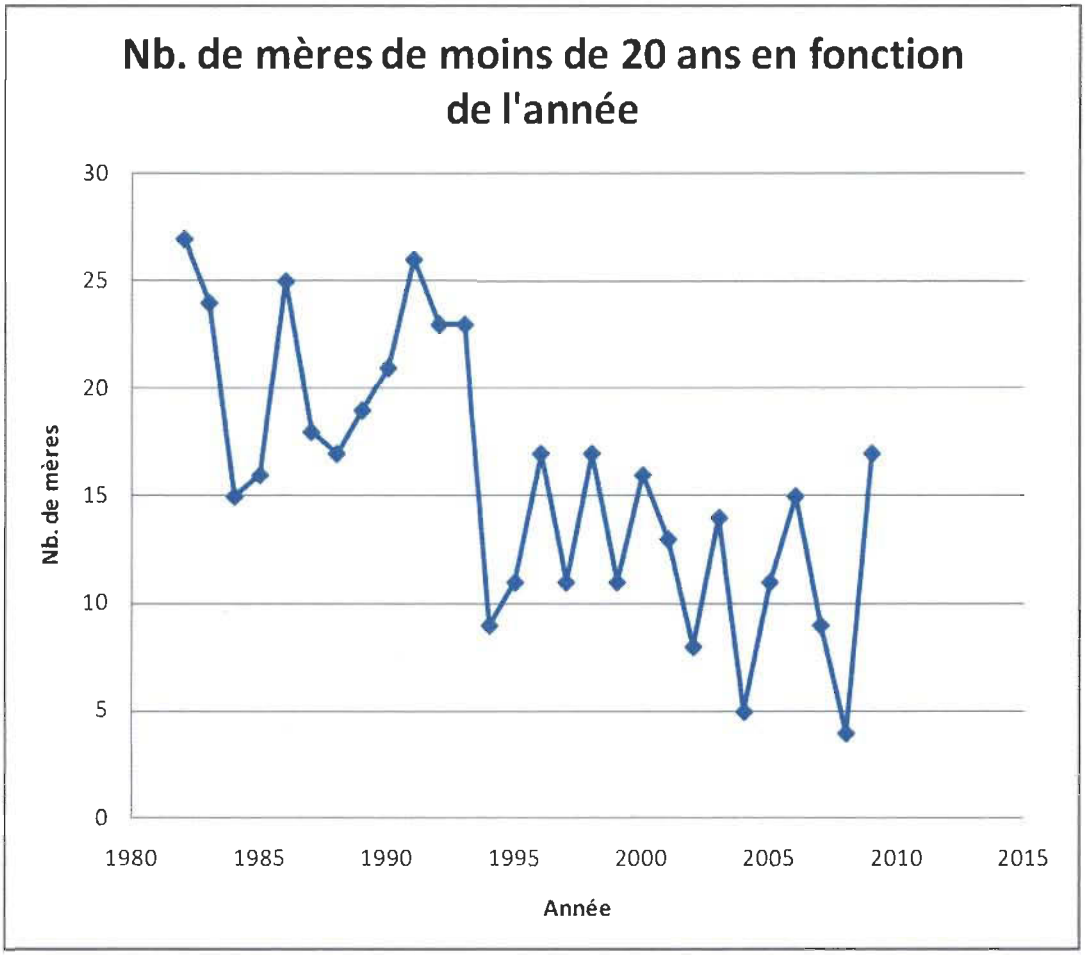
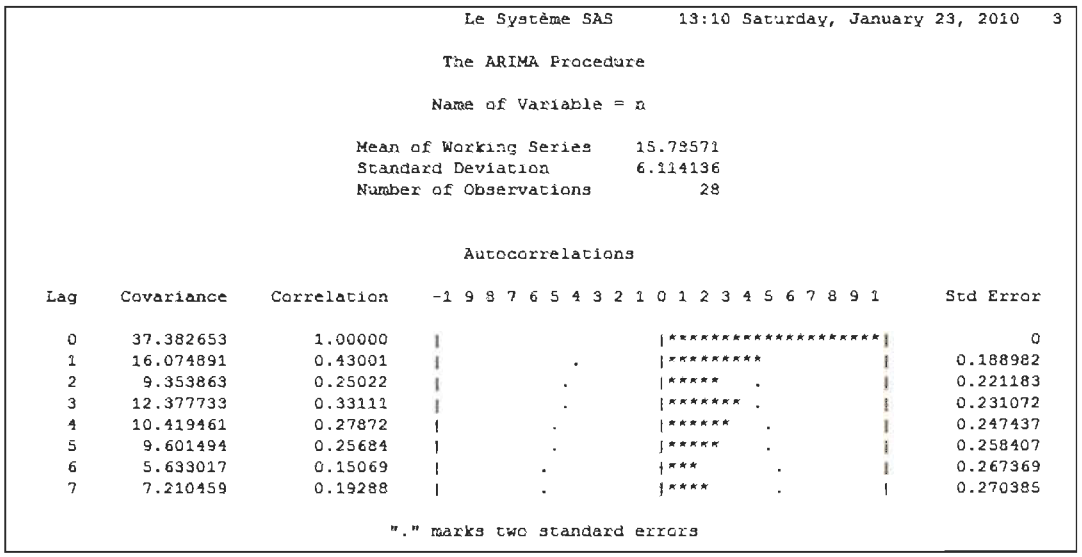


Figure 19 - Graphique des données brutes en fonction du temps de la série de l'exemple 6

3.3.2.3.1 Corrélogrammes des fonctions d'autocorrélation

Fonction d'autocorrélation

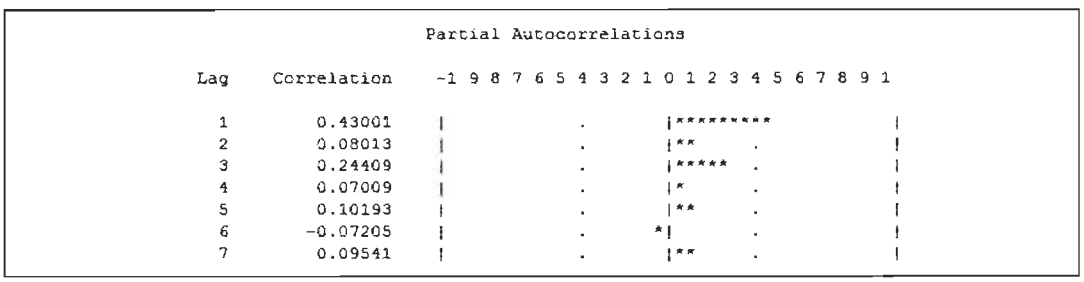


```
                                  Le Système SAS      13:10 Saturday, January 23, 2010   3

                                The ARIMA Procedure

                               Name of Variable = n

                          Mean of Working Series    15.79571
                          Standard Deviation        6.114136
                          Number of Observations          28


                                 Autocorrelations

Lag    Covariance     Correlation    -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1      Std Error

  0     37.382653         1.00000    |                    |********************|             0
  1     16.074891         0.43001    |                .   |*********           |      0.188982
  2      9.353863         0.25022    |                .   |*****    .          |      0.221183
  3     12.377733         0.33111    |                .   |*******  .          |      0.231072
  4     10.419461         0.27872    |                .   |******   .          |      0.247437
  5      9.601494         0.25684    |                .   |*****    .          |      0.258407
  6      5.633017         0.15069    |                .   |***      .          |      0.267369
  7      7.210459         0.19288    |                .   |****     .          |      0.270385

                              "." marks two standard errors
```

Figure 20 - Corrélogramme de la fonction d'autocorrélation associé à l'exemple 6

Nous remarquons que les valeurs des r_k s'atténuent rapidement après r_1. Cela indique que la série est stationnaire.

Fonction d'autocorrélation partielle



```
                              Partial Autocorrelations

                Lag    Correlation    -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1

                  1        0.43001    |                .   |*********           |
                  2        0.08013    |                .   |**       .          |
                  3        0.24409    |                .   |*****    .          |
                  4        0.07009    |                .   |*        .          |
                  5        0.10193    |                .   |**       .          |
                  6       -0.07205    |                .  *|         .          |
                  7        0.09541    |                .   |**       .          |
```

Figure 21 - Corrélogramme de la fonction d'autocorrélation partielle associé à l'exemple 6

Dans le cas de l'exemple 6, les trois candidats retenus sont une fois de plus : un modèle autorégressif d'ordre $p=1$, un modèle de moyenne mobile d'ordre $q=1$ et un modèle mixte d'ordre $(p,q)=(1,1)$. Ces choix s'expliquent par les pics en r_1 retrouvés dans les deux diagrammes des fonctions d'autocorrélation. Les modèles s'écrivent respectivement comme étant des modèles ARIMA d'ordre : $(p,d,q)=(1,0,0)$, $(p,d,q)=(0,0,1)$ et $(p,d,q)=(1,0,1)$ où d représente l'ordre de différenciation ici égal à 0 car la série était stationnaire dès le début. Dans la section suivante on procédera à la sélection du meilleur modèle parmi les trois candidats.

3.3.2.4 Exemple 7

L'étude de la stationnarité a été effectuée et il a été décidé que la série était stationnaire après avoir été transformée en extrayant le logarithme de chacune des valeurs tout en ayant effectué une différenciation d'ordre 1.

Tableau 12 - Valeurs de la série transformée log de la série de l'exemple 7

Valeurs de la série transformée log en fonction de l'année	
1982	2,772588722
1983	2,833213344
1984	2,48490665
1985	2,564949357
1986	2,944438979
1987	2,302585093
1988	2,197224577
1989	2,079441542
1990	2,302585093
1991	2,397895273
1992	2,48490665
1993	2,397895273
1994	2,079441542
1995	2,564949357
1996	1,609437912
1997	2,48490665
1998	2,302585093
1999	2,48490665
2000	2,197224577
2001	2,48490665
2002	2,079441542
2003	1,791759469
2004	1,098612289
2005	1,945910149
2006	1,386294361
2007	1,386294361
2008	1,609437912
2009	1,945910149

3.3.2.4.1 Comparaison des corrélogrammes pour différentes transformations

Fonction d'autocorrélation de la série initiale

Le Système SAS 12:41 Sunday, April 10, 2011 132

The ARIMA Procedure

Name of Variable = n

Mean of Working Series	46.67857
Standard Deviation	30.97586
Number of Observations	28

Autocorrelations

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	959.504	1.00000	|********************|	0
1	724.656	0.75524	|*************** . |	0.188982
2	603.759	0.62924	|************* . |	0.276507
3	481.993	0.50234	|********** . |	0.323632
4	364.531	0.37992	|******** . |	0.350375
5	286.327	0.29841	|****** . |	0.364790
6	197.386	0.20572	|**** . |	0.373407
7	93.389987	0.09733	|** . |	0.377433

"." marks two standard errors

Figure 22 – Fonction d'autocorrélation de la série initiale de l'exemple 7

Fonction d'autocorrélation de la série transformée log

Le Système SAS 12:41 Sunday, April 10, 2011 130

The ARIMA Procedure

Name of Variable = y

Mean of Working Series	3.650366
Standard Deviation	0.610253
Number of Observations	28

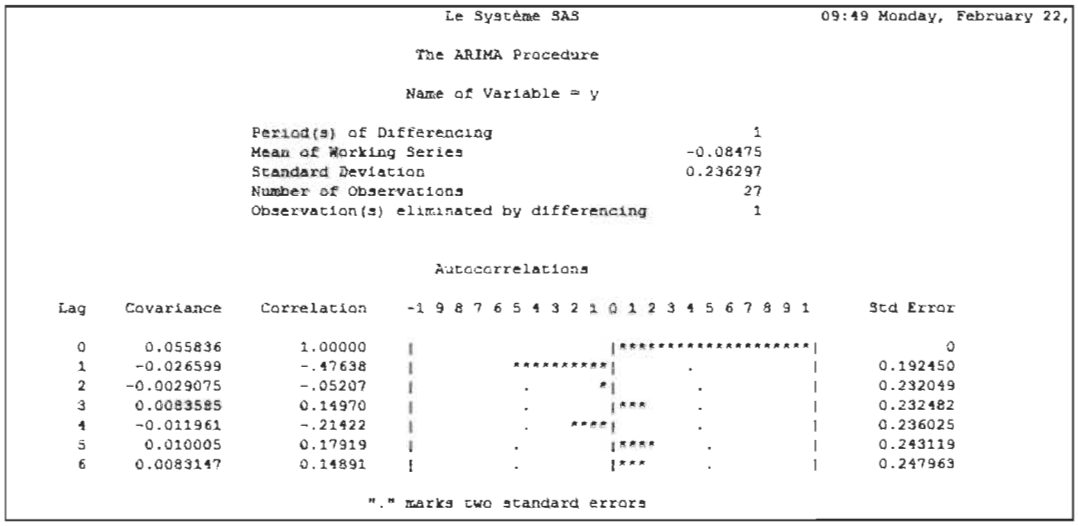
Autocorrelations

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	0.372409	1.00000	|********************|	0
1	0.294646	0.79119	|**************** . |	0.188982
2	0.250424	0.67244	|************* . |	0.283597
3	0.212010	0.56929	|*********** . |	0.335747
4	0.170271	0.45722	|********* . |	0.368613
5	0.142847	0.38358	|******** . |	0.388340
6	0.112918	0.30321	|****** . |	0.401643
7	0.068369	0.18359	|**** . |	0.409736

"." marks two standard errors

Figure 23 – Fonction d'autocorrélation de la série transformée log de l'Exemple 7

Fonction d'autocorrélation de la série transformée log et différenciée



Le Système SAS 09:49 Monday, February 22,

The ARIMA Procedure

Name of Variable = y

Period(s) of Differencing	1
Mean of Working Series	-0.08475
Standard Deviation	0.236297
Number of Observations	27
Observation(s) eliminated by differencing	1

Autocorrelations

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	0.055836	1.00000	|********************|	0
1	-0.026599	-.47638	*********| . |	0.192450
2	-0.0029075	-.05207	. *| . |	0.232049
3	0.0083585	0.14970	. |*** . |	0.232482
4	-0.011961	-.21422	. ****| . |	0.236025
5	0.010005	0.17919	. |**** . |	0.243119
6	0.0083147	0.14891	. |*** . |	0.247963

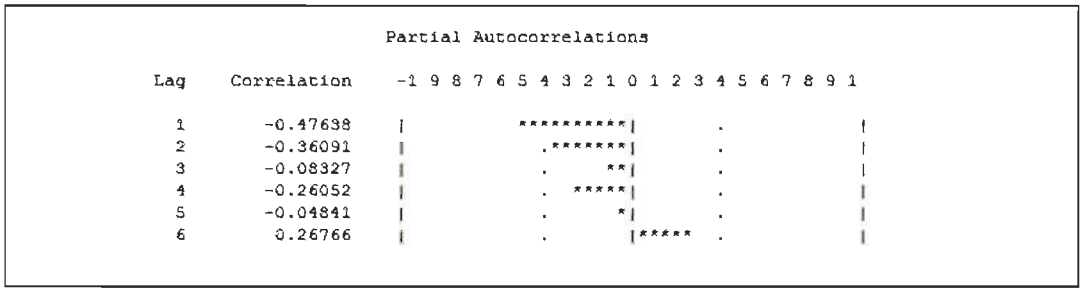
"." marks two standard errors

Figure 24- Fonction d'autocorrélation de la série transformée (log et différenciation) de l'exemple 7

Nous remarquons que la série ne semble pas stationnaire dès le départ à l'aide de l'étude du corrélogramme de la fonction d'autocorrélation de la série initiale. On assiste à une atténuation plutôt lente des valeurs de r_k. Ensuite, après avoir effectué la comparaison des valeurs du r^2 associées à la régression linéaire simple de la série initiale et de la série transformée log il a été décidé que la transformation logarithmique améliorait le modèle de façon significative. En effet, les valeurs de r^2 sont respectivement : 0.7818 et 0.9168.

En ce qui concerne les candidats potentiels exprimés dans le diagramme de la fonction d'autocorrélation de la série transformée log et différenciée, on retient un modèle autorégressif d'ordre 1 en raison du pic en r_1. Afin d'établir les autres candidats potentiels il faut s'en remettre à l'étude du diagramme de la fonction d'autocorrélation partielle.

Fonction d'autocorrélation partielle de la série transformée (log et d=1)



Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	-0.47638	**********| . |
2	-0.36091	.*******| . |
3	-0.08327	. **| . |
4	-0.26052	. *****| . |
5	-0.04841	. *| . |
6	0.26766	. |***** . |

Figure 25 - Fonction d'autocorrélation partielle de la série transformée log et différenciée de l'exemple 7

Une fois de plus, on émettra l'hypothèse qu'un modèle de moyenne mobile d'ordre 1 serait plausible en se basant sur la présence d'un pic en r_1.

Les candidats potentiels de l'exemple 7 peuvent se résumer par les modèles suivants : ARIMA d'ordre : $(p,d,q)=(1,1,0)$, $(p,d,q)=(0,1,1)$ et $(p,d,q)=(1,1,1)$.

3.4 Validation de l'adéquation du modèle retenu

Une fois les principaux candidats identifiés, un examen de certaines conditions s'applique afin de s'assurer que le modèle retenu est adéquat et que le meilleur modèle parmi les candidats proposés a été sélectionné. Ces conditions se résument à la vérification de la signification des paramètres du modèle, la comparaison des probabilités de la section *Autocorrelation of Check residuals*, des valeurs de l'estimation de l'erreur type et de l'indicateur RMRES2.

3.4.1 Test de Ljung-Box

Il existe plusieurs tests qui permettent de vérifier l'adéquation du modèle. Parmi les plus connus et utilisés, citons notamment les statistiques de Ljung-Box et Box-Pierce. Pour l'analyse des séries du projet nous avons retenu la vérification des probabilités associées à la statistique Ljung-Box car il est entendu que cette dernière est l'une des méthodes de vérification les plus efficaces [5] (p.496). Ces deux statistiques se rapportent aux fonctions d'autocorrélation associées aux résidus. En effet, les résidus ne devraient pas être fortement corrélés entre eux car le principe sous-jacent à l'application de la

méthodologie est que la modélisation ARIMA explique principalement la relation entre les valeurs de la série.

La statistique Ljung-Box [5] (p.497) se calcule de la manière suivante :

$$Q^* = n'(n'+2)\sum_{l=1}^{K}(n'-l)^{-1}r_l^2(\hat{a}).$$

Les probabilités de la section *Autocorrelation of Check residuals* sont en fait les probabilités d'erreur de type 1 (ou autrement appelé : erreur α qui représente l'erreur de rejeter l'hypothèse H_0 lorsqu'elle est vraie pour un α fixé, généralement 5%) sur la statistique Ljung-Box. On s'assure donc que les probabilités ne sont pas plus petites que 5%. On compare les probabilités de chacun des modèles et on retient celui pour lequel elles sont les plus élevées [5] (p.497).

Les probabilités associées à chacun des candidats de l'exemple 4 seront présentées et le choix du modèle final sera expliqué à la section 3.4.3.

3.4.2 Standard Error Estimate et indicateur RMRES2

La statistique t associée à chacune des variables indépendantes (dans notre cas elles représentent les paramètres des modèles) mesure l'importance additionnelle de la variable indépendante. Il peut arriver que la statistique t n'arrive pas à faire ressortir l'effet d'une variable indépendante prise individuellement. De plus, mentionnons l'effet possible de multicollinéarité qui existe lorsque les variables indépendantes sont reliées entre elles et donnent ainsi de l'information redondante. Pour ces deux raisons, en plus de vérifier les statistiques t, il est préférable de procéder à l'examen d'une mesure qui nous donne un aperçu global de ces interactions. À cette fin, on regarde la valeur de l'erreur type (*Standard Error Estimate*). Mentionnons aussi qu'à une erreur type plus petite est souvent associé un intervalle de prédiction plus court donc des erreurs de prévision moins grandes [5] (p.495). C'est pour avoir une idée sur la grandeur des erreurs de prévisions que nous avons créé l'indicateur RMRES2. Il s'agit d'abord de calculer les écarts entre les valeurs prédites et les valeurs réelles du modèle pour les années de 1982 à 2009 pour lesquelles les données sont disponibles. Ensuite, on calcule la moyenne des écarts

précédemment établis que l'on avait mis au carré. Finalement, il suffit d'extraire la racine carrée du résultat obtenu lors de la deuxième étape. Nous comparons les trois valeurs obtenues en tenant compte des autres facteurs mentionnés pour sélectionner le meilleur modèle.

L'équation de l'erreur type [5] (p.496) est :

$$s = \sqrt{\frac{SEE}{n - n_p}} = \sqrt{\frac{\sum_{t=1}^{n} (y_t - \hat{y}_t)^2}{n - n_p}}.$$

3.4.3 Modèles retenus pour les séries des exemples 4,5,6 et 7

Pour la série de l'exemple 4, toutes les étapes précédemment discutées seront détaillées. Pour les exemples 5,6 et 7 les résultats seront présentés et commentés. Les sorties SAS associées ont été mises en annexe[1] pour éviter d'alourdir le texte. Le principe est exactement le même que pour l'exemple 4 seulement les résultats diffèrent.

[1] Respectivement : annexes B, C et D

3.4.3.1 Candidats potentiels de l'exemple 4

3.3.2.3.1 Candidat 1 : ARIMA $(p,d,q)=(1,0,0)$

Conditional Least Squares Estimation

Paramètre	Estimation	Erreur standard	Valeur du test t	Pr. Approx. > |t|	Retard
MU	7.82041	1.24191	6.30	<.0001	0
AR1,1	0.53660	0.17250	3.11	0.0045	1

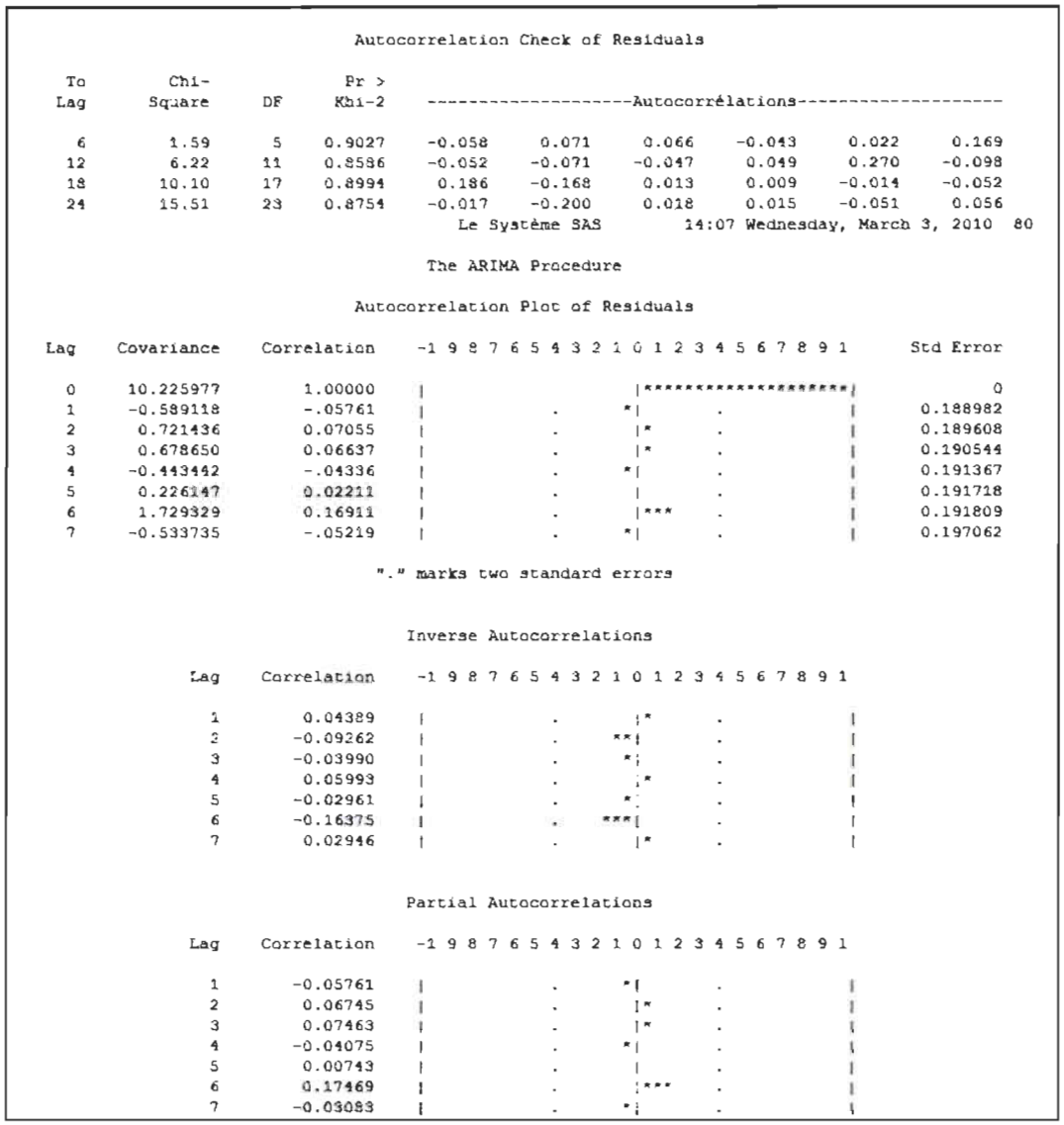
Constant Estimate	3.624012
Variance Estimate	10.22598
Std Error Estimate	3.197808
AIC	146.4836
SBC	149.148
Number of Residuals	28

* AIC and SBC do not include log determinant.

Figure 26 – Estimation des paramètres pour le candidat 1 de l'exemple 4

La décision d'inclure ou non la constante MU repose sur deux conditions : à savoir si la constante est significative et si elle améliore le modèle. La seconde condition est vérifiée par l'observation des probabilités associées à la statistique Ljung-Box.

Dans le cas du candidat 1 qui est un modèle ARIMA $(p,d,q)=(1,0,0)$, on a constaté que la constante MU était significative à 5% et qu'elle améliorait le modèle. En effet, elle faisait passer les probabilités de 0.8692, 0.7629, 0.9015 et 0.7592 à 0.9027, 0.8586, 0.8994 et 0.8754. Le paramètre AR1,1 (ϕ_1) est lui aussi significatif à 5%. Les estimations respectives pour les deux paramètres sont : 7.82041 et 0.53660. Dans la figure 22 à la page suivante, on retrouve aussi les diagrammes des fonctions d'autocorrélation reliées aux résidus et on constate qu'il n'y a pas de pic (à partir de r_1) ce qui est souhaitable car les résidus ne sont pas supposés être corrélés entre eux. En ce qui concerne l'erreur type et l'indicateur RMRES2 associés au candidat 1, les résultats sont respectivement : 3.197808 et 4.5309222566.



Autocorrelation Check of Residuals

To Lag	Chi-Square	DF	Pr > Khi-2	Autocorrélations					
6	1.59	5	0.9027	-0.058	0.071	0.066	-0.043	0.022	0.169
12	6.22	11	0.8586	-0.052	-0.071	-0.047	0.049	0.270	-0.098
18	10.10	17	0.8994	0.186	-0.168	0.013	0.009	-0.014	-0.052
24	15.51	23	0.8754	-0.017	-0.200	0.018	0.015	-0.051	0.056

Le Système SAS 14:07 Wednesday, March 3, 2010 80

The ARIMA Procedure

Autocorrelation Plot of Residuals

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	10.225977	1.00000	|********************|	0
1	-0.589118	-.05761	| . *| . |	0.188982
2	0.721436	0.07055	| . |* . |	0.189608
3	0.678650	0.06637	| . |* . |	0.190544
4	-0.443442	-.04336	| . *| . |	0.191367
5	0.226147	0.02211	| . | . |	0.191718
6	1.729329	0.16911	| . |*** . |	0.191809
7	-0.533735	-.05219	| . *| . |	0.197062

"." marks two standard errors

Inverse Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	0.04389	| . |* . |
2	-0.09262	| . **| . |
3	-0.03990	| . *| . |
4	0.05993	| . |* . |
5	-0.02961	| . *| . |
6	-0.16375	| . ***| . |
7	0.02946	| . |* . |

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	-0.05761	| . *| . |
2	0.06745	| . |* . |
3	0.07463	| . |* . |
4	-0.04075	| . *| . |
5	0.00743	| . | . |
6	0.17469	| . |*** . |
7	-0.03083	| . *| . |

Figure 27 – Probabilités (Check of Residuals) et diagrammes d'autocorrélation des résidus du candidat 1

3.3.2.3.2 Candidat 2 : ARIMA $(p,d,q)=(0,0,1)$

Conditional Least Squares Estimation

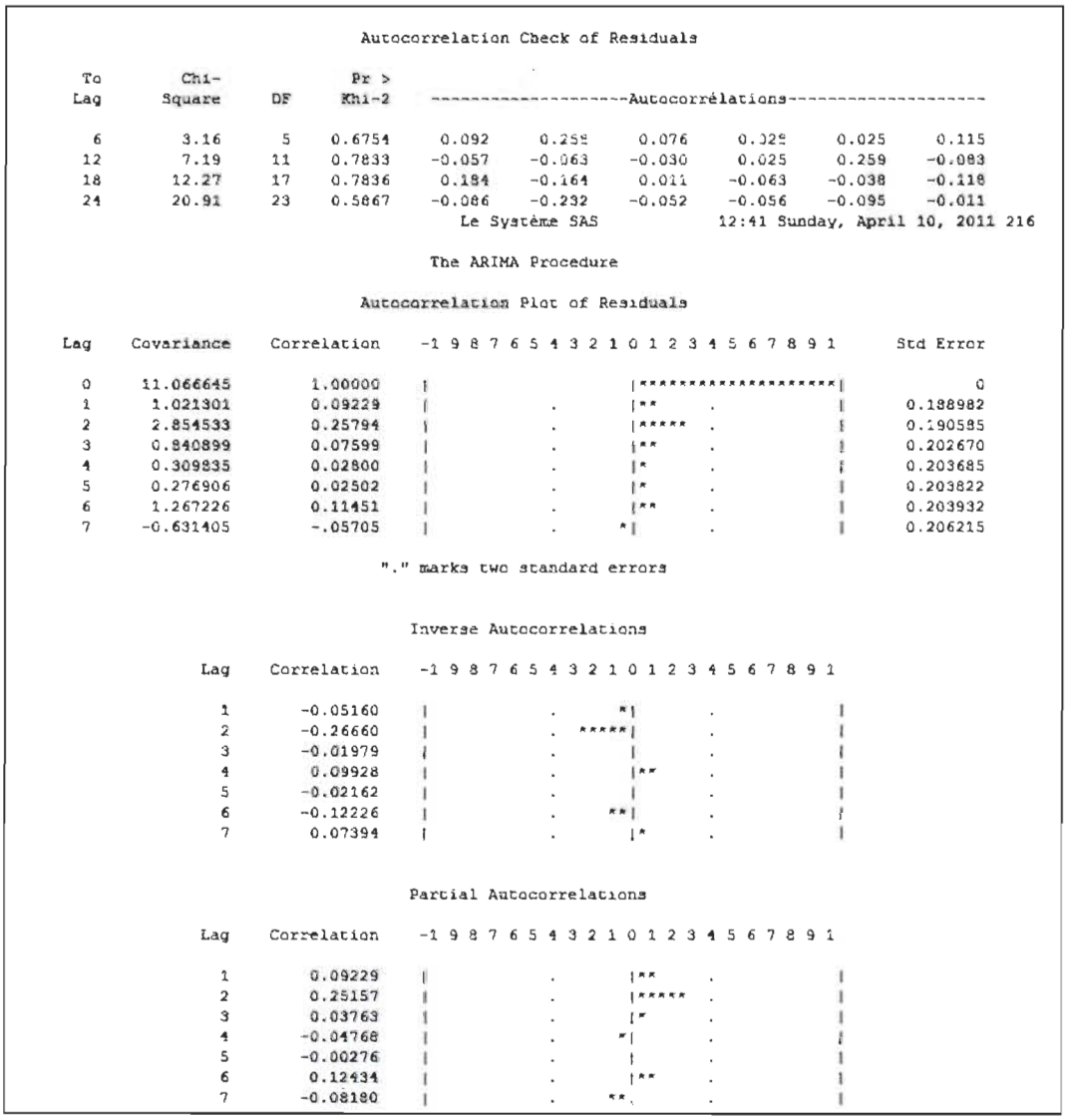
Paramètre	Valeur estimée	Erreur type	Valeur du test t	Approx. de Pr > |t|	Retard
MU	7.25737	0.86627	8.38	<.0001	0
MA1,1	-0.39483	0.18131	-2.18	0.0387	1

Constant Estimate	7.25737
Variance Estimate	11.06664
Std Error Estimate	3.326657
AIC	148.6957
SBC	151.3601
Number of Residuals	28

* AIC and SBC do not include log determinant.

Figure 28 – Estimation des paramètres pour le candidat 2 de l'exemple 4

Dans le cas du candidat 2 qui est un modèle ARIMA $(p,d,q)=(0,0,1)$, on a constaté que la constante MU était aussi significative à 5% et qu'elle améliorait le modèle. En effet, elle faisait passer les probabilités de 0.0001, 0.0001, 0.0001 et 0.0001 à 0.6754, 0.7833, 0.7836 et 0.5867. Le paramètre MA1,1 (θ_1) est lui aussi significatif à 5%. Les estimations respectives pour les deux paramètres sont : 7.25737 et -0.39483. Dans la figure 24 à la page suivante, on retrouve aussi les diagrammes des fonctions d'autocorrélation reliées aux résidus et on constate qu'il n'y a pas de pic (à partir de r_1) ce qui est souhaitable car les résidus ne sont pas supposés être corrélés entre eux. En ce qui concerne l'erreur type et l'indicateur RMRES2 associés au candidat 2, les résultats sont respectivement : 3.326657 et 4.7558067313.



Autocorrelation Check of Residuals

To Lag	Chi-Square	DF	Pr > Khi-2	Autocorrélations					
6	3.16	5	0.6754	0.092	0.258	0.076	0.029	0.025	0.115
12	7.19	11	0.7833	-0.057	-0.063	-0.030	0.025	0.259	-0.083
18	12.27	17	0.7836	0.184	-0.164	0.011	-0.063	-0.038	-0.118
24	20.91	23	0.5867	-0.086	-0.232	-0.052	-0.056	-0.095	-0.011

The ARIMA Procedure

Autocorrelation Plot of Residuals

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	11.066645	1.00000	|********************|	0
1	1.021301	0.09229	|** . |	0.188982
2	2.854533	0.25794	|***** . |	0.190585
3	0.840899	0.07599	|** . |	0.202670
4	0.309835	0.02800	|* . |	0.203685
5	0.276906	0.02502	|* . |	0.203822
6	1.267226	0.11451	|** . |	0.203932
7	-0.631405	-.05705	| . *|	0.206215

"." marks two standard errors

Inverse Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	-0.05160	| . *| . |
2	-0.26660	| . *****| . |
3	-0.01979	| . | . |
4	0.09928	| . |** . |
5	-0.02162	| . | . |
6	-0.12226	| . **| . |
7	0.07394	| . |* . |

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	0.09229	| . |** . |
2	0.25157	| . |***** . |
3	0.03763	| . |* . |
4	-0.04768	| . *| . |
5	-0.00276	| . | . |
6	0.12434	| . |** . |
7	-0.08180	| . *| . |

Figure 29 - Probabilités (Check of Residuals) et diagrammes d'autocorrélation des résidus du candidat 2

3.3.2.3.3 Candidat 3 : ARIMA $(p,d,q)=(1,0,1)$

Conditional Least Squares Estimation

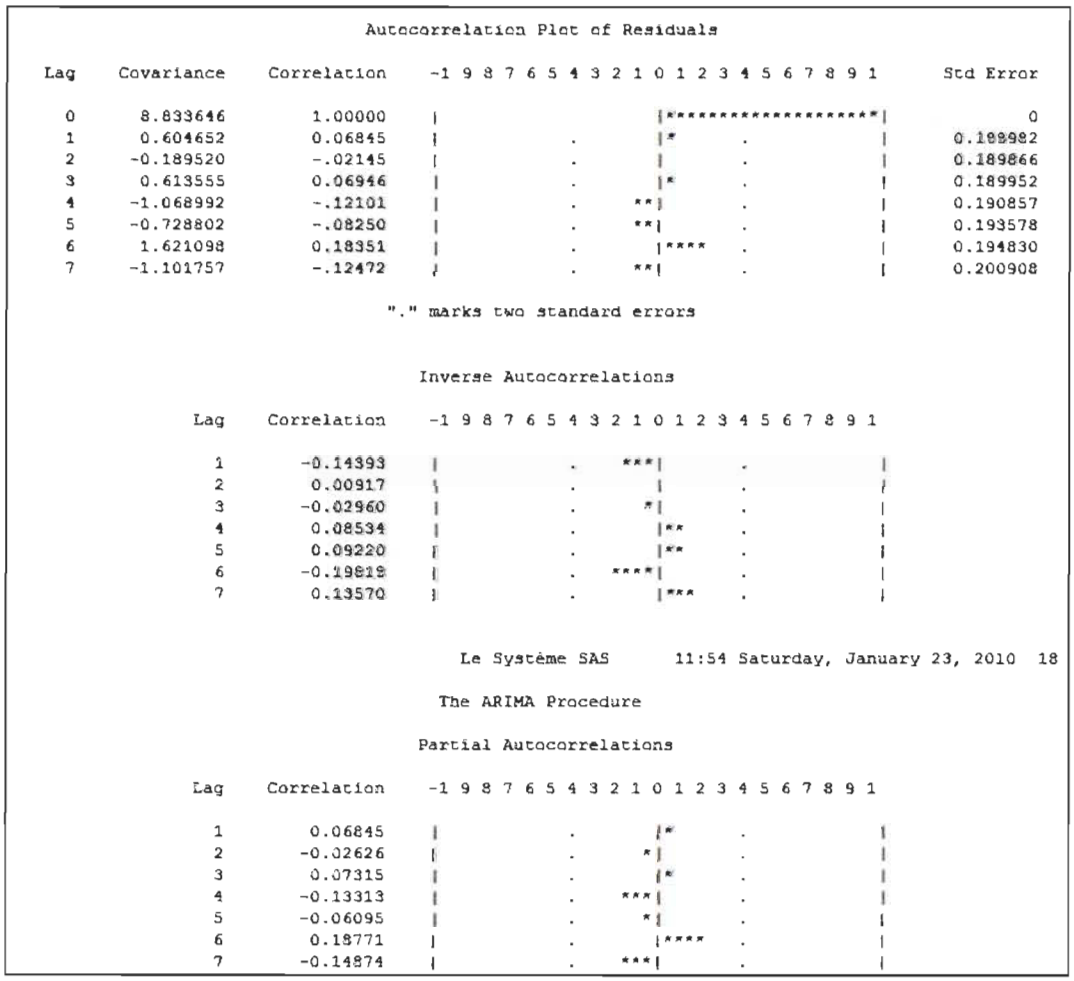
Paramètre	Valeur estimée	Erreur type	Valeur du test t	Approx. de Pr > |t|	Retard
MU	13.88339	2.63060	5.28	<.0001	0
MA1,1	0.46735	0.19026	2.46	0.0213	1
AR1,1	1.00000	0.04720	21.19	<.0001	1

Constant Estimate	1.355E-6
Variance Estimate	8.833646
Std Error Estimate	2.972145
AIC	143.2873
SBC	147.2839
Number of Residuals	28

* AIC and SBC do not include log determinant.

Figure 30 – Estimation des paramètres pour le candidat 3 de l'exemple 4

Dans le cas du candidat 3 qui est un modèle ARIMA $(p,d,q)=(1,0,1)$, on a constaté que la constante MU était significative à 5% mais qu'elle n'améliorait pas nécessairement le modèle. En effet, les probabilités du modèle avec MU sont : 0.6681, 0.4950, 0.4508 et 0.3574 et celle du modèle sans MU sont : 0.9243, 0.7794, 0.9573 et 0.8596. Toutefois, la sortie SAS du modèle sans MU indique que le paramètre MA1,1 ne serait plus significatif à 5% et cela reviendrait à tester un modèle autorégressif d'ordre 1 qui a déjà été vérifié comme étant le candidat 1. Nous garderons donc la constante MU pour le candidat 3. Les paramètres AR1,1 (ϕ_1) et MA1,1 (θ_1) sont significatifs à 5%. Les estimations respectives pour les trois paramètres sont : 13.88339, 0.46735 et 1.00000. Dans la figure 29 à la page suivante, on retrouve aussi les diagrammes des fonctions d'autocorrélation reliées aux résidus et on constate qu'il n'y a pas de pic (à partir de r_1) ce qui est souhaitable car les résidus ne sont pas supposés être corrélés entre eux. En ce qui concerne l'erreur type et l'indicateur RMRES2 associés au candidat 3, les résultats sont respectivement : 2.972145 et 3.525732912.



Autocorrelation Plot of Residuals

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	8.833646	1.00000	|********************|	0
1	0.604652	0.06845	|* . |	0.188982
2	-0.189520	-.02145	| . |	0.189866
3	0.613555	0.06946	|* . |	0.189952
4	-1.068992	-.12101	| . **| . |	0.190857
5	-0.728802	-.08250	| . **| . |	0.193578
6	1.621098	0.18351	| . |**** . |	0.194830
7	-1.101757	-.12472	| . **| . |	0.200908

"." marks two standard errors

Inverse Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	-0.14393	| . ***| . |
2	0.00917	| . | . |
3	-0.02960	| . *| . |
4	0.08534	| . |** . |
5	0.09220	| . |** . |
6	-0.19819	| . ****| . |
7	0.13570	| . |*** . |

Le Système SAS 11:54 Saturday, January 23, 2010 18

The ARIMA Procedure

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	0.06845	| . |* . |
2	-0.02626	| . *| . |
3	0.07315	| . |* . |
4	-0.13313	| . ***| . |
5	-0.06095	| . *| . |
6	0.18771	| . |**** . |
7	-0.14874	| . ***| . |

Figure 31 - Probabilités (Check of Residuals) et diagrammes d'autocorrélation des résidus du candidat 3

Discussion sur les candidats de l'exemple 4

Il a d'abord été question de vérifier que les paramètres de chacun des candidats retenus sont significatifs. Le cas échéant, le candidat concerné était automatiquement rejeté. Ensuite, les comparaisons des trois facteurs d'indice ont été effectuées. Ces trois facteurs sont les probabilités associées à la statistique Ljung-Box, les valeurs de l'erreur type et les valeurs de l'indicateur RMRES2 associées à chaque candidat.

Les paramètres de tous les candidats ont été trouvés significatifs. Les probabilités de chaque candidat ont été comparées et il a été trouvé que c'est le candidat 1 : ARIMA $(1,0,0)$ qui a les probabilités les plus élevées. Non seulement ce sont les plus élevées mais elles sont aussi élevées en soi : 0.9027, 0.8586, 0.8994 et 0.8754. Notons qu'elles sont toutes supérieures à 80% ce qui peut être considéré comme relativement bon.

Les comparaisons des valeurs pour l'erreur type et le RMRES2 donnent les mêmes résultats que la comparaison des probabilités. Le candidat 1 possèdent les valeurs les plus petites qui sont respectivement : 3.197808 et 4.5309222566 alors que pour le candidat 2 elles étaient de 3.326657 et 4.7558067313 et pour le candidat 3 ; 2.972145 et 3.525732912. Après l'examen de toutes ces conditions nous pouvons en conclure que le candidat 1 : ARIMA $(1,0,0)$ est le meilleur modèle pour la série de l'exemple 4.

3.4.3.2 Candidats potentiels de l'exemple 5

Pour l'exemple 5, les candidats[2] potentiels retenus lors de la phase d'identification sont les modèles ARIMA d'ordre : $(p,d,q)=(1,1,0)$, $(p,d,q)=(0,1,1)$ et $(p,d,q)=(1,1,1)$.

Pour le candidat 1, on a constaté que la constante MU n'était pas significative à 5% donc on la rejette automatiquement du modèle.

Le paramètre AR1,1 (ϕ_1) est significatif à 5% et son estimation est de -0.54857. On remarque aussi l'absence de pic (à partir de r_1) dans les diagrammes d'autocorrélation des résidus. Pour ce qui est de l'erreur type et de la valeur de l'indicateur RMRES2 associées au candidat 1, les résultats sont respectivement : 3.224615 et 3.4544394914.

Pour le candidat 2, on a constaté que la constante MU était significative à 5% et qu'elle améliorait le modèle. En effet, elle faisait passer les probabilités de 0.7300, 0.9819, 0.9790 et 0.9679 à 0.7710, 0.9522, 0.9869 et 0.9855. Le paramètre MA1,1 (θ_1) est significatif à 5%. Les estimations respectives des paramètres sont : -0.43887 et 1.00000. On remarque aussi l'absence de pic (à partir de r_1) dans les diagrammes d'autocorrélation des résidus. En ce qui concerne l'erreur type et l'indicateur RMRES2 associés au candidat 2, les résultats sont respectivement : 2.866213 et 2.8299938231.

Le candidat 3, qui est un modèle mixte, n'est pas retenu car l'un de ses paramètres n'est pas significatif à 5%.

---

[2] Sorties SAS reliées à l'exemple 5 se retrouvent à l'annexe B

Discussion sur les candidats de l'exemple 5

Encore une fois il faut d'abord vérifier que les paramètres de chacun des candidats retenus sont significatifs. Ensuite, on procède aux comparaisons des probabilités associées à la statistique Ljung-Box, des valeurs de l'erreur type et de l'indicateur RMRES2.

Dans le cas du candidat 1, on a exclu la constante MU car elle n'était pas significative. Pour le candidat 2, la constante MU a été retirée car elle n'était pas significative à 5%. Le candidat 3 a été automatiquement rejeté car le paramètre AR1,1 n'était pas significatif à 5%. La course pour le meilleur modèle se joue donc entre les candidats 1 et 2.

Les probabilités de chaque candidat ont été comparées et il a été trouvé que c'est le candidat 2 : ARIMA (0,1,1) qui a les probabilités les plus élevées. Non seulement ce sont les plus élevées mais elles sont aussi élevées en tant que tel : 0.7710, 0.8522, 0.9869 et 0.9855. Notons que trois d'entre elles sont supérieures à 85% ce qui peut être considéré comme relativement bon.

Les comparaisons des valeurs pour l'erreur type et le RMRES2 donnent les mêmes résultats que la comparaison des probabilités. Le candidat 2 possèdent les valeurs les plus petites qui sont respectivement : 2.866213 et 2.8299938231 alors que pour le candidat 1 elles étaient de 3.232751 et 3.3783792663.

Après l'examen de toutes ces conditions nous pouvons en conclure que le candidat 2 : ARIMA (0,1,1) est le meilleur modèle pour la série de l'exemple 5.

3.4.3.3 Candidats potentiels de l'exemple 6

Pour l'exemple 6, les candidats[3] potentiels retenus lors de la phase d'identification sont les modèles ARIMA d'ordre : $(p,d,q)=(1,0,0)$, $(p,d,q)=(0,0,1)$ et $(p,d,q)=(1,0,1)$.

Pour le candidat 1, on a constaté que la constante MU était significative à 5% et est estimée à 16.34102. Elle améliorait le modèle en faisant passer les probabilités de 0.6738, 0.6885, 0.7542 et 0.1817 à 0.7366, 0.6198, 0.6806 et 0.3246.

Le paramètre AR1,1 (ϕ_1) est significatif à 5% et son estimation est de 0.44108. On remarque aussi l'absence de pic (à partir de r_1) dans les diagrammes d'autocorrélation des résidus. Pour ce qui est de l'erreur type et de la valeur de l'indicateur RMRES2 associées au candidat 1, les résultats sont respectivement : 5.7180332 et 5.9912786705.

Pour le candidat 2, on a constaté que la constante MU était significative à 5% et qu'elle améliorait le modèle. En effet, elle faisait passer les probabilités de 0.0001, 0.0001, 0.0001 et 0.0001 à 0.4590, 0.4985, 0.3732 et 0.1794. Le paramètre MA1,1 (θ_1) est significatif à 5%. Les estimations respectives des paramètres sont : 16.07922 et -0.44437. On remarque aussi l'absence de pic (à partir de r_1) dans les diagrammes d'autocorrélation des résidus. En ce qui concerne l'erreur type et l'indicateur RMRES2 associés au candidat 2, les résultats sont respectivement : 5.782475 et 6.1062413098.

Le candidat 3 est un modèle mixte tous ses paramètres sont significatif à 5%. L'estimation de AR1,1 est 1.00000, celle de MA1,1 est 0.75155 et celle de MU est 19.66188. MU améliorait le modèle en faisant passer les probabilités de 0.8559, 0.5695, 0.8574 et 0.5722 à 0.9246, 0.5805, 0.8684 et 0.6192. L'erreur type estimée est de 5.537851 alors que le RMRES2 valait 5.9650984309.

---

[3] Sorties SAS reliées à l'Exemple 6 se retrouvent à l'annexe C

Discussion sur les candidats de l'exemple 6

Dans le cas de l'exemple 6, tous les paramètres des modèles étaient significatifs à 5%.

Sur le plan de la comparaison des probabilités de chaque candidat, on trouve que c'est le candidat 3 : ARIMA (1,0,1) qui a les probabilités les plus élevées. Notons que deux d'entre elles sont supérieures à 85% ce qui est très bien.

Les comparaisons des valeurs pour l'erreur type et le RMRES2 donnent les mêmes résultats que la comparaison des probabilités. Le candidat 3 possèdent les valeurs les plus petites avec respectivement comme valeurs: 5.537851 et 5.9650984309. On remarque toutefois que ces dernières mesures sont plus élevées que pour les autres exemples analysés.

Après l'examen de toutes ces conditions nous pouvons en conclure que le candidat 3 : ARIMA (1,0,1) est le meilleur modèle pour la série de l'exemple 6.

3.4.3.4 Candidats potentiels de l'exemple 7

Pour l'exemple 7, les candidats[4] potentiels retenus lors de la phase d'identification sont les modèles ARIMA d'ordre : $(p,d,q) = (1,1,0)$, $(p,d,q) = (0,1,1)$ et $(p,d,q) = (1,1,1)$.

Pour le candidat 1, on a constaté que la constante MU était significative à 5% et qu'elle améliorait le modèle. En effet, elle faisait passer les probabilités de 0.1343, 0.1189, 0.1320 et 0.1328 à 0.1956, 0.1402, 0.1978 et 0.4879. Le paramètre AR1,1 (ϕ_1) est significatif à 5%. Les estimations respectives des paramètres sont : -0.08354 et -0.47653. On remarque aussi l'absence de pic dans les diagrammes d'autocorrélation des résidus. En ce qui concerne l'erreur type et l'indicateur RMRES2 associés au candidat 1, les résultats sont respectivement : 0.2159 et 0.2665450171 (plus faibles qu'à l'ordinaire car la série a été transformée avec la fonction logarithmique).

Pour le candidat 2, on a constaté que la constante MU était significative à 5% et qu'elle améliorait le modèle. En effet, elle faisait passer les probabilités de 0.1148, 0.0801, 0.0926 et 0.0927 à 0.7710, 0.9522, 0.9869 et 0.9855. Le paramètre MA1,1 (θ_1) est significatif à 5%. Les estimations respectives des paramètres sont : -0.08039 et 0.66393. On remarque aussi l'absence de pic (à partir de r_1) dans les diagrammes d'autocorrélation des résidus. En ce qui concerne l'erreur type et l'indicateur RMRES2 associés au candidat 2, les résultats sont respectivement : 0.200033 (faible car c'est la série transformée log) et 0.2739308096. Les estimations respectives pour les trois paramètres sont : -0.08070, 0.61366 et -0.10150. Il n'y a pas de pic dans les diagrammes des fonctions d'autocorrélation reliées aux résidus. En ce qui concerne l'erreur type et l'indicateur RMRES2 associés au candidat 1, les résultats sont respectivement : 0.200033 et 0.2739308096.

Le candidat 3 a été rejeté car l'un de ses paramètres n'est pas significatif à 5%.

---

[4] Sorties SAS reliées à l'Exemple 7 se retrouvent à l'annexe D

Discussion sur les candidats de l'exemple 7

On commence par vérifier que les paramètres de chacun des candidats retenus sont significatifs. Après quoi, on enchaîne avec les comparaisons des probabilités associées à la statistique Ljung-Box, des valeurs de l'erreur type et de l'indicateur RMRES2.

Pour les deux premiers candidats, tous les paramètres étaient significatifs à 5%. Les probabilités de chaque candidat ont été comparées et il a été trouvé que c'est le candidat 2 : ARIMA (0,1,1) qui a les probabilités les plus élevées. On remarque qu'elles sont très faibles (0.1956, 0.1402, 0.1978 et 0.4879).

Les comparaisons des valeurs pour l'erreur type et le RMRES2 ne donnent pas les mêmes résultats que la comparaison des probabilités. Le candidat 2 possède la valeur de l'indicateur RMRES2 la plus petite qui est égal à 0.200033 alors que le candidat 1 avait une valeur de 0.2159. Pour l'erreur type, le candidat 1 a une valeur inférieure à celle du candidat 2 et ces valeurs sont respectivement : 0.2665450171 et 0.2739308096.

Après l'examen de toutes ces conditions nous ne pouvons pas en conclure que le candidat 2 : est le meilleur modèle car les probabilités qui lui sont rattachées sont trop faibles. La série de l'exemple 7 est un cas où la méthodologie n'a pas pu être appliquée et dans lequel il a fallu se tourner vers des alternatives. Nous traiterons ces autres types de modélisation dans la section suivante.

3.5 Équation du modèle et prévisions

Les équations de chacun des modèles des exemples ainsi que la sortie SAS résumant le modèle seront présentées dans cette section. On retrouve ici les notations abordées dans le chapitre 2 lors de l'exposition de la théorie sous-jacente à la méthodologie Box-Jenkins.

3.5.1 Exemple 4

$$\begin{aligned} X_t &= \mu + \phi_1 X_{t-1} + a_t \\ &= 7.820409 - 0.5366 X_{t-1} + a_t. \end{aligned}$$

Model for variable n

Estimated Mean 7.820409

Autoregressive Factors

Factor 1: 1 - 0.5366 B**(1)

Prévisions pour la variable n

Obs	Forecast	Std Error	Intervalle de confiance à 95 %	
29	5.7724	3.1978	-0.4972	12.0380
30	6.7204	3.6291	-0.3925	13.8333
31	7.2301	3.7441	-0.1081	14.5684
32	7.5037	3.7765	0.1019	14.9056
33	7.6504	3.7858	0.2303	15.0706

Les prévisions pour les cinq prochaines années sont les valeurs arrondies à l'unité : 6,7,8,8 et 8.

3.5.2 Exemple 5

$$X_t = \mu + a_t - \theta_1 a_{t-1}$$
$$= -0.43887 - 1a_{t-1} + a_t.$$

Model for variable n

Estimated Mean -0.43887
Period(s) of Differencing 1

Moving Average Factors

Factor 1: 1 - 1 B**(1)

The ARIMA Procedure

Prévisions pour la variable n

Obs	Forecast	Std Error	Intervalle de confiance à 95 %	
29	3.7117	2.8662	-1.9059	9.3294
30	3.2729	2.8662	-2.3448	8.8905
31	2.8340	2.8662	-2.7837	8.4517
32	2.3951	2.8662	-3.2225	8.0128
33	1.9563	2.8662	-3.6614	7.5739

Les prévisions pour les cinq prochaines années sont les valeurs arrondies à l'unité : 4,4,3,3 et 2.

3.5.3 Exemple 6

$$X_t = \mu + \phi_1 X_{t-1} + a_t - \theta_1 a_{t-1}$$
$$= 19.66188 + 1.00000 X_{t-1} + a_t - 0.75155 a_{t-1}.$$

Model for variable y

Estimated Mean -0.0807
Period(s) of Differencing 1

Autoregressive Factors

Factor 1: 1 + 0.1015 B**(1)

Moving Average Factors

Factor 1: 1 - 0.61366 B**(1)

Prévisions pour la variable y

Obs	Forecast	Std Error	Intervalle de confiance à 95 %	
29	2.5935	0.2036	2.1944	2.9925
30	2.5092	0.2117	2.0942	2.9241
31	2.4288	0.2239	1.9901	2.8676
32	2.3481	0.2350	1.8876	2.8086
33	2.2674	0.2456	1.7861	2.7487

Le Système SAS 09:49 Monday, February 22,

Obs	FORECAST	expo	expoL95	expoU95
1	2.59345	13.3759	8.97431	19.9362
2	2.50919	12.2949	8.11916	18.6183
3	2.42885	11.3458	7.31592	17.5954
4	2.34811	10.4657	6.60357	16.5867
5	2.26741	9.6543	5.96613	15.6225

Les prévisions pour les cinq prochaines années sont donc les valeurs arrondies à l'unité près dont on a extrait l'exponentielle car on avait utilisé la transformation logarithmique : 14,13,12,11 et 10.

3.6 Discussion sur les résultats de la modélisation ARIMA

Sur les 57 séries qui ont pu être modélisées avec la méthodologie ARIMA il est intéressant de noter que 29 séries étaient des modèles autorégressifs d'ordre 1 avec ou sans différence, 20 étaient des modèles de moyennes mobiles d'ordre 1 avec ou sans différence et seulement 4 étaient des modèles mixtes d'ordre (1,0,1) ou (1,1,1). On constate que la majorité des séries de type 1 ont été modélisées avec un modèle autorégressif d'ordre 1 avec ou sans différence alors que la majorité des modèles des séries de type 2 correspondaient plutôt à des modèles de moyennes mobiles d'ordre 1 avec ou sans différence. Ajoutons aussi que dans les deux différents cas de séries la constante MU améliorait le modèle la plupart du temps. Dans le cas particulier de la série de type 2, la transformation logarithmique améliorait souvent le modèle. Dans l'ensemble des séries, une différenciation d'ordre 1 était souvent nécessaire afin de rendre le modèle stationnaire.

3.7 Alternatives aux modèles ARIMA

3.7.1 Régression linéaire

3.7.1.2 Théorie et conditions d'application

Pour un rappel de la théorie portant sur la régression linéaire, on invite le lecteur à se rapporter à la section 1.3.1 du chapitre 1. Pour les manipulations à l'aide du logiciel SAS, il s'agit de s'en remettre au chapitre 3 du livre : *Forecasting time series* de Bowerman/O'Connell (2006) à la section 3.9. La régression linéaire a été la première alternative vers laquelle nous nous sommes tournés. Nous l'appliquions dans les cas où les résidus n'étaient pas corrélés entre eux afin de satisfaire les hypothèses de la régression linéaire. Pour ce faire, on procédait à la vérification de la valeur au test du Durbin-Watson avec les valeurs de référence $d_{L,\alpha}$ et $d_{U,\alpha}$ qui avaient pour valeurs respectives : 1.32844 et 1.4589 pour un nombre de données de 28 avec deux degrés de liberté. Lorsque la valeur est inférieure à $d_{L,\alpha}$ il y a présence d'autocorrélation positive dans les résidus et on ne peut pas appliquer la régression linéaire. Lorsque la valeur est supérieure à $d_{U,\alpha}$ mais pas supérieure à 2, on peut conclure qu'il n'y a pas d'autocorrélation dans les résidus. Avec une valeur supérieure à 2, on peut soupçonner de l'autocorrélation négative dans les résidus et on doit donc tester si $4-d<d_{L,\alpha}$ ou $4-d>d_{U,\alpha}$ avec les mêmes conclusions que le test pour l'autocorrélation positive. Un exemple de l'application de cette méthode suit.

3.7.1.3 Exemple 8

Le Système SAS

The REG Procedure
Model: MODEL1
Dependent Variable: y

Durbin-Watson D	1.909
Number of Observations	28
1st Order Autocorrelation	0.025

Figure 32 – Valeur du Durbin-Watson de l'exemple 8

On remarque que la valeur $d = 1.909 > d_{(1,\alpha}$ et elle est inférieure à 2. Il n'y a donc pas d'autocorrélation positive dans les résidus. Nous pouvons donc appliquer la régression linéaire et obtenir les prévisions rattachées au modèle.

The REG Procedure
Model: MODEL1
Dependent Variable: y

Statistiques de sortie

Obs.	Variable dépendante	Valeur	Erreur standard Prédiction de la moyenne	95% prédiction CL		Résidus
1	3.6636	3.5283	0.1048	2.9045	4.1520	0.1353
2	3.1355	3.4827	0.0992	2.8629	4.1026	-0.3473
3	3.4340	3.4372	0.0936	2.8210	4.0534	-0.003250
4	3.4340	3.3917	0.0883	2.7789	4.0046	0.0423
5	3.1355	3.3462	0.0831	2.7364	3.9560	-0.2107
6	3.2189	3.3007	0.0781	2.6937	3.9077	-0.0818
7	3.0910	3.2552	0.0734	2.6507	3.8597	-0.1642
8	2.9444	3.2097	0.0691	2.6073	3.8121	-0.2653
9	3.5264	3.1642	0.0651	2.5637	3.7647	0.3622
10	3.1355	3.1187	0.0616	2.5198	3.7176	0.0168
11	2.7081	3.0732	0.0587	2.4755	3.6708	-0.3651
12	3.3322	3.0277	0.0563	2.4309	3.6244	0.3045
13	2.9957	2.9822	0.0547	2.3861	3.5783	0.0136
14	2.9957	2.9367	0.0539	2.3409	3.5324	0.0591
15	3.4340	2.8911	0.0539	2.2954	3.4869	0.5428
16	2.8332	2.8456	0.0547	2.2495	3.4417	-0.0124
17	3.4340	2.8001	0.0563	2.2034	3.3969	0.6339
18	2.9957	2.7546	0.0587	2.1570	3.3523	0.2411
19	3.0445	2.7091	0.0616	2.1102	3.3080	0.3354
20	2.6391	2.6636	0.0651	2.0631	3.2641	-0.0246
21	2.3026	2.6181	0.0691	2.0158	3.2205	-0.3155
22	2.4849	2.5726	0.0734	1.9681	3.1771	-0.0877
23	2.0794	2.5271	0.0781	1.9201	3.1341	-0.4476
24	2.5649	2.4816	0.0831	1.8718	3.0914	0.0834
25	2.0794	2.4361	0.0883	1.8232	3.0489	-0.3566
26	2.3026	2.3906	0.0936	1.7744	3.0068	-0.0880
27	2.0794	2.3451	0.0992	1.7252	2.9649	-0.2656
28	2.5649	2.2996	0.1048	1.6758	2.9233	0.2654
29	.	2.2540	0.1106	1.6261	2.8820	.
30	.	2.2085	0.1165	1.5761	2.8410	.
31	.	2.1630	0.1224	1.5259	2.8002	.
32	.	2.1175	0.1284	1.4754	2.7597	.
33	.	2.0720	0.1345	1.4246	2.7194	.

Sum of Residuals	0
Sum of Squared Residuals	2.10861
Predicted Residual SS (PRESS)	2.40528

3.7.2 Régression de Poisson

3.7.2.1 Théorie

Pour une revue plus complète de la théorie qui sous-tend la régression de poisson on renvoie le lecteur à un ouvrage plus spécifique [16].

Une hypothèse fondamentale de la régression de poisson est que la moyenne et la variance soient égales. Nous pouvons vérifier cette condition à l'aide de la statistique de Pearson. Afin que l'on puisse considérer qu'il n'y a pas de surdispersion des données, ce qui aurait pour conséquence que la moyenne et la variance ne seraient pas égales, il faut que le résultat au test de Pearson donne environ 1.

Lorsqu'il y a présence de surdispersion on peut se tourner vers le modèle de la binomiale négative. Cela n'a pas été notre cas alors nous n'aborderons pas ce modèle.

3.7.2.2 Exemple 9

Le Système SAS

The GENMOD Procedure

Informations sur le modèle

Data Set	WORK.POIS1R908
Distribution	Poisson
Link Function	Log
Dependent Variable	n
Number of Observations Read	20
Number of Observations Used	15
Missing Values	5

Critère pour évaluer la qualité de l'ajustement

Critère	DF	Valeur	Valeur/DF
Deviance	13	13.1582	1.0122
Scaled Deviance	13	13.1582	1.0122
Pearson Chi-Square	13	13.9020	1.0694
Scaled Pearson X2	13	13.9020	1.0694
Log Likelihood		9.2994	

Figure 33 – Valeur du test Pearson de l'exemple 9

La valeur du test de Pearson est de 1.0694. Cette valeur peut être considérée comme près de 1. Nous pouvons donc procéder à la production des prévisions qui découlent de ce modèle.

		0.9462984	0.8779614	0.9089581	0.9797078	0.9138725

Le Système SAS 11:23 Monday,

The GENMOD Procedure

Observation Statistics

Observation	n	Pred Reschi	Xbeta Resdev	Std StResdev	HessWgt StReschi	Resraw Reslik
9	3	3.3575601	1.2112146	0.1456022	3.3575601	-0.35756
		-0.195136	-0.198763	-0.206238	-0.202475	-0.205973
10	5	3.4318266	1.2330927	0.1537589	3.4318266	1.5681734
		0.8465087	0.7919001	0.8261224	0.883091	0.8308902
11	1	3.5077358	1.2549708	0.166535	3.5077358	-2.507736
		-1.338962	-1.582887	-1.665996	-1.409265	-1.642783
12	4	3.5853241	1.2768489	0.1829654	3.5853241	0.4146759
		0.2190002	0.21497	0.2291618	0.2334582	0.2296817
13	2	3.6646285	1.298727	0.2021611	3.6646285	-1.664629
		-0.869566	-0.952333	-1.03281	-0.943049	-1.01987
14	3	3.7456871	1.3206051	0.2234103	3.7456871	-0.745687
		-0.385293	-0.399271	-0.442803	-0.427301	-0.439947
15	5	3.8285387	1.3424832	0.246182	3.8285387	1.1714613
		0.598703	0.5715108	0.6521576	0.6831869	0.6594874
16	.	3.9132228	1.3643613	0.2700912	.	.
		.	.	.	.	.
17	.	3.9997801	1.3862394	0.2948615	.	.
		.	.	.	.	.
18	.	4.088252	1.4081175	0.3202931	.	.
		.	.	.	.	.
19	.	4.1786808	1.4299956	0.3462403	.	.
		.	.	.	.	.
20	.	4.2711098	1.4518737	0.3725954	.	.
		.	.	.	.	.

Figure 34 – Prévisions rattachées au modèle de régression de Poisson de l'exemple 9

Les prévisions pour les cinq prochaines années sont donc les valeurs arrondies à l'unité près des observations 16 à 20 : 4,4,5,5 et 5.

3.7.3 Modélisation des résidus avec Proc Arima

3.7.3.1 Théorie

Pour la théorie portant sur la modélisation des résidus avec la méthode PROC ARIMA il faut se référer à la section 6.6 du chapitre 6 du livre : *Forecasting time series* de Bowerman/O'Connell (2006). Résumons la méthode en disant qu'un modèle de moyenne mobile d'ordre 1 sera appliqué pour modéliser les résidus et nous donner un modèle global pour faire la prévision. Cette méthode a été appliquée à toutes les séries montrant de l'autocorrélation positive dans les résidus.

3.7.3.2 Exemple 10

```
                    Procédure REG
                  Modèle : MODEL1
              Variable dépendante : y

Durbin-Watson D                    1.108
Nombre d'observations                 28
Autocorrélation de 1er ordre       0.264
```

Figure 35 – Valeur du test du Durbin-Watson pour l'exemple 10

On constate que $d = 1.108 < d_{L,\alpha}$ ce qui indique qu'il y a présence d'autocorrélation positive dans les résidus. Nous appliquons donc la méthode PROC ARIMA à la série de l'exemple 10.

```
                          Le Système SAS          10:42 Sunday, March 30, 2008  90

                        The ARIMA Procedure

                  Conditional Least Squares Estimation

                          Erreur                                  Pr.
Paramètre  Estimation    standard  Valeur du test t  Approx. > |t|  Retard  Variable    Shift

AR1,1       1.00000     0.03020             33.11         <.0001       1  n                0
NUM1        0.08419   0.0073525             11.45         <.0001       0  financiere       0
```

Figure 36 – Estimation des paramètres du modèle de l'exemple 10

On remarque que tous les paramètres sont significatifs à 5%.

Autoregressive Factors

Factor 1: 1 - 1 B**(1)

Input Number 1

Input Variable	financiere
Overall Regression Factor	0.084187

Prévisions pour la variable n

Obs	Forecast	Std Error	95Limites de confiance %		Actual	Residual
1	166.8583	14.5726	138.2966	195.4200	171.0000	4.1417
2	171.0842	14.5726	142.5225	199.6459	129.0000	-42.0842
3	129.0842	14.5726	100.5225	157.6459	122.0000	-7.0842
4	122.0842	14.5726	93.5225	150.6459	125.0000	2.9158
5	125.0842	14.5726	96.5225	153.6459	108.0000	-17.0842
6	108.0842	14.5726	79.5225	136.6459	101.0000	-7.0842
7	101.0842	14.5726	72.5225	129.6459	87.0000	-14.0842
8	87.0842	14.5726	58.5225	115.6459	69.0000	-18.0842
9	69.0842	14.5726	40.5225	97.6459	70.0000	0.9158
10	70.0842	14.5726	41.5225	98.6459	77.0000	6.9158
11	77.0842	14.5726	48.5225	105.6459	85.0000	7.9158
12	85.0842	14.5726	56.5225	113.6459	76.0000	-9.0842
13	76.0842	14.5726	47.5225	104.6459	65.0000	-11.0842
14	65.0842	14.5726	36.5225	93.6459	78.0000	12.9158
15	78.0842	14.5726	49.5225	106.6459	72.0000	-6.0842
16	72.0842	14.5726	43.5225	100.6459	56.0000	-16.0842
17	56.0842	14.5726	27.5225	84.6459	75.0000	18.9158
18	75.0842	14.5726	46.5225	103.6459	57.0000	-18.0842
19	57.0842	14.5726	28.5225	85.6459	59.0000	1.9158
20	59.0842	14.5726	30.5225	87.6459	60.0000	0.9158
21	60.0842	14.5726	31.5225	88.6459	61.0000	0.9158
22	61.0842	14.5726	32.5225	89.6459	68.0000	6.9158
23	68.0842	14.5726	39.5225	96.6459	52.0000	-16.0842
24	52.0842	14.5726	23.5225	80.6459	71.0000	18.9158
25	71.0842	14.5726	42.5225	99.6459	55.0000	-16.0842
26	55.0842	14.5726	26.5225	83.6459	54.0000	-1.0842
27	54.0842	14.5726	25.5225	82.6459	50.0000	-4.0842
28	50.0842	14.5726	21.5225	78.6459	70.0000	19.9158
29	70.0842	14.5726	41.5225	98.6459	.	.
30	70.1684	20.6087	29.7760	110.5607	.	.
31	70.2526	25.2404	20.7823	119.7229	.	.
32	70.3368	29.1451	13.2134	127.4602	.	.
33	70.4210	32.5852	6.5551	134.2869	.	.

Figure 37 – Prévisions rattachées au modèle de modélisation des résidus de l'exemple 10

Cet exemple bouclait la section sur les alternatives utilisées lorsque la modélisation ARIMA n'aboutissait pas à des résultats concluants. Un bilan des analyses ainsi qu'une ouverture du sujet feront l'objet de la conclusion de ce mémoire.

CONCLUSION

Depuis les années soixante-dix les techniques de modélisation des séries chronologiques ont beaucoup évolué et se sont perfectionnées. Parmi les plus connues et utilisées s'impose la méthodologie Box-Jenkins. Elle s'avère être un outil puissant et performant pour faire de la prévision. C'est pourquoi nous les modèles ARIMA comme approche pour le projet du le Ministère de la Santé et des Services Sociaux qui consistait à élaborer des modèles à des fins de prévision. Nos deux principaux buts étaient de voir si les modèles ARIMA étaient appropriés pour modéliser les séries de l'un des programmes du ministère et d'en sortir les prévisions pour les cinq prochaines années. Ce processus constitue donc un outil d'aide à la décision pour les différents aspects de gestion rattachés à ce programme particulier du Ministère. Ce mémoire débute donc avec une introduction sur les séries chronologiques afin de faciliter la compréhension de la nature des données en jeu dans les différentes séries du programme. Ensuite, le cadre théorique de la méthodologie Box-Jenkins a été détaillé afin de s'assurer que les concepts nécessaires à son application sont maîtrisés. Enfin, le troisième chapitre présente des exemples de séries qui ont pu être modélisées avec des modèles ARIMA. Il présente aussi certaines difficultés rencontrées lors des analyses ainsi que les alternatives utilisées pour les surmonter. Le chapitre se termine avec l'énoncé des résultats de l'ensemble de la modélisation des séries. Étant donné que seulement 57 des 188 séries à modéliser ont utilisé la méthodologie Box-Jenkins, il faut en convenir que ce n'est peut-être pas la meilleure approche pour le contexte du programme SIPPE du Ministère. Une perspective intéressante serait d'utiliser des modèles particuliers où l'on tient compte de la nature discrète des données des séries en question.

Dans le futur, ce mémoire pourrait servir de point de départ pour un travail d'exploration et d'élaboration de d'autres outils d'aide à la décision basé sur la prévision à l'aide de séries chronologiques. Les modèles concluants ici élaborés pourraient aussi être mis à jour en combinaison avec d'autres approches. L'utilisation des modèles à espaces d'état semble être une approche avec beaucoup de potentiel.

RÉFÉRENCES

[1] Statistique Canada, (2009) « Indicateurs financiers du bilan national, glossaire » Sur le site de Statistique Canada [En ligne]. (Page consultée le 14 avril 2011) http://www.statcan.gc.ca/pub/13-010-x/2009001/article03-fra.htm

[2] Ministère de la Santé et des Services sociaux, (2009) « Éco-Santé Québec » Sur le site du Ministère de la Santé et des Services sociaux [En ligne] (Page consultée le 14 avril 2011) http://www.msss.gouv.qc.ca/statistiques/stats_sss/index.php?id=143,248,0,0,1,0

[3] Éditeur officiel du Québec, (2011) « Loi sur la santé publique » Sur le site des Publications du Québec [En ligne] (Page consultée le 14 avril 2011) http://www2.publicationsduquebec.gouv.qc.ca/dynamicSearch/telecharge.php?type=2&file=/S_2_2/S2_2.html

[4] Ministère de la Santé et des Services sociaux, (2011) « Surveillance » Sur le site du Ministère de la Santé et des Services sociaux [En ligne] (Page consultée le 14 avril 2011) http://www.msss.gouv.qc.ca/sujets/santepub/environnement/index.php?surveillance

[5] BOWERMAN, Bruce L., and Richard T. O'Connell (1993) *Forecasting and time series: an applied approach* (3rd Edition) North Scituate, Mass. : Duxbury Press

[6] SEN, Ashish K. and Muni S. Srivasti (1990) *Regression analysis: theory, methods and applications* New York, Springer-Verlag

[7] ANDERSON, T. W. (1994) *The Statistical Analysis of Time Series* New York, J. Wiley

[8] WILLIAMS, Evan J (1959) *Regression Analysis* New York, J. Wiley

[9] DRAPER, Norman R and Harry Smith (1966) *Applied Regression Analysis* New York, Wiley

[10] GRAYBILL, Franklin A. (1961) *An Introduction to Linear Statistical Models* New York, McGraw Hill

[11] U.S. Census Bureau, (2011) « The X-12-ARIMA Seasonal Adjustment Program» Sur le site du U.S. Census Bureau [En ligne] (Page consultée le 14 avril 2011) http://www.census.gov/srd/www/x12a

[12] BOWERMAN, Bruce L., Richard T. O'Connell and Anne Koehler (2006) *Forecasting and time series: an applied approach* (4th Edition) Belmont, CA Thomson Brooks/Cole

[13] ALLISON, Paul David. (1999) *Logistic regression using the SAS system theory and application* Safari Tech Books Online Cary, N.C. SAS Institute

BIBLIOGRAPHIE

Livres

ALLISON, Paul David. (1999) *Logistic regression using the SAS system theory and application* Safari Tech Books Online Cary, N.C. SAS Institute

ANDERSON, T. W. (1994) *The Statistical Analysis of Time Series* New York, J. Wiley

BOSQ, Denis, et Jean-Pierre Lecoutre (1992) *Analyse et prévision des séries chronologiques : méthodes paramétriques et non paramétriques* Paris, Masson

BOWERMAN, Bruce L., and Richard T. O'Connell (1993) *Forecasting and time series: an applied approach (3rd Edition)* North Scituate, Mass. : Duxbury Press

BOWERMAN, Bruce L., Richard T. O'Connell and Anne Koehler (2006) *Forecasting and time series: an applied approach (4th Edition)* Belmont, CA Thomson Brooks/Cole

BOX, George P., and Gwilym M. Jenkins (1976) *Time Series Analysis: forecasting and control (Revised Edition)* Toronto, Holden-Day

BRILLINGER, David R (1975) *Time series: data analysis and theory* New York, Holt, Rinehart and Winston

BROCKWELL, Peter J. and Richard A. Davis (1996) *Introduction to Time Series and Forecasting* New York, Springer-Verlag

BROCKWELL, Peter J. and Richard A. David (1991) *Time Series: Theory and Methods (2nd Edition)* New York, Springer-Verlag

DIGGLE, Peter J. (1990) *Time Series: A Biostatistical Introduction* Oxford, Clarendon Press

DRAPER, Norman R and Harry Smith (1966) *Applied Regression Analysis* New York, Wiley

DURBIN, J. and S. J. Koopman (2001) *Time Series Analysis by State Space Methods* Oxford, Oxford University Press

GRAYBILL, Franklin A. (1961) *An Introduction to Linear Statistical Models* New York, McGraw Hill

SEN, Ashish K. and Muni S. Srivasti (1990) *Regression analysis: theory, methods and applications* New York, Springer-Verlag

WILLIAMS, Evan J (1959) *Regression Analysis* New York, J. Wiley

Sites Internet

Éditeur officiel du Québec, (2011) « Loi sur la santé publique » Sur le site des Publications du Québec [En ligne] (Page consultée le 14 avril 2011)

http://www2.publicationsduquebec.gouv.qc.ca/dynamicSearch/telecharge.php?type=2&file=/S_2_2/S2_2.html

Ministère de la Santé et des Services sociaux, (2009) « Éco-Santé Québec » Sur le site du Ministère de la Santé et des Services sociaux [En ligne] (Page consultée le 14 avril 2011)

http://www.msss.gouv.qc.ca/statistiques/stats_sss/index.php?id=143,248,0,0,1,0

Ministère de la Santé et des Services sociaux, (2011) « Surveillance » Sur le site du Ministère de la Santé et des Services sociaux [En ligne] (Page consultée le 14 avril 2011)

http://www.msss.gouv.qc.ca/sujets/santepub/environnement/index.php?surveillance

Statistique Canada, (2009) « Indicateurs financiers du bilan national, glossaire » Sur le site de Statistique Canada [En ligne]. (Page consultée le 14 avril 2011)

http://www.statcan.gc.ca/pub/13-010-x/2009001/article03-fra.htm

U.S. Census Bureau, (2011) « The X-12-ARIMA Seasonal Adjustment Program» Sur le site du U.S. Census Bureau [En ligne] (Page consultée le 14 avril 2011)

http://www.census.gov/srd/www/x12a

ANNEXE A

xt	mt	wt=xt-mt	bar(wj)	xrépétés	dt=xt-st	t	dt-49à60	Prévisions 49à60
153	187,083333	-34,0833333	17,25	16,0364583	136,963542	1	141,95618	**157,9926383**
189	189,916667	-0,91666667	62,2395833	61,0260417	127,973958	2	141,191	**202,2170417**
221	189,416667	31,5833333	70,6770833	69,4635417	151,536458	3	140,42582	**209,8893617**
215	186,125	28,875	65,625	64,4114583	150,588542	4	139,66064	**204,0720983**
302	181,375	120,625	62,2708333	61,0572917	240,942708	5	138,89546	**199,9527517**
223	175,833333	47,1666667	10,1770833	8,96354167	214,036458	6	138,13028	**147,0938217**
201	176,208333	24,7916667	-18,34375	-19,5572917	220,557292	7	137,3651	**117,8078083**
173	183,25	-10,25	-43,2395833	-44,453125	217,453125	8	136,59992	**92,146795**
121	188,583333	-67,5833333	-67,1770833	-68,390625	189,390625	9	135,83474	**67,444115**
106	190,958333	-84,9583333	-60,59375	-61,8072917	167,807292	10	135,06956	**73,26226833**
86	186,166667	-100,166667	-60,6666667	-61,8802083	147,880208	11	134,30438	**72,42417167**
87	176,458333	-89,4583333	-23,65625	-24,8697917	111,869792	12	179,45	**154,5802083**
228	168,625	59,375		16,0364583	211,963542	13		
283	161,166667	121,833333		61,0260417	221,973958	14		
255	155,625	99,375		69,4635417	185,536458	15		
238	153,208333	84,7916667		64,4114583	173,588542	16		
164	155,208333	8,79166667		61,0572917	102,942708	17		
128	162,375	-34,375		8,96354167	119,036458	18		
108	165,375	-57,375		-19,5572917	127,557292	19		
87	160,25	-73,25		-44,453125	131,453125	20		
74	158,375	-84,375		-68,390625	142,390625	21		
95	159,166667	-64,1666667		-61,8072917	156,807292	22		
145	161,291667	-16,2916667		-61,8802083	206,880208	23		
200	166	34		-24,8697917	224,869792	24		
187	168,291667	18,7083333		16,0364583	170,963542	25		
201	168,5	32,5		61,0260417	139,973958	26		
292	167,875	124,125		69,4635417	222,536458	27		
220	166,708333	53,2916667		64,4114583	155,588542	28		
233	162,958333	70,0416667		61,0572917	171,942708	29		
172	156,291667	15,7083333		8,96354167	163,036458	30		
119	151,875	-32,875		-19,5572917	138,557292	31		
81	152,916667	-71,9166667		-44,453125	125,453125	32		
65	149,916667	-84,9166667		-68,390625	133,390625	33		
76	146,333333	-70,3333333		-61,8072917	137,807292	34		
74	146,208333	-72,2083333		-61,8802083	135,880208	35		
111	144,541667	-33,5416667		-24,8697917	135,869792	36		
170	145	25		16,0364583	153,963542	37		
243	147,458333	95,5416667		61,0260417	181,973958	38		
178	150,375	27,625		69,4635417	108,536458	39		
248	152,458333	95,5416667		64,4114583	183,588542	40		
202	152,375	49,625		61,0572917	140,942708	41		
163	150,791667	12,2083333		8,96354167	154,036458	42		
139	146,916667	-7,91666667		-19,5572917	158,557292	43		
120	137,541667	-17,5416667		-44,453125	164,453125	44		
96	127,833333	-31,8333333		-68,390625	164,390625	45		
95	117,916667	-22,9166667		-61,8072917	156,807292	46		
53	107	-54		-61,8802083	114,880208	47		
94	99,625	-5,625		-24,8697917	118,869792	48		

ANNEXE B

Candidat 1 : ARIMA (1,1,0)

Conditional Least Squares Estimation

Paramètre	Valeur estimée	Erreur type	Valeur du test t	Approx. de Pr > |t|	Retard
MU	-0.37694	0.40165	-0.94	0.3570	0
AR1,1	-0.56647	0.16664	-3.40	0.0023	1

Constant Estimate	-0.59047
Variance Estimate	10.45068
Std Error Estimate	3.232751
AIC	141.9047
SBC	144.4964
Number of Residuals	27

* AIC and SBC do not include log determinant.

Conditional Least Squares Estimation

Paramètre	Valeur estimée	Erreur type	Valeur du test t	Approx. de Pr > |t|	Retard
AR1,1	-0.54856	0.16520	-3.32	0.0027	1

Variance Estimate	10.39814
Std Error Estimate	3.224615
AIC	140.8276
SBC	142.1235
Number of Residuals	27

* AIC and SBC do not include log determinant.

Autocorrelation Check of Residuals

To Lag	Chi-Square	DF	Pr > Khi-2	Autocorrélations					
6	2.97	5	0.7042	-0.089	-0.214	0.012	0.106	-0.156	0.040
12	4.05	11	0.9684	0.021	0.009	0.087	-0.101	-0.029	0.064
18	7.57	17	0.9749	-0.084	-0.013	0.164	0.118	-0.063	-0.026
24	16.92	23	0.8131	0.205	0.015	-0.166	0.003	0.079	-0.039

Autocorrelation Plot of Residuals

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	10.398143	1.00000	|********************|	0
1	-0.922955	-.08876	| . **| . |	0.192450
2	-2.220606	-.21356	| . ****| . |	0.193960
3	0.129593	0.01246	| . | . |	0.203483
4	1.101143	0.10590	| . |** . |	0.202520
5	-1.617390	-.15555	| . ***| . |	0.204552
6	0.414705	0.03988	| . |* . |	0.206896

"." marks two standard errors

Le Système SAS 16:44 Wednesday, April 13, 2011 95

The ARIMA Procedure

Inverse Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	0.10490	| . |** . |
2	0.20651	| . |**** . |
3	0.07366	| . |* . |
4	-0.04474	| . *| . |
5	0.12662	| . |*** . |
6	-0.04323	| . *| . |

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	-0.08876	| . **| . |
2	-0.22329	| . ****| . |
3	-0.03239	| . *| . |
4	0.05952	| . |* . |
5	-0.15035	| . ***| . |
6	0.04686	| . |* . |

Candidat 2 : ARIMA (0,1,1)

Conditional Least Squares Estimation

Paramètre	Valeur estimée	Erreur type	Valeur du test t	Pr. Approx. > |t|	Retard
MU	-0.43887	0.07049	-6.23	<.0001	0
MA1,1	1.00000	0.09637	10.38	<.0001	1

Constant Estimate	-0.43887
Variance Estimate	8.215178
Std Error Estimate	2.866213
AIC	135.4063
SBC	137.998
Number of Residuals	27

* AIC and SBC do not include log determinant.

Autocorrelation Check of Residuals

To Lag	Chi-Square	DF	Pr > Khi-2	Autocorrélations					
6	2.54	5	0.7710	0.050	0.171	-0.025	0.117	-0.167	0.043
12	4.52	11	0.9522	-0.005	-0.006	-0.002	-0.147	-0.120	-0.073
18	6.72	17	0.9869	-0.147	-0.018	0.084	0.074	0.006	-0.036
24	10.76	23	0.9855	0.075	-0.056	-0.120	-0.005	0.068	-0.037

Autocorrelation Plot of Residuals

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	8.215178	1.00000	|**********|**********|	0
1	0.406377	0.04952	| . |* . |	0.192450
2	1.404694	0.17099	| . |*** . |	0.192921
3	-0.207569	-.02527	| . *| . |	0.198455
4	0.961275	0.11701	| . |** . |	0.198574
5	-1.370340	-.16681	| . ***| . |	0.201111
6	0.355356	0.04326	| . |* . |	0.206172

"." marks two standard errors

Inverse Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	-0.08457	| . **| . |
2	-0.11649	| . **| . |
3	0.00076	| . | . |
4	-0.10059	| . **| . |
5	0.16589	| . |*** . |
6	-0.02813	| . *| . |

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	0.04952	| . |* . |
2	0.16895	| . |*** . |
3	-0.04197	| . *| . |
4	0.09420	| . |** . |
5	-0.17284	| . ***| . |
6	0.03004	| . |* . |

Candidat 3 : ARIMA (1,1,1)

Conditional Least Squares Estimation

Paramètre	Valeur estimée	Erreur type	Valeur du test t	Approx. de Pr > |t|	Retard
MU	-0.43775	0.08083	-5.42	<.0001	0
MA1,1	1.00000	0.11177	8.95	<.0001	1
AR1,1	0.06758	0.22176	0.30	0.7632	1

Le Système SAS 12:41 Sunday, April 10, 2011 324

The ARIMA Procedure

Constant Estimate	-0.40816
Variance Estimate	8.535452
Std Error Estimate	2.92155
AIC	137.3367
SBC	141.2242
Number of Residuals	27

* AIC and SBC do not include log determinant.

Autocorrelation Check of Residuals

To Lag	Chi-Square	DF	Pr > Khi-2	Autocorrélations					
6	2.80	4	0.5920	-0.027	0.171	-0.046	0.129	-0.178	0.052
12	4.35	10	0.9301	-0.009	-0.006	0.013	-0.139	-0.102	-0.053
18	6.37	16	0.9836	-0.138	-0.017	0.085	0.067	0.004	-0.040
24	10.43	22	0.9819	0.082	-0.054	-0.120	-0.005	0.067	-0.034

Autocorrelation Plot of Residuals

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	8.535452	1.00000	|********************|	0
1	-0.227794	-.02669	| . *| . |	0.192450
2	1.461498	0.17123	| . |*** . |	0.192587
3	-0.390398	-.04574	| . *| . |	0.198145
4	1.103807	0.12932	| . |*** . |	0.198536
5	-1.515320	-.17753	| . ****| . |	0.201632
6	0.441752	0.05176	| . |* . |	0.207340

"." marks two standard errors

Inverse Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	-0.02603	| . *| . |
2	-0.12074	| . **| . |
3	-0.01160	| . | . |
4	-0.09191	| . **| . |
5	0.15777	| . |*** . |
6	-0.01187	| . | . |

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	-0.02669	| . *| . |
2	0.17064	| . |*** . |
3	-0.03854	| . *| . |
4	0.10169	| . |** . |
5	-0.16690	| . ***| . |
6	0.01257	| . | . |

ANNEXE C

Candidat 1 : ARIMA (1,0,0)

Conditional Least Squares Estimation

Paramètre	Valeur estimée	Erreur type	Valeur du test t	Approx. de Pr > |t|	Retard
MU	16.34102	1.86824	8.75	<.0001	0
AR1,1	0.44108	0.17678	2.50	0.0193	1

Constant Estimate	9.133283
Variance Estimate	32.69589
Std Error Estimate	5.718032
AIC	179.0285
SBC	181.6929
Number of Residuals	28

* AIC and SBC do not include log determinant.

Autocorrelation Check of Residuals

To Lag	Chi-Square	DF	Pr > Khi-2	Autocorrélations					
6	2.59	5	0.7633	-0.034	-0.043	0.199	0.092	0.152	-0.027
12	9.02	11	0.6198	0.160	-0.056	0.068	0.196	-0.093	-0.233
18	13.81	17	0.6806	-0.024	0.056	-0.198	-0.124	-0.097	-0.055
24	25.51	23	0.3246	-0.080	0.021	-0.064	-0.248	0.097	0.042

The ARIMA Procedure

Autocorrelation Plot of Residuals

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	32.695889	1.00000	|********************|	0
1	-1.119067	-.03423	| . *| . |	0.188982
2	-1.390786	-.04254	| . *| . |	0.189203
3	6.490482	0.19851	| . |**** . |	0.189545
4	3.021571	0.09241	| . |** . |	0.196830
5	4.964861	0.15185	| . |*** . |	0.198379
6	-0.871296	-.02665	| . *| . |	0.202482
7	5.229303	0.15994	| . |*** . |	0.202607

"." marks two standard errors

Inverse Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	0.05498	| . |* . |
2	0.12125	| . |** . |
3	-0.19196	| . ****| . |
4	-0.05467	| . *| . |
5	-0.18140	| . ****| . |
6	0.02374	| . | . |
7	-0.12966	| . ***| . |

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	-0.03423	| . *| . |
2	-0.04376	| . *| . |
3	0.19610	| . |**** . |
4	0.10740	| . |** . |
5	0.18465	| . |**** . |
6	-0.04297	| . *| . |
7	0.14465	| . |*** . |

Candidat 2 : ARIMA (0,0,1)

Conditional Least Squares Estimation

Paramètre	Valeur estimée	Erreur type	Valeur du test t	Approx. de Pr > |t|	Retard
MU	16.07922	1.55764	10.32	<.0001	0
MA1,1	-0.44437	0.18310	-2.43	0.0225	1

Constant Estimate	16.07922
Variance Estimate	33.43701
Std Error Estimate	5.782475
AIC	179.6561
SBC	182.3205
Number of Residuals	28

* AIC and SBC do not include log determinant.

Autocorrelation Plot of Residuals

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	33.437015	1.00000	|********************|	0
1	0.562730	0.01683	| . | . |	0.188982
2	4.697591	0.14049	| . |*** . |	0.189036
3	7.550765	0.22582	| . |***** . |	0.192729
4	3.814329	0.11408	| . |** . |	0.201958
5	7.412294	0.22169	| . |**** . |	0.204246
6	-1.191005	-.03562	| . *| . |	0.212665
7	7.401857	0.22137	| . |**** . |	0.212878

"." marks two standard errors

Inverse Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	0.03396	| . |* . |
2	-0.01934	| . | . |
3	-0.19039	| . ****| . |
4	-0.06046	| . *| . |
5	-0.13217	| . ***| . |
6	0.08967	| . |** . |
7	-0.12649	| . ***| . |

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	0.01683	| . | . |
2	0.14025	| . |*** . |
3	0.22593	| . |***** . |
4	0.10206	| . |** . |
5	0.17906	| . |**** . |
6	-0.10870	| . **| . |
7	0.14083	| . |*** . |

Candidat 3 : ARIMA (1,0,1)

Conditional Least Squares Estimation

Paramètre	Valeur estimée	Erreur type	Valeur du test t	Approx. de Pr > |t|	Retard
MU	19.66188	3.67468	5.35	<.0001	0
MA1,1	0.75155	0.24808	3.03	0.0056	1
AR1,1	1.00000	0.10807	9.25	<.0001	1

Constant Estimate	2.3E-6
Variance Estimate	30.66779
Std Error Estimate	5.537851
AIC	178.1373
SBC	182.1339
Number of Residuals	28

* AIC and SBC do not include log determinant.

Autocorrelation Plot of Residuals

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	30.667792	1.00000	|**********|	0
1	1.278020	0.04167	| . *. |	0.188982
2	-3.821690	-.12462	| . **. |	0.189310
3	2.251053	0.07340	| . *. |	0.192217
4	-0.124790	-.00407	| . . |	0.193216
5	1.988597	0.06484	| . *. |	0.193219
6	-0.458714	-.01496	| . . |	0.193995
7	2.992100	0.09756	| . **. |	0.194036

"." marks two standard errors

Inverse Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	-0.11244	| . **| . |
2	0.18577	| . |**** . |
3	-0.13112	| . ***| . |
4	0.07576	| . |** . |
5	-0.12403	| . **| . |
6	0.05004	| . |* . |
7	-0.11969	| . **| . |

The ARIMA Procedure

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	0.04167	| . |* . |
2	-0.12657	| . ***| . |
3	0.08605	| . |** . |
4	-0.02890	| . *| . |
5	0.08940	| . |** . |
6	-0.03615	| . *| . |
7	0.12859	| . |*** . |

ANNEXE D

Candidat 1 : ARIMA (1,1,0)

Conditional Least Squares Estimation

Paramètre	Valeur estimée	Erreur type	Valeur du test t	Approx. de Pr > |t|	Retard
MU	-0.08354	0.02843	-2.94	0.0070	0
AR1,1	-0.47653	0.17585	-2.71	0.0120	1

Constant Estimate	-0.12335
Variance Estimate	0.046613
Std Error Estimate	0.2159
AIC	-4.23397
SBC	-1.6423
Number of Residuals	27

* AIC and SBC do not include log determinant.

Autocorrelation Check of Residuals

To Lag	Chi-Square	DF	Pr > Khi-2	Autocorrélations					
6	7.35	5	0.1956	-0.172	-0.284	0.073	-0.138	0.244	0.165
12	16.02	11	0.1402	-0.290	0.042	0.002	-0.076	0.172	-0.263
18	21.67	17	0.1978	0.099	0.159	-0.205	-0.026	0.006	-0.093
24	22.54	23	0.4879	0.064	0.024	-0.052	0.010	0.012	0.006

Le Système SAS 12:41 Sunday, April 10, 2011 232

The ARIMA Procedure

Autocorrelation Plot of Residuals

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	0.046613	1.00000	|********************|	0
1	-0.0080122	-.17189	| . ***| . |	0.192450
2	-0.013330	-.28383	| . ******| . |	0.198055
3	0.0034004	0.07295	| . |* . |	0.212587
4	-0.0064197	-.13772	| . ***| . |	0.213512
5	0.011351	0.24352	| . |***** . |	0.216777
6	0.0076745	0.16464	| . |*** . |	0.226683

"." marks two standard errors

Inverse Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	0.27318	| . |***** . |
2	0.29859	| . |****** . |
3	-0.02267	| . | . |
4	0.01154	| . | . |
5	-0.21511	| . ****| . |
6	-0.13710	| . ***| . |

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	-0.17189	| . ***| . |
2	-0.32292	| . ******| . |
3	-0.05418	| . *| . |
4	-0.25882	| . *****| . |
5	0.19884	| . |**** . |
6	0.17774	| . |**** . |

Candidat 2 : ARIMA (0,1,1)

Conditional Least Squares Estimation

Paramètre	Valeur estimée	Erreur type	Valeur du test t	Approx. de Pr > |t|	Retard
MU	-0.08039	0.01378	-5.83	<.0001	0
MA1,1	0.66393	0.14960	4.44	0.0002	1

Constant Estimate	-0.08039
Variance Estimate	0.040013
Std Error Estimate	0.200033
AIC	-8.35592
SBC	-5.76425
Number of Residuals	27

* AIC and SBC do not include log determinant.

Autocorrelation Check of Residuals

To Lag	Chi-Square	DF	Pr > Khi-2	Autocorrélations					
6	3.54	5	0.6180	-0.051	-0.030	0.110	-0.068	0.228	0.166
12	11.21	11	0.4260	-0.221	0.014	-0.025	-0.136	0.139	-0.274
18	16.51	17	0.4881	0.062	0.060	-0.209	-0.086	-0.061	-0.118
24	17.07	23	0.8057	0.035	0.007	-0.056	0.001	0.005	0.011

The ARIMA Procedure

Autocorrelation Plot of Residuals

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	0.040013	1.00000	|********************|	0
1	-0.0020461	-.05114	. *| . .	0.192450
2	-0.0011978	-.02994	. *| . .	0.192953
3	0.0044090	0.11019	. |** . .	0.193125
4	-0.0027326	-.06829	. *| . .	0.195439
5	0.0091066	0.22759	. |***** . .	0.196321
6	0.0066517	0.16624	. |*** . .	0.205861

"." marks two standard errors

Inverse Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	0.09051	. |** .
2	0.04857	. |* .
3	-0.07437	. *| .
4	0.01334	. | .
5	-0.22638	. *****| .
6	-0.16310	. ***| .

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	-0.05114	. *| .
2	-0.03364	. *| .
3	0.10733	. |** .
4	-0.05907	. *| .
5	0.23333	. |***** .
6	0.19035	. |**** .

Candidat 3 : ARIMA (1,1,1)

Conditional Least Squares Estimation

Paramètre	Valeur estimée	Erreur type	Valeur du test t	Pr. Approx. > |t|	Retard
MU	-0.08070	0.01449	-5.57	<.0001	0
MA1,1	0.61366	0.23963	2.56	0.0172	1
AR1,1	-0.10150	0.30176	-0.34	0.7395	1

Constant Estimate	-0.08889
Variance Estimate	0.041461
Std Error Estimate	0.203619
AIC	-6.4988
SBC	-2.61129
Number of Residuals	27

* AIC and SBC do not include log determinant.

Autocorrelation Check of Residuals

To Lag	Chi-Square	DF	Pr > Khi-2	Autocorrélations					
6	7.35	5	0.1956	-0.172	-0.284	0.073	-0.138	0.244	0.165
12	16.02	11	0.1402	-0.290	0.042	0.002	-0.076	0.172	-0.263
18	21.67	17	0.1978	0.099	0.159	-0.205	-0.026	0.006	-0.093
24	22.54	23	0.4879	0.064	0.024	-0.052	0.010	0.012	0.006

The ARIMA Procedure

Autocorrelation Plot of Residuals

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	0.046613	1.00000	|********************|	0
1	-0.0080122	-.17189	***| .	0.192450
2	-0.013230	-.28383	******| .	0.198055
3	0.0034004	0.07295	|* .	0.212587
4	-0.0064197	-.13772	***| .	0.213512
5	0.011351	0.24352	|***** .	0.216777
6	0.0076745	0.16464	|*** .	0.226683

"." marks two standard errors

Inverse Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	0.27318	|***** . |
2	0.29859	|****** . |
3	-0.02367	| . |
4	0.01154	| . |
5	-0.21511	| ****| . |
6	-0.13710	| ***| . |

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	-0.17189	| ***| . |
2	-0.32292	| ******| . |
3	-0.05418	| *| . |
4	-0.25882	| *****| . |
5	0.19884	| . |**** . |
6	0.17774	| . |**** . |