REVUE DE STATISTIQUE APPLIQUÉE

LAURENT FERRARA
DOMINIQUE GUEGAN

Analyse d'intervention et prévisions. Problématique et application à des données de la RATP

Revue de statistique appliquée, tome 48, n° 2 (2000), p. 55-72
<http://www.numdam.org/item?id=RSA_2000__48_2_55_0>

Rev. Statistique Appliquée, 2000, **XLVIII** (2), 55-72

ANALYSE D'INTERVENTION ET PRÉVISIONS. PROBLÉMATIQUE ET APPLICATION À DES DONNÉES DE LA RATP

Laurent Ferrara*, Dominique Guegan**

* *Université de Paris XIII, CNRS UMR 7539 - RATP, Département Commercial, LAC A73, 54 Quai de la Rapée, 75599 Paris Cedex 12, e-mail : Laurent.Ferrara@ratp.fr*

** *Université de Reims, UPESA 6059 - ENSAE-CREST, Timbre J340, 3 Avenue Pierre Larousse, 92245 Malakoff Cedex (France), e-mail : guegan@ensae.fr*

RÉSUMÉ

Les séries chronologiques de trafic ou de ventes de titres de la RATP sont souvent perturbées par des événements spéciaux. Lors de la modélisation d'une série, l'analyse d'intervention (Box et Tiao (1975)) prend en compte ces interventions extérieures et fournit une mesure de l'impact de celles-ci sur la série. Nous analysons les effets de cette technique d'intervention pour différentes fonctions et nous appliquons cette approche sur des séries réelles.

Mots-clés : *Série chronologique, analyse d'intervention, mesure d'impact.*

ABSTRACT

RATP's time series of traffic or ticket sales are often affected by particular events. When modelling a time series, the intervention analysis (Box and Tiao (1975)) allows to take into account the various external interventions and to give a measure of their impact on the series. We analyze the effects of this intervention technique for various functions and we apply this approach to real series.

Keywords : *Time series, intervention analysis, measure of impact.*

1. Introduction

Lorsqu'on essaie de modéliser des séries chronologiques à caractère économique, on est amené à tenir compte d'événements de nature diverse, extérieurs au modèle, qui viennent perturber les séries. L'effet de ces événements se fait sentir soit

par la présence d'un ou plusieurs points dits aberrants (*outliers*), qui occasionnent une rupture ponctuelle dans la série, soit par un changement sensible dans l'évolution de la série.

Dans le cas des séries de ventes de titres et de trafic de la RATP, ces événements peuvent prendre pour forme aussi bien une grève du personnel dont l'effet se répercute ponctuellement sur la série de trafic, qu'une promotion publicitaire pour un certain type de titre de transport, dont l'impact se répercute pendant plusieurs mois sur la série des ventes de ce produit. De même, des modifications d'infrastructure ou certaines mesures prises par la Régie (mesures anti-pollution, anti-fraude, ...) ont un impact, plus ou moins durable dans le temps, sur certaines séries de ventes ou de trafic.

Prendre en compte l'effet d'un événement extérieur sur une série permet d'améliorer la modélisation de cette série, mais surtout de mesurer l'impact de cet événement, ce qui peut être intéressant pour quantifier l'effet d'une mesure ou d'une promotion publicitaire. Cette mesure de l'impact pourra également être utilisée pour établir des séries corrigées des valeurs aberrantes, avec lesquelles il est souvent préférable de travailler. Le but étant bien sûr d'améliorer la modélisation d'une série pour pouvoir fournir de meilleures prévisions sur cette série.

Box et Tiao (1975) sont à l'origine de la théorie dite de «*l'analyse d'intervention*» qui permet de prendre en compte différentes interventions extérieures au modèle lors de la modélisation d'une série chronologique. L'apport de l'analyse d'intervention à une modélisation de série de type, par exemple, SARIMA (Box et Jenkins (1970)), se situe au niveau de l'information disponible au praticien pour modéliser cette série. En effet, l'approche de Box et Jenkins (1970) utilise simplement l'information quantitative contenue dans les données, alors que l'analyse d'intervention permet d'ajouter de manière additive une information de type qualitatif, par le biais de variables binaires exogènes.

De nombreux auteurs se sont intéressés par la suite à l'analyse d'intervention et ses applications et ont complété les travaux de Box et Tiao; nous citerons en particulier Bhattacharyya et Layton (1979), Tiao (1985), Tsay (1986), Brockwell et Davis (1987), Chang *et al.* (1988) et Box, Jenkins et Reinsel (1994).

Le but de notre étude est de mettre en évidence l'intérêt de la théorie de l'analyse d'intervention pour la modélisation des séries de trafic et de ventes de titres de la RATP. Pour cela nous précisons les différents types d'intervention que l'on peut envisager, nous les illustrons à partir de simulations et d'exemples d'applications possibles, puis nous traitons en détail une série de données réelles.

Cet article se décompose donc en trois parties; nous présentons dans le premier paragraphe les bases de la théorie de l'analyse d'intervention. Puis, dans le second paragraphe, nous nous intéressons à certains types de modèles d'intervention que nous développons et dont nous montrons l'intérêt à l'aide d'exemples réels et de simulations. Enfin nous donnerons dans le dernier paragraphe un exemple d'application : comment utiliser l'analyse d'intervention pour mesurer l'impact des mesures anti-pollution et d'une grève des agents sur le trafic du Métro pendant le mois d'octobre 1997.

2. Le Modèle

La théorie de l'analyse d'intervention développée par Box et Tiao (1975) permet de prendre en compte, lors de la modélisation SARIMA d'une série chronologique, des interventions extérieures au modèle. On apporte ainsi au modèle statistique une information supplémentaire de type qualitatif, qui est intégrée de manière additive au modèle à l'aide de variables déterministes exogènes de type binaire. On espère ainsi fournir une «*meilleure*» modélisation en terme d'ajustement du modèle aux données, grâce à l'utilisation d'un ensemble informationnel plus grand.

Nous allons suivre ici l'approche développée par Box et Jenkins (1970) pour la modélisation SARIMA d'une série. On rappelle qu'une suite de variables aléatoires $(N_t)_{t\in z}$ est modélisée par un processus SARIMA $(p\ d\ q)(P\ D\ Q)_s$, si elle vérifie le système suivant :

$$(1-B)^d(1-B^S)^D\phi(B)\Phi(B^S)N_t = C + \theta(B)\Theta(B^S)\varepsilon, \qquad (2.1)$$

où B est l'opérateur défini sur les variables aléatoires de la manière suivante : $B(X_t) = X_{t-1}$ et $B^b(X_t) = X_{t-b}$ pour $b > 1$, où $\phi(B)$ est un polynôme en B de degré p tel que $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p$, où $\Phi(B)$ est un polynôme en B de degré P tel que $\Phi(B) = 1 - \Phi_1 B - \Phi_2 B^2 - \ldots - \Phi_P B^P$, où $\theta(B)$ est un polynôme en B de degré q tel que $\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + ... + \theta_q B^q$, où $\Theta(B)$ est un polynôme en B de degré Q tel que $\Theta(B) = 1 + \Theta_1 B + \Theta_2 B^2 + \ldots + \Theta_Q B^Q$, où $(\varepsilon_t)_{t\in Z}$ est un processus bruit blanc gaussien de variance σ^2, S représente la saisonnalité de la série, C est une constante et $(d, D) \in I\!N^2$.

Les différences d'ordre d et les différences saisonnières d'ordre D doivent permettre de rendre la suite de variables aléatoires faiblement stationnaire.

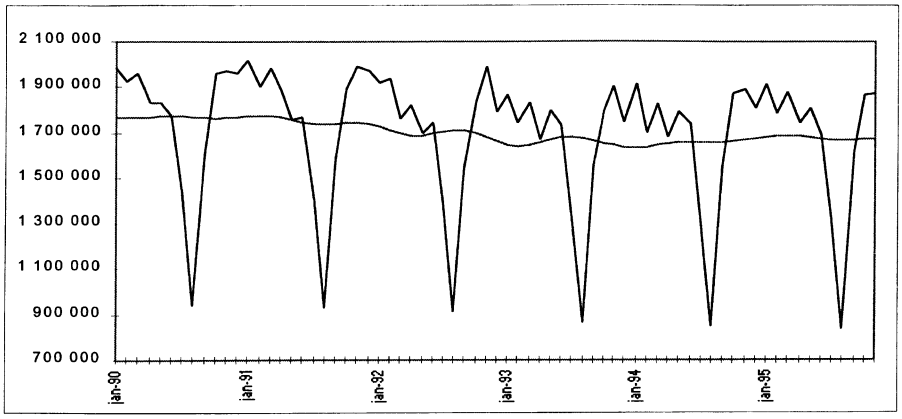
Pour illustrer une série que l'on peut ajuster par un modèle SARIMA, nous donnons (graphique 1) l'exemple de la série des ventes mensuelles de Cartes Orange Mensuelles (COM), de janvier 1990 à novembre 1995 sur l'ensemble du réseau RATP. Une étude effectuée sur cette série a permis de la modéliser par un processus SARIMA (2 0 1)(0 1 1), avec une saisonnalité S de 12 mois.

Nous allons maintenant faire une présentation théorique du modèle d'intervention de Box et Tiao (1975), et nous présentons au paragraphe suivant des simulations et des exemples réels originaux.

On note $(X_t)_{t\in Z}$ la suite de variables aléatoires à modéliser, perturbée par une intervention extérieure. Le modèle d'intervention proposé par Box et Tiao se présente alors sous la forme suivante :

$$X_t = C + \frac{\omega(B)B^b}{\delta(B)}\xi_t + N_t \qquad (2.2)$$

où $(N_t)_{t\in Z}$ est supposé suivre un modèle SARIMA sans constante de la forme (2.1), $\omega(B)$ est un polynôme en B de degré s tel que $\omega(B) = 1 - \omega_1 B - \omega_2 B^2 - \ldots - \omega_s B^s$, $\delta(B)$ est un polynôme en B de degré r tel que $\delta(B) = 1 - \delta_1 B - \delta_2 B^2 - \ldots - \delta_r B^r$ et b est un entier qui représente un retard à déterminer.



GRAPHIQUE 1
Evolution mensuelle des ventes de COM de janvier 1990 à novembre 1995.

La fonction déterministe $\delta^{-1}(B)\omega(B)B^b\xi_t$, représente l'effet de l'intervention qui vient s'ajouter de manière additive au bruit $(N_t)_{t\in Z}$, elle est appelée *fonction d'intervention.*

Dans l'équation (2.2), la suite de variables aléatoires $(\xi_t)_{t\in Z}$ représente l'effet d'une intervention extérieure à la date t', mis sous la forme d'une variable déterministe qui prend pour valeur 1 ou 0 suivant la présence ou l'absence de l'intervention. Cette variable est en général modélisée par deux classes de fonctions :

- une fonction en forme de saut

$$\begin{aligned} \xi_t = S_t^{(t')} &= 0 \qquad \text{si } t < t' \\ &= 1 \qquad \text{si } t \geqslant t' \end{aligned}$$

- une fonction en forme d'impulsion :

$$\begin{aligned} \xi_t = P_t^{(t')} &= 0 \qquad \text{si } t \neq t' \\ &= 1 \qquad \text{si } t = t' \end{aligned}$$

On remarque cependant que grâce à l'égalité suivante : $(1-B)S_t^{(t')} = P_t^{(t')}$, on peut toujours passer d'un saut à une impulsion.

Plus généralement, la série chronologique peut être perturbée par k interventions de natures différentes. Avec les notations précédentes, le modèle d'intervention (2.2) a alors une représentation plus générale donnée par :

$$X_t = C + \sum_{j=1}^{k} \frac{\omega_j(B)B^{b_j}}{\delta_j(B)}\xi_t^{(T_j)} + N_t \tag{2.3}$$

où, pour $j = 1, \ldots k, \omega_j(B)$ est un polynôme en B de degré l_j, $\delta_j(B)$ est un polynôme en B de degré r_j et b_j est un entier qui représente un retard à déterminer.

Une hypothèse fondamentale lors de l'utilisation de l'analyse d'intervention est que la structure du modèle, par exemple SARIMA, soit la même avant et après l'intervention.

Ainsi, après avoir déterminé la date d'intervention (problème sur lequel nous revenons à la fin du paragraphe 3), on fixe alors les deux sous-ensembles de données correspondant à l'évolution du processus avant et après l'intervention. On ajuste ensuite le même modèle sur chacun de ces deux sous-ensembles. Dans notre cadre, comme nous nous intéressons aux processus linéaires, nous chercherons à ajuster un processus SARIMA à l'aide des outils classiques que sont les fonctions d'autocorrélation et d'autocorrélation partielle.

En ce qui concerne la forme de la fonction d'intervention, il n'existe pas de méthode automatique fiable permettant de la déterminer. Cependant Box et Tiao (1975) ont proposé différents types de fonctions permettant de s'adapter à la forme graphique que prend la série, suite à l'effet de l'intervention extérieure, d'où l'importance d'une analyse graphique ou géométrique de la série à étudier. Cette analyse graphique nécessite donc une approche locale de la série qui s'éloigne de l'analyse souvent globale utilisée quand on fait une modélisation paramétrique d'un processus. Dans le paragraphe suivant, nous allons nous intéresser à quelques types de fonctions d'intervention que l'on rencontre en pratique. Une fois le modèle d'intervention correctement spécifié, l'estimation des paramètres du modèle se fait alors par la méthode des moindres carrés non linéaire qui nécessite l'utilisation d'une des méthodes de gradient, qui sont des algorithmes itératifs de minimisation.

Pour effectuer nos calculs nous avons utilisé le logiciel RATS 4.3 qui utilise l'algorithme de Gauss-Newton. De plus ce logiciel permet l'estimation simultanée de l'ensemble des paramètres du modèle d'intervention.

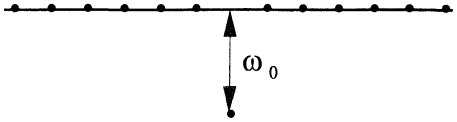
On se propose de présenter dans le paragraphe suivant trois types classiques de fonctions d'intervention que nous illustrerons. Nous tr aitons en détail un exemple de série au paragraphe 4.

3. Exemples et Simulations

Nous avons souligné dans le paragraphe 2 que le choix *a priori* d'une fonction d'intervention est assez subjectif car il dépend de la forme graphique et des propriétés géométriques de la série. Cependant, il existe plusieurs fonctions d'intervention «type», qui permettent de modéliser la plupart des phénomènes d'intervention que nous avons rencontrés dans les séries chronologiques de trafic ou de vente de titres de transport de la RATP. Dans ce paragraphe, nous allons présenter trois types de modèles d'intervention différents, utilisant chacun une fonction d'impulsion comme variable exogène déterministe. Afin de mieux visualiser le phénomène, nous avons illustré chaque type d'intervention d'un schéma local et d'une simulation de série. Pour souligner l'intérêt de chacun de ces trois modèles d'intervention, nous fournissons pour chacun des cas un exemple original de série sur lequel il semble intéressant d'ajuster le modèle d'intervention concerné.

3.1. Modèle avec intervention ponctuelle

Supposons que l'on observe $X_1, \ldots, X_n, \ldots$ une trajectoire. On suppose qu'elle est la réalisation d'un processus SARIMA $(X_t)_{t\in Z}$. Supposons d'autre part qu'à la date unique $t = t'$ l'évolution de cette trajectoire soit perturbée par une intervention extérieure, c'est l'exemple de la présence d'un point aberrant dans la série, que l'on a schématisé par le graphique 2.



GRAPHIQUE 2
Représentation locale de l'intervention.

Le modèle d'intervention s'écrit alors sous la forme suivante :

$$X_t = C + \omega_0 P_t^{t'} + N_t \tag{3.1}$$

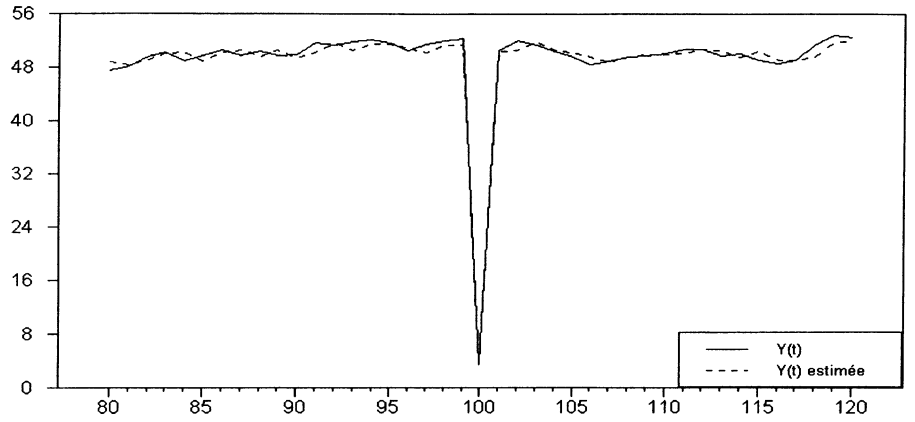
où $(N_t)_{t\in Z}$ suit un processus SARIMA défini par (2.1) sans constante et où $P_t^{t'}$ est une impulsion au temps $t = t'$. Le paramètre ω_0 représente l'impact de l'intervention extérieure sur la série. Son estimation permet de donner une mesure de cet impact. Dans la littérature anglo-saxonne ce modèle est appelé de type AO (*additive outlier*) (Box, Jenkins et Reinsel (1994)).

Nous examinons à l'aide d'une simulation l'influence d'un tel impact. Pour cela, on a simulé une série $(Y_t)_{t\in Z}$ générée par le système suivant;

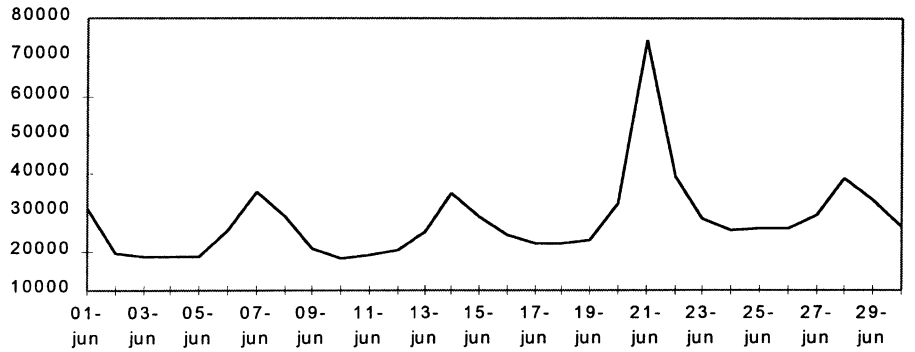
$$Y_t = C + \omega_0 P_t^{100} + \frac{(1+\theta_1 B)}{(1-\phi_1 B)}\varepsilon_t, \quad t = 1, \ldots 500, \quad (*)$$

où $C = 50$, $\omega_0 = -49$, et où $P_t^{100} = 1$ si $t = 100$ et 0 sinon. La série $(\varepsilon_t)_{t\in Z}$ des perturbations aléatoires est obtenue en effectuant un tirage aléatoire parmi les réalisations d'une variable gaussienne de moyenne nulle et de variance égale à 1. La valeur des deux paramètres du processus SARIMA est arbitrairement choisie, mais ces paramètres sont tels que le processus créé soit stationnaire ($|\phi_1| < 1$) et inversible ($|\theta_1| < 1$). On choisit donc $\phi_1 = \theta_1 = 0.5$. La série obtenue est représentée sur le graphique 3.

Considérons maintenant à titre d'exemple une série observée sur laquelle ce type de modélisation peut s'appliquer. Le graphique 4 représente la série du trafic journalier sur le Métro pour les voyageurs possédant un titre de transport «Paris-Visite», pour le mois de juin 1997. Ce titre de transport, destiné principalement aux touristes visitant la Capitale, permet d'effectuer plusieurs trajets par jour sur le réseau de la RATP avec le même titre. En raison de l'afflux de touristes en fin de semaine, on observe une périodicité de sept jours avec un pic le samedi. Partout en France a lieu le 21 juin la «Fête de la Musique». A cette occasion la RATP édite un titre spécial



GRAPHIQUE 3
Simulation de la série $Y(t)$.



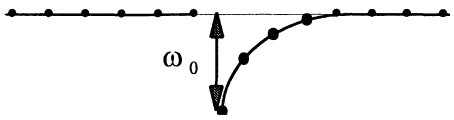
GRAPHIQUE 4
Evolution du trafic journalier pour le titre «Paris-Visite» dans le Métro, pour le mois de juin 1998.

permettant d'effectuer un voyage aller-retour dans la soirée du 21 juin au 22 juin, sur l'ensemble du Métro, du RER et du réseau SNCF. Pour des raisons informatiques, ce titre était codé en juin 1997 comme un titre de transport «Paris-Visite» et ne pouvait donc pas être distingué des autres titres «Paris-Visite» classiques utilisés ce même jour.

Après une étude attentive de la série, on constate qu'elle peut être modélisée à l'aide d'un modèle muni d'une intervention ponctuelle à la date du dimanche 21 juin pour obtenir l'impact sur le trafic du Métro de l'opération sur le titre «Paris-Visite», liée à la Fête de la Musique. La mesure de cet impact sera fournie par la valeur estimée du paramètre ω_0.

3.2. Modèle avec intervention à effet rémanent

Nous supposons maintenant, en utilisant les mêmes notations qu'au début du paragraphe 3.1., que l'intervention au temps $t = t'$ vient perturber l'évolution de la



GRAPHIQUE 5
Représentation locale de l 'intervention.

série avec un effet qui décroît progressivement dans le temps, de façon exponentielle. Cette intervention est schématisée par le graphique 5.

Le modèle d'intervention s'écrit alors sous la forme suivante :

$$X_t = C + \frac{\omega_0}{1 - \lambda B} P_t^{t'} + N_t \tag{3.2}$$

où $(N_t)_{t \in Z}$ suit un processus SARIMA de la forme (2.1) sans constante et où $P_t^{t'}$ est une impulsion au temps $t = t'$. Le paramètre ω_0 représente l'impact de l'intervention extérieure sur la série (son estimation fournira une mesure de cet impact) et λ est un paramètre strictement compris entre 0 et 1 qui mesure la vitesse de décroissance de l'effet de l'intervention. En clair, si λ est proche de 0, cela signifie que l'impact est quasiment ponctuel (on se ramène alors au cas précédent) et si λ est proche de 1, cela signifie que l'impact se prolonge dans le temps. Dans la littérature anglo-saxonne ce modèle est appelé de type TC (*temporary or transient change*) (Box, Jenkins et Reinsel (1994)).

Il est à noter que pour modéliser un choc dont l'effet se répercute dans le temps, le modèle suivant est parfois utilisé :

$$Y_t = C + \frac{\theta(B)}{\phi(B)} \omega_0 P_t^{t'} + N_t$$

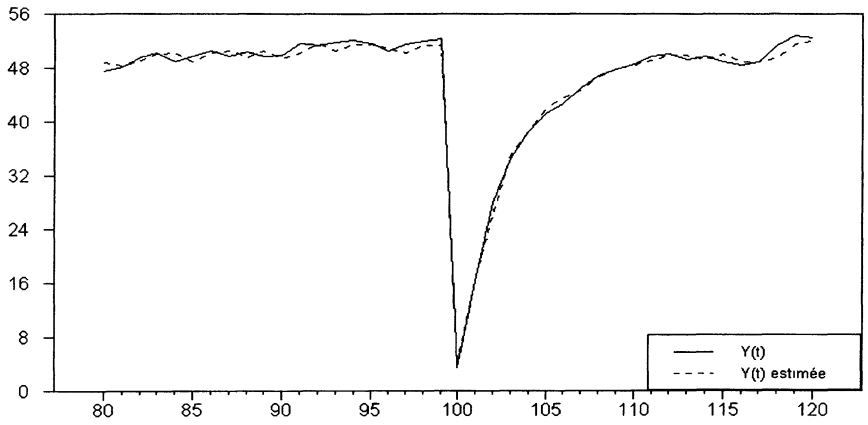
où $(N_t)_{t \in Z}$ suit un processus SARIMA $(p\,0\,q)(0\,0\,0)$ de la forme (2.1) sans constante, où $P_t^{t'}$ est une impulsion au temps $t = t'$ et $\phi(B)$ et $\theta(B)$ sont les polynômes en B de la structure SARIMA. Dans la littérature anglo-saxonne ce modèle est appelé de type IO (*Innovation Outlier*) (Box, Jenkins et Reinsel (1994)).

Nous allons montrer sur une simulation l'effet rémanent de ce type d'intervention décrit en (3.2). Pour cela, on a simulé une série $(Y_t)_{t \in Z}$ générée par le système suivant;

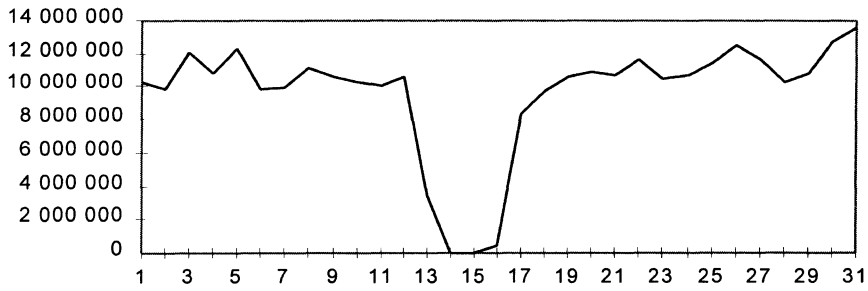
$$Y_t = C + \frac{\omega_0}{1 - \lambda B} P_t^{100} + \frac{(1 + \theta_1 B)}{(1 - \phi_1 B)} \varepsilon_t, \quad t = 1, \ldots, 500.$$

où $C = 50$, $\omega_0 = -49$, $\lambda = 0.7$ et où $P_t^{100} = 1$ si $t = 100$ et 0 sinon. La série $(\varepsilon_t)_{t \in Z}$ des perturbations aléatoires est obtenue en effectuant un tirage aléatoire parmi les réalisations d'une variable gaussienne de moyenne nulle et de variance égale à 1. La valeur des paramètres du processus SARIMA est la même que dans le modèle (*). La série que l'on obtient est représentée sur le graphique 6.

Considérons maintenant à titre d'exemple une série observée sur laquelle ce type de modélisation peut s'appliquer.



GRAPHIQUE 6
Simulation de la série $Y(t)$.



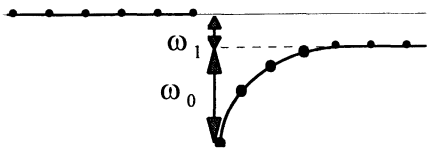
GRAPHIQUE 7
Évolution des ventes hebdomadaires de Billets sur le réseau ferré du 1er septembre 1995 au 4 avril 1996.

La série des ventes de billets sur le réseau ferré a été fortement perturbée, comme l'ensemble des séries de trafic ou de vente de titres de la RATP, par les grèves des agents de la RATP qui ont eu lieu durant le mois de décembre 1995 sur la totalité du réseau. A la fin de la grève, la reprise des ventes de billets s'est alors effectuée de manière progressive. Le graphique 7 représente l'évolution des ventes hebdomadaires de billets sur l'ensemble du réseau ferré entre le 1er septembre 1995 et le 4 avril 1996.

Cette série peut être modélisée à l'aide d'un modèle avec intervention à effet rémanent à partir de la semaine 14. La mesure de l'impact de la grève sur les ventes hebdomadaires de Billets sera fournie par la valeur estimée du paramètre ω_0.

3.3. Modèle avec intervention à effet rémanent et changement de niveau

Nous supposons maintenant, en utilisant les mêmes notations qu'au début du paragraphe 3.1., que l'intervention au temps $t = t'$ vient perturber l'évolution de la série avec un effet rémanent suivi d'un changement de niveau. Cette intervention est schématisée par le graphique 8.

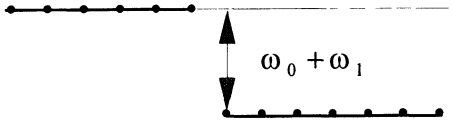


GRAPHIQUE 8
Représentation locale de l 'intervention.

Le modèle d'intervention s'écrit alors sous la forme suivante :

$$X_t = C + \left(\frac{\omega_0}{1-\lambda B} + \frac{\omega_1}{1-B}\right) P_t^{t'} + N_t \tag{3.3}$$

où $(N_t)_{t\in Z}$ suit un processus SARIMA de la forme (2.1) sans constante et où $P_t^{t'}$ est une impulsion au temps $t = t'$. Le paramètre ω_1 représente la différence de niveau consécutive au choc, ω_1 sera donc positif si le niveau observé après le choc est supérieur au niveau observé avant le choc, et négatif sinon. La somme de ω_0 et ω_1 représente l'impact de l'intervention extérieure sur la série, son estimation fournira une mesure de l'impact de l'intervention. De même que précédemment, λ mesure la vitesse de décroissance de l'effet de l'intervention. Dans la littérature anglo-saxonne ce modèle est appelé de type LS (level shift) (Box, Jenkins et Reinsel (1994)). Il se peut cependant que la série soit affectée par un changement de niveau consécutif à un choc, sans pour autant que l'effet rémanent du choc décroisse progressivement. Nous illustrons ce cas par le graphique 9.



GRAPHIQUE 9
Représentation locale de l'intervention.

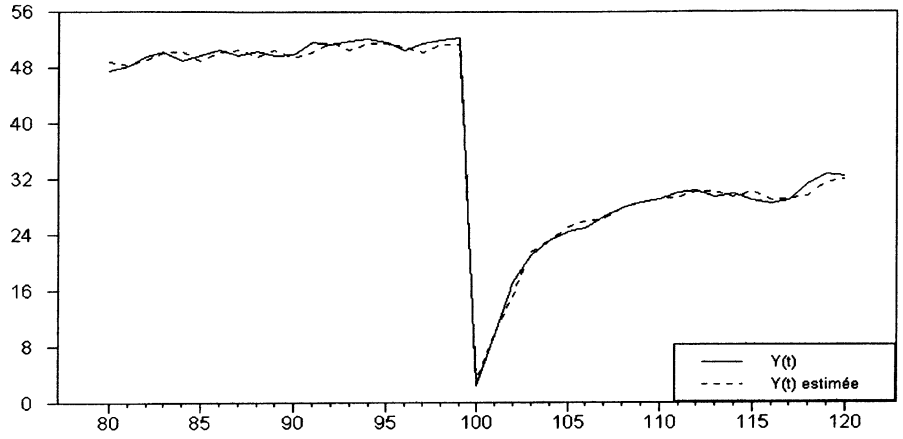
Cette situation, qui est plus spécifique que celle décrite en (3.3), est alors modélisée par l'équation suivante ($\lambda = 1$) :

$$Y_t = C + \left(\frac{\omega_0 + \omega_1}{1-B}\right) P_t^{t'} + N_t. \tag{3.4}$$

Nous allons montrer sur une simulation l'effet de ce type d'intervention décrit en (3.3). Pour cela, on a simulé une série $(Y_t)_{t\in Z}$ générée par le système suivant;

$$Y_t = C + \left(\frac{\omega_0}{1-\lambda B} + \frac{\omega_1}{1-B}\right) P_t^{100} + \frac{(1+\theta_1 B)}{(1-\phi_1 B)}\varepsilon_t, \quad t = 1, \ldots, 500,$$

où $C = 50, \omega_0 = -30, \omega_1 = -20, \lambda = 0.7$ et où $P_t^{100} = 1$ si $t = 100$ et 0 sinon. La série $(\varepsilon_t)_{t\in Z}$ des perturbations aléatoires est obtenue en effectuant un tirage aléatoire



GRAPHIQUE 10
Simulation de la série $Y(t)$.

parmi les réalisations d'une variable gaussienne de moyenne nulle et de variance égale à 1. La valeur des paramètres du processus SARIMA est la même que dans le modèle (*). La série que l'on obtient est représentée sur le graphique 10.

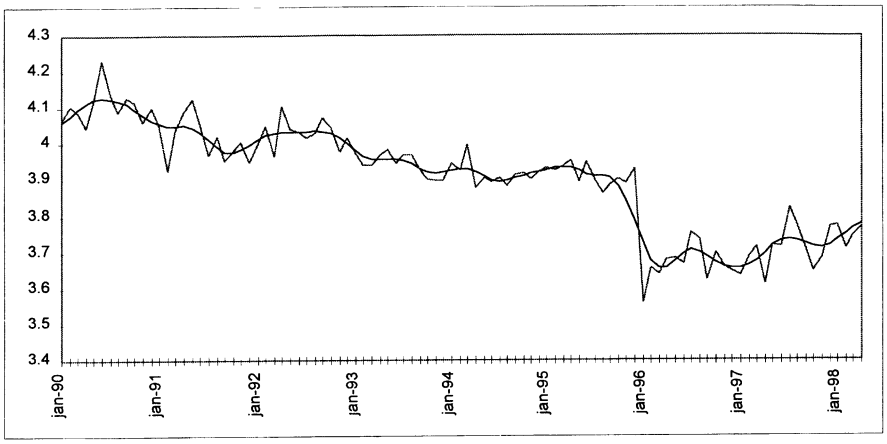
Considérons maintenant à titre d'exemple une série observée sur laquelle cette modélisation peut s'appliquer.

La série mensuelle de trafic par jour ouvrable sur l'ensemble du Métro a été fortement perturbée par les grèves de décembre 1995, mais aussi par les attentats et les alertes à la bombe qui ont eu lieu dans Paris durant les mois de juillet à octobre 1995. Les grèves ont perturbé le trafic de manière ponctuelle au mois de décembre 1995, mais on observe que par la suite, la reprise du trafic s'est effectuée à un niveau inférieur à celui observé avant les grèves. Le graphique 11 représente cette série désaisonnalisée ainsi que sa tendance.

Cette série peut être modélisée à l'aide d'un modèle avec intervention à effet rémanent et changement de niveau à partir du mois décembre 1995. La mesure de l'impact de la grève sur le trafic du Métro pour le mois de décembre 1995 sera fournie par la valeur estimée des paramètres $(\omega_0 + \omega_1)$. La mesure de la baisse de trafic due à ces événements sera fournie par la valeur estimée du paramètre ω_1.

Remarque :

Dans le cadre des exemples que nous avons proposés, les dates des données aberrantes sont connues au début de l'étude. Mais il se peut que dans certains cas, ces dates soient inconnues ou connues de manière imprécise; il faut alors détecter les points aberrants dans la série d'étude. Plusieurs approches ont été développées pour détecter les points aberrants d'une série et caractériser le type d'intervention à utiliser (AO, IO, LS ou TC), à l'aide de procédures itératives. Pour une description approfondie de ces procédures de détection que nous ne considérons pas dans ce papier, voir Denby et Martin (1979), Tiao (1985), Chang *et al.* (1988), Chen et Liu (1990), Box, Jenkins et Reinsel (1994).



GRAPHIQUE 11
Évolution mensuelle du trafic sur le Métro par jour ouvrable du mois de janvier 1990 au mois d'avril 1998.

4. Application à des données de trafic de la RATP

On se propose de mettre en pratique l'analyse d'intervention sur un exemple original. On s'intéresse donc à la série de trafic journalier sur le Métro, entre le 1er septembre 1997 et le 31 octobre 1997. Le graphique 12 représente cette série que l'on notera $(Y_t)_{t\in Z}$ par la suite.

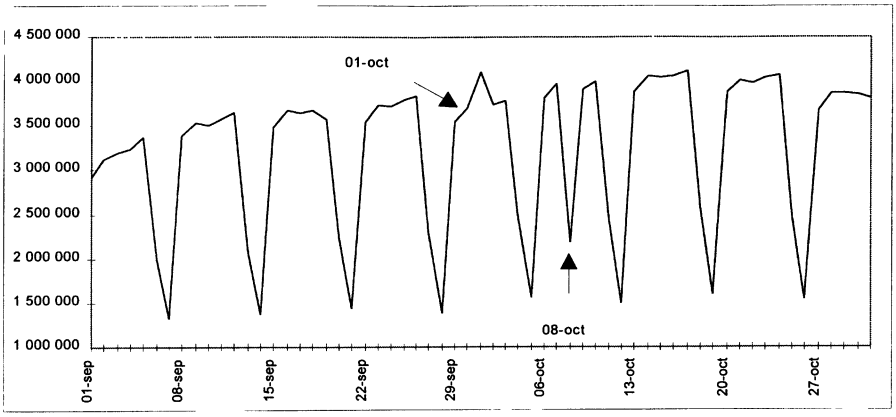
On se propose de modéliser dans un premier temps cette série, l'ensemble d'apprentissage de la série étant constitué par les données du 1er septembre au 26 octobre 1997, puis on effectuera ensuite des prévisions sur la période du lundi 27 au vendredi 31 octobre 1997.

Une première analyse graphique permet de constater la présence de deux points aberrants dans la série aux dates du 1er octobre et 8 octobre. L'historique des événements particuliers de la RATP révèle que ces dates correspondent aux événements suivants :

Mercredi 1er octobre : alerte à la pollution de niveau 3 en Ile de France, ce qui a entraîné des restrictions de circulation (plaques d'immatriculation paires ou impaires) ainsi que la gratuité de tous les modes de transport en commun.

Mercredi 8 octobre : grève partielle des agents de la Régie sur l'ensemble du réseau de la RATP.

Il semble donc judicieux d'utiliser un modèle à deux interventions ponctuelles aux dates du 1er octobre et du 8 octobre pour modéliser cette série, c'est-à-dire le modèle défini par l'équation (2.3), avec $k = 2$. On suppose de plus que ces deux événements ponctuels ont peu de chance de modifier durablement les habitudes des voyageurs du Métro, on néglige ainsi la présence éventuelle d'un effet rémanent de ces deux événements sur la série. Notre étude s'effectue en 5 étapes :



GRAPHIQUE 12
Évolution du trafic journalier sur le Métro du 1er septembre 1997 au 31 octobre 1997.

Étape 1 : Recherche de la structure du modèle SARIMA

En utilisant les fonctions d'autocorrélation et d'autocorrélation partielle sur l'ensemble des données d'apprentissage (voir graphes 14 et 15 en appendice), on retient un modèle de la forme SARIMA (1 0 0)(1 1 1) sans constante et de saisonnalité égale à 7.

Étape 2 : Estimation du modèle SARIMA

On estime alors le modèle obtenu par la méthode des moindres carrés non linéaire, et on obtient les résultats suivants :

$$(1 - \underset{(0.1681)}{0.2909}\, B)(1 - B^7)(1 + \underset{(0.2313)}{0.1043}\, B^7)Y_t = (1 - \underset{(0.2281)}{0.7571}\, B^7)\varepsilon_t$$

Les paramètres de la partie AR, saisonnière et nonsaisonnière, sont significativement égaux à 0, d'après le test de Student avec un risque α de 5%, mais on suppose que cela est dû aux 2 points aberrants de la série.

Afin de comparer la qualité d'ajustement des modèles aux données, on utilise comme critère à minimiser le RMSE (Root Mean Squared Error), défini par :

$$RMSE = \sqrt{\frac{1}{T-p}\sum_{t=1}^{T}(X_t - \hat{X}_t)^2}$$

où T est le nombre d'observations, p est le nombre de paramètres du modèle, $(X_t)_{t\in Z}$ est la série à modéliser et $(\hat{X}_t)_{t\in Z}$ est la série estimée. On obtient ici une valeur du RMSE égale à 349 580 que l'on essaiera de diminuer par la suite.

Étape 3 : Structure de la fonction d'intervention

On suppose donc que la série est générée par un processus qui suit l'équation (2.3) avec k égal à 2. Si on note P_t^{0110} la variable qui vaut 1 à la date du 1er octobre

et 0 sinon, et P_t^{0810} la variable qui vaut 1 à la date du 8 octobre et 0 sinon, le modèle est alors donné par :

$$Y_t = \omega_1 P_t^{0110} + \omega_2 P_t^{0810} + \frac{(1+\theta_1 B^7)}{(1-\phi_1 B)(1-B^7)(1-\phi_2 B^7)}\varepsilon_t$$

Étape 4 : Estimation du modèle d'intervention

On estime les paramètres du modèle par la méthode des moindres carrés non linéaire, et on obtient les résultats suivants :

$$Y_t = \underset{(45593)}{396677}\, P_t^{0110} - \underset{(46196)}{1709838}\, P_t^{0810} + \frac{(1+\underset{(0.152)}{0.815}\, B^7)}{(1-\underset{(0.051)}{0.836}\, B)(1-B^7)(1+\underset{(0.073)}{0.908}\, B^7)}$$

Les paramètres du modèle sont tous significativement différents de 0, d'après le test de Student avec un risque α de 5%. La valeur du RMSE est égale 73 006, ce qui représente un gain de 79,2% sur la qualité d'ajustement du modèle aux données sans intervention.

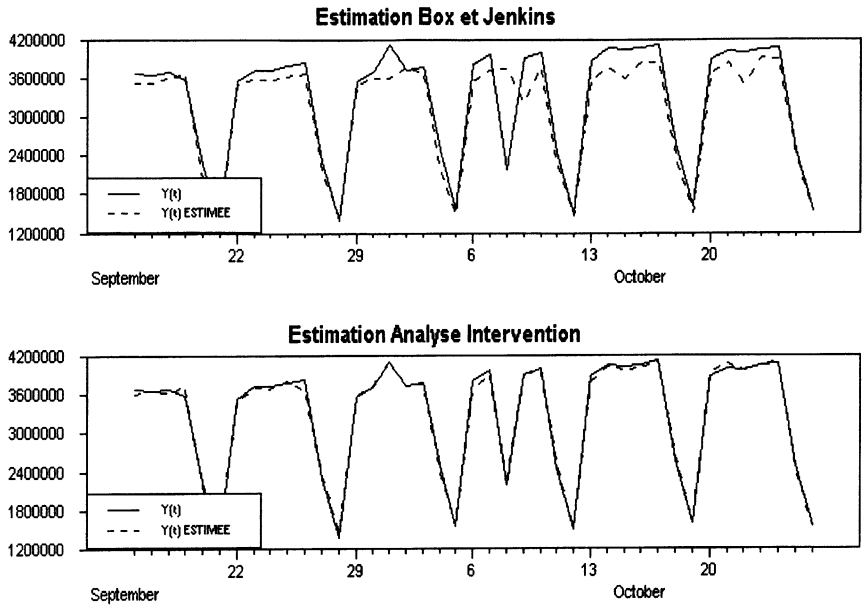
Ainsi, on estime que les mesures anti-pollution ont généré pour la journée du 1er octobre 1997 un apport de 396 677 voyageurs supplémentaires sur le Métro et que la grève des agents du mercredi 8 octobre a entraîné la perte de 1 709 838 voyageurs sur le Métro pour cette journée. La série corrigée de trafic sur le Métro s'obtient alors en retranchant l'impact estimé des événements :

Date	Série	Impact	Série Corrigée
Mercredi 1er Octobre	4 106 000	+ 396 677	3 709 323
Mercredi 8 Octobre	2 186 041	− 1 709 83 8	3 895 879

Nous donnons (graphique 13) les deux types d'estimations obtenues soit par la méthode classique de Box et Jenkins, soit par l'analyse d'intervention sur les deux dates.

Étape 5 : Impact sur les prévisions

Il est clair que la possible prise en compte d'événements extérieurs par l'analyse d'intervention améliore la modélisation d'une série, on peut alors se demander s'il en est de même pour la qualité des prévisions. On va donc essayer de prévoir le trafic journalier sur le Métro pour la période du lundi 27 octobre au vendredi 31 octobre 1997 en utilisant les deux différentes méthodes de modélisation. On utilise comme



(En trait plein, la série observée et en trait pointillé, la série estimée.)

GRAPHIQUE 13
Estimations du trafic journalier sur le Métro.

critère de comparaison le RMSE(f) défini par :

$$RMSE(f) = \sqrt{\frac{1}{h}\sum_{i=1}^{h}(X_{T+i} - \hat{X}_T(i))^2},$$

où h est l'horizon de prévision et $\hat{X}_T(i)$ est la prédiction obtenue pour X_{T+i} pour $i = 1, \ldots, h$.

Le tableau 1 donne les valeurs obtenues en prévision et permet de comparer, à l'aide des écarts relatifs, les données réelles et les données estimées par chacune des méthodes.

L'écart relatif journalier nous indique que les prévisions journalières sont de meilleure qualité (sauf pour le vendredi) lorsqu'on utilise l'analyse d'intervention. Sur l'ensemble des cinq jours ouvrables on observe également une diminution de l'erreur de prévision RMSE(f) de 81,1%, ainsi qu'une forte baisse de la moyenne des écarts relatifs. Ceci montre que l'analyse d'intervention permet une nette amélioration de la qualité des prévisions.

TABLEAU 1

Prévisions du trafic sur le Métro du lundi 27 octobre 1997 au vendredi 31 octobre 1997.

Dates	Données réelles	Box et Jenkins sans intervention	Analyse d'Intervention
lundi 27/10/97	3 679 491	3 571 211 (2.94%)	3 732 495 (−1.19%)
mardi 28/10/97	3 863 941	3 688 503 (4.54%)	3 848 360 (0.40%)
mercredi 29/10/97	3 864 114	3 440 616 (10.96%)	3 845 226 (0.49%)
jeudi 30/10/97	3 853 401	3 693 604 (4.15%)	3 872 965 (−0.51%)
vendredi 31/10/97	3 808 704	3 736 877 (1.88%)	3 886 603 (−2.04%)
Moyenne des écarts relatifs		**4.89%**	**− 0.57%**
RMSE(f)		**224 743**	**42 395**

(Les valeurs entre parenthèses sont les écarts relatifs journaliers).

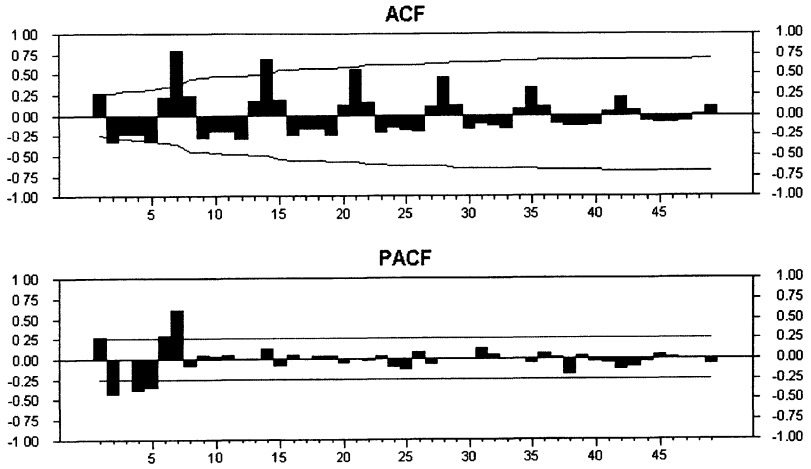
5. Conclusion

Cette étude avait pour objectif de présenter l'analyse d'intervention, de montrer les différents effets qu'elle permet de mesurer et de mettre en évidence son intérêt pour le praticien qui désire modéliser des séries à caractère économique, souvent perturbées par des phénomènes extérieurs. Nous avons souligné au travers de plusieurs exemples l'efficacité de cette méthode pour modéliser certaines séries d'intérêt pour la RATP et fournir une mesure de l'impact d'une ou plusieurs interventions extérieures sur la série. Or, pouvoir mesurer de manière fiable l'impact d'une grève ou d'une promotion sur un titre de transport fait partie du rôle du Département Commercial au sein de la RATP.

L'analyse d'intervention permet d'améliorer l'ajustement du modèle de Box et Jenkins aux données d'une série chronologique car elle prend en compte dans la construction du modèle un plus grand ensemble informationnel. De plus, l'application que nous avons traitée a permis de constater une nette amélioration sur les prévisions effectuées à partir d'une série perturbée lorsqu'on utilise l'analyse d'intervention de préférence au modèle classique de Box et Jenkins sans intervention. Cette propriété n'était pas évidente au départ, car on a constaté par expérience qu'un modèle fournissant le meilleur ajustement aux données d'une série (selon le critère du RMSE) ne fournit pas forcément les meilleures prévisions (selon le critère du RMSE(f)).

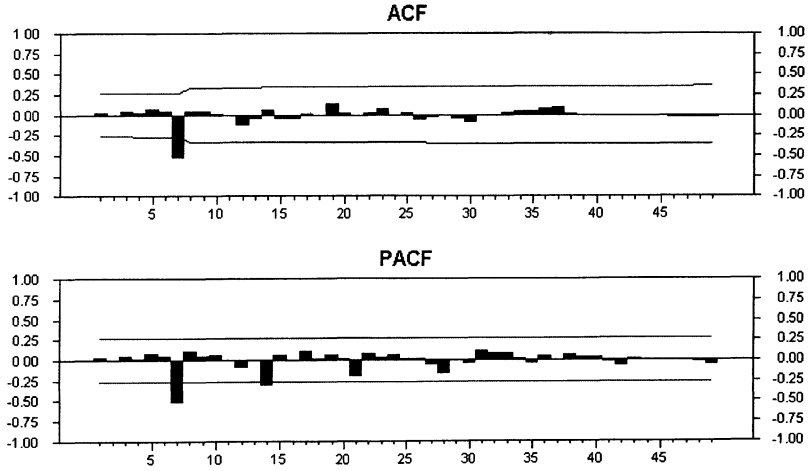
Cependant, les limites de l'analyse d'intervention sont atteintes en même temps que celles de la théorie de Box et Jenkins. Ainsi, il serait intéressant d'observer le comportement de cette théorie sur des processus non linéaires, en particulier sur des processus présentant des propriétés de longue mémoire, ce que nous comptons développer dans un prochain article. De même, il serait intéressant d'envisager une utilisation *a priori* de l'analyse d'intervention, qui permettrait de prévoir différents scénarii d'évolution d'une série consécutivement à un choc.

APPENDICE



GRAPHIQUE 14

Fonctions d 'autocorrélation (ACF) et d 'autocorrélation partielle (PACF) de la série de trafic journalier sur le Métro : $Y(t)$.



GRAPHIQUE 15

Fonctions d 'autocorrélation (ACF) et d'autocorrélation partielle (PACF) de la série de trafic journalier sur le Métro différenciée saisonnièrement : $(1 - B^7)Y(t)$.

Références

M.N. BHATTACHARYYA and A.P. LAYTON (1979), «Effectiveness of seat belt legislation on the Queensland road toll - an Australian case study in intervention analysis», *Journal of the American Statistical Association*, 74, 596–603.

G.E.BOX and G.M. JENKINS (1970), *Time Series Analysis, Forecasting and Control*, Holden-Day, San Francisco.

G.E. BOX, G.M. JENKINS and G.C. REINSEL (1994), *Time Series Analysis, Forecasting and Control (Third Edition)*, Prentice Hall, Englewood Cliffs, New Jersey.

G.E. BOX, and G.C. TIAO (1975), «Intervention analysis with applications to economic and environmentals problems», *Journal of the American Statistical Association*, 70, 70–79.

P.J. BROCKWELL and R.A. DAVIS (1987), *Time Series : Theory and Methods*, Springer-Verlag, New-York.

I. CHANG, G. TIAO and C. CHEN (1988), «Estimation of Time Series Parameters in the Presence of Outliers», *Technometrics*, 30, 2, 193–204.

C. CHEN and L.M. LIU (1990), «Joint Estimation of Models Parameters and Outlier Effects in Time Series», *Journal of the American Statistical Association*, 88, 421, 284–297.

L. DENBY and R. MARTIN (1979), «Robust Estimation of the First Order Autoregressive Parameter», *Journal of the American Statistical Association*, 74, 140–146.

G.C. TIAO (1985), «Autoregressive moving average models, intervention problems and outlier detection in time series», *Handbook of Statictics*, Vol. 5, E.J. HANNAN, P.R. KRISHNAIAH and M.M. RAO editions.

R.S. TSAY (1986), «Time series model specification in the presence of outliers», *Journal of the American Statistical Association*, 81, 132–141.