Statistique et analyse des données

Guy Melard

Modèles ARIMA pour des séries chronologiques non homogènes

Statistique et analyse des données, tome 4, n° 2 (1979), p. 41-50

<http://www.numdam.org/item?id=SAD_1979__4_2_41_0>

Statistique et Analyse des Données
2 - 1979 pp. 41, 50

MODELES ARIMA POUR DES SERIES CHRONOLOGIQUES NON HOMOGENES (*)

MELARD Guy, Université Libre de Bruxelles

RESUME

Cet article présente une ébauche d'étude comparée de deux approches pour la représentation de séries chronologiques comportant une tendance en dispersion, qui dérivent toutes deux de la méthode de Box et Jenkins. La première approche consiste à utiliser la méthode de Box et Cox pour déterminer la transformation à appliquer à la série afin de la rendre homogène. La seconde approche, due à l'auteur, se base sur une classe particulière de modèles ARIMA à coefficients dépendant du temps. Pour les deux approches, les spécifications sont validées au moyen d'une batterie de tests qui ont pour but de faire apparaître l'incapacité éventuelle du modèle à représenter la série à certaines époques de chaque année, lorsque certains niveaux de la variable sont atteints ou à certaines périodes de l'intervalle d'observation. L'étude comparée porte sur des séries économiques. Elle montre que l'une ou l'autre des deux approches convient le mieux selon le cas bien que la transformation de Box et Cox soit souvent préférable.

SUMMARY

In this paper, we outline a comparative study of two approaches, both derived from the Box-Jenkins method, for dealing with non-homogeneous time series, i.e. time series for which the variance is not constant. In the first approach, a transformation is applied in order to stabilize the variance, using the Box-Cox method. The second approach, due to the author, is based on a subclass of ARIMA models with time dependent coefficients. For both approaches the specifications are checked by means of a set of tests. Roughly speaking, these tests show the possible inability of a model to represent the time series in a certain period of each year, when a certain level is reached by the variate or during some part of the observation time interval. The comparative study bears on economic time series. It is shown that either approach is the most suitable, as the case may be, although the Box-Cox transformation is often preferable.

(*) *Contribution présentée au cours des Journées de l'Association des Statisticiens Universitaires, 22-26 mai 1978, Nice.*

1 - MODELE ARIMA SUR DONNEES TRANSFORMEES

1.1 - Principe

Considérons un modèle ARIMA (Box et Jenkins (1970), Anderson (1976)) :

$$\phi(B)\ X_t = \theta(B)\ \varepsilon_t$$

où B est l'opérateur de retard, $\phi(B)$ est l'opérateur autorégressif généralisé, de degré p, $\theta(B)$ est l'opérateur moyenne mobile, de degré q, $\{\varepsilon_t\}$ est un processus purement aléatoire normal de moyenne nulle et de variance σ^2; $\phi(0) = \theta(0) = 1$.

On suppose que les zéros des polynômes $\phi(B)$ et $\theta(B)$ sont, en module, supérieurs à 1, sauf qu'on peut admettre que $\phi(B)$ ait des zéros (connus) de module 1.

Beaucoup de séries chronologiques économiques ne sont pas susceptibles d'une représentation par un modèle ARIMA, notamment parce que la dispersion n'est pas homogène. Pour pouvoir appliquer la méthode de Box et Jenkins, il faut une transformation préalable F_h de la variable X :

$$\phi(B)\ [F_h(X_t)] = \theta(B)\ \varepsilon_t \qquad (h = 1, 2, 3; \text{ voir ci-dessous}).$$

1.2 - Estimation

Les polynômes $\phi(B)$ et $\theta(B)$ comportent des paramètres à estimer à partir de la série chronologique $(x_1, x_2, \ldots, x_n)$, de longueur n. On admet que F_h dépend de paramètres et on invoque la méthode de Box et Cox (1964) comme suit. La famille de transformation souvent utilisée si $X > 0$ est

$$F_1 : X \to \begin{cases} X^\lambda & \text{si } \lambda \neq 0 \\ \log X & \text{si } \lambda = 0\,. \end{cases}$$

Pour avoir continuité en le paramètre , on considère

$$F_2 : X \to \begin{cases} \dfrac{X^\lambda - 1}{\lambda} & \text{si } \lambda \neq 0 \\ \log X & \text{si } \lambda = 0. \end{cases}$$

Considérons les e_t qui s'obtiennent à partir de l'équation $\phi(B)\ y_t = \theta(B)\ e_t$ où $y_t = F_2(x_t)$, en fixant $e_0, e_{-1}, \ldots, e_{-q+1}, y_0, y_{-1}, \ldots y_{-p+1}$. Le jacobien vaut $\dot{x}^{n(\lambda-1)}$, où $\dot{x}$ est la moyenne

géométrique des x_t (t = 1,2, ..., n). Dans la plupart des programmes Box-Jenkins le paramètre λ n'est pas estimé conjointement avec les autres paramètres du modèle. L'article de Box et Cox contient pourtant les indications pour le faire et ceci a été appliqué par Ansley et al (1977) et Hermant (1977). Il s'agit de remplacer F_2 par F_3 définie par $F_3(X) = F_2(X)/(\dot{x}^{\lambda-1})$. En considérant $\dot{x}$ comme une constante, la fonction de vraisemblance prend la forme

$$L = (2\pi)^{-n/2}\ \sigma^{-n}\ \exp\left(-\frac{1}{2\sigma^2}\ \sum_{t=1}^{n} e_t^2\right)$$

ce qui permet d'employer l'algorithme des moindres carrés non linéaires.

2 - MODELES ARIMAG

2.1 - Principe

Partons du modèle ARIMA défini au § 1.1. On avait supposé que la variance de ε_t est σ^2. Au lieu que ε_t ait une variance constante, on admet qu'elle varie dans le temps, soit $\varepsilon_t = N(0, \sigma_t^2)$ où $\sigma_t > 0$ est une fonction certaine de t mais dépendant de paramètres inconnus. C'est ce que l'auteur a appelé un processus ARIMAG et pour lequel il a généralisé la méthode de Box et Jenkins. Voir à ce sujet Mélard (1977). Pour montrer l'intérêt de cette classe de processus, on peut remarquer les liens entre certains d'entre eux et quelques méthodes de prévision à court-terme. Goodman (1974) et Godolphin et Harrison (1975) ont montré que le lissage exponentiel d'ordre d sous-tend un processus ARIMA d'équation

$$\nabla^d X_t = (1 - \theta B)^d\ \varepsilon_t$$

où $\nabla = 1-B$. Les méthodes de démonstration partent du postulat que le processus générateur est de type ARIMA. En fait c'est d'un processus ARIMAG qu'il faudrait parler. A titre d'exemple considérons le lissage exponentiel simple. La précision $\hat{X}_t$ faite en (t-1) pour la valeur en t est définie par la relation

$$\hat{X}_t = \alpha X_{t-1} + (1 - \alpha)\ \hat{X}_{t-1}$$

où α est la constante de lissage. Notons que $\hat{X}_t - \hat{X}_{t-1} = \alpha(X_{t-1} - \hat{X}_{t-1})$. On postule que les erreurs de prévision $\varepsilon_t = X_t - \hat{X}_t$ sont des variables aléatoires de moyenne nulle et non corrélées entre elles. La différence $\nabla X_t = X_t - X_{t-1}$ peut s'écrire $\nabla X_t = (\hat{X}_t + \varepsilon_t) - (\hat{X}_{t-1} + \varepsilon_{t-1}) = (\hat{X}_t - \hat{X}_{t-1}) + \varepsilon_t - \varepsilon_{t-1} = \alpha(X_{t-1} - \hat{X}_{t-1}) + \varepsilon_t - \varepsilon_{t+1} = \alpha\varepsilon_{t-1} + \varepsilon_t - \varepsilon_{t-1} = (1 - (1-\alpha)B)\ \varepsilon_t$. L'hypothèse que σ_t^2 est constant n'est donc pas nécessaire.

2.2 - Estimation

Supposons fixées les valeurs initiales des $\hat{\varepsilon}_t = e_t$. La fonction de vraisemblance s'écrit :

$$L = (2\pi)^{-n/2} \left(\prod_{t=1}^{n} \sigma_t^{-1} \right) \exp \left(-\frac{1}{2} \sum_{t=1}^{n} (e_t/\sigma_t)^2 \right).$$

Supposons que σ_t soit paramétrisé de sorte que $\sigma_t = \sigma g_t$ où $\prod_{t=1}^{n} g_t = 1$ et que g_t ne dépend pas de σ, qui représente alors la moyenne géométrique des σ_t. Considérons σ comme fixé. Maximiser L revient alors à minimiser $\Sigma(e_t/g_t)^2$ par rapport aux paramètres contenus dans g_t, $\phi(B)$, $\theta(B)$. On estime ensuite σ^2 par la somme de carrés divisée par n. La méthode d'estimation de Box et Jenkins comporte donc une étape supplémentaire qui consiste à diviser e_t par g_t à chaque évaluation de la somme de carrés.

2.3 - Propriétés

Les conditions sont telles que les estimateurs ont des propriétés asymptotiques qui ne peuvent plus être déduites des théorèmes classiques sur les estimateurs du maximum de vraisemblance. Notons d'autre part que σ_t n'intervient pas dans l'expression de la prévision mais bien dans celle de l'écart-type associé.

3 - VALIDATION D'UN MODELE

3.1 - Introduction

Les tests de validation d'un modèle ARIMA concernent les résidus et principalement leur fonction d'autocorrélation, directement ou après transformation de Fourier. Ces tests ne suffisent pas car il est bien connu que la capacité prévisionnelle d'un modèle dépend aussi d'autres facteurs (par exemple Chatfield et Prothero (1973)). Des nouveaux tests semblent nécessaires tel le score-test de Pagan (1978) pour lequel l'hypothèse alternative est un modèle non linéaire du type

$$\phi(B)\ X_t = \theta(B)\ e_t + e_t\ .[\tilde{\theta}(B)\ e_{t-1}],$$

où $\tilde{\theta}(B)$ est un polynôme en B. Les tests que nous envisagerons ici sont les tests d'homogénéité (Mélard (1977-1977a)) pour lesquels l'alternative est moins spécifique.

3.2 - Les tests d'homogénéité

Le principe de base de ces tests est que si le modèle est correct sur $T = \{1, 2, \ldots, n\}$, il doit l'être sur T_1 et T_2 qui constituent une partition de T, définie indépendamment des observations. On ne respectera pas toujours cette règle, ce qui peut invalider les résultats. Les tests utilisent la proportion de points observés qui appartiennent aux intervalles de prévision d'horizon 1, calculés au niveau de confiance de 50 %. Si le modèle est correct, les proportions $\hat{p}_1$ et $\hat{p}_2$ correspondant à T_1 et T_2 ne doivent pas s'écarter de 0,5 de façon significative. Un test χ^2 à 2 degrés de liberté est donc réalisé. Les partitions utiles en pratiques sont les suivantes (n pair, égal à Ns, s pair) :

1° test "df" (début-fin) : $T_d = \{1, 2, \ldots, n/2\}$, $T_f = \{1 + n/2, \ldots, n\}$

2° test "bh" (hors-saison-en saison, par choix approprié des i_h)

$$T_b = \{i_1 + ks, i_2 + ks, \ldots, i_{s/2} + ks \mid \forall h, j \in \{1, \ldots, s/2\}, h \neq j : i_h \neq i_j$$

et $$i_h \in \{1, \ldots, s\} ; k = 0, 1, \ldots, N-1\} \quad , \quad T_h = T \setminus T_b$$

3° test "is" (inférieur-supérieur) : $T_i = \{t \mid \hat{x}_{t-1}(1) \leqslant x_o\}$, $T_s = T \setminus T_i$
où $\hat{x}_{t-1}(1)$ est la prévision faite en t-1 pour t, x_o fixé.

Outre la dépendance vis-à-vis des observations, des approximations sont commises : les e_t sur lesquels se base le test ne sont pas des variables aléatoires indépendantes - au mieux sont-elles faiblement autocorrélées; les paramètres, et notamment la variance des ε_t, sont estimés.

Ces tests sont donc de nature indicative mais il est prudent de tenir compte des résultats qu'ils fournissent en cas de rejet du modèle car ils peuvent révéler que l'hypothèse selon laquelle les ε_t sont identiquement distribués (à un facteur g_t près, éventuellement), normaux et indépendants est une hypothèse inacceptable pour les données.

4 - ETUDE COMPAREE

Il paraît évident que les deux approches, par modèle ARIMA avec transformation ou par modèle ARIMAG, ne peuvent pas servir indifféremment pour représenter une série donnée. En effet, dans le premier cas la dispersion est liée au niveau et dans le second cas elle est liée au temps. Dans cette étude nous examinons du point de vue statistique, sur base des tests d'homogénéité, dans quelle mesure ces deux approches sont acceptables pour représenter un ensemble de séries chronologiques économiques. Nous utiliserons parfois les statistiques de test pour choisir entre les deux approches mais ce n'est pas notre objectif principal. S'il fallait effectuer ce choix un critère raisonnable serait basé sur l'estimation de σ^2 en n'oubliant pas que dans un cas σ^2 est relatif au modèle transformé et dans l'autre il représente la moyenne géométrique des σ_t^2 .

Les notations suivantes sont employées :

α_K est le niveau de signification du test de Box et Pierce portant sur K autocorrélations résiduelles;

α_{bh} par exemple, est le niveau de signification du test d'homogénéité "bh" (de même pour "df" et "is").

4.1 - "Airline data" (Box et Jenkins (1970))

a) Modèle sur les logarithmes des données :

$$\nabla\nabla_{12} \log x_t = (1 - .40B)(1 - .61B^{12})\, e_t$$

$$\alpha_{20} = .37, \alpha_{df} = .04, \alpha_{bh} = .06, \alpha_{is} = .33$$

On constate que les intervalles de prévision sont trop larges au début et trop étroits à la fin; de plus ils sont trop larges en saison et trop étroits hors-saison.

b) Modèle ARIMAG :

$$\nabla\nabla_{12} x_t = (1 - .28B)(1 - .21B^{12})(.539 \exp(.0085t))\, e_t$$

$$\alpha_{20} = .19, \alpha_{df} = .52, \alpha_{bh} = .17, \alpha_{is} = .47$$

On voit une amélioration mais il y a des résidus réduits de plus de 3.

4.2 - Indice de production du charbon (276 observations)

Les 3 tests d'homogénéité rejettent tous les modèles; on constate des résidus importants.

4.3 - Indice de production d'électricité (276 observations)

α	"df"	"bh"	"is"
Box-Cox	.03	.03	.02
ARIMAG (g linéaire)	.07	.10	.09
ARIMAG (g exponentiel)	.28	.11	.29

En fait, aucun de ces modèles ne convient réellement car un examen plus détaillé révèle que :
1° dans le troisième cas, il y a 60 points en dehors des intervalles de prévision pour les mois à haut niveau de production, au lieu de la valeur attendue 65.5 mais 21 points sont en dessous et 39 au-dessus;
2° avec la transformation de Box et Cox, de paramètre -.04, le nombre correspondant de points est 53 seulement mais le partage est plus équilibré : 31 en dessous et 22 au-dessus.

4.4 - Indice de prix du porc (132 observations)

α	"df"	"bh"	"is"
logarithmes	.54	.44	.80
ARIMAG (g exponentiel)	.16	.08	.13

Le modèle ARIMAG convient moins bien.

4.5 - Stock de monnaie scripturale (276 observations)

α	"df"	"bh"	"is"
Box-Cox	.95	.95	.94
ARIMAG (g linéaire)	.01	.55	.00
ARIMAG (g exponentiel)	.46	.59	.33

L'hypothèse que l'écart-type des innovations ε_t varie linéairement est rejetée par les deux tests qui sont les plus appropriés à l'alternative intéressante. Les deux autres modèles sont acceptables, pour ce qui concerne les tests d'homogénéité, mais il est difficile de ne pas préférer le modèle avec transformation de Box et Cox.

4.6 - Indice des prix à la consommation des biens non durables

La série comporte 60 observations trimestrielles.

α	"df"	"bh"	"is"
Box-Cox	.07	.27	.07
ARIMAG (g linéaire)	.74	.80	.74
ARIMAG (g exponentiel)	.23	.79	.22

La fonction linéaire convient le mieux ici, sur base des 3 tests.

4.7 - Vingt-quatre séries trimestrielles (Gouzée (1976))

Ces séries, dont certaines sont "construites" appartiennent à un modèle trimestriel de la Belgique. Les résultats globaux sont présentés en annexe et sont basés sur le travail de Hermant (1977). Elle a estimé 3 types de modèles pour chaque série (parfois deux seulement) :

(N) modèle sur données non transformées

(G) modèle ARIMAG avec $g_t = \beta \exp(\gamma t)$

(BC) modèle avec transformation de Box et Cox, de paramètre λ.

Les estimations de γ et de λ sont données, accompagnées d'un intervalle de confiance à 95 %, mais seulement si cet intervalle ne contient pas 0, pour le premier, ou 1, pour le second. Rappelons que $\lambda = 1$ correspond aux données non transformées, $\lambda = 0$, aux logarithmes des données et $\lambda = -1$, aux inverses des données. Un modèle de chaque type a été sélectionné. Tous les modèles retenus sont acceptables au point de vue des autocorrélations. La colonne "modèles acceptables" indique les modèles qui ont passé les 3 tests d'homogénéité (à 10 %). Notons que le modèle (N) n'a été envisagé que si l'un au moins des modèles (G) et (BC) n'a pas été considéré comme intéressant, c'est-à-dire que le paramètre correspondant (γ ou λ) est non significatif (N.S. dans le tableau) au sens décrit ci-dessus. La dernière colonne contient des commentaires ou la référence à une note ci-dessous.

Notes : (1) cette série donne lieu à des résidus importants.

(2) le modèle sur la série non transformée est nettement moins bon pour les test "df" et "is" que le modèle Box-Cox pour lequel $\hat{\lambda} = -.18$, $(\hat{\lambda}^-, \hat{\lambda}^+) = (-1.5, 1.2)$.

En résumé du tableau :

	Total	Acceptés	refusés (cause)
Modèles (G) retenus	10	9	1 (test "bh")
Modèles (BC) retenus	17	16	1 (test "bh")
Modèles (N) retenus	14	12	2 (3 tests)

Pour les séries où les tests d'homogénéités permettent de choisir un modèle, un examen détaillé révèle que le modèle (G) est le meilleur une seule fois tandis que les modèles (BC) et (N) le sont 7 fois et 6 fois respectivement. Accessoirement, il faut remarquer que les valeurs de $\hat{\gamma}$ sont très voisines les unes des autres.

CONCLUSIONS

Il n'est pas encore possible de réaliser une classification des séries chronologiques sur base du type de transformation ou de fonction du temps à introduire dans le modèle ARIMA. Les premiers résultats montrent toutefois que la méthode de Box et Cox intégrée dans la méthode de Box et Jenkins conduit fréquemment à de bonnes représentations, mais que la classe des modèles ARIMAG doit être parfois envisagée. Les tests d'homogénéité constituent un instrument efficace pour cette étude comparée, surtout grâce aux interprétations qu'ils permettent, mais ceci ne doit pas faire oublier leurs défauts.

REMERCIEMENTS

Nous avons bénéficié d'un appui financier de l'Ecole de Commerce de l'Université Libre de Bruxelles. Nous sommes redevables à Monique Hermant de la programmation de la transformation de Box et Cox et des tests d'homogénéité, ainsi que des recherches à la base du paragraphe 4.7. Plusieurs de nos étudiants de l'année académique 1977-1978 ont contribué à la section 4. Nous remercions les deux lecteurs dont les suggestions ont permis une meilleure présentation des objectifs du § 4.

BIBLIOGRAPHIE

Anderson, O.D. (1976). Time Series Analysis and Forecasting. The Box-Jenkins approach. London : Butterworths.

Ansley, C.F., Spivey, W.A. and Wrobleski, W.J. (1977). A class of transformations for Box-Jenkins seasonal models. J. Roy. Statist. Soc. Ser. C. Appl. Statist., 26, 173-178.

Box, G.E.P. and Cox, D.R. (1964). An analysis of transformations. J. Roy. Statist. Soc. Ser.B, 26, 211-243.

Box, G.E.P. and Jenkins, G.M. (1970). Time Series Analysis : Forecasting and Control. San Francisco : Holden Day.

Chatfield, C. and Prothero, D.L. (1973). Box-Jenkins seasonal forecasting : problems in a case study. J. Roy. Statist. Soc. Ser. A, 136, 295-315.

Godolphin, E.J. and Harrison P.J. (1975). Equivalence theorems for polynomial-projecting predictors. J. Roy. Statist. Soc. Ser. B, 37, 205-215.

Goodman, M.L. (1974). A new look at higher-order exponential smoothing for forecasting. Operations Research, 22, 880-888.

Gouzée, N. (1976). Un nouveau modèle trimestriel belge : GG. Cahiers Economiques de Bruxelles, 70, 235-258.

Hermant, M. (1977). Les transformations de Box et Cox. Mémoire de licence spéciale en sciences mathématiques appliquées à la gestion. Université Libre de Bruxelles (non publié).

Mélard, G. (1977). Sur une classe de modèles ARIMA dépendant du temps. Cahiers du Centre d'Etudes de Recherche Opérationnelle, 19, n° 3-4, 285-295.

Mélard, G. (1977a). Prévisions de ventes. Communication au IVème Colloque International d'Econométrie Appliquée, Strasbourg (16-18 février 1977).

Pagan, A. (1978). Some simple tests for non-linear time series models. Core Discussion Paper 7812, Université Catholique de Louvain.

ANNEXE

	Nom de la série	(G) : $\hat{\gamma}$ ($\hat{\gamma}^-$, $\hat{\gamma}^+$)	(BC) : $\hat{\lambda}$ ($\hat{\lambda}^-$, $\hat{\lambda}^+$)	Modèles acceptables	Commentaires
1	Investissement fixe dans les secteurs du PIB réduit (prix courants)	N.S.	.21 (-.14,+.56)	N et BC	
2	Indice de prix du (1)	.015 (.004,.027)	-1.4 (-2.5,-0.3)	G et BC	
3	Investissement total (prix constants)	N.S.	N.S.	aucun	voir note (1)
4	Consommation privée totale (prix courants)	.026 (.018,.035)	-.74 (-.90,-.58)	BC	G rejeté par "bh"
5	Indice de prix des exportations	N.S.	-2.6 (-4.3,-0.8)	N et BC	
6	Produit national brut(prix constants)	N.S.	-.10 (-.46,+.26)	BC	
7	Produit national brut(prix courants)	.018 (.006,.030)	-.26 (-.60,+.08)	G et BC	
8	Chômage	N.S.	.02 (-.66,+.70)	N et BC	
9	Emploi total	N.S.	N.S.	N	
10	Index des salaires dans l'industrie	.014 (.000,.029)	.13 (-.35,+.61)	G et BC	
11	Investissement logement (prix constants)	N.S.	N.S.	N	
12	Investissement logement (prix courants)	N.S.	N.S.	N	
13	Indice de prix de (12)	.018 (.007,.030)	-1.1 (-1.7,-0.4)	G et BC	
14	Indice de prix de l'investissement public	.017 (.005,.030)	-.88 (-1.44,-.32)	G	BC rejeté par "bh"
15	Indice de prix de la consommation publique	.020 (.006,.034)	-1.5 (-2.5,-0.4)	G et BC	
16	Indice de prix de (17)	N.S.	N.S.	N	
17	Consommation privée de biens durables	.020 (.007,.034)	-.42 (-.78,-.07)	G et BC	
18	Indice de prix de la demande intérieure	N.S.	N.S.	N	voir note (2)
19	Consommation privée totale (prix constants)	.013 (.005,.022)	-.90 (-1.36,-.44)	G et BC	
20	Importations de biens et services (prix constants)	.016 (.003,.029)	-.19 (-.63,+.25)	G et BC	
21	Indice de prix de (7)	N.S.	-.16 (-1.3,1.0)	N et BC	
22	Taux de chômage	N.S.	.13 (-.58,+.84)	N et BC	
23	Emploi demandé par les secteurs du PIB réduit	N.S.	N.S.	N	
24	Prix implicite du capital	N.S.	-.04 (-.72,+.65)	N et BC	