La méthode de prévision de Box et Jenkins

Le cours suivant sur le **Box et Jenkins** qui suit est issu du support de la formation « **La prévision des Ventes** » que propose PREDICONSULT.

Il utilise des ressources utilisées dans le logiciel de prévision « Forecast Pro » que commercialise PREDICONSULT.

Box-Jenkins est une technique de prévision puissante, qui pour des données appropriées, surclasse fréquemment le lissage exponentiel. Néanmoins, les modèles de Box-Jenkins ont jusqu'ici été difficiles à identifier et très longs à construire. Ceci a réduit leur utilisation dans les prévisions dans les entreprises.

Les algorithmes d'automatisation comme ceux de Forecast Pro et Forecast Pro Batch permettent désormais aux prévisionnistes de construire rapidement et facilement des modèles de Box-Jenkins. Il en résulte une plus grande utilisation de ce type de modèles.

Dans la grande étude expérimentale de la précision des méthodes de prévision faite par Makridakis (1982), les modèles univariables de Box-Jenkins et ceux du lissage exponentiel étaient très proches quant à leurs performances. Dans l'idéal, un prévisionniste choisirait entre chaque modèle selon les caractéristiques des données. C'est précisément ce pour quoi le système expert de Forecast Pro a été conçu.

On construit les modèles de Box-Jenkins directement à partir de la fonction d'autocorrélation (ACF) des observations de la série chronologique. Par conséquent, une condition nécessaire au choix d'un modèle de Box-Jenkins est une stabilité raisonnable de la fonction d'autocorrélation. Si celles-ci sont peu stables et si la série est trop courte (disons moins de 40 points) pour permettre une estimation raisonnablement précise des autocorrélations, alors le lissage exponentiel est un meilleur choix. Ceci évite la principale difficulté de Box-Jenkins à ajuster un modèle complexe sur des corrélations exceptionnelles de faits isolés.

Box-Jenkins univariable ne prend pas en compte des variables explicatives. Si celles-ci présentent un intérêt important, alors une méthode multivariste, comme la régression dynamique est un meilleur choix.

Forecast Pro propose la procédure ARIMA (Autoregressions integrated Moving Average) univariste décrite par Box-Jenkins (1976). Ces modèles peuvent être totalement identifiés par le programme, mais il est également possible à l'utilisateur de construire interactivement un modèle ou encore de tester des variantes du modèle choisi par le système expert de Forecast Pro. Le programme propose le modèle multiplicatif saisonnier décrit par Box-Jenkins.

Cette section propose un résumé de la méthodologie statistique utilisée dans le programme. Ceux qui souhaiteraient approfondir le sujet se reporteront au livre *Applied Statistical Forecasting* (Goodrich 1989).

Implémentation de Box-Jenkins dans Forecast Pro

Identification automatique

Le programme détermine tout d'abord le niveau de différentiation nécessaire pour stationnariser la série ainsi que la saisonnalité. Il utilise une adaptation du test complet de Dickey-Fuller (1989). Il calcule donc les valeurs des paramètres d'un groupe de modèles candidats. Forecast Pro teste chaque modèle et choisit celui qui minimise le critère BIC.

Un ajustement exhaustif et l'examen de tous les modèles ARIMA d'ordre plus faible que le précédent nécessiteraient un temps de calcul inacceptable. Forecast Pro sur-ajuste tout d'abord un modèle *state space* et l'utilise pour générer plus rapidement des modèles de Box-Jenkins. Il arrive que cette approche manque le BIC minimum d'une faible qualité, mais il ne choisit jamais un mauvais modèle.

BFS a comparé sa modélisation Box-Jenkins automatique avec les résultats publiés de la M-competition, où un expert passait une heure pour identifier manuellement chaque modèle ARIMA. Forecast Pro surpasse l'expert à tout horizon de précision.

BSF recommande l'utilisation en standard de l'identification automatique. Utilisez l'identification manuelle uniquement quand un programme le suggère, ou lorsque vous avez de sérieuses raisons pour agir autrement.

Initialisation

Forecast Pro utilise la méthode de back forecasting pour initialiser les modèles de Box-Jenkins. Cette technique est décrite dans le livre de Box-Jenkins (1986).

Estimations des paramètres

Forecast Pro utilise la méthode des moindres carrés généralisés pour obtenir les estimations finales des paramètres. Si nécessaire, les paramètres sont ajustés pour assurer la stationnarité et l'invertibilité.

Terme constant

Par défaut, Forecast Pro utilise un terme constant seulement quand le modèle ARIMA ne met pas en jeu de différentiation. ceci a pour but d'éviter d'imposer des tendances déterministes qui peuvent amener à des erreurs très importantes pour des horizons importants. Vous pouvez, si vous le souhaitez, forcer le système. Dans ce cas Forecast Pro estimera la constante comme les autres paramètres, ce qui vous permettra de tester sa significativité statistique.

La méthode de Box-Jenkins

Deux concepts statistiques sont fondamentaux pour la compréhension de la modélisation Box-Jenkins et la régression dynamique : stationnarité et autocorrélation.

Stationnarité

Une série chronologique est stationnaire (au sens large) quand elle reste en équilibre statistique avec une moyenne constante, une variance constante et des autocorrélations constantes. Un processus stationnaire peut être représenté de façon optimale par un modèle ARIMA.

Malheureusement la majorité des séries industrielles et économiques ne sont pas stationnaires. Il y a de nombreuses formes de non-stationnarité, les suivantes sont particulièrement importantes :

Moyenne non stationnaire

La moyenne n'est pas constante, mais oscille lentement sans direction affirmée.

La série montre une tendance ou un cycle. La tendance n'est pas constante mais varie lentement.

Variance non stationnaire

La série chronologique est hétéroscédastique, c'est-à-dire que la variance n'est pas constante, mais oscille lentement sans direction affirmée.

Ces cas peuvent être traités en stationnarisant la série. Une moyenne non stationnaire est éliminée par *différentiation* de la série. Une variance non stationnaire est éliminée en appliquant une *transformation puissance de Box-Cox*.

Fonction d'autocorrélation

En accord avec la théorie statistique du modèle ARIMA, une série chronologique peut être décrite par la distribution liée de ses observations Y1, Y2,..., Yn. Cette distribution est caractérisée par son vecteur *moyenne* et par sa fonction d'*autocovariance*.

L'autocovariance de Y_t et Y_{t+m} est définie par :

$$\gamma_m = \mathrm{cov}(Y_t, Y_{t+m}) = E\left[(Y_t - \mu)(Y_{t+m} - \mu)\right],$$

où E représente l'opérateur espérance mathématique, cov représente la covariance et μ est l'espérance mathématique de Y_t. Remarquons que la fonction d'autocovariance est une fonction de l'*intervalle* de temps m et non

du temps en absolu. C'est une hypothèse implicite que l'autocovariance ne dépend pas du temps t. En d'autres termes, la série est *stationnaire*. Si ce n'est pas le cas, alors la fonction d'autocovariance n'est pas définie.

Remarquons que γ_0 est identique à la variance σ_Y^2. On calcule la fonction d'autocorrélation en divisant chaque terme de la fonction d'autocovariance par la variance σ_Y^2.

$$\rho_m = \frac{E(Y_t - \mu)(Y_{t+m} - \mu)}{\sigma_Y^2}$$

La fonction d'autocovariance est une valeur théorique décrivant une distribution statistique. En pratique, on ne dispose que d'estimations de vraies valeurs. La formule généralement utilisée est la suivante :

$$\rho_m = E\left[\frac{1}{T}\sum_{t=1}^{T-m}(Y_t - \overline{Y})(Y_{t+m} - \overline{Y})\right]$$

où $\overline{Y}$ est la moyenne de l'échantillon. L'estimation de la fonction d'autocorrélation sur l'échantillon est donnée par :

$$r_m = \frac{c_m}{c_0}$$

L'erreur attachée à cette estimation peut être importante, en particulier quand ces autocorrélations sont élevées. Les estimations sont également fortement inter-corrélées. Dans ces conditions, il est indispensable de faire très attention en prenant pour significative une corrélation particulière de la fonction d'autocorrélation calculée sur l'échantillon par examen visuel.

La fonction d'autocorrélation affichée dans Forecast Pro inclut des lignes pointillées à 2σ, où σ est l'écart-type approché du coefficient d'autocorrélation des valeurs observées, calculé par l'approximation de Bartlett (1946). La vitesse avec laquelle σ croît dépend des valeurs des autocorrélations de rang plus faible.

Description du modèle de Box-Jenkins

La méthode de Box-Jenkins modélise la fonction d'autocorrélation avec le minimum de paramètres. Comme Box-Jenkins inclut des termes, comme par exemple ceux de moyennes mobiles, il fournit, tout du moins en théorie, le modèle dynamique endogène optimum.

En conséquence, quand en définitive on choisit un modèle de régression dynamique, une analyse préliminaire par Box-Jenkins fournit des renseignements intéressants sur la dynamique du modèle. Comme la procédure est rapide et entièrement automatique, l'utilisateur ne doit pas s'en priver.

Le modèle de Box-Jenkins combine des termes autorégressifs (AR), une différenciation (I) et des termes moyennes mobiles (MA) pour former le modèle complet (ARIMA). Cette famille de modèles permet de représenter la structure des corrélations d'une série chronologique avec le minimum de paramètres à ajuster. Ces modèles sont donc très satisfaisants du point de vue statistique et sont susceptibles de fournir des prévisions d'excellente qualité.

La notation utilisée sera cohérente avec celle du lissage exponentiel.

N	Nombre de points de l'historique
m	horizon de la prévision
p	nombre de périodes par an
Y_t	valeur observée au temps t
∇_t	opérateur de différentiation
∇_s	opérateur de différentiation saisonnier
B	opérateur de différentiation
ϕ_i	coefficient d'autorégression (intervalle i)
$\phi(B)$	polynôme d'autorégression d'ordre p
Φ_i	polynôme d'autorégression saisonnier (intervalle i)
$\Phi(B^s)$	polynôme d'autorégression saisonnier d'ordre p_S
θ_i	coefficient de moyenne mobile (intervalle i)
$\theta(B)$	polynôme de moyenne mobile d'ordre q
Θ_i	coefficient de moyenne mobile saisonnier (intervalle i)
$\Theta(B^s)$	polynôme de moyenne mobile saisonnier d'ordre q_S
$\hat{Y}_t(m)$	prévision au temps $t + m$ faite à partir du temps t
e_t	erreur de prévision à un intervalle $Y_t - Y_{t-1}$
ε_t	choc aléatoire distribué normalement

Différentiation

Si, en moyenne, une série chronologique n'est pas stationnaire, il est tout d'abord nécessaire de la stationnariser par différentiation. Pour décrire le processus de différentiation nous utiliserons l'opérateur arrière, définit ainsi :

$$BY_t = Y_{t-1}$$

$$B^m Y_t = Y_{t-m}$$

Cet opérateur sera utilisé tout au long de la discussion du processus ARMA. Par exemple, l'opérateur de différentiation est défini comme

$$\nabla = (1-B)$$

Processus autorégressifs

Le modèle AR(p) est défini par l'équation

$$Y_t - \phi_1 Y_{t-1} - \phi_2 Y_{t-2} - \ldots - \phi_p Y_{t-p} = \varepsilon_t \qquad (1)$$

dans laquelle on voit que la variable dépendante est en fait régressée sur ses valeurs passées. On peut représenter cette équation en terme d'opérateur arrière B sous la forme

$$(1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p) Y_t = \varepsilon_t \qquad (2)$$

en notant $\phi(B)$ le polynôme en B, on obtient la forme abrégée

$$\phi(B) Y_t = \varepsilon_t \qquad (3)$$

Processus moyenne mobile

Le processus moyenne mobile MA(q) est défini par

$$Y = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \theta_q \varepsilon_{t-q} \qquad (4)$$

ou, sous bien abrégée

$$Y_t = \theta(B) \varepsilon_t \qquad (5)$$

Un processus purement moyenne mobile MA(q) n'est en réalité jamais observable dans le monde réel. Il décrit le processus très improbable pour lequel les autocorrélations sont différentes de zéro pour q intervalles et zéro ensuite.

Les termes moyennes mobiles sont utilisés en pratique en conjonction avec la différentiation et des termes autorégressifs. Dans ce cas, ils deviennent très utiles. Ils introduisent un lissage des données exactement comme le fait le lissage exponentiel.

Processus ARMA et ARIMA

Le processus autorégressif moyenne mobile ARMA (p,q) combine les processus AR(p) et MA(q). Sous forme abrégée ceci s'écrit

$$\phi(B)Y_t = \theta(B)\varepsilon_t \qquad (6)$$

D'où le processus AR(p) est un processus ARMA (p,0), et le processus MA(q) est un processus ARMA (0,q).

Toute série stationnaire est susceptible d'être modélisée par un processus ARMA (p,q). Toute série chronologique qui peut être rendue stationnaire en différenciant d fois peut être modélisée par un processus ARIMA (p,d,q). Le modèle ARIMA (p,d,q) est défini par l'équation suivante

$$\phi(B)(1-B)^d Y_t = \theta(B)\varepsilon_t \qquad (7a)$$

Il s'agit du modèle le plus général de Box-Jenkins non saisonnier.

Tendance déterministe

Par défaut Forecast Pro ne met pas de terme constant dans un modèle ARIMA sauf quand il n'y a pas différentiation. Pour introduire une constante, mettez la pseudo-variable _CONST dans le tableau des données juste sous le nom des séries que vous voulez prévoir. L'équation du modèle prend la forme (7b)

$$\phi(B)(1-B)^d Y_t = \theta(B)\varepsilon_t + c \qquad (7b)$$

L'effet du terme constant est d'introduire une tendance déterministe dans votre modèle, en plus des autres propriétés. Si vous avez différencié une fois, la tendance est linéaire, si vous avez différencié deux fois, la tendance est quadratique.

Ce n'est pas en général très désirable car cela extrapole la tendance globale de l'historique dans le futur, même si la tendance actuelle est très faible. Ceci entraîne en général une précision très faible des prévisions pour les horizons élevés. BFS a confirmé cet effet en testant les 111 séries de Makridakis.

Modèles saisonniers

L'équation (7a) est utilisable pour modéliser des séries saisonnières, à condition que les polynômes prennent en compte une ou plusieurs périodes. Cela signifie que soit p, soit q, soit les deux doivent être égal ou plu grand que la longueur d'une période s. Comme tous les termes intermédiaires seront également inclus, ceci va entraîner des modèles sur-dimensionnés qui contiennent des coefficients inutiles qu'il faut pourtant estimer. Ce phénomène est en général néfaste pour la valeur prédictive du modèle.

D'une autre façon, on peut considérer une version saisonnière du modèle (7) dans lequel l'opérateur arrière B est remplacé par son équivalent saisonnier B^S. L'équation s'écrit alors

$$\Phi(B^s)(1-B^s)^D Y_t = \Theta(B^s)\varepsilon_t \qquad (8)$$

dans lequel les polynômes θ et Θ sont respectivement d'ordre P et Q. On appelle ce modèle ARIMA (P,D,Q). Il met en relation les observations d'une période donnée à celles de la même période les années précédentes, mais pas aux observations de périodes plus récentes.

Le modèle saisonnier le plus général inclut simultanément des modèles ARIMA saisonniers et simples. L'équation suivante décrit le modèle ARIMA *multiplicatif saisonnier*.

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D Y_t = \theta(B)\Theta(B^s)\varepsilon_t \qquad (9)$$

On le symbolise par ARIMA (p,d,q)(P,D,Q).

Choisir l'ordre des données

La partie la plus difficile dans la modélisation Box-Jenkins est de décider quel modèle ARIMA (p,d,q) ajustera au mieux les données, c'est à dire d'identifier le degré de différentiation d, l'ordre p de AR, l'ordre q de MA. Une grande partie du travail de Box-Jenkins (1976) est dévolue à ce problème d'identification. Le système expert de Forecast pro identifie le modèle *automatiquement*. Il n'est donc pas nécessaire de lire la partie suivante qui décrit la procédure *originale* de Box-Jenkins.

La procédure originale de Box-Jenkins est basée sur des analyses graphiques et numériques des fonctions d'autocorrélation et d'autocorrélation partielle. C'est une procédure de reconnaissance de structure qui nécessite beaucoup d'agilité et est longue à apprendre. Nous ne discuterons que le cas non saisonnier.

Degré de différentiation

La procédure d'identification commence par la détermination du degré de différentiation d qui est nécessaire pour rendre stationnaire les données observées Y_t. C'est fait par l'examen de la fonction d'autocorrélation r_k.

On examine les premières valeurs de la fonction d'autocorrélation des données brutes Y_t, si elles décrivent relativement rapidement, aucune différentiation n'est nécessaire, à savoir d=0. Si ce n'est pas le cas, on remplace les données brutes par leurs différences premières ∇Y_t et on répète le processus. Si la fonction d'autocorrélation des valeurs différenciées décroît rapidement, d=1. Si non, les données sont différenciées une seconde fois pour obtenir $\nabla^2 Y_t$. On répète le processus jusqu'à ce que, pour une certaine valeur d, la fonction d'autocorrélation des valeurs différenciées décroisse rapidement. En pratique d est rarement supérieur à 2.

Une fois que le degré de différentiation est déterminé, le reste de l'analyse s'effectue sur les données stationnarisées $\nabla^d Y_t$. Si d vaut zéro, il s'agit des données brutes.

Ordre d'autorégression

L'ordre p d'autorégression est déterminé par inspection de la fonction d'autocorrélation partielle $\hat{\phi}_{kk}$. Nous allons expliquer cette notation singulière et la définition de cette fonction à partir d'un exemple.

Supposons que le processus soit entièrement autorégressif (q=0). Alors une stratégie rationnelle pour déterminer p serait de calculer une régression de Y_t sur son premier intervalle, pour ses deux valeurs précédentes et ainsi de suite jusqu'à ce que le dernier terme introduit s'avère non significatif. C'est réalisé par un test statistique sur $\hat{\phi}_{kk}$, qui est défini comme le coefficient de Y_{t-k} dans la régression sur $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-1}$ (à savoir le k ième coefficient de la régression).

En fait, un algorithme récursif est utilisé au lieu de faire autant de régressions. Un graphe des 48 premiers intervalles est affiché permettant de déterminer l'ordre du processus AR. Si le processus est ARIMA (p,d,0) alors la fonction d'autocorrélation partielle décroît de façon abrupte après le p ième intervalle.

Ordre moyenne mobile

Un processus purement moyenne mobile ARIMA (0,d,q) présente les mêmes aspects pour la fonction d'autocorrélation que le processus autorégressif ARIMA (p,d,0) pour sa fonction d'autocorrélation partielle. En d'autres termes si le processus est ARIMA(0,d,q) alors r_k est grand pour $k<q+1$ et petit pour $k>q$. On se sert donc de la fonction d'autocorrélation pour le processus MA de la même manière que la fonction d'autocorrélation partielle pour le processus AR.

Les fonctions r_k et $\hat{\phi}_{kk}$ montrent des structures semblables pour respectivement ARIMA (0,d,q) et ARIMA (p,d,o). Au lieu de décroître de façon soudaine à p et q respectivement, ces fonctions décroissent de façon amortie ou avec des oscillations. C'est par l'examen des deux fonctions que le prévisionniste (expert) peut déterminer les ordres de processus purement AR et MA.

Les processus mixtes ARIMA (p,d,q) sont plus complexes. Ni la fonction d'autocorrélation, ni la fonction d'autocorrélation partielle ne s'éteignent brutalement. En fait, les autocorrélations restent importantes pour $k \leq q+1$ et décroissent exponentiellement ensuite. La fonction d'autocorrélation partielle reste grande pour $k<p+1$ et décroît pour $k>p$. Le processus mixte ARIMA est très difficile à identifier par un non expert.