Initiation aux séries chronologiques : approche descriptive

On observe N relevés sucessifs d'une variable quantitative Y. Il s'agit de relevés équidistants dans le temps donc les valeurs observées peuvent être indicées par des numéros t allant de 1 à N, qui correspondent aux rangs d'observation de ces N relevés. On peut donc écrire la série des valeurs de la façon suivante : $y_t, \quad t \in \{1, \cdots, N\}$. On distingue généralement :

- un mouvement de longue durée (appelée trend) notée (f_t)
  Le trend schématise la tendance générale du phénomène.
- une composante saisonnière notée (S_t). Elle traduit les variations dites saisonnières qui résultent d'évènements se répétant à l'identique de période en période. On pose $S_t = c_i$ pour $t = i+kp$ où p est la longueur de la période. c_i est appelé coefficient saisonnier (de la saison i), i variant de $1, \cdots, p$.
- la composante résiduelle (e_t) correspondant à des mouvements perturbateurs irréguliers et imprévisibles.

Traditionnellement, on considère deux types de schéma :

- le schéma additif : $y_t = f_t + S_t + e_t$
- le schéma multiplicatif : $y_t = f_t \times S_t + e_t$

	schéma additif	schéma multiplicatif
série ajustée	$\widehat{y}_t = \widehat{f}_t + \widehat{c}_i$	$\widehat{y}_t = \widehat{f}_t \times \widehat{c}_i$
série corrigée des variations saisonnières	$y_t^* = y_t - \widehat{c}_i$	$y_t^* = y_t / \widehat{c}_i$
prévision à l'horizon h	$\widehat{y}_{t+h} = \widehat{f}_{t+h} + \widehat{c}_i$	$\widehat{y}_{t+h} = \widehat{f}_{t+h} \times \widehat{c}_i$

Moyennes mobiles $(M_{t,p})$ d'ordre p : elles ne sont calculables que pour les numéros de relevés t tels que $k < t < N - k$ où k est l'entier tel que $p = 2k + 1$ ou bien $p = 2k$).

Ordre	$p = 2k + 1$	$p = 2k$
Moyenne mobile	$M_{t,p} = \dfrac{1}{2k+1} \displaystyle\sum_{i=-k}^{i=k} y_{t+i}$	$M_{t,p} = \dfrac{1}{2k} \Big[(\displaystyle\sum_{i=-k+1}^{i=k-1} y_{t+i}) + (y_{t-k} + y_{t+k})/2 \Big]$

Initiation aux séries chronologiques : outils probabilistes

Bruit blanc : On appelle bruit blanc un processus aléatoire $(X_t, t \in T)$ stationnaire au second ordre, centré, de fonction d'autocovariance γ telle que $\gamma(s,t) = \sigma^2 \delta_{s,t}$.

Bruit blanc fort : On appelle bruit blanc fort une famille $(X_t, t \in T)$ de variables aléatoires réelles i.i.d. de variance σ^2.

Marche aléatoire : On appelle marche aléatoire un processus $(S_t, t \in \mathbb{N}^*)$ défini par

$$S_t = X_1 + X_2 + \cdots + X_t$$

où $(X_t, t \in \mathbb{N}^*)$ est un bruit blanc.

Remarquons que $\forall t \in \mathbb{N}^*, E(S_t) = 0, V(S_t) = t\sigma^2$ et $\forall h \in \mathbb{N}^*, \gamma(t, t+h) = t\sigma^2$.

Opérateur de retard : C'est un opérateur noté B dont l'effet sur un processus (Y_t) est de retarder ce processus dans le sens où $BX_t = X_{t-1}$. Pour un entier $k \geq 2$, on note $B^k = B \circ B^{k-1}$ les compositions successives de l'opérateur B.

Opérateur de filtrage linéaire : Soit $\{a_k\}$ une séquence de nombres réels, l'opérateur $\sum_k a_k B^k$ est appelé opérateur de filtrage linéaire.

A un processus (X_t), il fait correspondre le processus :

$$Y_t = \Big(\sum_k a_k B^k\Big) X_t = \sum_k a_k X_{t-k}.$$

Filtrage des processus stationnaire au second ordre : Soit $\{a_k, k \in \mathbb{Z}\}$ une suite absolument sommable (c'est-à-dire $\sum_{k \in \mathbb{Z}} |a_k| < \infty$), et soit $(X_t, t \in \mathbb{Z})$ un processus stationnaire au second ordre d'espérance μ_X et de fonction d'autocovariance g_X.

Alors le processus $(Y_t, t \in \mathbb{Z})$ tel que $Y_t = \sum_{k \in \mathbb{Z}} a_k X_{t-k}$ est stationnaire d'espérance

$$\mu_Y = \mu_X \sum_{k \in \mathbb{Z}} a_k \tag{1}$$

de fonction d'autocovariance g_Y telle que

$$g_Y(h) = \sum_{j \in \mathbb{Z}} \sum_{k \in \mathbb{Z}} a_j a_k g_X(h + k - j). \tag{2}$$

Processus AR(p) : Le processus $(X_t, t \in \mathbb{Z})$ est dit *autorégressif* d'ordre p (ou AR(p)) s'il est stationnaire au second ordre et s'il est solution de l'équation de récurrence :

$$X_t = a_1 X_{t-1} + \cdots + a_p X_{t-p} + Z_t \tag{3}$$

où (Z_t) est un bruit blanc de variance σ^2.

Existence des Processus AR(p) : L'équation récurrente :

$$X_t = a_1 X_{t-1} + \cdots + a_p X_{t-p} + Z_t$$

où (Z_t) est un bruit blanc de variance σ^2, admet une solution stationnaire au second ordre si et seulement si le polynôme :

$$P(z) = 1 - a_1 z - \cdots - a_p z^p \neq 0 \text{ pour } |z| = 1$$

Cas d'un AR(1) : $X_t = aX_{t-1} + Z_t$. Si $|a| < 1$, alors $g_X(h) = \dfrac{a^{|h|}}{1-a^2}\sigma^2$.

Processus MA(q) : Le processus $(X_t, t \in \mathbb{Z})$ est dit *moving average* ou *à moyenne ajustée* d'ordre q (ou MA(q)) s'il vérifie :

$$X_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q} \tag{4}$$

où (Z_t) est un bruit blanc de variance σ^2.

Si $(X_t, t \in \mathbb{Z})$ est MA(q), on a donc $E(X_t) = 0$ et sa fonction d'autocovariance vérifie :

$$g_X(h) = \begin{cases} \sigma^2 \displaystyle\sum_{k=0}^{|h|} \theta_k \theta_{k+|h|} & \text{si} \quad 0 \leq |h| \leq q \\ 0 & \text{sinon} \end{cases}$$

Processus ARMA(p, q) : Le processus $(X_t, t \in \mathbb{Z})$ est dit *autorégressif moving average* d'ordre (p, q) (ou ARMA(p, q)) s'il est stationnaire au second ordre et s'il vérifie :

$$X_t - a_1 X_{t-1} - \cdots - a_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q} \tag{5}$$

où (Z_t) est un bruit blanc de variance σ^2.

Existence des Processus ARMA(p, q) : on pose $P(z) = 1 - a_1 z - \cdots - a_p z^p$ et $Q(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$. On suppose que P et Q n'ont pas de zéros communs. Alors L'équation récurrente (5) admet une solution stationnaire au second ordre si et seulement si le polynôme

$$P(z) = 1 - a_1 z - \cdots - a_p z^p \neq 0 \text{ pour } |z| = 1.$$

Cette solution est unique et a pour expression : $X_t = \sum_{k\in\mathbb{Z}} \phi_k Z_{t-k}$

où (ϕ_k) est la suite des coefficients du développements en série de Laurent de $Q(z)/P(z)$ au voisinage du cercle unité.

Processus ARIMA(p,q,r) : Un processus (X_t) est dit ARIMA(p,q,r) si après r différentiations, le processus ainsi obtenu est ARMA(p,q). Avec les notations précédentes, on a $(I-B)^r Q(B)X_t = P(B)Z_t$.

ARIMA est l'abréviation de *AutoRegressive Integrated Moving Average*.

Stationnarité et différentiation : La série chronologique observée vérifie t-elle l'hypothèse de stationnarité au second ordre des modèles ARMA? Pour cela, on visualise cette série et si elle semble répondre aux critères de stationnarité (oscillations d'amplitudes pas trop variables autour d'une valeur constante au cours du temps), on pose $r = 0$. Sinon, il faut envisager une transformation stationnarisante. Par exemple, les variables économiques sont souvent considérées non stationnaires, mais à accroissements stationnaires. En présence d'une tendance, la meilleure méthode pour l'éliminer est de différencier la série : la différentiation d'ordre 1 transforme (X_t) en $(X_t - X_{t-1})$; la différentiation d'ordre r consiste à effectuer r différentiations d'ordre 1.

Diagrammes d'autocorrélation : On appelle *corrélogramme* le diagramme représentant les coefficients d'autocorrélation d'ordre 1, 2,$\cdots, k, \cdots$ de la série. L'autocorrélation d'ordre k est la corrélation entre la série et elle-même retardée de k relevés. On appelle *corrélogramme partiel* le diagramme représentant les coefficients d'autocorrélation partielle d'ordre 1, 2,$\cdots, k, \cdots$ de la série.

Notons ρ_k le coefficient d'autocorrélation d'ordre k, notons ρ_k^* celui de corrélation partielle d'ordre k. On a les formules suivantes :

$$\rho_1^* = \rho_1 \ , \qquad \rho_2^* = \frac{\rho_2 - \rho_1^2}{1-\rho_1^2} \ , \quad \cdots \quad \rho_k^* = \frac{\rho_k - \sum_{j=1}^{k-1} \Phi_{k-1,j}\rho_{k-j}}{1 - \sum_{j=1}^{k-1} \Phi_{k-1,j}\rho_j}$$

avec l'équation de récurrence : $\Phi_{k,j} = \Phi_{k-1,j} - \Phi_{k,k}\Phi_{k-1,k-j}$ et $\Phi_{k,k} = \rho_k^*$.

Le coefficient d'autocorrélation partielle d'ordre k traduit une corrélation conditionnelle entre la série initiale (X_t) et celle retardée de k relevés, soit (X_{t-k}). Le conditionnement correspond à la corrélation non expliquée par les relevés intermédiaires $(X_{t-1}, X_{t-2}, \cdots, X_{t-k-1})$. Le corrélogramme partiel sert donc à détecter des effets qui ne dépendent pas linéairement des relevés à petite distance (en temps).

Phase d'identification : Les outils principaux sont les tracés :

- de la série, avant et après différentiation(s) éventuelle(s),
- des corrélagrammes partiel et non partiel

Le nombre de différentiations effectuées afin d'atteindre la stationnarité d'ordre 2 fournit la valeur r.

La similitude en corrélogrammes observés et corrélogrammes théoriques permet de proposer un ou plusieurs couples (p, q) candidats.

Phase d'estimation et d'ajustement : Pour un triplet (p, q, r) donné, on effectue l'estimation des paramètres du modèle ARIMA(p, q, r) et on calcule le critère d'AKaike (ou AIC pour Akaike information criterion). D'une façon générale, on a : $AIC = 2k - 2\ln(L)$ où k est le nombre de paramètres du modèle et L la vraisemblance. Les modèles ayant beaucoup de paramètres (c'est-à-dire p, q et r grands) sont ainsi pénalisés. Le modèle retenu est celui ayant le plus faible AIC. La méthodologie AIC consiste à trouver le modèle expliquant le mieux les observations, avec un minimumm de paramètres. Les valeurs faibles pour p, q et r sont donc privilégiées.