Une introduction à la méthodologie de Box et Jenkins : l'utilisation de modèles ARIMA avec SPSS

Dominique Desbois

Une introduction à la méthodologie de Box et Jenkins : l'utilisation de modèles ARIMA avec *SPSS*

Dominique DESBOIS

INRA-SAE2 Nancy et SCEES
251 rue de Vaugirard, 75732 Paris Cedex 15
Fax : +33 1 49 55 85 00 E-mail : dominique.desbois@agriculture.gouv.fr

RÉSUMÉ :

Cette note initie l'utilisateur débutant à la mise en œuvre des procédures du module *Séries chronologiques* du logiciel *SPSS pour Windows* correspondant à la méthodologie de Box et Jenkins pour la modélisation à partir de processus ARIMA. Cette mise en œuvre concerne l'analyse et la prévision du prix des produits agricoles, en particulier celui du blé tendre. Le listage de chaque procédure d'analyse statistique est commenté par la présentation du formulaire et l'interprétation des résultats obtenus.

MOTS-CLÉS : Série chronologique, méthode de Box et Jenkins, processus ARIMA, prix des produits agricoles, logiciel statistique, mise en œuvre.

I) Introduction à l'analyse conjoncturelle des prix agricoles

Parmi les principales missions qui incombent à un service public de statistique agricole figure l'analyse de la conjoncture agricole au sein de laquelle l'étude des prix de marché des principaux produits agricoles occupe une place particulière. Cette étude de conjoncture suppose non seulement de dégager les principales tendances d'évolution des prix agricoles, d'estimer les variations saisonnières ou cycliques auxquelles ces prix sont soumis, mais également de pouvoir livrer des prévisions de court ou moyen terme concernant en particulier les indices de prix à la production (IPPAP). Ces prévisions portant sur des indices mensuels sont utilisées pour établir les comptes prévisionnels de l'agriculture, permettant une estimation en cours de campagne de la valeur globale de la production à un prix de référence intégrant le prix du marché mais aussi l'effet des différentes interventions (stockage, retrait, …) effectués par les organismes professionnels en charge de la régulation des marchés agricoles. Ces indices de prix mensuels sont par ailleurs transmis à Eurostat[1] pour être intégrés au système de banques de données statistiques européennes gérées par cet organisme.

Parmi l'ensemble des produits végétaux, les céréales occupent en France une place particulière puisque notre production, principalement blé tendre, maïs et orge, s'élevait en 1996 à une valeur de près de 7 milliards d'écus, représentant 32 % des quantités produites dans l'Union européenne. Au premier rang figure le blé tendre représentant plus de 50 % de notre production céréalière et classant la France en tête des producteurs européens. Vient ensuite le maïs grain pour lequel la France se classe également au premier rang des pays européens (43 % en volume de la production européenne 1998). Par contre, l'orge produit en France ne représente que 20 % du volume produit par l'Union européenne en 1998, ce qui classe la France au troisième rang derrière l'Allemagne et l'Espagne.

[1] Direction des Statistiques de la Commission européenne.

Les séries mensuelles du prix des céréales utilisées au SCEES[2], service statistique du ministère de l'Agriculture, pour l'analyse de la conjoncture sont fournies par l'ONIC[3], office interprofessionnel chargé de l'organisation du marché français des céréales. Elles sont constituées par un relevé de cotations effectué pour chacune des céréales auprès de sources variées : coopératives de producteurs, exportateurs, courtiers, etc. La base temporelle pour ces séries est la campagne de commercialisation. Annuelle pour les céréales, elle débute dès les premières récoltes et s'achève avant les récoltes de l'année suivante, soit pour le blé tendre du mois de juillet au mois de juin suivant.

Dans le cadre de la Politique agricole commune (PAC) régissant le marché européen des produits agricoles, les céréales bénéficient d'un encouragement à la production réalisé au moyen d'un système de soutien des producteurs par les prix. Ce système permet aux céréaliers européens de bénéficier d'un prix minimum garanti auquel l'organisme d'intervention chargé de la régulation du marché, l'ONIC en France, s'engage à effectuer les achats nécessaires à la résorption d'un excédent de production. Cependant, cette politique de soutien par les prix élaborée dans les années 60 dans le contexte de la mise en place du Marché commun (1962) a été progressivement victime de son succès. La progression des rendements céréaliers et l'extension des surfaces conjuguées au maintien d'un prix d'intervention élevé par rapport au niveau du marché mondial ont conduit progressivement à la constitution de stocks de plus en plus importants dont le coût financier devenait excessif en regard du budget communautaire. La réforme de la PAC introduite en 1993 vise à la maîtrise de la production au moyen d'une baisse progressive du prix d'intervention compensée par des aides transitoires concédées aux exploitants en échange d'un gel partiel des surfaces mises en culture.

Durant la période étudiée, allant de la campagne 1990/91 à la campagne 1998/99, nous disposons de relevés hebdomadaires des prix de marché du blé tendre de qualité meunière destiné à l'alimentation humaine. Du fait de la réforme de la PAC, le dispositif de soutien public par les prix des céréales a connu une évolution importante : les baisses sensibles entre juillet 1993 et juillet 1995 du prix d'intervention, déterminant majeur du niveau des prix sur le marché européen, invitent à considérer l'hypothèse d'une rupture dans la série. Pour tester cette hypothèse nous étudierons le rapport du prix de marché au prix d'intervention.

Les fonctionnalités du module *Séries chronologiques* de *SPSS pour Windows* vont nous permettre de construire une modélisation des variations de l'indice de prix mensuel du blé tendre rapporté au prix d'intervention, considéré comme un processus temporel soumis à des perturbations aléatoires.

II) Modélisation des séries chronologiques

II.1) Introduction aux modèles ARIMA

La classe des modèles ***ARIMA*** [Box et Jenkins, 1976] a été introduite pour reconstituer le comportement de **processus** soumis à des chocs aléatoires[4] au cours du temps : entre deux observations successives d'une série de mesures portant sur l'activité du processus, un événement aléatoire appelé **perturbation** vient affecter le

[2] Service central des enquêtes et études statistiques.

[3] Office national interprofessionnel des céréales.

[4] La notion de choc aléatoire ou d'impulsion a été introduite dans l'étude des séries temporelles en 1927 par Yule (cf. notice biographique dans ce même numéro de la revue de Modulad).

comportement temporel de ce processus et ainsi modifier les valeurs de la **série chronologique** des observations. Les modèles *ARIMA* permettent de combiner trois types de processus temporels : les processus autorégressifs (*AR-AutoRegressive*), les processus intégrés (*I-Integrated*), et les moyennes mobiles (*MA-Moving Average*). Dans le cas le plus général, un modèle *ARIMA* combine les trois types de processus aléatoires[5], la contribution de chacun d'eux étant précisée par la notation *ARIMA(p,d,q)*, où ***p*** est l'ordre du processus autorégressif *AR(p)*, ***d*** le degré d'intégration d'un processus *I(d)*, et ***q*** l'ordre de la moyenne mobile *MA(q)*.

Les processus auto-régressifs

Pour un processus autorégressif, chaque valeur de la série est une combinaison linéaire des valeurs précédentes de la série. Si la valeur de la série à l'instant *t*, Y_t, ne dépend que de la valeur précédente Y_{t-1} à une perturbation aléatoire près ε_t, le processus est dit autorégressif du premier ordre et noté *AR(1)* :

$$Y_t = \phi_1 \times Y_{t-1} + \varepsilon_t$$

Le coefficient ϕ exprime la force de la liaison linéaire entre deux valeurs successives. Un **processus autorégressif** où la valeur de la série à l'instant *t*, Y_t, dépend des *p* précédentes valeurs est dit d'ordre *p* et noté *AR(p)*. Ainsi un processus *AR(2)* s'écrit :

$$Y_t = \phi_1 \times Y_{t-1} + \phi_2 \times Y_{t-2} + \varepsilon_t$$

On peut dire qu'un processus autorégressif possède une « mémoire » au sens où chaque valeur est corrélée à l'ensemble des valeurs qui la précède. Par exemple, dans un processus AR(1), la valeur à l'instant *t*, Y_t, est fonction de la valeur précédenteY_{t-1}, elle-même fonction de la valeur Y_{t-2}, elle-même fonction de la valeur Y_{t-3}, etc. Si la valeur absolue du coefficient de régression ϕ_1 est inférieure à 1 (autrement dit si $-1\langle \phi_1 \langle +1$), l'effet de chaque perturbation aléatoire sur le système tend à décroître au cours du temps. Un processus autorégressif d'ordre *p*, *AR(p)*, pourra être noté comme un modèle *ARIMA(p,0,0)*.

Les processus intégrés

Le comportement des séries chronologiques peut être affecté par l'effet cumulatif de certains processus. Par exemple, l'état des stocks est modifié à chaque instant par les consommations et les approvisionnements, cependant le niveau moyen de ces stocks dépend essentiellement de l'effet cumulé des changements instantanés sur la période entre deux inventaires. Même si sur le court terme les valeurs du stock peuvent fluctuer avec des aléas importants autour de cette valeur moyenne, le niveau de la série sur le long terme demeurera inchangé. Une série chronologique déterminée par l'effet cumulatif d'une activité appartient à la classe des **processus intégrés**. Même si le comportement d'une série est erratique, les différences d'une observation à la prochaine peuvent être relativement faibles voire osciller autour d'une valeur constante pour un processus observé à différents intervalles de temps. Cette **stationnarité** de la série des différences pour un processus intégré est une caractéristique importante du point de vue de l'analyse statistique des séries chronologiques. Les processus intégrés constituent l'archétype des séries non stationnaires.

Un exemple de processus *I(1)*, intégré d'ordre *1*, est la marche aléatoire définie par : $Y_t = Y_{t-1} + \varepsilon_t$

5 La caractérisation des processus aléatoires a été effectuée par Kolmogorov en 1933.

où la perturbation aléatoire ε_t est un bruit blanc[6]. On utilise le terme de marche aléatoire car la valeur courante est définie comme une étape aléatoire à partir de la valeur précédente. La marche aléatoire est également un processus autorégressif d'ordre *1*, *AR(1)*, dont le coefficient de régression ϕ_1 est égal à 1. Ainsi, la marche aléatoire possède une « mémoire parfaite » mais limitée à l'observation précédente. Un processus est intégré d'ordre *1*, noté *I(1)*, si la série des différences premières est stationnaire. De même un processus est intégré d'ordre *2*, noté *I(2)*, si la série des différences secondes (les différences des différences) est stationnaire. Un processus intégré d'ordre *d*, *I(d)*, pourra être noté comme processus *ARIMA(0,d,0)*.

Les moyennes mobiles

La valeur courante d'un processus de moyenne mobile est définie comme une combinaison linéaire de la perturbation courante avec une ou plusieurs perturbations précédentes. L'ordre de la moyenne mobile indique le nombre de périodes précédentes incorporées dans la valeur courante. Ainsi, une moyenne mobile d'ordre *1*, *MA(1)*, est définie par l'équation suivante : $Y_t = \varepsilon_t - \theta_1 \times \varepsilon_{t-1}$

Pour une moyenne mobile, chaque valeur est une moyenne pondérée des plus récentes perturbations tandis que pour un processus autorégressif c'est une moyenne pondérée des valeurs précédentes. L'effet d'une perturbation aléatoire décroît tout au long de la série au fur et à mesure que le temps s'écoule dans un processus autorégressif tandis que dans une moyenne mobile la perturbation aléatoire affecte la série temporelle pour un nombre fini d'observations (l'ordre de la moyenne mobile) puis au-delà cesse brutalement d'exercer une quelconque influence.

II.2) La méthodologie de Box et Jenkins

Dans la méthodologie d'analyse des séries chronologiques synthétisée par Box et Jenkins en 1976, on utilise ces trois types de processus pour construire un modèle restituant le mieux possible le comportement d'une série temporelle selon une procédure en trois étapes : *identification*, *estimation* et *diagnostic*, qu'il convient de réitérer jusqu'à ce que le résultat soit jugé satisfaisant.

L'identification

La première étape dans la méthodologie proposée par Box et Jenkins concerne la décomposition retenue de la série chronologique selon les trois types de processus en spécifiant les trois paramètres *p*, *d* et *q* du modèle *ARIMA(p,d,q)*. On suppose à cet instant que toute composante saisonnière a été éliminée de la série chronologique, les modèles avec saisonnalité impliquant la spécification d'un autre ensemble de paramètres qui seront abordés ultérieurement.

L'identification des processus autorégressifs et de moyennes mobiles susceptibles d'expliquer le comportement de la série temporelle suppose de vérifier tout d'abord la stationnarité de la série puisque les processus de base, qu'ils soient autorégressifs ou de moyennes mobiles, sont essentiellement stationnaires en raison des contraintes pesant sur leurs paramètres. Un processus est dit faiblement stationnaire si son espérance et sa variance sont constantes et si sa covariance ne dépend que de l'intervalle de temps :

[6] Un bruit blanc est un processus stationnaire dont les accroissements sont indépendants et stationnaires. Le modèle du « bruit blanc » constitue la référence pour les résidus d'un modèle correctement spécifié.

$$E[Y_t] = m$$
$$V[Y_t] = \sigma^2$$
$$\text{cov}[Y_t, Y_{t+\theta}] = \gamma_Y(\theta)$$

Si la série n'est pas stationnaire – c'est à dire si la moyenne de la série varie sur le court terme ou que la variabilité de la série est plus élevée sur certaines périodes que sur d'autres – il convient de transformer la série pour obtenir une série stationnaire. La transformation la plus courante est la différenciation de la série, opération où chaque valeur de la série est remplacée par la différence entre cette valeur et celle qui la précède. Transformation logarithmique ou bien racine carrée peuvent être utilisées en situation d'hétéroscédasticité, où la variance de la série n'est pas constante et dépend des valeurs prises, par exemple avec une forte volatilité pour des valeurs élevées et une faible volatilité pour des valeurs faibles.

Une fois obtenue la stationnarité de la série, l'étape suivante consiste à analyser le graphe de la fonction d'autocorrélation (FAC) et celui de la fonction d'autocorrélation partielle (FAP) afin de déterminer les paramètres *(p,d,q)* du modèle.

Le paramètre *d* est fixé par le nombre de différenciations effectuées pour rendre la série stationnaire, en règle générale une différenciation suffit : $d \in \{0,1,2\}$.

Une fois ce paramètre fixé, il convient de spécifier l'ordre *p* du processus auto-régressif et *q* celui de la moyenne mobile. Les **corrélogrammes**, graphes de la fonction d'autocorrélation et de la fonction d'autocorrélation partielle permettent selon leurs aspects d'identifier correctement les paramètres *p* et *q* dont les valeurs n'excèdent pas deux en règle générale : $p \in \{0,1,2\}$ et $q \in \{0,1,2\}$.

La **fonction d'autocorrélation**, notée FAC, est constituée par l'ensemble des autocorrélations $\rho_k = corr(Y_t, Y_{t-k})$ de la série calculées pour des décalages d'ordre *k*, $k \in \{1,\ldots,K\}$. Le décalage maximum *K* admissible pour que le coefficient d'autocorrélation ait un sens se situe en général entre $\frac{n}{6} \le K \le \frac{n}{3}$, où n est le nombre d'observations temporelles. Pour $n \ge 150$, on prendra $K = \frac{n}{5}$.

Le **coefficient d'autocorrélation d'ordre k**, ρ_k, peut être estimé par :

$$r_k = \frac{\sum_{t=k+1}^{n}(y_t - \bar{y}_1)(y_{t-k} - \bar{y}_2)}{\sqrt{\sum_{t=k+1}^{n}(y_t - \bar{y}_1)^2 \sum_{t=k+1}^{n}(y_{t-k} - \bar{y}_2)^2}} \text{ avec } \bar{y}_1 = \frac{1}{n-k}\sum_{t=k+1}^{n} y_t \text{ et } \quad \bar{y}_2 = \frac{1}{n-k}\sum_{t=k+1}^{n} y_{t-k}$$

Sous l'hypothèse H_0 « $\rho_k = 0$ », la statistique $t_c = \frac{|r_k|}{\sqrt{1-r_k^2}}$ suit une loi de Student à $n-2$ degrés de libertés. Si la valeur calculée t_c est supérieure au quantile $\alpha/2$ d'une loi de Student à $n-2$ degrés de liberté $t_c \rangle t_{n-2}^{\alpha/2}$, alors l'hypothèse H_0 est rejetée au seuil α (test bilatéral).

La **fonction d'autocorrélation partielle**, notée FAP, est constituée par l'ensemble des autocorrélations partielles, le coefficient d'autocorrélation partielle mesurant la corrélation entre les variables entre Y_t et Y_{t-k}, l'influence de la variable Y_{t-k-i} étant contrôlée pour $i<k$.

Outre les coefficients de corrélation, les corrélogrammes affichent les intervalles de confiance à 95 %, qui permettent de déterminer quels sont les coefficients statistiquement significatifs à prendre en compte.

L'interprétation des corrélogrammes pour la spécification des processus *AR* et *MA* est généralement gouvernée par les règles suivantes :

- les processus autorégressifs d'ordre *p*, *AR(p)*, présentent une fonction d'autocorrélation dont les valeurs décroissent exponentiellement avec des alternances possibles de valeurs positives et négatives ; leur fonction d'autocorrélation partielle présente exactement *p* pics aux *p* premières valeurs du corrélogramme d'autocorrélation partielle ;
- les processus de moyenne mobile d'ordre *q*, *MA(q)*, présentent exactement *q* pics aux *q* premières valeurs du corrélogramme de la fonction d'autocorrélation et des valeurs exponentiellement décroissantes de la fonction d'autocorrélation partielle ;
- si la fonction d'autocorrélation décroît trop lentement, on conseille de différencier la série avant l'identification du modèle ;
- les processus mixtes de type *ARMA* peuvent présenter des graphes d'autocorrélation et d'autocorrélation partielle plus complexes à interpréter et nécessiter plusieurs itérations de type *identification-estimation-diagnostic*.

L'estimation

La procédure **Arima** du module *SPSS Séries chronologiques* permet selon un algorithme rapide d'estimation du maximum de vraisemblance [Mélard, 1984] d'estimer les coefficients du modèle que vous avez identifié au préalable en fournissant les paramètres *p*, *q* et *d*. L'exécution de la procédure ajoute de nouvelles séries chronologiques représentant les valeurs ajustées ou prédites par le modèle, les résidus (erreurs d'ajustement) et les intervalles de confiance de l'ajustement à votre fichier de données courant. Ces séries pourront être utilisées dans une nouvelle itération de type *identification-estimation-diagnostic*.

Le diagnostic

Dans cette étape finale du triptyque *identification-estimation-diagnostic* de la méthode de Box et Jenkins, les principales vérifications à effectuer portent sur les éléments suivants :

- les valeurs des fonctions d'autocorrélation et d'autocorrélation partielle de la série des résidus doivent être toutes nulles ; si les autocorrélations d'ordre 1 ou 2 diffèrent significativement de 0, alors la spécification *(p,d,q)* du modèle *ARIMA* est probablement inadaptée ; cependant, une ou deux autocorrélations d'ordre supérieur peuvent par aléas dépasser les limites de l'intervalle de confiance à 95 % ;
- les résidus ne doivent présenter aucune configuration déterministe : leurs caractéristiques doivent correspondre à celle d'un bruit blanc. Une statistique couramment utilisée pour tester un bruit blanc est le ***Q'*** de Box et Ljung, connue également comme la statistique de Box et Pierce modifiée. La valeur du *Q'* peut être vérifiée sur une base comprise entre un quart et la moitié des observations et ne doit pas être significative pour que l'hypothèse du bruit blanc puisse être conservée pour la série des résidus. Cette vérification peut facilement être effectuée en utilisant la procédure *SPSS Autocorrelation* qui donne la statistique de Box et Ljung ainsi que sa significativité à chaque pas du décalage dans le corrélogramme de la fonction d'autocorrélation.

Dans l'approche classique de Box et Jenkins, on examine également l'erreur-type des coefficients du modèle en vérifiant leur significativité statistique. Dans le cas d'un surajustement des données par un modèle trop complexe, certains coefficients peuvent ne pas être statistiquement significatifs et doivent donc être abandonnés.

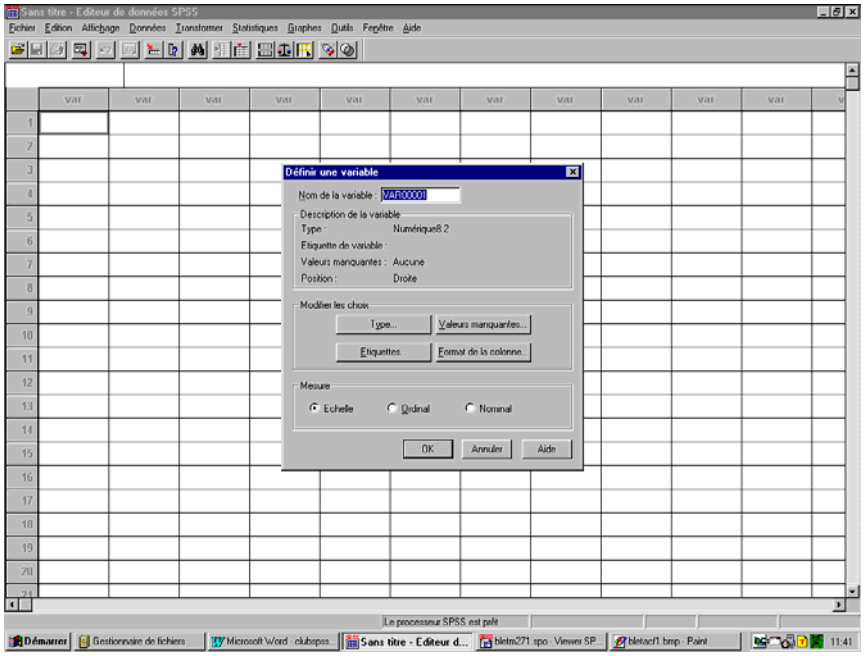
II.3) Utilisation de la procédure ARIMA

Les données, extraites d'un rapport de stage effectué au SCEES sur la prévision du prix des produits agricoles [Cadilhac et Martinot, 2000], concernent donc le rapport des moyennes mensuelles des relevés hebdomadaires du prix de marché au niveau de prix d'intervention fixé pour la campagne de commercialisation du blé tendre pour la période allant de la campagne 1990/91 à la campagne 1998/99. Les relevés hebdomadaires du prix de marché s'entendent « Départ Eure et Loire » (27).

Saisie des données

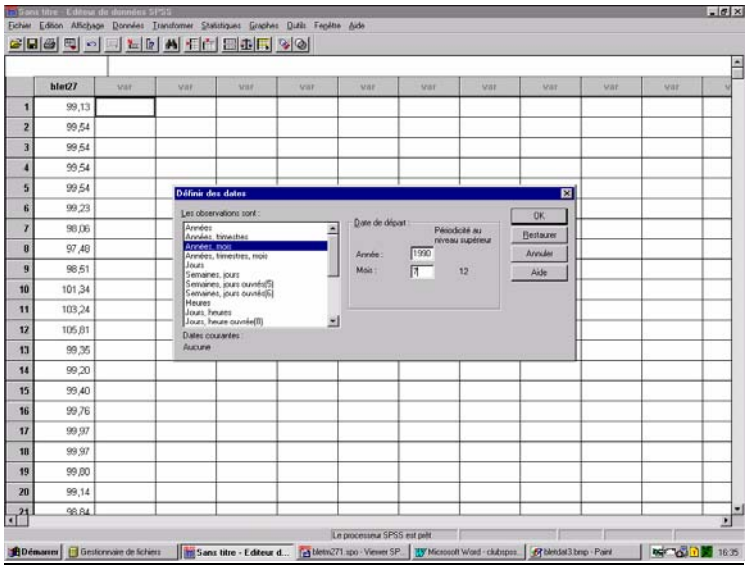
S'agissant d'une seule série mensuelle dont l'empan est limité, on peut utiliser les fonctionnalités de gestion des données offertes par *SPSS pour Windows* afin de créer le fichier des données qui sera exploité par la suite. Pour créer cette série temporelle, il convient de se positionner après le lancement du logiciel dans la fenêtre de l'éditeur des données et d'effectuer un double-clic sur l'entête de la première colonne (var).

Cette opération ouvre la boîte de dialogue Définir une variable comme suit :



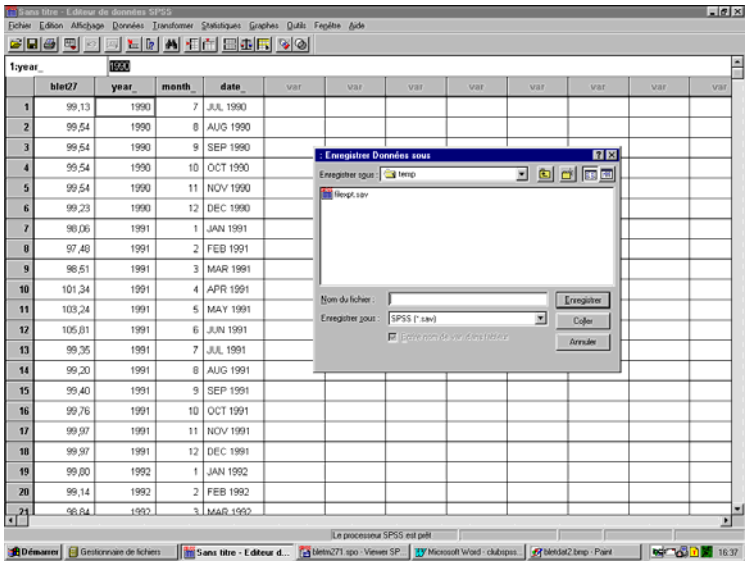
Il suffit alors d'indiquer le nom choisi pour la série dans la boîte textuelle Nom de la variable et de cliquer sur le bouton OK, pour que la première colonne de l'éditeur de données s'affiche avec le nom de variable choisi et que la première cellule de cette colonne s'affiche en surbrillance. On peut alors taper la première valeur numérique de la série (99,13) et valider cette saisie avec la touche Entrée pour passer à la ligne suivante, en répétant le procédé jusqu'à la dernière valeur de la série.

L'étape suivante consiste à affecter des dates et une périodicité à cette série en sélectionnant l'option Définir des dates du menu Données :



Dans la boîte de dialogue ainsi ouverte, il suffit d'indiquer le type de périodicité souhaitée (mensuelle) dans la liste définissant les observations (Les observations sont :) en choisissant l'item `Années, mois` puis en sélectionnant le contenu par défaut (1900) de la boîte textuelle `Année` qui vient s'afficher définir l'année initiale de la série (1990) par modification ainsi que le mois initial (juillet) selon le même procédé dans la boîte textuelle correspondante (Mois :), avant de cliquer sur le bouton OK pour valider le choix de ces paramètres définissant la base temporelle et la périodicité de la série. Il s'ensuit la création de trois variables générées par le système donnant l'année (year_), le mois (month_) puis la date (date_) associant une référence temporelle à chaque valeur de la série.

Pour sauvegarder l'ensemble de ces éléments, il convient d'utiliser l'option Enregistrer du menu Fichier qui ouvre la boîte de dialogue Enregistrer Données sous :



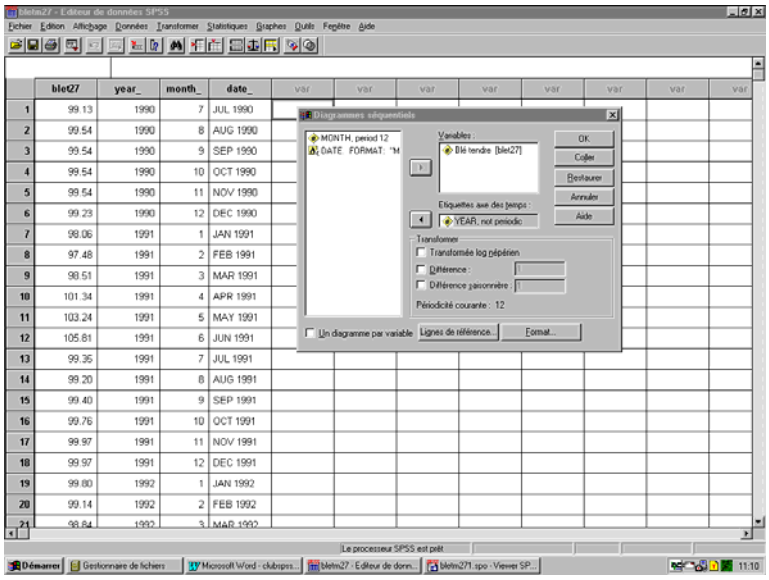
Il suffit alors de spécifier le Nom du fichier (blet27m.sav) dans la boîte textuelle prévue à cet effet. Le format du fichier ainsi créé est par défaut le format de gestion des données spécifique à *SPSS* (extension .sav).

Vérifier ou obtenir la stationnarité

Avant d'utiliser la procédure Arima, il faut tout d'abord examiner la série et vérifier sa stationnarité avec un graphique temporel. Si la moyenne de la série ou sa variance présente une variation au cours du temps, il faut alors différencier la série ou utiliser une transformation qui rende le produit stationnaire.

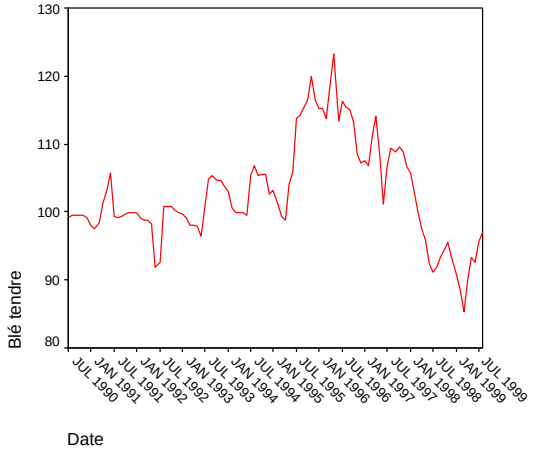
Pour obtenir un diagramme séquentiel, il suffit de choisir les options suivantes du menu : Graphes

Diagramme séquentiel



pour ouvrir la boîte de dialogue. Afin d'obtenir un graphique temporel, il suffit alors de transférer la variable blet27 (blé tendre) avec le bouton associé à la liste de variables et de sélectionner la variable year_ comme étiquette de l'axe des temps. On obtient ainsi un graphique temporel donnant l'allure générale de la série :

Figure 1 : graphique du prix de marché du blé tendre rapporté au prix d'intervention



Apparemment, cette série n'est pas stationnaire : elle semble présenter d'une part une rupture de tendance et d'autre part une volatilité des valeurs qui s'accroît au cours du temps. Ce graphe séquentiel permet une première approche intuitive du phénomène de rupture dans la série, introduit par la réforme de la PAC avec la baisse du prix d'intervention à partir de juillet 1993, en remarquant que la variabilité des valeurs de la série est plus importante après cette date. Il convient de préciser cette intuition en étudiant les corrélogrammes de la série selon deux bases temporelles distinctes : avant et après juillet 1993.

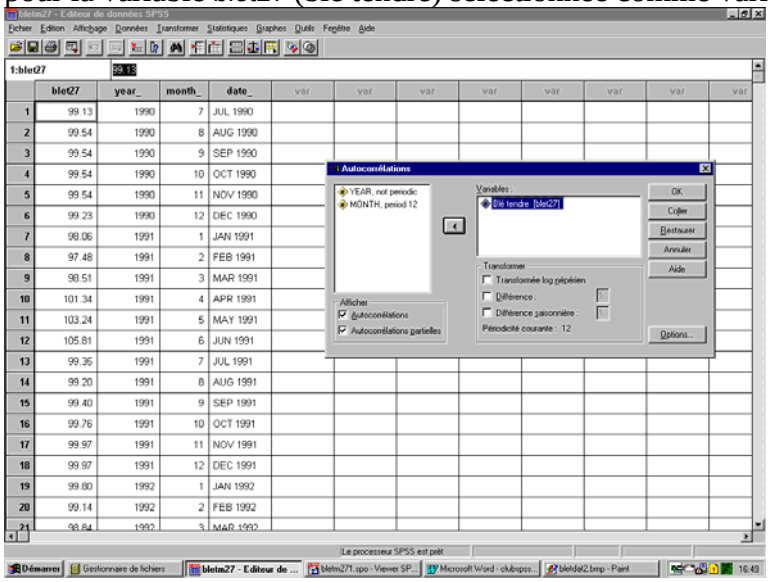
Analyse des fonctions d'autocorrélation

Pour obtenir le graphe de la fonction d'autocorrélation (FAC) et celui de la fonction d'autocorrélation partielle (FAP), il suffit de choisir à partir du menu général les options suivantes : Graphes

Séries chronologiques

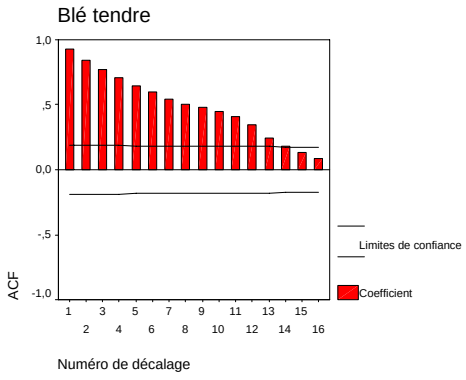
Autocorrélations…

Dans la boîte de dialogue ainsi ouverte, il convient alors de cocher les options d'affichage Autocorrélations et Autocorrélations partielles afin d'obtenir ces deux graphiques pour la variable blet27 (blé tendre) sélectionnée comme variable de l'analyse.

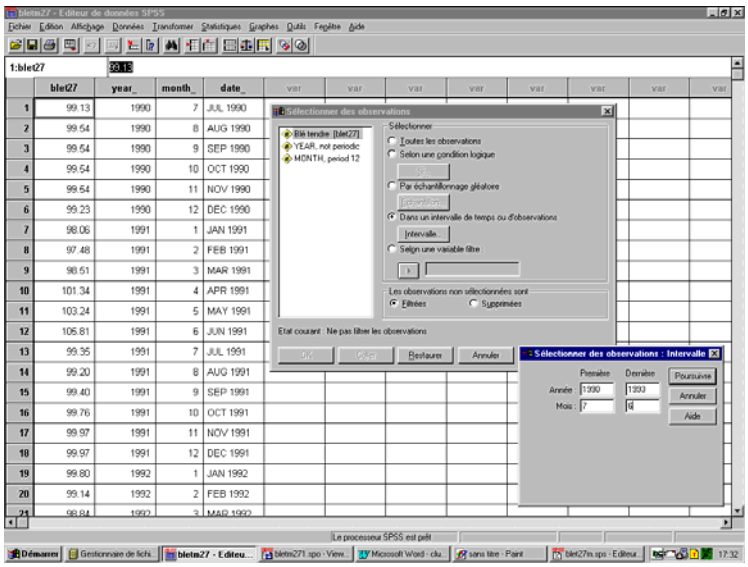


La décroissance lente du graphe de la fonction d'autocorrélation obtenu ci-après met clairement en évidence la non-stationnarité de la série.

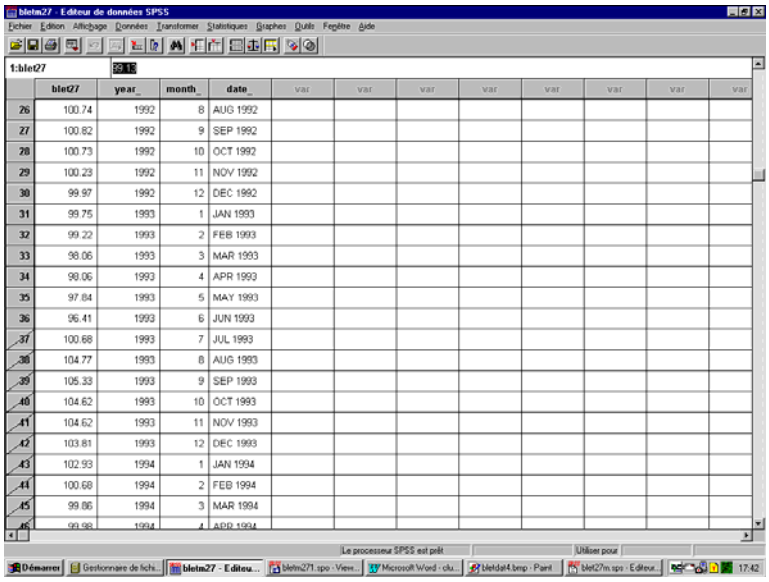
Figure 2 : corrélogramme du blé tendre (FAC – fonction d'autocorrélation)



L'étude de la série selon deux bases temporelles distinctes (avant et après juillet 1993) peut s'effectuer aisément en utilisant la fonction Sélectionner des observations… du menu Données, pour indiquer que la base temporelle d'analyse de la série est en premier lieu la période allant de juillet 1990 à juin 1993. Dans la boîte de dialogue ainsi ouverte, il suffit alors de choisir l'option Dans un intervalle de temps ou d'observations puis de spécifier, au moyen du bouton `Intervalle…`, la Première et la Dernière observation de l'intervalle d'étude choisi pour la variable blet27.



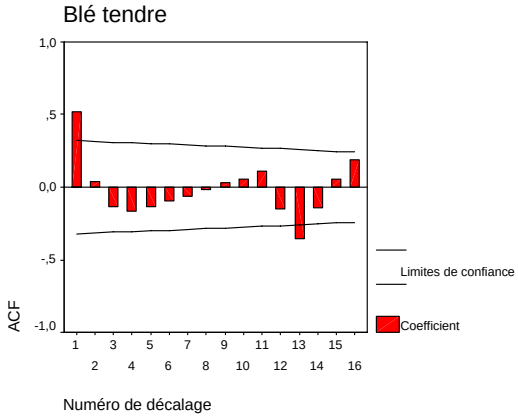
Les analyses ultérieures à ce filtrage des observations (l'option par défaut Filtrées n'a pas été modifiée pour les observations non sélectionnées) porteront seulement sur l'ensemble des observations sélectionnées, soit les 36 premières valeurs de la série. Les valeurs filtrées (inactivées) possèdent un marquage spécifique ainsi que le montre l'extrait ci-après du fichier des données.



Au vu du corrélogramme (fonction d'autocorrélation - FAC) obtenu après sélection de cette base temporelle, on peut affirmer que la série ne présente pas de

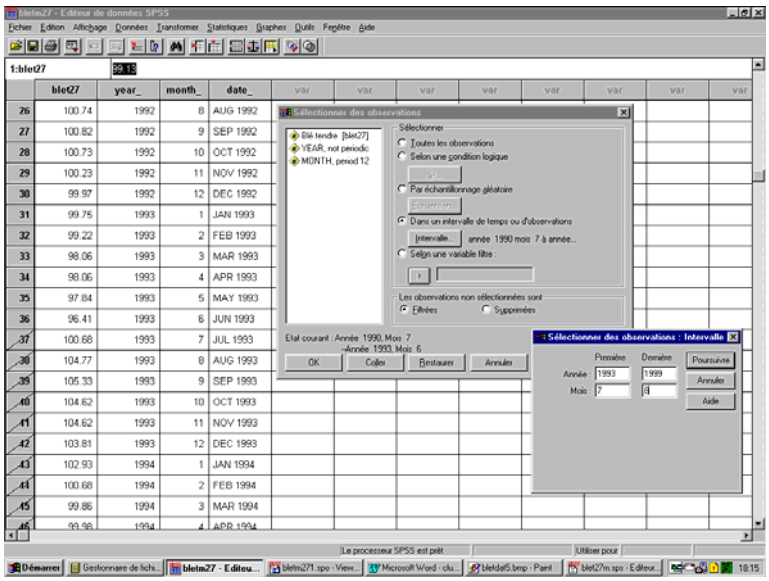
tendance marquée sur cette période allant de juillet 1990 à juin 1993. Une telle série possède une autocorrélation significative à l'ordre 1 et semble présenter une périodicité annuelle :

Figure 3 : corrélogramme (FAC) du blé tendre sur la base 1990-1993



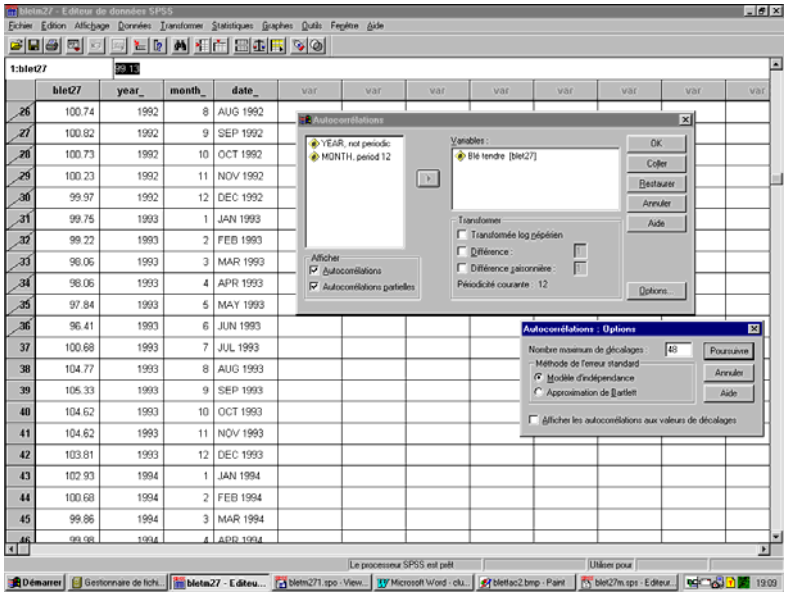
Si la modélisation d'une telle série peut permettre d'analyser les évolutions passées, elle ne répond cependant pas directement à notre objectif premier de prévision des prix agricoles comme aide à l'analyse conjoncturelle.

Puisque ces observations ont été simplement filtrées mais qu'elles n'ont pas été supprimées, on peut alors basculer sur l'autre partie de la série en modifiant la sélection courante comme suit :



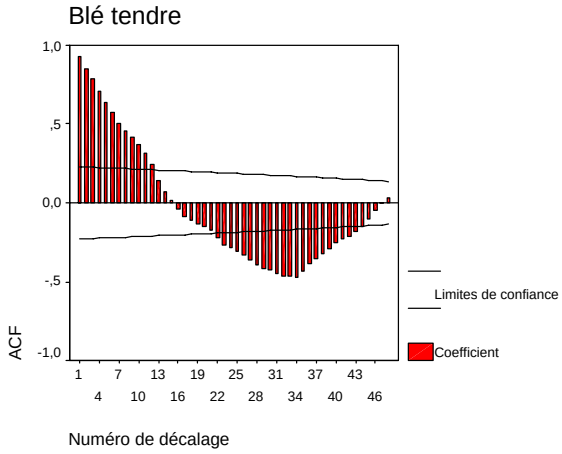
Après validation de ces choix, les observations actives vont du mois de juillet 1993 au mois d'août 1999. Les analyses porteront donc désormais sur les cinq dernières campagnes de commercialisation ainsi qu'en témoigne ci-après le marquage des observations effectué au niveau des numéros de séquence. Pour cette seconde analyse, on peut envisager d'étendre le calcul des coefficients d'autocorrélation avec des ordres de décalage plus élevés pour être en mesure de percevoir d'éventuels phénomènes périodiques. Il suffit d'augmenter l'ordre du décalage (par exemple, sur 48 mois au lieu

de 16) au niveau de la boîte d'options du module ACF de calcul de la fonction d'autocorrélation, ouverte par le bouton correspondant Options..., comme suit :



Examinons alors le graphe de la fonction d'autocorrélation obtenue sur la base temporelle allant du mois de juillet 1993 au mois d'août 1999 :

Figure 4 : corrélogramme (FAC) du blé tendre sur la base 1993-1999

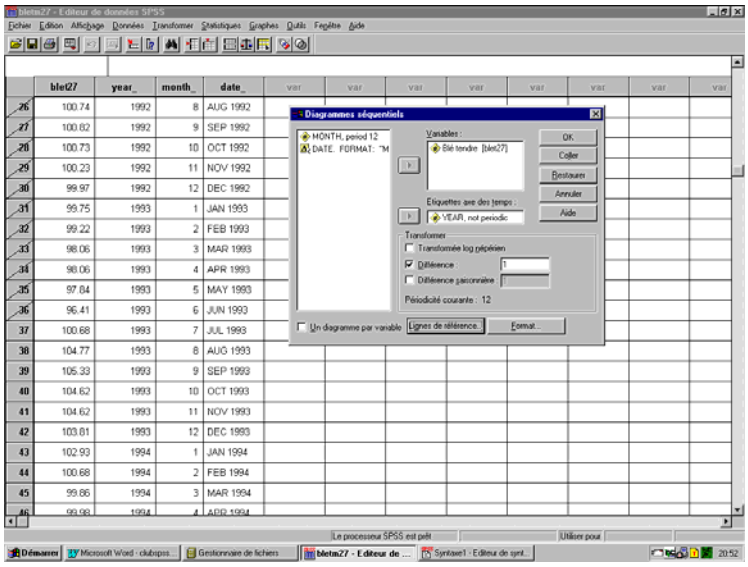


La décroissance linéaire de la fonction d'autocorrélation indique la présence d'une tendance dans la série : nous avons donc affaire sur cette période 1993-1999 pour la fixation du niveau des prix de marché relativement au prix d'intervention du blé tendre à un processus essentiellement non-stationnaire. Cela n'a rien d'étonnant : les chroniques économiques présentent souvent des éléments de non-stationnarité qui concernent soit leur moment du premier ordre (espérance mathématique) dus à une rupture de la série, une tendance ou la présence d'un cycle long, soit leur moment du second ordre (variance ou covariance) dus à des modifications de structure ou à une saisonnalité explosive.

Selon la terminologie introduite par [Nelson et Plosser, 1982], on distingue essentiellement deux types de non-stationnarité : une **non-stationnarité** de type **déterministe** (notée ***TS*** pour *Trend Stationnary*) et une **non-stationnarité** de type **aléatoire** (notée ***DS*** pour *Differency Stationnary*). Pour rendre stationnaire les processus

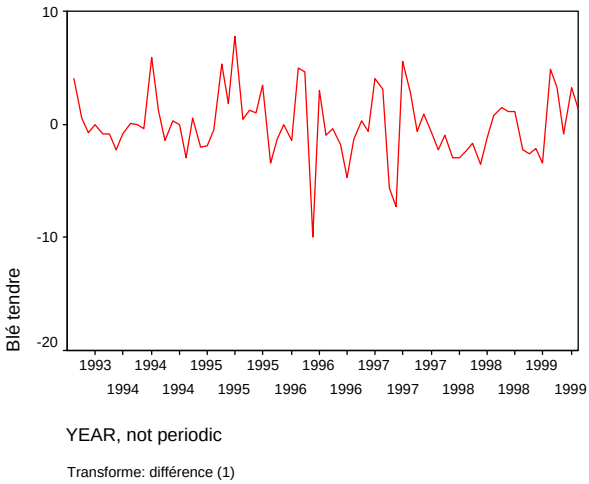
intégrés de type *TS*, on régresse selon une fonction polynomiale du temps $Y_t = f(t) + \varepsilon_t$ (le plus souvent linéaire, $Y_t = b_0 + b_1 t + \varepsilon_t$) en utilisant la méthode des moindres carrés ordinaires, les résidus ε_t constituant alors un processus ARMA stationnaire ou un bruit blanc. Pour stationnariser les processus intégrés de type *DS*, on préfère utiliser un filtre aux différences $[I - B]^d (Y_t) = \mu + \varepsilon_t$ (le plus souvent l'opérateur différences premières *I-B* défini par $[I - B](Y_t) = Y_t - Y_{t-1} = \mu + \varepsilon_t$), les résidus ε_t constituant alors un processus ARMA stationnaire ou un bruit blanc, β une constante réelle exprimant la dérive du processus et *d* l'ordre de différenciation du filtre.

Dans l'hypothèse où nous avons affaire à un processus non-stationnaire de type aléatoire, une pratique éprouvée consiste à appliquer l'opérateur aux différences premières *I-B*, puis à examiner le corrélogramme de la série différenciée. Pour visualiser la série différenciée, il suffit de spécifier le choix d'une transformation Différence à l'ordre 1 dans la boîte de dialogue Diagramme séquentiel obtenue par l'option Diagramme séquentiel… du menu Graphes, comme suit :



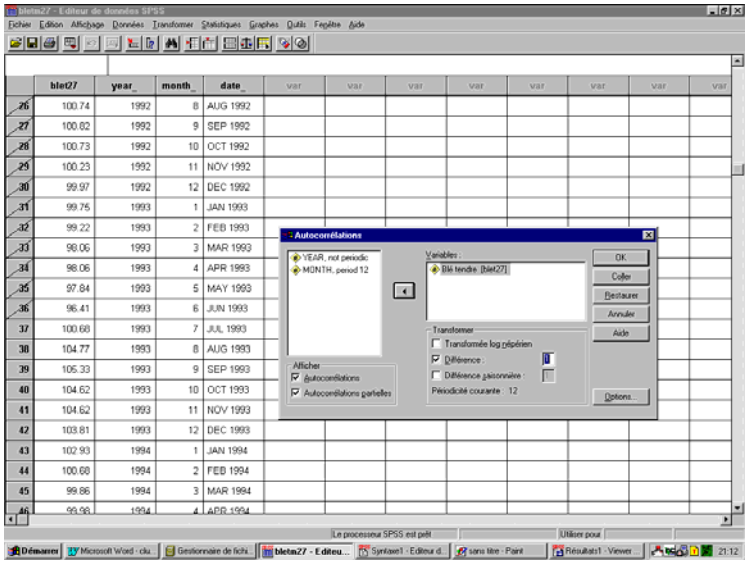
Le graphique ainsi obtenu montre l'effet de la différenciation : la série différenciée ne semble plus présenter de tendance et sa volatilité ne semble plus croître en fonction du temps.

Figure 5 : diagramme de la série des différences premières (blé tendre, 1993-1999)



On peut noter également que les valeurs de la série semblent se répartir de façon aléatoire de part et d'autre de l'axe des origines. La série différenciée semble donc stationnaire de moyenne nulle. Afin de conforter cette hypothèse, il convient cependant de tracer les corrélogrammes de la série différenciée.

Des options similaires permettent au niveau de la boîte de dialogue Autocorrélations permettent de spécifier que l'on souhaite calculer les autocorrélations et les autocorrélations partielles sur la série des différences premières, comme l'indique la spécification suivante :



Le graphe de la fonction d'autocorrélation comme celui de la fonction d'autocorrélation partielle ne présentant pas de pics significatifs aux premiers ordres de décalages, les corrélogrammes ci-après nous invitent à tester l'hypothèse de bruit blanc sur la série différenciée. On pourrait également envisager de tester la présence d'une saisonnalité car les autocorrélations au voisinage des décalages d'ordre k=12, 24 ou 36 dépassent légèrement l'intervalle de confiance à 95 %.

Figure 6 : corrélogramme (FAC) de la série différenciée (blé tendre, 1993-1999)

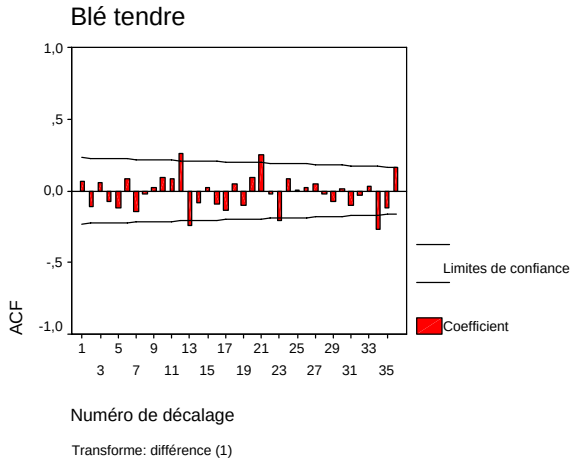
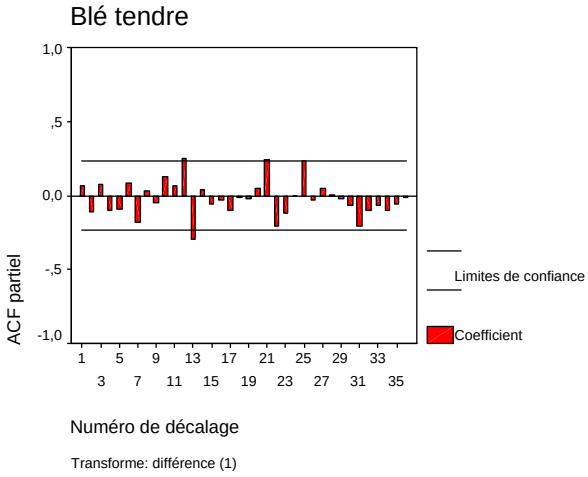


Figure 7 : corrélogramme (FAP) de la série différenciée (blé tendre, 1993-1999)

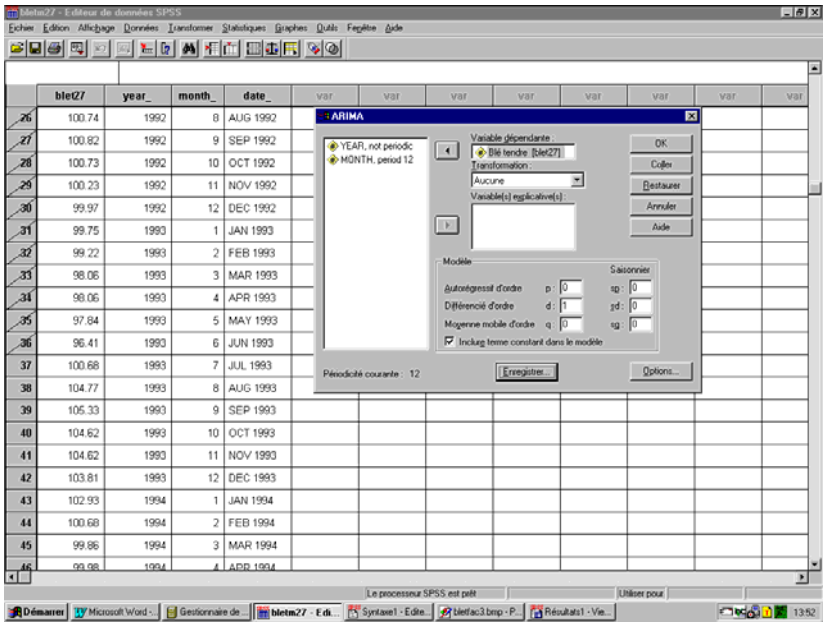


Estimation des coefficients du modèle

Dans l'hypothèse où nous nous sommes placés, le modèle à tester est un processus intégré d'ordre 1, noté *I(1)* ou *ARIMA(0,1,0)*, définissant une marche aléatoire : $[I - B](Y_t) = Y_t - Y_{t-1} = \mu + \varepsilon_t$.

Pour estimer les coefficients d'un modèle *ARIMA(0,1,0)*, il suffit de choisir à partir du menu général les options suivantes : Statistiques
Séries chronologiques
ARIMA

La boîte de dialogue ainsi ouverte permet de spécifier les paramètres du modèle ARIMA à estimer :



La Variable dépendante (à expliquer) est la variable Y_t, soit blet27 (blé tendre) et l'ordre de différenciation du filtre aux différences est $d = 1$. Le seul paramètre à estimer de la

modélisation portant sur la série différenciée $(Y_t - Y_{t-1}) = \mu + \varepsilon_t$ est la constante μ. On conserve donc l'option par défaut `Inclure terme constant dans le modèle`.

Les résultats standards donnés par la procédure Arima de *SPSS* sont les suivants :

```
>Warning # 16445
>Since there is no seasonal component in the model, the seasonality
of the
>data will be ignored.

MODEL:  MOD_7
Model Description:
Variable:   BLET27
Regressors: NONE
Non-seasonal differencing: 1
No seasonal component in model.

Parameters:
CONSTANT ________  < value originating from estimation >
95,00 percent confidence intervals will be generated.
Split group number: 1  Series length: 74
No missing data.
Melard's algorithm will be used for estimation.

Conclusion of estimation phase.
Estimation terminated at iteration number 0 because:
   No ARMA parameters were available for estimation.

FINAL PARAMETERS:
Number of residuals  73
Standard error       3,0300279
Log likelihood       -184,00826
AIC                  370,01651
SBC                  372,30697

            Analysis of Variance:
              DF  Adj. Sum of Squares    Residual Variance
Residuals     72            661,03699            9,1810694

           Variables in the Model:
                      B         SEB      T-RATIO   APPROX. PROB.
CONSTANT    -,04972603   ,35463795   -,14021632       ,88888060

Regressor Covariance Matrix:
             CONSTANT
CONSTANT    ,12576807

Regressor Correlation Matrix:
             CONSTANT
CONSTANT    1,0000000

The following new variables are being created:
  Name        Label
  FIT_1       Fit for BLET27 from ARIMA, MOD_7 CON
  ERR_1       Error for BLET27 from ARIMA, MOD_7 CON
  LCL_1       95% LCL for BLET27 from ARIMA, MOD_7 CON
  UCL_1       95% UCL for BLET27 from ARIMA, MOD_7 CON
  SEP_1       SE of fit for BLET27 from ARIMA, MOD_7 CON
```

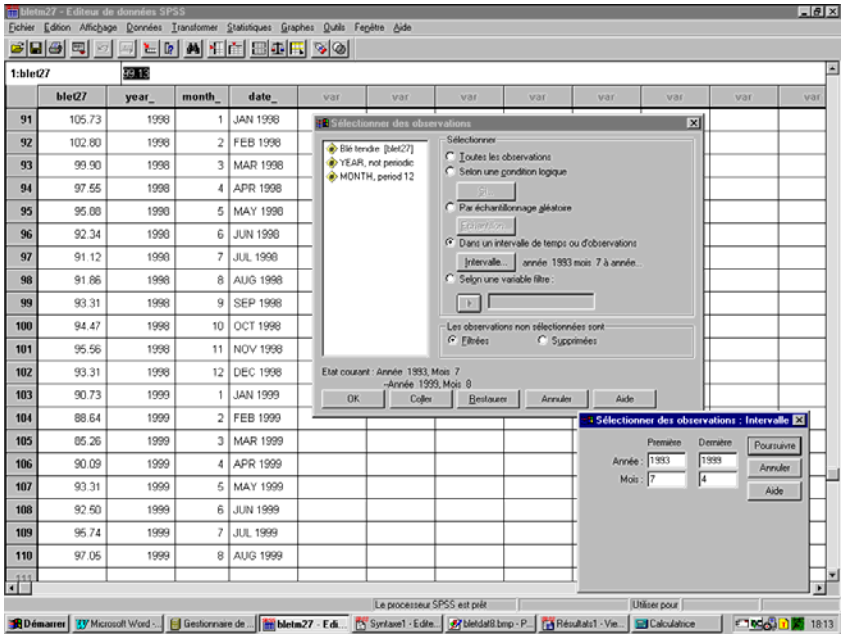
- un premier message d'avertissement signale qu'aucune composante saisonnière n'a été spécifiée dans ce modèle, l'estimation fournie ne prend donc en compte un éventuel caractère saisonnier des données ;
- la procédure Arima de *SPSS* affiche également un message avertissant que l'estimation est effectuée sans aucune itération puisque le modèle spécifié ne comporte pas de paramètre *ARMA* (autorégressifs ou de moyenne mobile) ;
- l'estimation de la constante, seul coefficient à estimer du modèle, vaut $\mu \approx -0{,}0497$ avec un écart-type estimé $\sigma_\mu \approx 0{,}3546$; la valeur de la statistique de Student calculée vaut donc $t_c = \frac{\mu}{\sigma_\mu} \approx -0{,}1402$, correspondant à un seuil de rejet de l'hypothèse nulle de $\alpha \approx 0{,}89$; ce niveau de risque de première espèce de près de 90 % chances de se tromper en rejetant H_0 conduit à conserver l'hypothèse de la nullité de ce coefficient.

Selon cette estimation, le modèle serait une marche aléatoire sans dérive ($\mu = 0$) soit $Y_t - Y_{t-1} = \varepsilon_t$. Cependant, pour confirmer un tel modèle, il convient de vérifier l'hypothèse selon laquelle les résidus ε_t forment un bruit blanc.

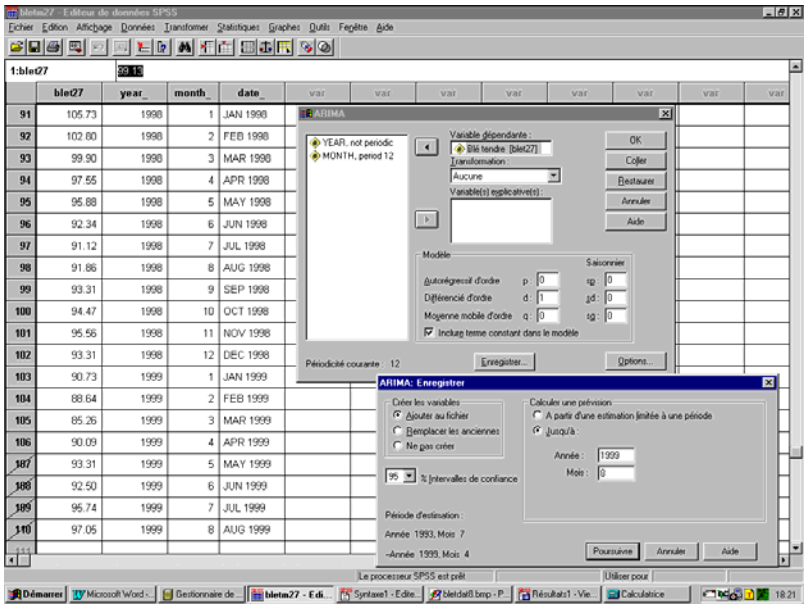
Validation du modèle

Rappelons qu'un **bruit blanc** est un processus $\{\varepsilon_t, t \in Z\}$ strictement stationnaire si et seulement si, sur n'importe quel sous-ensemble de temps $\{t_1, t_2, \ldots, t_n\}$, les variables aléatoires $\{\varepsilon_{t_1}, \varepsilon_{t_2}, \ldots, \varepsilon_{t_n}\}$ sont **indépendantes et identiquement distribuées** (**iid**). Donc, tous les coefficients d'autocorrélations de ce processus doivent être nuls $\rho_k = 0, \forall k \geq 1$. Le corrélogramme empirique des réalisations d'un tel processus aléatoire apparaît donc comme plat (coefficients empiriques d'autocorrélation proches de 0 $r_k \approx 0$, aux fluctuations d'échantillonnage près). Il existe une batterie très complète de tests spécifiquement construits pour valider ou invalider l'hypothèse du bruit blanc [Bourbonnais et Terraza, 1998]. D'autre part, un certain nombre de vérifications peuvent être également réalisées graphiquement.

Si le modèle est correctement spécifié, l'ajustement réalisé selon ce modèle (variable FIT_1) doit suivre les évolutions de la série empirique (blet27); une première vérification graphique consiste à projeter sur un même diagramme les valeurs observées et les valeurs prédites sur un intervalle de validation. Pour ce faire, il convient de réserver avant estimation cet intervalle de validation qui dans le cas présent porte sur le court terme du mois de mai 1999 au mois d'août 1999, soit 4 mois, en définissant l'intervalle d'estimation du mois de juillet 1993 au mois d'avril 1999 comme suit :

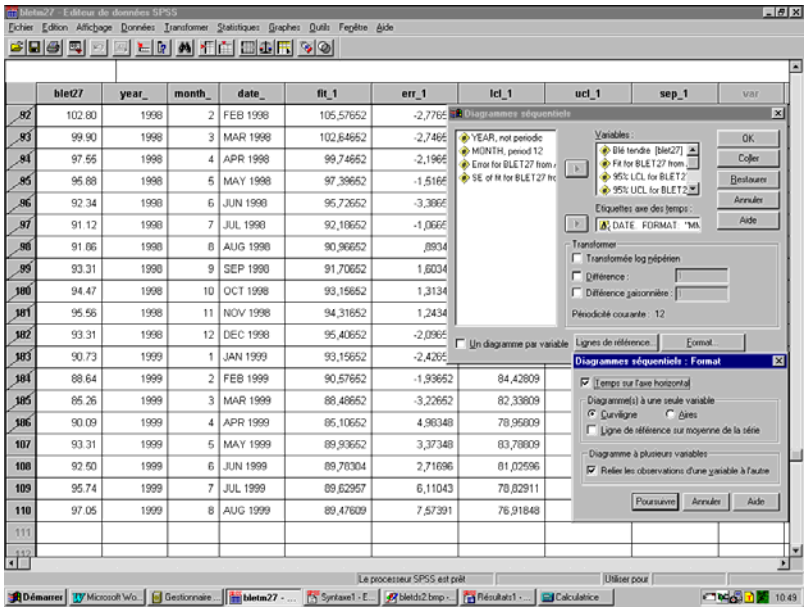


On peut constater sur la capture d'écran suivante que les quatre dernières observations ont été inactivées. Il suffit alors de réitérer l'estimation du modèle sur cette nouvelle base temporelle, en spécifiant dans la boîte d'options d'enregistrement obtenue avec le bouton Enregistrer… que le calcul de la prévision n'est plus limitée à une période (option par défaut) mais s'étend jusqu'en août 1999, comme suit :



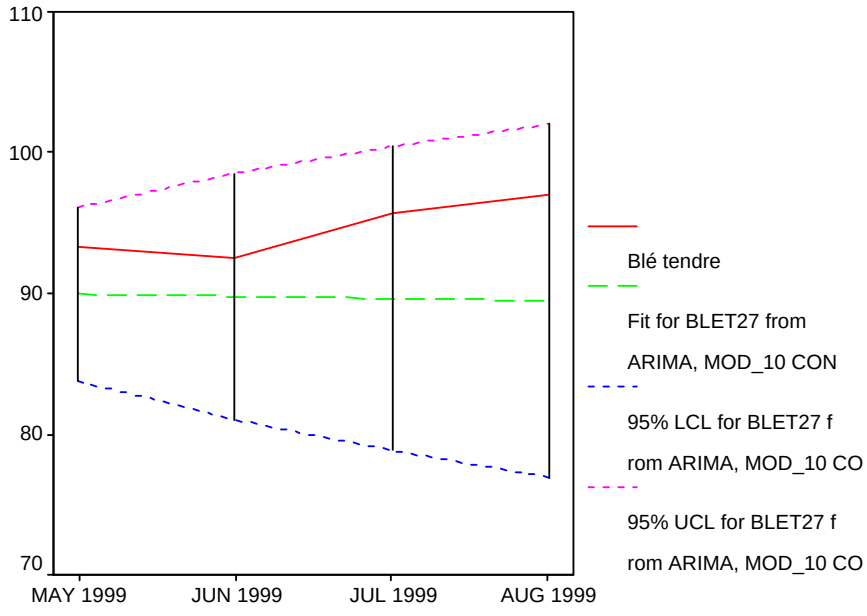
Bien qu'on obtienne alors une valeur de la constante $\mu \approx -0{,}1535$, le test de Student effectué montre que l'on peut conserver l'hypothèse de nullité pour ce coefficient du modèle.

En sélectionnant désormais l'intervalle de validation (mai 1999-août 1999), un diagramme séquentiel comprenant la série observée (blet27), la prévision fournie par l'ajustement (fit_1), la limite inférieure (lcl_1) respectivement supérieure (ucl_1) de l'intervalle de confiance à 95%, et en reliant les observations d'une variable à l'autre, comme suit :



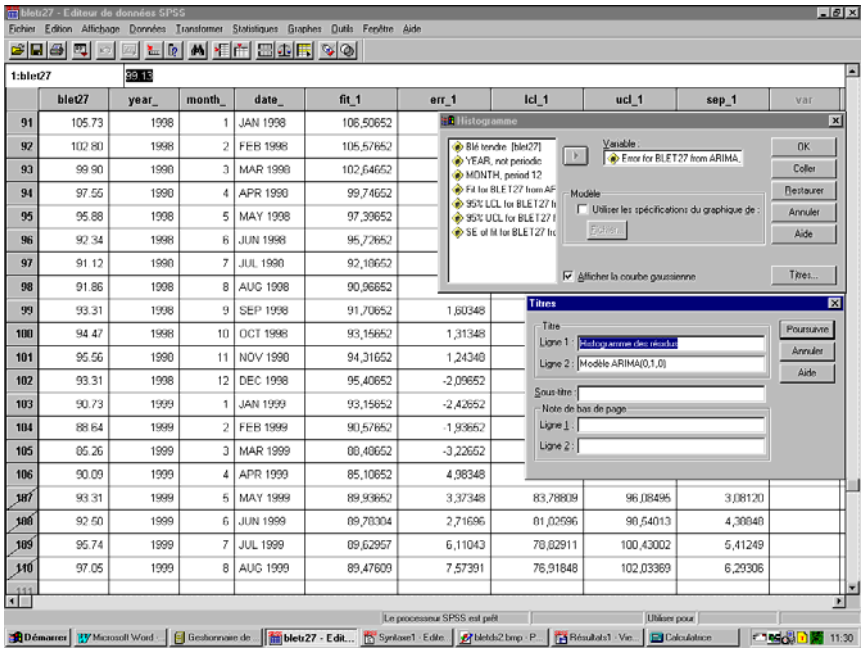
On obtient le diagramme suivant, où l'on peut constater visuellement que l'évolution observée de la série demeure dans l'intervalle de confiance à 95 % estimé à partir de l'ajustement :

Figure 8 : prévisions du modèle ARIMA(0,1,0), blé tendre, 1993-1999

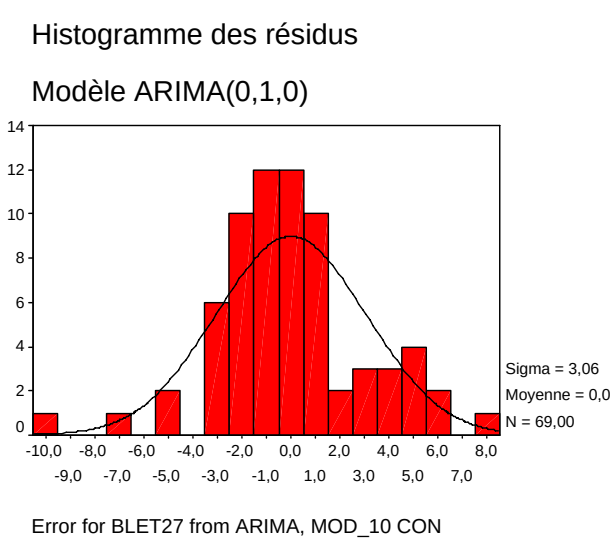


En sélectionnant à nouveau la base d'estimation du modèle (juillet 1993-avril 1999), on peut également observer la distribution des résidus autour de leur moyenne et ainsi avoir une première épreuve graphique de l'hypothèse de résidus gaussiens de

moyenne nulle. Il suffit d'utiliser l'option Histogramme… du menu Graphes, pour la variable err_1 des résidus du modèle en cochant l'option Afficher la courbe gaussienne et éventuellement spécifier un titre à ce graphique rappelant les paramètres du modèle, comme suit :



On obtient alors le graphique suivant qui donne également le nombre d'observations, la moyenne de la série des résidus ainsi que l'écart-type :



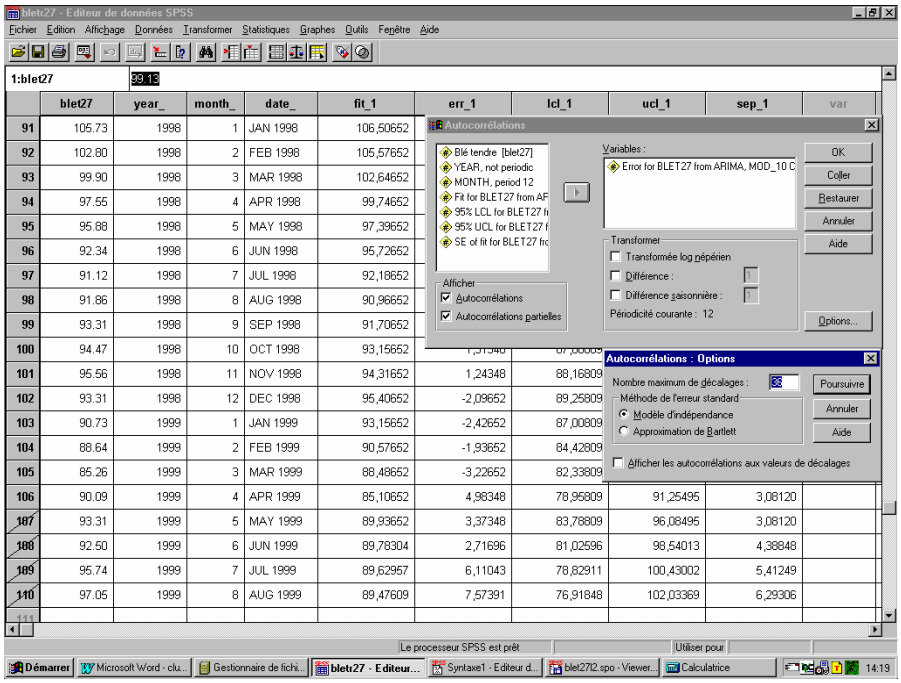
Pour un nombre d'observations suffisamment grand ($n \rangle 30$), la distribution des résidus suit une loi normale $N\left(\bar{\varepsilon}, \frac{\sigma_\varepsilon}{\sqrt{n}}\right)$ de moyenne $\bar{\varepsilon}$ et d'écart-type $\frac{\sigma_\varepsilon}{\sqrt{n}}$, dont le tracé figure sur l'histogramme.

Si ces résidus suivent un bruit blanc alors la fonction d'autocorrélation des résidus ne doit pas receler d'autocorrélations significativement différentes de 0. Selon [Box et Pierce, 1970], la statistique de Quenouille $Q = n\sum_{k=1}^{K} r_k^2(\varepsilon_t)$, fonction de la somme des carrés des autocorrélations, suit sous l'hypothèse nulle H_0 : « $\rho_1 = 0 \quad \rho_2 = 0 \quad \ldots \quad \rho_{k-1} = 0 \quad \rho_k = 0$ », c'est à dire en l'absence d'autocorrélation, une distribution du $\chi^2_{\{\upsilon\}}$ à $\upsilon = K - (p+q)$ degrés de libertés, où K figure le nombre de décalages considérés, p et q sont les ordres, respectivement du processus autorégressif *AR(p)* et de la moyenne mobile *MA(q)* du modèle ARIMA étudié. Le test de Box et Pierce permet de rejeter l'hypothèse H0 au seuil α si la statistique Q est supérieure au quantile $1-\alpha$ de la loi du $\chi^2_{\{\upsilon\}}$ correspondante.

Par la suite, on a montré [Ljung et Box, 1978] que la distribution de Q ne suit pas exactement celle du χ^2, principalement en raison de l'approximation effectuée pour estimer la variance des autocorrélations. Ljung et Box ont donc proposé une correction de la statistique de Quenouille : $Q' = n(n+2)\sum_{k=1}^{K} \frac{r_k^2(\varepsilon_t)}{n-k}$ dont la distribution suit plus fidèlement celle du $\chi^2_{\{K-(p+q)\}}$, bien que sa variance soit supérieure à celle d'un χ^2.

Comme pour le test de Box et Pierce, le test de Ljung et Box permet de rejeter H_0 si à un niveau de risque donné (par exemple 5 %), la statistique Q' est supérieure au quantile correspondant de la loi du χ^2 (dans ce cas $\chi^2_{\{\upsilon\}}[0{,}95]$).

Pour effectuer ce test de Ljung et Box, il suffit d'étudier la série des résidus (err_1)en utilisant la procédure Autocorrélations... de l'option Séries chronologiques du menu Graphes, spécifiée comme suit :



La sortie semi-graphique des résultats de la procédure donne à gauche du graphe de la fonction d'autocorrélation l'estimation du coefficient d'autocorrélation et son erreur-type, puis à droite la valeur calculée de la statistique de Ljung et Box ainsi que le risque de première espèce (colonne Prob.) :

```
MODEL:  MOD_16.
Variable:  ERR_1        Missing cases:  1     Valid cases:  69

Autocorrelations:   ERR_1   Error for BLET27 from ARIMA, MOD_10 CON

     Auto- Stand.
Lag  Corr.   Err. -1  -.75  -.5 -.25   0   .25  .5   .75   1   Box-Ljung  Prob.
                 +----+----+----+----+----+----+----+----+
  1   ,044   ,118                .    I*   .                      ,137   ,711
  2  -,105   ,117                .  **I    .                      ,949   ,622
  3   ,029   ,116                .    I*   .                     1,013   ,798
  4  -,066   ,115                .   *I    .                     1,339   ,855
  5  -,096   ,114                .  **I    .                     2,038   ,844
  6   ,099   ,113                .    I**  .                     2,796   ,834
  7  -,141   ,112                 .***I   .                     4,376   ,736
  8  -,035   ,112                 .  *I   .                     4,476   ,812
  9   ,010   ,111                 .   *   .                     4,485   ,877
 10   ,095   ,110                 .   I** .                     5,235   ,875
 11   ,100   ,109                 .   I** .                     6,075   ,868
 12   ,278   ,108                 .   I***.**                  12,710   ,390
 13  -,222   ,107                ****I   .                     17,034   ,198
 14  -,061   ,106                 .  *I   .                    17,361   ,237
 15   ,052   ,105                 .   I*  .                    17,604   ,284
 16  -,075   ,104                 .  *I   .                    18,118   ,317
 17  -,105   ,103                 . **I   .                    19,165   ,319
 18   ,063   ,102                 .   I*  .                    19,550   ,359
 19  -,100   ,101                 . **I   .                    20,526   ,364
 20   ,109   ,100                 .   I** .                    21,722   ,356
 21   ,248   ,099                 .   I***.*                   28,003   ,140
 22  -,049   ,098                 .  *I   .                    28,254   ,167
 23  -,182   ,097                ****I   .                     31,785   ,105
 24   ,068   ,096                 .   I*  .                    32,284   ,120
 25   ,004   ,095                 .   *   .                    32,286   ,150
 26   ,052   ,094                 .   I*  .                    32,595   ,174
 27   ,053   ,093                 .   I*  .                    32,926   ,200
 28  -,059   ,091                 .  *I   .                    33,338   ,224
 29  -,083   ,090                 . **I   .                    34,176   ,233
 30   ,035   ,089                 .   I*  .                    34,331   ,268
 31  -,098   ,088                 . **I   .                    35,581   ,261
 32  -,003   ,087                  .  *   .                    35,583   ,303
 33   ,051   ,086                  .  I*  .                    35,935   ,333
 34  -,290   ,085               ***.**I   .                    47,747   ,059
 35  -,063   ,083                  . *I   .                    48,312   ,066
 36   ,121   ,082                  .  I**.                     50,495   ,055
```

On peut vérifier sur ce graphique qu'aucun coefficient d'autocorrélation n'apparaît comme significativement différent de 0 au seuil de 5 %.

Au vu de l'ensemble de ces résultats, il semble donc qu'on puisse valider le modèle de marche aléatoire à dérive nulle que nous avions proposé. Conséquence pratique de ce modèle : la meilleure prévision possible à l'instant *t* est l'observation de la série à l'instant *t-1* et les changements observés dans cette série sont purement aléatoires (bruit blanc) et de moyenne nulle sur le long terme même si sur le court terme les différences observées d'un mois à l'autre peuvent paraître importantes en valeur absolue. Cette conséquence se déduit en fait du caractère aléatoire et stationnaire du processus de « bruit blanc » modélisant la série des différences.

Une telle conclusion peut désarçonner quelque peu le néophyte mais surprendra moins le conjoncturiste habitué au comportement des séries de prix pour des stocks de marchandises sur un marché où opère une multiplicité d'opérateurs et où l'information sur les prix est publique.

III) Références

Box G.E.P. et Jenkins G.M. 1976. *Time Series Analysis : Forecasting and Control*, Holden-Day, San Francisco.

Box G.E.P. et Pierce D.A. 1970. « Distribution of Residual Autocorrelations in Autoregressive Moving Average Time Series Models », *Journal of the American Statistical Association*, vol. 65 .

Bourbonnais R. et Terraza M. 1998. *Analyse des séries temporelles en économie*, PUF, Paris, 274 p.

Cadilhac F. et Martinot A. 2000 « Projet de fin d'études : prévision de prix de produits agricoles », DAF / SCEES/ Ministère de l'Agriculture et de la Pêche, 129 p.

David M. et Michaud J.C. (1989) *La prévision, approche empirique d'une méthode statistique*, Masson.

Lejeune M. 1997, *Statistique, cours B7 : séries chronologiques*, Collection des cours du CNAM, 63 p.

Ljung G.M. et Box G.E.P. 1978. « On a Measure of the Lack of Fit in Time Series Models », *Biometrika*, vol. 65, pp. 297-303.

Mélard G. 1984. « A Fast Algorithm for the Exact Likelihood of Autoregressive-Moving Average Models », *Applied Statistics*, vol. 33 n°1, pp. 104-119.

Nelson C.R. et Plosser C. 1982. « Trends and Random Walks in Macroeconomics Time Series : Some Evidence and Applications », *Journal of Monetary Economics*, vol. 10.

SPSS Inc. 1994. *SPSS Trends 6.1*, SPSS Inc., Chicago, 356 p.

Walter C. et Scheps R. 1998. « Marchés financiers, hasard et prévisibilité », *Les sciences de la prévision* , Seuil, Paris, pp. 125-146.

George Udny Yule
(18 février 1871- 26 juin 1951)