

LA DESAISONNALISATION : DES ORIGINES JUSQU'AUX NOUVEAUX LOGICIELS X12-ARIMA ET TRAMO-SEATS

Ketty Attal

Introduction

Le présent document avait initialement pour projet de présenter deux nouveaux logiciels de désaisonnalisation : X12-ARIMA et TRAMO-SEATS. Cependant, pour une meilleure compréhension de ces méthodes, il paraissait utile d'expliquer les évolutions en matière de désaisonnalisation qui ont conduit à celles-ci. Une première partie sera donc consacrée à une présentation générale des méthodes de désaisonnalisation, du point de vue historique et du point de vue de leur philosophie. La deuxième partie fera une description des deux nouveaux logiciels, qui reposent sur des principes très différents, mais qui ont en commun d'une part l'utilisation des modèles ARIMA, d'autre part le souci d'améliorer la "qualité" de la série qui doit être décomposée, par un traitement préalable assez poussé.

Un peu d'histoire

L'histoire des séries temporelles prend source dans l'Antiquité, avec l'observation des mouvements des astres. D'abord purement descriptive, l'étude des séries temporelles s'est ensuite enrichie d'une volonté d'analyse. Celle-ci consiste à isoler les différents mouvements qui entrent en jeu dans le mouvement général de la série. La désaisonnalisation, à savoir l'élimination des variations systématiques (saisonnnières) dans la série est un aspect de l'analyse des séries temporelles dont le développement a donc été étroitement lié aux progrès de celle-ci.

Les origines

Dès le 17^{ème} siècle, des astronomes ont fait remarquer qu'une série chronologique pouvait être considérée comme étant formée de composantes inobservables. La première formalisation de la décomposition d'une série remonte aux travaux de Fourier (1807) sur la décomposition d'une série périodique en une somme de fonctions trigonométriques. En 1847, Buys Ballot, météorologue, a recherché les

variations périodiques dans les séries et les estimait grâce à une analyse fondée sur un tableau dans lequel les données étaient arrangées en 12 (resp. 4) colonnes, pour chaque mois (resp. trimestre). Il modélisait la tendance par un polynôme et la saisonnalité par des indicatrices.

Au début du 20^{ème} siècle, le développement de la publication de statistiques économiques a orienté de nombreux travaux vers la construction de méthodes de désaisonnalisation. Deux types de modèles ont été définis :

- le modèle additif : $Y_t = T_t + C_t + S_t + I_t$

- le modèle multiplicatif : $Y_t = T_t \times C_t \times S_t \times I_t$

où T_t est la tendance, C_t le cycle, S_t la saisonnalité et I_t l'irrégulier.

L'écriture du modèle multiplicatif est ici très schématique puisqu'il a en réalité été défini de diverses façons :

par exemple $Y_t = T_t \times (1 + C_t) \times (1 + S_t) \times (1 + I_t)$.

Mills, en 1924, a défini chacune des composantes de la façon suivante :

T_t : tendance : mouvement lisse, régulier, de long terme.

S_t : variations saisonnières : périodiques, avec une période annuelle (12 mois ou 4 trimestres).

C_t : variations cycliques : périodiques de façon moins marquée mais néanmoins caractérisées par un degré important de régularité.

I_t : irrégulier : le résidu.

Les années 20-30 ont vu l'émergence des grands courants de la désaisonnalisation

Dans les années 20-30, les schémas de décomposition définis plus haut étaient bien acceptés et d'autres concepts ont été fixés : l'idée que la saisonnalité varie dans le temps ; la nécessité de calculer la tendance et le cycle lorsqu'on estime la composante saisonnière ; l'impossibilité de décrire les tendances et les cycles par des formules mathématiques explicites ; la nécessité de traiter les points extrêmes. Le problème résidait surtout dans la façon de désaisonnaliser. Les travaux de

l'époque étaient inspirés par deux grandes méthodes, dont on donnera une description pour le cas d'un modèle multiplicatif.

La première, mise au point par Persons (1919), appelée méthode des "link relatives", était basée sur les médianes, pour chaque mois, des rapports entre les observations pour ce mois et celles du mois précédent. Il construisait des indices à partir de ces médianes, en déduisait des indices de tendance par lesquels il divisait les données initiales pour avoir les indices saisonniers. Il obtenait les indices saisonniers définitifs en les ajustant de façon que leur somme fasse 1. Les données cvs (corrigées de variations saisonnières) étaient les données initiales divisées par ces indices.

La seconde méthode, dite méthode des rapports à la moyenne mobile, a été introduite par Macauley (1930) et était utilisée par la Réserve Fédérale (US). Elle s'appuyait sur le calcul d'une moyenne mobile centrée d'ordre 12 pour obtenir une estimation de la tendance. Le rapport entre les données originales et cette estimation fournissait une première estimation des composantes saisonnières. Pour en éliminer l'irrégulier, on calculait les médianes (ou moyennes) de ces composantes pour chaque mois. Puis on ajustait ces nouveaux indices pour que leur somme fasse 1 et l'on obtenait ainsi les indices saisonniers définitifs. Ces deux méthodes pouvaient être adaptées au cas d'un modèle additif.

Malgré quelques critiques sur les moyennes mobiles, au moment de la découverte de l'effet Slutsky-Yule en 1927 (introduction par les moyennes mobiles de cycles artificiels dans l'irrégulier), l'idée la plus répandue dans les années 30 était bien celle qu'il ne fallait pas spécifier une forme fonctionnelle pour la tendance et la saisonnalité (par exemple un polynôme et des fonctions harmoniques ou indicatrices), et donc que les méthodes empiriques étaient préférables.

Quant au problème des valeurs extrêmes, qui agissent fortement sur les moyennes, il était parfois résolu en utilisant des médianes ou des moyennes tronquées.

D'autres types de critiques se faisaient toutefois entendre à l'égard des méthodes empiriques. Ainsi Snow (1923) estimait que la logique de la méthode de Persons n'était pas claire ; Fisher (1937) déplorait que l'on applique des méthodes empiriques ad hoc alors qu'il existait des outils mathématiques adéquats. Des recherches ont donc été faites pour élaborer des méthodes basées sur la modélisation. Celles-ci s'appuyaient en général sur une décomposition additive de la série initiale ou d'une transformation simple de cette série, et des modèles étaient posés pour la série initiale et pour chacune des composantes. Le modèle pour la série initiale était estimé à partir des données mais les modèles pour les composantes inobservables ne pouvaient être estimés qu'à condition de faire certaines hypothèses arbitraires. Les diverses méthodes différaient par le type de modèle posé et par les hypothèses faites sur les composantes.

Les développements des méthodes de désaisonnalisation ont été liés à ceux de l'informatique et des méthodes d'analyse des séries temporelles.

Après la seconde guerre mondiale, l'évolution des travaux sur la désaisonnalisation a été très liée au développement de l'informatique. Celle-ci, en accroissant la rapidité d'exécution des calculs, permet d'élaborer des méthodes plus sophistiquées qui purent être appliquées même si le nombre de séries à traiter était très important. Les méthodes les plus répandues ont été développées au bureau du Census (US). C'est là en particulier qu'a été mise au point par Julius Shiskin (1954) la méthode Census II (version informatique de la Census I) qui est à l'origine de la méthode X11 (1965). De plus, les traitements informatiques semblaient plus objectifs puisque dans certaines méthodes, des décisions faisant appel au jugement de l'utilisateur ont pu être en partie automatisées. Enfin, l'informatique a facilité l'utilisation de régressions visant à corriger les effets de jours ouvrables, c'est-à-dire les effets liés au nombre de jours travaillés dans le mois.

Les nouvelles méthodes d'analyse des séries temporelles se sont retrouvées dans les évolutions de la désaisonnalisation. L'utilisation de l'économétrie en liaison avec celle-ci s'est développée avec l'idée que la saisonnalité dans une variable économique ne peut être considérée comme un phénomène isolé, mais peut être reliée à la saisonnalité d'autres variables économiques avec laquelle elle est liée, et également que la saisonnalité elle-même peut contenir de l'information sur les relations entre les séries.

Par ailleurs, l'analyse spectrale était au départ utilisée pour chercher des périodicités exactes alors qu'on était conscient que les cycles économiques n'avaient pas une périodicité exacte : cette technique ne semblait donc pas adaptée au problème de décomposition des séries. Les progrès dans la compréhension de l'analyse spectrale ont finalement permis de l'appliquer à la désaisonnalisation. Wiener (1939,1941) et Kolmogorov (1949) puis Hannan (1967), Cleveland et Tiao (1976), Bell (1984) ont pu, par extraction de signal, estimer la composante saisonnière.

A partir de la publication des travaux de Box et Jenkins sur les modèles ARIMA (1970), et grâce aux progrès de l'informatique, l'utilisation de ces modèles s'est répandue et a été orientée vers deux types de méthodes. D'une part elle a constitué un développement important de X11 qui a évolué vers X11-ARIMA (Dagum, 1975). Dans cette nouvelle version, les modèles ARIMA sont utilisés pour prolonger la série initiale afin de limiter les révisions des estimations lorsque l'on possède un point supplémentaire. D'autre part, la modélisation ARIMA a aussi été introduite dans les méthodes de désaisonnalisation fondées sur la théorie de l'extraction du signal (Box, Hillmer et Tiao (1978), Burman (1980), Hillmer et Tiao (1983) et Hillmer, Bell et Tiao (1983)).

Les évolutions récentes

Aujourd'hui, les deux grandes philosophies de la désaisonnalisation, à savoir l'approche empirique ad hoc et l'approche par modélisation, inspirent diverses méthodes, dont certaines mêlent les deux. Parmi les méthodes actuellement les plus utilisées, on peut citer BV4 (Université technique de Berlin), Dainties (méthode officielle de la Commission Européenne), SABL (Laboratoires Bell), X11-ARIMA 88 (Statistique Canada), STL, STAMP. D'autres, mises au point plus récemment, pourraient connaître un certain succès : X12-ARIMA et TRAMO-SEATS.

Les principales critiques que l'on peut faire à chacune des deux approches sont difficiles à éviter. Ainsi, on reproche aux méthodes empiriques de ne pas s'appuyer sur la théorie statistique, ce qui rend particulièrement difficile, voire impossible, la connaissance des propriétés statistiques des estimateurs utilisés ; les méthodes basées sur les modèles sont satisfaisantes sur ce plan là mais on s'interroge sur la pertinence de la modélisation dans le cas de certaines séries très chahutées, et on invoque le manque de modèles adéquats et de théorie statistique pour les séries non stationnaires.

C'est pourquoi les améliorations que tentent d'apporter les nouvelles méthodes de désaisonnalisation ne concernent pas le principe même des méthodes existantes mais visent plutôt à corriger certains de leurs défauts. Les principales préoccupations sont tournées d'une part vers les problèmes de non symétrie des filtres en début et fin de série, et d'autre part vers l'élimination des divers effets perturbateurs qui influencent les résultats de la désaisonnalisation (points aberrants, changements de régime, effets de calendrier...).

Bilan sur les différents types de méthodes de désaisonnalisation

Schématiquement, les méthodes de désaisonnalisation peuvent être classées en deux grandes catégories : les méthodes non paramétriques et les méthodes paramétriques. Les méthodes non paramétriques, ou empiriques, permettent de décomposer la série en composantes inobservables par une procédure, souvent itérative, basée sur des lissages successifs. On peut résumer l'ensemble des lisseurs utilisés dans ces méthodes sous le nom de "régressions locales". Les régressions locales consistent à ajuster des polynômes, en général par les moindres carrés, sur des intervalles glissants (se décalant d'un point à chaque fois). Au centre de l'intervalle, la donnée lissée est la valeur, à cette date, du polynôme ajusté (la donnée lissée à la date suivante est obtenue par ajustement d'un polynôme sur l'intervalle suivant). On peut montrer que les régressions locales reviennent à appliquer des moyennes mobiles particulières lorsque les intervalles de temps sont réguliers. On peut toutefois

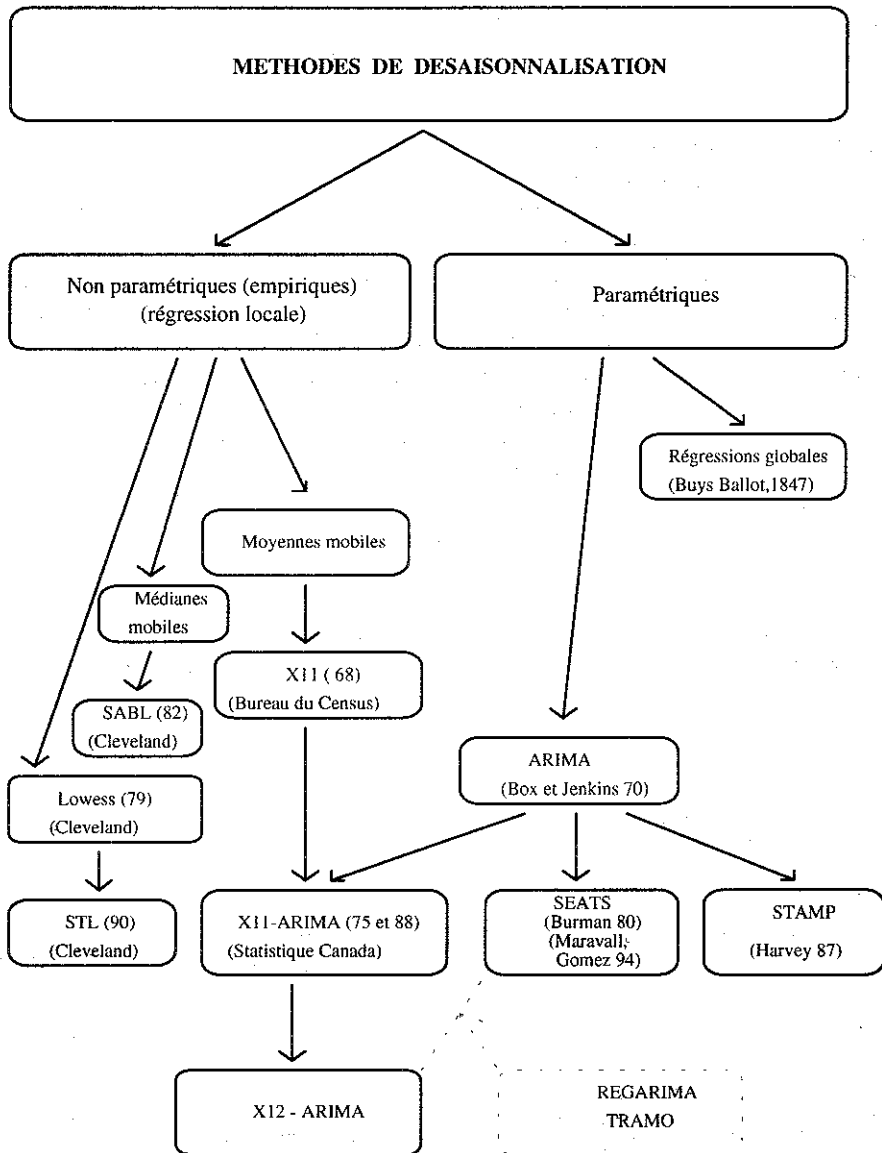
distinguer les méthodes purement basées sur les régressions locales (sans préoccupation des moyennes mobiles associées) et les méthodes utilisant directement les moyennes mobiles. Dans le premier groupe, on trouve STL (Cleveland, 1990), fondée sur le "lowess", à savoir le lissage robuste par régressions locales ; dans le second groupe, la plus célèbre est X11.

Les méthodes paramétriques peuvent elles aussi se diviser en deux grands ensembles : les méthodes par régression globale (inspirées par Buys Ballot) qui posent pour chaque composante, excepté l'irrégulier, une fonction déterministe du temps ; les méthodes basées sur des modèles stochastiques (non déterministes) : il s'agit principalement des modèles ARIMA (Box et Jenkins). Ces méthodes supposent les composantes inobservables mais modélisables par des ARIMA. Parmi celles-ci on distingue encore deux groupes : celles qui estiment les modèles des composantes à partir du modèle ARIMA de la série initiale (Burman, Bell, Hillmer, Maravall) - SEATS est la plus récente - et celles qui les estiment directement (Engle, Harvey, Todd) - par exemple la méthode STAMP -.

Les modèles ARIMA ont fait le pont entre les deux grandes approches de la désaisonnalisation. A la base même de celle-ci pour certaines méthodes par modélisation, ils ont eu à partir de 1975 un rôle important dans les méthodes par moyennes mobiles. En effet les filtres moyennes mobiles symétriques font perdre des points en début et fin de série. Ce problème est en général réglé en appliquant aux extrémités des moyennes mobiles asymétriques. Toutefois, pour réduire l'asymétrie, l'idée est venue de prolonger les séries brutes par des modèles ARIMA, ce qui a conduit au développement X11-ARIMA.

Enfin, le souci d'éliminer les effets perturbateurs qui nuisent à la qualité de la désaisonnalisation s'est manifesté dès les années 30 chez les partisans des diverses méthodes de l'époque. Puis dans les années 70-80, des auteurs ont tenté de prendre en compte les valeurs extrêmes dans la modélisation des séries (Fox (1972), Abraham et Box (1979), Denby et Martin (1979), Martin (1980), Chang (1982), Bell (1983), Hillmer, Bell et Tiao (1983)). Mais plus récemment la technique adoptée est la régression globale sur des variables particulières (de type indicatrices), qui permet de traiter toutes sortes de non-linéarités. Ce traitement préalable assez poussé des séries en amont de la désaisonnalisation constitue un deuxième pont entre les deux grandes philosophies de la désaisonnalisation : X11-ARIMA a été étendu à X12-ARIMA avec son étape RegARIMA, et SEATS peut être couplé avec TRAMO qui traite la série et la modélise par un ARIMA.

Le graphique de la page suivante résume cette présentation.



Deux nouveaux logiciels de désaisonnalisation : X12-ARIMA et TRAMO-SEATS

Deux méthodes de désaisonnalisation ont été mises au point récemment. X12-ARIMA (1994) fait partie des méthodes empiriques : elle succède à X11-ARIMA, en y apportant quelques développements importants. SEATS (1994) (Signal Extraction in ARIMA Time Series), conjugué avec TRAMO (Time Series Regression with ARIMA Noise, Missing Observations, and Outliers), s'inspire du programme élaboré par J.P. Burman (1980) et appartient au groupe des méthodes paramétriques, basées sur une modélisation globale de la série. Il ne s'agira pas ici de faire une comparaison empirique de ces logiciels, qui pourrait constituer une étude à part entière. On en fera plutôt une présentation, qui paraît intéressante à plusieurs titres. D'abord ces deux méthodes font l'objet de diverses études réalisées par un groupe de travail à EUROSTAT, qui laissent penser qu'elles pourraient être largement utilisées dans les années à venir. Ensuite parce qu'elles illustrent bien le bilan sur les différentes approches de la désaisonnalisation présenté précédemment : en opposition quant au principe même de la décomposition d'une série, ces méthodes ont toutefois en commun le souci de capter et d'éliminer les divers effets susceptibles de perturber la désaisonnalisation, d'utiliser au mieux l'information contenue dans la série et de fournir un diagnostic détaillé des résultats. Le présent papier s'attachera principalement à décrire comment X12-ARIMA et TRAMO-SEATS tentent de répondre à ces exigences.

1. Le principe des méthodes

Une décomposition d'une série temporelle n'est pas unique. Elle dépend de la définition des composantes, qui peut être plus ou moins précise. L'utilisateur, en fonction de ses objectifs, choisira la méthode qui lui convient le mieux.

1.1. X12-ARIMA

X12 est fondée sur le même principe que X11. Elle décompose la série selon un schéma additif ou multiplicatif en tendance, saisonnalité et irrégulier. Les composantes sont obtenues à l'issue d'un processus itératif basé sur des lissages par moyennes mobiles. Dans le cas d'un schéma multiplicatif, on peut le résumer de la façon suivante. Une moyenne mobile symétrique d'ordre 12 (pour une série mensuelle, 4 pour une série trimestrielle) fournit une première estimation de la tendance puisqu'elle a pour propriétés d'éliminer les saisonnalités d'ordre 12 et de réduire le bruit. La série initiale est alors divisée par cette estimation pour donner les rapports saisonnalité-irrégulier (SI). Cette dernière série est disposée en 12 colonnes, chacune correspondant à un mois. Une moyenne mobile pondérée (d'ordre 3x3 par exemple) est ensuite appliquée à chaque colonne pour éliminer l'irrégulier et fournir

des facteurs saisonniers provisoires. Ceux-ci sont alors redispés sous forme de série et normalisés. Les rapports SI sont divisés par ces facteurs saisonniers pour obtenir une estimation de la composante irrégulière. Les points extrêmes sont repérés sur la série de l'irrégulier : ils sont éliminés si leur écart à la moyenne est supérieur à 2,5 fois l'écart type, ou corrigés à l'aide d'une fonction de poids si leur écart à la moyenne est supérieur à 1,5 fois l'écart-type. Les étapes précédentes (sauf la première) sont réitérées en intégrant la correction des valeurs extrêmes. Une première estimation de la série désaisonnalisée est obtenue en divisant la série initiale par les nouveaux facteurs saisonniers. Une moyenne mobile de Henderson est alors appliquée à celle-ci pour une dernière estimation de la tendance. La série initiale est divisée par cette tendance et l'on fait subir aux rapports SI obtenus les étapes précédentes qui conduisent alors aux facteurs saisonniers et à l'irrégulier définitifs.

Dans X12, la définition des composantes n'est donc pas explicite. On pourrait dire brièvement que la saisonnalité doit être à peu près annulée par une moyenne mobile symétrique d'ordre égal à la périodicité de la série, et que les coefficients saisonniers sont normalisés, c'est-à-dire que leur somme sur n'importe quelle période de 12 mois doit faire approximativement 12. Cette propriété revient à dire que la somme des valeurs de la série brute sur un an doit être environ égale à la somme des valeurs de la série corrigée des variations saisonnières (CVS) sur la même année. La tendance finale est obtenue par application sur la série CVS d'une moyenne mobile de Henderson : elle répond donc au critère de lissage correspondant à ce type de moyennes mobiles (minimisation de la somme des carrés des différences troisièmes de la série). Enfin l'irrégulier est ce qui reste après élimination de la tendance et de la saisonnalité.

1.2. SEATS

SEATS définit plus précisément les composantes, à partir de leur densité spectrale :

- la tendance présente des pics spectraux à la fréquence 0,
- la composante saisonnière présente des pics spectraux à la fréquence correspondant à la périodicité,
- la composante cyclique représente les fluctuations périodiques de période supérieure à un an ; elle présente des pics spectraux à la fréquence correspondante, entre 0 et $2\pi/s$, où s est la périodicité,
- l'irrégulier est un bruit blanc, de spectre plat.

Comme la densité spectrale d'une série peut être paramétrée par l'intermédiaire du formalisme des processus ARIMA, le principe de la décomposition par SEATS est

le suivant : le logiciel part d'une série initiale modélisée par un modèle ARIMA. Il estime ce modèle et fournit un diagnostic détaillé de l'estimation. Puis il partitionne de façon additive le spectre en spectres associés aux différentes composantes par l'application d'un filtre de type Wiener-Kolmogorov : chaque composante est alors modélisée par un ARIMA. Afin d'identifier les composantes de façon unique, il suppose qu'elles doivent vérifier la condition "canonique" qui veut que chacune des composantes (excepté l'irrégulier) soit exempte d'irrégulier. Cela signifie qu'aucun bruit additif ne peut être extrait d'une composante autre que l'irrégulier. La variance de celui-ci est donc maximisée et au contraire, la tendance, la saisonnalité et le cycle sont aussi stables que possible. Bien qu'arbitraire (puisque toute autre décomposition exprimée par la canonique plus un bruit indépendant est admissible), cela permet d'éviter la contamination des composantes par du bruit, à moins qu'il n'y ait des raisons a priori de le faire.

Sur le principe de ces méthodes, on peut émettre un rapide jugement. Par rapport à X12-ARIMA, l'approche par modélisation (SEATS) est plus confortable, dans la mesure où les propriétés de la décomposition s'appuient sur la théorie statistique. On peut ainsi connaître la variance des estimateurs et effectuer des tests. Il est de surcroît plus aisé d'obtenir des prévisions à partir de composantes modélisées. Cependant, supposer a priori une forme particulière pour les éléments de la décomposition peut être jugé trop arbitraire, d'autant plus que dans la réalité, il n'est pas certain que l'on puisse modéliser de façon satisfaisante toute série par un modèle ARIMA.

2. La modélisation ARIMA dans les logiciels de désaisonnalisation

Dans le bilan sur les différentes méthodes de désaisonnalisation, on a vu que les modèles ARIMA étaient présents dans les deux grandes familles de méthodes. Ainsi X11 a été étendue à X11-ARIMA puis X12-ARIMA. On vient par ailleurs de montrer que SEATS s'appuie également sur ce type de modélisation. L'idée de modéliser la série initiale découle de motivations différentes suivant l'approche choisie pour la décomposition. Mais dans les deux cas, les travaux de Box et Jenkins ont fait apparaître cette famille de modèles comme une technique de modélisation et d'extrapolation très puissante et bien adaptée à un traitement en masse de séries.

L'écriture générale d'un modèle ARIMA saisonnier est la suivante :

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

où : Y_t est la série chronologique étudiée

ε_t est un bruit blanc

C est une constante

S est la périodicité de la série

B est l'opérateur retard, défini par : $BX_t = X_{t-1}$

ϕ est un polynôme en B de degré p

Φ est un polynôme en B^S de degré P

θ est un polynôme en B de degré q

Θ est un polynôme en B^S de degré Q

L'écriture schématique est $(p,d,q)(P,D,Q)$,

La modélisation ARIMA fournit des informations utiles sur la qualité des données brutes. En effet, le principe fondamental de la désaisonnalisation est l'existence d'un schéma de composition de la série. Si l'on ne peut ajuster à la série un modèle ARIMA, qui décrit la structure générale de la série en fonction de ses valeurs passées et de perturbations retardées, cela peut signifier que la série est presque purement aléatoire ou fortement perturbée par des éléments accidentels, de sorte que sa composante systématique n'est pas identifiable. On doit alors avoir un jugement critique sur la qualité des résultats de la décomposition.

2.1. Le rôle de la modélisation ARIMA dans X12-ARIMA

Le principal objectif de la modélisation ARIMA dans cette méthode est de répondre à l'exigence évoquée plus haut d'utiliser au mieux l'information contenue dans la série.

X11 apparaît comme le logiciel de référence pour la méthode de désaisonnalisation par moyennes mobiles. Le problème de la suppression de points en début et fin de série par les moyennes mobiles symétriques y est résolu par l'emploi de moyennes mobiles asymétriques, qui sont des moyennes mobiles de Musgrave. Celles-ci prolongent les moyennes mobiles de Henderson avec pour principe de minimiser les révisions des estimations des derniers points lorsque la donnée suivante sera disponible. Les autres moyennes mobiles symétriques sont prolongées par des moyennes mobiles asymétriques prédéfinies.

L'une des critiques qui étaient formulées à l'encontre de X11 était le fait que les estimations relatives aux observations les plus récentes n'étaient pas aussi fiables que les observations centrales, à cause de la nécessité d'appliquer des moyennes mobiles asymétriques en fin de série.

L'extension X11-ARIMA, développée en 1975 par Dagum, de même que X12-ARIMA, résout en partie ces problèmes : elle modélise la série initiale par un ARIMA et la prolonge d'un an au début et à la fin. Elle minimise l'erreur quadratique moyenne des estimations, et on peut montrer qu'elle minimise aussi les révisions qu'il faut apporter aux facteurs saisonniers lorsque la donnée suivante est disponible. En particulier la modélisation ARIMA est avantageuse dans les situations correspondant à des points de retournement de tendance car les modèles ARIMA captent bien les évolutions récentes.

Certes, le problème des estimations aux extrémités de séries n'est pas entièrement résolu puisqu'il faudrait bien plus d'un an supplémentaire de chaque côté pour symétriser totalement les moyennes mobiles. Cependant, l'estimation de la tendance-cycle pour la dernière observation est effectuée à l'aide d'une combinaison de coefficients des moyennes mobiles de Henderson et des coefficients de pondération du modèle ARIMA utilisés pour l'extrapolation, qui sont spécifiques à la série et peuvent donc saisir les mouvements les plus récents.

Enfin, l'addition d'une ou deux années de chiffres extrapolés, donc sans valeurs extrêmes, permet d'obtenir une meilleure estimation de l'écart-type de l'irrégulier, donc d'améliorer l'identification et le remplacement des valeurs extrêmes.

2.2. Le rôle de la modélisation ARIMA dans SEATS

Le rôle des modèles ARIMA dans SEATS est beaucoup plus rapide à définir. La décomposition de la série se fait à partir du modèle ARIMA de la série initiale et conduit à une modélisation ARIMA de chaque composante. La modélisation ARIMA est donc à la base même de la méthode, et, contrairement à X11-ARIMA, n'intervient pas simplement comme une amélioration d'une méthode qui pourrait fonctionner sans. Il faut noter que dans SEATS, les séries sont également prolongées car le filtre de Wiener-Kolmogorov fait "perdre" des points aux extrémités.

La modélisation ARIMA étant indispensable à la méthode de désaisonnalisation SEATS, il existe un modèle par défaut, le modèle Airline $(0,1,1)(0,1,1)$, dont l'expérience montre qu'il s'adapte bien à un grand nombre de séries, et qui sera utilisé si aucun autre n'a été spécifié.

2.3. Les procédures de modélisation ARIMA automatique dans X12-ARIMA et TRAMO

Dans X12-ARIMA et dans SEATS, l'utilisateur peut lui-même spécifier le modèle ARIMA qui lui semble le mieux adapté à la série. Mais il peut aussi avoir recours à une identification automatique du modèle. Ainsi, le logiciel TRAMO, qui peut être aisément couplé avec SEATS, propose une telle procédure. Dans X12-ARIMA, elle est directement intégrée dans le logiciel. Toutefois les procédures diffèrent dans les deux méthodes.

Dans TRAMO, l'identification automatique du modèle ARIMA se fait en deux temps. La première étape consiste en la détermination des facteurs de différenciation $(I - B)^d (I - B^s)^D$ et de la constante. Elle est réalisée par itération sur la séquence de modèles AR et ARMA (1,1), qui ont une structure multiplicative lorsque les données sont saisonnières. Cette procédure est basée sur les résultats de Tiao et Tsay (1983) et Tsay (1984). Les différences régulières et saisonnières sont d'ordre maximum respectivement 2 et 1.

Dans la deuxième étape, le logiciel procède à l'identification du modèle ARMA pour la série stationnaire. Elle se fait suivant la procédure de Hannan-Rissanen, avec quelques modifications. La recherche consiste à balayer les valeurs suivantes :

$$0 \leq (p, q) \leq 3 \text{ et } 0 \leq (P, Q) \leq 2$$

Elle est faite séquentiellement (pour des polynômes réguliers fixés, on obtient les polynômes saisonniers et vice-versa) et les ordres finals des polynômes sont choisis selon le critère BIC (Critère d'Information Bayésien), avec une contrainte possible au nom du principe de parcimonie et en faveur des modèles équilibrés (mêmes ordres AR et MA). Le critère BIC mesure l'écart entre la vraie loi des observations et le modèle proposé.

Dans X12-ARIMA, la modélisation automatique se fait suivant le même principe que dans X11-ARIMA : dans le programme sont prédéfinis cinq modèles saisonniers, qui ont été choisis parce qu'ils modélisaient de façon satisfaisante un grand nombre de séries aussi bien en termes d'ajustement sur le passé qu'en termes de projections pour les trois dernières années. Ces modèles sont testés et le meilleur est conservé. Il s'agit des modèles dont les parties non saisonnières sont (0,1,1), (0,1,2), (2,1,0), (0,2,2) et (2,1,2) et avec toujours la même partie saisonnière (0,1,1)_s. L'utilisateur a la possibilité d'introduire un ensemble de modèles qu'il souhaite tester automatiquement. Si l'utilisateur entre un nombre important de modèles, cela peut revenir à effectuer un balayage des paramètres, comme dans SEATS.

Les critères de qualité de la modélisation sont les suivants : la somme des carrés des erreurs de prévision ; le test du "portmanteau" (Ljung-Box) ; les prévisions "hors échantillon" ; si plusieurs modèles incorporés à l'option automatisée satisfont aux critères d'acceptation, le programme utilise celui qui donne les meilleurs résultats pour une année d'extrapolations rétrospectives.

3. Le traitement préalable des séries : couplage régression-ARIMA

3.1. Définition du modèle Reg-ARIMA

Les modèles ARIMA sont des modèles linéaires. Lorsqu'on les utilise pour modéliser une série, on suppose donc que la série est exempte de non-linéarités. Or dans la réalité, cette hypothèse n'est pas vérifiée pour un grand nombre de séries qui peuvent présenter des points aberrants, des changements brutaux de niveau, des effets de calendrier... Afin que la modélisation ARIMA soit pertinente, il convient donc de les éliminer préalablement.

Le traitement des non-linéarités dans les méthodes de désaisonnalisation se justifie à plusieurs titres. Le programme SEATS suppose explicitement que la série initiale en est exempte et n'effectue lui-même aucune correction. Dans X11-ARIMA, il existe bien des traitements pour certains types d'effets perturbateurs tels que les points aberrants, les effets de jours ouvrables et les effets de jours fériés mobiles. Mais, bien que la première étape du programme permette de réaliser quelques ajustements préalables, ils sont pour la plupart effectués au cours du processus de décomposition de la série. Or des tests empiriques tendent à prouver que l'on obtient de meilleurs résultats pour la désaisonnalisation lorsqu'ils sont effectués en amont du processus.

La solution proposée dans les deux cas est l'introduction d'une régression linéaire destinée à intégrer l'information que l'on possède sur la série et à capter les perturbations identifiables. Ce sont les résidus de cette régression qui seront alors modélisés par un ARIMA. Ce couplage régression-ARIMA constitue l'apport principal de X12-ARIMA par rapport à X11-ARIMA et de l'utilisation de TRAMO avec SEATS.

Le modèle Reg-ARIMA est défini de la même façon dans TRAMO et dans X12-ARIMA :

Si Y_t est la série initiale, alors le programme ajuste le modèle de régression suivant :

$$Y_t = \sum \beta_i X_{it} + z_t$$

où les β_i sont les coefficients de la régression, les X_{it} les variables explicatives qui sont soit fournies par l'utilisateur, soit prédéfinies par le programme (voir des exemples plus loin), et les z_t les résidus.

Il modélise ensuite les résidus de cette régression par un ARIMA :

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D z_t = \theta(B)\Theta(B^s)a_t + c$$

où a_t est un bruit blanc (i.i.d $(0, \sigma^2)$), et gaussien dans le modèle de TRAMO).

La modélisation de z_t par un ARIMA met en lumière le fait que les résidus d'une régression dans le cas de séries temporelles sont la plupart du temps autocorrélés ; faire l'hypothèse qu'ils ne le sont pas conduirait à des résultats inexacts.

La dénomination Reg-ARIMA est la combinaison de ces deux modèles :

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D (Y_t - \sum \beta_i X_{it}) = \theta(B)\Theta(B^s)a_t + c$$

Ce modèle s'interprète comme une généralisation d'un modèle de régression pur, où les résidus sont supposés être un bruit blanc, à un modèle de régression où les résidus suivent un processus ARIMA.

Les logiciels TRAMO et X12-ARIMA offrent à l'utilisateur la possibilité de choisir ses variables explicatives, s'il possède des informations sur la série qu'il cherche à modéliser, ou bien d'utiliser celles qui sont prédéfinies par le programme. Par exemple on peut introduire des variables exogènes dont on sait qu'elles sont liées à la variable étudiée. Les variables prédéfinies sont destinées à isoler des perturbations ou des effets interprétables :

- les points aberrants ("additive outliers" et "innovation outliers"),
- les ruptures de niveau ("level shifts"),
- les changements transitoires de niveau ("temporary changes"),
- les saisonnalités fixes,
- les changements de régime (modification de la saisonnalité),
- les effets de jours fériés,
- les effets de jours ouvrables.

Tous ces effets sont modélisés par des variables particulières (indicatrices par exemple) mais la méthodologie diffère parfois entre les deux logiciels.

3.2. Le traitement des points aberrants et des changements de niveau (par ruptures ou temporaires) :

- les variables :

TRAMO propose différentes variables d'intervention du type :

- 1. variables indicatrices (additive outliers),
- 2. toutes les séquences possibles de 1 et de 0,
- 3. $1/(1-\delta B)$ appliqué à toute séquence de 1 et de 0, où $0 < \delta \leq 1$,
- 4. $1/(1-\delta_s B^s)$ appliqué à toute séquence de 1 et de 0, où $0 < \delta_s \leq 1$
- 5. $1/(1-B)(1-B^s)$ appliqué à toute séquence de 1 et de 0.

Interprétation : Les variables indicatrices et les séquences de 1 et de 0 captent des perturbations ponctuelles ou répétitives. Le troisième type de variables peut s'écrire de la façon suivante :

si X_t est une séquence de 1 et de 0, alors

$$(1-\delta B)X_t = \left(\sum_{i=0}^{+\infty} \delta^i B^i \right) X_t = \sum_{i=0}^{+\infty} \delta^i X_{t-i}$$

Ainsi, à chaque instant t , la valeur de cette variable est égale à la valeur de X_t (soit 0 ou 1) plus les répercussions des valeurs 1 (perturbations) antérieures de X affectées d'un coefficient d'autant plus faible que ces perturbations sont éloignées dans le temps. Par exemple, si $X_t = 1$ et si les seules dates antérieures pour lesquelles X vaut 1 sont $(t-i)$ et $(t-j)$ alors $(1-\delta B)X_t = 1 + \delta^{t-i} + \delta^{t-j}$.

L'interprétation du quatrième type de variables est quasiment la même, sauf qu'à un instant t , seules les perturbations distantes de t d'un multiple de la période s sont répercutées.

Enfin, la dernière variable représente, à chaque instant t , le produit du nombre de perturbations précédentes et du nombre de perturbations précédentes distantes de t d'un multiple de la période.

X12-ARIMA modélise les points aberrants par des variables indicatrices, les ruptures de niveau et les changements transitoires de niveau par les variables suivantes :

- si une rupture de niveau survient à un instant t_0 , la variable

$$LS_t^{t_0} = \begin{cases} -1 & \text{pour } t < t_0 \\ 0 & \text{pour } t \geq t_0 \end{cases}$$

- si un changement transitoire de niveau a lieu entre t_0 et t_1 , la variable

$$TR_t^{(t_0, t_1)} = \begin{cases} -1 & \text{pour } t \leq t_0 \\ (t - t_0) / (t_1 - t_0) - 1 & \text{pour } t_0 < t < t_1 \\ 0 & \text{pour } t \geq t_1 \end{cases}$$

- **la méthodologie :**

S'ils ne sont pas spécifiés, les points aberrants et autres effets peuvent être détectés automatiquement par le programme. TRAMO et X12 utilisent pour cela des méthodologies différentes.

TRAMO a une approche similaire à celle de Chen et Liu (1993), avec quelques modifications. Les paramètres de la régression sont initialisés par les moindres carrés ordinaires et les paramètres du modèle ARMA sont alors estimés avec deux régressions, comme dans Hannan et Rissanen (1982). Puis le filtre de Kalman fournit la série des résidus, et une nouvelle estimation des paramètres de régression est obtenue. Pour chaque observation, des tests de Student sont effectués, pour chaque type de non linéarité comme dans Chen et Liu (1993). Les points aberrants sont modifiés un par un et à chaque fois, une nouvelle estimation des paramètres du modèle est obtenue. Une fois cette première séquence achevée, une régression multiple est effectuée et si des points aberrants sont détectés, le programme revient à la première séquence et itère jusqu'à ce qu'aucun point aberrant ne soit éliminé dans la régression multiple.

Dans X12-ARIMA, la détection des points aberrants est basée sur la méthodologie de Chang et Tiao (1983) avec des extensions et des modifications (Bell, 1983, 1994 et Otto et Bell, 1990). L'approche générale est similaire à la régression stepwise (GLS), où les variables candidates sont les variables précédemment décrites pour toutes les dates où la détection de points aberrants est effectuée. Le programme calcule la statistique de Student pour la significativité de chaque type de non-linéarité à chaque date, cherche celles qui sont significatives et ajoute les variables correspondantes au modèle. Les valeurs critiques des tests peuvent être définies par

l'utilisateur. Certains tests sont effectués pour repérer si les ruptures de niveau sont provisoires, ce qui est important pour la prévision.

3.3. Le traitement des saisonnalités fixes et des effets de calendrier

- les saisonnalités fixes

Dans TRAMO, des variables de type indicatrices peuvent permettre de prendre en compte ces effets.

X12-ARIMA propose deux types de variables périodiques : soit des variables indicatrices, soit des variables de Fourier (trigonométriques).

- les effets de jours fériés

Ces effets ont lieu lorsqu'il s'agit de jour fériés qui ne tombent pas toujours à la même date et peuvent survenir à l'un ou l'autre de deux mois (ou trimestres) consécutifs ; ils ne sont donc pas pris en compte dans la correction des variations saisonnières.

Dans les deux logiciels, des variables sont construites pour évaluer l'« effet Pâques ». X12-ARIMA traite aussi l'effet « Thanksgiving » et l'effet « Labor day ».

Le modèle utilisé par X12-ARIMA suppose un changement constant du niveau de l'activité journalière durant un nombre spécifié de jours avant le jour férié. Pour le Thanksgiving, il suppose un changement constant de l'activité journalière à partir d'un certain nombre de jours avant et jusqu'à un certain nombre de jours après le Thanksgiving.

- la correction des effets de jours ouvrables

TRAMO corrige ces effets en introduisant des variables spécifiques dans la partie régressive de la modélisation REGARIMA. En revanche, SEATS ne semble pas réaliser de traitements relatifs aux jours ouvrables.

Au contraire, dans X12-ARIMA, ils peuvent non seulement être pris en compte par des variables spécifiques dans la régression, mais sont à nouveau traités dans le processus itératif de décomposition de la série.

Dans l'étape REGARIMA, une option qui repose sur le critère d'Akaike (AIC) permet de déterminer si des variables de jours ouvrables doivent être introduites dans la régression. Plusieurs sortes de variables sont prédéfinies dans le programme pour modéliser les effets de jours ouvrables, en fonction du type de la série. Celle-ci peut en effet être de type flux (les données mensuelles sont des sommes de données journalières) ou de type stock (valeur à un jour donné du mois).

Dans certaines étapes de la décomposition proprement dite, le logiciel effectue un nouveau traitement des effets de jours ouvrables par régression de l'irrégulier sur des variables particulières. Cependant, les tests de Fisher pour ce type de régression ne sont pas toujours très fiables dans la mesure où l'irrégulier présente souvent de l'autocorrélation et de l'hétéroscédasticité. C'est pourquoi la modélisation des effets de jours ouvrables dans l'étape RegARIMA (sur la série initiale) semble préférable, sauf lorsque la tendance est particulièrement chahutée.

3.4. Autres traitements

Traitement des observations manquantes

Les deux logiciels offrent la possibilité de remplacer les observations manquantes. Pour cela, une valeur "bizarre" (par exemple -9999) est imputée aux dates où les observations sont manquantes et ces observations sont alors traitées comme des valeurs extrêmes, par introduction des indicatrices correspondantes. La valeur de remplacement est la différence entre la valeur imputée et le paramètre de régression estimé pour l'indicatrice. En réalité, il existe quelques contraintes : dans X12-ARIMA, le nombre d'observations manquantes ne doit pas être trop important ; dans TRAMO, lorsque certaines valeurs manquantes initiales ne sont pas estimables (paramètres libres), elles sont remplacées par des valeurs "bizarres" pour toute la suite des traitements et les estimations affectées par ces valeurs sont alors facilement repérables.

TRAMO propose une autre méthodologie de traitement des observations manquantes, basée sur l'espérance des valeurs manquantes conditionnellement aux données disponibles. Le principe est décrit dans Gomez et Maravall (1994).

Traitement des changements de régime, des effets "promotions"

Il peut arriver que la saisonnalité d'une série se modifie à partir d'une certaine date : on parle alors de "changement de régime". Dans X12-ARIMA, ce phénomène est traité en introduisant des variables de saisonnalité fixe sur chacune des deux périodes. Pour cela, le moment où se produit la rupture doit être connu de l'utilisateur.

Parfois, la série peut être l'objet d'une forte variation que l'on sait interpréter, et qui a lieu sur une période non habituelle (sinon le phénomène serait traité par l'ajustement saisonnier). Il peut s'agir par exemple de promotions pour écouler des stocks. X12-ARIMA propose un traitement de cet effet, à condition là encore que l'utilisateur détermine précisément la période en question.

4. Diagnostics

Les programmes X12-ARIMA et TRAMO-SEATS fournissent à chaque étape diverses statistiques que l'on ne présentera pas ici dans leur ensemble. On exposera quelques diagnostics relatifs à la qualité de la désaisonnalisation, qui constituent des nouveautés par rapport à des logiciels plus anciens.

Le premier élément de diagnostic est l'analyse du spectre de l'irrégulier. Celle-ci permet de détecter la présence de saisonnalité résiduelle et de tester la présence d'effets de jours ouvrables. Il se peut en effet que, dans une série déjà corrigée des variations saisonnières et des effets de jours ouvrables, ces éléments soient encore présents. Dans le cas d'une application directe d'un des logiciels sur la série, cela peut être dû à une mauvaise utilisation des procédures. Mais souvent, cette situation se produit lorsque la série n'a pas été traitée directement mais est l'agrégation d'un ensemble de séries qui, elles, ont été ajustées par les logiciels. En effet, dans certaines "sous-séries", les effets saisonniers et les effets de jours ouvrables sont parfois difficiles à détecter et leur correction n'est pas satisfaisante, ce qui peut créer des effets résiduels dans la série agrégée.

Le second élément important du diagnostic porte sur la stabilité de l'ajustement. En effet, on peut espérer que lorsqu'une nouvelle donnée est disponible, les estimations obtenues en intégrant cette nouvelle donnée ne diffèrent pas trop des estimations précédentes. A moins qu'il n'existe réellement une très grande variabilité dans la saisonnalité ou dans la tendance, des révisions importantes sont plutôt le signe d'une mauvaise qualité d'ajustement. Le principe du diagnostic de stabilité n'est pas nouveau mais la façon dont il est mis en œuvre dans X12-ARIMA est une innovation par rapport à X11-ARIMA. X12-ARIMA propose deux types de diagnostics de stabilité. Le premier consiste à faire tourner le programme sur au moins quatre subdivisions de la série ("sliding spans"). Pour un mois commun à deux intervalles, on compare les résultats obtenus sur les deux intervalles, aussi bien pour la série ajustée que pour les évolutions d'un mois à un autre. Le second type de diagnostics considère les révisions historiques : pour une date donnée, le programme analyse les différences entre les résultats obtenus par ajustement sur la période antérieure à cette date (c'est-à-dire la période pour laquelle cette date était la dernière de la série) et ceux obtenus sur la totalité de la série. Les comparaisons portent sur la série ajustée, les évolutions d'un mois à un autre, la tendance, les évolutions de la tendance d'un mois à un autre.

D'une façon plus générale, on attire l'attention sur le fait que les deux logiciels produisent de nombreuses statistiques et divers graphiques qui fournissent à l'utilisateur une information précieuse sur la qualité de l'ajustement et permettent de le guider pour apporter des améliorations si elle n'est pas satisfaisante.

BIBLIOGRAPHIE

- Bell W. R., Chen B., Findley D. F., Monsell B. C., Otto M. C. (1996) : New capabilities and methods of the X-12-ARIMA seasonal adjustment program.
- Bell W. R., Hillmer S.C. (1992) : « Issues involved with the seasonal adjustment of economic times series », *Modelling seasonality*,83-138.
- Bureau Of The Census (1996) : *X12-ARIMA Reference Manual* (Beta Version).
- Cleveland W. S. (1983) : *Seasonal and calendar adjustment*, Handbook of Statistics, 3, 39-92.
- Droesbeke J. J., Tassi P. (1990) : *Histoire de la statistique*, Que sais-je? PUF, Paris
- Dagum E. B. (1980) : *La méthode de désaisonnalisation X-11-ARMMI*, Statistique Canada.
- Fischer B. (1995) : *Decomposition of times series : comparing different methods in theory and in practice*, EUROSTAT.
- Gomez V., Maravall A. (1996) : *Programs TRAMO and SEATS Instructions for the User* (Beta Version).
- Hylleberg S. (1992). « The historical perspective », *Modelling seasonality*,15-25.