

CLASSIFICATION DE SÉRIES TEMPORELLES

APPLICATIONS À LA PRÉVISION ET À LA DÉSAISONNALISATION

Dominique LADIRAY

Insee, Département Statistiques de Court Terme

Version de travail (ne pas citer)

Introduction

L'analyse des données "à la française" et l'analyse des séries temporelles ont chacune une longue histoire mais curieusement leurs chemins ne se sont que rarement croisés, au moins jusqu'à un passé récent. Dans les vingt dernières années, avec la mise à disposition d'énormes bases de données temporelles, par exemple en biométrie et en météorologie, il y a eu une explosion d'intérêt pour l'exploration de ces données. Des centaines d'articles ont alors proposé des méthodes et algorithmes pour classer, indexer, segmenter et discriminer les séries temporelles.

De nombreuses méthodes de classification, mesures de similarité et algorithmes ont été développés au cours des ans, essentiellement pour des données d'enquêtes. Malheureusement, la plupart de ces mesures de similarité ne peuvent être directement utilisées sur des données temporelles. De nouvelles distances et de nouvelles stratégies ont été définies, certaines d'entre elles étant basées sur des outils ou résultats assez récents de l'analyse des séries temporelles : coefficients cepstrum, transformée par ondelettes, modèles markoviens cachés etc.

La première partie est consacrée à la présentation rapide des méthodes les plus utilisées en classification de séries temporelles.

Deux applications de ces méthodes seront ensuite proposées :

- une méthode de construction d'un modèle de prévision basée sur la classification de séries ;
- une méthode d'évaluation des performances de logiciels de désaisonnalisation à partir d'une base de séries reconstituées, les « séries Frankenstein ».

Cet exposé se veut essentiellement pédagogique : pour les justifications mathématiques, le lecteur est renvoyé à une large bibliographie proposée en dernière partie.

1. Classification de séries temporelles

La classification (Everitt, 1980 ; Saporta, 2006) a pour objet de regrouper les objets en classes dont les caractéristiques, et le contenu, sont déterminées non a priori comme dans l'analyse discriminante, mais par les données elles mêmes. Les objets d'une même classe se ressemblent donc et les objets de classes différentes ont au contraire peu de points en commun.

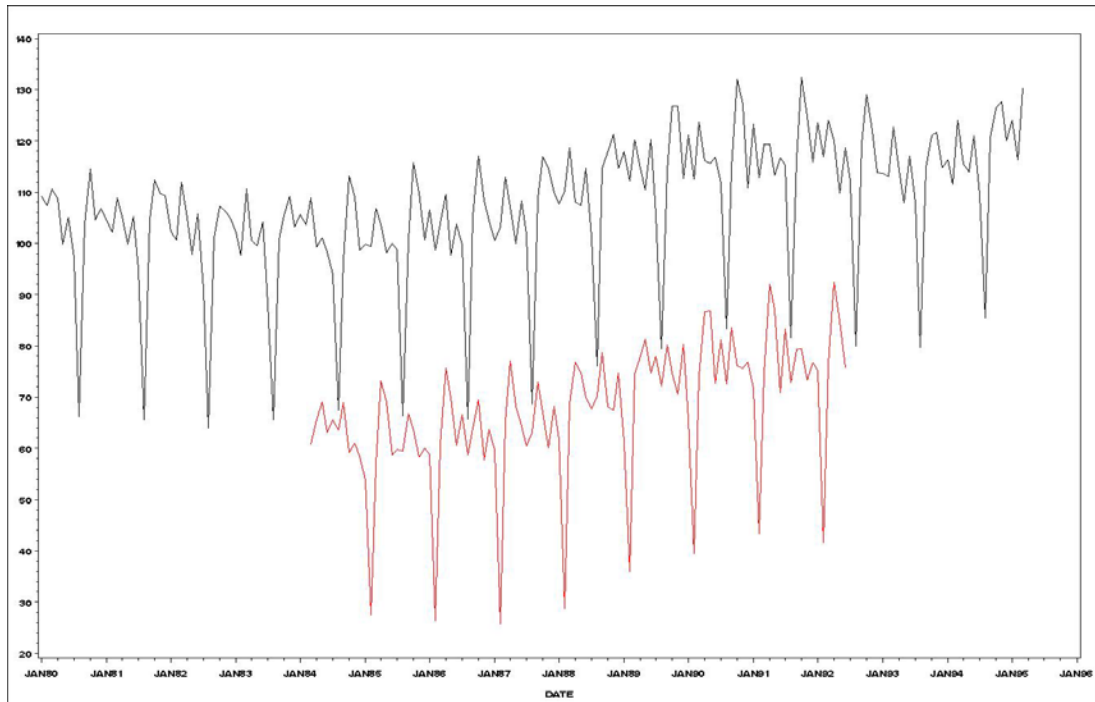
Toute méthode automatique de classification est ainsi basée sur une mesure de « similarité – dissimilarité » entre objets, une mesure de « similarité – dissimilarité » entre classes et une stratégie d'agrégation qui permettent de construire les classes.

De nombreuses méthodes de classification sont disponibles dans les logiciels statistiques standards : méthodes de partitionnement (K-means, nuées dynamiques etc.), cartes auto-organisatrices de Kohonen (« self organizing maps »), méthodes hiérarchiques descendantes et ascendantes etc. Les méthodes de classification ascendante hiérarchique (CAH) construisent, à partir des n individus de la population, des partitions successives emboîtées. Ces partitions ont la propriété de pouvoir être représentées dans un arbre de classification (« dendogram ») où la proximité entre deux individus ou deux classes peut se mesurer par la hauteur à laquelle ils ou elles se rejoignent pour former un seul groupe (voir figures 3 et 4).

Des centaines de distances ont été proposées pour classer des données d'enquêtes, parmi lesquelles la distance euclidienne est la plus populaire. Mais, lorsqu'il s'agit de classer des séries

temporelles, l'utilisation de la distance euclidienne, et de toute autre métrique de Minkowski, sur les données brutes peut conduire à des résultats peu intuitifs. En particulier, cette distance est très sensible aux effets d'échelle, à la présence de points atypiques ou manquants et ne permet pas de prendre en compte d'éventuels décalages temporels. Ainsi, la distance entre les deux séries de la figure 1 est grande alors qu'elles pourraient, dans certains cas du moins, être considérées comme « semblables » puisque la série noire (X_t) et la série rouge (Y_t) sont liées par la relation simple $Y_t = 0.75X_{t-6}$.

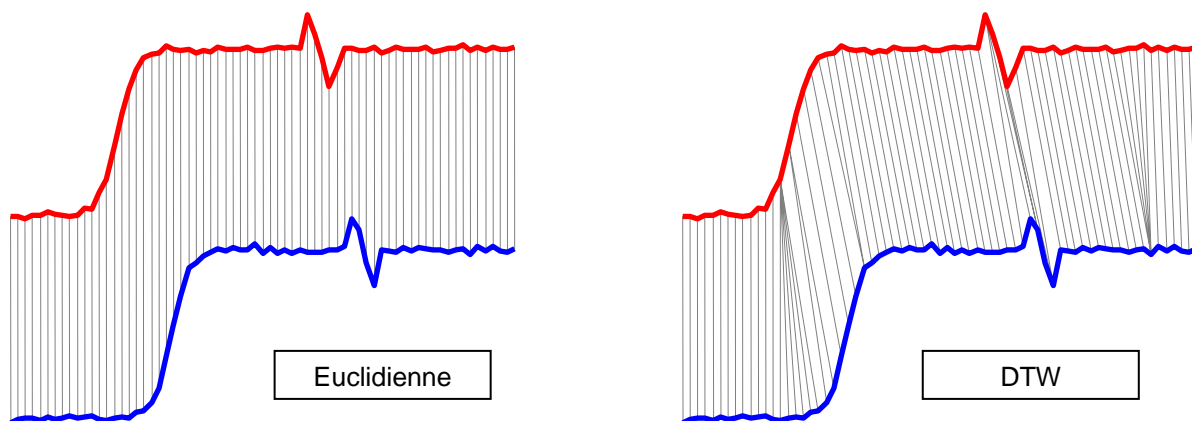
Figure 1: Deux séries temporelles « similaires »



1.1. Nouvelles distances et similarités

Une façon de résoudre ces problèmes est de définir de nouvelles distances et mesures de similarité : Dynamic Time Warping (DTW, Berndt & Clifford, 1994), Longest Common SubSequence (LCSS, Das et al., 1997) et Edit Distance on Real sequence (EDR, Chen et al., 2003). La figure 2 illustre les différences entre la distance euclidienne et la distance DTW qui considère un temps élastique, non linéaire.

Figure 2 : Distance euclidienne et Dynamic Time Wrapping

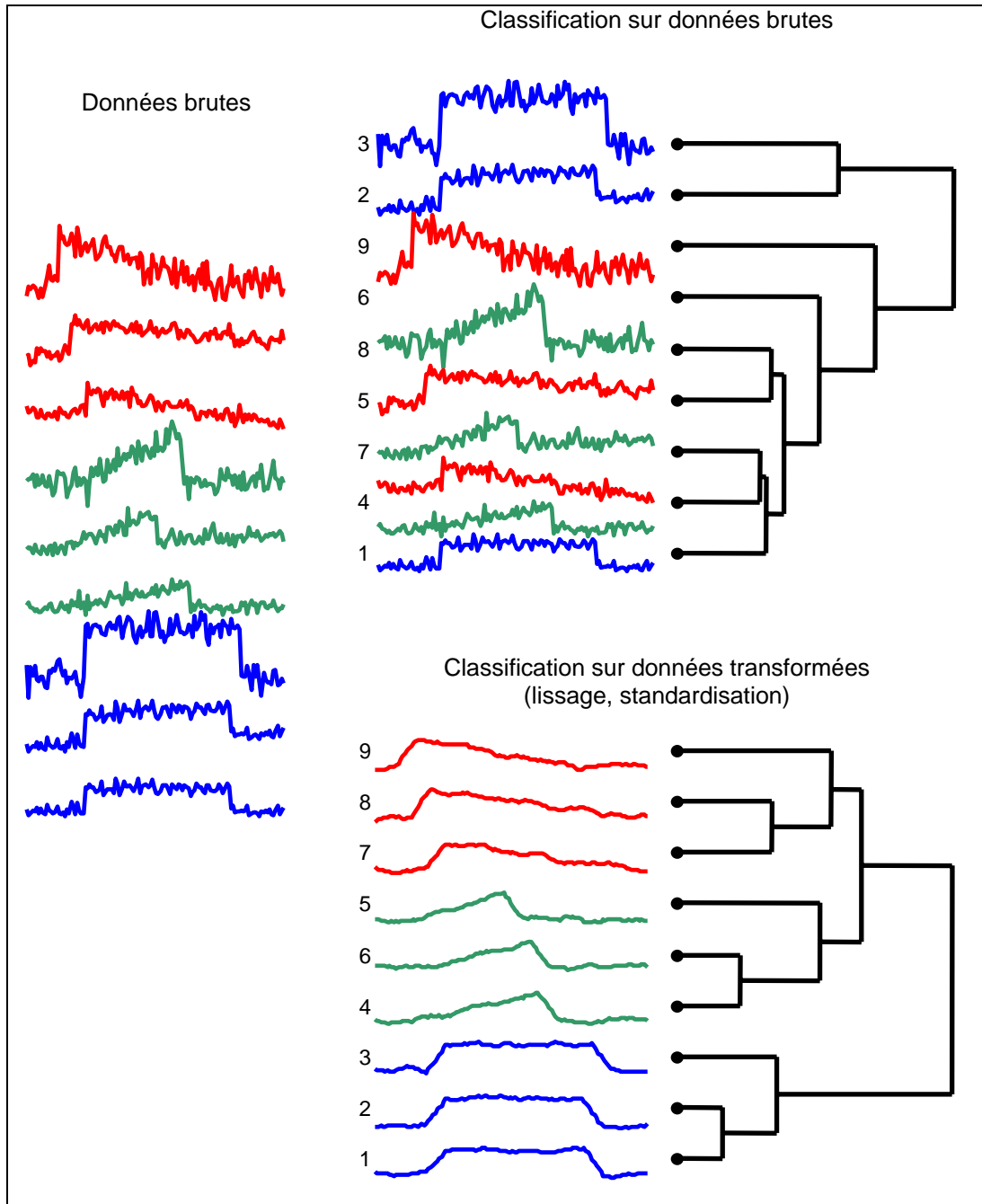


1.2. « Préparer » les données brutes

Une autre manière est de traiter les données brutes avant la classification en leur appliquant des transformations plus ou moins simples : standardisation, lissage, interpolation, stationnarisation etc. La figure 3 illustre l'importance et l'efficacité de la préparation des données.

Les mesures de similarité et distances évoquées ci-dessus sont calculées directement dans l'espace des séries elles-mêmes, dans l'espace des temps, mais les temps de calcul pour certaines de ces distances peuvent s'avérer trop importants pour de grosses bases de données et on s'oriente alors vers des techniques de réduction de dimension.

Figure 3 : Bien préparer ses données



1.3. Réduire la dimension du problème

L'idée est alors de projeter la série dans un autre espace, de dimension plus réduite, avec une transformation qui préserve les similarités et les distances ; on utilise alors un petit nombre de coefficients de la transformation pour faire la classification. Ces représentations des séries temporelles permettent le plus souvent et de traiter les problèmes évoqués ci-dessus et de considérablement réduire le temps de calcul.

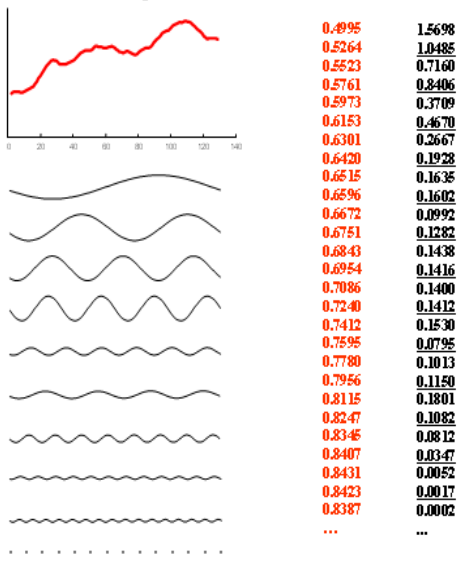
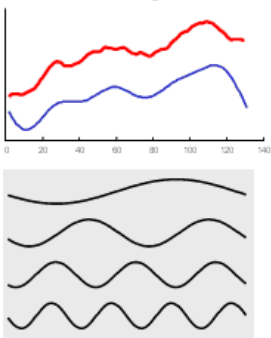
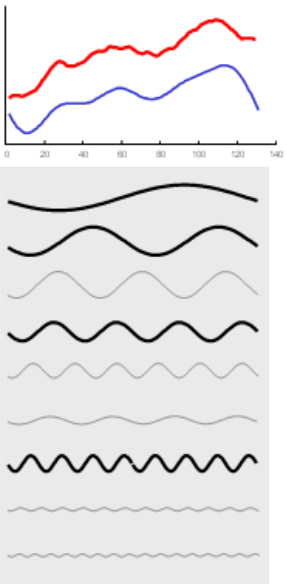
Un exemple simple et bien connu est d'utiliser la transformée de Fourier rapide. Toute fonction mathématique « sympathique » admet en effet une décomposition en sinus et cosinus de la forme :

$$X(t) = \sum_{k=1}^n [a_k \cos(\omega_k t) + b_k \sin(\omega_k t)]$$

Dans cette décomposition, les coefficients a_k et b_k représentent la contribution de la fréquence ω_k à la série X_t : les basses fréquences traduiront la tendance et le cycle de la série, les fréquences saisonnières la saisonnalité etc.

Le tableau 1 montre un exemple de réduction de la dimension d'une série temporelle par transformée de Fourier. Il suffit de retenir par exemple les plus importants coefficients de Fourier pour obtenir une représentation simplifiée de la série en quelques nombres.

Tableau 1 : un exemple de réduction de dimension par la transformée de Fourier rapide

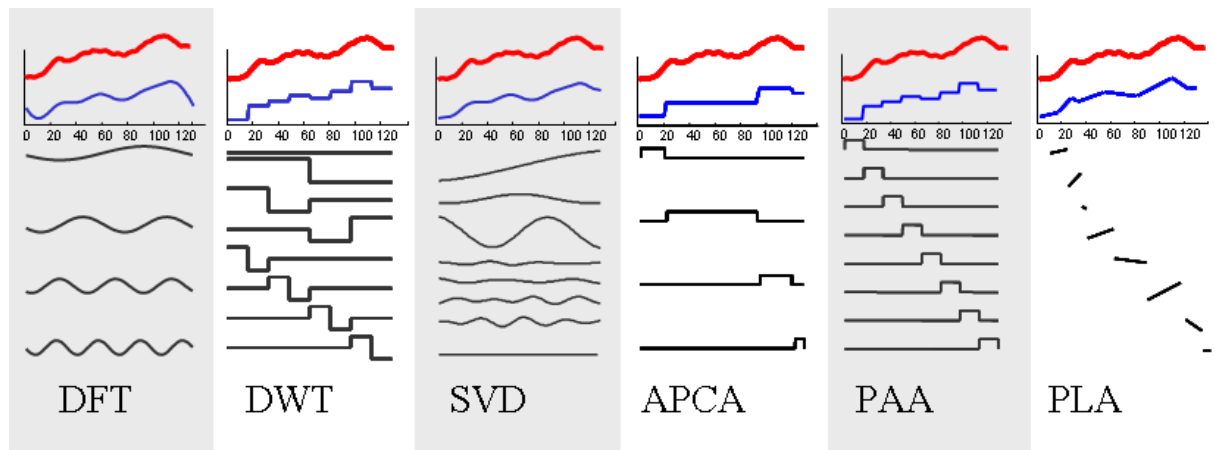
Série brute, coefficients de Fourier et fonctions associées	Décomposition sur les 4 premières fonctions	Décomposition sur les 4 plus importantes fonctions
		

Le même principe peut être appliqué à beaucoup d'autres transformations, certaines traitant directement le cas de séries non stationnaires :

- Fonctions d'autocorrélation directe, inverse, partielle (ACF, PACF, IACF) (Maballée and Maballée, 1911; Wang and Wang., 2000);
- Transformée de Fourier discrète (DFT) (Agrawal et al., 1993);
- Transformées par ondelettes, utilisant les bases de Daubechies, Haar (DWT) ou autres (Huntala et al., 1997);
- Polynômes de Chebyshev (Ng and Cai, 2004)

- Codage du Cepstrum (LPC), (Kalpakis et al., 2001);
- Décomposition en valeurs singulières via une analyse en composantes principales par exemple (Korn et al., 1997; Cleveland, 2004);
- Approximations linéaires par morceaux (Morikane et al., 2001);
- Approximations polynômiales par morceaux (Piecewise Aggregate Approximation, PAA - Keogh et al., 2000);
- Adaptive Piecewise Constant Approximation (APCA) (Keogh et al., 2001);
- Smooth Localized Complex Exponential model (SLEX) (Huang et al., 2004);
- Etc.

Voici quelques exemples de telles transformations:



1.4. Modéliser les séries

Une dernière idée est de modéliser les séries temporelles pour capturer et résumer leurs dynamiques. Deux séries seront alors similaires si les modèles ajustés le sont. Plusieurs idées sont exploitées dans la littérature :

- Un modèle autorégressif, ou même ARIMA, peut être ajusté aux séries. Ces modèles sont alors comparés en utilisant une métrique particulière (Piccolo (1990) ; Maharaj (2000) ; Xiong et Yeung (2002) ; Piccolo (2007), et Corduas et Piccolo (2008)).
- Les modèles markoviens cachés sont aussi assez populaires (Bicego et al., 2003; Smyth, 1997 ; Panuccio et al., 2002 ; Zend et Garcia-Frias, 2006).

2. Application à la désaisonnalisation

Dans le domaine de l'ajustement saisonnier, de nombreuses études portent sur la comparaison entre différentes méthodes ou logiciels, sur la valeur optimale des paramètres, sur la validation de nouvelles méthodologies etc. L'exercice est alors particulièrement difficile puisqu'il n'est pas possible, les « vraies » composantes étant inconnues, de mesurer les performances « réelles » des méthodes étudiées. Les auteurs testent alors leurs propositions sur quelques séries réelles ou par simulation (Hood et al., 2000; Lothian, 1978) et le risque d'un biais dû au choix des séries tests est sérieux. Ainsi, les études sur séries simulées reposent sur des modélisations a priori des séries ce qui peu favoriser telle ou telle méthode. La classification peut dans ce cadre être utilisée pour mettre en évidence des séries, tendances, cycles et saisonnalités types.

2.1. Les différents visages de la saisonnalité

A titre d'exemple, une classification ascendante hiérarchique est faite sur un échantillon de 1100 séries mensuelles issues de la base Euro-Indicateurs d'Eurostat. Ces séries sont ajustées par Tramo-Seats et X12-ARIMA et les spectres des 2200 composantes saisonnières sont alors classés. Ces spectres sont estimés par transformée de Fourier rapide et, comme le nombre de points du spectre dépend dans ce cas de la longueur de la série, des fonctions splines ont été utilisées pour

interpoler chaque spectre sur les 50 même fréquences. Un algorithme classique de classification, basé sur la distance euclidienne et la stratégie de Ward, est alors utilisé sur les spectres standardisés.

L'arbre résumant la classification (voir figure 4) suggère qu'il existe des types très différents de saisonnalité et est utilisé pour déterminer le nombre de classes. Les boîtes à moustaches présentées dans la figure 5 permettent de mesurer la dispersion dans chaque classe : les classes 1 et 5 apparaissent assez homogènes. Enfin, la figure 6 montre quelques saisonnalités caractéristiques de certaines classes. Bien entendu, l'étude pourrait être poursuivies pour voir par exemple si certaines formes de saisonnalités sont plus fréquentes dans certains secteurs économiques (production industrielle, commerce extérieur, balance des paiements etc.).

Figure 4 : Arbre de la CAH des 2200 spectres des composantes saisonnières

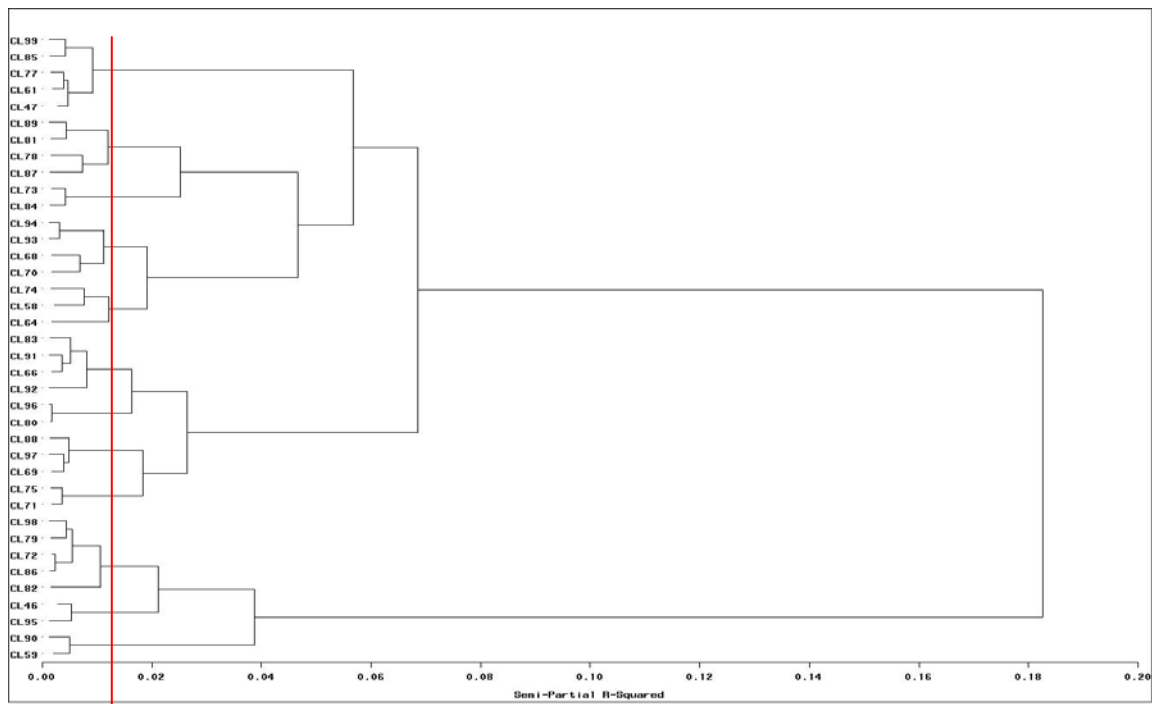


Figure 5 : Dispersion des séries dans les différentes classes

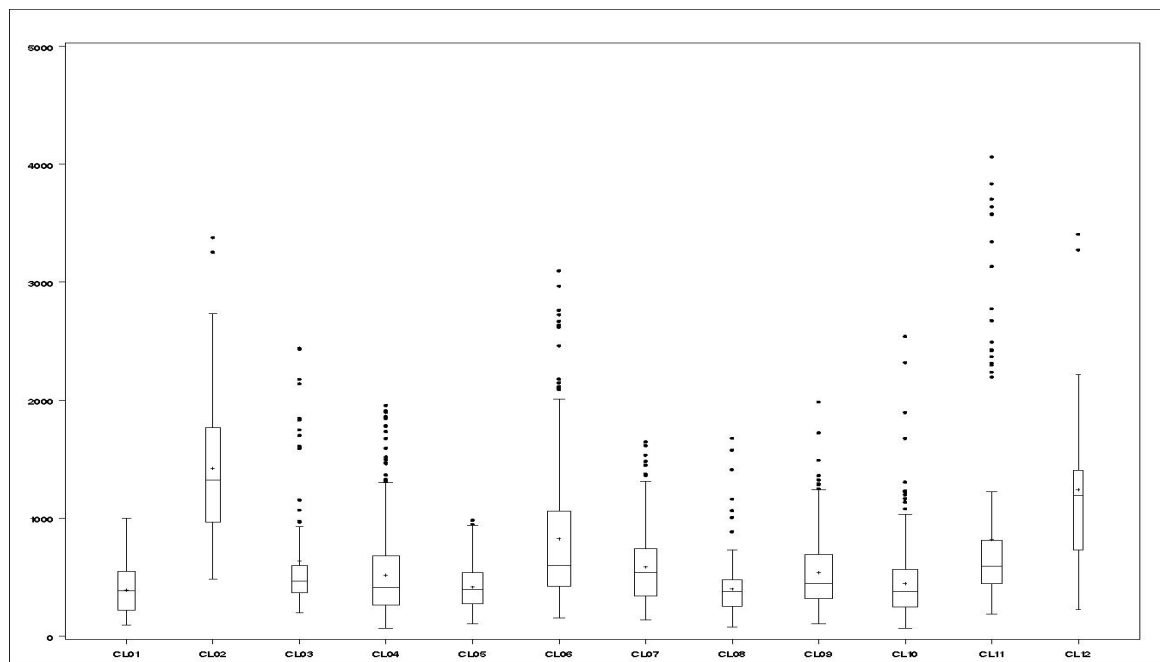
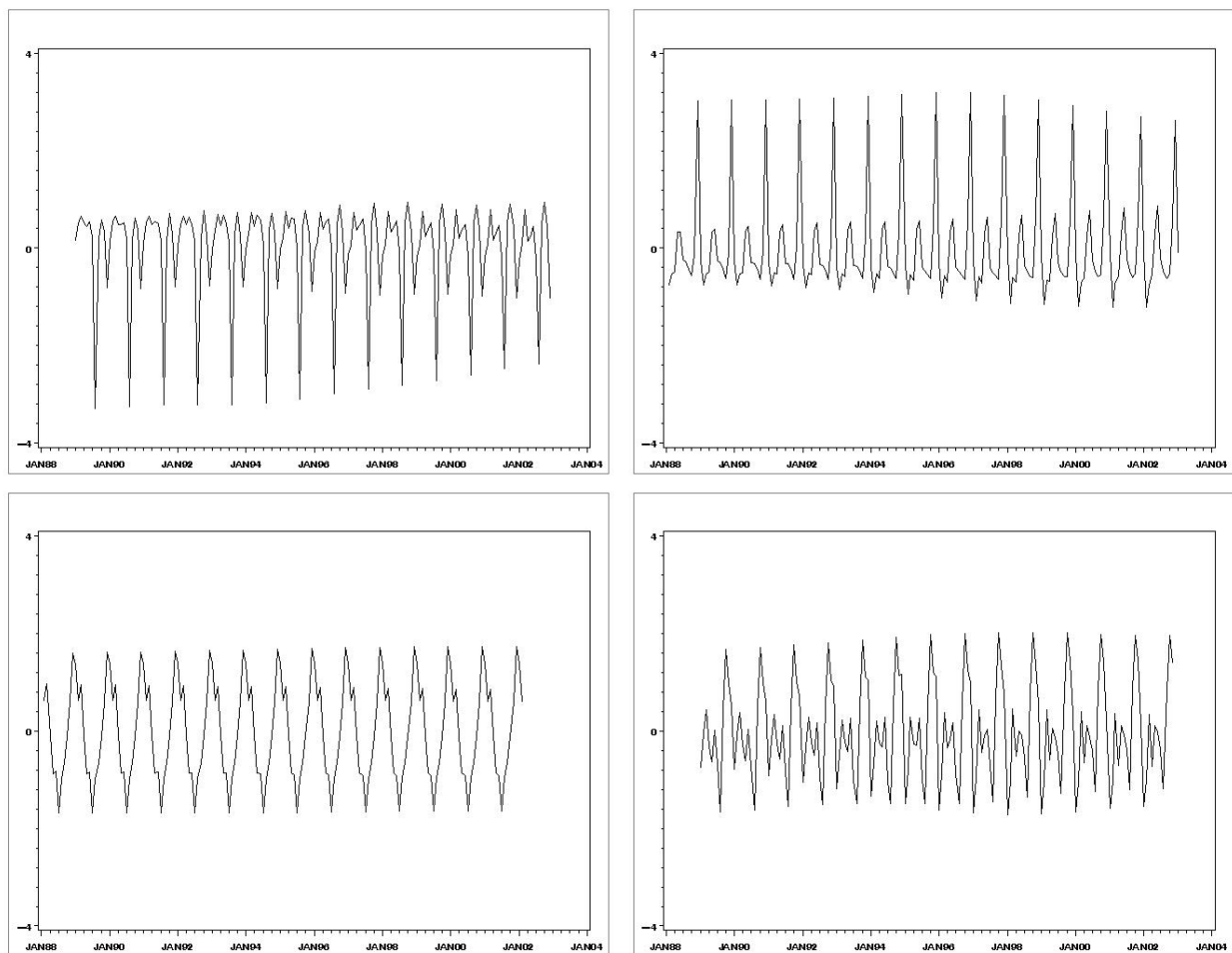


Figure 6 : Quelques saisonnalités mensuelles caractéristiques



2.2. Les séries « Frankenstein »

Évaluer la qualité des méthodes de désaisonnalisation est beaucoup plus difficile que d'évaluer celle des méthodes de prévision pour la raison simple que personne n'observera jamais la « vraie » série désaisonnalisée. Des jeux d'essai existent, par exemple les séries de la « compétition M2 » (Makridakis et al., 1993), pour évaluer les performances des méthodes et modèles de prévision.

Comme cela a déjà été mentionné, les méthodes d'ajustement saisonnier sont usuellement testées sur un petit nombre de séries réelles ou simulées et le risque d'un biais de sélection est réel. Il n'est par exemple pas évident que les saisonnalités exhibées dans le paragraphe précédent et issues de l'analyse d'un ensemble de séries européennes puissent être observées en Amérique du nord.

On pourrait cependant imaginer de construire un jeu d'essai de la façon suivante (Ladiray, 2004) :

- Choisir un large ensemble de séries temporelles, par exemple à partir de bases de données internationales comme la base « Main Economic Indicators » de l'OCDE ;
- Désaisonnaliser chaque série avec plusieurs méthodes (Tramo-Seats, X12-ARIMA, STAMP, BAYSEA, DECOMP, DAINTIES etc.) et exhiber ainsi des ensembles de saisonnalités, tendances-cycles, effets de calendriers ;
- Classer chacun des ensembles de composantes pour mettre en évidence des composantes saisonnières, des tendances-cycles, des effets de jours ouvrables « types » ;
- Reconstruire des « séries Frankenstein » à partir de ces composantes.

Un tel « jeu d'essai » est très utile, non seulement pour évaluer les performances d'une méthode mais aussi pour définir les valeurs par défaut des logiciels. En effet, les logiciels de désaisonnalisation ont généralement de nombreux paramètres dont des valeurs par défaut ont été fixées a priori pour

rendre l'utilisation de ces logiciels plus facile. D'où viennent ces valeurs par défaut ? Certaines d'entre elles sont fixées en fonction des propriétés statistiques des paramètres ; d'autres à la suite d'essais successifs ; d'autres enfin à partir d'un jeu plus ou moins restreint de données. Toute modification d'une méthode, par exemple l'utilisation de nouvelles moyennes mobiles dans X12-ARIMA, demande alors de nombreuses expériences. Les valeurs par défaut des paramètres doivent être à la fois « statistiquement correctes » et robustes, dans le sens où elles doivent être valables pour un grand nombre de séries temporelles. Les séries Frankenstein permettraient ainsi de déterminer plus facilement ces valeurs par défaut, éventuellement en se concentrant sur les séries d'un secteur particulier.

3. Application à la prévision

3.1. Le besoin d'estimations rapides

Malgré les efforts faits ces dernières années par l'ensemble du système statistique européen, un certain nombre d'indicateurs économiques européens clés restent publiés trop tardivement. De nombreux analystes économiques et acteurs institutionnels, dont la Banque Centrale Européenne, s'en sont plaints à de multiples reprises. La plupart des instituts de statistique européens ont beaucoup travaillé ces dernières années pour raccourcir leurs délais de publication et ne pourraient envisager de nouveaux progrès qu'au prix de modifications profondes, et donc coûteuses, de leurs systèmes de collecte et de production.

C'est pourquoi il semble assez naturel de se tourner vers des techniques économétriques d'estimation rapide. Le principe de ces méthodes est d'expliquer, puis de prévoir, une variable clé par des variables disponibles plus rapidement. Ainsi le PIB trimestriel européen est publié 45 jours après la fin du trimestre quand nombre de variables mensuelles sont déjà connues pour au moins deux mois du trimestre de référence.

Pour être utilisables, ces modèles économétriques doivent posséder au moins quatre qualités non indépendantes:

1. ils doivent être simples, c'est-à-dire ne faire intervenir qu'un nombre limité de variables,
2. ils doivent être interprétables : les relations qu'ils expriment doivent avoir un sens économique,
3. ils doivent être stables dans le temps, et en particulier ne doivent pas être remis en cause chaque mois,
4. et enfin, ils doivent avoir un bon pouvoir prédictif.

La construction de ces modèles fait donc intervenir à la fois une expertise économique (points 1 et 2) et une expertise statistique (points 3 et 4), ce qui pose en pratique bon nombre de problèmes.

3.2. Le problème de la sélection des variables explicatives

Dans l'exemple de l'estimation rapide du PIB de la zone euro, de nombreuses variables explicatives sont potentiellement candidates : l'indice de la production industrielle, les enquêtes de conjoncture, les indices de chiffres d'affaire, les données d'emploi, les prix à la consommation, les prix de l'énergie, les immatriculations, les statistiques de construction etc.

Il n'est pas difficile d'imaginer a priori une vingtaine de variables candidates, pour 13 pays de la zone euro et la zone elle-même, et avec éventuellement 2 retards. Nous arrivons donc à $20 \times (13+1) \times 3 = 840$ variables potentielles. Mais on peut construire plus de 20 milliards de modèles à 4 variables à partir de nos 840 variables. Plusieurs méthodes de sélection de variables sont souvent utilisées :

- La méthode usuelle de l'économiste: à partir de la liste des variables explicatives, il cherchera à bâtir une relation économiquement significative. Ensuite, celle-ci sera soumise aux exigences statistiques de stabilité et de qualité de la prévision. Il est fort probable que le praticien ne trouve pas directement le "bon modèle"; il cherchera sans doute à l'améliorer en incorporant des variables retardées Cette méthode est très consommatrice de temps et disons le tout net donne rarement de bons résultats.

- La méthode « General to specific » purement statistique (Hendry, 2000 ; Hendry et Krolzig, 2001) consiste à partir d'un modèle contenant un nombre important de variables. Une succession de tests statistiques permet alors d'éliminer les variables ou retards non significatifs pour aboutir à un modèle « statistiquement correct ». Dans la pratique, ces modèles sont souvent difficilement interprétables.
- Une méthode, particulièrement efficace mais consommatrice en temps de calcul, consiste à construire et évaluer tous les modèles possibles construits à partir d'un nombre limité de variables présélectionnées (James Mitchell, NIESR)
- L'analyse factorielle dynamique (Altissimo et alii, 2001 ; Doz et Lenglart, 1999) est aussi utilisée comme méthode de réduction : seuls les premiers facteurs sont considérés comme variables explicatives. Notons que dans cette méthode les facteurs sont déterminés indépendamment de la variable à expliquer. Cela entraîne un paradoxe amusant mais désagréable : si une des variables en entrée de l'analyse factorielle explique parfaitement la variable d'intérêt, elle sera « mise en moyenne » avec les autres dans le facteur principal et on passera ainsi à côté de la régression idéale !

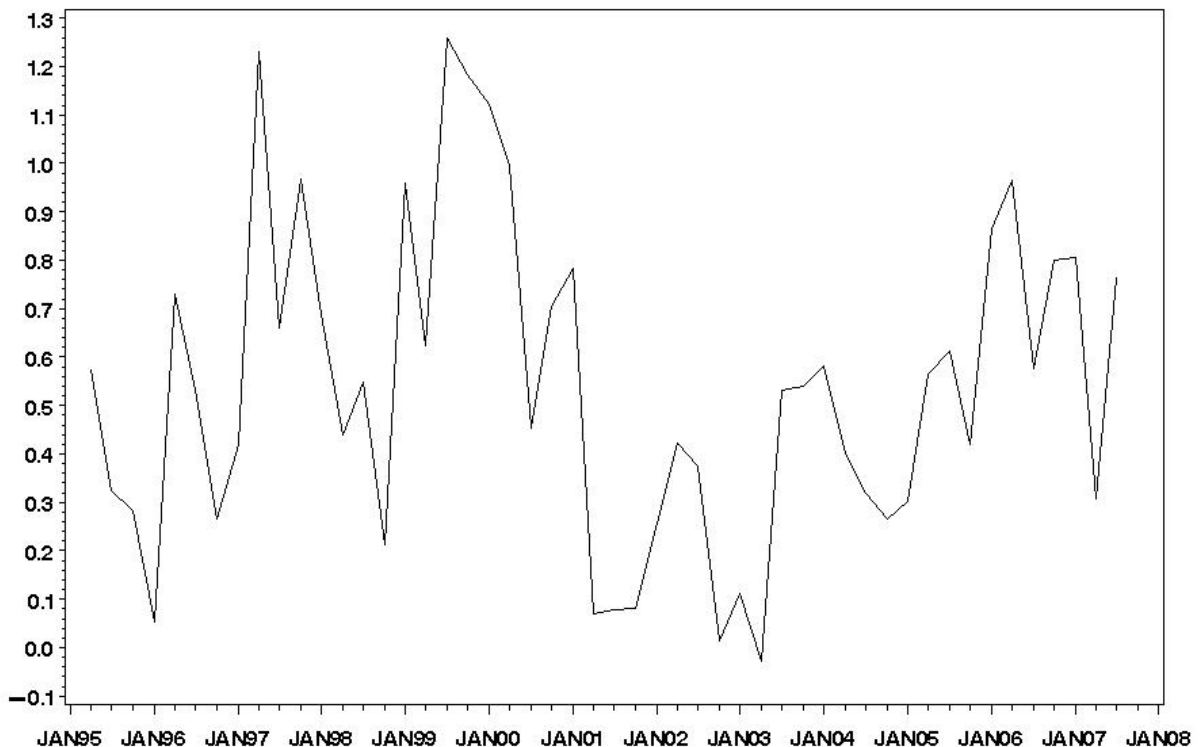
3.3. Une méthode basée sur la classification

La méthode proposée, même si elle met en œuvre des méthodes statistiques "confirmatoires", reste de philosophie exploratoire. Son but est de sélectionner, dans l'ensemble des modèles possibles, un certain nombre de modèles statistiquement corrects parmi lesquels l'économiste fera son choix. Elle procède en six étapes principales.

3.3.1. La variable à expliquer

On cherche à estimer le PIB trimestriel corrigé des variations saisonnières de la zone euro et plus précisément le taux de croissance trimestriel de cette variable représenté à la figure 7. Tramo-Seats ajuste automatiquement un modèle (0,1,1) à cette variable avec un écart-type du résidu égal à 0.29, ce qui donne une idée de la précision que l'on peut attendre d'un modèle de régression : tout modèle dont l'écart-type des résidus serait supérieur à 0.29 serait a priori rejeté.

Figure 7 : Taux de croissance trimestriel du PIB CVS-CJO de la zone euro



3.3.2. Les variables explicatives potentielles

Nous cherchons dans cet exemple à prévoir la valeur du 4^{ème} trimestre 2007 et ce 30 jours après la fin du trimestre. L'ensemble des variables explicatives potentielles est constitué de l'ensemble des séries disponibles dans la base Euro-indicateurs à la date du 30 janvier 2008. Un rapide tri permet de sélectionner 668 variables mensuelles dont nous connaissons au moins les valeurs des deux premiers mois du 4^{ème} trimestre 2007 et 162 variables trimestrielles. Les variables mensuelles sont trimestrialisées après prévision du mois manquant avec Tramo-Seats.

3.3.3. Réduction du nombre de variables candidates par classification

Une classification sur les 830 variables candidates est faite et six classes sont retenues. Ces classes sont alors représentées soit par les séries les plus liées à la variable à expliquer, soit par le ou les premiers facteurs dynamiques de la classe. Le lien avec la variable à expliquer peut se mesurer soit par des corrélations, soit par des tests de causalité.

In fine, si on retient 2 variables par classe et si on admet 2 retards dans le modèle, on obtient un nombre d'environ 60000 modèles possibles avec 4 variables explicatives.

3.3.4. Réduction du nombre de modèles potentiels

Il est alors possible de chercher les n meilleurs modèles en utilisant une méthode automatique de sélection de modèles, par exemple une régression « stepwise », et en incluant à ce stade les variables retardées. Plusieurs critères sont possibles (R², Cp de Mallows etc.) et une première évaluation non dynamique des modèles peut être faite à ce stade.

3.3.5. Evaluation statistique des modèles candidats

Dans cette étape, les quelques dizaines ou centaines de modèles présélectionnés sont évalués en profondeur en appliquant une méthode d'ajustement avec auto-corrélation des résidus (PROC AUTOREG de SAS). Divers tests statistiques évaluant la qualité de l'ajustement et la stabilité des modèles sont calculés : R², AIC, BIC, tests Reset, tests de Chow, statistiques de Durbin-Watson, tests Arch, tests de stationnarité tests de Godfrey, cohérence entre les signes des taux de croissance, erreur quadratique moyenne, erreur quadratique moyenne aux horizons 1 et 2 etc.

Coherence between growth rate signs

Cette évaluation permet donc de sélectionner les modèles qui, du point de vue du statisticien, sont les « meilleurs ».

3.3.6. Le choix du ou des modèles retenus

Le choix final du modèle est alors basé sur des critères statistiques et sur des « critères d'expert » :

- Le modèle a-t-il toutes les qualités statistiques requises ?
- Est-il suffisamment simple et robuste ?
- Est-il pertinent du point de vue économique et interprétable ?

En général, la réponse à cette dernière question est négative. Mais la méthode de classification utilisée va nous permettre de remédier aisément à ce problème. Le tableau 2 présente les meilleurs modèles au sens du R² et de l'erreur quadratique moyenne.

Quelques explications sont nécessaires pour comprendre le nom des variables :

- Le nom de la variable est la concaténation entre le nom de l'indicateur, le nom du pays et le retard considéré. Ainsi PIB_EA13_1 désigne le PIB de l'eurozone à 13 pays retardé d'une période ;
- Pour les indicateurs, ET_BAL désigne la balance commerciale, IO les nouvelles commandes dans l'industrie, UE les anticipations sur le chômage, IP l'indice de production industrielle, IT l'indice de chiffres d'affaires dans l'industrie etc.

Le second modèle fait intervenir 3 variables : IO_PL (nouvelles commandes dans l'industrie polonaise), IP_EA13 (indice de la production industrielle de la zone euro) et UE_NL (anticipations sur le chômage aux Pays-Bas). Si l'expert juge que ce modèle n'est « pas satisfaisant » du point de vue

économique, il peut alors facilement chercher dans les classes des variables proches des variables sélectionnées automatiquement qui seraient plus interprétables. Ainsi, la variable EXP_EA13 (exportations de la zone euro) est très proche de la variable IO_PL et de même, la variable UN_EA13 (chômage dans la zone euro) est elle très proche de la variable UE_NL.

On a donc un autre modèle incluant les 3 variables EXP_EA13, IP_EA13 et UN_EA13 qui s'avère avoir une erreur quadratique moyenne très acceptable de 0.155.

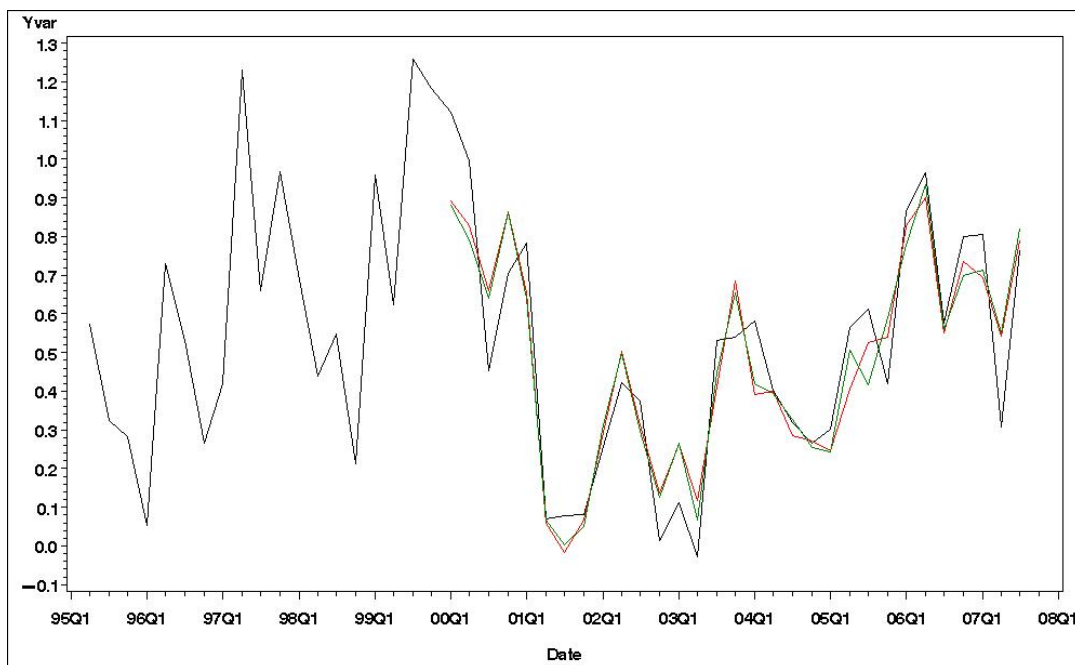
Les deux modèles donnent d'ailleurs des ajustements très comparables comme le montre la figure 8.

Il est enfin à noter que les modèles purement autorégressifs ne fonctionnent pas bien.

Tableau 2 : Les meilleurs modèles au sens de l'erreur quadratique moyenne

Rang	Modèle	R2	RMSE
1	ET_BAL_GR IO_PL IP_EA13 UE_NL	0,933	0,133
2	IO_PL IP_EA13 UE_NL	0,921	0,136
3	IO_PL IP_EA13 IT_RT_IE UE_NL	0,876	0,141
4	EPI_BU_PL IO_PL IP_EA13 UE_FR	0,869	0,144
5	IO_PL IP_EA13 IP_PT UE_NL	0,862	0,148
6	DIT_EA13 IO_PL IP_EA13 UE_NL	0,856	0,148
7	IO_PL IP_EA13 IT_RT_IE UE_FR	0,855	0,148
8	IO_PL IP_EA13 SV_PR_ES UE_FR	0,855	0,151
9	IOB_EA13 IO_PL IP_EA13 IT_RT_IE	0,851	0,153
100	PIB_EA13_1 PIB_EA13_2	0,257	0,296
101	PIB_EA13_1	0,226	0,296
102	PIB_EA13_2	0,040	0,298

Figure 8 : Ajustement des deux modèles retenus



Bibliographie

- [1] Agrawal, R., Faloutsos, C., Swami, A. (1993), Efficient Similarity Search in Sequence Databases. *Lecture Notes in Computer Science* 730, Pages 69-84 Springer Verlag,
- [2] Altissimo F., Bassanetti A., Cristadoro R., Forni M., Lippi M., Reichlin L. et Veronese G. (2001), « A Real Time Coincident Indicator of the Euro Area Business Cycle », document de travail n° 436, Service des études de la Banque d'Italie.
- [3] Berndt, D., Clifford, J. (1994). *Using dynamic time warping to find patterns in time series*. AAAI-94 Workshop on Knowledge Discovery in Databases.
- [4] Bicego, M., Murino, V., Figueiredo, M., 2003. Similarity-based clustering of sequences using hidden Markov models. *MLDM 2003, LNAI 2734*, 86–95.
- [5] Chen, L., Ozsu, M. T., Oria, V. (2003). *Robust and efficient similarity search for moving object trajectories*. Technical Report. CS-2003-30, School of Computer Science, University of Waterloo.
- [6] Cleveland, W. P., (2004), Stability and Consistency of Seasonally Adjusted Aggregates and Their Component Patterns, *Studies in Nonlinear Dynamics & Econometrics*: Vol. 8: No. 2.
- [7] Corduas, M., Piccolo, D., (2008). Time series clustering and classification by the Autoregressive metric. *Computational Statistics and Data Analysis* 52, 1860–1872.
- [8] Dagum, E. B. (1979), On the Seasonal Adjustment of economic Time Series Aggregates: A Case Study of the Unemployment Rate, *Counting the Labor Force, National Commission Employment and Unemployment Statistics*, Appendix, 2, 317-344, Washington.
- [9] Das, G., Gunopulos, D., Mannila, H. (1997) Finding similar time series. In *Proceedings of the 1st European Symposium on Principles of Data Mining and Knowledge Discovery*, pages 88–100.
- [10] Doz, C., Sylngart, F. (1999), Analyse factorielle dynamique : test du nombre de facteurs, estimation et application à l'enquête de conjoncture dans l'industrie, *Annales d'économie et de statistique*, n°52.
- [11] Everitt, B.S. (1980), *Cluster Analysis*, Second Edition, London: Heineman Educational Books Ltd.
- [12] Focardi, F. M. (2001), Clustering economic and financial time series: Exploring the existence of stable correlation conditions, Discussion Paper 2001-04, InterTek Group, Paris.
- [13] Galbraith, J. K., Jiaqing, L. (1999), *Cluster and Discriminant Analysis on Time Series as a Research Tool*, UTIP working paper, n°6, LBJ School of Public Affairs, University of Texas in Austin.
- [14] Ge, X., Smyth, P., (2000), Deformable Markov Templates for Time Series Pattern Matching, Technical Report N. 00-10, University of California at Irvine.
- [15] Hood, C. C., Ashley, J. D., Findley, D. F., (2000), *An empirical evaluation of the performance of TRAMO/SEATS on simulated series*, in the ASA Proceedings of the Joint Statistical Meetings.
- [16] Huang, H.-Y., Ombao, H., Stoffer, D. S., (2004), Discrimination and Classification of Nonstationary Time Series Using the SLEX Model, *Journal of the American Statistical Association*, Vol 99, 467, pp. 763-774.
- [17] Huhtala, Y., Kärkkäinen, J., Toivonen, H. (1999). Mining for similarities in aligned time series using wavelets. *Data Mining and Knowledge Discovery: Theory, Tools, and Technology*, SPIE Proceedings Series, Vol. 3695, pp 150-160.
- [18] Hendry, D.F. (2000): *Econometrics: Alchemy or Science?* Oxford University Press.
- [19] Hendry, D.F., Krolzig, H.-M. (2001): *Automatic econometric model selection*. London, Timberlake Consultants' Press.
- [20] Kakizawa, Y., Shumway, R.H., M. Taniguchi, M., (1998), Discrimination and Clustering for Multivariate Time Series”, *Journal of the American Statistical Association*, Vol. 93, n°441, 328-340.
- [21] Kalpakis, K., Gada, D., Puttagunta, V. (2001), “Distance measures for effective clustering of ARIMA time-series”. In proceedings of the IEEE Int'l Conference on Data Mining. San Jose, CA, Nov 29-Dec 2. pp 273-280.
- [22] Keogh, E. J., Chakrabarti, K., Pazzani, M. J., Mehrotra, S. (2000), Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases. *Knowledge and Information Systems*, vol. 3, pp 263-286
- [23] Keogh, E., Chakrabarti, K., Pazzani, M. & Mehrotra, S. (2001). Locally adaptive dimensionality reduction for indexing large time series databases. In *proceedings of ACM SIGMOD Conference on Management of Data*. pp 151-162.
- [24] Keogh, E., Kasetty, S., (2003), “On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration”, to appear in the *Data Mining and Knowledge Discovery Journal*.

- [25] Keogh, E., Pazzani, M. (1998), "An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback", In proceedings of the 4th Int'l Conference on Knowledge Discovery and Data Mining. New York, NY, Aug 27-31. pp 239-241
- [26] Korn, F., Jagadish, H., Faloutsos, C. (1997). Efficiently supporting ad hoc queries in large datasets of time sequences. *In proceedings of the ACM SIGMOD Int'l Conference on Management of Data.*
- [27] Ladiray D. (1997), *Using Business Survey Data To Forecast Employment*, 51^e session de l'Institut International de Statistique, Istanbul.
- [28] Ladiray, D., (2004), *Comparison of Seasonal Adjustment Methods and Softwares: Methodology and Results*, travail en cours, INSEE-CNAM, Paris.
- [29] Lin, J., Keogh, E., Lonardi, S. & Chiu, B. (2003), "A Symbolic Representation of Time Series, with Implications for Streaming Algorithms". In proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. San Diego, CA. June 13.
- [30] Lothian, J., (1978), The Identification and Treatment of Moving Seasonality in the X-11 Seasonal Adjustment Method, working paper 78-10-004, Time Series Research & Analysis Division, Statistics Canada, Ottawa.
- [31] Maballée, Colette et Berthe, (1911), Classification of Time Series and Forecasting: The SiNCiD Method, *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 3, 159-167.
- [32] Maharaj, E.A., 2000. Clusters of time series. *Journal of Classification* 17, 297–314.
- [33] Makridakis, S., Chatfield, C., Hibon, M., Lawrence, M., Mills, T., Ord, K, (1993), The M2-competition: A real-time judgmentally based forecasting study, *International Journal of Forecasting*, vol. 9(1), pages 5-22.
- [34] Morinaka, Y., Yoshikawa, M., Amagasa, T., (2001), The L-index: An Indexing Structure for Efficient Subsequence Matching in Time Sequence Databases, in *Proceedings of The Fifth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD2001)*, pp.51-60.
- [35] Ng, R. T., Cai, Y., (2004), Indexing Spatio-Temporal Trajectories with Chebyshev Polynomials. *Proceedings of SIGMOD 2004*
- [36] Panuccio, A., Bicego, M., Murino, V., 2002. A hidden Markov model-based approach to sequential data clustering. Proceedings of the Joint IAPR International Workshop on Structural Syntactic, and Statistical Pattern Recognition, 2002, 734–742.
- [37] Piccolo, D., 1990. A distance measure for classifying ARIMA models. *Journal of Time Series Analysis* 11, 153–164.
- [38] Patel, P., Keogh, E. Lin, J., Lonardi, S. (2002). Mining motifs in massive time series databases. In the *2nd IEEE International Conference on Data Mining*, Maebashi City, Japan
- [39] Saporta, G. (2006), *Probabilités, analyse des données et statistique*, Technip.
- [40] Wang, C., Wang, X. S. (2000), Supporting content-based searches on time series via approximation. *In proceedings of the 12th Int'l Conference on Scientific and Statistical Database Management. Berlin, Germany*, pp 69-81.
- [41] Xiong, Y., Yeung, D.-Y., (2002) *Mixtures of ARMA models for model-based time series clustering*. In: Proceedings of the IEEE International Conference on Data Mining.
- [42] Zeng, W., Garcia-Frias, J. (2006), A novel HMM-based clustering algorithm for the analysis of gene expression time-course data, *Computational Statistics and data Analysis*, 50, 2472 – 2494