ÉCONOMÉTRIE, PRÉVISION ET ANALYSE DES DONNÉES

Dominique LADIRAY
Institut National de la Statistique et des Études Économiques
15 boulevard Gabriel Péri, BP 100
92244 Malakoff Cedex
FRANCE
(dominique.ladiray@insee.fr)

1 Introduction

L'analyse des données et l'analyse des séries temporelles ont toutes deux une longue histoire mais curieusement leurs chemins se sont rarement croisés, au moins jusqu'à récemment. Dans les dix dernières années, avec la mise à disposition d'énormes ensembles de données temporelles, on a assisté à une explosion d'intérêt pour l'exploration de ces fichiers gigantesques. Des centaines de papiers ont alors présenté et diffusé des méthodes et algorithmes pour indexer, classer, discriminer, segmenter des séries temporelles. C'est à ces nouveaux liens que le présent travail s'intéresse en se concentrant sur le domaine de l'analyse de la conjoncture et la prévision à court terme des grands agrégats économiques.

Les raisons pour lesquelles l'économétrie des séries temporelles a tardé à incorporer dans ses méthodes des outils d'analyse des données sont sans doute plus subtiles et vont bien au-delà de la simple disponibilité d'importantes bases d'indicateurs économiques. Dans un article de 1989, Philip Mirowski ([31]) défend l'idée qu'il s'est écoulé une trentaine d'années avant que les concepts stochastiques de la Physique ne pénètrent en Économie. On pourrait sans doute en dire autant de l'Économétrie : le filtre de Kalman et les modèles « état-mesure » ont par exemple été développés au début des années 1960 mais n'ont été adoptés par les économètres qu'au milieu des années 80. Les retards avec les outils de l'analyse des données sont encore plus importants : il faut attendre la fin des années 80 pour voir apparaître, dans la théorie de la cointégration (Engle et Granger, 1987, [13]), les équations de l'analyse canonique et la fin des années 1990 pour voir une application de l'analyse en facteurs communs et spécifiques en analyse conjoncturelle. Bien entendu ces méthodes doivent être adaptées à la nature particulière des données temporelles et il a fallu mettre au point des algorithmes efficaces de calcul. Mais les raisons profondes du retard tiennent aux différences fondamentales de philosophies entre l'analyse des données et l'économétrie, ou plus exactement entre les analyses exploratoire et confirmatoire si chères à John Tukey ([44]). L'économètre s'accommode

mal de l'absence – toute apparente - de modèles dans les outils de l'analyse des données et une méthode n'est adoptée que lorsqu'elle est parée d'une panoplie complète : « Representation, estimation and testing » pour paraphraser le titre de l'article de Engle et Granger ([13]) sur la théorie de la cointégration.

Cet article présente quelques exemples d'utilisation de techniques d'analyse des données dans le domaine de la prévision économique de court-terme et les pistes que semblent suivre les économètres pour adapter ces outils à leurs problèmes. La seconde partie traite de l'analyse factorielle dynamique et des développements naturels qui se font ou se feront immanquablement autour de la régression PLS. La troisième partie aborde la classification sur séries temporelles, technique assez largement utilisée dans le domaine de la fouille de données chronologiques mais encore peu présente en prévision. Les parties 4 et 5 présentent deux applications possibles et assez différentes de la classification : en désaisonnalisation et en recherche de modèles de prévision.

2 Économétrie et analyse factorielle

L'un des objectifs essentiels de l'analyse de la conjoncture, outre celui de donner une évaluation de la situation économique actuelle, est de détecter aussi vite que possible les retournements de l'activité. Implicitement, le conjoncturiste se réfère à un « cycle des affaires » définit par Burns et Mitchell en 1946 ([6]) :

« Les cycles des affaires sont une sorte de fluctuations visible dans les agrégats économiques de pays dont la production s'organise essentiellement autour des entreprises : un cycle se compose de périodes d'expansion se produisant à peu près en même temps dans de nombreuses activités économiques, suivies par des périodes de récession toutes aussi générales, des contractions et redémarrages qui se fondent dans la phase d'expansion du cycle suivant ; la succession de ces changements est récurrente mais pas périodique ; la durée des cycles des affaires varie de un à dix ou douze ans»[1] [2].

Les enquêtes de conjoncture menées auprès des entreprises sont un élément essentiel du diagnostic conjoncturel en France mais aussi dans la plupart des pays européens (European commission, [14], [16]). Le nombre et la diversité des questions posées rendent souvent délicate l'interprétation des résultats obtenus ; il est alors assez naturel de chercher un résumé synthétique de cette information. Une façon simple de faire est de calculer une moyenne, simple ou pondérée, de plusieurs soldes d'opinion relatifs à l'activité économique. Ainsi, l'IFO allemand publie un « indicateur des affaires »,

[1] "*Business cycles are a type of fluctuation found in aggregate economic activity of nations that organise their work mainly in business enterprises: a cycle consists of expansions occurring at about the same time in many economic activities, followed by similarly general recessions, contractions and revivals which merge into expansion phase of the next cycle; the sequence of changes is recurrent but not periodic; in duration business cycles vary from more than one year to ten or twelve years*".

[2] Cette notion est cependant plus ancienne. Dans un article de 1919, Persons ([36]) définit les principales composantes d'une série temporelle dont : " *A wavelike or cyclical movement superimposed upon the secular trend; these curves appear to reach their crests during the periods of industrial prosperity and their troughs during periods of industrial depression, their rise and fall constituting the business cycle"*.

moyenne des opinions sur les productions passée et future des entreprises interrogées ; la DG ECFIN de la Commission Européenne publie un « indicateur de confiance dans l'industrie » moyenne pondérée des soldes d'opinion sur les carnets de commande, les stocks de produits finis et la production future ; l'indice NAPM américain est la moyenne des soldes sur les commandes, la production, les délais de livraisons, l'emploi et les stocks.

2.1 Analyse factorielle dynamique

De telles méthodes présentent l'avantage de la simplicité mais le choix des questions retenues et celui des poids de pondération demeurent ad hoc. Dans les années 90[3], l'analyse factorielle dynamique est apparue comme un cadre naturel pour résoudre ce problème (Altissimo et al, [1] ; Doz et Lenglart, [11], [12] ; Forni et Reichlin, [18] ; Forni et al, [19], Quah et Sargent, [40] ; Stock et Watson, [42], etc.). Dans ce cas, chaque variable peut être décomposée en deux composantes orthogonales entre elles : l'une commune à l'ensemble des séries, l'autre spécifique à la variable considérée. Le facteur commun fournit alors une estimation de l'indice composite recherché.

Le modèle de base peut se présenter simplement. Si :

- $y_{i,t}$ représente le solde d'opinion relatif à la question *i* pour la date *t*; *i* varie de 1 à *I* et *t* de 1 à *T*.
- $F_{j,t}$ représente la valeur du facteur commun *j* à la date *t*; *j* varie de 1 à *J*.
- $u_{i,t}$ représente la valeur de la composante spécifique *i* à la date *t*;

Alors le modèle s'écrit :

$$\forall i \in [1;I], \forall t \in [1;T] \quad y_{it} = \lambda_{i1} F_{1t} + ... + \lambda_{iJ} F_{Jt} + u_{it}$$
$$E(u_{it}) = 0, \quad E(F_{jt} u_{it}) = 0, \quad V(F_{1t},...,F_{Jt}) = \mathrm{Id},$$
$$V(u_{1t},...,u_{It}) = Diag(\sigma_1^2,...,\sigma_I^2) = \Sigma$$

Deux méthodes peuvent être utilisées pour estimer ce genre de modèles. La première se place dans le domaine des fréquences. Dans ce cas, la dynamique du modèle n'a pas besoin d'être spécifiée : les méthodes standard d'analyse factorielle (encore appelées analyse en composantes commune et spécifiques) peuvent être utilisées, de façon à décomposer la matrice de densité spectrale. La deuxième méthode relève plus directement du domaine temporel : dès lors que la dynamique des différentes composantes a été spécifiée, le modèle peut être mis sous une forme espace-état et estimé par le filtre de Kalman.

Doz et Lenglart ([11], [12]) montrent que les procédures standards d'analyse factorielle statique peuvent être utilisées dans ce contexte (au prix d'une perte d'efficacité) car elles fournissent des estimateurs convergents, même dans un cadre dynamique :

[3] La plus ancienne référence que je connaisse sur l'analyse factorielle dynamique est un travail de Gewege (1977, [20]).

« *Néanmoins, nous appliquons également aux données la technique standard de l'analyse factorielle. Certes, cette méthode n'est pas a priori appropriée dans un cadre d'analyse dynamique (elle a été créée au départ pour étudier des données individuelles). Mais nous montrons qu'elle fournit de estimateurs convergents des paramètres du modèle, même dans le cas où il y a présence d'autocorrélation temporelle des variables et où cette autocorrélation n'est pas prise en compte. En définitive, les programmes d'analyse factorielle standard peuvent donc être utilisés. De surcroît, ils offrent des éléments statistiques qui aident à choisir le nombre de facteurs communs à retenir. Ils offrent également des procédures de rotation d'axes qui facilitent l'interprétation des résultats obtenus lorsque plusieurs facteurs communs sont nécessaires pour décrire les données.*

Les résultats obtenus par l'une ou l'autre méthode sont toujours très proches, ce qui renforce leur crédibilité. »

2.2 Les errances de la pratique

L'analyse factorielle a dès lors été largement utilisée pour tenter d'estimer le fameux cycle des affaires et prévoir les points de retournement : la littérature regorge d'exemples de construction d'indicateurs coïncidents ou avancés. Peu habitués aux techniques d'analyse des données, les économètres sont alors tombés dans des « pièges » classiques.

Tout d'abord, il était tentant de mettre toutes les variables possibles et imaginables dans l'analyse, en espérant que l'analyse factorielle ferait automatiquement le travail de tri. Ainsi, en 2001, la Banque d'Italie (Altissimo et al, [1]) publie la méthodologie d'un indicateur coïncident mensuel du cycle des affaires de la zone Euro (EuroCOIN) défini comme le facteur commun d'une analyse factorielle sur 951 séries. Sur ces 951 séries, 258 apparaissent « avancées », 404 « coïncidentes » et 289 « retardées ». Comme la majorité des variables utilisées sont des variables liées à la production, le premier facteur commun est donc « naturellement » lié à la production, les variables monétaires étant tout aussi « naturellement » reléguées sur des axes secondaires.

Les vrais problèmes surgissent quand ce facteur commun est utilisé comme indicateur avancé et sert donc à prévoir une variable économique liée à la production comme le produit national brut ou l'indice de production industriel.

Quatre points méritent d'être mentionnés :

1. 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 !
2. Combien de facteurs doit-on retenir ? Doz et Lenglart ([11], [12]) montrent que les facteurs spécifiques de l'analyse peuvent apporter une information intéressante pour expliquer le cycle économique.
3. Les modèles sont en général difficilement interprétables du point de vue économique puisque chaque facteur est une combinaison linéaire de l'ensemble des variables de l'analyse factorielle.

4. La qualité de l'ajustement final dépend fortement du choix des variables prises en compte dans l'analyse factorielle. Comment faire ce choix ?

L'analyse factorielle, et l'analyse en composantes principales qui en est un cas particulier, sont nées au début du siècle dernier[4]. L'idée d'utiliser ces composantes principales dans des modèles de régression est venue assez vite et les problèmes ci-dessus ont été identifiés. La régression PLS (Tenenhaus, [43]), développée par Wold en 1966 ([46]), propose une solution aux deux premiers problèmes et a fait l'objet ces dernières années de nombreuses recherches : récemment, cette régression PLS a été adaptée au cas de séries temporelles (Preda et Saporta, [37] [38] [39]). Notons aussi que l'approche régression sur matrice de rang réduit a été étudiée par Cubbada (2004, [9]).

Les problèmes 3 et 4 se posent par exemple dans le cas de l'indicateur EuroCOIN. Nul doute que Altissimo *et al.* doivent rencontrer quelques soucis pour expliquer pourquoi cet indicateur monte ou descend. D'autres méthodes statistiques classiques pourraient être utilisées, même si elles ne sont pas toujours adaptées au cas temporel. La classification automatique vient immédiatement à l'esprit.

3 Classification de séries temporelles

La classification de séries temporelles est un problème qui a beaucoup occupé ces dernières années les statisticiens travaillant en médecine, biologie, météorologie, sismologie etc., domaines où d'énormes bases de données temporelles sont disponibles et doivent être analysées. Des dizaines de papiers ont été publiés et une bonne adresse pour commencer une recherche bibliographique sur la fouille de données temporelles est le site de Eamon Keogh (http://www.cs.ucr.edu/~eamonn/).

La classification a pour objectif de regrouper des objets dans des classes, définies à partir des données et non définies a priori, de telle sorte que les objets d'une même classe soient semblables et différents des objets des autres classes. Toute méthode de classification est donc basée sur un triplet :

- Une mesure de « similarité-dissimilarité » entre deux objets ;
- Une mesure de « similarité-dissimilarité » entre deux classes ;
- Et une stratégie d'agrégation des classes entre elles pour construire la partition.

De très nombreuses méthodes de classification ont été développées au fil des années pour les données individuelles et sont aujourd'hui disponibles dans les principaux logiciels de statistique : méthodes hiérarchiques (ascendante, descendante), méthodes non hiérarchiques (plus proches voisins, K-means etc.) etc.

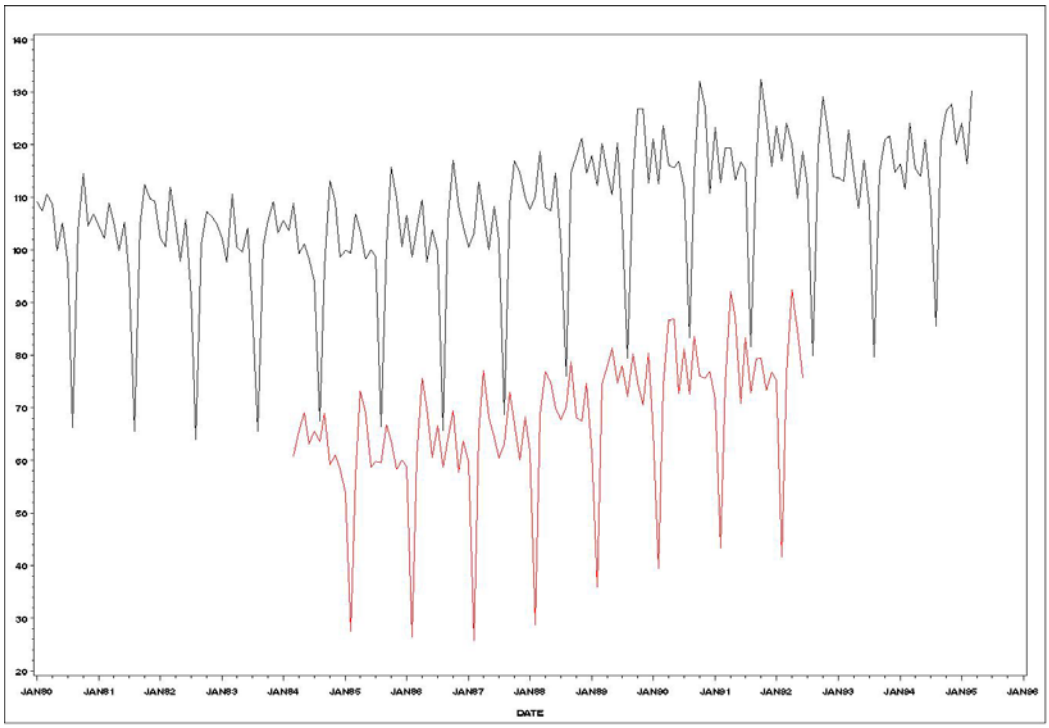
Des centaines de distances ont été proposées mais la plupart d'entre elles ne peuvent pas être utilisées directement sur des séries temporelles. Ainsi, la distance euclidienne, comme toutes les métriques de type Minkowski, donne des résultats « étranges » lorsqu'elle est appliquée sur des données temporelles brutes. En particulier, cette distance

[4] L'analyse factorielle a été proposée par Spearman en 1904 dans le cadre d'un facteur, puis généralisée à plusieurs facteurs par Garnett en 1919 (Fine, [17]).

est sensible aux problèmes d'unité et d'échelle, ne permet pas de comparer des séries de taille différentes, ne peut s'adapter aux décalages temporels et est très affectée par la présence de « non-linéarités » comme les valeurs manquantes, les points atypiques, les effets de calendrier etc.

Ainsi, la distance euclidienne entre les deux séries de la figure 1 peut-elle paraître artificiellement grande alors que ces deux séries seront dans certains cas considérées comme semblables : la série noire (X_t) et la série rouge (Y_t) sont en effet liées par la relation simple $Y_t = 0.75X_{t-6}$.

Figure 1 : des séries semblables ?



Outre ces problèmes de « niveau », de longueur et de décalage temporel, la taille des bases de données temporelles pose un réel problème de temps de calcul lorsqu'il s'agit de classer des centaines de milliers de séries.

3.1 Définir de nouvelles distances

De nouvelles mesures de similarité ont été récemment developpées : Dynamic Time Warping (DTW, Berndt & Clifford, 1994), Longest Common SubSequence (LCSS, Das et al., 1997), Edit Distance on Real sequence (EDR, Chen et al., 2003) etc.

Ces distances sont alors calculées peuvent alors être calculées directement sur les séries éventuellement transformées pour les rendre plus facilement comparables dans l'espace des temps. Les transformations les plus communes sont par exemples standardisation, lissage, désaisonnalisation, stationnarisation, interpolation etc.

Mais les calculs peuvent être très longs et il est alors plus efficace d'utiliser au préalable des méthodes de réduction du nombre de dimensions.

3.2 Changer d'espace de représentation

L'idée de base est de projeter les séries dans un espace en utilisant une transformation préservant les distances et de n'utiliser pour la classification qu'un faible nombre des coefficients de la transformation. Ainsi, on peut calculer le périodogramme de chaque série et l'interpoler sur un nombre réduit et défini a priori de fréquences.

Ces représentations des séries permettent le plus souvent de prendre en compte des distorsions observées dans l'espace des temps (décalage temporel par exemple) tout en réduisant fortement les temps de calcul. De nombreuses techniques de projection ou de décomposition ont été proposées, dont certaines sont directement applicables à des séries non stationnaires :

- Fonctions d'autocorrélation (ACF, PACF, IACF) (Maballée, 1911 ; Wang and Wang., 2000);
- Transformée de Fourier discrète, périodogramme (DFT) (Agrawal et al., 1993);
- Transformées par ondelettes avec bases de Daubechies ou de Haar (DWT), ou autres (Huntala et al., 1997);
- Polynômes de Chebyshev (Ng and Cai, 2004)
- Codage du Cepstrum (Linear Predictive Coding, LPC), (Kalpakis et al., 2001);
- Décomposition en valeurs singulières en utilisant par exemple une analyse en composantes principales (Korn et al., 1997; Cleveland, 2004);
- Smooth Localized Complex Exponential model (SLEX) (Huang et al., 2004);
- Différentes approximations par fonctions linéaires par morceaux
    - Piecewise Linear Approximation (Morikane et al., 2001);
    - Piecewise Aggregate Approximation (PAA) (Keogh et al., 2000);
    - Adaptive Piecewise Constant Approximation (APCA) (Keogh et al., 2001).

Le plus souvent, les algorithmes et distances usuels de classification pourront alors être utilisés sur les données ainsi transformées. Les paragraphes suivants montrent deux applications possibles en analyse de la conjoncture et en prévision.

4 Les différentes facettes de la saisonnalité

La désaisonnalisation est un traitement important en analyse des séries temporelles. Dès 1862, Jevons recommandait l'élimination de ces fluctuations périodiques :

« *Toute fluctuation périodique, qu'elle soit journalière, hebdomadaire, trimestrielle ou annuelle, doit être détectée et mise en évidence, non seulement pour l'étudier en tant que telle, mais aussi parce que ces variations périodiques doivent être évaluées et*

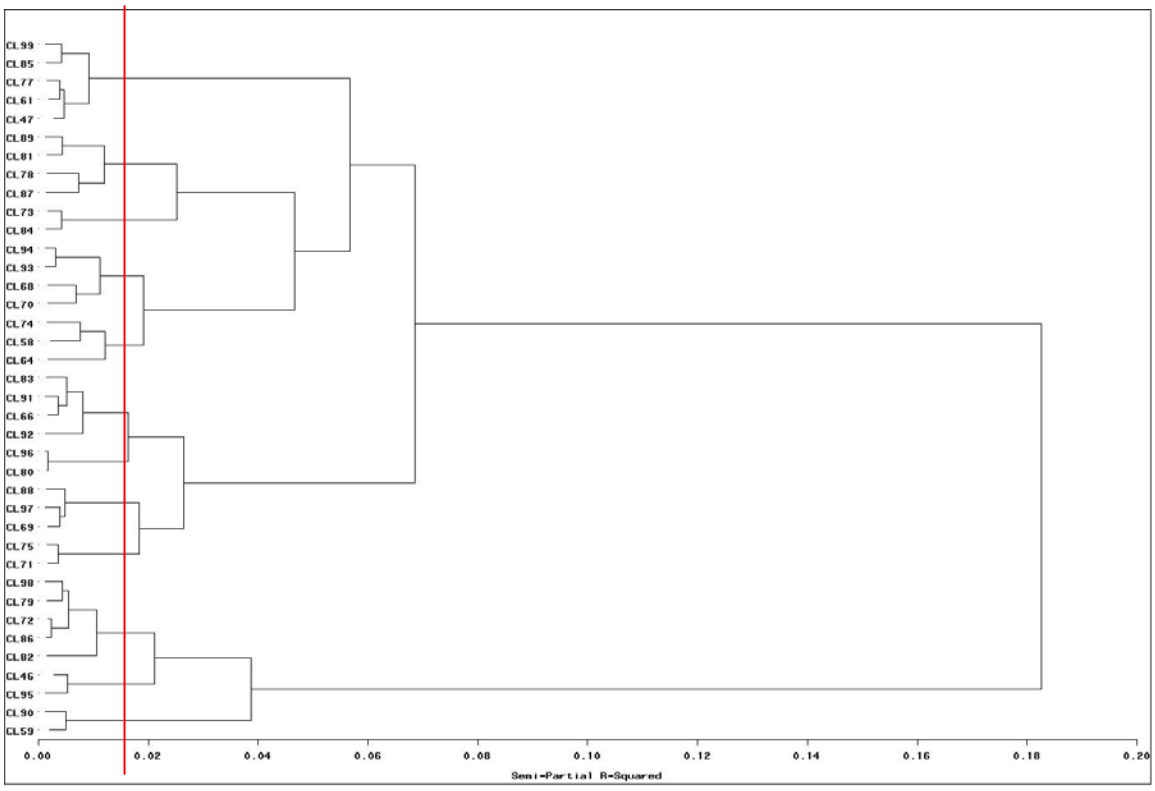
éliminées pour mieux faire ressortir celles qui, irrégulières ou non périodiques, sont probablement plus importantes et intéressantes »[5].

La correction des variations saisonnières est aussi une étape importante dans la prévision économique puisque le plus souvent, les modèles incorporent des variables dont le comportement saisonnier peut être très différent. Ainsi les secteurs des services et l'industrie ont des saisonnalités différentes mais l'emploi dans les services peut dépendre, via l'intérim, de l'activité industrielle.

De nos jours, la plus grande partie des corrections des variations saisonnières sont faites avec l'un des logiciels Tramo-Seats ou X-12-Arima. Ces logiciels possèdent des dizaines de paramètres qui permettent à l'utilisateur d'adapter sa désaisonnalisation aux caractéristiques de la série étudiée. Dans la pratique, ces logiciels sont utilisés pour ajuster des milliers de séries et les utilisateurs font confiance, pour la très large majorité de ces séries, aux valeurs par défaut des paramètres.

La classification permet de mettre en évidence les grandes familles de saisonnalités que l'on peut trouver dans les séries économiques.

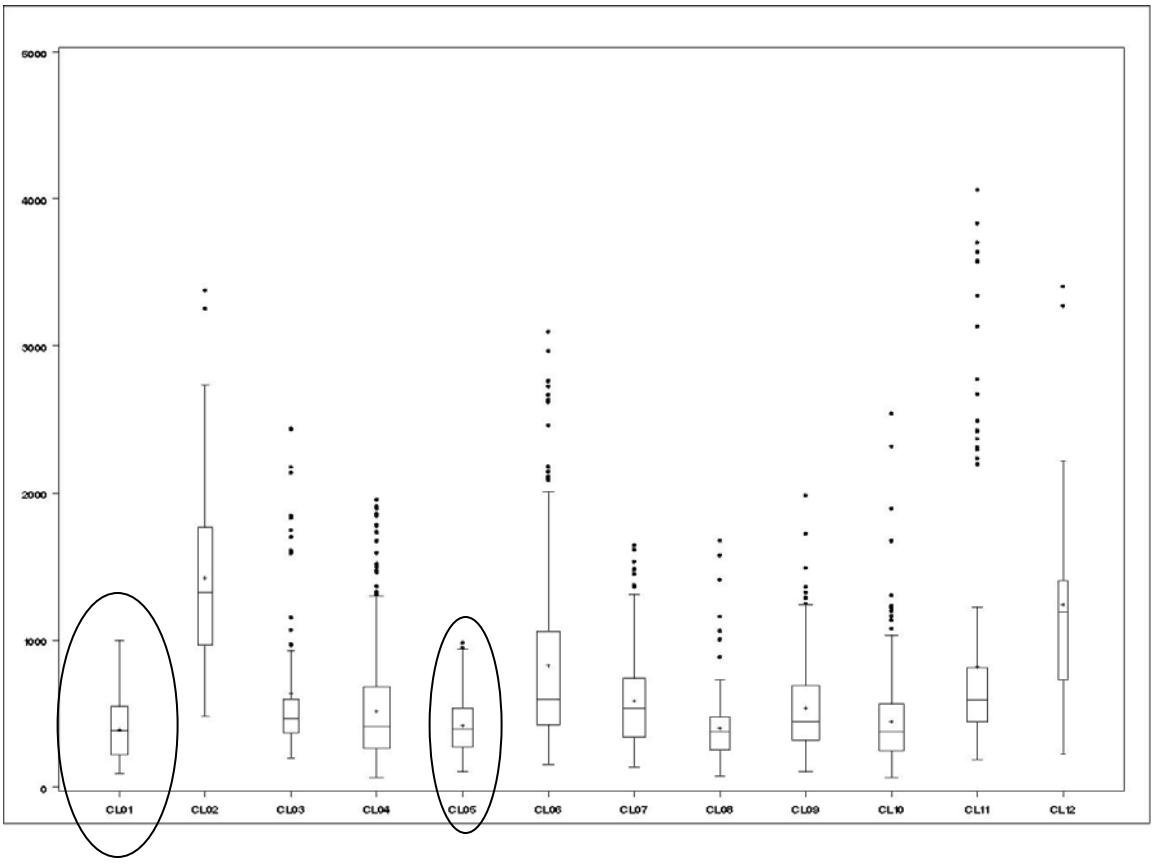
Figure 2 : le dendogramme révèle des formes de saisonnalités très différentes



[5] "*Every kind of periodic fluctuations, whether daily, weekly, quarterly, or yearly, must be detected and exhibited not only as a subject of study in itself, but because we must ascertain and eliminate such periodic variations before we can correctly exhibit those which are irregular or non-periodic and probably of more interest and importance*"

A titre d'exemple, 1100 séries mensuelles, de plus de 5 ans, on été extraites de la base de données Euro-Indicateurs disponible sur le site d'Eurostat. Ces séries ont été désaisonnalisées avec Tramo-Seats et X-12-Arima. Les spectres des 2200 composantes saisonnières ont été estimés par une transformée de Fourier rapide et, comme le nombre de points du spectre obtenu dépend de la longueur de la série, ces spectres ont été interpolés à l'aide de fonctions splines et évalués sur les mêmes 50 fréquences. Enfin, les 2200 spectres ont été standardisés pour éviter tout effet d'échelle. Une classification ascendante hiérarchique, avec stratégie de Ward, a enfin été réalisée sur ces spectres. L'arbre de la figure 2 résume la classification et montre qu'il existe des formes très variées de saisonnalités. Une représentation en 12 classes a été choisie. La figure 3 montre la dispersion des spectres des saisonnalités à l'intérieur de chaque classe : les classes 1 et 5 semblent par exemple très homogènes.

Figure 3 : Les boxplots montrent la dispersion des saisonnalités dans chaque classe et donc leur homogénéité.



Enfin, les figures 4 à 6 montrent quelques exemples de saisonnalités caractéristiques.

- La figure 4 montre des saisonnalités très semblables ;
- La figure 5 montre des saisonnalités décalées extraites de séries de taille variable ;
- La figure 6 montre des saisonnalités « inversées » mais très similaires.

Figure 4 : Des saisonnalités très semblables.

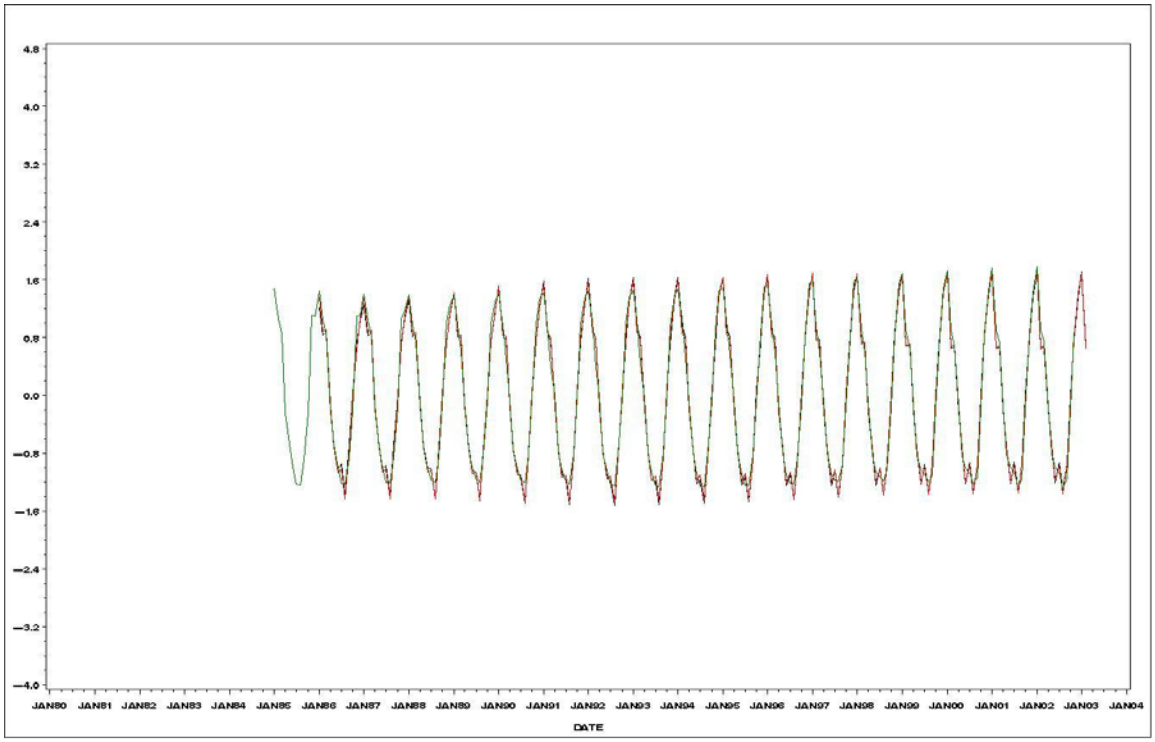


Figure 5 : Des saisonnalités de longueurs différentes et décalées.

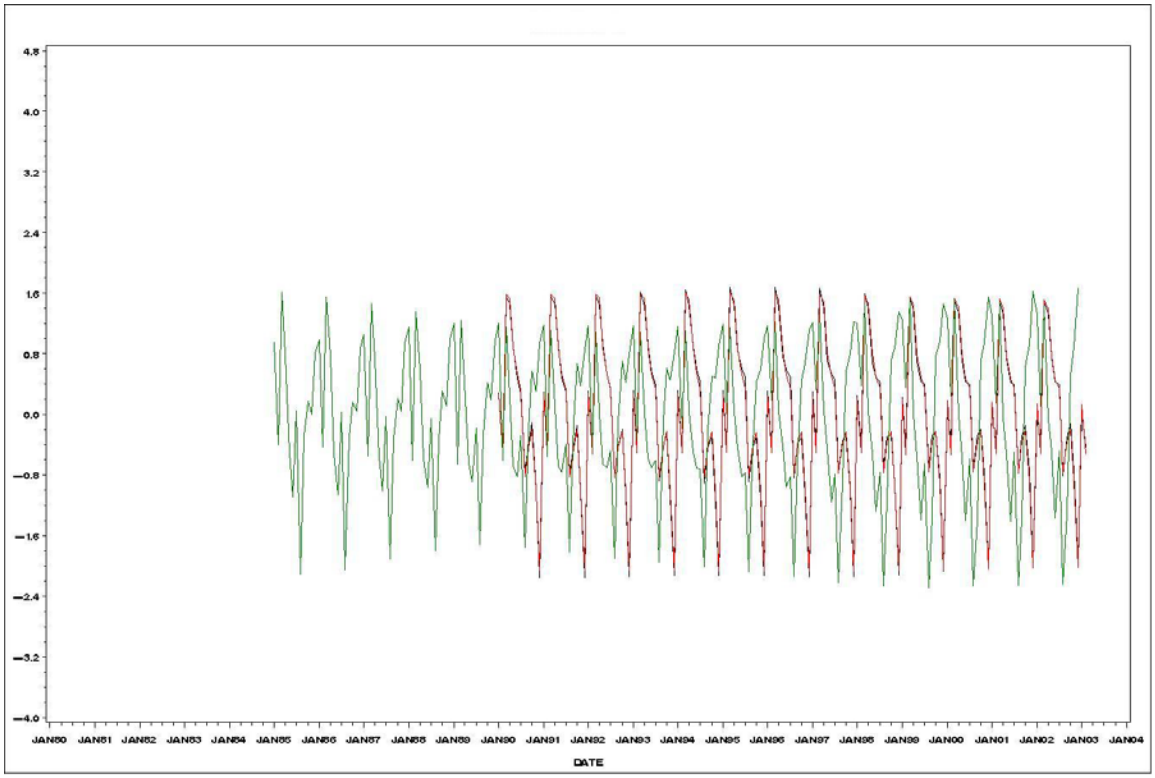
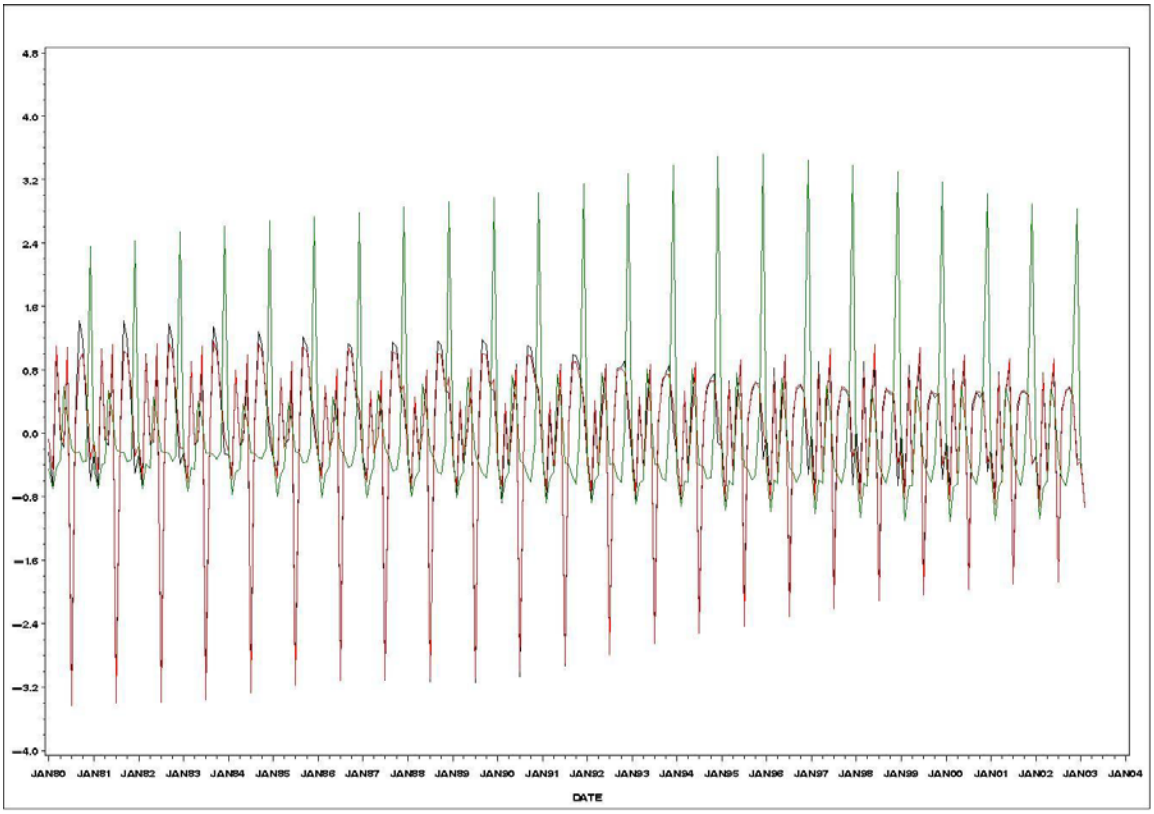


Figure 6 : Des saisonnalités inversées mais de même forme.



5 Prévision et classification

La division des enquêtes de conjoncture de l'INSEE réalise chaque année plus de 60 enquêtes auprès des professionnels des différents secteurs d'activité de l'économie française. Les questions posées, essentiellement qualitatives, permettent de connaître l'opinion des entrepreneurs sur la situation et sur les perspectives de leur entreprise et ce dans des domaines très divers: activité, demande, stocks, emploi, trésorerie, concurrence, investissements .....

Ces renseignements, recueillis le plus souvent à un rythme mensuel ou trimestriel, sont très utiles à l'analyse conjoncturelle puisqu' ils sont disponibles très rapidement (les résultats sont obtenus avant la fin du mois de réalisation de l'enquête) et qu'ils permettent d'anticiper, avec parfois quelques mois d'avance, le sens de variation des principaux indicateurs quantitatifs de l'activité économique qui, eux, ne seront connus que plus tard.

Mettre en concordance les résultats qualitatifs des enquêtes et l'évolution quantitative des indicateurs d'activité relève de procédures d'étalonnages, qui visent à établir des relations économétriques entre ces grandeurs dans une optique de prévision. 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 à chaque nouvelle enquête,

4. et enfin, ils doivent avoir un bon pouvoir prédictif.

5.1 A la recherche du modèle

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.

Ainsi, pour étalonner pour la première fois une variable, deux stratégies sont a priori possibles:

- Celle 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"; par ailleurs, il cherchera sans doute à l'améliorer en incorporant des variables retardées ....
- Celle du statisticien qui consisterait ici à chercher des modèles statistiquement corrects. Malheureusement, la combinatoire du problème est rapidement décourageante. Ainsi, l'enquête trimestrielle d'activité dans l'industrie contient une quarantaine de variables. Si on suppose a priori que des retards sur un an peuvent intervenir, on obtient alors environ 160 variables explicatives candidates. Il existe alors environ 670 000 modèles à trois variables et plus de 26000000 de modèles à quatre variables parmi ces variables candidates.

La pratique, qui est le plus souvent un compromis entre ces deux attitudes orthogonales, conduit toujours à de longs tâtonnements.

La méthode développée ici, même si elle met en œuvre des méthodes statistiques "pointues", 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.

Cette recherche se fait en cinq étapes principales:

- Dans un premier temps, les variables explicatives candidates sont regroupées en un petit nombre de classes homogènes, chaque classe comportant des variables qui "se ressemblent" et qui donc apportent une information de même nature. Cette classification se fait indépendamment de la variable à expliquer. Cette étape repose sur l'utilisation de méthodes de classification sur séries temporelles.
- La variable à expliquer intervient dans un second temps: on cherche dans chaque classe la variable, éventuellement retardée, la plus liée à la variable endogène. Cette recherche est faite sur la base d'études de corrélation (Spearman, Pearson) ou de tests de causalité (Hsiao).
- Dans un troisième temps, on recherche les meilleurs modèles incluant un nombre limité de ces variables sélectionnées. Ces modèles sont exhibés par des procédures de régression de type "stepwise" (les critères R square, Adjusted R square, Mallows' Cp peuvent être utilisés).

- Ensuite, ces modèles sont évalués quant à leur stabilité, leur pouvoir prédictif etc. Différents indicateurs statistiques sont alors calculés pour juger de leur stabilité (tests du CUSUM, de CHOW, de Ploberger-Cramer, indices de conditionnement ....) et de leur pouvoir prédictif (R-square, erreur quadratique moyenne en prévision à divers horizons ...)
- Enfin, il reste à choisir, parmi les modèles « statistiquement corrects » un modèle interprétable au sens économique.

5.2 Un exemple simple

L'exemple présenté ici date un peu (Ladiray, 1997, [29]) mais illustre bien le potentiel d'une méthode mixte exploratoire-confirmatoire.

L'objectif était de prévoir l'emploi dans le secteur des services, mesuré par les comptes nationaux trimestriels, en fonction des résultats des enquêtes de conjoncture. La variable à expliquer est le taux de croissance trimestriel de l'emploi et les variables potentiellement explicatives sont les opinions des chefs d'entreprises de l'industrie et du secteur des services, ce qui représente 48 variables.

1. Classification ascendante hiérarchique sur les variables

Le tableau 1 montre la répartition des variables en 7 classes. Cette répartition traduit en particulier une « opposition » entre variables « réelles » (production, carnets de commandes, stocks etc.) et « financières » (prix, salaires etc.).
Par nature, les variables très corrélées entre elles (carnets de commandes et carnets de commandes à l'exportation par exemple) se retrouvent dans les mêmes classes.

Tableau 1 : constitution des 7 classes

CLUSTER1	CLUSTER2	CLUSTER3	CLUSTER4	CLUSTER5	CLUSTER6	CLUSTER7
MG	CAPA	TPXEPA	CAXPR	DIOS	DEMCS	CS
OSC	CAPR	TPXPA	CSSK	DIREC	PGP	
OSCD	SALPR	TPXPRE		TXSAL	TDEPA	
OSCDE	REPA	VPXPA			TDEPRE	
OSCE	REPR	VPXPRE			TDPRE	
OSD	PVPA				TPPRE	
OSDE	PVPR				TSK	
OSSK	TRES				VPXEPA	
TDL	SALPA					
TDPA	CAXPA					
TPDT	CAPRO					
TPPA	DITRE					
TRDT	GTE					
	GTEP					
	MAPS					
	TU					

2. Le choix de la « meilleure » variable explicative

Dans chaque classe, on va chercher la « meilleure » variable explicative définie comme la variable, éventuellement retardée, qui est la plus « liée » à la variable à

expliquer. Ce lien peut être mesuré par des coefficients de corrélation (Pearson, Kendall, Spearman) mais aussi par des mesures de causalité (Granger, Hsiao).

3. La recherche des modèles

Le nombre des variables explicatives candidates est maintenant assez faible. Dans cet exemple, nous avons 7 classes, 4 retards potentiels (une année) et donc 28 variables explicatives potentielles soit environ 23000 modèles avec 4 variables ou moins. Tous ces modèles peuvent facilement et rapidement être examinés, selon un critère à déterminer, par des procédures de régression stepwise.

Les R2 ajustés sont présentés dans le tableau 2.

4. L'évaluation statistique des modèles et le choix du modèle final

La dernière étape de la procédure automatique concerne l'évaluation statistique des modèles sélectionnés, en termes de stabilité du modèle et de précision des prévisions. Les résultats de quelques statistiques sont présentés dans le tableau 2.

Les modèles sélectionnés par l'analyse des corrélations de Spearman (tableau 2) paraissent corrects, surtout en termes de stabilité. Malheureusement, ils restent difficiles à expliquer d'un point de vue économique.

L'utilisation de tests de causalité de Hsiao pour sélectionner les variables explicatives donne des résultats plus agréables (tableau 2, seconde partie). En particulier, les modèles 2 et 3 paraissent satisfaisants puisqu'ils lient l'évolution de l'emploi dans les services au chiffre d'affaires passé du secteur (CAPA), l'évolution des stocks dans l'industrie (variables CS ou TSK) et des difficultés de recrutement (DIOS). L'apparition simultanée de la variable DIOS et de cette même variable retardée d'un trimestre (DIOS_1) laisse supposer un effet de niveau, traduit par la variable DIOS elle même, et un effet de variation qui sera traduit éventuellement par la variable DIOS différenciée une fois. C'est ce qui a été fait dans le modèle finalement retenu.

Tableau 2 : Meilleurs modèles sélectionnés pour étalonner le glissement trimestriel de l'emploi dans les services et quelques statistiques de qualité.

METHODE=spearman , 7 CLASSES

MODEL	_IN_	_ADJRSQ_	EQMP	EQMP2	COND	CHOW	DW1	DW4	PK95	CUSUM_AV	CHOW_AV1
CAPA TPXPR_1 CAXPR TPPRE_2	4	0.704	0.40257	0.19749	34.3	0	2.60	2.45	0	0	1
OSSK_2 CAPA TPXPR_1 TPPRE_2	4	0.696	0.47194	0.26308	58.9	0	2.72	2.19	0	0	0
OSSK_3 CAPA TPXPR_1 TPPRE_2	4	0.694	0.46733	0.24115	41.3	0	2.47	2.19	0	0	1
OSSK_1 CAPA TPXPR_1 TPPRE_2	4	0.684	0.50881	0.26464	35.6	0	2.60	2.40	0	0	0
CAPA TPXPR_1 CAXPR_1 TPPRE_2	4	0.678	0.51339	0.24026	33.9	0	2.70	2.13	0	0	0
OSSK_3 CAPA TPPRE_1 TPPRE_2	4	0.673	0.46660	0.34412	37.3	0	2.57	2.01	0	0	1
CAPA TPXPR_1 TPPRE_2 TPPRE_3	4	0.670	0.49292	0.22221	45.2	0	2.44	2.24	0	0	1
OSSK CAPA TPXPR_1 TPPRE_2	4	0.670	1.40081	0.28951	35.1	0	2.62	2.32	0	0	0
CAPA TPXPR_1 DIREC TPPRE_2	4	0.670	1.04426	0.23867	30.8	0	2.63	2.46	0	0	0
CAPA TPXPRE TPXPR_1 TPPRE_2	4	0.669	0.48657	0.21629	51.0	0	2.61	2.34	0	0	0
CAPA CAXPR_1 TPPRE_1 TPPRE_2	4	0.669	0.78870	0.32813	32.9	0	2.78	1.99	0	0	1
CAPA TPXPR_1 TPPRE_2	3	0.668	0.50020	0.22601	26.0	0	2.49	2.34	0	0	0

METHODE=hsiao , 7 CLASSES

MODEL	_IN_	_ADJRSQ_	EQMP	EQMP2	COND	CHOW	DW1	DW4	PK95	CUSUM_AV	CHOW_AV1
CAXPR_1 DIOS_1 DIOS_2 TSK	4	0.708	0.72250	0.35957	16.2	0	2.83	2.55	0	0	1
CAPA DIOS DIOS_1 TSK	4	0.690	0.56341	0.27076	16.1	0	2.58	2.45	0	0	1
CAPA DIOS DIOS_1 CS	4	0.686	0.97175	0.28013	20.0	0	2.53	2.45	0	0	1
CAPA CAXPR DIOS TSK_2	4	0.685	0.46896	0.21489	13.7	0	2.71	2.38	0	0	1
CAPA CAXPR_1 DIOS_1 TSK	4	0.682	0.59830	0.31036	12.5	0	3.04	2.35	0	0	1
CAPA DIOS TSK_2 CS	4	0.680	0.87749	0.24574	12.6	0	2.85	2.53	0	0	1
OSC CAPA DIOS TSK_2	4	0.678	0.85595	0.26981	54.7	0	2.77	2.29	0	0	1
CAPA DIOS_1 DIOS_2 TSK	4	0.678	0.58717	0.31417	17.2	0	2.56	2.38	0	0	1
CAPA CAXPR_1 TSK	3	0.678	0.44048	0.31955	10.5	0	3.04	2.17	0	0	0
CAPA CAXPR TSK TSK_2	4	0.675	0.45376	0.28092	13.3	0	2.96	2.42	0	0	2
CAPA CAXPR_1 TSK TSK_2	4	0.675	0.53511	0.31738	12.0	0	3.18	2.23	0	0	0
CAPA CAXPR_1 TSK CS	4	0.674	0.61446	0.32426	13.9	0	3.18	2.18	0	0	0
CAPA DIOS DIOS_1 TSK_2	4	0.673	0.60983	0.23017	17.4	0	2.64	2.48	0	0	0
CAPA DIOS DIOS_1 DIOS_2	4	0.672	0.62356	0.28555	23.2	0	2.43	2.24	0	0	0
CAPA CAXPR CAXPR_1 TSK	4	0.672	0.58993	0.32123	14.2	0	2.95	2.13	0	0	1

Pour bien lire les tableaux:

EQMP (respectivement EQMP2) est l'écart quadratique moyen des erreurs de prévision sur toute la période d'estimation (respectivement sur les deux dernières années). Il s'agit bien là d'indicateurs dynamiques qui synthétisent les erreurs que l'on aurait faites à l'époque avec le modèle testé.

COND est l'indice de conditionnement de la matrice des variables explicatives. Plus il est élevé, plus les colinéarités entre variables sont importantes. Une règle empirique indique que la valeur 15 est un seuil important.

DW1 est la valeur du Durbin-Watson à l'ordre 1. Pour les modèles sélectionnés (en gris), le DW de 2.5 indique une autocorrélation des résidus à prendre en compte dans l'estimation des paramètres de la régression, ce qui a été fait pour le modèle proposé.

Les autres tests sont des tests de stabilité. Ainsi CHOW indique le nombre de ruptures, au sens de CHOW, détectées.

6 Conclusion

L'analyse de données temporelles a donc fait des progrès très significatifs ces dernières années, notamment dans les domaines où la mise à disposition de très grandes bases de données a amené les statisticiens à faire évoluer les méthodes de fouille des données.

Ces méthodes ont cependant du mal à pénétrer le monde des économètres et celui des prévisions économiques de court terme. C'est dommage dans la mesure où les économètres rencontrent, sur des données de type temporel, les mêmes problèmes et difficultés que les statisticiens ont rencontré sur les données d'enquête. La façon dont ces problèmes ont été abordés, étudiés et pour certains d'entre eux résolus est une mine de renseignements pour permettre d'améliorer les méthodes économétriques actuelles.

En particulier, la classification, la régression PLS et l'analyse discriminante devrait bientôt faire leur entrée dans la panoplie des outils de l'économètre …. Et dans les principaux journaux économétriques.

Les progrès seront sans aucun doute fulgurants ainsi que le montrent les titres de quelques articles trouvés, ça et là, depuis le début de l'année :

- Aminghafari, M., Cheze, N., Poggi, J-M. (2006), Multivariate denoising using wavelets and principal component analysis, *Computational Statistics and Data Analysis*, Vol. 50, n° 9, pp 2381-2398
- Bair E., Hastie, T., Paul, D., Tibshirani, R., (2006), Prediction by Supervised Principal Components, *Journal of the American Statistical Association*, Vol. 101, n°. 473, pp.119-137
- Ombao, Ringo Ho (2006), Time-dependent frequency domain PCA of multichannel non-stationary signals, *Computational Statistics and Data Analysis.*
- Pena, Poncela (2006), Nonstationary dynamic factor analysis, Journal of *Statistical Planning and Inference*
- Raftery, A.E., Dean, N., (2006), Variable selection for Model-based Clustering, *Journal of the American Statistical Association*, Vol. 101, n° 473, pp.168-178

7 Bibliographie

[1] Agrawal, R., Faloutsosa, C., Swami, A. (1993), Efficient Similarity Search in Sequence Databases. *Lecture Notes in Computer Science* 730, Pages 69-84 Springer Verlag, 1993

[2] Altissimo, F., Bassanetti, A., Cristadoro, R., Forni, M., Lippi, M., Reichlin, L., Veronese, G. (2001), "A real time coincident indicator of the euro area business cycle", document de travail 436, Banque d'Italie, Rome.

[3] Aminghafari, M., Cheze, N., Poggi, J-M. (2006), Multivariate denoising using wavelets and principal component analysis, *Computational Statistics and Data Analysis*, Vol. 50, n° 9, pp 2381-2398

[4] Bair E., Hastie, T., Paul, D., Tibshirani, R., (2006), Prediction by Supervised Principal Components, *Journal of the American Statistical Association*, Vol. 101, n°. 473, pp.119-137

[5] Berndt, D., Clifford, J. (1994). *Using dynamic time warping to find patterns in time series.* AAAI-94 Workshop on Knowledge Discovery in Databases.

[6] Burns, A.F., Mitchell, W.C., (1946), *Measuring Business Cycles*, National Bureau of Economic Research, Cambridge, MA.

[7] 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.

[8] Cleveland, W. P., (2004), Stability and Consistency of Seasonally Adjusted Aggregates and Their Component Patterns, *Studies in Nonlinear Dynamics & Econometrics*: Vol. 8: No. 2, Article 15.

[9] Cubbada, G. (2004), *A Reduced Rank Regression Approach to Coincident and Leading Indexes Building*, Economics & Statistics Discussion Papers, University of Molise, Dept. SEGeS

[10] 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.

[11] Doz, C., Lenglart, F. (1997), Analyse factorielle et modèles à composantes inobservables, *INSEE Méthodes* n° 69-70-71, INSEE, Paris.

[12] Doz, C., Lenglart, 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°54, INSEE, Paris.

[13] Engle, R.F., Granger, C.W.J. (1987). Cointegration and Error: Representation, Estimation, and Testing, *Econometrica*, 55, 251-276

[14] European Commission (1997), "The Joint Harmonized EU Program of Business and Consumer Surveys", *European Economy*, 6, Bruxelles.

[15] European Commission, (2000), Business Climate Indicator for the Euro Area, presentation paper, Directorate General Economic and financial affairs, Bruxelles.

[16] European Commission (2001), "Business and Consumer Surveys: Results", *European Economy Supplement B*, 8-9, Bruxelles.

[17] Fine J. (1992), « Modèles fonctionnels et structurels », dans *Modèles pour l'analyse des données multidimensionnelles*, éditeurs Droesbecke, Fichet, Tassi, Economica

[18] Forni, M., Reichlin, L. (1998), Let's get real: a dynamic factor analytical approach to disaggregated business cycle, *Review of Economic Studies*.
[19] Forni, M., Hallin, M., Lippi, M., Reichlin, L. (2001), Coincident and Leading Indicators for the Euro Area, *Economic Journal*, n° 471, Vol. 111, pp. 62-85.
[20] Geweke, J. (1977), .The Dynamic Factor Analysis of Economic Time Series., in D.J. Aigner and A.S. Golberger (eds.): *Latent Variables in Socio-Economic Models*, Amsterdam, North-Holland, Ch. 19.
[21] 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.
[22] 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.
[23] Jevons, W.S. (1862), On the Study of Periodic Commercial Fluctuations, Investigations in currency and finance, London: Macmillan, 1884.
[24] Johansen, S. (1988), Statistical analysis of cointegrating vectors, *Journal of Economic Dynamics and Control*, 12, pp. 231-254
[25] 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.
[26] 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
[27] 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.
[28] 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*.
[29] Ladiray, D. (1995), *Using Business Survey Data for Quantitative Forecasts*, 22nd CIRET conference, Singapore
[30] Maballée, Colette et Berthe, (1911), Classification of Time Series and Forecasting: The SiNCiD Method, *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 3, 159-167.
[31] Mirowski, P. (1989), The Probabilistic Counter-Revolution, or how Stochastic Concepts came to Neoclassical Economic Theory, *Oxford Economic Papers*
[32] 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.
[33] Ng, R. T., Cai, Y., (2004), Indexing Spatio-Temporal Trajectories with Chebyshev Polynomials. *Proceedings of SIGMOD 2004*
[34] Ombao, Ringo Ho (2006), Time-dependent frequency domain PCA of multichannel non-stationary signals, *Computational Statistics and Data Analysis.*
[35] Pena, Poncela (2006), Nonstationary dynamic factor analysis, Journal of *Statistical Planning and Inference*

[36] Persons, W.M. (1919), Indices of Business Conditions, *Review of Economic Statistics* n°1, pp 5-107

[37] Preda, C., Saporta, G. (2002), Régression PLS sur un processus stochastique, *Revue de Statistique Appliquée*, vol. 50.

[38] Preda, C., Saporta, G. (2005), Clusterwise PLS regression on a stochastic process, *Computational Statistics and Data Analysis*, Vol. 49, n° 1, pp. 99-108.

[39] Preda, C., Saporta, G. (2005), PLS regression on a stochastic process, *Computational Statistics and Data Analysis*, Vol. 48, n° 1, pp. 149-158.

[40] Quah D., Sargent, T.J. (1993), A dynamic index model for large cross-sections, in *Business cycles, indicators and forecasting*, J.H. Stock and M.W. Watson Ed., University of Chicago Press

[41] Raftery, A.E., Dean, N., (2006), Variable selection for Model-based Clustering, *Journal of the American Statistical Association*, Vol. 101, n° 473, pp.168-178

[42] Stock, J.H., Watson, M.W. (1993), A procedure for predicting recessions with leading indicators: econometric issues and recent experience, in *Business cycles, indicators and forecasting*, J.H. Stock and M.W. Watson Ed., University of Chicago Press.

[43] Tenenhaus M. (1998), *La régression PLS, théorie et pratique*, Technip.

[44] Tukey, J.W. (1980), We need both exploratory and confirmatory statistics, *The American Statistician*, Vol. 34, n°1, pp. 23-25

[45] 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.

[46] Wold H. (1966), « Estimation of Principal Components and Related Models by Iterative Least Squares », in *Multivariate Analysis*, ed. P. R. Krishnaiah, New York: Academic Press, pp. 391-420.