Modélisation de séries temporelles multivariées par équations structurelles via la segmentation

Christian Derquenne[1]

[1] *EDF - Recherche et Développement - 1, avenue du Général de Gaulle - 92140 Clamart*

Résumé. Nous proposons une méthode pour construire un modèle complexe pour séries temporelles irrégulières et multivariées. Pour cela, afin de tenir compte de non linéarité, de rupture, de volatilité entre celles-ci, nous les normalisons à l'aide de leurs versions segmentées. Puis, nous avons utilisé une approche exploratoire à l'aide d'un modèle à équations structurelles libre pour établir des liens entre ces variables normalisées.

Mots-clés. Séries temporelles, modèles à équations structurelles, segmentation.

Abstract. We propose a method to build a complex model for irregular and multivariate time series. For that, to take account of non-linearity, breakdowns, and volatility between them, we normalize with their segmented versions. Then, we used an exploratory approach using a free structural equations model linking these standardized variables.

Keywords. Time series, cointegration, structural equations modeling, segmentation.

1 Contexte et problème

Dans de nombreuses applications en finance, en environnement, en management de l'énergie, en fiabilité, ... les données peuvent être observées sous forme de séries temporelles univariées ou multivariées. Pour des données univariées, il est d'usage de stationnariser la série par différenciation, puis d'appliquer par exemple, un modèle ARMA ou un modèle de cointégration sur les données résiduelles, si l'on dispose de prédicteurs temporels. Dans le cas multivarié, des approches par modèle espaces-états ou par modèle de cointégration multivarié sont utilisés. Lorsqu'il y a des prédicteurs pour modéliser une réponse univariée ou multivariée, une structure du modèle mettant en lien certaines variables peut être imposée par l'expert. Cette structure est représentée sous la forme d'un graphe (modèle graphique) correspondant aux équations du modèle choisi. Ces modèles graphiques sont également utilisés pour des données non temporelles et sont nommés : modèles à équations structurelles (SEM). Ils sont composés d'un modèle externe (ou de mesure) qui lie les variables observables à des variables latentes (inobservables) et d'un modèle interne (ou de structure) liant certaines variables latentes entre elles.

Deux approches statistiques sont généralement utilisées pour estimer ces modèles. La méthode LISREL (LInear Structural RELations) (Jöreskog, 1979) qui modélise la matrice de covariances entre les variables endogènes et exogènes en fonction de la structure imposée par l'expert du domaine. L'estimateur du maximum de vraisemblance est généralement utilisé. La seconde méthode est l'approche PLS (Partial Least Squares) introduite par Wold H. (1982) ; elle modélise les variables observées en partant de la structure des individus, à l'aide de l'estimateur des moindres carrés. Quel que soit le type de données : statique, dynamique, univarié, multivarié,

ces modèles théoriques sont proposés par un expert. A l'opposé, lorsque l'on ne fixe pas un modèle a priori, les groupes de variables observées (modèle externe), les liens entre les variables latentes et leurs orientations (modèle interne), alors le modèle est dit "libre" (Derquenne et al., 2003). L'approche de construction du modèle est exploratoire. Par ailleurs, les relations entre les variables observées qu'elles soient statiques ou dynamiques ne sont pas toujours linéaires. Par conséquent, l'usage des approches précédentes n'est plus valable de façon raisonnable.

Dans le cas des séries temporelles, l'évolution des phénomènes peut être non linéaire ou linéaire mais avec une variabilité non constante, avec rupture. Pour pallier ce problème, il est possible de découper la série sous forme de segments linéaires, à l'aide de méthodes de segmentation (Arlot, 2010, Lavielle, 2006, Derquenne, 2010). Le découpage peut être très fructueux dans la recherche de liens entre séries chronologiques. La série temporelle peut alors être stationnarisée par morceaux en la normalisant à l'aide d'informations fournies par la segmentation (Derquenne, 2013). Les relations entre ces variables transformées pourront alors être analysées et utilisées pour construire des modèles espaces-états ou des modèles à équations structurelles car non seulement elles présentent, par construction, l'avantage d'être stationnarisées, mais aussi de fournir soit une relation linéaire, soit une évolution présentant une absence complète de lien.

Dans ce papier, nous présentons briévement la méthode de segmentation utilisée, puis nous formalisons les modèles à équations structurelles figés et libres reposant sur la segmentation, enfin nous appliquons cette démarche à un exemple d'application réel, mais anonymisé. Nous concluons alors sur les apports, les limites et les perspectives de ce travail.

2 Modélisation des séries temporelles : SEM et segmentation

2.1 Segmentation et stationnarisation de séries temporelles

Nous avons introduit une méthode de segmentation de séries temporelles (Derquenne, 2010) complètement non-supervisée qui permet de réduire la complexité par rapport à d'autres méthodes, mais surtout propose des solutions de segmentation de la série contenant des segments croissants, décroissants, constants et des dispersions différents. Elle contient deux phases principales : la préparation des données offrant une première segmentation des données et la modélisation des segments à l'aide d'un modèle linéaire gaussien hétéroscédastique par adaptations successives. Afin d'améliorer cette méthode, nous avons introduit, dans une phase préalable, une meilleure prise en compte des composantes de variance de la série à l'aide d'une transformation adéquate des données (Derquenne, 2011), ainsi qu'une approche par méta-segmentation (Derquenne, 2012) qui consiste à agréger les "meilleures" segmentations.

Soient $(X_1, ..., X_p)$, p séries temporelles de taille T et $(\tilde{X}_1, ..., \tilde{X}_p)$ leurs transformations sous forme de segments. Chaque $\tilde{X}_j$ possède m_j segments de longueurs respectives T_{jk}, avec $\sum_{k=1}^{m_j} T_{jk} = T$. Grâce à la segmentation, chaque estimation de $X_{j(t)}$ a la forme suivante : $\tilde{X}_{j(t)} = \sum_{k=1}^{m_j} (\alpha_{jk}^{(0)} + \alpha_{jk}^{(1)} t).1_{[t \in \tau_{jk}]}$ où τ_{jk} correspond au segment numéro k de la segmentation de X_j. Enfin, on désigne par $(X_1^*, ..., X_p^*)$, les séries temporelles stationnarisées grâce aux segmentations respectives, telles que $X_{j(t)}^* = ((X_{j(t)} - \tilde{X}_{j(t)})/s_{jk}).1_{[t \in \tau_{jk}]}$, où $s_{jk} = \sqrt{\sum_{t \in \tau_{jk}} (X_{j(t)} - \tilde{X}_{j(t)})^2 / T_{jk}}$.

2.2 Modèles à équations structurelles pour séries temporelles

Soient $(Y_1, ..., Y_q)$, q réponses et $(X_1, ..., X_p)$, p prédicteurs sous forme de séries temporelles de taille T. On note ΔX_i, la série X_i différenciée. Un modèle de cointégration multivarié est de la forme : $\Delta Y_{1,t} = \alpha_{11}\Delta X_{1,t-1} + \alpha_{12}\Delta X_{1,t} + \epsilon_{1,t}$ et $\Delta Y_{2,t} = \beta_1 \Delta Y_{1,t} + \alpha_{21}\Delta X_{2,t-1} + \alpha_{22}\Delta X_{2,t} + \epsilon_{1,t}$, où les α's sont les coefficients liant les prédicteurs aux réponses, β_1 celui qui lie les deux réponses et les ϵ's sont des bruits blancs indépendants. Son modèle graphique (fig. 1-a) est :

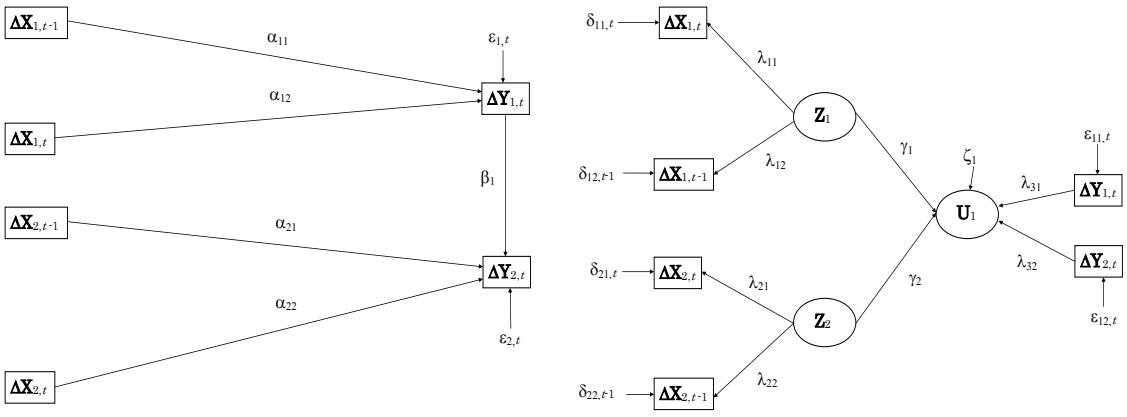

Figure 1: (a) Modèle de cointégration (b) Modèle à équations structurelles

Une autre approche consiste à utiliser des modèles à équations structurelles. Le modèle de mesure lie les variables observables (X, Y) à des variables latentes ou inobservables (Z, U) ; le modèle de structure lie certaines variables latentes entre elles (cf. fig. 1-b). Dans l'exemple suivant, le modèle de mesure est de la forme : $\Delta X_{1,t} = \lambda_{11} Z_1 + \delta_{11,t}$; $\Delta X_{1,t-1} = \lambda_{12} Z_1 + \delta_{12,t-1}$; $\Delta X_{2,t} = \lambda_{21} Z_2 + \delta_{21,t}$; $\Delta X_{2,t-1} = \lambda_{22} Z_2 + \delta_{22,t-1}$; $\Delta Y_{1,t} = \lambda_{31} U_1 + \epsilon_{11,t}$; $\Delta Y_{2,t} = \lambda_{32} U_1 + \epsilon_{12,t}$, alors que le modèle de structure est : $U_1 = \gamma_1 Z_1 + \gamma_2 Z_2 + \zeta_1$. Dans ce cas, chaque série temporelle (variable manifeste) est le reflet de sa variable latente Z_1, Z_2 ou U_1. L'hypothèse d'unidimensionnalité de chaque bloc de variables manifestes est donc requise. En d'autres termes, ces variables sont dépendantes et complémentaires les unes par rapport aux autres. Les loadinds λ's permettent de modéliser les variables manifestes à l'aide de leur variable latente associée, les ϵ's et les δs correspondent aux erreurs de mesure, les coefficients γ's lient la variable latente endogène U_1 à ses variables latentes exogènes Z_1 et Z_2. Enfin, ζ_1 est l'erreur de prédiction de U_1. L'ajustement est réalisé par la méthode LISREL ou par l'approche PLS.

Un des problèmes majeurs dans la modélisation de séries temporelles, munis de prédicteurs, est qu'ils soient fortement liés entre eux pour expliquer un ou plusieurs phénomène(s). Il apparait alors de la multicolinéarité entre variables candidates à l'explication et cela peut provoquer des signes de coefficients opposés à ceux de la relation bivariée entre le prédicteur et la réponse, ou encore rendre artificiellement grande la variance des coefficients, provoquant une diminution de la statistique de test, et par conséquent la non signification d'apport statistique du prédicteur à la variable à expliquer. Dans ce cas, les modèles de cointégration deviennent inutilisables si l'on

désire garder les prédicteurs choisis par l'expert afin de comprendre la formation de la réponse.

L'utilisation des séries normalisées permet alors d'utiliser des modèles de cointégration mais pour des configurations plus simples que celles possibles dans les modèles à équations structurelles. Deux voies sont raisonnables pour une analyse statistique plus fine, l'expert fixe son modèle théorique, alors la méthode d'estimation SEM a pour objectif de confirmer ou d'infirmer ses hypothèses. A l'opposé, il arrive dans de nombreux cas que les relations entre les variables ne soient pas complètement connues. Dans ce cas, une approche exploratoire sera préférée. Il s'agira alors de construire un modèle "libre" à l'aide de l'information disponible dans les données.

Le principe de construction de ce type de modèle se déroule en cinq étapes : (i) construction des blocs de variables, (ii) estimation des variables latentes, (iii) établissement des liens statistiques entre ces variables latentes, (iv) orientation des liaisons entre les variables latentes et (v) application de la méthode LISREL ou de l'approche PLS pour estimer le modèle libre proposé. La première étape revient à classifier les séries temporelles standardisées ou différenciées à l'aide de l'analyse factorielle exploratoire (ACP avec rotation oblique, par exemple). Cela permet d'obtenir des groupes unidimensionnels de variables (toutes les valeurs propres sont inférieures à 1, sauf la plus grande). L'estimation des variables latentes consiste à sélectionner pour chaque bloc de variables, la première composante principale. Les liens entre cet ensemble de premières composantes principales sont construits à l'aide des coefficients de corrélation linéaire partielle. Un lien entre deux variables latentes sera jugé significatif si la p-valeur du test est inférieure à un seuil fixé (par exemple, 0,01). Puis, l'orientation des liens est généralement laissé à l'expert, bien qu'il soit possible d'optimiser un modèle structurel en maximisant, par exemple, un R^2 global. Enfin, le modèle libre proposé est estimé à l'aide de LISREL ou de l'approche PLS.

3 Application de la méthode

Nous disposons de 6 prédicteurs $(X_1, ..., X_6)$ et de 7 réponses $(Y_1, ..., Y_7)$ sous forme de séries temporelles. De nombreux traitements ont été effectués sur ces données réelles anonymisées, tels que des modèles de cointégration, des modèles à équations structurelles figés et libres. Nous présentons seulement les résultats de deux modèles libres. Le premier est construit sur les séries différenciées ; le second est appliqué sur des courbes temporelles standardisées par segmentation. De plus, nous avons ajouté à ses 14 séries temporelles, leurs séries retardées en $t-1$. Pour chaque modèle, la classification des variables a été appliquée séparément sur l'ensemble des X_t, X_{t-1}'s et des Y_t, Y_{t-1}'s afin de garder un aspect explicatif. Pour chacun des deux modèles, les mêmes nombres de groupes pour les X's et les Y's ont été obtenus, 6 et 8, respectivement (cf. table 1). Nous constatons que les contenus des groupes des deux modèles sont relativement similaires, autant pour les X' que pour les Y's. La seule différence est le regroupement de séries avec leurs séries retardées pour le modèle par segmentation : $G_5^{(X)}$ et $G_1^{(Y)}$.

Les figures 2-a et 2-b affichent les deux modèles à équations structurelles libres construits à l'aide de l'approche PLS. Pour le modèle sur données différenciées, les variables latentes associées aux groupes des prédicteurs et des réponses sont respectivement nommées $\mathbf{V}_g$ et $\mathbf{W}_g$, où g est le numéro du groupe associé ; pour le second modèle, nous avons $\mathbf{Z}_g$ et $\mathbf{U}_g$. Les liens entre variables latentes ont été établis seulement si la p-valeur du test du coefficient de corrélation partielle était inférieure à 0,01 afin d'éviter des modèles trop complexes. 27 et 23 arrêtes, respectivement, ont

été obtenues pour les modèles sur séries différenciées et normalisées. De plus, les variables latentes $\mathbf{W}_7$, $\mathbf{W}_8$ et $\mathbf{U}_5$ n'ont pas de lien significatif avec les autres. Cela correspond à la série Y_6 et son retard. Dans le premier modèle quatre variables latentes ($\mathbf{W}_2$,$\mathbf{W}_3$,$\mathbf{W}_4$,$\mathbf{W}_5$) sont des cibles (elles ne pointent sur aucune autre variable latente), alors que dans le modèle sur séries normalisées par segmentation, seules deux cibles ont été trouvées : $\mathbf{U}_3$ et $\mathbf{U}_8$ (les R^2 associés sont encadrés).

Deux cibles sont communes aux deux modèles : $\mathbf{W}_5$ et $\mathbf{U}_3$ résumant $Y_5^{(t)}$ et $Y_7^{(t)}$. Celles-ci correspondent à des réponses très importantes en termes métier. Cependant les variables latentes qui l'expliquent (t de Student sur les fléches) dans les deux modèles ne contiennent pas les mêmes variables : $X_2^{(t)}$, pour le modèle sur séries différenciées et $X_3^{(t)}, X_4^{(t)}, X_5^{(t)}$ pour le modèle sur séries normalisées. Ces trois dernières sont apparues plus cohérentes que la première en termes métier. Les cibles $\mathbf{W}_4$ et $\mathbf{U}_8$ sont également très porteuses d'information avec leurs variables : $Y_2^{(t)}, Y_4^{(t)}$ et $Y_2^{(t)}$, respectivement. Ce qui n'est pas le cas pour la cible $\mathbf{W}_2$. Pour conclure, le modèle structurel sur données différenciées a tendance à sur-évaluer (t de Student et R^2, plus élevés) des liens entre variables latentes par rapport au modèle utilisant des séries normalisées par segmentation, ce qui peut être symptomatique, par exemple d'une mauvaise prise en compte de relations non linéaires.

Grp. X	**Modèle avec diff.**	**Modèle avec segm.**	**Grp.** Y	**Modèle avec diff.**	**Modèle avec segm.**
$G_1^{(X)}$	$X_3^{(t-1)}, X_4^{(t-1)}, X_5^{(t-1)}$	$X_3^{(t-1)}, X_4^{(t-1)}, X_5^{(t-1)}$	$G_1^{(Y)}$	$Y_5^{(t-1)}, Y_7^{(t-1)}$	$Y_4^{(t-1)}, Y_4^{(t)}$
$G_2^{(X)}$	$X_3^{(t)}, X_4^{(t)}, X_5^{(t)}$	$X_6^{(t)}, X_7^{(t)}$	$G_2^{(Y)}$	$Y_1^{(t)}, Y_3^{(t)}$	$Y_1^{(t-1)}, Y_3^{(t-1)}$
$G_3^{(X)}$	$X_2^{(t-1)}, X_6^{(t-1)}, X_7^{(t-1)}$	$X_3^{(t)}, X_4^{(t)}, X_5^{(t)}$	$G_3^{(X)}$	$Y_2^{(t-1)}, Y_4^{(t-1)}$	$Y_5^{(t)}, Y_7^{(t)}$
$G_4^{(X)}$	$X_2^{(t)}, X_6^{(t)}, X_7^{(t)}$	$X_1^{(t-1)}, X_1^{(t)}$	$G_4^{(X)}$	$Y_2^{(t)}, Y_4^{(t)}$	$Y_5^{(t-1)}, Y_7^{(t-1)}$
$G_5^{(X)}$	$X_1^{(t)}$	$X_2^{(t-1)}, X_2^{(t)}$	$G_5^{(X)}$	$Y_5^{(t)}, Y_7^{(t)}$	$Y_6^{(t-1)}, Y_6^{(t)}$
$G_6^{(X)}$	$X_1^{(t-1)}$	$X_6^{(t-1)}, X_7^{(t-1)}$	$G_6^{(X)}$	$Y_1^{(t-1)}, Y_3^{(t-1)}$	$Y_1^{(t)}, Y_3^{(t)}$
n.a.	n.a.	n.a.	$G_7^{(X)}$	$Y_6^{(t-1)}$	$Y_2^{(t-1)}$
n.a.	n.a.	n.a.	$G_8^{(X)}$	$Y_6^{(t)}$	$Y_2^{(t)}$

Table 1: Groupes de variables pour les deux modèles

4 Synthèse et voies de recherche

Dans ce papier, nous avons construit un modèle pour comprendre comment se formaient les liaisons entre prédicteurs et réponses multivariées pour séries temporelles irrégulières. Tout d'abord, afin de tenir compte de non linéarité, de rupture, de volatilité entre celles-ci, nous avons choisi de segmenter ces séries sous forme de segments linéaires permettant d'éliminer de la non stationnarité, en standardisant la série brute grâce à ceux-ci. Puis, nous avons utilisé une approche exploratoire à l'aide de modèles à équations structurelles libres pour établir des liens entre ces variables normalisées. Cette dernière permet de construire des blocs (groupes) homogènes de variables, puis de les résumer sous forme de variables latentes, de rechercher les liens significatifs entre celles-ci et enfin d'appliquer l'approche PLS sur le modèle libre proposé.

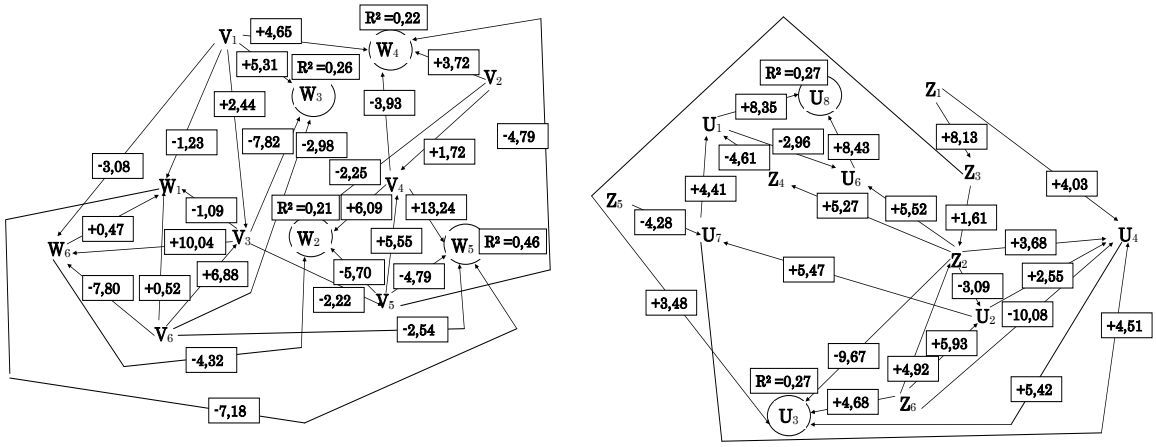

Figure 2: (a) SEM libre sur données différenciées (b) SEM libre sur données normalisées

Dans l'application, la comparaison des résultats des deux modèles (séries différenciées *vs* séries normalisées) a permis de montrer une plus grande cohérence des résultats en termes de stabilité des coefficients, de liens significatifs et d'aspects métiers pour le modèle sur séries normalisées à l'aide de la segmentation. Nos travaux futurs consisteront en la comparaison avec d'autres méthodes, notamment les modèles espaces-états et la prévision de séries temporelles contenues dans les cibles, soit avec l'approche modèle libre, soit avec l'approche modèle fixé par l'expert.

Bibliographie

[1] Arlot, S. & Celisse, A. (2010): Segmentation of the mean of heteroscedastic data via cross-validation, *Statistics and Computing*, pp. 1-20.

[2] Derquenne, Ch. et Hallais, C. (2003): Une méthode alternative à l'approche PLS : comparaison et application aux modèles conceptuels marketing, RSA, LII, 37-72.

[3] Derquenne, C. (2010): An Explanatory Segmentation Method for Time Series, *in Proceedings of Compstat 2010*, Y. Lechevallier & G. Saporta (eds.), 1^{st} Edition, pp. 935-942.

[4] Derquenne, C. (2011): Segmentation of Time Series with Heteroskedastic Components, 58^{th} *World Statistical Congress of ISI*, Dublin, Ireland.

[5] Derquenne, C. (2012): Meta-segmentation of time series for searching a better segmentation, *in Proc. of Compstat 2012*, Limassol, Cyprus, pp. 191-204.

[6] Derquenne, C. (2013): Clustering of time series via a segmentation approach, IFCS 2013, *Program and Book of Abstracts*, Tilburg, The Netherlands p. 134.

[7] Jöreskog, K.G. & Sörbum, D. (1979): Advance in Factor Analysis and Structural Equation Models, Abt Books, Cambridge.

[8] Lavielle, M. and Teyssière, G. (2006): Détection de ruptures multiples dans des séries temporelles multivariées. *Lietuvos Matematikos Rinikinys*, Vol **46**.

[9] Wold, H. (1982): Soft Modelling : The Basic Design and Some Extensions, in Jöreskog K.G. and Sörbum D., Editors, Systems under Indirect Observation, 2, 1-54, N-H, Amsterdam.