Modélisation de séries temporelles multiples et multidimensionnelles

Mireille Gettler-Summa *, Bernard Goldfarb*, Laurent Schwartz**, Jean Marc Steyaert**, Frédérique Lefaudeux***

*CEREMADE Université Paris Dauphine
1 Pl Du Mal De Lattre de Tassigny - 75016 - Paris France
summa@ceremade.dauphine.fr, goldfarb@dauphine.fr
**LIX Ecole Polytechnique Paris 91128 - Palaiseau - France
steyaert@lix.polytechnique.fr, laurent.schwartz@polytechnique.fr
***Isthma,14 rue du Soleillet - 75020 Paris - France
lefaudeux@isthma.fr

Résumé : On présente ici une recherche de modélisation de séries temporelles multiples et multidimensionnelles extraites de données de sites officiels. La difficulté réside d'une part dans la construction des bases de données en raison des différents formats initiaux, des incohérences et des données manquantes, d'autre part dans le grand nombre de variables, endogènes et exogènes, et dans la multiplicité des entrées admissibles pour le problème. Les séries temporelles exogènes sont de plus munies d'une partition a priori. On présente dans cette recherche une approche pour la réduction des variables et des solutions de modélisation de ces données complexes que l'on construit à partir d'adaptation de solutions classiques au contexte temporel multidimensionnel.

Mots clés : séries temporelles multiples, codage, réduction de dimension, modélisation, épidémiologie du cancer, variables latentes et séries temporelles

Abstract: The most relevant elements in this paper are the automatic extraction of temporal data from Official databases and the modelization attempt of some multiple time series by exogenous other multiple time series. The results are applied on to an Epidemiological problem of modeling cancer rates incidence over twenty years, for different countries all over the world. Many issues come up when getting the data: most of the data bases are not available in the same format, some data bases are limited in terms of the number of lines that are allowed for a single query, and after importing the data, one needs to have coherence and continuity over time for each variable. The variables may cover various domains and their definition may have changed over time: expert knowledge is needed to achieve the final attribute coding and validate the retained data. A pre processing phase is then carried on: splines functions for smoothing atypical values and for filling the remaining missing data by interpolation, temporal transformation such as 5th order sum over past years lagged variables in the cancer data base. As an example the epidemiological data consists at that point in a complex set of data: multiple (25 countries in the example), multidimensional (socio economy, nutrition, health care, environment, standardized cancer rates etc.) time series (twenty one years). In order to reduce the data dimension, an exploratory phase builds and discovers the factor blocks that will be introduced in the models. Factors are computed with the Varimax rotation method because most of the variables are highly correlated. Grouping is also performed through clustering approaches for complex time series and the partition is one of the exogenous variable for the modelization phase. A generalized LISREL approach for multidimensional time series is finally performed: as an example, ecology, socio economy, nutrition, health care, style of life and environment are the latent variables of the epidemiological study whereas death cancer rates are the endogenous variables.

Keywords: multiple temporal series, coding, dimension reduction, modeling, epidemiology of cancer, latent variables and series

1. Introduction

Le contenu de ce travail est la présentation, à partir d'une recherche appliquée à des données allant de 1980 à 2000, de la modélisation de séries temporelles multiples et multidimensionnelles afin d'extraire des connaissances sur les relations entre plusieurs grandes bases de données mondiales officielles. Il fait suite à une approche exploratoire sur les variables endogènes, [Gettler-Summa & al. (2007)], développant une classification de séries multidimensionnelles multiples dans le cadre de l'Analyse de Données fonctionnelles [Ramsay & Silverman (1997)]

La difficulté de cette recherche consiste d'une part dans la constitution des tableaux complexes soumis à l'analyse, à partir des formats des bases de données officielles disponibles, d'autre part dans la nécessité de réduire le nombre de variables exogènes, et enfin dans la modélisation de séries temporelles multiples (endogènes) par d'autres séries temporelles multiples (exogènes), les variables exogènes étant munies d'un partition a priori que l'on prend en compte pour le traitement.

2. Constitution de la base de données

2.1. Extraction des données

Les données proviennent de requêtes effectuées à partir de plus de vingt sites officiels mondiaux dont : Laborstat, World Bank, World Bank Educational data base, World Bank Health and Nutrition, Fao Terrastat, IEA, Fao Acquastat, Wold Health Organisation, FDA, IARC, US-DAS, OECD, [WHO (2006)]. Les résultats des requêtes, ponctuelles dans le temps, sont ensuite organisés en plusieurs tableaux à trois entrées, pays en lignes, variables en colonnes, et variation temporelle dans une cellule selon le format Delta Suite®. Les différents tableaux naissent de distinctions imposées par les experts, ici un tableau pour chaque sexe et pour chacun des treize localisations de cancer étudiées en variables endogènes de la modélisation.

Voici par exemple sur la figure 1 ci-dessous un extrait du tableau d'une variable endogène, taux de cancer du poumon : en lignes les pays, en colonnes les dix classes d'âge de décès de cinq ans en cinq ans à partir de 40 ans, dans une cellule la série temporelle des ASR, taux standardisés classiques de l'épidémiologie, pour les décès par cancer du poumon de 1980 à 2000.

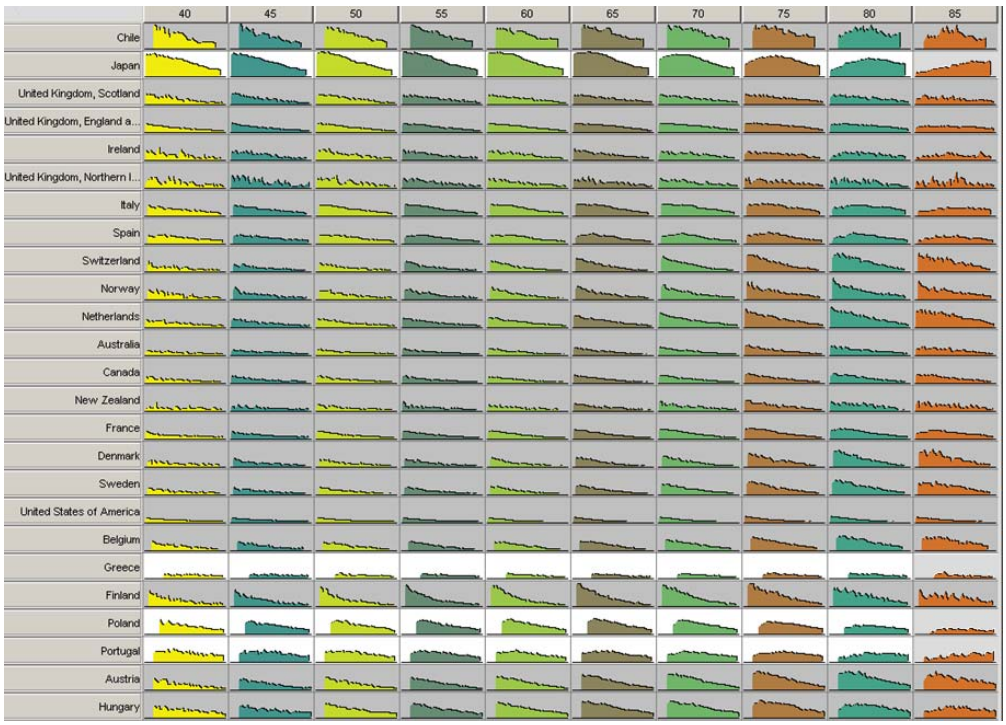

Figure 1 : Table des taux standardisés de décès par cancer du poumon de 1980 à 2000 pour vingt cinq pays et pour dix classes d'âge.

2.2. Codage des données

La représentation finale des données passe par le traitement des données aberrantes, des données manquantes, et par des transformations propres aux séries temporelles.

Pour détecter les données hors normes, il a été choisi de lisser en continu les données initialement discrètes par des splines cubiques a cinq nœuds, dérivables jusqu'à l'ordre 3 stade où on accepte une discontinuité éventuelle. On calcule alors l'écart entre la valeur observée et la valeur sur le spline : si cet écart est supérieur au double de l'écart type, on remplace par la valeur obtenue avec le spline.

Pour les données manquantes, plusieurs situations se présentent classiquement : transversalement pour une variable et une seule date mais pour toutes les observations, ou longitudinalement pour une seule et même observation, pour tout un sous ensemble d'observations, ou encore ponctuellement sur une date ou bien sur tout un intervalle temporel. Ici on a choisi d'éliminer les pays qui présentaient une des variables entièrement manquantes, comme par exemple l'Argentine qui n'avait pas de données économiques, de santé ou de nutrition sur toute une période. Quand aux cas restants qui étaient environ 10% des cellules, on vérifie si elles sont ou non MCAR (missing completly at random) ; si oui on vérifie ensuite si elles sont MAR (non liées aux variables endogènes) ou bien MNAR (liées à la variable endogène). Dans notre recherche, nous n'avons que des MCAR que l'on comble par interpolation spline.

Certaines séries temporelles exogènes sont connues par les experts comme ayant un lien décalé dans le temps avec les variables endogènes à modéliser. On travaille alors en recodant par des moyennes mobiles dont on ajuste la fenêtre sur la connaissance a priori du domaine. Ici on a ainsi recodé avec un décalage de 10 ans la 'consommation annuelle de cigarette pro capita des plus de 15 ans' ou encore 'l'émission annuelle de CO2 en tonnes pro capita'.

2.3. Format des données finales

Le problème de modélisation se pose finalement sur des tableaux de séries multiples multidimensionnelles sans données manquantes où les séries exogènes sont munies d'une partition a priori, et

les variables endogènes sont multiples. Ils ont le format de la figure 2 ci-dessous qui illustre notre sujet de recherche.

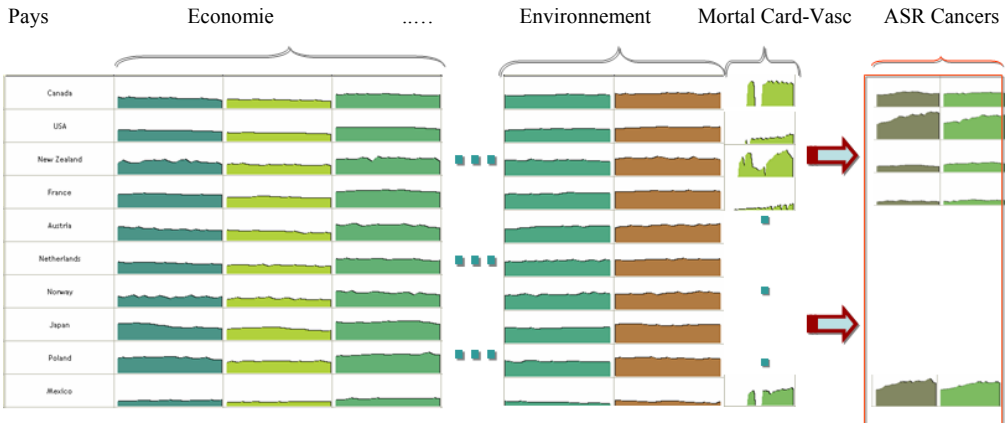

Figure 2 :
Tableau de séries multiples multidimensionnelles de variables endogènes modélisant les taux standardisés de décès par cancer du poumon de dix classes d'âge de 1980 à 2000, pour les pays.

Vingt cinq pays sont retenus, sur vingt et une années, pour plus de quatre vingt variables 'séries chronologiques' exogènes partitionnées sémantiquement en quatre groupes : santé, environnement, économie et nutrition, plus une variable particulière d'évolution de la mortalité par maladie cardio-vasculaires dans le pays, maladie en 'concurrence' (au sens des risques compétitifs), dans le temps et pour la mortalité, avec les cancers. Une autre variable exogène particulière dans notre problème, nominale, provient d'un partitionnement automatique réalisé sur les séries temporelles multiples multidimensionnelles d'évolution de chaque site de cancer, simultanément pour dix classes d'âge et pour tous les pays dans une étude exploratoire préliminaire sur les vingt et une années. Les résultats de la classification de ces données complexes sont présentés dans la figure 3, les parangons de chaque classe sont soulignés. On a illustré sur deux classes d'âge, une classe jeune, 45-49 ans et une classe âgée, 70-74 ans, les évolutions des parangons, pour évoquer ce qui caractérise chaque classe.

3. Stratégie de réduction du nombre des variables

Quelle que soit la méthode de modélisation envisagée pour notre recherche, les variables exogènes étant plus de quatre vingt et les variables endogènes, pour un même site de cancer, étant au nombre de dix puisqu'il y a dix classes d'âge, une stratégie de réduction du nombre de variables s'impose.
On a choisi une méthode d'Analyse en Composantes Principales avec Rotation compte tenu des corrélations entre certaines variables, et ceci à l'intérieur de chaque groupe sémantique des séries chronologiques exogènes.
Pour les séries chronologiques endogènes, on fait un choix expert de quelques classes d'âge qui formeront isolément chacune une seule variable 'complexe' au sens de la fouille de Données complexes: l'évolution pour cette classe d'âge du taux de mortalité, et ce sera la variable exogène du modèle. Cette variable n'en reste pas moins une variable complexe puisqu'il s'agit d'une série temporelle sur vingt années.

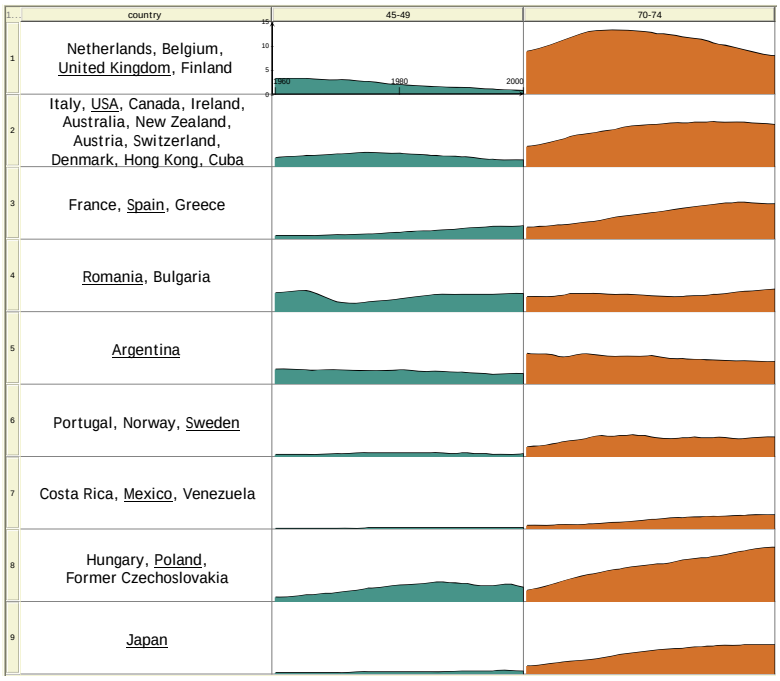

Figure 3 : Tableau des 9 classes issues d'une partition automatique sur les séries multiples multi-dimensionnelles de variables endogènes modélisant les taux standardisés de décès par cancer du poumon de dix classes d'âge de 1980 à 2000, pour les pays.

Voici quelques résultats de la réduction de chaque groupe de variable.
Pour l'emploi, la table 1 de l'ACP avec rotation suggère que l'on garde quatre facteurs, ceux de valeurs propres supérieures à 1 ; le premier facteur par exemple est construit par l'emploi en milieu rural et par le secteur banque et assurance (à l'opposé), le deuxième par construction et grossiste contre détaillant et dans une moindre mesure banque et assurance.
Pour l'économie, la table 2 de l'ACP avec rotation suggère que l'on garde deux facteurs, toujours de valeurs propres supérieures à 1 ; le premier facteur est construit par le budget santé pro capita, le PIB pro capita et le ratio économie souterraine / PIB, le deuxième par l'index de Gini et le budget santé pro capita.
Pour l'environnement, la table 3 de l'ACP avec rotation suggère que l'on garde deux facteurs, toujours de valeurs propres supérieures à 1 ; le premier facteur est construit par l'émission de monoxyde de carbone et d'azote, le deuxième par les émissions de soufre.

Pour la nutrition, on garde quatre facteurs, toujours de valeurs propres supérieures à 1; le premier facteur est construit par la quantité de glucides, de consommation de céréales et de lipides (à l'opposé), le deuxième par les consommations d'alcool. On montre sur la table 4 les coordonnées sur les deux premiers axes des principales variables qui y ont contribué.

	Valeur propre	Proportion	Cumulée
1	3.91140720	0.3556	0.3556
2	2.98287599	0.2712	0.6268
3	1.34071173	0.1219	0.7486
4	1.07662214	0.0979	0.8465
5	0.51161972	0.0465	0.8930
6	0.40646327	0.0370	0.9300
7	0.28269142	0.0257	0.9557
8	0.18783644	0.0171	0.9727
9	0.14208940	0.0129	0.9857
10	0.09398088	0.0085	0.9942
11	0.06370182	0.0058	1.0000

Table 1 : Valeurs propres, pourcentage d'inertie et inerties cumulées de l'ACP avec rotation sur les séries temporelles d'emploi

	Valeur propre	Proportion	Cumulée
1	2.87621917	0.4109	0.4109
2	1.69262645	0.2418	0.6527
3	0.90162994	0.1288	0.7815
4	0.65890158	0.0941	0.8756
5	0.47408316	0.0677	0.9434
6	0.21359594	0.0305	0.9739
7	0.18294376	0.0261	1.0000

Table 2 : Valeurs propres, pourcentage d'inertie et inerties cumulées de l'ACP avec rotation sur les séries temporelles d'économie

	Valeur propre	Proportion	Cumulée
1	3.09284597	0.6186	0.6186
2	1.04226891	0.2085	0.8270
3	0.53963318	0.1079	0.9349
4	0.20254216	0.0405	0.9755
5	0.12270977	0.0245	1.0000

Table 3 : Valeurs propres, pourcentage d'inertie et inerties cumulées de l'ACP avec rotation sur les séries temporelles d'environnement

	Factor1	Factor2
Alcohol consumption (l/cap)	-0.17852	0.87466
Daily Calorie supply /capita	0.07361	0.29176
Daily Alcohol supply (cal/capita)	-0.16185	0.89145
Daily Carbohydrates supply (g/capita)	0.88832	-0.19998
Daily Total Fat supply (g/capita)	-0.66274	0.15958
Yearly Cereals consumption (kg/capita)	0.72561	-0.29888
Yearly Fruits supply (kg/capita)	-0.55164	-0.21518
Yearly Vegetables supply (kg/capita)	0.12406	0.15083
Yearly Fish supply (kg/capita) run	-0.13873	-0.06954

Table 4 : Coordonnées sur les deux premiers facteurs de rotation de l'ACP sur les séries temporelles de nutrition, pour les séries de contributions majeures

4. Modélisation des séries temporelles endogènes

Le problème de la modélisation est difficile même après avoir réduit à une seule variable 'série chronologique' la variable endogène. La particularité longitudinale des données nécessite une généralisation des méthodes usuelles que l'on applique en analyse multidimensionnelle, [Serban et al. (2006)].

La modélisation à l'aide de variables latentes par l'analyse des covariances initialement conçue pour un dispositif d'observation à une date donnée relie des variables endogènes observées à des variables exogènes observées par un système de 3 équations

une représentation des variables exogènes latentes η par les variables exogènes observées y en tenant compte d'erreurs ε de représentations soit $y = \Lambda_y \eta + \varepsilon$,

une représentation des variables endogènes latentes ξ par les variables exogènes observées x en tenant compte d'erreurs ε de représentations soit $x = \Lambda_x \xi + \delta$,

un modèle interne associant les variables latentes exogènes aux variables latentes endogènes, en tenant compte d'erreurs de modélisation $\eta = B\eta + \Gamma\xi + \zeta$

Une première généralisation est possible à des observations répétées sur quelques dates, soit T1, T2, …, Tk, lorsque le nombre de dates d'observations est limité (quelques unités) ; elle met en jeu une démultiplication des variables observées et des variables latentes et impose d'ajouter une variable latente par date schématisant l'effet temps. Il est impossible de dépasser 5 ou 6 dates d'observation, compte tenu de la complexité du modèle correspondant [Loehlin (2004)].

Une seconde étape de généralisation, peut être réalisée en regroupant des dates par période. Par exemple des observations sur les années 1981 à 2000 (20 années) peuvent être regroupées en 4

périodes [égales ou non selon les connaissances préalables] P1, P2, P3 et P4 qui seront cette fois considérées comme ouvrant un effet inter-période. On peut également détecter des groupes de périodes par classification automatique, en transposant la matrice initiale, et en classifiant les dates qui sont en lignes du nouveau tableau.
Dans ce contexte, on perdra la séquentialité temporelle de l'effet intra-période, au profit de la séquentialité temporelle de l'effet inter-période ; autrement dit l'effet inter-période est bien pris dans sa dimension temporelle intrinsèque, alors que l'effet intra-période sera traité comme un effet latent ordinaire.
Une généralisation plus large [Bollen & Curran (2006)] est obtenue par le modèle des courbes latentes, mais ce cadre est limité en pratique par le nombre de dates d'observations. Son intérêt réside essentiellement dans la prise en compte dans le modèle des variations temporelles des individus par une fonction « latente » sous-jacente associée à une erreur de modélisation temporelle.
La question étudiée ici fait évidemment évoquer une généralisation étendue de ce modèle, dans laquelle les variables observées (tant exogènes qu'endogènes) sont traitées comme des fonctions en représentation vectorielle, et non plus comme des valeurs. Les variables latentes correspondantes (exogènes et endogènes) sont alors aussi traitées comme des fonctions.

Plusieurs approches ont été testées sur le tableau complexe de données épidémiologiques comme PLS ou LISREL.
La figure 4 montre un graphe de relations, ici pour l'estimation interne, construit pour une approche structurelle dans la modélisation des évolutions de cancer du poumon pour une classe d'âge.

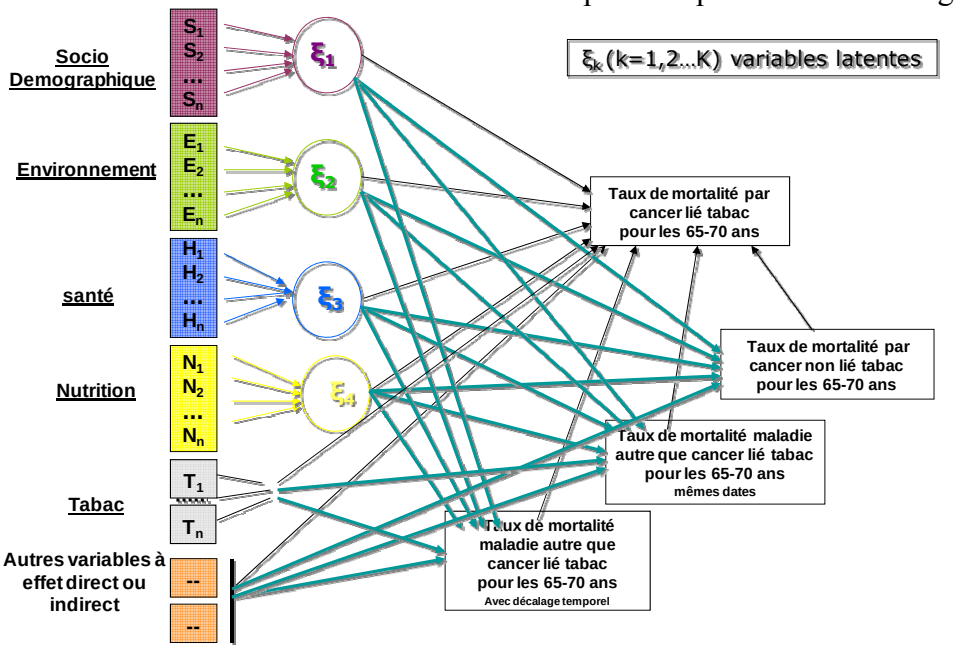

Figure 4 : Graphe des relations structurelles pour l'estimation interne d'une approche LISREL pour modéliser les taux de mortalité du cancer du poumon

La table 5 montre les significativités obtenues pour quelques facteurs pour deux variables endogènes différentes, les ASR du cancer du poumon et les ASR tous cancers. Les variables dans l'ordre descendant sont : consommation annuelle de cigarettes pro capita, consommation annuelle de tabac en kilos pro capita, quatre classes issues de la classification automatique, les facteurs indicés de la réduction d'environnement, économie, emploi et nutrition. On voit par exemple que c'est la consommation globale de tabac qui compte et non pas la consommation de cigarettes pour reconstruire les deux ASR de cancers, on voit aussi que le facteur 3 environnemental est significatif pour tous cancers mais pas pour cancer du tabac.

Variables endogènes		Poumons	Tous cancers
CIG_DC_P15_99_ma5_l10	Cigs consumption (nb/capita)	0.0663	0.2372
TOB_DC_P15_99_ma5_l10	Tobacco consumption (kg/capita)	<.0001	<.0001
cluster1		0.1180	<.0001
cluster2		<.0001	<.0001
cluster3		0.0003	0.0329
cluster6		0.9806	0.6413
cluster8		<.0001	0.0004
env1		<.0001	<.0001
env2		0.0232	0.1817
env3		0.2451	0.0001
eco1		<.0001	<.0001
eco2		<.0001	<.0001
work1		<.0001	<.0001
work2		<.0001	<.0001
work3		0.6784	0.0187
work4		0.0003	0.7697
diet1		0.0028	0.0010
diet2		<.0001	0.0004
diet3		<.0001	<.0001
diet4		<.0001	0.0774

Table 5 :

Tableau de significativité des cofacteurs pour les ASR du cancer du poumon et de tous cancers.

5. Conclusion

Nous avons entrepris une approche théorique de la modélisation de séries temporelles multiples et multidimensionnelles par des variables 'séries temporelles' munies d'une partition a priori sur les variables exogènes. Cette recherche n'est cependant pas finie et nous présentons dans ce travail en fait un traitement où le temps est 'mis à plat'. Une généralisation d'une approche de type LISREL aux séries chronologiques multidimensionnelles est en cours et ses résultats devront être comparés aux résultats présentés ici.

Par ailleurs on fait ici l'hypothèse de séries temporelles sans césures fréquentes et définies aux mêmes dates. Dans les cas contraires, il faut envisager tous les pré-traitements classiques nécessaires à l'étude des séries chronologiques.

Dans l'application épidémiologique du cancer, les données de 2000 à 2004 sont maintenant disponibles et une méta-analyse est en cours pour valider les modèles obtenus.

Références

Bollen K.A., Curran P.J. (2006) *Latent curve models, a structural equation perspective*; Wiley Interscience, J. Wiley & sons.

Forey, B., Hamling, J., P. Lee P., Wald, N. (2002). *International smoking statistics*. London: Oxford University Press.

Gettler Summa, M., Steyaert JM, Vautrain F., Weitkunat R. (2007). *A new clusteringmethod for times series for discover geographical cancer trends from 1960 to 2000* Annals of epidemiology Vol.17, N°.9

Loehlin J.C. (2004) *Latent variable models*, 4th ed ; Lawrence Erlbaum associates editors, Mahwah NJ.

Ramsay, J.O. , Silverman B.W. (1997). *Functional Data Analysis*, New York: Springer.

Serban, M., Brockwell A., Lehoczky J., Srivastava S. (2006). *Modelling the dynamic dependence structure in multivariate financial time series*. Journal of time series analysis, vol28, N°5, 763-782

WHO (2006). Geneva: WHO. *Cancer*. Fact sheet No. 297