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