MULTIPLE TIME SERIES:
NEW APPROCHES AND NEW TOOLS IN DATA MINING
APPLICATIONS TO CANCER EPIDEMIOLOGY

Mireille Gettler Summa*, Frédérick Vautrain****
Laurent Schwartz**, Mathieu Barrault****
Jean Marc Steyaert***, Nicolas Hafner****

* Centre de recherche en Mathématiques de la Décision
Université Paris Dauphine 1 Pl. du Ml de Lattre de Tassigny 75016 Paris France
summa@ceremade.dauphine.fr
** Service de Radiothérapie, Hôpital Pitié Salpêtrière 47 boulevard de l'Hôpital Paris
*** LIX – Ecole Polytechnique
steyaert@polytechnique.fr
**** Isthma, 14-16 rue Soleillet 75020 Paris France
vautrain@isthma.fr, barrault@isthma.fr, hafner@isthma.fr

Résumé Des résultats innovants en fouille de données complexes fournissent des approches originales pour les épidémiologistes qui bénéficient de traitements interactifs pour aborder leurs données conjointement sous toutes leurs entrées et en tirer des résultats. L'étude présente des algorithmes qui travaillent sur des espaces multidimensionnels de fonctions (ici des chroniques ou bien encore des distributions discrètes ou discrétisées à support fini) avec moins de perte d'information que dans les codages habituels par agrégation, quantiles ou autres ; ils ont été implémentés dans le logiciel DELTA Suite : chaque cellule d'une table étudiée contient une donnée complexe (par exemple une série temporelle). Delta Suite est utilisé ici dans deux études épidémiologiques de l'évolution des cancers dans le temps et dans l'espace: en un premier temps pour la visualisation simultanée et l'exploration des chroniques de taux de mortalité par cancer sur cinq entrées conjointes (géographiques, temporelles, âge, sexe et pathologies) puis dans un deuxième temps pour la comparaison géographique des courbes d'évolution des cancers du poumon pour 51 pays et 21 années par généralisation des approches de classification automatique.

Mots clés: logiciel pour la fouille de données, séries temporelles multidimensionnelles, base de données en épidémiologie du cancer, classification pyramidale complexe

Abstract. Innovating data mining tool for complex data provide new and comprehensive viewpoints to the epidemiologist who can derive original results and perform interactive treatments. New algorithms working in a multidimensional space on curves (such as a set of multiple time series in our study) or on discrete distributions, with les loss of information as it is the case with more classical encoding techniques (e.g. by data aggregation: means, quintiles etc.) have been studied and have been implemented in Delta Suite software. Each cell of the table under study is a function (in this case time series). Delta Suite software is applied for comparing epidemiological trends w.r.t. time, on two illustrative studies of the WHO data:
- the simultaneous visualization and exploration of the time series for cancer death rates on many entries, geographical information, temporal data, age, sex and pathologies.
- the geographic comparison of lung cancer evolutions over 21 years by building automatic classifications.

Key words: Data Mining software, multiple time series, cancer epidemiology data base, complex pyramidal clustering