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

1 INTRODUCTION

1.1 A new exploratory approach for time series

The data of the death certificates that the World Health Organization (WHO) provides officially on its servers contain five main entries: death main cause, sex, age class, certificate death country (state or province in some cases), death year. These five variables define therefore a multidimensional space in which the statistical investigations are performed; a great number of multidimensional arrays (usually 2 entries, but quite often 3 and even 4) (Huang 1999) can be built in order to perform Multiple Array Analysis (Pardoux 2001).

Such an approach is seldom used in epidemiology where the main tools are based on the juxtaposition of mono valued statistics (Groupe d'étude et de réflexion interrégional 1996). Innovating data mining tools for complex data (Gettler- Summa Pardoux 2000) provide new and comprehensive viewpoints to the epidemiologist who can derive original results and perform interactive treatments. They allow working in a multidimensional space on curves (such as a set of multiple time series in our study) instead of sets of single curves (Last and al. 2004) or on discrete distributions (Bock Diday 2000), with less loss of information as it is the case with more classical encoding techniques (e.g. by data aggregation: means, quintiles etc.). Therefore the data can be kept and integrally treated. These original methods have been applied using Delta Suite software which was developed by the Research Centre for Decision Mathematics of Paris Dauphine and by Isthma Company (France) 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 evolution over 30 years by applying generalized automatic clustering algorithms.

1.2 Questions in Cancer epidemiology methodology

It appears that the overall cancer death rate has been about constant in the past 30 years, both in Europe and in the United States, with the notable exception of mortality among children and young adults. At the same time, there has been a global rise in mortality by lung cancer and slightly less important rises of melanoma and brain tumor, whereas one observes a concomitant decrease for cancers of the stomach, esophagus and head-and-neck (Gettler Summa and al. 2001).

Smoking explains certainly the increased lung cancer death rate, but in the meanwhile there is a partial decrease in other malignancies such as head-and-neck tumors which are also mainly caused by tobacco. The reasons for these multiple and apparently contradictory shifts remain largely unknown. The goal of this paper is to address these cancers in cancer epidemiology using statistical and mathematical techniques rarely used in the medical community.

2 Data description

2.1 The initial array

The data analyzed in the present work come from a data base of deaths caused by cancer which has been extracted from files obtained from the World Health Organization (WHO1981-2000). We extracted the number of deaths caused by cancer for 122 countries, both sexes, 21 age-classes (ranging from birth to 100+), 13 cancer types and 50 years of data. The original data were organized in 1,521,828 lines, 63% of which had missing data and 14,074 of which were containing out of the range figures. Odd data were processed by the 'missing value' menu in Delta software which is described further. 574,200 lines array have been eventually validated.

2.2 Initial data pre-processing

Since we want to compare country evolutions, we have to normalize the data over the age-classes; for this purpose, we compute new homogenous ASRs (Average Standardized Ratios) that are commonly used in epidemiology (National Center for Health Statistics 1998)

Since we want to compare temporal evolutions between countries, we have to synchronize the various data in order to have the same periods of time for all the countries.

We have to reconstruct a common denomination for the various types of cancer since the nomenclature, ICD, varies along time (at least 4 different ones in the past 50 years).

The definition of the age-classes, initially given in the database to be of 5 year duration, has to be modified according to the cancer type in order to be medically significant.

91 countries have been retained globally at the beginning of our study since their data can be exploited on common periods of time. Furthermore, we have proved by a Khi² adjustment test, that, for some countries, the variations of the collected data along any period of time follow a Gaussian law: what is effectively measured (in the case of Iceland, e.g.) is more the average value of mortality caused by lung cancer than real fluctuations. We have decided not to take these countries into account.

2.2.1 Expert pathology classifications

Death cause denominations on death certificates have varied during the period of data collecting. These evolutions can be found in the 'International Classification of Diseases’ (ICD): for instance the hierarchical presentation of some lung diseases is no longer the same since the appraisal of new diseases such as AIDS which implies that we have to put them in a different pre-existing group. We have taken into account 5 classifications: the 6th ICD from 1949 to 1957, the 7th ICD from 1958 to 1967, the 8th ICD from 1968 to 1978, the 9th ICD from 1979 to 1998, the 10th ICD from 1999 (Anderson and al. 2001).

Comparison indices Ci, for the i-th disease, have been computed for the transitions from one classification to the next one; they are applied as weights on data in order to obtain a global final array of normalized and homogeneous data: $Ci = D_i^{P1}/ D_i^{P2}$ where D_I^P is the number of deaths caused by disease i in the classification period P.

2.2.2 Standardized rates

The populations do not present the same age distributions according to the country. But statistics show that age has a great influence on mortality by cancer. It is thus necessary in our study to define and use a notion of age-normalized ratio and to reconstruct for each population a set of new figures corrected in such a way that they propose for each country a unique normalized age structure. The standard time scales deal with 5 years age-classes and they can vary according to the projection period and to the type of the reference population (worldwide, European, other). New distributions are presently being worked out up to year 2020. These ratios are modified by so-called direct or indirect adjustments in epidemiology.

3 DELTA: New Approaches in Data Mining

The DELTA platform, allows incorporating variation on a single data: probability distribution, functional variation etc.: multiple time series are characteristic examples of complex data, where variation consists of time evolution. Delta allows managing, editing and analyzing such complex data in a multidimensional statistical framework.

3.1 DELTA METRIC: data manager and editor in the context of epidemiology

DELTA allows a number of operations on a classical data array in order to build new data arrays, significantly more complex since it is possible to define functions as entries of these new arrays.

It is composed of a graphic editor on complex data which allows a visual exploratory analysis, which is completed by statistical numerical indices (regression coefficients etc.) and data management options (sorting, filtering etc.)

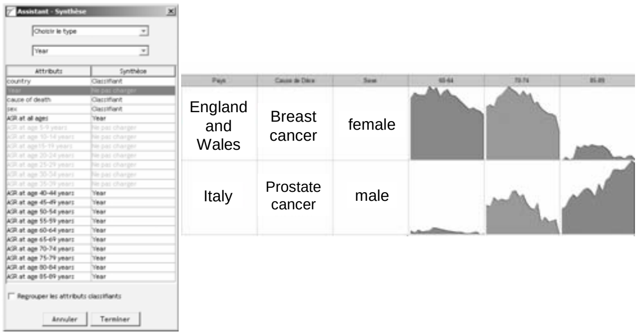
3.1.1 The SYNTHESIS operator: taking into account the variations of basic data and building the complex data base

The initial data analyzed in this study concern 51 countries and 21 years of reporting over 9 pathologies and 16 age classes for both sexes. Pathologies are: malignant tumors of breast, prostate, stomach, esophagus, head and neck (lips, mouth, larynx, pharynx …), lung (trachea, bronchus …), bladder, pancreas, brain..

country	Year	cause of death	sex	ASR at all ages		ASR at age 45 49 years		ASR at age 85 89 years
England and Wales	1980	Malignant neoplasm of stomach	female	17.91536	…	0.24453	…	0.98982
England and Wales	1980	Malignant neoplasm of stomach	male	26.86425		0.56563		1.33101
…								
Italy	1999	Malignant neoplasm of trachea, bronchus and lung	female	19.02513	…	0.43480	…	0.42265
Italy	1999	Malignant neoplasm of trachea, bronchus and lung	male	92.82309		1.57175		2.36752

TAB. 1 – *Initial data*

We have built time series by synthesizing the observed data according to time (Year); we thus obtained new statistical units where the variables are temporal curves (functions of the time variable), such as shown below.



Assistant - Synthèse

Choisir le type

Year

Attributs	Synthèse
country	Classifiant
Year	Ne pas charger
cause of death	Classifiant
sex	Classifiant
ASR at all ages	Year
ASR at age 5-9 years	Ne pas charger
ASR at age 10-14 years	Ne pas charger
ASR at age 15-19 years	Ne pas charger
ASR at age 20-24 years	Ne pas charger
ASR at age 25-29 years	Ne pas charger
ASR at age 30-34 years	Ne pas charger
ASR at age 35-39 years	Ne pas charger
ASR at age 40-44 years	Year
ASR at age 45-49 years	Year
ASR at age 50-54 years	Year
ASR at age 55-59 years	Year
ASR at age 60-64 years	Year
ASR at age 65-69 years	Year
ASR at age 70-74 years	Year
ASR at age 75-79 years	Year
ASR at age 80-84 years	Year
ASR at age 85-89 years	Year

□ Regrouper les attributs classifiants

Pays	Cause de Décès	Sexe	60-64	70-74	85-89
England and Wales	Breast cancer	female			
Italy	Prostate cancer	male			

FIG. 1 - *Automatic generation of time series for countries, diseases and males, females*

One can equally synthesize raw data according to the variable country, thus obtaining new variables in the shape of discrete distributions defined on the domain of countries or on the variable sex, yielding binary distributions; the age classes could also be used as the synthesis variable.

The synthesis process produces an array of complex data in which statistical units are described in terms of generalized variables: continuous functions, intervals, distributions, lists, discrete mapping. Figure 2 presents a sample of an array where several variables are expressed and synthesized: ‘’number of deaths during the year’’ function variable (gives the curve for the evolution of the number of deaths year per year between 1980 and 2000), ‘’ Raw Death Ratio’’ distribution (number of deaths in an age class divided by the population of this class -a percentage), ‘’ Raw Death Ratio by sex’’ distribution.

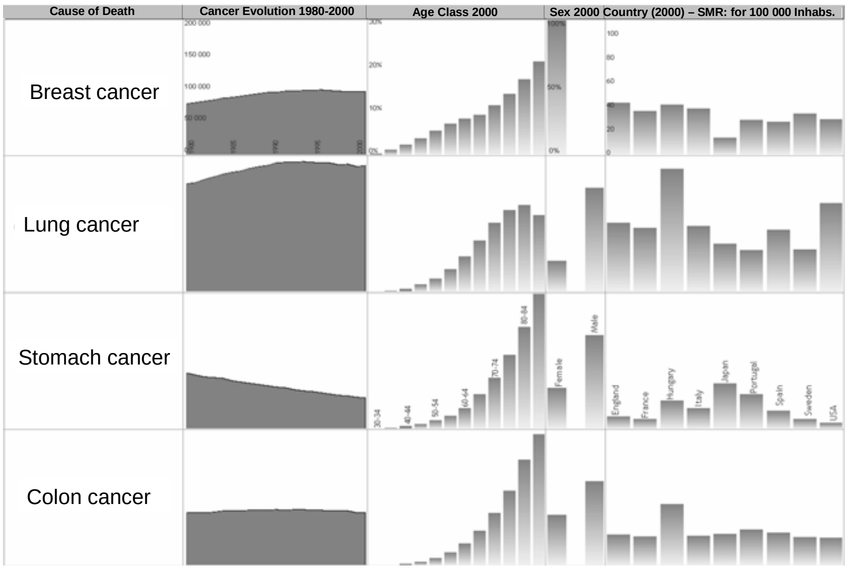


FIG. 2 - *Joint edition of variables from various types (discrete applications, curves)*

3.1.2 Options for the re-encoding of complex data

The encoding of data is a key step in any statistical analysis: the methods, hence the results ultimately depend on the values that have been retained to represent the initial information.

In the case of a time series, it is usual, during the exploratory phase, to smoothen, deseasonalize or decompose the series. DELTA offers these operators on any function obtained during the SYNTHESIS process. Another re-encoding allows making more precise a visual feeling on the variations of a curve, by means of elementary numerical operators. A polynomial approximation of degree 1 was performed (higher degrees can be used) the coefficients of which (steepness ‘a’ and origin step ‘b’) characterize each statistical unit and allow an easy interpretation of the cancer growth rate. Other re-encoding options have been implemented according to the type of the complex data: polynomial approximation, Fourier transformation, regression, BSplines and other smoothing options etc.

3.1.3 Filtering options

With DELTA, the filtering options are queries that are specified from a dedicated predicate editor. They are necessary for any interactive exploratory search.

For example the result of a double filter onto the curves of the initial data, in order to select all the countries and diseases such that their death curves for the age class 50-54 are above the similar death curves for the age class 80-84, reveals that this situation occurs for 13 countries over 18, mainly for head and neck, esophagus and lung cancers.

3.1.4 The multiple time series editor

DELTA contains a complex data worksheet editor for a visual screening of the data.

Graphic scales Four modes are possible to render a cell:

- "Local" mode, each cell has its own scale (and cannot be visually compared to another).
- "Global" mode, the cells of the same column have a common scale and can be compared to the others of the same column. But it is not possible to compare the columns.
- " Common" mode, all the cells have the same scale and can be compared to the others.
- "Manual" mode, the user may emphasize some visual effects. The scales are defined by the user at convenience on both axes.

Missing data management Missing data are dealt with by DELTA and are denoted ‘’n/a’’.

In some circumstances it is possible to re-encode them by a number of smoothing and interpolating functions, especially on discrete variables. 36 countries have thus been interpolated and/or extrapolated, depending on the reason for which data were missing (no census data for a year, delayed ICD, etc.).

The Cohort approach The data manager allows to build, among other arrays, cohorts and to visualize them (Breslow 1987). Let us show an example on the WHO data.

Let us imagine tracking persons born between 1916 and 1920. Those who died in 2000 pertain to the death statistics of the 80-84 age class, whereas a death in 1980 would put them in the age class 60-64. Figure 3 exhibits 8 cohorts specified in rows: people born during 4 periods (1920-1924, 1925-1929, 1930-1934 and 1935-1939), two countries (England and Japan), 4 diseases (head and neck, prostate, stomach, lung); in each cell, a discrete mapping: raw death ratios per disease. So doing, one can read, at the intersection of row England-1930-1934 and of column age-class-60-64, the quotient of the number of deaths of this age class during the years 1990-1994 and the population of this age class at this period.

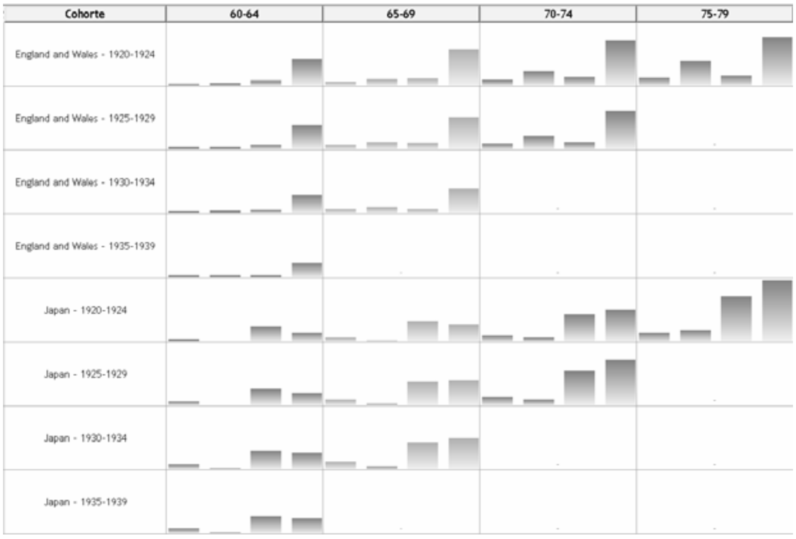


FIG. 3 - *Visualization of age class cohorts*

3.1.5 Statistical results and their interpretation of a study Countries/Cancers/Years

We now present the major conclusions of a case study that will illustrate DELTA functionalities for elementary statistical mining. This study works out statistics on 51 industrialized countries, with a western style of life, based on 21 years of data concerning 10 types of cancer and 16 age classes (each class being of 5 years duration).

The conclusions are issued from the analysis of graphs similar to figure 1. They first come from a visual analysis of the graphics and then are formally validated by the computation of various numerical indicators obtained from the curves, using DELTA tools.

From the resulting figures, one can assert the following principal facts:

- the epidemiology of ''Cancer'' is NOT the epidemiology of each ''type of cancer''; for instance ''all cancers'' for a given country shows some trends that are contradictory according to the pathology (lung cancer is increasing while stomach cancer decreases); for a given pathology, very different trends can be observed depending on the country (for the lung cancer one can see an important increase in Hungary contrasting with an equivalent decrease in England).
- The proportion of oldest people is the highest for prostate cancer and smallest for lung cancer; it increases globally with years, especially for lung cancer.
- Lung cancer remains the most important in the number of deaths; stomach cancer has diminished in a significant manner.
- All tobacco linked cancers do not follow the same evolutions

3.2 DELTA METRICS multidimensional time series analyzer in the context of epidemiology

In order to estimate in a same analysis the influence of the various dimensions (age, country, sex, pathology) on the trajectories, a multidimensional approach for curve analysis is necessary. DELTA METRICS was used to perform factorial analyses and unsupervised classifications. We present in this study the clustering phase results.

3.2.1 Automatic clustering for the geographic analysis of lung cancer from 1969 to 1998

DELTA contains several tools for performing unsupervised classifications: one can use a 'k-means' approach or build hierarchical classifications for time series. These algorithms have been designed on recent results and generalize previously known methods for quantitative or qualitative data to the new universe of functions (Ramsey 1997).

Methodology. One of the main methodological elements which distinguish the classical classification methods of mono-valued data from the classification of functional data (probability distributions, time series, etc.) is the definition of the metrics over the data space.

This metrics can be defined in two steps. In the first step one computes the dissimilarities between the curves representing two distinct statistical units on a single variable. In the second step, one computes the dissimilarities aggregates over all the variables.

Dissimilarities computation. The classical dissimilarities have been implemented in DELTA, but they have been adapted to the structure of the data dealt with on the platform.

In the case of discrete applications, three distances between the graphs have been implemented: the L_1 distance, the L_2 distance and the L_{max} distance.

In the case of continuous functions, defined on subsets of RxR, values intermediate between the sampling points which are obtained by a linear interpolation have been used.

When the envelop of the sample abscissa is not the same for the two mappings, we use by default the intersection of the supports. (Other options are possible, such as giving the value 0 when the mapping is not defined).

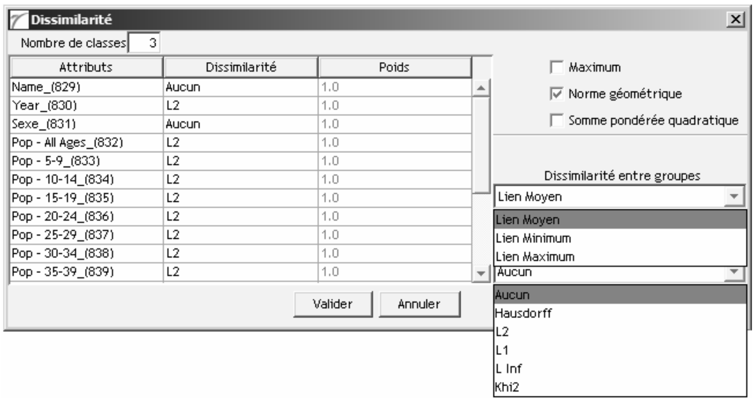
$$
\begin{aligned}
&Diss(D_1, D_2) = \\
&\frac{1}{Xs - Xe}\int_{Xs}^{Xe}\left(Y^1(t) - Y^2(t)\right)^2 dt \\
&= \frac{1}{Xs - Xe}\sum_i \int_{X_i}^{X_{i+1}} \left(\left(Y^2_{i+1} - Y^2_i\right)\frac{t - X^2_i}{X^2_{i+1} - X^2_i} + Y^2_i - \left(Y^1_{i+1} - Y^1_i\right)\frac{t - X^1_i}{X^1_{i+1} - X^1_i} - Y^1_i \right)^2 dt \\
&Xs = Max\left(Min\left(X_1\right), Min\left(X_2\right)\right) \\
&Xe = Min\left(Max\left(X_1\right), Max\left(X_2\right)\right)
\end{aligned}
$$

For interval data (the value is a real interval) the Hausdorff distance was implemented

In the case of finite discrete distributions for the same variable, to build a hierarchical classification, the formula used for each variable is inspired from the chi2 distance.

Hierarchical classification of complex data with DELTA. DELTA provides tools for building ascending hierarchical classification of complex data. The algorithm presents the remarkable feature that the final result of the aggregation process proposes an order on the terminals which optimizes the distance to the initial matrix of their dissimilarities (Pak 2005).

The menu in Figure 4 presents the various possibilities at the three levels of formula choice for the algorithm of hierarchical classification for complex data. The aggregation links currently implemented are δ_{Max} , δ_{Min} and the mean link. The distances between statistical units for the same variable are listed at 3.2.2.1 and the aggregation functions between variables can be the maximum, the geometric mean or the weighted quadratic sum.



Dissimilarité

Nombre de classes 3

Attributs	Dissimilarité	Poids
Name_(829)	Aucun	1.0
Year_(830)	L2	1.0
Sexe_(831)	Aucun	1.0
Pop - All Ages_(832)	L2	1.0
Pop - 5-9_(833)	L2	1.0
Pop - 10-14_(834)	L2	1.0
Pop - 15-19_(835)	L2	1.0
Pop - 20-24_(836)	L2	1.0
Pop - 25-29_(837)	L2	1.0
Pop - 30-34_(838)	L2	1.0
Pop - 35-39_(839)	L2	1.0

□ Maximum

☑ Norme géométrique

□ Somme pondérée quadratique

Dissimilarité entre groupes: Lien Moyen

Aucun

FIG. 4 - *Parameters for the automatic classification*

Main results. We now present and comment a hierarchical classification of the ASRs concerning lung cancer for 51 countries over 21 years.

The classes of curves have been computed by the "Minimum link" method and yield a 8 classes partitions; the countries are ordered as it can be seen on the following dendrogram.

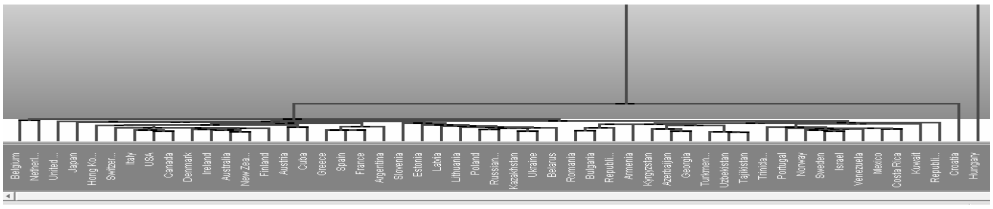


FIG. 5 - *Classification of countries*

We have chosen eight homogeneous groups among all possible partitions:

- Netherlands, Belgium, United Kingdom (value: high then medium, trend: decreasing)
- Italy, USA, Canada, Ireland, Australia, New Zealand, Austria, Switzerland, Finland, Denmark, Hong Kong (value: medium-high, trend: steady then decreasing)
- France, Spain, Greece, Argentina, Cuba (value: medium, trend: steady)
- Russian Federation, Estonia, Poland, Kazakhstan, Ukraine, Lithuania, Latvia, Belarus, Slovenia, Romania, Bulgaria, Rep. of Moldavia, Armenia (value: medium then high, trend: increasing)
- Trinidad & Tobago, Tajikistan, Uzbekistan, Turkmenistan, Georgia, Azerbaijan, and Kyrgyzstan (Value: very low)
- Portugal, Norway, Israel, Sweden (low value, trend: steady then slightly increasing)
- Kuwait, Costa Rica, Mexico, Venezuela (Value : Low; trend: slightly decreasing)
- Rep. of Korea, Croatia, Hungary (Highly increasing)

Japan has particular features: low and steady values for 40-74 then slightly increasing (value and trend).

We have next computed the mean curve for each class, then the distances (see 3.2.2.1) of each time series to the mean of its class and the paragon for the curves of each class is the country, which has the closest time series to the mean (height paragons plus Japan on Figure 6)

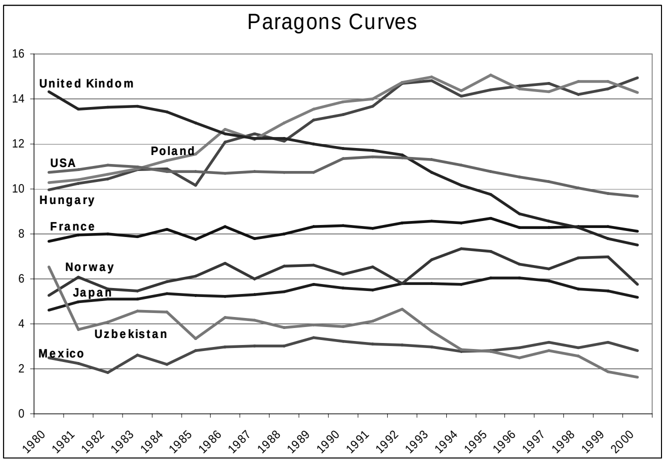


FIG. 6 - *9 typical time series for lung cancer epidemiology 1980-2000*

In addition to the trend typology being classified in 9 groups, the major observation is that the paragon curves of the "western style" countries migrate towards an interval of values much smaller in 2000 than twenty years before, with the "exceptions" of Hungary and Russia. This observation is coherent with and refines the result shown in Figure 7: the inter-geographic variation coefficient of lung cancer for these countries decreases between 1980 and 2000.

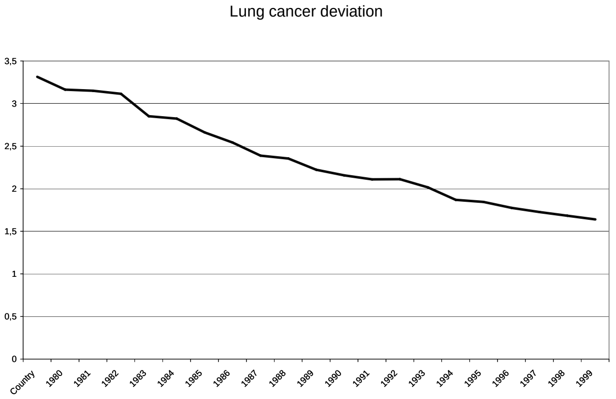


FIG. 7 - *Evolution of Lung cancer inter-geographical deviation between 1980 and 2000*

4 CONCLUSION AND PERSPECTIVES

From a methodological point of view, this paper demonstrates the bonus that the tools of multidimensional analysis bring into the domain of epidemiology on huge mass of information in the exploratory phases. Furthermore, the epidemiological data are temporal by essence and DELTA SUITE software gives the opportunity to take time into account. The main innovation presented in these studies is the fact that each cell of the table under study is a function (in this case time series). We have thus been able to study and to compare epidemiological trends w.r.t. time in several dimensions (geographic, age, disease type, sex) and, particularly, to build typologies for countries in terms of the mortality evolutions of various cancer types since 20 years.

The results show stability and at the same time some clear evolutions in the mortality induced by cancer. If the global mortality, all the sites being put together, stands almost unchanged, the mortality by sites shows big variations. The exploratory data mining of large volumes of mortality data by cancer in the world is therefore useful to reveal the links between various factors or groups of factors. It should also allow estimating the respective weights of these factors on mortality by cancer. In order to achieve this objective, we have to develop new specific modules for the treatment of complex time series that will extend DELTA capabilities: quantitative modeling, forecasting tools in particular. As a natural extension in the epidemiological study, we should add a number of new variables addressing new potentially relevant domains: life style, food, growth speed, psychological trends, etc.

5 BIBLIOGRAPHY

Anderson R.N., Arialdi M.N.,Hoyert D.L., Rosenberg H.M. (2001) *Comparability o f cause of death between ICD-9 and ICD-10: preliminary estimates* National Vital Statistics Reports Volume 49, Number 2

Breslow NE, Day NE. (1987) *Rates and Rate Standardization The Design and Analysis of Cohort Studies* Statistical Methods in Cancer Research, Vol. II, IARC Scientific Publications No. 82, Lyon, International Agency for Research on Cancer 48-79.

Bock H.-H, Diday E.(2000) Analysis of Symbolic Data , Editions Springer, 2000

Huang F. (1999) *Knowledge Discovery in Data Mining highly multiple time series of astronomical observations* CTAC99 Camberra Australia

Gettler-Summa M. , Pardoux C.(2000) *Symbolic Approaches for Three-way Data Analysis of Symbolic Data* , Editions Springer, 342-352

Gettler Summa M., Goupil-Testu F., Goupil J., A.Sasco A., Schwartz L., Vautrain F., (2001) Application de l'Analyse de données aux causes de décès de 1950 à 1997 ; position de la mortalité par tumeur Actes du colloque de la Société Francophone de Classification,

Groupe d'étude et de réflexion interrégional (1996) *L'analyse des données évolutives Méthodes et applications,* Edition technip

Last M., Kandel A., Bunke H. (2004) *Data Mining in time series data bases* World scientific

National Center for Health Statistics. (1998) *Report of the workshop on age adjustment*. Vital and Health Statistics. Series 4. No. 30. December 1998.

Pak Kutluhan Kemal (2005) *Classifications Hiérarchique et Pyramidale Spatiales; Nouvelles Techniques d'Interprétation*, Thèse de Doctorat Université Paris dauphine

Pardoux. C. (2001) *Analyse exploratoire d'une suite de tableaux de données indicés par le temps.* La Revue de MODULAD.. n° 47. 67-102.

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

WHO (1981-2000) International histological classification of tumours, 2nd ed. Geneva, World Health Organization

Aknowlegments to FX Conseil and Phillips Morris International