# A new clustering method for time series to discover geographical cancer trends from 1960 to 2000

Mireille Gettler Summa*, Laurent Schwartz**, Jean Marc Steyaert**, Frédérick Vautrain***, Rolf Weitkunat****
*CNRS France, **Ecole Polytechnique France, ***Isthma France, ****Philip Morris International R&D Switzerland

## BACKGROUND

Long-term trends in cancer mortality can provide important information on etiologic factors. As the spectrum of risk factors, both known and unknown, should not have a completely random geographical distribution, it can be expected that a limited number of trend patterns should be identifiable across countries.

In general, previous research on cancer mortality trends has been focused on specific geographic areas, a specific gender, specific age groups, and specific cancer types. In the present work a generalization of multivariate methods applied to matrices of time series is used to analyze the largest possible subset of the WHO international cancer mortality data. Multivariate analysis was undertaken separately for a selection of both tobacco-related and unrelated cancer, as well as for both types of cancer, in order to reduce the complexity of the available information. The main aim was to identify, for each of these groups of cancers, clusters of countries with similar patterns of cancer trends across age groups.

## APPROACH

In order to identify a limited number of trend patterns across countries, a new pyramidal clustering approach was developed for time series matrices. As indicated above, multivariate analyses were undertaken in order to reduce the complexity of the available information. The term Time Series Matrices Clustering (TSMC) is proposed to denote this approach. Since mortality rates were not available for the whole period of investigation in all countries, analyses were conducted across 31 countries along 41 years as well as across 52 countries along 21 years.

## METHODS

### CLUSTERING TIME SERIES MATRICES

TSMC was implemented using the DELTA Metrics© software. The method is based on assessing the proximity of multiple multivariate time series by computing new indices of dissimilarity which allows merging multiple Euclidean distances computed between pairs of curves. Part of the framework is the Functional Data Analysis. Geometrical aggregations of integration-based distances were computed (L2 to account for unlagged pairs of time series) and the minimal linkage ultra metrics was used to build the pyramid on the initial functions joined to the first-difference functions tables.

### DATA

Cancer mortality data from initially 122 countries were extracted from the World Health Organization statistical database (WHOSIS) in March 2005. This database contains absolute numbers of deaths officially reported by WHO member states. Standardized age ratios were computed, according to the 7th, 8th, 9th, and 10th ICD (Segi reference population) for all cancers, tobacco unrelated cancers, and lung cancer (2004 IARC monograph). Due to missing data and geopolitical changes only 31 countries could be considered in the 1960 to 2000 and only 52 in the 1980 to 2000 analyses. Age groups from 40 to 74 years were considered (seven 5-year age groups).

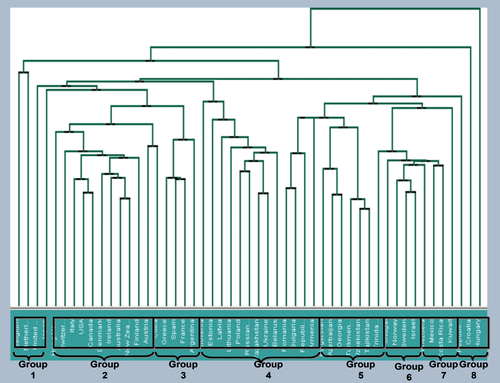
### COMPUTATION OF TIME SERIES SIMILARITY

**Similarity in Time**
- correlation based distance
- Lp norm etc

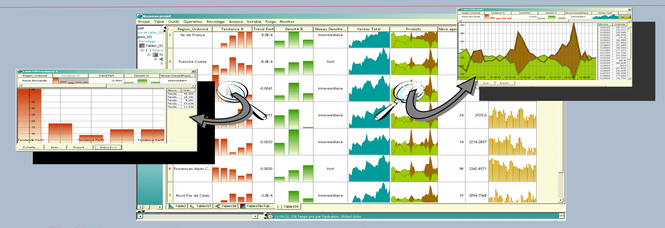
**Similarity in shape**
- Differential
- Dynamic Time Warping transformations (Berndt & Clifford 1994)

**Similarity in change**
- SARIMA parameters etc

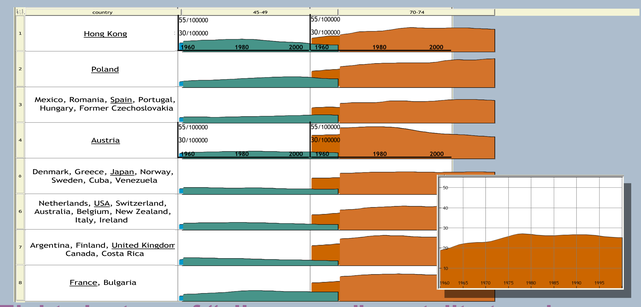
### GENERALIZED PYRAMIDISATION




## RESULTS



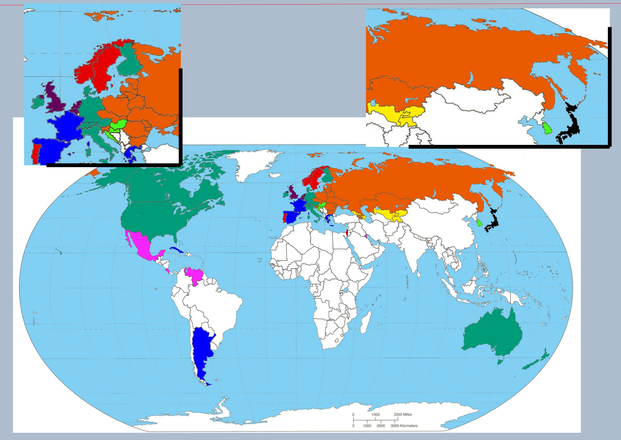
Editing a time series matrix.



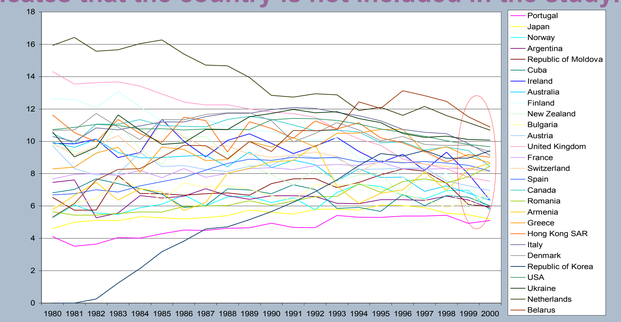

Eight clusters of “all cancers” mortality trends (males, 40-74 years, 1960-2000).




Nine clusters of “lung cancer” mortality trends (males, 40-74 years, 1960-2000).



Nine clusters of lung cancer mortality trends (males, 40-74 years, 1980-2000). Countries of the same color are in the same lung cancer mortality trend cluster. White indicates that the country is not included in the study.




Lung cancer mortality in 28 countries during the period of 1980-2000 (males, 65-69 years of age).


Nine clusters of “tobacco unrelated cancers” mortality trends (males, 40-74 years, 1980-2000).

Underlined countries are typical representatives of their cluster with respect to trends over the whole period and for all age classes.

## CONCLUSIONS

Most of the groupings are stable across the two periods of investigation.

Clusters of cancer mortality trends often correspond to geographical proximities.

There is a converging pattern of cancer mortality trends in most Western countries.

Clusters of countries for all cancers and for lung cancer are not identical; results corresponding to the latter are more heterogeneous.

Trends regarding tobacco-related and unrelated cancers combined largely reflect changes in lung cancer trends.