Clustering Trajectories of a Three-Way Longitudinal Dataset

Mireille Gettler Summa, Bernard Goldfarb, Maurizio Vichi

Clustering Trajectories of a Three-Way Longitudinal Data Set

Mireille Gettler Summa, Bernard Goldfarb, and Maurizio Vichi

Abstract: Longitudinal data are widely used information for repeated observations of the same units over a period of time in order to investigate developmental trends across life span of units. Each object depicts, in the space of the features and of time, a trajectory describing its changes over time. Here trajectories are modeled according to three features: trend, velocity and acceleration. Clustering trajectories of a longitudinal data set is an important issue to assess similarities in the histories of the observed units that we fully discuss in this chapter. Starting from the Tucker model, widely used in psychometrics, we consider the optimal partition of trajectories that minimizes a distance accounting for trend, for velocity and for acceleration of trajectories. A Sequential Quadratic Programming algorithm is proposed to solve the clustering problem and its performance is evaluated by simulation.

Key words: Trajectories, trend, velocity acceleration, time series clustering, three way data table, factorial analysis, complex data

1. Introduction

In recent years, growing attention has been paid to the study of multivariate-multioccasion phenomena analyzed through a set **X** of *IJT* values corresponding to *J* variables, observed on a set of *I* units, on *T* different occasions (different times, places, etc.). The three-way array **X** is organized according to three modes: units, variables and occasions. The most widely collected three-way array is given when, together with units and variables, different time occasions are considered. The temporal repeated observation of the units allows us to evaluate the dynamics of the phenomenon differently from the classical case of a multivariate or cross-sectional (two-way) data set. There are several major advantages, over conventional cross-sectional or univariate time-series data sets, when we use *three-way longitudinal data*: the researcher has a large amount of data to describe the phenomenon increasing the degrees of freedom and reducing co-linearity among explanatory variables. This allows us to make inferences about the dynamics of change from cross-sectional evidence.

The three-way longitudinal data may be the result of the following type of observation. (i) Repeated recurring surveys, with no overlap of units, e.g., a survey organization repeats a

survey on a defined topic, generally with regular time intervals. No overlaps of the sample units are required at different times. Examples of these surveys are given by the repeated analyses made by all Central Bureaux of Statistics. (ii) Repeated surveys with partial overlap of units. Also these surveys are repeated at regular intervals. The survey design includes rotating units to allow variance reduction, i.e., units are included in the analysis a number of times, and then rotated out of the survey outcome. iii) *Longitudinal surveys, with no rotation of units*. A set of units is followed over time with a survey designed with this aim. These are called *panel data* and in the current work we will refer to this type of observation.

In addition let us now suppose that units are heterogeneous, i.e., the population, from which the data are observed at time t, is composed of G homogeneous disjoint subpopulations.

Panel data are usually from a small number of observations over time (short time series) on a usually large number of cross-sectional units like individuals, households, firms, or governments, and frequently characterize economic, demographic and social phenomena.

The chapter is organized as follows. Section 2 briefly lists the notation used; while, Section 3 describes three features of a trajectory: trend, velocity and acceleration. Section 4 describes dissimilarities between trajectories, while Section 5 illustrates the model used for clustering and the algorithm proposed. Section 6 is devoted to the application on the lung cancer data.

2. Notation

For the convenience of the reader the notation and terminology used is listed here.

I, J, T	number of units, variables and occasions, respectively;
G, Q	number of classes, components for variables, respectively;
$C_1, C_2, \ldots, C_G$	G clusters of units;
$\mathbf{X} = [x_{ijt}]$	$(I \times J \times T)$ three-way data array; where x_{ijt} is the value of the j^{th} variable observed on the i^{th} object at the t^{th} time. On each occasion, the variables are supposed commensurate, if this is not the case the data are supposed standardized;
$\mathbf{X}_{I,JT}$	$(I \times JT)$ matrix $[\mathbf{X}_{..1}, \mathbf{X}_{..2}, \ldots, \mathbf{X}_{..T}]$, i.e., the matricized version of $\mathbf{X}$ with frontal slabs $\mathbf{X}_{..t} = [x_{ijt}]_{I \times J}$ next to each other. It is column standardized;
$\mathbf{E} = [e_{ijt}]$	$(I \times J \times T)$ three-way arrays of residual terms;

$\mathbf{E}_{I,JT}$ ($I \times JK$) matrix [$\mathbf{E}_{..1}$,…, $\mathbf{E}_{..K}$], i.e., the matricized version of $\underline{\mathbf{E}}$ with frontal $\mathbf{E}_{..k}$ =$[e_{ijk}]_{I\times J}$ slabs next to each other;

$\mathbf{U} = [u_{ig}]$ ($I \times G$) membership function matrix defining a partition of units, into G classes, where $u_{ig} = 1$ if the i^{th} object belongs to class g, $u_{ig} = 0$ otherwise. Matrix $\mathbf{U}$ is constrained to have only one nonzero element per row;

I_g cardinality of cluster C_g, i.e. $I_g = \left|C_g\right| = \sum_{i=1}^{I} u_{ig}$;

$\overline{\mathbf{X}} = [\bar{x}_{gjt}]$ ($G \times J \times T$) three-way centroid array, where $\bar{x}_{gjt}$ is the centroid value of the j^{th} variable obtained on the g^{th} cluster at the t^{th} occasion;

$\overline{\mathbf{X}}_{G,JT}$ ($G \times JT$) centroids matrix, i.e., matricized version of the centroid array $\underline{\overline{\mathbf{X}}}$, with frontal slabs $\overline{\mathbf{X}}_{..k} = [\bar{x}_{gjk}]_{G\times J}$ next to each other;

$\mathbf{x}_i$, $\mathbf{u}_i$, $\boldsymbol{e}_i$, column vectors representing the i^{th} row of $\mathbf{X}$, $\mathbf{U}$ and $\mathbf{E}$ respectively;

$\bar{\mathbf{x}}_g$ g^{th} row of $\overline{\mathbf{X}}$, specifying the centroid vector of the g^{th} class of the partition of the I objects.

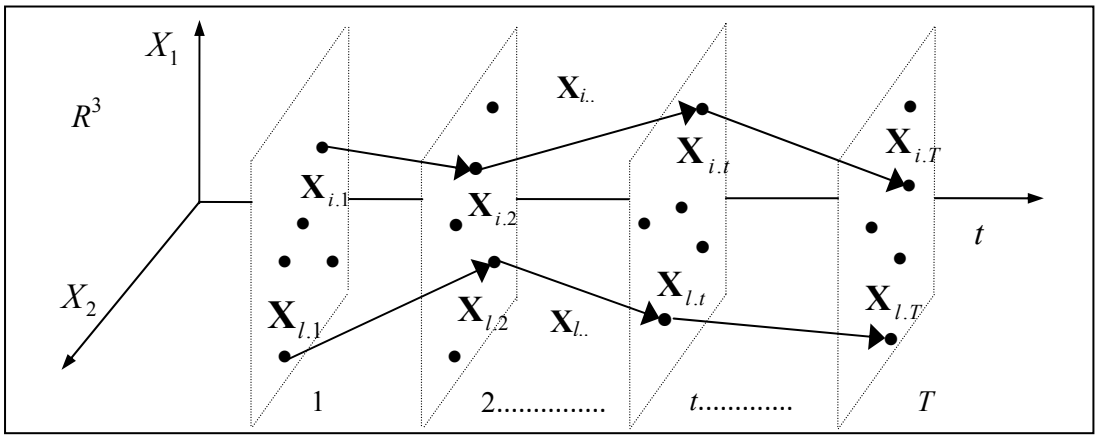
This chapter deals with the problem of partitioning trajectories of a three-way longitudinal data set into classes of homogeneous trajectories.

3. Trajectories

A time trajectory describes a nonlinear curve in the J+1 dimensional space that has several characteristics; specifically we consider: trend, velocity and acceleration (D'Urso & Vichi 1998).

For each object i, $\mathbf{X}_{i..} \equiv \left\{x_{ijt} : j = 1,...,J; t = 1,...,T\right\}$ describes a *time trajectory* of the ith object according to the J examined variables. The trajectory $\mathbf{X}_{i..}$ is geometrically represented by T-1 segments connecting T points $\mathbf{X}_{i.t}$ of R^{J+1} (Figure 1).

Figure 1: Two time trajectories in R^3



Trend is the basic characteristic of a trajectory indicating the tendency of the *J*-variate objects along different time points.

Velocity and acceleration are two trajectories' characteristics strongly describing changes of the shape of trajectories. For example in R^2, velocity of each segment of the trajectory is the slope of the straight line passing through it: if velocity is negative (positive) the slope will be negative (positive) and the angle made by each segment of the trajectory with the positive direction of the t-axis will be obtuse (acute). Geometrically, acceleration of each pair of segments of trajectory represents their convexity or concavity. If acceleration is positive (negative) the trajectory of the two segments is convex (concave).

For each time trajectory $\mathbf{X}_{i..}$, the velocity of evolution of an object *i* in the interval from *t* to *t*+1, denoted $s_{t,t+1}$, is, for the *j*th variable

$$v_{ijt,t+1} = \frac{x_{ijt+1} - x_{ijt}}{s_{t,t+1}}. \qquad (1)$$

In particular: $v_{ijt,t+1} > 0$ ($v_{ijt,t+1} < 0$) if object *i*, for the *j*th variable, presents an increasing (decreasing) rate of change of its position in the time interval from *t* to *t*+1; $v_{ijt,t+1} = 0$ if the object *i* for the *j*th variable does not change position from *t* to *t*+1.

Acceleration measures the variation of velocity of $\mathbf{X}_{i..}$ in a fixed time interval.

For each time trajectory $\mathbf{X}_{i..}$, the acceleration of an object *i* in the interval from *t* to *t*+2, denoted $s_{t,t+2}$, is, for the *j*th variable

$$a_{ijt,t+2} = \frac{v_{ijt+1,t+2} - v_{ijt,t+1}}{s_{t,t+2}}. \qquad (2)$$

Of course acceleration must be computed on two *contiguous* time intervals [t, t+1], [t+1, t+2]. In particular: $a_{ijt,t+2} > 0$ ($a_{ijt,t+2} < 0$) if the object i, for the jth variable, presents an increasing (decreasing) variation of velocity in the time interval from t to t+2; $a_{ijt,t+2} = 0$ if object i, for *the jth* variable, does not change velocity from t to t+2.

Therefore, the basic information of a trajectory can by organized into three three-way matrices $\mathbf{X} = [\mathbf{X}_{..t} = [x_{ijt}]_{I \times J},\ t = 1,\ldots,T]$; $\mathbf{V} = [\mathbf{V}_{..t,t+1} = [v_{ijt,t+1}]_{I \times J},\ t = 1,\ldots,T\text{-}1]$ and $\mathbf{A} = [\mathbf{A}_{t,t+2} = [a_{ijt,t+2}]_{I \times J},\ t = 1,\ldots,T\text{-}2]$ respectively for trend, velocity and acceleration, where:

$$\mathbf{V}_{..t,t+1} = \frac{1}{s_{t,t+1}}(\mathbf{X}_{..t+1} - \mathbf{X}_{..t}); \qquad \mathbf{A}_{..t,t+2} = \frac{1}{s_{t,t+2}}\left(\frac{1}{s_{t+1,t+2}}(\mathbf{X}_{..t+2} - \mathbf{X}_{..t+1}) - \frac{1}{s_{t,t+1}}(\mathbf{X}_{..t+1} - \mathbf{X}_{..t})\right).$$

4. Dissimilarity between trajectories

A dissimilarity between trends of objects $\mathbf{X}_{i\ldots}$ and $\mathbf{X}_{l..}$ is evaluated according to a measure of distance between $\mathbf{X}_{i.t}$ and $\mathbf{X}_{l.t}$, for $t = 1,\ldots,T$,

$$_1\delta(i,l) = \pi_1 \sum_{t=1}^{T} \left\| \mathbf{X}_{i.t} - \mathbf{X}_{l.t} \right\|^2 = \pi_1 \sum_{t=1}^{T} tr\left[(\mathbf{X}_{i.t} - \mathbf{X}_{l.t})'(\mathbf{X}_{i.t} - \mathbf{X}_{l.t}) \right] \tag{3}$$

where π_1 is a suitable weight to normalize distances .

A dissimilarity between velocities of objects $\mathbf{X}_{i\ldots}$ and $\mathbf{X}_{l..}$ in a time interval, is evaluated according a measure of distance between $\mathbf{V}_{i.t,t+1} = (v_{i1t,t+1}, \ldots, v_{iJt,t+1})'$ and $\mathbf{V}_{l.t,t+1}$, $t = 1,\ldots,T\text{-}1$;

$$_2\delta(i,l) = \pi_2 \sum_{t=1}^{T-1} \left\| \mathbf{V}_{i.t,t+1} - \mathbf{V}_{l.t,t+1} \right\|^2 = \pi_2 \sum_{t=1}^{T-1} tr\left[\left(\mathbf{V}_{i.t,t+1} - \mathbf{V}_{l.t,t+1}\right)'\left(\mathbf{V}_{i.t,t+1} - \mathbf{V}_{l.t,t+1}\right) \right] \tag{4}$$

where π_2 is a suitable weight to normalize the velocity dissimilarity,

A dissimilarity between accelerations of objects $\mathbf{X}_{i\ldots}$ and $\mathbf{X}_{l..}$ in a time interval, is evaluated according to a measure of distance between $\mathbf{A}_{i.t,t+2} = (a_{i1t,t+2}, \ldots, a_{iJt,t+2})'$ and $\mathbf{A}_{l.t,t+2}$, $t = 1,\ldots,T\text{-}2$,

$$_3\delta(i,l) = \pi_3 \sum_{t=1}^{T-2} \left\| \mathbf{A}_{i.t,t+2} - \mathbf{A}_{l.t,t+2} \right\|^2 = \pi_3 \sum_{t=1}^{T-2} tr\left[\left(\mathbf{A}_{i.t,t+2} - \mathbf{A}_{l.t,t+2}\right)'\left(\mathbf{A}_{i.t,t+2} - \mathbf{A}_{l.t,t+2}\right) \right] \tag{5}$$

where π_3 is a suitable weight to normalize the acceleration dissimilarity.

A dissimilarity between two trajectories that takes into account trend, velocity and acceleration is thus formalized as the sum of the three individual dissimilarities

$$d(i,l) = \pi_1 \sum_{t=1}^{T} \left\| \mathbf{X}_{i.t} - \mathbf{X}_{l.t} \right\|^2 + \pi_2 \sum_{t=1}^{T-1} \left\| \mathbf{V}_{i.t,t+1} - \mathbf{V}_{l.t,t+1} \right\|^2 + \pi_3 \sum_{t=1}^{T-2} \left\| \mathbf{A}_{i.t,t+2} - \mathbf{A}_{l.t,t+2} \right\|^2 \quad (6)$$

5. The Clustering Problem

For clustering the trajectories we minimize the following loss function with respect to binary variable matrix **U**, and continuous variables matrices $\overline{\mathbf{X}}_{..t}$, $\overline{\mathbf{V}}_{..t,t+1}$ and $\overline{\mathbf{A}}_{..t,t+2}$, where we add a feature of a dimensionality reduction of the variables via the orthonormal projection matrix **BB'**

$$\text{Min} \quad \pi_1 \sum_{t=1}^{T} \left\| \mathbf{X}_{..t} - \mathbf{U}\overline{\mathbf{X}}_{..t}\mathbf{BB'} \right\|^2 + \pi_2 \sum_{t=1}^{T-1} \left\| \mathbf{V}_{..t,t+1} - \mathbf{U}\overline{\mathbf{V}}_{..t,t+1}\mathbf{BB'} \right\|^2 + \pi_3 \sum_{t=1}^{T-2} \left\| \mathbf{A}_{..t,t+2} - \mathbf{U}\overline{\mathbf{A}}_{..t,t+2}\mathbf{BB'} \right\|^2 \quad (7)$$

subject to [P1]

$\mathbf{B'B}=\mathbf{I}_J$

$u_{ig} \in \{0, 1\}$ $(i=1,\ldots,n;\ g=1,\ldots,G)$, (8)

$$\sum_{g=1}^{G} u_{ig} = 1 \quad (i=1,\ldots,n). \quad (9)$$

where matrices $\overline{\mathbf{X}}_{..t}$, $\overline{\mathbf{V}}_{..t,t+1}$ and $\overline{\mathbf{A}}_{..t,t+2}$ are the matrices of the *G* consensus trajectories including trend, velocity and acceleration information. In problem [P1] the observed trajectories are classified in G consensus trajectories and their location in the space is identified. Furthermore, we suppose we consider a dimensionality reduction specified by the orthonormal projection matrix **BB'**.

The quadratic problem [P1] in the continuous variables $\overline{\mathbf{X}}_{..t}$, $\overline{\mathbf{V}}_{..t,t+1}$ and $\overline{\mathbf{A}}_{..t,t+2}$ and binary **U** is solved here by using the sequential quadratic algorithm (SQP) (Powell 1983). It is well known that the partitioning of *n* objects in *k* clusters is a NP-hard problem in the class of the NP-complete problems (Krivánek & Morávek 1986), therefore the problem of clustering

trajectories which is a three-way extension is also NP-Hard and no guarantee to find the optimal solution is available. Therefore to increase the chance to find the optimal solution a multistart procedure is applied which consists of starting the algorithm from different random solutions and retaining the best solution.

6. Application

Cancer mortality data, initially from 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 for the years 1980-2000. For these years, the WHO database includes cause-of-death statistics, coded according to two former versions of the ICD (International Classification of Diseases, version 9 from 1979 to 1998, and version 10 from 1999); the years of transition between the two versions exhibit large differences among countries. We accounted for these changes in disease classification by using specific transition data sets available on the WHO website since 2005. In this chapter we focus on lung cancer, because the mortality data were available across a sufficient number of countries, age bins and years.

Data related to the age below five (especially for the age below one year) and above 89 are reported in a heterogeneous and incomplete way across countries. In the present analysis, only age groups from 40 to 74 were considered (leading to seven 5-year age bins: 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74). To account for differences in country- and period-specific variations in age distributions, the mortality data with respect to lung cancer were directly age-standardized according to Segi's world population [REF?], for men.

We exclude 23 countries because there were clearly visible outliers from the local (i.e. country-specific) trend (e.g. Brazil 1971, 1982; Chile 1988; Portugal 1988, 1989). Moreover, in order to assess the degree of annual variation of the data, coefficients of variation were computed, and the countries above the 80th percentile were excluded.

Information on mortality for the period of 1980 to 2000 was missing for more than five years in 47 countries. Therefore these countries were not considered in the present analysis. For the remaining 51 countries with less than five years of missing data, imputation was undertaken by interpolation (spline interpolation when possible, otherwise linear interpolation).

For the years up to 1998, the lung cancer absolute frequencies were provided in the WHO database, whereas for the years from 1999 (ICD 10) they had to be computed by integrating across absolute frequencies for all specific sites.

We aim at categorising the evolution of lung cancer mortality in the past 21 years from 1980 to 2000 in the selected 51 countries; and we expect to uncover some general trends in the clusters. The resulting data array is a three way (51 × 21 × 7) table, 51 countries, 21 years, 7 age bins.

We present some results for partitions from 2 up to 11 clusters, and some more detailed results for the 2-cluster partition and for the 11-cluster partition.

The algorithm indicates the 2-cluster partition as the optimum one according to the pseudo F criterion (REF? 'Optimal f' [f OR F? BUT CF. TABLE 1] concerns the final reduced space) as Table 1 shows.

clusters	Pseufo F	Optimal f
2	238,6807	27,2302
3	183,8518	24,2847
4	146,6301	23,0354
5	124,0362	21,9418
6	107,8434	21,114
7	96,334	20,3309
8	85,9706	19,9132
9	77,8542	19,5201
10	71,4224	19,1255
11	67,3236	18,4785

Table 1

Figure 1 shows the consensus trajectories for the two cluster partition.

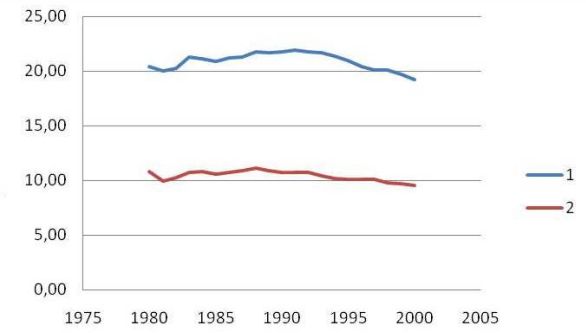


Figure 1

Figure 1b represents the two consensus trajectories in the first factorial plane.

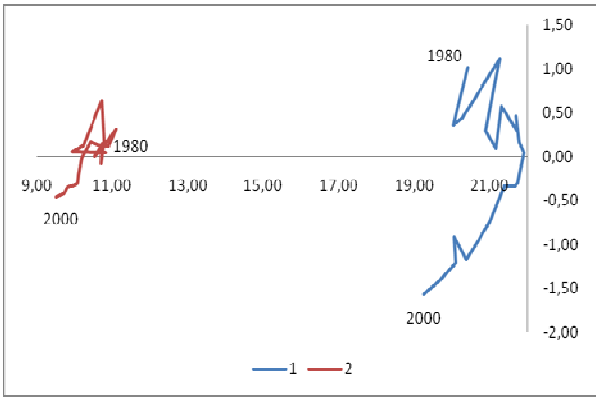


Figure 10b [FIG. 1, RIGHT PANEL?]

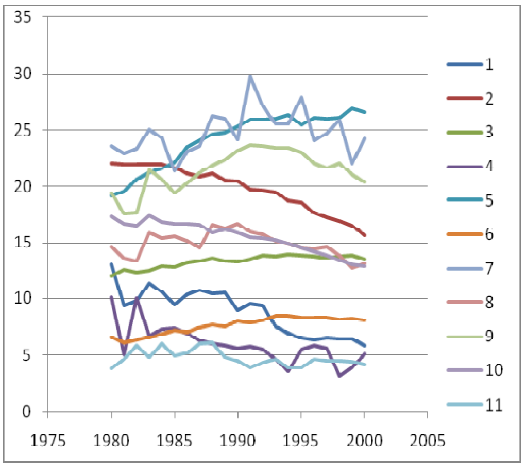


Figure 2

Figure 2 shows the eleven consensus trajectories of the 11-cluster partition on the first principal component whereas Figure 3 selects the less "erratic" ones: three clusters, cluster 4, cluster 7 and cluster 11 appear to be sparse (less than two elements) and could be considered as outliers (countries Kuwait, cluster 4; Estonia, cluster 7; Trinidad and Costa Rica, cluster 11).

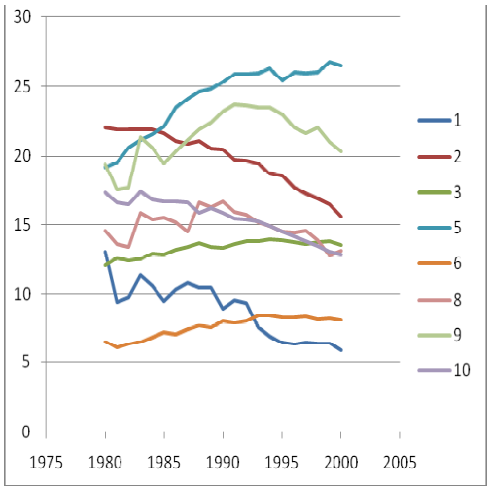


Figure 3 [CAPTION?]

Figure 4 shows a geographical map restricted to Europe with the country memberships of the selected eight clusters.

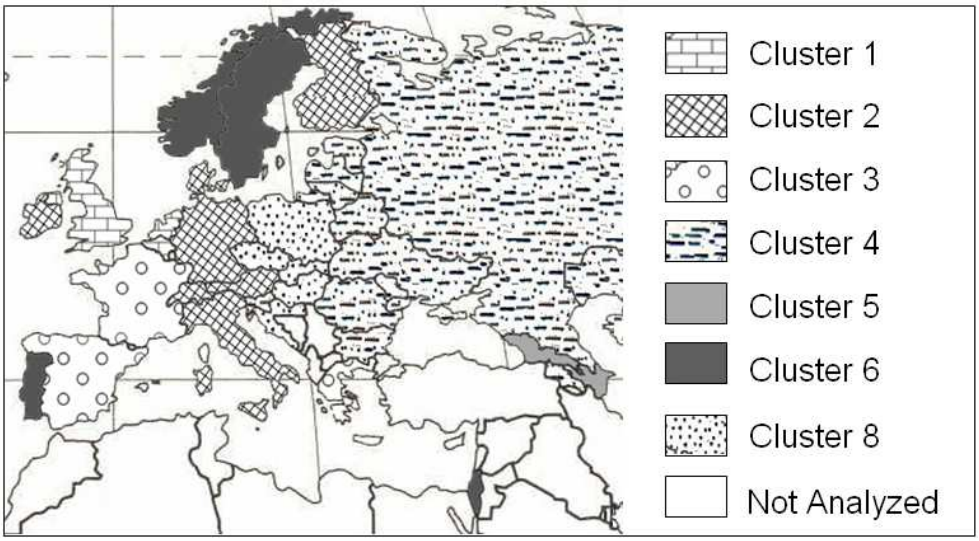


Figure 4 [CAPTION?]

7. Conclusions

We still have a lack of theoretical elements for solutions during the approximation step, particularly in the weighting step for the different components in the distance formula. We are evaluating a coefficient based on the Mahalanobis distance, but it needs to be adjusted at each step and a global satisfying procedure has not yet been found because sequential quadratic programming is quite computationally demanding. A more specific algorithm is needed and we are studying a new version of a fast coordinate descent algorithm.

The results of this application on the lung cancer data can be compared to those of more complete studies on cancer evolution typologies [3] [4].

Some convergence can be pointed out and particularly the similarity between the 8-cluster partitions (mainly for European countries), even if not exactly identical everywhere.
Nevertheless, we also observe some discrepancies:

- geographical proximities appear to be respected by the clustering procedure when it deals only with the values, but they are less apparent when including velocity and acceleration (e.g. Hong Kong is associated with Australia, but also with Austria)

- the 2-cluster partition was expected to group countries according to "western style of life" (with a convergent decreasing pattern on values) as opposed to the complementary group of countries; we do not obtain the same finding in the new approach (for example France is associated with Turkmenistan). It seems that the partition in two clusters is mainly linked to the different levels of variables. Probably for this partition levels of variables are more important than contiguity of countries.

-

Further work should focus on the elaboration of interpretation procedures for the resulting clusters in the approach presented in this chapter. How can the axes of figure 0bis[??] be labeled in order to reveal why cluster 2 has a larger range of coordinates on the second axis than cluster 1 along the 21 years? What are the main effects on the groupings: common ranges both for values and velocities, or whatever else? Moreover, supplementary (illustrative) variables could be introduced in order to propose hints for explaining the between-cluster differences.
[COULD SOME FINAL SENTENCE OR TWO BE INCLUDED TO STRESS THE NOVELTY AND IMPORTANCE OF THIS WORK?]

References

[1]D'Urso, P., Vichi M. (1998). Dissimilarities between Trajectories of a Three-Way Longitudinal Data Set, In*: Advances in Data Science and Classification*, A. Rizzi, M. Vichi, H.H.Bock (Eds), *Studies in Classification, Data Analysis, and Knowledge Organization*, Springer-Verlag, Heidelberg, 585-592.

[2]Gabrielson, E. (2006). Worldwide trends in lung cancer pathology. Respirology, 11, 533-538.

[3]Gettler-Summa, M., Schwartz, L., Steyaert, J. M., Vautrain, F., Barrault, M., & Hafner, N. (2006). Multiple time series: New approaches and new tools in data mining applications to cancer epidemiology, MODULAD Journal 34, 37-46**,** INRIA

[4]Kriványek, M. & Morávek, J. (1986), NP-Hard Problems in Hierarchical-Tree Clustering, *Acta Informatica*, 23, 311-323.

[5]Levi, F., Lucchini, F., Negri, E. Zatonski, W., Boyle, P., & LaVecchia, C. (2004). Trends in cancer mortality in the European Union and accession countries, 1980-2000. Annals of Oncology, 15, 1425-1431

[6]Powell, M.J.D. (1983): Variable Metric Methods for Constrained Optimization, in *Mathematical Programming: The State of the Art*, A. Bachem, *et al*. eds, Springer Verlag, 288-311.