Constructing social and economic indicators for EU countries using dynamic classification: case studies

I. Gertsbakh[1], I. Yatskiv[2] & O. Platonova[2]
[1]*Department of Mathematics, Ben Gurion University, Israel*
[2]*Department of Computer Science,*
Transport & Telecommunication Institute, Latvia

Abstract

This paper presents applications of the dynamic classification algorithm (DCA) described in Gertsbakh and Yatskiv (Dynamic Classification: Economic Welfare Growth in EU During 1995-2004. Proceedings of International Conference "Data Mining 2006", 11-13 July 2006, Prague. WIT Press. 2006, p.53-62) to the development of three socio-economic indicators: National Health Index (NHI), Population Mobility Index (PMI) and Logistic Performance Index (LPI). In each of these three cases we work with the respective data for EU-25 countries. The essence of the DCA is to transform a multidimensional vector to a scalar on the basis of combining cluster and discriminant analysis. In Gertsbakh and Yatskiv, the DCA was used for obtaining a measure of socio-economic welfare and its dynamics for EU-25 countries over the period 1995-2004. The output of DCA is a collection of time series (graphs) representing the dynamics and the pattern of the scalar socio-economic indicator for each country in the considered time period.

Keywords: cluster analysis, Fisher discriminant function, time series, socio-economic indicators.

1 Introduction: the data and the algorithm

The data. The row data are given as a three-dimensional array of entries having the following form:

$$\mathbf{y}(i; j = 1,...,k; t) = [x_{i1}(t),...,x_{ij}(t),...,x_{ik}(t)], \quad (1)$$

WIT Transactions on Information and Communication Technologies, Vol 40, © 2008 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA080151

where i is the object number, $i = 1, ..., N$; j is the parameter number, $j = 1, ...,K$; t denotes the year, $t = 1, ..., T$. The total number of entries in (1) is $(N{\cdot}K{\cdot}T)$. The "position" of an object in a particular year is described by a point (1) in K-dimensional space.

Since the original data (1) consist of indices of various dimensions and of highly differing ranges, we operate with *standardized* data: we replace $x_{ij}(t)$ by its standardized value $x_{ij}^{\circ}(t) = (x_{ij}(t) - m_j)/s_j$, where m_j and s_j are the averages and standard deviations of the j-th coordinate which were calculated from the training sample. The training sample is chosen as a median year in case of T being odd or as an average of two "middle" years if T is even.

Clustering. We apply the following principal clustering algorithms to the training sample: Ward's, Complete Linkage, Iterative k-means. We single out three clusters of countries. The first cluster (denoted as "L") contains countries with the "worst" values of variables. The second cluster (denoted as "H") contains the countries with the "best" variable values. The remaining countries constitute the cluster "M" (middle).

Since the results of cluster analysis have some small variations from method to method, we use a "robust" decision, i.e. the L and H clusters are taken as the *intersection* of the corresponding clusters produced by different methods.

Finding the Fisher vector. Calculate the coordinates of the centers of the L and H clusters and the vector $\mathbf{c} = [c_1,...,c_K]$ connecting their centers as having coordinates $c_j = c_j^H - c_j^L, j = 1,...K.$

Compute the Fisher vector $\mathbf{f}$, which provides the maximal separation of L and H, according to the following formula, see [2]:

$$\mathbf{f} = [f_1 ,...,f_k] = \mathbf{W}^{-1}\mathbf{c}' \tag{2}$$

where $\boldsymbol{W}$ is the pooled variance-covariance matrix computed for H and L clusters. If the coordinates of vector $\mathbf{f}$ have "noncorrect" signs (according to their influence of the corresponding coordinate on the resulting index), it is replaced by a vector $\mathbf{f}^*$ by introducing a suitable convex combination of the Fisher vector and the vector connecting the centres of groups H and L.

Formula for the Index. The index $I(i,t)$ is defined according to eqn.

$$I(i,t) = A\sum_{j=1}^{j=K} w_j x^0(i,j,t) + B = 10 \cdot \frac{\sum_{j=1}^{K} f_j^*(x_{ij(t)} - \bar{x}_j^L)/s_j}{\sum_{j=1}^{K} f_j^*(\bar{x}_j^H - \bar{x}_j^L)/s_j} \tag{3}$$

We choose the constants A and B to provide that for the original (nonstandarized) center coordinates of groups L and H, $\bar{x}_j^L$ and $\bar{x}_j^H$, $j=1,...,K$, the Index would be equal 0 and 10, respectively.

Geometrically, our procedure means, up to a linear transformation, projecting the point in K-dimensional space on the line determined by the Fisher vector. This vector guarantees the best separation of the extreme clusters according to the Mahalanobis distance $\mathbf{D}^2$, see [2].

2 National Health Index (NHI)

The following six principal parameters were used for the NHI for EU-25 in 1993-2003.

1) Life expectancy at birth in years, the average for men and women [LIFE_0];
2) Life expectancy at the age 65 (the average number of further years for a. person at age 65), average for men and women [LIFE-65];
3) Number of infant deaths per 1000 live births [INFANT];
4) Number of medical practitioners per 100,000 population [MEDIC];
5) Health care expenditure as a share of GDP % [HEALTH];
6) Total health expenditure per head of population in PPP USD [EXP].

The first three parameters characterize the health as a whole, and the last three the expenditures for maintaining the national health system. All data are taken from Eurofound [3]. The average values of these parameters for the training year 1998 for EU-25 are presented in the Table 1. Cluster analysis results on the basis of Ward's method are presented on Fig 1.

Table 1: The coordinates of cluster centroids and the common centroid for the training sample.

Variable	LIFE_0	LIFE_65	INFANT	EXP	HEALTH	MEDIC
Common centroid	75.96	16.36	6.60	1426.12	5.76	317.24
Cluster H	78.40	17.50	4.80	2257.85	7.58	323.02
Cluster L	71.45	14.42	10.33	495.28	4.50	318.64

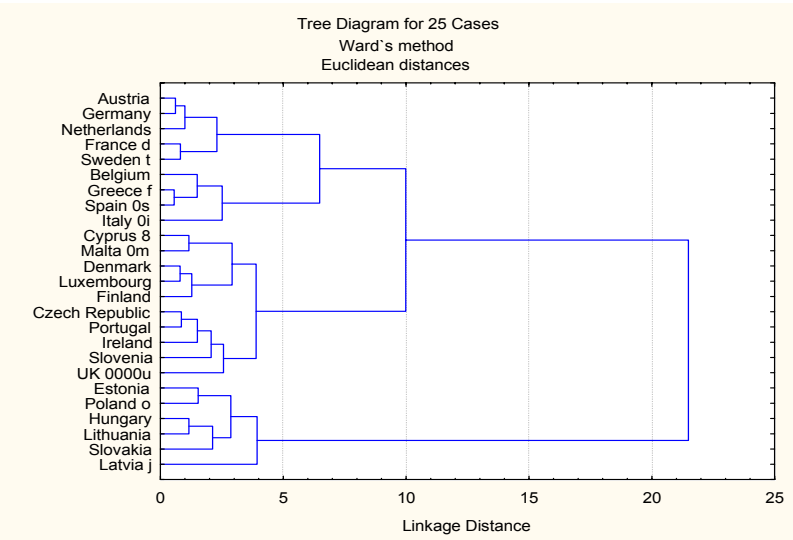

Figure 1: The results of classification by Ward's method.

Eventually, cluster L has 6 countries: Estonia, Latvia, Lithuania, Poland and Slovakia. Cluster H contains 9 countries: Austria, Belgium, France, Germany, Greece, Italy, Netherlands, Spain, and Sweden.

The uncorrected $\mathbf{f}$ and corrected $\mathbf{f}^*$ Fisher vector coordinates are given in table 2. The Mahalanobis distance for corrected function equals 112.47, which is quite satisfactory.

Table 2: The uncorrected and corrected Fisher vectors.

Variable	LIFE_0	LIFE_65	INFANT	EXP	HEALTH	MEDIC	D^2
$\mathbf{f}$	0.95	1.18	0.237	3.03	-0.26	-0.18	166.3
$\mathbf{f}^*$	1.17	1.35	-0.13	2.96	0.10	-0.14	112.47

The final formula for the NHI computed by eqn. (3) as follows:

NHI = 0.286*LIFE_0+0.724*LIFE_65 - 0.0371*INFANT+0.00307*EXP+
+ 0.0519*HEALTH - 0.00117*MEDIC – 31.88

The time series for the NHI for clusters H and L are presented on Figure 2 and Figure 3, respectively.

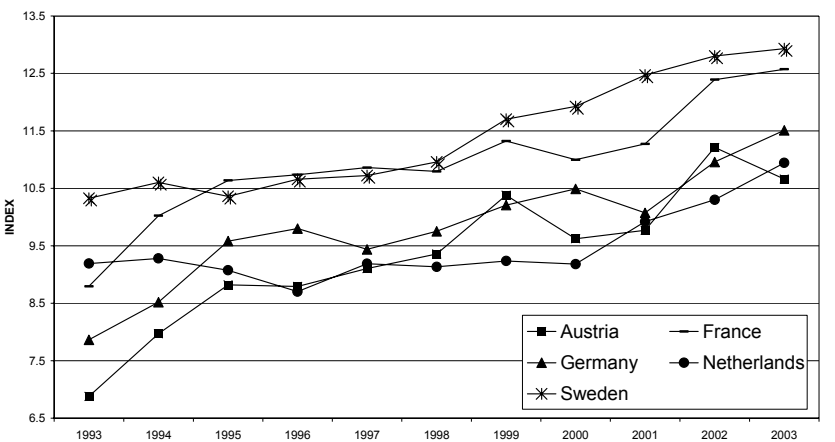

Figure 2: Time series of NHI values for countries of the cluster H.

3 Population Mobility Index (PMI)

We apply our DCA for estimating the Population Mobility Index (PMI), for 23 of EU-25 countries over the period 1996-2004. The data for two counties (Malta and Cyprus) are missing.

The following three principal parameters were used for the PMI:

- Persons killed in road accidents per thousand inhabitants [KILLED];
- Passenger cars per 1 000 inhabitants [CARS];
- Total length of motorways in km per 1000 square kilometers [WAYS].

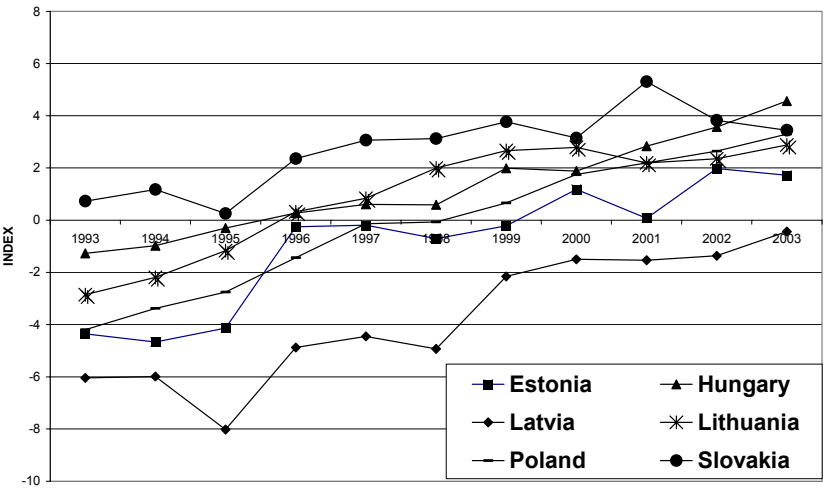

Figure 3: Time series of NHI values for countries of the cluster L.

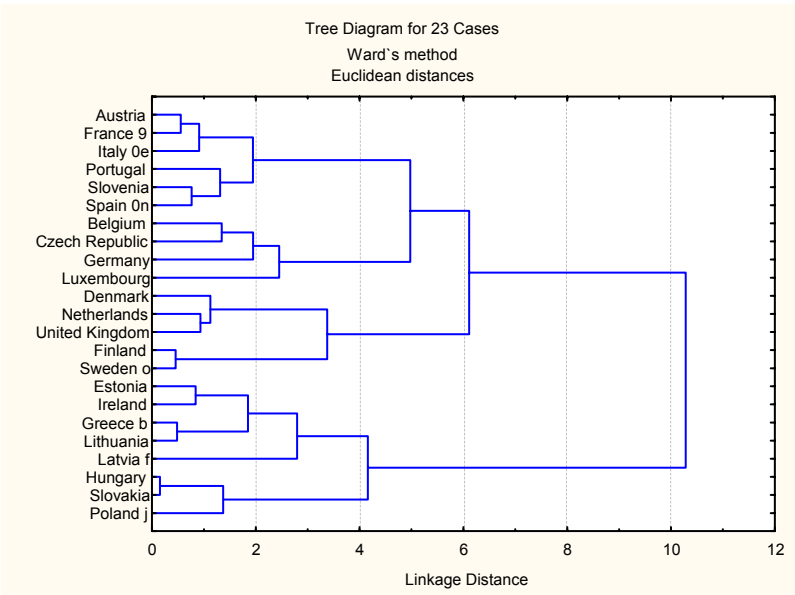

Figure 4: The results of classification by Ward's method.

The data were taken from the sources: EUROSTAT [4] and UNECO [5].

Cluster analysis results for the training year 2000 method is presented in Fig 4. Eventually, the following 8 countries were selected as cluster H: Austria, Belgium, France, Germany, Italy, Portugal, Slovenia, and Spain. Cluster L has 6 countries: Estonia, Latvia, Lithuania, Poland, Greece, Hungary and Slovakia.

The coordinates of cluster centroids and average values of parameters for the training year 2000 are presented in Table 3.

The Fisher vector is presented in Table 4. It does not need correction since all coordinate signs are "correct" from point of view of their influence on the PMI.

The final formula for the PMI computed by eqn. (3) is as follows:

$$PMI = -1.908*KILLED + 0.043*CARS + 0.0293*WAYS - 12.99.$$

4 Logistic Performance Index (LPI)

We applied DCA to the analysis of the development of country's transportation logistic system. This index termed as Logistic Performance Index (LPI) has been suggested by the World Bank (WB). The expert working group of this bank carried out a survey in which took part about 800 professionals in logistics all over the world, operators and agents of the largest logistics companies [6].

Table 3: The coordinates of cluster centroids and the common centroid for the training sample.

Variable	KILLED	CARS	WAYS
Cluster H	0.12	486.96	79.03
Cluster L	0.17	274.86	51.37
Common centroid	0.13	400.65	74.42

Table 4: The Fisher vector.

Variable	KILLED	CARS	WAYS
$\mathbf{f}^*$	-0.039	2.167	0.55

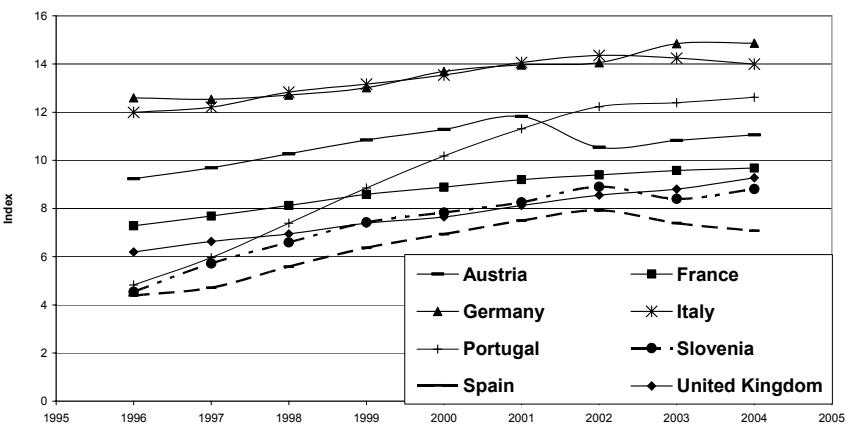

Figure 5: Time series of PMI values for countries of cluster H.

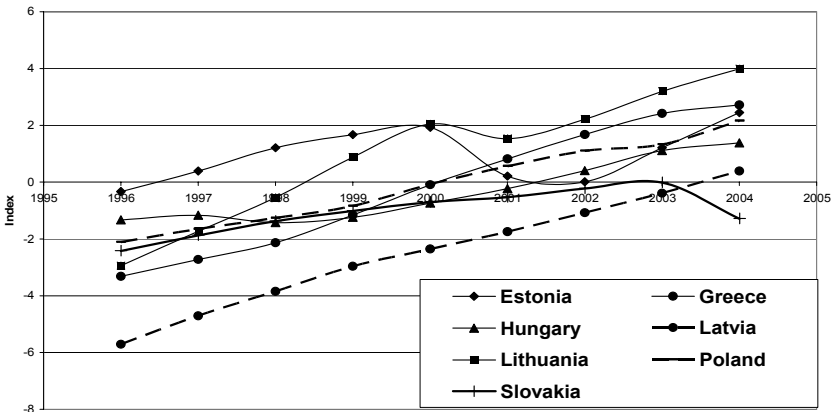

Figure 6: Time series of PMI values for countries of cluster L.

Each expert who took part in the survey was asked to evaluate the situation in the following seven logistics related areas, for the countries with which his company have had intensive business relations.

- Efficiency of the clearance process by customs and other border agency [Customs];
- Quality of transport and information technology infrastructure for logistics [Infrastr];
- Easy and affordability of arranging international shipments [Internat];
- Competence of the local logistic industry [Logist];
- Ability to track and trace of international shipments [Track];
- Domestics logistics costs [Domestics];
- Timeliness of shipments in reaching destination [Timelin].

From the data presented by each respondent from a particular country, eight counties with most intensive and profitable transportation/supply connections were chosen. Logistics development has been evaluated from 1 to 5, where 1 corresponds to the lowest and 5 – to the highest (best) logistics development level.

We compared the results obtained by our DCA with the results given by WB. The latter were presented in the source available [6] only for the year 2007. We, therefore acted as if this is our training sample

The data given by the WB in [6] present seven parameter values for EU-23 countries (Cyprus and Malta were missing). The *domestic logistics costs* were excluded from further analysis as nonsignificant in [6] and also in our analysis.

The cluster H contains 9 countries: Austria, Belgium, France, Germany, Greece, Italy, Netherlands, Spain, and Sweden. The cluster L has 6 countries: Hungary, Estonia, Latvia, Lithuania, Poland, Slovenia and Slovakia.

Table 5: The coordinates of cluster centroids and the common centroid for year 2007.

Variable	Customs	Infrastr	Internat	Logist	Track	Timelin
Common centroid	3.32	3.45	3.41	3.51	3.55	3.95
Cluster H	3.79	3.99	3.77	3.99	4.02	4.27
Cluster L	2.77	2.81	3.06	2.98	2.96	3.53

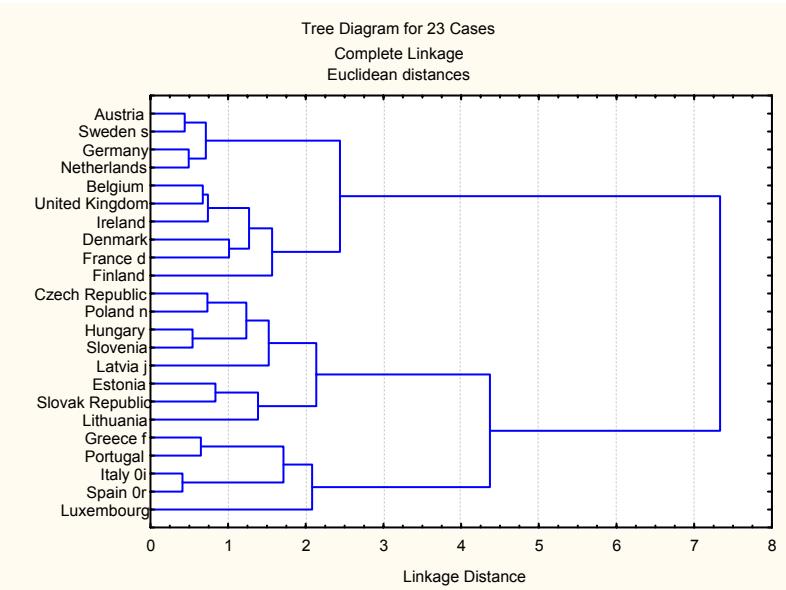

Figure 7: The results of classification by complete linkage method.

Table 6: The uncorrected and corrected Fisher vectors.

Variable	Customs	Infrastr	Internat	Logist	Track	Timelin	D^2
f	1.64	-1.01	0.26	2.05	1.45	-0.40	79.50
$\mathbf{f}^*$	1.79	0.01	0.81	2.06	1.68	0.43	21.90

The final formula for the **LPI** computed by eqn. (3) is as follows:

LPI = 2.604* CUSTOMS + 0.0157* INFRASTR + 1.517* INTERNAT +
+3.015* LOGIST + 2.394* TRACK + 0.848* TIMELIN – 30.97

The report [6] mentions that the LPI was originally obtained using the method of principal components. Table 7 demonstrates that the LPI obtained by our DCA is practically identical to that derived by the WB. For comparison of the DCA and WB, the DCA data were linearly transformed to the scale [1, 5]. Out of

23 counties, 13 have identical LPI values and only for one country the ranks differ by two units. The correlation between the LPI's is 0.999.

The approach offered in [6] works good if there is a significant correlation between the initial parameters. Contrary to the approach in [6], the DCA has no such limitation.

Table 7: The values of LPI on the basis of two algorithms: WB and DCA.

COUNTRY	Lithuania	Slovakia	Estonia	Latvia	Poland	Czech Republic	Slovenia	Hungary	France
LPI_WB	2.78	2.92	2.95	3.02	3.04	3.13	3.14	3.15	3.76
LPI_DCA	2.8	2.95	2.95	3.02	3.08	3.14	3.07	3.12	3.75
COUNTRY	Finland	Denmark	Belgium	Ireland	UK	Austria	Sweden	Germany	Netherl.
LPI_WB	3.82	3.86	3.89	3.91	3.99	4.06	4.08	4.1	4.18
LPI_DCA	3.85	3.88	3.87	3.94	3.99	4.04	4.06	4.1	4.16

5 Conclusions

We demonstrated how to obtain scalar indices characterizing three different areas, health, transportation and logistics, by using the suggested DCA. The last example compares the Logistic Performance Index (LPI) obtained by our method with the LPI obtained by using the method of principal components (MPC) [6]. The MPC works well because there is a significant correlation between the original parameters. In the case when such correlation is no significant, the MPC will not produce dimension reduction and will be therefore no efficient. Contrary to that, the DCA does not impose such limitation, and this makes the DCA in some sense more universal.

References

[1] I. Gertsbakh, I. Yatskiv. Dynamic Classification: Economic Welfare Growth in EU During 1995-2004. Proceedings of International Conference "Data Mining 2006", 11-13 July 2006, Prague. WIT Press. 2006, p.53-62

[2] Johnson, R.A., & Wichern, D.W., *Applied Multivariate Statistical Analysis*, 4th ed, Prentice Hall: New York, 1999.

[3] http://www.eurofound.europa.eu/index.htm

[4] Statistical Databases *http://epp.eurostat.ec.europa.eu*

[5] UNECO *http://w3.unece.org/*

[6] J.-V. Arvis, M.-A.MUstra, J.Panzer, L.Ojala, T. Naula. Connecting to compete: Trade Logistics in the Global Economics. The Logistics Performance Index and Its Indicators. The World Bank. 2007.