MEASURING DISSIMILARITY IN LEARNING CAPITAL USING TIME TRAJECTORIES

Paolo Mariani, Mauro Mussini

Department of Economics, Quantitative Methods and Management, University Milano Bicocca, Milan, Italy

Emma Zavarrone[1]

Department of Consumption, Behaviour and Corporate Communication "Giampaolo Fabris", IULM University, Milan, Italy

***Abstract**. This article focuses on measuring dissimilarities in learning capital accumulation for university students across their academic path. We use time trajectories of learning capital accumulation to detect which phases of students' academic path are more fruitful in terms of knowledge accumulation. This is carried out using an approach consisting of two steps: first, we synthesise the information on learning capital accumulation by using an indicator of the knowledge accumulated by each individual; second, we handle three-way data of the type "individuals × variables × times" by rearranging their structure in a way that allows to perform a joint analysis of them. We apply the time trajectories analysis to a cohort of Italian university students enrolled in an Italian faculty of Economics in 2002-2003.*

***Keywords**: Time trajectories, Learning capital, Graduates.*

1. INTRODUCTION

The concept of human capital is debated from various points of view: macro or microeconomics, knowledge or experience, formal or informal. One of them refers to knowledge accumulation through the attendance of university study programmes, say academic learning capital (hereafter, *alc*) (Civardi and Zavarrone, 2006, 2012; Fabbris *et al*., 2011). Learning capital is related to themes that play a central role in Horizon 2020 (Horizon 2020), such as wellbeing and social inclusion, since investments in education can generate both economic returns and job opportunities.

[1] Paolo Mariani, email: paolo.mariani@unimib.it;
Mauro Mussini, email: mauro.mussini1@unimib.it;
Emma Zavarrone, email: emma.zavarrone@iulm.it.

Moreover, learning capital also has a wide range of non-economic benefits, since it tends to improve health (itself a form of human capital), and to lower risks of criminal activity (Lochner and Moretti, 2004; Lleras-Muney, 2005). These aspects lead to optimize the investments in learning capital and to make the education system more efficient and effective through the reduction of early drop out and the adjustment of educational offer to labour market needs. Within this context, the analysis of learning capital accumulation process is crucial since it allows to assess whether academic educational objectives are fulfilled by minimising both dropout and dilation of the time required to get a degree. In this article, we focus attention on the effectiveness of academic study programmes by linking the measurement and representation of the dissimilarities in the *alc* accumulation with the lengthening of the time for degree completion.

This article aims at highlighting the differences in *alc* accumulation by using time trajectories (Carlier, 1986; D'Urso and Vichi, 1998; D'Urso, 2000; Coppi and D'Urso, 2001) which illustrate the accumulation process. The *alc* accumulation is measured by using a synthetic indicator (Civardi and Zavarrone, 2006) that provides information on knowledge accumulated by each individual at different points in time. In our case, the observed time series is not long enough to fulfil the requirements for the application of multi-way techniques (Rizzi and Vichi, 1995; Kroonenberg, 2007). To overcome this issue, we rearrange the data in a way which allows to perform a joint analysis of them. We then follow an *analyse conjointe de tableaux quantitatifs* procedure (Escofier and Pagés, 1994; Bolasco, 1999), and obtain a principal component analysis (Kroonenberg and De Leeuw, 1980) where the *alc* is collapsed in the first principal component. In addition, we show that our approach is well suited to synthesise the dissimilarities among the *alc* accumulation by study programme. This feature may be of interest for study educational planners since they can obtain information about the phases of the various study programmes where the *alc* accumulation is problematic.

The article is organised as follows. Section 2 introduces the methodology to perform our analysis. In Section 3, we study the differences in *alc* accumulation for a cohort of students enrolled in an Italian faculty of Economics in the academic year 2002-2003 and graduated up to year 2008. Section 4 concludes.

2. THREE-WAY ANALYSIS

The multi-way technique is a method to analyse complex data structure, such as the situation where observations are collected on a number of variables at several time points (Kroonenberg, 1993). The multi-way technique rationale is based on a particular decomposition of the total variability: the variability within groups on one side (where the groups are the different waves) and the variability between groups (due to time). This second part is modelled through a linear regression where the different times represent the observations of the covariate. It follows that a time series of few observations is inadequate as it does not allow the regression model to be estimated. To overcome this issue, the three-way data can be handled as a tall combination mode matrix of dimensions (individuals × occasions) × variables; in other words, we rearrange the data in order to have as many rows referred to an individual as the occasions experienced by that individual, obtaining the so-called tall matrix. Applying a principal component analysis on rearranged data matrix, this structure can be used to explain data and obtain a time trajectory for each individual across the variables. This allows one to study a phenomenon in both a structural way – that is by fixing basic relations among the variables we are interested in – and a dynamic way – in order to capture change and development of the variables in accordance with the occasions we are referring to. In matrix notation, we analyse three-way data matrices $\mathbf{X}_{ijk}$ (individuals × variables × occasions), with i=1,…, I individuals, j =1,…, J variables, k=1,…, K occasions.

The explorative analysis can be applied if:

- Three-way data matrices have the same variables and individuals for each occasion;
- Three-way data matrices have the same individuals but different variables for one or more occasions (all variables depend on a specific k occasion);
- Three-way data matrices have the same variables but different individuals for one or more occasions (all individuals depend on a specific k occasion).

In our empirical analysis, we face a situation which is conforming with the first one, so we can operate on data matrix $\mathbf{X}_{ijk}$ directly.

For a given number of occasions, k, the information on the variable j for the individual i across time can be represented by the time trajectory of the variable j for the individual i (D'Urso and Vichi, 1998). More specifically, $T_{ik}(x_j)$ denotes the sequence of the values of the variable j observed for the individual i across the time (i.e., the k occasions), that is the time trajectory for the i-th individual.

Time trajectories can be compared by measuring the dissimilarities in their respective positions in the data space. We refer to one of the various measures of dissimilarity proposed by D'Urso and Vichi (1998), which compares the positions of the trajectories of the individuals i and l as follows

$$d_{il}^2 = \sum_{k=1}^{K} \sum_{j=1}^{J} \left(x_{ijk} - x_{ljk} \right)^2 , \qquad (1)$$

where *K* and *J* are the numbers of occasions and variables, respectively. As noted in D'Urso and Vichi (1998), the measure in (1) gauges cross-sectional dissimilarity since the instantaneous position of the *i*-th trajectory is compared to the position of *l*-th trajectory at the same instant.

3. APPLICATION

Time trajectories are applied to capture dissimilarities in the path of *alc* accumulation for the students enrolled in eight different university study programmes in the Faculty of Economics of the Bicocca University in Milan. We start by a brief but necessary description of data, since they are of a local character and may not be known.

3.1 DATA DESCRIPTION

The data come from the Bicocca University archives and regard the cohort of students (N=334) who enrolled in the Faculty of Economics of the University in the academic year 2002-2003 and obtained a first-level degree within six years (i.e., from 2003 to 2008), that is a three-year longer period than the legal duration (3 years) of the study programme, in order to observe also the trajectories of the students who did not pass all their exams within the due time. Tab. 1 shows the distributions of graduates by study programme.

Table 1: Graduates by study programmes, 2005-2008, Bicocca University

Study programmes	n	%
International business (IB)	11	3.3
Economics and social sciences (ESS)	14	4.2
Economics for banks, insurance and financial institutions (EBIF)	53	15.9
Economics and business administration (EBA)	79	23.6
Economics and business (EB)	136	40.7
Economics for tourism (ET)	34	10.2
Economics and management of public administration (EMPA)	6	1.8
Economics, statistics and information for business(ESIB)[a]	1	0.3
Total	334	100.00

[a] Since the study programmes has only one graduate, the study programmes is not mentioned in the following tables even if the graduate is included in the analysis.

Tab. 2 reports the distribution of graduates by study programme and by year of graduation. We note that a large proportion of students (45.51 per cent) graduated one year after the programme's legal duration, whereas only 22.75 per cent of students graduated within the programme's legal duration. From Tab. 2, we notice that Economics and Business and Economics and Management of Public Administration are the study programmes with the higher percentages of students who graduated on time, however the latter study programme has only six students who enrolled in 2002-2003 and achieved their degree (see Tab. 1).

Table 2: Percentages of graduates by study programmes and year of graduation, 2005-2008, Bicocca University

Study	**Year of graduation**				
programme	**2005**	**2006**	**2007**	**2008**	**Total**
IB	18.18	45.45	36.36	0.00	100
ESS	21.43	78.57	0.00	0.00	100
EBIF	9.43	49.06	32.08	9.43	100
EBA	24.05	54.43	17.72	3.80	100
EB	29.41	33.09	30.15	7.35	100
ET	14.71	52.94	20.59	11.76	100
EMPA	33.33	50.00	16.67	0.00	100
Total	22.75	45.51	25.15	6.59	100

3.2 EMPIRICAL STRATEGY

In our application, starting from the data referred to single graduates, we extend the analysis to the *alc* accumulation observed for each study programme. The analysis we perform can be synthesised as follows:

- Information on *alc* accumulation per graduate is restructured in a tall matrix;
- Once graduates have been partitioned by study programme and year of graduation, we calculate the mean *alc* for every year of the legally required time, therefore, each mean *alc* is assumed as the representative *alc* of the respective group of graduates defined by study programme and year of graduation;
- Principal component analysis is carried on the tall matrix and the first component is then used to synthesise the information on mean *alc*.
- The time trajectories of the first principal component of the mean *alc* are compared among the various groups of graduates, and the cross-sectional dissimilarities of the *alc* accumulation are computed by study programme and year of graduation.

We exploit the information collected at Students secretariat on registration number, exam name, mark (*m*), credits (*c*) and exam date for each graduate.[2] We build a dataset in which the information is organised on a 12 quarters based time to have a matrix of 334 rows (the graduates) and 15 columns: one for the registration number of graduates, the year of degree, the study programme and the 12 *alc* values. Each cell displays the amount of learning capital held by each graduate per quarter, where academic learning capital (alc_{ik}) for graduate *i* at quarter *k* is measured as

$$alc_{ik} = \sum_{h=1}^{p_{ik}} m_{ikh} c_{ikh} \,, \tag{2}$$

where p_{ik} denotes the number of exams passed by graduate *i* at quarter *k*. Then, the alc_{ik} are aggregated by each of the three years that constitute the programme's legal duration.[3] Table 3 shows some descriptive statistics for the *alc* accumulated over the programme's legal duration by quarter and year.

Table 3: descriptive statistics for the *alc* accumulated over the programme's legal duration by quarter and year

year	quarter	mean	max	standard deviation
2003	1st	270.70	675.00	173.68
	2nd	339.24	1039.50	203.61
	3rd	332.14	1005.00	213.76
	4th	61.27	774.00	95.94
2004	1st	438.26	1138.50	225.40
	2nd	388.36	1116.00	205.34
	3rd	315.93	990.00	190.36
	4th	31.53	1635.00	157.35
2005	1st	380.00	1105.50	220.01
	2nd	529.24	1423.50	276.52
	3rd	410.42	1453.50	252.87
	4th	24.60	1651.50	148.59

[2] The initial database consists of $p_{i.}$+ 1 records for graduate *i*, where $p_{i.}$ is the total number of exams passed by graduate *i* and 1 denotes the final oral exam. Each of $p_{i.}$ records is composed of registration number, exam name, mark, credits and date; the final oral exam record is composed of registration number, graduation mark, credits and date.

[3] The maximum value of *alc* accumulated in the legally required time is 5400, and it is obtained by the product between the maximum mark (30) and the number of credits that a student must achieve to get a degree (180).

Initially, we have a three-way matrix which collects the *alc* accumulated within the legally required time (i.e., the three years from 2003 to 2005) and the information on the year of degree (i.e., the four years from 2005 to 2008) for the 334 graduates, therefore, we have a $334 \times 3 \times 4$ matrix. The three-way matrix is then restructured by study programme and year of graduation: for each of the eight study programmes, graduates are partitioned into four groups by year of degree; then, the mean *alc* of each group is calculated for every year of the legally required time (Wedel and Kamakura, 2000).We achieve a 32×3 tall matrix where the 32 rows are the groups of graduates by study programme and year of graduation, whereas the 3 columns report the mean *alc* per year of legally required time.

3.3 MAIN RESULTS

Tab. 4 shows the amount of mean *alc* accumulated over the three-year legal duration (i.e., 2003-2005) by year of graduation. We can see that the *alc* accumulation in the 2003-2005 period decreases when the year of degree increases. We stress that programmes on: Economics and Management of Public Administration; International Business; Economics and Social Sciences are less important because the number of enrolled students is small (see Tab. 1). Therefore, even though we report also the results for these study programmes, our discussion only refers to the remaining study programmes. Among them, Economics for Banks, Insurance and Financial institutions is the study programme showing the best performance in terms of accumulation of *alc*.

Table 4: Mean *alc* accumulated over the programme's legal duration (from 2003 to 2005) by study programme and year of graduation

Study programme	2005	2006	2007	2008
IB	4681.50	3276.30	2493.75	.
ESS	3878.00	3784.50	.	.
EBIF	4640.40	3750.69	3032.03	2658.90
EBA	4508.21	3585.70	2748.75	2041.00
EB	4491.83	3591.77	2622.84	2324.25
ET	4334.30	3685.67	2878.43	2359.13
EMPA	4908.00	4151.00	1944.00	.
Total	4491.75	3648.95	2619.97	2345.82

From Tab. 5, we observe that the first principal component (PCA1) explains the 83.0% of variance, therefore, it retains most of the information on the mean *alc* accumulated over the 2003-2005 period.

Table 5: Principal component analysis

Component	Initial eigenvalues			KMO test	Bartlett's test		
	Total	% Var.	% Cum. Var.		Chi-squared	df	p-value
1	2.50	83.00	83.00	0.685	47.624	3	<0.001
2	0.38	13.00	96.00				
3	0.12	4.00	100				

We extract the PCA1 of the mean *alc* accumulation over the three years of the programme's legal duration. Tab. 6 shows the component loadings of the PCA1.

Table 6: Component loadings of the PCA1

mean *alc*	PCA1
2003	0.920
2004	0.953
2005	0.862

Tab. 7 shows the values concerning the scores of the PCA1 for each study programme by year of graduation, which are calculated using the tall matrix as defined above.

Table 7: PCA1 by study programme and year of degree

Study programme	Year of degree			
	2005	2006	2007	2008
IB	1.41	-0.17	-1.07	–
ESS	0.50	0.40	–	–
EBIF	1.36	0.37	-0.45	-0.87
EBA	1.22	0.19	-0.76	-1.56
EB	1.20	0.18	-0.91	-1.24
ET	1.00	0.29	-0.61	-1.21
EMPA	1.65	0.82	-1.66	–

Time trajectories in Fig. 1 are obtained by linking the points defined by the PCA1 scores with respect to the years of graduation (Summa *et al*., 2011). Only the time trajectories of the four study programmes with students who graduated until 2008 are drawn: EBIF, EBA, EB and ET. As expected, the PCA1 score decreases as the year of graduation increases, since graduates who accumulated more *alc* during the programme's legal duration are more likely to get a degree at the end of 2005. Even though the trends are similar, the trajectories indicate different typologies of accumulation. ET shows the smallest range between the PCA1 score of the graduates who got a degree in 2005 and that of the graduates who got a degree in 2008. Students who graduated in EBA in 2008 are those with the lowest value of PCA1 score in the due time. Students who enrolled in EBIF accumulated more *alc* than those enrolled in the remaining study programmes regardless the year of graduation, since the respective trajectory lies above the others.

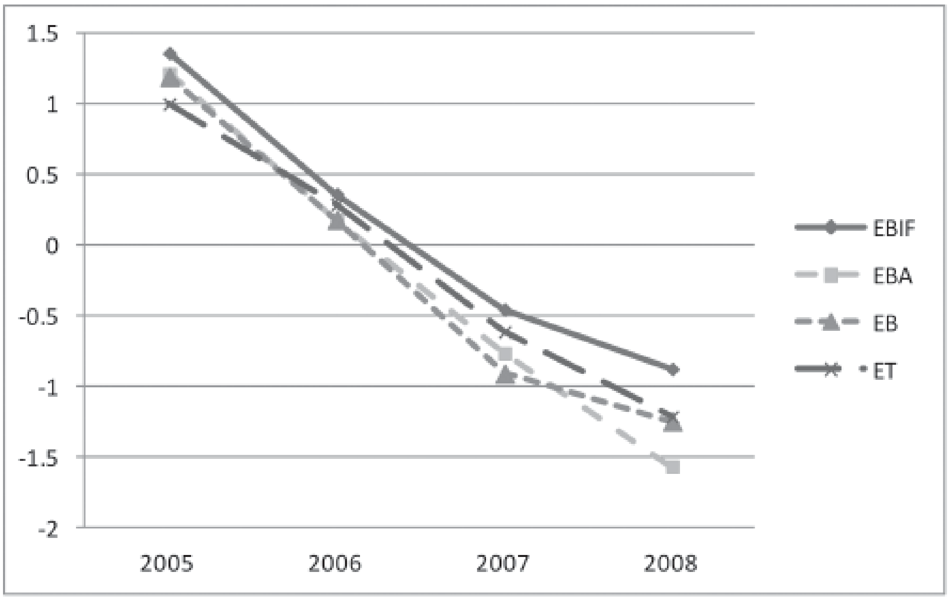

Figure 1: Trajectories of the PCA1 score of *alc* accumulation by study programme

Tab. 8 reports the cross-sectional dissimilarities calculated for each pair of trajectories shown in Fig. 1. This measure synthesises the distance between two trajectories: the larger the dissimilarity value, the larger the gap in *alc* accumulation among the graduates of the two programmes under consideration. The largest dissimilarity is between the trajectory of EBIF and that of EBA, whereas the smallest dissimilarity is between the trajectory of EB and the one of EBA. The values of dissimilarity between the trajectory of ET and the remaining trajectories seem to suggest that the path of *alc* accumulation of the graduates in ET is intermediate.

Table 8: Dissimilarities between the time trajectories of the PCA1 score of *alc* accumulation by study programme

	EBIF	EBA	EB	ET
EBIF	–	0.626	0.405	0.274
EBA	0.626	–	0.123	0.204
EB	0.405	0.123	–	0.138
ET	0.274	0.204	0.138	–

4. CONCLUDING REMARKS

The paper presents an approach to represent the *alc* accumulation by using time trajectories. Time trajectories provide a graphic representation of the process of *alc* accumulation from a longitudinal standpoint, emphasising the phases in which there are the main contributions in terms of *alc* accumulation. Moreover, the fore mentioned methodology can be applied to very general contexts since it enables the fitting to linear as well as non-linear functions of *alc* accumulation.

In our empirical analysis, the attention is paid to the different paths of *alc* accumulation of the graduates of various study programmes of an Italian faculty of Economics. The analysis of the time trajectories shows that the *alc* accumulation process is influenced by study programme and year of graduation. Generally speaking, the information provided by the time trajectories analysis may be useful from two points of view: first, it can be used for partitioning the accumulation of learning capital across the time span for getting a degree; second, the investigation of the *alc* accumulation path may be useful to plan the study programmes and to provide the academic governance with comparative information for improving the effectiveness of their study programmes.

Future research will be devoted to extending the analysis by including the information describing students before they enrol at university. This gives the academic governance the opportunity to take into account the knowledge starting point from which each student begins his university path.

ACKNOWLEDGEMENTS

The authors thank the editor and two anonymous reviewers for their constructive comments on earlier versions of the manuscript.

REFERENCES

Bolasco, S. (1999). *Analisi multidimensionale dei dati. Metodi, strategie e criteri d'interpretazione*, Carocci, Roma: 256-257.

Carlier, A. (1986). Factor analysis of evolution and cluster methods on trajectories. *COMPSTAT86*. Physica-Verlag, Heidelberg: 140-145.

Civardi, M. and Zavarrone, E. (2006). Estimating university human capital through education evaluation. In: Fabbris L. (Ed.) *Effectiveness of University Education in Italy*. Physica-Verlag, Heidelberg: 369-380.

Civardi, M. and Zavarrone, E. (2012). Capitale umano e traiettorie di crescita individuali. In: Tronti L. (Ed.), *Capitale umano. Definizioni e misurazioni*. Cedam-Kluwer, Milano: 113-128.

Coppi, R. and D'Urso, P. (2001). The geometric approach to the comparison of multivariate time trajectories. In: Borra S., Rocci R., Vichi M. and Schader M. (Eds.) *Advances in Data Science and Classification*. Springer-Verlag, Heidelberg: 93-100.

D'Urso, P. (2000). Dissimilarity measures for time trajectories, *Journal of the Italian Statistical Society*, (9): 53-83.

D'Urso, P. and Vichi, M. (1998). Dissimilarities between trajectories of a three-way longitudinal data set. In: Rizzi A., Vichi M. and Bock H.H. (Eds.) *Advances in Data Science and Classification*. Springer-Verlag, Heidelberg: 585-592.

Escofier, B. and Pages, J. (1994). Multiple factor analysis (AFMULT package), *Computational Statistics & Data Analysis*, (18): 121-140.

Fabbris, L., Boccuzzo, G., Martini, M.C. and Scioni, M. (2011). A participative process for the definition of a human capital indicator. In: Ingrassia S., Rocci R. and Vichi M. (Eds.) *New Perspective in Statistical Modeling and Data Analysis*, Springer-Verlag, Heidelberg: 39-47.

Horizon 2020. *The EU Framework Programme for Research and Innovation*. Available at http://ec.europa.eu/research/horizon2020/index_en.cfm

Kroonenberg, P.M. (1993). *Principal Component Analysis of three-mode data*. DSWO Press, Leiden.

Kroonenberg, P.M. (2007). *Applied Multiway Data Analysis*. Wiley, New York.

Kroonenberg, P.M. and De Leeuw, J. (1980). Principal Component Analysis of three-mode data by means of alternating least squares algorithm, *Psychometrika*, (65): 69-97.

Lleras-Muney, A. (2005). The relationship between education and adult mortality in the United States, *The Review of Economic Studies*, (72): 189-221.

Lochner, L. and Moretti, E. (2004). The effect of education on crime: Evidence from prison inmates, arrests, and self-reports, *American Economic Review*, (94): 155-189.

Rizzi, A. and Vichi, M. (1995). Representation, synthesis, variability and data preprocessing of three way data set, *Computational Statistics & Data Analysis*, (19): 203-222.

Summa, M.G., Goldfarb, B. and Vichi, M. (2011). Clustering trajectories of a three-way longitudinal data set. In: Summa M.G., Bottou L., Goldfarb B., Murtagh F., Pardoux C. and Touati, M. (Eds.) *Statistical Learning and Data Science*. Chapman and Hall/CRC.

Wedel, M. and Kamakura, W. (2000). *Market Segmentation*. Kluwer Academic Publishers, Boston.