A distribution curves comparison approach to analyze the university moving students performance

L'analisi delle preformance degli studenti universitari in mobilità, secondo un approccio basato sul confronto tra curve di distribuzione

Giovanni Boscaino, Giada Adelfio, Gianluca Sottile

Abstract Nowadays in Italy we observe a one-directional migration flow of university students, typically from the South to the North. It represents the new millennium migration flow: people migrate looking for better job opportunities already during their educational path, believing that northern universities may provide more opportunities for being more successful. This paper aims to study the performance of those Sicilian students that move to the northern universities, to take the second level degree, in comparison with those remain in Sicily. We want to test the empirical evidence that shows a similar performance between the two groups of students. We use different measures of performance and follow a new methodology based on the comparison among the distribution curves. Results seems to confirm our idea, highlighting some difference.

Abstract *Da qualche anno anche l'Italia sta assistendo a un flusso migratorio di studenti, che da un povero Sud si dirige verso un più ricco Nord. Se un tempo la migrazione avveniva nel momento di cercare lavoro, adesso questa è anticipata da studenti che ritengono di avere meggiore successo se conseguono il titolo al Nord. L'obiettivo di questo studio è verificare l'ipotesi secondo la quale gli studenti che restano per iscriversi alla Laurea Magistrale hanno un percorso simile rispetto a quelli che si iscrivono a un percorso magistrale del Nord. Considerate diverse misure di performance, e svolto i confronti attraverso una nuova procedura di raffronto tra curve di distribuzione, i risultati non mostrano sostanziali differenze.*

Key words: performance, student migration, distribution comparison

---

Giovanni Boscaino, Giada Adelfio, Gianluca Sottile
Dipartimento di Scienze Economiche, Aziendali e Statistiche, viale delle Scienze, Edificio 13, Palermo (Italy) e-mail: giovanni.boscaino; giada.adelfio; gianluca.sottile @unipa.it

1 Introduction

The intra-European students' mobility is one of the targets of the Bologna Process to promote EU integration and foster a higher quality educational system. As known, the individual student mobility is a great opportunity to gain experience, to increase the knowledge of other cultures, to enhance human capital formation, networks, and relational abilities. Literature offers a lot of papers on the international student mobility [7] [4] [11] and its determinants, and they underline the influence of the socio-economic and cultural conditions of both the areas of origin and the destination. Actually, other factors could influence the wish of moving abroad, such as the quality of the receiving universities and the migration costs.

Studies on internal student mobility are rather scant on UK [8] and on Netherlands [13], where domestic student mobility flows are almost entirely unidirectional – from South to North – mirroring the internal economic migration. This is the case of Italy too: in fact, the migration flows from the South to the North of Italy, historically due to the different job opportunities, has changed in students migration flows. Probably, this is due to the expectation of students who think they will be more successful if they graduate in a more economically developed area [3].

Literature offers several studies about student performance, mainly devoted to find its determinants, and often based on different measures [16] [1] [5]. On the other hand, it is not so common to find papers about the performance of the university moving students [12] [3] [6]. Our research focuses on this new point of interest, and in particular on students that enroll in a university second level degree at institutions outside the students region of origin.

Comparison is made using a new method for curve clustering [15], since we compare the distribution curves (marginal and conditional to some given socio-demographic characteristics) of students, both in terms of performance and time to the degree.

2 The new method for clustering of curves

The clustering of distribution curves here performed is based on a new method to find similarities of curves in a quantile regression coefficient modeling framework, also multivariate, in which the effect of covariates on a response variable is represented by curves in the space of percentiles [9] [10]. In particular, let y be the response of a dependence model problem, in [15] the authors first estimate the regression coefficient functions $\beta_1(p \mid \boldsymbol{\theta}), \ldots, \beta_q(p \mid \boldsymbol{\theta})$, namely effects curves, and then assess if these q curves, that describe the effects of each covariate on the response, can be clustered based on similarities of effects, as a variable selection procedure.

The proposed clustering approach is based on a new measure of dissimilarity that uses both the shape of a curve and its distance with respect to other curves:

- the *shape* of a curve is evaluated using its second derivative. Moreover, two different curves are similar in shape if, at any given point, the signs of the second derivatives are concordant;
- the *distance* between two curves is evaluated as their with respect to other curves. Two curves are said close if their distance at any given point is lower than a fixed value.

Let i and i' be two different curves, $\boldsymbol{p} \in (0,1)$ the vector of percentiles. Then

$$d^{ii'}_{\text{shape}}(\boldsymbol{p}) = I(\text{sign}(\beta''_i(\boldsymbol{p} \mid \boldsymbol{\theta})) \times \text{sign}(\beta''_{i'}(\boldsymbol{p} \mid \boldsymbol{\theta})) = \mathbf{1})$$
$$d^{ii'}_{\text{distance}}(\boldsymbol{p}) = I(|\beta_i(\boldsymbol{p} \mid \boldsymbol{\theta}) - \beta_{i'}(\boldsymbol{p} \mid \boldsymbol{\theta})| \leq f(\alpha, \text{dist}(\boldsymbol{p}))),$$

where $f(\cdot,\cdot)$ is a cut-off function, that depends on a probability value α, and on dist($\boldsymbol{p}$), the vector of the distances between two curves across all percentiles. Finally, the proposed dissimilarity measure between two curves is defined as

$$d(i,i') = 1 - \int_0^1 \left[d^{ii'}_{\text{shape}}(p) \cdot d^{ii'}_{\text{distance}}(p) \right] \mathrm{d}p. \tag{1}$$

In [15], the new measure is used, to account for their concordance at each point, and any hierarchical clustering method can be applied. The method has been developed in the R package `clustEff` [14]. The proposed approach is very flexible and can be generalized to different contexts.

3 Data, Methodology and results

Thanks to a partnership with the Italian Ministry of Education and Research (MIUR) and four Italian Universities (University of Cagliari, University of Palermo, University of Siena, University of Torino), we refer to the whole dataset of the students enrolled in 2008 at any Degree Course offered by any Italian University, and followed for 8 years. In such a way, it is possible to follow the whole career of the students and their eventual mobility along the Italian area.

The dataset is very large, more than 100,000 records, where the record is the student. Many variables are available: gender, age, nationality, residence, diploma mark, high school diploma type, university and degree course attended at each year of observation, residence, average mark and total credits gained at each year.

In particular, the statistical unit is the student that graduates at the first level degree and that enrol in a second level degree course. The comparison

regards three groups of students: i) the group of the Southern graduates that enrol in one of the universities in the South of Italy; ii) the group of the Southern graduates that enrol in the Northern university; iii) the group of resident in the North of Italy and still studying in the North.

Our specific question is "is there any difference in performance between moving and unmoving students?". We refer to the Southern students as a first step analysis of the consistent migration flow from the South to the North of Italy (fig. 2). The chords represent the migration flows, linking the leaving and destination regions. As thicker is the chord, as more significant is the migration flow. The Fig. 1 highlights that main attractive universities are in the Centre-North of Italy, in particular in Lombardia, Piemonte, Emilia Romagna and Lazio.

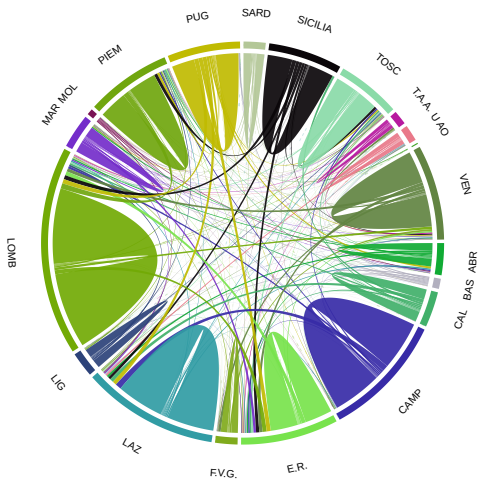

Fig. 1 Chord Diagram: migration student flows within Italy

The performance, in this paper, is measured accounting for two different perspectives: i) marks, ii) gained educational credits. In order to account both for the information of marks and credits, we use a new indicator that combines marks and credits in a unique measure, keeping unchanged the mark-scale [2]. In such a way, there is no need to analyze the two distributions of marks and credits because the new measure weights each mark by the correspondent exam credit.

We report the analysis of the distribution of the transformed mark conditioned to the High School Diploma Type (binary variable "Lyceum", "No-Lyceum"), and our procedure identifies four clusters, together with interesting results, suggesting the necessity of further analysis (fig. 2). Indeed, Northern No-Lyceum students still studying in the North perform worse

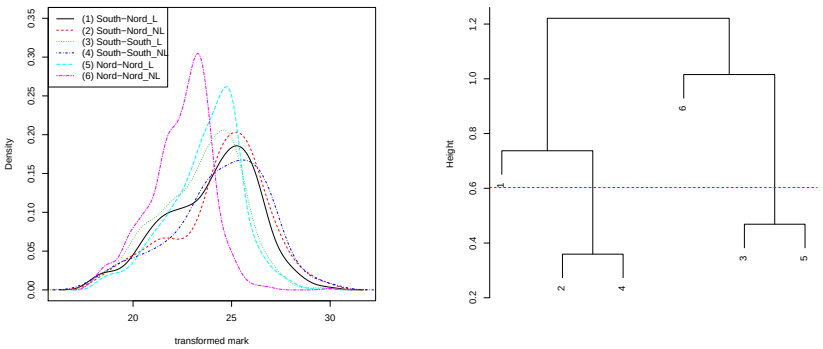

Fig. 2 Output of the proposed algorithm: the distribution curves (on the left) and the corresponding dendrogram (on the right)

than others and stay separate from the other students. Southern and Northern Lyceum student performance is similar (density curves are in the same cluster) when they do not migrate, and stay studying in their macro-area of residence (i.e. South or North, respectively). The third cluster is the one of the Lyceum Southern students migrating in the North, that perform on average a slightly better than the students of the two previous clusters. Finally, No-Lyceum Southern student performance does not change both if they migrate to the North or they continue their studies in the Southern Universities, and they seem to perform better, on average, than the students of the three previous cluster.

We perform the analysis of the distribution of the transformed mark conditioned to the gender too but, although the clustering approach identifies three clusters, results do not show difference in performance.

These results, just provisional and partial, suggest the use of the proposed approach such as an exploratory method for studying this kind of dependence, based on the comparison of distribution functions. Indeed, further analysis (not here reported for the sake of brevity) considering a quantile regression model approach for the student performance, confirms our exploratory analysis results and highlights some differences with respect to other variables, like topic area of the degree course and going into detail of the geographical areas for students of the three groups.

Acknowledgements

This paper has been supported by the national grant of the Italian Ministry of Education University and Research (MIUR) for the PRIN-2015 program “Prot. 20157PRZC4. PI: G. Adelfio".

References

1. Adelfio, G., G. Boscaino, and V. Capursi. A new indicator for higher education student performance. Higher Education 68 (5), 653-668 (2014).
2. Adelfio, G.; Boscaino, G., and Capursi, V. Further considerations on a new indicator for higher education student performance. Proceedings XLVIII Scientific Meeting of the Italian Statistical Society (2016)
3. Attanasio, M.; Enea, M. La mobilitÃă studentesca In Italia: un’analisi dei flussi dal Sud d’Italia verso il Centro-Nord. Proceedings of Popdays 2017. (2017)
4. Beine, M.; Bertoli, S; and Fernandez-Huertas Moraga, J. A practitionersâĂŹ guide to gravity models of migration, The World Economy, vol 6, n- 4, 496-512. Wiley Blackwell. (2016)
5. Birch, E. R. and P. W. Miller. Student outcomes at university in Australia a quantile regression approach. Australian Economic Press 45 (1), 1-17.(2006)
6. Boscaino, G.; Vassallo, P.. La migrazione studentesca dalla laurea triennale alla laurea magistrale. Proceedings of Popdays 2017. (2017)
7. Caruso, R.; de Wit, H.. Determinants of Mobility of Students in Europe. Empirical Evidence for the period 1998-2009. Journal of Studies in International Education, vol 19, n. 3, 265-282. Sage Publishing. (2015)
8. Faggian, A.; McCann, P.; and Sheppard, S.. Human Capital, Higher Education and Graduate Migration: An Analysis of Scottish and Welsh Students. Urban Studies. Volume: 44 issue: 13, 2511-2528 (2007)
9. Frumento, P., Bottai, M., Parametric modeling of quantile regression coefficient functions, Biometrics, 72, 74-84 (2015)
10. Frumento, P. (2017). qrcm: Quantile Regression Coefficients Modeling. R package version 2.0, https://CRAN.R-project.org/package=qrcm.
11. Kahanec, M.; and Kralikova R.. Higher Education Policy and Migration: The Role of International Student Mobility. CESifo DICE Report 9(4):20-27 (2011)
12. Ordine, P.; Rose, G.. StudentsâĂŹ mobility and regional disparities in quality and returns to education in Italy. Giornale degli Economisti e Annali di Economia, vol. 66, n. 2, 149-176 (2007)
13. SÃă, C., Florax, R.J.G.M.; Rietveld, P.. Determinants of the regional demand for higher education in the Netherlands: A gravity model approach. Regional Studies, 38, 375-392 (2004)
14. Sottile, G., Adelfio, G. (2017). clustEff: Clusters of Effects Curves in Quantile Regression Models. R package version 0.1.2 GPL (General Public Licence): https://CRAN.R-project.org/package=clustEff
15. Sottile, G., Adelfio, G. Clusters of effects curves in quantile regression models. Submitted (2018)
16. Van Bragt, C. A. C.; Bakx, A. W. E. A.; Bergen, T. C. M.; and Croon, M. A.. Looking for students personal characteristics predicting study outcome. Higher Education 61, 59-75. (2011)