AN APPLICATION OF CART ALGORITHM TO ADMINISTRATIVE DATA: ANALYSIS OF YOUTH INITIAL EMPLOYMENT TRAJECTORIES

Rocco Ilaria[1,2]

[1] Labour Market Observatory of Veneto Lavoro, Italy,
(e-mail: ilaria.rocco@venetolavoro.it)

[2] Department of Statistical Sciences, Sapienza University of Rome, Italy

ABSTRACT: This work presents an application of the classification and regression tree (CART) algorithm to administrative data on employment in the Veneto region, an area in northern Italy with a strong economy and a dynamic labour market. This data, derived from the national stream of declarations due by employers to notify each activation, termination, extension, or transformation of employment relationships, allows to investigate the occupational condition of young people who enter for the first time in the regional labour market: we classified their initial work trajectories and then we analysed the individual and working features characterising each identified class. In the current socio-economic context, an insight on youth working conditions is crucial to facilitate the successful design of labour and education policies.

KEYWORDS: classification and regression tree, administrative data, youth, labour market.

1 Introduction

Tree-based methods, that find application in many disciplinary fields, from economics (Williams et al. 1987; Keely and Tan 2008; Manasse and Roubini 2009; Galletta 2016; Bilton et al. 2017), to engineering, medicine, biology, and marketing (De'ath and Fabricius 2000; Dacko et al. 2016), are useful statistical techniques for exploring patterns in complicated datasets if assumptions of linear models are somewhat violated (De'ath and Fabricius 2000; Frisman et al. 2008) or if response or explanatory variables present outliers, missing and unbalanced values (Low and Lai 2016). The classification and regression tree (CART) algorithm, introduced by Breiman et al. (1984), is a non-parametric approach without distributional assumptions that allows to handle datasets containing variables of categorical, scale, and ordinal measurement types (Wałęga and Wałęga, 2021).
This work aims to present an application of the CART algorithm to administrative data on employment in the Veneto region, a territory in northern Italy with a strong economy and a particularly dynamic labour market. Using this method we will classify the initial trajectories of young people into the labour market, and then we

will explore the individual and working characteristics associated to each identified class (through multinomial regressions).

2 Data source

The data used come from the database derived from the Labour Information System of Veneto (SILV). It is an administrative archive that collects the stream of declarations ("Compulsory Communications") due by employers to notify the events of activation, termination, extension, or transformation of each employment relationship. This database, that ensures in a timely manner a constant updating of the information, has as reference universe all the subordinate and para-subordinate employments activated by regional enterprises, both public and private; it also allows to monitor the work experiences like the non-curricular internships that are particularly relevant for the young component of the labour supply.
Moreover, this data source offers the possibility to observe the dynamics of the regional labour market since at least 2008, the year of the computerization of the national information system (in the Veneto region this process started in the '90s and the data collected since 2000 onwards can be considered reliable).
The wealth of data analytical details (information is available for single worker and company) opens broad possibilities for exploring in depth the characteristics of young workers and, in a longitudinal perspective, their trajectories in the labour market.

3 Application

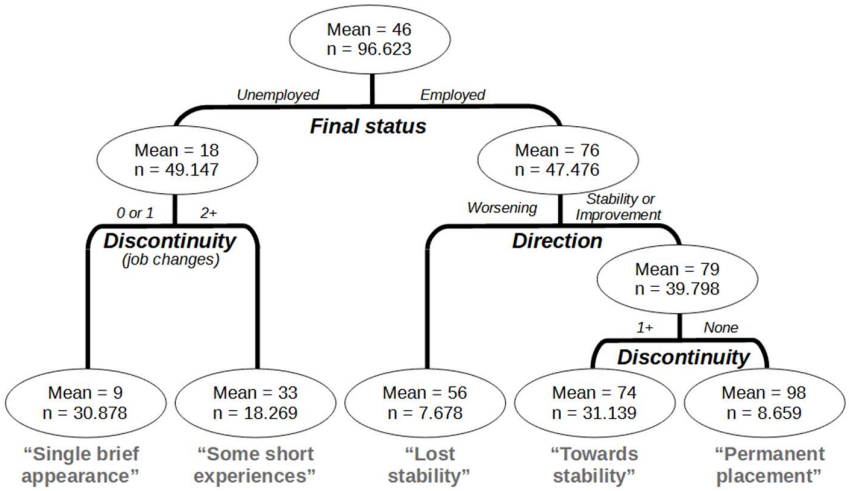
The population of interest includes the young people aged between 15 and 29 years old that were hired for the first time by a firm located in the Veneto region in 2007 (n=97,000). The working histories of these subjects were followed for 12 years after their first entrance into the labour market and five key indicators were selected to describe the main characteristics of their initial trajectories:

- the "initial status", i.e. the type of the first labour contract (stable vs fixed-term);
- the "final status", i.e. the employment condition after a 12-year follow-up, (employed vs unemployed);
- the career "direction", determined by comparing the prevalent contract in the first and in the last trimester (i.e. stability, improvement, or worsening);
- the "discontinuity" of the working paths (i.e. number of job changes);
- the "saturation rate" (i.e. the percentage of days worked in the observed period).

The CART model, built using R software, divides the entire sample (the initial parent node) into smaller, homogeneous groups (child nodes) based on a dependent variable, that in this case is the "saturation rate". The other four key indicators listed above were included in the model as predictor variables. As illustrated in Figure 1, the "final status" is the first variable that best splits data into homogeneous subgroups most

relevant to the outcome of interest. Also the "discontinuity" of the working paths has a crucial role, both among subjects employed at the end of the follow-up period and among the unemployed ones.

Figure 1: Tree plot for the "saturation rate"



The whole population was split into five classes that were named according to their main identifying characteristics as listed below:

- young people that have no employment contract open at the end of the follow-up period were classified in two groups according to the number of jobs they changed (often carrying out low-skilled professions):
  - *"Single brief appearance"* group has the lowest mean "saturation rate"; its members, mainly men and foreigners, were employed for a single short period;
  - *"Some short experiences"* group has a bit higher "saturation rate" and includes young people that were employed for short periods in at least two firms;
- young people that have a contract open at the end of the follow-up period were classified in three groups with an increasing "saturation rate":
  - *"Lost stability"* group is characterised by working careers that start with a stable contract and then move to a fixed-term job; this class shows a high percentage of apprentices and construction workers;
  - *"Towards stability"* group, on the contrary, comprises workers that start with a fixed-term job, in many cases before the age of 20, and then reach the contractual stability, usually changing firm;
  - *"Permanent placement"* includes careers that start with a stable contract which continues for the whole observed period; the subjects in this class, mainly aged

between 25 and 29 years old and often graduates, show the highest presence of qualified profiles, both in intellectual and technical professions.
The results of this analysis represent a preliminary exploration of the participation of young people in the labour market; a deep insight into their trajectories and conditions is crucial to facilitate the successful design of labour and education policies.

References

BILTON, P., JONES, G., GANESH, S., & HASLETT, S. 2017. Classification trees for poverty mapping. *Computational Statistics and Data Analysis*, **115**, 53–66.

BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R. A., & STONE, C.J. (Eds.). 1984. *Classification and regression trees (the wadsworth statistics/probability series)*. New York:: Chapman and Hall.

DACKO, M., ZAJAC, T., SYNOWIEC, A., OLEKSY, A., KLIMEK-KOPYR, A., & KULIG, B. 2016. New approach to determine biological and environmental factors influencing mass of a single pea (Pisum sativum L.) seed in Silesia region in Poland using a CART model. *European Journal of Agronomy*, **74**, 29–37.

DE' ATH, G., & FABRICIUS, K.E. 2000. ClassifIcation and regression trees: A powerful yet simple technique for ecological data analysis. *Ecology*, **81**(11), 3178–3192.

FRISMAN, L., PRENDERGAST, M., LIN, H.J., RODIS, E., & GREENWELL, L. 2008. Applying classification and regression tree analysis to identify prisoners with high HIV risk behaviors. *Journal of Psychoactive Drugs*, **40**(4), 447–458.

GALLETTA, S. 2016. On the determinants of happiness: A classification and regression tree (CART) approach. *Applied Economics Letters*, **23**(2), 121–125.

KEELY, L.C., & TAN, C.M. 2008. Understanding preferences for income redistribution. *Journal of Public Economics*, **92**(5–6), 944–961.

LOW, C.T., & LAI, P.C. 2016. Personal factors influencing the perception of quality of life in Hong Kong–a classification tree approach. *Procedia Environmental Sciences*, **36**, 70–73.

MANASSE, P., & ROUBINI, N. 2009. "Rules of thumb" for sovereign debt crises. *Journal of International Economics*, **78**(2), 192–205.

WILLIAMS, M.A., JOSKOW, A.S., JOHSON, R.L., & HURDLE, G.J. 1987. Explaining and predicting airline yields with non-parametric regression trees. *Economics Letters*, **24**(1), 99–105.

WALEGA, G., WALEGA, A. 2021. Over-indebted Households in Poland: Classification Tree Analysis. *Social Indicator Research,* **153**, 561–584.