R for Labour Market Policies

Gloria Ronzoni[1,*], **Ettore Colombo**[1], **Matteo Fontana**[1]

1. CRISP (Interuniversity research centre on public services), University of Milan Bicocca
*Contact author: gloria.ronzoni@crisp-org.it

Keywords: labour market, Talend integration, RMySQL , 3 millions workers

The project "Labour Market Observatory of Lombardia" started in 2005 with a collaboration between CRISP research centre and Regione Lombardia. The Observatory is aimed at gathering, updating and analyzing data and useful information that effectively investigates the efficacy of employment policies, educational system, professional training, further education and the regional labour market trends.

Here we propose an application of labor market monitoring at regional level dealing with longitudinal data: it deals with a classification of Lombardia workers' careers with the aim to identify the trend of contractual profiles (stable, improve or worse condition). The period of interest is since January, 2000 to June, 2009 and 3 millions workers are involved.

To achieve this goal we used a set of 'open source' programs like R, MySQL (Data Storage DBMS) and Talend Open Studio (ETL instrument). The RMySQL package allowed R to communicate with MySQL using the SQL language. The data set used contained 7 millions of rows with a set of 20 quantitative and qualitative variables, so the large amount of data required the implementation of a cyclic R algorithm that allowed to work with events relative to 100.000 workers each time.

Concerning the statistical methodology, the complexity of information didn't permit to use employment status approaches based on panels tipically used in the literature, so our purpose was to use an optimal scaling approach (**smacof** package, MDS approach) that determined, for each contractual typology, a weight of stability based on the working duration time. This quantification allowed to compute a career stability index for each worker and define his career condition.

Time of elaboration (less than one day) was widely reduced throught the joint use of R and MySQL. At the end of the algorithms run the resulting data set contained 3 millions of workers' careers clustered.

In the final part of this work we propose some analysis of the resulting data set and in particular an application of multiple correspondence analysis (**ca** package) that shows how workers' profiles are related to the associated qualitative variables, such as demographic and social information.

References

Lovaglio P.G. (2008). Analisi classificativa longitudinale dei percorsi lavorativi della provincia di Milano, in M. Mezzanzanica e P.G. Lovaglio (Eds), Numeri al lavoro, il sistema statistico del mercato del lavoro: metodologie e modelli di analisi, Quaderni dellosservatorio del mercato del lavoro, 3, Franco Angeli, Milano, pp59-81, ISBN 13: 9788846496195.