Symposium on Big Data in Finance, Retails and Commerce: Statistical and Computational Challenges
Lisbon, 2-3 November 2017, Portugal

Mismatch between jobs and skills in the EU

Joao Sollari Lopes (Speaker)

Instituto Nacional de Estatística, joao.lopes@ine.pt

Marco Moura

Instituto Nacional de Estatística, marco.moura@ine.pt

Sónia Quaresma

Instituto Nacional de Estatística, sonia.quaresma@ine.pt

Abstract

In order to study the mismatch between the available skills of labour force and the skills required by labour market, we developed the concept of Labour Market Attractiveness. This concept consisted of the combination of a set of variables from 6 Eurostats datasets on different subjects (i.e. Demographics; Earnings structure; Education and training; Life conditions; Employment and unemployment; and National accounts). The impact of this combined dataset on Skills Mismatch, as well as on Labour Market Mobility and Emigration, was assessed using various data mining techniques, particularly, clustering analysis, model selection analysis using multivariate regression and weighted network correlation analyses. We showed that Labour Market Attractiveness is able to form consistent clusters at country-level, which can be well defined using only the variables "Youth Unemployment" and "GDP". Furthermore, from this combined dataset we defined 6 Eigenvariables, namely, "Unemployment", "Poverty", "Ageing Population", "Secondary Education (Adults)", "Employment" and "Earnings structure". Considering these Eigenvariables, we found that: Skills Mismatch is negatively associated to "Employment" and "Secondary Education (Adults)", while being positively associated to "Poverty" and "Unemployment"; Labour Market Mobility is associated to "Earnings structure" and "Employment"; and Emigration is negatively associated to "Secondary Education (Adults)" and "Ageing Population". From model selection, we showed that: Skills Mismatch is best explained by "Proportion of employed youth" and "Proportion of employed youth working at NACE M-N";

Labour Market Mobility is best explained by “Proportion of employed adults working at NACE L and K”; and Emigration is best explained by “Proportion of employed adults working at NACE K”, “Proportion of employed youth with Higher Education”, “Proportion of employed youth working at NACE B-E and O-Q” and “Total Population size”.

Keywords: Labour Market Attractiveness, Labour Market Mobility, Skills Demand, Skills Supply.

Introduction The European Big Data Hackathon took place in March 2017 - in parallel with the conference New Techniques and Technologies for Statistics (NTTS). This event was organised by the European Commission (Eurostat) and gathered 22 teams from 21 European countries. The aim was to compete for the best data product combining official statistics and Big Data to support policy makers in a pressing policy question, namely, “How to tackle the mismatch between jobs and skills at regional level in Europe?”. Indeed, the mismatch between the available skills of the labour force and the skills required by the labour market entail significant economic and social costs for individuals and firms [1]. Furthermore, a strong education and an efficient development of skills are essential for thriving in the emerging new economy and fast-changing labour market [1]. Nonetheless, a survey from 2014 showed that skills mismatch (i.e. over-qualification, under-qualification) remains at 45% in the European Union [2]. This led to the publication of the European Union Guidelines for the employment policies of the Member States in 2015, which called for enhancing labour supply, skills and competences [3]. In order to better support policy makers in solving this skills mismatch problem, the data product was required to be supported by relevant data, statistical analysis and visualization. Teams were also invited to use provided datasets (including European Employment Services (EURES) data on jobseekers and on job vacancies [4]), and additional publicly available data sources with international applicability (e.g. Eurostat online database [5]).

The development of our data product was focused on three main ideas: 1) combine official statistics data from Eurostat at NUTS2 level (i.e. solid data sets known for being well-structured, clean and accurate, but also characterized by a morose collect and release process) with real-time unstructured Big Data; 2) explore the notion of Labour Market Attractiveness as an important factor in the mismatch between skills demand and supply, in labour market mobility, and in emigration; 3) create a flexible, interactive and user-friendly product that allows for customization of the answers in order to be used either by policy makers or

by both citizens searching for help on jobseeking and enterprises looking for particular labour market characteristics.

The definition of Labour Market Attractiveness has to be considered carefully, thus, our approach should be seen has a first-step towards a more mature definition. We considered 17 variables from 6 Eurostat datasets, namely "reg_demo" for demographics data, "earn" for earnings structure data, "edtr" for data on education and training, "ilc" for life conditions information, "employ" for employment/unemployment data, and "na10" for national accounts data. These variables were broken by several categorical levels (e.g. "age groups", "level of education", "qualifications", "occupations") originating more than 70 variables. Several data mining techniques were then considered to analyse this compiled Labour Market Attractiveness dataset. Using the datasets we calculated distances between regions and visualize those using social networks algorithms. We further clustered the regions using a Partition Around Medoids method on those distances creating a categorical variable with grouping information [6]. This created variable, along with collected variables on Skills mismatch, Labour Market Mobility and Emigration, were used separately as dependent variables on model selection using multivariate linear and non-linear regression analyses with the Labour Market Attractiveness dataset as independent variables [7]. We further constructed Eigenvariables from the considered set and performed Weighted Correlation Network Analysis on the dependent variables [8].

In our work we assumed two major simplifications in the construction of the skills mismatch indicator, however, these simplifications do not affect our product in terms of proof-of-concept and can be dropped in later developments. The first one was to use previously cleaned and treated data on job vacancies and education attainment from the Eurostat's "labour" and "edtr" data sets, respectively. Instead, a better approach would be to use the freshly collected EURES data provided, but the use of this data would have two caveats: a) the cleaning and structuring of the data requires a considerable expertise on the subject; b) the normalizing of the data, using for example marginal calibration techniques, requires detailed demographic data at the required regional level in order to successfully capture the populations considered. The second simplification was to use an ad hoc mapping between qualifications (classified using ISCED-F 13) and the cross between occupations (defined using ISCO-08) and economic activity (defined using NACE Rev. 2). Nonetheless, a formal mapping will be released in mid-2017 by European Skills, Qualifications and Occupations (ESCO) from the EC.

Statistical analyses were carried out at country-level and at NUTS1- and NUTS2-level. They were performed in R using libraries "cluster", "glmulti", "Hmisc", "MASS", "nnet", "sna" and "WGCNA".

Conclusions At country-level, the compiled Labour Market Attractiveness dataset is able to form consistent clusters (clusters separation between 2.31 and 4.68). Moreover, these clusters can be well defined even using only a data subset of "Youth unemployment" and "GDP" ($R_{McFadden}$ = 0.84). Furthermore, the Labour Market Attractiveness dataset can be reduced to 6 Eigenvariables, namely, "Unemployment", "Poverty", "Ageing Population", "Secondary Education (Adults)", "Employment" and "Earnings structure". Considering these Eigenvariables, we found that Skills Mismatch is very negatively associated to "Employment" (ρ = -0.69, p = 0.058) and moderately negatively associated to "Secondary Education (Adults)" (ρ = -0.38, p = 0.352), while being moderately associated to "Poverty" (ρ = 0.38, p = 0.352) and "Unemployment" (ρ = 0.36, p = 0.385). Labour Market Mobility is strongly associated to "Earnings structure" (ρ = 0.59, p = 0.002) and moderately associated to "Employment" (ρ = 0.35, p = 0.082). Lastly, Emigration is very negatively associated to Secondary Education (Adults) (ρ = -0.50, p = 0.007) and moderately negatively associated to Ageing Population (ρ = -0.36, p = 0.063). Finally, the model selection using multivariate linear regression shows that for Skills Mismatch the most important explanatory variables are "Proportion of employed youth" (Importance = 0.84) and "Proportion of employed youth at NACE M-N" (Importance = 0.78). For Labour Market Mobility they are "Proportion of employed adults at NACE L" (Importance = 1.00), "Proportion of employed adults at NACE K" (Importance = 0.98) and "Proportion of employed youth at NACE B-E" (Importance = 0.67). Regarding Emigration the variables are "Proportion of employed adults at NACE K" (Importance = 0.93), "Proportion of employed youth with Higher Education" (Importance = 0.89), "Proportion of employed youth at NACE O-Q" (Importance = 0.86), "Total population size" (Importance = 0.74) and "Proportion of employed youth at NACE B-E" (Importance = 0.55).

References

[1] https://ec.europa.eu/commission/publications/skills-education-and-lifelong-learning-european-pillar-social-rights_en

[2] CEDEFOP (2015) *Skills, qualifications and jobs in the EU: the making of a perfect match?* Publications Office of the European Union, Luxembourg.

[3] Council Decision (EU) 2015/1848 of 5 October 2015

[4] https://ec.europa.eu/eures/public/homepage

[5] http://ec.europa.eu/eurostat/data/database

[6] Reynolds, A.P. et al. (1992) Clustering rules: A comparison of partitioning and hierarchical clustering algorithms. *J Math. Model. Algorithms* 5(4), 475-504.

[7] Calcagno, V., de Mazancourt, C. (2010). glmulti: an R package for easy automated model selection with (generalized) linear models. *Journal of statistical software* 34(12), 1-29.

[8] Langfelder, P., Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis *BMC Bioinformatics* 9(1), 559.