MACHINE LEARNING METHODS FOR ESTIMATING THE EMPLOYMENT STATUS IN ITALY

In recent decades, National Statistical Institutes have focused on producing official statistics by exploiting multiple sources of information (multi-source statistics) rather than a single source, usually a statistical survey. The growing importance of producing multi-source statistics in official statistics has led to increasing investments in research activities in this sector.

In this context, one of the research projects addressed by the Italian National Statistical Institute (Istat) concerned the study of methodologies for producing estimates on employment rates in Italy through the use of multiple sources of information, survey data and administrative sources. The data comes from the Labour Force (LF) survey conducted by Istat and from several administrative sources that Istat regularly acquires from external bodies. The "quantity" of information is very different: those coming from administrative sources concern about 25 million individuals, while those coming from the LF survey refer to an extremely limited number (about 330,000) of individuals. The two measures do not agree on employment status for about 6% of the units from the LF survey.

One proposed approach uses a Hidden Markov model to take into account the deficiencies in the measurement process of both survey and administrative sources. The model describes a measurement process as a function of a time-varying latent state (in this case the employment category), whose dynamics is described by a Markov chain defined over a discrete set of states. At present, the implementation phase for the production process of statistics on employment through the use of HM models is coming to an end in Istat.

The present work describes the use of Machine Learning methods to predict the individual employment status. This approach is based on the application of *decision tree* and *random forest* models, that are predictive models, usually used to classify instances of large amounts of data. In the work, the obtained results will be described, together with their usefulness in this application context. The models have been applied through the use of the software R.