Integration of administrative sources and survey data through Hidden Markov Models for the production of labour statistics

Danila Filipponi, Ugo Guarnera, Roberta Varriale

1 Introduction

The work presented in 27 aprile 2017 by Danila Filipponi and Ugo Guarnera dealt with the integration of administrative sources and the Italian Labour Force Survey (Lfs) data through Hidden Markov Models for the production of labour statistics. It was proposed to use an unsupervised approach considering the target variables as latent (unobserved) variables, and to model the measurement processes through the distributions of the observed variables conditional on the latent variables. Within this general framework, appropriate models are Hidden Markov Models (HMM). In fact the methodological choices have to take into account that the variable of interest is categorical and the data are longitudinal. A simulation study was presented where the methodology was applied in different scenarios. The main goal of the study was to assess the robustness of the methodology with respect to departure from the model assumptions. In particular, the interest was in evaluating the robustness of the measurement error parameter estimates and of the aggregate estimates based on prediction of the latent variable given the observed measurements.

2 Developments

After the comments and suggestions from the Advisory board, we made additional advances in our work.

2.1 Hidden Markov Model: application and results

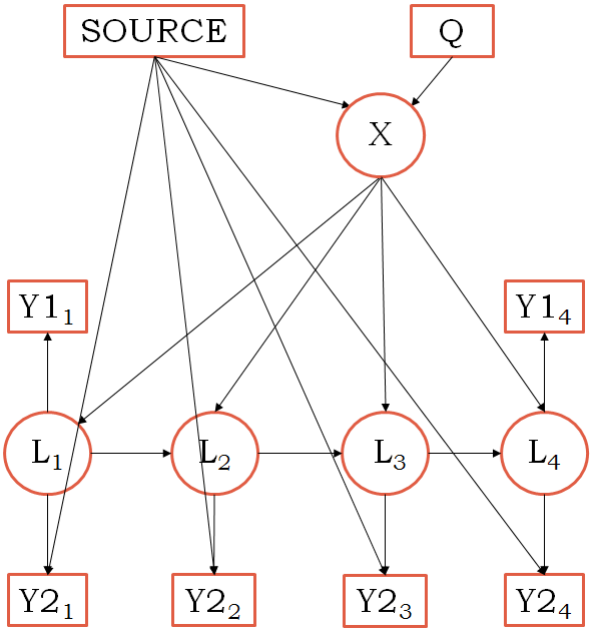
We focused our work on the application of HMM to our data.

The HMM is fitted on a person-linked combined dataset containing monthly employment status measured by administrative sources and by Labour Force Survey (LFS). The administrative data contains individual scores for the complete population per month, while the LFS is administered twice a year, with three months in between. Of course, the LFS is only administered on a sample of the population. When the hidden Markov model is applied, 12 time-points, one for every month, are considered, spanning a time-frame of one year.
Figure 1 shows a graphical representation of the model for 4 time points, where circles represent latent variables and rectangles observed measures. $Y_{1,1}, ..., Y_{1,T}$ denotes the response variables originating from the LFS and $Y_{2,1}, ..., Y_{2,T}$ denote the response variables originating from the administrative source. $L_1, ..., L_T$ represents the latent Markov variable measuring the 'true' employment scores (employed/not employed). The variable X is a discrete (latent) random effect, that is included in the model to account for individual heterogeneity, so $\mathbf{L}$ follows a first order Markov chain conditional on X. *Source* and $\mathbf{Q}$ represent covariates used in the model. In particular, the *Source* covariate specifies from which administrative source the $Y_{2,t}$ score originates.

The main assumptions of the model are:

- heterogeneity of the Markov process in the population. We assume the presence of three subpopulations differing in terms of the latent status development (initial and transition probabilities) and

Figure 1: Hidden Markov model used to estimate 'true employment rates' per month in Italy, 4 time points.



we used a latent variable X to capture them. The different trajectories of $\mathbf{L}$ can be described as never employed ($x = 1$), always employed ($x = 2$) and moving between employment and unemployment ($x = 3$). In order to correctly identify the three components of the latent process we used the covariate *Source* with some restrictions;

- the transition probabilities do not change over time within each components;
- the classification errors of the indicators are mutually independent and independent over time;
- different miss-classification errors for the administrative sources. To this purpose we used the variable *Source* as covariate in the measurement model;
- we introduce some restriction in the measurement model in order to assume equal to zero the miss-classification error in Lfs when the indicator $Y_{1,t} = 1$ (employed);
- the missing values due to the panel construction are assumed to be Missing Completely At Random;
- we estimated different models for each region in order to take into account spatial heterogeneity.

We based the choice between different models on the Information Criteria AIC and BIC and a deep knowledge of the phenomenon. We compared the results obtained through the model with those obtained with both the Lfs data using traditional methods and administrative information. Some of them are reported in Figures 2 and 3.

2.2 Software

We started to use Software Latent GOLD v.5.1 (Vermunt J.V. and Magidson, J., 2015) for the estimation and evaluation of the HMMs. This was due to the fact that the package MlEst (R software) has some limitations in the model specification and we needed a more flexible tool.

Figure 2: Employment by Region. Year 2015

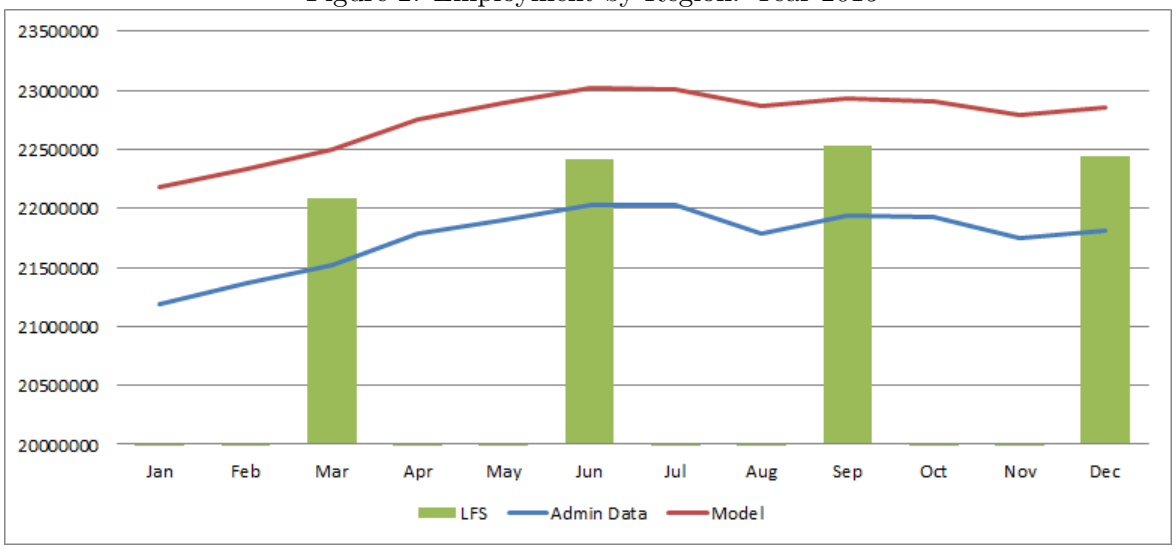
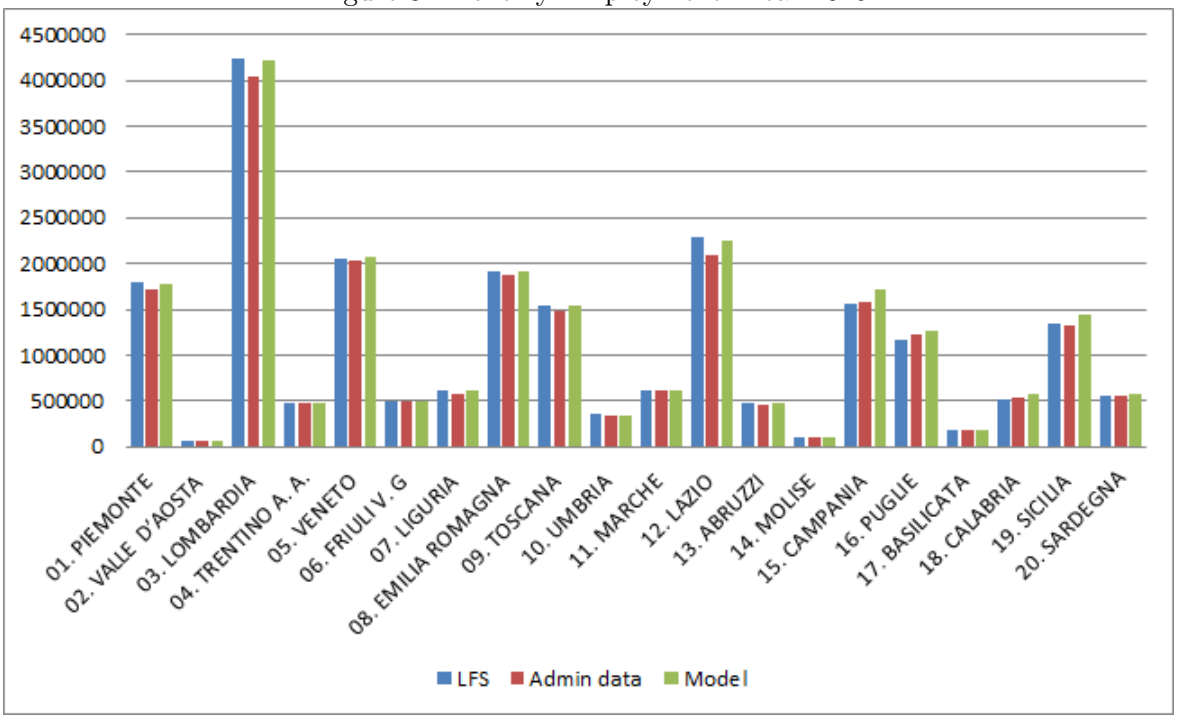


Figure 3: Monthly Employment. Year 2015



2.3 Project ISTAT-CBS: Hidden Markov models to estimate employment rates

Finally, we supervised the work of a PhD student from Tilburg university on the Project Agreement ISTAT-CBS (ROSE Network collaboration): *Hidden Markov models to estimate employment rates.*
The project deals with the extension of Multiple Imputation of Latent Classes (MILC) method to longitudinal data, in order to evaluate the variability of population estimates for different subgroups obtained by the application of HMM.