Hidden Markov Models to Estimate Italian Employment Status

Keywords: *Hidden Markov Models, Labour force statistics, longitudinal data*

1 Introduction

The increased availability of large amount of administrative information at the Italian Institute of Statistics (Istat) makes it necessary to investigate new methodological approaches for the production of estimates, based on combining administrative data with statistical survey data. Traditionally, administrative data have been used as auxiliary sources of information in different phases of the production process such as sampling, calibration, imputation. In order to take into account deficiencies in the measurement process of both survey and administrative sources, a more *symmetric approach* with respect to the available sources can be adopted. A natural strategy, according to this approach, is to consider 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. In this context, Latent Class Analysis (LCA) is traditionally considered as a method to identify a categorical latent variable using categorical observed variables, which a longitudinal extension is the Hidden Markov Model (HMM). Examples on the use of latent models in Official Statistics are becoming common (Oberski [1]; Boeschoten [2]) and several applications can be found in the field of employments research (Biemer [3]; Magidson [4]; Pavlopoulos [5]).
In this paper we show the use of a latent model for estimating employment rates in Italy using both Labour Force Survey(LFS) and administrative data. Here, the use of HMM is particularly suitable since, as for many (European) countries, employment administrative data are collected on a monthly bases, while the LFS data contains a rotating panel structure The aim is both to show which are the potentiality of this methodology and the possible problems and research topics.

2 The informative context

The main sources available for the production of labour statistics are the Italian Labour Force Survey (Lfs) and administrative sources. The Italian LFS is a continuous survey carried out during every week of the year. Each quarter, the LFS collects information on almost 70,000 households in 1,246 Italian municipalities for a total of 175,000 individuals (representing 1.2% of the overall Italian population). The LFS provides quarterly estimates of the main aggregates of labour market (employment status, type of work, work experience, job search, etc.), disaggregated by gender, age and territory (up to regional detail). Administrative data relevant for the labour statistics come mainly from social security and fiscal authority. Data are organized in an information system having a linked employer-employees structure. From this data structure it is possible to obtain information on the statistical unit of interest, i.e., the worker. The available data are linked at person level and the resulting dataset contain 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.

Table 1: Cross-classification of the employment status measured by Lfs and AD. LFS data, Year 2015. Italy

Lfs \ AD	Out	In	Total
Not Employed	59.9	2.9	62.8
Employed	3.2	34.0	37.2
Total	63.1	36.9	100.0

Table 1 shows the miss classification between the employment status estimated by LFS and Administrative sources. The *symmetric approach* relies on the assumption that both statistical and administrative data could be affected by measurement errors and then the off diagonal values (2.9% and 3.2%) can be miss classification errors of both sources. The main goal of this analysis is twofold: (i) to evaluate the accuracy of the available administrative sources, in order to define their use into the statistical production, (ii) to produce statistics on the employment status by small geographical domains in order to fulfill the population census requirements.

3 The Hidden Markov Model: the Italian case

Within the general framework described above 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.
Figure 1 depicts the HMM for estimating the Italian employment status. Following the conventions, circles represent latent variables and rectangles observed variables; arrows connecting latent and/or observed variables represent direct effects, which do not need to be linear.

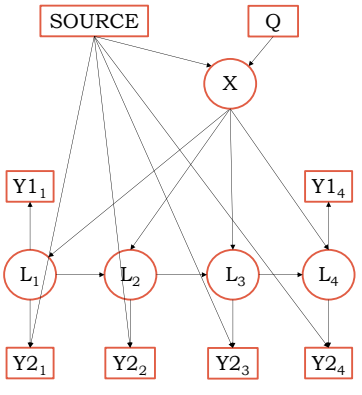

Figure 1: HMM to estimate the monthly employment status in Italy

$Y_{1,t}$ and $Y_{2,t}$ represent the response variables for the LFS and the administrative source, respectively, where $t = 1, ..., T$, and $T = 12$. The vectors $Y_{1,1:T} = (Y_{1,1}, ..., Y_{1,T})$ and $Y_{2,1:T} = (Y_{2,1}, ..., Y_{2,T})$ contains the measures originating from the two sources over the period $t = 1$ to $t = T$. The number of categories of the two response variables is equal, namely two, with (1) unemployed and (2) employed.

Variables L represent the "true" target phenomenon. These are the variables that we would observe if data were error free. According to the HMM modeling, the latent variable at time t, L_t takes values on a finite set of size r, with the set $(1, 2, .., r)$. In

this work, the number of latent states is equal to two ($r = 2$), with (1) unemployed and (2) employed. For a given final time T, the values $(l_1, ..., l_T)$ represent the realization of an unobserved random process $L_{1:T} = (L_1, ..., L_T)$ at discrete times $1, .., T$. We assume that the stochastic process $L_{1:T}$ is a first order Markov process, that is $P(L_{(t+1)}|L_1, L_2, \ldots, L_t) = P(L_{(t+1)}|L_t)$. The law of this process is specified through the initial and transition probabilities.
The latent variable X is used to capture the population heterogeneity of the Latent Markov process. In particular, the different components (classes) of the latent variable X represent the different trajectories of $L_{1:T}$, with different initial state and transition probabilities.
In the model, we also assume the effect of some time-invariant covariates (Q and $Source$), associated to the latent variables X and to the measures $Y_{2,t}$ through statistical models. In particular, the covariate $Source$ represents the admnistrative source from which we extract the administrative information: (1) no signs of employment from the administrative sources, (2) social security data: employees, (3) social security data:outworkers and (4) fiscal data; and it will help in the identification of the different components of the latent variable X.
In example, if a person does not have signs of employment from the administrative sources ($Source = 1$) it is likely that the person is unemployed and will stay unemployed, while if the employment signals are from a civil servant source ($Source = 2$) is likely that the person will stay employed and if the employment signals are coming from outworkers social security data ($Source = 3$) is likely that the person will change the employment status during the reference period. Restrictions used to link in a deterministic way the type of administrative source and the measurement process will be described in the full paper.
In the following, we refer to the vector Q to represent all the other covariates used in the model, that are sex, age, income, education and information on retirement.
The distribution of the observed indicators, given the covariates, is:

$$\mathrm{P}_{Y_{1,1:T},Y_{2,1:T}|Q_1,Q_2}(\mathbf{y}_1, \mathbf{y}_2|q_1, q_2) = \sum_{x=1}^{3}\sum_{l_1=1}^{2}\sum_{l_2=1}^{2}\cdots\sum_{l_T=1}^{2}\phi_{x|q,source}\pi_{l_1|x}\prod_{t=2}^{T}\pi_{l_t|l_{t-1},x}\prod_{t=1}^{T}{\psi^{(1)}_{y_{1,t}|l_t}}^{\delta_t}\psi^{(2)}_{y_{2,t}|l_t,source}, \tag{1}$$

where $\pi_{l_1|x}$ represent the initial probabilities, $\pi_{l|l_{t-1},x}$ represent the transition probabilities, $\psi^{(j)}_{y_{j,t}|l_t,source}$ with $j = 1, 2$ represent the conditional response probabilities and δ_t indicates whether and observation for $Y_{1,t}$ is present at time-point t. The parameter $\phi_{x|q,source}$ represent the effect of covariates on the latent variable X.
In the model estimation, we assume that: (i) the classification errors are assume to be conditionally independent (contemporary and serial independence), (ii) the amount of classification error within the indicators does not change over time, (iii) the transition probabilities do not change over time, (iv) the missing values due to the panel construction are Missing Completely At Random and missing values due to attrition are Missing At Random .
In the application to the Italian labour market, we estimate one model for each region in order to consider spatial heterogeneity. In order to reduce the parameters of the estimates and satisfy some initial hypotheses, various constraints have been introduced; in particular absence of "false positives" in the employment status of LFS.
For the model estimation, we used the Software Latent GOLD v.5.1

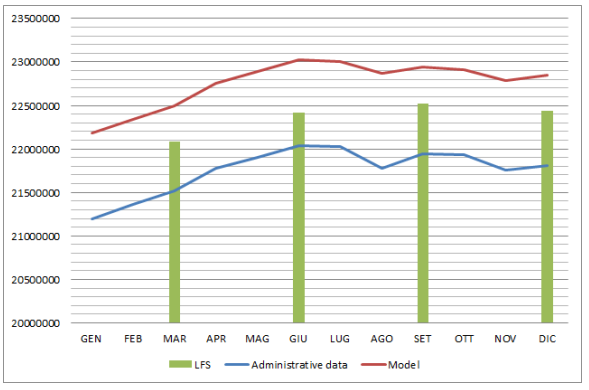

Figure 2: Monthly and quarterly estimate of the employment status in Italy. Year 2015

4 Results and conclusions

Based on the estimation of HMM parameters (initial and transition probabilities, size of the latent classes of X, parameters linked to covariate effects), using Bayes theorem one can derive the distribution of the latent variables conditional on the available information (posterior distribution) and use the expectations from this distribution to obtain predictions of the true values for each unit. From this, we can compute the estimation of aggregates based on a posterior probabilities.
Table 2 reports both the aggregate estimate of the employment status at the national level based on the information from LFS, AD and the predicted a posterior probabilities of the HMM. Estimates of the employment status at different level of aggregation of the entire population can be produced. Furthermore, one of the possible usages of latent models in Official statistics is to assess the quality of the available sources, through the estimation of the parameters of the measurement model (measurement error).
Despite of the quantity of information that can be extracted from the application of HMM, important issues should be taken into account in order to use HMM results for Official Statistics. Further research is needed in order to derive the accuracy of the final aggregates and extensions in order to account for possible departure from the Markov assumption.

References

[1] D. L Oberski. Total survey error in practice. *In P. P. Biemer et al. (Eds.), Total survey error (chap. 16 Estimating error rates in an administrative register and survey questions using a latent class model). New York: Wiley*, 2015.

[2] Oberski D. De Waal T. Boeschoten, L. Estimating classification errors under edit restrictions in composite survey-register data using multiple imputation latent class modelling (milc). *Journal of Official Statistics*, 33(4):921–962, 2017.

[3] P. P. Biemer. In analysis of classification error for the revised current population survey employment questions. *Survey Methodology*, 30(2):127–140, 2004.

[4] Vermunt J. K. Tran B. Magidson, J. Using a mixture latent markov model to analyze longitudinal us employment data involving measurement erro. *New trends in psychometrics*, page 235–242, 2009.

[5] Vermunt J.K. Pavlopoulos, D. Measuring temporary employment. do survey or register data tell the truth? *Survey Methodology*, 41(1):197–214, 2015.