

New Techniques and Technologies for Statistics 2019



| ROBERTA VARRIALE - ISTAT |

Estimate employment rates in Italy using both Labour Force Survey (LFS) and administrative data.

Application of an Hidden Markov Model: a longitudinal extension of Latent Class Analysis (method to identify a categorical latent variable using categorical observed variables)

Italian Labour Force Survey (LFS)

- Continuous survey carried out during every week of the year
- 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)
- Sample survey: each quarter, the LFS collects information on almost 70,000 households in 1,246 Italian municipalities for a total of 175,000 individuals (1.2% of the overall Italian population)
- Per each unit, we observe maximum two occasions (with three months in between)

Administrative Data (AD)

- Mainly from social security and fiscal authority
- Data organized in an information system having a linked employer-employees structure
- Individual employment status for the complete population per month
- Different data sources with different characteristics and quality

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

LFS\AD	Not present (Not employed)	Present (Employed)	Total
Not employed	60.2	2.6	62.8
Employed	3.5	33.8	37.2
Total	63.7	36.3	100.0

The diagram illustrates the proposed model architecture. It features a sequence of latent variables $L_1, L_{t-3}, L_{t-2}, L_{t-1}, L_t, L_{t+1}, \dots, L_T$ represented by red circles. Each L_t is connected to $Y_{1,t}$ (red rectangle) and $Y_{2,t}$ (red rectangle). A **SOURCE** node (red rectangle) is connected to all $Y_{2,t}$ nodes. A Q_1 node (red rectangle) is connected to L_t and L_{t+1} . The diagram also shows connections between L_{t-3} and $Y_{2,t-3}$, L_{t-2} and $Y_{2,t-2}$, L_{t-1} and $Y_{2,t-1}$, and L_{t+1} and $Y_{2,t+1}$. The $Y_{1,t}$ node is connected to $Y_{1,t+1}$.

Circles: latent variables

Rectangles: observed variables

Arrows connecting latent and/or observed variables: direct effects (not necessarily linear)

VARIABLES

t : unit of time ($t=1, \dots, 12$; month)

L_t : latent employment status (0 = not employed; 1 = employed)

Y_1 : observed employment status, from LFS

Y_3 : observed employment status, from AD

X: latent subpopulations (1 = never employed; 2 = stable employers (S); 3 = not stable employers (M))

COVARIATES

Q1: retirement status, student, earnings, age, gender

Source: type of administrative source (1 = No source; 2 = Employees (social security data); 3 = Self-employers, with time information (social security data);

HYPOTHESES

(i) the responses are locally independent (CIA) and independent over time, given the latent variables, (ii) the measurement error of 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.

CONSTRAINTS

- $Y_{2,t} = 0$ (unemployed) when the statistical unit is not present in administrative sources ($s = 1$)
- $Y_{2,t} = 1$ (employed) when the statistical unit is self-employed with information from administrative sources that are not related to time ($s = 4$)
- no false positives in the data coming from the LFS

SOFTWARE:

LatentGOLD v.5.1 has been used (Vermunt and Magidson, 2015)

Initial probabilities

	l_1	
x	0	1
2 (S)	0.01	0.99
3 (M)	0.65	0.35

Transition probabilities

		l_t	
		0	1
2 (S)	1	0.78	0.22
	2	0.00	1.00
3 (M)	1	0.81	0.19
	2	0.18	0.82

AD Measurement error

SOURCE		$Y_{2,t}$	
		Not employed	Employed
Employees	Not employed	0.99	0.01
	Employed	0.05	0.95
Self-employers with time information	Not employed	0.95	0.05
	Employed	0.02	0.98

The adopted HMM was used to evaluate the accuracy of available administrative sources, to define their use into the statistical production, and to produce estimates on employment status at different level of aggregation of the entire population. The methodology seems to be promising.

Important issues:

- further research is needed to derive the accuracy of the final aggregates and extensions need to be evaluated to account for possible departures from the Markov assumption;
- introduce dependence between the measures under investigation, to take into account, for example, the tendency to conceal undeclared work in the LFS;
- to evaluate the opportunity of applying the estimate and prediction phase directly on the entire population (in the present application, HMM was estimated on LFS data and values obtained by marginal imputation have been used to build a synthetic micro-data file for the entire population and the estimates of employment rate for different domains).

F. Bartolucci, A. Farcomeni, and F. Pennoni. *Latent Markov models for longitudinal data*. Chapman and Hall/CRC, 2012.

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

H. J.K. Vermunt and J. Magidson. *Technical guide for LatentGOLD5.0: Basic, advanced, and syntax*, 2015