Hidden Markov model to analyze NEET and youth unemployment comparing EU countries over time

Fulvia Pennoni

Department of Statistics and Quantitative Methods
University of Milano-Bicocca
Email: fulvia.pennoni@unimib.it

joint work with *B. Bal-Domańska*
Wroclaw University of Economics and Business, Poland

Outline

- Introduction and conceptual framework
- Data
- Hidden Markov model
- Results
- Conclusions

Introduction

- Young people belong to one of the groups most sensitive to fluctuations of the labor markets
- We consider those who are Not in Employment, Education, or Training (NEET), aged 15 to 29 and youth unemployment (YU) concerning those aged between 15 and 24
- Both (despite differences) NEETs and YU are positioned in an overall debate on youth unemployment (Eurofound, 2012)
- We account fof:
  - (*i*) a dynamic classification of the European countries on the basis of NEETs and YU separately through a statistical model tailored for the analysis of longitudinal data;
  - (*ii*) identification of the country's patterns of changes in both response variables;
  - (*iii*) understanding to what extent these patterns are due to the countries' economy and labor market conditions

Introduction

- Numerous factors influence the activity of young people on the labor market:
  - The economic structure and its growth as well as the productivity, the balance of the trade in goods and services, the inflation, the general unemployment, the labor costs, the jobs offered;
  - The legal regulations and the social policy tools
  - The educational models and the vocational training systems
  - The technological change, for example, the human resources employed in science and technology
  - The cultural and social conditions; the globalization processes; the demographic transformation (e.g., population aging, migrations, fertility rates, youth cohort sizes)

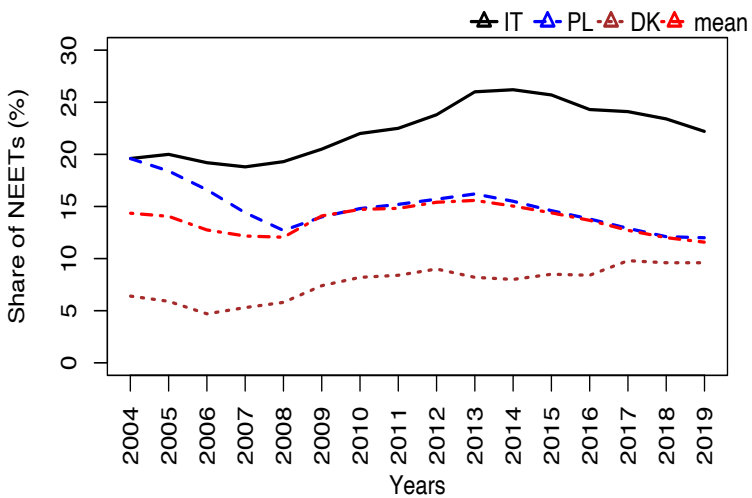
Data and conceptual framework

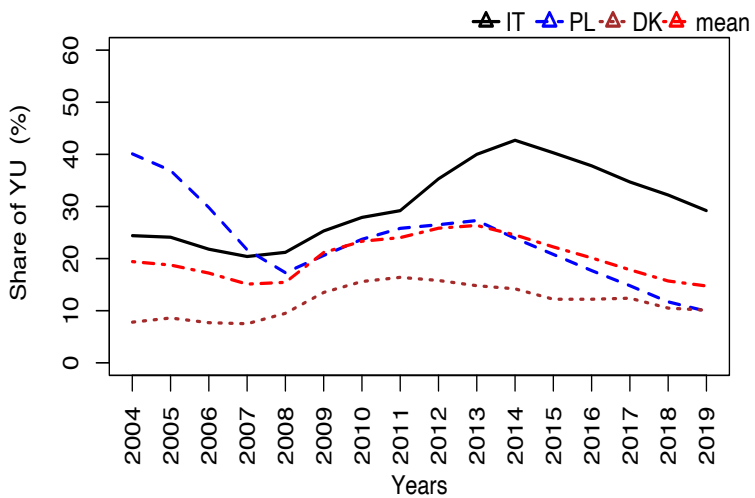
- The empirical analyses are carried out through data collected by Eurostat[1] covering a period of sixteen years from 2004 to 2019
- We focus on the differences between the EU economies through exogenous macroeconomic variables which reflect the countries' economy and the labor market status:
  - Dgdp: Gross domestic product at market prices chain linked volumes, index 2005=100
  - Hrst: Persons employed in science and technology as percentage of the active population
  - El: Early leavers from education and training
  - Pedu: Participation rate in education and training during the last four weeks before the survey
  - Part: Part-time employment as percentage of total employment
  - Temp: Temporary contracts as percentage of total employment
  - LCI: Labor Cost Index expressed as a share of non-wage costs **(compensation of employees plus taxes minus subsidies)**

[1] **Retrieved from** https://ec.europa.eu/eurostat

Data structure

- A focus on the observed values of NEETs and YU for three countries namely Italy (IT), Poland (PL) and Denmark (DK) and the average values of the 28 countries over time





Descriptive analysis: Average value of the covariates

Year	*dgdp*	*part*	*temp*	*hrst*	*pedu*	*el*	*lci*
2004	96.196	12.954	9.246	25.575	9.014	15.143	22.607
2005	96.196	13.154	9.464	26.064	9.068	14.414	22.607
2006	105.043	13.289	9.514	26.589	9.279	13.968	22.607
2007	110.371	13.318	9.589	27.107	9.154	13.632	21.704
2008	111.757	13.429	9.257	27.704	9.486	13.450	21.704
2009	105.654	14.179	9.075	28.029	9.639	12.721	21.704
2010	107.336	14.646	9.521	28.125	9.825	12.204	21.704
2011	109.321	14.857	9.725	29.350	10.161	11.518	21.589
2012	109.132	15.096	9.536	29.857	10.232	11.100	21.589
2013	109.743	15.325	9.732	30.186	10.682	10.418	21.589
2014	112.411	15.375	10.125	30.650	10.607	9.904	21.550
2015	116.793	15.364	10.439	30.954	10.818	9.846	21.455
2016	120.161	15.357	10.500	31.575	10.875	9.479	21.368
2017	124.654	15.161	10.429	32.243	11.321	9.411	21.304
2018	128.989	14.807	10.229	32.971	11.568	9.193	21.293
2019	132.893	14.696	9.861	33.675	11.821	8.993	20.468

Proposed hidden Markov model

- We denote:
  - n: sample size,
  - T: number of time occasions;
  - Y_{it}: continuous response variable for subject i at occasion t, $i = 1, \ldots, n$, $t = 1, \ldots, T$;
  - $\mathrm{U}_i = (U_{i1}, \ldots, U_{iT})$: latent process affecting the distribution of the response variables

- The distribution of the latent process is assumed to follow a first-order Markov chain with state-space $1, \ldots, k$, where k is the number of latent states

- Under the local independence assumption, the response vectors $(\mathrm{Y}_{i1}, \ldots, \mathrm{Y}_{iT})$ are assumed to be conditionally independent given the latent process $\boldsymbol{U}_i$ and the elements Y_{ijt}, $j = 1 \ldots, r$ of Y_{it} are conditionally independent given U_{it}

- A conditional Gaussian distribution is assumed for the response variables

$$f(\boldsymbol{Y}_{it} = \boldsymbol{y} | U_{it} = u) \sim N(\boldsymbol{\mu}_u, \Sigma)$$

Parametrization

- Parameters of the latent model are:
  - The initial probabilities which are parameterized as

$$\log \frac{\pi_{iu|\boldsymbol{x}}}{\pi_{i1|\boldsymbol{x}}} = \beta_{0u} + \boldsymbol{x}'_{i1}\boldsymbol{\beta}_{1u}, \quad u = 2, \ldots, k,$$

  - The transition probabilities which are parameterized as

$$\log \frac{\pi_{i,u|\bar{u},\boldsymbol{x}}}{\pi_{i,|\bar{u},\boldsymbol{x}}} = \gamma_{0u} + \boldsymbol{x}'_{it}\boldsymbol{\gamma}_{1u}, \quad u = 1, \ldots, k, \ = 1, \ldots, k+1, \ \neq u,$$

Maximum log-likelihood estimation

- Let $\boldsymbol{y}_{it} = (\boldsymbol{y}_{it})'$ denote the vector of responses for a sample of n independent units

- Assuming independence between units the log-likelihood referred to the observed data can be written as

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(\boldsymbol{y}_i) = \sum_{i=1}^{n} \log \sum_{u_i} \left(\prod_{t=1}^{T} f(\boldsymbol{y}_{it} | u_{it}) \right) \left(\pi_{iu|\boldsymbol{x}} \prod_{t=2}^{T} \pi_{i,u|\bar{u},\boldsymbol{x}} \right),$$

- where
  - $\boldsymbol{\theta}$ is the vector of all the model parameters
  - $f(\boldsymbol{y}_i)$ is the manifest distribution of all response observed variables

Expectation-Maximization algorithm

- The EM algorithm is based on the complete-data log-likelihood denoted as $\ell^*(\boldsymbol{\theta})$

- It alternates two steps until a suitable convergence criterion is reached:

E-step: compute the posterior expected value of the dummy variables given the observed data and the current value of the parameters by means of suitable recursions

M-step: update the estimates of $\boldsymbol{\theta}$ by maximizing the expected value of $\ell^*(\boldsymbol{\theta})$ obtained at the E-step

Other features of the estimation

- Initialization of the model parameters: we combine deterministic and random initializations of the EM algorithm to overcome the problem of multimodality of the log-likelihood function

- Selection of k: the number of latent states are selected by using the Bayesian Information Criterion (BIC, Schwarz, 1978)

- Prediction of the sequence of latent states: local decoding is performed to estimate the subject specific sequence of latent states, which is based on the estimated posterior probabilities of U_{it} directly provided by the EM algorithm

Results

- Results of the fitting of the two models for NEETs and YU with an increasing number of latent states ranging from 1 to 5

NEET

k	$\hat{\ell}_k$	m	BIC_k
1	-1,343.602	2	2,693.869
2	-1,141.111	24	2,362.195
3	-1,017.790	60	2,235.513
4	-958.309	110	2,283.161
5	-899.924	174	2,379.652

YU

k	$\hat{\ell}_k$	m	BIC_k
1	-1,649.265	2	3,305.195
2	-1,479.582	27	3,049.134
3	-1,346.322	68	2,919.234
4	-1,307.804	125	3,032.133
5	-1,288.177	198	3,236.131

- The *BIC* index suggests that three latent states are preferable for the parsimony principle

Results

- The estimated expected averages are the following whereas the estimated common variance $\hat{\sigma}^2$ is 4.177 for NEETs and 18.902 for YU

	$u = 1$	$u = 2$	$u = 3$
NEET	8.441	13.501	20.357
YU	12.644	21.883	37.968

- The three states define clusters of countries whose average shares of NEETs and YU are very different

- They are defined as the group of the best countries (1st), of intermediate countries (2nd) and of vulnerable countries (3rd)

Results: initial and transition probabilities

- Estimated averaged initial and transition probabilities for NEETs

	NEET		
	$u = 1$	$u = 2$	$u = 3$
$\hat{\pi}_{u,\boldsymbol{x}}$	0.298	0.382	0.320
$\hat{\pi}_{u\|1,\boldsymbol{x}}$	0.499	0.437	0.064
$\hat{\pi}_{u\|2,\boldsymbol{x}}$	0.117	0.794	0.089
$\hat{\pi}_{u\|3,\boldsymbol{x}}$	0.073	0.349	0.578

- The 1st group of countries, about 30% in 2004, shows low values of this rate

- We estimate that from 2005 to 2019, the 44% of countries move from the 1st (best) to the 2nd (intermediate) cluster

Results: initial and transition probabilities

- Estimated averaged initial and transition probabilities for YU

	YU		
	$u = 1$	$u = 2$	$u = 3$
$\hat{\pi}_{u,\boldsymbol{x}}$	0.399	0.490	0.111
$\hat{\pi}_{u\|1,\boldsymbol{x}}$	0.833	0.086	0.081
$\hat{\pi}_{u\|2,\boldsymbol{x}}$	0.204	0.687	0.109
$\hat{\pi}_{u\|3,\boldsymbol{x}}$	0.067	0.523	0.410

- We notice that the 1st group of countries, about 40% in 2004, characterized by the lowest average value of YU, shows the highest persistence probability (0.83) from 2005 to 2019

Results: the effect of the macroeconomics variables

- Estimates of the logit regression parameters of the initial probabilities to the other latent states with respect to the 1st state for NEET and YU (the reference state is the first, (significant †at 10%, *at 5%, **at 1%)

	NEET		YU	
Effect	$u = 2$	$u = 3$	$u = 2$	$u = 3$
intercept	62.066**	-46.67**	-35.828**	0.383**
dgdp	-0.395	0.857	0.055	0.642
part	0.392	-0.692	-0.551	-1.108
temp	-0.078	-0.148	-0.214	0.088
hrst	-0.525	0.263	0.023	-2.175
pedu	-1.756	-4.268**	0.553	-0.505
el	0.122	-0.107	0.311	-2.560
lci	-	-	1.234†	0.831

- At the initial period, countries with high values of *pedu* tend to belong to the 1st group with respect to the 3rd for NEETS
- Countries with high values of *lci* tend to belong to the 2nd group of intermediate countries

Results: the effect of the macroeconomics variables

- Estimates of the logit regression parameters of the transition probabilities

NEET						
Effect $\bar{u}, u$	1-2	1-3	2-1	2-3	3-1	3-2
intercept	31.554	25.793	-27.816	15.005	-28.32†	-7.297
dgdp	-0.090	-0.377	0.039	-0.097	0.037	0.041
part	-0.293	-1.525	-0.188	0.005	-0.219	-0.116
temp	-0.072	1.289	-0.045	-0.046	-0.015	0.027
hrst	-0.679	0.199	0.677	-0.190	0.142	0.047
pedu	-0.181	-2.240†	-0.011	-0.170	1.396	0.338
el	0.396	1.363	0.170	-0.070	-0.988	-0.101
YU						
Effect $\bar{u}, u$	1-2	1-3	2-1	2-3	3-1	3-2
intercept	12.534	63.777**	-13.345†	39.615*	-50.314**	-15.700
dgdp	-0.012	-0.590*	0.045*	-0.345*	0.074	0.143
part	0.007	-0.772	-0.208*	0.142	-2.117	-0.230
temp	0.131	-2.541*	-0.311*	0.235*	-0.310	-0.114
hrst	-0.440**	0.747	0.522*	-0.151	1.520	0.318
pedu	0.041	-1.407	-0.156	-0.163	-0.569	0.726
el	-0.149	-0.187	-0.072	-0.112	1.563	-0.070
lci	-0.124	0.188	-0.099	-0.212	-0.265	-0.461

Results: the effect of the macroeconomics variables

- The odds to transit from the best (1st) to the vulnerable group (3rd) is equal to exp(-2.24) = 0.11 for a point increase in *pedu*, all the other macroeconomic variables held fixed

- An increase of people employed in science and technology (*hrst*) disfavors the transition from the 1st (best) to the 2nd (intermediate) group, all the other variables held fixed (the odds is exp(-0.44) = 0.64), and it favors the transition from 2nd to the 1st group (the odds is exp(0.52) = 1.68)

- Temporary contracts (*temp*) disfavor the transition from the 1st to the 3rd group, thus revealing as important factors for virtuous countries

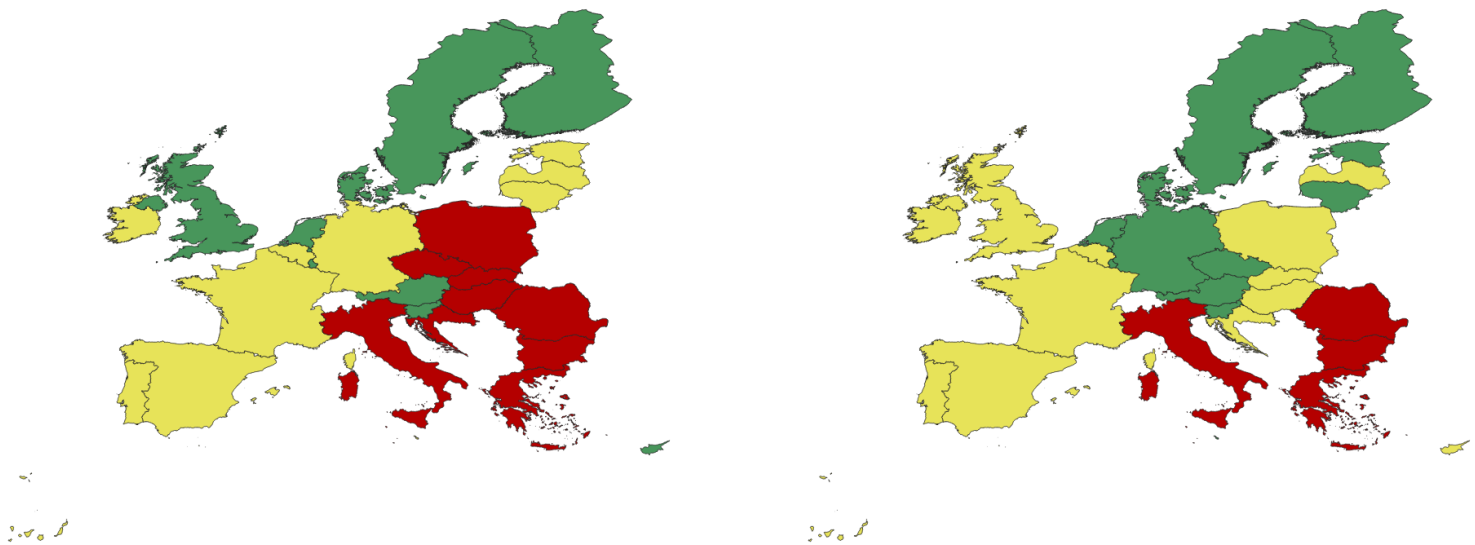
Results: local decoding for NEETs

- We predict the patterns of each country over time through the most likely sequence of latent states on the basis of the maximum posterior probabilities bounds
- For NEET we estimate that ten countries held their position in the same group all over the period, six of which belong to the 1st group
- The countries improving their position from the vulnerable (3rd) up to the best (1st) cluster are CZ, DE, EE, LT, and MT

Results: local decoding for YU

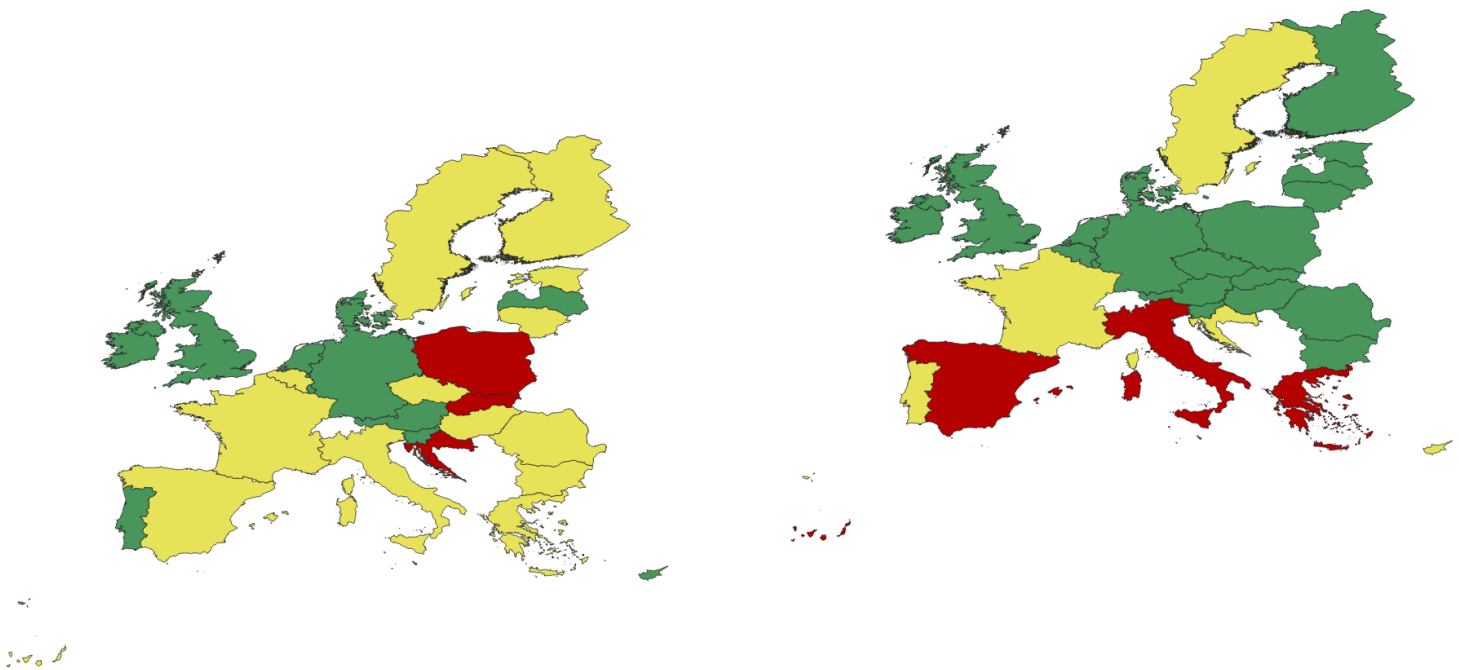
- For YU we estimate that eight countries held their position in the same group all over the period
- The countries improving their position from the worst (3rd) or the intermediate (2nd) group up to the best (1st) are BG, CZ, EE, LT, HU, PL, RO, SK, and FI
- We show the country maps depicting the decoded states for NEETs and YU in 2004 (top panel) and in 2019 (bottom panel) with colors green 1st state (best), yellow 2nd state (intermediate), red 3rd state (vulnerable)

Results: decoded state for NEETs 2004 and 2019



- In 2004 BG, CZ, EL, IT, HU, PL, RO, SK are allocated in the 3rd cluster (in red) of vulnerable countries, and at the end of 2019, only BG, IT, EL and RO, remain in this cluster

Results: decoded state for YU 2004 and 2019



- In 2004, countries allocated in the vulnerable group are HR, PL, and SK, and in 2019 they are EL, ES, and IT

Conclusions

- We propose an approach based on a hidden Markov model for continuous longitudinal data to characterize NEETs and YU shares among countries in a dynamic fashion
- We consider some macroeconomic variables which may influence the countries allocation to the latent states
- The proposed approach allows us to characterize differences among countries and to evaluate changes while comparing the countries' dynamics
- We observe that countries with coordinated markets which transfer a significant part of gross domestic product to active labor market policies are best performing

Main References

- Pennoni, F., Bal-Domańska, B. (2021). NEETs and youth unemployment: a longitudinal comparison across European countries. *submitted*
- Bartolucci F, Farcomeni A, Pennoni F. *Latent Markov Models for Longitudinal Data*. Boca Raton, FL: Chapman and Hall/CRC. 2013.
- Bartolucci, F., Pandolfi, S., and Pennoni, F. (2017). LMest: An R package for latent Markov models for longitudinal categorical data. *Journal of Statistical Software*, **81**:1-38
- Dempster, A. P., Laird, N. M., Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm (with discussion). *Journal of the Royal Statistical Society B*, **39**: 1–38.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**: 461-464.
- Zucchini W, MacDonald IL, Langrock R. *Hidden Markov Models for Time Series: An Introduction Using R*. New York: Springer-Verlag. 2017.