# Latent Mixed Markov Models for the Production of Population Census Data on Employment

D.Filipponi, R.Varriale

Statistics Sweden and Örebro University Summer School
August 22 2023

# outline

# General Context

- Starting from 2018, the Italian Statistical Institute (ISTAT) has initiated a **Permanent Population Census** to provide annual updates of traditional Census data.

- The Permanent Census is designed to ensure the regular estimate of population counts and key socio-economic indicators in accordance with Eurostat's census regulations, covering various geographical levels.

# General Context

Permanent Census has been possible through the integration of information from different data sources focused on the same target phenomena:

- different and detailed administrative data sources
- Permanent Census survey planned to collect all the mandatory information (according to Eurostat requirements) and to overcome the deficiencies of administrative sources
- other Social Survey

# Available information on employment

In this work we deal with the estimation of the employment status of the Italian resident over 15 years - for different geographical levels - using all the available information

- Labour Force Survey data (LFS)
- Permanent Census data (PC)
- Labour Register (LR)

# Available information on employment: LFS

- Continuous survey carried out during every week of the year
- In each quarter, information is collected about approximately 1.2% of the overall Italian population
- Two-stage sampling design (municipalities, households) with stratification of first-stage units, rotated panel scheme
- Output: quarterly estimates of the main aggregates of labour market, by gender, age and territory

# Available information on employment: PC

- The PC does not involve anymore all Italian households, but only a sample of them
- Every year collects information on about 1,400,000 resident households located in 2,800 Italian municipalities
- Two-stage sampling design (municipalities, households) with stratification of first-stage units

# Available information on employment: LR

Italian Labour Register is realised in ISTAT trough the integration of different administrative sources (AD)

- Social Security data
- Fiscal data
- Italian Statistical Business Register

# Available information on employment: LR

Data are organized in an information system having a linked employer-employees structure.

- Statistical unit LR: job position (individual + employer)
- Population LR: regular job positions of all individuals that have at least an administrative data source signal

From LR, information on the employment status of each individual is derived consistently with the International Labour Organization (ILO) definitions.

# Available information on employment

Table 1: Available information on employment: Administrative Data (AD), Labour Force Survey (LFS) and Permanent Census (PC)

| Unit | AD | | | LFS | | | | PC |
|---|---|---|---|---|---|---|---|---|
| | Week 1 | ... | Week 52 | Week 1 | Week 2 | ... | Week 52 | Week 40 |
| 1 | X | X | X | . | X | . | . | X |
| 2 | X | X | X | X | . | . | . | . |
| 3 | X | X | X | . | X | . | . | X |
| 4 | X | X | X | . | . | X | X | X |
| 5 | X | X | X | . | . | . | X | . |
| 6 | X | X | X | . | X | . | X | X |
| 7 | X | X | X | X | . | X | . | . |
| 8 | X | X | X | . | . | . | . | X |
| 9 | X | X | X | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . |
| $n$ | X | X | X | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . |
| N | X | X | X | . | . | . | . | X |

# Measurement Error in Data Sources

- Discrepancies between different data sources are typically observed
- Identical units yield different values
- Editing aims to correct erroneous values, but inconsistencies persist due to measurement errors.

# Measurement Error in Data Sources

- Surveys: Inadequate questionnaire design, data collection procedures, interviewer and respondent effects
- Register data: Unique errors like specification error, administrative delay, and errors during data entry.

# Measurement Error in Data Sources

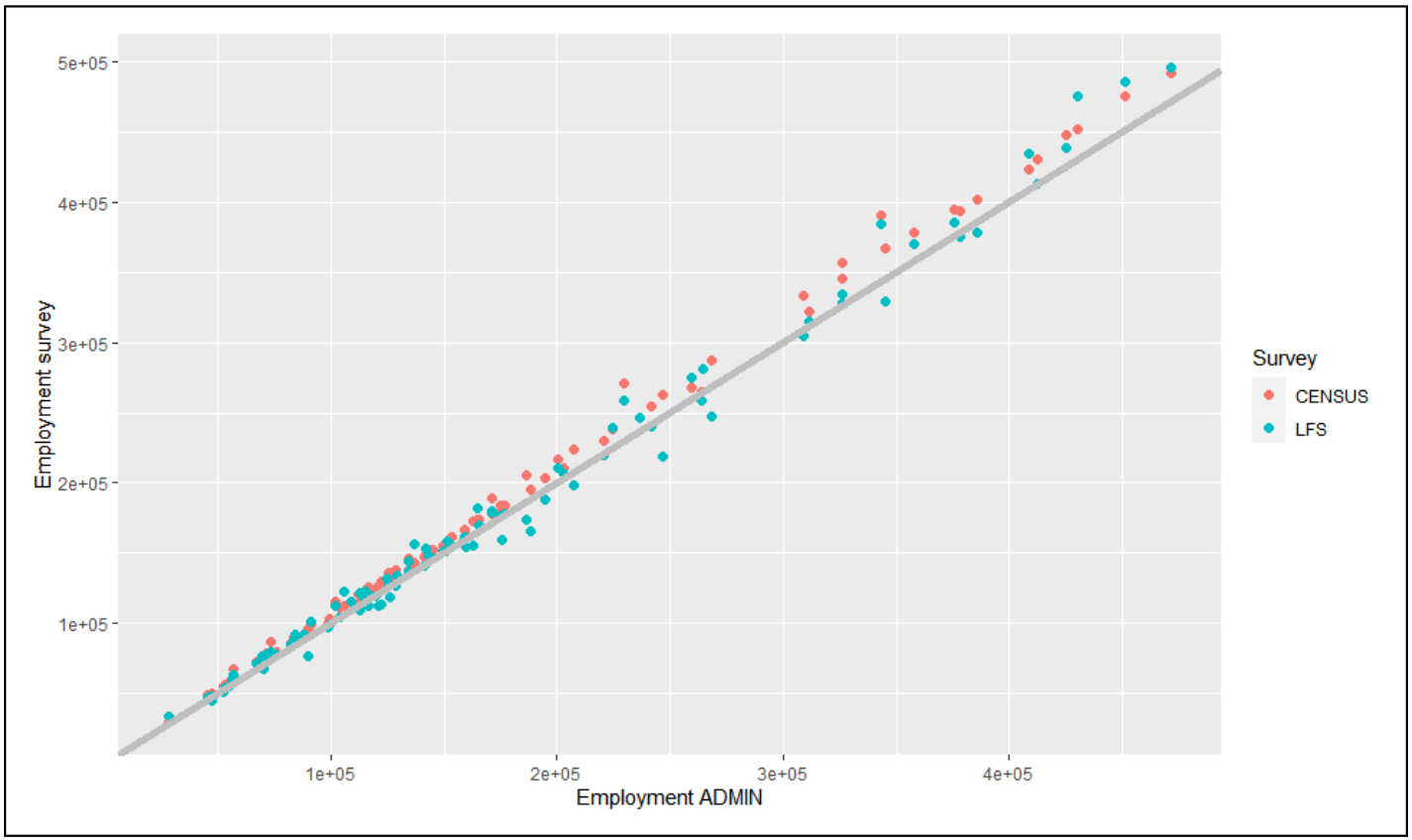
Table 2: Cross-classification of employment status measured by LFS and AD (IV quarters 2019) and Census and AD (October 2019). Italy

| LFS \ AD | Out (Not empl.) | In (Empl.) | Total |
|---|---|---|---|
| Not employed | 61.7 | 2.7 | 64.4 |
| Employed | 3 | 32.6 | 35.6 |
| Total | 64.7 | 35.3 | 100.0 |

| Census \ AD | Out (Not empl.) | In (Empl.) | Total |
|---|---|---|---|
| Not employed | 58.5 | 2.7 | 61.2 |
| Employed | 4 | 34.8 | 38.8 |
| Total | 62.5 | 38.5 | 100.0 |

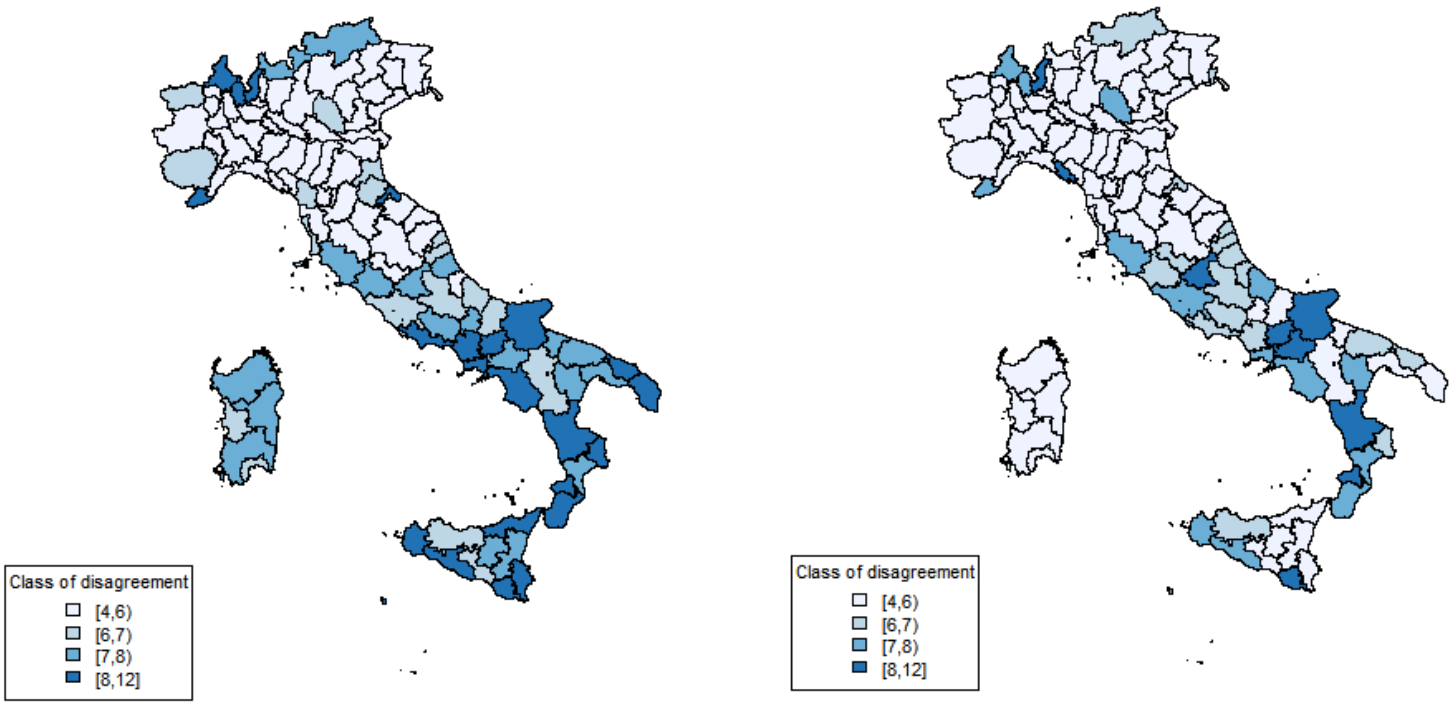
# Measurement Error in Data Sources

Figure 1: Employment LFS, Census, and AD by province. Year 2019



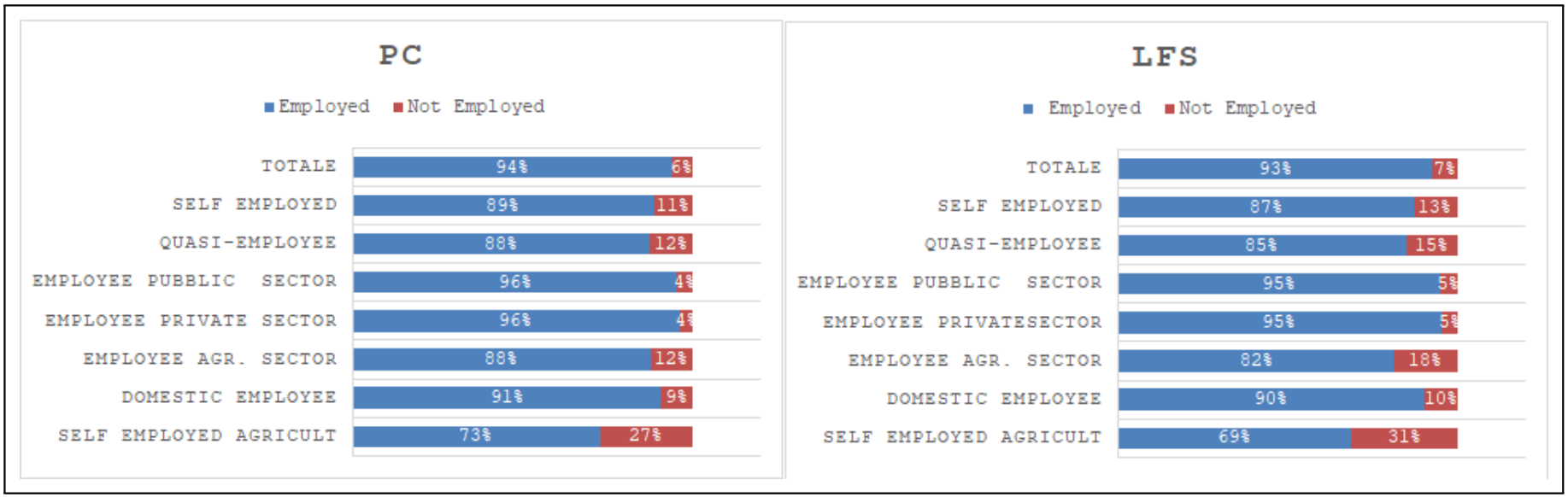
# Measurement Error in Data Sources

Figure 2: Employment measures by Census/LFS and out of Admin sources



# Measurement Error in Data Sources

Figure 3: Employment measured by Admin sources and missed by surveys by type of administrative source. Year 2019



# Methodological Setup

- Handle the multisource information
- Employ a unit-level modeling approach
- Operate in a non-supervised manner: each source is an imperfect representation of the target phenomenon, with none of the sources being entirely errors free
- Account for the longitudinal structure of the data

# LMMs for measurement errors

Latent Markov models (LMMs) fall within the broader category of latent class models (LCMs) and offer a compelling approach to reconciling inconsistent data sources in official statistics

- LMMs enable the estimation and correction of classification errors without requiring error-free benchmarking data
- LMMs incorporate multiple indicators, allowing simultaneous error correction across all available sources
- LMMs are designed to estimate and correct measurement errors in longitudinal categorical data

# LMMs for measurement errors

Basic setup of LMMs:

- Latent model: it exists an unobserved (latent) true path $L$ on $T$ time points, which follow a (first-order) Markov process with $L$ states

$$P(L) = P(L_1, \ldots, L_T) = P(L_1)P(L_2|L_1)\ldots P(L_T|L_{T-1})$$

- Measurement model: the observed variable, $Y_t$, $t = 1, .., T$, is independently generated at each time point $t$, from the true value $L_t$. The generation process is governed by the probability

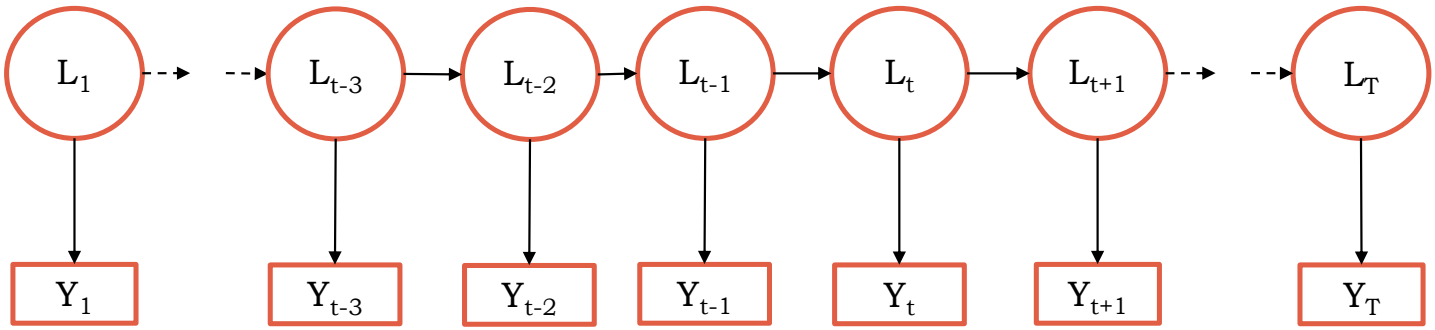
$$P(Y_t|L_t)$$

# LMMs for measurement errors

Combining the assumptions regarding the latent and the measurement model, the observed path are marginalised over the true data:

$$P(Y) = \sum_{l_1=1}^{L} \sum_{l_1=1}^{L} \dots \sum_{l_T=1}^{L} P(L_1) \prod_{t=2}^{T} P(L_t/L_{t-1}) \prod_{t=0}^{T} P(Y_t/L_t)$$

# LMMs for measurement errors: Basic Setup



# LMMs for Employment Estimation: Formalisation

## Unit of time

- $t \in (1, \ldots, 12)$

## Observed Variables:

- $Y_{1:T}^{(1)}$: binary vector of (possibly missing) values of the employment status at times $1, \ldots, T$ resulting from the LFS
- $Y_{1:T}^{(2)}$: binary vector of values of the employment status at times $1, \ldots, T$ resulting from AD
- $Y_{1:T}^{(3)}$: binary vector of values of the employment status at time $t = 10$ resulting from the Census

# LMMs for Employment Estimation: Formalisation

Observed Variables:

- $Q$, $S$: covariates
- $Q$: retirement, student, earnings, age, gender, municipalities, ...
- *Source* (AD): No source; Employees; Self-employers (time information); Self-employers (no time information)

# LMMs for Employment Estimation: Formalisation

## Latent variables

- $L_{1:T} = (L_1, \ldots, L_t)$: latent variable vector measuring the employment scores over time
- $X$ discrete (latent) random effect: is used to capture the population heterogeneity

  x=1 : Never employed (no labour force)
  x=2 : always employed (permanent jobs)
  x=3 : moving between employment and unemployment

- $L_{1:T}$ is first order Markov Chain conditional on $X$
- the different components of $X$ represent the different trajectories of $L_{1:T}$.

# LMMs for Employment Estimation: Formalisation

Main elements of the latent model:

- $\tau_j^x \doteq P(L_1 = j|X = x)$: initial probabilities of $L_{1:T}$
- $\pi_{k|j}^x \doteq P(L_t = k|L_{t-1} = j, X = x)$: transition probabilities of $L_{1:T}$
- $\phi_{x|q,s}$: conditional probability of $X$ - $P(X = x|Q = q, S = s)$

where $x = 1, 2, 3$, $j, k \in \{0, 1\}$, and $(q, s)$ are realizations from the variables $Q$ and $S$

Constraints:

- sub-population $X = 2, 3$, $L$ is Markov Chain
- sub-population $X = 1$, $L$ is degenerate with $\tau_1^1 = 0$, $\pi_{1|0}^1 = 0$

# LMMs for Employment Estimation: Formalisation

Main elements of the measurement model:

- $\psi^{(1)}_{j|i} \doteq P(Y^{(1)}_t = j | L_t = i)$
- $\psi^{(2)}_{j|i,s} \doteq P(Y^{(2)}_t = j | L_t = i, S = s)$
- $\psi^{(3)}_{j|i} \doteq P(Y^{(3)}_t = j | L_t = i)$

where $t = 1, \ldots, 12;\ (i, j) \in \{0, 1\}; s \in \{1, 2, 3, 4\}$

Constraints:

"no false positive" data are present in the LFS data: $\psi^{(1)}_{1|0} = 0$

AD measurement error linked to the "quality" of the source: $\psi^{(2)}_{1|i,1} = 0$ for $i = 1, 2$

# LMMs for Employment Estimation: Formalisation

The probability distribution of the manifest variables conditional on the covariates can be obtained by marginalizing with respect to the latent variables as follows:

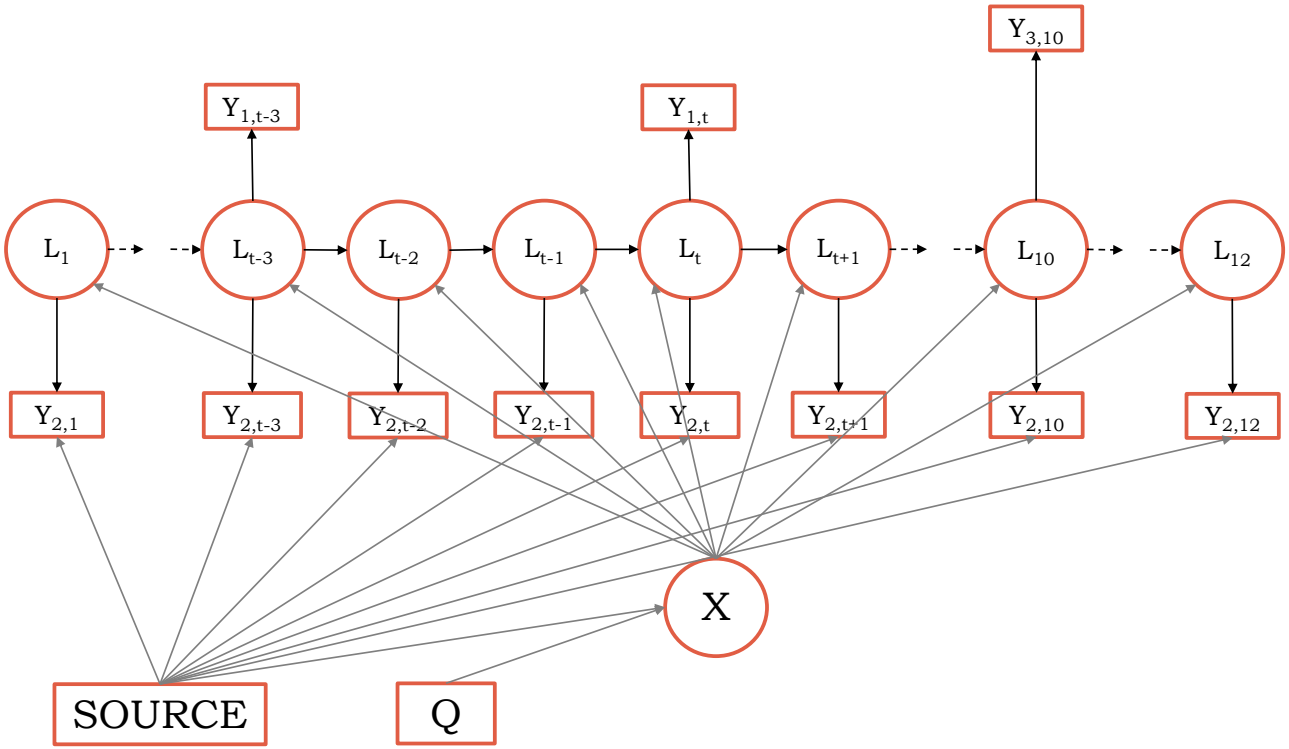
$$\mathrm{P}(Y_{1:T}^{(1)} = \mathbf{u},\ Y_{1:T}^{(2)} = \mathbf{v},\ Y_{1:T}^{(3)} = \mathbf{z}|\ Q = q, S = s) =$$

$$\sum_{x=1}^{3}\sum_{l_1=1}^{2}\sum_{l_2=1}^{2}\cdots\sum_{l_T=1}^{2}\phi_{x|q,s}\ \tau_{l_1}^{x}\ \prod_{t=2}^{T}\pi_{l_t|l_{t-1}}^{x}\prod_{t=1}^{T}\left(\psi_{u_t|l_t}^{(1)}\right)^{\delta_t^1}\psi_{v_t|l_t,s}^{(2)}\left(\psi_{z_t|l_t}^{(3)}\right)^{\delta_t^2}$$

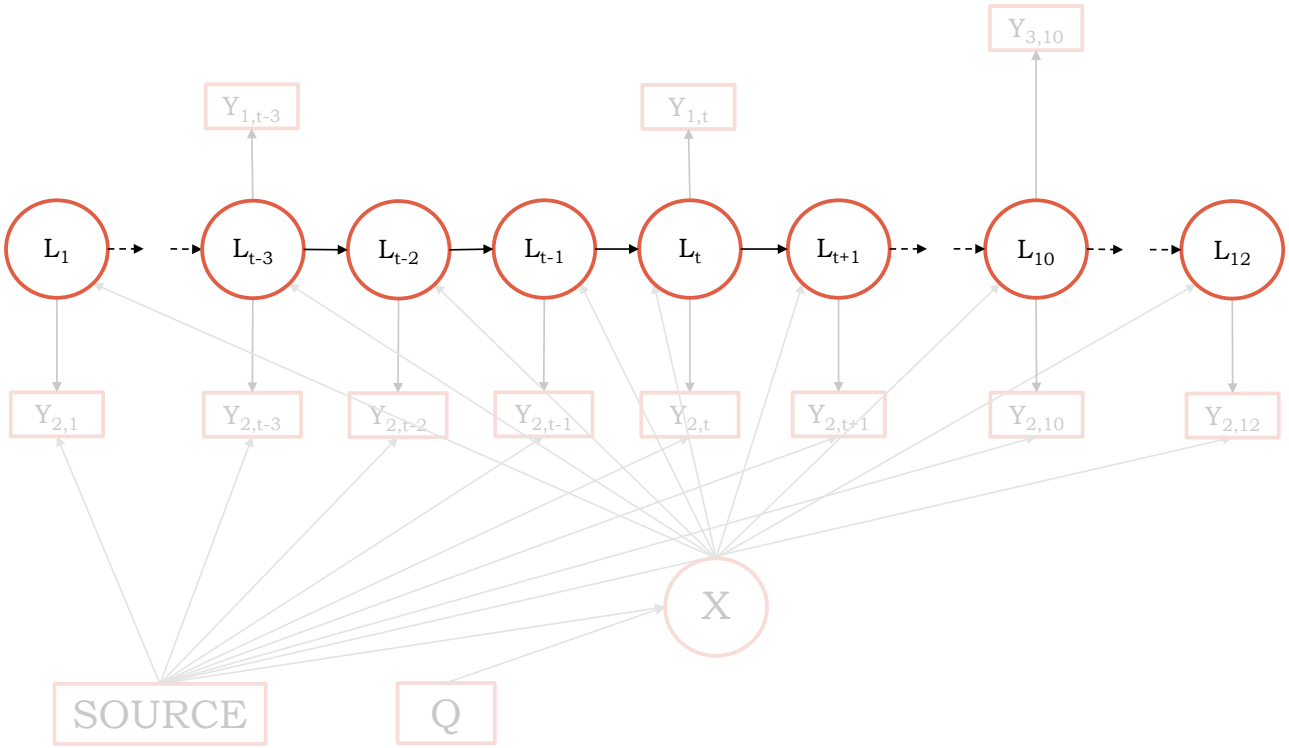
where:

- $\delta_t$ are indicators variable taking value 1 if at time $t$ the LFS/Census measure are available and 0 otherwise
- $\mathbf{u} = (u_1, \ldots, u_T)$, $\mathbf{v} = (v_1, \ldots, v_T)$, $\mathbf{z} = (z_1, \ldots, z_T)$ are realizations from the random vector $Y_{1:T}^{(1)}$, $Y_{1:T}^{(2)}$, $Y_{1:T}^{(3)}$, respectively

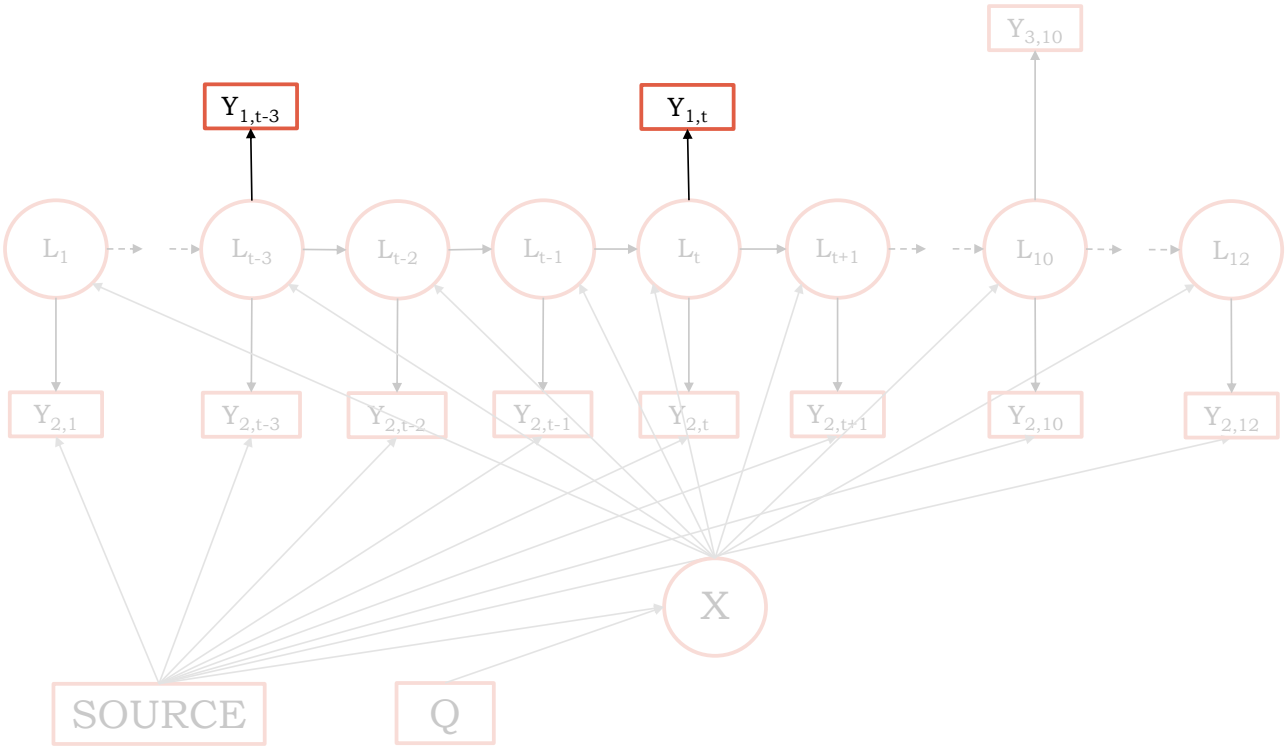
# LMMs for Employment Estimation: Graphical representation



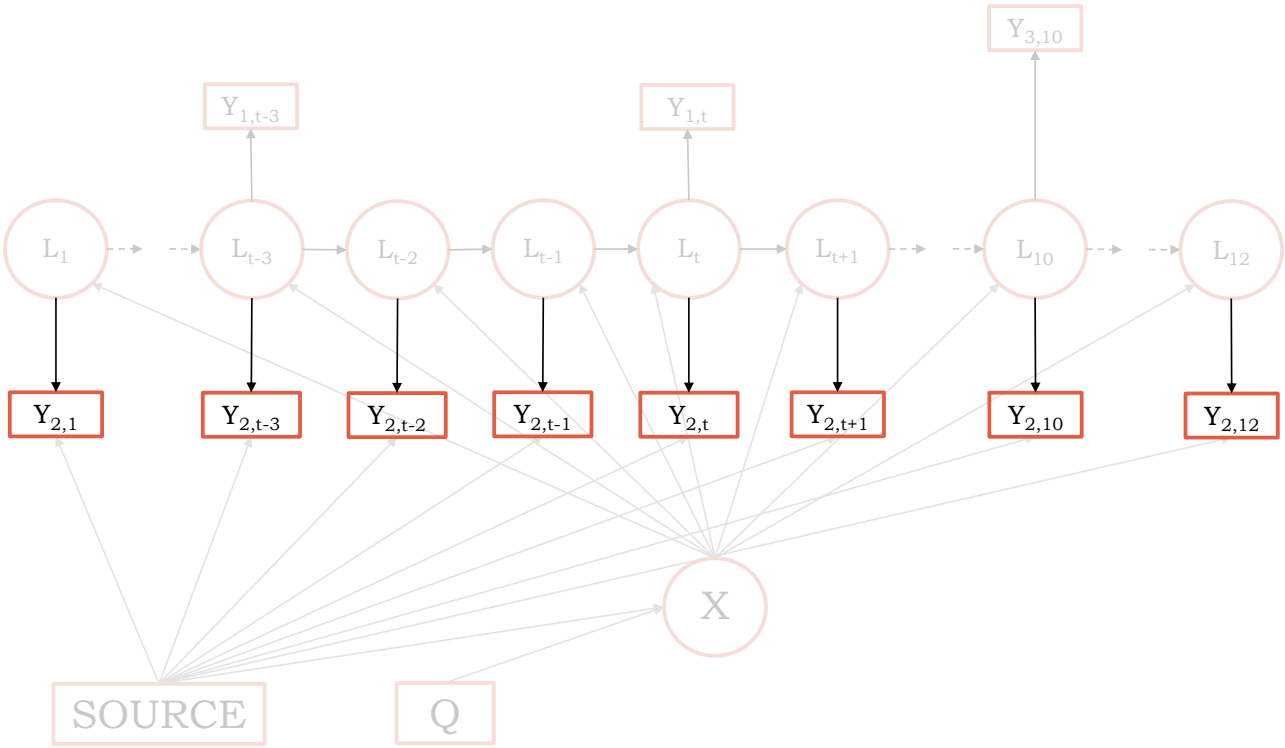
# Graphical representation: latent variables $L$



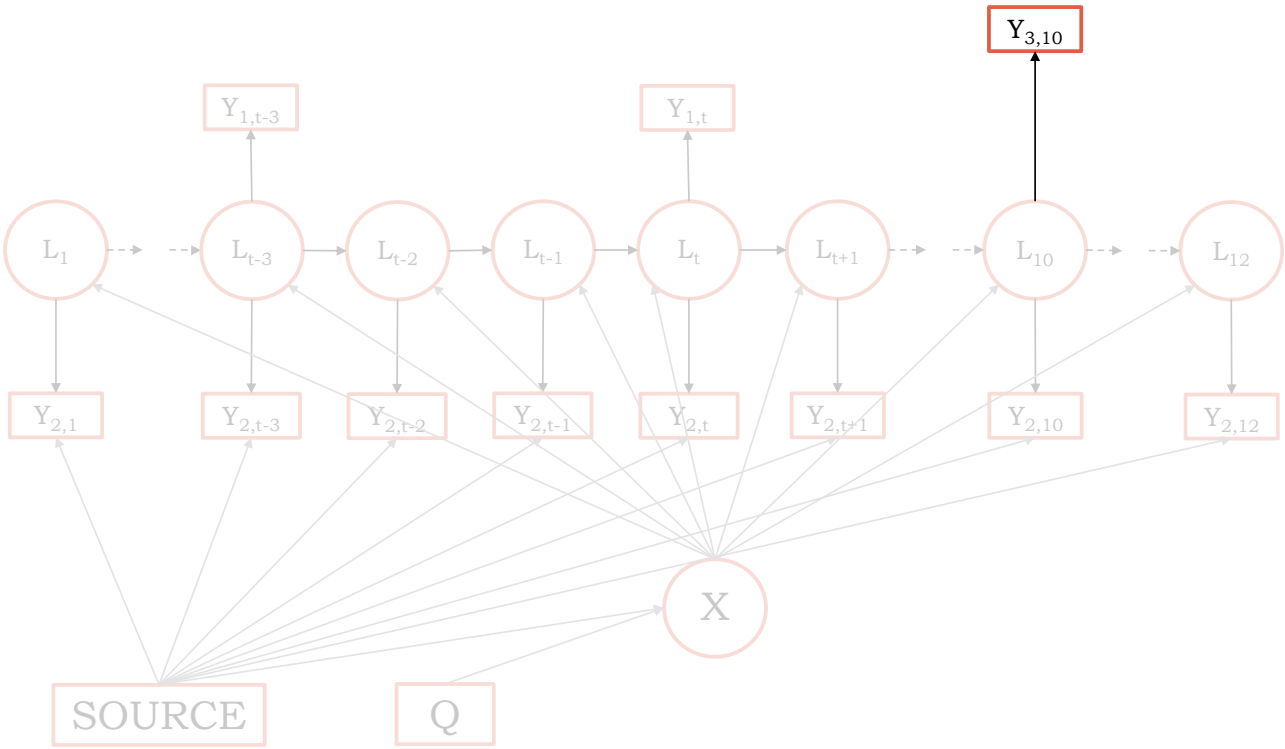
# Graphical representation: manifest variable LFS



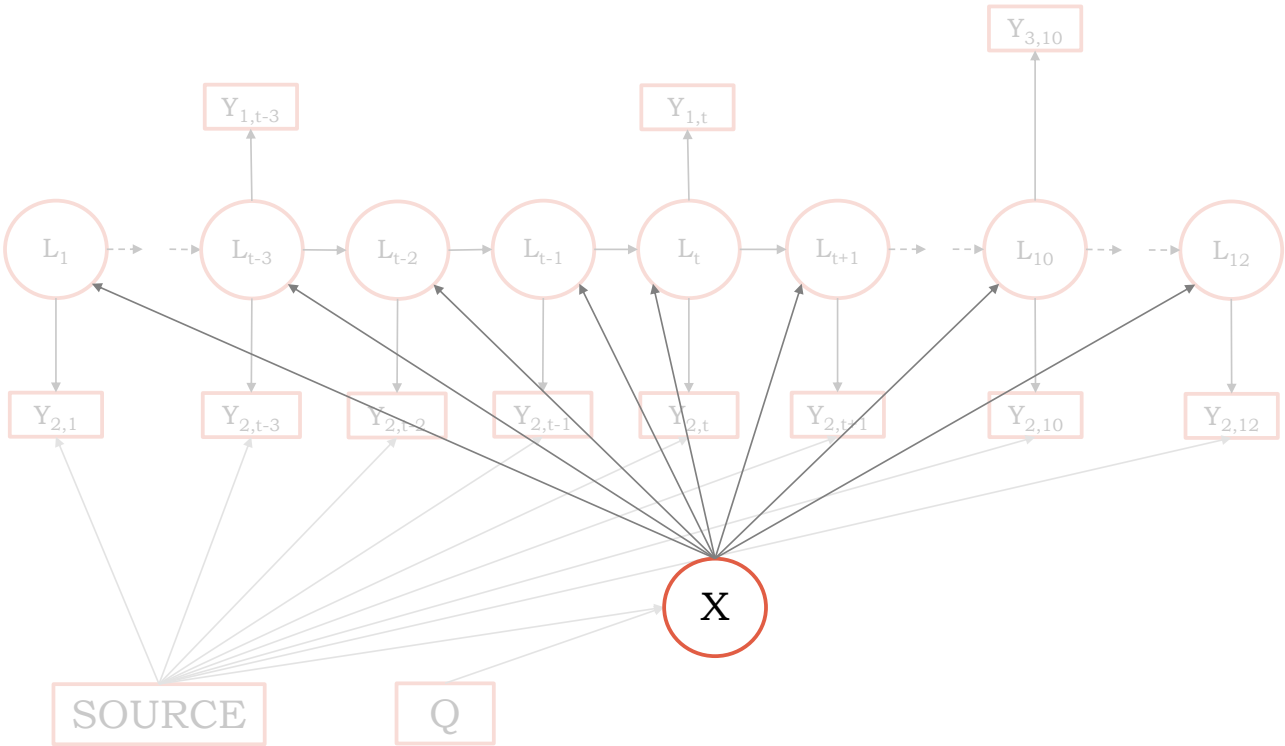
# Graphical representation: manifest variable AD



# Graphical representation: manifest variable PC

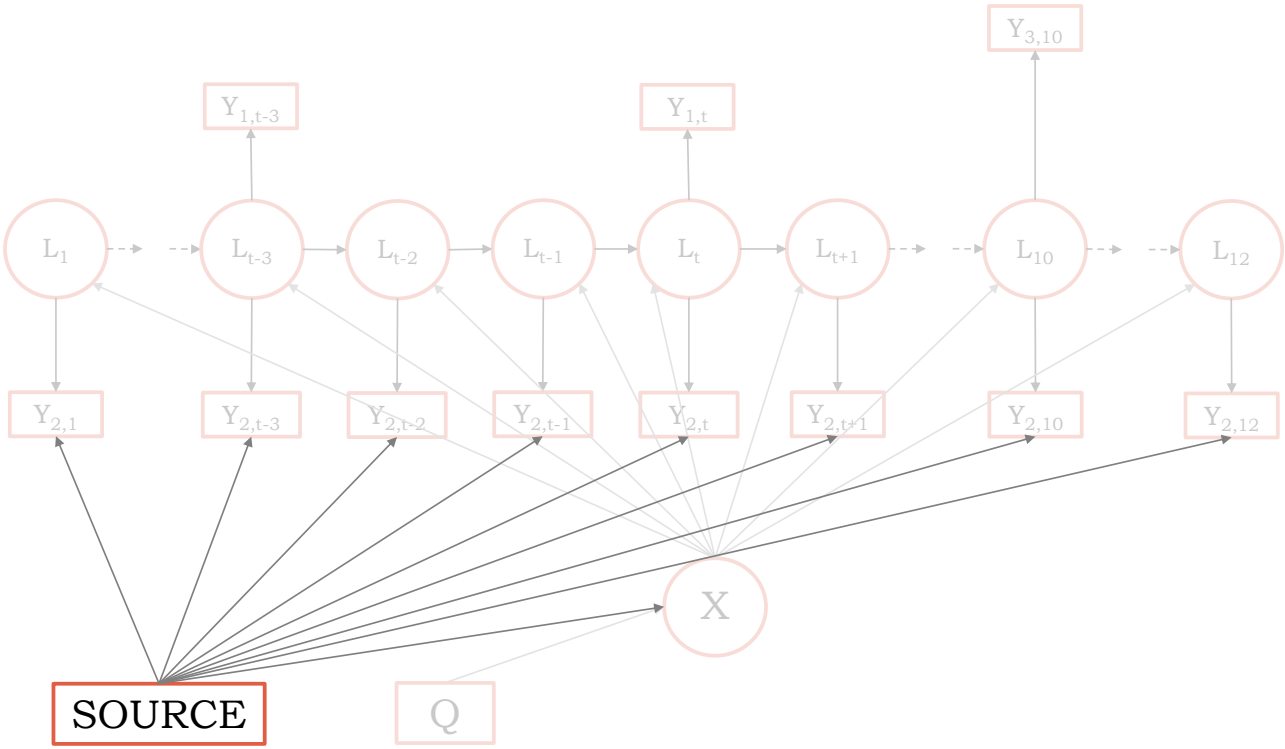


# Graphical representation: latent variable $X$



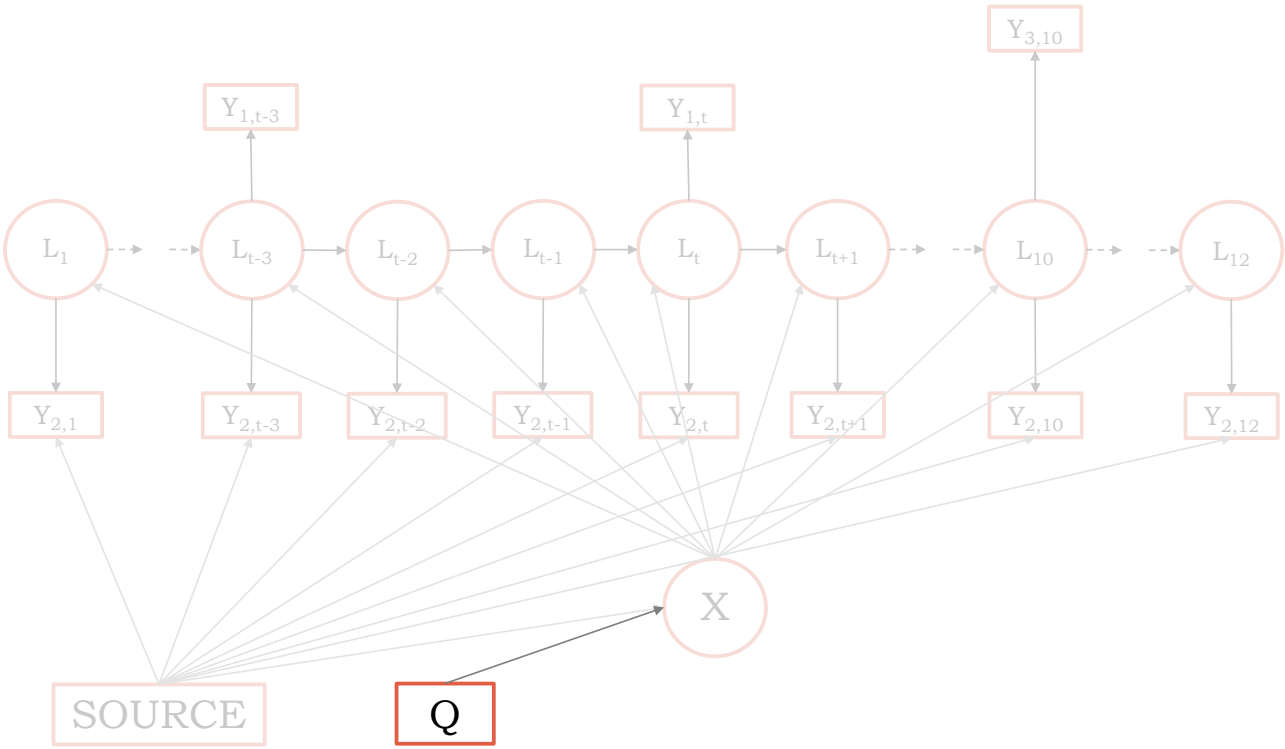

1= Never employed, 2= Stayers (employed), 3= Movers

# Graphical representation: covariate *Source*



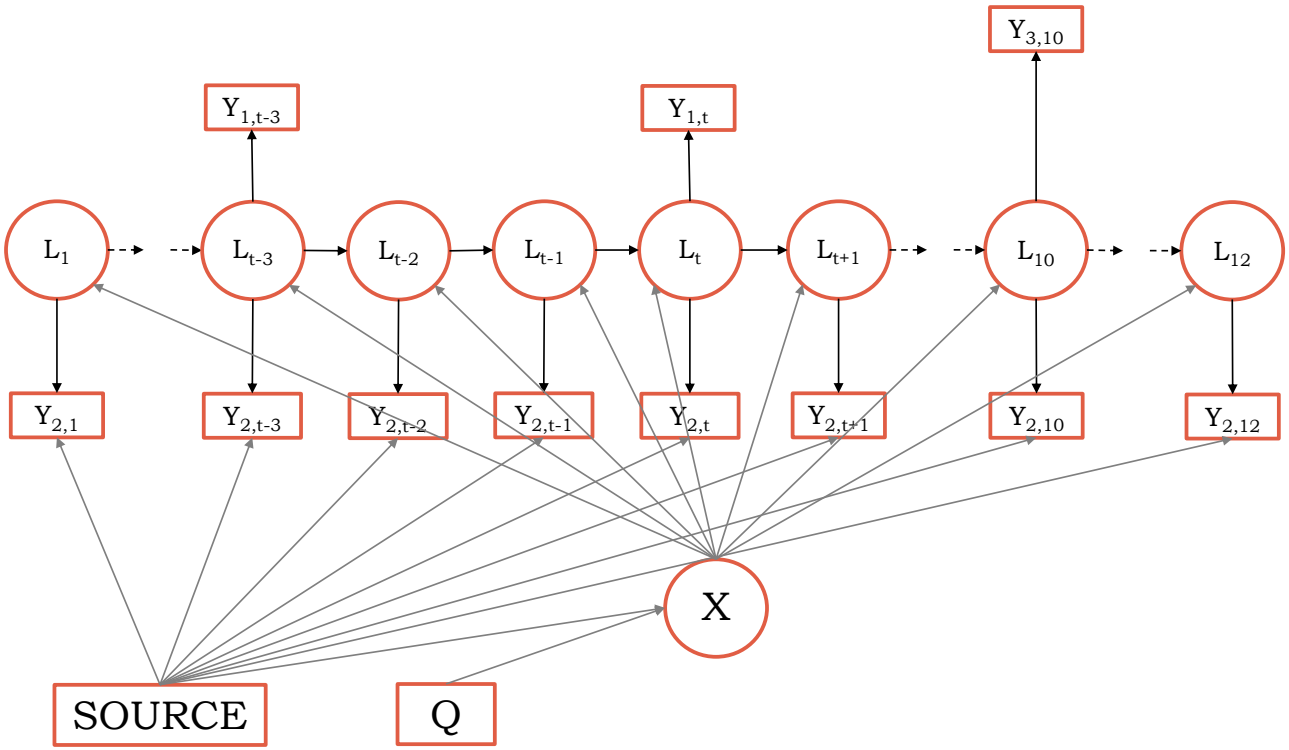

1= No source, 2= Employees, 3= Self-employers (time), 4= Self-employers (no time)

# Graphical representation: covariate $Q$




$Q$ : retirement, student, earnings, age, gender

# Graphical representation: complete model



# LMMs for Employment Estimation: Parameters Estimation

- For each each region a different model has been estimated using LFS and Census surveyed data and admin data
- ML estimates are obtained using EM algorithm. In order to be able to deal with applications involving large numbers of time points, the E-step computations use a the Baum-Welch algorithm, originally proposed by Baum et al. (1970)
- The final models have been chosen among different model using model selections criteria.

# LMMs for Employment Estimation: Results (selected)

Table 3: Classification errors (LFS), $\psi^{(1)}_{j|i}$

| $L_t$ | $Y_t^{(1)}$ 0 | 1 |
|---|---|---|
| 0 | 1.00 | 0.00 |
| 1 | 0.07 | 0.93 |

Table 4: Classification errors (PC), $\psi^{(3)}_{j|i}$

| $L_t$ | $Y_t^{(3)}$ 0 | 1 |
|---|---|---|
| 0 | 0.98 | 0.02 |
| 1 | 0.06 | 0.94 |

LF and PC surveyed data, Marche. Year 2019

# LMMs for Employment Estimation: Results (selected)

Table 5: Classification errors (AD), $\psi^{(2)}_{j|i,s}$

| Source | $L_t$ | $Y_t^{(2)}$ 0 | $Y_t^{(2)}$ 1 |
|---|---|---|---|
| No source | 0 | 1.00 | 0.00 |
| | 1 | 1.00 | 0.00 |
| Employees | 0 | 0.99 | 0.01 |
| | 1 | 0.01 | 0.99 |
| Self-employers (time information) | 0 | 0.99 | 0.01 |
| | 1 | 0.01 | 0.99 |
| Self-employers (no time information) | 0 | 0.98 | 0.02 |
| | 1 | 0.00 | 1.00 |

LF and PC surveyed data, Marche. Year 2019

# LMMs for Employment Estimation: Results (selected)

Table 6: Latent class size, $\phi_{x|q,s}$; Initial probabilities, $\tau_j^x$

| X | $P(X = x)$ | $L_1$ | |
|---|---|---|---|
| | | 0 | 1 |
| Never employed | 0.52 | 1 | 0 |
| Stayers (employed) | 0.39 | 0.01 | 0.99 |
| Movers | 0.09 | 0.73 | 0.27 |

LF and PC surveyed data, Marche. Year 2019

# LMMs for Employment Estimation: Results (selected)

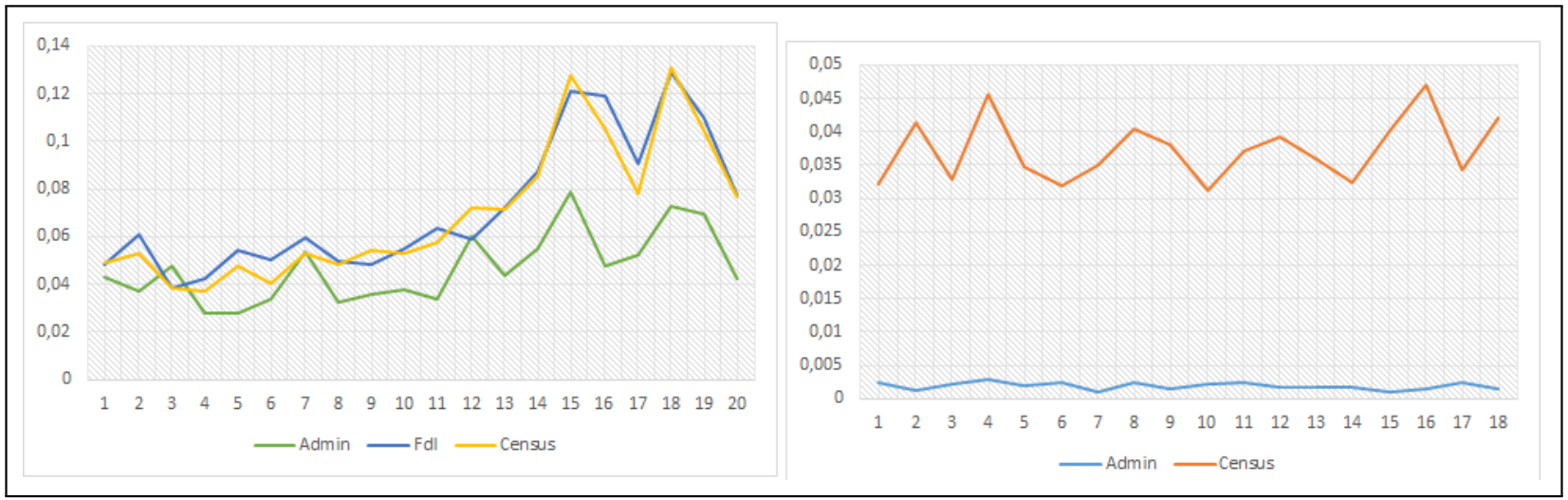
Table 7: Latent class size, $\phi_{x|q,s}$; Transition probabilities, $\pi^x_{k|j}$

| X | $P(X = x)$ | $L_{t-1}$ | $L_t$ 0 | $L_t$ 1 |
|---|---|---|---|---|
| Never employed | 0.52 | 0 | 1 | 0 |
| | | 1 | 0.5 | 0.5 |
| Stayers (employed) | 0.39 | 0 | 0.88 | 0.12 |
| | | 1 | 0.01 | 0.99 |
| Movers | 0.09 | 0 | 0.82 | 0.18 |
| | | 1 | 0.18 | 0.82 |

LF and PC surveyed data, Marche. Year 2019

# LMMs for Employment Estimation: Results (selected)

Figure 4: Measurement model parameters by regions. $\psi_{0|1}^{(k)}$ and $\psi_{1|0}^{(k)}$ for $k = 1, 2, 3$



# LMMs for Employment Estimation: predictions

Forecasts of the true variable for individual units of the complete population have been randomly generated from the marginal posterior distribution

$$\mathrm{P}(L_t = y | Y^{(1)}_{1:T} = \mathbf{u},\ Y^{(2)}_{1:T} = \mathbf{v},\ Y^{(3)}_{1:T} = \mathbf{z}\ Q = q, S = s) =$$
$$\frac{\mathrm{P}(L_t = y,\ Y^{(1)}_{1:T} = \mathbf{u},\ Y^{(2)}_{1:T} = \mathbf{v},\ Y^{(3)}_{1:T} = \mathbf{z}|\ Q = q, S = s)}{\mathrm{P}(Y^{(1)}_{1:T} = \mathbf{u},\ Y^{(2)}_{1:T} = \mathbf{v},\ Y^{(3)}_{1:T} = \mathbf{z}|\ Q = q, S = s)}$$

obtained by marginalizing

$$\mathrm{P}(X, L_{1:T},\ Y^{(1,2,3)}_{1:T}) = \phi_{x|q,s}\ \tau^x_{l_1} \prod_{t=2}^{T} \pi^x_{l_t|l_{t-1}} \prod_{t=1}^{T} \left(\psi^{(1)}_{u_t|l_t}\right)^{\delta^1_t} \psi^{(2)}_{v_t|l_t,s} \left(\psi^{(3)}_{z_t|l_t}\right)^{\delta^2_t}$$

An efficient way to compute the posterior membership probabilities is the forward recursion algorithm

# Multiple Imputation using Latent Class Models (MILC)

The uncertainty of the target estimate is measured out using the Multiple Imputation Latent Class (MILC) method. The MILC procedure comprises five steps:

- Generate $m$ nonparametric bootstrap samples from the original data set containing indicators and covariates used for LMM estimation. Each bootstrap sample is obtained by sampling from the observed frequency distribution
- Fit the LMM on each of the $m$ bootstrap samples

# Multiple Imputation using Latent Class Models (MILC)

- Create one imputation for the latent variable $L$ using parameters obtained from the $m$-th bootstrap sample
- Obtain estimates of interest from each imputation
- Pool the estimates obtained from all imputations using the pooling rules defined by Rubin

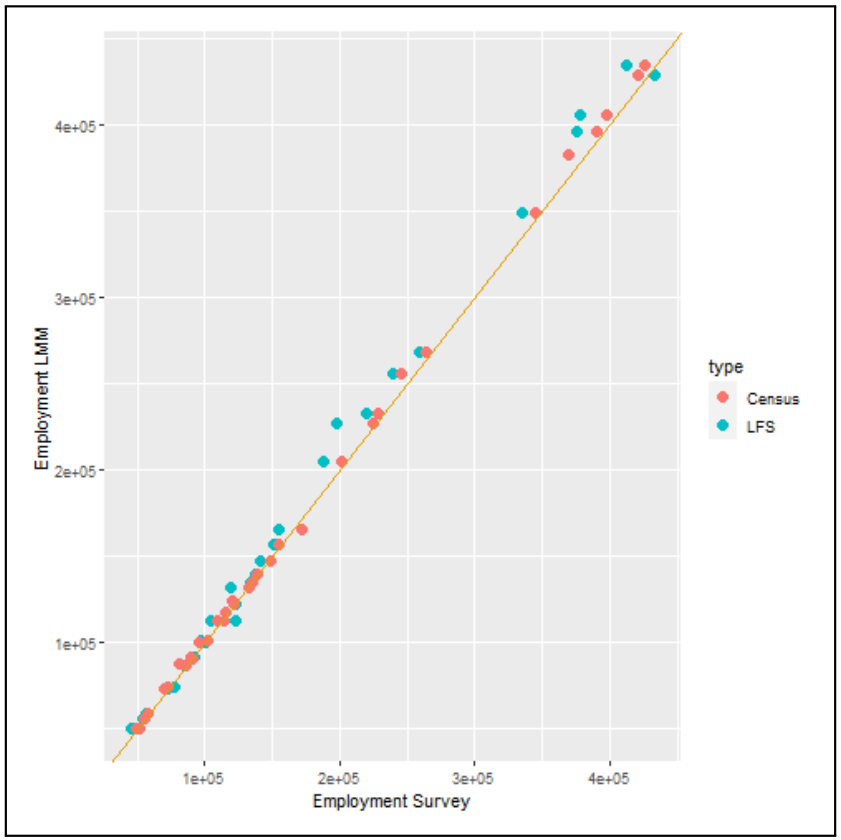
# Results

| region | Employment proportion | cv |
|---|---|---|
| 01 PIEMONTE | 0.425 | 0.0011 |
| 02 VALLE D'AOSTA | 0.443 | 0.0039 |
| 03 LOMBARDIA | 0.442 | 0.0008 |
| 04 TRENTINO A. A. | 0.474 | 0.0014 |
| 05 VENETO | 0.448 | 0.0007 |
| 06 FRIULI V. G | 0.429 | 0.0013 |
| 07 LIGURIA | 0.406 | 0.0015 |
| 08 EMILIA ROMAGNA | 0.450 | 0.0006 |
| 09 TOSCANA | 0.430 | 0.0011 |
| 10 UMBRIA | 0.412 | 0.0023 |
| 11 MARCHE | 0.428 | 0.0015 |
| 12 LAZIO | 0.398 | 0.0013 |
| 13 ABRUZZI | 0.386 | 0.0020 |
| 14 MOLISE | 0.367 | 0.0035 |
| 15 CAMPANIA | 0.320 | 0.0007 |
| 16 PUGLIE | 0.336 | 0.0013 |
| 17 BASILICATA | 0.365 | 0.0026 |
| 18 CALABRIA | 0.317 | 0.0019 |
| 19 SICILIA | 0.302 | 0.0014 |
| 20 SARDEGNA | 0.365 | 0.0014 |

Table 8: Employment Proportion and Coefficient of Variation by Italian Region

# Results

Figure 5: Employment estimate by LFS, Census and LMM by province



# Conclusions and future research

Concerning LMM,

- Introducing random effect into the latent model to account for the variability of small areas
- Improve the evaluation of the accuracy
- Use sampling weights

Concerning Istat production process,

- Analysis of coherence of the results with other short term statistics (like LFS) and national account (for the estimation of non regular jobs)
- Implement a "system" producing statistics on employment, through the use of multiple data sources (survey and administrative data)

# REFERENCES

Bartolucci, F., Farcomeni, A. & Pennoni, F. (2013), Latent Markov Models for Longitudinal Data, Chapman and Hall/CRC press.

Boeschoten, L., Filipponi, D., & Varriale, R. (2020), Combining Multiple Imputation and Hidden Markov Modeling to Obtain Consistent Estimates of Employment Status, Journal of Survey Statistics and Methodology, https://doi.org/10.1093/jssam/smz052

Filipponi, D., Guarnera, U., & Varriale, R. (2021). Latent Mixed Markov Models for the Production of Population Census Data on Employment. In Perna, C., Salvati, N., Schirippa Spagnolo, F. (Eds.) Book of short papers SIS 2021, 112-117, Pearson.

Pavlopoulos, D., & Vermunt, J. (2015). Measuring temporary employment. do survey or register tell the truth? Survey Methodology, 41 (1), 197-214.

Vermunt, J. K., & Magidson, J. (2016). Technical guide for Latent GOLD 5.1: Basic, advanced, and syntax. Statistical Innovations Inc.,