Multi-source data: new approaches for non-standard employment statistics.

The Dutch and Italian experience

D.Filipponi[1], S.Loriga[1], M.Garnier-Villarreal[2,3], D.Pavlopoulos[2,3], R.Varriale[4], R. Stoel[3]

[1]ISTAT, [2]Vrije Univ. Amsterdam, [3]Statistics Netherlands, [4]Sapienza Univ. of Rome

SECOND WORKSHOP ON METHODOLOGIES FOR OFFICIAL STATISTICS, 6*th* of December, 2023

General context

- In official statistics, different data sources are available to measure the same phenomenon and discrepancies between data are typically observed
- Increasingly widespread production of statistics through the use of multi-source data
- Statistics based on multi-source data bring new methodological problems

Research project among Istat, CBS and Vrije Universiteit Amsterdam:

A Life Course Dynamics Approach for Non-Standard Employment
Integration of administrative sources and survey data for the production of labour statistics[1]

[1] *Istat Framework agreement ACP/40/2022; Data have been processed in CBS environment*

General context

- This work explores the impact of measurement errors on employment contract data and mobility trends over time
- Cross-country comparison: we compare Italian and Dutch employment data from 2016 to 2021
- Integrated information from the Labour Force Survey and the Employment Register
- Multiple-group hidden Markov models to estimate error-corrected employment trajectories to:
  - analyze the impact of measurement errors within each country
  - facilitate meaningful cross-country comparisons (analyse variation in some parameters of the hidden Markov model between countries)

Available information on employment: data sources

Both in Italy and The Netherlands,

- Labour Force Survey (LFS): survey data
- Employment Register (ER): administrative data

LFS

Main common characteristics:

- Continuous survey, carried out during all the year
- Interviews are referred to all the weeks of each quarter
- Representative of the national population aged 15 and older
- Quarterly rotation scheme

Differences

- Sampling design
- Panel waves (see Figures)

Data structure, IT

	2017				*2018*				*2019*			
LFS sample group	quarter 1	quarter 2	quarter 3	quarter 4	quarter 1	quarter 2	quarter 3	quarter 4	quarter 1	quarter 2	quarter 3	quarter 4
1	x	x	.	.	x	x	.	.	.	.	.	.
2	.	x	x	.	.	x	x	.	.	.	.	.
3	.	.	x	x	.	.	x	x	.	.	.	.
4	.	.	.	x	x	.	.	x	x	.	.	.
5	.	.	.	.	x	x	.	.	x	x	.	.
6	.	.	.	.	.	x	x	.	.	x	x	.
7	.	.	.	.	.	.	x	x	.	.	x	x
8	.	.	.	.	.	.	.	x	x	.	.	x
9	.	.	.	.	.	.	.	.	x	x	.	.
10	.	.	.	.	.	.	.	.	.	x	x	.
11	.	.	.	.	.	.	.	.	.	.	x	x
12	.	.	.	.	.	.	.	.	.	.	.	x

Data structure, NL

	2017				*2018*				*2019*			
LFS sample group	**quarter 1**	**quarter 2**	**quarter 3**	**quarter 4**	**quarter 1**	**quarter 2**	**quarter 3**	**quarter 4**	**quarter 1**	**quarter 2**	**quarter 3**	**quarter 4**
1	x	x	x	x	x	.	.	.	.	.	.	.
2	.	x	x	x	x	x	.	.	.	.	.	.
3	.	.	x	x	x	x	x	.	.	.	.	.
4	.	.	.	x	x	x	x	x	.	.	.	.
5	.	.	.	.	x	x	x	x	x	.	.	.
6	.	.	.	.	.	x	x	x	x	x	.	.
7	.	.	.	.	.	.	x	x	x	x	x	.
8	.	.	.	.	.	.	.	x	x	x	x	x
9	.	.	.	.	.	.	.	.	x	x	x	x
10	.	.	.	.	.	.	.	.	.	x	x	x
11	.	.	.	.	.	.	.	.	.	.	x	x
12	.	.	.	.	.	.	.	.	.	.	.	x

Employment Register (ER) data, Italy

- Managed internally by Istat
- It is built by integrating administrative data collected mainly by social security and tax authorities
- Harmonized data is organized with an employer-employee linkage structure, representing the basis for extracting information about the "worker" coherent with the International Labour Office (ILO) definitions
- It contains weekly and monthly information, depending on the original data sources

Employment Register (ER) data, The Netherlands

- Dutch ER is administered by the Institute for Employee Insurance
- It contains information on the labor market and income for all insured workers in The Netherlands
- It contains monthly information, but employers typically submit relevant data only once a year

Data

- We dispose of two distinct indicators related to the employment contract, derived respectively from LFS and ER, for Italy and the Netherlands
- Both indicators have three different levels:
  - PE: employees with permanent contracts
  - FT: employees with fixed-term contracts
  - OT: others, such as self-employed and non-employed individuals
- The analysis is based on quarterly data and focuses on individuals aged between 25 and 55
- To reduce the size of the dataset, a 10% sample of units was randomly selected from the original data in both countries. The sample was stratified by the month of the first interview in LFS to ensure participation from all LFS cohorts.

Data

- Additional information: gender, educational level, and whether the interview was conducted via proxy, etc.
- Italian data
  - From 2017 to 2021, with 20 data points
  - The number of LFS interviews conducted by each individual is maximum 4
- Dutch data
  - From 2016 to 2019, with 16 data points
  - The number of LFS interviews conducted by each individual is maximum 5

Data

Table: Distribution of employment categories in ER and LFS and discrepancy between the two sources. Italy, years 2017-2021

Employment contract LFS \ ER	Permanent	Temporary	Other	All
		Row Percentages		
Permanent	91.07	2.73	6.19	100.00
Temporary	20.48	60.28	19.24	100.00
Others	5.36	3.26	91.38	100.00
All	44.67	8.34	47.00	100.00
		Column Percentages		
Permanent	90.51	15.08	6.01	45.08
Temporary	4.07	66.59	3.74	9.03
Others	5.42	18.33	90.25	45.89
All	100.00	100.00	100.00	100.00

The overall missclassification rate is 11.6%

Data

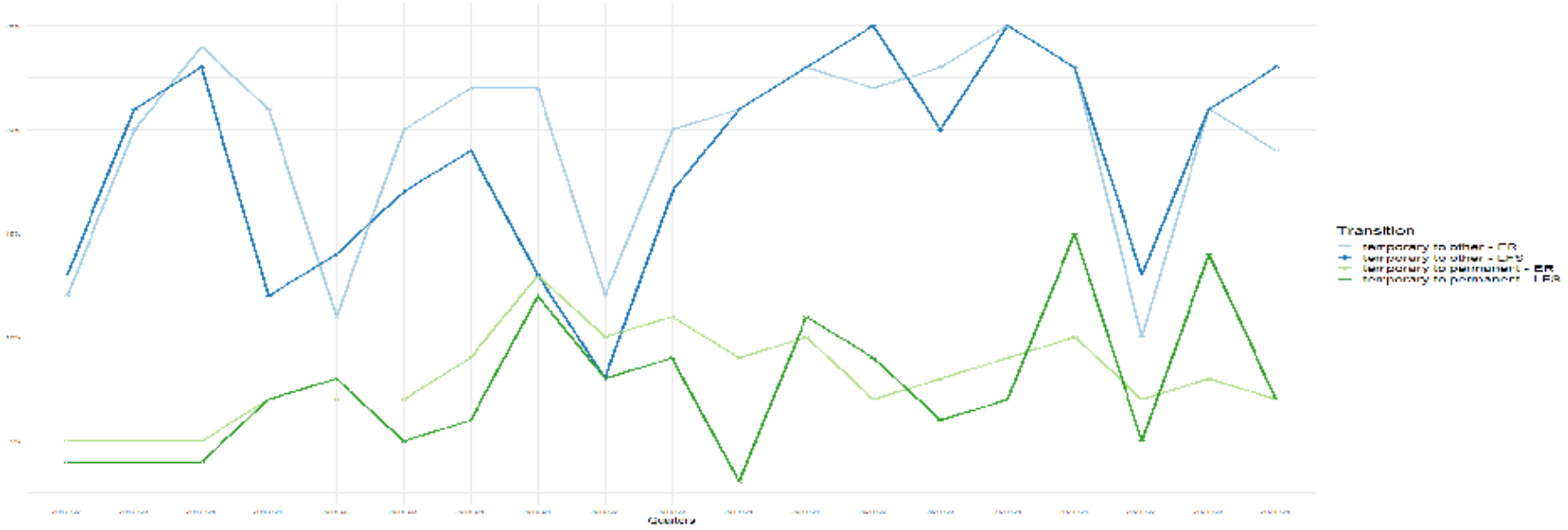
Table: Distribution of employment categories in ER and LFS and Discrepancy between the two sources. The Netherlands, years 2016-2019

Employment contract LFS \ ER	Permanent	Temporary	Other	All
		Row Percentages		
Permanent	78.84	13.95	7.21	100.00
Temporary	10.21	78.09	11.70	100.00
Others	3.38	4.07	92.55	100.00
All	47.90	18.27	33.83	100.00
		Column Percentages		
Permanent	95.86	45.56	14.64	63.80
Temporary	2.37	48.72	4.54	10.83
Others	1.77	5.72	80.82	25.37
All	100.00	100.00	100.00	100.00

The overall missclassification rate is 17.6%

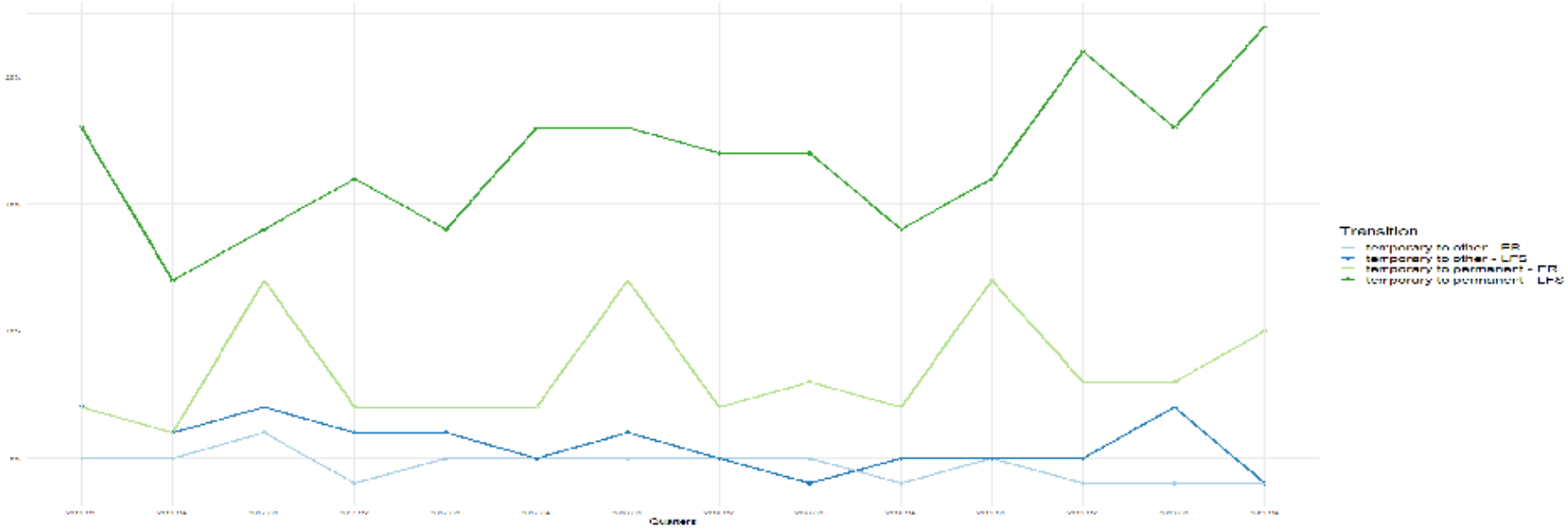
Data

Figure: Observed transition flow in LFS data and ER data by quarters. Italy, years 2017-2021



Data

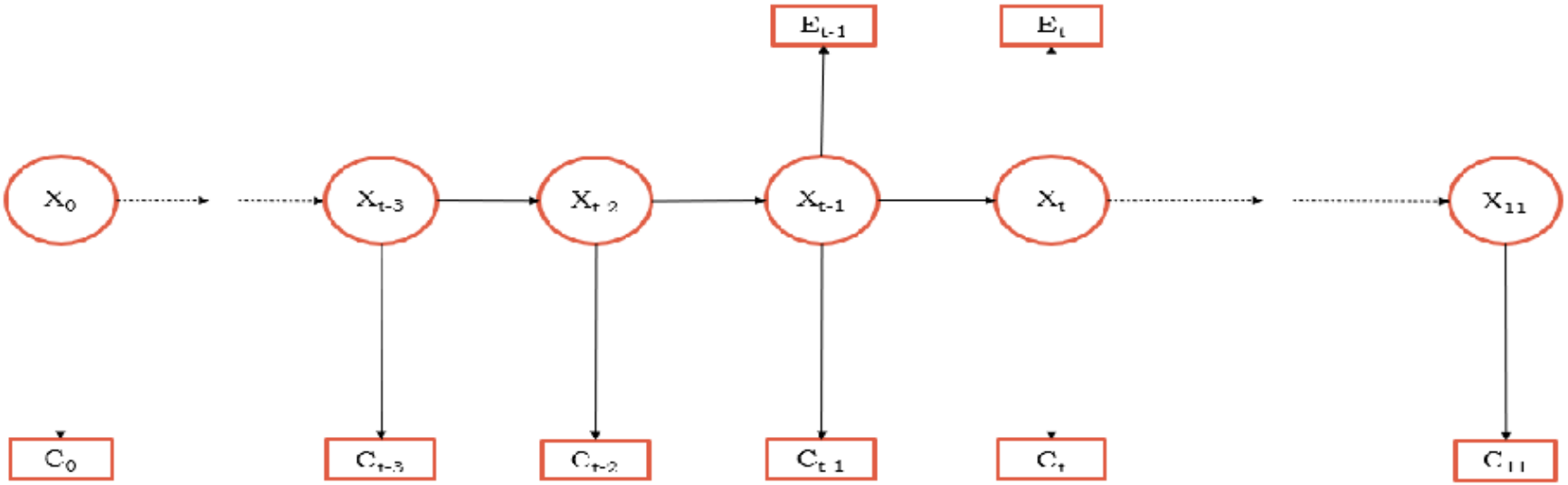
Figure: Observed transition flow in LFS data and ER data by quarters. The Netherlands, years 2016-2019



The hidden Markov model (HMM)

- HMMs to estimate the employment status, by taking into account the longitudinal structure of the data and the deficiencies in the measurement process of both LFS and AD
- The individual employment status is considered as a discrete latent variable, having 3 categories: employees with permanent contract (PE), employees with fixed-term contract (FT), individuals not included in paid employment (OT, self-employed or not employed)
- The measurement process is described by the distributions of the manifest variable conditional on the latent variable
- In the recent literature, parallel work on Italian and Dutch data

Basic HMM, graphical representation (one country)



Model extensions

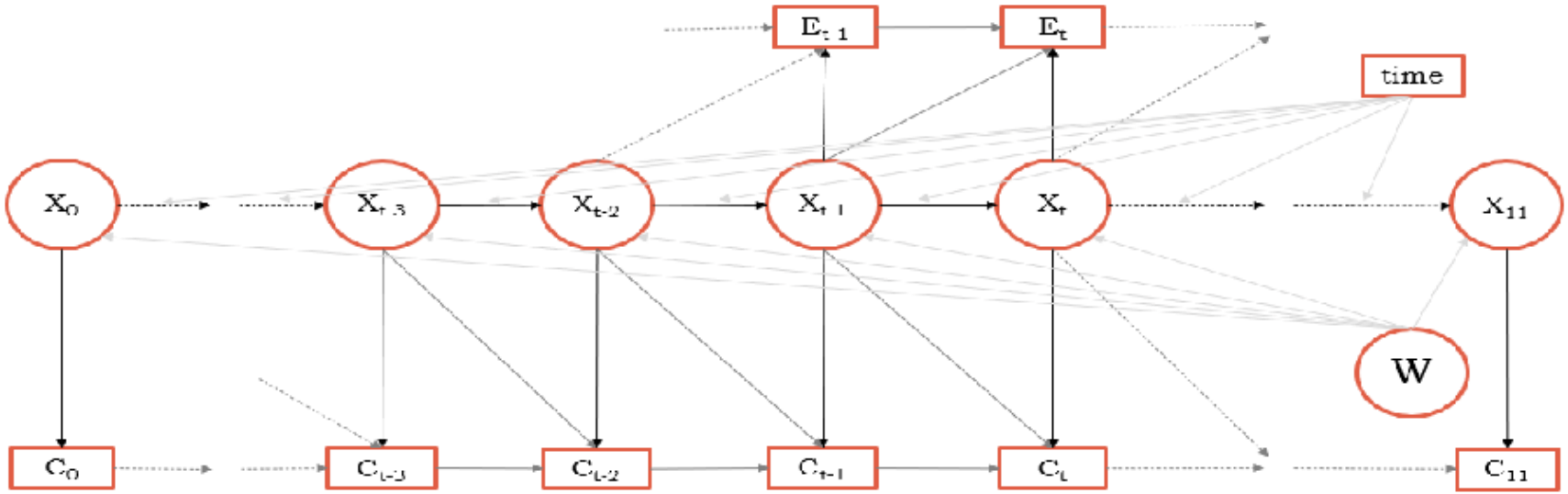
- We allow the latent transition probabilities to be time-heterogeneous (introducing a dependence on a quadratic specification of time)
- We relax the basic assumption of the Independent Classification Error (ICE), i.e. the observed states are independent of one another within and between time points.

Model extensions

We relax ICE in two ways:

- We allowed the response from the survey to depend on covariates
- We allow the error probabilities in both the survey and the register data to depend on the lagged observed and lagged true contract type
  - Same error: we estimate an extra error parameter for the case where an error was made in $t-1$ and it can be repeated in t
  - One error: we estimate an extra error parameter when an error was made in $t-1$

Extended HMM, graphical representation (one country)



Extented HMM (one country)

$$P(\mathbf{C}_i = \mathbf{c}_i, \mathbf{E_i} = \mathbf{e}_i | t, \mathbf{W_i}) = \sum_{x_0=1}^{3} \sum_{x_1=1}^{3} \dots \sum_{x_T=1}^{3} P(X_{ik0} = x_0 | \mathbf{W_i})$$

$$\prod_{t=1}^{T} P(X_{it} = x_t | X_{i(t-1)} = x_{t-1}, t, t^2, \mathbf{W_i})$$

$$\prod_{t=0}^{T} P(C_{it} = c_t | X_{it} = x_t, X_{i(t-1)} = x_{t-1}, C_{i(t-1)} = c_{t-1})^{\delta_{it}^1}$$

$$\prod_{t=0}^{T} P(E_{it} = e_t | X_{it} = x_t, X_{i(t-1)} = x_{t-1}, E_{i(t-1)} = e_{t-1})^{\delta_{it}^2}$$

Extented HMM (one country)

$$P(\mathbf{C}_i = \mathbf{c}_i, \mathbf{E_i} = \mathbf{e}_i | t, \mathbf{W_i}) = \sum_{x_0=1}^{3} \sum_{x_1=1}^{3} \dots \sum_{x_T=1}^{3} P(X_{ik0} = x_0 | \mathbf{W_i})$$

$$\prod_{t=1}^{T} P(X_{it} = x_t | X_{i(t-1)} = x_{t-1}, t, t^2, \mathbf{W_i})$$

$$\prod_{t=0}^{T} P(C_{it} = c_t | X_{it} = x_t, X_{i(t-1)} = x_{t-1}, C_{i(t-1)} = c_{t-1})^{\delta_{it}^1}$$

$$\prod_{t=0}^{T} P(E_{it} = e_t | X_{it} = x_t, X_{i(t-1)} = x_{t-1}, E_{i(t-1)} = e_{t-1})^{\delta_{it}^2}$$

Extented HMM (one country)

$$P(\mathbf{C}_i = \mathbf{c}_i, \mathbf{E_i} = \mathbf{e}_i | t, \mathbf{W_i}) = \sum_{x_0=1}^{3} \sum_{x_1=1}^{3} \dots \sum_{x_T=1}^{3} P(X_{ik0} = x_0 | \mathbf{W_i})$$

$$\prod_{t=1}^{T} P(X_{it} = x_t | X_{i(t-1)} = x_{t-1}, t, t^2, \mathbf{W_i})$$

$$\prod_{t=0}^{T} P(C_{it} = c_t | X_{it} = x_t, X_{i(t-1)} = x_{t-1}, C_{i(t-1)} = c_{t-1})^{\delta_{it}^1}$$

$$\prod_{t=0}^{T} P(E_{it} = e_t | X_{it} = x_t, X_{i(t-1)} = x_{t-1}, E_{i(t-1)} = e_{t-1})^{\delta_{it}^2}$$

The multiple-group HMM

- Multiple group HMM: a statistical technique that allows researchers to investigate differences across groups, by enabling specification of HMM with group-specific or equal parameter estimates across groups
- Two extremes:
  - Baseline: equal model parameters across groups
  - Unrestricted: model parameters governing the initial state probabilities, latent transition probabilities, and measurement error probabilities are considered specific to each group k

The multiple-group HMM

- Quarterly data from 2017 to 2019 for both countries
- Estimates of the relevant model parameters can be obtained via Maximum likelihood estimation using the Expectation-Maximization algorithm
- Software Latent GOLD v.6.0 (Vermunt and Magidson, Upgrade Manual for Latent GOLD Basic, Advanced, Syntax, and Choice. Version 6.0, Statistical Innovations Inc., 2021)

Model selection

The final model selection occurred in two steps

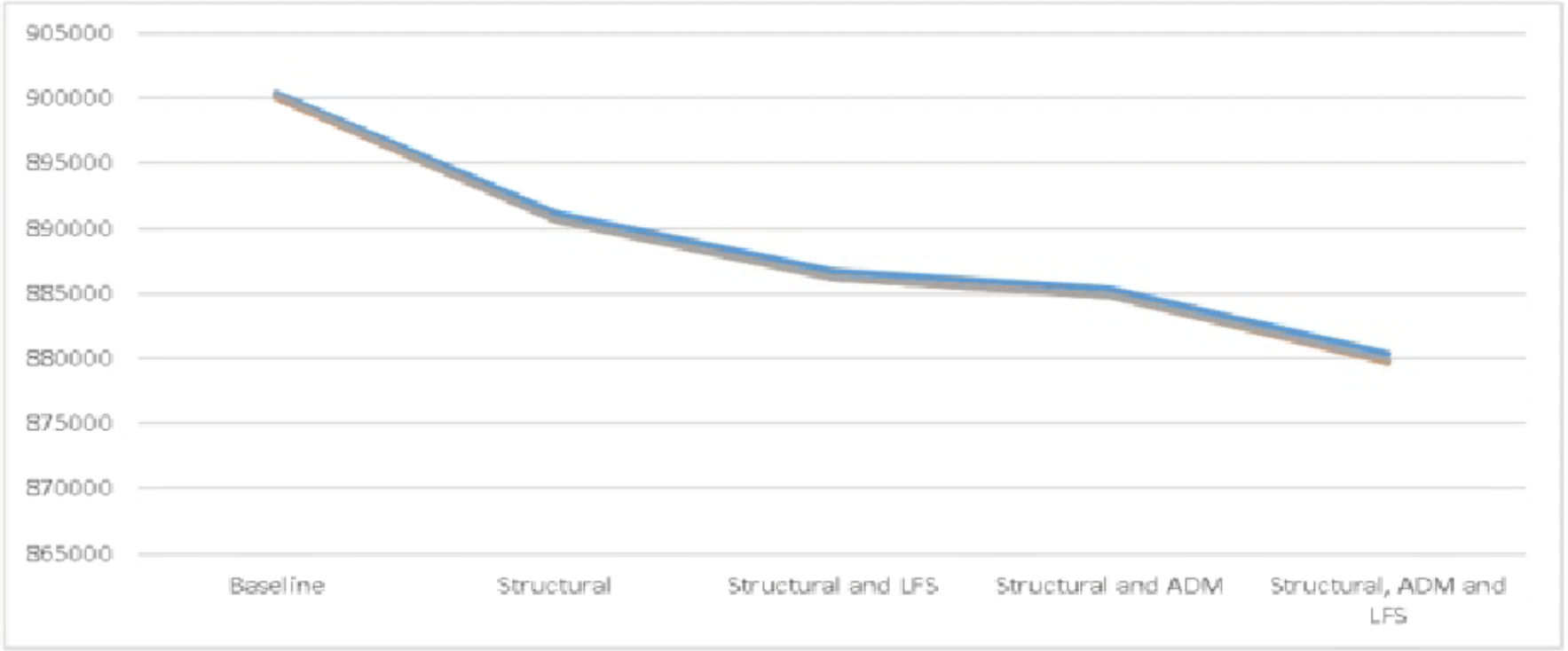
1. We chose the model based on parameter invariance across the two countries
2. We focused on specifying the measurement error component

Model selection, step 1

We considered 5 models:

- (a) Baseline: invariance of parameters in both the structural model and the measurement error part
- (b-e) Heterogeneity in the structural model *plus*
  - (b) invariance in the measurement error of both indicators (LFS, ER)
  - (c) invariance in measurement error of only ER indicator
  - (d) invariance in measurement error of only LFS indicator
  - (e) heterogeneity in measurement error of both indicators

Model selection, step 1



BIC % change	-1.0%	-0.5%	-0.2%	-0.6%
AIC % change	-1.1%	-0.5%	-0.2%	-0.6%
AIC3 % change	-1.0%	-0.5%	-0.2%	-0.6%

Model selection, step 2

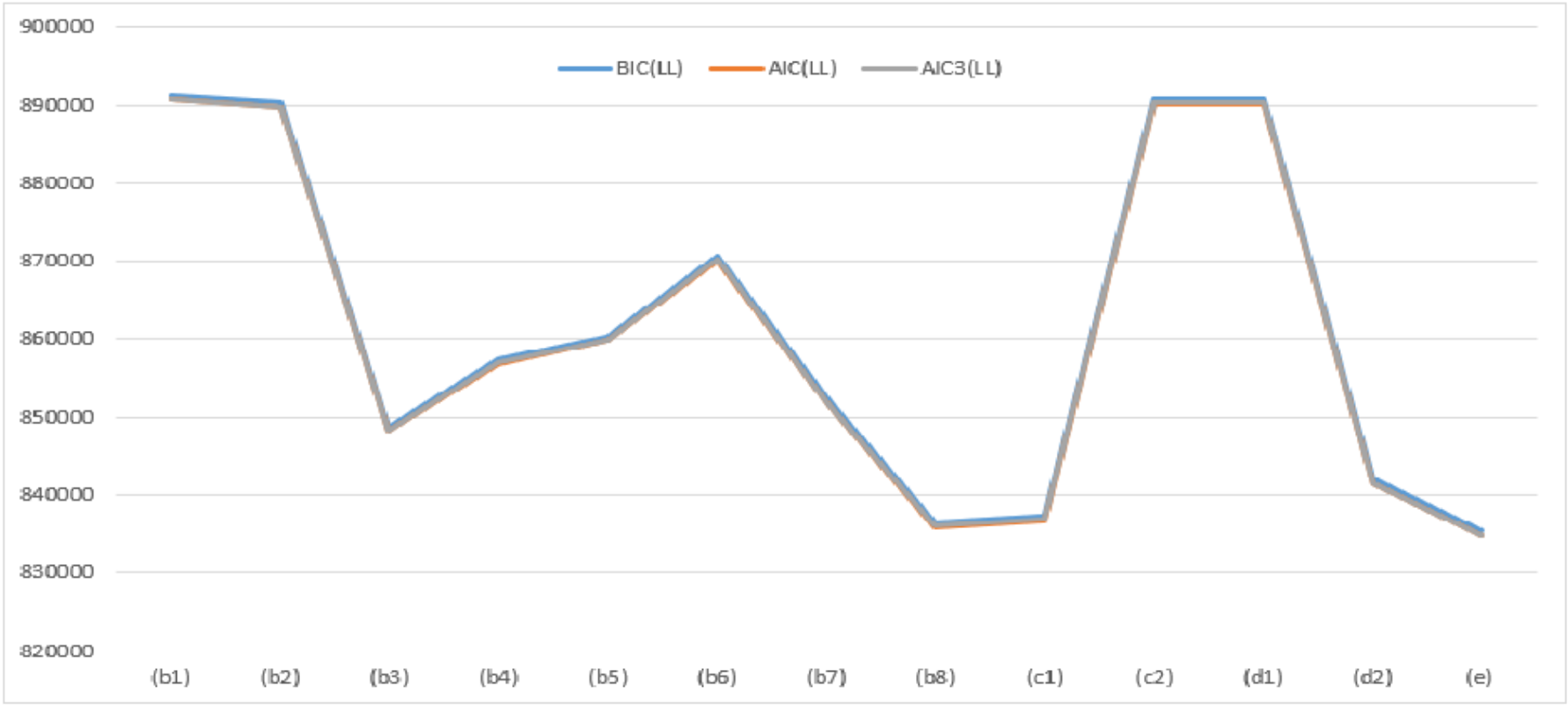
We considered 12 models:

- Models (b1)-(b8): Models with different configurations of measurement error, but equal parameters in the two countries, both LFS and ER indicators
- Model (c1): Correlated error LFS (an error IT same error NL) - Correlated error ER (same error)
- Model (c2): Correlated error LFS (same error) - Correlated error ER (an error NL same error IT)
- Model (d1): Correlated error LFS (same error) - Correlated error ER (an error IT same error NL)
- Model (d2): Correlated error LFS (an error NL same error IT) - Correlated error ER (same error)
- Model (e): Same type of measurement error configuration (same error) with different parameters for the two countries

Model selection, step 2

- (b1) Baseline model: random measurement error in LFS and ER indicator
- (b2) Correlated error LFS (age proxy) - Random error ER
- (b3) Correlated error LFS (same error) - Random error ER
- (b4) Random error LFS - Correlated error ER (same error)
- (b5) Correlated error LFS (an error) - Random error ER
- (b6) Random error LFS - Correlated error ER (an error)
- (b7) Correlated error LFS and ER (an error)
- (b8) Correlated error LFS and ER (same error)

Model selection, step 2



Model selection, step 2

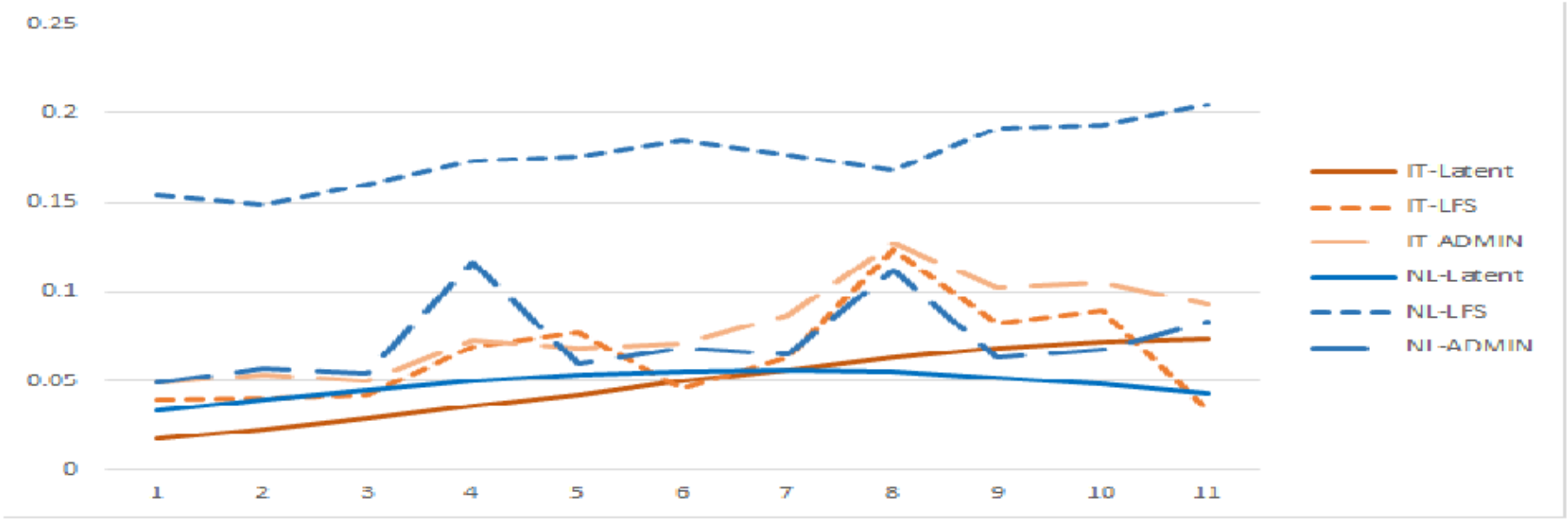
Figure: Conditional probabilities for error repetition, years 2017-2019

				Contract LFS in t		
	Contract LFS in t−1	Latent contract in t	Latent contract in t−1	permanent	temporary	other
Italy	permanent	temporary	temporary	0.8114	0.1717	0.0169
Italy	permanent	other	other	0.8164	0.0053	0.1783
Italy	temporary	permanent	permanent	0.1608	0.8379	0.0013
Italy	temporary	other	other	0.0154	0.7588	0.2259
Italy	other	permanent	permanent	0.3033	0.0017	0.6951
Italy	other	temporary	temporary	0.1168	0.3247	0.5585
Netherlands	permanent	temporary	temporary	0.9709	0.0265	0.0026
Netherlands	permanent	other	other	0.9791	0.0006	0.0203
Netherlands	temporary	permanent	permanent	0.1599	0.8388	0.0013
Netherlands	temporary	other	other	0.0138	0.7839	0.2023
Netherlands	other	permanent	permanent	0.0507	0.0003	0.949
Netherlands	other	temporary	temporary	0.1107	0.3078	0.5815

				Contract ADMIN in t		
	Contract ADMIN in t−1	Latent contract in t	Latent contract in t−1	permanent	temporary	other
Italy	permanent	temporary	temporary	0.8828	0.1116	0.0056
Italy	permanent	other	other	0.7847	0.0016	0.2137
Italy	temporary	permanent	permanent	0.2704	0.7284	0.0012
Italy	temporary	other	other	0.0037	0.2786	0.7177
Italy	other	permanent	permanent	0.6706	0.0024	0.3269
Italy	other	temporary	temporary	0.0221	0.6984	0.2795
Netherlands	permanent	temporary	temporary	0.8697	0.1241	0.0062
Netherlands	permanent	other	other	0.8764	0.0009	0.1226
Netherlands	temporary	permanent	permanent	0.0928	0.9067	0.0004
Netherlands	temporary	other	other	0.0015	0.7162	0.2824
Netherlands	other	permanent	permanent	0.0643	0.0002	0.9355
Netherlands	other	temporary	temporary	0.017	0.5395	0.4434

Transition flow

Figure: Observed transition flow from temporary to permanent contracts in LFS and ER data and estimated flow by quarters. Italy-Netherlands, years 2017-2019



Conclusion and future work

- Confirms measurement error threatens mobility estimates
- Differences between IT-NL in this are not large

- Next steps
  - further discussion on model selection
  - split 'other' to self-employed and non-employed
  - longer time period
  - discussion with subject matter experts

This work is in progress...

Comments? Questions? Suggestions?

D.Filipponi, dafilipp@istat.it
S.Loriga, siloriga@istat.it
M.Garnier-Villarreal, m.garniervillarreal@vu.nl
D.Pavlopoulos, d.pavlopoulos@vu.nl
R.Varriale, roberta.varriale@uniroma1.it
R. Stoel, r.stoel@cbs.nl