Using EUROMOD with population administrative data for Estonia

Alari Paulus

ISER, University of Essex
Praxis Centre for Policy Studies

EUROMOD Workshop
University of Milan
25-26 Sep 2019

Previously on 'EUROMOD annual meeting' ...

- What?
  A plan to use Estonian administrative data for EUROMOD
- Why?
  To improve precision, level of detail and timeliness of input data
- When?
  A feasibility study completed in spring 2018, dataset construction started in Oct 2018
- Who?
  MoF (initiative), Statistics Estonia (data), Praxis (know-how)

This season

Register-based EUROMOD input dataset for Estonia:

- First cross-sectional dataset finalised in spring 2019
- Combines information from 17 administrative data sources
- Annual incomes for 2017
- Covers whole population (1.3 mln)
- Stored on Stat. Estonia servers
  - full dataset accessible in their secure room
  - a random 1% subsample accessible over VPN

Future plans for developments

- Annual (cross-sectional) datasets 2013-2018 $\rightarrow$ in progress
- Monthly (panel) data $\rightarrow$ model restructuring needed?
- Extended scope of the model
  - e.g. indirect taxes*, property taxes, state pensions
- Dynamic and/or behavioural elements
  - e.g. LM adjustments, labour supply, tax compliance*, macro effects
- A web-based user interface for general public access

Plan for today

- Share overall experiences
- Data validation (key aspects)
- (Top) income analysis – work in progress

Main problems and surprises

- Major challenges: residential status and household structures
- Data revisions/updates more time-consuming than expected
  - ca 9 hours to re-generate the whole dataset
  - many registers, large sample, code optimisation
- Validation/simulations also time-consuming
  - generally run a single system at once (ca 5 min)
  - a subsample not always a substitute
- Next level for macrovalidation
  - most external estimates internalised
  - no sampling error – expect 100% match with 'controls'
  - discovered and fixed mistakes in SILC income data

Distribution of households by household size

Hh size	Reg 2017 #	Reg 2017 %	SILC 2017 #	SILC 2017 %	PopCen 2011 #	PopCen 2011 %
1	226,665	39.95	236,988	39.64	190,592	34.26
2	140,663	24.79	168,510	28.19	167,972	30.20
3	87,627	15.45	87,903	14.70	96,252	17.30
4	60,697	10.70	74,183	12.41	65,094	11.70
5	28,507	5.02	20,518	3.43	23,714	4.26
6	12,471	2.20	7,285	1.22	7,999	1.44
7+	10,718	1.89	2,456	0.41	4,636	0.83
Total	567,348	100.00	597,843	100.00	556,259	100.00

SILC vs register data: variables

- UDB SILC + variables from national SILC (e.g. detailed benefits)
- Most of income data already from registers

		Register-based		
		Yes	No	Total
SILC-based	Yes	101	28	129
	No	42	-	42
	Total	143	28	171

SILC vs register data: variables

- UDB SILC + variables from national SILC (e.g. detailed benefits)
- Most of income data already from registers

		Register-based		
		Yes	No	Total
SILC-based	Yes	101	28	129
	No	42	-	42
	Total	143	28	171

- Added vars (42): mostly detailed income components, assets
- Excluded vars (28): mostly survey-specific and not relevant
- Excluded incomes: interest income, income from non-registered self-employment, (voluntary) private transfers, in-kind incomes
- Problematic: occupation, industry, work hours

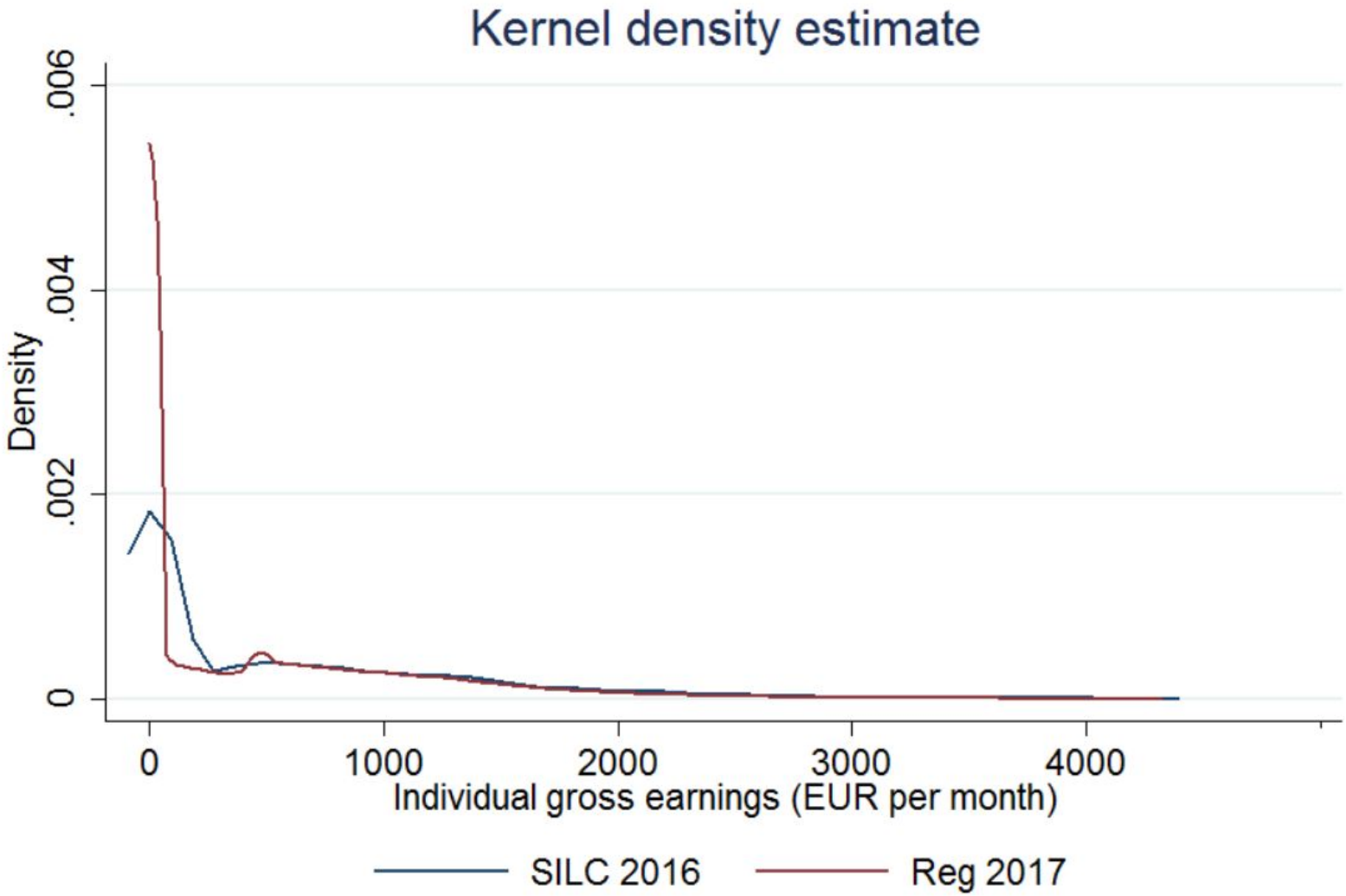
Macrovalidation

- Employed/unemployed – small discrepancies (5-6%) due to different definitions
- Market incomes – n/a (only aggregates of individual tax reports published, not that of employer reports)
- Non-simulated benefits/benefits
  - very small discrepancies ($\pm$2%)
  - some large discrepancies – different definitions/units
- Simulated benefits
  - some discrepancies still – due to residency, hh structures, annual data, simulation errors? $\rightarrow$ to investigate further
  - new take-up calibrations (subsistence benefit 33%, needs-based family benefit 20%) to match total recipients
- Simulated taxes – very good (PIT 99%, SIC 98-103%)
- Income inequality (Gini) – huge gap (SILC 0.304 vs Reg 0.393)

Research questions

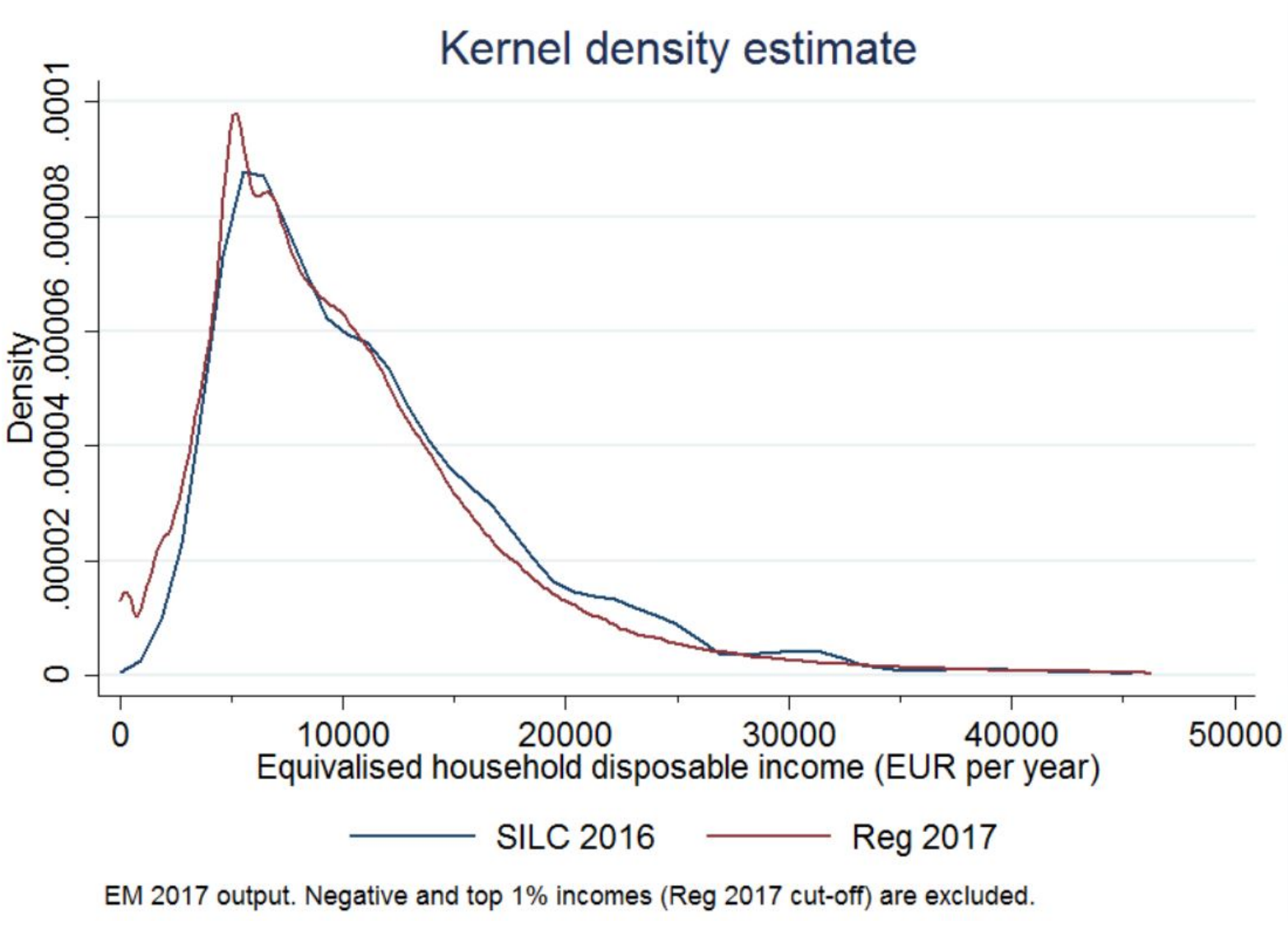
- What is causing the inequality gap?
- To what extent SILC and register-based household income distributions overlap?
- How well does SILC capture the tales of the distribution?
- What income sources are more prevalent in the tales?
- How much does it matter for fiscal and distributive analysis?

Distribution of individual gross earnings



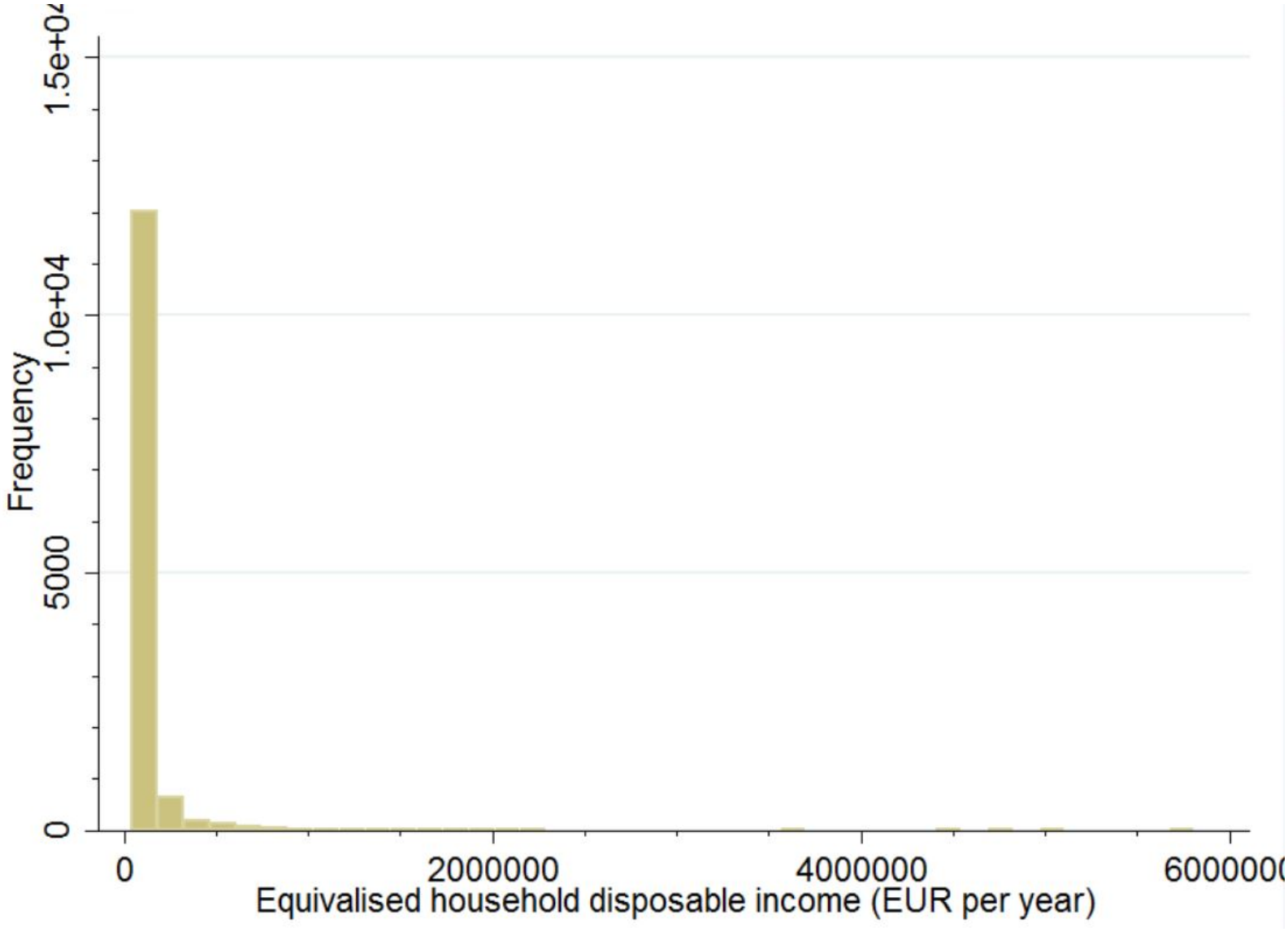
EM 2017 output. Negative and top 1% incomes (Reg 2017 cut-off) are excluded.

Distribution of equiv. hh disposable income



EM 2017 output. Negative and top 1% incomes (Reg 2017 cut-off) are excluded.

Distribution of equiv. hh disp. income (Reg): top 1%

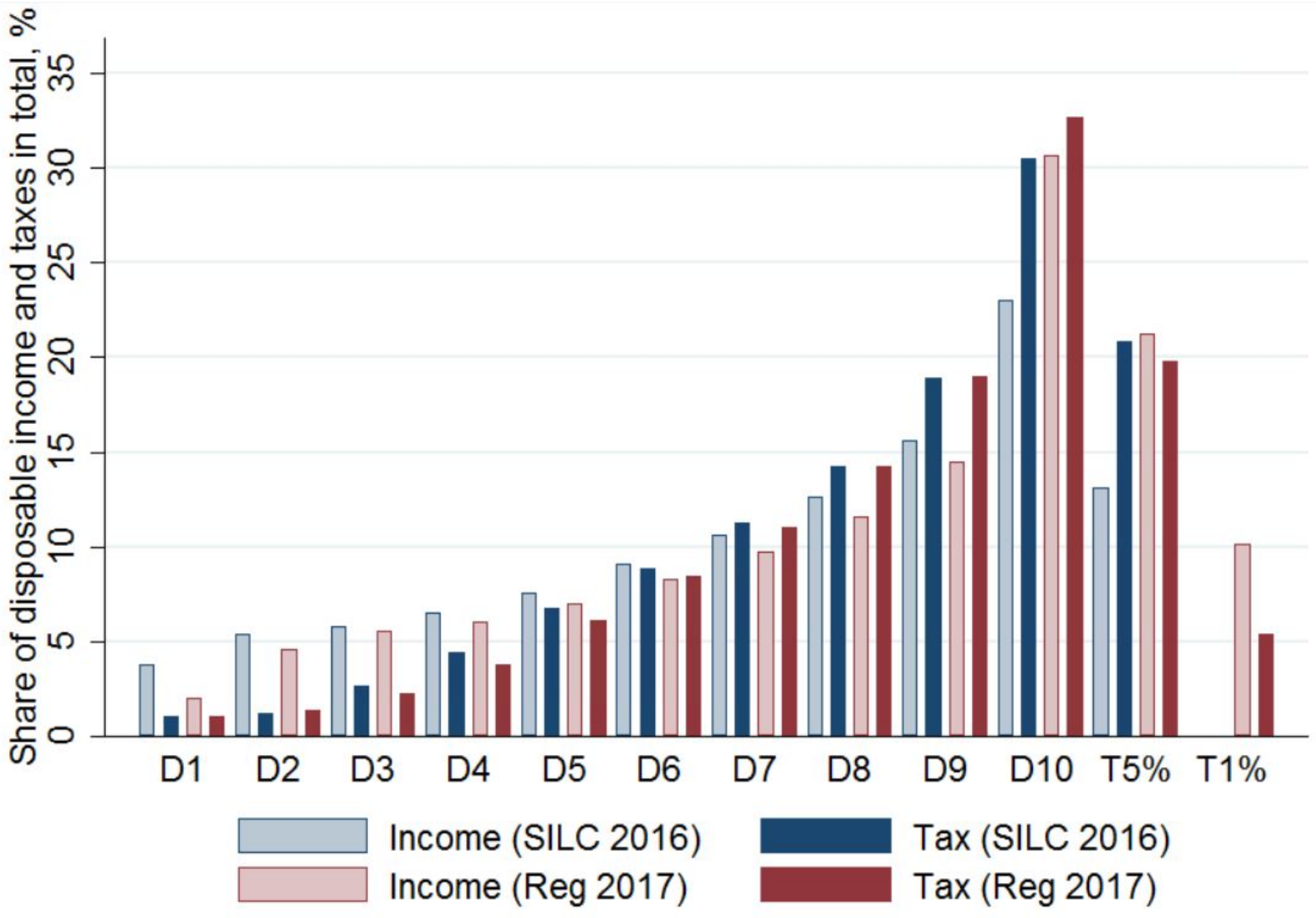


Income inequality (Gini)

	Market income	Disposable income
SILC 2017 (ee_2017_c2, FYA=0)		
all sample	0.46038	0.30444
non-negative incomes	0.45790	0.30417
positive incomes	0.37374	0.30417
Reg 2017 (ee_2017_d1, FYA=0)		
all population	0.54235	0.39271
non-institutionalised population	0.54100	0.39188
incomes capped at T1 (censored)	0.51181	0.35632
incomes below T1 (truncated)	0.49937	0.34092
non-negative incomes below T1	0.49886	0.33828
positive incomes below T1	0.41255	0.33261

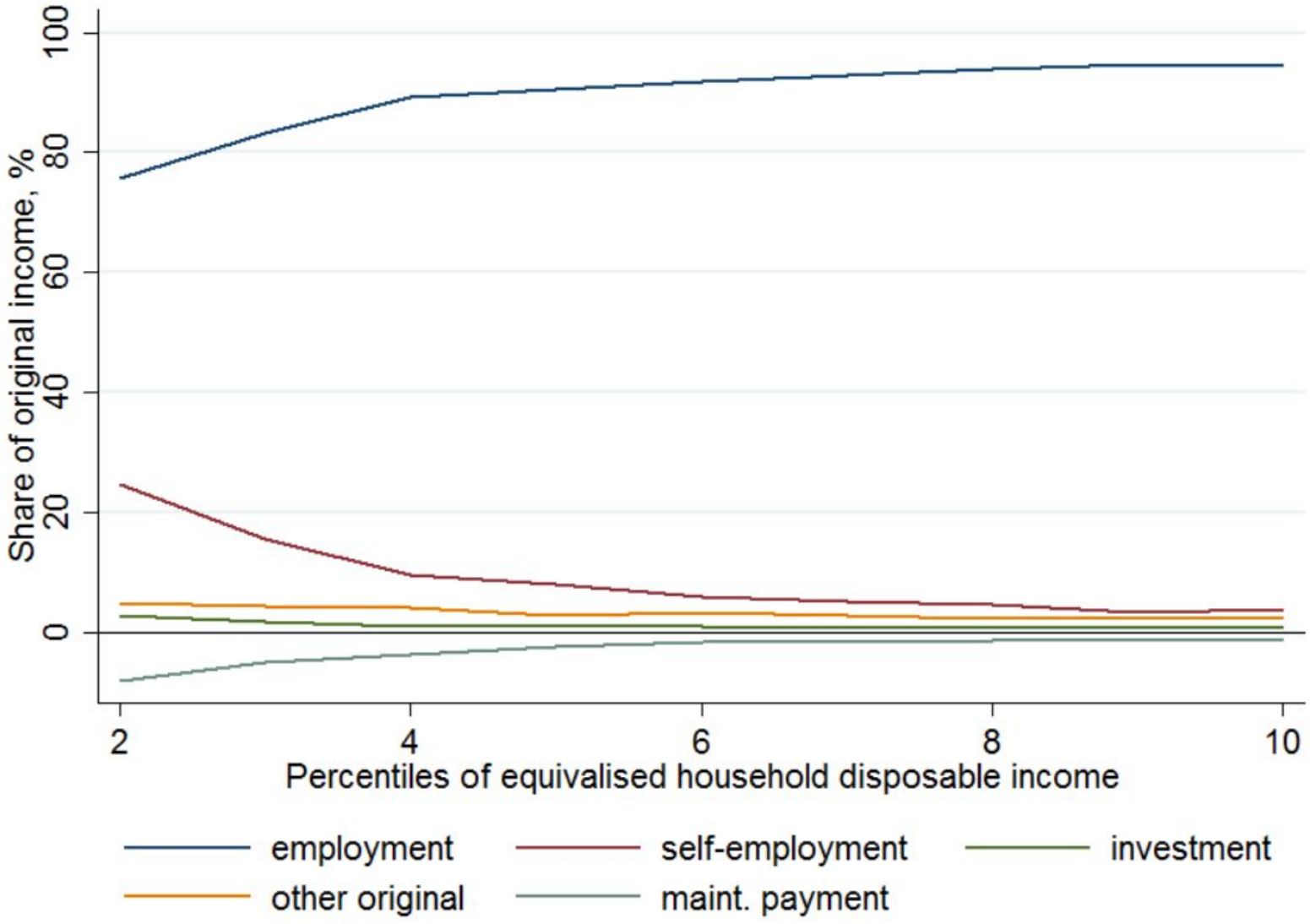
T1 = income cut-off level for the top 1% (register data, equivalised, annual): 52,706 EUR (original income) and 46,241 EUR (disposable income).

Share of income and taxes by income groups

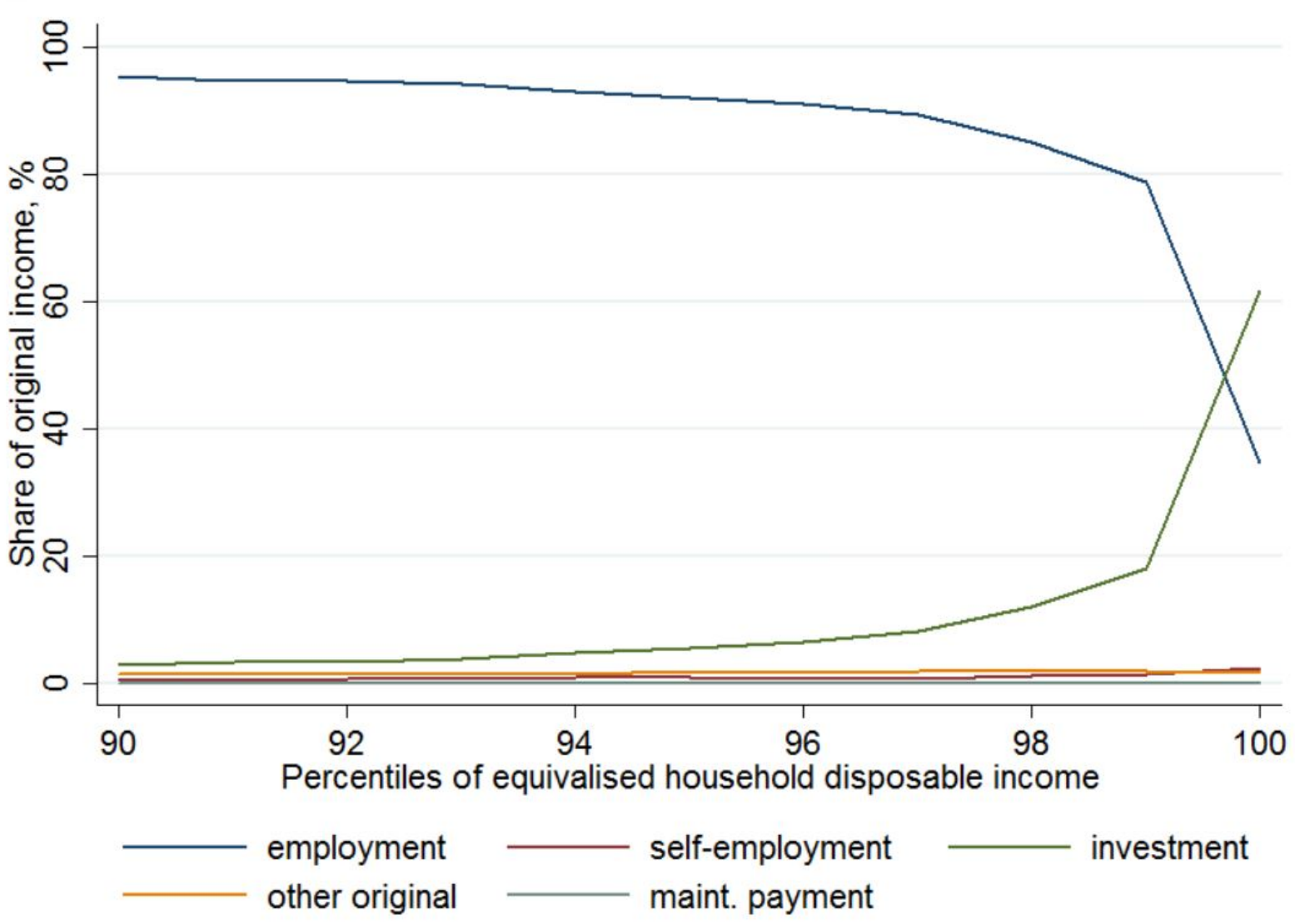


EM 2017 output.

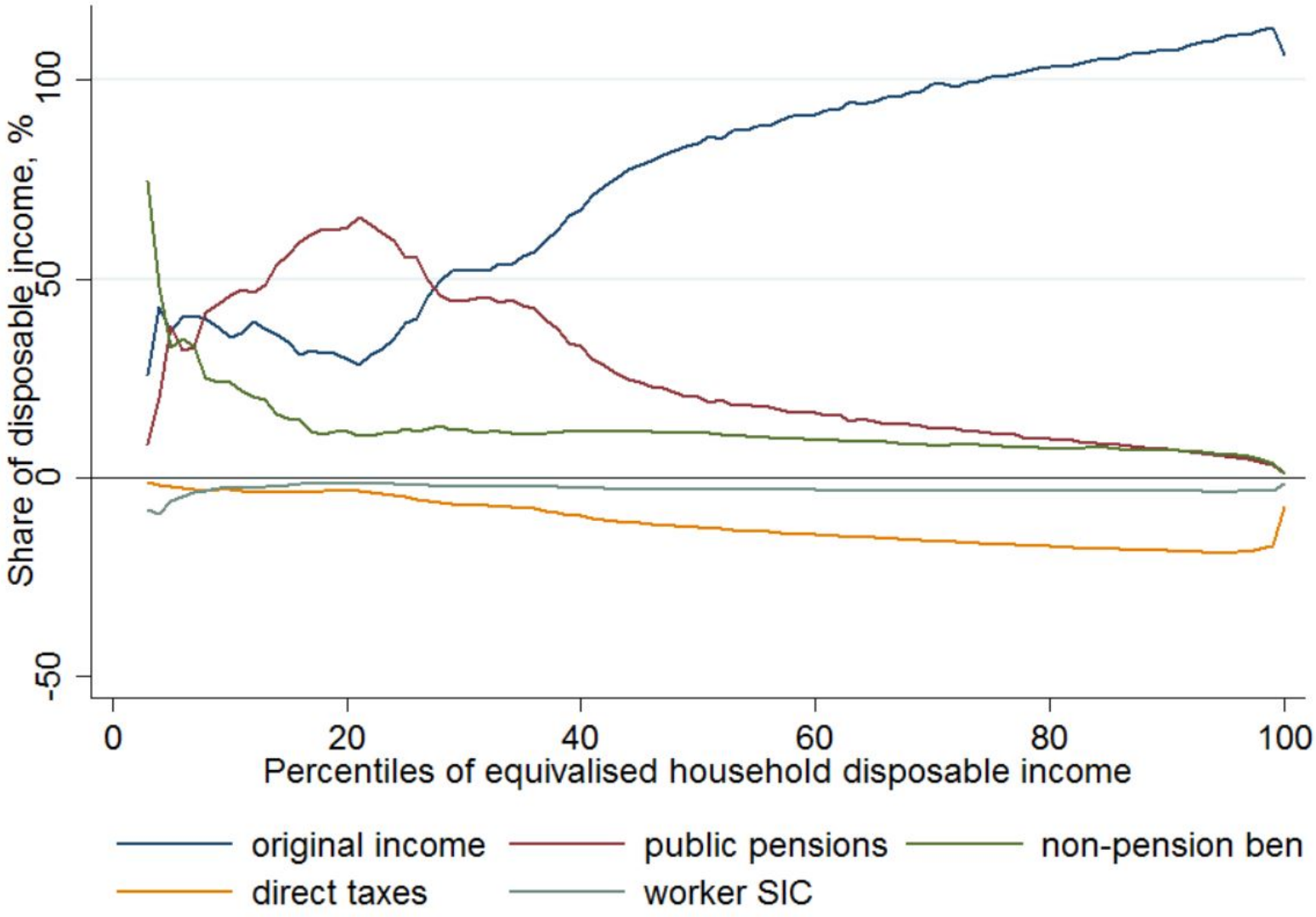
Composition of original income (Reg): bottom 10%



Composition of original income (Reg): top 10%



Composition of disp. income by percentiles (Reg)



Preliminary conclusions

- Population register data very promising but not straightforward
  - access and technical requirements
  - residents and household structures
  - information not collected
- Compared to SILC-based estimates
  - improved precision of fiscal aggregates
  - discrepancies at the bottom and in particular at the top (1%)
  - large differences in income inequality
  - tax system regressive at the very top
- Implications for survey data/sample
  - difficult to spot data/simulation problems if large sampling error
  - missing rich can substantially bias results

Thank you!

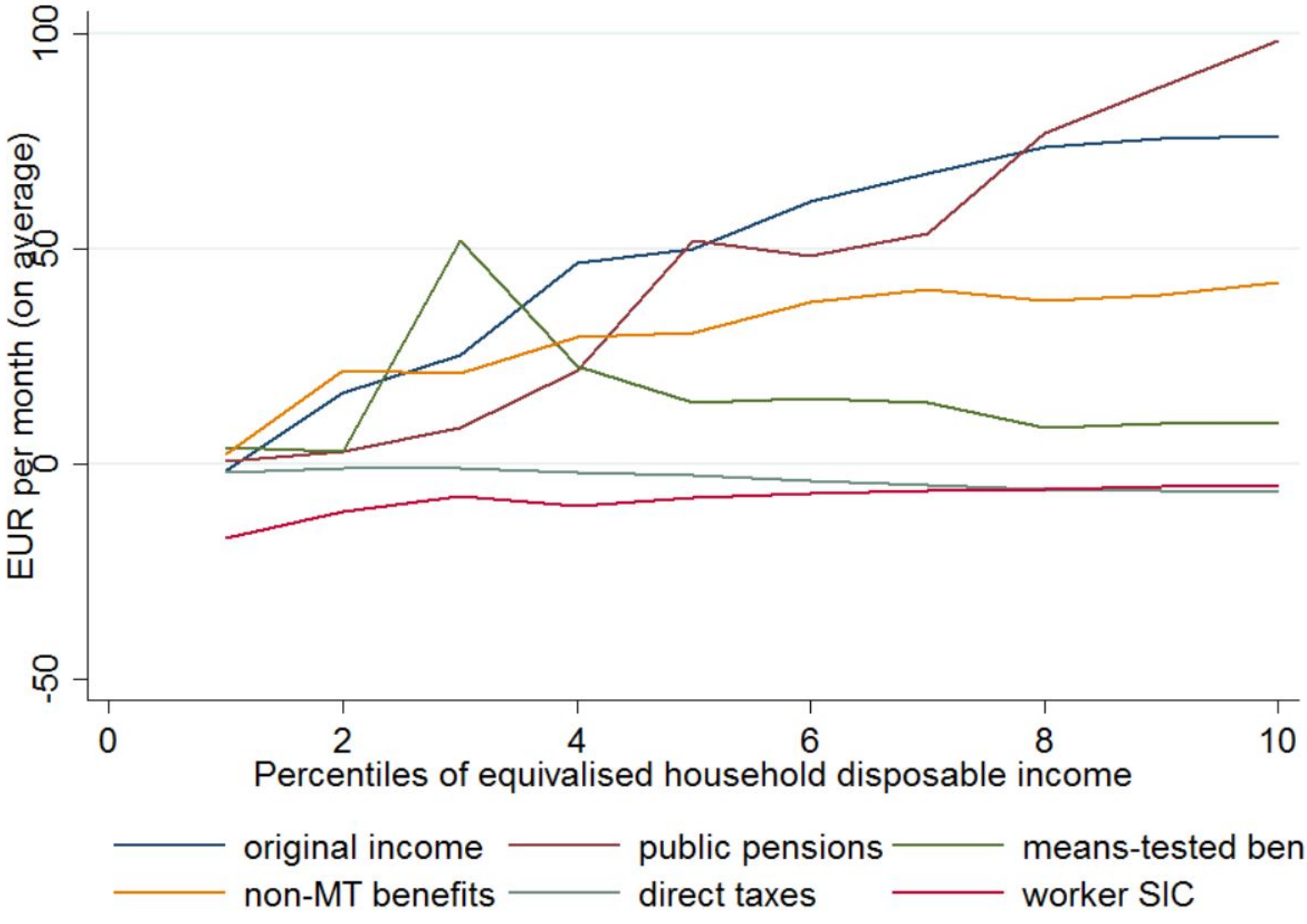
... to be continued

Share of income and taxes in top percentiles (Reg)

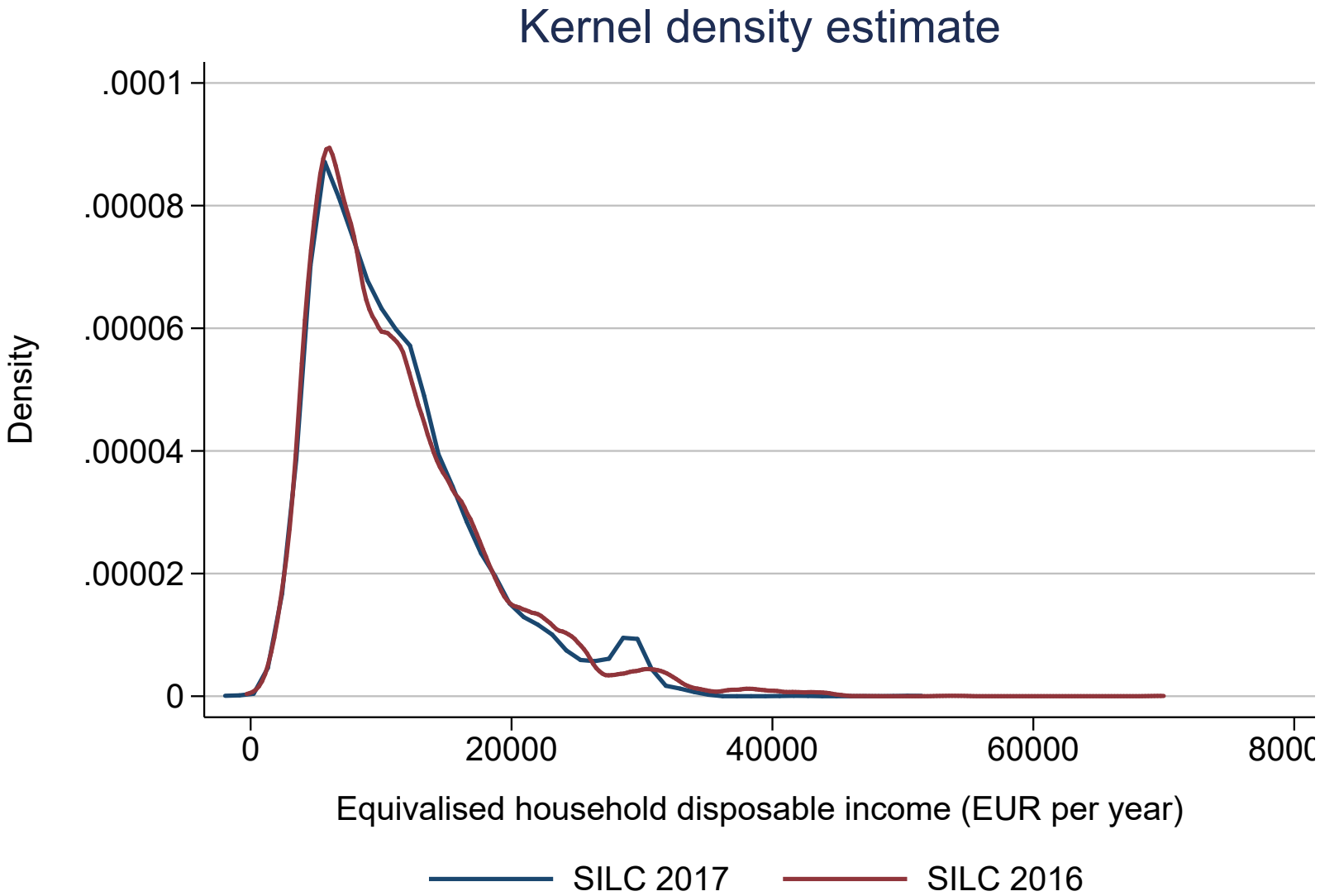
	Individual gross earnings			Equiv. disp. hh income		
Top	Cut-off	Share of income	Share of SIC	Cut-off	Share of income	Share of SIC/IT
10%	23,027	31.1	30.1	19,270	30.7	32.6
5%	30,000	19.9	19.5	24,599	21.2	19.8
1%	51,888	7.0	7.2	46,241	10.2	5.4

Notes: percentile cut-offs in EUR per year; share of income/SIC/IT as a percentage of the corresponding total.

Composition of disp. income (Reg): bottom 10%



SILC 2017 vs SILC 2016



EM 2017 output.