Nowcasting the household income distribution

Fabrice Murtin
OECD Statistics and Data Directorate

February 4th 2020
UNCTAD Technical Workshop

Background and motivations

- Since 1998, regular OECD data collection on income distribution and poverty (OECD IDD) based on national sources and comparable definitions
- Strong internal and external demand for IDD (e.g. COPE & its reports, *How's Life?*, Inclusive Growth, Economic Surveys, G20, etc.)
- However, despite annual collection, information is not timely: average lag is 2-3 years…
- This limits the possibility to use distributive information in macro-economic assessments where timeliness is key requirement (e.g. *Economic Outlook, Going for Growth*)

Background and motivations

- Project seeks to *nowcast* household income by decile (in year T) in as many OECD countries as possible based on contemporaneous information through reduced-form econometrics

- Once methodology has been thoroughly tested, estimates could be released regularly by the OECD in various forms (NAD household dashboard, MDLS, *How's Life?*, *ad hoc* statistical briefs, G20 documents)

- In the very short term: working paper and feedback from experts in a variety of *fora*

OECD Better Life Initiative

Nowcasting: basic principles

- Objective: Construct a predictive model that can be evaluated by out-of-sample (OOS) performance
- Parsimony: a complicated model increases in-sample fit (R2) but may decrease OOS
- Credibility: meaningful coefficients
- Specificity: the model must be decile-specific and possibly country-specific

The dependent variable

- Average equivalised household disposable income per decile from IDD
- We consider two income series per country (waves 6 and 7)
- Linear interpolations used to cover gaps up to 3 years for countries lacking annual surveys and for earlier (pre-2000s) periods
- We also considered and tested a model to predict individual income's components for each decile (i.e. wage, self-employment income, capital income, transfers received, taxes paid) but model performed less well than model for total income

Explanatory variables

- We created a group of 30+ predictors, drawn from national accounts and other official sources, that are timely and available for most countries
- Examples: GDP, unemployment rate, mean net household disposable income (SNA), self-employment rate, wage rate, hours worked per worker, long-term interest rates, house prices, property income, share prices, current transfer received by households, taxes on business and on different kinds of households etc…

The predictive models

- We consider several predictive algorithms routinely used in AI:

  Random forest, Gradient boosting, Neural network, SVM

- We compare the results with those obtained from a log-linear model with variable selection (LASSO)
- LASSO model: For each decile we predict the growth rate of real household disposable income (defl. PCD):

$$\Delta_{t,t-1} log\, \boldsymbol{y} = \Delta_{t,t-1} \boldsymbol{X}.\beta_1 + \Delta_{t-k,t-k-1} log \boldsymbol{y}\, .\beta_2 + \boldsymbol{\varepsilon}$$

- Performance: we evaluated 1 year-ahead out-of-sample performance against observed growth rates and a naive 'random walk' model (forecasted growth=last observed growth)

OECD Better Life Initiative

Out-of-sample performance (growth rates)

OOS correlation between predicted and observed growth rates

	Decile 1	Deciles 2 to 9	Decile 10	All deciles
LASSO	0.59	0.79	0.17	0.60
ANN	-0.12	0.19	0.59	0.09
SVM	-0.04	-0.05	0.30	0.00
DRF	-0.39	0.00	0.17	-0.10
GBM	0.29	0.34	-0.21	0.25

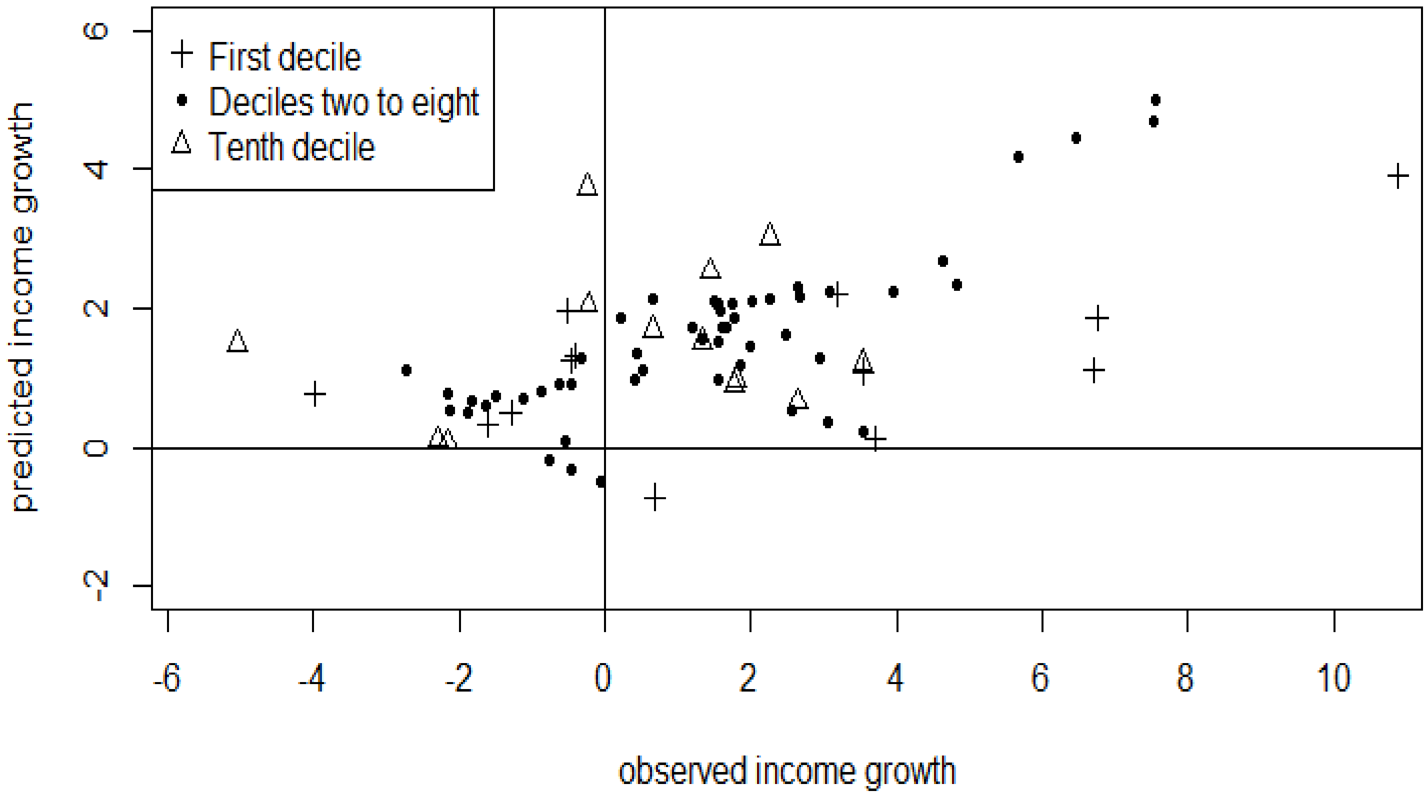
Estimated model (LASSO)

- All coefficients have the same sign across all income deciles (and all variables are 'correctly' signed)

	Disposable income growth rate for decile:									
	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Control for lags	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Change in/growth of:										
Unemployment rate	-	-	-	-	-	-	-	-	-	
GDP		+	+	+	+	+	+		+	+
Wage rate		+	+	+	+	+	+	+	+	+
Average tax wedge	-	-	-	-	-	-	-	-	-	-
Disbursement of government		+	+	+	+	+	+	+	+	
Current transfers	+									
Net current household receipts									-	
Self-employment rate										+
Share price										+
Disposable income									+	

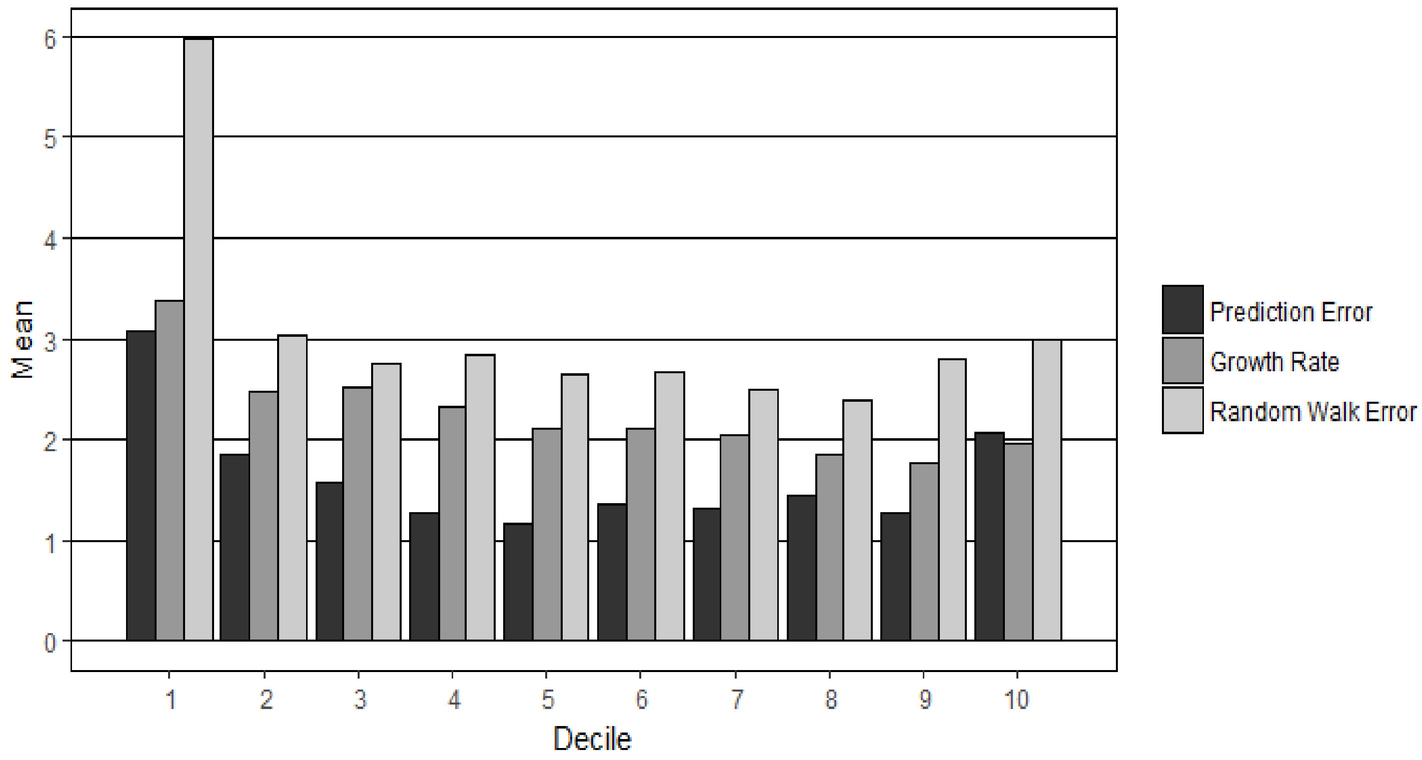
Estimated model

- Average OOS correlation in 2014 (across all deciles in 13 countries) is 0.59
- Fails to capture tails (and negative growth)



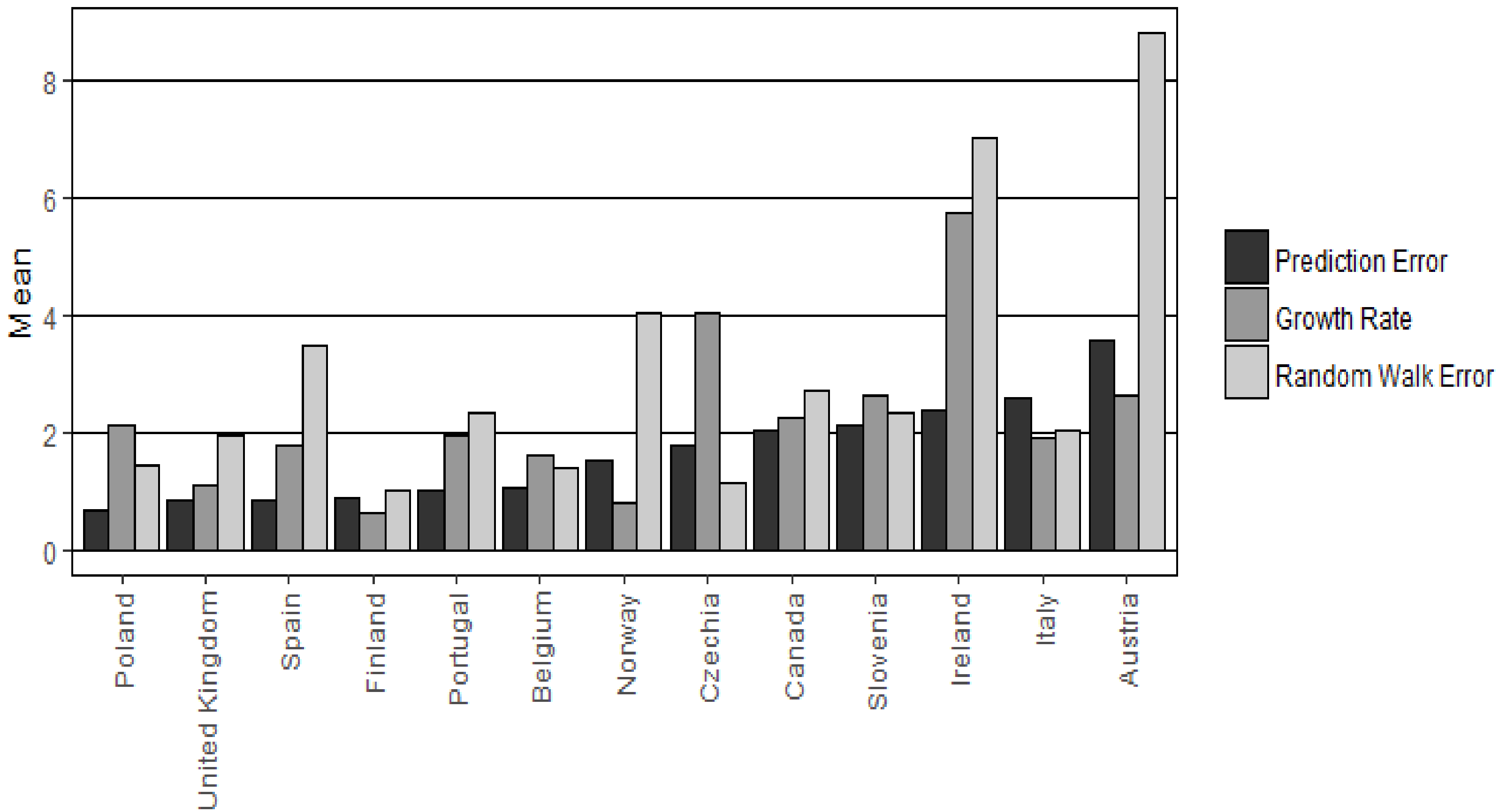
OECD Better Life Initiative

Out-of-sample performance (growth rates)



The estimated model easily beats the random walk
But D1 and D10 are the hardest to predict

Out-of-sample performance (growth rates)



OECD Better Life Initiative

Conclusions

- Nowcasting real changes in household income for various deciles is difficult because:
  - A complicated model is unstable
  - A simplistic model is inaccurate
- More research is needed to:
  - Better model the tails of the distribution
  - Better capture 'regime changes' (large deviations)
  - Better account for country heterogeneity
- On country heterogeneity:
  - other methodologies (microsimulations) may outperform regression-models but are difficult to implement in a consistent way and are much more demanding in terms of information
  - predicting the distribution from NA totals

OECD Better Life Initiative

Thank you!