EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS

2024 ESTORIL - PORTUGAL

Instituto Nacional de Estatística
Statistics Portugal

eurostat

The conference is partly financed
by the European Union

EUROPEAN CONFERENCE ON
QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Estimating Non-Regular Earings for Micro Organizations

Gergely Attila Kiss, Hungarian Central Statistical Office,gergely.kiss@ksh.hu

Beáta Horváth, Hungarian National Bank

István Balázs, Hungarian Central Statistical Office

Instituto Nacional de Estatística
Statistics Portugal

eurostat

The conference is partly financed
by the European Union

Agenda

1. Introduction
2. Data Chjaracteristics and Panel Structure
3. Process of estimation
4. Outlier detection
    1. First Stage
    2. Second stage
    3. Raw aggregates
5. Sample Adjustments
    1. Weighting and Calibration
    2. Comparison
6. Summary

Introduction

- Target of the method
  - Non-Regular earnings are need for some earnings statistics
- Reason for developing
- Used data
  - Surveys are too burdenful for both sides
  - Administrative data is very helpful vs unknown sample characteristics at the end
- Difficulties for estimating
- Generalizable solution?

Data Characteristsics

- Sources are from:
  - National Tax Authority
  - Hungarian State Treasury
- Cross sectional years cover 4.7 million observations.
- Merge 4 year on unique identifiers results in 1.6 million observations
  - These are people who didn't change ISCO codes and organization, join or exit the labor market
  - Obviously this should leave out two age class the young and the elderly
  - Another difference we observe is some ISCO codes tend to be under represented (phisycal jobs)
    - Either coming from the previous step and these are the ISCO codes the have young or elderly mainly working in them.
    - There could be ISCO categories where people jump between organizations or change ISCO categories.
- Finally keep only micro organizations

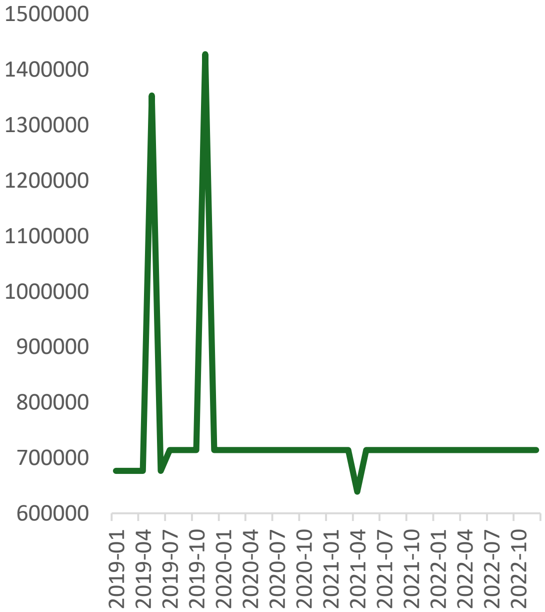
Heterogeneity analysis I

Age categories	Panel	Year 2019	Year 2020	Year 2021	Year 2022
< 25	0.5%	14.8%	10.4%	10.4%	10.7%
25-35	12.0%	22.7%	22.3%	22.3%	21.9%
35-45	25.9%	28.3%	26.9%	25.7%	24.7%
45-55	37.1%	22.6%	25.1%	26.2%	26.7%
55-65	23.3%	10.9%	14.0%	14.0%	14.3%
65 <	1.3%	0.6%	1.24%	1.5%	1.8%

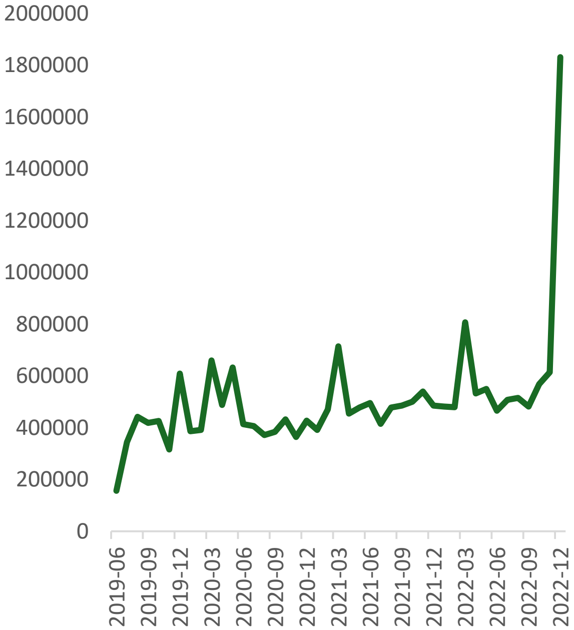
Heterogeneity analysis II

ISCO categories	Panel	Year 2019	Year 2020	Year 2021	Year 2022
C0	0.0%	1.0%	0.7%	0.8%	0.0%
C1	10.8%	6.9%	7.1%	7.0%	7.4%
C2	18.0%	14.7%	15.1%	15.8%	15.5%
C3	16.4%	15.2%	15.5%	16.4%	15.7%
C4	6.6%	7.1%	7.0%	6.9%	7.2%
C5	10.7%	10.7%	10.7%	10.4%	10.6%
C6	0.8%	0.7%	0.7%	0.7%	0.7%
C7	12.4%	9.3%	9.2%	8.7%	8.7%
C8	14.6%	13.2%	13.0%	12.7%	12.9%
C9	9.5%	21.1%	21.0%	20.4%	21.2%

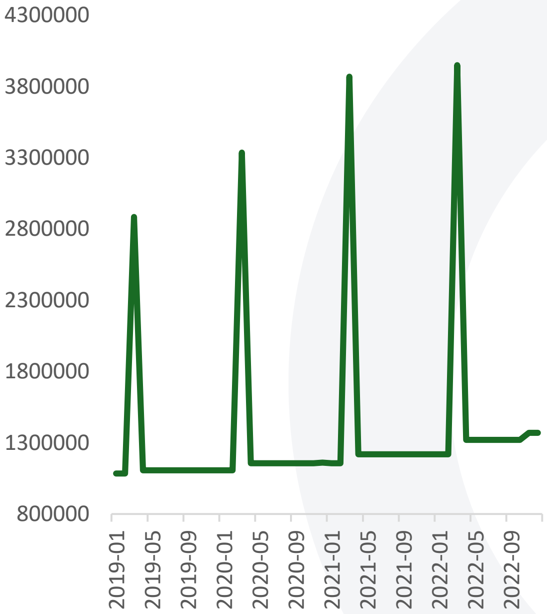
Typical earnings time series



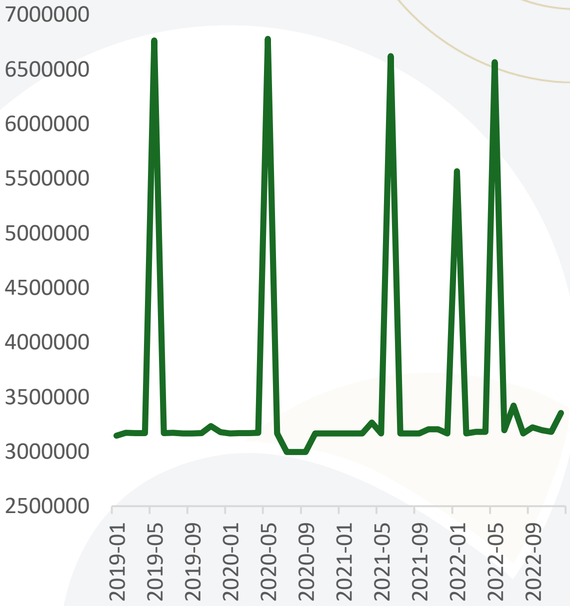
Panel A: Peaks



Panel B: Fluctuations

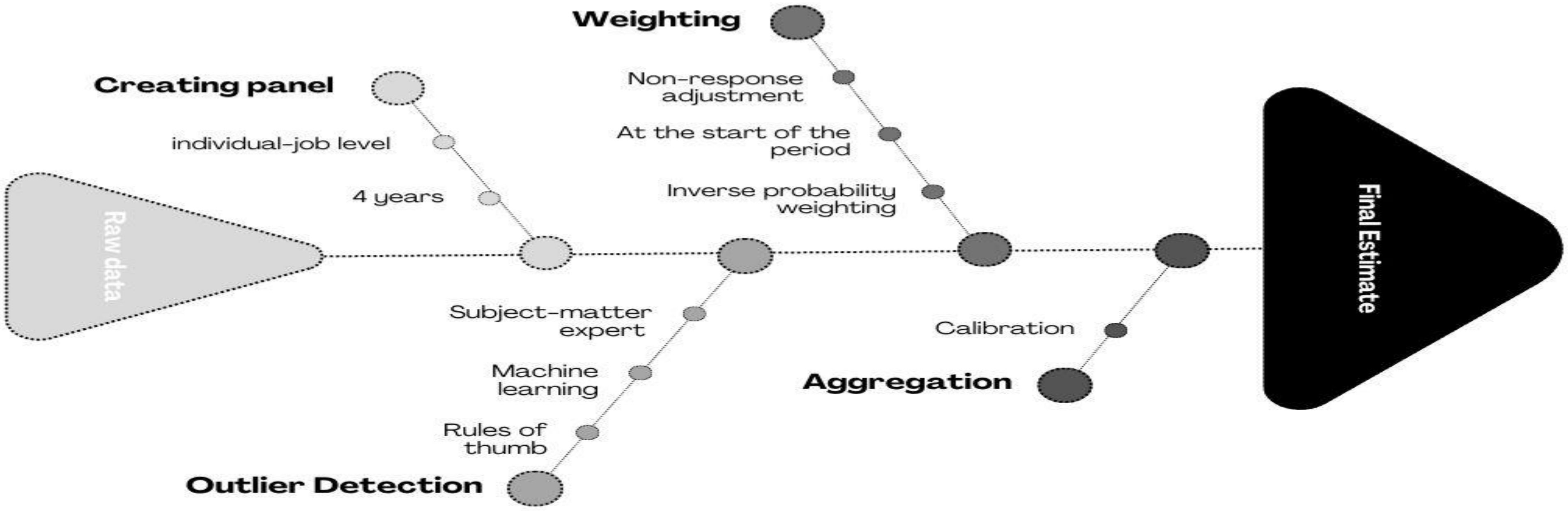


Panel C: Peaks with level-shift



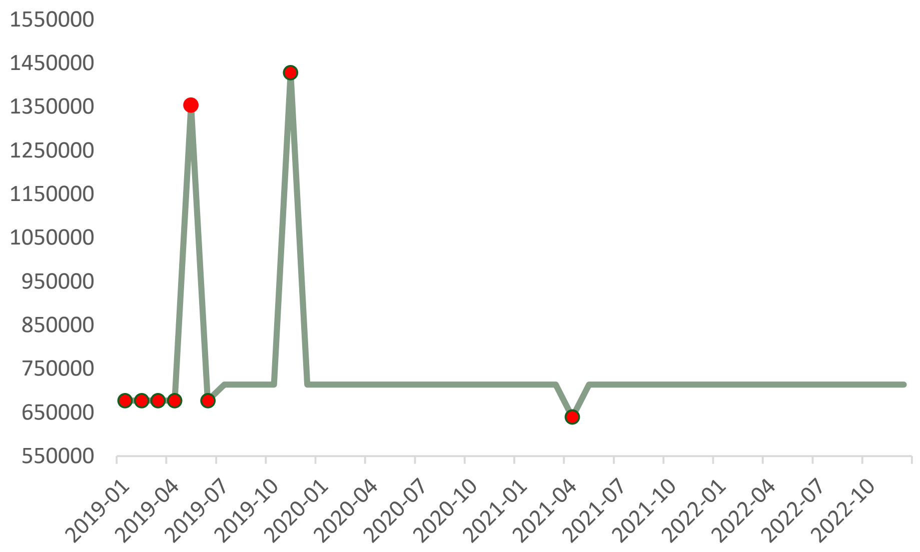
Panel D: Peaks with fluctuations

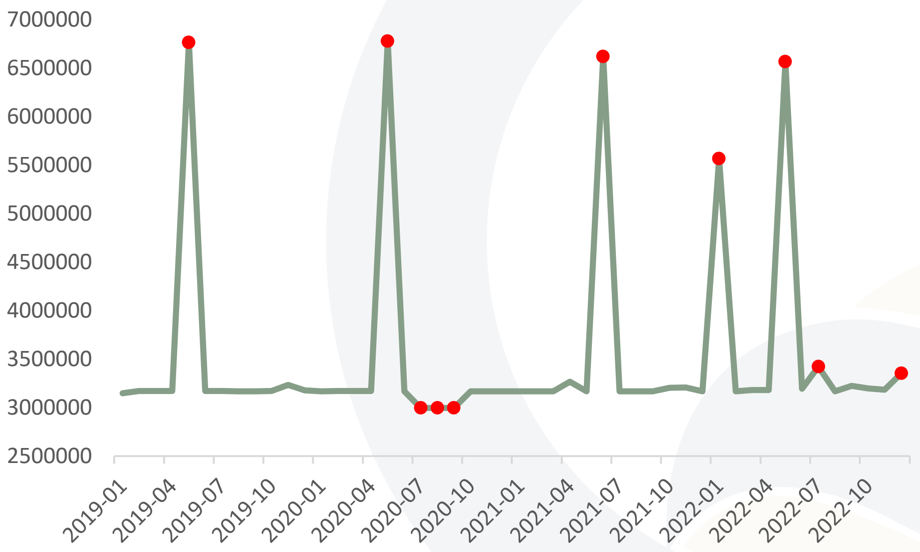
Process of estimation



First stage of outlier detection

- This stage uses mainly random forest based isolation to detect outliers in the previously seen time series in a unsupervised fashion.





Second stage of outlier detection

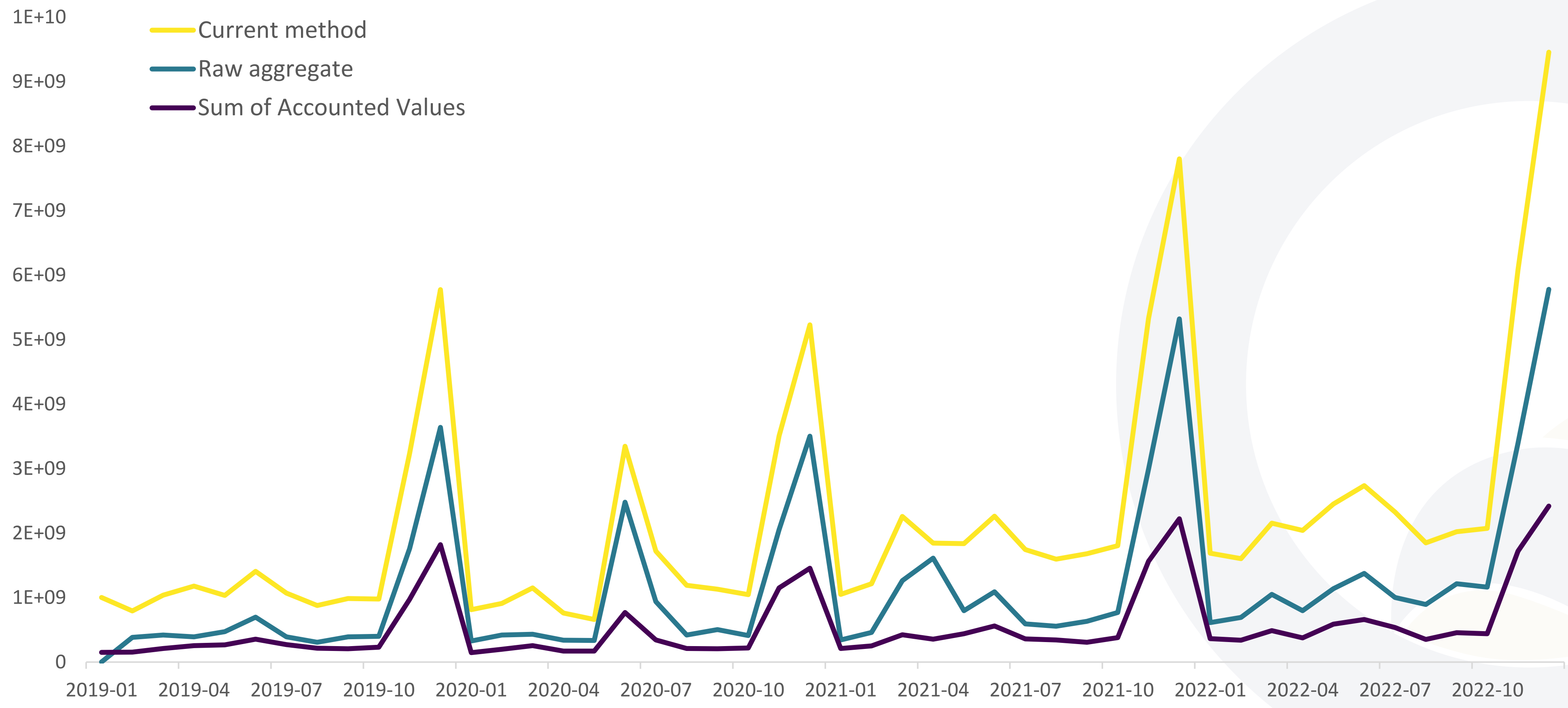
- This stage uses mainly subject matter expert knowledge to define filters after the first stage to be more precise on classifying outlier points as non-regular earning values.

1. Retroactive Raises: commonly used in Hungary

$$D_{RR} = \left(\frac{w_t - \overline{w}_{t-1,t-n}}{m_t * \overline{w}_{t-1,t-n}} < p \right)$$

2. Regulatory one-time payouts: These are not included by definition, subject matter experts identify when and in what amount it happend to which ISCO categories.
3. Small and Regular Fluctuations: Usually should be fluctuation in earnings due to per piece rate or performance wages.
   - Generally small fluctionations < 70-80 EUR, in January larger ones.

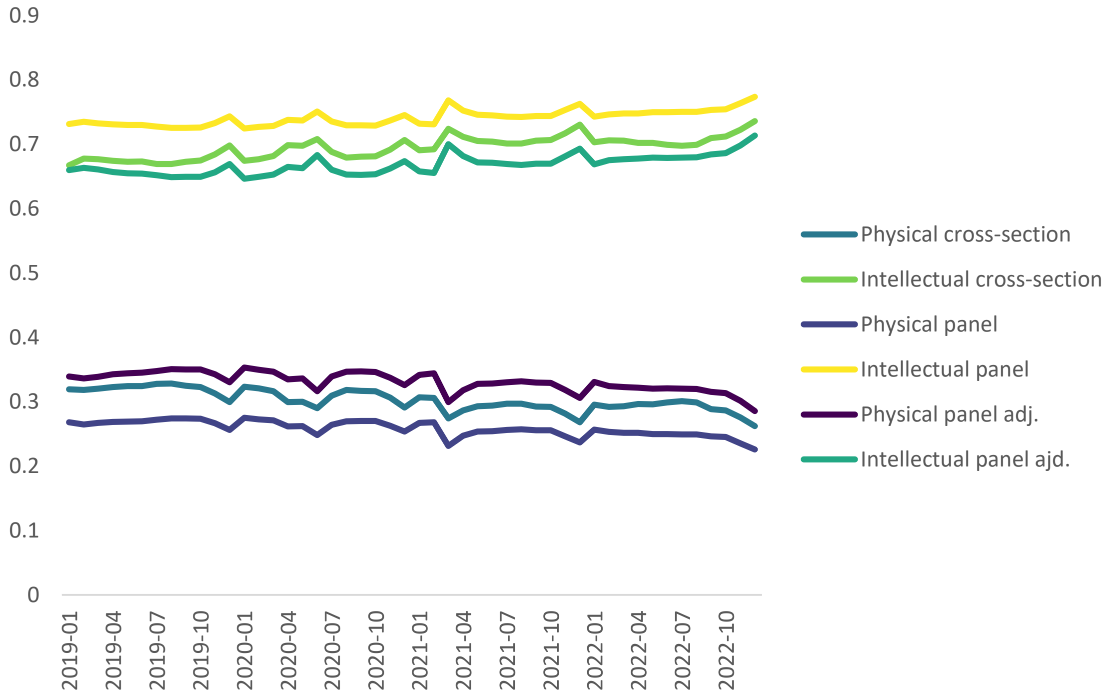
Raw aggregates



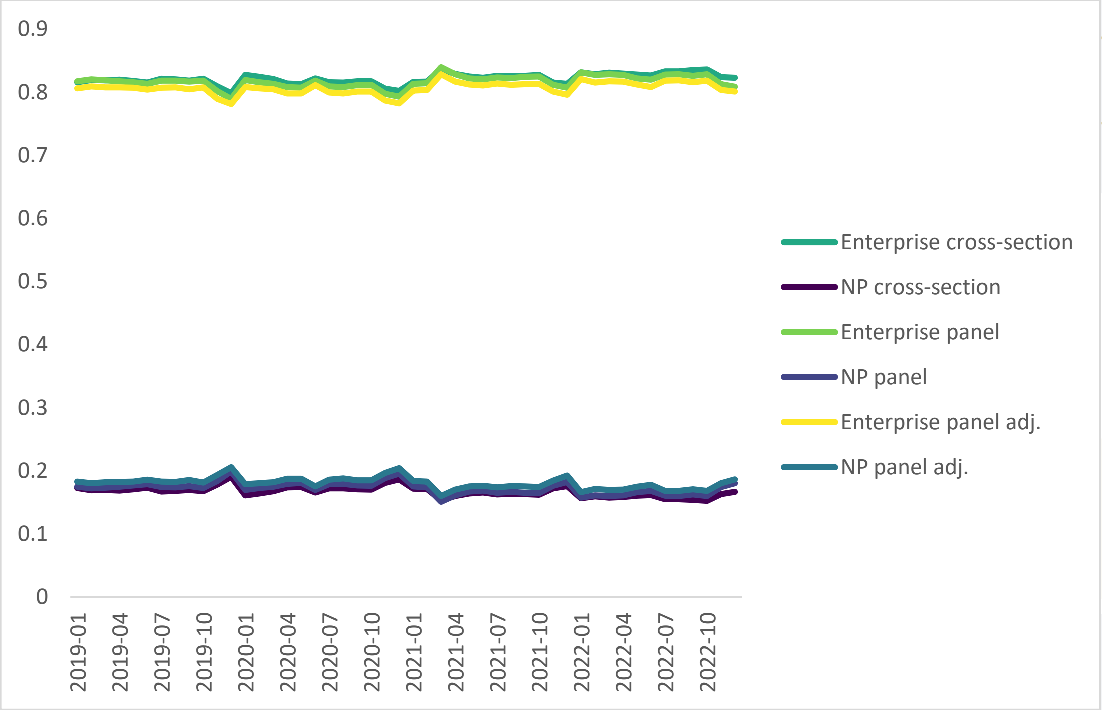
Weighting & Calibration

- Our ides is to make this similar to non-response adjustment.
  - Understanding differences between the panel samples and cross-sectional years.
  - Then use inverse probabilty weighting to correct the sample distributions
  - Calculated on the 1.6 million observations and not just for Micro organizations
- Targeted currently to achieve similarity to cross-sections in currently published important categories:
  - Type of organizations (entr., non-proit and gov. Body)
  - Type of job (physical, intellectual, unknown)
- Weights are calcuated from they 2019 cross-sectional data with logistic regression
  - Used variables: ISCO (4 digit), gender of employee and age category
- Calibration is done on each cross-sectional data using observation numbers
  - By variables: NACE (2 digit) and County codes

Comparison of aggregated series

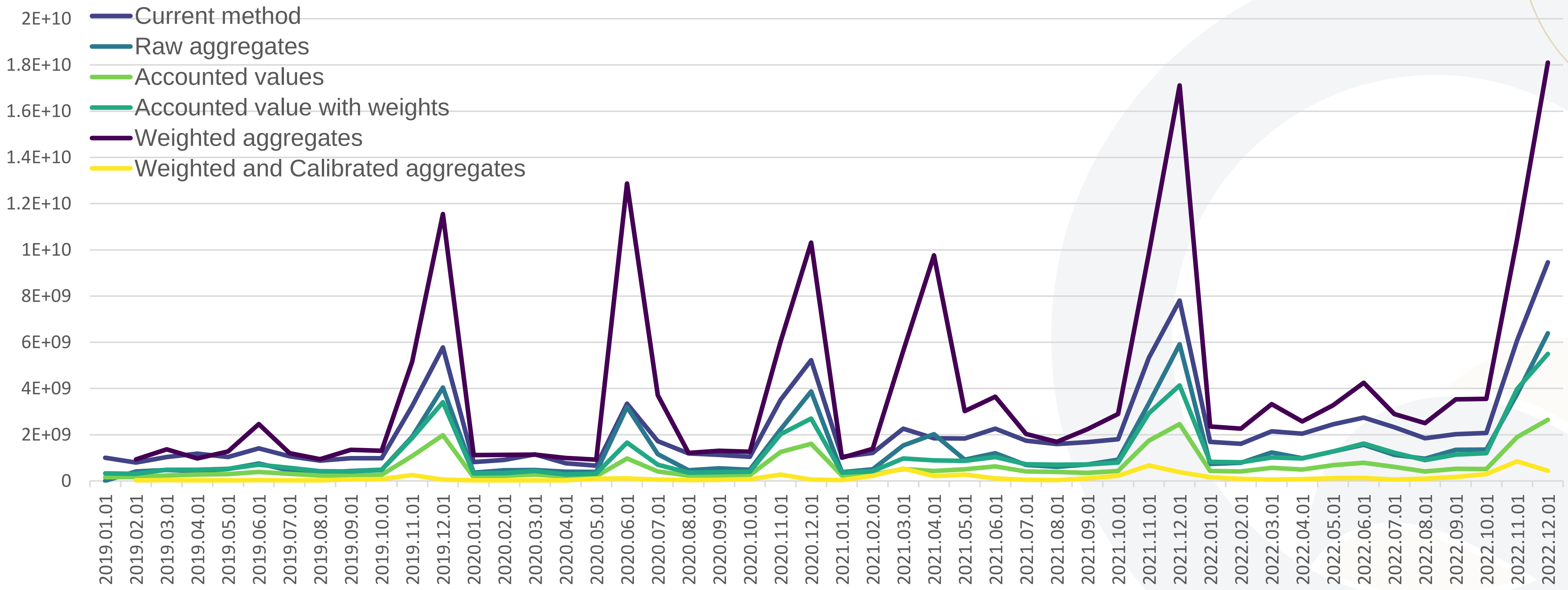


Panel A: Comparison of aggregated series by occupation types



Panel B: Comparison of aggregated series by organization types

Comparison of all series



Instituto Nacional de Estatística
Statistics Portugal
eurostat
The conference is partly financed by the European Union

Summary

- Bottom up appraoch to reach Macro indicators.
- The problem of using unknown populations due to „BIG" data.
- The weighted results still have to be corrected to reach a level of index continuity and do not have a large revision.
- Possible improvements:
  - Calibration technique seems to be off
  - Outlier filtering not strict enough at peaks?
  - Weights should be corrected yearly?
  - Is there still a regulatory payout we did not find at 2021.04?
- Future of the method: If method provides acceptable results with small fluctuations from month-to-month, the method is easy generalizable to national level.

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS 2024 ESTORIL - PORTUGAL

Thank you for your attention!

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed
by the European Union