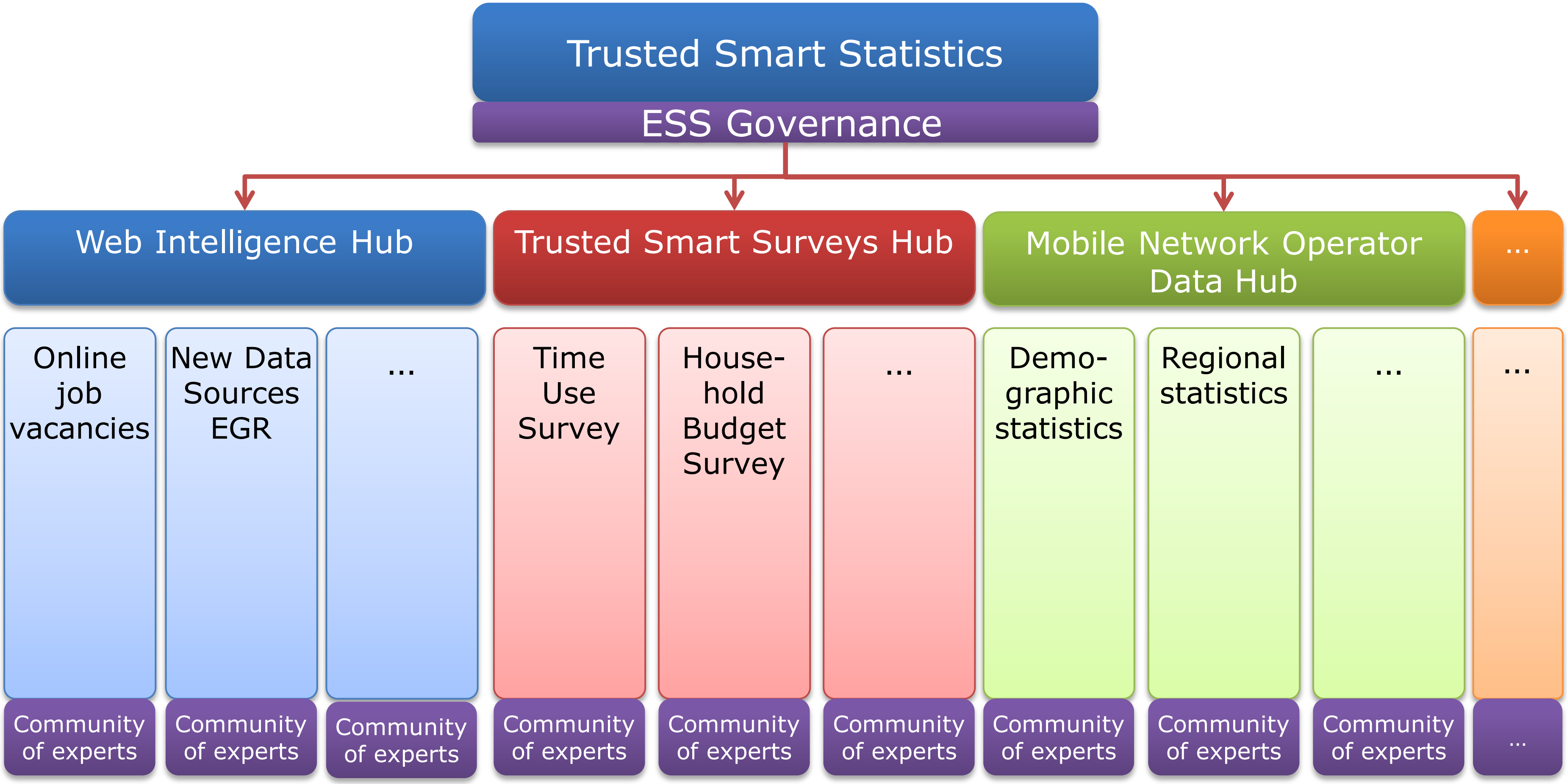
European
Commission

Trusted Smart Statistics - Web Intelligence Hub Update on developments on the use of OJA in official statistics,

Big Data for Labour Market Intelligence

6 December 2022

Trusted Smart Statistics

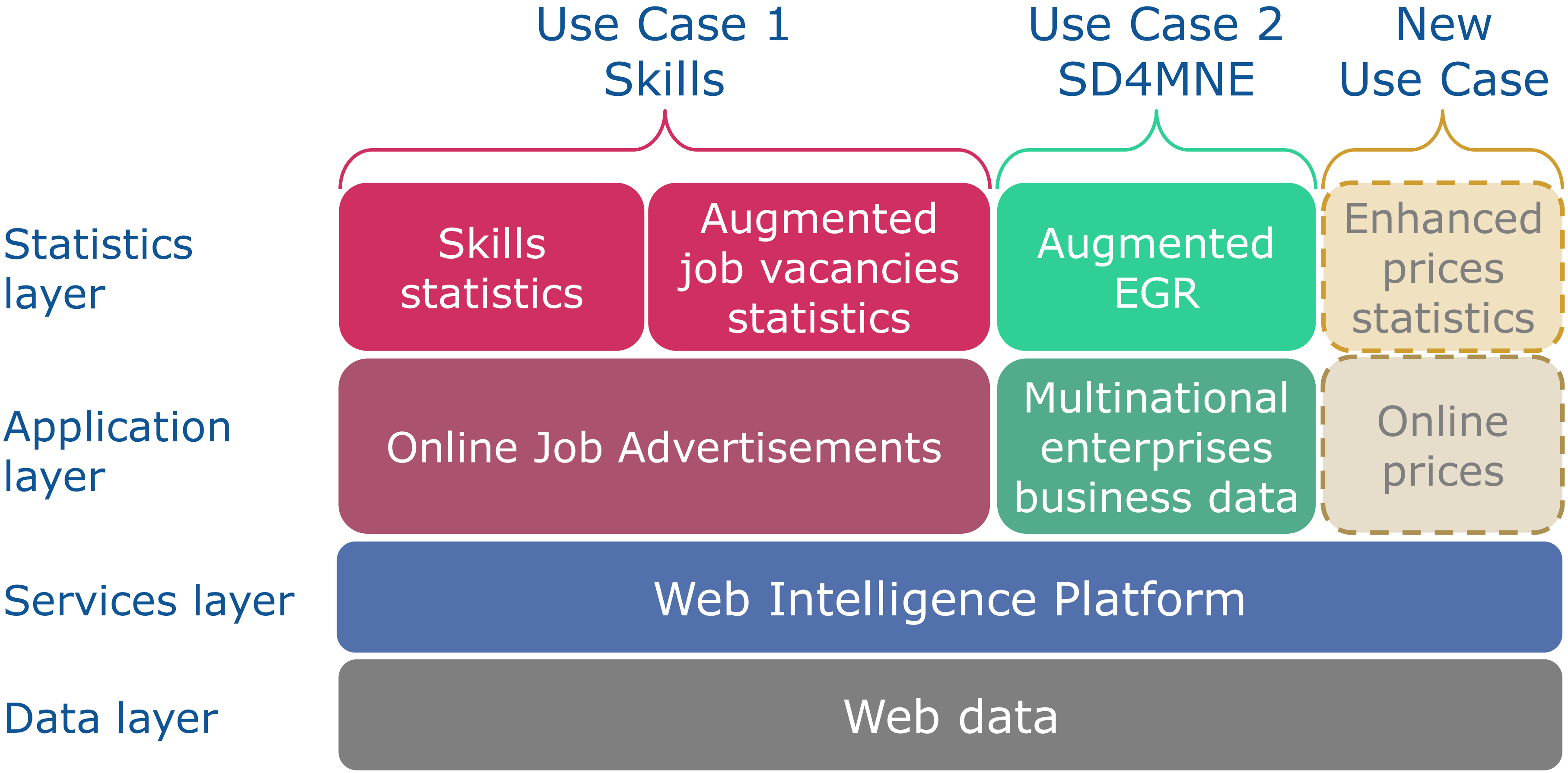


Trusted Smart Statistics

ESS Governance

- Web Intelligence Hub
  - Online job vacancies – Community of experts
  - New Data Sources EGR – Community of experts
  - … – Community of experts
- Trusted Smart Surveys Hub
  - Time Use Survey – Community of experts
  - House-hold Budget Survey – Community of experts
  - … – Community of experts
- Mobile Network Operator Data Hub
  - Demo-graphic statistics – Community of experts
  - Regional statistics – Community of experts
  - … – Community of experts
- …
  - … – …

Web Intelligence Hub



Use Case 1
Skills

Use Case 2
SD4MNE

New
Use Case

Statistics layer

Skills statistics

Augmented job vacancies statistics

Augmented EGR

Enhanced prices statistics

Application layer

Online Job Advertisements

Multinational enterprises business data

Online prices

Services layer

Web Intelligence Platform

Data layer

Web data

Web Intelligence Hub

- Know more about Trusted Smart Statistics and the Web Intelligence Hub:
  - **YouTube**
  - **ETF Open Space channel**
  - **Case study: Eurostat and Smart Statistics, by Fernando Reis (Eurostat), 8 June 2021, EN**
  - **https://youtu.be/bAUT52D1MpM**

Updates on OJA

- Labour market concentration
  - **Competition in urban hiring markets: evidence from online job advertisements**
  - **https://ec.europa.eu/eurostat/web/products-statistical-working-papers/-/ks-01-21-430**
- Validation and continuous improvement
- Linking to business register and NACE evaluation
- OJA gold standard for occupation
- OJA time-series

OJA GOLD STANDARD FOR OCCUPATION

Purpose of collecting labelled data

1 Monitor and improve the quality of the OJA data production

- Evaluation of classifiers → perform quality checks of the **data classified**
- Measurement of the accuracy of the **classifiers**
- Improvement of “ontologies” / dictionaries

2 Provide annotated OJA data

HOW?

- Gold standard: benchmarking annotators
- Monitor quality of the automatic classification process
- Train Machine Learning models

OJA-NLP dataflow

- **OJA-NLP data flow (released yearly):**
  - **Set of raw data extracted for a sample of ads:**
    - **Job description full text**
    - **Job title**
    - **Classified data**
    - **Additional information used to classify (matched token, dictionary terms…)**
  - **Stratified sample:**
    - for each **language** (25 languages, excluding very small languages,
    - at least **50 ads for each category**),
    - stratify by: occupation, contract, salary, working hours, education, economic activity, experience.
  - **Period covered: Oct 2021 – Jan 2022**

Data annotation: Definition

- **Process of labelling the data → various formats: text, video or images.**
- **Refers to the human classification of raw data → annotated (labelled) data.**
- **Enrich a linguistic data collection with annotations (NLP, Semantic Web technologies, LOD Linked Open Data Cloud, …) → machine learning, quality purposes**
  - *Corpora - datasets of natural language*
  - *Annotated corpus* - single set of data annotated with the same specification:

1. audit samples;
2. train, validation and testing of machine learning models;
3. estimate the precision of the classifiers.

Data annotation: Quality

- Quality of annotated OJAs
  - **Measurements:**
    - ✓ **Accuracy** → how close a label is to the truth.
    - ✓ **Consistency** → degree to which multiple annotations on various training items agree with one another.
  - **Standard methods to assure annotated data of high quality:**
    - ✓ Use of gold standards, **consensus**, and **auditing**.

Data annotation: Gold standard

- Quality of data annotation

  - **Use of gold standards (or Benchmarks) method**
    - ✓ Gold standard: sample carefully built by experts with a very high precision.
    - ✓ Expensive: require the intensive team work of highly qualified experts.
    - ✓ Small in relative terms and are normally not be sufficient for training ML models.
    - ✓ Ideal for **benchmarking annotators** used to obtain other types of annotated data.

Gold standard for occupation of OJA data

❑ Prepare samples for annotation

- Extract reproducible samples
- (country, language) pair

❑ Variable to annotate: ‘Occupation 4D’ (and lower levels if necessary)

❑ Variables included in the sample:

- *oja_id,*
- *title,*
- *full job description,*
- *occupation,*
- *match_token,*
- *dictionary term*

Specific guidelines: How to label OJA data?

Labels used in the annotation process:

1. Labels = ISCO-08 labels (for occupation projects)
2. Metadata labels
   - ✓ Correct
   - ✓ Incorrect
   - ✓ Comment
   - ✓ No reference to occupation in the description
   - ✓ Impossible to classify at 4th level
   - ✓ Wrong language
   - ✓ Not a job ad
   - ✓ Job description missing
   - ✓ Multiple ISCO labels
   - ✓ Misspelling

First 4 countries: Preliminary results

- Nice set of countries-languages:
  - **AT, IT, SI, BG**
  - **It covers Latin, Germanic and Slavic languages**
  - **Results are broadly confirmed for a few more countries for which we recently received data**
  - **But they are still preliminary: typically one annotator per country, and dealing with a very complex classification system**
- Aggregate accuracy rate (correct/[correct+incorrect]) at ISCO-4d is 44.9%
  - **Not bad for a classification with 400 categories (a random assigning would score around 0.25%)**
  - **This is stratified, so rarer and more difficult occupations get more weight – The accuracy rate increases by some 10-15pp when weighting by frequency**

OJA TIME-SERIES

Relevance of OJA time series analysis

- Answering many relevant question: What was the impact of covid on the labour market? How is skill demand changing? Etc.
- Offering external validation to OJA indicators by testing its relationship to other relevant series
- Augmenting other indicators through nowcasting
  - **OJA data are more timely than other series, so they could be used for nowcasting**
  - **E.g. nowcasting of job vacancies in the latest quarter**
- Flows or stocks of OJA time series could be more relevant, depending on the type of analysis

Issues with OJA time series analysis

- OJA data ingestion happens through scraping or through API download and its regularity is less than ideal.
- Scrapers can fail for reasons like changes in website structure or temporary overload;
- The market for online ads changes => need to update the list of data sources
- This leads to the following three problems / solutions:
  1. Missing data -> Estimated through a survival analysis model
  2. 0-to-N peaks -> Even allocation of scraped ads to days within the scraping interval
  3. Incomparable source sets -> Chaining

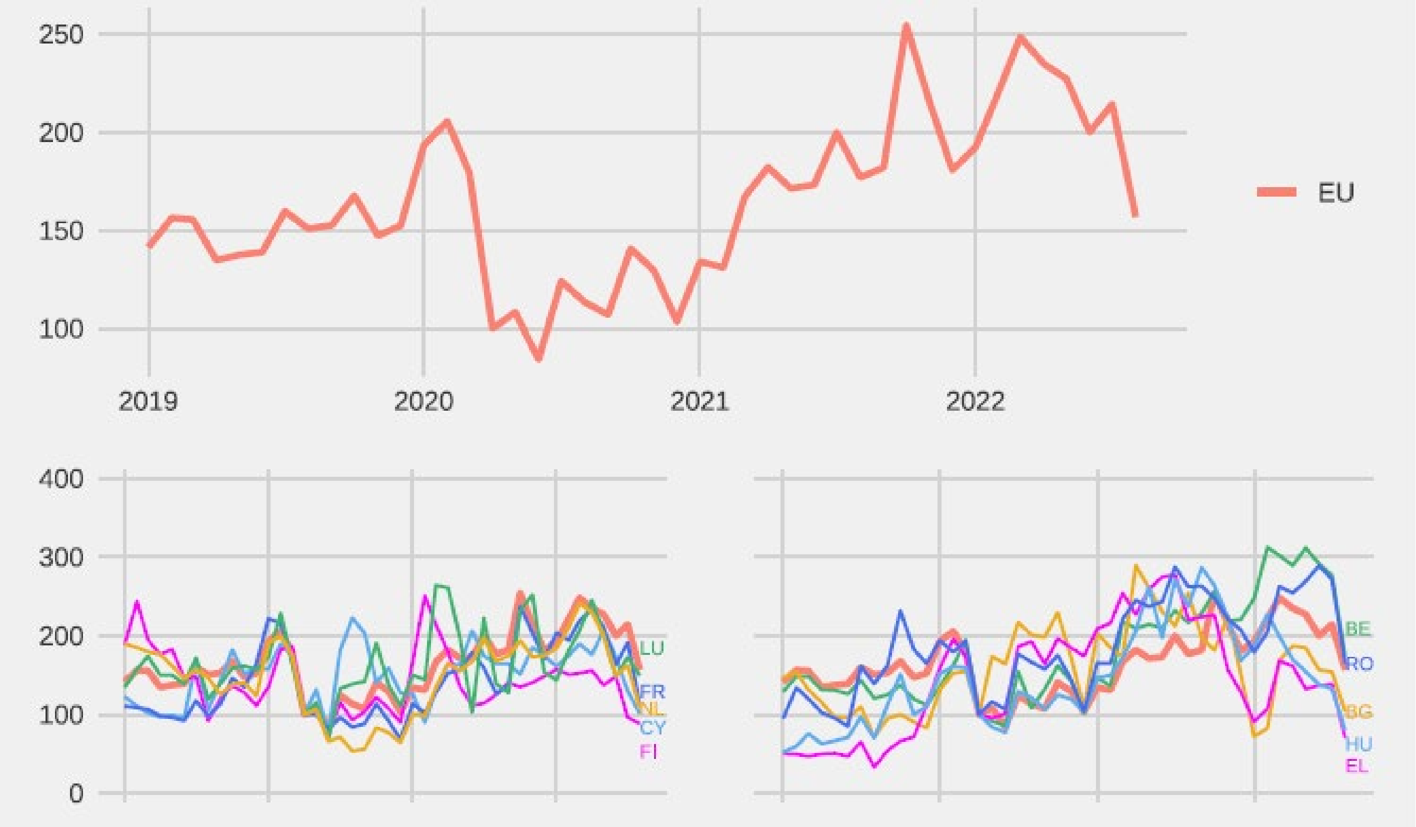
R implementation

- Queries are executed in SQL from the cloud by using the DBI and noctua R packages (Ascheri et al 2022, https://www.revistadestatistica.ro/wp-content/uploads/2022/03/RRS-1_2022_1.pdf)
- Processing is slow: 200m ads divided into some 30 countries X 40 occupations X 400 sources X 1600 dates ≈ 800m cells
- The script is divided into 10 steps loading/saving intermediary input/output
- Other metadata, like discarded ads at each step and a "EU tracker" projecting the EU total at each step, are saved at each step

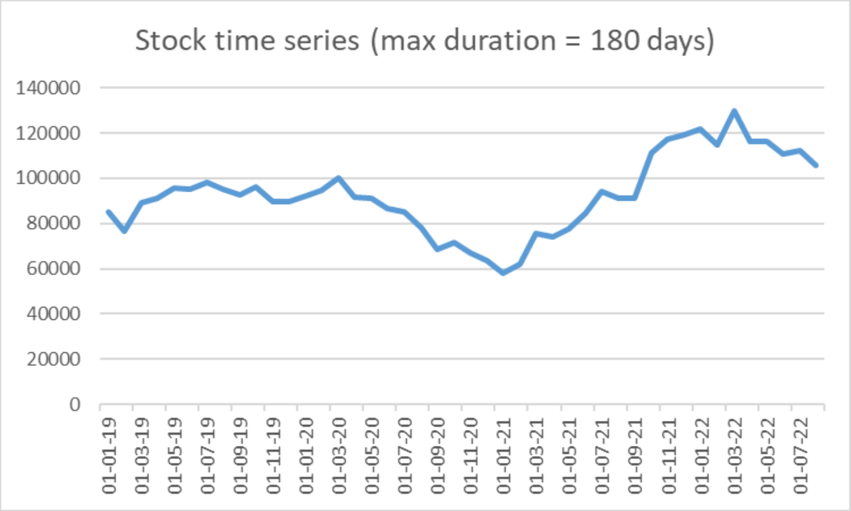
Implementation: Users

- A table with a set of weights by occupation / country / month will be made available to users
  - **So they can replicate easily the OJA time series and do their own analysis**
- The weight for each cell is calculated as the number of posted ads estimated through the time series method and the number of ads found in the OJA dataset
  - $w_{o,c,m} = N_{o,c,m}^{time-series} / N_{o,c,m}^{raw}$
- Users will only need to merge their OJA queries with the time-series-weight table and will be good to go

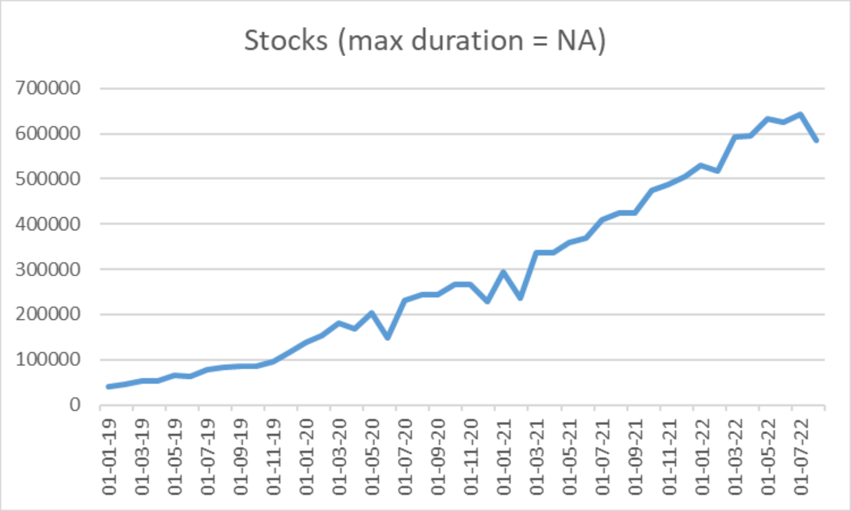
Country series: Flows



Stocks

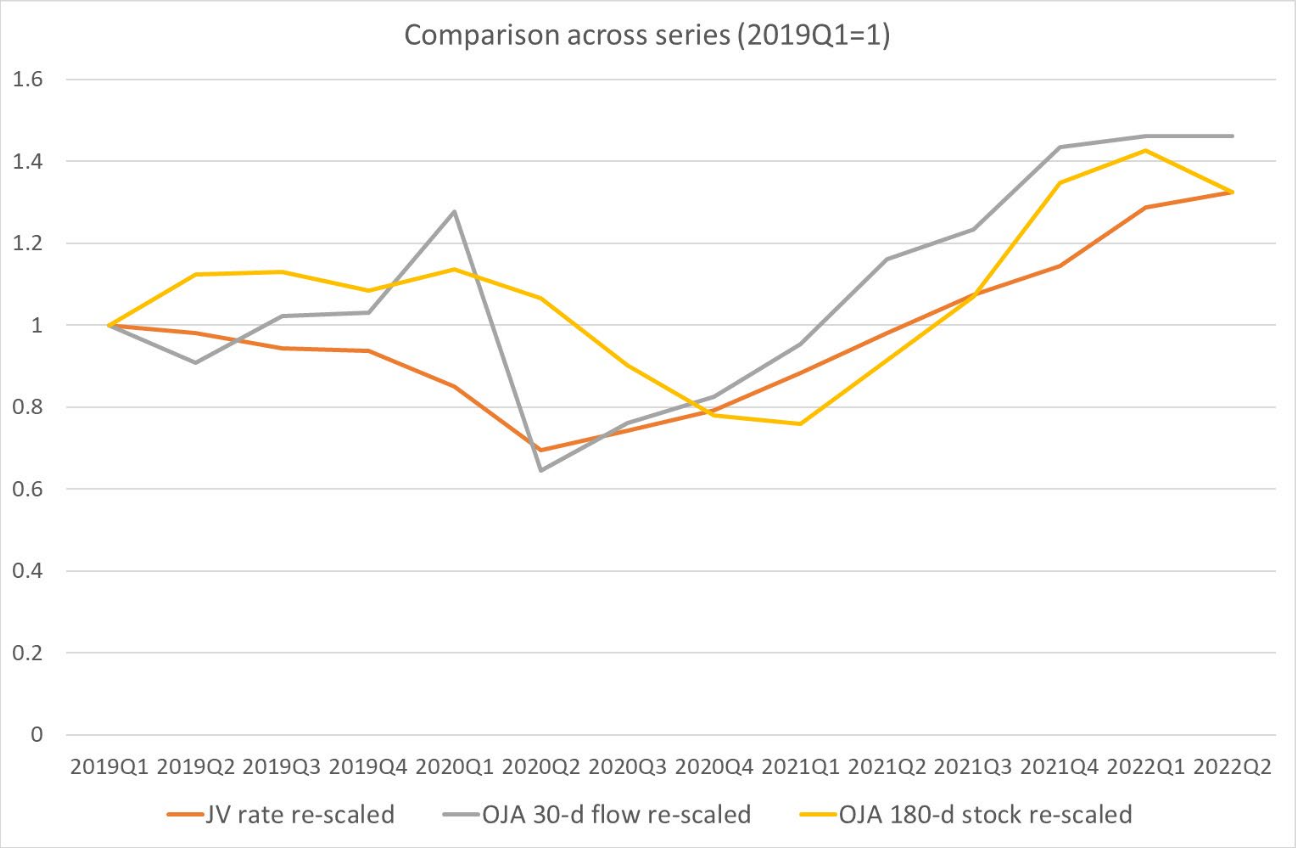


Stock time series (max duration = 180 days)



Stocks (max duration = NA)

Comparison of JV and OJA series



Results will be presented at NTTS

- New Technologies and techniques for Statistics 2023
- https://ec.europa.eu/eurostat/cros/content/NTTS2023_en

European Commission

Thank you for your attention

Fernando Reis

Eurostat

✉ fernando.reis@ec.europa.eu

https://github.com/reisfe/

https://twitter.com/reisfe/

https://linkedin.com/in/reisfe/