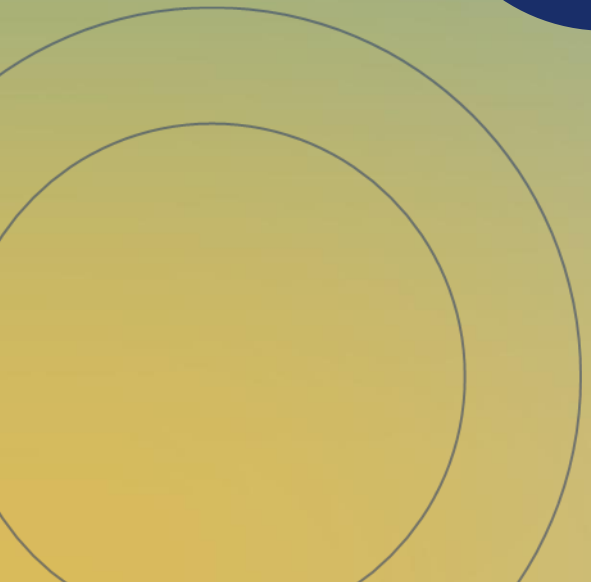
EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS

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

Integration of administrative and survey data in a Short-Term Business Statistics with statistical learning algorithms

Sandra Barragán, David Salgado, Sergio Pardina, Esther Puerto

S.G. for Methodology and Sampling Design

Statistics Spain

June 2024

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

- Introduction
- Input Quality
- Methodology
- Results
- Conclusions

The conference is partly financed by the European Union

Introduction: motivation

- Use case in a short-term business statistics: Services Sector Activity Indicators (SSAI)
  - Monthly
  - Target: turnover
- Use of Administrative registers (monthly VAT):

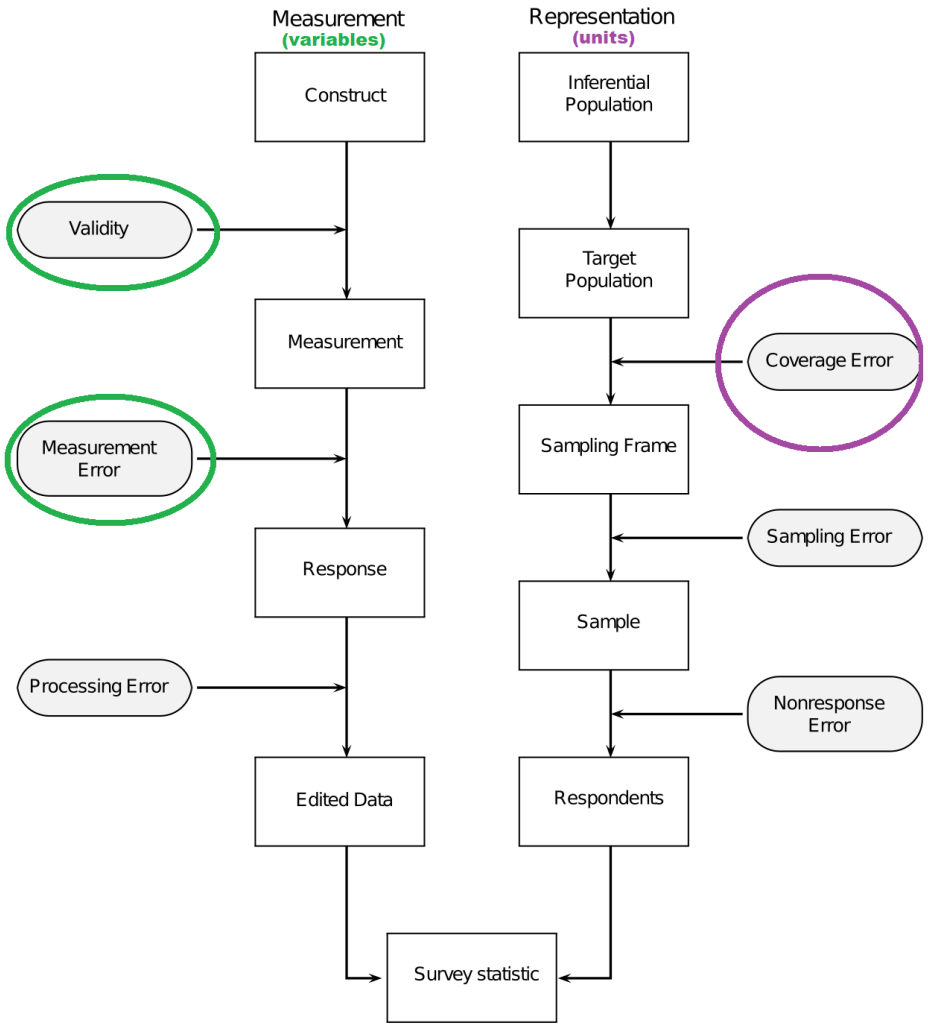
Quality dimensión	Goal	Reality in Admin Reg (monthly)
Burden reduction	Remove questionnaire	Only large enterprises
Timeliness	m+30d	Available in m+35d

The conference is partly financed by the European Union

Introduction: the problem

- In the administrative registers the errors are not kept under control due to the external mechanism of data generation.

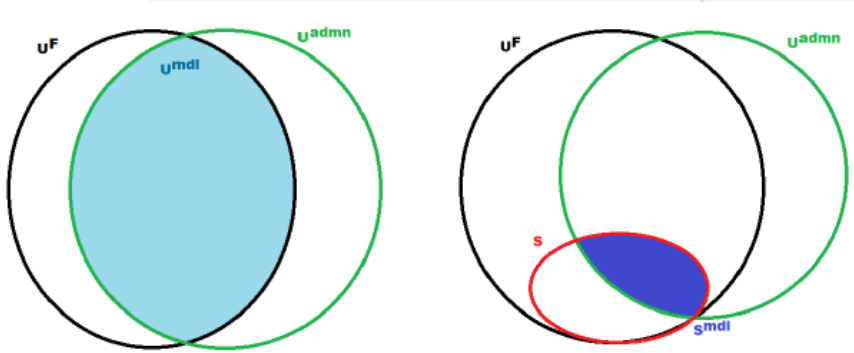
Total Error model (see [Groves and Lyberg, 2010] and the two-phase life-cycle model by Zhang [2012])



The conference is partly financed by the European Union

Introduction: notation

- U the finite population of analysis.
- $U_F \subset SBR$ the frame population from the Statistical Business Register (SBR).
- U^{adm} the set of business units contained in the tax register.
- $U^{mdl} = U^{adm} \cap U_F$ the set of statistical units in the tax register.
- $S^{mdl} = S \cap U^{mdl}$ the set os statistical units in the training data set.
- y is the statistical variable of interest (turnover in this use case).
- y^{stat} denotes the values obtained with the survey.
- y^{adm} denotes the values obtained with the administrative register.



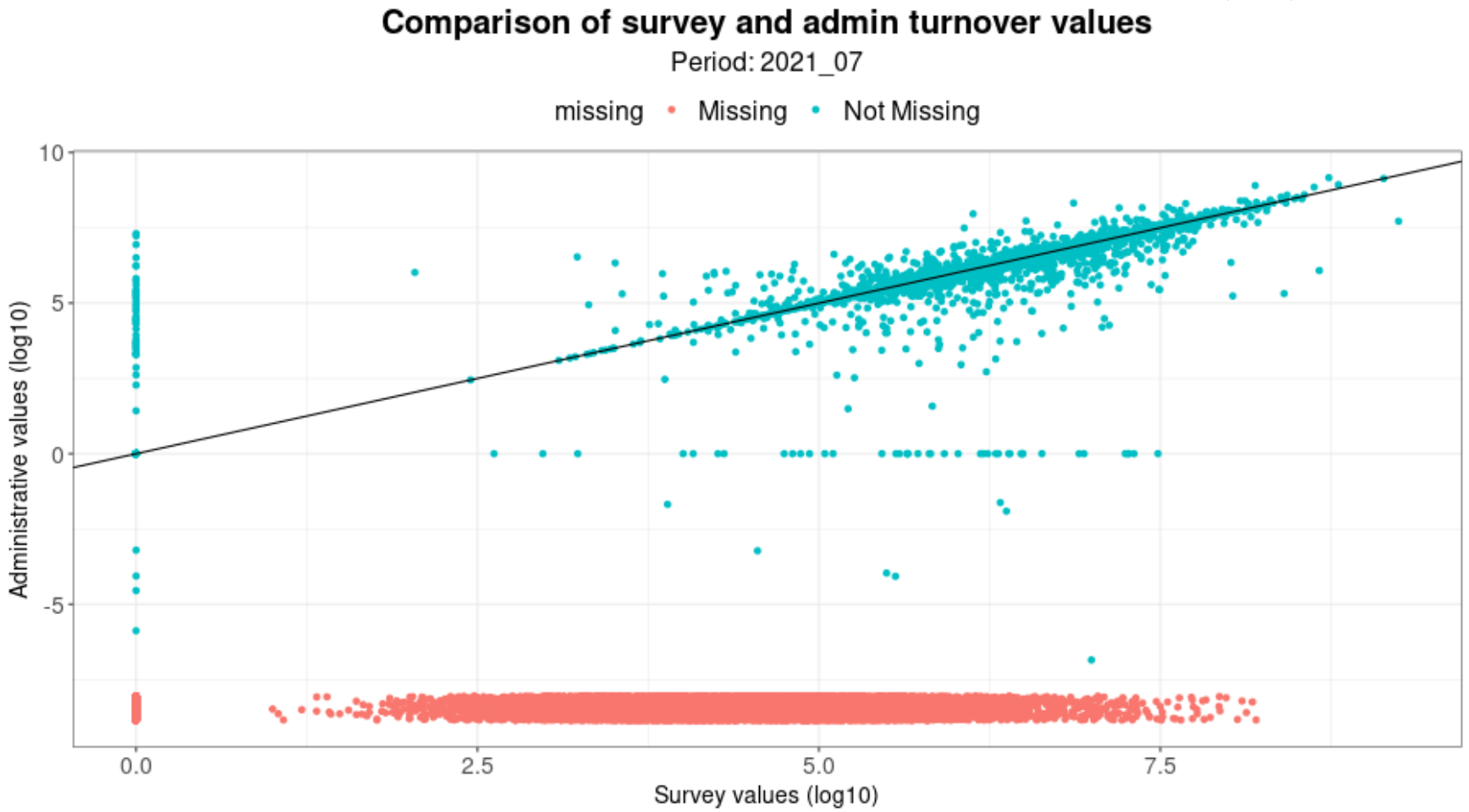
The conference is partly financed by the European Union

Input quality: microdata

For details, see the session of this conference Q2024:

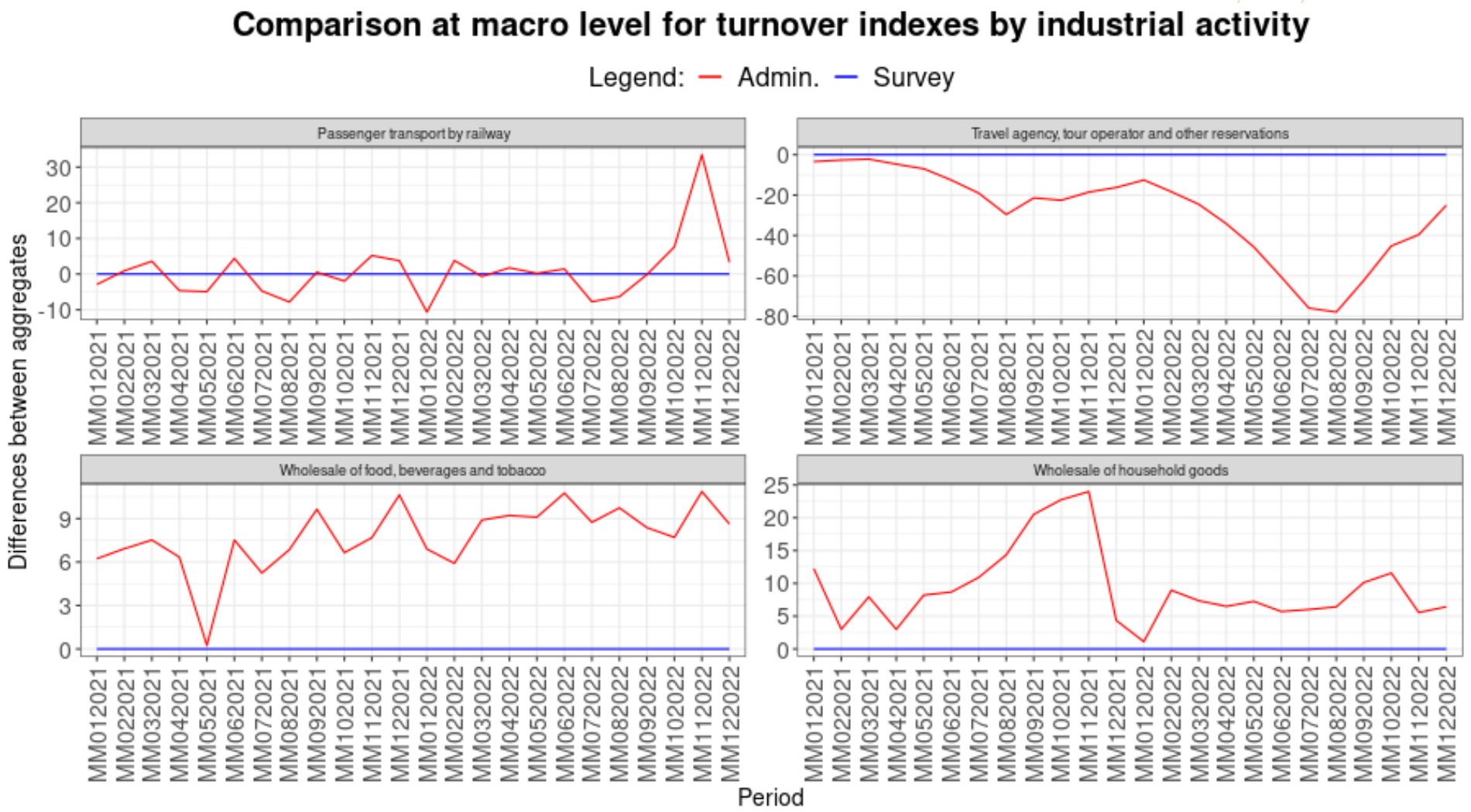
Measuring the quality of administrative sources: at macro level with novel indicators and micro level with distributions comparison. (Nieto et al.)

Session 7 - Quality of administrative data, June 5, 2024, 14:00-15:30



INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

Input quality: macrodata



Methodology

- REPRESENTATION: coverage error $k \in S^{mdl} = S \cap U^{adm}$
- MEASUREMENT: Validity and measurement errors. The statistical variable is predicted by the synthetic target $y^{\circ} = f\left(y^{adm}; \mathbf{x}\right) + \epsilon$

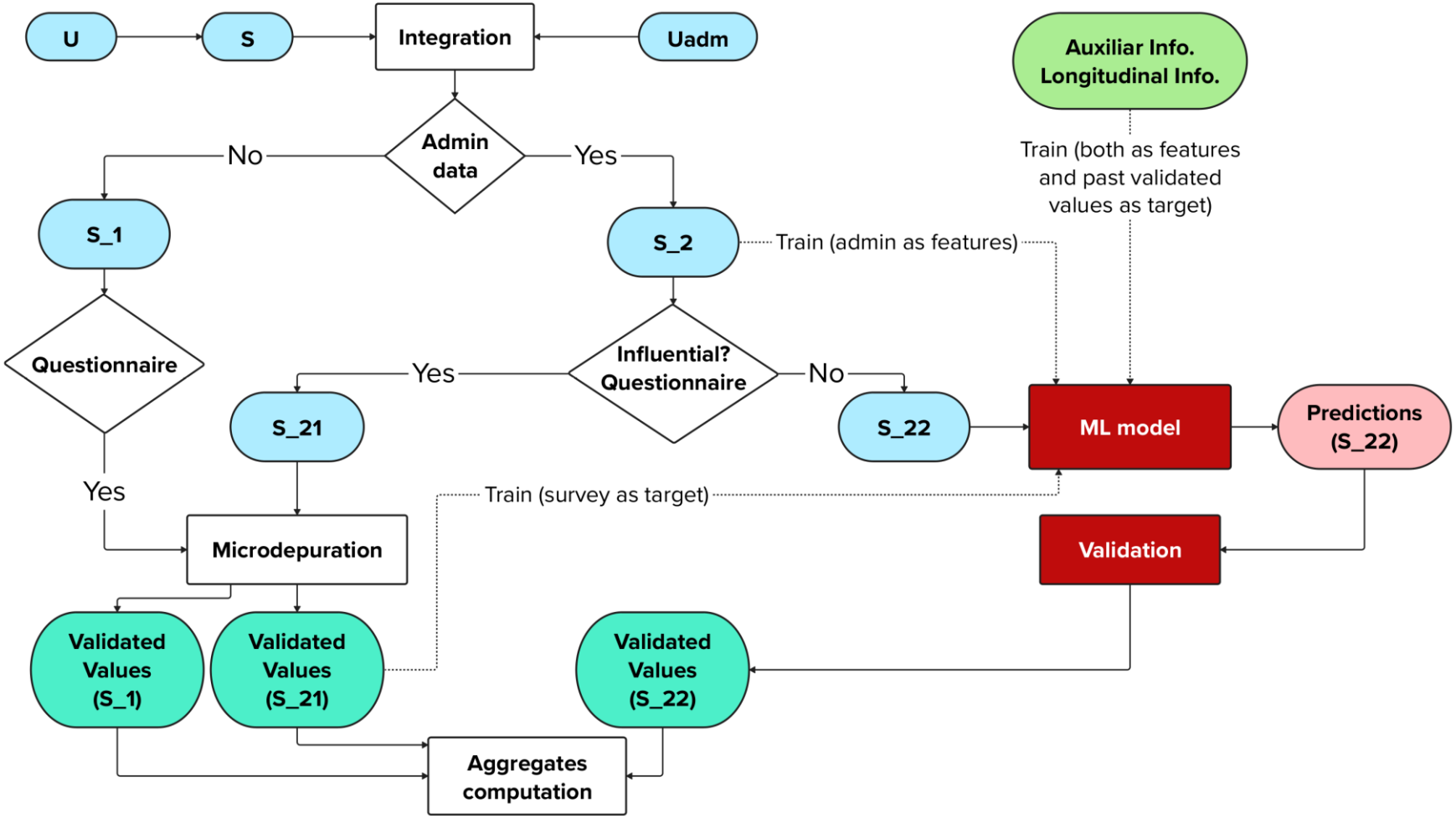
  which is calculated as follows:

$$\hat{y}_{kt}^{\circ} = \hat{f}\left(y_{kt}^{adm}; x_k\right) \; k \in S_{22} \subset S^{mdl}$$

Then, the predictions $\hat{y}_{kt}^{\circ}$ are validated to be used in the computation of the aggregates and in the training dataset of the following reference periods $t+1, t+2 \ldots$

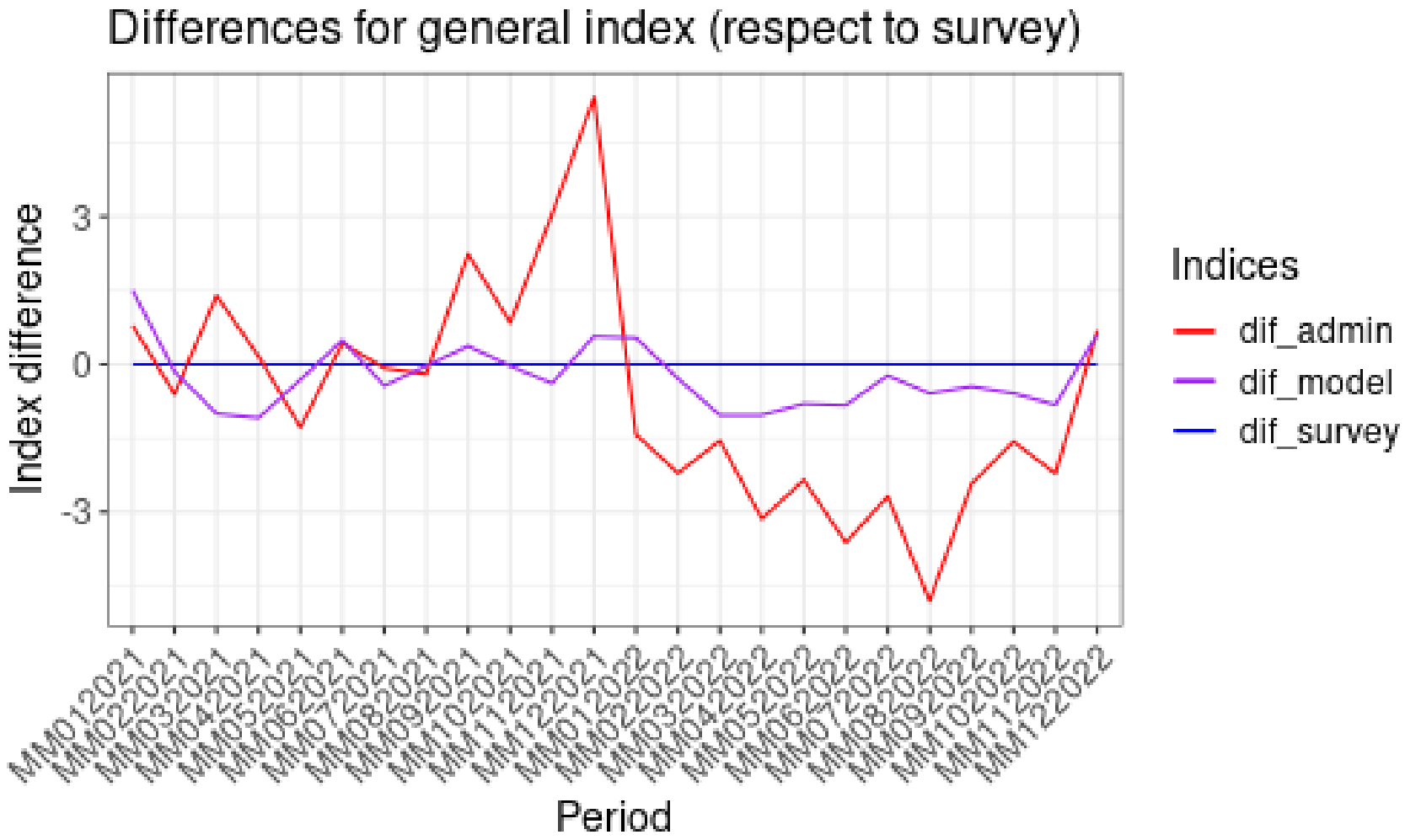
The conference is partly financed by the European Union

Methodology: proposed process



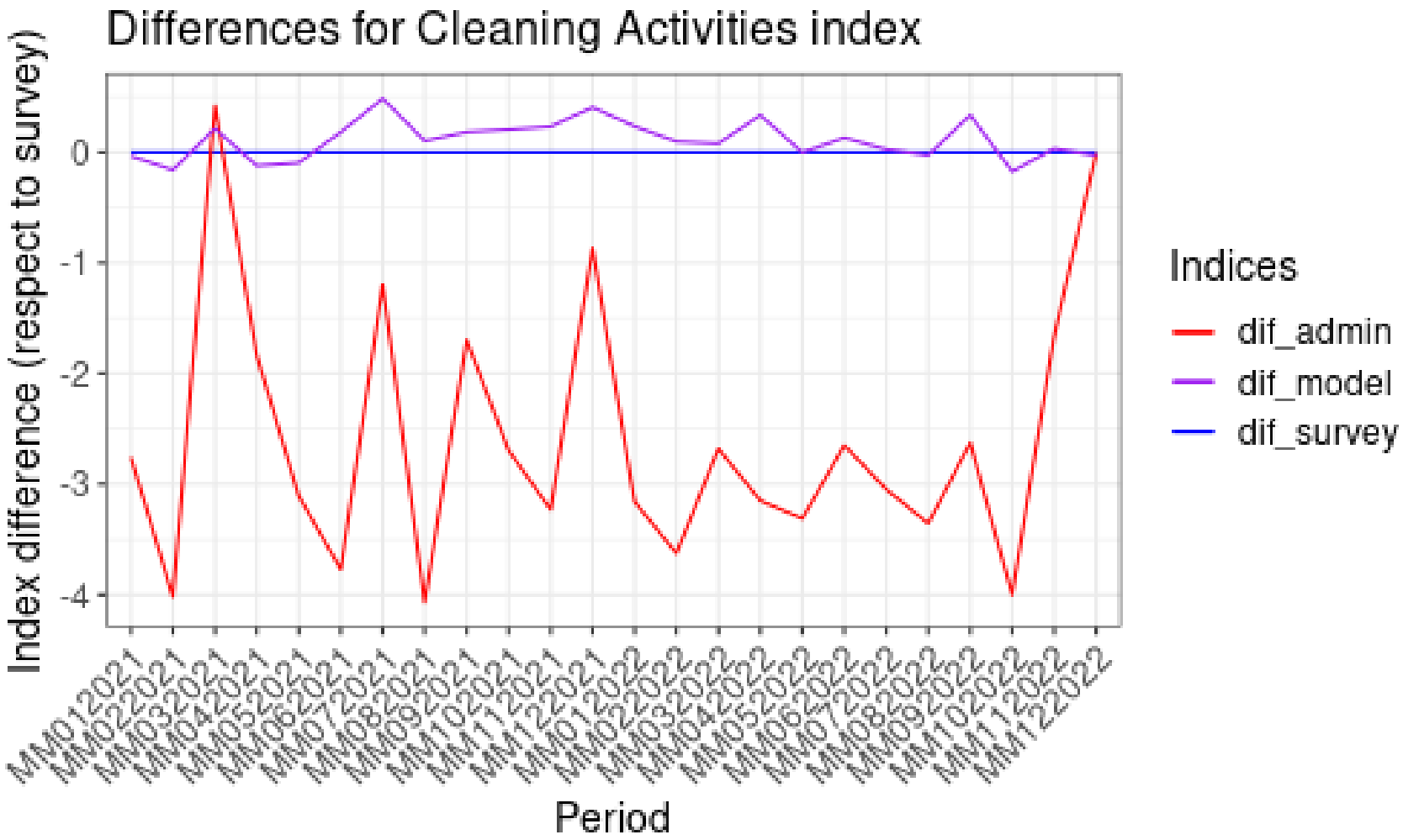
The conference is partly financed by the European Union

Results



The conference is partly financed by the European Union

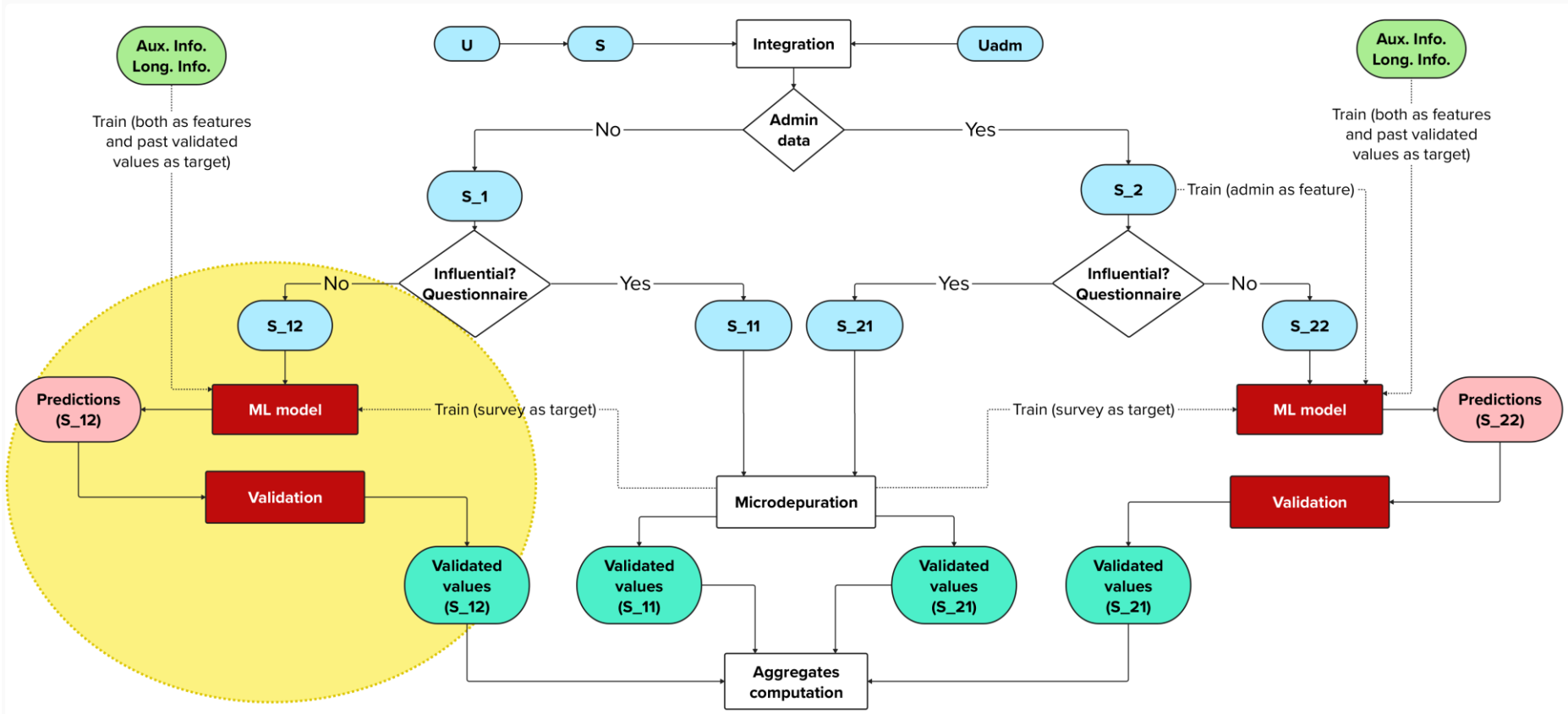
Results



The conference is partly financed by the European Union

Conclusions: future work

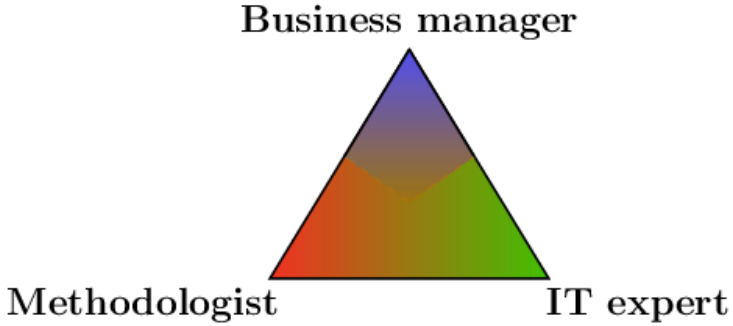
- Indicator to evaluate the quality of data sources across multiple dimension
- Estimate variances and MSE when combine sampling designs and model predictions
- Accuracy (lack of measurement errors) and response burden reduction



The conference is partly financed by the European Union

Conclusions

- The differences observed in the exploration of the input quality cautiously **discourage** the use of the administrative values by **mere substitution** as the statistical values.
- End-to-end statistical production process **integrating administrative data with survey data** in a probability sample where synthetic values are computed **using a statistical learning model** by using the administrative data as regressors as well as longitudinal information.
- To account for validity and measurement errors in administrative data using statistical learning models aiming at response burden reduction, a **selection of units** maintaining the model quality is advised.



EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Integration of administrative and survey data in a Short-Term Business Statistics with statistical learning algorithms

Sandra Barragán, David Salgado, Sergio Pardina, Esther Puerto

S.G. for Methodology and Sampling Design

Statistics Spain

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

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly
financed by the European Union