



INSTITUTO NACIONAL DE ESTATÍSTICA STATISTICS PORTUGAL



The conference is partly financed by the European Union





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

> Sandra Barragán, David Salgado, Sergio Pardina, Esther Puerto





The conference is partly financed by the European Union

S.G. for Methodology and Sampling Design Statistics Spain

June 2024



INSTITUTO NACIONAL DE ESTATÍSTICA STATISTICS PORTUGAL

eurostat O

The conference is partly financed by the European Union

- Introduction
- Input Quality
- Methodology
- Results
- Conclusions





INSTITUTO NACIONAL DE ESTATÍSTICA STATISTICS PORTUGAL

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



INSTITUTO NACIONAL DE ESTATÍSTICA STATISTICS PORTUGAL

eurostat 🖸

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])





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.



eurostat

financed by the Europe

INSTITUTO NACIONAL DE ESTATÍSTICA Statistics Portugal





INSTITUTO NACIONAL DE ESTATÍSTICA STATISTICS PORTUGAL

Comparison of survey and admin turnover values

Period: 2021 07

eurostat O

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



Input quality: macrodata

Passenger transport by railway Travel agency, tour operator and other reservations 30 -20 20 -40 10 Differences between aggregates 0 -60 80 MM112022 M042022 M052022 M062022 M072022 MM082022 MM092022 102022 122021 MM022022 MM032022 MM042022 MM052022 MM062022 MM072022 MM092022 2022 MM022022 MM032022 MM012022 MM082022 MM102022 MM012022 IM042021 IM052021 12021 1202 M03202 MM12202 22022 MM03202 MM06202 MM07202 MM08202 MM09202 MM10202 M02202 **MM04202** MM05202 MM06202 MM07202 MM08202 MM09202 MM10202 MM01202 M02202 **MM01202** ---MM Ż Ż Ň MM \geq \geq \geq \geq Wholesale of food, beverages and tobacco Wholesale of household goods 25 20 9 15 6 10 3 5 0 MM052022 -MM062022 -MM072022 -MM112022-MM122022-MM032022 -MM042022 -MM082022 -MM092022 -12022 MM112021 MM122021 MM022022 2022 MM012022 MM022022 MM032022 MM042022 MM082022 MM092022 MM102022 MM012022 MM052022 MM062022 MM072022 MM102022 MM072021 MM082021 MM062021 MM092021 MM102021 MM01202 MM04202 MM05202 MM01202 MM02202 MM03202 MM04202 MM05202 MM06202 MM07202 MM08202 MM09202 MM10202 MM11202 MM12202 MM02202 MM03202 Ň --Ž M

Period

Comparison at macro level for turnover indexes by industrial activity

STATISTICS PORTUGAL

Legend: - Admin. - Survey

INSTITUTO NACIONAL DE ESTATÍSTICA

The conference is partly

financed by the European Union

eurostat O





eurostat 🖸

The conference is partly financed by the European Union

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 $y^{\circ} = f(y^{adm}; \mathbf{x}) + \epsilon$ which is calculated as follows:

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

Then, the predictions \hat{y}_{kt}° 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...



EUros

eurostat 🖸

The conference is partly financed by the European Union

Methodology: proposed process







INSTITUTO NACIONAL DE ESTATÍSTICA Statistics Portugal

eurostat 🖸

The conference is partly financed by the European Union

Results

Differences for general index (respect to survey)







eurostat 🖸

The conference is partly financed by the European Union

Results





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



INSTITUTO NACIONAL DE ESTATÍSTICA

STATISTICS PORTUGAL

The conference is partly

financed by the European Union

eurostat O







financed by the European Unior

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.





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





The conference is partly financed by the European Union





INSTITUTO NACIONAL DE ESTATÍSTICA STATISTICS PORTUGAL



The conference is partly financed by the European Union