Measuring the quality of administrative sources: at macro level with novel indicators and micro level with distributions comparison

Alicia Nieto[1], Sandra Barragán[1], Alba Rodríguez[1], Soledad Saldaña[1], David Salgado[1]

[1]S.G. for Methodology and Sampling Design, Statistics Spain, Spain

Abstract

In the production of official statistics there are three main data sources: surveys, administrative registers, and (privately held) digital sources. The use of administrative sources is lately increasing, however there is a lack of control in the quality of these new sources. The administrative data can be used in different ways, the most challenging is to use them as primary source of data, directly or indirectly to compute the target aggregates.

The advantages of administrative data as a primary source are widely known improving different quality dimensions (reduction of response burden, cost savings, increase of granularity, etc.), but the disadvantages must be considered carefully. In terms of the representation and measurement lines in the Total Survey Error paradigm and the Two-Phase Life-Cycle model by Zhang, errors both related to units and to variables are present. Coverage errors arise when identifying units in the target population and validity errors proliferate because of the differences between concepts for statistical and administrative purposes. Therefore, a need to measure the quality of input data emerges as a consequence of the data generation process lying out of the control of NSIs.

In official statistics several quality and performance indicators are used but the focus is on measuring the quality of the output, and most of them have been designed to be used when using survey data as input. So, there is a need to broaden the list of quality indicators to provide room for those quality measures of multisource statistics and even more in the case of statistics based only on administrative data.

At Statistics Spain we are carrying out an exercise to measure the quality of the administrative data used in several short-term statistics of different domains/characteristics. In this work we present the proposal to measure the quality of the input with some indicators for the administrative data source. Moreover, we take advantage of the access to both administrative and survey variables for a part of the sample to directly compare the distributions of the target variable under study.

The ultimate goal is to provide objective measures of the direct use of administrative values without further treatment to gain some knowledge about the quality of the final estimates in comparison with fully survey-based traditional results. This analysis may help us decide regarding the need of treatment of administrative sources to keep under control their disadvantages and ensure the quality of admin-based final outputs.

Keywords: quality indicators, administrative data, survey-admin comparison

1. Introduction

For several years now, Statistics Spain (INE) has been promoting the use of new data sources in its production system. A first motivation arises from the legal point of view, since the National Statistical Plan (the main instrument for organizing the official statistical production of the Spanish Public Administration and specifically of Statistics Spain) establishes among its general strategic lines the use of new sources of information based on the intensification of the use of administrative records. In this line, the latest modification of the Spanish National Statistics Act, as of July 2022, promotes the reuse of administrative data for official statistical production purposes by stating that this data will be considered to the extent feasible as the primary data source. In addition, the law explicitly recognizes the legal support for national statistical authorities to collect administrative data from ministry departments, public organisms and public entities belonging to the Spanish Public Administration.

This paper presents our efforts to replace the traditional fully survey-based statistics, monthly conducted using a stratified probabilistic sampling design, with a combination of survey and administrative data from the Spanish National Tax Agency. These administrative registers provide monthly data on purchases, sales, revenues, and VAT deductions for large companies (monthly billing over 6 million euros).

The following ongoing work takes advantage of the exceptional circumstance by which we have access to both survey and admin legal-unit-level microdata from 2019 to the present, an unlikely scenario in common production conditions that allows us to conduct the thorough comparison proposed in this paper.

2. Data description and comparison

We shall use the Service Sector Activity Indicators (SSAI) short-term business statistics as our use case. We provide a concise description of both data sources, namely survey microdata including the main characteristics of the statistical business register used as the population frame and the Tax Register providing the administrative information.

2.1. Survey data

We shall identify:

1. The **target population** for year *y*, denoted by U^y and defined as the set of all companies that existed in Spain throughout the year *y*. As indicated by the notation, the target population is updated on an annual basis.

2. The **population frame** for the **Service Sector Activity Indicators** for year *y*, denoted by U_F^y and defined as **t**he subset of companies that existed in Spain throughout the year *y* extracted from the central statistical business register (DIRCE) with economic activity section codes G, H, I, J, L, M, and N according to CNAE-2009, the Spanish national official adaptation of NACE Rev.2. As indicated by the notation, the population frame is updated on an annual basis for sample selection purposes.

3. The monthly **probabilistic sample** *s* for month m and year *y*, denoted by $s_{my}^{surv} \subset U_F^y$ and defined as the set of companies selected according to a probabilistic sampling design *p(·)* from the population frame U_F^y.

For each sampling unit $k \in s_{my}^{surv}$, information is available on the value $z_{k,my}^{surv}$ of its monthly turnover.

2.2. Administrative data

The Spanish National Tax Agency collects data for tax purposes about monthly VAT returns via a web service. While large companies are legally required to file these returns, in practice, a number of them do not, and many others that are not legally obliged do. Additionally, the administrative classification of a company according to its economic activity may differ from the statistical classification in some aspects, further complicating the linkage.

We shall identify:

1. The **administrative population** for year *y* and month *m*, denoted by U_{my}^{adm} and defined as the set of companies contained in the Tax Register in month *m* and year *y*.

2. The **administrative population** for the **Service Sector Activity Indicators** for year *y* and month *m*, denoted by $U_{my}^{SSAI} \subset U_{my}^{adm}$ and defined as the set of companies from the Tax Register whose economic activity has been classified in the service sector (section codes G, H, I, J, L, M, and N) according to tax criteria and whose information is available for month *m* and year *y*.

For each administrative unit $k \in U_{my}^{adm}$, information is available on the value $z_{k,my}^{adm}$ of its monthly turnover.

2.3.Data comparison: survey vs. administrative data

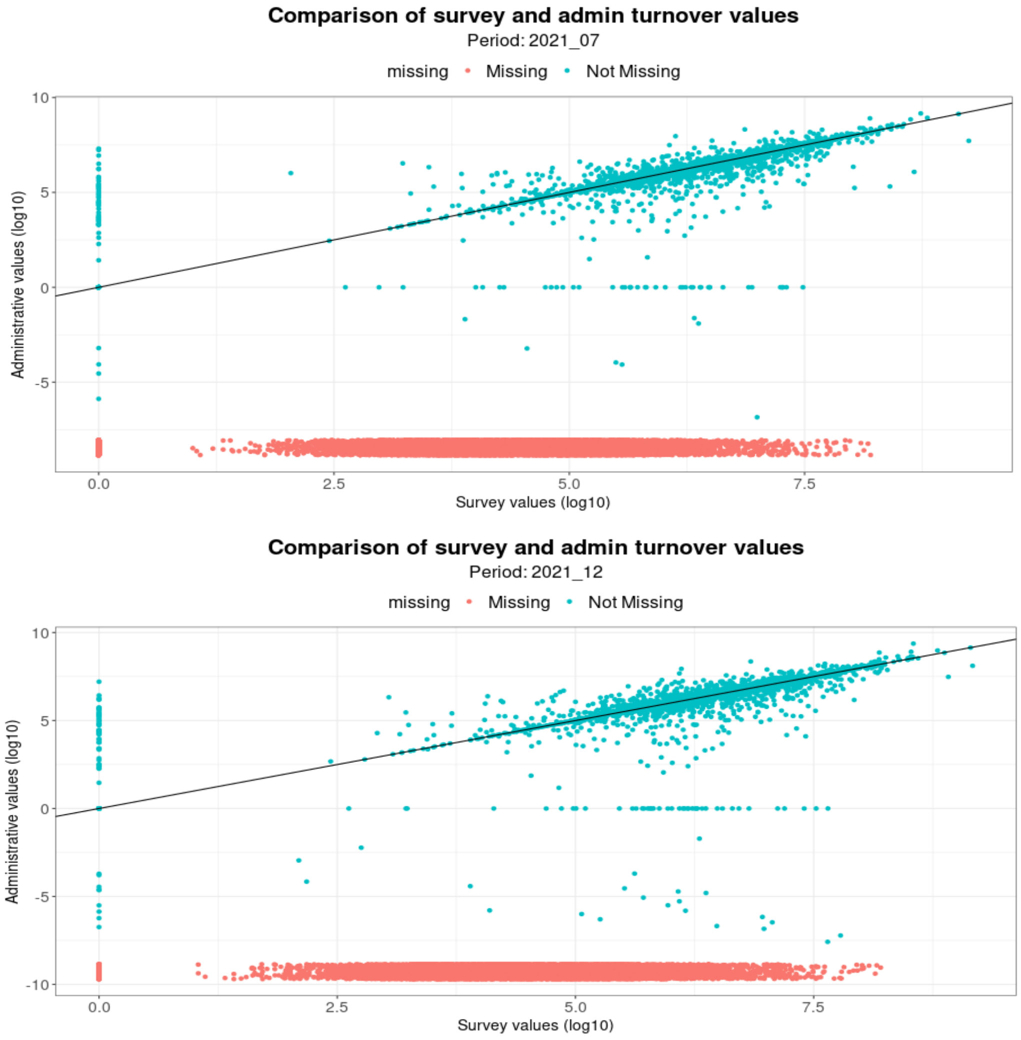
Our ongoing entire study focuses on analyzing the differences, similarities, and compatibilities between the survey values $\{z^{surv}_{k,my}\}_{k\in s^{surv}_{my}}$ and the administrative values $\{z^{adm}_{k,my}\}_{k\in U^{adm}_{my}}$ for years 2019 to 2022.

To show some specific issues we shall occasionally focus on July and December, 2021, as illustrative examples, since their behaviour diverges greatly from using survey data or administrative records.

2.3.1. Unit-level microdata

In order to make the comparison $\{z^{surv}_{k,my}\}_{k\in s^{surv}_{my}}$ and $\{z^{adm}_{k,my}\}_{k\in U^{adm}_{my}}$ we start by making a scatterplot as in figure 1.

Figure 1: In orange, those companies for which information is available in the survey but it is missing for the administrative record. In blue, those companies for which information is available for both sources.



Key points for these scatterplots (both are just illustrative examples):

1. **Missing values for the administrative source:** Focusing on the orange area (turnover values for companies with survey data but no administrative information), we can see that these missing values for the administrative source do not only happen for companies with small turnover, but also for those with monthly turnover even greater than 10^7 €. Although the weights of each company would need to be analyzed to see how much they influence on the final indices, such high missing turnover values can be expected to lead to a significant underestimation compared to the estimate using survey information.

2. **Zero values:** Both scatterplots highlight the values of the lines X=0 and Y=0, corresponding to companies whose monthly turnover is zero.

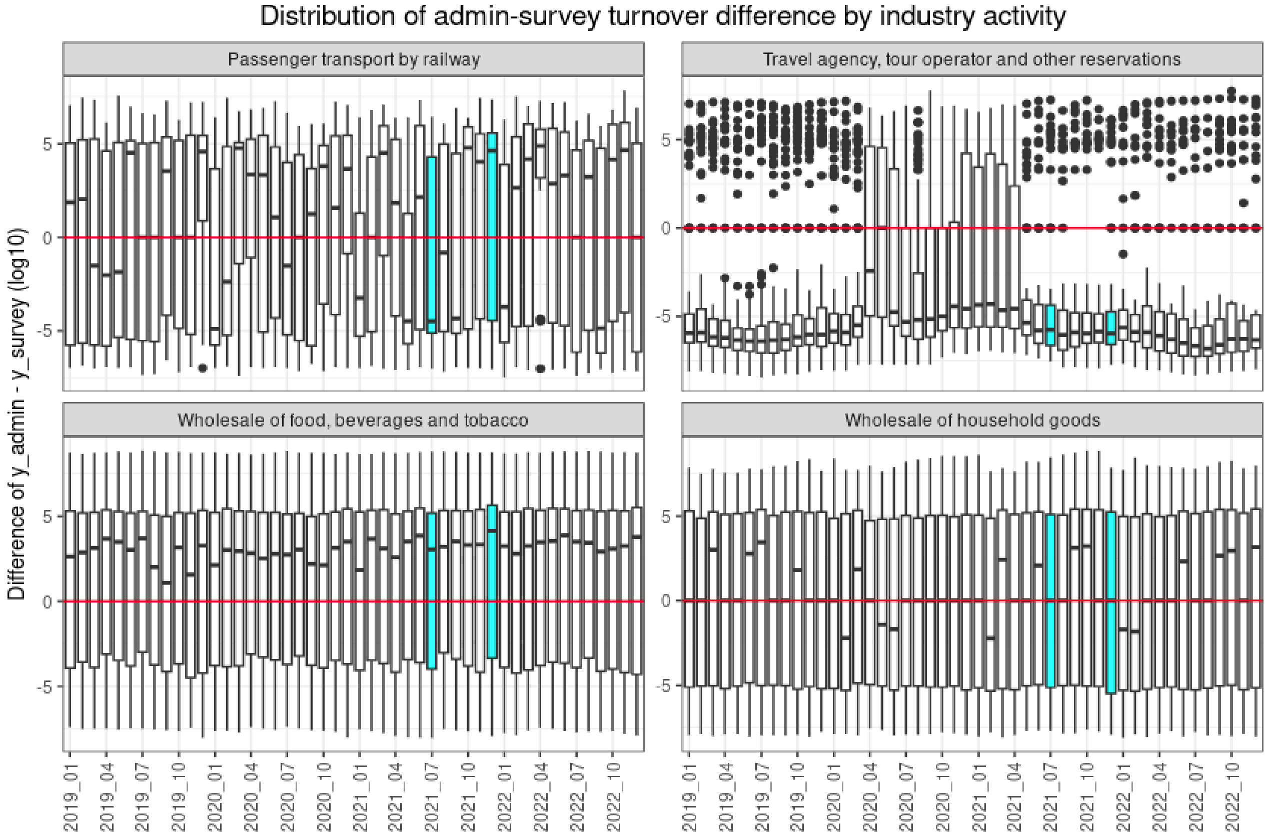
   While it may seem a priori valid that certain companies have zero turnover for some months (we can think of seasonal activities), there are many values showing this behaviour, which invites us to think about measurement errors in this data. We are currently investigating the cause of this phenomenon and its potential impact on the overall aggregate.

3. **Negative turnover:** In addition to the numerous zeroes we just discussed, it can be seen that a set of companies at the survey-admin intersection have **negative turnover** in the administrative source. This leads us to suspect that there may be slight differences between the statistical and administrative definitions and/or collection of the turnover, or that there exist measurement errors whose origin is still unknown to us (tax adjustment per periods, etc.).

Having qualitatively understood the overall comparative behaviour of the target variable in both sources, we move on to a preliminary assessment of the difference of both turnover values variable in each **activity sector**. The motivation resides in the final index computation, which is a compound Laspeyres index which needs elementary indices for each activity sector and NUTS2 region. This assessment is now carried out at the sampling unit level without consideration of aggregating, estimation, and weighting procedures.

Figure 2, which shows the distributions of the difference $z^{adm} - z^{surv}$ for 4 illustrative economic sectors, exemplifies the absence of a systematic pattern in the difference in turnover values across activity sectors. Some sectors exhibit a systematic overestimation of turnover (bottom-left subfigure), while others display alternating overestimation and underestimation (top-left and bottom-right subfigures). Still, others, though generally overestimated, exhibit an undeniably unpredictable pattern (top-right subfigure).

Figure 2: Administrative-survey turnover difference on a logarithmic scale for four selected economic sectors due to their heterogeneity among the thirty-six existing ones (by time period).



The comparison between these monthly distributions have been also addressed using the Kolmogorov—Smirnov distance between the empirical distributions of survey and admin values per sector (see figure A1 in appendix).

Despite the lack of homogeneity in the behaviour of the turnover difference values $z^{adm} - z^{surv}$ across activity sectors, we observe that empirical distributions of both sources behave similarly, as they exhibit small Kolmogorov-Smirnov distances (systematically less than 0.25 except for outliers[1]). Remember that this comparison has only been conducted for units *k* with both survey and admin data, i.e. $k \in s_{my}^{surv} \cap U_{my}^{adm}$. In the appendix we include also quantile-quantile plots for the pairs of distributions of $\{z_{k,my}^{surv}\}_{k \in s_{my}^{surv} \cap U_{my}^{adm}}$ and $\{z_{k,my}^{adm}\}_{k \in s_{my}^{surv} \cap U_{my}^{adm}}$ (figures A2 and A3) and of $\{z_{k,my}^{synth}\}_{k \in s_{my}^{surv}}$ and $\{z_{k,my}^{adm}\}_{k \in s_{my}^{surv}}$, with $z_{k,my}^{synth} = z_{k,my}^{adm}$ if $k \in s_{my}^{surv} \cap U_{my}^{adm}$ and $z_{k,my}^{synth} = z_{k,my}^{surv}$, otherwise (figures A4 and A5). Differences at the sampling unit level

[1] Outliers corresponding to a distance of 1, which are due to exceptional circumstances, occurred with taxis in Spain during 2019.

are explicit in the left tails due to the negative values in the administrative source; in the rest, the resemblance of quantiles is notorious.

2.3.2. Aggregated-level macrodata

After conducting the initial analysis of unit-level microdata (which, in summary, has shown similar behaviour between survey and admin sources), we proceed to compare aggregated turnovers and indices by activity sector using information from both sources.

As a first comparative assessment we compute the SSAI indices at all levels of aggregation using only survey data and using directly administrative values whenever possible for units $k \in s_{my}^{surv} \cap U_{my}^{adm}$. Figure 3 shows the difference between the national index I_{my}^{surv} using survey data and I_{my}^{adm} using admin data whenever possible. We observe that administrative information under- or over-estimates in an unpredictable manner relative to the survey-based index. Upon closer inspection of the selected use cases, while the values for July 2021 are nearly identical, the divergence for December 2021 is more pronounced.

When comparing by activity sector (Figure 4 shows 4 selected examples), the behaviour becomes even more chaotic. In fact, a comparison of Figures 2 and 4 reveals that the behaviour of the microdata by activity sector diverges significantly from that of the corresponding aggregates, discrepancy that stems from the influence of sampling weights, index weighting, and the whole estimation procedure.

The estimated quantiles of the turnover *at the population level* for both same pairs as before have been also computed and represented as quantile-quantile plots (see figures A6 to A9 in the appendix). The effect of sampling weights and index weighting are visible. Despite similarities at the micro level, at the aggregated level discrepancies arise visibly.

Figure 3: In red the national index difference $I_{my}^{adm} - I_{my}^{surv}$; in blue, the benchmark values I_{my}^{surv}.

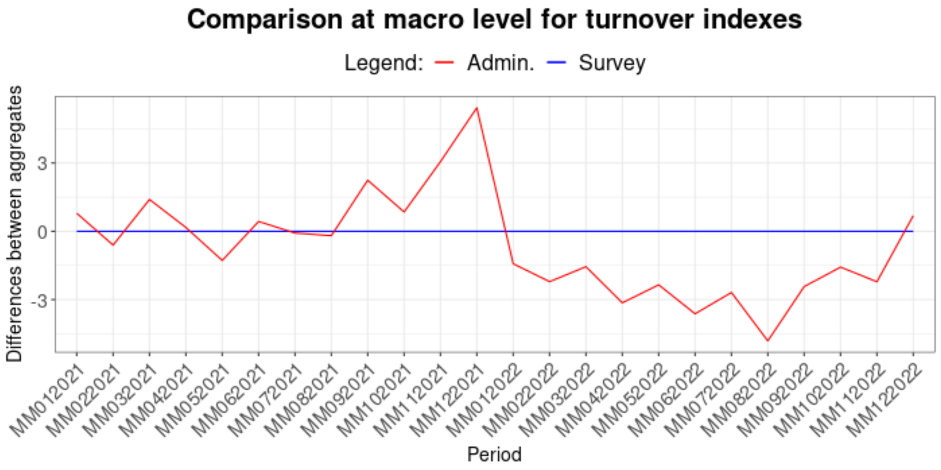
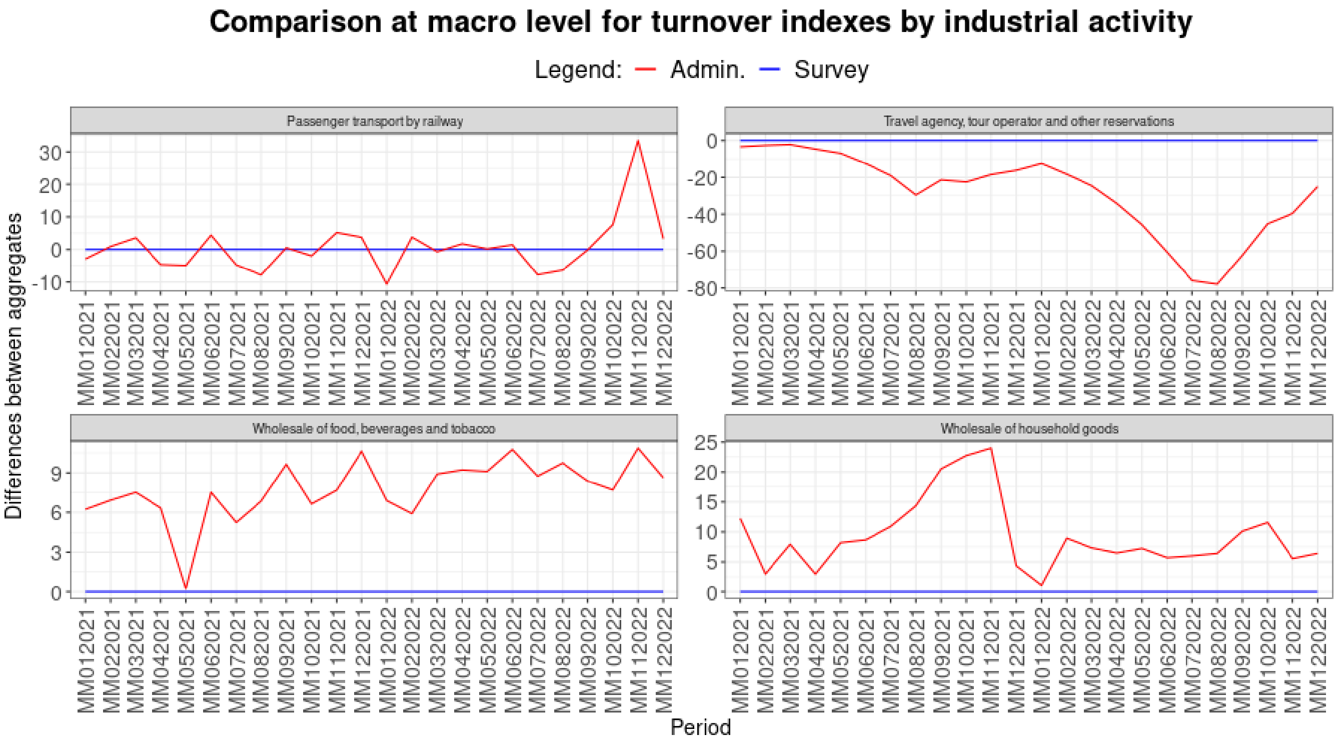


Figure 4: In red the national index difference $I_{my}^{adm} - I_{my}^{surv}$; in blue, the benchmark values I_{my}^{surv}, by sector.



3. Can we anticipate the performance of a data source with input quality indicators?

As anticipated in the previous section, observing Figure 3 for the months of July and December, 2021, we found that while the final aggregate result was almost identical for the former, the divergence was more pronounced for the latter. We find it legitimate to enquire whether a set of indicators, possibly graphically represented, exists to anticipate the quality a data source regarding the final target aggregated results.

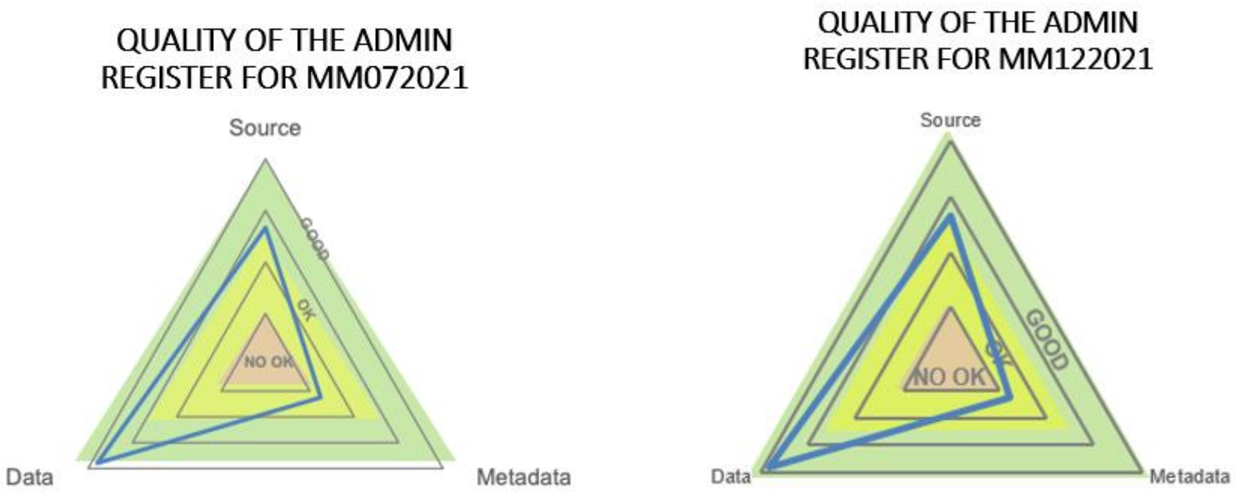
As a first immediate option we have focused on the quality framework proposed by Daas et al. [1, 2] to compute indicators to be compared with results presented above. In particular, we have concentrated on indicators for their proposed hyperdimensions: source, data, and metadata, mostly consisting of calculating the percentage of updated data from the analysed source (admin) of the previous time period in the current period. A visual summary is represented in figure 5.

This summary is further developed within each hyperdimension in figures A10 and A11 in the appendix. Our preliminary conclusion so far is that this graphical summary and thus the underlying set of indicators do not allow us to discern between favourable or unfavourable data quality conditions in the admin data source.

One would expect the summary diagram to show "NOT OK" for the data from December 2021 and a "GOOD" or “OK” for the data from July 2021, and yet the diagrams in figure 5 are practically indistinguishable.

This highlights the amount of work that still lies ahead to achieve a fast, accurate, efficient, and appropriate way to measure the quality of using an administrative source as input when replacing a sample.

Figure 5. Graphical summary of quality indicators for the administrative source according to the hyperdimensions by Daas et al (2009, 2011). July, 2021 (left); December, 2021 (right)



References

1. Piet Daas, Saskia Ossen, Rachel Vis-Visschers and Judit Arends-Tóth (2009) Statistics Netherlands. *Checklist for the Quality evaluation of Administrative Data Sources*.

2. Piet Daas, Saskia Ossen. BLUE-Enterprise and Trade Statistics (2011). *Report on methods preferred for the quality indicators of administrative data sources*.

Appendix

Figure A1: Monthly distributions of Kolmogorov-Smirnov distances for all economic sectors. Illustrative selected examples of July and December, 2021 are marked in blue.

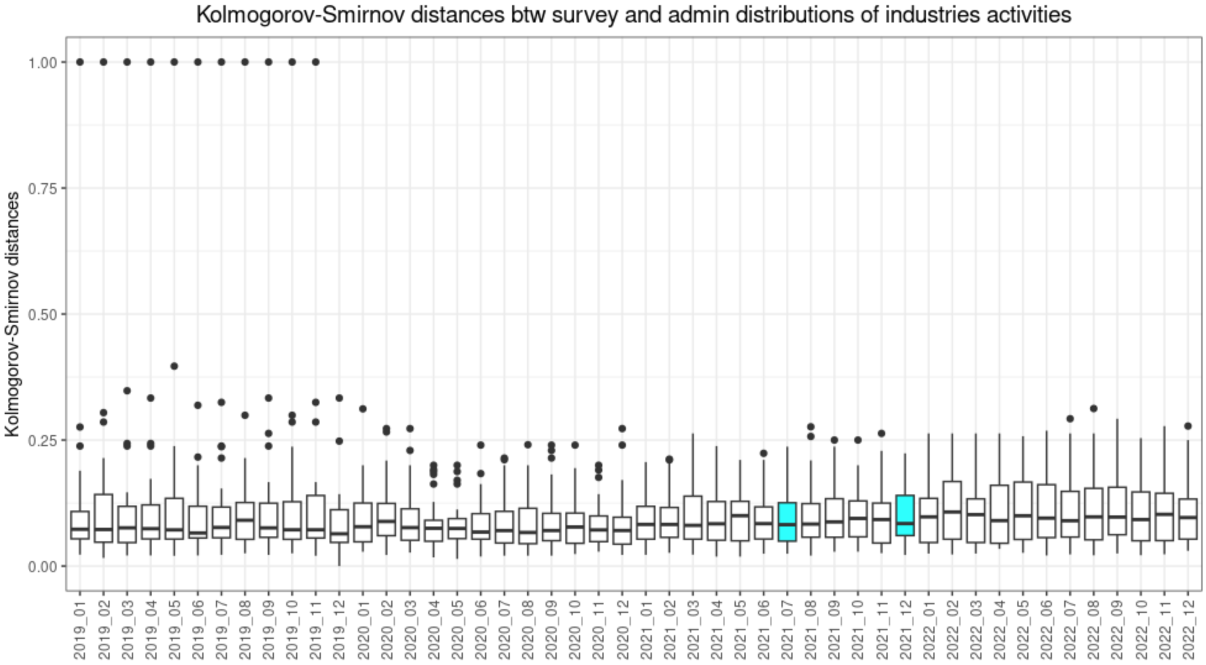


Figure A2. QQ plot for turnover values $\{z_{k,my}^{surv}\}_{k\in s_{my}^{surv}\cap U_{my}^{adm}}$ and $\{z_{k,my}^{adm}\}_{k\in s_{my}^{surv}\cap U_{my}^{adm}}$. July, 2021.

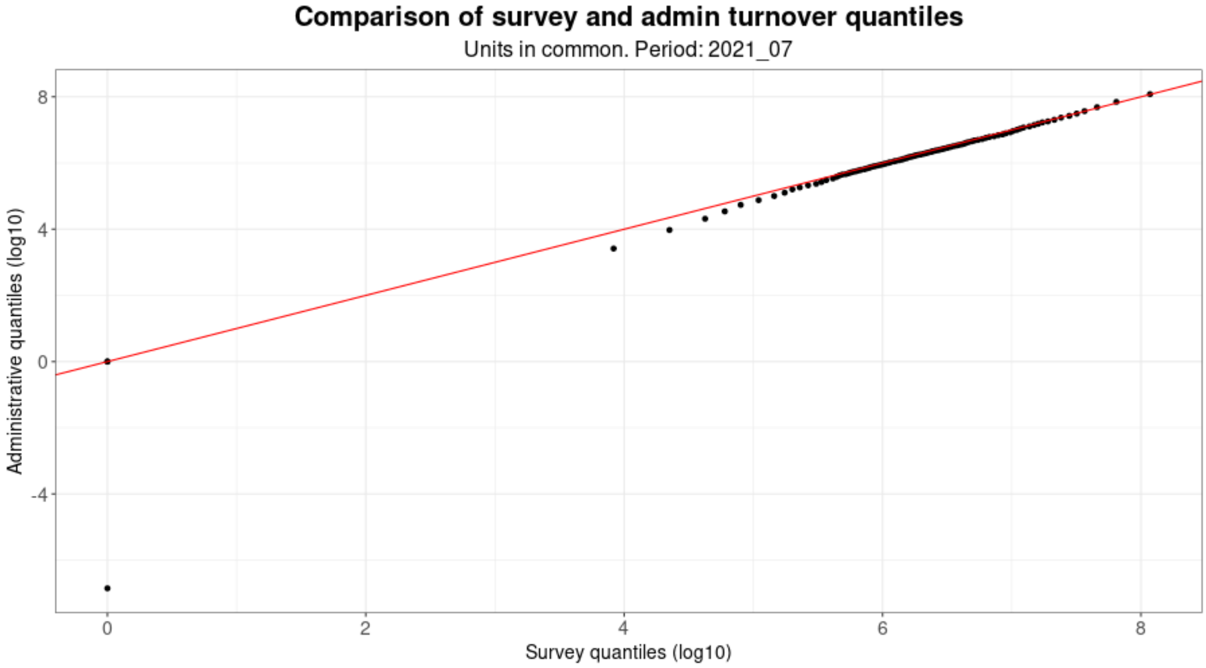


Figure A3. QQ plot for turnover values $\{z_{k,my}^{surv}\}_{k\in s_{my}^{surv}\cap U_{my}^{adm}}$ and $\{z_{k,my}^{adm}\}_{k\in s_{my}^{surv}\cap U_{my}^{adm}}$. December, 2021.

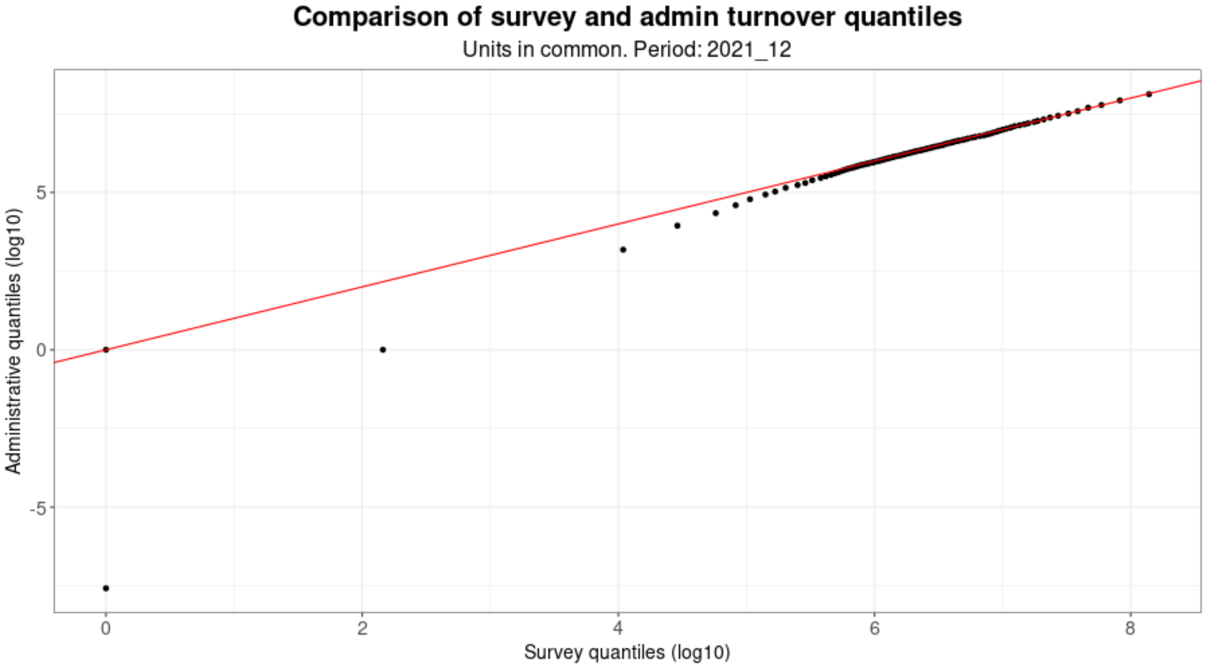


Figure A4. QQ plot for turnover values $\{z_{k,my}^{synth}\}_{k\in s_{my}^{surv}}$ and $\{z_{k,my}^{adm}\}_{k\in s_{my}^{surv}}$. July, 2021

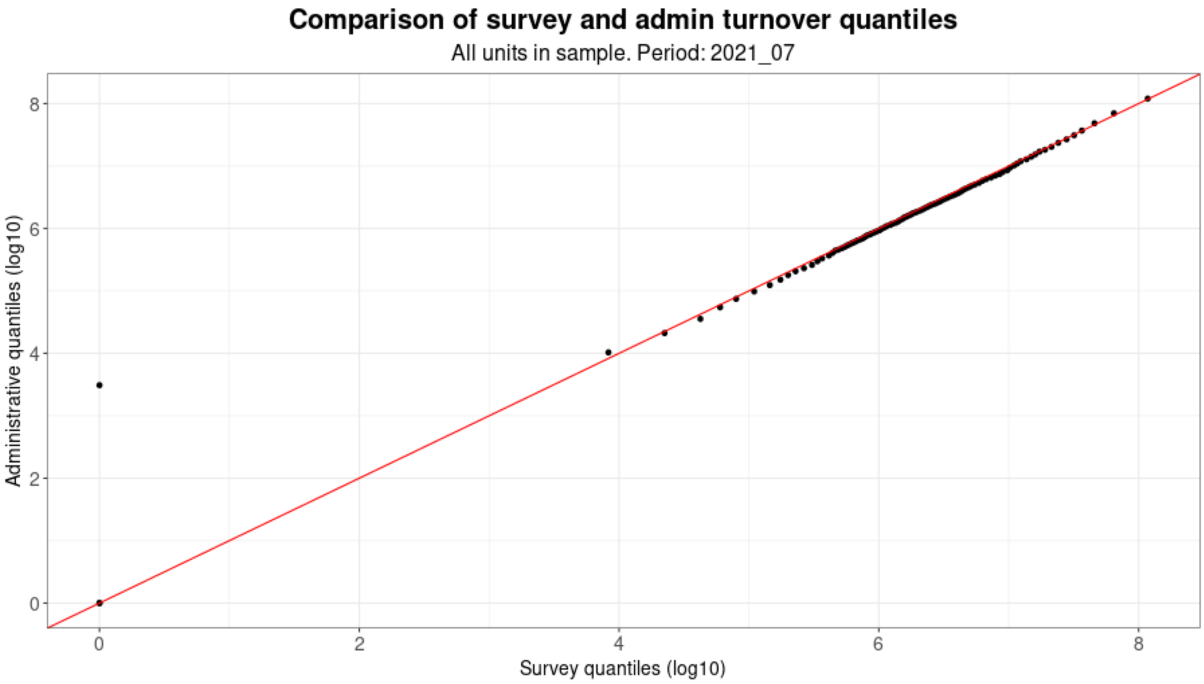


Figure A5. QQ plot for turnover values $\{z_{k,my}^{synth}\}_{k\in s_{my}^{surv}}$ and $\{z_{k,my}^{adm}\}_{k\in s_{my}^{surv}}$. December, 2021

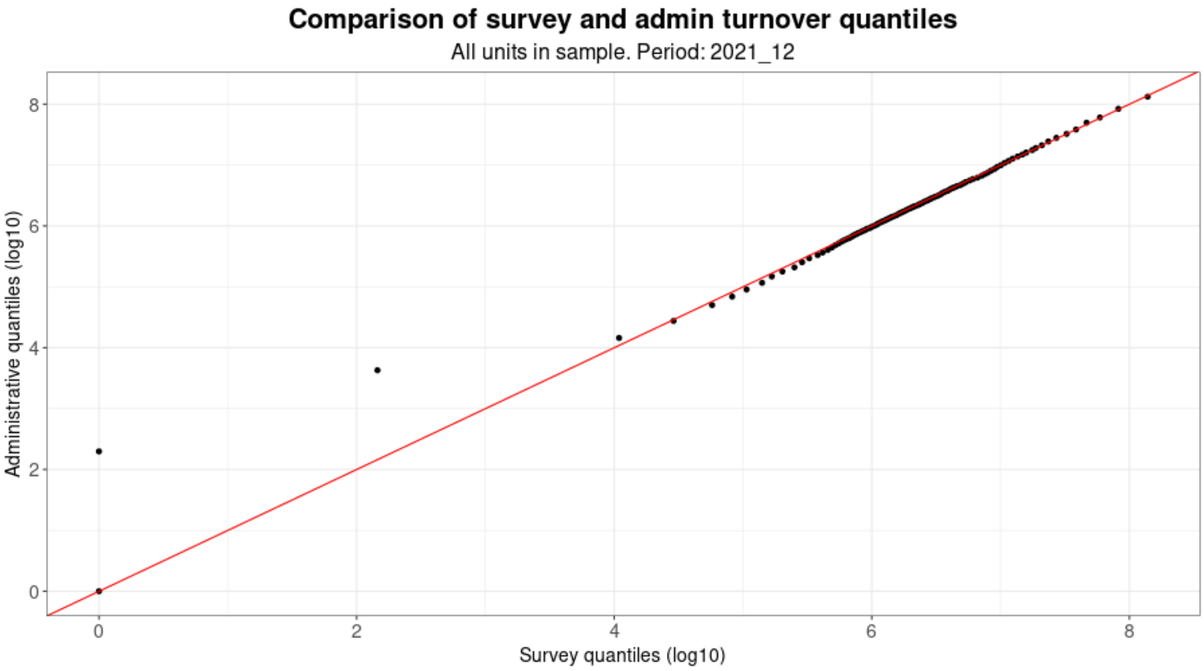


Figure A6. QQ plot for population-estimated turnover values from $\left\{z_{k,my}^{surv}\right\}_{k\in s_{my}^{surv}\cap U_{my}^{adm}}$ and $\left\{z_{k,my}^{adm}\right\}_{k\in s_{my}^{surv}\cap U_{my}^{adm}}$. July, 2021.

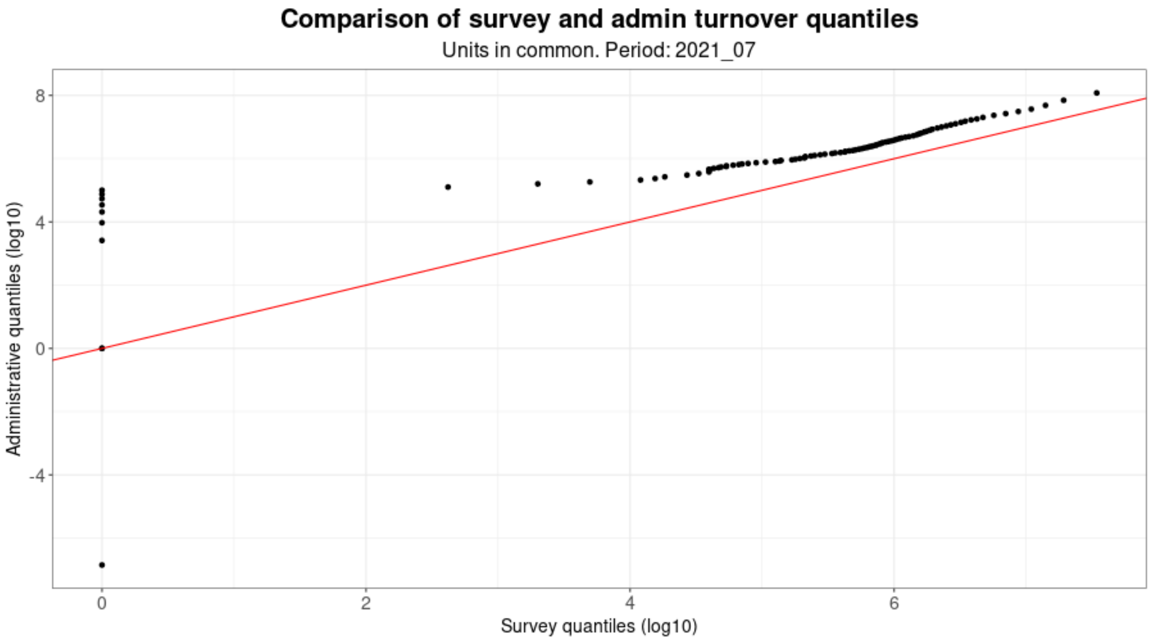


Figure A7: QQ plot for population-estimated turnover values from $\left\{z_{k,my}^{surv}\right\}_{k\in s_{my}^{surv}\cap U_{my}^{adm}}$ and $\left\{z_{k,my}^{adm}\right\}_{k\in s_{my}^{surv}\cap U_{my}^{adm}}$. December, 2021.

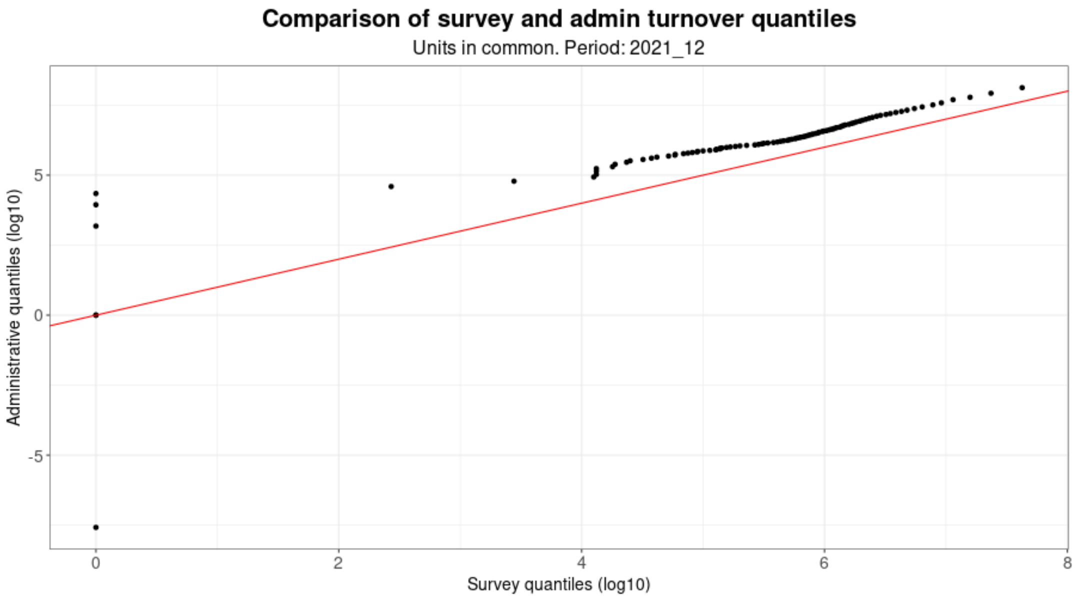


Figure A8. QQ plot for population-estimated turnover values from $\{z_{k,my}^{synth}\}_{k\in s_{my}^{surv}}$ and $\{z_{k,my}^{adm}\}_{k\in s_{my}^{surv}}$. July, 2021.

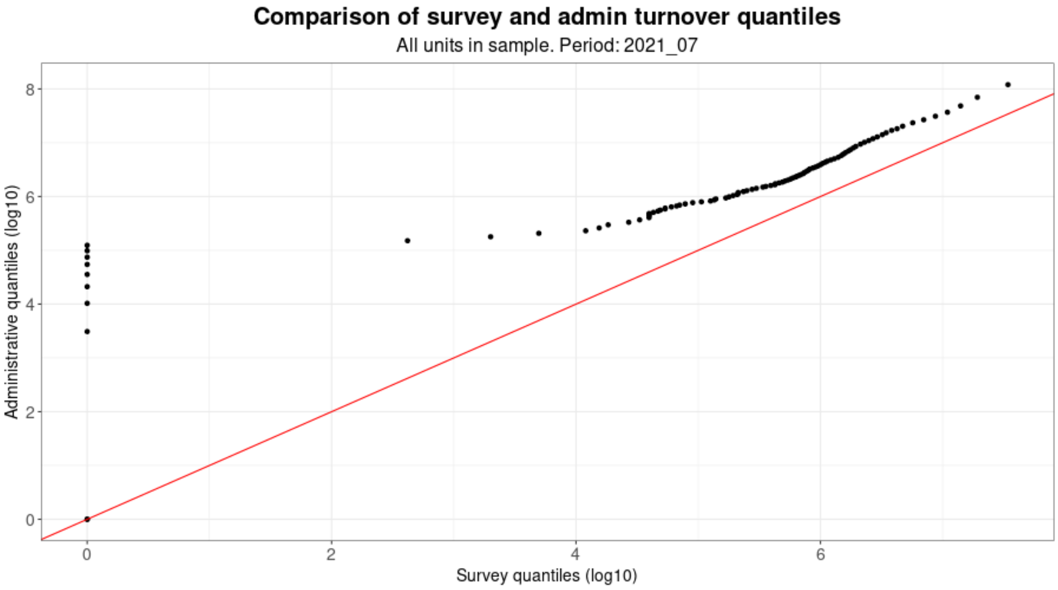


Figure A9: QQ plot for population-estimated turnover values from $\{z_{k,my}^{synth}\}_{k\in s_{my}^{surv}}$ and $\{z_{k,my}^{adm}\}_{k\in s_{my}^{surv}}$. December, 2021.

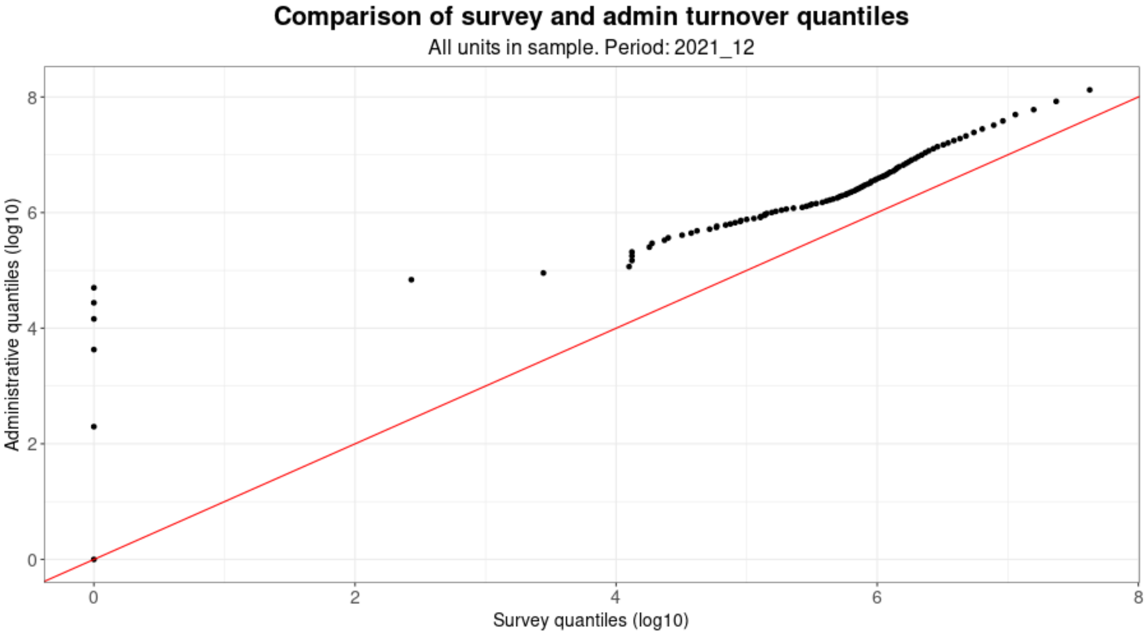


Figure A10. Graphical summaries of the preliminary proposal for quality indicators for administrative sources, indicating for each of the hyperdimensions - source, metadata, and data - whether the administrative source is a possible substitute for the survey (in this case, SSAI) or not, for the data of the time period **July 2021.**

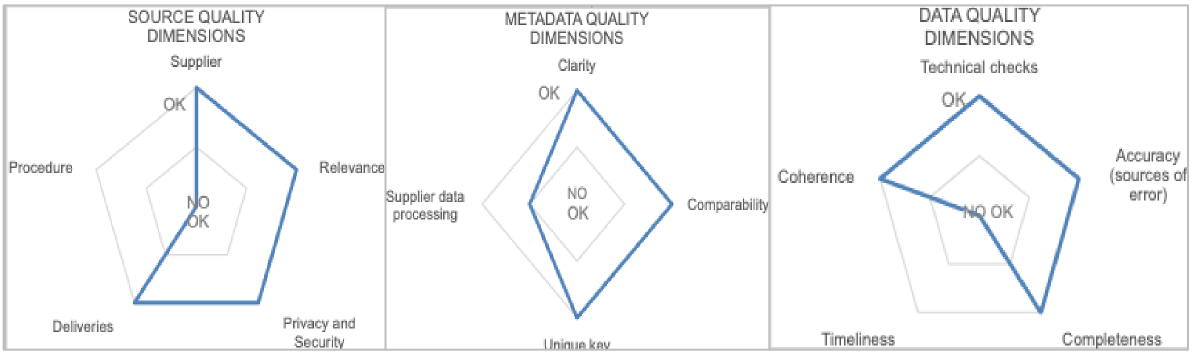
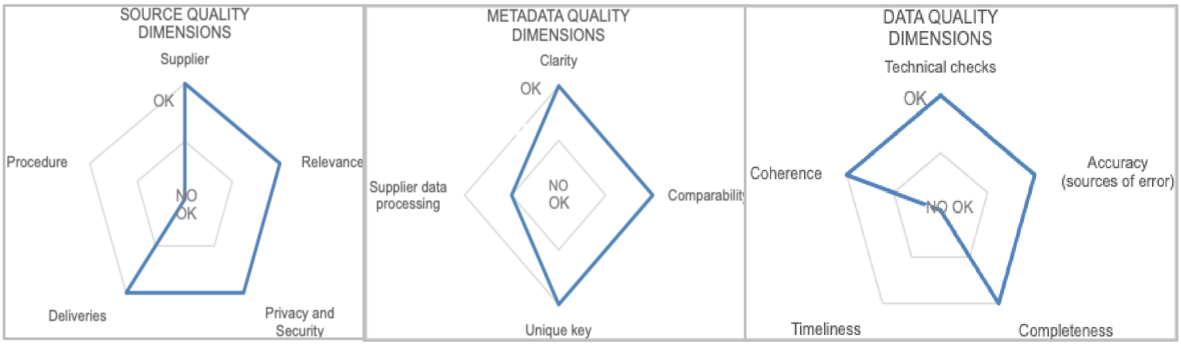


Figure A11. Graphical summaries of the preliminary proposal for quality indicators for administrative sources, indicating for each of the hyperdimensions - source, metadata, and data - whether the administrative source is a possible substitute for the survey (in this case, SSAI) or not, for the data of the time period **December 2021.**


SOURCE QUALITY DIMENSIONS
Supplier
OK
Relevance
Privacy and Security
Deliveries
Procedure
NO OK
METADATA QUALITY DIMENSIONS
Clarity
OK
Comparability
Unique key
Supplier data processing
NO OK
DATA QUALITY DIMENSIONS
Technical checks
OK
Accuracy (sources of error)
Completeness
Timeliness
Coherence
NO OK