Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Quality indicators for representation in administrative data

by Natalie Shlomo and Myong Sook Kim

Release date: December 23, 2025

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

• Statistical Information Service	1-800-263-1136
• National telecommunications device for the hearing impaired	1-800-363-7629
• Fax line	1-514-283-9350

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, the Agency has developed standards of service which its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

© His Majesty the King in Right of Canada, as represented by the Minister of Industry, 2025

Use of this publication is governed by the Statistics Canada Open Licence Agreement.

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, December 2025
Vol. 51, No. 2, pp. 509-532
Statistics Canada, Catalogue No. 12-001-X

Quality indicators for representation in administrative data

Natalie Shlomo and Myong Sook Kim[1]

Abstract

National Statistical Institutes (NSIs) are directing resources into advancing the use of administrative data in official statistics. Administrative data, however, are not developed for the purpose of producing statistics rather as a result of an event or transaction relating to administrative procedures of organizations, public administrations and government agencies. Therefore, it is essential to check the quality of the administrative data with respect to sources of error, particularly representativeness to the target population. In this paper, we utilize the strength of probability-based reference samples or censuses that can be used to detect the lack of representativeness in administrative data and introduce quality indicators based on distance metrics and representativity indicators (R-indicators). We demonstrate their application with a simulation study and discuss a real application applied on a UK Office for National Statistics (ONS) administrative dataset.

Key Words: Distance metrics; Propensity scores; R-indicators.

1. Introduction

National Statistical Institutes (NSIs) are directing resources into advancing the use of administrative data in their national statistical systems. This is a top priority for the UK Office for National Statistics (ONS) as they are undergoing transformations in their statistical systems to make more use of administrative data for future censuses and population statistics. Administrative data are defined as secondary data sources since they are produced by other agencies as a result of a transaction or event relating to administrative procedures of organizations, public administrations and government agencies. On the other hand, administrative data are enhanced by the increasing use of information and communication technology and are becoming more readily available in electronic form. Therefore, they have the potential to become increasingly important data sources for the production of official statistics by significantly reducing the cost and burden of response and improving the efficiency of statistical production systems.

Ingesting administrative data into statistical systems at NSIs is not without costs and it is vital to understand where potential errors can arise. The NSIs have invested in developing quality frameworks for ingesting external data sources, such as administrative data, into their statistical systems. Many of these frameworks are largely qualitative in nature. One of the first quality frameworks for administrative data in the European context was developed by Daas, Ossen, Vis-Visschers and Arends-Toth (2009). The framework provides a checklist for the quality evaluation of administrative data sources based on “hyper-dimensions” according to the source, metadata, and data. Each hyper-dimension may have several dimensions and a quality indicator can be measured using either qualitative or quantitative methods. For example, the quality dimension of item missing data in a variable can be expressed as the proportion of missing data. Since then, many NSIs have developed similar quality frameworks for administrative data. An example is the quality assurance of administrative data based on the patient register data at the UK

1. Natalie Shlomo, Social Statistics Department, School of Social Sciences, University of Manchester, Oxford Road, Manchester M13 9PL, United Kingdom. E-mail: natalie.shlomo@manchester.ac.uk; Myong Sook Kim, Social Statistics Department, School of Social Sciences, University of Manchester, Oxford Road, Manchester M13 9PL, United Kingdom. E-mail: myongsook.kim@postgrad.manchester.ac.uk.

National Health Service. This data has been explored by the ONS for producing population statistics. The authors propose steps that can be taken to assess the quality of this data through processing and validation checks, as well as propose actions to improve the data ingestion (Office for National Statistics (ONS), 2016). More recently, a Task Force of European Statisticians assessing the quality of administrative sources for use in censuses published a set of guidelines that were endorsed by the 69th plenary session of the Conference of European Statistics in 2021 (United Nations Economic Commission for Europe (UNECE), 2021). The guidelines set out four hyper-dimensions: source stage, data stage, process stage and output stage with quality dimensions listed under each hyper-dimension that can be assessed qualitatively or quantitatively. Following on from this task force, Dunstan and Correia (2021) developed a bespoke quality framework for assuring the quality of administrative data for use in the 2021 Census of England and Wales. Here the quality framework is set out in two hyper-dimensions: source stage and data stage. The advantage of this framework is that all dimensions defined in each of the hyper-dimensions are assessed using a Likert-type scale, for example (1) "does not meet" (2) "partially meets" (3) "meets". Thus the scores can be aggregated to obtain overall quantitative quality scores for each hyper-dimension (Spector, 1992). Many NSIs are devoting resources to developing quality frameworks for ingesting administrative data following international principles, such as those set out in the United Nations (UN) (2019). The commonality of these quality frameworks is that they include the basic elements of definitions of the principles of quality: Relevance, Accuracy and Reliability, Timeliness and Punctuality, Coherence and Comparability, Accessibility (Eurostat, 2019).

On a conceptual basis, the life cycle of integrated statistical microdata from a quality perspective was shown in Zhang (2012) where possible sources of error are identified when using administrative data, depending on whether it is a single data source or integrated with other data sources such as survey data. For a single source administrative data, one of the main sources of error is representation to the target population of interest and it is this quality dimension that is the focus of this research.

We propose quality indicators that allow NSIs to assess quantitatively if the administrative data is representative to the target population and to identify sub-groups that may not be covered in the administrative data. Moreover, administrative data often comes in streams and the NSI needs to know when the administrative data is sufficiently complete. Therefore, there is a need to evaluate administrative data on a continuous basis to decide when the data is complete for production.

To assess the quality of administrative data, NSIs should take advantage of timely census data or high-quality probability-based (weighted) surveys to compare distributions between this source of auxiliary data and the administrative data. Since administrative data is ingested on a more frequent basis compared to a census that is conducted every 5 or 10 years, NSIs should be utilizing their survey data collections to assess the quality of administrative data. For example, large data collections, such as the Labour Force Survey, can be used to assess the quality of administrative data with respect to representativeness. Whilst it is known that these surveys might suffer from high non-response, they are generally of good quality with respect to the steps taken to mitigate nonresponse bias during data collection and post-survey adjustments that compensate for nonresponse. Moreover, NSIs typically carry out post-stratification and calibrate the survey design weights adjusted for nonresponse to known population benchmarks, usually by sex, age and geographic regions. These population benchmarks are typically derived annually by NSIs using principles of demographic accounting between census years and are treated as “true” values.

Another example of high-quality survey data that can serve the purpose of assessing the representativeness of administrative data is a survey data collection that is collected for the sole purpose of correcting for coverage. In many countries, these coverage surveys are mandatory (similar to a census), hence should be of high-quality and can be used for a quality assessment of administrative data. Note that coverage surveys are typically used for dual-system estimation to account for under-coverage in a census, for example, by enabling the estimation of coverage weights to the census. Therefore, these types of coverage surveys provide good estimates of distributions in the population for evaluating the quality of administrative data. In this paper, we assume that we have access to a high-quality probability-based reference sample to use for the assessment of representativeness in administrative data.

We use univariate and bivariate distributions obtained from a census or a weighted probability-based reference survey and compare them to distributions in the administrative data on a common set of variables. We can use distance metrics to assess the difference between distributions. In addition, we propose a Representativity (R-) Indicator that is designed for quantifying the representativeness of population groups in administrative data by estimating a propensity score and studying its variation. We view the propensity score as a balancing score whereby conditional on the propensity score, the distribution of observed baseline covariates should be similar between the administrative data and the weighted survey data or census (Austin, 2011).

The proposed approach uses aggregate-level auxiliary information in the form of variable and category-level frequency tables as the foundation for the estimation of the propensity scores based on a linear regression function. Note that this approach is similar to the weight calibration approach and other linear-based survey quality indicators developed, for example, in Särndal and Lundström (2008). This approach avoids the need for individual record-level data to estimate propensity scores as shown, for example, in Chen, Li and Wu (2020). Also, as is common in the estimation of coverage probabilities to correct for under- and over- coverage in a Census (Alho, 1990; Office for National Statistics (ONS), 2022), record-level linkage between the administrative data and the auxiliary data source is not needed as we do not aim to estimate an individual coverage probability. Here, we aim to propose quality indicators to identify representativeness on a group-level rather than on an individual-level, and therefore we have the advantage that the application is faster and simpler computationally and prevents the introduction of possible linkage error. A disadvantage of this approach is the need to use a linear regression model to estimate the propensity scores rather than a logistic or probit regression model, leading to estimated propensity scores that could be outside of the [0, 1] interval. However, given that the main aim is to investigate the variation of the propensity scores to derive quality indicators, we proceed with this practical and easy-to-implement approach.

We show how these quantitative quality measures can be used to assess representativeness in administrative data and can provide a more direct assessment of quality rather than the previous use of check lists and score cards that are largely qualitative. Furthermore, estimating the propensity score in the administrative data allows for adjusting estimates derived from the administrative data by weighting according to the inverse of the propensity score, similar to inverse propensity weighting for survey samples (IPW) discussed in Kim and Haziza (2014) and more recently in Chen et al. (2020) and references therein for nonprobability surveys. To use the IPW approach for non-probability samples, Chen et al. (2020) list the assumptions needed to produce unbiased estimates: (1) The inclusion indictor of the administrative data is independent of any target variable given a set of covariates $\mathbf{X}$; (2) All units have a non-zero propensity score to be included in the administrative data;(3) Inclusion indicators for units i and j are independent given x_i and x_j for $i \neq j$. For the case of administrative data, NSIs will harmonize and combine multiple administrative data sources with the aim of ensuring coverage to a target population (see the application in Section 4). Those missing in the administrative data, or have erroneous or multiple addresses, are generally those that may not engage in government transactions. For example, young people and students may be listed in alternative addresses. We nevertheless make the assumption that any systematic deviations from the target population can be explained by covariates $\mathbf{X}$ in the data and the assumptions of Chen et al. (2020) for the case of a non-probability sample will also hold in the setting of administrative data.

Section 2 describes the proposed quality indicators to assess representativeness of the administrative data. Section 3 demonstrates the proposed quality indicators on (simulated) administrative data including an interpretation of the findings. Section 4 discusses an application using a real administrative dataset implemented within the secure servers at the ONS. We conclude in Section 5 with a discussion.

2. Quality indicators

In many quality assessment frameworks for administrative data across NSIs, there is little use of external auxiliary census data or high-quality survey data collected by the agency to quantify representativeness in administrative data.

Administrative data typically have the following errors: individuals listed in the wrong address; missing records because individuals may not interact with the agency; duplicate records where individuals may be listed in different addresses. Furthermore, there can be erroneous records, for example, deaths and emigrants who have left the country. NSIs use best practices to deduplicate and clean the administrative data prior to their use in statistical systems. They can remove deaths and recorded emigrants from the administrative data, although unrecorded emigrants are hard to identify and remove. In this work, we assume that the number of records in the administrative data is M and the number of (weighted) records in the reference data is N, and it is possible for M to be greater than N.

In this section, we show how this type of auxiliary data can be used to develop quality indicators for administrative data. The R-code software (R Core Team, 2022) for the quality indicators and a user manual is available on GitHub (Kim and Shlomo, 2023).

2.1 Distance metrics

In this section, we compare univariate and bivariate distributions obtained from the administrative data to external population auxiliary aggregated data, either obtained directly from a census or estimated from a high-quality probability-based random survey.

Denote variable k having categories $h, h = 1, 2 \ldots H$ in the administrative dataset having M individuals. Let $\Delta_{h,i}^{k}$ be the $0-1$ indicator for individual i being a member of category h in variable k. We can then calculate the counts for category $h: m_h^k = \sum_{i=1}^{M} \Delta_{h,i}^{k}$ and note that $\sum_{h=1}^{H} m_h^k = M$. We can also obtain the probability distribution of variable k having category $h, h = 1, 2 \ldots H$ and calculate: $p_h^k = \frac{m_h^k}{M}$. The variable k can also represent a cross-tabulation of two or more variables, for example age group $\times$ sex.

Now assume we have equivalent estimates of these distributions from a census, or alternatively from a large probability-based random sample of size n where every individual i in the sample has an associated survey weight w_i. The survey weights typically are adjusted for nonresponse and calibrated to known population benchmarks and hence sum to the known population size N. In this case, $n_h^k = \sum_{i=1}^{n} w_i \Delta_{h,i}^{k}$ and $\sum_{h=1}^{H} n_h^k = N$ and the equivalent probability distribution is $q_h^k = \frac{n_h^k}{N}$.

We can now define a variety of distance metrics to assess deviations between the distributions $\{p_h^k, h = 1, \ldots, H\}$ and $\{q_h^k, h = 1, \ldots, H\}$. In this research, we show two distance metrics for variable k (dropping the indicator for variable k): the Indicator of Dissimilarity (ID) (Duncan and Duncan, 1955):

$$\mathrm{ID} = \frac{1}{2} \sum_h |\, p_h - q_h \,|$$

and the Hellinger's Distance (HD):

$$\mathrm{HD} = \frac{1}{\sqrt{2}} \sqrt{\sum_h \left(\sqrt{p_h} - \sqrt{q_h}\right)^2}.$$

There are subtle differences between distance metrics. For example, Hellinger's Distance places more weight on the smaller proportions compared to the larger proportions whereas the Indicator of Dissimilarity treats all proportions equally. We also looked at the Kullback-Leibler (KL) divergence, but this is related to the Hellinger's Distance and is not a "true" distance metric (does not satisfy the triangle inequality). We have included the calculation of distance metrics in the R-code on GitHub (Kim and Shlomo, 2023). This will facilitate more empirical work for future recommendations.

The two distance metrics, HD and ID, are bounded by 0 for a perfect agreement in the distribution and 1 for a perfect disagreement. A perfect disagreement occurs when both distributions have 100% of their mass in different cells of the distribution. For this research, we define 1 minus the distance metric, denoted $(1-\mathrm{HD})$ and $(1-\mathrm{ID})$, so that 1 represents a perfect agreement and 0 a perfect disagreement.

For small $(1-\mathrm{HD})$ or $(1-\mathrm{ID})$ we can determine which category of the variable is causing the low distance metric and attempt to rectify the lack of representativeness by identifying further data sources to supplement the administrative data. Note that these distance measures can quantify differences in univariate and bivariate statistics but to look at a more detailed multivariate assessment of variables we turn now to the use of R-indicators in Section 2.2.

2.2 R-indicators

The R-indicator and its related partial R-indicators were originally designed to assess the representativeness of responses from a survey and are particularly useful as an objective function in the optimization of adaptive survey designs (Schouten, Cobben and Bethlehem, 2009; Schouten and Shlomo, 2017). The R-indicators measure the contrast between those who are missing and not missing in the data and identify those groups that are not represented in the collected data. In the survey-based setting, response propensities are estimated using a logistic regression model in the survey microdata based on covariates available from sample frame, auxiliary data and paradata. Under the logistic regression model (Binomial Distribution), the maximum standard deviation of the estimated response propensities is 0.5 and hence in the survey-based literature, the R-indicator is defined as $\hat{R}_{\hat{\rho}} = 1 - 2 \times \text{SD}_{\hat{\rho}}$ where $\text{SD}_{\hat{\rho}}$ is the standard deviation of the estimated response propensities. This allows for an R-indicator bounded between (close to) 0 and 1. See Schouten et al. (2009) for more details.

Ouwehand and Schouten (2014) developed R-indicators to assess Short Term statistics from business surveys and their representativeness to the administrative data source of the Business VAT registry in the Netherlands. Here the administrative data was used as the auxiliary data source. The authors used the standard approach of estimating R-indicators based on estimating propensity scores according to the sample-based literature mentioned above.

In this work, we develop the R-indicator and partial R-indicators to assess the representativeness of an administrative dataset compared to a target population and use a different approach. Recent research by Bianchi, Shlomo, Schouten, Da Silva and Skinner (2019) adapts the R-indicator to the case where we wish to assess the representativeness of survey data but only population-based aggregate auxiliary information are available instead of sample-based frame information. This approach was successfully applied in Shlomo, Luiten and Schouten (2022) where the authors assessed the representativeness of the EU Statistics on Income and Living Conditions Surveys (EU-SILC) across multiple countries in the European Union. We draw upon this research and utilize population-based aggregate auxiliary information obtained from a timely census or from weighted sample counts calculated from a large high-quality probability-based random sample.

To calculate the population-based R-indicator, denote the inclusion indicator r_i equal to 1 for all units in the administrative dataset (denoted by "A"). We have information available on the values $\mathbf{x}_i = (x_{1,i}, x_{2,i}, \ldots, x_{K,i})^T$ of a vector of K auxiliary variables $\mathbf{X}$, for example, sex, age group, geographical region, ethnic minority group and employment status. Therefore, each $x_{k,i}$ is a binary indicator variable. We also assume that values of $\mathbf{x}_i$ are observed for all individuals in the administrative dataset so that $\{x_i; i \in A\}$ is observed.

Assume we know $\mathbf{x}_i$ at the aggregate level based on an external dataset, either from a census or a probability-based random sample. To calculate the estimates of the propensity scores, we need the population total $\sum_U \mathbf{x}_i$ and population cross-products $\sum_U \mathbf{x}_i \mathbf{x}_i^T$ where U represents the population. This information is known as the population-based auxiliary information. Assuming that we have available a probability- based random sample, denoted s, where each individual i in the sample has a survey weight w_i that is adjusted and calibrated to the population size N, we can consider two types of calculations to estimate the population cross-products $\sum_U \mathbf{x}_i \mathbf{x}_i^T$ (Bianchi et al., 2019):

Type 1: If the probability-based random sample is large, or if we have access to timely census data, we can estimate the population-based auxiliary information by: $\sum_U \mathbf{x}_i \approx \sum_s w_i \mathbf{x}_i$ and $\sum_U \mathbf{x}_i \mathbf{x}_i^T \approx \sum_s w_i \mathbf{x}_i \mathbf{x}_i^T$.

Type 2: If the probability-based random sample is small and estimating the cross-products from the sample is not feasible, we can rely on the marginal information as follows:

Step 1: Calculate an estimate for the mean in the population from the sample: $\bar{X}_U = \frac{\sum_U \mathbf{x}_i}{N} \approx \frac{\sum_s w_i x_i}{N}$.

Step 2: Estimate the cross-products in the population by: $N\hat{S}_{XX} + N\bar{X}_U \bar{X}_U^T$ where $\hat{S}_{XX} = \left(\sum_{i=1}^M d_i\right)^{-1} \sum_{i=1}^M d_i (\mathbf{x}_i - \bar{X}_U)(\mathbf{x}_i - \bar{X}_U)^T$ is estimated from the administrative data A of size M, and d_i is an adjustment weight for the administrative data, referred as the representativeness weight below, such that $\sum_{i=1}^M d_i = N$.

We define the propensity score for a unit in the administrative dataset as the conditional expectation of the indicator variable r_i given the values of specified covariates. We denote $\rho_X(\mathbf{x}_i)$ by ρ_i.

In the population-based setting, we model the propensity scores under an identity (linear) link function where the true propensity scores satisfy: $\rho_i = \mathbf{x}_i^T \boldsymbol{\beta},\ i \in A.$ Assuming that the administrative data has a probability selection mechanism (albeit with unknown probabilities) we can write the conditional expectation of the indicator r_i given the auxiliary $\mathbf{x}_i$ as: $E(r_i \mid \mathbf{x}_i) = \rho_X(\mathbf{x}_i) \equiv \rho_i$. Using the linear link function allows the use of "plug-in" estimates directly into the formula for estimating propensity scores. For the linear probability model, the estimate of ρ_i in the administrative data A is given by: $\hat{\rho}_i^{\text{OLS}} = \mathbf{x}_i^T \hat{\boldsymbol{\beta}} = \mathbf{x}_i^T \left(\sum_A d_i \mathbf{x}_i \mathbf{x}_i^T\right)^{-1} \sum_A d_i \mathbf{x}_i,\ i \in A,$ where d_i is the representativeness weight for unit i.

We use the "plug-in" estimators to approximate $\hat{\rho}_i^{\text{OLS}}$ by $\hat{\rho}_i^P = \mathbf{x}_\mathrm{i}^\mathrm{T} \left(\sum_U \mathbf{x}_i \mathbf{x}_i^T\right)^{-1} \sum_A d_i \mathbf{x}_i,\ i \in A.$ Note that $\hat{\rho}_i^P$ is computed only on the set of individuals in the administrative data A.

The estimation of $\hat{\rho}_i^{\text{OLS}}$ takes a similar approach to the calibration-based method for estimating propensity scores for a non-probability sample when only population totals are known by solving the estimating equations $\sum_{\text{NP}} \frac{x_i}{\pi(x_i, \theta)} - \sum_U x_i = 0$ as shown in Chen, Li, Rao and Wu (2022). The authors point out that the calibration-based estimators and the maximum pseudo-likelihood estimators for the parameters of the model as shown in Chen et al. (2020), are not equivalent in a mathematical sense, but both estimators are consistent assuming the parametric form for estimating propensity scores.

In our setting of assessing the representativeness in administrative data, we will use a single representative weight: $d_i \equiv d = [M/N]^{-1}$ and equal for all individuals $i \in A$ (recall that M is the size of the administrative data and N is the size in the population). By applying d to all records in the administrative data, we ensure equal overall totals. We do not calculate more detailed strata-level representative weights as the aim is to identify those variables and their categories that are contributing to the lack of representativeness.

To obtain the R-indicator, we need to estimate the population-level variance of the propensity scores: $S_\rho^2 = \frac{1}{N-1}\sum_U (\rho_i - \overline{\rho})^2$. In the population-based setting, based on the available administrative data, we can estimate this variance of the propensity scores as follows: $\hat{S}_{\hat{\rho}^P}^2 = \frac{1}{N-1}\sum_A d_i(\hat{\rho}_i^P)^{-1}(\hat{\rho}_i^P - \hat{\overline{\rho}})^2$ where $\hat{\rho}_i^P$ is estimated as above and $\hat{\overline{\rho}}$ is the overall average propensity score. Note that $(\hat{\rho}_i^P)^{-1}$ adjusts for any bias related to lack of representation in the administrative data. As per Bianchi et al. (2019) we estimate the variance of the estimated propensity scores by: $\hat{S}_{\hat{\rho}^P}^2 = \frac{N}{N-1}\left\{\frac{1}{N}\sum_A d_i\hat{\rho}_i^P - \left(\frac{1}{N}\sum_A d_i\right)^2\right\}$. This format allows implementing a sample size bias correction, however this bias correction is not necessary for assessing representativeness of large administrative datasets.

If there is no variation in the propensity scores, the variance would be (close) to 0 and the administrative data are representative to the target population. The larger the variation, the less representative is the administrative data on the set of auxiliary variables.

Using the identity link function results in estimated propensity scores that can be larger than the value of 1, particularly because the representativeness weight d is close to 1 and propensity scores are likely to be high. However, since we are only interested in the predicted values to assess the variation of the propensity scores rather than inference and interpretation of odds ratios and effect sizes, we continue to use the identity link function to develop administrative-based quality indicators and accept that there will be propensity scores greater than 1.

In the survey-based literature and under a logistic regression function, the estimated R-indicator is defined as $\hat{R}_{\hat{\rho}^P} = 1 - 2\hat{S}_{\hat{\rho}^P}$ (Schouten et al., 2009) given that the maximum standard deviation of the propensity scores under the logistic regression is 0.5. However, in the administrative data setting, we define the estimator for the R-indicator as: $\hat{R}_{\hat{\rho}^P} = 1 - \hat{S}_{\hat{\rho}^P}$. This is because we base our predictions on the Normal Distribution to estimate propensity scores for the administrative data and we expect the average propensity score to be high implying smaller standard deviations compared to using the logistic regression modelled under the Binomial Distribution. If propensity scores are completely random, we can expect a maximum standard deviation of 1 under the Normal Distribution and the R-indicator (close to) 0. If propensity scores are all equal, the standard deviation under the Normal Distribution is 0 and the R-indicator is (close to) 1. We note that the R-indicator is a function of the auxiliary variables in the regression model. If we had a different set of auxiliary variables, we would obtain a different R-indicator. Moreover, the aim of the R-indicator is to assess representativeness in the form of "deviations" from a target population and not for the purpose of inference. We currently do not implement diagnostics on the fit of the propensity score model, but this could be a topic of future work.

The unconditional partial R-indicators are also mentioned in Bianchi et al., 2019. These indicators measure the amount of variation of the propensity scores between the categories of a variable. The larger the between-category variation is, the stronger the impact of the variable on a lack of representativeness. As earlier, let x_k be one of the components of the vector $\mathbf{X}$. The variable x_k is categorical with H categories (we drop the index k). Let m_h denote the weighted size in the administrative data in category h for $h = 1, 2, \ldots, H$. That means $m_h = \sum_{i \in A} d \Delta_{h,i}$ where $\Delta_{h,i}$ is the $0-1$ indicator for participating unit i being a member of category h and $\sum_{h=1}^{H} m_h = N$ given the definition of d as the representativeness weight. Define $\hat{\bar{\rho}}_h$ the average propensity score in category h of x_k for the units in the administrative dataset and $\hat{\bar{\rho}}$ the overall average propensity score based on the estimated population-based propensity scores $\hat{\rho}_i^P$. The estimate for the unconditional partial R-indicator for variable x_k is: $R_U(x_k) = \sqrt{\frac{1}{N} \sum_{h=1}^{H} m_h (\hat{\bar{\rho}}_h - \hat{\bar{\rho}})^2}$. The larger the value of the partial R-indicator, the stronger the association of the variable with a lack of representativeness in the administrative dataset. By computing and comparing the unconditional partial indicators for a set of variables, it can be established for which variables the relationships are strongest in contributing to the lack of representativeness. The unconditional partial R-indicator at the category level h for variable x_k is $R_U(x_k^h) = \sqrt{\frac{m_h}{N}} \, (\hat{\bar{\rho}}_h - \hat{\bar{\rho}})$ and can assume positive and negative values. Note that at the category-level, a negative sign represents under-representation and a plus sign represents over-representation.

Finally, we note that when producing estimates from the administrative dataset, one should weight each individual i by its inverse propensity score $(\hat{\rho}_i^P)^{-1}$ to adjust for bias arising from lack of representation in the estimates. The weighted distributions from the administrative data will be close to the benchmarked population distributions, particularly for those variables that are included in the model for estimating the propensity scores. This will be shown in Section 3.

2.3 Variance estimation of the R-indicators

There are two sources of variation for the proposed quality indicators and IPW estimates calculated from the administrative data: (1) due to the large size of administrative datasets representing a target population, they will generally have small variances for estimates derived from the administrative data, although there will be additional sources of variation if the administrative data have undergone data processing procedures such as coverage adjustments and imputation; (2) if the population benchmarks that are used to estimate the propensity scores are based on weighted estimates calculated from a high-quality random reference survey, we can expect additional variation due to the estimation of these benchmarks.

When univariate and bivariate counts are available from a census or known population totals and these are used as "plug-ins" for estimating propensity scores, we need to focus on the source of variation in (1) above. For this case, to estimate the variance of the R-indicators, distance metrics as well as IPW estimates derived from the administrative data, we can use a standard resampling approach, such as the Bootstrap:

- Sample with replacement from the administrative data;

- Implement any data processing features that were carried out on the original data;
- Estimate the propensity scores for B replicates according to the known population totals;
- Calculate the estimator of interest and derive the estimate for the variance by multiplying the sum of squares across the replicates by 1/B.

We can also include in the administrative data a set of bootstrap replication weights to facilitate variance estimation for users of the data.

As mentioned in (2) above, the uncertainty in estimates of indicators and IPW estimates may increase when we use weighted sample counts from a reference sample as the benchmarks for the estimation of the propensity scores. In that case, the variance is $O(1/n)$ instead of $O(1/N)$ where n is the sample size of the survey. Wu (2022) provides analytical expressions of the variance for IPW estimates related to nonprobability sampling which can be relevant in this case. In addition, Opsomer and Erciulescu (2021) present an approach to obtain replication weights from bootstrapping for variance estimation following a sample-based calibration that accounts for the extra variation due to benchmarks obtained from survey samples.

Alternatively, if the microdata of the survey sample, including the relevant variables, design variables and survey weights, are available, we can use a two-step bootstrap approach. Chen et al. (2022) show that standard with replacement bootstrap procedures applied separately to the sample and to the administrative data can provide good coverage for confidence intervals when the reference probability sample is selected by single-stage equal or unequal probability sampling designs.

In our context of assessing the quality of large administrative datasets using a survey reference sample for the population benchmarks, it is possible that the microdata for these surveys may not be available to apply the two-step bootstrap approach. In this situation, using very large probability-based surveys for the population benchmarks, such as the Labour Force Survey, American Community Survey, Coverage Surveys, etc. may have a small impact on any additional variation due to their large sample sizes. Moreover, the calculation of survey weights in these high-quality survey data typically includes the step of post-stratification where survey weights are calibrated to known population totals as mentioned in Section 1 and therefore do not have a variance. Therefore, the standard bootstrap procedure described above, assuming that the benchmarks are "true" population counts, could be applied in this case.

In the simulation study, we show the simulation standard errors resulting from drawing multiple random samples to produce the population benchmarks. We calculate the propensity scores on the administrative dataset for each of the samples and then show the simulation standard errors in the tables and confidence intervals in the figures in Section 3. As expected, the simulation standard errors are small. We also demonstrate some results of a two-step bootstrap sampling approach as set out in Chen et al. (2022) to show their similarity to the simulation standard errors in Table 3.5 of Section 3.

3. Demonstration of quality indicators and simulation

To demonstrate the proposed quality indicators, we use a simulated dataset of census microdata containing 6 anonymized UK Local Authorities across three regions of the UK produced from the 2001 UK Census by putting together many published UK census tabular data. This created a (synthetic) dataset of 1,163,659 individuals. The simulated variables on the dataset are: Geography (Local Authority) (6 categories), Sex (2 categories), Age Group (14 categories), Ethnic Group (16 categories), Marital Status (6 categories), Economic Status (10 categories).

From this dataset we carry out 2 steps (1) generate an administrative dataset treated as fixed (2) draw B = 40 random samples without replacement from the census to use as the probability-based reference samples to estimate the population benchmarks for the propensity score modelling. For the random samples, we draw a 1:50 sample and assume no nonresponse, i.e. each individual in the sample has a survey weight of 50. We repeat the simulation B = 40 times on each of the random samples and show in the results section the average outcomes and the simulation standard errors (confidence intervals) derived from the sample replicates.

3.1 Generating the administrative data

From information by the Office for National Statistics (ONS) experiences, we define 4 types of groups with associated probabilities: Group 1: The individual is not represented at all in the administrative data and are deleted; Group 2: The individual has moved to another geographic location and the geographical location is changed in the data; Group 3: the individual is recorded in the administrative data and in the correct location and therefore there is no change; Group 4: The individual has a duplicate record in another geographic location and therefore is copied to the other location as well (a duplicate).

For each group, we define a probability which varies across strata defined by sex, ethnic group (White, Non-White) and age group (below 30, 31-44, 45 and over) as shown in Table 3.1.

Table 3.1
Strata probabilities to generate administrative data

Age Group	Ethnic Group	Sex	p1	p2	p3	p4
below 30	White	Male	0.45	0.15	0.3	0.1
31-44	White	Male	0.2	0.2	0.5	0.1
45 and over	White	Male	0.1	0.15	0.7	0.05
below 30	White	Female	0.4	0.15	0.35	0.1
31-44	White	Female	0.15	0.15	0.55	0.15
45 and over	White	Female	0.1	0.1	0.75	0.05
below 30	Non-White	Male	0.37	0.15	0.43	0.05
31-44	Non-White	Male	0.15	0.2	0.6	0.05
45 and over	Non-White	Male	0.05	0.13	0.8	0.02
below 30	Non-White	Female	0.33	0.15	0.47	0.05
31-44	Non-White	Female	0.1	0.18	0.67	0.05
45 and over	Non-White	Female	0.05	0.1	0.83	0.02

Note: p1 – individual deleted from the administrative data; p2 – individual moved to another geography; p3 – individual included in administrative data and in the correct location; p4 – individual duplicated in another geography.

We randomly select the affected individuals in each of the strata and carry out the appropriate action in each of the groups to produce the generated administrative data. Following this step, we obtain a final administrative data set with $M = 1{,}026{,}969$ records (smaller than the original dataset of size $N = 1{,}163{,}659$).

3.2 Results

In Figure 3.1 we demonstrate the comparison of some univariate and bivariate distributions of the average of the B = 40 weighted survey counts (and their simulation confidence intervals) and the administrative data described in Section 3.1.

Figure 3.1 (a) Distributions of Age Group, (b) Geography (Local Authority) and (c) Geography (Local Authority) × Sex, comparing the average weighted survey counts (B = 40) and the administrative data (with simulation confidence intervals)

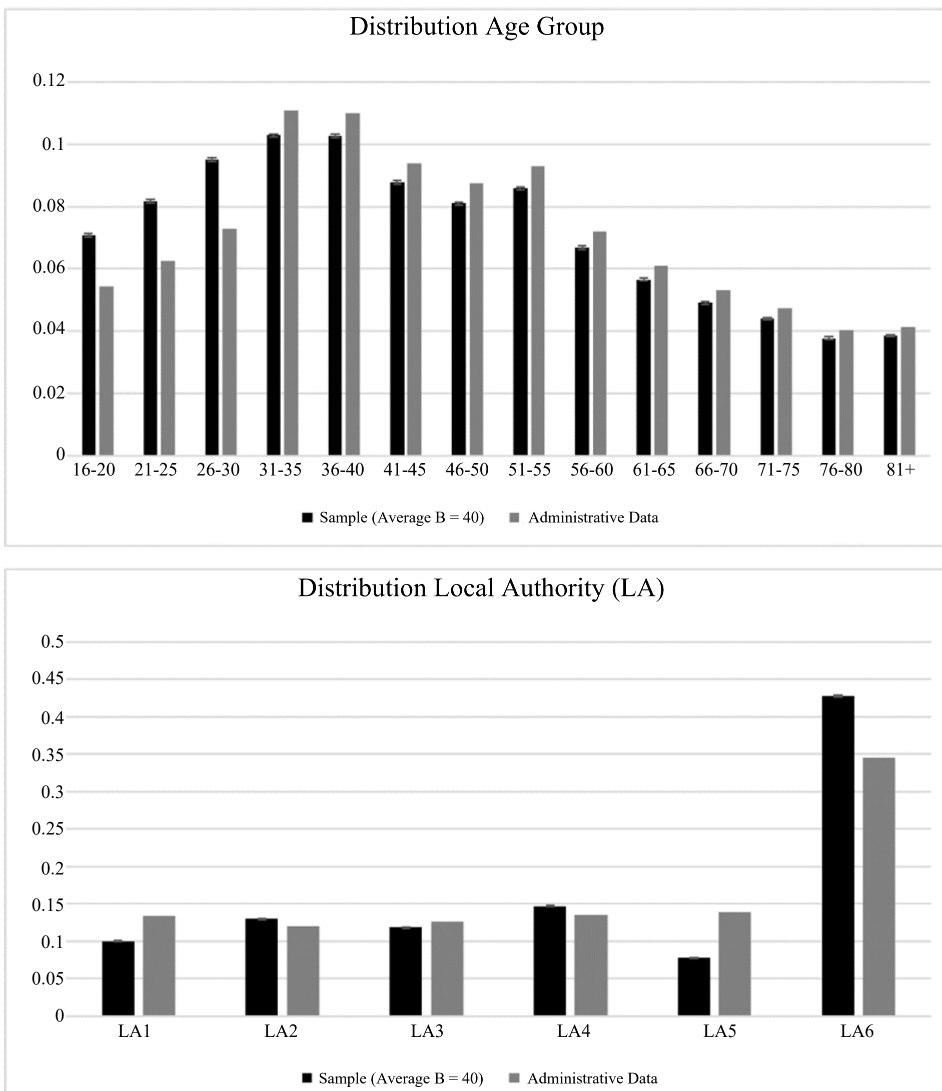
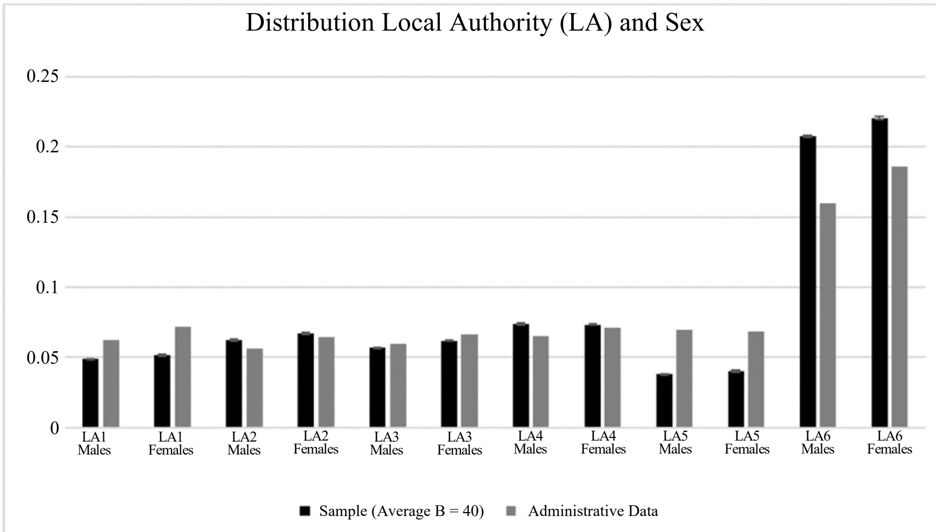


Figure 3.1 (continued) (a) Distributions of Age Group, (b) Geography (Local Authority) and (c) Geography (Local Authority) × Sex, comparing the average weighted survey counts (B = 40) and the administrative data (with simulation confidence intervals)



The approach we used to generate the administrative data was carried out differently for young ages and older ages where young people tend to migrate away from home more, eg. university housing. Hence, we see in Figure 3.1 under-representativeness in younger ages compared to older ages in the first plot (a). We also arbitrary selected the Local Authority LA6 for moving individuals from one Local Authority (LA) to another, hence we see a large discrepancy between the administrative data count and the weighted survey count for LA6 in the second plot (b) of Figure 3.1. Finally, we also show the impact on a bivariate distribution as can be seen in the last plot (c) of Figure 3.1 where males were slightly impacted more than females with respect to registering in the wrong geographical location LA6.

We now proceed to the quality indicators described in Section 2. In Table 3.2 we provide the distance metrics Indicator of Dissimilarity $(1-\mathrm{ID})$ and Hellinger's Distance $(1-\mathrm{HD})$ for univariate and bivariate distributions crossed with geography.

From Table 3.2 we can see that the Geography and Age Group variables have the smallest $(1-\text{distance}$ metrics) compared to other univariate distributions. This outcome impacts on all bivariate distributions when using the Geography variable, particularly when crossed with Age Group with 14 categories. The distance metrics also reveal that the number of categories impact on the distance metrics.

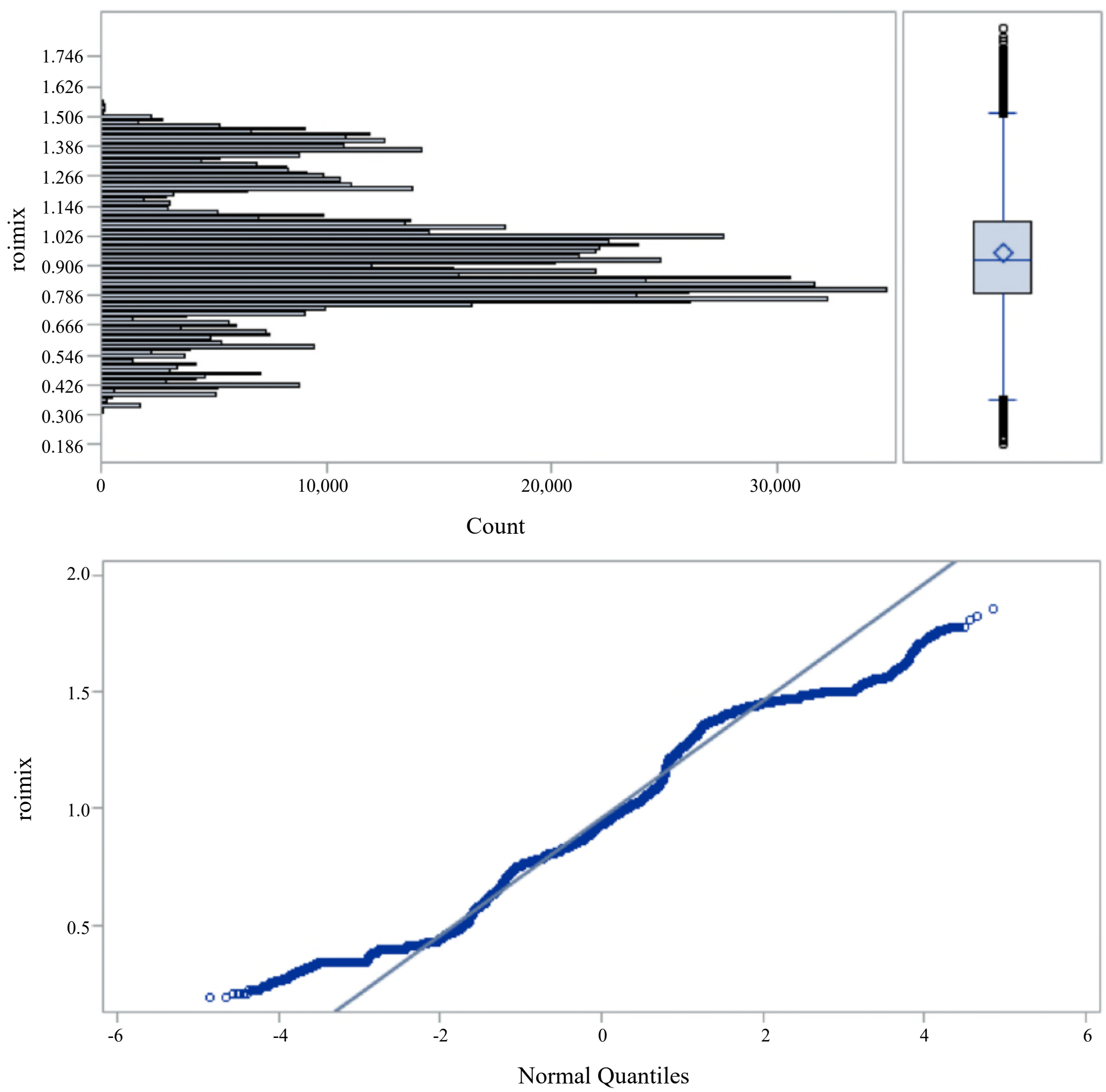
We turn now to the estimation of propensity scores and to the R-indicator calculations. Figure 3.2 shows the histogram of the estimated propensity scores (labeled "roimix") and against the Normal Distribution for the first sample (B = 1) and the plots are similar for all samples.

Table 3.2
Indicator of Dissimilarity (1 - ID) and Hellinger's Distance (1 - HD) (as described in Section 2.1) for a range of univariate and bivariate distributions

Variable	(1 - ID) Indicator of Dissimilarity (standard error)	(1 - HD) Hellinger's Distance (standard error)
Geography (6)	0.8973 (0.00049)	0.9103 (0.00043)
Age Group (14)	0.9424 (0.00047)	0.9500 (0.00038)
Ethnic Group (16)	0.9955 (0.00018)	0.9909 (0.00025)
Marital Status (6)	0.9603 (0.00049)	0.9688(0.00037)
Economic Status (10)	0.9732 (0.00047)	0.9738 (0.00032)
Geography/Sex (12)	0.8972 (0.00048)	0.9091 (0.00043)
Geography/Age Group (84)	0.8836 (0.00049)	0.8900 (0.00041)
Geography/Ethnic Group (96)	0.8900 (0.00047)	0.9034 (0.00042)
Geography/Marital Status (36)	0.8924 (0.00054)	0.9016 (0.00042)
Geography/Economic Status (60)	0.8927 (0.00048)	0.9014 (0.00042)

Note: Simulation standard errors in parentheses.

Figure 3.2 Histogram and QQ plot of the estimated propensity scores (labeled "roimix") for sample 1



As can be seen, in Figure 3.2, we have deviations from the Normal Distribution in the "tails" as would be expected when estimating propensity scores using the linear regression, with a good fit in the middle of the distribution. We also see propensity scores greater than 1 given the high representativeness indicator d of the administrative data showing the expected large propensities to be included in the administrative data.

For the R-indicator representing the overall variation of the propensity scores, we obtain the values (simulation standard error in parentheses):

Type 1 (population covariance): R = 0.7228 (0.0013).

Type 2 (mixture of administrative data and population means): R = 0.7413 (0.0010).

The R-indicator is bounded by 1 and (close-to) 0 and hence this value is slightly high. Therefore, we can say with confidence that we have above average overall representativeness of the administrative data compared to the auxiliary population (weighted sample) benchmarks.

Table 3.3 contains the partial variable-level R-indicators. Here, the higher the partial variable-level R-indicator, the more it is contributing to the lack of representativeness.

Table 3.3
Variable- level partial R-indicators (simulation standard errors)

Variable	Partial R-Indicator: Type 1	Partial R-Indicator: Type 2
Geography (6)	0.2869 (0.00218)	0.2106 (0.00093)
Sex (2)	0.0145 (0.00094)	0.0234 (0.00087)
Age Group (14)	0.1005 (0.00081)	0.1312 (0.00105)
Ethnic Group (16)	0.0288 (0.00069)	0.0230 (0.00064)
Marital Status (6)	0.0617 (0.00088)	0.0794 (0.00096)
Economic Status (10)	0.0534 (0.00068)	0.0674 (0.00086)

We see in Table 3.3 that all of the partial variable-level R-indicators under both Type 1 and Type 2 calculations are significantly different from 0 because the confidence intervals do not contain 0.

The Geography and Age Group variables have larger partial variable-level R-indicators for both Type 1 and Type 2 calculations and are contributing the most to lack of representativeness. This was evident in the distance metrics in Table 3.2 although for the partial variable-level R-indicators we account for the multivariate lack of representativeness compared to the calculation of distance metrics on single distributions. The results agree with the way we generated the administrative data.

Next, we look more closely at the Geography and Age Group variables with respect to their categories through the partial categorical- level R-indicators in Tables 3.4a and 3.4b, respectively.

For the Geography variable in Table 3.4a, LA5 has the highest over-representation in the administrative data for both types of calculations, whilst LA6 has the highest under-representation. For Age Group in Table 3.4b, the younger age groups have high under-representation compared to the older age groups for both types.

Next, we show in Table 3.5 the results of implementing a two-step bootstrap approach discussed in Section 2.3 to estimate standard errors for the R-indicator and variable-level partial R-indicators (Table 3.3). We resampled with replacement both the sample and the administrative data for B = 60, calculated the propensity scores and R-indicators and resulting standard errors of the R-indicators. Table 3.5 shows some slight differences between the simulation standard errors and the bootstrap standard errors although they have similar magnitude.

Table 3.4a
Categorical-level partial R-indicators for Geography (simulation standard errors)

Geography	Category-level Partial R-indicator: Type 1	Category-level Partial R-indicator: Type 2
LA1	0.0858 (0.00122)	0.0821 (0.00072)
LA2	-0.0441 (0.00080)	-0.0241 (0.00081)
LA3	-0.0065 (0.00096)	0.0200 (0.00079)
LA4	-0.0464 (0.00076)	-0.0272 (0.00078)
LA5	0.2270 (0.00231)	0.1437 (0.00081)
LA6	-0.1382 (0.00089)	-0.1229 (0.00076)

Note: LA = Local Authority.

Table 3.4b
Categorical-level partial R-indicators for Age Group (simulation standard errors)

Age Group	Category-level Partial R-indicator: Type 1	Category-level Partial R-indicator: Type 2
16-20	-0.0455 (0.00070)	-0.0619 (0.00110)
21-25	-0.0520 (0.00073)	-0.0675 (0.00106)
26-30	-0.0545 (0.00073)	-0.0723 (0.00107)
31-35	0.0211 (0.00093)	0.0216 (0.00078)
36-40	0.0187 (0.00099)	0.0194 (0.00081)
41-45	0.0180 (0.00092)	0.0173 (0.00078)
46-50	0.0105 (0.00100)	0.0190 (0.00086)
51-55	0.0127 (0.00085)	0.0207 (0.00075)
56-60	0.0114 (0.00113)	0.0170 (0.00092)
61-65	0.0113 (0.00116)	0.0158 (0.00097)
66-70	0.0108 (0.00102)	0.0157 (0.00087)
71-75	0.0089 (0.00074)	0.0139 (0.00059)
76-80	0.0068 (0.00122)	0.0124 (0.00102)
81+	0.0044 (0.00084)	0.0116 (0.00066)

Table 3.5
Comparison of two-step bootstrap standard error estimation to the simulation standard errors for R-indicators and Partial Variable-level R-indicators

Variable	Partial R-Indicator: Type 1		Partial R-Indicator: Type 2	
	Simulation Standard Error	Two-step Bootstrap Standard Error	Simulation Standard Error	Two-step Bootstrap Standard Error
R-indicator	0.0013	0.0010	0.0010	0.0007
Geography (6)	0.00218	0.00145	0.00093	0.00064
Sex (2)	0.00094	0.00082	0.00087	0.00082
Age Group (14)	0.00081	0.00060	0.00105	0.00087
Ethnic Group (16)	0.00069	0.00066	0.00064	0.00060
Marital Status (6)	0.00088	0.00062	0.00096	0.00077
Economic Status (10)	0.00068	0.00064	0.00086	0.00090

In summary, the findings of the R-indicator and their partial R-indicators are as expected given the way we generated the administrative data for this simulation study. We show here both types of calculations of the R-indicator and the partial R-indicators depending on the auxiliary information available for the variables of interest: Type 1 with full bivariate counts known or Type 2 with only the marginal counts known. The partial categorical-level R-indicators show higher standard errors under the full bivariate information in the Type 1 calculations, compared to using only the marginal information in the Type 2 calculations, although this was not evident for the variable-level partial R-indicators.

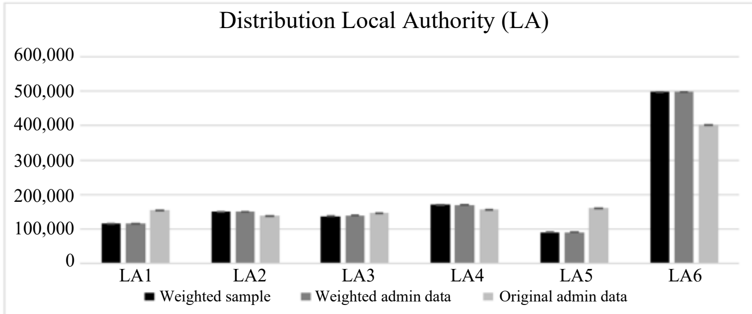
Based on all the partial categorical level R-indicators, we can build a profile of individuals that are under and over-represented in the administrative data taking into account the multivariate nature of the lack of representativeness. For example, we see that the younger age groups in LA6 are under-represented and it is important to look for alternative administrative data sources that can correct for this under-representation.

3.3 Adjusting for lack of representation

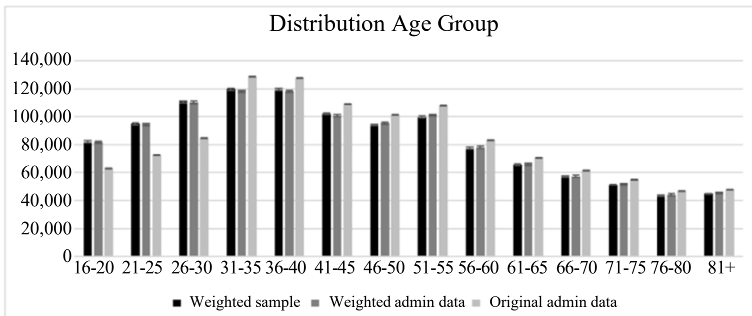
For the calculation of the R-indicators, we estimated propensity scores of being represented in the administrative data. We can now use the inverse of these propensity scores to weight the administrative data and to improve the estimation of distributions in the administrative data. We note that given we used a linear regression model to estimate the propensity scores, the weight is similar to the "g-weight" in standard generalized regression estimators (Särndal and Lundström, 2008). In Figure 3.3, we show distributions for some of the variables of the weighted sample counts, the inverse propensity weighted administrative data according to the Type 1 calculation and the original administrative data where we use a single correction factor according to the representativeness weight d to ensure that the total of the administrative data equals the weighted sample size: Correction factor C = 1163650/1026969 = 1.133. Under each of the figures, we also provide the $(1-\mathrm{HD})$ distance metrics comparing (1) the weighted sample distribution to the inverse propensity weighted administrative data distribution, and (2) the weighted sample distribution to the original administrative data distribution.

In summary, we see that weighting the administrative data by the inverse of the propensity scores improves the estimation of distributions compared to the original administrative data. This holds true particularly for those variables that were used to produce the simulated administrative data where there were large differences between the original administrative data and weighted sample counts, such as Age Group and Geography. The weighting also corrects the estimation of distributions in the administrative data for those variables that were not included in the production of generating the administrative data. For example, Economic Activity was not used to generate the administrative data, but nevertheless the $(1-\mathrm{HD})$ was improved when using the inverse propensity weights compared to the original administrative data: $(1-\mathrm{HD}) = 0.9962$ to $(1-\mathrm{HD}) = 0.9738$, respectively. Note that the variables shown in Figure 3.3 were included in the model to estimate the propensity scores and therefore it is important to account for all variables of interest when estimating propensity scores.

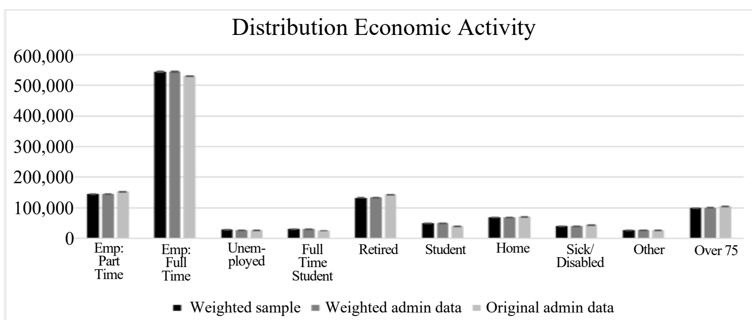
Figure 3.3 Comparison of weighted sample counts, inverse-propensity weighted administrative data (Type 1) and original administrative data for the variables



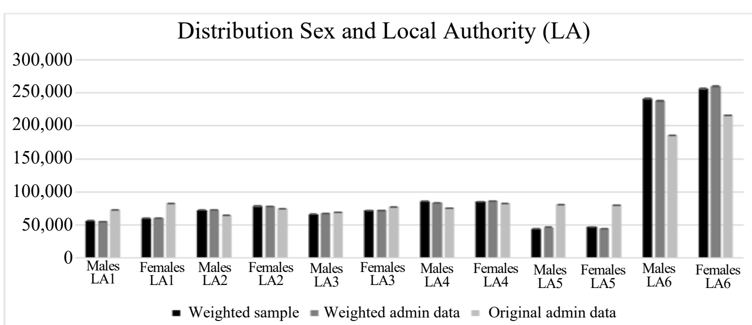
(1 - HD) weighted sample, weighted admin data = 0.9981 (0.00005)
(1 - HD) weighted sample, original admin data = 0.9103 (0.00043)



(1 - HD) weighted sample, weighted admin data = 0.9956 (0.00010)
(1 - HD) weighted sample, original admin data = 0.9500 (0.00038)



(1 - HD) weighted sample, weighted admin data = 0.9962 (0.00010)
(1 - HD) weighted sample, original admin data = 0.9738 (0.00032)



(1 - HD) weighted sample, weighted admin data = 0.9911 (0.00034)
(1 - HD) weighted sample, original admin data = 0.9091 (0.00043)

Notes: - (a) Geography, (b) Age Group, (c) Economic Activity and (d) Geography × Sex (below each figure are the distance metrics (1 - HD) for (1) the weighted sample vs the weighted administrative data, and (2) the weighted sample vs the original administrative data (simulation standard error in parenthesis)).
- HD = Hellinger's Distance.

4. Application

Working together with the team at the ONS, we have produced the proposed quality indicators using the R-code available on Github (Kim and Shlomo, 2023) according to the Type 2 calculation. The aim was to assess the representativeness of a combined administrative data source titled: Housing by Ethnicity Dataset (ABHED). This dataset is highly confidential and is not allowed outside of the secure servers of the ONS, therefore we worked collaboratively to produce the proposed quality indicators where the team at the ONS applied the R-code within their secure environment. The ABHED administrative data was developed by combining three separate admin-based composite data sources: Admin-based ethnicity dataset version 3.0 (ABED), admin-based household estimates version 3.0 and Admin-based housing stock version 1.0 (ABHS). In this application carried out on secure servers within the ONS, the team were able to use the Census 2021 for England and Wales as the comparative population auxiliary data. The application focused on four variables found within both the ABHED and the Census 2021 datasets: sex, age group, accommodation type and ethnicity and was carried out for one Local Authority. Full details of the work carried out on assessing the quality of the administrative data appear in the report by the Office for National Statistics (ONS) (2023).

For the univariate and bivariate distributions directly compared to the population auxiliary data through distance metrics, it was found that ethnicity and accommodation type deviated the most compared to the other variables at the univariate level according to the Hellinger's Distance. This also resulted in lower distance metrics when examining the bivariate distributions with ethnicity and accommodation type.

Moving to the R-indicator assessment, the overall R-indicator was 0.7565 showing good overall representativeness. Next, the partial R-indicators were calculated to check which variables, and which categories, are contributing to the lack of representativeness. Ethnicity had the largest variable-level partial R-indicator ($R_U(x_k) = 0.0346$) with the category "Other" having a large categorical-level partial R-indicator of $R_U(x_k^h) = -0.1764$, showing that this group is under-represented. This could be due to different definitions between the administrative data source compared to the auxiliary population benchmark from the Census 2021.

Accommodation type also had a large variable-level partial R-indicator ($R_U(x_k) = 0.0172$) with the categories Terraced Housing showing a large over-representation ($R_U(x_k^h) = 0.0798$) as well as Commercial Building ($R_U(x_k^h) = 0.0558$). The categories Purpose-built Flat had a large under-representation ($R_U(x_k^h) = -0.0643$) as well as Semi-detached ($R_U(x_k^h) = -0.0487$). The more evenly spread of the category-level partial R-indicators for Accommodation type ensured that the variable-level partial indicator for Accommodation was smaller than that of Ethnicity and contributed less to the lack of representativeness. The complete details of the partial R-indicators are in Table 4.1 reprinted from the report in Office for National Statistics (ONS) (2023).

Note that R-indicators account for the multivariate lack of representativeness compared to the distance metrics and provide more information on the impact of these variables when combined.

Table 4.1
Variable and categorical-level partial R-indicators for the ONS administrative data: Housing by Ethnicity Dataset (ABHED)

Variable	Partial R-indicator
Sex	0.0005
1. Female	0.0161
2. Male	-0.0167
Ethnicity	0.0346
1. White	0.0364
2. Mixed	0.0038
3. Asian	-0.0444
4. Black	-0.0118
5. Other	-0.1764
Accommodation type	0.0172
1. Detached	0.0041
2. Semi-detached	-0.0487
3. Terraced	0.0798
4. In a purpose-built block of flats or tenement	-0.0643
5. Part of a converted or shared house, including bedsits	-0.0265
6. Part of another converted building, for example, former school, church or warehouse	0.0217
7. In a commercial building, for example, in an office building, hotel or over a shop	0.0558
8. A caravan or other mobile or temporary structure	-0.0012
Age group	0.0059
1. 0-15	0.0470
2. 16-24	0.0241
3. 25-34	-0.0392
4. 35-44	-0.0344
5. 45-54	-0.0119
6. 55-64	-0.0065
7. 65-74	0.0002
8. 75+	0.0155

Note: Reprinted from Office for National Statistics (ONS) (2023).

The application did not include variance estimation for the quality indicators. However, given that the ONS were able to use the 2021 Census counts for population benchmarks to estimate the R-indicators, we anticipate small standard errors as seen in the simulation study presented in Section 3. Future work will include adding the two-step bootstrap variance approach in the Github package (Kim and Shlomo, 2023).

In summary, results of this real application show that the quality assessment to understand the lack of representativeness in the administrative data ABHED was able to point to variables and their categories that were less representative compared to the general population benchmarks obtained from the 2021 Census.

5. Discussion

It is important to assess the representativeness of administrative data prior to introducing their use in national statistical systems. The proposed quality indicators are motivated by the fact that the NSIs should make more use of their high-quality survey data where best practices are used to ensure robust data collection and survey weights are calculated to mitigate non-response biases and calibration to known population totals. The high-quality survey data can be used to assess the quality of administrative data, particularly when there are no census data available. The proposed quality indicators in this paper should help towards the goal of a quantitative assessment of representativeness of administrative data as they illuminate the variables and their categories that are contributing to the lack of representativeness. We showed that the derived quality indicators do a good job of assessing representativeness but depend on the variables that are included in the model to estimate the propensity scores. Moreover, we show that taking the inverse of the propensity scores as a weight in the administrative data improves the estimation of distributions.

Moreover, we have proposed a two-step bootstrapping approach in Section 2.3 to estimate standard errors for the indicators and estimates derived from IPW and demonstrate the approach in the simulation study presented in Section 3. The simulation study showed that the standard errors based on the replicate random samples were small (although we did not account for nonresponse in the samples drawn which may have increased the variation in the sample benchmarks) and the bootstrap standard errors were similar to the simulation standard errors for the R-indicators. We conclude that the quality indicators provide a reliable indication on the representativeness of the administrative data based on the covariates in the propensity score model.

The R-code with a user manual is available on Github (Kim and Shlomo, 2023). The R package takes as input the administrative dataset and a probability-based reference sample with appropriate survey weights and produces the quality indicators as shown in this paper.

In consultation with the ONS, to account for the very large administrative datasets that they hold (some having over 50 million individuals), we have developed new R-code in the Github platform for estimating the propensity scores. The new R-code can estimate the propensity scores from frequency tables of counts spanned by the variables in the propensity score model rather than on the individual microdata. This avoids the need to carry out the estimation of the propensity scores on microdata and reduces considerably the computing resources needed to calculate the quality indicators.

Acknowledgements

This research was funded by the grant: “Methodological Advancements on the Use of Administrative Data in Official Statistics”, Economic and Social Research Council (ESRC), ES/V005456/1, United Kingdom.

References

Alho, J.M. (1990). Logistic regression in capture-recapture models. *Biometrics*, 46(3), 623-635.

Austin, P.C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46(3), 399-424.

Bianchi, A., Shlomo, N., Schouten, B., Da Silva, D.N. and Skinner, C. (2019). Estimation of response propensities and indicators of representative response using population-level information. *Survey Methodology*, 45(2), 217-247. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2019002/article/00009-eng.pdf.

Chen, Y., Li, P., Rao, J.N.K. and Wu, C. (2022). Pseudo empirical likelihood inference for nonprobability survey samples. *The Canadian Journal of Statistics*, 50(4), 1166-1185.

Chen, Y., Li, P. and Wu, C. (2020). Doubly robust inference with non-probability survey samples. *Journal of the American Statstistical Association*, 115(532), 2011-2021.

Daas, P., Ossen, S., Vis-Visschers, R. and Arends-Toth, J. (2009). *Checklist for the Quality Evaluation of Administrative Data Sources*. Discussion paper (09042), Statistics Netherlands, The Hague, Netherlands, Available from: https://ec.europa.eu/eurostat/documents/64157/4374310/45-Checklist-quality-evaluationadministrative-data-sources-2009.pdf/24ffb3dd-5509-4f7e-9683-4477be82ee60 [accessed 25/02/2024].

Duncan, O.D. and Duncan, B. (1955). A methodological analysis of segregation indexes. *American Sociological Review*, 20(2), 210-217. https://doi.org/10.2307/2088328.

Dunstan, S. and Correia, S. (2021). *Framework for Assuring the Quality of Administrative Data for Use in the 2021 Census of England and Wales*. Working Paper EAP144, UK Statistics Authority. United Kingdom, Available at: https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2021/05/EAP144-Framework-quality-assurance-of-admin-data.pdf [accessed 25/02/2024].

Eurostat (2019). *Quality Assurance Framework of the European Statistical System, Version 2*. Available at: https://ec.europa.eu/eurostat/documents/64157/4392716/ESS-QAF-V1-2final.pdf/bbf5970c-1adf-46c8-afc3-58ce177a0646 [accessed 14/04/2025].

Kim, J.-K. and Haziza, D. (2014). Doubly robust inference with missing data in survey sampling. *Statistica Sinica*, 24.1, 375-394.

Kim, M.S. and Shlomo, N. (2023). *Quality Indicators for Administrative Data User Manual for R Package on GitHub*. Available at: https://github.com/sook-tusk/qualadmin [accessed 25/02/2024].

Office for National Statistics (ONS) (2016). *Patient Register: Quality Assurance of Administrative Data used in Population Statistics*. Office for National Statistics, United Kingdom, Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/patientregisterqualityassuranceofadministrativedatausedinpopulationstatisticsdec2016 [accessed 25/02/2024].

Office for National Statistics (ONS) (2022). *Coverage Estimation for Census 2021 in England and Wales*. ONS website, methodology article, (released 9 November 2022). Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/coverageestimationforcensus2021inenglandandwales [accessed: 11/07/2024].

Office for National Statistics (ONS) (2023). *Quality Indicators for Representativeness in Administrative Data: R-indicators and Distance Metrics*. Office for National Statistics, United Kingdom. Available at: https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/qualityindicatorsforrepresentativenessinadministrativedatarindicatorsanddistancemetrics#:~:text=R%2Dindicators%20and%20distance%20metrics%20use%20variables%20from%20the%20census,auxiliary%20data%20in%20this%20paper [accessed 25/02/2024].

Opsomer, J.D. and Erciulescu, A.L. (2021). Replication variance estimation after sample-based calibration. *Survey Methodology*, 47(2), 265-277. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2021002/article/00006-eng.pdf.

Ouwehand, P. and Schouten, B. (2014). Measuring representativeness of short-term business statistics. *Journal of Official Statistics*, 30(4), 623-649.

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.R-project.org/.

Särndal, C.-E. and Lundström, S. (2008). Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator. *Journal of Official Statistics*, 24(2), 167-191.

Schouten, B., Cobben, F. and Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, 35(1), 101-113. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2009001/article/10887-eng.pdf.

Schouten, B. and Shlomo, N. (2017). Selecting adaptive survey design strata with partial R-indicators. *International Statistical Review*, 85(1), 143-163.

Shlomo, N., Luiten, A. and Schouten, B. (2022). Representativeness of 2011 EU-SILC responses and response rates over time. In *Improving the Measurement of Poverty and Social Exclusion in Europe: Reducing Nonsampling Errors*, (Eds., P. Lynn and L. Lyberg), Publications Office of the European Union. Available at: https://op.europa.eu/en/publication-detail/-/publication/798e3ef9-fe65-11ec-b94a-01aa75ed71a1/language-en [accessed 15/04/2025].

Spector, P.E. (1992). *Summated Rating Scale Construction*. Sage.

United Nations (UN) (2019). *United Nations National Quality Assurance Frameworks Manual for Official Statistics*. Department of Economic and Social Affairs Studies in Methods Series ST/ESA/STAT/SER.M/100. Available at https://unstats.un.org/unsd/methodology/dataquality/references/1902216-UNNQAFManual-WEB.pdf [accessed 25/02/2024].

United Nations Economic Commission for Europe (UNECE) (2021). *Guidelines for Assessing the Quality of Administrative Sources for Use in Censuses*. Conference of European Statisticians Task Force, United Nations, Geneva, Switzerland. Available at: https://unece.org/sites/default/files/2021-10/ECECESSTAT20214_WEB.pdf [accessed 25/02/2024].

Wu, C. (2022). Statistical inference with non-probability survey samples. *Survey Methodology*, 48(2), 283-311. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2022002/article/00002-eng.pdf.

Zhang, L. (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, 66, 41-63.