Audit sampling as a quality standard for multisource official statistics

Li-Chun Zhang[1,2]

[1] *Statistisk sentralbyrå (lcz@ssb.no)*
[2] *University of Southampton (L.Zhang@soton.ac.uk)*

Abstract

Designed surveys through sampling or census are the standard approach to official statistics, where the targets are descriptive summaries of a given population. Official statistics are also commonly produced by combining relevant administrative registers, such as in the Nordic countries since the 1960s. The scope of non-survey data sources are being extended to include various so-called big-data sources, although so far relatively few multisource statistics of this kind have been credited as official statistics. *Trustworthy* evaluation of multisource official statistics is a fundamental issue for creating a new quality assurance standard. In this paper, audit sampling inference will be explained, illustrated and promoted to this end.

Keywords: Descriptive inference, survey sampling, auditing, census, social media index, register household

1 Introduction

A large portion of official statistics are aimed at descriptive summaries of a real finite population, such as the total, mean or quantiles of some specific values associated with the given population units, as well as any functions of such quantities. Although conceptually any such target parameter can be obtained through an error-free census of the target population, errors are unavoidable in practice despite the huge costs that may be required of the census survey operation. Over the 20th century survey sampling has been established as the standard approach to descriptive official statistics, which requires the sample to be selected under a probability sampling design. Survey sampling is more agile and costs much less than census survey. The inference of the associated statistical uncertainty is primarily grounded in the known sampling design; see Hansen (1987), Smith (1994), Kalton (2002), Rao (2005, 2011), Beaumont and Haziza (2022) for reviews and appraisals.

Since the 1960s, it has become increasingly common to produce official statistics based on relevant administrative registers, as in the Nordic countries; see e.g. Nordbotten (1966, 2010), UNECE (2007), Thygesen (2010). To satisfy the high quality required of official statistics, it is typically necessary to have a complete population frame, such as the Central Population Register, as well as the possibility to combine data from multiple registers generally (Zhang, 2012). In other words, a well coordinated *system* of statistical registers (Wallgren and Wallgren, 2014) is the key both to enable register-based statistics on a broad range of topics and to ensure that the statistics are fit-for-purpose.

Table 1: Multiple data sources for official statistics at present

Non-survey data	Example
Register	vital event, diagnosis wage, income tax, VAT, welfare payment
Transaction	scanner data price, point-of-sales receipt bankcard or giro payment B2B or B2P invoice property sales contract
Remote sensing, fixed	smart meter reading weather station reading traffic loop signal
Remote sensing, mobile	satellite image, drone image airborne laser scanning maritime AIS, lorry tracking signal mobile phone signal
Internet	web page, social media post
Survey data	census probability sample non-probability sample

Currently, the scope of multisource statistics are being extended to include various ‘big data’ or ‘new’ sources. Zhang and Haraldsen (2022) summarise the non-survey data sources available at present, which are reproduced here in Table 1 alongside the survey data, where the broad types of non-survey data sources are given in italics. Multisource statistics can be produced by combing data from two or more sources exemplified in Table 1. Di Zio et al. (2017) provide a synopsis of statistical methods for combining multiple sources of administrative and survey data. A similarly high-level overview covering the other non-survey data sources is yet to be compiled.

Regarding the quality of multisource official statistics, the ESSnet project KOMUSO has investigated the combination of administrative and survey data. The deliverables are available at the CROS portal (`https://cros-legacy.ec.europa.eu/content/essnet-quality-multisource-statistics-komuso_en`), including both guidelines for multisource statistics and multisource frames for social statistics, as well as a repository of statistical methods for measuring the output quality of multisource statistics. See also Yung et al. (2022) for a recent proposal of quality framework for statistical algorithms, which is relevant to making use of many other non-survey data.

Of the many quality dimensions (or aspects), we shall focus on statistical uncertainty (or accuracy) in this paper.

1.1 Approaches to descriptive inference

It is essential to recognise from the beginning the descriptive nature of the targets for official statistics, whenever this is the case, in contrast to analytic targets such as the life expectancy (of a hypothetical cohort of individuals) or a model that can be used to understand the given population. We shall use the term descriptive inference (Smith, 1983) to emphasise this epistemological distinction when our interest lies with the descriptive targets.

As mentioned before, design-based descriptive inference of any population parameter is the prevalent practice in survey sampling, where the uncertainty of estimation is evaluated with respect to repeated sampling under the given sampling design, while all the other values involved in estimation are treated as constants associated with the given population. In contrast, by model-based descriptive inference, the uncertainty of estimation would be evaluated with respect to an *assumed* statistical model of the relevant population values, while the selected sample is typically treated as fixed (Valliant et al., 2000).

The validity of design-based descriptive inference is assured by the *known* sampling design for the given finite population, "whatever the unknown properties of the population" (Neyman, 1934). Although models are necessary when the observations for inference are not obtained by probability sampling, model-based descriptive inference may be invalid if the assumed model is misspecified in any respects that matter to the task at hand, whether or not the available observations are obtained by probability sampling.

Of course, models may still be necessary in practical survey sampling for dealing with the non-sampling errors. For instance, Kalton (2002) notes that, "Whenever there are missing data, models are needed in the survey analysis... An important feature to note about all these compensation procedures is that they are general-purpose strategies, intended to enable analysts to perform any form of analysis. The models underlying these compensation procedures are developed to this end..." In other words, models are accepted as necessary practical remedies in such situations, which nevertheless does not negate the validity of design-based descriptive inference, nor does it imply generally the validity of fully model-based descriptive inference.

Finally, there are many model-assisted methods in survey sampling, where models are formulated to motivate the use of available auxiliary information in addition to the sampling design, in order to improve the efficiency of inference, but the properties of the estimators are still evaluated only with respect to the sampling design (e.g. Särndal et al, 1992; Breidt and Opsomer, 2017). Such methods will not be further discussed in this paper, because our focus here is the *validity* of descriptive inference approaches.

1.2 Audit sampling inference for multisource statistics

Multisource statistics may or may not involve survey sampling at all. To keep a sharp focus on the central thrust of this paper, we shall concentrate on the case where multisource statistics can be produced *without* any designed survey sampling. The approach of audit sampling inference, which is to be elaborated in this paper, is applicable as well when survey sampling is one of the data sources. However, it would require us to discuss the relevant technical means, which are more complicated than what is necessary for our focal case; so we kindly refer the interested readers to Zhang et al. (2023).

For an example that falls in the focus here, consider register-based statistics of the highest level of education. As remarked by Zhang (2012), to make use of the relevant registers of school enrolment records and examination results, integration with the Central Population Register is necessary, in order to delineate the target population as well as to reconcile any potentially overlapping or conflicting information in the various registers. What can one say about the statistical uncertainty of the resulting multisource statistics?

One possibility is to calculate some relevant indicators for the underlying data generation process and the final outputs, albeit without providing any estimated biases, variances or confidence intervals of the disseminated figures. Another option, which is less common in practice, is to explicitly model what are considered as the most important data generation mechanisms, in order to obtain various model-based uncertainty measures of the disseminated figures. For instance, let there be K distinct categories of the highest education level, such that a simple model is to assume that each individual in the target population follows independently the K-nomial distribution given the relevant covariates. However, since any assumed model cannot be entirely correct, what is the validity of such model-based descriptive inference?

To answer this last question, one needs validation data that are external to the data that have been used to produce the disseminated figures in the first place; otherwise some degree of circular reasoning would be unavoidable. For instance, one can perform cross-validation, whereby the population is split into a training set and a test set many times, such that the errors of a model fitted to the training set can be observed in the test set and used to generate some uncertainty estimates. Nevertheless, the final uncertainty estimates, obtained by combing the results from different test sets, would not directly refer to the disseminated figures that are produced based on the model fitted to the whole population (not just any subset of it). Moreover, the assumption of independence between the individuals still cannot be validated by such a cross-validation approach.

By resorting to design-based audit sampling inference, or simply *auditing*, one can avoid these conundrums for valid, model-based descriptive inference. As formulated by Zhang (2021), "Wherever the goal of survey sampling is to produce a point estimate of some target parameter of a given finite population, auditing aims not to estimate the target parameter itself but some chosen error measure of any given estimator of the target parameter, which may be biased due to failure of the underlying model assumptions or other favourable conditions that are necessary."

To recapitulate, for valid descriptive inference of the errors of multisource statistics based on non-survey data, one needs additional validation data. Next, provided the validation data are obtained under a known probability sampling design, auditing inference can yield design-based inference of these errors (as descriptive targets). Finally, design-based auditing inference is valid for the given population, regardless the models or algorithms that have been used to produced the multisource statistics that are being assessed.

As remarked by De Waal et al. (2020), the KUMOSO project has developed a number of "quality measures and methods to calculate them for separate steps, or building blocks, in the statistical production process. We hope that in the, hopefully near, future, an all-encompassing theory or framework to base quality measures for multisource statistics upon will be developed. Such an all-encompassing theory or framework should be able to handle several different types of error sources at the same time and, preferably, use the same statistical theory to treat these error sources." Auditing inference does provide a general and valid design-based framework, and one can apply it to the final statistical outputs directly. The approach is as universally applicable as survey sampling is for descriptive inference of finite populations.

Given the direction of travel, one can expect an ever increasing uptake of register-based statistics and model-based statistics based on non-survey data sources, where designed surveys are not directly needed for producing a range of fit-for-purpose official statistics on a continuous basis. Auditing can save cost in this context, as long as audit sampling can be conducted less frequently or with a smaller sample size than that is needed for producing the target statistics directly based on survey sampling.

Thus, we believe that auditing provides a feasible and trustworthy *quality assurance standard*, the adoption of which can be important for upholding the high quality required of official statistics and the public trust in them.

In the rest of the paper, we shall first explain the gist of auditing inference in some generic settings in Section 2, where non-survey multisource statistics may suffer from representation or measurement errors. Three real cases will be discussed in Section 3, to enhance the conceptual transition from survey sampling to auditing, as well as to provide some historic traces of the idea. Some summary final remarks will be given in Section 4.

2 Auditing inference in generic settings

Here we outline the gist of auditing inference in several generic settings, where non-survey multisource statistics may suffer representation or measurement errors. Let us start by introducing the following notations.

Denote by $U = \{1, ..., N\}$ the target population. Denote by $y_U = \{y_i : i \in U\}$ the values of interest, which are associated with each population unit. Denote by B a known set of units of the size N_B. For any $i \in B$, one either observes y_i directly or x_i in case measurement errors may be present. Both the units in B and the values associated with them may result from combining and processing multiple non-survey data sources (exemplified in Table 1).

Let $\bar{Y} = \sum_{i \in U} y_i / N$ be the population mean that is the descriptive target of interest. Let $\bar{y}_B = \sum_{i \in B} y_i / N_B$ be the corresponding mean in the set B if y_i is available, or $\bar{x}_B = \sum_{i \in B} x_i / N_B$ in the presence of measurement errors.

Finally, denote by s a sample of units which, depending on the setting, is either selected from U or $U \cup B$ under a known probability sampling design.

2.1 The selection error

Meng (2018) considers big data selection error in the setting $B \subset U$, as Figure 1 illustrates, where B is *not* a probability sample but measurement errors are absent. In this setup, a selection error exists if $\bar{y}_B \neq \bar{Y}$.

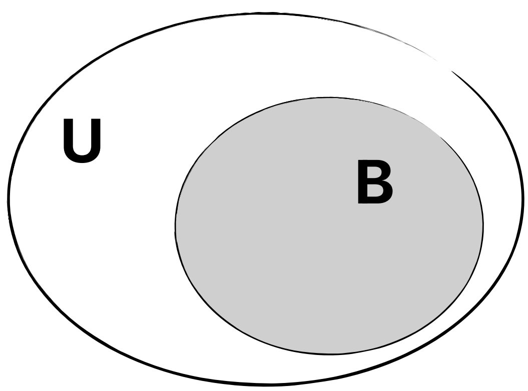

Figure 1: Setting of selection error of set B (shaded) from population U.

Meng (2018) discusses the selection error in terms of three factors, referred to as "data quantity" (i.e. the amount of data N_B), "problem difficulty" (i.e. the variation in the target variable y_i), and "data quality" (i.e. the correlation between the target variable y_i and the unit inclusion in B), the last of which is termed as the data defect correlation (ddc) regarding $\bar{y}_B$.

Meng (2022) observes that the ddc echos the unified criterion for selection error proposed in Zhang (2019), which is given as follows. Let

$$\delta_{iB} = \mathbb{I}(i \in B \mid i \in U)$$

indicate the B-set inclusion of any population unit. We have

$$\bar{y}_B - \bar{Y} = \frac{1}{N_B}\sum_{i\in U}\delta_{iB}y_i - \frac{1}{N}\sum_{i\in U}y_i = \frac{N}{N_B}\rho_U$$

where

$$\rho_U = \frac{1}{N}\sum_{i\in U}\delta_{iB}y_i - \big(\frac{1}{N}\sum_{i\in U}\delta_{iB}\big)\big(\frac{1}{N}\sum_{i\in U}y_i\big)\ . \tag{1}$$

Clearly, there would be no selection error, if the finite-population correlation ρ_U between y_i and δ_{iB} is 0. Zhang (2019) builds a non-parametric asymptotic (NPA) non-informativeness assumption on ρ_U, which permits a unified criterion for evaluating both the quasi-randomisation and super-population modelling approaches in case one would like to adjust the observed B-set mean $\bar{y}_B$.

For the purpose of auditing the selection error of $\bar{y}_B$, which is our topic here, it is evident that all the terms in (1) are either known or can be estimated given a probability sample s from U, where both y_i and δ_{iB} are treated as constants over sampling. Moreover, by virtue of the known audit sampling design of s, design-based inference of the selection error $\bar{y}_B - \bar{Y}$ (i.e. as a descriptive target) is valid regardless how B has been generated. Finally, other functions of the error, such as $(\bar{y}_B - \bar{Y})^2$, can also be estimated by auditing.

2.2 Auditing a model of selection mechanism

Consider still the problem of selection error (Figure 1). One common approach to derive an adjusted estimator of $\bar{Y}$ (i.e. instead of $\bar{y}_B$) is to introduce a selection model of the B-set inclusion indicator δ_{iB}, denote by

$$p_i = E_M(\delta_{iB} \mid y_i)$$

for each $i \in U$, where E_M denotes model expectation. Notice that although one would usually let p_i depend on other relevant covariates in addition to y_i, for simplicity we do not make them explicit in the notation here.

For the purpose of auditing the selection model that yields $\{p_i : i \in U\}$, we start by observing the identify

$$Cov_M(\delta_{iB}, y_i) = E_M Cov_M(\delta_{iB}, y_i \mid y_U) + Cov_M(p_i, y_i \mid y_U)\ .$$

The first term on the right-hand side vanishes because $Cov_M(\delta_{iB}, y_i \mid y_U) \equiv 0$. Given a sufficiently large population size N, the two remaining Cov_M-terms are essentially equal to the respective finite-population covariances because, under the model-based framework, $\{(y_i, \delta_{iB}) : i \in U\}$ are assumed to form an independent and identically distributed (IID) sample generated by some true joint distribution of (y_i, δ_{iB}). We have then

$$\sum_{i\in U} \frac{\delta_{iB} y_i}{N} - \big(\sum_{i\in U} \frac{\delta_{iB}}{N}\big)\big(\sum_{i\in U} \frac{y_i}{N}\big) \doteq \sum_{i\in U} \frac{p_i y_i}{N} - \big(\sum_{i\in U} \frac{p_i}{N}\big)\big(\sum_{i\in U} \frac{y_i}{N}\big)$$

In other words, the selection model would be perfectly compatible with the given (U, B), provided we have

$$\frac{1}{N}\sum_{i\in U} \delta_{iB} y_i - \frac{1}{N}\sum_{i\in U} p_i y_i = 0 \tag{2}$$

and

$$\frac{1}{N}\sum_{i\in U} \delta_{iB} - \frac{1}{N}\sum_{i\in U} p_i = 0 \tag{3}$$

Provided (3) is an estimating equation for the selection model parameters, which is common, it would be satisfied by the estimated $\{\hat{p}_i : i \in U\}$, i.e. $\sum_{i\in U} \delta_{iB} = \sum_{i\in U} \hat{p}_i$, and auditing would only need to be concerned with (2). For instance, one can examine a design-based estimate of $\sum_{i\in U} \delta_{iB} y_i/N - \sum_{i\in U} \hat{p}_i y_i/N$, or one can develop a significance test for (2) where the test-statistic distribution is generated by repeated audit sampling under the given design.

2.3 The error of selection and coverage

Consider now the problem of selection error in combination with coverage error, as illustrated in Figure 2, where B is *not* a probability sample, and $B \setminus U \neq \emptyset$ but the joint subset $BU = B \cap U$ is unknown. Measurement errors are absent. In this setup, an error due to selection *and* coverage exists if $\bar{y}_B \neq \bar{Y}$.

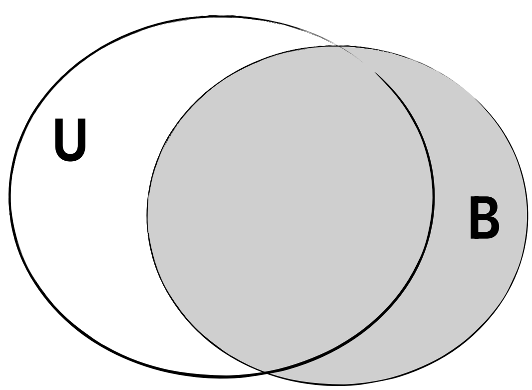

Figure 2: Setting of selection and coverage error of set B (shaded), unknown joint subset $B \cap U$ with target population U.

Let $\bar{y}_{BU}$ be the mean among the units in BU, belonging to both B and U. We have $\bar{Y} = \bar{y}_B$ if $\bar{Y} = \bar{y}_{BU}$ and $\bar{y}_B = \bar{y}_{BU}$. In other words, the selection and coverage error of $\bar{y}_B$ can be decomposed into the two selection errors of $\bar{y}_{BU}$ in terms of $BU \subset U$ and $BU \subset B$, respectively. Let $\delta_{iU} = \mathbb{I}(i \in U \mid i \in B)$ in addition to $\delta_{iB} = \mathbb{I}(i \in B \mid i \in U)$ defined earlier. Auditing the selection and coverage error of $\bar{y}_B$ amounts then to audting ρ_U given by (1) earlier and auditing

$$\rho_B = \frac{1}{N_B} \sum_{i \in B} \delta_{iU} y_i - \big(\frac{1}{N_B} \sum_{i \in B} \delta_{iU}\big)\big(\frac{1}{N_B} \sum_{i \in B} y_i\big)$$

i.e. as long as audit sampling of s from $U \cup B$ is feasible.

2.4 The error of representation and measurement

Consider still the setting of Figure 2, but suppose that measurement errors exist in addition such that only $\bar{x}_B$ is available but not $\bar{y}_B$ directly. In this setup, an error due to representation (pertaining to both selection and coverage) and measurement exists if $\bar{x}_B \neq \bar{Y}$.

One can view $\bar{x}_B - \bar{Y}$ as a descriptive parameter of the given union $B \cup U$, with associated values $\{x_i : i \in B\}$ for domain B and $\{y_i : i \in U\}$ for domain U, respectively. Similarly for any functions of it, such as $(\bar{x}_B - \bar{Y})^2$. However, since $\bar{x}_B$ is known, auditing such a descriptive parameter essentially requires only design-based estimation of $\bar{Y}$ given an audit sample s from U.

The reader is therefore entitled to wonder how auditing differs to survey sampling in this setting after all. Indeed, the same question would be pertinent for ρ_U as well, as long as $\bar{Y}$ is the only unknown quantity in (1).

To answer this question, let us recall an earlier remark that auditing can save cost, "as long as audit sampling can be conducted less frequently or with a smaller sample size than ... survey sampling", and consider it in the example of register-based statistics of the highest education level.

First, less frequent surveying is a fact in the Nordic countries, because the question about the highest education level is omitted in all the social surveys in these countries, such as the Labour Force Survey or EU-SILC, while at the same time this register variable can be used as auxiliary information for reducing the sampling error or the nonresponse bias of other survey variables at the estimation stage.

Next, should any Nordic country undertake an auditing of these statistics, the audit sampling design would surely differ to that of survey sampling in a country that does not have the same register capacity. For instance, the overall sample size can be smaller because the register variable provides strong auxiliary information in any case, and one could allocate a much larger part of the sample to the subpopulations where the register data are perceived to have a relatively low quality, such as the more recent immigrants.

Finally, we note that the difference between auditing and survey sampling will again manifest itself in the case illustrations to be discussed next.

3 Case discussions

Zhang (2021) defines audit sampling inference generally, where it is applied to the transaction-based proxy expenditure weights for Consumer Price Index. The idea of auditing of non-survey multisource statistics has actually a long tradition in official statistics, in which context it is often referred to as quality surveys. For instance, quality surveys have been common in the Nordic countries in connection with register-based censuses (e.g. Axelson et al., 2016), and Statistics Canada used to conduct a semi-annual survey to measure the errors in the Business Register (e.g. Lorenz and Laniel, 1992). Below we discuss three other cases of auditing in more details.

3.1 U.S. census coverage error

The U.S. Census Bureau traditionally conducts a post-enumeration survey (PES), and derives a so-called dual system estimator (DSE) of the population sizes by combining the census and the PES. Figure 3.1 illustrates the setting of target population U, census enumeration B and the PES sample s that is taken from $B \cup U$. If the DSE results are disseminated as the final population size estimates, then one should rightly view them as multisource official statistics. If the census counts are disseminated as official statistics, while the differences between the DSE results and the census counts are disseminated as estimates of the census net coverage errors, then the PES serves indeed the purpose of auditng the census coverage error.

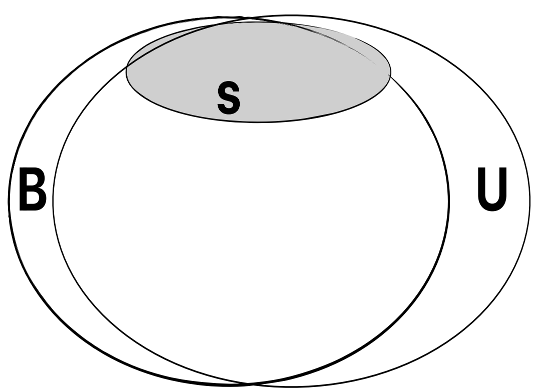

Figure 3: Population U, census enumeration B, PES sample s (shaded)

The debate between the two points of view, both statistically and legally, went on for decades. The decisions not to adjust the 1980 and 1990 censuses, i.e. not to accept the DSE results as the official population sizes, were upheld by the legal system. In January 1999, the U.S. Supreme Court ruled that the DSE adjusted numbers may not be used for apportioning Congress, i.e. the allocation of congressional seats to states. This finally sealed the PES as audit sampling instead of survey sampling. A headline of The Washington Post reads, "High Court Rejects Sampling In Census".

Several important lessons can be learned from this case. First, although one may readily accept the validity of design-based descriptive inference, the practically unavoidable non-sampling errors can still cast doubts on the results. As remarked by Freedman and Wachter (2001), "from a technical perspective, sampling is not the issue. The crucial questions are about the size of processing errors, and the validity of statistical models for missing data, correlation bias, and homogeneity — in a context where the margin of allowable error is relatively small." Nevertheless, these obstacles for accepting the DSE results as official statistics apparently do not prevent one from accepting the *same* estimates for the purpose of auditing.

Next, in retrospect, one can see that the case of the U.S. population census does provide a long-standing example of auditing in official statistics, although the conceptual transition from survey sampling may have been obscured to many, perhaps partly because the technical argument has often been framed as a question: what could be wrong with either the census or DSE.

Table 2: Net coverage error of U.S. census (Source: `census.gov`)

Census	1990	2000	2010	2020
Net coverage error (%)	-1.61	0.49	0.01	-0.24
Standard error (%)	0.20	0.20	0.14	0.25
Hypothetical standard error (%)	*0.30*	*0.30*	*0.22*	*0.38*

Finally, Table 2 shows the overall census net coverage error estimates from 1990 to 2020, together with the estimated standard errors, which can be openly obtained from the Census Bureau's online repository. In the bottom line of the table we have listed the corresponding hypothetical standard errors in italics, based on a conjectured 50% reduction of the PES sample size assuming that the variance is inversely proportional to the sample size. The question we would like to ask the reader is the following one. Suppose these hypothetical figures are proposed as the accuracy benchmarks for the PES auditing design, would you consider them to be acceptable for the purpose of quality assessment? If the answer is yes, then one can save cost; indeed, one might have saved costs historically, had the view of PES as auditing been crystal to all.

3.2 Bias of Social Media Index

Daas et al. (2015) use selected Dutch social media messages (e.g. Facebook, Twitter, LinkedIn, Google+) with sentiment classification (positive, neutral, or negative) to construct a Social Media Index (SMI), which aimed to emulate the Consumer Confidence Index (CCI) that is produced by Statistics Netherlands on the basis of a monthly survey. Figure 4 shows that the two indices resemble each other over the given 27 months, denoted by SMI_t and CCI_t for $t = 1, ..., 27$, where the empirical correlation coefficient between them is 0.88.

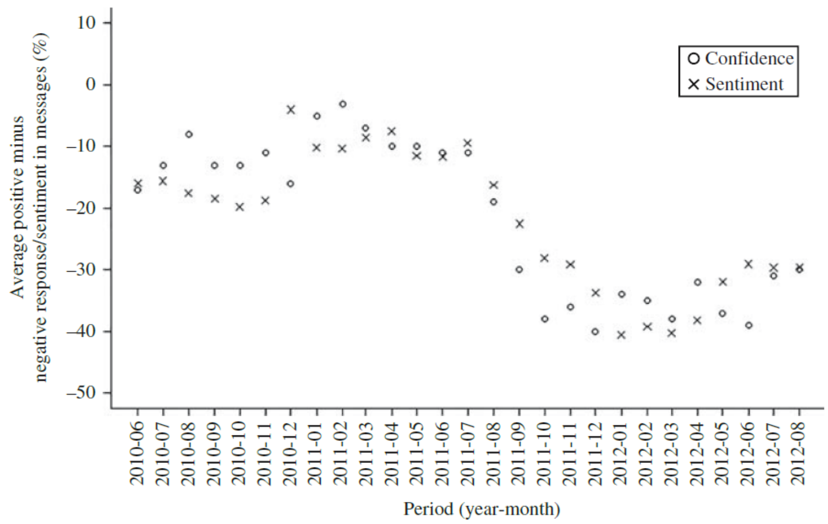

Figure 4: Monthly SMI (cross) and CCI (circle), by Daas et al. (2015, Fig. 4)

Patone and Zhang (2020) study the descriptive validity of SMI. Denote by θ_t the target of CCI in month t, for which the CCI is assumed to be approximately unbiased. Consider the SMI as an estimator with its own expectation and variance. Denote by ξ_t the expectation of SMI_t in month t. Given the large number of messages underlying each SMI_t, its variance is practically negligible compared to that of CCI_t, such that

$$\text{CCI}_t - \text{SMI}_t \doteq \text{CCI}_t - \xi_t \qquad \text{and} \qquad E_p(\text{CCI}_t - \text{SMI}_t) \doteq \theta_t - \xi_t$$

where E_p denotes expectation over repeated sampling for the CCI.

Treating $\theta_t - \xi_t$ as a descriptive parameter of the monthly Dutch population, Patone and Zhang (2020) develop formally a test statistic for

$$\mathcal{H}_0 : \ \theta_t - \xi_t = \mu \quad \text{vs.} \quad \mathcal{H}_1 : \ \theta_t - \xi_t \neq \mu$$

where the distribution of the test statistic is generated by repeated sampling for the CCI.

One could only obtain the CCI from the homepage of Statistics Netherlands but not its sampling variances directly. A plot of the 95% confidence interval of CCI over 2000 - 2014 is available in van den Brakel et al. (2017), from which one can gauge the coefficient of variation (CV) of CCI_t to be between 0.01 and 0.34 over the 27 months in consideration here. Patone and Zhang (2020) let $\eta = \text{CV}(\text{CCI}_t)$ be a constant over time, which ranges from 0.05 to 0.5. Given each η, the sampling variances of CCI_t are calculated as $(\eta\text{CCI}_t)^2$. Applying the test yields then the p-value of $\mathcal{H}_0$ accordingly.

As can be seen from Figure 5, the p-value is very close to 0, say, if $\eta \leq 0.2$ and it would only exceed 0.05 for $\eta > 0.37$ (marked by the horizontal line). In other words, viewing the Consumer Confidence Survey as an audit of the SMI,

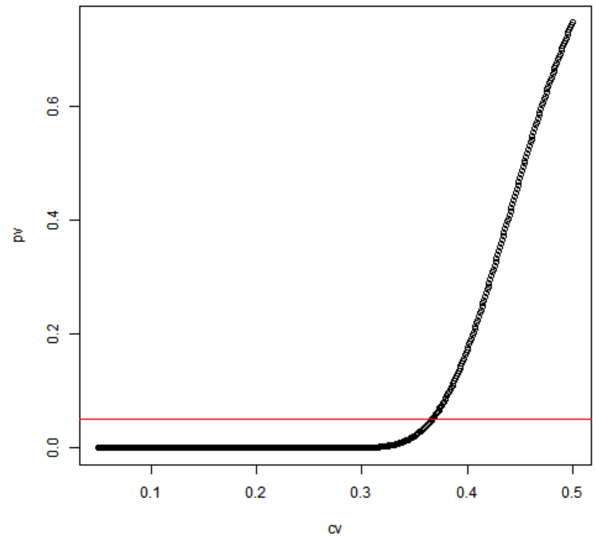

Figure 5: P-value of $\mathcal{H}_0$ given $\mathrm{CV}(\mathrm{CCI}_t)$, by Patone and Zhang (2020, Fig. 3)

one may reject the null hypothesis that the SMI deviates from the target of CCI by a constant over time.

3.3 The case of Norwegian register households

Statistical Household Register (HR) can be produced from combining multiple administrative registers of Population, Dwelling, Employment, Education, Post, etc. Household unit errors arise if people in the same household are allocated to different register households, or people in different households are allocated to the same register household. Zhang (2011) illustrates the unit error problem of register households with the fictive data in Table 3. As can be seen, there are 4 households both in reality and according to the HR, with quite similar household and associated individual characteristics. Only Lena is assigned the wrong household in the HR (marked as bold in Table 3),

Next, Table 4 compares different household counts (by size and type) in the Municipality Kongsvinger, which are based on the Central Population Register, the 2001 census of housing and a proxy HR, respectively. The proxy HR differs to the HR in production because it does not use the data from the 2001 census; see Zhang (2011) for details. The households compiled from the Central Population Register use only the family relationships by birth or marriage, which is the reason why the counts differ notably to the census results.

For a more rigorous and general assessment of the unit errors in the HR, Zhang (2011) introduces the allocation matrix to formally represent the data

Table 3: Fictive data at Storgata 99, by Zhang (2011, Tab. 1)

			REALITY				
Dwelling	Family	Household	Person	Name	Sex	Age	Income
H101	1	1	1	Astrid	Female	72	y_1
H102	2	2	2	Geir	Male	35	y_2
H102	2	2	3	Jenny	Female	34	y_3
H102	2	2	4	Markus	Male	5	y_4
H201	3	3	5	Knut	Male	29	y_5
H201	4	**3**	6	Lena	Female	28	y_6
H202	5	4	7	Ole	Male	28	y_7
			HOUSEHOLD REGISTER				
Dwelling	Family	Household*	Person	Name	Sex	Age	Income
H101	1	1	1	Astrid	Female	72	y_1
H101	2	2	2	Geir	Male	35	y_2
H101	2	2	3	Jenny	Female	34	y_3
H101	2	2	4	Markus	Male	5	y_4
H101	3	3	5	Knut	Male	29	y_5
-	4	**4**	6	Lena	Female	28	y_6
-	5	4	7	Ole	Male	28	y_7

of any given block of individuals. For the fictive data in Table 3, we have

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad A^* = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

where A refers to the true matrix and A^* that according to the HR.

Any error in the register-based statistics of households or their attributes, such as income or wealth, can be expressed as a function of finite-population joint distribution of (A, A^*), denoted by $f_U(A, A^*)$. By audit sampling, one can obtain a design-based estimate of $f_U(A, A^*)$, denoted by $\hat{f}_U(A, A^*)$, and any error of interest as a function of $\hat{f}_U(A, A^*)$.

Table 5 provides an example of auditing inference of the errors in the register household statistics (by size) for Kongsvinger. Specifically, the expectations of census results are calculated here conditional on the register, i.e. according to $\hat{f}_U(A \mid A^*)$, as well as the root squared errors of prediction (RSEP), from which one can derive the interval estimates of household counts and compare them to the register-based statistics. Notice that technically it is possible to derive the estimates either with or without taking into account the estimation uncertainty of $\hat{f}_U(A \mid A^*)$, as illustrated in Table 5.

There is a tacit assumption for the approach of Zhang (2011) reviewed above.

Table 4: Household counts in Kongsvinger, by Zhang (2011, Tab. 2)

Source: Central Population Register							
	Household size						
Household Type	1	2	3	4	5	6+	Total
Single	4143	0	0	0	0	0	4143
Couple without Children	0	1505	0	0	0	0	1505
Couple with Children	0	0	766	965	279	51	2061
Single Adult with Children	0	557	250	63	13	1	884
Others	0	4	0	0	0	0	4
Total	4143	2066	1016	1028	292	52	8597

Source: Census 2001							
	Household size						
Household Type	1	2	3	4	5	6+	Total
Single	3051	0	0	0	0	0	3051
Couple without Children	0	1845	0	0	0	0	1845
Couple with Children	0	0	826	966	283	61	2166
Single Adult with Children	0	433	197	58	10	1	699
Others	0	41	37	26	17	15	136
Total	3051	2319	1060	1080	310	77	7897

Source: Proxy Household Register							
	Household size						
Household Type	1	2	3	4	5	6+	Total
Single	3050	0	0	0	0	0	3050
Couple without Children	0	1791	0	0	0	0	1791
Couple with Children	0	0	811	977	281	55	2124
Single Adult with Children	0	418	190	52	10	1	671
Others	0	60	60	44	42	23	229
Total	3050	2269	1061	1073	333	79	7865

The joint distribution f_U is defined for each pair of allocation matrices (A, A^*), which requires blocking the individuals by judgement *a priori*, assuming that the individuals in the different blocks cannot possibly belong to the same household in reality. (It is easily ensured that blocking does not separate the individuals in the same register household.) This is similar to blocking in record linkage of large files. However, since the true blocks are not all known, blocking can induce errors in the inference.

A more satisfactory solution to blocking is to base the sampling design on graph sampling techniques (Zhang, 2022). Denote by $P = \{i_1, .., i_N\}$ all the individuals relevant to the target household population, and $R = \{k_1, ..., k_M\}$ the register households, and $H = \{h_1, ..., h_{M'}\}$ the households in reality. Define undirected simple graph $G = (U, L)$, with nodes $U = P \cup R \cup H$, edges $(ki), (ih) \in L$ if person i belongs to register household k and household h. The graph setting is illustrated in Figure 6, where the nodes P, R are known but not H, and the edges between P and R are known but not those between P and H.

Table 5: Auditing register household errors, by Zhang (2011, Tab. 4)

	Household size					
	1	2	3	4	5	6+
Proxy Household Register	3050	2269	1061	1073	333	79
Census	3051	2319	1060	1080	310	77
Conditional Expectation of Census	3100	2314	1053	1063	317	81
RSEP without estimation uncertainty	30	17	10	8	6	5
RSEP including estimation uncertainty	38	20	10	8	6	5

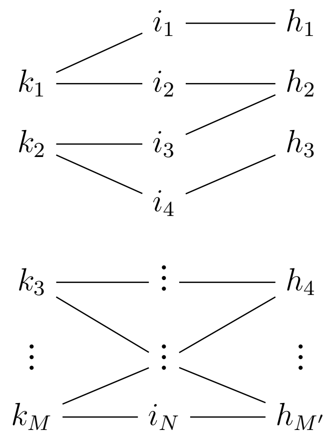

Figure 6: Illustration of setting for graph audit sampling, register households $R = \{k_1, ..., k_M\}$, individuals $P = \{i_1, .., i_N\}$, households $H = \{h_1, ..., h_{M'}\}$

Each true block of individuals is then a component in the graph G defined above. By graph sampling, starting from any initial sample of nodes selected from $R \cup P$, denoted by s_0, one would observe the edges incident to each node in s_0, such that any adjacent nodes not in s_0 can be included which form the first wave of snowball sample, denoted by s_1. Snowball sampling wave by wave amounts to repeating this incident observation procedure (Zhang, 2022), until no new nodes can be added in this way.

The obtained sample graph would contain a subset of all the components of G, each of which must have one or more nodes selected in the initial sample s_0. Let Ω denote all the components of G, which are unknown in advance, and let Ω_s denote the components observed in the sample graph. For each $\kappa \in \Omega$, there exists a paired allocation matrices (A_κ, A^*_κ), such that the target population distribution $f_U(A, A^*)$ is just the empirical distribution of $\{(A_\kappa, A^*_\kappa) : \kappa \in \Omega\}$, with the point mass $1/|\Omega|$ on each element of Ω.

Snowball sampling yields a sample of (A_κ, A^*_κ) for any $\kappa \in \Omega_s$, each of which corresponds to a block of individuals in reality. Various estimators of $f_U(A, A^*)$ can now be constructed according to the graph sampling theory (Zhang, 2022), which are unbiased with respect to repeated graph sampling.

4 Final remarks

Great cost reduction can be achieved by fit-for-purpose multisource official statistics when these can replace (repeated) survey sampling that would have been necessary otherwise. There are also huge potentials for increasing the scope or frequency of the relevant statistical outputs.

No matter how good the available non-survey data and the required models (or algorithms) actually are in such cases, descriptive inference based on them cannot be validated without additional data.

Audit sampling inference (Zhang, 2021) provides a general design-based statistical framework to error evaluation, where the uncertainty of the assessment is grounded in the known audit sampling design, just like survey sampling is general and valid for descriptive inference, "irrespectively of the unknown properties of the target population studied" (Neyman, 1934).

We have explained the gist of auditing inference in several generic settings, where non-survey multisource statistics may suffer from selection, coverage or measurement errors. We have also shown that the idea actually has a long tradition in official statistics, as well as many important applications. However, the awareness of the conceptual transition from survey sampling (for producing the statistical outputs) to audit sampling (for error evaluation) may still need to be enhanced among the practitioners, in order to capitalise on the cost saving the transition can offer to the enterprise of official statistics.

We believe auditing should be adopted as a quality assurance standard for descriptive official statistics that can be produced using whichever data sources, models or algorithms. The validity and generality of such a standard can be important for upholding the high quality required of official statistics and the public trust in them.

References

[1] Axelson, M., Holmberg, A., Jansson, I., Werner, P. and Westling, S. (2016). A Register- Based Census: The Swedish Experience. In *Administrative Records for Survey Method- ology (eds. A.Y. Chun, M. Larsen, G. Durrant, J.P. Reiter).* Wiley.

[2] Beaumont, J.-F. and Haziza, D. (2022). Statistical inference from finite population samples: A critical review of frequentist and Bayesian approaches. *The Canadian Journal of Statistics*, 50:1186-1212.

[3] Breidt, F. J. and Opsomer, J. D. (2017). Model-assisted survey estimation with modern prediction techniques. *Statistical Science*, 32:190-205.

[4] Daas, P.J., Puts, M.J., Buelens, B., and van den Hurk, P.A. (2015). Big data as a source for official statistics. *Journal of official statistics*, **31**, 249-262.

[5] De Waal, T., van Delden, A., and Scholtus, S. (2020). Commonly used methods for measuring output quality of multisource statistics. *Spanish Journal of Statistics*, 2:79-107. doi:https://doi.org/10.37830/SJS.2020.1.05

[6] Di Zio, M., Zhang, L.-C. and De Waal, T. (2017). Statistical methods for combining multiple sources of administrative and survey data. *The Survey Statistician*, **76**, 17-26.

[7] Freedman, D.A. and Wachter, K.W. (2001). Census adjustment: Statistical promise or illusion? *Society*, 39:26-33. https://doi.org/10.1007/BF02712617

[8] Hansen, M. (1987). Some History and Reminiscences on Survey Sampling. *Statistical Science*, 2:180-190.

[9] Kalton, G. (2002). Models in practice of survey sampling. *Journal of official statistics*, 18:129-154.

[10] Kiær, N. (1895). Den repræsentative Undersøgelsesmethode [The representative method of statistical surveys]. *Samfunnsøkonomiske studier (trykt utg.) 27, 1976.*

[11] Lorenz, P. and Laniel, N. (1992). Measuring the quality of the Business Register – Methodology and results. *Proceedings of the Survey Research Methods Section*, Washington.

[12] Meng, X.L. (2022). Comments on "Statistical inference with non-probability survey samples" – Miniaturizing data defect correlation: A versatile strategy for handling non-probability samples. *Survey Methodology*, 48:339-360.

[13] Meng, X.L. (2018). Statistical paradises and paradoxes in big data (i): Law of large populations, big data paradox, and 2016 US presidential election. *The Annals of Applied Statistics*, 12:685-726.

[14] Neyman, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society, pp. 558-625.*

[15] Nordbotten, S. (2010). The statistical archive system 1960-2010: A summary, presented at the Nordic Statistical Meeting, Copenhagen, August 11-14, 2010. http://nordbotten.com/articles/K%C3%B8benhavn%202010.pdf

[16] Nordbotten, S. (1966). Om "arkivstatistiske systemer", presented at the Nordic Statistical Meeting, Copenhagen, June 16-18, 1966. [English translation "On statistical file systems II", *Statistisk tidskrift, vol. 2, pp. 114-125, 1967.*]

[17] Patone, M. and Zhang, L.-C. (2020). On two existing approaches to statistical analysis of social media data. *International Statistical Review*, 89:54-71.

[18] Rao, J. N. K. (2011). Impact of frequentist and Bayesian methods on survey sampling practice: A selective appraisal. *Statistical Science*, 26:240-256.

[19] Rao, J. N. K. (2005). Interplay between sample survey theory and practice: An appraisal. *Survey Methodology*, 31:117-138.

[20] Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model-Assisted Survey Sampling*. Springer, New York.

[21] Smith, T. M. F. (1994). Sample surveys 1975–1990; an age of reconciliation? (with discussion). *International Statistical Review*, 62:5-34.

[22] Smith, T.M.F. (1983). On the validity of inferences from non-random sample. *Journal of the Royal Statistical Society, Series A*, 146:394-403.

[23] Thygesen, L. (2020). The importance of the archive statistical idea for the development of social statistics and population and housing censuses in Denmark, presented at the Nordic Statistical Meeting, Copenhagen, August 11-14, 2010. `https://www.dst.dk/extranet/staticsites/Nordic2010/pdf/bf7d6701-5b9f-4888-adc2-a45ce8debf87.pdf`

[24] UNECE (2007). *Register-based statistics in the Nordic countries: review of best practices with focus on population and social statistics*. United Nations Publication, ISBN 978-92-1-116963-8.

[25] Valliant, R., Dorfman, R. M., and Royall, R. M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. Wiley, New York.

[26] van den Brakel, J., Söhler, E., Daas, P. and Buelens, B. (2017). Social media as a data source for official statistics: the Dutch consumer confidence index. *Survey Methodology*, 43:183-210.

[27] Wallgren, A. and Wallgren, B. (2014). *Register-based Statistics: Statistical Methods for Administrative Data, 2nd edn*. Wiley.

[28] Yung, W., Tam, S-M, Buelens, B., Chipmand, H., Dumpert, F., Ascari, G. Rocci, F., Burger, J. and I. Choi (2022). *A quality framework for statistical algorithms*, Statistical Journal of the IAOS, 38:291-308.

[29] Zhang, L.-C. (2022). *Graph Sampling.* CRC Press.

[30] Zhang, L.-C. (2021). Proxy expenditure weights for Consumer Price Index: Audit sampling inference for big-data statistics. *Journal of the Royal Statistical Society, Series A*, **184**, 571-588.

[31] Zhang, L.-C. (2019). On valid descriptive inference from non-probability sample. *Statistical Theory and Related Fields*, **3**, 103-113. DOI:10.1080/24754269.2019.1666241

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

[33] Zhang, L.-C. (2011). A unit-error theory for register-based household statistics. *Journal of Official Statistics*, **27**, 415-432.

[34] Zhang, L.-C., Sanguiao-Sande, L. and Lee, D. (2023). Design-based predictive inference. *To be submitted.*

[35] Zhang, L.-C. and Haraldsen, G. (2022). Secure big data collection and processing: Framework, means and opportunities. *Journal of the Royal Statistical Society, Series A*, 185:1541-1559. DOI:10.1111/rssa.12836