Proxy expenditure weights for Consumer Price Index: Audit sampling inference for big-data statistics

Li-Chun Zhang[1,2,3]

[1] *University of Southampton (email: L.Zhang@soton.ac.uk)*
[2] *Statistics Norway*
[3] *University of Oslo*

Abstract: Purchase data from retail chains can provide proxy measures of private household expenditure on items that are the most troublesome to collect in the traditional expenditure survey. Due to the inevitable coverage and selection errors, bias must exist in these proxy measures. Moreover, given the sheer amount of data, the bias completely dominates the variance. To investigate the potential of *replacing* costly and burdensome surveys by non-survey big-data sources, we propose an audit sampling inference approach, which does not require linking the audit sample and the big-data source at the individual level. It turns out that one is unable to reject a null hypothesis of unbiased big-data estimation at the chosen size, because the audit sampling variance is too large compared to the bias of the big-data estimate. For the same reason, audit sampling fails to yield a meaningful mean squared error estimate. We propose a novel accuracy measure that is generally applicable in such situations. This can provide a necessary part of the statistical argument for the uptake of non-survey big-data sources, in replacement of traditional survey sampling. An application to disaggregated food price indices is used to demonstrate the proposed approach.

Keywords: privacy protection, survey burden and cost, proxy source effect, evaluation coverage

1 Introduction

The Consumer Price Index (CPI) measures the rate at which the prices of consumption goods and services change from month to month. It has substantial financial implications for governments and businesses, as the CPI is often specified in legislation and long-term contracts for adjusting payments for the effects of inflation. It is also a key statistic for

socio-economic policy-making, in which context it is of interest to calculate the CPI for specific subpopulations, such as pensioners, students, households with small children.

In practice, the CPI is calculated as a weighted average of the price changes (or indices) for a specified set of *aggregates* of consumer items. The weight of an aggregate is the share of the relevant consumer items in the household total expenditure, which reflect their relative importance in household consumption. The price indices for the aggregates and the weights of the aggregates are based on data obtained from separate sources. In this paper we focus on the estimation of the expenditure shares, to be referred to as the *CPI weights*. The price indices for the aggregates will be treated as given constants, when studying the statistical properties of the resulting CPI.

The CPI weights are traditionally estimated from the Consumer Expenditure Survey (CES). Due to the limited sample size, the CPI weights calculated specifically for different subpopulations have relatively large sampling variances. The CES is extremely burdensome due to its diary component, where the sample household needs to keep a diary of all relevant purchases typically over a two-week period. In many western countries, the CES has currently a high nonresponse rate, and is known to suffer from various misreporting errors; see e.g. Frickr et al. (2015), Battistin and Padula (2016), and Bee et al. (2012).

After the Norwegian CES was discontinued in 2012, the CPI has been calculated using *proxy weights* compiled from retailer turnovers that are available to the System of National Accounts. Under this *supplier-data approach*, one cannot connect the items of transaction to different consumer subpopulations, in order to calculate the CPI weights for any specific subpopulation. However, provided the connection is feasible, the relevant purchase data can provide proxy CPI weights for the subpopulations of interest. Unlike the CES-based weights, these proxy weights can be considered to have virtually zero sampling variance for practical purposes because of the sheer amount of data that can be made available. But they are generally biased due to a number of errors that are unavoidable in reality. In particular, these include coverage errors caused by the discrepancy between the available purchases and the entire household consumption of the target subpopulation, and selection errors from the available purchases because, for various technical reasons, one is not able to code and classify all the items in them.

In such a situation, where bias completely dominates variance, modelling the intrinsic variability of the proxy weights would be fruitless, as long as it cannot capture the bias. Additional observations of expenditure are necessary to investigate the extent to which the proxy weights may be biased. The thrust of this paper is to develop a general *audit sampling inference* approach to the following three relevant questions in this context:

I. Under which *condition* is the price index based on proxy CPI weights unbiased?

II. How to *test* the potential bias of index resulting from the proxy CPI weights?

III. How to *measure* the accuracy of the price index based on proxy CPI weights?

As will be easily made clear later, it is possible for the CPI to be unbiased, even when it is based on weights that are biased themselves. Question I is nevertheless important, because it allows one to shift the focus from the bias of proxy weights to that of the resulting CPI, which is what really matters. We shall refer to the condition for unbiased CPI as the *validity condition* for proxy weights. Smith (1983) and Zhang (2019) use the term in the same sense, whereby the observed data can be used validly for estimation, although they are not obtained via a known probabilistic design. Whereas Question I can be answered analytically, additional data are necessary in order to address Questions II and III. The available CES data will be treated as an audit sample in this respect.

We emphasise the following. Firstly, the audit sampling approach is developed from the perspective that non-survey big-data (including administrative registers) can possibly *replace* survey sampling altogether, in routine production of official statistics. The matter of using transaction-based proxy weights for the Norwegian CPI provides a case-in-point that is relevant to other countries as well. Secondly, the proposed approach is generally applicable, where confidence and accuracy measures for big-data statistics are required to construct the statistical argument for the transition. The inference of uncertainty is based on the audit sampling distribution, which is valid whether or not the validity condition for the big-data statistics is satisfied. Thirdly, in case the transition to big-data statistics does take place, audit sampling inference can be applied from time to time to assess their performance. How frequent and how accurate the auditing inference needs to be depends chiefly on the user demands and the available resource in a given context.

It is important to notice that audit sampling serves a fundamentally different purpose to that of survey sampling. Wherever the goal of survey sampling is to produce a point estimate of some target parameter of a given finite population, audit sampling aims not to estimate the target parameter itself, but some chosen accuracy measure of any given estimator of the target parameter, which may be potentially biased due to failure of the underlying assumptions or other favourable conditions that are necessary.

Mean squared error (MSE) is a common choice of accuracy measure. However, as will be demonstrated later, MSE estimation can easily produce negative (hence unusable) results, unless the audit sampling variance is small compared to the bias of the big-data statistic. It is unattractive to simply increase the audit sample size in such situations, which means audit sampling would be relatively more costly in a relatively favourable setting where the big-data statistic has only a small bias. Instead, we propose and develop a novel accuracy measure for big-data statistics, where bias completely dominates variance, which is generally

applicable based on audit sampling and overcomes the problem of limited audit sample size. To the best of our knowledge the proposal is brand new to the literature.

Notice that there is a long tradition in survey sampling, where auxiliary information including non-survey proxy measures can be used to improve the efficiency, by appropriate weighting adjustment or prediction modelling; see e.g. Särndal et al. (1992) and Valliant et al. (2000), respectively. One can approach the CES from such a perspective, where the target parameters are the yearly expenditure shares of a given household population, and the purchase data that yield the proxy CPI weights are relevant auxiliary data. Nevertheless, we are not aware of any existing practice where the two sources are combined in this way. A major obstacle is privacy concern, against linking individual observations in the CES to the purchase data from the retailers. Another is the extra cost and burden required to collect the required auxiliary data from the businesses.

The proposed audit sampling approach has also very different characteristics to the traditional application of statistical techniques for auditing. Audit sampling techniques for Accounting (e.g. Neter, 2011) can be relevant, if one treats the expenditure measures derived from purchase data as the *book (proxy) amounts*, and use sampling from them to obtain a sample of *audited (correct) amounts*, e.g. in order to estimate the accounting error of total book amount and to analyse the individual book amount errors. But this approach is ineffective for assessing the under-coverage error of the total book amount. Moreover, in the present context, it would require taking a sample from the purchase data directly, which may not be feasible due to the same objections above against privacy, cost and burden. In contrast, under our approach, the CES constitutes a *separate* sample from the target universe, which is not linked to the purchase data at the individual level, but allows one to assess the combined effect of all the errors in the purchase data.

The rest of paper is organised as follows. In Section 2, we introduce the data for this study and the objective of age-group specific food price indices motivated by the available data. In Section 3 we clarify the validity condition for unbiased CPI based on biased purchase data proxy weights, and develop tests for the bias of the resulting CPI when the validity condition is not fully satisfied. In Section 4, we start by showing that the estimation of MSE runs into troubles, because the CES audit sample size is not sufficiently large. We then propose and develop a novel accuracy measure for big-data statistics, and demonstrate the approach by applying it to the age-group food price indices. A summary of some final remarks are given in Section 5.

2 Data and target index

We describe below the data for this study, the choice of target index and some relevant overall considerations. The available purchase data are obtained from some of the largest supermarket chains in Norway. These pertain to the 11 COICOP groups 111 - 119, 121 and 122, according to the classification of individual consumption by purpose (COICOP), developed by the United Nations Statistics Division and adopted by the Eurostat. Together these 11 commodity groups at the 3-digit level constitute the first 2-digit category Food and non-alcoholic drinks in Table 1, henceforth simply referred to as *food*.

Table 1: Household expenditure, total in NOK. (Source: ssb.no)

	1998 - 2000		2012	
	Total	%	Total	%
Consumption in all	*280078*	*100*	*435507*	*100*
01 Food, non-alcoholic drinks	33499	12,0	51429	11,8
02 Alcohol, tobacco	8114	2,9	11717	2,7
03 Clothing, shoes	16278	5,8	23618	5,4
04 Housing, household energy	71278	25,4	135982	31,2
05 Furniture, household articles	17321	6,2	24495	5,6
06 Health	7717	2,8	11421	2,6
07 Transport	56832	20,3	81574	18,7
08 Post, telecommunication	5610	2,0	8253	1,9
09 Culture, recreation	33634	12,0	43347	10,0
10 Education	869	0,3	985	0,2
11 Restaurant, hotel, etc.	11379	4,1	15557	3,6
12 Other goods or services	17547	6,3	27129	6,2

As the target index we shall consider age-group specific food price indices, as examples of disaggregated CPIs. To compute these target indices, we treat the 11 commodity groups at the 3-digit level as the aggregates of consumption items, for which we need the associated price indices and CPI weights. The audit sampling inference approach to be developed later will be applied to these age-group food price indices.

Aggregate price indices Figure 1 shows the price indices for the chosen 11 aggregates published by Statistics Norway, over the 36 months of 2015 - 2017, denoted by $T = 36$. They are based on prices in the scanner data collected directly from all the major supermarket chains, which arise from scanning the bar codes for individual products at electronic points of sale in retail outlets, including detailed information about quantities and values of goods sold as well as their prices. In this study we use these 11×36 price indices as possible food aggregate price indices one is likely to encounter in practice, and treat them

as given constants. Scanner data constitute a rapidly expanding source of price data for CPI purposes. For information regarding index methodology based on scanner data, we refer to the website of Ottawa Group (`http://www.ottawagroup.org`).

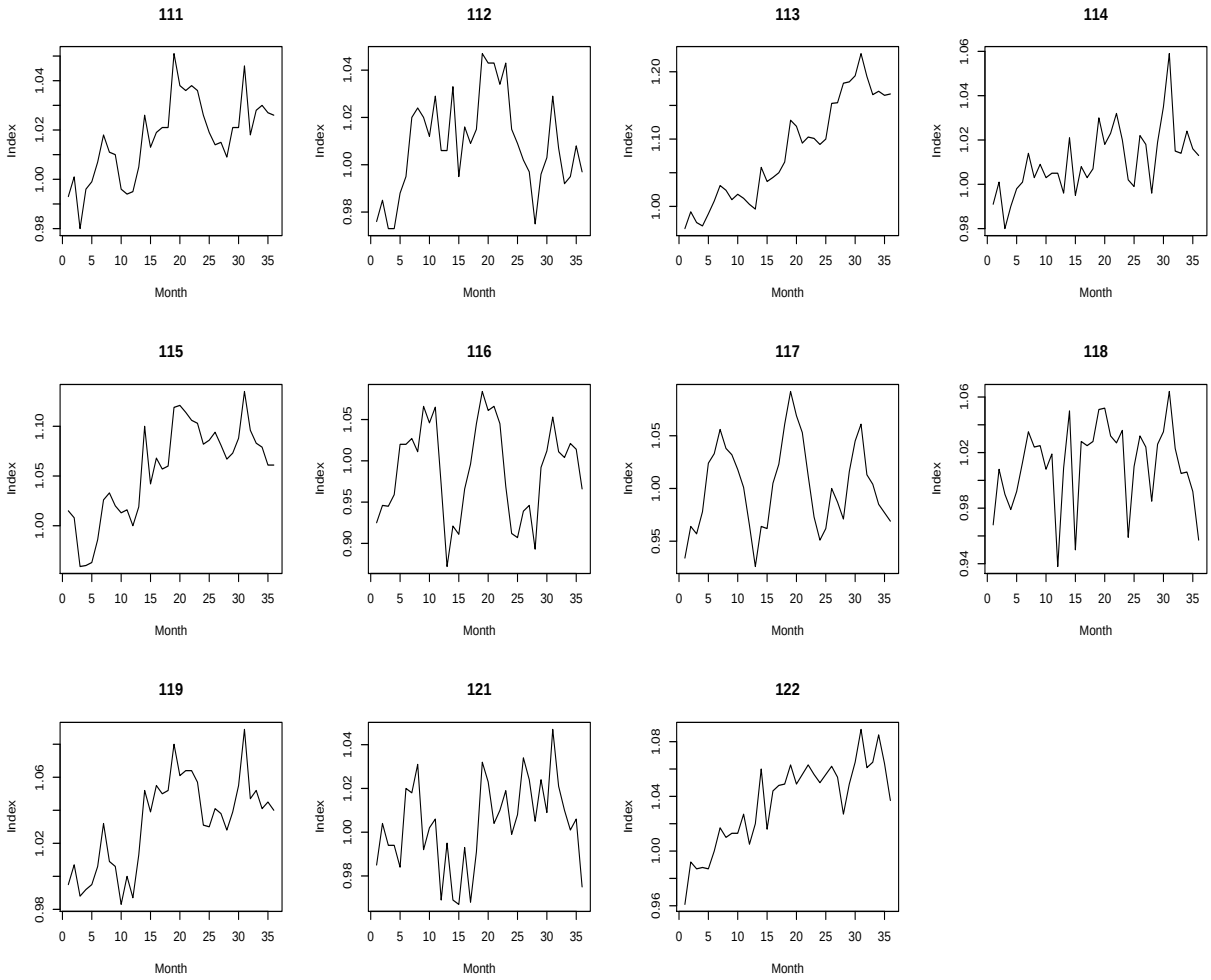

Figure 1: Price indices for 11 food aggregates over 36 months (Source: Statistics Norway)

Aggregate CPI weights The CPI weights for the aggregates are traditionally estimated based on the expenditure data collected in the CES. Food items are the most burdensome CES diary component in this respect. Throughout the first decade of this millennium, the sample size of the Norwegian CES is about 7000 households, and the response rate is about 50%. The diary data are collected over a period of two weeks for each respondent household. For this study, we shall ignore the nonresponse effects, as if the respondent households were a simple random sample by design, treat the CES-based CPI weights as unbiased estimators of the true CPI weights, and only calculate their sampling variances

as the associated uncertainty of estimation.

For disaggregation of the food price index, we consider four groups by the age of household head: up to 25, 26 - 40, 41 - 67, 68 and above, denoted by $g = 1, ..., 4$. As can be seen in Table 1, the *total* expenditure share of the 11 food aggregates decreased by only 0.2% over the decade, which on average is less than 0.02% each year. The weights based on the CES in 2012 further breaks down the 11 yearly food aggregate weights by age groups. These are given in Table 2, for all ages as well as for each age group. Clearly, these CES-based age-group specific weights are associated with large relative standard errors.

Table 2: Weights of food aggregates based on CES and transaction data (proxy).

COICOP	All		Up to 25		26 - 40		41 - 67		68 and above	
(3-digit)	CES	Proxy	CES	Proxy	CES	Proxy	CES	Proxy	CES	Proxy
111	.136	.150	.152	.166	.147	.157	.136	.147	.123	.124
112	.195	.138	.177	.118	.181	.135	.200	.146	.172	.126
113	.059	.041	.035	.023	.044	.035	.060	.045	.090	.067
114	.161	.168	.161	.148	.162	.169	.154	.169	.178	.181
115	.017	.020	.015	.012	.016	.017	.020	.022	.031	.032
116	.076	.084	.055	.067	.067	.083	.073	.085	.092	.100
117	.102	.104	.100	.092	.097	.100	.099	.107	.101	.113
118	.093	.090	.107	.106	.111	.088	.097	.088	.078	.085
119	.059	.079	.073	.086	.063	.088	.055	.073	.050	.072
121	.025	.030	.020	.026	.024	.024	.027	.031	.038	.043
122	.076	.096	.103	.156	.089	.103	.079	.087	.046	.056

For proxy weights of the food aggregates from transaction data, one possibility is via the retailer loyalty members, whose membership IDs are registered together with the purchases. Some obvious issues of this option include a relatively large under-coverage error of loyalty members and their registered purchases, confidentiality restrictions and burden on the businesses for data extraction. An alternative is via the card transactions, since almost all purchases are paid by card (or other digital means) in Norway, and a handful payment services account for most of the transactions. The separate data of purchases and card transactions can be linked to each other based on non-personal features such as time, outlet and amount, connecting thus the cardholder to the linked purchases. Standard disclosure control methods are applied to preserve data confidentiality, whereby expenditure data by age groups can be extracted from the linked data, without revealing the cardholder identity or the time, place and other details of individual purchases.

For this study, we have fully anonymised expenditure data based on extractions provided by the largest card payment service and some of the largest supermarket chains in Norway. The data pertain to a single weekday in September of 2016, consisting of 0.8

million transactions, broken down according to the age of the cardholder. These proxy weights are given in Table 2. More data can be acquired if the proxy weights are to be used in routine production. Thus, for this study we shall consider the proxy weights to have practically zero variance compared to the sampling variance of the CES weights.

Regarding the set-up For audit sampling inference we shall use the CES in 2012 as an audit sample, and treat the CES-based weights as unbiased estimates of the corresponding true CPI weights for 2012. Setting aside the coverage and selection errors of the available transaction data with respect to all household purchases, the proxy CPI weights do not refer to exactly the same subpopulations as those identified in the CES for two reasons. First, the available transaction data refer to a different time point than the CES. Second, the proxy weights are broken down by the age of the card holder, instead of the age of the household head in the CES. Thus, these proxy weights are *necessarily* biased for the true CPI weights in 2012. Should the proxy weights nevertheless found to be advantageous compared to the CES-based weights, it would only strengthen the plausibility of adopting concurrent proxy weights in the routine production of CPI.

3 Testing proxy source effect

3.1 Proxy source effect and validity condition

Let p_{it} be the price index for aggregate i in month t, where $i = 1, ..., m$ and $m = 11$, and $t = 1, ..., T$ and $T = 36$, as shown in Figure 1. Let $\boldsymbol{w}_g = (w_{1g}, ..., w_{mg})^\top$ be the true weights of the aggregates for age-group g, where $\sum_{i=1}^m w_{ig} = 1$ for $g = 1, ..., 4$. The target food price index for age group g in month t is given as

$$P_{gt} = \sum_{i=1}^{m} w_{ig} p_{it} = \boldsymbol{p}_t^\top \boldsymbol{w}_g$$

where $\boldsymbol{p}_t = (p_{1t}, ..., p_{mt})^\top$. Denote by $\boldsymbol{w}_g^* = (w_{1g}^*, ..., w_{mg}^*)^\top$ the big-data proxy weights, where $\sum_{i=1}^m w_{ig}^* = 1$ for $g = 1, ..., 4$, which are given in Table 2. Denote by $P_{gt}^* = \boldsymbol{p}_t^\top \boldsymbol{w}_g^*$ the age-group food index based on the proxy weights. The bias of P_{gt}^*, or *proxy source effect*, due to the proxy weights is given by

$$\Delta_{gt} = P_{gt}^* - P_{gt} = \boldsymbol{p}_t^\top (\boldsymbol{w}_g^* - \boldsymbol{w}_g) \tag{1}$$

It is clear from the expression (1) that one does *not* need to have $\boldsymbol{w}^* = \boldsymbol{w}$, in order to avoid the proxy source effect for P_{gt} that is the target of estimation. Of course, the bias of

P^*_{gt} would most likely be small, if $\boldsymbol{w}^*$ is close to $\boldsymbol{w}$. More specifically, the validity condition can be stated as follows. Now that $\sum_{i=1}^m (w^*_{ig} - w_{ig}) \equiv 0$ by definition, for any (g,t), one can consider Δ_{gt} as the *empirical* covariance between p_{it} and $w^*_{ig} - w_{ig}$, over $i = 1, ..., m$. Thus, the proxy weights are valid, i.e. yielding unbiased index P^*_{gt}, provided p_{it} is empirically uncorrelated with the bias of the proxy weights $w^*_{ig} - w_{ig}$ across the aggregates. That is, whether or not an aggregate has a higher than average price index is not related to whether or not its proxy weight is higher than the corresponding true weight.

3.2 Tests for bias

We have $\Delta_{gt} = 0$ in (1) for a particular month t, if the proxy weights are valid for P_{gt}. Since price indices are used to measure how prices change over time, it is helpful to make greater use of $\boldsymbol{p} = \{\boldsymbol{p}_t : t = 1, ..., T\}$ when testing the bias of proxy weights. Given any g, one can postulate a simple structural relationship

$$P_{gt} = \gamma_g + \beta_g P^*_{gt}$$

between $\{P^*_{gt} : t = 1, ..., T\}$ and $\{P_{gt} : t = 1, ..., T\}$ over time. Next, let $\hat{\boldsymbol{w}}_g = (\hat{w}_{1g}, ..., \hat{w}_{mg})^\top$ be the CES-based weights, for $g = 1, ..., 4$. Given unbiased $\hat{\boldsymbol{w}}_g$ over repeated sampling, we can write $\widehat{P}_{gt} = \boldsymbol{p}_t^\top \hat{\boldsymbol{w}}_g = P_{gt} + e_{gt}$, where $E(e_{gt}) = 0$ and $V(e_{gt}) = \boldsymbol{p}^T V(\hat{\boldsymbol{w}}_g) \boldsymbol{p}_t^T$ is the sampling variance of $\hat{P}_{gt}$. Combing the two equations, we have

$$\widehat{P}_{gt} = \gamma_g + \beta_g P^*_{gt} + e_{gt} \ . \tag{2}$$

Using the model (2), we consider below three tests for the null hypothesis

$$\mathcal{H}_0 : \ \Delta_{gt} = 0 \quad \text{for given } g \text{ and all } t = 1, ..., T.$$

In each case, we choose a target parameter θ, possibly vector-valued, for which an unbiased estimator $\hat{\theta}$ can be given based on the CES, and use the Wald test statistic

$$X(g,g) = (\hat{\theta} - \theta)^\top \hat{V}(\hat{\theta})^{-1} (\hat{\theta} - \theta) \ \sim \ \chi^2_\kappa$$

with suitable degree of freedom κ. Moreover, to explore the sensitivity of the test, we combine $\widehat{P}_{gt}$ and P^*_{ht} for different age groups, where $g \neq h$, to define another test statistic, denoted by $X(g,h)$, which should result in rejection of the null hypothesis. The test is shown to be lacking power, if $X(g,h)$ fails to reject in such a set-up.

First, assuming $\beta_g \equiv 1$, the model (2) reduces to $\widehat{P}_{gt} = \gamma_g + P^*_{gt} + e_{gt}$. Let $H_0 : \gamma_g = 0$

under the reduced model, and $\theta = \gamma_g = 0$, and regress $\widehat{P}_{gt}$ on P^*_{gt} to yield

$$\hat{\theta} = \hat{\gamma}_g = \bar{\widehat{P}}_g - \bar{P}^*_g = \sum_{t=1}^{T} \widehat{P}_{gt}/T - \sum_{t=1}^{T} P^*_{gt}/T = \bar{\boldsymbol{p}}^\top \hat{\boldsymbol{w}}_g - \bar{\boldsymbol{p}}^\top \boldsymbol{w}^*_g$$

where $\bar{\boldsymbol{p}} = \sum_{t=1}^{T} \boldsymbol{p}_t/T$, and $V(\hat{\theta}) = \bar{\boldsymbol{p}}^\top V(\hat{\boldsymbol{w}}_g)\bar{\boldsymbol{p}}$, and $\kappa = 1$. To check the sensitivity of the test, one can regress $\widehat{P}_{gt}$ on P^*_{ht} for $g \neq h$, and define $X(g,h) = V(\hat{\bar{P}}_g)^{-1}(\hat{\bar{P}}_g - \bar{P}^*_h)^2$. Since $\hat{\bar{P}}_g - \bar{P}^*_h$ is a biased estimator of θ, unless $\bar{P}_g = \gamma_h + \bar{P}^*_h$ happens to be the case under H_0 here, $X(g,h)$ should lead one to reject H_0 if the test does not lack power.

Next, as long as $\beta_g = 1$, it may be worthwhile to explore P^*_{gt} in practice, even if $\gamma_g \neq 0$ and $\Delta_{gt} \neq 0$ over time. Let $H_0 : \beta_g = 1$, and $\theta = \beta_g = 1$, and

$$\hat{\theta} = \hat{\beta}_g = \frac{\sum_{t=1}^{T}(P^*_{gt} - \bar{P}^*_g)(\widehat{P}_{gt} - \bar{\widehat{P}}_g)}{\sum_{t=1}^{T}(P^*_{gt} - \bar{P}^*_g)^2} = \boldsymbol{d}_g^\top \hat{\boldsymbol{w}}_g$$

where $\kappa = 1$, and $V(\hat{\theta}) = \boldsymbol{d}_g^\top V(\hat{\boldsymbol{w}}_g)\boldsymbol{d}_g$, and

$$\boldsymbol{d}_g = \sum_{t=1}^{T}(P^*_{gt} - \bar{P}^*_g)(\boldsymbol{p}_t - \bar{\boldsymbol{p}}) / \sum_{t=1}^{T}(P^*_{gt} - \bar{P}^*_g)^2 \ .$$

To check the sensitivity, one can regress $\widehat{P}_{gt}$ on P^*_{ht} for $g \neq h$, giving $\hat{\theta}' = \boldsymbol{d}_h^\top \hat{\boldsymbol{w}}_g$, and

$$X(g,h) = \left(\boldsymbol{d}_h^\top V(\hat{\boldsymbol{w}}_g)\boldsymbol{d}_h\right)^{-1}(\hat{\theta}' - 1)^2 \ .$$

Finally, let $H_0 : (\gamma_g, \beta_g) = (0,1)$ under (2), and $\theta = (\gamma_g, \beta_g)^\top$, where $\hat{\beta}_g = \boldsymbol{d}_g^\top \hat{\boldsymbol{w}}_g$, and

$$\hat{\gamma}_g = \bar{\widehat{P}}_g - \hat{\beta}_g \bar{P}^*_g = \boldsymbol{\ell}_g^\top \hat{\boldsymbol{w}}_g \quad \text{and} \quad \boldsymbol{\ell}_g = \bar{\boldsymbol{p}} - \bar{P}^*_g \boldsymbol{d}_g \ .$$

We have $\kappa = 2$, and $V(\hat{\gamma}_g) = \boldsymbol{\ell}_g^\top V(\hat{\boldsymbol{w}}_g)\boldsymbol{\ell}_g$, and $Cov(\hat{\gamma}_g, \hat{\beta}_g) = \boldsymbol{\ell}_g^\top V(\hat{\boldsymbol{w}}_g)\boldsymbol{d}_g$. Again, to check the sensitivity, one can regress $\widehat{P}_{gt}$ on P^*_{ht} for $g \neq h$, giving

$$X(g,h) = (\hat{\gamma}', \hat{\beta}' - 1)V(\hat{\theta}')^{-1}(\hat{\gamma}', \hat{\beta}' - 1)^\top$$

where $\hat{\theta}' = (\hat{\gamma}', \hat{\beta}')^\top$, and $\hat{\gamma}' = \boldsymbol{\ell}_h^\top \hat{\boldsymbol{w}}_g$ with $\boldsymbol{\ell}_h = \bar{\boldsymbol{p}} - \bar{P}^*_h \boldsymbol{d}_h$ and $\hat{\beta}' = \boldsymbol{d}_h^\top \hat{\boldsymbol{w}}_g$, and $V(\hat{\gamma}') = \boldsymbol{\ell}_h^\top V(\hat{\boldsymbol{w}}_g)\boldsymbol{\ell}_h$ and $V(\hat{\beta}') = \boldsymbol{d}_h^\top V(\hat{\boldsymbol{w}}_g)\boldsymbol{d}_h$ and $Cov(\hat{\gamma}', \hat{\beta}') = \boldsymbol{\ell}_h^\top V(\hat{\boldsymbol{w}}_g)\boldsymbol{d}_h$.

Figure 2 shows four scatter plots of $\widehat{P}_{gt}$ vs. P^*_{ht}, where $(g,h) = (4,4)$, $(4,1)$, $(1,1)$ and $(1,4)$. As can be seen, $\widehat{P}_{4t}$ and P^*_{4t} for the fourth age group (top-left in the figure) appear to be scattered around the unity slope (dashed line), as well as $\widehat{P}_{1t}$ and P^*_{1t} for the first age

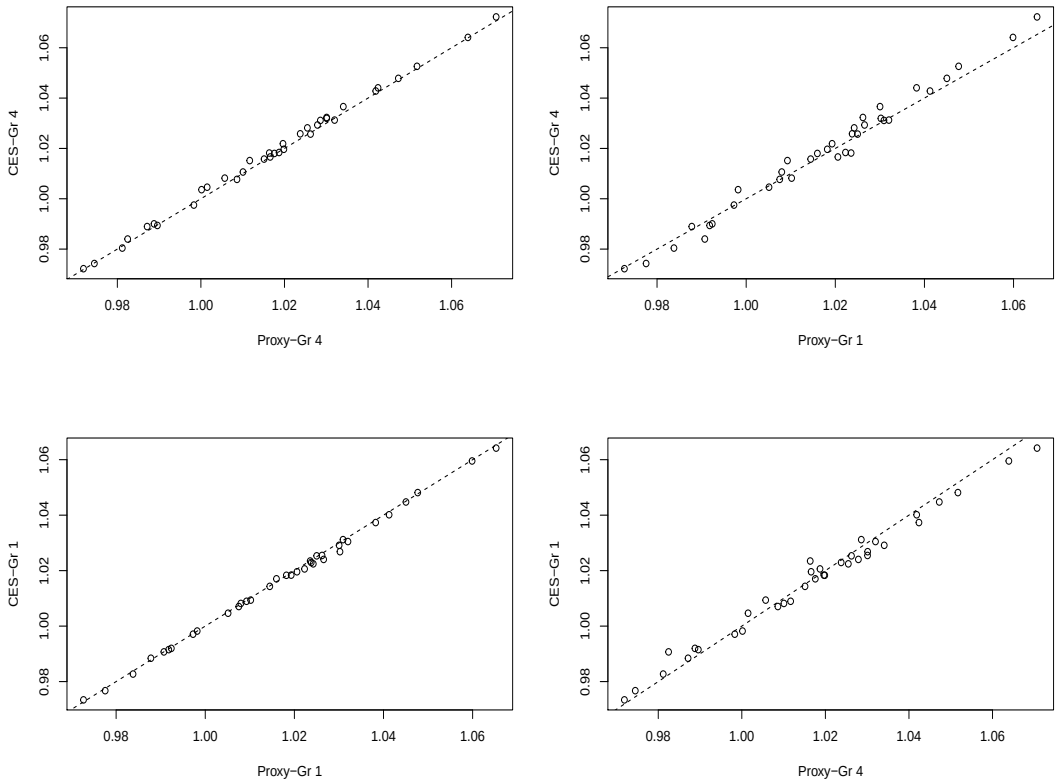

Figure 2: Scatter plots of age-group (Gr) food indices, CES-based vs. proxy weights.

group (bottom-left). In contrast, the points $(P^*_{1t}, \widehat{P}_{4t})$ appear to scatter around a different slope to the unity (top-right). Similarly for $\widehat{P}_{1t}$ vs. P^*_{4t} (bottom-right).

Table 3: Testing $H_0 : \gamma_g = 0$ assuming $\beta_g \equiv 1$, for $g = 1, 4$.

Age group $g = 1$			Age group $g = 4$		
Statistic	Test statistic	P-value	Statistic	Test statistic	P-value
$X(1,1)$	0.0000744	0.993	$X(4,4)$	0.03803	0.970
$X(1,4)$	0.0000643	0.994	$X(4,1)$	0.03666	0.971

Table 3 gives the test results for $H_0 : \gamma_g = 0$ assuming $\beta_g \equiv 1$, for $g = 1, 4$. For example, in the case of $X(4,4)$, we are testing the bias of using the proxy weights that ideally should be close to the true weights for that age group. The p-value is close to 1, so the observed mean difference $\widehat{\bar{P}}_4 - \bar{P}^*_4$ is very small compared to the sampling variance of $\hat{\boldsymbol{w}}_4$. But the test lacks sensitivity, since one cannot reject based on $X(4,1)$ either, where the p-value is about the same, although one can observe in Figure 2 that the source effect is larger in this case. Similarly with the results for age group $g = 1$ based on $X(1,1)$ and $X(1,4)$.

Table 4: Testing unity slope $H_0 : \beta_g = 1$, for $g = 1, 4$.

Age group $g = 1$			Age group $g = 4$		
Statistic	Test statistic	P-value	Statistic	Test statistic	P-value
$X(1,1)$	0.0318	0.859	$X(4,4)$	0.045	0.832
$X(1,4)$	2.62	0.106	$X(4,1)$	6.82	0.009

Table 4 gives the test results for unity slope $H_0 : \beta_g = 1$ alone. In the case of $X(4,4)$, the p-value is 0.832, and $\hat{\beta}_4$ deviates little to unity relatively to the uncertainty of $\hat{\boldsymbol{w}}_4$. Similarly with $X(1,1)$. This test is more sensitive, because the p-value is 0.009 based on $X(4,1)$ and 0.106 based on $X(1,4)$, where the former is significant at the 5% level.

Table 5: Testing $H_0 : (\gamma_g, \beta_g) = (0, 1)$, for $g = 1, 4$.

Age group $g = 1$			Age group $g = 4$		
Statistic	Test statistic	P-value	Statistic	Test statistic	P-value
$X(1,1)$	1.63	0.442	$X(4,4)$	0.45	0.799
$X(1,4)$	122	0.000	$X(4,1)$	110	0.000

Table 5 gives the test results for $H_0 : (\gamma_g, \beta_g) = (0, 1)$, where the two coefficients are tested jointly. The p-value is 0.799 based on $X(4,4)$ for $g = 4$, and 0.442 based on $X(1,1)$ for $g = 1$, such that H_0 cannot be rejected for either age group. Moreover, the power of this test is high and the results provide stronger corroboration to the proxy CPI weights, where the p-value is 0.000 either based on $X(4,1)$ or $X(1,4)$.

The tests above are used to investigate each age group on its own. It is possible to test the biases for all the indices jointly, e.g. $H_0 : (\gamma_g, \beta_g) = (0, 1)$ for *all* $g = 1, ..., 4$. Moreover, if there are many subpopulations at the same time instead of the 4 age groups here, one may also consider extending separate linear regression models to a single random-effects model, and develop tests under it. In reality though, the source effect is unlikely to be zero, as long as the proxy weights $\boldsymbol{w}_g^*$ are not exactly equal to the true $\boldsymbol{w}_g$, such that it is sensible to treat any non-rejection result as an intermediate step in the investigation.

4 Accuracy of proxy weights food index

To address Question III in Section 1, we will treat $\widehat{P}_{gt}$ as an unbiased audit sample estimator of P_{gt}. Let $\widehat{V}(\widehat{P}_{gt})$ be an unbiased estimator of the variance of $\widehat{P}_{gt}$. Given zero variance of P_{gt}^*, an unbiased estimator of the MSE of P_{gt}^* is

$$\text{mse}(P_{gt}^*) = (P_{gt}^* - \widehat{P}_{gt})^2 - \widehat{V}(\widehat{P}_{gt}) \ . \tag{3}$$

However, for the data of this study, the MSE estimate is negative in *all* the 36 months, due to the relatively large sampling variance of $\widehat{P}_{gt}$, as indicated by the test results based on $\widehat{\bar{P}}_g - \bar{P}^*_g$ reported earlier. This reveals an important drawback of using MSE as the uncertainty measure in the present setting: unless the audit sample is sufficiently large, indeed much larger than the actual CES, unbiased MSE estimation may fail to yield any meaningful accuracy measure, when P^*_{gt} has a relatively small bias.

Below we propose a new accuracy measure, discuss its properties, and demonstrate empirically the advantages of this novel proposal for big-data statistics.

4.1 Evaluation coverage

The concept of *evaluation coverage* as an accuracy measure is as follows. Let θ_0 be a target scalar parameter value. Let A_α be an imaginary autonomous *evaluation* confidence interval, which is centred at θ_0 and covers it with a probability α of choice, say $\alpha = 0.95$. Let θ^* be a constant value in the parameter space. In the application later, we will let $\theta_0 = P_{gt}$ be the true food index for age group g in month t, and $\theta^* = P^*_{gt}$ the corresponding proxy weights index, given any combination (g, t). To assess how good θ^* is as an estimator of θ_0, we now calculate the probability that θ^* is covered by A_α. We shall call this probability $c(\theta^*) = \Pr(\theta^* \in A_\alpha)$ the *evaluation coverage* of θ^* by A_α.

Notice that $c(\theta^*)$ does not depend on audit sampling; it is a measure with respect to the joint distribution of the estimator θ^* to be assessed and the hypothetical evaluation confidence interval A_α, where the distribution of A_α depends on the true parameter θ_0. In this respect, it is simply an alternative accuracy measure to $\mathrm{MSE}(\theta^*) = E[(\theta^* - \theta_0)^2]$, which is based on the distribution of θ^* and the true parameter θ_0.

Suppose $c(\theta_0) = 0.95$ and $c(\theta^*) = 0.93$, which means the confidence interval A_α that covers θ_0 in 95% of the times would cover the estimate θ^* in 93% of the times. In other words, if one treats θ^* as the true value, then the 95% confidence interval A_α would fail to cover it only 2% more often than when one correctly holds θ_0 to be the truth. As will be shown later, the evaluation coverage of a constant θ^* by A_α reaches its maximum value α if $\theta^* = \theta_0$, otherwise it decreases as $|\theta^* - \theta_0|$ increases. In this way, the evaluation coverage can provide a measure of the accuracy of θ^*, which achieves its maximum value when $\theta^* = \theta_0$, and decreases monotonously towards 0 as the bias of θ^* increases to infinity. This is appropriate when θ^* is based on big data, and it is associated with an unavoidable bias but only a negligible variance for all practical purposes.

Moreover, let $\tilde{\theta}$ be an unbiased estimator of θ_0 based on a probability sample of size n, such that its variance decreases as n increases. One can compare the big-data estimator θ^* and the finite-sample estimator $\tilde{\theta}$ according to their respective evaluation coverages, where

$c(\tilde{\theta}) = \Pr(\tilde{\theta} \in A_\alpha)$, and the probability is evaluated with respect to the joint distribution of $\tilde{\theta}$ and A_α. One may consider θ^* to be better than $\tilde{\theta}$ if $c(\theta^*) > c(\tilde{\theta})$, and vice versa. Indeed, the sample size n at which $c(\theta^*) = c(\tilde{\theta})$ can be considered the break-even point, which indicates the cost that is required if one opts to estimate θ_0 based on a designed probability sample, instead of based on the available big data.

Of course, one can make the comparison between θ^* and $\tilde{\theta}$ based on their respective MSEs. In practice, though, one needs to estimate the MSE, which can be difficult if it requires a very large (hence costly) probability audit sample, as noticed previously in connection with (3). In contrast, as we will explain and demonstrate below, one can estimate $c(\theta^*)$ based on an audit sample and obtain a meaningful comparison between θ^* and $\tilde{\theta}$, even when the same sample fails to yield a positive estimate of MSE(θ^*). This is an important advantage of evaluation coverage over MSE.

Definition Below we define the evaluation coverage and describe its properties formally. Let A_α be a $100\alpha\%$ autonomous normal *evaluation* confidence interval of θ_0, given by

$$A_\alpha = (Z_A - \omega,\ Z_A + \omega)\ , \tag{4}$$

which is of the width 2ω, where

$$Z_A \sim N(\theta_0, \sigma^2_{\alpha,\omega}) \qquad \text{and} \qquad \sigma_{\alpha,\omega} = \frac{\omega}{\kappa_\alpha} \qquad \text{and} \qquad \kappa_\alpha = \Phi^{-1}(\frac{1+\alpha}{2})$$

and Φ denotes the cumulative distribution function of the standard normal distribution. The interval A_α is said to be autonomous, because it is an imaginary confidence interval, independent of any actual observations available or collected to estimate θ_0. Let $\tilde{\theta}$ be an estimator of θ_0. The evaluation coverage of $\tilde{\theta}$ is the probability

$$c(\tilde{\theta}) = \Pr(\tilde{\theta} \in A_\alpha) \tag{5}$$

which is evaluated with respect to the joint distribution of $(\tilde{\theta}, Z_A)$, where the two are independent of each other by definition. Instead of the target parameter θ_0 itself, audit sampling aims ultimately is to estimate $c(\tilde{\theta})$, which is an accuracy measure of $\tilde{\theta}$.

We have $c(\theta_0) = \alpha$ by definition, i.e., the evaluation coverage of the true parameter value θ_0 is α. Otherwise, the evaluation coverage of a given estimator $\tilde{\theta}$ generally varies with the choice of ω, which determines the width of A_α. The choice of ω affects the stringency of evaluation. For instance, in the context of price index, if one chooses $\omega = 0.01$, then the interval A_α will correspond to an evaluation precision of $\pm 1\%$ on either side of the true index. As will be illustrated in Section 4.2, one can also relate the choice to the sampling

variance of the CES, in which case the accuracy of the proxy weights will be measured against a 'yardstick' that can be related to the tangible precision of the CES.

Some properties of the evaluation coverage as an accuracy measure are given below as Results 1 - 3, the proofs of which are given in Appendix A.

Result 1 For any constant θ^* in the parameter space, we have $0 < c(\theta^*) \leq \alpha$, where the maximum is attained iff $\theta^* = \theta_0$.

Indeed, the further away from θ_0 is a zero-variance (big-data) point estimate, the lower is its evaluation coverage, for any choice of (α, ω). For any two $\theta^* \neq \theta'$ in the parameter space, if $|\theta' - \theta_0| > |\theta^* - \theta_0|$, then $c(\theta') < c(\theta^*)$.

Result 2 If $\tilde{\theta}^* \sim N(\theta^*, \tau^2)$ where $\theta^* \in (\theta_0 - \omega, \theta_0 + \omega)$, then $c(\tilde{\theta}^*) < c(\theta^*)$.

In other words, extra variance reduces the evaluation coverage, as long as the absolute bias is less than ω, i.e., θ^* is not too far away from θ_0. Moreover, if $\theta^* \in (\theta_0 - \omega, \theta_0 + \omega)$, $\tilde{\theta}_1^* \sim N(\theta^*, \tau_1^2)$ and $\tilde{\theta}_2^* \sim N(\theta^*, \tau_2^2)$ where $\tau_1^2 < \tau_2^2$, then $c(\tilde{\theta}_1^*) > c(\tilde{\theta}_2^*)$. However, extra variance could increase the evaluation coverage if the bias is sufficiently large. To see why, let θ^* be so far away from θ_0 that its evaluation coverage is virtually zero, then an estimator that is centred on θ^* but has a large variance can actually be much closer to θ_0 from time to time, which increases its chance of being covered by A_α. In contrast, using MSE as the criterion, the estimator with variance would always be worse due to the extra variance. This is an example of the difference between the two accuracy measures.

Result 3 If $\tilde{\theta}^* \sim N(\theta^*, \tau^2)$ and $\tilde{\theta}' \sim N(\theta', \tau^2)$, then $c(\tilde{\theta}') < c(\tilde{\theta}^*)$ if $|\theta' - \theta_0| > |\theta^* - \theta_0|$.

Result 3 is complementary to Result 2, that is, given two estimators with the same variance, the one with less absolute bias has a higher evaluation coverage.

Estimation The evaluation coverage can communicate in an intuitive and meaningful way the accuracy of any estimator, including a zero-variance big-data estimate. To estimate the evaluation coverage of θ^* that is a constant in the parameter space, we observe firstly

$$c(\theta^*) = \Phi\big(\frac{\theta^* - \theta_0}{\sigma_{\alpha,\omega}} + \kappa_\alpha\big) - \Phi\big(\frac{\theta^* - \theta_0}{\sigma_{\alpha,\omega}} - \kappa_\alpha\big) \tag{6}$$

by (5), where only θ_0 is unknown. Given an unbiased estimate $\hat{\theta}_0$, we obtain

$$\hat{c}(\theta^*) = \Phi\big(\frac{\theta^* - \hat{\theta}_0}{\sigma_{\alpha,\omega}} + \kappa_\alpha\big) - \Phi\big(\frac{\theta^* - \hat{\theta}_0}{\sigma_{\alpha,\omega}} - \kappa_\alpha\big) \ .$$

It is thus clear that one only needs the point estimate $\hat{\theta}_0$ derived from an audit sample to calculate $\hat{c}(\theta^*)$. The audit sampling variance of $\hat{\theta}_0$ affects only the variance of $\hat{c}(\theta^*)$. This means the estimation of $c(\theta^*)$ is not as critically constrained by the audit sampling variance as when estimating MSE(θ^*).

Meanwhile, let $\tilde{\theta}$ be an estimator (of θ_0) with expectation θ' and variance τ^2. Due to the independence between $\tilde{\theta}$ and Z_A by definition, where Z_A is given in (4), we have

$$\tilde{\theta} - Z_A \sim N(\theta' - \theta_0, \nu^2) \qquad \text{where} \quad \nu^2 = \sigma^2_{\alpha,\omega} + \tau^2 \ ,$$

as long as $\tilde{\theta}$ can be assumed to have the normal distribution $N(\theta', \tau^2)$. It follows that the evaluation coverage of $\tilde{\theta}$ by (5) is given as

$$c(\tilde{\theta}) = \Phi(\frac{\omega}{\nu} + \frac{\theta' - \theta_0}{\nu}) - \Phi(-\frac{\omega}{\nu} + \frac{\theta' - \theta_0}{\nu}) \ . \tag{7}$$

In cases $\tilde{\theta}$ is a hypothetical unbiased estimator, one only needs to stipulate its variance τ^2, in order to calculate $c(\tilde{\theta})$. In cases $\tilde{\theta}$ is an existing estimator that can be considered as unbiased for θ_0, one only needs to estimate τ (and ν), in order to estimate $c(\tilde{\theta})$. In either case, audit sampling is not needed to estimate $c(\tilde{\theta})$. In cases where it is too optimistic to assume unbiased $\tilde{\theta}$, e.g. when it is based on an estimated working model instead of some known distribution, one may use the realised value of $\tilde{\theta}$ as θ' in (7), instead of ignoring its potential bias, which would yield a lower estimated value of $c(\tilde{\theta})$ by Result 3 above. Audit sampling is needed to obtain $\hat{\theta}_0$ in such situations.

4.2 Application to age-group food index

Let us now estimate the evaluation coverage of the age-group food price indices based on proxy weights. We have one set of proxy weights $\boldsymbol{w}^*_g$ and one set of CES-based weights $\hat{\boldsymbol{w}}_g$, for each $g = 1, ..., 4$. Applying these to $\boldsymbol{p}_t$ for any given month t, we obtain P^*_{gt} and $\widehat{P}_{gt}$, respectively. Let $\theta_0 = P_{gt}$ and $\theta^* = P^*_{gt}$. The evaluation coverage of P^*_{gt} is given by (6),

$$c^* := c(P^*_{gt}) = \Phi\big(\frac{P^*_{gt} - P_{gt}}{\sigma_{\alpha,\omega}} + \kappa_\alpha\big) - \Phi\big(\frac{P^*_{gt} - P_{gt}}{\sigma_{\alpha,\omega}} - \kappa_\alpha\big) \ .$$

Using the CES as an audit sample, we obtain $\hat{\theta}_0 = \widehat{P}_{gt}$ and the estimator

$$\hat{c}^* = \Phi\big(\frac{P^*_{gt} - \widehat{P}_{gt}}{\sigma_{\alpha,\omega}} + \kappa_\alpha\big) - \Phi\big(\frac{P^*_{gt} - \widehat{P}_{gt}}{\sigma_{\alpha,\omega}} - \kappa_\alpha\big) \ .$$

The variance of $\hat{c}^*$ can be approximated as follows:

$$\hat{c}^* \approx c^* + \sigma_{\alpha,\omega}^{-1}\Big(\phi\big(\frac{P_{gt}^* - P_{gt}}{\sigma_{\alpha,\omega}} + \kappa_\alpha\big) - \phi\big(\frac{P_{gt}^* - P_{gt}}{\sigma_{\alpha,\omega}} - \kappa_\alpha\big)\Big)(\widehat{P}_{gt} - P_{gt})$$

$$V(\hat{c}^*) \approx \frac{V(\widehat{P}_{gt})}{\sigma_{\alpha,\omega}^2}\Big(\phi\big(\frac{P_{gt}^* - P_{gt}}{\sigma_{\alpha,\omega}} + \kappa_\alpha\big) - \phi\big(\frac{P_{gt}^* - P_{gt}}{\sigma_{\alpha,\omega}} - \kappa_\alpha\big)\Big)^2 \tag{8}$$

where $V(\widehat{P}_{gt}) = \boldsymbol{p}_t V(\hat{\boldsymbol{w}}_g)\boldsymbol{p}_t^\top$, and ϕ denotes the probability density function of the standard normal distribution. Only the audit-sample point estimate $\hat{\theta}_0 = \widehat{P}_{gt}$ is needed to estimate $c(P_{gt}^*)$, whereas its variance $V(\widehat{P}_{gt})$ affects only the variance of $\hat{c}(P_{gt}^*)$. Based on the CES, the estimate of $\mathrm{MSE}(P_{gt}^*)$ by (3) is negative and unusable for any of the 36 months, whereas the same audit sample is able to provide meaningful estimates when evaluation coverage $c(P_{gt}^*)$ is adopted as the accuracy measure of P_{gt}^*.

Fixing α at 0.95, one can vary the stringency of evaluation in terms of ω. The average estimated standard error of the CES-based index $\widehat{P}_{4t}$ is 0.029 over the 36 months, for age group $g = 4$. As the first choice we set $\omega = 2 \cdot 0.029$, in which case $A_{0.95}$ can be considered as a 95% confidence interval of the true index P_{4t}, which has the same width as that based on the traditional CES. The corresponding estimated $\hat{c}(P_{4t}^*)$ tells us how often the proxy weights index P_{4t}^* is covered by this $A_{0.95}$. For comparison, we calculate by (7) the evaluation coverage of a hypothetical unbiased index $\widetilde{P}_{4t}$ that is of the same precision as the CES-based index. The results are shown in the top panel of Figure 3 and summarised in Table 6. The estimated evaluation coverages of P_{gt}^* and $\widetilde{P}_{gt}$ for the other age groups are also summarised in the upper part of Table 6.

Take first age group $g = 4$. With $\omega = 0.058$, the evaluation coverage of P_{4t}^* varies from 0.948 to 0.950 over the 36 months, which is very close to $c(P_{4t}) = 0.95$ for the true index. It compares favourably to the hypothetical unbiased index $\widetilde{P}_{4t}$, whose evaluation coverage varies from 0.828 to 0.849. The confidence interval of $c(P_{4t}^*)$ are marked by vertical lines in Figure 3, which is seen to fluctuate in width before reaching a different level towards to the end. Still, the evaluation coverage of $\widetilde{P}_{4t}$ is out of these confidence intervals throughout the whole period. The results suggest that the bias of P_{4t}^* is small enough for it to outperform the hypothetical unbiased index $\widetilde{P}_{4t}$. Indeed, the standard error of $\widetilde{P}_{4t}$ needs to be reduced to about 1/9 of that stipulated here, in order to achieve about the same median and mean evaluation coverages as P_{4t}^* over the 36 months. But such an increase of the sample size is unthinkable in reality due to the associated cost.

The results are similarly in favour of the proxy weights index P_{gt}^* for the other age groups, $g = 1, 2, 3$. The hypothetical unbiased index has the lowest evaluation coverage for age group $g = 1$, due to the small subsample size of this age group in the CES.

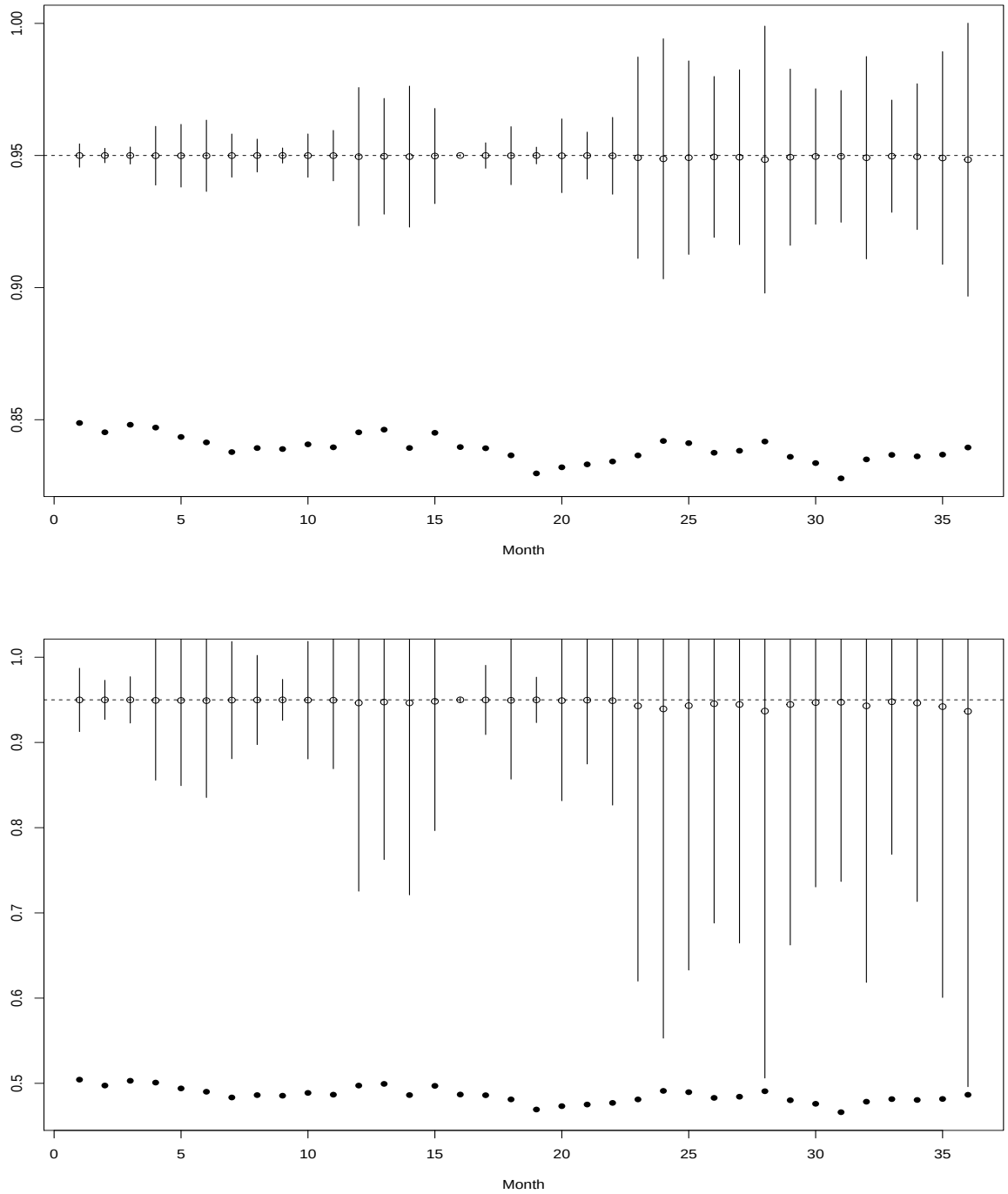

Figure 3: Evaluation coverage by $A_{0.95}$ of food index for age group $g = 4$ over 36 months: $\hat{c}(P_{4t}^*)$ of proxy weights index (circle), 95% confidence interval of $c(P_{4t}^*)$ marked by vertical line, $c(\widetilde{P}_{4t})$ of hypothetical unbiased index of same precision as CES-based index (solid), $c(P_{4t}) \equiv 0.95$ of true index (horizontal dashed line). Top: $\omega = 0.058$; bottom: $\omega = 0.02$.

Table 6: Evaluation coverage by $A_{0.95}$ of food index for age group $g = 1, ..., 4$ over 36 months: proxy weights index P^*_{gt} or hypothetical unbiased $\widetilde{P}_{gt}$ of same precision as CES.

	$\omega = 0.058$							
Summary	P^*_{1t}	$\widetilde{P}_{1t}$	P^*_{2t}	$\widetilde{P}_{2t}$	P^*_{3t}	$\widetilde{P}_{3t}$	P^*_{4t}	$\widetilde{P}_{4t}$
Minimum	0.948	0.563	0.949	0.900	0.949	0.925	0.948	0.828
Median	0.950	0.579	0.950	0.904	0.950	0.927	0.950	0.839
Mean	0.950	0.580	0.950	0.904	0.950	0.927	0.950	0.839
Maximum	0.950	0.597	0.950	0.908	0.950	0.929	0.950	0.849
	$\omega = 0.02$							
Summary	P^*_{1t}	$\widetilde{P}_{1t}$	P^*_{2t}	$\widetilde{P}_{2t}$	P^*_{3t}	$\widetilde{P}_{3t}$	P^*_{4t}	$\widetilde{P}_{4t}$
Minimum	0.937	0.227	0.945	0.643	0.944	0.759	0.937	0.466
Median	0.949	0.236	0.949	0.658	0.949	0.772	0.949	0.486
Mean	0.949	0.237	0.949	0.659	0.949	0.773	0.947	0.486
Maximum	0.950	0.247	0.950	0.674	0.950	0.785	0.950	0.504

Next, the results are given in Figure 3 and Table 6, where ω is reduced to 0.02, i.e., the width of $A_{0.95}$ is reduced to about 1/3 of that used above. The same bias of P^*_{4t} leads then to a lower evaluation coverage for the same level $\alpha = 0.95$, which varies now from 0.937 to 0.950 over the 36 months. However, the reduction of evaluation coverage is much greater for the hypothetical unbiased $\widetilde{P}_{4t}$, which varies from 0.466 to 0.504. In other words, subjected to the increased stringency of evaluation, the proxy index P^*_{4t} compares even more favourably to the CES-based index. The results are similarly in favour of the proxy weights index P^*_{gt} for the other age groups $g = 1, 2, 3$, as ω is reduced.

Finally, Figure 3 shows that the width of the confidence interval of $c(P^*_{4t})$ increases quite fast as ω is reduced from 0.058 to 0.02. The reason is clear from (8). While the variance $V(\widehat{P}_{4t})$ due to audit sampling remains the same, the other two terms are affected by the change in ω: in the denominator $\sigma_{\alpha,\omega} = \omega/\kappa_\alpha$ is reduced proportionally with ω; in the numerator the squared $\phi\big(\frac{P^*_{4t}-\widehat{P}_{4t}}{\sigma_{\alpha,\omega}} + \kappa_\alpha\big) - \phi\big(\frac{P^*_{4t}-\widehat{P}_{4t}}{\sigma_{\alpha,\omega}} - \kappa_\alpha\big)$ is increased, because the reduction of $\sigma_{\alpha,\omega}$ increases the asymmetry of $\frac{P^*_{4t}-\widehat{P}_{4t}}{\sigma_{\alpha,\omega}} \pm \kappa_\alpha$ around 0. The two effects amplify each other, increasing $\sqrt{V(\hat{c}^*)}$ more quickly than a rate proportional to 0.058/0.02.

The audit sample size does affect the variance of the estimator of evaluation coverage. To improve its precision, one may need to use a larger audit sample, although in the context of CPI it will be hard to obtain approval for an audit sample size that is considerably larger than the traditional CES. Meanwhile, it seems more reasonable to place relatively higher confidence in the comparison results between the proxy weights index P^*_{gt} and the hypothetical index $\widetilde{P}_{gt}$, where one has ignored the potential bias of the latter. In any case, this is clearly an important issue for future research and application.

5 Summary remarks

We have developed an audit sampling inference approach for big-data statistics, which is privacy-preserving as it does not require linking the audit sample and big data at the individual level. It consists of three elements: (I) clarification of the validity condition by which a big-data estimator can be unbiased, (II) testing of the bias in case the validity condition does not hold exactly, and (III) accuracy measure of a biased big-data estimator. The inference is valid with respect to the auditing sampling distribution, regardless of the method and condition by which the big-data estimator is produced. The proposed approach is applied to age-group food price indices. Based on the data of this study, one can conclude that proxy CPI weights derived from the transaction data can replace the relevant diary component that is the most burdensome part of the traditional CES.

A difficult challenge arises, when the audit sampling variance is relatively large compared to the bias of big-data estimate, which results in non-rejection of the bias as well as negative (hence unusable) MSE estimate. We develop the evaluation coverage as a novel accuracy measure, which is not severely constrained by the audit sample size. This provides the means to assess the big-data bias against the costs of alternative and possibly unbiased estimation methods. The evaluation coverage is flexible to apply, where one can either use an existing survey on the same topic or, if such a survey does not exist, undertake separate agile audit sampling with a relatively small sample size and low cost.

Acknowledgement I would like to thank two anonymous referees and the Joint Editor for constructive suggestions, which helped to improve the presentation.

A Proof of Result 1, 2 and 3

Let θ_0 be the true scalar parameter value. Let $Z \sim N(\theta_0, \sigma^2)$ be a normally distributed random variable. The shortest $100\alpha\%$ confidence interval of θ_0 is $A_{\alpha,\omega} = Z \pm \omega$ with $\omega = \kappa_\alpha \sigma$, where σ is a short-hand for $\sigma_{\alpha,\omega}$, and $c(\theta_0) = \Pr(\theta_0 \in A_{\alpha,\omega}) = \alpha$, and κ_α is the $(1+\alpha)/2$ quantile of $N(0,1)$. Results 1 - 3 are given Proof-1 to 3 below, respectively.

Proof-1 Let Φ be the CDF of $N(0,1)$. Write $c^* = \alpha(\theta^*)$. We have

$$c^* = \Pr\big(\theta^* - \kappa_\alpha\sigma \le Z \le \theta^* + \kappa_\alpha\sigma\big) = \Pr\big(\frac{\theta^* - \theta_0}{\sigma} - \kappa_\alpha \le \frac{Z - \theta_0}{\sigma} \le \frac{\theta^* - \theta_0}{\sigma} + \kappa_\alpha\big)$$

$$c^* - \alpha = \Phi(\kappa_\alpha + \frac{\theta^* - \theta_0}{\sigma}) - \Phi(-\kappa_\alpha + \frac{\theta^* - \theta_0}{\sigma}) - \Phi(\kappa_\alpha) + \Phi(-\kappa_\alpha)$$

$$= \begin{cases} \big(\Phi(\kappa_\alpha + \frac{\theta^*-\theta_0}{\sigma}) - \Phi(\kappa_\alpha)\big) - \big(\Phi(-\kappa_\alpha + \frac{\theta^*-\theta_0}{\sigma}) - \Phi(-\kappa_\alpha)\big) < 0 & \text{if } \theta^* > \theta_0 \\ 0 & \text{if } \theta^* = \theta_0 \ \square \\ -\big(\Phi(\kappa_\alpha) - \Phi(\kappa_\alpha + \frac{\theta^*-\theta_0}{\sigma})\big) + \big(\Phi(-\kappa_\alpha) - \Phi(-\kappa_\alpha + \frac{\theta^*-\theta_0}{\sigma})\big) < 0 & \text{if } \theta^* < \theta_0 \end{cases}$$

Proof-2 Let $\nu^2 = \sigma^2 + \tau^2$. Without loss of generality, suppose $\theta^* \in [\theta_0, \theta_0 + \kappa_\alpha\sigma)$. We have

$$\begin{aligned} c(\tilde{\theta}^*) &= \Pr\big(-\kappa_\alpha\sigma \le Z - \tilde{\theta}^* \le \kappa_\alpha\sigma\big) \\ &= \Pr\left(-\frac{\kappa_\alpha\sigma}{\nu} + \frac{\theta^* - \theta_0}{\nu} \le \frac{Z - \tilde{\theta}^*}{\nu} + \frac{\theta^* - \theta_0}{\nu} \le \frac{\kappa_\alpha\sigma}{\nu} + \frac{\theta^* - \theta_0}{\nu}\right) \\ &= \Phi(\frac{\kappa_\alpha\sigma}{\nu} + \frac{\theta^* - \theta_0}{\nu}) - \Phi(-\frac{\kappa_\alpha\sigma}{\nu} + \frac{\theta^* - \theta_0}{\nu}) \\ &< \Phi(\kappa_\alpha + \frac{\theta^* - \theta_0}{\sigma}) - \Phi(-\kappa_\alpha + \frac{\theta^* - \theta_0}{\sigma}) = c(\theta^*) \end{aligned}$$

since $\kappa_\alpha\sigma + (\theta^* - \theta_0) \ge 0$ and $-\kappa_\alpha\sigma + (\theta^* - \theta_0) \le 0$. $\square$

Proof-3 Let $\nu^2 = \sigma^2 + \tau^2$. We have

$$\begin{aligned} c(\tilde{\theta}') - c(\tilde{\theta}^*) = &\big[\Phi(\frac{\kappa_\alpha\sigma}{\nu} + \frac{\theta' - \theta_0}{\nu}) - \Phi(\frac{\kappa_\alpha\sigma}{\nu} + \frac{\theta^* - \theta_0}{\nu})\big] \\ &- \big[\Phi(-\frac{\kappa_\alpha\sigma}{\nu} + \frac{\theta' - \theta_0}{\nu}) - \Phi(-\frac{\kappa_\alpha\sigma}{\nu} + \frac{\theta^* - \theta_0}{\nu})\big] \end{aligned}$$

Due to symmetry $c(\tilde{\theta}^*) = c\big(\tilde{\theta}^* + (2\theta_0 - \theta^*)\big)$, we only need to consider the situation of $\theta' > \theta^* > \theta_0$. Then, the interval $(-\frac{\kappa_\alpha\sigma}{\nu} + \frac{\theta^*-\theta_0}{\nu}, -\frac{\kappa_\alpha\sigma}{\nu} + \frac{\theta'-\theta_0}{\nu})$ is closer to 0 than $(\frac{\kappa_\alpha\sigma}{\nu} + \frac{\theta^*-\theta_0}{\nu}, \frac{\kappa_\alpha\sigma}{\nu} + \frac{\theta'-\theta_0}{\nu})$, where the two have the same length. The result follows from

$$\Phi(-\frac{\kappa_\alpha\sigma}{\nu} + \frac{\theta' - \theta_0}{\nu}) - \Phi(-\frac{\kappa_\alpha\sigma}{\nu} + \frac{\theta^* - \theta_0}{\nu}) > \Phi(\frac{\kappa_\alpha\sigma}{\nu} + \frac{\theta' - \theta_0}{\nu}) - \Phi(\frac{\kappa_\alpha\sigma}{\nu} + \frac{\theta^* - \theta_0}{\nu}) > 0. \square$$

References

[1] Battistin, E. and M. Padula (2016). Survey instruments and the reports of consumption expenditures: evidence from the consumer expenditure surveys.

Journal of the Royal Statistical Society, Series A, **179**, 559-581.

[2] Bee, A., Meyer, B. D. and Sullivan, J. X. (2014) The validity of consumption data: are the consumer expenditure interview and diary surveys informative? In *Improving the Measurement of Consumer Expenditures (eds. C. Carroll, T. Crossley and J. Sabelhaus).* Chicago: University of Chicago Press.

[3] Fricker, S., Kopp, B., Tan, L. and R. Tourangeau (2015). A review of measurement error assessment in a U.S. household expenditure survey. *Journal of Survey Statistics and Methodology*, **3**, 67-88.

[4] Neter, J. (2011). Statistics in Auditing. In *Encyclopedia of Statistical Sciences, eds. S. Kotz, C.B. Read, N. Balakrishnan, B. Vidakovic and N.L. Johnson.* John Wiley & Sons, Inc. `https://doi.org/1.1002/0471667196.ess2553.pub3`

[5] 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.

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

[7] Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference.* New York: John Wiley.

[8] Zhang, L.-C. (2019). On valid descriptive inference from non-probability sample. *Statistical Theory and Related Fields*, `DOI:1.1080/24754269.2019.1666241`