Estimating Sampling Errors in Consumer Price Indices

Paul A. Smith

S3RI and Dept. of Social Statistics & Demography, University of Southampton, Southampton, UK
E-mail: p.a.smith@soton.ac.uk

Summary

Several different approaches have been proposed for the calculation of sampling errors in consumer price indices (CPIs), but only one is in regular use providing published estimates. The history of the development of sampling error estimation for CPIs is reviewed, and the development in the USA (where most research has taken place) is summarised. The proposed approaches for sampling error estimation are set out, with their properties and their different conceptualisations of the sampling error. The frameworks for the measurement of errors more widely in the CPI are also reviewed, and the evidence for the importance of the sampling error among the other error sources is presented. Some opportunities and directions for further work are provided.

Key words: Consumer price index; mean squared error; variance components; replicate variance estimation; jackknife; bootstrap; model-based variance estimates.

1 Introduction

The consumer price index (CPI) is an important economic indicator with a wide use, which encompasses macroeconomic policymaking to uprating of costs and benefits. It is a widely scrutinised statistic, not least because of the effect it has directly on the lives and budgets of individuals. Although the origins of the CPI depended on the foresight and interest of some key people, and on others for further development, the production of a national CPI has generally needed the resources of the state (O’Neill et al., 2017; Smith & Ralph, 2021). In the UK, a CPI was introduced (under a different name) with the First World War, but it was not until after the Second World War that a well-designed system was instituted, including regular collections for prices, regular updating of the basket and regular surveys as a source of weights (Ralph et al., 2020, Chapter 3).

Consumer price indices generally have a complex structure for data collection because of the different components, most of which have been collected through surveys (although recent opportunities to use scanner data and web-scraped price information may potentially change this). This has made it a significant challenge to produce estimates of sampling variances for CPIs, although the need for them has been acknowledged for a long time. The focus of this review is the published CPI, although the methods discussed can be extended to other, related indices. In recent years, the weights for some CPIs have not come directly from a survey, but from a survey combined with other sources through a balancing process within the national accounts. This presents further challenges for assessing variances.
The sampling structure for prices does not have common terminology in different countries. Here, we will refer to the highest level of sampling as areas (in the USA, these are also described as PSUs and cities, as they are the basis of city indices; in the UK, they are locations; and in Italy, they are municipalities). The largest areas are included with certainty, and these are called self-representing areas. Within areas, prices are collected from outlets (also called establishments, stores or shops). Commodities are classified (possibly in more than one hierarchical layer) into strata called groups (item strata) from which representative items are selected (these are called entry-level items in the USA). A price collector collects the price for a product (called an item in the USA).

1.1 Historical Development of Variance Estimation for Price Indexes

Jevons (1869) had already considered the error in deriving indices from the prices of a small number of commodities and produced an estimate of the error in the change in his index of the value of gold (measured by the change in wholesale1 prices). Edgeworth (1888) continued this line of thought as part of a series of reports for the British Academy for the Advancement of Science, and he already concluded (p. 320) that one should ‘Take more care about the prices than the weights’. Indeed, his investigations were quite detailed, and it seems right to follow Balk (1987) and quote a little more detail of his conclusions (with Balk’s highlighting in italics; the bold emphasis is Edgeworth’s) by way of setting up the remainder of this review:

“The error [of a price index] is found to depend in a definite manner upon six distinct circumstances. The erroneousness of the result is greater, the greater the inaccuracy of the data: the weights and the (comparative) prices. The erroneousness of the result is also greater, the greater the inequality of the weights, and the greater the inequality of the price returns. Lastly, the result is more accurate, the greater the number of the data and the smaller the number of omitted articles. These circumstances are not all equally operative. Other things being the same, the inaccuracy of the price returns affects the result more than inaccuracy of the weights; and the inequality of the price-returns more than the inequality of the weights’. (Edgeworth, 1888, pp. 316–317)”

Notwithstanding these early calculations of the variances of wholesale price indices, when the state took over collection and extended the range of price quotes available, the designs became considerably more complicated, and implementing a regular index calculation was the main priority. Therefore, there was a period when there was little development in the production of quality measures for national price indices. Nonetheless, it was not long before Bowley (1926, 1928) published further, considering the effects of correlation between price changes on the sampling error of an index, and the error components more widely. Bowley calculated sampling errors with autocorrelation on the index published by the Statist, for which all the prices (47 quotes) were published and which provided a tractable model system. This approach does not seem to have been used on the cost of living index produced by the Board of Trade in the UK, which was based on many more prices. Bowley also considered errors in index numbers more widely, and we return to this topic in Section 5.

There was renewed interest in the calculation of sampling variances at least from Mudgett (1951, Chapter 6; see also Wilkerson, 1967, section 2). One of the challenges is that price indices are derived from multiple sources, including price collection surveys, centralised price collections, household budget surveys and administrative data. The first attempts at estimating components of variance due to specific parts of the sampling were in Sweden in 1953 and 1958 (see Dalén & Ohlsson, 1995), and these followed work in the late 1940s on non-sampling error in the Swedish CPI by von Hofsten (see the reanalysis in McCarthy, 1961,
section V). There were also forays into aspects of the variance of price indices by several authors—Banerjee (1956, 1959, 1960) and Adelman (1958).

However, the pressure to develop sampling error estimates was greatest in the USA, where Adelman (1958) evaluated the cost of a suitable system, and Kruskal & Telser (1960) and McCarthy (1961) were openly critical of the lack of quality measures. McCarthy (section VII) developed an outline replication design, which would allow variances and components of variance to be estimated, and this led to a redesign of the US CPI with the introduction of replicate samples from December 1963. These facilitated the construction of sampling variances, which accounted for most of the components of sampling. The replicate sample areas are used for both price collection and the budget survey (Leaver & Valliant, 1995 p. 554), which means that the effect of variance in both parts can be included in the same half-sample estimates. This allowed estimates of variances to be made for changes in the index from 1964 onwards (Wilkerson, 1967), and the USA was therefore the first country to produce good evidence for the overall sampling variability of its CPI. The Bureau of Labour Statistics (BLS) has generally continued this approach to calculating sampling variances since then, with some developments, particularly an update that added probability sampling for all the components of the CPI from 1978 (BLS, 2015, Chapter 17) and some occasional experimentation to investigate the properties of alternative methods. As well as using the replicate samples for calculating the variance of the index, BLS has also used them to estimate variance components—which part of the sampling has the highest variance—as an aid to optimising the CPI sample (Baskin & Johnson, 1995 extended using scanner data by Leaver & Larson, 2003). A more detailed history of the progress of variance estimation work in the USA is given in Section 4, as it will be more comprehensible after the different variance estimators have been introduced.

The UK undertook some initial studies in support of a review of the possibility of producing regional price indices, where the effect of the sampling error might be more pronounced, initially in Department of Employment (1971, Appendix 1), where the effects of sampling errors in the prices and the weights were considered separately. This work was later extended by Fowler (1973), particularly with regard to the effect of sampling error in the weights. More details of the developments in the UK are given in Smith & O'Donoghue (in prep).

There was a general increase in interest in the measurement of the quality of CPIs during the later 1980s, with developments taking place in the Netherlands, Italy and Sweden. This led to several sessions at the biennial meetings of the International Statistical Institute.

The Dutch undertook some work to estimates the variance of the CPI, mainly between 1985 and 1991, using initially balanced half sample methods (Balk & Kersten, 1986; Balk, 1989), and later moving to a design-based approach using Taylor linearisation (Balk, 1991). Unfortunately, the details of these latter derivations are unpublished.

At the same time, Biggeri & Giommi (1987) undertook a review of strategies for estimating the sampling error of a CPI, coupled with some initial work on the variability due to estimating household expenditure weights in the Italian CPI. They suggested focusing on a sampling structure where the household budget survey (providing the weights) and the price collection were undertaken in the same sample areas, but were unable to match up areas at this level, and had to start with higher-level aggregates. Later, D’Alò et al. (2006) undertook further work in Italy to estimate the variance components [from sampling areas (municipalities), outlets, items and residual] for two groups of commodities; this allowed model variances for these groups to be produced, but they were not at that stage combined into an overall model variance.

The beginnings of work in Sweden also coincided with these developments elsewhere, with Andersson et al. (1987b) undertaking some preliminary studies on parts of the error. But this was followed by further development work to estimate sampling errors, where Dalén & Ohlsson (1995) used the cross-classified sampling of outlets and items as the basis for a design-
Based estimator of the variance attributable to price sampling. They claim this as the first calculation of sampling errors from a truly design-based estimator of the sampling variance, although Balk (1991) also seems to have taken a design-based approach (see the next section for more discussion of design-based and model-based approaches to variance estimation). Norberg (2004) evaluated this estimator in a simulation study and found that it was effective as long as there were indeed strong outlet and item effects, but that if these effects were weak, the variance was overestimated. A similar procedure was used in Finland (Jacobsen, 1997 in ILO et al., 2004, 5.99). Dalén (1995) also considered other error components for the Swedish CPI in a complete framework.

In the UK, further work was undertaken as part of a review of the Retail Prices Index in the 1990s, with unpublished reviews of variance component estimation in support of sample allocation in 1995 and of variance estimation in 1998 (for more details, see Smith & O’Donoghue, in prep).

In France, Ardilly & Guglielmetti (1992) derived variances accounting for the multiple stages of aggregation in prices in the estimation of the CPI and based on a two-stage design for the price collection. As in most applications, they needed to make some simplifying assumptions about the sampling and explored different approaches to imputing the variances where there were insufficient prices to estimate them directly. They commented that the weights might be important but did not take account of the variation in the weights in variance estimates (except where they were temporarily zero because no prices were collected).

In Luxembourg, some variances were calculated in 1998, based on a model that allowed for some correlation between price changes in the same establishment and based on several simplifying assumptions about the sampling (see ILO et al., 2004, 5.94–5.97, for a sketch of the approach and Dalén & Muelteel, 1998, in ILO et al., 2004, for details).

In parallel with these developments in Europe, there was continued refinement of existing methods and development of alternative approaches in the USA; these are documented in Section 4.

In Italy, the sample design was due to be made fully probabilistic from 2006, to support variance estimation. The proposed design contained some interesting features, with municipalities selected by balanced sampling and outlet samples for different item groups selected with positive coordination to maximise the outlet sample overlap (Biggeri & Falorsi, 2006; D’Alò et al., 2006). This proposal was not actually implemented, however, and the Italian CPI continues with a largely non-probability design.

The final entry in developments of variance estimates for a CPI belongs to Norway, where Zhang (2010) developed a model-based procedure, using similar estimators to Kott (1984) and Valliant (1991) for Lowe (Laspeyres-type) indexes and also extending these to Paasche-type and, perhaps more interestingly, to elementary aggregate models corresponding to the Carli, Jevons and Dutot indices. The corresponding estimates are regularly calculated but remain unpublished.

1.2 Overview

The main (published) developments in calculating variances and errors for CPIs have been made in rather few countries—the USA, Sweden, Italy, the Netherlands, France, Luxembourg, and the UK. There are also applications of these approaches in Finland, an example from Brazil (Fava, 2007), which uses an estimator of the variance of the arithmetic mean of prices to approximate the variance of a geometric mean, and one from Iraq (Fatah & Ahmed, 2012) apparently based on the US approach. Several papers originated from India in the 1950s and 1960s and/or
were published in Sankhyā, but this does not seem to have been translated into an estimate of the variance of the CPI in India.

Andersson et al. (1987b) say that 'the Swedish CPI has not been regarded, by its users, as an estimate of an unknown parameter. Rather, there seems to be a widespread agreement that the published value of the CPI is, by definition, the truth'. This seems to be a general situation for CPIs, particularly because there is no single optimum choice of aggregation formula(e) to use. In most cases where error calculations have been made, there are therefore challenges in explaining exactly what components of the quality and (for variances) what type of variance are being measured.

Some further work has been performed on the sample size and allocation requirements for price indices, and this has often involved some estimation of the components of variance due to different parts of the design (e.g. Baskin 1992, 1993; Baskin & Johnson 1995; D’Alò et al., 2006). This has sometimes resulted in estimates of sampling variances or components thereof as a by-product of the main objective.

The remainder of the paper consists of a discussion about design-based and model-based approaches to variance estimation for price indices in Section 2. Section 3 gives an overview of the methodology for different approaches to calculating sampling errors (and most of these approaches can be applied to deal with variability in either the prices or the weights). Section 4 provides a more detailed history of the development of variance estimates for the CPI in the USA, where much of the research into different approaches has taken place. Section 5 considers the range of types of errors that affect price indices and the frameworks and approaches that have been suggested to combine them. Here, there is also consideration of how to combine variance estimates due to different elements of the CPI sampling processes. Section 6 gives a discussion of the effectiveness of different approaches and highlights some areas where further research is needed to support quality measurement in national CPIs.

2 Design-based and Model-based Variance Estimation

The tension between model-based and design-based approaches to thinking in survey statistics has lessened as it has been realised that some problems can only be approached through the use of models, but the different approaches have both been used in the calculation of errors in the CPI. The distinction affects, at least to some degree, how the resulting statistics can be interpreted. The strict definition of a (design-based) sampling error is the difference between the values obtained if the whole population could be sampled and the values obtained through a sampling process. Von Hofsten (1959) argues that the calculation of a sampling error for a price index is essentially impossible, because various parts of the sampling cannot be sensibly defined. This has not prevented multiple approaches to calculating such errors, and McCarthy (1961) makes a strong case for the existence and value of the sampling error concept for price indices. But there is certainly an element of truth that a range of assumptions and simplifications must be made to construct a suitable process for estimating the sampling error of a price index, and we document these in Section 3. Von Hofsten’s line of thinking leads naturally to the model-based approach, and we return to this in the succeeding text.

A design-based approach does not in fact need the whole population to be specified, but does need probabilistic selection at each stage, and the selection probabilities. The calculations do not require information on the distributions of the values (of the prices or the weights) in the population, although in practice, long-tailed distributions containing outlying values may have an important effect on the estimates. These estimates have an interpretation in terms of the error
arising from the sampling processes for prices and for the weighting information. Including both components together in an evaluation is challenging.

An intermediate approach is to seek to calculate the same design-based sampling error but to use an approach with a model-based justification to calculate it. The replication methods in the BLS's CPI fall into this category, and Kott (1983) gives a model-based justification for why this approach produces satisfactory results. In some sense, variance estimates that work well under both model-based and design-based approaches are ideal, because they have natural interpretations in both contexts. Jackknife and bootstrap approaches are similar to the replication methods and attempt to approximate the sampling distribution of the required statistic.

Several authors have espoused a model-based approach to estimating price indices, generally of the Lowe index (or Laspeyres type, following the notation of Zhang, 2010). Kott (1984) seems to have been the first to take this approach, setting out a superpopulation model under pps sampling in a simplified unstratified design where the prices are assumed to be ‘nearly homogeneous’ and using a modified (model-based) version of the HT estimator of the Laspeyres-type index. He considers cases with and without autocorrelation in the price trends. Valliant & Miller (1989) and Valliant (1991) set up model-based estimators built on a simple autoregressive model for price change, the former for a one-stage design and the latter for a two-stage design with rotation. Both models admit a range of model-based estimators, and Valliant (1991) in particular identifies some such estimators, which also have design-based interpretations, and the corresponding variance estimators can therefore be regarded as approximations for the design-based estimators too.

Zhang (2010) chooses an explicitly model-based approach and uses the residual variance to capture the extent to which the observed data vary around a specified model. Zhang's models are designed to be appropriate to the chosen form of the index but do not have the property of going to zero as the sample size increases towards the population size. In this sense, the variances produced do not have the same interpretation as variances due to the sampling process, unless we make very specific assumptions about how the model also captures the sampling information. Some of Kott's (1984) estimators are consistent, and some are not; but as Kott points out, ‘[in the superpopulation framework] consistency is an odd property to require of an estimator based on a sample of a specified size’. Nevertheless, such an approach can be much more straightforward to calculate and be a good indicator of the variability in the price index. Kott (1984) takes this approach further and sets out the theory for a superpopulation approach to the design of a price index; adopting such a strategy would make the variance estimate and the design line up more clearly than in Zhang's approach where the design and variance estimation come from different paradigms. V

Norberg (2004) undertook a simulation study involving both design-based and model-based estimators of the variance. His model-based estimators are of the variance component type used by Baskin & Johnson (1995). He found that the model-based variance estimator generally worked well with synthetic data but did not always produce results when based on simulations derived from real data. He concluded ‘[t]his estimator, however, is not practical for complex situations like this’, and overall preferred one of the random-group estimators.

3 Methods for Calculation of Sampling Errors

This section lists the methods that have been proposed for the calculation of sampling errors in CPIs and discusses their strengths and weaknesses.
3.1 Design-based Approaches with Taylor Linearisation

Several authors have used standard results from sampling theory as a basis for deriving variance estimators accounting for the complex designs used in price index surveys. Andersson et al. (1987b) use this approach to assess the variance due to sampling outlets in the Swedish CPI; in their particular application, sampling is \( \pi ps \), and they make the (commonly used) simplifying assumption that sampling is with replacement (see their section 3.4 for detailed expressions). Ardilly & Guglielmetti (1992) also take a design-based approach to the variance of the French CPI.

As an example of the kinds of expression, which result, in the USA, CPI Taylor linearisation was for some time used to give an approximation for the variance of the Laspeyres index between times \( s \) and \( t \):

\[
\text{var} \left( \hat{I}_{t,s} \right) \approx \left( \hat{I}_{t,s} \right)^2 \left( \frac{\text{var} \left( \sum_{i,m} \hat{w}_{im} \hat{I}_{im,0}^s \right)}{\hat{w}_{im} \hat{I}_{im,0}^t} + \frac{\text{var} \left( \sum_{i,m} \hat{w}_{im} \hat{I}_{im,0}^s \right)}{\hat{w}_{im} \hat{I}_{im,0}^t} - 2 \frac{\text{cov} \left( \sum_{i,m} \hat{w}_{im} \hat{I}_{im,0}^s, \sum_{i,m} \hat{w}_{im} \hat{I}_{im,0}^t \right)}{\hat{w}_{im} \hat{I}_{im,0}^t} \right),
\]

where the \( \hat{w}_{im} \) are weights and \( \hat{I}_{im,0}^s \) are component indices for item stratum \( i \) and index area \( m \) (Leaver & Valliant, 1995). The variances and covariances were estimated with replicate samples (see Section 3.2), but in principle, these could also be derived analytically using Taylor expansions of ratios; nevertheless, such a procedure becomes quite involved. There is a suggestion that Taylor linearisation may underestimate empirical variances for smaller sample sizes (see, e.g. Andersson et al., 1987a), and this underestimation and/or the errors of approximation may add up in the different components of the formula.

Valliant (1991) also uses Taylor linearisation to derive variances, but in a superpopulation (model-based) framework. He makes the argument that these variances can also be considered as the design-based variances because the index formulae can also be derived from a two-stage cluster sample of outlets and prices within them. Valliant notes that for long-term price changes, the number of covariances to be estimated can become very large.

Dalén & Ohlsson (1995) take the cross-classified design, which arises from the independent selection of items and outlets and use this to derive a design-based estimator, again using Taylor linearisation to deal with the ratio form of the price index. Their equations (3.3)–(3.6) give the form of this estimator, which contains many terms and is not recapitulated here. They deal only with the variation in the prices, with further thinking about how the variation in the weights should be incorporated in Dalén (1995). Skinner (2015) extends the cross-classified sampling results and in particular gives a bootstrap estimator (see Section 3.3), which is computationally easier, although it has not yet been applied to price indices.

3.1.1 Strengths and weaknesses

Design-based approaches can be more efficient than replication approaches when the latter require the collapsing together of strata to make estimates, and the criticism that they need re-deriving for different estimators is less pertinent for price indices where the form of the estimator does not change (Valliant, 1991). So, many of the standard criticisms of Taylor linearisation are muted in this case. The estimators may nevertheless be extremely complicated, and require several layers of approximation, whose aggregate effect is not clear without detailed investigation. Dalén & Ohlsson (1995) nevertheless produce a valid design-based estimator, and Norberg (2004) shows that this is effective when the cross-classification involved reflects strong
effects in the data. There are real challenges in how to adapt the design-based approach to properly take account of imputation, and also how to deal with procedures for dealing with the transience of products, such as product replacement rules and adjustments for quality change.

3.2 Replication-based Approaches

An approach that avoids the tedious calculation of variance estimators is to use replicate samples (also known as balanced half samples, balanced repeated replication). This is the basis of the longest-standing regular approach to calculation of standard errors for a CPI, from the USA (e.g. Wilkerson, 1967; Leaver, 1990). Leaver et al. (1991) included the contributions of both the variance of the weights and the variance of the prices to the overall estimates, facilitated by having common sample areas for the two surveys (Leaver & Valliant, 1995 p. 554). Replicates have also been suggested for individual elements of the variance including the weights (Balk & Kersten, 1986), and Koop (1986) outlines an approach to combining the variability of weights and prices even when the surveys take place in independently sampled areas.

Each sample stratum $h = 1, \ldots, H$ is divided into two parts, $h_a$ and $h_b$, called half samples. A series of replicates is constructed by choosing one of the half samples from each stratum, appropriately reweighted to give the correct population estimate, and using the variation among these replicates as a basis for estimating the variance. It is not necessary to use all $2^H$ replicates, although in general, the larger the number used the better the estimate obtained. The process can be made more efficient by use of a Hadamard matrix, a matrix containing 1 and −1 entries with orthogonal columns (Wolter, 2007), which makes the procedure balanced (in the replicates). The procedure is to take a column of the matrix and to take $h_a$ in stratum $h$ if the $h^{th}$ element of the column is 1 and $h_b$ if it is −1. This set of half samples is used to make an estimate (of any statistic of interest, but in this case the index). Then the process is repeated with the next column; label the index estimated from the $a^{th}$ replicate by $I_a$. If there are $A$ replicates, the variance estimator is then

$$\text{var}(I) = \frac{1}{A} \sum_{a=1}^{A} (I_a - \bar{I})^2,$$

where $\bar{I} = \frac{1}{A} \sum_{a=1}^{A} I_a$. Biggeri & Giommi (1987) give three further estimators in their outline of the method.

There are practical challenges in setting up a collection to operate this way. In the USA, the whole collection process for prices was replicated with selection of two sets of prices within self-representing metropolitan areas or with paired selections of non-self-representing areas. In particular, a different sample of representative items was chosen in each replicate (by the same sampling procedure); the half samples need not contain the full detail, so it is sufficient if the union of the samples provides sufficient detail for the calculation of the national index. In an ideal situation, these replicate sample prices are collected independently, and then all of the variation in the collection procedures is accounted for in the replication variance. This is a singular advantage of this procedure, particularly with respect to price indices with their relatively complex data collection processes, that it can measure the variability due to non-probability but repeatable sampling procedures. For example, there is no need to derive variances accounting for replacement indices or quality change as long as the procedures are used in the same way in both replicates.

If the replicate sampling can be repeated in further stages of the sample design, then the variances induced by sampling in the different stages can be estimated, and the information...
used to improve the efficiency of the sampling process. This was implemented in the USA where a few cities had such additional replication, and this was used in decomposing the variance (Wilkerson, 1967), although with limited success because the variance components themselves have large variances. Later work on variance decomposition has relied on models (see Section 3.4).

Outside the USA, where no other country uses a design with real replicates, the process is more usually approximated by dividing a single sample into two pieces, which therefore have a small negative dependence, which would not be present if the replicates were selected separately with replacement. If the sampling fractions are small, the difference from ignoring this dependence should be negligible, and this is the procedure used by Balk & Kersten (1986). In the USA, the procedure changed around 2012 from having two separately selected half samples to a single sample with prices randomly allocated to replicates (Shoemaker & Marsh III, 2011); the results were substantially similar, although in general a little lower because the new method improved the balance between the two half samples and eliminated one component of initial weight variation. It is however essential to ensure that price quotes are retained in the random group to which they are originally assigned. Redoing the randomisation period to period leads to substantial overestimates of the variance of changes relative to the original balanced half samples method.

3.2.1 Strengths and weaknesses

The replicate samples approach has definite advantages in estimating the variance due to all the components of the sampling procedure, whether probability based or not, as long as they are repeatable in the different replicates. Therefore, variability due to imputation (Leaver & Larson, 2002), product replacement and quality adjustment are all included as long as the replicates are set up properly. There is a cost to running a system that really implements the replicates, but even in the USA, this has now been replaced by forming replicates from a single sample (Shoemaker & Marsh, 2011), and the same approach has been used in other countries (e.g. Balk & Kersten, 1986). And because there is no need to fit or assume a suitable model, there is no difficulty in asserting the objectivity of the resulting variance estimates.

If the sampling procedure generates small samples, replication can produce large variances. Shoemaker (2009) reported an unusual estimate of the variance of the US CPI, which was traced to a single cell within the housing category, where a particular set of circumstance led to a small and heterogeneous sample. One of the two housing replicates as a consequence contributed approximately half of all the variance in the national CPI in this month. Shoemaker considered several alternative estimators, including jackknife and bootstrap estimators (see Section 3.3), and while some were better in the affected month, they showed outlying variance estimates in other months, which were not seen in the replicate samples approach.

3.3 Jackknife and Bootstrap

Jackknife and bootstrap methods have become increasingly popular for variance estimation, and various developments have been made, which enable them to be used in ever more complex designs. The jackknife is already in use in the USA for special item categories, which do not have replicates and therefore do not fit in the random-group method (BLS, 2015, p. 39).

Biggeri & Giommi (1987) outline a jackknife procedure and four associated estimators (similar to those proposed for the replication variances; see Section 3.2). Leaver & Cage (1997) describe a jackknife approach for the US CPI where items are grouped into seven strata for the self-representing areas, each with 32 clusters and an eighth stratum for the non-
self-representing areas with 12 clusters. Each replicate involves deleting one cluster in one stratum, recalculating the weights in this stratum to produce an estimate from the remaining clusters and then using the new estimate together with the other strata (unchanged) to make a replicate estimate. Leaver and Cage point out that the stratified jackknife estimator assumes equal expected price (or price change) among the clusters within a stratum, so the resulting variances are on average expected to overestimate the true variances for items where this assumption does not hold. However, they find both underestimation and overestimation relative to random groups, although Shoemaker (2009) indeed finds that jackknife variance estimates are generally slightly higher than the stratified random-group method. Shoemaker (2009) and Klick & Shoemaker (2019) use a slightly different implementation of the jackknife, as a (conservative) upper bound for the variance estimates.

The bootstrap is in principle also available, with resampling of the clusters within the strata with replacement from the available clusters. The only evidence that this has been used is once again from the USA where Shoemaker (2009) investigated both jackknife and bootstrap estimators as possibilities for dealing with an unusual variance estimate (see also Section 3.2). Jackknife and bootstrap produced practically the same pattern of variance estimates, with the jackknife slightly higher than the bootstrap.

3.3.1 Strengths and weaknesses

The jackknife and bootstrap act in quite a similar way to the replicate samples once the sampling system has been set up in this way. Essentially, they replace the choice of half sample in each stratum. Therefore, they are able to take account of the variance including such processes as imputation, product replacement and quality adjustment, as long as these are recalculated in the jackknife or bootstrap replicate. However, this is a considerable complication, and it is not clear whether it is performed in the US implementations.

Otherwise, the properties are rather similar to the stratified replicates approach, including the potential susceptibility to unusual observations. The bootstrap provides some additional flexibility in that over many replicates it allows the distribution of the variance estimates to be constructed, which may demonstrate the susceptibility to outliers.

Andersson et al. (1987a) in a slightly different context find that the sampling distribution of the variance estimates over replicates is spread much wider than the replication equivalent; it is not clear how far this result can be generalised to single price indices.

3.4 Model-based Approaches

Kott (1984), Valliant & Miller (1989), Valliant (1991) and Zhang (2010) all make use of a model-based framework for price index estimation. They all operate with (at least) the Laspeyres-type index and use an estimator of the form \( \sum_{i} w_i p_{i}^{0,t} \), a weighted sum of price relatives where 0 in the superscript represents the base period and \( t \) the current period. Zhang starts with models which motivate the elementary aggregates (for Carli, Dutot and Jevons), and builds up the price relatives from these components. The motivating model for the Carli (perhaps the simplest if not the most realistic for modern implementations of the CPI) is

\[
\hat{r}_{ij}^{0,t} = \theta_{ti} + \eta_{tij}
\]

with a common parameter \( \theta \) for the ratio of prices between the base period 0 and \( t \) (0 < \( t \)), and price-specific error \( \eta \) with var \( (\eta_{tij}) = \sigma_{\eta}^2 \) and cov \( (\eta_{tij}, \eta_{tik}) = 0 \) for \( j \neq k \). With this model, the Carli is the best linear unbiased estimator of \( \theta \), but as Zhang points out, this does not mean that model (1) is suitable for the price data to hand. He prefers a robust model-based variance
estimator, which does not depend on the distribution of $\eta$ but instead only on the property that $E(\eta_{tij}) = 0$. This estimator is

$$v_i = \frac{1}{n_i (n_i - 1)} \sum_{j=1}^{n_i} \left( \frac{1}{n_i} \sum_{k=1}^{n_i} r_{ij}^{0,t} - \frac{1}{n_i} \sum_{k=1}^{n_i} r_{ij}^{0,t} \right)^2$$

(4)

(see Zhang, 2010 for estimators corresponding with the Dutot and Jevons model assumptions). Zhang goes on to provide additional expressions, which allow for the calculation of variances when chaining the elementary aggregates and also for the chaining of higher-level indices.

Valliant (1991) uses a more complex model, which allows for different parameters for different outlets $h$ and correlation of prices within outlets. He also introduces an autoregressive error, which allows for outlet-specific autocorrelation in time in the model errors for individual price relatives:

$$r_{hij}^{t,0} = \alpha_{th} + \omega_{thi} + \epsilon_{thij}$$

$$\epsilon_{thij} = \rho_h \epsilon_{(t-1)hij} + \xi_{thij}$$

(5)

with parameter $\alpha$ for outlet $h$, a random parameter $\omega$ for commodity $i$ within outlet $h$, and with price-specific errors $\epsilon$, with $E(\omega_{thi}) = E(\epsilon_{thij}) = 0$, $E(\omega_{thi} \omega_{t'h'i'}) = 0$ if $thi \neq t'h'i'$, $E(\omega_{thi}^2) = \sigma_{\omega h}^2$, $E(\epsilon_{thij} \epsilon_{t'h'i'j'}) = 0$ if $thi j \neq t'h'i' j'$, $E(\xi_{thij}^2) = \sigma_{\xi h}^2$ and $-1 < \rho_h < 1$. The additional flexibility of this model suggests that it may be able to fit better, although there is no summary of the model fit in Valliant’s paper (where the results are largely based on a simulated population, which is nevertheless constructed from actual US price data). This approach is quite closely related to the stochastic approach to determining the appropriate form of an index number. This model leads to a range of possible (L-type) index estimators (Valliant (1991) gives seven possibilities), most of which do not correspond directly with the classical ways of constructing a CPI, although some of the estimators are similar. These estimators lead to different variance estimators, some of which have good design properties as well as model properties and can be interpreted as approximate design-based estimators.

As well as these explicitly model-based estimators, a number of authors have used models to estimate components of the sampling variance due to the different stages of the multistage design, particularly in the USA (Baskin 1992, 1993; Baskin & Johnson, 1995) and also in Italy (D’Alò et al., 2006). Fixed effects (analysis of variance) models can be used but are susceptible to producing negative estimates of variance, so random-effect models and restricted maximum likelihood or Bayesian estimation have been investigated.

The US design has four stages of sampling, and the price relatives $r$ can be related to the stages through a random-effect model:

$$r_{ijkl}^{t,s} = \mu_{ijkl}^{t,s} + \alpha_i^{t,s} + \beta_{ij}^{t,s} + \gamma_{ik}^{t,s} + \epsilon_{ijkl}^{t,s}$$

(6)

where $\mu$ is a fixed effect and $\alpha, \beta, \gamma$ and $\epsilon$ are mutually independent random effects with mean 0 and variances $\sigma_{\alpha}^2$, $\sigma_{\beta}^2$, $\sigma_{\gamma}^2$ and $\sigma_{\epsilon}^2$ (in principle, the variances also depend on $t,s$, but the model is simplified so that it is not time dependent, giving time-averaged values). $i, j, k$ and $l$ label the sampling of PSUs, outlets, items and products, respectively. These models seem to be best fitted by restricted maximum likelihood, although this does give different variance component estimates than analysis of variance (Baskin & Johnson, 1995). The variance components are generally used in sample design to ensure that the greatest sample sizes are introduced at the most variable stages, but they can also be used to estimate the overall sampling error of the index, and the process is sketched by d’Alò et al. (2006 section 4).
3.4.1 Strengths and weaknesses

The key advantage of these model frameworks is that they do not require a probability sampling mechanism to justify the estimates; Zhang (2010) argues that this is the only realistic framework for most CPIs because they use purposive sampling. The method is also relatively simple—in this respect similar to the strictly design-based approach of Section 3.1 in that setting up the model and deriving the appropriate estimators and variance estimators may involve some detailed algebra, but once it is set up, it can be used relatively straightforwardly for a period. Some elements of the adjustment of the prices in an index, such as quality adjustment and imputation, are model based and can be included in the model structure if desired (at the cost of extra complexity). Hedonic modelling methods, which are also widely used for some items where it is difficult to account for quality differences, also fit naturally within the model-based framework (Johnson, 1975, provides an early example, and there seems to have been little research on this problem since then).

The disadvantages are that it is necessary to check regularly that the model continues to fit the data. And there is the question of what is an appropriate model. The model variance captures the average difference of the observations from the model, so the choice of model makes a difference to the estimate of the variance. This makes it more difficult to justify a particular model and to claim objectivity for the variance estimates. But this has not been a particular hindrance in other parts of official statistics based on models, such as small area estimation. In this sense, Zhang’s approach of using a model as a basis to motivate the existing approaches and then calculating a robust variance, seems more likely to be widely acceptable. However, a model-based approach is not impossible, and indeed, it would be possible to construct a completely model-based price index system (as suggested by Kott, 1984).

The model-based sampling error also responds in some way to von Hofsten's (1959) criticism that there is no such thing as a sampling error, in that it does not arise from the sampling process. Including all the elements that led von Hofsten to consider that sampling errors were too challenging may also be difficult in a modelling framework, but at least such a framework will treat all its elements consistently.

4 The Development of Variance Estimates in the US Consumer Price Index

The main push towards calculating variance estimates for the US CPI came during the 1950s, with Mudgett (1951) setting out the theory and providing a critique of BLS’s compilation procedures. Adelman (1958) undertook a local study and calculated the variances of her price index, comparing with the BLS’s index for the same area. Kruskal & Telser (1960) criticised the absence of studies into the variability of the CPI. McCarthy (1961) wrote a staff paper in support of a review of the CPI methods, an impressive paper that not only set out the case for calculating sampling variances for the CPI but also made the best use of available data to calculate approximate error measures. The bias component was calculated from Swedish data, and estimates of the variance due to sampling commodities and due to sampling both cities and outlets were calculated from detailed data made available by BLS for the review. This was the first, if rather approximate, attempt to calculate the errors for a national price index. McCarthy also set out a framework for a half-sample approach to sampling both cities and commodities, which would enable variances to be estimated accounting for the complexities of price collection. He specified an additional use of both commodity half samples within some paired cities to enable the components of the variance due to different parts of the sampling to be estimated. This approach was implemented in the US CPI from December 1963. Almost all of the subsequent development work has been undertaken by employees of BLS; although many of these
have respected academic reputations, it is interesting that there has not been much direct interest from the academic community. This is, however, partly attributable to a lack of pressure from stakeholders, despite the many uses of detailed components of CPIs.

The new sampling structure led to the first substantive estimates of sampling variances for a national index—originally in some American Sociological Association conference proceedings (Wilkerson, 1964) and later in Wilkerson (1967). These demonstrated that the variances were relatively small and suggested that focus on improving accuracy should be in other areas.

There therefore seems to have been a hiatus in the production of sampling errors, until Weber (1980) described the sampling procedure incorporating the half-sample selections in detail and also set out the variance estimation methods. This was to be built into a variance estimation system, although the implementation of the calculations is not described. Weber explained that the theory for the application of half samples was not completely worked out for complex statistics, but Kott (1983) provided a (model-based) superpopulation justification for the approach. Kott also made the first foray into designing a CPI as a model-based index (Kott, 1984), although this approach has not been taken up in a national CPI.

The model-based approach was developed further in a series of papers by Richard Valliant, who set out a model (summarised in Section 3.4) for the evolution of prices and used it as the basis for a series of model-based estimators (some also with good design-based properties), first under a single-stage design (Valliant & Miller, 1989) and then a two-stage design (Valliant 1991). Valliant (1999) also provided a more general view of the use of models in price indices, including the model-based approach, its relation to the stochastic approach to index numbers, and the estimation of variance components and how they are used in efficient sample allocation.

Despite Weber's description of the variance estimation system, it was not until Leaver (1990) that the next set of variance estimates, for 1978–1986, were produced, more than 20 years after Wilkerson's initial estimates. The whole price selection had been made probabilistic from the redesign introduced for 1978 (BLS, 2015, Chapter 17), so this was a natural starting point. Leaver's estimates were conditional on the weights (i.e. treating the weights as fixed constants) and so underestimated the total variance. They corroborated the analytical deduction of Valliant & Miller (1989) and Valliant (1991) that the standard error of the index would grow with distance from the base period, although the variances of changes in the index were approximately constant over a given lag. Leaver et al. (1991) extended the approach to unconditional variances (i.e. accounting for the sampling variation in the weights), again using Taylor linearisation to combine the different variance and covariance estimates. The variances of the index level were essentially unaffected, and the variances of 12-month change increased by between 6% and 20% over the conditional variance, again in line with Edgeworth's (1888) early results on the relatively small impact of the variance of the weights. Leaver & Swanson (1992) extended the calculation of estimates to 1987–1991, taking account of the redesign from 1987, which had used variance information to rebalance the sample towards selecting additional outlets rather than additional items. These estimates also included a new component from the estimation of the base year expenditures in each replicate, which had not been included in the earlier estimates, and incorporated covariances between indices for higher-level geographical areas. The estimates for this period are slightly higher than those for 1978–1986, and also more variable, but otherwise, the pattern of the estimates tells much the same story.

Leaver & Valliant (1995) summarised the conditional/unconditional variance estimation work and proposed an estimator, which used the replicates without the need for the Taylor linearisation, which had more degrees of freedom and therefore produced more stable variance estimates. Baskin & Leaver (1996) investigated a jackknife variance estimator and found that it had slightly larger empirical bias and empirical variance than the Taylor linearisation
estimator. They also examined the use of the geometric mean in place of the Laspeyres estimator for housing but found that this had little impact on the estimated variance. Leaver & Cage (1997) extended this to the whole CPI and to superlative indices, again finding little difference in variance estimates under the different estimators. From 1999, the CPI switched to using the geometric mean estimator, in response to the recommendations of the Boskin report (Boskin et al., 1996). Leaver and Cage also continued research into jackknife variance estimation, finding that it underestimated variance relative to the stratified random-group estimator in some item groups and overestimated in others.

One important effect identified in the series of papers by Leaver and co-authors is the impact that imputation has on the variance. This was investigated further by Leaver & Larson (2002), who calculated a set of variances using imputation across the whole index (and therefore not accounting for the imputation variance) compared with a set where imputation was performed within replicates (and accounting for imputation variance). Imputation variance accounted for 0–10% of total variance in 1-month changes in fresh fruit and vegetables but was somewhat higher for citrus fruits, which are particularly seasonal. Patterns for variance of longer-term changes were counterintuitive and are difficult to interpret.

Decomposing the variance into components, which can be used to improve the sampling, has also been an important activity in the USA. Several papers describe the periodic redesigns. Baskin undertook a programme of research on how these components could best be estimated (Baskin, 1992,1993; Baskin & Johnson, 1995), preferring restricted maximum likelihood estimation, although it produced some differences compared with standard analysis of variance. It was also taken up for an analysis of the new housing sample in the CPI by Shoemaker (2002). In principle, the models underlying variance components can be used to produce overall sampling errors (see Section 3.4), but this seems not to have been performed in USA, where replicate errors have been available.

The USA began experimentation with scanner data in the early 2000s. Leaver & Larson (2001) undertook a study on different index estimators from scanner prices of cereals and calculated their jackknife variances. There was little difference in variances for different estimators, in line with previous results. They also calculated imputation variance for the scanner data cereals index, essentially the same as that of Leaver & Larson (2002). Here, the rate of missingness was low and the imputation variance constituted only ~0.2% of total variance. Leaver & Larson (2003) repeat the variance components work of Baskin and co-authors (see previous paragraph) with scanner data and using a log-linear model; between (outlet-)chain and between item-type variances are the principal sources of variation.

Developments in recent years have been more piecemeal. The USA introduced a chained superlative price index in 2002, and Shoemaker (2003) investigated its sampling variance using the stratified random-group method. The variances of the superlative index were generally slightly higher than for the main index. Shoemaker (2009) and Klick & Shoemaker (2019) worked with jackknife variance estimators (see Section 3.3), to investigate outlying variance estimates and differences between indices for different (special) populations, respectively.

Around 2012, the BLS moved away from the selection of two replicates directly and instead used a single sample. The price quotes were then randomly allocated to replicates so that variances could continue to be produced using the half-sample procedures, which had been in use for variance estimation since 1964. The allocation of prices to replicates had to remain stable in order to obtain similar estimates (Shoemaker & Marsh III, 2011).
5 Total Quality Estimation for Consumer Price Indices

There has been a lot of work by economists on the conceptual basis of consumer prices, and the sorts of biases which arise because the concepts are not well matched by the data and methodological construction (see, e.g. Wynne & Sigalla, 1994; Moulton, 1996). Dalén (1995) contrasts this with the more usual approach of survey statisticians who are concerned with sampling errors and other kinds of non-sampling errors.

5.1 Mean Squared Error Estimators of Consumer Price Index Error

Biggeri & Giommi (1987) set out a structure for the classification of errors within a price index based on a breakdown of the mean squared error (mse) as

\[ \text{mse} \left( \hat{I} \right) = E \left( \hat{I} - E \left( \hat{I} \right) \right)^2 + \left( E \left( \hat{I} \right) - I \right)^2 + \left( I - I^* \right)^2 + 2 \left( E \left( \hat{I} \right) - I \right) (I - I^*). \] (7)

where \( \hat{I} \) is the estimated index, \( I \) is the defined goal (‘true value’) of the index and \( I^* \) is the ideal goal (these elements had already been described by McCarthy, 1961, p. 211). The first term on the right-hand side of (7) represents the total variance, including both sampling variance (as discussed so far in this paper) and measurement variance. The second term represents the (squared) bias derived from sampling and measurement. The third term represents the (squared) bias capturing how well the specified form of the index captures some ideal index number; this has often been operationalised as how well a given index approximates a superlative index, but this has been a longstanding area of debate with no universally agreed solution, and we therefore do not consider this element further here. The final term captures the interaction between the bias terms. Andersson et al. (1987b) use the same structure, with the first two terms collapsed together to form the mse of \( \hat{I} \) with respect to the target index \( I \).

Balk (1989) provides a more specific formula for the mse under the assumption that the samples for the expenditure survey, items and outlets are all independent. Dalén (1995) goes into more detail, setting out:

- two biases—one from estimation of the weights (but curiously not considering any component of variance from this estimation) and one the aggregate of the biases in the estimation of the price relatives;
- two variances—related to two sampling stages, one from selection of item groups and one from selection of prices.

5.2 Error Components

Partial enumerations of the errors, which potentially affect price indices, have been provided by Edgeworth (1888), Bowley (1928), Morgenstern (1963 Chapter 10), Biggeri & Giommi (1987), Dalén (1995) and BLS (2015). Economists have also considered many specific error sources, and these are mostly wrapped up here in ‘measurement error’, rather than considered separately, because we focus on the statistical errors. They are summarised (also with a simple error taxonomy) in ILO et al. (2004, chapter 11). A synthesis of these sources of error, summarising and extending the component classification of Biggeri and Giommi, and showing which among the selected authors considered which errors, is shown in Table 1. It is clear that the range of potential error sources in a CPI has grown with time as there has been more detailed consideration of the processes for producing an index.
Table 1. Synthesis classification of error sources summarised and extended from the framework of Biggeri & Giommi (1987), using selected works

| Combined classification (derived from Biggeri & Giommi 1987 with additions) | Edgeworth (1888) | Bowley (1928) | Biggeri & Giommi (1987) | Dalén (1995) | BLS (2015) |
| --- | --- | --- | --- | --- | --- |
| Component A: sampling error and error variance $E\left(\hat{E} - E\left(\hat{E}\right)\right)^2$ | | | | | |
| A1 representative item sample error | | | | | |
| A2 point of purchase sample error | | | | | |
| A3 point of time sample error | | | | | |
| A4 expenditure weight sample error | | | | | |
| A5 product sample error | | | | | |
| A6 measurement error variance | | | | | |
| Component B: bias $E\left(\hat{E} - \hat{E}^*\right)^2$ | | | | | |
| B1 frame errors* | | | | | |
| B2 coverage errors* | errors from unrepresented item groups | (d) omission of relevant classes | B2 2.2 nonrepresented consumption 2.7 (part) undercoverage coverage error |
| B3 concept error for prices | errors in price relatives | (e) noncoincidence of the field of investigation with the objective | B3 2.8 conceptual ambiguities in certain item groups estimation error (part) |
| B4 homogeneity error (of products between periods) | | | B4 | |
| B5 nonresponse bias | | (a) inaccuracy of price relatives | B5 2.7 (part) nonresponse nonresponse error | | 

Continues
Table 1. Continued

| Combined classification (derived from Biggeri & Giommi 1987 with additions) | Edgeworth (1888) | Bowley (1928) | Biggeri & Giommi (1987) | Dalén (1995) | BLS (2015) |
|---|---|---|---|---|---|
| B6 | measurement error (includes substitution and quality adjustment) | B6 | 2.5 quality adjustment errors 2.6 errors in the recorded price | response error, estimation error (part) |
| B7 | concept error for weights | errors in weights (c) errors in weights | B7 | 2.1 high-level weight errors |
| B8 | processing error | | | processing error |
| B9 | mismatch of prices and weights | (f) inappropriateness of the relative to the weight | | } estimation error (part) |
| B10 | selection bias | | | 2.7 (part) selection bias |

Component C: target bias \((l - l^*)^2\)

| C1 | calculated index is not target index | C1 | 2.3 elementary aggregate bias | estimation error (part) |
| C2 | lack of characteristicity | C2 |
| C3 | lack of consistency in aggregation | C3 |
| C4 | low-level substitution bias | | 2.4 low-level substitution bias |

*Could be split into B1a and so forth as (a) representative items, (b) points of purchase, (c) households and (d) products to match the component A breakdown.
Sampling variances are only one component of the accuracy of a CPI, and there is a general feeling among authors (e.g. Wilkerson, 1967; Biggeri & Giommi, 1987; Dalén, 1995) that sampling errors are generally small relative to the other types of error in a price index. Specific studies on these other types of errors are as follows:

- **Non-response error**—Kersten (1985) calculates bounds on the bias in a price index induced by non-response in the expenditure survey, which provides the weights.
- **Imputation error**—quite large proportions of prices are not collected in any period because the products are unavailable (see, e.g. BLS, 2015, table 3). Leaver & Larson (2002) give some examples of imputation rates and investigate the proportion of the variance due to imputation in the US CPI.
- **Formula error**—this is a vexed question because there is no gold standard with which to compare a formula, but the general approach has been to try to approximate a superlative index formula as closely as possible (see, e.g. Mudgett, 1951, pp. 47–51; Dorfman et al., 2006). In view of the continuing debate particularly in the UK about elementary aggregates, we will not consider a formula error further here.
- **Measurement error**—many of the kinds of errors, which are of concern to economists, come under this category, including new items and products, and dealing with quality change, but it also includes measurement errors of the survey type resulting from the practical difficulties of the field operations. Andersson et al. (1987b) undertook a small study comparing list prices and outlet prices, which forms one component of measurement error. Dalén (1995) gave an evaluation of quality adjustment error for clothing in the Swedish CPI. The effect of these types of biases on the CPI has been addressed more frequently by economists than statisticians, and Chapter 11 of the CPI manual (ILO et al., 2004) provides an overview and many further references. Much recent attention has turned to web-scraped and transaction/scanner data, and these present their own measurement problems, for example, in automated classification of products, although may also avoid some of the measurement errors of observation. Little attention seems to have been directed to assessing these errors while the form of an index based on such information remains to be resolved.
- **Coverage error**—many CPI sampling operations are effectively cut-off samples with parts of the population of interest excluded. Scanner data offer an opportunity to examine the effect of these exclusions, and Brunetti et al. (2018) make such calculations for Italy, where sampling in the main CPI is restricted to the main provincial towns and uses only a sample of the most-sold products. They find only some differences, mainly due to sampling towns only, and concentrated in the south of Italy.
- **Rounding error**—a component of error that is rarely considered and rarely important in official statistics, but in the context of a CPI where both the index level and the percentage changes are regularly reported to only one decimal place, the relative error can be larger and more important. In this respect, it is notable that the UK publishes 12-month inflation rates calculated from the rounded index values, in contradiction of best practice for rounding, for the purposes of presentational consistency. The only known evaluation of rounding error is by the BLS (Williams, 2006), where the error in the monthly change in an index rounded to one decimal place is of a similar magnitude to the sampling error. Rounding and sampling error also interact—for example, Wilkerson (1967) says ‘in fact, a real change of only .1 per cent in the monthly CPI is significant but, because a change of this size in the published figures, one cannot be sure that any particular .1 per cent change is significant’. When the sampling error is near to or less than the effect of rounding, there will therefore be a danger of over-interpretation.
of changes in inflation, and the presentation of sampling error information must account for this.

- Correlated price collector error—Andersson et al. (1987b) analysed this error in the Swedish CPI, and although the correlation is theoretically positive, they found many negative estimates, suggesting that the parameters were not well identified. The authors drew no conclusion about variance inflation from correlated price collector error but noted that estimates were relatively low, <1.1 for most items.

These various components of error can give rise to biases and variances. It is generally challenging in a survey context to measure biases, often requiring some special study to obtain less biased or unbiased measures against which to evaluate. But in price indices, there may be no agreement even on what the right target parameter is (see also the quotation in Section 1.2); Dalén (1995) suggests that these biases should therefore be called bias risks, which ‘show the sensitivity of the index estimate to different, but not unreasonable, index definitions’.

5.3 Producing an Overall Quality Statement for a Consumer Price Index

There are relatively few proposals for the theoretical form of the total error in the CPI, and only Dalén (1995) makes an attempt at numerical evaluation of a range of these variances and biases for the same index. But even he declines to combine all the biases together to give an overall impression of the mean squared error, because the estimates of the biases are themselves erratic, changing from year to year, and it is not clear that a combined version would be a satisfactory guide to the overall quality of the CPI.

Perhaps the ideal strategy to obtain a credible overall quality measure would be to introduce a system, which produces estimates of the various quality components as part of the standard operation of a CPI. Then the variation in the quality measures could also be assessed, and a smoothed version of the total error indicator produced, which would have more general application.

6 Discussion and Future Research Directions

It is interesting that most of the research on estimation of sampling errors of CPIs, in particular, has been undertaken by NSIs (occasionally in collaboration with academics), whereas work on biases in CPIs has seen more academic activity in the economic sphere (ILO et al., 2004, Chapter 11). Indeed many of the references cited here are from conference proceedings and have not been extended to articles in refereed journals. Part of the reason for this is likely to be the institutional environment, which puts less reward on academic publishing and presents many operational pressures on staff time. However, even allowing for this the topic seems under-represented. Perhaps some of the research has taken place in small steps, which do not make satisfactory full papers, but it is difficult to see why these have not eventually been combined into academic publications. Perhaps this paper can in some measure redress this balance.

One of the most challenging aspects of calculating price indices and their variances is the number of data sources from which they are constructed. The situation is becoming more complex as new sources are being included in calculation of the CPI, such as scanner data and web-scraped prices. Alternative data sources such as tax information are also available on the expenditure side, and we return to the way these different data sources can be used together in Section 6.3. One can therefore regard the CPI as composed of ‘different separate surveys, each covering different aspects of the index’ (Biggeri & Falorsi, 2006).
6.1 Choosing a Variance Estimation Approach

It seems possible to draw some general conclusions from the history of research on sampling variances for CPIs. Several researchers have considered the impact of variability in the estimation of the weights used in the production of price indices, generally derived from a household survey, although some countries now adjust these through the national accounting framework. There is general agreement that the effect of variability in the weights is relatively minor compared with the variability in the prices, as already noted by Edgeworth (1888).

There seem to be some clear leaders among the four main approaches described in Section 3. Most national CPIs incorporate probability sampling in some stages of selection of prices and in the calculation of the weights, and the replication-based approaches provide a natural way to estimate this variation and the variation due to non-probability but replicable parts of the procedures. This seems to be the current frontrunner and is the only method in use for regular publication of CPI variances; variations of the method, which seek to create the replicates through application of the jackknife or bootstrap, seem possible but are not in regular use, and there is some limited evidence that the variance of the variance estimate is larger with these procedures.

Model-based procedures have some attractions, particularly in allowing other model-based procedures for non-response and quality adjustment to be included seamlessly, but it is less easy to explain their genesis to users and less easy to explain what they actually mean. More work to produce and validate variance estimates using these methods is needed to provide the evidence from which to judge their usefulness.

Except in special situations, Taylor linearisation seems too complicated for practical use as a single method, although it clearly supports the use of other methods by simplifying the estimation of variances of complex statistics.

6.2 Presentation of Quality Measures

Valliant (1992) investigated the smoothing of sampling error estimates over time and concluded that smoothed series were more in line with user expectations of the smoothness of standard errors and had negligible impact on the coverage of confidence intervals in a simulation study. On this basis, it would seem reasonable to consider the smoothing of variance estimates, particularly for subseries where the sampling variability of the variance might be expected to be greatest.

In the USA, sampling errors are calculated each month with the index but published a year in arrears for the whole year alongside the detailed CPI estimates publication (Shoemaker, 2003). These publications include a description of the sampling and the methods used to calculate the sampling errors, which probably help users (those with sufficient sophistication to understand and use sampling errors) to interpret how the measures fit with the index estimation. Reed & Rippy (2012) give an accessible introduction to the errors in the CPI, intended for public consumption.

6.3 Further Directions for Research in Quality Measurement for Consumer Price Indices

Further development of the model-based approach would be valuable as a way to evaluate the quality information that it can provide. The fitting and evaluation of models in the model-based approach to price indices is a first topic that needs attention. It is one of the helpful list of topics where further research on the use of models in price indices would be beneficial given by Valliant (1999), and all these areas still seem to be open for new research. It would also be interesting and important to adapt variance estimation to account for extra variance due to
• quality adjustment—suggested by von Hofsten (1959) and Leaver & Valliant (1995). This may be in part already covered by the half-sample-type approaches;
• price imputation (Leaver & Valliant, 1995); and
• hedonic price measurement.

The processing of the weights through the national accounting framework also presents some challenges for quality estimation, although in general, it is expected to improve the quality of the estimated weights by triangulation from different sources. Estimating the variances of outputs balanced in this way is already a challenging problem (Mushkudiani et al., 2020), before the extra complexity of inclusion in a price index is added, and this is a topic needing further study.

There were some interesting ideas about the redesign of the Italian CPI suggested in Biggeri & Falorsi (2006) and D'Alò et al. (2006), including positive coordination of outlet samples for different commodities, which would then mean that they were no longer independent. It would be interesting to examine whether Dalén & Ohlsson's (1995) approach could be extended to this kind of design. Biggeri and Falorsi and D'Alò et al. also propose balanced sampling of municipalities, and there are procedures for estimating the variance of balanced samples, so these could conceivably be incorporated, although this may be something that can be more straightforwardly accommodated by selecting balanced replicates and using the balanced half-sample approach. It would also be worth extending this idea, and investigating whether the variance in the CPI could be reduced through balanced sampling of municipalities/cities/areas rather than random sampling, measured through the replicate sample approach.

The USA has rotation in price collection (Valliant, 1991, implies 20% rotation in areas, outlets and items each year, although other work suggests that areas are replaced less frequently following each census). These rotations form panels from which even lower-level subindices could be constructed. This type of structure lends itself to modelling using state space models incorporating the rotation structure, and this could be used in a time series type analysis of the variances in a way that was excluded by Baskin & Johnson (1995). In principle, this approach could be used to estimate the long-term path of the variance of the overall CPI.

The use of non-probability samples is widespread in the construction of CPIs, with many examples of cut-off type sampling and more recently information from scanner data and webscraping being included in national indices. These form a range of surveys with different probability and non-probability designs, which are combined to make the final CPI. Much recent research has been directed at obtaining valid estimates and inferences from non-probability samples (see, e.g. Elliott & Valliant, 2017; Zhang, 2019; Rao, 2020, in press for summaries). But examples of applications of these approaches in components of price indices are needed, and beyond that the methods for combining quality measures from the non-probability components with those from probability components needs to be worked out.

In short, there are many opportunities for further research on the quality assessment of CPIs.

Notes

1 Now known as producer prices.
2 A discovery that has regularly been remade by other workers unaware of this early result.

Acknowledgements

This work was supported by the Office for National Statistics (ONS) under contract PU-16/0031-6.009 and by Horizon 2020 Framework Programme funding as part of the
MAKSWELL project G.A. no. 770643. I am grateful to the members of the Advisory Panel on Consumer Prices—Technical and to two referees for their comments on a draft version of this article.

References

Adelman, I. (1958). A new approach to the construction of index numbers. *Rev. Econ. Stat.*, **40**, 240–249.

Andersson, C., Forsman, G. & Wretman, J. (1987a). Estimating the variance of a complex statistic: a Monte Carlo study of some approximate techniques. *J. Off. Stat.*, **3**, 251–265.

Andersson, C., Forsman, G. & Wretman, J. (1987b). On the measurement of errors in the Swedish consumer price index. *Bull. Int. Stat. Inst.*, **52**, 155–171.

Ardilly, P. & Guglielmetti, F. (1992). Optimisation de l’échantillon pour le calcul de l’indice des prix à la consommation. *Insee Méthodes*, 29-30-31, 71-123. (in French) http://www.jms-insee.fr/1991/S01_3_ACTE_ARDILLY&x02010;GUGLIELMETTI_JMS1991.PDF

BLS (2015). BLS handbook of methods, chapter 17, consumer price indexes. Available from https://www.bls.gov/opub/hom/pdf/homch17.pdf. Accessed 10 January 2020.

Balk, B.M. (1987). Introductory remarks. *Bull. Int. Stat. Inst.*, **52**, 134–135.

Balk, B.M. (1989). On calculating the precision of consumer price indices. *Bull. Int. Stat. Inst. 47th* session, Paris, contributed papers, book 1.

Balk, B.M. (1991). Estimating the precision of a consumer price index; some experiences from the Netherlands. *Bull. Int. Stat. Inst. 48th* session, Cairo, contributed papers, book 1, 51–52.

Balk, B.M. & Kersten, H.M.P. (1986). On the precision of consumer price indices caused by the sampling variability of budget surveys. *J. Econ. Social Measure.*, **14**, 19–35.

Banerjee, K.S. (1956). A note on the optimal allocation of consumption items in the construction of a cost of living index. *Econometrica*, **24**, 294–295.

Banerjee, K.S. (1959). Precision in the construction of cost of living index numbers. *Sankhyā*, **21**, 393–400.

Banerjee, K.S. (1960). Calculation of sampling errors for index numbers. *Sankhyā*, **22**, 119–130.

Baskin, R.M. (1992). Hierarchical Bayes estimation of variance components for the U.S. consumer price index. *Proc. Survey Res. Methods Sect. Am. Stat. Assoc.*, **1992**, 716–719.

Baskin, R.M. (1993). Estimation of variance components for the U.S. consumer price index via Gibbs sampling. *Proc. Survey Res. Methods Sect. Am. Stat. Assoc.*, **1993**, 808–813.

Baskin, R.M. & Johnson, W.H. (1995). Estimation of variance components for the US Consumer Price Index. *Proc. Sect. Surv. Res. Methods Am. Stat. Assoc.*, **1995**, 126–131.

Baskin, R.M. & Leaver, S.G. (1996). Estimating the sampling variance for alternative forms of the U.S. consumer price index. *Proc. Survey Res. Methods Sect. Am. Stat. Assoc.*, **1996**, 192–197.

Biggeri, L. & Falorsi, P.D. (2006). A probability sample strategy for improving the quality of the consumer price index survey using the information of the business register. In *Joint UNECE/ILO Meeting on Consumer Price Indices, 10–12* May 2006. Geneva. https://www.unece.org/fileadmin/DAM/stats/documents/cececes/ge/22/2006/mtg2/wp.12.e.pdf

Biggeri, L. & Giommi, A. (1987). On the accuracy and precision of the consumer price indices: methods and applications to evaluate the influence of the sampling of households. *Bull. Int. Stat. Inst.*, **52**, 137–154.

Boskin, M.J., Duberger, E.R., Gordon, R.J., Griliches, Z. & Jorgenson, D. (1996). Toward a more accurate measure of the cost of living. Available from www.ssa.gov/history/reports/boskinrpt.html

Bowley, A.L. (1926). The influence on the precision of index-numbers of correlation between the prices of commodities. *J. R. Stat. Soc.*, **89**, 300–319.

Bowley, A.L. (1928). Notes on index numbers. *Econ. J.*, **38**, 216–237.

Brunetti, A., Fatello, S., Polidoro, F. & Simone, A. (2018). Improvements in Italian CPI/HICP deriving from the use of scanner data. In *49th Meeting of the Italian Statistical Society*, 20–22 June 2018. Palermo. http://meetings3.sis-statistica.org/index.php/sis2018/50th/paper/view/1484/32

D’Alò, M., Di Consiglio, L., Falorsi, S. & Solari, F. (2006). Estimation of variance for consumer price index. In *Proceedings of the Scientific Conference of the Italian Statistical Association*, pp. 469–472, 14–16 June 2006: Turin.

Dalén, J. (1995). Quantifying errors in the Swedish consumer price index. *J. Official Stat.*, **11**, 261–276.

Dalén, J. & Ohlsson, E. (1995). Variance estimation in the Swedish consumer price index. *J. Business Econ. Stat.*, **13**, 347–356.

Department of Employment (1971). *Proposals for Retail Prices Indices for Regions, Cmnd 4749*. HMSO: London.

Dorfman, A.H., Lent, J., Leaver, S.G. & Wegman, E. (2006). On sample survey designs for consumer price indexes. *Survey Methodol.*, **32**, 197–216.
Sampling Errors in Consumer Price Indices

23

Edgeworth, F.Y. (1888). Memorandum by the secretary, Mr F. Y. Edgeworth, on the accuracy of the proposed calculation of index numbers. Rep. British Assoc. Adv. Sci., 1888, 188–219. (Reprint with explanatory introduction available from http://www.hetwebsite.net/het/texts/edgeworth/edgewpapers/edgew11.pdf accessed 20 March 2020).

Elliott, M.R. & Valliant, R. (2017). Inference for nonprobability samples. Stat. Sci., 32, 249–264.

Fatah, K.S. & Ahmed, S.O. (2012). Variance estimates for price changes in the consumer price index for Kurdistan Region of Iraq (January–December, 2009). J. Basrah Res. (Sci.), 38, 80–91.

Fava, V.L. (2007). A precisão dos índices de preços. Revista Economía, 8, 39–63.(in Portuguese).

Fowler, R.F. (1973). Further Problems of Index Number Construction, Studies in Official Statistics, Research Series No 5. Her Majesty's Stationery Office: London.

ILO, IMF, OECD, UNECE, Eurostat & The World Bank (2004). Consumer Price Index Manual. International Labour Office: Geneva. Available from http://www.ilo.org/wcmsp5/groups/public/---dgreports/---stat/documents/presentation/wcms_331153.pdf

Jevons, W.S. (1869). The depreciation of gold. Economist, 1869, 530–532.[also reprinted in Journal of the Statistical Society of London 32 438–463].

Johnson, D.G. (1975). The accuracy of regression based cost indices. J. R. Stat. Soc. Ser. A, 138, 411–422.

Kersten, H.M.P. (1985). Nonresponse assessment of a consumer price index. J. Business Econ. Stat., 3, 336–343.

Klick, J. & Shoemaker, O. (2019). Measures of variance across CPI populations. Proc. Joint Stat. Meet., 2019, 2242–2253. Available from https://www.bls.gov/osmr/research-papers/2019/pdf/st190080.pdf (accessed 10 January 2020).

Koop, J.C. (1986). Estimating variance of a consumer price index and some comments on inference. J. Official Stat., 2, 74–76.

Kott, P.S. (1983). Estimating the variances of price indexes by half sampling: why it works. Proc. Surv. Res. Methods Sect. Am. Stat. Assoc., 1983, 349–354.

Kott, P.S. (1984). A superpopulation theory approach to the design of price index estimators with small sampling biases. J. Business Econ. Stat., 2, 83–90.[with corrections 3 100].

Kruskal, W.H. & Telser, L.G. (1960). Food prices and the Bureau of Labor Statistics. J. Business, 33, 258–279.

Leaver, S.G. (1990). Estimating variances for the U.S. consumer price index for 1978–1986. Proc. Survey Res. Methods Sect. Am. Stat. Assoc., 1990, 290–295.

Leaver, S.G. & Cage, R.A. (1997). Estimating the sampling variance for alternative estimators of the U.S. consumer price index. Proc. Survey Res. Methods Sect. Am. Stat. Assoc., 1997, 740–745.

Leaver, S.G., Johnstone, J.E. & Archer, K.P. (1991). Estimating unconditional variances for the U.S. consumer price index for 1978–1986. Proc. Survey Res. Methods Sect. Am. Stat. Assoc., 1991, 614–619.

Leaver, S.G. & Larson, W.E. (2001). Estimating variances for a scanner-based consumer price index. Proc. Gov. Stat. Sect. Am. Stat. Assoc., 2001.

Leaver, S.G. & Larson, W. (2002). Assessing the impact of imputation on the sampling variance of the U.S. consumer price index. Proc. Survey Res. Methods Sect. Am. Stat. Assoc., 2002, 2013–2017.

Leaver, S.G. & Larson, W. (2003). Estimating components of variance of price change from a scanner-based sample. Proc. Survey Res. Methods Sect. Am. Stat. Assoc., 2003, 2318–2325.

Leaver, S.G. & Swanson, D.C. (1992). Estimating variances for the U.S. consumer price index for 1987–1991. Proc. Survey Res. Methods Sect. Am. Stat. Assoc., 1992, 740–745.

Leaver, S.G. & Valliant, R. (1995). Statistical problems in estimating the U.S. consumer price index. In Business Survey Methods, Eds. B. Cox, D. Binder, N. Chinnappa, A. Christianson, M. Colledge & P. Kott, Chapter 28, John Wiley: New York, pp. 543–566.

McCarthy, P.J. (1961). Sampling considerations in the construction of price indexes with particular reference to the United States consumer price index. In The Price Statistics of the Federal Government, Ed. G. Stigler, National Bureau of Economic Research: Washington DC, pp. 197–232.

Morgenstern, O. (1963). On the Accuracy of Economic Observations, Ed. 2. Princeton University Press: Princeton.

Moulton, B.R. (1996). Bias in the consumer price index: what is the evidence? J. Econ. Perspect., 10, 159–177.

Mudgett, B.D. (1951). Index Numbers. Wiley: New York.

Mushkudiani, N., Pannekoek, J. & Zhang, L.-C. (2020). Uncertainty measures for economic accounts. Econ. Syst. Res., 32(4), 476–501. https://doi.org/10.1080/09535314.2020.1792843

Norberg, A. (2004). Comparison of variance estimators for the consumer price index. In 8th Ottawa Group Meeting, 23–25 August 2004. Helsinki. http://www.stat.fi/og2004/norbergp.pdf (accessed 3 Mar 2020).

O’Neill, R., Ralph, J. & Smith, P.A. (2017). Inflation—History and Measurement. Palgrave MacMillan: Basingstoke. https://doi.org/10.1007/978-3-319-64125-6

Ralph, J., O’Neill, R. & Smith, P.A. (2020). The Retail Prices Index: A Short History. Palgrave Pivot: Cham, Switzerland. https://doi.org/10.1007/978-3-030-46563-6

International Statistical Review (2021)
© 2021 The Authors. International Statistical Review published by John Wiley & Sons Ltd on behalf of International Statistical Institute.
Rao, J.N.K. (2020), in press. On making valid inferences by integrating data from surveys and other sources. *Sankhyā Ser. B.* https://doi.org/10.1007/s13571-020-00227-w

Reed, S.B. & Rippy, D.A. (2012). Consumer price index data quality: how accurate is the US CPI? *Beyond Numh.*, 1, 1–8.

Shoemaker, O.J. (2002). Estimation and analysis of variance components for the revised CPI housing sample. *Proc. Surv. Res. Methods Sect. Am. Stat. Assoc.*, 2002, 3208–3212.

Shoemaker, O.J. (2003). Estimation and comparison of chained CPI-U standard errors with regular CPI-U results (2000–2001). *Proc. Surv. Res. Methods Sect. Am. Stat. Assoc.*, 2003, 3847–3853. Available from https://www.bls.gov/osmr/research-papers/2003/pdf/st030120.pdf (accessed 22 March 2020).

Shoemaker, O.J. (2009). The impact of high variances at the lowest aggregate levels on the CPI’s all-US–all-items variance. *Proc. Surv. Res. Methods Sect. Am. Stat. Assoc.*, 2009, 2835–2843.

Shoemaker, O.J. & Marsh III, F. (2011). Revising replicate selection in the CPI variance system. *Proc. Gov. Stat. Sect. Joint Stat. Meet., 2011*, 4068–4079. Available from https://www.bls.gov/osmr/research-papers/2011/pdf/st110080.pdf (accessed 30 March 2020).

Skinner, C.J. (2015). Cross-classified sampling: some estimation theory. *Stat. Probab. Lett.*, 104, 163–168.

Smith, P.A. & O’Donoghue, J. in prep. Estimating components of the variance of the UK Consumer Price Index.

Smith, P.A. & Ralph, J. (2021). Measuring the general level of prices in the UK in the long-nineteenth century: from individual innovation to state production. *Romance, Revolution & Reform*, 3, 10–35. https://www.rrrjournal.com/issue-3-general-level-of-prices

Valliant, R. (1991). Variance estimation for price indexes from a two-stage sample with rotating panels. *J. Business Econ. Stat.*, 9, 409–422.

Valliant, R. (1992). Smoothing variance estimates for price indexes over time. *J. Official Stat.*, 8, 433–444.

Valliant, R. (1999). Uses of models in the estimation of price indexes: a review. *Proc. Surv. Res. Methods Sect. Am. Stat. Assoc.*, 1999, 94–102.

Valliant, R. & Miller, S. (1989). A class of multiplicative estimators of Laspeyres price indexes. *J. Business Econ. Stat.*, 7, 387–394.

von Hofsten, E. (1959). Price indexes and sampling. *Sankhyā*, 21, 401–403.

Weber, W.L. (1980). A system of variance estimation for the U.S. consumer price index. *Proc. Surv. Res. Methods Sect. Am. Stat. Assoc.*, 1980, 628–633.

Wilkerson, M. (1964). Measurement of sampling error in the consumer price index: first results. *Proc. Business Econ. Stat. Sect. Am. Stat. Assoc.*, 1964, 220–30.

Wilkerson, M. (1967). Sampling error in the consumer price index. *J. Am. Stat. Assoc.*, 62, 899–914.

Williams, E. (2006). The effects of rounding on the consumer price index. *Monthly Labor Rev.*, 129(10), 80–89.

Wolter, K. (2007). *Introduction to Variance Estimation*, Ed. 2. Springer Science and Business Media: New York.

Wynne, M.A. & Sigalla, F.D. (1994). The consumer price index. *Fed. Reserve Bank Dallas Econ. Rev.*, 2, 1–22.

Zhang, L.-C. (2010). A model-based approach to variance estimation for fixed-weights and chained price indices. In *Official Statistics: Methodology and Applications in Honour of Daniel Thorburn*, Eds. M. Carlson, H. Nyquist & M. Villani, Stockholm University & Statistics Sweden: Stockholm, pp. 149–166.

Zhang, L.-C. (2019). On valid descriptive inference from non-probability sample. *Stat. Theory Related Fields*, 3, 103–113.

[Received April 2020, Revised December 2020, Accepted January 2021]

**Supporting Information**

Supporting information may be found in the online version of this article.