Herzberg, Julika; Knetsch, Thomas A.; Schwind, Patrick; Weinand, Sebastian

Quantifying Bias and Inaccuracy of Upper-Level Aggregation in the Harmonised Index of Consumer Prices for Germany and the Euro Area

Review of Income and Wealth

Provided in Cooperation with:
John Wiley & Sons

Suggested Citation: Herzberg, Julika; Knetsch, Thomas A.; Schwind, Patrick; Weinand, Sebastian (2022) : Quantifying Bias and Inaccuracy of Upper-Level Aggregation in the Harmonised Index of Consumer Prices for Germany and the Euro Area, Review of Income and Wealth, ISSN 1475-4991, Wiley, Hoboken, NJ, Vol. 69, Iss. 3, pp. 581-605, https://doi.org/10.1111/roiw.12602

the review of income and wealth

Review of Income and Wealth
Series 69, Number 3, September 2023
DOI: 10.1111/roiw.12602

QUANTIFYING BIAS AND INACCURACY OF UPPER-LEVEL AGGREGATION IN THE HARMONISED INDEX OF CONSUMER PRICES FOR GERMANY AND THE EURO AREA

BY JULIKA HERZBERG, THOMAS A. KNETSCH, PATRICK SCHWIND

AND

SEBASTIAN WEINAND*

Deutsche Bundesbank, Directorate General Statistics

Current measurement practices of the Harmonised Index of Consumer Prices (HICP) produce an upward bias of about one-ninth of a percentage point in German inflation due to changing consumption being disregarded and preliminary data being used in the compilation of expenditure weights. The statistical uncertainty produced by these sources of mismeasurement can be illustrated by an interdecile range of about one-quarter of a percentage point. The annual updating of the quantity component of the weights, implemented in 2012, has reduced the substitution component, making the disregard of changing consumption virtually a non-issue for the euro area HICP. The measurement of the German HICP is impaired by the extrapolation of expenditure weights. The use of preliminary national accounts data since 2012 has not led to an improvement. This source of mismeasurement is likely to be relevant for the euro area HICP as well but cannot be quantified due to data constraints.

JEL Codes: C43, E31

Keywords: inflation measurement, substitution bias, updating of expenditure weights, HICP

1. INTRODUCTION

The Harmonised Index of Consumer Prices (HICP) attracts much attention as a measure of inflation in Europe. With any consumer price index (CPI), the HICP shares the property of proneness to measurement error (ILO, IMF, OECD, UNECE, Eurostat, World Bank, 2020, Chapter 12). This is properly taken into account by users. A prominent example is the European Central Bank (ECB)'s

Note: The authors thank Andreas Dietrich for testing the seasonality and providing seasonal and calendar factors. The comments of Ludwig von Auer, Edgar Brandt, Martin Eiglsperger, Elisabeth Falck, Annette Fröhling, Jens Mehrhoff, Jan Menz, Johannes Hoffmann, and Elisabeth Wieland are gratefully acknowledged. The authors are greatly indebted to Mick Silver for his extensive review. The authors also thank an anonymous referee and the editor for their helpful comments.
Disclaimer: The authors accept the responsibility for any shortcomings. The views expressed in this paper are solely those of the authors and should not be interpreted as reflecting the views of the Deutsche Bundesbank or the Eurosystem.
*Correspondence to: Sebastian Weinand, Deutsche Bundesbank, Directorate General Statistics, Wilhelm-Epstein-Strasse 14, 60431 Frankfurt am Main, Germany (sebastian.weinand@bundesbank.de).

monetary policy target. In its 2021 monetary policy strategy statement, the ECB's Governing Council confirmed that the HICP "remains the appropriate price measure for assessing the achievement of the price stability objective" and considered price stability "best maintained by aiming for two per cent inflation over the medium term" (ECB, 2021c). While the deliberations in the recent strategy review took account of the interrelations between the equilibrium real interest rate, the effective lower bound of the nominal interest rate and the inflation buffer, those factors calling for a sufficient inflation buffer, which had been already identified in the 2003 strategy review, were reaffirmed (ECB, 2021a). In particular, the ECB is still committed to providing a safety margin against deflation risks, considering regional inflation differentials and potential measurement bias (ECB, 2004, p. 51).

Measurement issues arise from a partial conflict of interests: namely that price indices should ensure like-for-like comparisons over time, whereas changes in supply and demand conditions should adapt as comprehensively and in as timely a manner as possible. The HICP measurement rules, inter alia, prescribe viable and Europe-wide harmonized solutions for aggregating individual prices and updating the basket of goods and services as consumption patterns change and/or new products are introduced, as well as for the implementation of new distribution channels and amended product characteristics, with the caveat that the HICP must be released promptly on a monthly basis to fulfill its policy purposes.

Of the main potential sources of HICP measurement bias, it is the impact of quality changes, new products, and new outlets that is almost impossible to quantify without access to micro price data.[1] This is also true of lower-level aggregation, i.e., the aggregation of prices at product levels for which expenditure weights are not available.[2]By contrast, changes in consumption behavior over time and its relation to variations in relative prices are an issue for HICP measurement at the upper level of aggregation. These effects can be studied using publicly available disaggregate price indices and corresponding information about expenditure weights. The bias induced by disregarding substitution has been widely discussed for a fixed-basket price index that aims to approximate a cost-of-living index (COLI). However, it is also relevant for the HICP, though it is conceptualized as a cost-of-goods index (COGI).[3] The HICP is a chain-linked Laspeyres-type index based on weights that are annually "updated to make them representative for the [previous calendar] year" (EU, 2020, Art. 3, 1. (b)).

This paper is aimed at providing insights into how far the HICP meets the criterion of representativity, allowing it to be regarded as sufficiently reliable.[4] In particular, the study reveals empirical evidence on the substitution (or

[1]See ECB (2014, pp. 40–42) for an overview of potential measurement issues in consumer price indices.

[2]The lower-level aggregation yields elementary price indices. These are also subject to potential bias. A recent study on elementary index bias is Gábor-Tóth and Vermeulen (2019), for instance.

[3]The HICP manual explicitly states that "it measures the changing cost of a *fixed basket of products* at different sets of prices over time"(Eurostat, 2018, Section 2.2.1, italics in original).

[4]In this context, reliability requires that "a price index should be as accurate as possible in its measurement of price movements and should not be subject to significant bias" (Camba-Mendez, 2003, p. 33). According to the 2003 review of the ECB's monetary policy strategy, reliability is among the

representativity) bias in the HICP.[5] Yet the bias is not the only criterion to be scrutinized here. The scope is widened to include inaccuracy. In general, this criterion is meant to measure uncertainty surrounding the HICP figures as a result of any source of error (e.g., sampling variability, lacking information, simplifying assumptions, or compilation practices). In this context, the focus is on the inaccuracy resulting from expenditure weights estimated using preliminary national accounts data.

The evaluation is designed to measure both the bias and inaccuracy of the HICP against a superlative price index with full-information expenditure weights. Admittedly, this benchmark is a tough criterion. The data needed to calculate this benchmark are only available with a significant delay. Full-information expenditure weights can therefore only be calculated retrospectively. In addition, it is not guaranteed that the benchmark will reflect the "true" aggregate price development. Compared with the HICP, however, it is considered to be closer to the unobservable "truth," given that it is based on more, and, particularly at the point in time when the HICP is compiled, unknown information. In the specific HICP context, the bias and inaccuracy measures can be decomposed into a *substitution component* (i.e., official versus superlative index formula) and a *data vintage component* (i.e., official or real-time versus full-information weights).

In the terminology of index number theory, the HICP is a fixed-basket price index according to Lowe (1823) because price and weight reference periods do not coincide. Balk and Diewert (2012) and Armknecht (2015) are recent examples making the case for this nomenclature. With the annual updating of weights, the HICP differs from a multi-year fixed-basket Laspeyres price index, like the national CPI in Germany. The analytical framework of this study is enlarged for the purpose of making comparisons between these index types as well. In this setup, the decomposition consists of three factors where the additional factor is called *annual updating component*.[6]

The theory and practice of CPI measurement have been developed in line with the steady and intense discourse among compilers, users, and scholars over decades. In this context, the US CPI has by far served as the main subject of research. Looking at the more recent past, it might come as a surprise that the HICP has not been selected as often for methodological and empirical work. Its relevance for policymakers in Europe is indisputable, and it has by now amassed a sufficiently long data history. In addition, researchers have recently paid more attention to weighting and updating procedures than to index formulae and their theoretical underpinning. While the latter may be considered inapplicable to the HICP, the former are key for the enhancement of the upper-level aggregation practices of any CPI.

criteria that a price index used as a target for monetary policy must fulfill. The others are credibility, comparability, periodicity, and timeliness as well as consistency with the European Union Treaty.

[5]The term "representativity bias" is conceptually more appropriate than "substitution bias" but is not used in this paper as the emphasis is intended to be on the economic substance of the phenomenon rather than statistical technicalities.

[6]The purely price-updated weights are different from the full-information weights used in the benchmark price index because the full-information set comprises previous and later household budget surveys and revised or even final national accounts.

In this paper, bias and inaccuracy of upper-level aggregation are measured in the HICP for Germany and the euro area. Ideally, the results for the euro area would certainly have been given more attention, because the monetary policy objective refers to inflation in the currency union as a whole. Due to data constraints, however, the examination for the euro area is limited to the substitution component. The full-fledged analysis is carried out for the German HICP. All results are documented for the all-items HICP, covering the monthly year-on-year rates from January 1997 to December 2019. The calculations are based on the price index series of the product groups or classes and the respective series of expenditure weights.[7] The set of 83 price series since the year 2000 and 78 before 2000 is the most disaggregated level to approach with publicly available price data. It generally matches the degree of detail in comparable studies and accounts for the conclusion drawn by previous research that the higher the number of disaggregate price indices included, the more meaningful evidence on the substitution bias tends to become (Manser and McDonald, 1988).

The investigation of the HICP bias and inaccuracy before and after 2012 is of particular interest. In January 2012, a major change in measurement practices came into force. While weights had only been updated on the basis of price information until then, the new regulation prescribed the use of detailed household expenditure data from preliminary national accounts. This innovation in HICP measurement was introduced with the aim of mitigating the substitution bias (ECB, 2012). However, owing to the recourse on preliminary data and the impossibility to incorporate later revisions, it entails the risk of impairing accuracy.[8]

The remainder of this paper is structured as follows. In the next section, the related empirical literature on upper-level aggregation issues is briefly summarized. In Section 3, the evaluation framework is sketched out. Section 4 gives an overview of the results for Germany and the euro area. In Section 5, the results are put into a broader perspective, also discussing potential implications for HICP compilers and users. In Section 6, conclusions are drawn, and possibilities for future research are discussed.

2. Literature

Theoretical and empirical research in price statistics has dealt with aspects of upper-level aggregation issues in CPI measurement for quite a long time. The subject has been influenced substantively by the theory of index numbers.

[7]This refers to the 3-digit or 4-digit level of the (E)COICOP classification. See Eurostat (2018), Chart 3.1, for instance.

[8]The annual updating of weights tends to reduce the substitution bias because the distance between the current period and the base period is shortened to a minimum. The updating requires the latest available national accounts data to be used, while comprehensive household budget surveys (as the primary source of weights) are only conducted at multi-year intervals (Eurostat, 2018, Section 3.5). At the time of their incorporation into HICPs, the figures on the household expenditures broken down by consumption purpose are compiled on the basis of incomplete information. While national accounts are later revised by incorporating delayed information, HICP weights remain fixed because national accounts revisions are not considered a reason to adjust weights (Eurostat, 2018, Section 10.4.4).

Diewert (1976)'s seminal work on superlative indices paved the way for renewed interest in CPI measurement toward the end of the last century. The pitfalls of measurement principles at the upper level of aggregation have been both studied in specific individual contributions and included in broad-based empirical assessments of CPI compilation in all steps of its production chain. The most notable landmark study is probably the report of the Boskin Commission, formally called the "Advisory Commission to Study the Consumer Price Index." The commission estimated the total bias of the US CPI to be about 1.1 percentage points per annum, only 0.15 of which was due to upper-level substitution, 0.25 due to lower-level substitution, 0.1 due to outlet substitution and 0.6 due to new products and quality change (Boskin *et al.,* 1998, Table 1).

Lebow and Rudd (2003) surveyed the literature on the US CPI measurement bias and produced new results. According to their estimates, the upper-level substitution bias was about 0.3 percentage point. This amounts to a doubling in terms of percentage points and even more of a difference in relative terms, because the sum of the sources of measurement bias totaled 0.9 percentage point. The present study is similar to that of Lebow and Rudd (2003) in two respects. First, they explicitly addressed the impact of weighting on CPI measurement, though they stressed the role of different sources (consumer expenditure survey versus personal consumption expenditure) rather than the reporting status (vintages) of the data used for the derivation of weighting schemes, which is the focus of this paper. Second, Lebow and Rudd (2003) proposed a formal decomposition of the bias resulting from the difference between the published CPI and the true COLI. As in this paper, they factored out the impacts of the index formula and the weighting schemes.

Greenlees and Williams (2010) reconsidered the upper-level aggregation bias of the US CPI after the time interval for the updating of weights had shortened from 10 years to 2 years (taking effect in January 2002). Although the more frequent adjustment of expenditure weights was expected to reduce the upward bias, they did not find an improvement vis-à-vis Lebow and Rudd (2003)'s result of 0.3 percentage point per annum. Two-thirds of the total bias were due to the price-updating (i.e., the difference between the price-updated CPI and a true Laspeyres index), whereas the Laspeyres-Törnqvist difference accounted for one-third. By contrast, Armknecht and Silver (2014) proved that the measurement bias of the post-2002 CPI amounted to 0.16 percentage point per annum, thus confirming the Boskin Commission's estimate of the upper-level aggregation bias.

Silver and Ioannidis (1994) analyzed the (mis-)measurement of nine European CPIs. The similarities between their work and this paper are not limited to the fact that the substitution component is measured by the difference between a Laspeyres index and a Törnqvist index. Under the notion of "untimely weights" (i.e., weights compiled on the basis of outdated information from household budget surveys conducted only at intervals of several years), Silver and Ioannidis (1994) addressed a source of mismeasurement that is close to the data vintage component in this paper.[9] In contrast to the overwhelming majority of the literature at that time, they looked

[9]Silver and Ioannidis (1994, p. 552) suggest calculating indices "using survey period weights in the base period to which they relate, as opposed to when they come available." This is one idea considered in

not only at the bias (or mean deviation) but also the mean absolute deviation and the root mean squared error—another similarity their work shares with this paper.

In the late 1990s, the report of the Boskin Commission prompted some research on CPI measurement outside the US. The other countries where researchers, statisticians, and/or central bank economists were engaged with this topic at the time included Canada (Crawford, 1998) and the UK (Cunningham, 1996; Baxter, 1997), as well as with Germany (Hoffmann, 1998), France (Lequiller, 1997), and Portugal (Neves and Sarmento, 1997), some countries which are now part of the euro area. Since then, interest in this topic has decreased somewhat. In particular, no broad-based attempt to study HICP mismeasurement has been made so far. The ECB (2014, p. 42) concluded that "it [was] not possible to estimate measurement bias in the euro area HICP."

3. Methodology

According to EU (2016), the HICP is an annually chain-linked Laspeyres-type index. In Art. 2 (14) of the same regulation, a Laspeyres-type index is defined as a Lowe index:

$$P^{o}_{\text{HICP}}(y,m) = \sum_{i=1}^{I} w_i^{o}(y-1,12) \times \frac{p_i(y,m)}{p_i(y-1,12)}\,, \tag{1}$$

where $p_i(y,m)$ is the price of good i ($i = 1, \ldots, I$) in year y ($y = 1, \ldots, Y$) and month m ($m = 1, \ldots, 12$). According to further legal specification (EU, 2020), the weight reference period is the previous year, implying that expenditure weights "are adjusted to reflect the prices of the price reference period" (EU, 2016, Art. 2 (14)) which is December of the previous year $(y-1,12)$. Therefore, the weight of the official index is $w_i^{o}(y-1,12) = w_i(y-1) \times p_i(y-1,12)/p_i(y-1)$ (superscript o for "official"), where $w_i(y-1)$ and $p_i(y-1)$ indicate the average expenditure share and price of good i in year $y-1$.

In measurement practice, however, quantity information has often been more outdated than formally prescribed by this regulation because national accounts as a major source for the derivation of weights are available only until $y-2$ when updates are made. From 2012 to 2020, the weights of the German HICP have been compiled according to the formula: $w_i^{o}(y-1,12) = w_i(y-2) \times p_i(y-1,12)/p_i(y-2)$, suggesting that the weight reference year is $y-2$ de facto (see Online Appendix B for details).

Nonetheless, the quantity information in expenditure shares has been more up-to-date since 2012 than in the pre-2012 period when the weights of the German HICP were calculated using the quantity information of the national CPI's base year by. Therefore, the pure price updating of expenditure weights bridged over a longer time span then, namely $w_i^{p}(y-1,12) = w_i(by) \times p_i(y-1,12)/p_i(by)$ (superscript p

the compilation of full-information weights in this paper. The other general similarity is the use of interpolation techniques to derive weights in years between two surveys though Silver and Ioannidis (1994) do not rely on national accounts data for that purpose.

for "price-updated") with $y - 2 > by$. To examine the effect of the 2012 methodological change, it is worth compiling Lowe indices according to the pre-2012 updating practice, denoted by $P_L^p(y,m)$, and compare them with the official HICP, $P_L^o(y,m)$, since 2012. The deviation is the *annual updating effect*. Before 2012, it is zero by construction, because official weights in the HICP were purely price-updated, i.e., $w_i^o = w_i^p$.

The core of the empirical analysis is the comparison between the official HICP and a measure of "true" inflation. As the "true" inflation is unknown, it is necessary to choose a reference which is also a price index, but which exhibits characteristics that give rise to a close proximity to the "truth." Let the reference be defined by a superlative price index with full-information weights denoted by P_S^f.

The choice of a superlative index means that "true" inflation is assumed to be best proxied by a COLI, which turns out to be at odds with the HICP concept. However, as superlative indices not only have a sound underpinning in economic theory but also fulfill symmetry (i.e., an equal-handed treatment of prices and quantities in both periods of the price comparison), they are "likely to be seen as desirable, even when the CPI is not meant to be a cost of living index" (ILO, IMF, OECD, UNECE, Eurostat, World Bank, 2020, para. 1.151). Of the set of superlative indices, the Törnqvist, Fisher, and Walsh indices are the standard candidates chosen (see Online Appendix A for details). Eventually, the choice proves to be of marginal empirical importance. Therefore, in the following section, only the results for the Törnqvist index are reported. The results for the Fisher and Walsh indices are presented in Online Appendix A.

The full-information weights are constructed on the basis of the most comprehensive information set available. For the German HICP, this means that, for every year in which a household budget survey was conducted (e.g., 2000, 2005, 2010, and 2015), full-information weights are derived from this most detailed and reliable information. For the years in between, the weights are interpolated using the latest vintage of national accounts data (see Online Appendix B for details). Before 2000 and from 2016 onwards, the extrapolation implies that the weights in these subperiods are less reliable than the interpolated weights. Toward the end of the time span considered in this paper, national accounts have only been partially revised, implying that the full-information weights are likely to be systematically closer to the official HICP weights.

The deviation of the HICP from the reference may be expressed by the following ratio:

$$\frac{P_L^o}{P_S^f} = \frac{P_L^o}{P_S^o} \times \frac{P_S^o}{P_S^f}, \tag{2}$$

where P_S^o denotes a superlative index with the official HICP weights. The decomposition makes it possible to separate the *substitution effect* P_L^o/P_S^o from the *data vintage effect* P_S^o/P_S^f.

The factoring can be refined further by taking on board the *annual updating effect* P_L^p/P_L^o. In precise terms, the three-factor decomposition

$$\frac{P_L^p}{P_S^f} = \frac{P_L^p}{P_L^o} \times \frac{P_L^o}{P_S^o} \times \frac{P_S^o}{P_S^f} \tag{3}$$

allows for a rigorous analysis of the trade-off between an annual update of weights (potentially suffering from limited reliability induced by the use of preliminary national accounts data) and a multi-year fixed-basket weighting scheme (reflecting the most reliable and detailed information on consumption behavior).

Bias and inaccuracy metrics need to be interpretable in percentage points per annum. For this purpose, monthly year-on-year price relatives are aggregated.[10] For the two Lowe indices (with official weights and purely price-updated weights), this means

$$P_L^x(y,m) = \sum_{i=1}^{I} w_i^x(y-1) \times \frac{p_i(y,m)}{p_i(y-1,m)}\,, \quad x = o,p\,, \tag{4}$$

and for the superlative Törnqvist index with full-information and official weights

$$P_S^x(y,m) = \prod_{i=1}^{I} \left[\frac{p_i(y,m)}{p_i(y-1,m)}\right]^{\frac{1}{2}\left[w_i^x(y-1)+w_i^x(y)\right]}, \quad x = f,o\,, \tag{5}$$

where $w_i^x(y-1) \equiv w_i^x(y-1,12)$, for notational convenience. Consequently, $P_L^x(y,m)$ and $P_S^x(y,m)$ represent monthly factors measuring the approximate year-on-year changes of the aggregate price indices.

The performance of HICP measurement is assessed in terms of bias and inaccuracy. The mean deviation (MD) quantifies the measurement bias in the period under analysis. It is defined by

$$MD = \frac{1}{T}\sum_{t=1}^{T} \ln\left(P_L^x(t)/P_S^f(t)\right) \quad x = o,p\,. \tag{6}$$

[10]Following the construction principle of the HICP in Equation (1), price relatives referring to December of the previous year are aggregated and then chained. This procedure can be applied retrospectively to different index methods, resulting also in different year-on-year percentage changes that could be used for bias and inaccuracy analysis. For the superlative indices, these year-on-year percentage changes, however, actually rely not only on the weights of the current and the previous period (y and $y-1$), but also on the weights of the period before the previous one ($y-2$). Therefore, metrics do not consistently fit the definition of superlative indices that use current and previous period weights only. In the present analysis, monthly year-on-year price relatives (instead of December price relatives) are therefore already derived at the disaggregate level. The weighted aggregates of these price relatives consistently stick to the definition of the corresponding indices. Moreover, they are scaled as aggregate year-on-year percentage changes that are consistently comparable over time. Nevertheless, two drawbacks of this approach should be kept in mind. First, it is not possible to "back-transform" these year-on-year price relatives into a throughout index. Second, the year-on-year percentage changes of the Lowe price index with official weights do not completely replicate officially published HICP rates. However, the observed deviations are on average small.

The logarithmic transformation of the price index ratios $P_L^x(t)/P_S^f(t)$ ensures that *MD* may be interpreted as the measurement bias measured as a percentage of the "true" price index. Inaccuracy is measured by the mean squared deviation

$$MSD = \frac{1}{T}\sum_{t=1}^{T}\left[\ln\left(P_L^x(t)/P_S^f(t)\right)\right]^2 \quad x = o, p \tag{7}$$

and the root mean squared deviation, $RMSD = \sqrt{MSD}$. *RMSD* reflects the uncertainty entailed by the year-on-year HICP rate of change in percentage points.[11] The interdecile range, $IDR = P_{90} - P_{10}$, which reports the data range between the 10th and 90th percentiles, and the interquartile range, $IQR = P_{75} - P_{25}$, serve as additional measures of dispersion in the accuracy analysis. A significant advantage of both measures is their robustness against outliers (Welch, 2001).

The logarithmic transformation of price index ratios in Equations (6) and (7) allows an additive decomposition of both the mean deviation and the mean squared deviation. In precise terms, *MD* can be decomposed by plugging Equation (3) into (6), resulting in

$$MD = \frac{1}{T}\sum_{t=1}^{T}\underbrace{\ln\left(P_L^p(t)/P_L^o(t)\right)}_{=u(t)} + \frac{1}{T}\sum_{t=1}^{T}\underbrace{\ln\left(P_L^o(t)/P_S^o(t)\right)}_{=s(t)} + \frac{1}{T}\sum_{t=1}^{T}\underbrace{\ln\left(P_S^o(t)/P_S^f(t)\right)}_{=v(t)},$$

$$= \bar{u} + \bar{s} + \bar{v}, \tag{8}$$

where $\bar{u} = \frac{1}{T}\sum_{t=1}^{T} u(t)$ is the annual updating component, $\bar{s} = \frac{1}{T}\sum_{t=1}^{T} s(t)$ the substitution component, and $\bar{v} = \frac{1}{T}\sum_{t=1}^{T} v(t)$ the data vintage component. Similarly, *MSD* can be decomposed by plugging Equations (3) into (7), yielding

$$MSD = \frac{1}{T}\sum_{t=1}^{T}u^2(t) + \frac{1}{T}\sum_{t=1}^{T}s^2(t) + \frac{1}{T}\sum_{t=1}^{T}v^2(t) + COV\,, \tag{9}$$

where $COV = 2\left(COV_{su} + COV_{sv} + COV_{uv}\right)$ is the sum of covariance terms.[12]

4. Results

The bias and inaccuracy metrics refer to inflation, i.e., the year-on-year HICP percentage change, and are expressed in percentage points. The analysis is carried out on the basis of monthly observations. For the whole period under analysis, the euro area HICP covers the 19 countries that currently make up the currency union

[11] For the formation of an uncertainty interval, considering that the bias is typically nonzero, (sample) variance and standard deviation are to be favored, with the latter being measured in percentage points. Both measures can be found in Online Appendix A.

[12] A decomposition by Equation (2) results in $MD = \bar{s} + \bar{v}$ and $MSD = \frac{1}{T}\sum_{t=1}^{T} s^2(t) + \frac{1}{T}\sum_{t=1}^{T} v^2(t) + 2COV_{sv}$.

(fixed composition) to avoid statistical breaks due to changes in the territorial coverage. In total, 276 monthly observations are available.

The HICP has changed in terms of coverage and measurement standards. Particularly in the initial period of the HICP, changes were implemented with a higher frequency. In 2000 and 2001, for instance, its coverage was extended and further harmonized by including expenditure in the areas of health, education, social production services and insurance, as well as hospital services and some services within homes (Destatis, 2000; ECB, 2000; 2001). These extensions increased the HICP coverage by approximately 5 percent of household consumption expenditure, in total. While the analysis pays particular attention to the 2012 methodological change by, for instance, referring to split samples, an explicit treatment of the 2000 and 2001 changes is not provided. Therefore, their potential impact should be borne in mind when interpreting results (see also Online Appendix B).

For the German HICP, it is feasible to compile a superlative price index with full-information weights. The full analysis comprises the measurement of deviations between the official HICP and this benchmark as well as between factors, which reflect data vintage and substitution effects, and the impact of annual updating. The results will be shown in the first part of this section. The second part is restricted to a comparison of the substitution components of the German and euro area HICPs. Data gaps hamper the construction of full-information weights for the euro area HICP and, therefore, the calculation of the data vintage component.

4.1. Full-Fledged Analysis of German HICP

Evolution of Deviations Over Time

Figure 1 displays the logarithmic deviation according to Equation (2) and its decomposition into two components in the period of January 1997 to December 2019. The monthly deviations range from −0.2 to about 0.5 percentage point. The overwhelming number of realizations is in positive territory. Exceptions are found for single months or rather short periods in 1998, 2015, 2016, and 2019 as well as during a major part of the year 2000.

The substitution component is positive in almost all the realizations. In the very few cases where a negative value occurs, the deviation is very small in absolute value. The dominance of positive substitution components does not come as a surprise because the sign is expected in line with the theory of consumer substitution.

The monthly series of the data vintage component bears more striking features. Except 1999 and 2000 the monthly deviations are almost entirely positive; their size is relatively large. In addition, clusters of substantial realizations such as those in 2008, 2009, and 2015 turn out to be associated with larger realizations of the other effects. This is particularly detrimental, as the partial components of the total deviation tend to have a reinforcing rather than a compensating effect.

The 2008 and 2009 cluster of large positive realizations of both the substitution and the data vintage components might be related to strong relative price shifts induced by the sharp economic recession in that year. A fixed-basket price index tends to perform worse under these circumstances than in "normal" times. The real-time compilation of weights entails a higher risk of mismeasurement in terms of more substantial data vintage components. The methodologically built-in

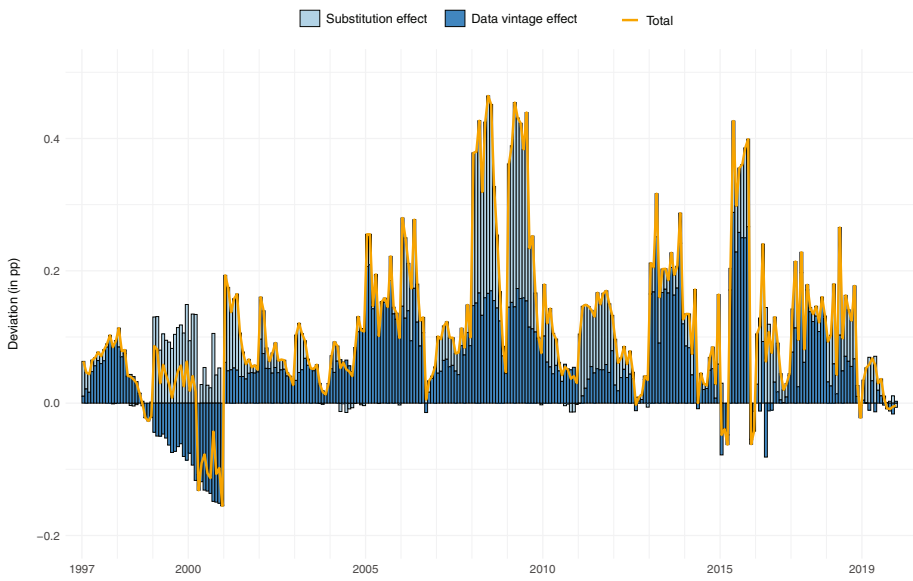

Figure 1. Monthly Deviations of German HICP According to Equation (2) [Colour figure can be viewed at wileyonlinelibrary.com].

dependence on past consumption patterns (or lagged adjustment to a rapid change) is the underlying source of error in both cases.

By contrast, a statistical break is the reason for the very large data vintage components in 2015. According to HICP rules, the package holiday prices are chain-linked via December, producing severely distorted year-on-year HICP rates for package holidays during the year 2015 on account of the seasonal pattern changing due to adjusted measurement practices (Dietrich *et al.,* 2021; Deutsche Bundesbank, 2019a). In purely arithmetical terms, the interplay between a tremendously increased price change (an average of 16.5 percent in 2015 after the extraordinary revision compared with −0.3 percent before the revision) and the substantial difference between the official HICP weight (3.7 percent) and the full-information weight (2.8 percent) contribute significantly to the data vintage effect.[13]

The monthly deviations shown in Figure 1 are far from being normally distributed around a positive mean. The histogram in Figure 2 reveals that a comparatively large share of probability mass is located around the central moments of the distribution. A striking feature of the empirical distribution is the cluster of very large realizations. As detailed below, these can be regarded in part, as being justified from an economic point of view, but can also partly be seen as outliers in the sense of obvious mismeasurement. The empirical distribution of the monthly deviations are markedly skewed to the right (sample skewness 1.060) and leptokurtic (sample excess kurtosis 1.581). The monthly deviations turn

[13]By a back-of-the-envelope calculation, the distortionary effect can be estimated at $16.5 \times (3.7 - 2.8)/100 = 0.15$ percentage points, which is more than the average data vintage component in 2015. Without the distortionary effect, the average data vintage component in 2015 would thus be slightly negative.

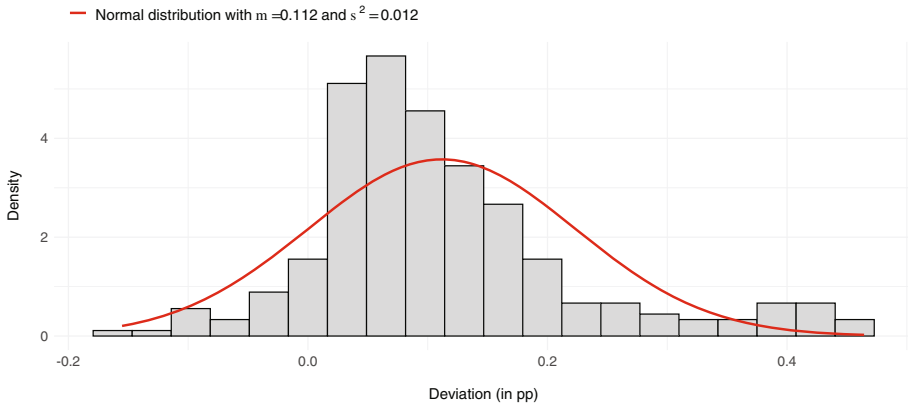

Figure 2. Distribution of Monthly Deviations of German HICP According to Equation (2), 1997–2019 [Colour figure can be viewed at wileyonlinelibrary.com].

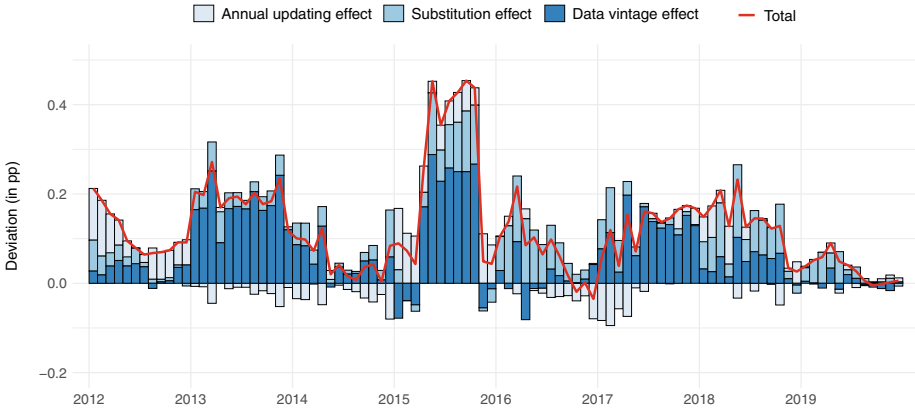

Figure 3. Monthly Deviations of German HICP According to Equation (3) [Colour figure can be viewed at wileyonlinelibrary.com].

out to be serially correlated. An autoregressive moving-average (ARMA) model that properly captures the serial correlation structure of deviations according to Equation (2) is presented in Online Appendix C.[14]

Figure 3 displays the logarithmic deviation according to Equation (3) and its decomposition into three components in the subperiod since January 2012. The add-on of the annual updating component does not alter the main observations regarding the time series of the total deviation. This is due to the fact that positive and negative realizations of the annual updating component are more or less equal in numerical terms and small overall in absolute value. Three clusters of visible realizations are identified, being in positive territory in 2012 and 2015 and negative between mid-2016 and mid-2017.

[14]An ARMA structure modeling the serial correlation of the deviations according to Equation (2) differs only slightly in terms of estimated coefficients, but does not change with regard to lag orders.

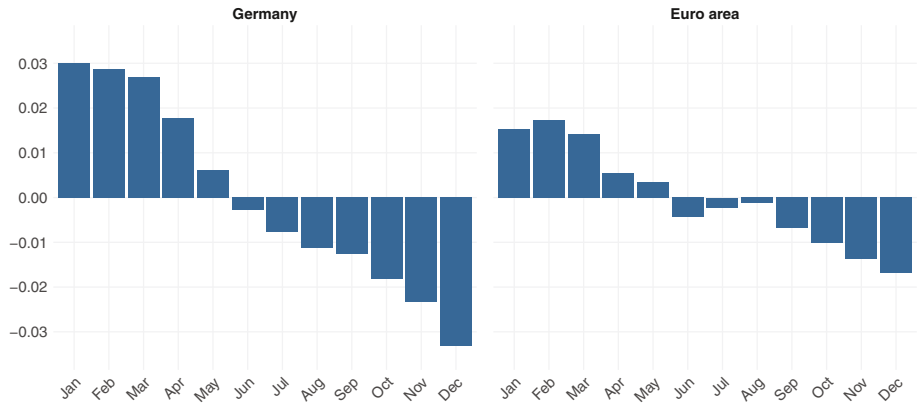

Figure 4. Monthly Averages of Estimated Seasonal Factors for Substitution Components in German and Euro Area HICPs, 1997–2019 [Colour figure can be viewed at wileyonlinelibrary.com].

Visual inspection suggests that a seasonal pattern appears in the substitution component. Formal tests for the period from January 1997 to December 2019 confirm that seasonal and calendar effects are present in the substitution component, whereas they are absent in the data vintage and annual updating components. The seasonal factors tend to decline from January to December (see Figure 4). This makes sense against the backdrop that the longer the distance between the base and the current period of the price index, the larger the substitution bias is expected to be.[15]

Bias

Considering the period from January 1997 to December 2019, the official year-on-year HICP rates of change are, on average, about one-ninth of a percentage point higher than the annual percentage changes of the full-information Törnqvist price index. Strictly speaking, this follows on from the two-factor decomposition "Equation (2)" in Table 1. The three-factor decomposition does not yield another conclusion either, given that the effect of annual updating on the mean deviation is positively signed (as expected) but marginal in magnitude. The substitution component and the data vintage component are positive, too. For the former, this is theoretically to be expected, as the Törnqvist formula accounts for adjustments in consumption related to changes in relative prices while the Lowe index does not. However, the theory makes no predictions with regard to the sign of the data vintage effect. The evidence shows that the data vintage effect exhibits a positive sign and is about the same size as the substitution effect.

The results for the whole period mask differences between the mean deviation before and after the methodological change in 2012. The introduction of the

[15]It is worth recalling that substitution should not be interpreted literally, as since 2012 the quantity adjustment of official weights lags two years behind and, thus, does not reflect output responses to relative price movements in the current year. Before 2012, official weights were only price-updated. However, the representativity of base year weights generally weakens the longer the distance between the base and the current period.

TABLE 1
MD AND *RMSD* FOR GERMAN HICP, 1997–2019

Metric	Period	Total		Components		
		Equation (2)	Equation (3)	$\bar{v}$	$\bar{s}$	$\bar{u}$
MD	Before 2012	0.111	0.111	0.053	0.059	0.000
	Since 2012	0.113	0.122	0.068	0.044	0.010
	Total	0.112	0.115	0.058	0.054	0.003
RMSD	Before 2012	0.160	0.160	0.091	0.093	0.000
	Since 2012	0.154	0.161	0.109	0.062	0.049
	Total	0.158	0.160	0.098	0.083	0.029

Notes: *MD* measured as a percentage of the Törnqvist price index with full-information weights; *RMSD* in percentage points per annum.

annual updating of weights only marginally affects the upper-level aggregation bias, amounting to about one-ninth of a percentage point before and after 2012. The substitution effect declines slightly from 0.06 to 0.04 percentage point. Before 2012, the annual updating effect is zero, because official HICP weights were purely price-updated at that time. Therefore, the pre-2012 substitution component measures the bias induced by the non-adjustment of consumption patterns over 5 years. It could be argued that a fair comparison should counteract this effect with the sum of the substitution and annual updating components in the post-2012 period. Yet the two components, amounting to 0.05 percentage point, are almost the same size as the pre-2012 substitution component.

The positive sign of the annual updating component confirms the hypothesis that the more timely the weights, the less the Lowe index suffers from overstating "true" price developments. Its small magnitude, however, might come as a surprise. The data vintage component and the substitution component are predominant sources of measurement bias both before and after 2012. While the introduction of the annual updating increased the detrimental contribution of the data vintage component from 0.05 to 0.07 percentage point, it has also gained significance in relative terms as it accounts for almost 56 percent of the total bias in the post-2012 period.

Inaccuracy

The *RMSD* is about one-sixth of a percentage point when the entire time span is analyzed. Splitting it into pre-2012 and post-2012 subperiods reveals that the adjustment in the updating procedure of weights has hardly any effect on the *RMSD*.

As the mean deviation is shown to be positive, the root mean squared deviation cannot be taken to form an uncertainty interval around the HICP rates; instead, the variance and the standard deviation are used, taking account of the mean deviation. The standard deviations are substantially smaller (see Online Appendix A for the results).

Since 2012, the data vintage component makes the largest contribution to inaccuracy. The *RMSD* of this component is about one-tenth of a percentage point over the whole period. In the post-2012 phase, it is almost 0.2 percentage

TABLE 2
Decomposition of *MSD* for German HICP, 1997–2019

Equation	Period	Components			
		$\bar{v}$	$\bar{s}$	$\bar{u}$	Cov.
(2)	Before 2012	32.6	33.9		33.5
	Since 2012	49.9	16.0		34.1
	Total	38.3	28.0		33.7
(3)	Before 2012	32.6	33.9	0.0	33.5
	Since 2012	45.8	14.7	9.3	30.2
	Total	37.2	27.2	3.3	32.3

Notes: Components as percentage of *MSD*. Törnqvist price index with full-information weights used as the reference for "true" inflation.

point higher than in the years between 1997 and 2011, on average. However, the *RMSD* of the substitution component, which had been 0.09 percentage point before the 2012 methodological change, declines by about one-third. The annual updating component introduced by this methodological change produces volatility in nearly the same order of magnitude as the substitution component in the post-2012 period.

A look at the *MSD* decompositions (reported in Table 2) gives further insights,[16] particularly regarding the relative weights of the components, because these metrics are additive in the components including covariance terms. The subsequent exposition refers to the threefold *MSD* decomposition.[17] In the post-2012 phase, the data vintage component makes up nearly half of the overall variance, whereas the substitution makes up one-seventh and the annual updating component one-eleventh. The decomposition of the pre-2012 *MSD* exhibits a quite different pattern. The data vintage component, the substitution component, and the covariance term cover nearly the same share of the overall *MSD*.

Given the non-normal distribution of the monthly deviations, it is worth analyzing their dispersion by means of interquartile and interdecile ranges. As reported in Table 3, half of the central probability mass of the deviations observed over the total sample spreads over an interval with a length of one-ninth of a percentage point. To capture 80 percent of the probability mass, the length of the interval has to be doubled. The full-sample results are irrespective of whether the deviations according to equation (2) or equation (3) are considered. However, the comparison between the dispersion measures calculated separately for the pre-2012 and post-2012 periods reveals two notable findings. First, the 2012 adjustment of measurement practices widened the interquartile range and, second, the widening was indeed noticeable as regards the deviation between the official HICP and the full-information superlative index. In precise terms, the methodological change increased the interquartile range from one-tenth to almost one-seventh of a percentage point.

[16]Results for the variance decompositions are reported in Online Appendix A.

[17]The results for the other metrics and decompositions are fully tabulated in Table A2. They seem to be rather indifferent in qualitative terms.

TABLE 3
IQR AND *IDR* FOR GERMAN HICP, 1997–2019

	IQR		IDR	
Period	Equation (2)	Equation (3)	Equation (2)	Equation (3)
Before 2012	0.100	0.100	0.238	0.238
Since 2012	0.137	0.123	0.237	0.213
Total	0.110	0.108	0.246	0.237

Notes: Metrics in percentage points per annum. Törnqvist price index with full-information weights used as the reference for "true" inflation.
Abbreviations: IQR: Interquartile Range; IDR: Interdecile Range.

Weight Profiles of Selected Products

The mathematical reason behind the data vintage component and the annual updating component is that different weighting schemes are applied. Therefore, a comparison of weights may help to dissect and further understand these components. In general, the components result from an interplay of weight differences with price changes averaged over all 83 goods. An overview of the weight profiles of six selected products is therefore only partial, but may nonetheless provide useful insights into the common features and differences between full-information, official, and purely price-updated weights.

In Figure 5, the weight profiles of meat, footwear, telephone and telefax equipment, electricity, package holidays, and actual rentals for housing are plotted. These items are chosen because they either represent specific product categories or markedly affect HICP developments due to their high weight or high volatility.[18] The plots display the weights $w^x(y-1)$, $x = f, o, p$, where the time axis $y-1$ indicates the price reference period, which is December of the year $y-1$. Thus, the official and the purely price-updated weights coincide from 1996 to 2010 and deviate only after the 2011 introduction of the annual updating of weights.

The reporting years of the household budget surveys (2005, 2010, and 2015) are the cornerstones of the full-information weights. The weight profiles for meat, footwear, and telephone and telefax equipment seem to be characterized by a smooth transition from one base year to the next. From an economic point of view, this makes sense, because the household expenditure for these products is expected to adjust smoothly, even if it is not at all stable. The statistical lesson is that interpolation or extrapolation with national accounts data generally has the potential to incorporate fluctuations of a shorter duration in the full-information weights. On one hand, their economic substance is ensured because they originate from final, or at least revised, national accounts. On the other hand, the amount of money that households spend for electricity, package holidays, and rents varies

[18]Meat is an example for food, footwear for a traditional industrial product, and telephone and telefax equipment for a good of "predominantly electronic character" (Eurostat, 2018, Annex 12.9). Electricity is selected from the supply services. Actual rentals are chosen because of their high weight, and package holidays have attracted the most attention in the German HICP for several reasons (e.g., volatility, seasonality, and revision) for quite a while (Deutsche Bundesbank, 2017, 2019a, 2019b).

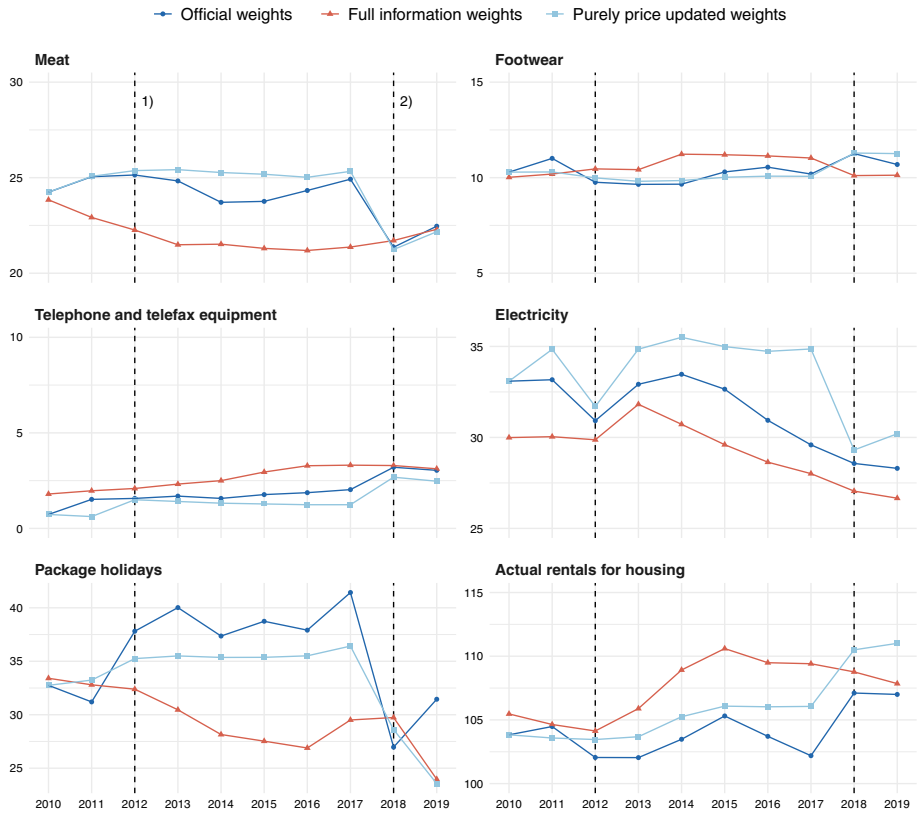

Figure 5. Weight Profiles (in ‰) of Selected Products, 2010–2019.

Notes: On the time axis, year $y - 1$ indicates the price reference period which is December of year $y - 1$. In HICP compilation, the weights of year $y - 1$ are applied to the indices of year y. (1) Price reference year where official and purely price-updated weights are derived from the 2010 household budget survey for the first time; weights before this date derived from the 2005 household budget survey. (2) Price reference year where official and purely price-updated weights are derived from the 2015 household budget survey for the first time [Colour figure can be viewed at wileyonlinelibrary.com].

appreciably from one year to another. This points to the risk of mismeasurement entailed by the real-time derivation of HICP weights.[19]

As it typically takes 2 years for consumption expenditures derived from a new household budget survey to be incorporated into price statistics, the official HICP weights for the year of the survey and the following year are still updates based on the previous survey. Only from the second year on are the official weights based on the latest survey.[20] This appears to be less of a problem for products such as footwear and telephone and telefax equipment, whose expenditure shares are rather stable over time. For products with a more volatile weight profile, the lag tends to cause

[19]For a more in-depth discussion on the volatility of HICP weights, see Eiglsperger and Schackis (2009).

[20]The weights of the years 2017 and 2018 are an exception. The incorporation of the results of the 2015 household budget survey was postponed for 1 year.

belated shifts. Examples include the adjustment of official and purely price-updated weights of meat, electricity, package holidays, and rental from 2017 to 2018 as well as of electricity and package holidays from 2011 to 2012.

The weight profiles of meat illustrate that it is impossible in practice to use real-time updating techniques to properly capture structural shifts. The expenditure share of meat declined from 2.4 percent in 2010 to 2.1 percent in 2015, according to the respective household budget surveys. The wisdom of hindsight makes it possible to adequately model this transition retrospectively by means of full-information weights. The real-time compilation of weights, however, had failed to capture this structural shift until the 2015 survey was considered. This happened belatedly in 2018 with an abrupt correction.

The evidence from the six selected products suggests that full-information weights tend to differ more strongly from official weights than the official from the purely price-updated weights.[21] This explains why the data vintage component exceeds the annual updating component in magnitude. A reason for the almost entirely positive sign of the realization of the data vintage components might be that the differences between the full-information weights and the official weights seem to be very persistent. In years when the results of new household budget surveys were incorporated into HICP weights, alignments can be observed. Nonetheless, this line of reasoning is based on empirical evidence that cannot be generalized across all relevant dimensions. On one hand, the trend can generally be explained by the lagged consideration of survey information in official weights. On the other hand, the weight-updating procedures currently used in HICP compilation are shown to be largely incapable of predicting the full-information weights. Under the hypothesis that the latter are a good proxy for the weighting pattern underlying "true" inflation, this evidence and its adverse effects on HICP measurement in terms of large data vintage components might prompt statisticians to conceive methodological improvements in this area. Some example strategies will be explained in Section 5.1.

4.2. Substitution Components of German and Euro Area HICPs

The substitution component of the euro area HICP shares many qualitative features with that of the German HICP (see Figure 6). First, the realizations are almost entirely nonnegative, as expected in line with theory. Second, they are subject to a seasonal profile (see Figure 4). Third, the largest substitution components are observed in 2008 and 2009, the years of the global financial crisis and the Great Recession. However, the results for the euro area turn out to be less pronounced.

The substitution component averages 0.04 percentage point, which is about one-quarter smaller than its German counterpart (see Table 4). Its volatility is scaled down by almost one-third. The seasonal factors of the euro area substitution component are subject to a more moderate decay over the course of the year than the substitution component of the German HICP.

The most important result is attributable to the implications of the annual updating of weights introduced in 2012 as a general requirement for the HICP in

[21]This also holds for a comparison that is based on the weights of all goods.

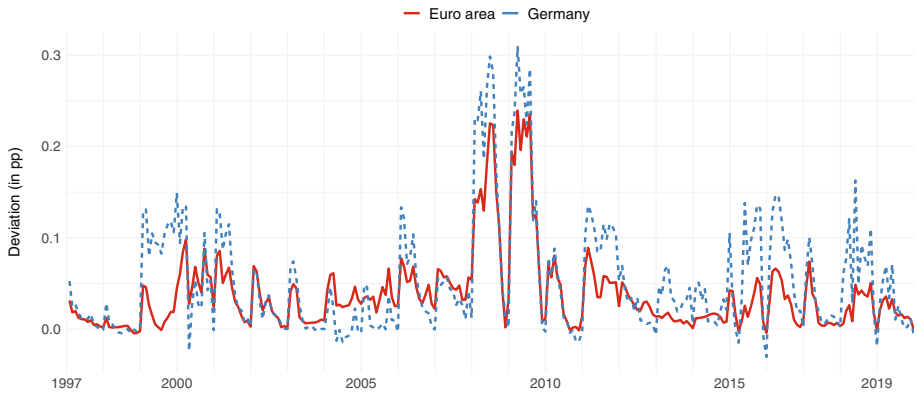

Figure 6. Monthly Substitution Components of German and Euro Area HICPs [Colour figure can be viewed at wileyonlinelibrary.com].

TABLE 4
Metrics for Substitution Components of German and Euro Area HICPs, 1997–2019

Metric	Period	Germany	Euro Area
MD	Before 2012	0.059	0.047
	Since 2012	0.044	0.022
	Total	0.054	0.039
MSD	Before 2012	0.009	0.005
	Since 2012	0.004	0.001
	Total	0.007	0.003
RMSD	Before 2012	0.093	0.069
	Since 2012	0.062	0.028
	Total	0.083	0.058

all EU Member States. As a consequence of this change, the average substitution component of the euro area HICP halved to a virtually negligible bias of 0.02 percentage points per annum. The volatility measures also declined considerably. While disregarding substitution has contributed to the bias and, in the case of the German HICP, to inaccuracy to a marked, albeit smaller extent, it has become a non-issue for the euro area HICP.

5. Discussion

The analysis sheds light on the upper-level aggregation in CPI measurement, considering the potential detrimental effects of the real-time computation of expenditure weights. The CPIs under review are the HICPs for Germany and the euro area in the period from January 1997 to December 2019. For these price indices, no evidence on measurement bias and inaccuracy stemming from upper-level aggregation is found in the more recent literature. However, it is possible to compare the results of this study with the evidence for the US CPI. As detailed in Section 2, the existing literature points to an upper-level aggregation bias between 0.15 and 0.3 percentage point per annum. Taking this as a reference value, the mean deviations reported in this paper generally appear to be of a plausible magnitude.

Two main results of the paper may be interpreted as good news for the HICP. First, the use of the Laspeyres formula does not, per se, induce a marked measurement bias at the upper level of aggregation. Second, the annual updating of weights introduced in 2012 was a step forward toward marginalizing the substitution part of the bias. These conclusions are attributable to the potential disregarding of changing consumption patterns. The paper, however, also conveys the bad news that the substitution component is less detrimental for inflation mismeasurement at the upper level of aggregation than the compilation of expenditure weights on the basis of incomplete information. On average, the data vintage component has contributed 0.07 percentage point to the measurement bias of the German HICP since 2012. Compared with the pre-2012 period, this is a deterioration both in absolute terms and relative to the total upper-level aggregation bias.

The subsequent section is devoted to a discussion of potential ways in which statistical offices may seek to enhance HICP compilation as regards the real-time updating of weights. For the time being, HICP users have to accept the shortcomings in this area. Potential implications for the interpretation of HICP figures are outlined in Section 5.2.

5.1. Options for Improving HICP Compilation

Any strategy for mitigating the impact of the data vintage component in real-time HICP compilations boils down to the question of how to appropriately and sufficiently reliably predict the "true" or final expenditure weights using the data available at the time when HICP weights need to be compiled. An obvious conjecture is that the unfavorable interplay of the dependence on lagged, but still rather preliminary, national accounts data (including price-updating) and the impossibility of revising HICP weights may be a crucial cause of inferior performance. Potential options for improvement begin by relaxing, at least, one of the two elements.

In its recommendations, the Boskin Commission laid out two options (Boskin *et al.,* 1998, pp. 12–13). Its first proposal, which actually addresses the issue of compiling the current-period weights (required for the compilation of the preferred superlative price index), was to further elaborate on the real-time performance of weight-updating procedures by extrapolating survey-based household expenditure data on the basis of timely information. Its second suggestion was to publish, as a complement to the timely CPI, a second price index that would incorporate revised, or even final, data and would thus be released with a considerable lag and, in principle, an authorized level of subsequent revision.

In 2002, the Bureau of Labor Statistics (BLS), the producer of the US CPI, introduced a revision-prone supplemental index (C-CPI), with weights being monthly updated. At the upper level of aggregation, its final version is based on the Törnqvist formula. Apparently, the existence of this index has not compromised the public's perception of the Laspeyres-type index being the headline measure. However, the situation in the US differs from that in Europe. The BLS "has long accepted the COLI as the measurement objective" (Greenlees and Williams, 2010, p. 747), making it easier to advertise the supplemental index as the most theoretically appealing approximation (even though its final values are only provided

with a reporting delay of 10–12 months) while keeping the timely and revision-free headline index unscathed for policy purposes. On the contrary, it is still a convincing argument that an official price index should require no revision, because it is usually adopted for use in indexation provisions. The prominent role of the HICP in the ECB's monetary policy is considered an obstacle to establish a competing all-items price index in Europe.

At first glance, it turns out that the weight-updating part of the first option has already been implemented into HICP measurement standards. The results of this paper, however, highlight that the annual updating of weights is not sufficient to eliminate measurement bias. The introduction of the biennial updating of expenditure weights has not resolved the timeliness issue of the US CPI either. Proposals to mitigate the real-time problem in the calculation of weights have been made in the literature. They include the application of the Lloyd–Moulton index formula, which approximates a superlative index for a specific (estimated) elasticity of substitution (Shapiro and Wilcox, 1997) and the use of hybrid indices (Armknecht and Silver, 2014). However, these approaches are inappropriate for HICP practice as they would either require, e.g., the estimation of elasticities of substitution or imply inconsistency of aggregation. It would therefore be worth revisiting the Boskin Commission's suggestion that the potential of scanner data for inferring quantity information in a timely manner be established. In recent years, more and more transaction data have become available, allowing statistical offices to obtain information about both prices and quantities. According to Eurostat (2017, p. 3), in 2017, one-fifth of EU countries were already using scanner data in the compilation of their HICPs, though its use is generally restricted to food and beverages as well as personal and home care products.

The broader availability of transaction volumes does not necessarily ensure a better extrapolation of expenditure weights because these data sources are mostly constrained to specific product categories, outlets, and/or enterprises. This raises questions such as the following: do they lie fully within the scope of the HICP?[22] Are they representative of consumer spending in the specific segments? How can they be properly integrated into the full spectrum of the final consumption expenditures?

High-quality information about households' expenditure might be expected to become available more quickly if the new data sources were to be reconciled with data from a continuous consumer survey. New techniques such as "home scans" might facilitate its implementation at low costs and without unduly bothering reporting entities. Diewert and Fox (2022) argue that a continuous consumer survey is very much required for producing meaningful price indices in turbulent times such as the coronavirus crisis. National accountants and price statisticians may unite their efforts to speed up the compilation of a detailed consumption pattern.[23]

[22] With scanner data, it is usually impossible to distinguish whether the recorded purchases are made by "non-residents or residents living in institutional households" (Eurostat, 2018, Sect. 3.3.6), as these are not covered by the HICP domain.

[23] It should be kept in mind that small sample sizes of consumer surveys could induce higher volatility in the derivation of weights. This could be even more a problem when still applying chain-linking.

5.2. Implications for the Use of the HICP

With the results of this paper, it is not possible to directly quantify the upper-level aggregation of the euro area HICP, as the analysis is silent on the effect of the use of preliminary data in the compilation of expenditure weights. The evidence for the German HICP implies that since 2012, the data vintage component is quantitatively more relevant than the substitution component. Therefore, the marginal post-2012 estimate for the latter in the euro area HICP cannot be interpreted as an all-clear signal regarding bias and inaccuracy in the upper-level aggregation practices. By contrast, against the backdrop of the achieved harmonization of weight updating procedures in Europe, it seems more logical to assume that the euro area HICP may suffer from this source of mismeasurement to a comparable extent. If this line of reasoning were accepted, a small positive margin would appear to be justified to account for mismeasurement at the upper level of aggregation in the euro area HICP rate.

According to the evidence for the German HICP, the bias is of a magnitude of around one-ninth of a percentage point. Using a bootstrap procedure (see Online Appendix C), confidence bands for the (point) estimate of the bias are performed. With a probability of 90 percent, the "true" upper-level aggregation bias falls into the rather narrow interval between 0.099 and 0.125 percentage point.

Looking at the inaccuracy metrics, the main message is that, from a statistical point of view, small variations in the year-on-year HICP rates are not clear-cut indications for changes in inflation dynamics. With respect to upper-level aggregation effects, this can be seen from the dispersion measures. Disregarding changing consumption and using preliminary data in the compilation of expenditure weights cause a statistical uncertainty surrounding the German HICP, which may be illustrated by an interquartile range of 0.110 and an interdecile range of 0.246 percentage point. The 90 percent confidence bands for the interquartile and the interdecile ranges are [0.093; 0.127] and [0.212; 0.278], respectively. This suggests a fairly accurate estimation of the measurement uncertainty at the upper level of aggregation.

6. Conclusions

The HICP may suffer from mismeasurement owing to changing consumption patterns being disregarded and preliminary data being used in the compilation of expenditure weights. Mismeasurement at the so-called upper level of aggregation is studied in terms of bias and inaccuracy. This is analyzed using monthly disaggregated price index series and the respective weights from January 1997 to December 2019. The year-on-year percentage rates of the official HICP are evaluated against "true" inflation, which is assumed to be represented by a superlative (Törnqvist) price index with full-information weights.

This paper provides partial evidence on HICP mismeasurement. Previous research has shown that the measurement issues at the upper level of aggregation

Moreover, in contrast to national accounts, continuous consumer surveys could bear the risk of being non-representative in practice.

are quantitatively less relevant than potential pitfalls at the lower level of aggregation, in quality adjustment procedures, and the timely consideration of new products and distribution channels. It is beyond the scope of this paper to put the results into the perspective of a broad-based (and up-to-date) assessment of HICP mismeasurement. The HICP coverage is taken for granted in this analysis. Of course, the omission of the cost of owner-occupied housing (OOH) in the HICP can be regarded as another, or even the major, source of mismeasurement.[24]

For the German HICP, the bias and inaccuracy measures are decomposed into two or three components, i.e., substitution, data vintage, and, possibly, annual updating. The main contribution to the bias stems from the use of preliminary data in the updating of expenditure weights. The substitution component is strictly positive, as expected in line with the theory, but very small in magnitude. The 2012 introduction of the annual updating of weights reduced the measurement bias induced by disregarded substitution, however, at the expense of worsening the data vintage component of the same magnitude. As a consequence, the impact of the introduction of the annual updating on the total bias was neutral. Apart from bias, inaccuracy turns out to be a relevant performance criterion for HICP measurement. A variance decomposition of the inaccuracy reveals that it is mainly driven by the data vintage effect rather than the substitution effect. This is even more pronounced for the period after the methodological change in 2012.

For the euro area HICP, data availability limits the analysis to uncover the effects of disregarded substitution. This source of mismeasurement is less detrimental here than in the German HICP. Since the 2012 methodological change, the substitution component has virtually become a non-issue. As in the German case, the impact of the use of preliminary national accounts in the updating of weights might be the more relevant upper-level measurement issue in the HICP for the euro area, too. The calculation of full-information weights for the euro area HICP is a much more complex exercise.[25] Finding solutions to the challenges emerging in this multi-country context seems to justify a separate paper of their own.

The compilers of price statistics may learn from the results of the full-fledged analysis for the German HICP that an annual updating of weights using quantity information from preliminary national accounts can reduce the substitution bias. However, the task of extrapolating reliable expenditure weights must not be "transferred" to national accountants. The adverse repercussions might be more severe in price statistics than in national accounts, as while preliminary data are regularly revised in the latter case, this exerts a permanent effect in the former.

From the results of the paper, two major conclusions are drawn as regards the use of the HICP as a measure of "true" inflation. First, it still appears justified to

[24]In the ECB's recent monetary policy strategy review, it was mentioned that, among the areas for further improvement, "the integration of OOH into the HICP remains outstanding" (ECB, 2021b, p. 8).

[25]Applying the calculation scheme used for Germany is likely to fail because real-time national accounts data for household consumption expenditures are not available for the euro area—neither for all euro area countries, nor in the required breakdown. In addition, many euro area countries do not publish a national CPI from which HICP-consistent base year weights can be inferred, like Germany does. A way to solve this problem might be to refer to the household budget surveys conducted in all euro area countries. However, this would require enormous efforts in terms of collecting, processing, and calculating data.

assume a small positive margin when accounting for mismeasurement at the upper level of aggregation. Second, upper-level measurement errors turn out to be large enough to be considered a relevant source when assessing the precision of the HICP.

References

Armknecht, P., "Fixed basket methods for compiling consumer price indexes," *American International Journal of Contemporary Research*, 5, 97–106, 2015.

Armknecht, P. and M. Silver, "Post-Laspeyres: The case for a new formula for compiling consumer price indexes," *Review of Income and Wealth*, 60(2), 225–244, 2014. https://doi.org/10.1111/roiw.12005

Balk, B. M. and W. E. Diewert, "The consumer price index and its substitution bias," in W. E. Diewert (ed), *Price and productivity measurement*, Trafford On Demand Pub, Victoria 2012.

Baxter, M., "Implications of the US Boskin report for the UK retail price index," *Economic Trends*, 527, 56–62, 1997.

Boskin, M. J., E. R. Dulberger, R. J. Gordon, Z. Griliches and D. W. Jorgenson, "Consumer prices, the consumer price index, and the cost of living," *Journal of Economic Perspectives*, 12(1), 3–26, 1998. https://doi.org/10.1257/jep.12.1.3

Camba-Mendez, G., "The definition of price stability: Choosing a price measure," *Background studies for the ECB's evaluation of its monetary policy strategy*, European Central Bank, Frankfurt am Main, 31–43, 2003.

Crawford, A., "Measurement biases in the Canadian CPI: An update," *Bank of Canada Review. Spring*, 1998, 38–56, 1998.

Cunningham, A., "Measurement bias in price indices: An application to the U.K.'s RPI." *Bank of England Working Paper*, 1996.

Destatis, "Erweiterter Erfassungsbereich für den Harmonisierten Verbraucherpreisindex," *Wirtschaft und Statistik*, 148, March 2000.

Deutsche Bundesbank. "The volatility of the traditional core inflation rate in Germany," *Monthly Report*, 49–50, November 2017.

Deutsche Bundesbank. "Dampening special effect in the HICP in July 2019," *Monthly Report*, 57–59, August 2019a.

Deutsche Bundesbank. "The revision of the sub-index for package holidays and its impact on the HICP and core inflation," *Monthly Report*, 8–9, March 2019b.

Dietrich, A., Eiglsperger, M., Mehrhoff, J., and Wieland, E., Chain-linking over December and methodological changes in the HICP: View from a central bank's perspective, ECB Statistics Paper Series, No 40, Working Paper. February 2021.

Diewert, W.E. and Fox, K.J., "Measuring real consumption and consumer price index bias under lockdown conditions," Canadian Journal of Economics/Revue canadienne d'économique, 55, 480-502, 2022.

Diewert, W. E., "Exact and superlative index numbers," *Journal of Econometrics*, 4(2), 115–145, 1976. https://doi.org/10.1016/0304-4076(76)90009-9

ECB, "Extended and further harmonised coverage of the HICP," *Monthly Bulletin*, 3, 23, 2000.

ECB, "Changes in the coverage and methods for computation of the Harmonised Index of Consumer Prices," *Monthly Bulletin*, 2, 24–5, 2001.

ECB, "The monetary policy of the ECB. European Central Bank," *Frankfurt am Main*, 2nd ed., 2004.

ECB, "New standards for HICP weights," *Monthly Bulletin*, 3, 36–9, 2012.

ECB, "Potential measurement issues in consumer price indices," *Monthly Bulletin*, 4, 40–2, 2014.

ECB, An overview of the ECB's monetary policy strategy. https://www.ecb.europa.eu/home/search/review/pdf/ecb.strategyreview_monpol_strategy_overview.en.pdf. Accesssed: 5 April 2022, 2021a.

ECB, "Inflation measurement and its assessment in the ECB's monetary policy strategy review," *Occasional Paper Series*, 265, ECB, 2021b.

ECB, The ECB's monetary policy strategy statement, 2021c. https://www.ecb.europa.eu/home/search/review/pdf/ecb.strategyreview_monpol_strategy_statement.en.pdf. Accessed: 5 April 2022.

Eiglsperger, M. and Schackis, D. Weights in the harmonised index of consumer prices: Selected aspects from a user's perspective. *11th Ottawa Group Meeting*, 2009.

EU, *Regulation (EU) 2016/792 of the European Parliament and the Council of 11 May 2016*, Official Journal of the European Union, Luxembourg, 2016.

EU, *Commission Implementation Regulation (EU) 2020/1148 of 31 July 2020*, Official Journal of the European Union, Luxembourg, 2020.

Eurostat, Harmonised index of consumer prices (HICP): Practical guide for processing supermarket scanner data. September 2017.

Eurostat, Harmonised index of consumer prices (HICP), Methodological Manual. November 2018.
Gábor-Tóth, E. and P. Vermeulen, "Elementary index bias: Evidence for the euro area from a large scanner dataset," *German Economic Review*, 20(4), 618–656, 2019. https://doi.org/10.1111/geer.12182
Greenlees, J. S. and E. Williams, "Reconsideration of weighting and updating procedures in the US CPI," *Jahrbücher für Nationalökonomie und Statistik*, 230(6), 741–758, 2010.
Hoffmann, J. "Problems of inflation measurement in Germany: Economic Research Group Deutsche Bundesbank," *Discussion Paper 1/1998*, 1998.
ILO, IMF, OECD, UNECE, Eurostat, World Bank, *Consumer price index manual: Concepts and Methods*, International Monetary Fund, Washington DC, 2020.
Lebow, D. E. and J. B. Rudd, "Measurement error in the consumer price index: Where do we stand?" *Journal of Economic Literature*, 41(1), 159–201, 2003. https://doi.org/10.1257/002205103321544729
Lequiller, F., "Does the French consumer price index overstate inflation?" *INSEE Série des documents de travail de la Direction des Etudes et Synthèses Économiques*, G 9714, 2–46 , 1997.
Lowe, J., *The present state of England in regard to agriculture, trade, and finance*, Second ed., Longman, Hurst, Rees, Orme and Brown, London, 1823.
Manser, M. E. and R. J. McDonald, "An analysis of substitution bias in measuring inflation, 1959-85," *Econometrica*, 56(4), 909–930, 1988. https://doi.org/10.2307/1912704
Neves, P. D. and L. M. Sarmento, "The substitution bias of the consumer price index," *Banco de Portugal Economic Bulletin*, June, 25–33, 1997.
Shapiro, M. D. and D. W. Wilcox, "Alternative strategies for aggregating prices in the CPI," *Review of the Federal Reserve Bank of St. Louis*, 79(3), 113–25, 1997.
Silver, M. and C. Ioannidis, "The measurement of inflation; untimely weights and alternative formulae: European evidence," *The Statistician*, 43(4), 551–562, 1994. https://doi.org/10.2307/2348139
Welch, F., *The causes and consequences of increasing inequality*, Vol 2, University of Chicago Press, Chicago, IL, 2001.

Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's web site:

Appendix A: Superlative indices
Appendix B: Derivation of weights
Appendix C: Bootstrapping confidence bands