Empirical findings on upper-level aggregation issues in the HICP*

Julika Herzberg Thomas A. Knetsch Dilyana Popova

Patrick Schwind Sebastian Weinand

May 6, 2022

Abstract

We analyse potential mismeasurement of the Harmonised Index of Consumer Prices (HICP) at the upper level of aggregation, focusing on two sources of measurement error: the choice of index formula (representativity component) and the reliability of weights (data vintage component). While the former is well-known in the literature and captures the fact that a Laspeyres-type index such as the HICP suffers from a systematic overestimation of inflation due to the disregard of changes in consumption patterns, less attention has been paid to the latter so far. HICP weights are annually updated based on national accounts. When used, these data are only granted a preliminary status. The use of final data is expected to yield more reliable weights and, thus, a better estimate of inflation. With national accounts vintage data, we calculate bias and inaccuracy metrics in order to analyse mismeasurement at the upper level of aggregation in the HICPs for Germany, France, Italy, Spain and the Netherlands, as well as for the country group, representing 82% of the euro area HICP, over the period from 2012 to 2019. Measured in terms of annual HICP rates of this country group, the total upper-level aggregation bias falls short of one-tenth of a percentage point. The representativity and data vintage components contribute to the overall bias in quite similar shares. As expected by theory, the representativity component is positive for all countries considered. Data vintage components are positive in all countries but the Netherlands. The uncertainty surrounding HICP inflation due to upper-level aggregation issues is also small.

Keywords: Inflation measurement · Representativity bias · Updating of weights

JEL classification: E31 · C43

[*] Deutsche Bundesbank, Directorate General Statistics, Wilhelm-Epstein-Strasse 14, 60431 Frankfurt am Main, Germany, julika.herzberg[@]bundesbank.de, thomas.knetsch[@]bundesbank.de, dilyana.popova[@]bundesbank.de, patrick.schwind[@]bundesbank.de, sebastian.weinand[@]bundesbank.de. This paper is an updated version of a background study prepared for the workstream "Inflation Measurement" of the Monetary Policy Strategy Review carried out by the Eurosystem in 2020 and 2021. We thank workstream participants for useful comments and suggestions. 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.

1 Introduction

As confirmed in its recent monetary policy strategy review, the Governing Council of the European Central Bank (ECB) considers the Harmonised Index of Consumer Prices (HICP) as "the appropriate price measure for assessing the achievement of the price stability objective" (ECB, 2021b, p. 1). According to the adjusted monetary policy aim, the year-on-year percentage change of the HICP (henceforth referred to as inflation) is targeted to be at 2% over the medium term, suggesting that the Governing Council sees the need for an inflation buffer above zero. The existence of a measurement bias is amongst the reasons for this inflation buffer.

Measurement issues generally arise at various stages of HICP compilation. Mismeasurement can thus stem from different sources, including the disregard of changes in consumption patterns, belated introduction of new products, untimely account of new distribution channels and improper adjustment for quality changes. When it comes to the aggregation of individual price changes over the basket of goods and services, a distinction is made between the lower and the upper level. At the lower level, prices are aggregated without any weighting information while, at the upper level, households' expenditure shares are applied to form price indices.

Designed as a Laspeyres-type index, the HICP generally measures the aggregate price change of a fixed basket of goods and services (cost-of-goods index or COGI). In strict terms, HICP weights are only representative for the base period in the price comparison. A change in the consumption patterns from the base period to the comparison period may induce a source of mismeasurement which is henceforth called representativity bias. A further source of mismeasurement at the upper level of aggregation stems from using preliminary national accounts data in the annual updating of weights. As the HICP is not allowed to be revised in order to take account of new releases in the national accounts, a data vintage effect may impair inflation measurement, provided that national accounts are expected to converge to "true" consumption patterns from earlier to later releases.

The focus of this paper is on estimating the extent of HICP mismeasurement at the upper level of aggregation, thereby separating out the effects of imperfect representativity and the use of preliminary data in weight compilation. While theory suggests that the representativity effect is positive, the sign of the data vintage effect is generally unknown up front. Apart from the bias, it is worth looking at the root mean squared deviation and the interdecile range as measures of inaccuracy. The analysis is similar to what Herzberg et al. (2021) recently studied for Germany. In general, the same formal evaluation framework is applied. In this paper, we consider the HICPs of the five largest euro area countries (Germany, France, Italy, Spain and the Netherlands), as well as of the country group (henceforth called Big-5 aggregate). Given that the Big-5 aggregate covers more than four-fifths of the euro area HICP, the empirical findings give insight into upper-level aggregation issues of the ECB's key inflation measure.[1]

In order to carry out both the representativity and the data vintage parts of HICP mismeasurement at the upper level of aggregation for this group of countries, the benchmark index against which the HICP is evaluated needs to be adjusted with regard to the weight concept. Instead of the full-information weights compiled by Herzberg et al. (2021), the superlative index is constructed on the basis of weighting schemes using final, or at least revised, information from national accounts (henceforth called final NA weights). The benchmark index is assumed to better proxy the "true" aggregate price development than the HICP, as it is formed on a symmetric weighting using timely and more mature information about households' consumption expenditures.

We consider the period from January 2012 to December 2019. The start of the sample is chosen because the annual updating of HICP weights became mandatory in 2012. The sample terminates by the end of 2019 because we would like to ensure that final NA weights are calculated using national accounts data which incorporate information available at statistical offices with a lag of, at least, two years. Given that national accounts revisions are usually frontloaded, these data may be considered sufficiently close to a final status. In addition, we provide evidence for the representativity component over the whole HICP history starting in 1997 and terminating by the end of 2021. This allows us to take a long-run, albeit partial, view on HICP upper-level mismeasurement, including recession periods such as the Financial Crisis 2008/2009 and the COVID-19 pandemic 2020/2021.

Our main conclusions are the following. Measured in terms of annual HICP rates, the total upper-level aggregation bias of the Big-5 aggregate clearly falls short of one-tenth of a percentage point. The representativity and the data vintage components contribute to the overall bias in quite similar shares. As expected, the representativity component is positive for all countries under consideration. The representativity effect for the euro area HICP is even markedly smaller than that for the Big-5 aggregate. Data vintage components are positive in all countries but the Netherlands. Owing to a negative data vintage component, the overall upper-level aggregation bias is negative for the Dutch HICP. Theoretical considerations and a comparison on the basis of German data let us conclude that the data vintage effect may be interpreted as a lower bound if calculated using final NA weights instead of taking the universe of information into account.

The uncertainty surrounding HICP inflation due to upper-level aggregation issues is small, too. The interdecile range of the deviations between the HICP and the benchmark amounts to about one-tenth of a percentage point for the Big-5 aggregate. For the individual countries, we evidence wider interdecile ranges, suggesting that contrary developments in country HICPs tend to balance each other out in the aggregate. Consid-

[1] It is not feasible for us to carry out the analysis on the basis of a total representation of the euro area, as it has been impossible to gather national accounts vintage data for all euro area countries.

ering the root mean squared error, the results for the Big-5 aggregate as well as for the individual countries confirm that the representativity and data vintage components make fairly equal contributions to HICP inaccuracy at the upper level of aggregation.

Against the backdrop of existing evidence reported mainly for the US consumer price index (CPI), the upper-level aggregation bias of the HICP turns out to be a relatively small number. In the report of the Boskin Commission (Boskin et al., 1998), for instance, only 0.15 of the 1.1 percentage points total bias per annum was found to be due to upper-level substitution, while a much larger portion was due to the introduction of new products and quality changes. Later research on the subject by Greenlees and Williams (2010) found the upper-level bias to be more prevalent, amounting to 0.3 percentage points per annum. More recently, Armknecht and Silver (2014) found evidence in the post-2002 US CPI that the Boskin Commission's findings on the presence of measurement bias still hold, with an upper-level aggregation bias of 0.16 percentage points.[2] Silver and Ioannidis (1994) paid attention to potential mismeasurement caused by the use of "untimely weights" and, thus, considered a phenomenon which is quite similar to the data vintage effect studied in this paper. This is in fact not the only similarity. In addition, they expanded the range of statistical metrics by looking also at the root mean squared error, for instance, and they considered European CPIs in their empirical investigation.

The remainder of this paper is structured as follows. In the next section, the evaluation framework is sketched out. In Section 3, empirical results are presented. In Section 4, conclusions are drawn.

2 Methodology

In this section, we describe the evaluation framework. We start with a brief explanation of key HICP construction principles. We follow up with the exposition of weight concepts and index formulae. Finally, we introduce the statistical metrics which are employed to measure HICP mismeasurement at the upper level of aggregation.

2.1 Upper-level aggregation principles of the HICP

The HICP is designed as a chain-linked Laspeyres-type index where weights are updated at the beginning of each calender year and kept constant throughout (EU, 2020).[3] In

[2] While plenty of research has studied the impact of upper-level aggregation bias in the US, empirical evidence for the HICP is rarely available. The report of the Boskin Commission can be credited for later on sparking further research interest in CPI measurement bias outside the US. Among the current EU members, in the nineties the topic was studied with respect to inflation in France (Lequiller, 1997), Portugal (Neves and Sarmento, 1997) and Germany (Hoffmann, 1998).

[3] In the corresponding academic literature the HICP is often referred to as a Lowe index since weight and price reference periods are different from each other (Lowe, 1823).

formal terms, it may be written as:

$$P^o_{\text{HICP}}(y,m) = \sum_{i=1}^{I} w^o_i(y-1,12) \cdot \frac{p_i(y,m)}{p_i(y-1,12)} \ ,$$

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 $(m = 1,\ldots,12)$. The weight of good i applied to the HICP (henceforth called "official weight" and marked by superscript "o") in year y is denoted by $w^o_i(y-1,12)$, as it refers to the price reference period which is December of the previous year. For notational convenience, however, we write $w^o_i(y-1) \equiv w^o_i(y-1,12)$ in the remainder.

According to Eurostat (2020, p. 2), Article 3.1 of the HICP Implementing Regulation (EU, 2020) means that "the expenditure shares used for the HICP in year t should be representative of year $t-1$". On the basis of these expenditure shares (referring to annual household consumption expenditure data from the national accounts), HICP weights result from an obligatory price update to December, i.e. $w^o_i(y-1) = w_i(y-1) \cdot p_i(y-1,12)/p_i(y-1)$ where $w_i(y-1)$ and $p_i(y-1)$ indicate the average expenditure share and price of good i in year $y-1$ respectively.

In the measurement practice of the time period under consideration, however, the information about consumption expenditures was often more outdated than formally prescribed by regulation because national accounts were available only until $y-2$ when updates were made. HICP legislation force statistical offices to review and update the expenditure shares of $y-2$ to make them representative of year $y-1$, implying a freedom of choice as regards the options "to-price-update" or "not-to-price-update" from $y-2$ to $y-1$ (Eurostat, 2018, Sections 3.5 and 8.2.3). As far as we are aware, the statistical offices of Germany, France, Italy and Spain generally made use of the price-updating from 2012 to 2020,[4] whereas the Dutch statistical office assumed consumption expenditures of $y-2$ and $y-1$ to be the same in relative terms.

2.2 Derivation of final NA weights

In the period under review, the weights of the HICP in calendar year y are generally formed on the basis of the first releases of households' consumption expenditures for the year $y-2$. As time goes by, the information content of national accounts data becomes more adequate along the time and vintage dimensions. First, households' consumption expenditures for the year $y-1$ could be used instead of (price-updated) expenditure values for the year $y-2$. Second, final or at least revised data could be used instead of first releases. Weights formed on the basis of this information content are called final NA weights (henceforth indicated by superscript "f") and promise to be generally closer

[4] In the French HICP, price-updating was applied as a general rule, while the possibility of adjusting to the previous year's expenditures was retained for exceptional cases where significant changes were identified.

to the (unknown) "true" expenditure shares needed to compile the best aggregate price index possible.

We pinpoint the difference between the official and final NA weights by comparing the updating formulae of the two weight concepts:[5]

$$w_i^o(y-1) = \bar{w}_i(y-\xi) \cdot \frac{c_i(y-2;y-1)}{c_i(y-\xi;y-1)} \cdot \frac{p_i(y-1)}{p_i(y-2)} \cdot \frac{p_i(y-1,12)}{p_i(y-1)} \tag{1a}$$

$$w_i^f(y-1) = \bar{w}_i(y-\xi) \cdot \frac{c_i(y-1;\infty)}{c_i(y-\xi;\infty)} \cdot \frac{p_i(y-1,12)}{p_i(y-1)} \tag{1b}$$

where $c_i(y;v)$ is households' consumption expenditure of good i in year y as it is reported in the national accounts vintage released in year v; the final vintage is denoted by $v=\infty$. According to the release calendar of national accounts, detailed consumption expenditures for the year $y-2$ are available by the end of $y-1$.

Both weight updating formulae have in common that the extrapolation with consumption expenditures is anchored by, say, a base weight $\bar{w}_i(y-\xi)$ which is derived from the universe of information about consumption patterns but is available only with a time lag of $\xi > 2$ years. The base weight is hypothetical but may be well approximated using HBS information. Eq. (1a) describes the formula employed in the "to-price-update" option; in the "not-to-price-update" option, the factor $[p_i(y-1)]/[p_i(y-2)]$ does not appear. In Eq. (1b), this factor is missing, too, because the final vintage comprises information about consumption expenditures of every weight reference period by definition.

The knowledge of base weights is not needed for the calculation of final NA weights. By substituting Eq. (1a) in Eq. (1b), we obtain the following expression for final NA weights:

$$w_i^f(y-1) = w_i^o(y-1) \cdot \frac{c_i(y-1;\infty)}{c_i(y-\xi;\infty)} \Big/ \left[\frac{c_i(y-2;y-1)}{c_i(y-\xi;y-1)} \cdot \frac{p_i(y-1)}{p_i(y-2)}\right] \tag{2}$$

If we assume $c_i(y-\xi;y-1) = c_i(y-\xi;\infty)$ for ξ sufficiently large because of the frontloading of current revisions, we end up with the relation:

$$\frac{w_i^f(y-1)}{w_i^o(y-1)} = \frac{c_i(y-1;\infty)}{c_i(y-2;y-1) \cdot [p_i(y-1)/p_i(y-2)]}, \tag{3}$$

suggesting that the ratio between the official and final NA weight of some good i is equal to the ratio between the final release for households' consumption expenditure of good i in the weight reference period $y-1$ and the price-updated first release referring to one year prior. In the stylised vintage dataset displayed in Tab. 1, the entries used as final

[5] For the sake of better readibility, equations are simplified in two respects. First, the national accounts breakdown of households' consumption expenditures is not as detailed as needed for the HICP. Hence, the updating of weights is regularly impossible to be made using the same expenditure category (as displayed in these equations) but a broader one. Second, updated weights need to be scaled such that they altogether sum up to unity. This scaling factor is omitted in the equations.

	vintage available by end of year										
reporting	block A		block B					block C			
period	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
y_0	×	×	×	×	×	×	×	×	×	×	×
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2009	×	×	×	×	×	×	×	×	×	×	×
2010	o	×	×	×	×	×	×	×	×	×	×
2011		o	×	×	×	×	×	×	×	×	f
2012			o	×	×	×	×	×	×	×	f
2013				o	×	×	×	×	×	×	f
2014					o	×	×	×	×	×	f
2015						o	×	×	×	×	f
2016							o	×	×	×	f
2017								o	×	×	f
2018									o	×	f
2019										×	f
2020											×

Note: Entries in the vintage dataset are denoted by "×" with two exceptions. The first releases which are employed in the calculation are marked by "o". The entries which are taken as final releases are marked by "f".

Table 1: Stylised vintage dataset.

releases are denoted by "f" and the first releases by "o".

Apart from current revisions which result from capturing late incoming information, national accounts are subject to benchmark revisions. In multi-year intervals, often every five years and harmonised among European countries, conceptual and methodological enhancements are introduced. In addition, benchmark revisions are often an occasion initiating the account of new data sources. Benchmark revisions generally alter the complete time series of households' consumption expenditures. This generally impairs the comparability across the vintage dimension. Within the period under consideration, we have to account for the benchmark revisions in 2013 and 2018. In Tab. 1, the various accounting regimes are denoted by "A", "B" and "C".

Vintages before and after a benchmark revision are generally made comparable by a vintage transformation. This is carried out following Knetsch and Reimers (2009). In particular, we run bivariate cointegrating regressions:

$$\ln c_i(y, 2017) = \alpha_i^B + \beta_1^B \ln c_i(y, 2021) + \epsilon_i^B(y) \tag{4a}$$

$$\ln c_i(y, 2012) = \alpha_i^A + \beta_1^A \ln c_i(y, 2021) + \epsilon_i^A(y) \tag{4b}$$

where α_i^R, β_i^R $(R = A, B)$ are regression coefficients and ϵ_i^R are covariance stationary residuals. The samples cover the periods from y_0 to 2017 in block "B" and from y_0 to 2011 in block "A".

The quality of vintage transformation functions is checked using the coefficient of determination (R^2). If $R^2 \geq 0.8$, the first releases of the blocks "A" and "B" are made comparable with the final releases, applying the following transformation:

$$\hat{c}_i(y; y+1) = \exp(\alpha_i^A)[c_i(y; y+1)]^{\beta^A}, \quad y = 2010, 2011, \tag{5a}$$

$$\hat{c}_i(y; y+1) = \exp(\alpha_i^B)[c_i(y; y+1)]^{\beta^B}, \quad y = 2012, ..., 2016. \tag{5b}$$

Otherwise, first releases of blocks "A" and "B" remain untransformed and the latest vintages before the respective benchmark revision are taken as final releases.[6]

The vintage data sets available for the five countries under consideration are not entirely homogeneous but differ in the number of available vintages, the time length of the data in each vintage and lack of information for some positions. Data are retrieved from the ECB's Statistical Data Warehouse and double-checked with, and occasionally complemented by, information from the national accounts data repositories of the national central banks of the countries under review. The very few remaining gaps in the vintage data sets are filled by estimates.[7]

It is worth clarifying the difference between the final NA weights used in this analysis and the full-information weights introduced by Herzberg et al. (2021).[8] The two concepts have in common that information is used irrespective of the time when it becomes available. However, they differ in terms of the data sources out of which timely and/or more mature information can be taken. While final NA weights are restricted to timely and the most mature national accounts, the data base underlying full-information weights is principally the universe of available information. In this context, the information from the multi-year household budget surveys (HBS) is the most relevant additional source. In particular, full-information weights take into account HBS information for the years the surveys refer to, while (final) national accounts data are applied for interpolation in the years in between. By contast, the HBS results enter in the calculation of final NA weights at the same date and in the same way as in the calculation of HICP weights.

2.3 Index formulae

In the following analysis, monthly year-on-year price relatives are aggregated to make summary metrics interpretable as a source of mismeasurement of inflation, i.e. comparable to the year-on-year percentage change of a price index and measured in percentage points

[6] This alternative induces the data vintage effect to be systematically distorted downward because the impact of later current revisions is not captured. Hence, it is no more than a surrogate in the rare cases where the estimation of proper vintage transformation functions fail.

[7] The vintage data sets and detailed meta information are available upon request.

[8] Herzberg et al. (2021) calculate full-information weights for the German HICP. It goes beyond the scope of this study to calculate full-information weights for the other countries. For comparability, final NA weights serve as a uniform basis for computing the data vintage effects in this paper. The empirical differences between final NA weights and full-information weights are highlighted in the case of Germany in a digression in the next section.

(see Herzberg et al., 2021, footnote 10). Thus, the aggregate price relative representing the HICP is defined as:

$$P_L^o(y,m) = \sum_{i=1}^{I} w_i^o(y-1) \cdot \frac{p_i(y,m)}{p_i(y-1,m)}\,. \tag{6}$$

We indicate this index with subscript "L" because it is of a Laspeyres type and superscript "o" because it is based on official weights.

The benchmark index against which P_L^o is evaluated is designed by a superlative index which symmetrically incorporates the weights of both the base period and the comparison period. Superlative indices are Fisher, Törnqvist or Walsh indices, for example. In this study, we restrict the exposition of results to the Törnqvist index, where arithmetic averages of the value shares in the two periods are used as weights:[9]

$$P_{T\ddot{o}}^x(y,m) = \prod_{i=1}^{I} \left[\frac{p_i(y,m)}{p_i(y-1,m)}\right]^{\frac{1}{2}\left[w_i^x(y-1)+w_i^x(y)\right]}, \quad x = f, o\,. \tag{7}$$

In our analysis, we calculate Törnqvist indices using official and final NA weights.

2.4 Bias and inaccuracy metrics

HICP mismeasurement at the upper level of aggregation is evaluated using a number of statistical metrics building on the deviation of the Laspeyres-type index based on official weights (P_L^o) from the Törnqvist index based on final weights ($P_{T\ddot{o}}^f$). In order to disentangle representativity and data vintage effects, the deviation is decomposed in the following way:

$$\frac{P_L^o}{P_{T\ddot{o}}^f} = \frac{P_L^o}{P_{T\ddot{o}}^o} \cdot \frac{P_{T\ddot{o}}^o}{P_{T\ddot{o}}^f}\,. \tag{8}$$

The first term relates a Laspeyres-type index with a Törnqvist index, both using official HICP weights. This is a measure of the representativity effect. The second term consists of two Törnqvist indices, the one based on official and the other on final NA weights. This ratio captures the data vintage effect.

We focus on bias and inaccuracy to evaluate the quality of the current HICP measurement. The mean deviation (MD) captures the measurement bias. It is defined by:

$$MD_{\text{Total}} = \frac{1}{T}\sum_{t=1}^{T} \ln\left(P_L^o(t)/P_{T\ddot{o}}^f(t)\right). \tag{9}$$

According to Eq. (8), the measurement bias can be additively decomposed into a repre-

[9] In Herzberg et al. (2021), statistical mismeasurement metrics are reported using Fisher, Törnqvist or Walsh indices, confirming the well-known result that metrics are rather insensitive to the choice of the superlative index formula.

sentativity component and a data vintage component:

$$\begin{aligned} MD_{\text{Total}} &= MD_{\text{Representativity}} + MD_{\text{Data vintage}} \\ &= \frac{1}{T}\sum_{t=1}^{T} \ln\left(P_L^o(t)/P_{T\ddot{o}}^o(t)\right) + \frac{1}{T}\sum_{t=1}^{T} \ln\left(P_{T\ddot{o}}^o(t)/P_{T\ddot{o}}^f(t)\right). \end{aligned} \tag{10}$$

The root mean square deviation (*RMSD*) is used as a measure of the statistical uncertainty surrounding HICP inflation. It is separately calculated for the total effect as well as for the representativity and the data vintage sources of mismeasurement:

$$RMSD_{\text{Total}} = \sqrt{\frac{1}{T}\sum_{t=1}^{T} \ln\left(P_L^o(t)/P_{T\ddot{o}}^f(t)\right)^2} \tag{11a}$$

$$RMSD_{\text{Representativity}} = \sqrt{\frac{1}{T}\sum_{t=1}^{T} \ln\left(P_L^o(t)/P_{T\ddot{o}}^o(t)\right)^2} \tag{11b}$$

$$RMSD_{\text{Data vintage}} = \sqrt{\frac{1}{T}\sum_{t=1}^{T} \ln\left(P_{T\ddot{o}}^o(t)/P_{T\ddot{o}}^f(t)\right)^2} \tag{11c}$$

In contrast to the bias, $RMSD_{\text{Representativity}}$ and $RMSD_{\text{Data vintage}}$ do not sum up to the total *RMSD*.

An additional uncertainty measure is the interdecile range:

$$IDR_{\text{Total}} = P_{90} - P_{10} \ , \tag{12}$$

where P_{10} and P_{90} are the 10th and the 90th percentiles of $\ln\left(P_L^o(t)/P_{T\ddot{o}}^f(t)\right)$.

3 Results

The empirical study is based on price indices and weights for 76 product groups, containing COICOP positions at the two, three or four digit level. In terms of number and breakdown of detailed HICP data, the data sets are uniform for the five countries under consideration.

The full-fledged analysis covers the period from January 2012 to December 2019. Hence, the empirical results rely on 96 monthly observations. The bias and inaccuracy measures are reported in Section 3.1. Section 3.2 provides information about the weight profiles of selected product groups, helping interpret the data vintage component. In Section 3.3, we finally report results for the representativity component over the whole HICP history starting in 1997 and terminating by the end of 2021. The long sample consists of 300 monthly observations.

3.1 Bias and inaccuracy

Bias and inaccuracy metrics are calculated on the basis of the logarithmic deviation of a Laspeyres-type index based on official weights from the benchmark index, which is a Törnqvist price index using final NA weights. From a mathematical perspective, the logarithmic deviation is measured as a percentage of the benchmark index. Owing to the construction of the price indices, monthly deviations and bias estimates may be interpreted as mismeasurement in HICP inflation, measured in percentage points, and inaccuracy metrics as uncertainty measures surrounding HICP inflation, also measured in percentage points.

Deviation over time. Fig. 1 displays the monthly deviations of the logarithmic deviation for all five countries as well as the Big-5 aggregate from January 2012 to December 2019 and their decomposition according to Eq. (8). Overall, the time series profiles observed for Germany, France, Italy and Spain exhibit quite similar characteristics. These are found in the Big-5 aggregate, too. In the case of the Netherlands, however, the pattern is distinctly different.

In total over all countries, we observe that the monthly deviations range from -0.16 to 0.31 percentage points. The ranges vary from one country to another. For instance, in the case of Germany, the range is from -0.03 to 0.31 which is about double the size of the range observed for France (from -0.01 to 0.17). In the case of the Netherlands, the range of deviations spreads from -0.16 to 0.14 and is thus more or less centered around zero. The deviations of the Big-5 aggregate fluctuate in a comparatively small range which is almost completely located above the zero line.

In the Big-5 aggregate as well as in Germany, France, Italy and Spain, there are only very few realisations which are located in the negative territory. By contrast, the majority of monthly deviations are negative in the case of the Netherlands.

As regards the decomposition of total deviations, we observe consistently positive contributions from the representativity effect for all countries and the country group. A striking feature is that the Netherlands differ systematically from the other countries in terms of the sign of the data vintage component. While it is mostly positive in the case of Germany, France, Italy and Spain as well as the Big-5 aggregate, the opposite appears for the Netherlands. This implies that, in arithmetical terms, representativity and data vintage components are typically compensating each other in the case of the latter, while they are reinforcing each other in the former group of countries.

Bias. We estimate the total upper-level aggregation bias by averaging the monthly deviations over the complete sample from January 2012 and December 2019 (see Tab. 2). For the Big-5 aggregate, the bias is positive but small. It falls short of one-tenth of a percentage point. It is also positive for Germany, France, Italy and Spain, whereas it is

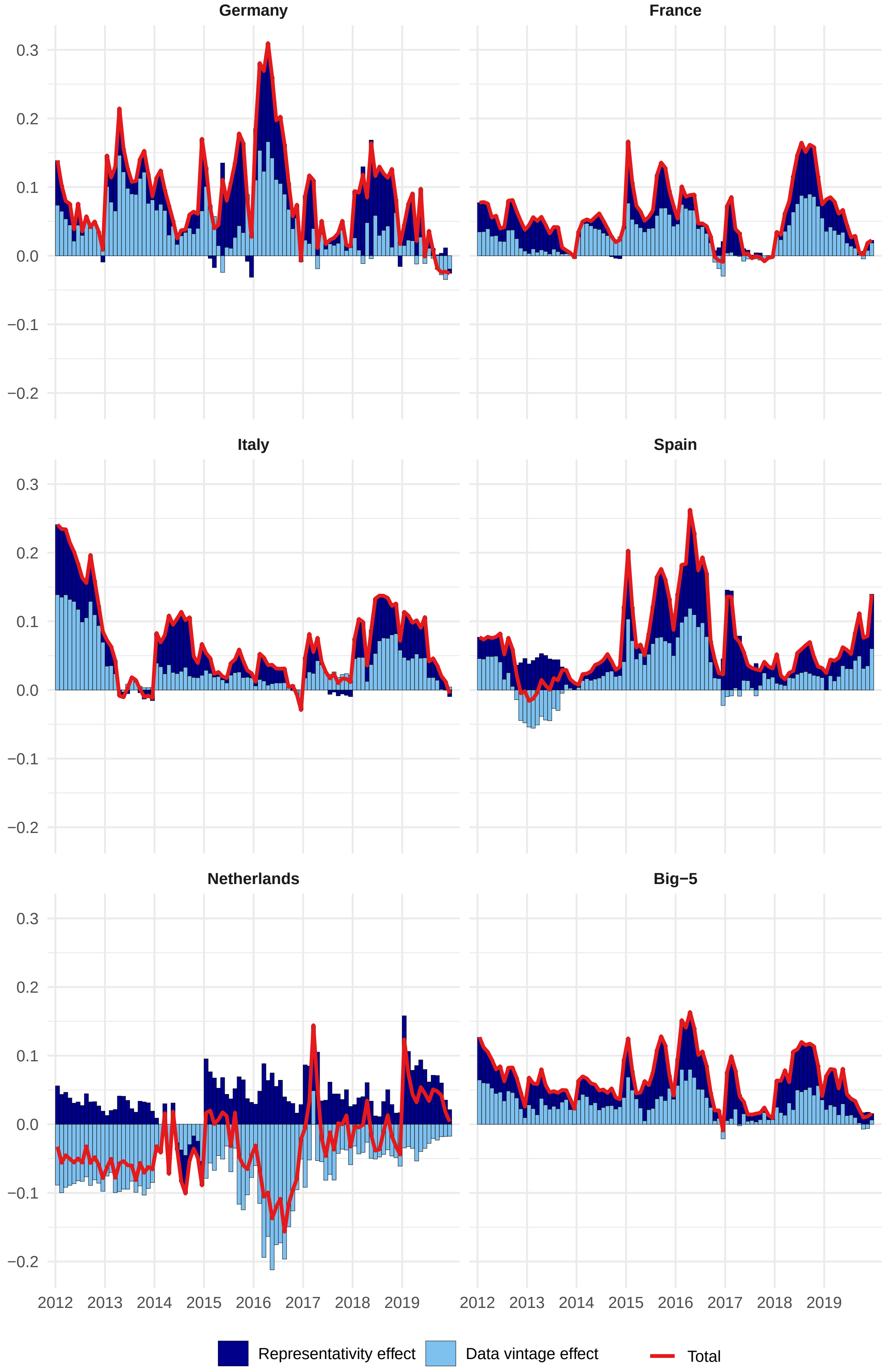

Figure 1: Monthly deviations (in percentage points p.a.) in the year-on-year change rates between official Laspeyres-type index and superlative Törnqvist index with final NA weights, decomposed into representativity and data vintage effects.

negative in the case of the Netherlands. Amongst the countries reporting a postive bias, the largest is observed for Germany and the lowest for France.

	Representativity	**Data vintage**	**Total**
Germany	0.044	0.046	0.090
France	0.027	0.029	0.056
Italy	0.031	0.036	0.068
Spain	0.042	0.025	0.067
Netherlands	0.040	-0.069	-0.028
Big-5	0.037	0.030	0.066
Euro Area	0.022	-	-

Table 2: *MD* (in percent of a Törnqvist index with final NA weights), 2012-2019.

The representativity component of the bias is positive for the Big-5 aggregate and all individual countries in this group. The data vintage component of the bias is also positive in the Big-5 aggregate as well as in Germany, France, Italy and Spain. For the Netherlands, however, the data vintage contribution to the bias is negative and – in absolute terms – strongest among all countries under review.

In the Big-5 aggregate as well as in Germany, France and Italy, the representativity and data vintage components contribute to the total bias in roughly equal shares.[10] For Spain, the data vintage component is significantly smaller than the representativity component. In the case of the Netherlands, we observe that, in absolute terms, the (positive) representativity contribution to the bias is about two-thirds the size of the (negative) data vintage contribution.

The systematic difference between the Netherlands and the remaining countries under review deserves some explanation. It turns out to be related to the fact that the Dutch statistical office chose the not-to-price-update option whereas the other statistical offices generally applied the price-updating one (recall Section 2.2). At first glance, the waiver of price updating from $y-2$ to $y-1$ turns out to be advantageous as the resulting negative data vintage effects "correct" the positive representativity bias while, with price updating, data vintage effects tend to reinforce mismeasurement at the upper level of aggregation. Theory suggests that it is a matter of own and cross-product price elasticities how the expenditures of individual goods and services respond to price changes. The assessment of whether the presence or absence of price updating is justified, thus, generally requires (empirical) knowledge about the responsiveness of all goods and services on prices. In addition, the internal consistency of statistical procedures may also be an aspect to be considered given that, according to HICP weight updating rules, the expenditures for the annual average $t-1$ have to be price-updated to December $t-1$.

[10] Regarding Germany, the size of the data vintage component relative to the representativity component differs from the results found in Herzberg et al. (2021). The reason for the lower data vintage presented in this note is differences in the derivation of the alternative weights as described in Section 2.2.

Inaccuracy. The uncertainty surrounding inflation measurement due to upper-level aggregation issues is measured by two statistical metrics, the total *RMSD* and the interdecile range (see Tab. 3). As a key takeaway from this analysis, it might be worth memorising that the interdecile range of the Big-5 aggregate is about one-tenth of a percentage point. Looking at the country results, both total *RMSD* and the interdecile range are highest for the German HICP. While Italy and Spain follow quite closely, the distance to France is more marked. Total *RMSD* and interdecile range for the Dutch HICP do not allow us to sort the Netherlands uniformly in this country ranking. While the total *RMSD* is lowest, the interdecile range is only second-smallest.

	RMSD			***IDR***
	Representativity	Data vintage	Total	
Germany	0.062	0.063	0.112	0.153
France	0.035	0.039	0.070	0.115
Italy	0.043	0.051	0.091	0.143
Spain	0.052	0.045	0.088	0.149
Netherlands	0.052	0.081	0.059	0.125
Big-5	0.043	0.036	0.075	0.097
Euro Area	0.028	-	-	-

Table 3: *RMSD* and *IDR* (in percentage points p.a.), 2012-2019.

For the Big-5 aggregate as well as for all individual countries except the Netherlands, we observe that the *RMSD* of the representativity component and the *RMSD* of the data vintage component do not differ markedly in size. With the same exception, they roughly sum up to the total *RMSD*, suggesting that the covariance term obviously does not play a recognisable role. For the Netherlands, however, the data vintage component yields a comparatively high *RMSD*. This does not result in a high total *RMSD* thanks to its compensating impact relative to the representativity component.

3.2 Weight profiles

For a thorough understanding of the data vintage effect, it may help to take a closer look at the weight estimates. As it is impossible to present weight profiles for all product categories according to each concept and country or country group, we focus on three pieces of evidence which illustrate some key insights regarding commonalities and differences in weight profiles.

First, in order to obtain a impression about the disparity between official and final NA weights across countries and over time, we report the absolute difference between the two weight concepts averaged over all product categories (see Tab. 4). The absolute differences between official and final weights are, on average, smaller in the Big-5 aggregate than for the individual countries. This is one factor explaining why the data vintage effects tend to be higher in each individual country than in the group. In the country dimension, we observe that the mean absolute weight difference is largest for the Netherlands, followed

	Germany	France	Italy	Spain	Netherlands	Big-5
2011	0.64	0.42	0.51	0.60	0.51	0.35
2012	0.59	0.42	0.61	0.65	0.83	0.33
2013	0.84	0.39	0.46	0.82	1.02	0.34
2014	0.61	0.80	0.54	0.71	1.10	0.44
2015	0.60	0.77	0.46	0.54	1.11	0.43
2016	0.63	0.36	0.34	0.72	1.10	0.34
2017	0.56	0.36	0.40	0.47	1.03	0.31
2018	0.62	0.40	0.39	0.52	0.42	0.34
2019	0.34	0.27	0.32	0.43	0.33	0.20
Average	0.60	0.47	0.45	0.61	0.83	0.34

Table 4: Mean absolute difference (in ‰-points) between offical and final NA weights.

by Spain and Germany, while Italy and France are the countries for which the lowest values are reported. Looking along the time dimension, we find some support for the hypothesis that the two weight concepts deviate the less, the closer the year is to the present.[11]

Second, Figs. A.1 through A.4 display the time profiles of official and final NA weights of four product categories for the five countries under review as well as the Big-5 aggregate. The product categories chosen are "Meat" (ECOICOP: 0112), "Garments" (0312), "Actual rentals for housing" (041) and "Electricity" (0451).[12] Overall, the plots let us conclude that official and final weights do not differ substantially.

Third, we broaden the comparison of weight concepts. For the German HICP, we take full-information weights into consideration. The main message is that final NA weights are more similar to official weights than to full-information weights. This implies that estimates of the data vintage effect using final NA weights may be regarded as a lower bound of the "true" impact of weight uncertainty on HICP mismeasurement. Details of this analysis are found in the following digression.

Digression. In Fig. 2, full-information weights of the selected product categories are added to official and final NA weights in the case of Germany. Looking at the plots for "Electricity" and "Garments", we find no clear indication that final NA weights are systematically closer to official weights than full-information weights. In the case of "Actual rentals for housing" and "Meat", however, empirical support is given towards this hypothesis. The plots for these product categories further reveal that the three weight types cluster together in 2018, which is the year when the 2015 HBS results were considered in HICP weights for the first time.

Both observations point to implications of the differing construction principles of final

[11] This hypothesis is justified by the fact that, towards the end of the sample, the final NA weights are effectively only revised NA weights and should, thus, be rather close to official weights.

[12] These categories are chosen since each of them represents one of the main categories of the HICP (food, energy, services and non-energy industrial goods).

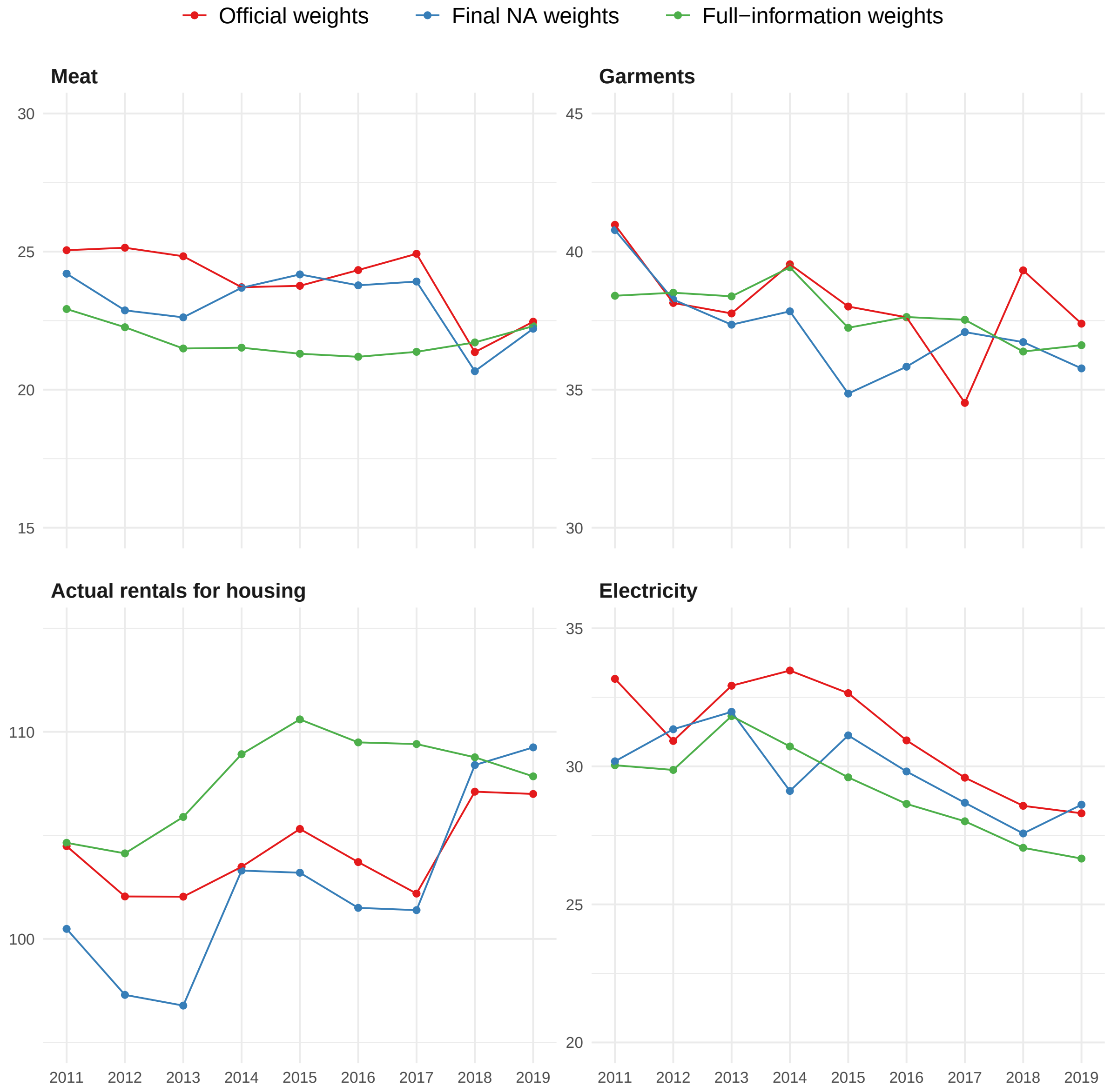

Figure 2: Official, final NA and full-information weights (in ‰) in German HICP.

NA and full-information weights as regards the consideration of HBS information and the role of national accounts. While the results of the 2015 HBS determine the full information weights in 2015 (and through interpolation and extrapolation with national accounts also the years before and after), they affect final NA weights only from 2018 onward. This means that just altering the national accounts reporting status in weight calculations is usually insufficient to "correct" a potential large deviation of official weights from the pattern implied by the HBS unknown at this date.

If we accept the view that the best estimate of representative household consumption patterns is generally derived using HBS information, a closer proximity to the unknown "true" inflation is achieved with price indices based on full-information weights rather than with price indices based on final NA weights.

In Fig. 3, the time series of the data vintage effect resulting from the use of full-

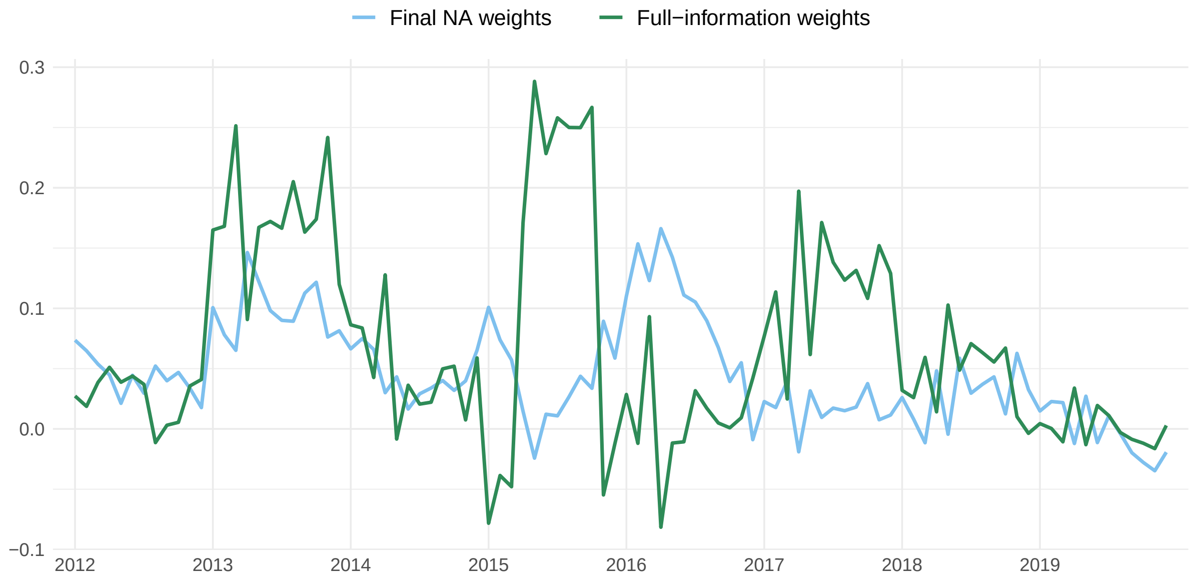

Figure 3: Data vintage effects (in percentage points p.a.) of Germany based on final NA and full-information weights.

information weights are plotted vis-à-vis their final NA weights counterpart for the German HICP. In both cases, the overwhelming number of monthly realisations lie above the zero line. However, marked differences are visible over the entire sample under review. There are more and longer time intervals where the data vintage effect calculated on the basis of full-information weights exceeds the data vintage effect calculated on the basis of final NA weights. On average, the data vintage effect amounts to 0.07 percentage points in the case of full-information weights and to 0.05 percentage points in the case of final NA weights.

3.3 Long-run evidence on the representativity component

The full-fledged analysis cannot be extended beyond the year 2012 because it is infeasible to calculate final NA weights. In principle, it would be very cumbersome but not per se impossible to gather the vintage data sets for the years prior to 2012. The infeasibility rather comes as a consequence of the flexibility statistical offices were granted in the compilation of HICP weights of that time. From the outset of the HICP until 2011, European regulation had imposed minimum standards while the harmonised weight updating procedure with a systematic use of national accounts data was implemented only in 2012 (Eiglsperger and Schackis, 2009; ECB, 2012). Of course, without this weight updating rule, final NA weights are impossible to construct, as they result from plugging in timely and more mature national accounts data in a weight updating formula which was given birth as a general standard only with the 2012 methodological change.

A look at the representativity part of upper-level aggregation over the entire HICP history is worthwhile nonetheless. Complementing the evidence of Herzberg et al. (2021),

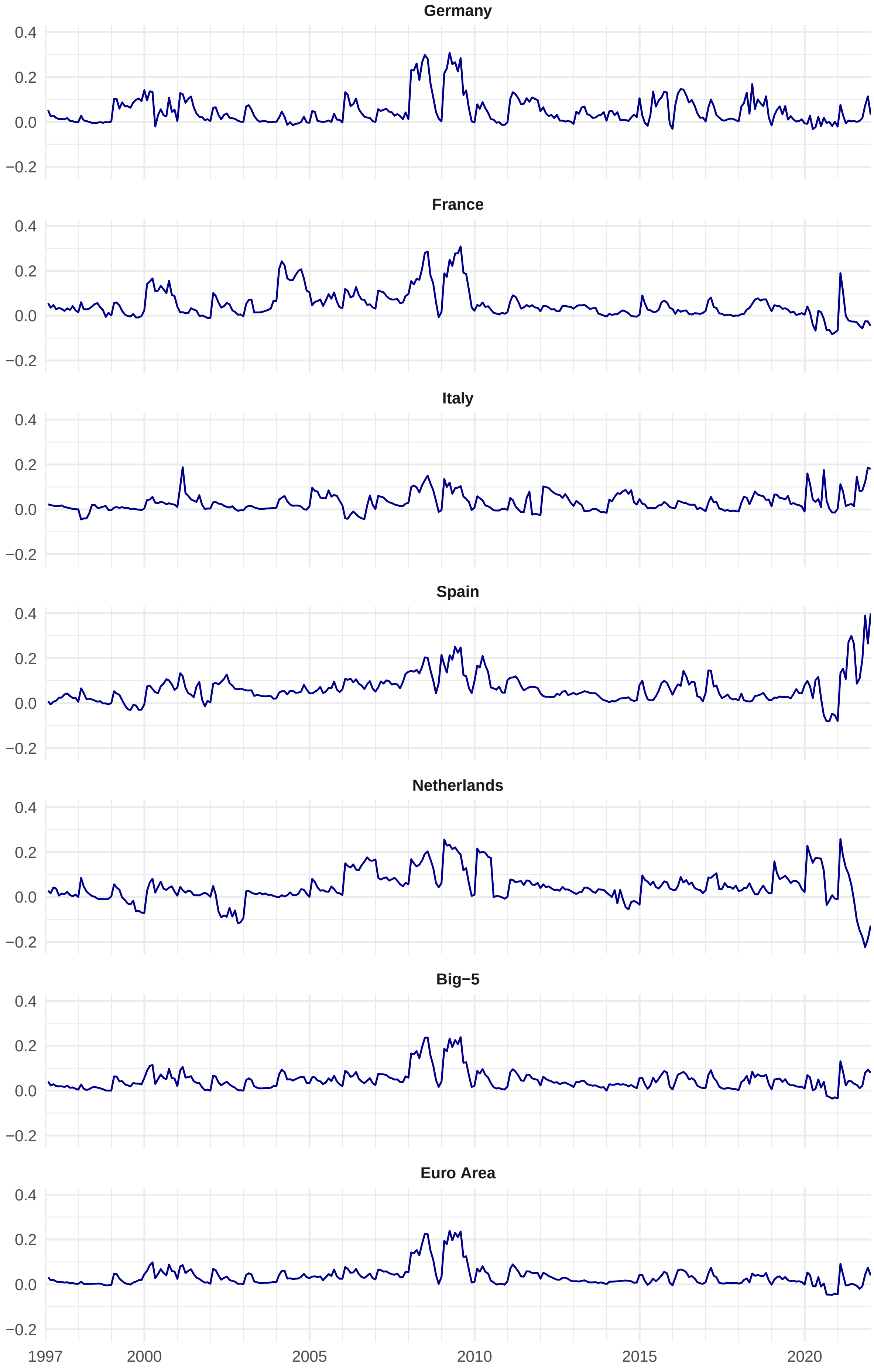

Figure 4: Representativity component (in percentage points p.a.) of several HICPs.

we report the results for bias and inaccuracy measures for the Big-5 aggregate and all individual countries in this group. Two issues are in the spotlight. The first is the identification of cyclical patterns in upper-level HICP mismeasurement. The second addresses the question whether, and to what extent, the 2012 methodological change has led to a reduction in the representativity bias.

In Fig. 4, the time series of the monthly logarithmic deviation of the HICP from the Törnqvist index using official weights are plotted for the euro area, the Big-5 aggregate and all individual countries in this group over the period from 1997 to 2021. The plots for the euro area and the Big-5 aggregate seem to closely resemble each other. Across individual countries, some differences in the time profiles are observed. With (local) maxima uniformly detected in the years of the Financial Crisis 2008/2009, however, a visible commonality is worth reporting, too. Abstracting from the COVID-19 pandemic, inflation mismeasurement due to imperfect representatitivity peaked at about one-quarter of a percentage point in the euro area and the Big-5 aggregate. With three-tenths of a percentage point, the peak was markedly higher in Germany and France whereas, in Italy, it was lower at around one-sixth of a percentage point. The multitude and size of relative price shifts which appear in economically turbulent times such as the Financial Crisis may well explain why the representativity bias was above average.

For the COVID-19 pandemic, it is difficult to derive common features for the evolution of the representativity component over time. The monthly deviations observed for the years 2020 and 2021 are not extraordinary compared with long-run patterns in the case of the euro area, the Big-5 aggregate, Germany and – abstracting from a single outlier – also France. In the case of Italy, monthly deviations turn out be moderately more volatile in the COVID-19 pandemic than in the Financial Crisis. As regards the Spanish and the Dutch HICP, sizeable and strongly oscillating realisations indicate that the COVID-19 pandemic seems to have become an even more severe challenge. Further questions arise because of the higher frequency of negative deviations observed in these years.

In Tab. 5, the mean deviations, averaged over the complete sample as well as the subsamples “Before 2012” and “Since 2012” are reported. The representativity bias is smaller for the euro area than for the Big-5 aggregate, suggesting that mismeasurement due to imperfect representativity is generally more of a problem in the larger euro area countries than the smaller ones.

	Germany	**France**	**Italy**	**Spain**	**Netherlands**	**Big-5**	**Euro Area**
Before 2012	0.057	0.073	0.026	0.071	0.049	0.056	0.047
Since 2012	0.037	0.018	0.039	0.056	0.039	0.036	0.019
Total	0.049	0.051	0.031	0.065	0.045	0.048	0.036

Table 5: *MD* (in percent of a Törnqvist index) of representativity component, 1997-2021.

The results confirm that the 2012 methodological change reduced the size of the representativity bias in the country groups and all countries under review except Italy. In the case of the euro area, the bias has been more than half the size since 2012 as it had been before. The most significant progress was made in the French HICP whereas it was rather small in the Dutch HICP. The counterintuitive result reported for Italy seems to be due mainly to the extraordinarily low mean deviation in the pre-2012 period while the estimate for the representativity bias in the period since 2012 turns out to fit the results of the other countries well. There might have been special factors in the compilation of the Italian HICP in the pre-2012 period that dampened the size of the representativity bias.

The results for the *RMSD* of the representativity component, reported in Tab. 6, reveal that the uncertainty surrounding the HICP was reduced by the 2012 methodological change. This was quite substantial in the case of the euro area. Progress in the French HICP contributed the most while a more moderate decline is observed for the German and the Dutch HICP. In these countries, the pre-2012 levels had been comparatively high. In the Italian case, we note a small increase from a low pre-2012 level, whereas the pre-2012 level of Spain was already high.

	Germany	**France**	**Italy**	**Spain**	**Netherlands**	**Big-5**	**Euro Area**
Before 2012	0.091	0.100	0.047	0.089	0.088	0.075	0.069
Since 2012	0.057	0.042	0.057	0.094	0.080	0.045	0.030
Total	0.079	0.082	0.051	0.091	0.085	0.065	0.057

Table 6: *RMSD* (in percentage points p.a.) of representativity component, 1997-2021.

4 Conclusion

We present evidence on HICP mismeasurement at the upper-level of aggregation for the five largest euro area countries and the country group as a whole. Our results show that neither the Laspeyres formula nor the updating of weights with preliminary national accounts data are major sources of bias or inaccuracy. For the Big-5 aggregate, the mean deviation of the HICP from a superlative index, based on weights being updated with timely and more mature national accounts data, clearly falls short of one-tenth of a percentage point and the interdecile range of the deviations has a length of about one-tenth of a percentage point.

The upper-level of aggregation is one stage of HICP production where mismeasurement might occur. Hence, the results of this study enlighten only a part of a multi-faceted picture and, according to existing knowledge in this field, this part is likely to be quantitatively less important than other sources of mismeasurement. In ECB (2021a, chapter 3), a comprehensive evaluation framework is sketched out and patchy evidence is presented. However, an overall assessment in the style of the Boskin Report is still lacking for the HICP.

As regards mismeasurement at the upper level of aggregation, it is generally insufficient to look at the representativity component only. As already argued by Herzberg et al. (2021), the use of more current weights through the annual updating procedure implemented in 2012 has come at the cost of relying on preliminary national accounts data. This paper is an attempt to quantify this additional source of mismeasurement for a country group representing more than four-fifths of the euro area HICP. However, the benchmark can only be specified in terms of final NA weights, implying that the estimates of the data vintage components may be interpreted as a lower bound for the "true" impact of preliminary data in weight updates (which would be better proxied using full-information weights as carried out in Herzberg et al. (2021) for the German HICP).

The results of this paper let us conclude that the representativity and data vintage components contribute in fairly equal parts to the total upper-level aggregation bias und inaccuracy since 2012. The starting date for the full-fledged analysis is forced by the concept of final NA weights which only makes sense to be applied in the period since the implementation of the annual updating of weights. With regard to the effects of this methodological change, we can therefore provide partial evidence, namely that the representativity component has been reduced. Knowledge about weight updating practice in the pre-2012 era when weight compilation methods were rather non-harmonised, apart from imposing some minimum standards, would be needed to extend the analysis in this direction.

The rules for the updating of HICP weights still includes some discretion. The freedom of choosing between the "to-price-update" and the "not-to-price-update" options turns out to have a bearing on the data vintage effect. This conclusion may be drawn from the observation that the results for the Netherlands ("not-to-price-update" option) differ systematically from the results for the other countries under review ("to-price-update"option). It might be worthwhile to further study the impact of this freedom of choice on HICP measurement in order to check whether an initiative for a harmonisation of this aspect is justified by empirical evidence. The solution might be neither of the two options but a third one which has been implemented temporarily in reaction to the strong consumption shifts during the COVID-19 pandemic (Eurostat, 2020). Perpetuating these weight updating rules would automatically solve this issue.

In the paper, we document first COVID-19 evidence for the representativity component. While the COVID-19 pandemic turns out to be non-critical for the euro area HICP in this perspective, differences are notable across countries. This may indicate that upper-level aggregation issues are generally a concern. This does not come as a surprise given the tremendous challenges which price statisticians faced by this seminal event. A thorough study incorporating many important aspects which are not at all tackled in this paper (e.g. price imputations) is needed to fully capture and explain potential mismeasurement at the upper level of aggregation during the COVID-19 pandemic.

References

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

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

ECB (2012). New standards for HICP weights, Monthly Bulletin. April 2012, Box 3, 36-39.

ECB (2021a). Inflation measurement and its assessment in the ECB's monetary policy strategy review. Occasional Paper Series 265.

ECB (2021b). The ECB's monetary policy strategy statement. hiips://www.ecb.europa.eu/home/search/review/pdf/ecb.strategyreview_monpol_strategy_statement.en.pdf. Accessed: 5 April 2022.

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

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

Eurostat (2018). *Harmonised Index of Consumer Prices (HICP): Methodological Manual*. European Union, Luxembourg.

Eurostat (2020). Guidance on the compilation of HICP weights in case of large changes in consumer expenditures. hiips://ec.europa.eu/eurostat/documents/10186/10693286/Guidance-on-the-compilation-of-HICP-weights-in-case-of-large-changes-in-consumer-expenditures.pdf. Accessed: 1 December 2021.

Greenlees, J. S. and Williams, E. (2010). Reconsideration of weighting and updating procedures in the US CPI. *Jahrbücher für Nationalökonomie und Statistik*, 230(6):741–758.

Herzberg, J., Knetsch, T., Schwind, P., and Weinand, S. (2021). Quantifying bias and inaccuracy of upper-level aggregation in HICPs for Germany and the euro area. Discussion Paper 06/2021, Deutsche Bundesbank.

Hoffmann, J. (1998). Problems of inflation measurement in Germany. Discussion Paper 01/1998, Deutsche Bundesbank.

Knetsch, T. A. and Reimers, H.-E. (2009). Dealing with Benchmark Revisions in Real-Time Data: The Case of German Production and Orders Statistics. *Oxford Bulletin of Economics and Statistics*, 71(2):209–235.

Lequiller, F. (1997). Does the French consumer price index overstate inflation? *INSEE Série des documents de travail de la Direction des Etudes et Synthèses Économiques*, 3.

Lowe, J. (1823). *The present state of England in regard to agriculture, trade, and finance.* Second Edition. London: Longman, Hurst, Rees, Orme and Brown.

Neves, P. D. and Sarmento, L. M. (1997). The substitution bias of the consumer price index. *Banco de Portugal Economic Bulletin*, pages 25–33.

Silver, M. and Ioannidis, C. (1994). The measurement of inflation; untimely weights and alternative formulae: European evidence. *The Statistician*, 43(4):551–562.

A Weight profiles

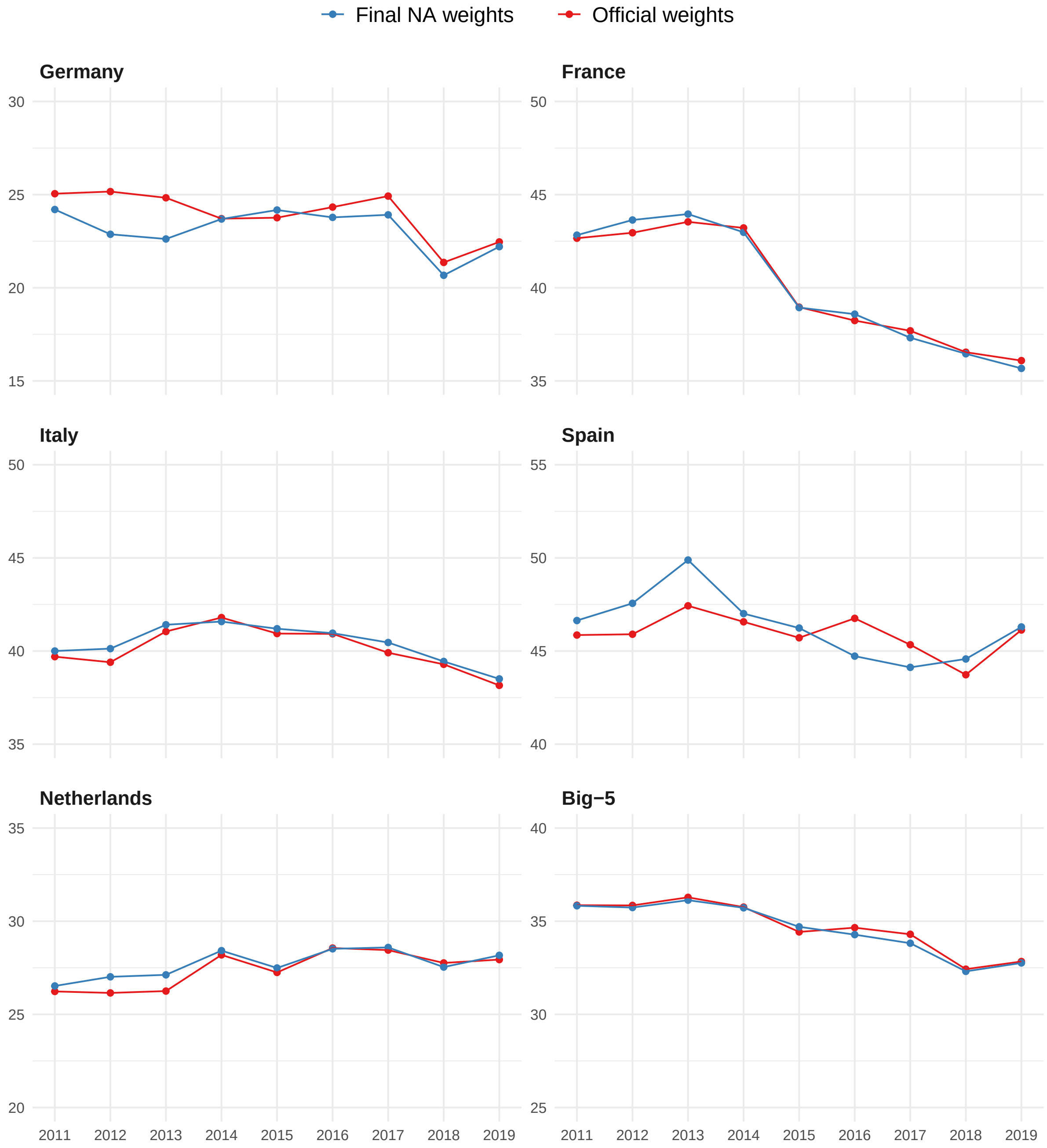

Figure A.1: Weight profiles for "Meat".

Note: 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. The vertical axis displays the weight (in ‰) of *meat* in the HICP of the respective country.

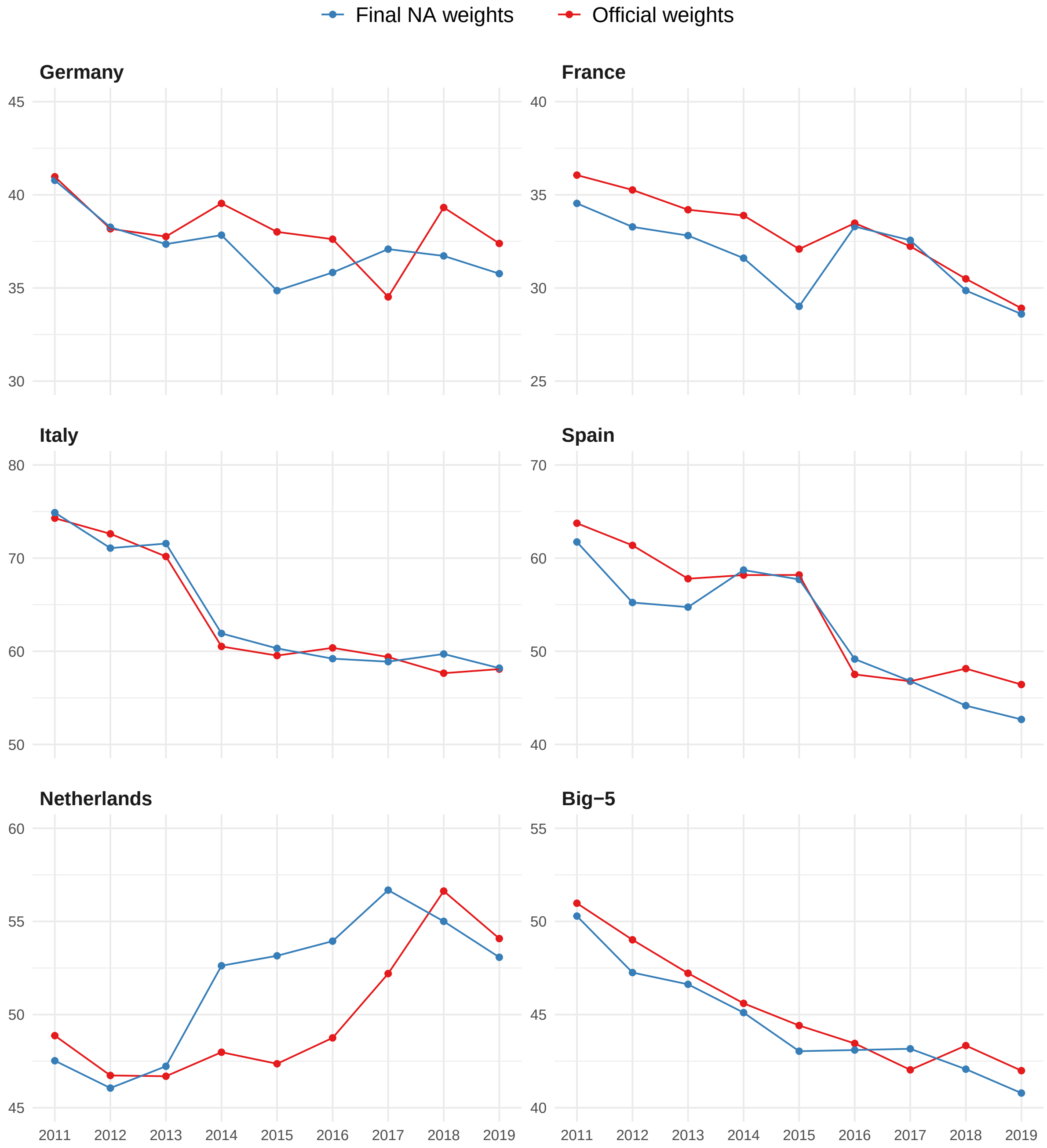

Figure A.2: Weight profiles for “Garments”.

Note: 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. The vertical axis displays the weight (in ‰) of *garments* in the HICP of the respective country.

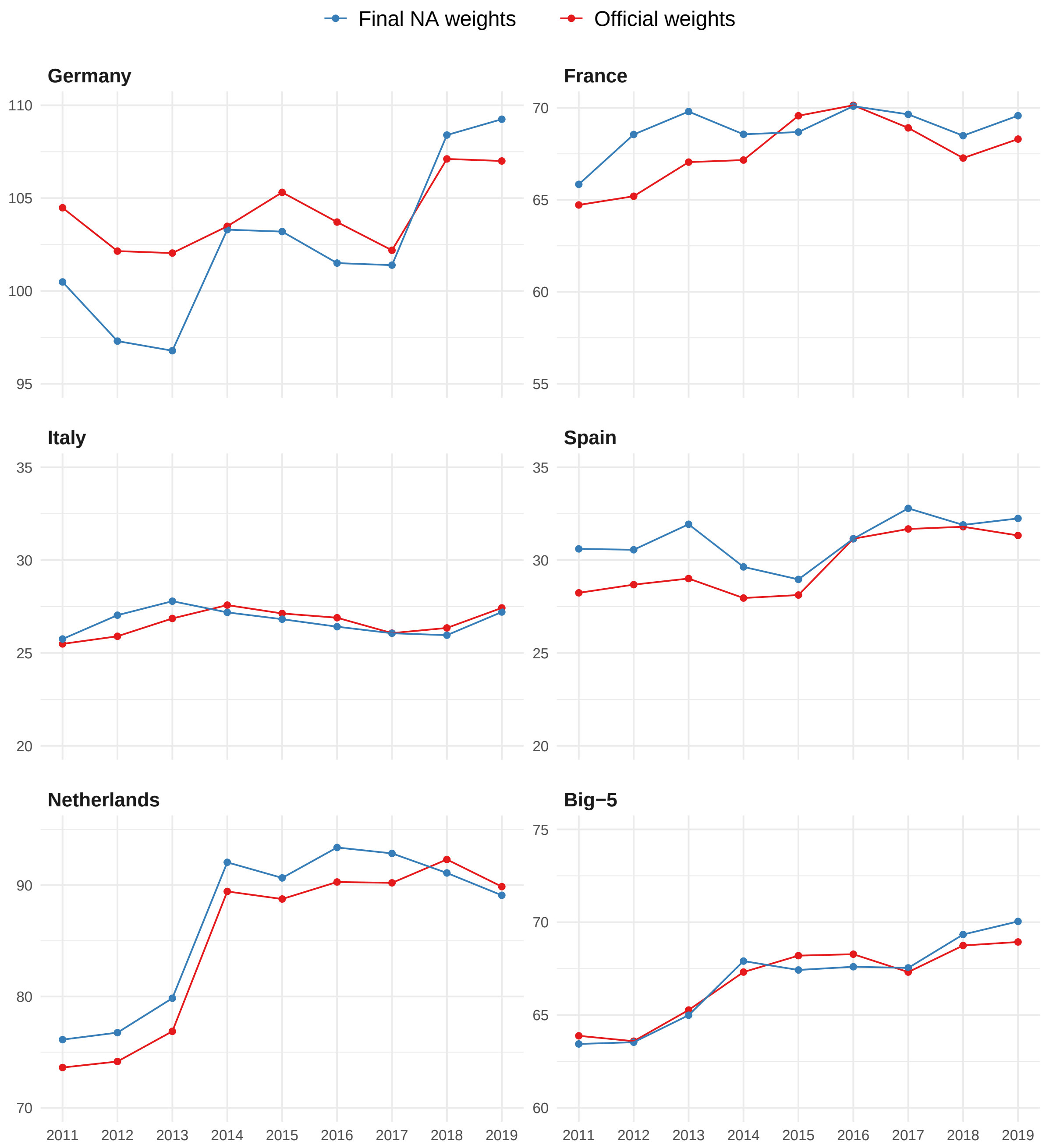

Figure A.3: Weight profiles for "Actual rentals for housing".

Note: 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. The vertical axis displays the weight (in ‰) of *actual rentals for housing* in the HICP of the respective country.

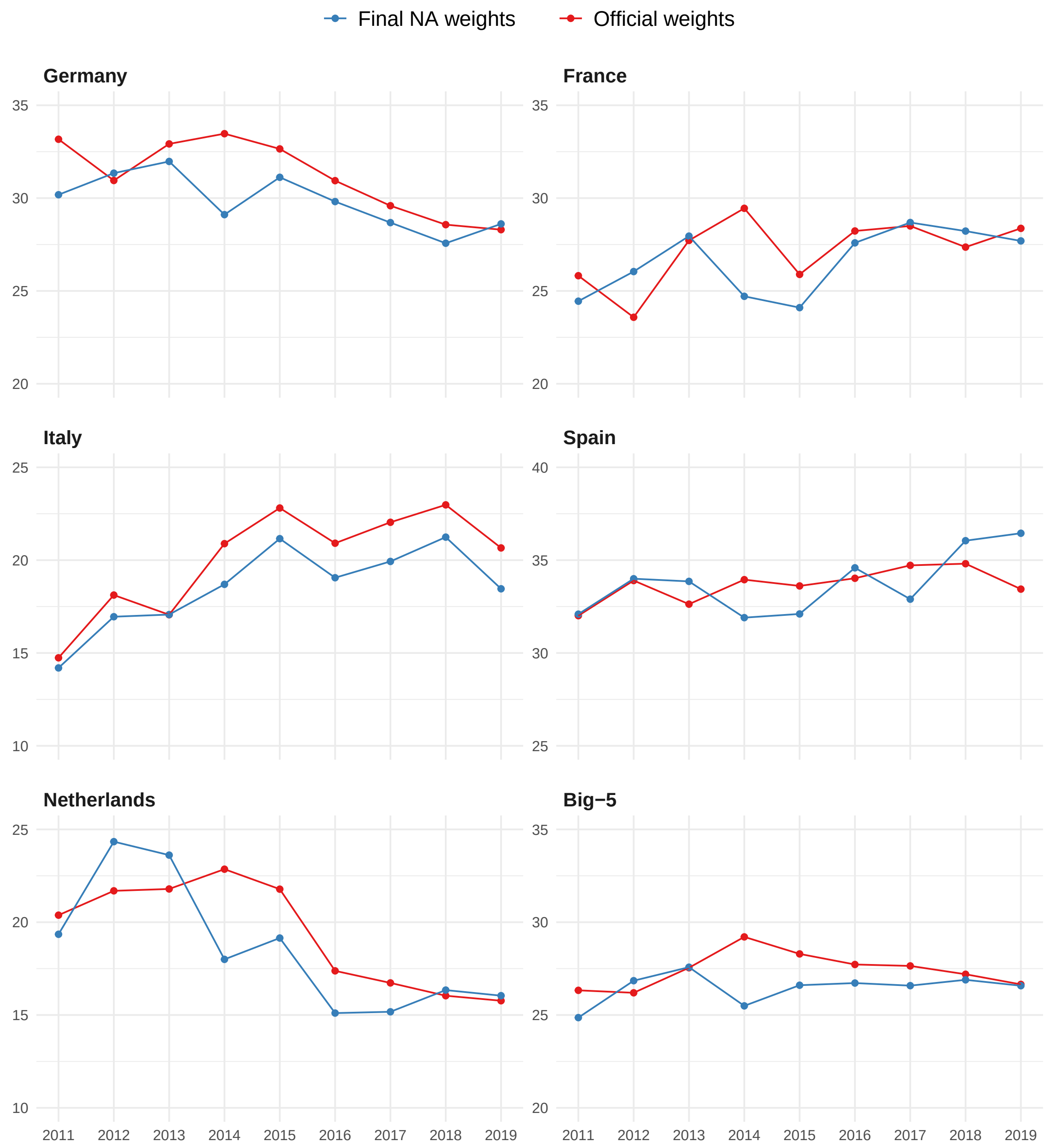

Figure A.4: Weight profiles for "Electricity".

Note: 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. The vertical axis displays the weight (in ‰) of *electricity* in the HICP of the respective country.