STATISTICS IN TRANSITION new series, September 2022
Vol. 23, No. 3, pp. 1–25, DOI 10.2478/stattrans-2022-0027
Received – 23.02.2022; accepted – 09.06.2022

Assessing the effect of new data sources on the Consumer Price Index: a deterministic approach to uncertainty and sensitivity

Jacek Białek[1], **Tomasz Panek**[2], **Jan Zwierzchowski**[3]

ABSTRACT

One of the greatest challenges facing official statistics in the 21st century is the use of alternative sources of data about prices (scanned and scraped data) in the analysis of price dynamics, which also involves selecting the appropriate formula of the price index at the elementary group (5-digit) level. When consumer price indices of goods and services are constructed, a number of subjective decisions are made at different stages, e.g. regarding the choice of data sources and types of indices used for the purpose of estimation. All of these decisions can affect the bias of consumer price indices, i.e. the extent to which they contribute to the overall uncertainty about the resulting index values. By measuring how robust consumer price indices are, one can assess the impact that the decisions made at the different stages of index construction have on the index values. This assessment involves analysing uncertainty and sensitivity. The purpose of the study described in the article was to determine how much and in which direction the consumer price index changes when including scanner and scraped data in the analysis, in addition to the data on prices collected by enumerators. The impact of these new data sources was assessed by analysing uncertainty and sensitivity under the deterministic approach. To the best of the authors' knowledge, it is a novel application of robustness analysis to measure inflation using new data sources. The empirical study was based on data for February and March 2021, while scanner and scraped data about selected categories of food products were obtained from one retail chain operating hundreds of points of sale in Poland and selling products online. It was found that the choice of a data source has the most significant impact on the final value of the index at the elementary group level, while the choice of the aggregation formula used to consolidate different data sources is of secondary importance.

Key words: price indices, scraped data, scanner data, robustness analysis, inflation.

JEL: C43, E31

[1] University of Lodz, Department of Statistical Methods, E-mail: jacek.bialek@uni.lodz.pl; Statistics Poland, Department of Trade and Services, Poland. E-mail: J.Bialek@stat.gov.pl. ORCID: https://orcid.org/0000-0002-0952-5327.

[2] Warsaw School of Economics, Institute of Statistics and Demography, Poland. E-mail:tompa@sgh.waw.pl. ORCID: https://orcid.org/0000-0002-1034-7222.

[3] Warsaw School of Economics, Institute of Statistics and Demography, Poland. E-mail:jzwier@sgh.waw.pl.

1. Introduction

Traditionally, data used to measure inflation are provided by enumerators who collect information about prices and characteristics of selected products in randomly selected shops located in the so-called price survey regions (there are 207 price survey regions in Poland, where enumerators visit about 35,000 shops). Pandemic-related difficulties affecting the traditional method of price data collection, such as the requirement of social distancing and shop closures, and the growing volume of online shopping have provided stronger motivation to intensify work on the use of alternative data sources, such as scanner data and scraped data.

For the purpose of this article, **scanner data** are defined as detailed data about consumer products obtained by scanning bar codes at electronic points of sale (CPI Manual, 2004). The list of barcodes most commonly used by retail chains includes GTIN (Global Trade Item Number) or its European version – EAN (European Article Number), PLU (Price Look-Up) and SKU (Stock Keeping Unit). Product codes, including the code assigned by a given chain store, together with the product label, are used for classifying products using 5-digit ECOICOP codes and below this level of aggregation (Chessa, 2015; 2016, Białek and Beręsewicz, 2021). One of the benefits of using scanner data is that they contain information about the level of consumption even at the lowest level of aggregation.

Scraped data are collected automatically from websites by a special computer programme called a scraper. The programme collects "raw" data, which are then cleaned and formatted to enable further analysis. Scraped data can be collected with greater frequency than data from other sources (usually they are collected daily), which is useful for understanding consumer behaviour patterns and data variability. One should bear in mind, however, that scanner data represent actual prices, while scraped data represent 'merely' offered prices, without providing any information about the level of consumption.

One of the biggest challenges facing official statistics in the 21st century is the use of alternative sources of data about prices (scanner and scraped data) in the analysis of consumer price dynamics, which also involves choosing the appropriate formula of the price index at the elementary group (5-digit) level (Chessa, 2015; 2016, de Haan et. al., 2021). Decisions about whether or not to include a given data source in the measurement of inflation, as well as the choice of a price index formula, can have a measurable impact on the bias of consumer price indices (Saisana, Saltelli and Tarantola, 2005; Sharpe and Salzman, 2004; Nardo et al., 2005 and Nardo et al., 2011), i.e. they contribute to the overall uncertainty about the resulting index values.

When applied to consumer price indices, **robustness analysis** can help to assess the impact of decisions made at different stages of index construction on the value of the index; the process involves uncertainty analysis and sensitivity analysis.

The underlying idea of **uncertainty analysis** is to construct a model linking input variables representing sources of uncertainty (assumptions made at different stages of constructing a consumer price index), determine distribution functions of input variables, and, based on this information, determine the distribution of the consumer price index or output variables used for measuring the impact of changes in the underlying assumptions of the index on its value. Uncertainty analysis itself consists in analysing parameters of the distribution of the consumer price index.

The goal of **sensitivity analysis** is to assess what share of the variance of the consumer price index is due to each of the identified sources of uncertainty (each initial assumption). This is achieved by decomposing the variance into variances explained by particular input variables representing types of assumptions made at different stages of index construction. Sensitivity analysis is therefore closely connected with uncertainty analysis. By combining these two types of analysis, one can measure how robust consumer price indices are when one changes assumptions regarding their construction, in other words, one can analyse the impact of these assumptions on the value of the consumer price index, which in turn affects estimates of consumer price fluctuations. In practice, steps taken during uncertainty and sensitivity analysis of consumer price indices depend on what sources of uncertainty (stages of index construction) and what assumptions concerning these sources (variants of solutions adopted at these stages) are made during a particular analysis.

The purpose of the study described in the article was to determine how much and in which direction the consumer price index changes as a result of including scanner and scraped data, in addition to price data collected by enumerators. The impact of these new data sources was assessed by analysing uncertainty and sensitivity under the so-called **deterministic approach**. To the best of the authors' knowledge, it is a novel application of robustness analysis to measure inflation using new data sources. The empirical study was based on data for February and March 2021, while scanner and scraped data about selected categories of food products were obtained from one retail chain operating hundreds of points of sale in Poland and selling products online. It was found that the choice of a data source has the biggest impact on the final value of the index at the elementary group level, while the choice of the aggregation formula used to consolidate different data sources is of secondary importance.

The article has the following structure: section 2 provides a description of the study sample and the main stages during which data scanned and scraped by Statistics Poland are prepared for the analysis of consumer price dynamics in cooperation with one retail chain. Section 3 presents a description of the research method, in particular uncertainty

and sensitivity analysis, and includes a list of index formulas used in the study. Section 4 is devoted to the analysis of the results, while section 5 contains a summary of the study and conclusions.

2. Study sample

For the last three years, a consortium consisting of Statistics Poland, the Institute of Computer Sciences of the Polish Academy of Sciences and Warsaw School of Economics has been running a project called *InstatCeny*, with the goal of exploiting alternative sources of data to calculate Consumer Price Index (CPI). This article presents preliminary results of a study based on data about food products that have been obtained for the project. It may be treated as some continuation of the previously started research on these group of products (Białek et al., 2021). Given the reference period of available data (Statistics Poland has been scraping price data about food products only since the start of 2021), price indices were calculated using three sources of data about prices observed in February and March 2021. Seven elementary groups of food products were selected: rice, raw and whole milk, fresh low fat milk, yoghurt, beverages and other milk products, sugar and coffee. Data collected by enumerators in 207 price survey regions included information about each representative of the selected elementary groups, in each case, the actual price was recalculated to a fixed unit of measurement (e.g. price per litre for milk, price per kilogram for rice, etc.). In the case of scanner and scraped prices, the list of product representatives used to create subcategories of elementary groups were extended to include 3 new items: yoghurt flavoured with chocolate and fruit, powdered sugar, ground coffee. As these subcategories were sufficiently well represented and homogenous, they were included in the estimation of price changes in the corresponding ECOICOP elementary groups, despite certain discrepancies with respect to the list of representatives in the classification.

2.1. Acquisition of scanner data

The manner in which scanner data are acquired differs depending on the retail chain that supplies them. Statistics Poland uses secure (encrypted) transfers and, in the case of one retail chain – direct downloads by means of an application programming interface (API). What types of variables are provided also depends on the supplier; as regards scanner data used in the study, in addition to codes enabling product classification, the transferred *csv* data frame contained information about product ID, unit of sale, transaction date, selling price, total, a flag relating to discounts, sales and promotions and the amount of VAT. The seven elementary groups of products

provided by the retail chain and used in the study amounted to 32MB of scanner data per month.

2.2. Acquisition of scraped data

Data for the *InstatCeny* project are scraped using Python scripts that rely on the Selenium package (Białek et al., 2021). Scrapers developed by Statistics Poland have been running since the start of 2021 and scraped data are saved and archived in the form of JSON files. The range of variables included in scraped data is very similar to that found in files transferred directly by the retail chain. It turned out that products shown on the retail chain's website represent between 40 and 90% of products of the same category that can be found in the chain's stores. For example, at the start of 2021, 33 products were classified as rice in scanner data, compared to 27 in scraped data. In the case of coffee, the ratio was 275 to 152. The reason why not all products found on shelves are included in the online offering is that websites probably only feature the most popular products. The elementary groups of products provided by the retail chain and used in the study amounted to 4MB of scraped data per month.

2.3. Preparation of data from alternative sources for processing

After scanner and scraped data had been cleaned (i.e. standardising names, removing incorrect data and unusual prices), all products were classified into appropriate elementary groups and the 6-digit codes. Products found in both alternative sources were classified into categories on the basis of product labels and previously created dictionaries of key words and phrases. The process was performed using the *data_selecting* and *data_matching* functions from the *PriceIndices* R package (Białek, 2021). Text labels were compared using the Jaro-Winkler distance (Jaro, 1989; Winkler, 1990), with the threshold distance (above which two labels were regarded as different) set to 0.02. Next, the sample of scanned products was filtered in order to remove products with extreme price fluctuations (3% of cases) as well as those with relatively low levels of sales (up to 25% of products, depending on the category). In the case of scraped data, only extreme price changes were filtered out, with cut-off values equal to 0.25 and 3 for the ratio of March to February prices, which had a negligible effect on the sample size (only two products from the yoghurt category were removed). It should be noted that the "monthly price" is determined differently for each data source. In the case of scanner data, it is defined as the ratio of total sales of a product in value to its total sales in quantity, which is known as unit sales. In the case of data scraped each month (regardless of whether or not the product was sold), the monthly price is calculated as the average of all observations scraped in a given month.

Figure A.1 in the appendix shows an illustrative comparison of monthly prices for March 2021 from the three data sources regarding the four most numerously

represented product categories (coffee beans, natural yoghurt, long-grain rice, wheat flour). In general, scraped prices are characterised by the highest level of variability, while prices collected by enumerators are the least variable. However, this pattern does not hold for all categories of products: e.g. prices of long-grain rice from scanner data show the biggest fluctuations. The relatively smallest number of price outliers were observed in scanner data, which is not surprising given that these data were subjected to three kinds of filters, as described above. The biggest amount of noise was found in scraped data, despite the application of the price outlier filter at the level of GTIN code. As regards price outliers in data collected by enumerators, three such cases were recorded with respect to long-grain rice, which is a kind of exception. There is no doubt, however, that differences in average prices obtained from the three sources can be considerable, with scraped prices, on average, tending to be higher than those from the other two sources (again with the exception of long-grain rice).

3. Method description

The analysis of consumer price indices involved three sources of uncertainty, representing three kinds of decisions made during index construction: what data source with consumer prices is used, what formula is used for aggregating indices of 6-digit subgroups into 5-digit elementary groups within each data source, what formula is used for aggregating indices of price changes within elementary groups based on different data sources into total indices for each of the elementary groups. The purpose of the analysis was to assess the robustness of consumer price indices calculated for 7 elementary groups of products within the food division of the COICOP classification.

Consumer price indices for the selected elementary groups were estimated using six different sources of information about consumer prices of products classified into these elementary groups in February and March 2021:

- consumer prices surveyed by enumerators,
- data from the IT system of the retail chain (scanner data),
- online prices of the retail chain (scraped data),
- consumer prices surveyed by enumerators combined with retailer's online prices,
- consumer prices surveyed by enumerators combined with retailer's scanner data,
- all three data sources combined.

Consumer price indices for the selected elementary groups are calculated by aggregating price indices of 6-digit subgroups within each elementary groups. Data collected by enumerators and those from online price listings do not contain information about quantities of products purchased within 6-digit categories of each subgroup (or about the share of each subgroup in the total value of products sold within each elementary group). As a result, consumer price indices for 6-digit subgroups were

aggregated into elementary groups' indices (5-digit level) using unweighted geometric mean, known as the Jevons index (CPI Manual, 2004):

$$ {}_{J}I_{G} = \sqrt[n]{\prod_{i=1}^{n} I_{G,g_i}}, \qquad G=1,2,\ldots,7, \tag{1} $$

where:

I_{G,g_i} , - price indices for the *i*-th subgroup of the subgroup *G*-th elementary group,

${}_{J}I_{G}$ - the Jevons index for *G*-th elementary group.

In contrast, scanner data contain information not only about prices but also about amounts of products purchased within particular subgroups of each elementary group. These amounts were also used as weights in the process of aggregating subgroups of elementary groups for the other data sources. In this way, in addition to using the unweighted Jevons formula, price indices for subgroups could also be aggregated by employing weighted indices, namely the Laspeyres, Paasche, Fisher and Törnqvist indices (CPI Manual, 2004):

$$ {}_{L}I_{G} = \sum_{i=1}^{n} w_{G,g_i}^{0} I_{G,g_i}, \qquad G=1,2,\ldots,7, \tag{2} $$

$$ {}_{P}I_{G} = \frac{1}{\sum_{i=1}^{n} w_{G,g_i}^{1} \frac{1}{I_{G,g_i}}}, \qquad G=1,2,\ldots,7, \tag{3} $$

$$ {}_{F}I_{G} = \sqrt{{}_{L}I_{G} \cdot {}_{P}I_{G}}, \qquad G=1,2,\ldots,7, \tag{4} $$

$$ {}_{T}I_{G} = \prod_{i=1}^{n} (I_{G,g_i})^{\frac{w_{G,g_i}^{0} + w_{G,g_i}^{1}}{2}}, \qquad G=1,2,\ldots,7, \tag{5} $$

where:

${}_{L}I_{G}$, ${}_{P}I_{G}$, ${}_{F}I_{G}$, ${}_{T}I_{G}$ - price indices proposed by Laspeyres (1871), Paasche (1874), Fisher (1922) and Törnqvist (1936), respectively, for the *G*-th elementary group,

$w_{G,g_i}^{0}, w_{G,g_i}^{1}$ – weight of the *i*-th subgroup of the *G*-th elementary group in the base period (February 2021) and in the reference period (March 2021).

The weight for a given subgroup in the base period is calculated by dividing the total value of products in this subgroup sold in the base period by the total value of sold products in all subgroups of a given elementary group. The same method is used to calculate weights for subgroups in the reference period (weight values are presented in Table A.1 in the Appendix).

When price indices are calculated using more than one source of data, the price index for each elementary group is calculated by aggregating price indices for these elementary groups that were calculated separately on the basis of each data source. This

aggregation was performed using the Young index or the geometric Young index (Białek, 2017):

$$ {}_{Y}I_G = (I_G^N)^{w_N^\tau} + (I_G^W)^{w_W^\tau} + (I_G^S)^{w_S^\tau}, \qquad G=1,2,\ldots,7, \tag{6} $$

$$ {}_{YG}I_G = (I_G^N)^{w_N^\tau}(I_G^W)^{w_W^\tau}(I_G^S)^{w_S^\tau}, \qquad G=1,2,\ldots,7, \tag{7} $$

where:

w_N^τ , w_W^τ , w_S^τ – index weight for the G-th elementary group, established on the basis of a period more distant than the base period (in our case, it was 2020), calculated using prices surveyed by enumerators, scraped prices and scanner data, respectively.

${}_{Y}I_G$, ${}_{YG}I_G$ - the Young index or the geometric Young index for the G-th elementary group.

In the case of price indices calculated on the basis of more than one data source, shares of purchases within particular elementary group from a given source in total sales from all sources combined were used as weights. Shares of products purchased online were estimated on the basis of information obtained from the Household Budget Survey conducted by Statistics Poland. Shares of purchases in chain stores were obtained from databases maintained by Passport GMID, Euromonitor International as well as domestic market surveys conducted by Statistics Poland (Table A.2 in the Appendix).

Price indices for the analysed subgroups are shown in Table A.3 in the Appendix. In the analysis, the consumer price index for a given elementary group is a composite index, based on consumer price indices for its subgroups using the different data sources. When price data come from more than one source, the composite index is calculated in two steps. The first step consists in calculating price indices for each elementary group using data from each source. In the second step, price indices for a given elementary group calculated in the first step are aggregated into a single price index for this particular elementary group, which takes into account information from all data sources.

From a formal perspective, the dependence of the consumer price index for a given elementary group on input variables representing the assumptions made during index construction can be expressed using the following model:

$$ I_G = f_{rs}\left(\boldsymbol{I}_{G,g}^Z, \boldsymbol{w}_{G,g}^S, \boldsymbol{w}_G^Z\right), \qquad G=1,2,\ldots,7, \tag{8} $$

where:

$\boldsymbol{I}_G$ – value of the consumer price index for the G-th elementary group,

$\boldsymbol{I}_{G,g}^Z$ – a vector of price indices for subgroups of the G-th elementary group calculated on the basis of different sources of price data (Z=N,W,S),

$\boldsymbol{w}_{G,g}^{S}$ – a vector of weights of price indices for subgroups of the *G*-th elementary group calculated on the basis of scanner data (Laspeyres, Paasche, Fisher and Törnqvist indices),

$\boldsymbol{w}_{G}^{Z}$ - a vector of weights of price indices for subgroups of the *G*-th elementary group calculated on the basis of different sources of price data (Z=N,W,S),

f_{rs} – a function transforming consumer price indices for subgroups calculated on the basis of sources of price data and by applying weights assigned to price indices for subgroups and elementary groups obtained from the *s*-th source of price information (*s*-th combination of sources of price data when data come from different sources) and the *r*-th index formula (the *r*-th combination of index formulas when data come from more than one source).

The purpose of uncertainty and sensitivity analysis was to examine changes in the values of the composite index of consumer prices identified at the level of elementary groups, which result from changes of assumptions made during the estimation of the index.

3.1. The main idea of the analysis of uncertainty

Uncertainty analysis aims at quantifying the variability of the output that is due to the variability of the input. One can distinguish two main groups of methods for analysing uncertainty, namely probabilistic and deterministic methods. Probabilistic methods involve simulations based on different assumptions regarding the construction of a composite index (in our study, the consumer price index for a given elementary group), which are treated as inputs of uncertainty. The simulation model should reflect the probabilistic nature of the phenomenon of interest (Saisana, Saltelli and Tarantola, 2005; OECD, 2008, Panek 2016).

The most commonly used probabilistic method of uncertainty analysis is the Monte Carlo method (Saisana, Saltelli and Tarantola, 2005; OECD, 2008; Panek, 2016). The Monte Carlo approach involves a multivariate assessment of the proposed model with quasi-randomly selected parameters (input variables describing assumptions made during the construction of the composite index). The procedure consists of a few steps. In the first step, we determine the probability distribution function of each input variable (each assumption) X_k, k=1,2,…,z. In our study, the first input variable (X_1) represents the choice of a source of data about consumer prices, the second input variable (X_2) represents index formulas used for aggregating subgroup indices within each data source into elementary group indices, and the third input variable (X_3) refers to index formulas used for aggregating elementary group indices calculated on the basis of different data sources into composite indices for each elementary group. All these assumptions are random variables with discrete distributions. We then randomly

generate N combinations of independent input variables $\boldsymbol{X}^l$, l=1,2,…,N. The set $X^l = X_1^l, X_2^l, \ldots, X_k^l$ of combinations of input variables constitutes a sample. Such a sample can be generated using different sampling methods, such as simple random sampling, stratified sampling, or quasi-random sampling (Saltelli, Chan and Scott, 2000). For each sample (combination of assumptions), the model is assessed by calculating values of input variables (in our study, values of the consumer price index). The sequence $\boldsymbol{Y}^l$, can be used to estimate empirical probability distribution functions of particular input variables and their In deterministic methods we do not use the simulation model. Instead of the simulation model, the value of a composite index (in our study, the consumer price index for a given elementary group) for all possible N combinations of independent assumptions $\boldsymbol{X}^l$. for a given elementary group, is calculated (see Table A.4 in the Appendix). Values of this index determine its distribution, which is used to assess how robust the consumer price index for a given elementary group is to changes of its underlying assumptions regarding the choice of data sources and index formulas used in the process of aggregation.

3.2. The main idea of the analysis of sensitivity

The purpose of the analysis of sensitivity is to determine how particular assumptions (input variables) underlying a given price index values.

When a few sources of uncertainty are simultaneously taken into account when modelling a composite index, a nonlinear model can be used. In the analysis of sensitivity, good results have been achieved by applying methods based on the analysis of variance (Chan et al., 2000; Saltelli et al., 2008; Panek, 2016). As in the case of uncertainty analysis, depending on data characteristics, sensitivity analysis can be performed by applying probabilistic or deterministic methods. In the probabilistic approach, sensitivity indices are usually estimated by means of Sobol's method (1993), modified by A. Saltelli (2002). Sobol's method, as already mentioned, uses quasi-random sampling to determine distributions of input variables. Sensitivity of a composite index (the consumer price index for a given elementary group) to different parameters (input variables, i.e. assumptions made during its construction) is assessed on the basis of sensitivity indices, which are calculated after decomposing the total output variance $D^2(Y)$ of the output variable Y (in our study, the consumer price index):

$$D^2(Y) = \sum_{k=1}^{z} V_k + \sum_{k=1}^{z} \sum_{\substack{k'=1 \\ k<k'}}^{z} V_{k,k'} + \ldots + V_{1\ldots z}, \tag{9}$$

where:

$$V_k = D_{X_k}^2\left[E_{X_{-k}}(Y|X_k)\right], \tag{10}$$

$$V_{kk'} = D_{X_k X_{k'}}^2\left[E_{X_{-kk'}}(Y|X_k, X_{k'})\right] - D_{X_k}^2\left[E_{X_{-k}}(Y|X_k)\right] - D_{X_{k'}}^2\left[E_{X_{-k'}}(Y|X_{k'})\right] \tag{11}$$

The first term of equation (11) provides assessment of the direct impact of the input variable X_k, (in our example the input variable represents the choice of a particular method at a given stage of constructing a composite index) on the total output variance of the Y output variable. The second term of equation (11) represents the impact of the interaction between the k-th and k'-th input variable on the total variance of the Y output variable (the indirect impact of variable X_k on the total output variance of the Y output variable).

The direct impact of input variables on the values of a composite index (assuming there are no interactions between input variables) is measured by first-order sensitivity indices:

$$S_k = \frac{V_k}{V} = \frac{D^2_{X_k}\left[E_{X_{-k}}(Y|X_k)\right]}{D^2(Y)}, \qquad k=1,2,\ldots,z. \tag{12}$$

The model without interactions between input variables is known as an additive model. In this case, the sum of all first-order sensitivity indices is equal to 1 ($\sum_{k=1}^{z} S_k = 1$).

In the case of a non-additive model, one needs to estimate higher-order sensitivity indices, which measure the interaction effects among the set of input variables. However, in practice, they are rarely calculated, since given a model with k input variables, the number of sensitivity indices that would have to be estimated equals 2^k-1. For this reason, the impact of interactions between input variables on the variance of the output variable is calculated indirectly. In the first step, we calculate total sensitivity indices which measure the total impact of the input variable X_k on the on the total output variance of the Y output variable, i.e. the direct impact as well as that resulting from interactions between all possible combinations of other input variables:

$$S_k^T = \frac{D^2(Y) - D^2_{X_{-k}}\left[E_{X_k}(Y|X_{-k})\right]\left[E_{X_k}(Y|X_{-k})\right]}{D^2(Y)} = \frac{E_{X_{-k}}\left(D^2_{X_k}(Y|X_{-k})\right)}{D^2(Y)}. \tag{13}$$

The analysis of sensitivity takes into account three types of assumptions, which enable us to calculate three total sensitivity indices:

$$S_1^T = \frac{D^2(Y) - D^2_{X_{-k}}\left[E_{X_k}(Y|X_{-k})\right]\left[E_{X_k}(Y|X_{-k})\right]}{D^2(Y)} = S_1 + S_{12} + S_{13} + S_{123}\ , \tag{14}$$

$$S_2^T = S_2 + S_{23} + S_{23} + S_{123}\ , \tag{15}$$

$$S_3^T = S_3 + S_{13} + S_{23} + S_{123}\ , \tag{16}$$

Ultimately, the impact of interactions between input variables (their indirect impact) on the variance of the output variable is calculated as a difference between their

indirect and indirect impacts ($S_k^T - S_k$). Considerable differences between S_k^T and S_k indicate the significant role of interactions between the k-th input variable in shaping the value of the Y output variable, which in turn indicates a high degree of correlation between input variables. The analysis of interactions between input variables helps to understand the model structure. Variance-based methods of sensitivity analysis, both for independent and dependent input variables were described by Saltelli et al. (2008).

Under the deterministic approach we do not use quasi-random sampling to determine distributions of input variables. Instead of applying quasi-random sampling, sensitivity indices are calculated using values of price indices for particular elementary groups, estimated for each combination of sources of price data and index formulas.

4. Results

4.1. Uncertainty analysis

We opted for the deterministic approach because the number of assumptions used in the study was relatively small and it was possible to analyse all possible combinations.

Values of price indices for particular elementary groups, which were calculated for each combination of each combination of sources of price data and index formulas are presented in Table A.4 in the Appendix. For each elementary group, Figure 1 shows:

- the value of the index estimated by means of the method currently used by Statistics Poland (price data are surveyed by enumerators, consumer price indices for subgroups are aggregated using the Jevons index),
- the value of the index estimated on the basis of combined data sources and preferred aggregation formulas (consumer price indices for subgroups were aggregated into the price index for their respective elementary groups using the Törnqvist index, consumer price indices for a given elementary group, calculated on the basis of different data sources were aggregated into a single index using the geometric Young index),
- the minimum and maximum values of each index.

It should be noted that, owing to the limited availability of data, the robustness analysis of consumer price indices includes changes that were observed over one month (between February and March 2021). However, in view of the short period covered by the analysis, obtained results cannot constitute a basis for drawing conclusion of a general nature. For that purpose it is necessary to base the analysis on a longer period.

Given high levels of inflation in Poland both in 2000 and in 2021, differences in annual estimates of inflation based on price indices calculated from different sources of data are likely to be much higher than those calculated on the basis of monthly data

(assuming that the trend of changes observed between February and March 2021 was to continue throughout the year, annual price changes would be 12 times bigger).

The biggest range of consumer price indices (difference between the maximum and minimum estimates of price indices), resulting from various combinations of different data sources and different formulas was obtained for the sugar elementary group – 5.1 percentage points.

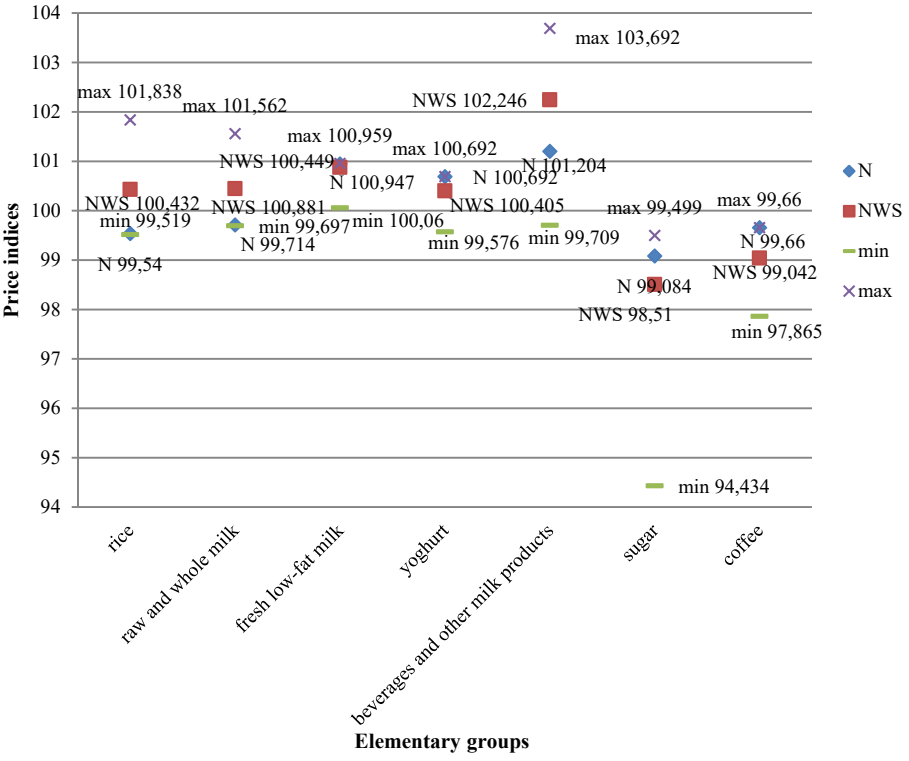

Figure 1. Results of uncertainty analysis for consumer price indices

Source: authors' work based on data in Table A.4.

The smallest range of consumer price indices can be observed for the elementary group of fresh low fat milk – 0.9 percentage point.

4.2. Sensitivity analysis

In the study described above, the analysis of sensitivity was conducted using the deterministic approach in which sensitivity indices were calculated on the basis of values of consumer price indices for particular elementary groups, estimated for each combination of data sources and index formulas used. The analysis involved seven

variants of data sources, five index formulas for aggregating subgroups into indices for elementary groups and two formulas for aggregating indices calculated from different data sources. This resulted in a total of 70 combinations of assumptions.

In the deterministic model, total model variance V is estimated as variance from all calculated index values for all combinations of assumptions. In order to determine the variance component directly associated with a given assumption V_k (formula 10), it is necessary to calculate the mean index value for each possible value of this assumption. The number of mean values depends on the number of assumption variants. The variance of these means constitutes the estimator of V_k.

In order to calculate $E_{X_{-k}}\left(D^2_{X_k}(Y|X_{-k})\right)$ (formula 13) for the k-th assumption, one needs to consider all possible combinations of the remaining assumptions (denoted as k). For each such combination, one calculates the variance of values obtained for the final index. The construction of indices obtained from a given combination will differ only with respect to assumption k. The mean value calculated from all these variances constitutes the estimator of $E_{X_{-k}}\left(D^2_{X_k}(Y|X_{-k})\right)$.

Results of these calculations are presented in Table 1. The choice of a particular data source accounted directly for as much as 85% or more of total variability for all data sources, except for the yoghurt elementary group. The impact of the index formula used for aggregating indices for subgroups into elementary group indices accounted for only 0.09% of variability for fresh whole milk, 3% for beverages and other milk products, 4% for fresh low fat milk, 5.4% for coffee, 6.9% for rice, 9.3% for sugar and 16.5% for yoghurt. For all elementary groups, the impact of the index formula used for aggregating indices derived from various data sources into one composite index was negligible – less than 0.0001% of total variance.

For all elementary groups and all assumptions considered in the study, total impacts of interactions (13) are higher than their direct impacts (10). The bigger the differences between these two values, the more interactions between the assumptions contribute to the total variability of the final indices. The yoghurt elementary group is particularly noteworthy in this respect since the total impacts for the first two assumptions exceed 90%, while their direct impacts are equal to 21.9% and 16.5%, respectively, which implies that in the case of this elementary group these two assumptions strongly interact with each other. In other words, when one of them is replaced, values of the final index are not greatly affected. However, certain combinations of these assumptions can result in relatively disparate values of the index.

Table 1. Sensitivity indices of consumer price indices

Product elementary groups	Values of sensitivity indices in %								
	price data source			index formulas – subgroups			index formulas – elementary groups		
	S_1	S_1^T	$S_1^T - S_1$	S_2	S_2^T	$S_2^T - S_2$	S_3	S_3^T	$S_3^T - S_3$
Rice	88.22	107.09	18.87	6.88	14.51	7.64	0.00004	0.00035	0.00031
Raw and whole milk	99.82	114.89	15.07	0.09	0.22	0.13	0.00001	0.00008	0.00007
Fresh low fat milk	93.44	110.34	16.89	4.06	8.08	4.02	0.00000	0.00000	0.00000
Yoghurt	21.89	96.01	74.12	16.52	96.24	79.73	0.00000	0.00007	0.00007
Beverages and other milk products	96.69	111.51	14.82	3.04	4.08	1.04	0.00006	0.00036	0.00031
Sugar	85.45	104.28	18.83	9.32	17.93	8.61	0.00000	0.00001	0.00000
Coffee	84.42	108.75	24.33	5.44	19.20	13.77	0.00001	0.00006	0.00005

Source: authors' calculations based on data obtained from Statistics Poland and from a retail chain.

5. Conclusions

In the case of four elementary groups (fresh low fat milk, yoghurt, sugar and coffee) inflation indicators obtained by applying the method currently used by Statistics Poland are higher than measures produced by applying the preferred method and lower in the case of the remaining three elementary groups (rice, fresh and whole milk and beverages and other milk products). In comparison with the preferred method, the official method underestimates the rate of inflation to the most (1 p.p.) in the elementary group of beverages and other milk products. The maximum overestimation of the rate of inflation (0.6 p.p.) can be observed in the coffee elementary group.

The results indicate that consumer price indices calculated for the elementary groups of interest are characterised by a relatively low robustness to changes of a data source about consumer prices and by a relatively high robustness to changes of index formulas used for calculating price indices at the level of subgroups and elementary groups.

For all elementary groups of interest, the first assumption, i.e. the choice of a given data source, has the biggest impact on the final value of the price index. Which index formula is used for aggregating indices for subgroups into elementary group indices

has much less influence on the final results. The effect of choosing a particular index for aggregating indices derived from different sources is negligible.

The authors are fully aware that the results are only a preliminary assessment of the impact of new data sources on the measurement of changes in consumer prices. These conclusions cannot be generalised because, for one thing, they are based on data from only two months and, secondly, they only refer to selected elementary groups of food products. Nonetheless, the study is an important starting point for further, more comprehensive studies into the robustness of price indices in the measurement of inflation.

The relatively low stability of price indices as a consequence of adopting different assumptions underlying their construction means that one needs to be very cautious when attempting to include new data sources in the measurement process. The danger associated with such attempts is that changes manifested in price indices, rather than reflect actual price changes, may well be merely the result of including new data sources.

Acknowledgement

The article is the outcome of the project entitled "The construction of an integrated system of price statistics" (INSTATCENY), funded by the National Centre for Research and Development (NCBiR) (1st edition of the GOSPOSTRATEG competition, No. 1/382525/14 / NCBR / 2018). Authors would like to thank the National Centre for Research and Development for financing this publication.

References

Białek J., (2017). Approximation of the Fisher price index by using Lowe, Young and AG Mean indices. *Communications in Statistics – Simulation and Computation*, 46(8), pp. 6454–6467.

Białek, J., Dominiczak-Astin, A. and Turek, D., (2021). Porównanie cen i wskaźników cen konsumpcyjnych: tradycyjna metoda uzyskiwania danych a źródła alternatywne. *Wiadomości Statystyczne. The Polish Statistician*, 66(9), pp. 32–69.

Białek, J., (2021). PriceIndices – a New R Package for Bilateral and Multilateral Price Index Calculations. *Statistika – Statistics and Economy Journal*, Vol. 2/2021, pp. 122–141, Czech Statistical Office, Praga.

Białek, J., Beręsewicz, M., (2021). Scanner data in inflation measurement: From raw data to price indices. *Statistical Journal of the IAOS,* Vol. 37, pp. 1315–1336.

Chan, K., Tarantola, S., Saltelli, A. and Sobol, I. M., (2000). Variance based methods, in: A. Saltelli, K. Chan i M. Scott (eds.). *Sensitivity analysis*, Wiley, New York, pp. 167–197.

Chessa, A., (2015). Towards a generic price index method for scanner data in the Dutch CPI. *In 14th meeting of the Ottawa Group*, Tokyo, pp. 20–22.

Chessa, A., (2016). A new methodology for processing scanner data in the Dutch CPI. *Eurostat review of National Accounts and Macroeconomic Indicators*, Vol. 1, pp. 49–69.

International Labour Office, (2004). *Consumer price index manual: Theory and practice*, Geneva.

De Haan, J., Hendriks, R. and Scholz, M., (2021). *Price measurement using scanner data: Time-product dummy versus time dummy hedonic indexes*. Review of Income and Wealth 67(2), pp. 394–417.

Fisher, I., (1922). *The making of index numbers: a study of their varieties, tests, and reliability*, Number 1, Houghton Mifflin.

Jaro, M., (1989). Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, Vol. 84, pp. 414–420.

Laspeyres, K., (1871). IX. Die berechnung einer mittleren waarenpreissteigerung. *Jahrbücher für Nationalökonomie und Statistik,* 16(1), pp. 296–318.

Nardo, M., Saisana, M., Saltelli, A. and Tarantola, S., (2011). Tools for composite indicators building. *Paperback – European Commission*, Dictus Publishing.

Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., Hoffman, A. and Giovannini, E., (2005). Handbook on constructing composite indicators: Methodology and user guide. OECD, *Statistics Working Paper*.

OECD, (2008). Handbook on constructing composite indicators. Methodology and user guide. *OECD Publications*, Paris.

Panek, T., (2016). *Quality of Life – from conception to measurement*. Warsaw School of Economic Press, Warsaw.

Paasche, H., (1874). Über die preisentwicklung der letzten jahre nach den hamburger börsennotirungen. *Jahrbücher für Nationalökonomie und Statistik*, pp. 168–178

Saisana, M., Saltelli A. and Tarantola S., (2005). Uncertainty and sensitivity techniques as tools for the analysis and validation of composite indicators. *Journal of the Royal Statistical Society*, A., Vol. 168(2), pp. 307–323.

Saltelli, A., (2002). Making best use of model valuations to compute sensitivity indices. *Computer Physics Communications*, Vol. 145, pp. 280–297.

Saltelli A., Chan K. and Scott, E. M. (red.), (2000). *Sensitivity analysis*. John Wiley & Sons, New York.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M. and Tarantola, S., (2008). *Global sensitivity analysis*. The primer. John Wiley & Sons, Chichester.

Sharpe, A., Salzman, J. (2004). *Methodological choices encountered in the construction of composite indices of economic and social well-being. Center for the Study of Living Standards*, Ottawa, CAN.

Sobol, I. M., (1993). Sensitivity analysis for non-linear mathematical models. *Mathematical Modelling and Computational Experiment*, Vol. 1, pp. 407–414.

Törnqvist, L., (1936). The Bank of Finland's consumption price index. *Bank of Finland Monthly Bulletin*, pp. 1–8.

Winkler, W., (1990). String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 354–359.

Appendix

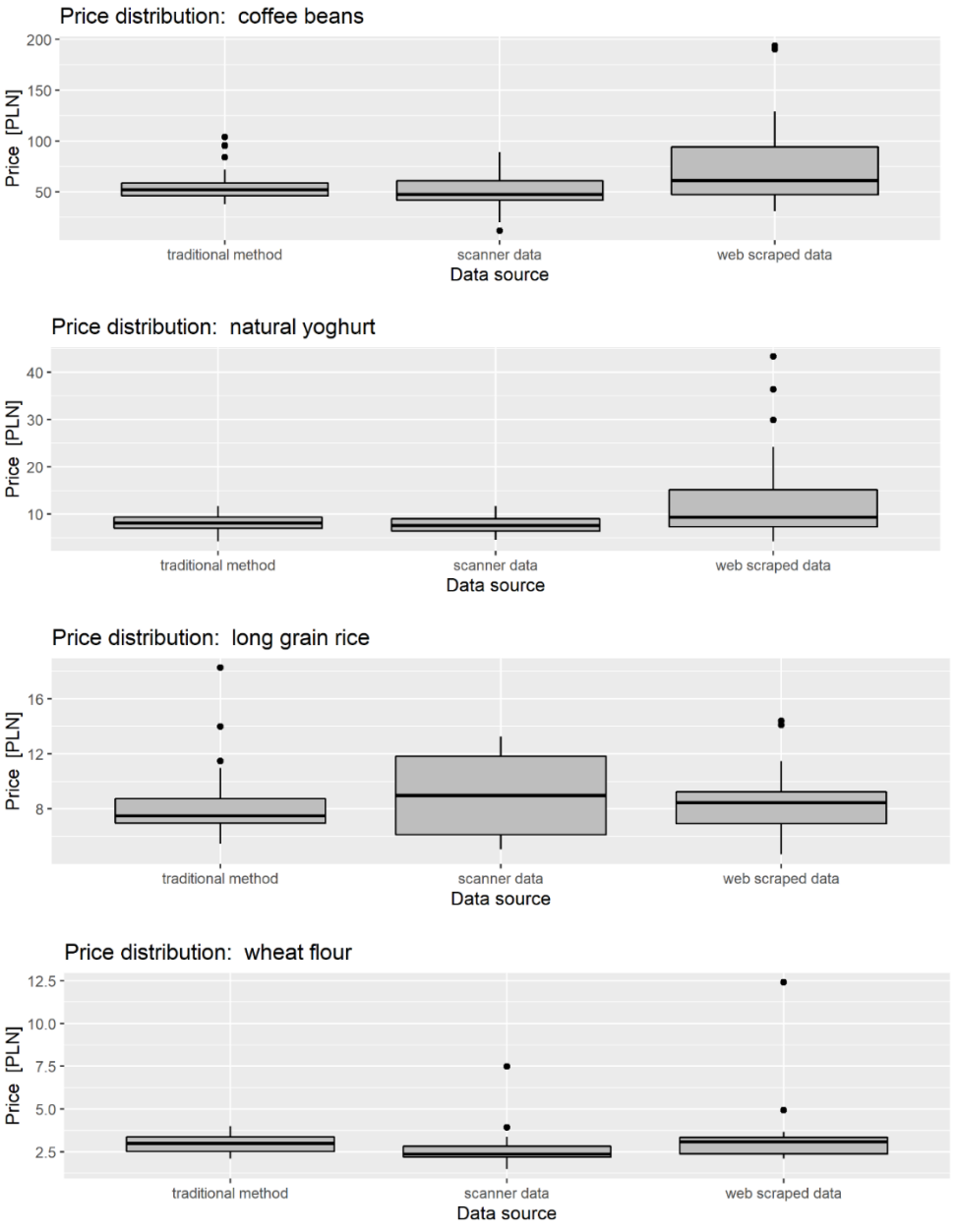

Figure A.1. Box plots for prices of selected categories of food products (based on 3 data sources for March 2021).

Source: authors' calculations based on data obtained from Statistics Poland and from a retail chain.

Table A.1. Weights of subgroups (within their respective elementary groups)

Elementary groups and subgroups	Weights	
	February 2021	March 2021
RICE		
long grain rice	0.654	0.678
white rice	0.346	0.322
RAW AND WHOLE MILK		
UHT whole milk	0.474	0.499
pasteurised whole milk	0.526	0.501
FRESH LOW FAT MILK		
UHT low fat milk	0.420	0.421
goat milk	0.034	0.036
pasteurised low fat milk	0.546	0.544
YOGHURT		
Actimel	0.100	0.097
fruit flavoured yoghurt	0.317	0.307
chocolate and nuts flavoured yoghurt	0.006	0.006
drinkable yogurt	0.282	0.294
natural yoghurt	0.294	0.297
BEVERAGES AND OTHER MILK PRODUCTS		
kefir	0.211	0.224
buttermilk	0.096	0.101
monte	0.229	0.201
homogenised cheese	0.464	0.473
SUGAR		
cane sugar	0.136	0.126
white sugar	0.790	0.778
powdered sugar	0.075	0.096
COFFEE		
instant coffee	0.083	0.073
coffee beans	0.546	0.512
ground coffee	0.371	0.415

Source: authors' calculations based on data obtained from a retail chain.

Table A.2. Elementary group weights depending on the place / mode of purchase

Elementary groups	Weights		
	price survey data (shops excluding chain stores)	scanner data (chain stores)	web scraped data
Rice	34.65	64.36	0.99
Fresh whole milk	34.86	64.73	0.41
Fresh low fat milk	34.85	64.72	0.43
Yoghurt	34.85	64.72	0.43
Beverages and other milk products	34.84	64.71	0.45
Sugar	39.82	59.72	0.46
Coffee	41.28	57.00	1.72

Source: authors' calculations based on data obtained from databases maintained by Passport GMID, *Euromonitor International and domestic market surveys conducted by Statistics Poland.*

Table A.3. Price indices of consumer products by price data source and index formula, February 2021-March 2021

Data source and combinations of index formulas	Indices (February 2021=100)						
	rice	fresh whole milk	fresh low fat milk	yoghurt	beverages and other milk products	sugar	coffee
All data sources							
Jevons x Young	101.023	100.425	100.692	99.966	102.809	99.323	99.473
Jevons x Geom. Young	101.017	100.424	100.692	99.964	102.802	99.323	99.472
Laspeyres x Young	100.498	100.484	100.882	100.430	102.391	98.511	99.048
Laspeyres x Geom. Young	100.495	100.483	100.882	100.430	102.383	98.511	99.046
Paasche x Young	100.372	100.418	100.880	100.381	102.121	98.510	99.039
Paasche x Geom. Young	100.370	100.416	100.880	100.381	102.116	98.509	99.038
Fisher x Young	100.435	100.451	100.881	100.406	102.256	98.511	99.043
Fisher x Geom. Young	100.433	100.449	100.881	100.405	102.250	98.510	99.042
Tornqvist x Young	100.434	100.451	100.881	100.405	102.252	98.510	99.043
Tornqvist x Geom. Young	100.432	100.449	100.881	100.405	102.246	98.510	99.042

Table A.3. Price indices of consumer products by price data source and index formula, February 2021-March 2021 (cont.)

Data source and combinations of index formulas	Indices (February 2021=100)						
	rice	**fresh whole milk**	**fresh low fat milk**	**yoghurt**	**beverages and other milk products**	**sugar**	**coffee**
Price survey data and scanner data							
Jevons x Young	101.034	100.421	100.695	99.966	102.821	99.333	99.501
Jevons x Geom. Young	101.028	100.419	100.695	99.965	102.814	99.333	99.501
Laspeyres x Young	100.504	100.480	100.885	100.432	102.403	98.530	99.061
Laspeyres x Geom. Young	100.502	100.478	100.885	100.432	102.396	98.530	99.060
Paasche x Young	100.378	100.413	100.883	100.383	102.132	98.528	99.052
Paasche x Geom. Young	100.376	100.412	100.883	100.383	102.127	98.528	99.051
Fisher x Young	100.441	100.446	100.884	100.407	102.267	98.529	99.056
Fisher x Geom. Young	100.439	100.445	100.884	100.407	102.261	98.529	99.055
Tornqvist x Young	100.441	100.446	100.884	100.407	102.264	98.529	99.056
Tornqvist x Geom. Young	100.438	100.445	100.884	100.407	102.258	98.529	99.055
Price survey data and web scraped data							
Jevons x Young	99.550	99.735	100.936	100.681	101.190	99.064	99.588
Jevons x geom. Young	99.550	99.735	100.936	100.681	101.190	99.064	99.588
Laspeyres x Young	99.530	99.718	100.948	100.200	100.754	98.715	99.528
Laspeyres x geom. Young	99.530	99.718	100.948	100.200	100.754	98.714	99.527
Paasche x Young	99.527	99.734	100.948	100.171	100.750	98.706	99.528
Paasche x geom. Young	99.527	99.734	100.948	100.171	100.750	98.705	99.528
Fisher x Young	99.529	99.726	100.948	100.185	100.752	98.711	99.528
Fisher x geom. Young	99.529	99.726	100.948	100.185	100.752	98.709	99.527
Tornqvist x Young	99.529	99.726	100.948	100.185	100.752	98.711	99.528
Tornqvist x geom. Young	99.529	99.726	100.948	100.185	100.752	98.709	99.527

Table A.3. Price indices of consumer products by price data source and index formula, February 2021-March 2021 (cont.)

Data source and combinations of index formulas	Indices (February 2021=100)						
	rice	**fresh whole milk**	**fresh low fat milk**	**yoghurt**	**beverages and other milk products**	**sugar**	**coffee**
Scanner and web scraped data							
Jevons x Young	101.809	100.806	100.556	99.577	103.667	99.482	99.341
Jevons x geom. Young	101.808	100.806	100.556	99.577	103.666	99.482	99.340
Laspeyres x Young	101.015	100.906	100.841	100.552	103.258	98.345	98.674
Laspeyres x geom. Young	101.015	100.906	100.841	100.552	103.258	98.344	98.674
Paasche x Young	100.825	100.795	100.838	100.492	102.847	98.347	98.661
Paasche x geom. Young	100.825	100.795	100.838	100.492	102.847	98.346	98.661
Fisher x Young	100.920	100.850	100.839	100.522	103.052	98.346	98.667
Fisher x geom. Young	100.920	100.850	100.839	100.522	103.052	98.345	98.667
Tornqvist x Young	100.919	100.850	100.839	100.522	103.048	98.346	98.667
Tornqvist x geom. Young	100.919	100.850	100.839	100.522	103.047	98.345	98.667
Price survey data							
Jevons x Young	99.540	99.714	100.947	100.692	101.204	99.084	99.660
Jevons x geom. Young	99.540	99.714	100.947	100.692	101.204	99.084	99.660
Laspeyres x Young	99.522	99.697	100.959	100.202	100.768	98.763	99.579
Laspeyres x geom. Young	99.522	99.697	100.959	100.202	100.768	98.763	99.579
Paasche x Young	99.519	99.713	100.958	100.173	100.763	98.756	99.577
Paasche x geom. Young	99.519	99.713	100.958	100.173	100.763	98.756	99.577
Fisher x Young	99.520	99.705	100.958	100.188	100.765	98.760	99.578
Fisher x geom. Young	99.520	99.705	100.958	100.188	100.765	98.760	99.578
Tornqvist x Young	99.520	99.705	100.958	100.188	100.765	98.760	99.578
Tornqvist x geom. Young	99.520	99.705	100.958	100.188	100.765	98.760	99.578

Table A.3. Price indices of consumer products by price data source and index formula, February 2021-March 2021 (cont.)

Data source and combinations of index formulas	Indices (February 2021=100)						
	rice	fresh whole milk	fresh low fat milk	yoghurt	beverages and other milk products	sugar	coffee
Scanner data							
Jevons x Young	101.838	100.801	100.559	99.576	103.692	99.499	99.385
Jevons x geom. Young	101.838	100.801	100.559	99.576	103.692	99.499	99.385
Laspeyres x Young	101.033	100.901	100.846	100.556	103.283	98.374	98.685
Laspeyres x geom. Young	101.033	100.901	100.846	100.556	103.283	98.374	98.685
Paasche x Young	100.840	100.790	100.842	100.496	102.869	98.377	98.671
Paasche x geom. Young	100.840	100.790	100.842	100.496	102.869	98.377	98.671
Fisher x Young	100.937	100.846	100.844	100.526	103.076	98.376	98.678
Fisher x geom. Young	100.937	100.846	100.844	100.526	103.076	98.376	98.678
Tornqvist x Young	100.936	100.846	100.844	100.525	103.071	98.376	98.678
Tornqvist x geom. Young	100.936	100.846	100.844	100.525	103.071	98.376	98.678
Web scraped data							
Jevons x Young	99.885	101.522	100.060	99.757	100.104	97.319	97.865
Jevons x geom. Young	99.885	101.522	100.060	99.757	100.104	97.319	97.865
Laspeyres x Young	99.831	101.562	100.098	99.987	99.715	94.493	98.292
Laspeyres x geom. Young	99.831	101.562	100.098	99.987	99.715	94.493	98.292
Paasche x Young	99.823	101.521	100.098	99.978	99.709	94.434	98.339
Paasche x geom. Young	99.823	101.521	100.098	99.978	99.709	94.434	98.339
Fisher x Young	99.827	101.542	100.098	99.983	99.712	94.464	98.316
Fisher x geom. Young	99.827	101.542	100.098	99.983	99.712	94.464	98.316
Tornqvist x Young	99.827	101.542	100.098	99.983	99.712	94.462	98.316
Tornqvist x geom. Young	99.827	101.542	100.098	99.983	99.712	94.462	98.316

Source: authors' calculations based on data obtained from Statistics Poland and from a retail chain.

Table A.4. Price indices of consumer products for subgroups, February 2021 – March 2021

Elementary groups and subgroups	Indices (February 2021=100)		
	price survey data	scanner data	web scraped data
RICE			
long grain rice	99.48	99.14	99.71
white rice	99.60	104.61	100.06
RAW AND MILK			
UHT whole milk	100.07	99.14	100.8
pasteurised whole milk	99.36	102.49	102.25
FRESH LOW FAT MILK			
UHT low fat milk	100.84	100.57	100
goat milk	100.95	100	100
pasteurised low fat milk	101.05	101.11	100.18
YOGHURT			
Actimel	103.25	106.16	100.15
fruit flavoured yoghurt	99.61	100.49	100.59
chocolate and nuts flavoured yoghurt	-	92.31	98.75
drinkable yogurt	99.04	99.62	100.34
natural yoghurt	100.92	99.79	98.97
BEVERAGES AND OTHER MILK PRODUCTS			
kefir	101.90	101.45	100.01
buttermilk	102.01	102.14	101.31
monte	101.08	110.91	100
homogenised cheese	99.84	100.59	99.11
SUGAR	**99.08**	**99.50**	**97.32**
cane sugar	99.54	100.83	101.17
white sugar	98.63	97.81	93.02
powdered sugar	-	99.88	97.94
COFFEE			
instant coffee	99.77	101.17	96.56
coffee beans	99.55	98.27	98.13
ground coffee	-	98.74	98.92

Source: authors' calculations based on data obtained from Statistics Poland and from a retail chain.