Remarks on Price Index Methods for the CPI Measurement Using Scanner Data

Jacek Białek [1] | *University of Lodz, Lodz, Poland*

Abstract

Scanner data are a quite new data source for statistical agencies and the availability of electronic sales data for the calculation of the Consumer Price Index (CPI) has increased over the past 16 years. Scanner data can be obtained from a wide variety of retailers (supermarkets, home electronics, Internet shops, etc.) and provide information at the level of the barcode, i.e. the Global Trade Item Number (GTIN, formerly known as the EAN code). One of new challenges connected with scanner data is the choice of the index formula which should be able to reduce the chain drift bias and the substitution bias. In this paper, we compare several price index methods for CPI calculations based on scanner data. In particular, we consider bilateral index methods with chained versions of direct weighted and unweighted indices, and also selected multilateral index methods, i.e. the quality adjusted unit value method (QU method) and its special case (the Geary-Khamis method), the augmented Lehr method, the so called "real time index", the GEKS method and the CCDI method. We consider different weighting schemes in quantity weights on the price index. We compare all these methods using a real scanner data set obtained from one supermarket chain.. The main aim of the paper is to show how big differences among bilateral and multilateral indices may rise while using real scanner data sets. In particular our results lead to the conclusion that the choice of the multilateral formula and the weighting scheme does matter in inflation measurement. It is shown that differences between values of all discussed formulas may exceed several percentage points even in the case of only one homogeneous group of products.[2]

Keywords

Scanner data, Consumer Price Index, superlative indices, elementary indices, chain indices, QU-GK index, Geary-Khamis method, real time index, GEKS, bilateral indices, multilateral indices

JEL code

C43, E31

INTRODUCTION

Scanner data mean transaction data that specify turnover and numbers of items sold by GTIN (barcode, formerly known as the EAN code). Scanner data have numerous advantages compared to traditional survey data collection because such data sets are much bigger than traditional ones and they contain complete transaction information, i.e. information about prices and quantities.

[1] Department of Statistical Methods, University of Lodz, 90-255 Lodz, 3/5 POW Street, Poland. E-mail: jacek.bialek@uni.lodz.pl. Central Statistical Office in Poland, Department of Trade and Services, Warsaw, Poland.

[2] This is a modified and improved version of the paper titled "Comparison of Price Index Methods for the CPI Measurement Using Scanner Data" which was presented during the 16th *Meeting of the Ottawa Group on Price Indices,* Rio de Janeiro, 8–10 May 2019.

In other words, scanner data contain expenditure information at the item level (i.e. at the barcode or the GTIN level), which makes it possible to use expenditure shares of items as weights for calculating price indices at the lowest (elementary) level of data aggregation.

Scanner data from two supermarkets were introduced in the Dutch CPI in 2002 and, in January 2010, the number of supermarkets providing the scanner data was extended to six. The Dutch CPI was re-designed (de Haan, 2006; Van der Grient and de Haan, 2010; de Haan and Van der Grient, 2011). In 2017, scanner data of ten supermarkets chains were used and at present surveys are not carried out anymore for supermarkets, i.e. scanner data from other retailers (for instance, from do-it-yourself stores or from travel agencies) are used in the Dutch CPI (Chessa, 2015). Until 2015, four EU countries were using scanner data (the Netherlands, Norway, Sweden, and Switzerland). The number of countries that make use of scanner data in their CPI has been growing, i.e. in April 2016, the number of EU countries increased to seven (Belgium, Denmark and Iceland started to use such data sets) and at present, some of national statistical institutes (NSIs) consider starting to use scanner data. Some other countries consider using scanner data in their CPI calculation in the nearest future (or have just started using it), for instance: the French National Statistical Institute (INSEE) launched in 2010 a pilot project in order to get some insights into the suitability of these data for CPI purposes, the Statistics Portugal was awarded in 2011 a Eurostat grant to undertake the initial research on the exploitation of scanner data, in Luxembourg, collaboration was put in place with several retailers who agreed to transmit every month their data to the IT system (STATEC) and scanner data have been introduced in the regular production from January 2018. In January 2018, in Poland, the project titled "INSTATCENY" began and its main aim is to create the new methodology of CPI measurement based on data from different (traditional and untraditional) sources, including scanner data and web-scraped data. In 2017, the Eurostat provided *Practical Guide for Processing Supermarket Scanner data,* which is commonly available on website: <*https://ec.europa.eu/eurostat/web/hicp/overview*>. In the above-mentioned guide, we can read: "This guide describes the situation in 2017. It will need to be updated as the use of scanner data develops and broadens". In fact, the methodology for CPI (or HICP) construction using scanner data has strongly evolved over the last few years (see for instance: Ivancic et. al., 2011; Krsnich, 2014; Griffioen and Bosch, 2016; de Haan et al., 2016; Chessa and Griffioen, 2016; Chessa, 2017; Diewert and Fox, 2017). One of new challenges connected with scanner data is the choice of the index formula which should be able to reduce the chain drift bias and the substitution bias.

In this paper, we compare several price index methods for CPI calculations based on scanner data. The main aim of the paper is to show how big differences among bilateral and multilateral indices may rise while using real scanner data sets. In particular it is shown that the choice of the multilateral formula and the weighting scheme does matter in inflation measurement. The paper is organised as follows: Section 1 presents main advantages, disadvantages and challenges connected with using scanner data. Section 2 describes a selected bilateral and multilateral index method which can be used in the case of scanner data and this Section also discusses updating and weighting problem connected with multilateral methods; Section 3 proposes some price index modifications; Section 4 presents the results from our simulation study and examines the influence of the price and quantity behaviour on differences between the discussed index methods; Section 5 continues the comparison of bilateral and multilateral methods; it presents the empirical study based on real scanner data sets obtained from one supermarket and the e-commerce platform *allegro.pl*; last section lists the main conclusions.

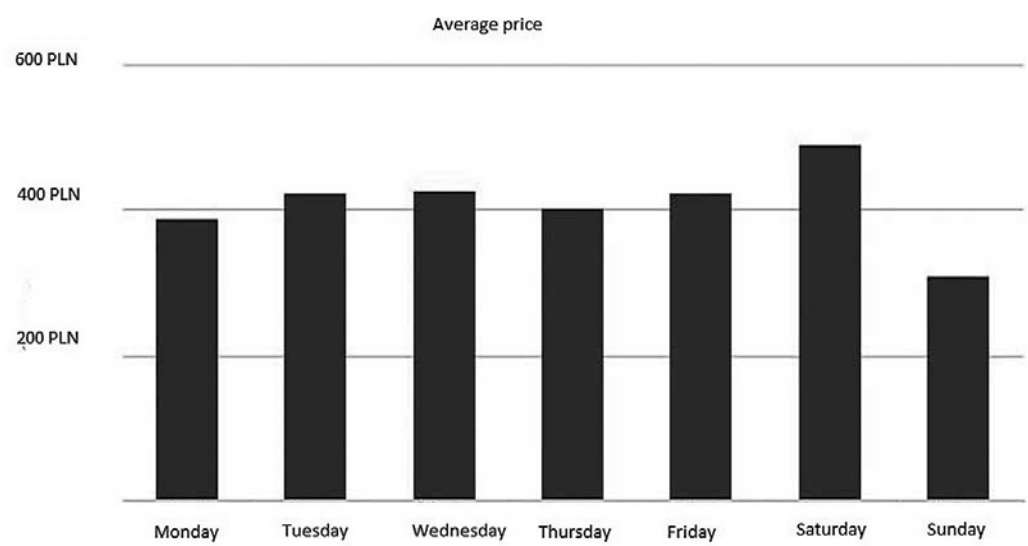
1 SCANNER DATA: ADVANTAGES, DISADVANTAGES AND CHALLENGES

One of the main advantages of using scanner data is the fact that these data sets allow for the level of elementary aggregate to be taken down to lower levels, as the information about prices and quantities (thus also about weights) is available. Scanner data sets are huge and may provide some additional

information about products (such as the following attributes: size, colour, package quantity, etc.). These attributes may be useful in aggregating items into homogeneous groups. Besides, obtaining scanner data is much cheaper than obtaining CPI data in the traditional manner. In the Eurostat's *Practical Guide for Processing Supermarket Scanner data* from 2017, we can read (page 9): "In the traditional price collection, price collectors have to trust intuition and common sense and it may happen that prices are collected as long as the item is available even though it is no longer representative. In scanner data the representativeness is guaranteed".

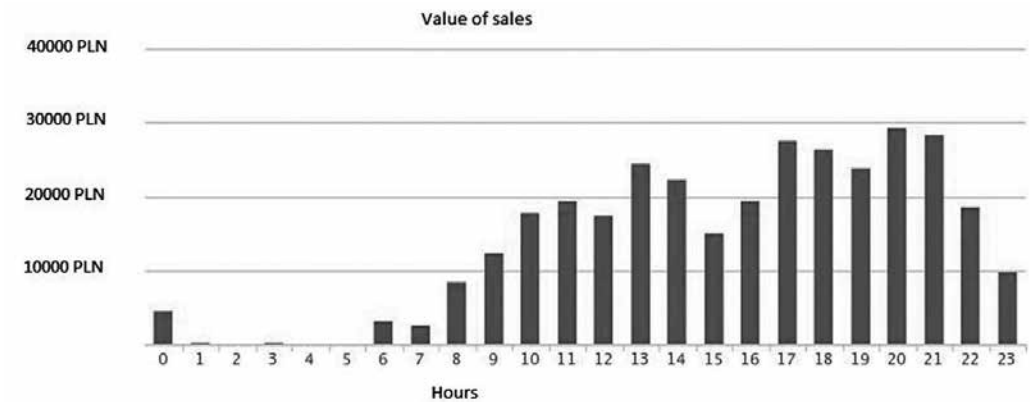
To list disadvantages of using scanner data, we should start with the substantial dependence of the NSI on a retailer. In fact, in the traditional CPI measurement, a price collector must only receive from a given retailer the permission to visit the outlet. In the case of scanner data, there must be a legal contract between the NSI and a given retailer which allows the NSI to fully control and monitor the data. The second disadvantage (or rather a new challenge) is the fact that methodology for sampling of scanner data is still poorly developed, i.e. there are open questions about sampling of regions (if retailers use regional prices) or sampling of outlets (if the outlets differ from one another with respect to opening hours or offered goods). Also the choice of the time interval for aggregating of scanner data may be problematic, since many product prices change periodically but the price cycle may equal to a quarter, a month, a week or even a day. In fact, observing scanner data, it is easy to confirm the known fact that prices are often higher on Saturday (see Figure 1), and each day the value of sales (see Figure 2), the number of transactions (see Figure 3) and the price (see Figure 4) are the highest in the afternoon or in the evening. Moreover, the distribution of the expenditure on a given product may strongly depend on the day, for instance, the expenditure fluctuations on Monday and on Friday may differ from each other significantly (see Figure 5). The same remark could be repeated for prices of products (see Figure 6). All figures which confirm our above-mentioned remarks, i.e. Figures 1–6, are obtained for the homogeneous group of child safety seats (sample EANs: 5902581655226, 5902533903429, or 5902581652850) and the analysed scanner data set comes from the net portal *allegro.pl* (the *TradeWatch* tool) and concerns the time interval: 22.11.2015–16.12.2018.

Figure 1 Daily average price of child safety seats



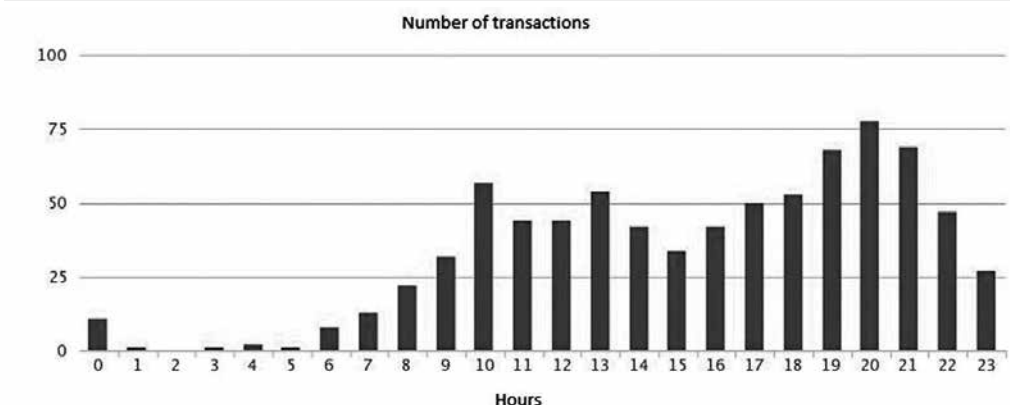
Source: <*TradeWatch.pl*>

Figure 2 Average value of sales of child safety seats depending on hours



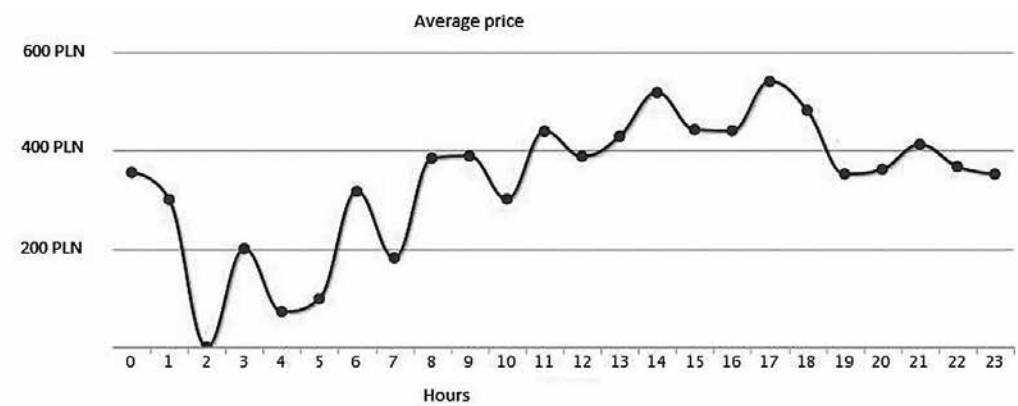
Source: <*TradeWatch.pl*>

Figure 3 Average number of transactions concerning child safety seats depending on hours



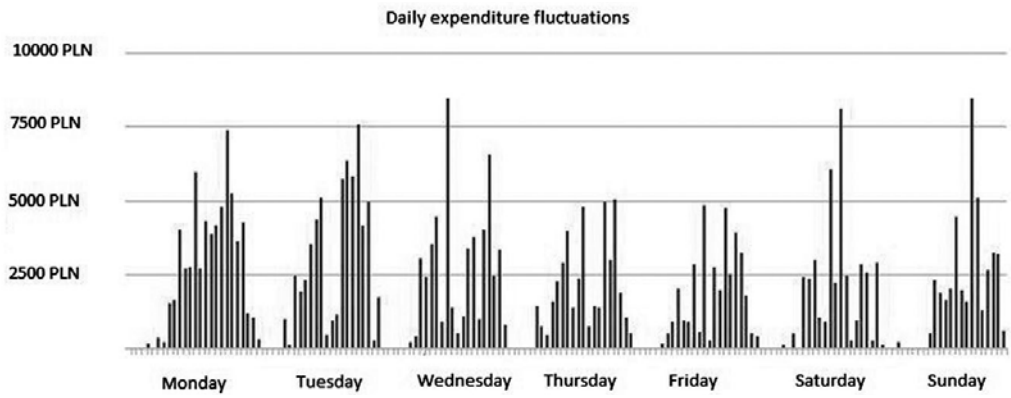
Source: <*TradeWatch.pl*>

Figure 4 Average price of child safety seats depending on hours



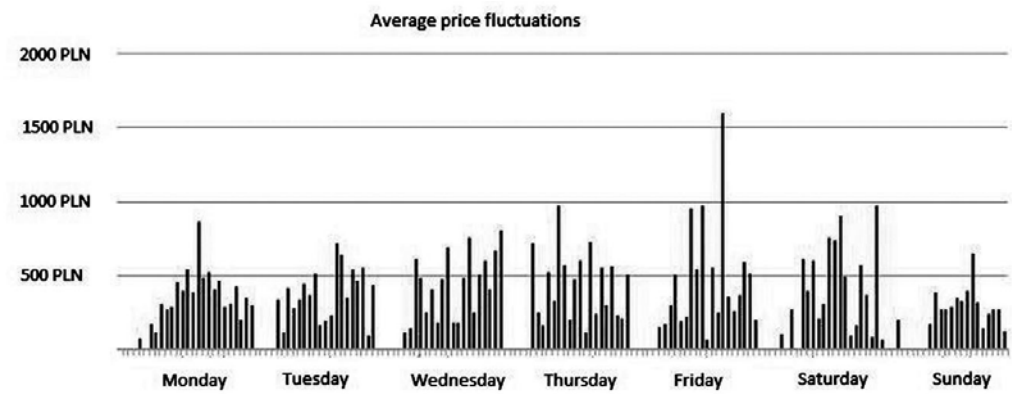
Source: <*TradeWatch.pl*>

Figure 5 Fluctuations of the average expenditure on child safety seats depending on days



Source: <*TradeWatch.pl*>

Figure 6 Fluctuations of the average price of child safety seats depending on days



Source: <*TradeWatch.pl*>

The first challenge connected with scanner data concerns item codes. As it was mentioned above, the GTIN (formerly known as the EAN) is the name currently used for coding in the case of scanner data. Nevertheless, the following codes may also be used: price look-up (PLU) and stock-keeping units (SKUs). PLU codes are shorter than GTINs and SKUs can be slightly more generic than GTINs. In practice, when NSIs use scanned data from different retailers with different code systems, some problems with identifying products may rise. Moreover, scanner data may contain data on transactions between the retailer and other business, which should be verified and detected (such transactions should be excluded from CPI calculations). The next challenge is detecting items which were returned within the given period after the purchase. Since typically, 10 000–25 000 item codes are used in the supermarket, a huge challenge is to create the appropriate, preferably automatic (or at least almost fully automatic) IT system which is able to go through with the above-mentioned detections and which takes into consideration seasonal goods, replacements, as well as disappearing and appearing item codes in the sample. Finally, one of new challenges connected with scanner data is the choice of the index formula which should be able to reduce the chain drift bias and the substitution bias. In the paper, we focus on the choice of the price index formula.

2 INDEX METHODS FOR CPI CALCULATIONS USING SCANNER DATA

Most statistical agencies use bilateral index numbers in the CPI measurement, i.e. they use indices which compare prices and quantities of a group of commodities from the current period with the corresponding prices and quantities from a base (fixed) period. In multilateral methods, we collect information about prices and quantities of a group of commodities from T periods and next we calculate a sequence of price indices for these T periods. Although Ivancic, Diewert and Fox (2011) have suggested that the use of multilateral indices in the scanner data case can solve the chain drift problem, most statistical agencies using scanner data still make use of the monthly chained Jevons index (Chessa et al., 2017). Since the elementary Jevons price index belongs to bilateral (direct) index methods, we start our description of possible methods with these methods. Following Chessa et al. (2017), let us denote the sets of homogeneous products belonging to the same product group in months 0 and t by G_0 and G_t respectively, and let $G_{0,t}$ denote the set of matched products in both moments 0 and t. A product may refer to a single item (GTIN) or to a sub-group of items (GTINs) having the same characteristics, and thus being in the same homogeneity group. In the next part of the paper, we consider the second scenario, i.e. a homogeneous group of different GTINs but having identical characteristics. We also consider a month as a time period over which scanner data are aggregated. In fact, one month is the longest interval among time intervals recommended by Eurostat for the scanner data aggregation (see *Practical Guide for Processing Supermarket Scanner data,* 2017, page 13) although, the same document on the same page states: "Most commonly, scanner data are collected weekly, i.e. all transactions taking place during a week are aggregated".

2.1 Bilateral index methods

2.1.1 Unweighted formulas

A recommendation of the European Commission concerning the choice of the elementary formula at the lowest level of data aggregation can be found on website: <*http://www.ilo.org/public/english/bureau/stat/download/cpi/corrections/annex1.pdf*> and it is as follows: "For the HICPs the ratio of geometric mean prices or the ratio of arithmetic mean prices are the two formulae which should be used within elementary aggregates. The arithmetic mean of price relatives may only be applied in exceptional cases and where it can be shown that it is comparable". In other words, if expenditure information is not available, the European Commission recommends the Jevons (1865) price index (see also Diewert, 2012; or Levell, 2015), which can be written as follows:

$$P_J^{0,t} = \prod_{i \in G_{0,t}} \left(\frac{p_i^t}{p_i^0}\right)^{\frac{1}{N_{0,t}}}, \tag{1}$$

where p_i^τ denotes the price of the i-th product at the time $\tau \in \{0,t\}$ and $N_{0,t} = card\ G_{0,t}$. On the other hand, the same recommendation takes also into consideration ("in exceptional cases") the Carli (1804) price index, which can be written as follows:

$$P_C^{0,t} = \frac{1}{N_{0,t}} \sum_{i \in G_{0,t}} \frac{p_i^t}{p_i^0}, \tag{2}$$

In our research, we consider only the first Formula (1) together with its monthly chained version which is denoted here by $P_{CH-J}^{0,t}$.

2.1.2 Weighted formulas

Since scanner data contain information about the expenditure, it is possible in their case to calculate weighted bilateral indices. *Superlative* price indices, firstly proposed by Diewert (1976), are the most

recommended index formulas for the scanner data case (as base formulas). Following Chessa et al. (2017), we consider the Törnqvist (1936) price index, which is given by:

$$P_T^{0,t} = \prod_{i \in G_{0,t}} \left(\frac{p_i^t}{p_i^0}\right)^{\frac{s_i^0 + s_i^t}{2}}, \tag{3}$$

where s_i^0 and s_i^t denote the expenditure shares of matched products in months 0 and t.

Other commonly known superlative price indices are the Fisher price index (1922) and the Walsh price index (1901). Their formulas, denoted by $P_F^{0,t}$ and by $P_W^{0,t}$ respectively, can be written as follows:

$$P_F^{0,t} = \sqrt{P_{La}^{0,t} \cdot P_{Pa}^{0,t}}, \tag{4}$$

$$I_W^{0,t} = \frac{\sum_{i \in G_{0,t}} p_i^t \cdot \sqrt{q_i^0 q_i^t}}{\sum_{i \in G_{0,t}} p_i^0 \cdot \sqrt{q_i^0 q_i^t}}, \tag{5}$$

where q_i^0 and q_i^t denote quantities of matched products in months 0 and t, $P_{La}^{0,t}$ and $P_{Pa}^{0,t}$ denote the Laspeyres price index (1864) and the Paasche price index (1874) respectively (see Section 2.2.1). In the next part of the paper only the Fisher and the Törnqvist price indices are taken into consideration.

2.2 Multilateral index methods

Multilateral index methods have their genesis in comparisons of price levels across countries or regions. These methods satisfy the transitivity, which is a desirable property for spatial comparisons due to the fact that the results are independent of the choice of base country (region). Commonly known methods are the GEKS method (also known as the EKS method – see Gini (1931), Eltetö and Köves (1964), Szulc (1964), the Geary-Khamis (GK) method (Geary, 1958; Khamis, 1972), the CCDI method (Caves, Christensen and Diewert, 1982; Inklaar and Diewert, 2016) or the real time index method (Chessa, 2015)). In this paper, we consider most of these methods but the problem of the best choice of the multilateral formula seems to be still open.

2.2.1 The quality adjusted unit value index and the Geary-Khamis (GK) method

The term "Quality adjusted unit value method" (shortened to the "QU method") was introduced by Chessa (see, for instance, Chessa, 2015, 2016). The QU method is a family of unit value based index methods with the above-mentioned Geary-Khamis (GK) method as a special case. According to the QU method, the price index $P_{QU}^{0,t}$ which compares the period t with the base period 0 is defined as follows:

$$P_{QU}^{0,t} = \frac{\sum_{i \in G_t} p_i^t q_i^t \Big/ \sum_{i \in G_0} p_i^0 q_i^0}{\sum_{i \in G_t} v_i q_i^t \Big/ \sum_{i \in G_0} v_i q_i^0}, \tag{6}$$

where the numerator in (6) is the measure of the turnover (expenditure) change between the two considered months and the denominator in (6) is a weighted quantity index. Note that both the turnover index and

the weighted quantity index are transitive, and thus the price index $P_{QU}^{0,t}$ is also transitive (Chessa et al., 2017). Note also that the quantity weights v_i are the only unknown factors in Formula (6) and these factors convert sold quantities q_i^0 and q_i^t into "common units" $v_i q_i^0$ and $v_i q_i^t$. Prices of products, p_i^0 and p_i^t, are converted into "quality adjusted prices" p_i^0 / v_i and p_i^t / v_i. If the considered consumption segment is homogeneous, then product quantities can be summed (factors v_i are equal for all products) and the index $P_{QU}^{0,t}$ simplifies to the unit value index (the nominator of (6)). If the above-mentioned consumption segment is not homogeneous, then the unit value index must be adjusted. Note also that the formula $P_{QU}^{0,t}$ defines a family of price indices. In fact, limiting considerations to products sold in both moments 0 and t, and setting v_i equal to the product prices in the current period t, the Formula (6) leads to the Laspeyres index:

$$P_{La}^{0,t} = \frac{\sum_{i \in G_{0,t}} p_i^t q_i^0}{\sum_{i \in G_{0,t}} p_i^0 q_i^0}. \tag{7}$$

Similarly, if we consider the group of products $G_{0,t}$ and if the quantity weights v_i are set equal to the prices in the base period (month) 0, then the formula $P_{QU}^{0,t}$ simplifies to the Paasche price index, i.e.

$$P_{Pa}^{0,t} = \frac{\sum_{i \in G_{0,t}} p_i^t q_i^t}{\sum_{i \in G_{0,t}} p_i^0 q_i^t}. \tag{8}$$

In other words, different choices of factors v_i lead to different prices index formulas. In the GK method, the weights v_i are defined as follows:

$$v_i = \sum_{z=0}^{T} \varphi_{i,GK}^z \frac{p_i^z}{P_{QU}^{0,z}}, \tag{9}$$

where:

$$\varphi_{i,GK}^z = \frac{q_i^z}{\sum_{\tau=0}^{T} q_i^\tau}, \tag{10}$$

and [0, T] is the entire time interval of the product observations (typically T = 12, see Diewert and Fox, 2017). Please note that Formulas (6), (9) and (10) lead to a set of equations which should be solved simultaneously. The above-mentioned solution can be found iteratively (Maddison and Rao, 1996; Chessa, 2016) or as the solution to an eigenvalue problem (Diewert, 1999). An interesting alternative method for obtaining this solution can be also found in Diewert and Fox (2017).

2.2.2 The augmented Lehr index

The Lehr method is similar to the Geary-Khamis method (see Section 2.2.1, Formula (6) with weights defined in (9)) but it does not use the complex iterative method. The quality adjusted factors v_i are defined here as follows:

$$v_i = \frac{p_i^0 q_i^0 + p_i^T q_i^T}{q_i^0 + q_i^T}. \tag{11}$$

The immediate conclusion from (11) is that the Lehr index uses only data from months 0 and T, and in fact this is a bilateral index. Nevertheless, we can change the formula of the quality adjustment factors, and thus, similarly to multilateral methods, we take into considerations all available information from the interval, i.e. (see Loon and Roels, 2018):

$$v_i = \frac{\sum_{\tau=0}^{T} p_i^\tau q_i^\tau}{\sum_{\tau=0}^{T} q_i^\tau}. \tag{12}$$

In the next part of the paper, the augmented Lehr index, i.e. the index constructed as in (6) with quantity weights defined in (12), will be denoted by $P_{AL}^{0,t}$ and the above-mentioned factors will be signified by v_i^{AL}. In other words, the considered augmented Lehr index can be written as follows:

$$P_{AL}^{0,t} = \frac{\sum_{i\in G_t} p_i^t q_i^t \,/ \sum_{i\in G_0} p_i^0 q_i^0}{\sum_{i\in G_t} v_i^{AL} q_i^t \,/ \sum_{i\in G_0} v_i^{AL} q_i^0} \tag{13}$$

2.2.3 The real time index

Let us note that price imputations are not needed when prices from each month of the current year are included in weights v_i. Taking typically value $T = 12$, Chessa (2015) suggests defining these weights by including product prices and quantities from each month of the current year and the base month December of the previous year (there are 13 months together). However, as the same author admits, in practice, we can use prices and quantities of all 13 months only in the final month of the year, and thus some updating method is needed for v_i calculations each month. Although there are several methods for updating quantity weights (see for instance Krsinich, 2014), we focus on an interesting and quite easy for implementation method proposed by Chessa (2015). He suggests the following procedure of calculating *the real time index*: (1) For the current year, we use a time window with December of the previous year as the fixed base month and the window is enlarged each month with the current month; (2) The price index of the current month t is calculated by using the updated quantity weights according to a special algorithm. In particular, this algorithm needs some initial values of price indices $P_{QU}^{0,\tau} : 0 \le \tau \le t$ and it repeats updating weights $v_i = \sum_{z=0}^{t} \varphi_{i,GK}^z \frac{p_i^z}{P_{QU}^{0,z}}$ and next updating values of price indices $P_{QU}^{0,\tau} : 0 \le \tau \le t$ (according to (6)) until the difference between indices from the last two iterations is small enough. Chessa (2015) recommends a method for calculating initial indices. Moreover, he sets the stop criterion at 0.001 and assumes the maximum absolute difference between the price index vectors as a distance measure. Nevertheless, in our study, we set the stop criterion at 0.0001 and we use the Euclidean distance for comparisons of two successive iterations. Steps (1) and (2) are repeated until December of the current year and after that the base month is shifted to December of the current year. In this way, the whole procedure may be repeated in the subsequent year. For more details, see also Chessa (2016).

2.2.4 The GEKS method

Let us consider a time interval $[0,T]$ of observations of prices and quantities which will be used for the GEKS index construction. The GEKS price index between months 0 and t is an unweighted geometric mean of $T+1$ ratios of bilateral price indices $P^{\tau,t}$ and $P^{\tau,0}$ which are based on the same price index formula. The bilateral price index formula should satisfy the time reversal test, i.e. it should satisfy the condition $P^{a,b} \cdot P^{b,a} = 1$. Typically, the GEKS method uses the superlative Fisher price index and in such case the GEKS formula can be written as follows:

$$P_{GEKS}^{0,t} = \prod_{\tau=0}^{T} \left(\frac{P_F^{\tau,t}}{P_F^{\tau,0}}\right)^{\frac{1}{T+1}}. \tag{14}$$

The GEKS formula based on the Jevons price index is also considered in this paper, i.e.

$$P_{JGEKS}^{0,t} = \prod_{\tau=0}^{T} \left(\frac{P_J^{\tau,t}}{P_J^{\tau,0}}\right)^{\frac{1}{T+1}}. \tag{15}$$

2.2.5 The CCDI method

The GEKS method for making international index number comparisons between countries comes from Gini (1931) but it should be mentioned that it was derived in a different manner by Eltetö and Köves (1964) and Szulc (1964). Feenstra, Ma and Rao (2009), and also De Haan and var der Grient (2011) suggested that the Törnqvist price index formula (see (3)) could be used instead of the Fisher price index in the Gini methodology. Caves, Christensen and Diewert (1982) used the GEKS idea with the Törnqvist index as a base in the context of making quantity comparisons across production units (the CCD method) and Inklaar and Diewert (2016) extended the CCD methodology to making price comparisons across production units. Thus, in the paper of Diewert and Fox (2017), the multilateral price comparison method that uses the GEKS method based on the Törnqvist price index is called the CCDI method. The corresponding CCDI price index can be expressed as follows:

$$P_{CCDI}^{0,t} = \prod_{\tau=0}^{T} \left(\frac{P_T^{\tau,t}}{P_T^{\tau,0}}\right)^{\frac{1}{T+1}}. \tag{16}$$

2.2.6 Other methods

In the literature, we can find some other multilateral index methods which are not considered in this paper. The Country-Product Dummy (CPD) method proposed by Summers (1973) has been adapted for spatial price comparisons to the time domain and now it is known as the Time Product Dummy (TPD) method (de Haan and Krsinich, 2014). The multilateral hedonic method is closely related to the TPD method, i.e. its model parameters (known as "item fixed effects") are not estimated for items (as in the TPD method) but they are estimated for the characteristics of items (attributes). Both the TPD method and the above-mentioned hedonic method do not simplify to a unit value index when all products are homogeneous and they are flawed with regard to their use of turnover in constructing weights (Chessa, 2015). Some other methods can be encountered in the paper of Haan et al. (2016), for instance, the so-called "Cycle Method" (see also Willenborg, 2010, 2017; Willenborg and Van der Loo, 2016).

2.3 Alternative weighting schemes in the QU method

In the classical form, the GK method uses quantity shares as weight in the construction of v_i. In the litera-ture, we can find at least two other weighting schemes in quantity weights for the GK price index. The first variant was proposed by Hill (2000) and it assumes that deflated prices, i.e. $p_i^z / P_{QU}^{0,z}$, are weighted by the ratio of the turnover share of the *i*-th product in the month *z* (denoted here by s_i^z) and the sum of turnover shares of the same product over different months. In the paper of Chessa (2016), this variant is referred to as the "QU-TS" method but we use here the shortened notation "TS", i.e. we denote the above-mentioned weights for deflated prices as $\varphi_{i,TS}^z$. In other words, in the TS method, weights $\varphi_{i,GK}^z$ are replaced by weights calculated as follows:

$$\varphi_{i,TS}^z = \frac{s_i^z}{\sum_{\tau=0}^{T} s_i^\tau}, \tag{17}$$

and the final quantity weights are computed as follows:

$$v_i = \sum_{z=0}^{T} \varphi_{i,TS}^z \frac{p_i^z}{P_{QU}^{0,z}}. \tag{18}$$

The other weighting scheme assumes that deflated prices in months with sales receive equal weight, and thus it is denoted here by the EW method (in Chessa (2016), this method is referred to as the "QU-EW" method). In other words, in the considered weighting scheme, we use the following weights for deflated prices:

$$\varphi_{i,EW}^z = \frac{\delta_i^z}{\sum_{\tau=0}^{T} \delta_i^\tau}, \tag{19}$$

where $\delta_i^z = 1$ if $q_i^z > 0$ and $\delta_i^z = 0$ otherwise. Analogically to (17), in the EW method, the final quantity weights can be written as:

$$v_i = \sum_{z=0}^{T} \varphi_{i,EW}^z \frac{p_i^z}{P_{QU}^{0,z}}. \tag{20}$$

In the next part of the paper, we will use different notations for quantity weights defined in (9), (18) and (20), i.e. these weights, connected with the GK, TS and EW methods, will be signified by v_i^{GK}, v_i^{TS} and v_i^{EW} respectively. Similarly, the corresponding multilateral indices, which compare the time moment *t* with the time moment 0, will be denoted by $P_{GK}^{0,t}$, $P_{TS}^{0,t}$ and $P_{EW}^{0,t}$ respectively. In the paper we also suggest considering a different system of weights based on observed and available expenditures, namely:

$$\varphi_{i,EX}^z = \frac{p_i^z q_i^z}{\sum_{\tau=0}^{T} p_i^\tau q_i^\tau}, \tag{21}$$

which allows us to calculate the final quantity weights in the QU method as follows:

$$v_i = \sum_{z=0}^{T} \varphi_{i,EX}^z \frac{p_i^z}{P_{QU}^{0,z}}. \tag{22}$$

We will denote these quantity weights by v_i^{EX} and the corresponding QU index, i.e. the index defined in (6) but using weights v_i^{EX} instead of weights v_i, by $P_{EX}^{0,t}$.

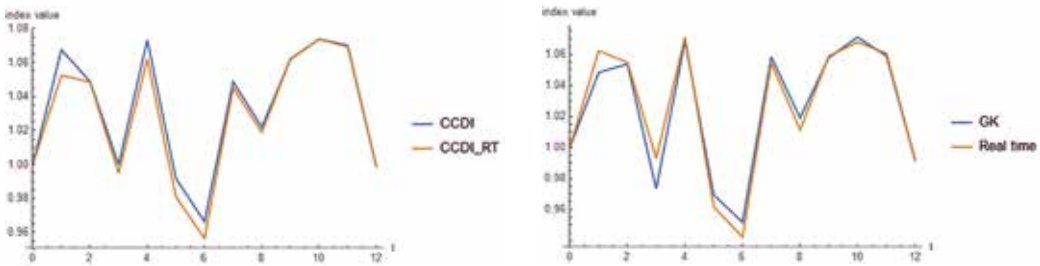
3 EMPIRICAL STUDY

Poland is at the beginning of the way to the regular and official use of scanner data in the CPI measurement. Statistics Poland has started to cooperate with three supermarkets but they do not provide scanner data in a regular way. Moreover, there is no IT system for combining and analysing different data sources from different retailers (supermarkets) written in different file formats. Nevertheless, some experiments on real scanner data sets are being done by using the R package and Mathematica software. In the following empirical study, we consider data sets come from one supermarket chain and they concern the following group of products: plain flour (COICOP group: 011121), milk 3.2% (COICOP group: 011411) and rice (COICOP group: 011111). In this case, we have a 13-month time series (Dec. 2014–Dec. 2015).

Figure 7 presents a comparison of two selected multilateral indices calculated over the whole period of 13 months (i.e. the **CCDI** and **GK** indices when a full window is available) with the corresponding indices calculated over the "currently" available window (i.e. for the current time moment t, the available time window is $[0,t]$ – see the **CCDI_RT** and the **real time** indices). Figure 8 presents a comparison of the **GEKS** index with the **CCDI** and **JGEKS** indices calculated over the whole period of 13 months. Figure 9 presents all considered multilateral indices together with the **chained Jevons** index calculated for the fully available time window. All calculations were done in the Mathematica 11 software.

Figure 7 Comparison of selected multilateral indices (CCDI, GK) for fully and "currently" available time windows (calculated for plain flour, milk and rice)

a) plain flour



b) milk

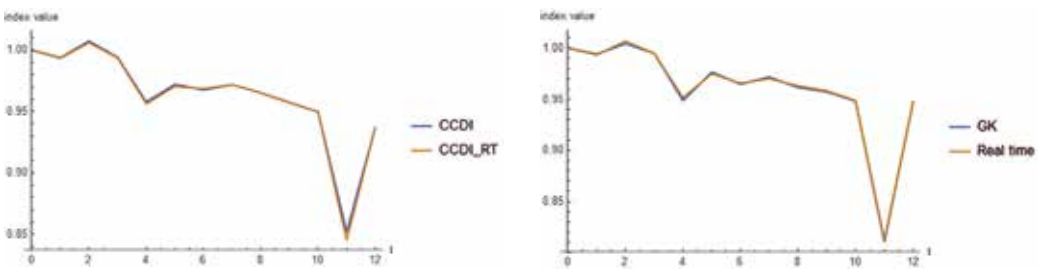
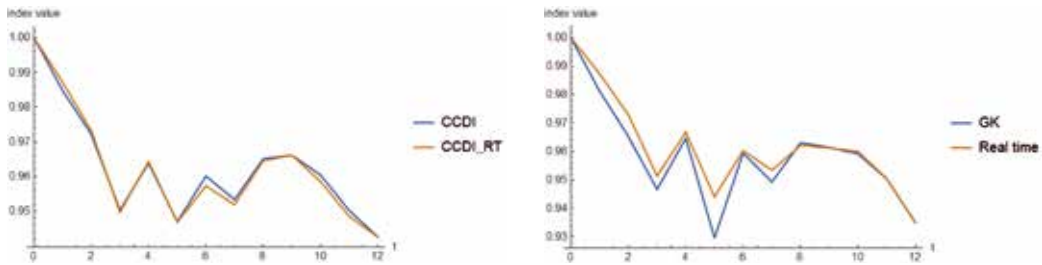


Figure 7 (continuation)

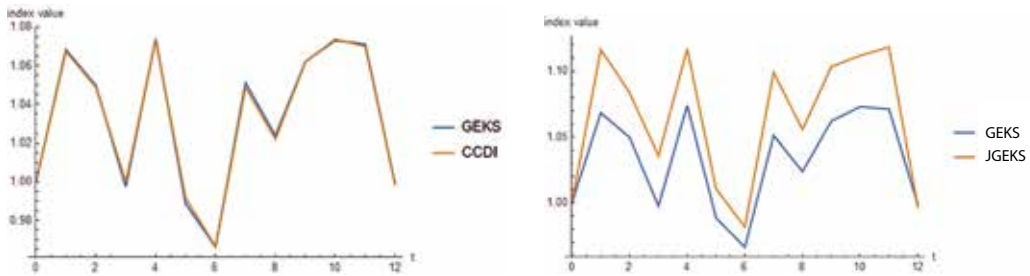
c) rice



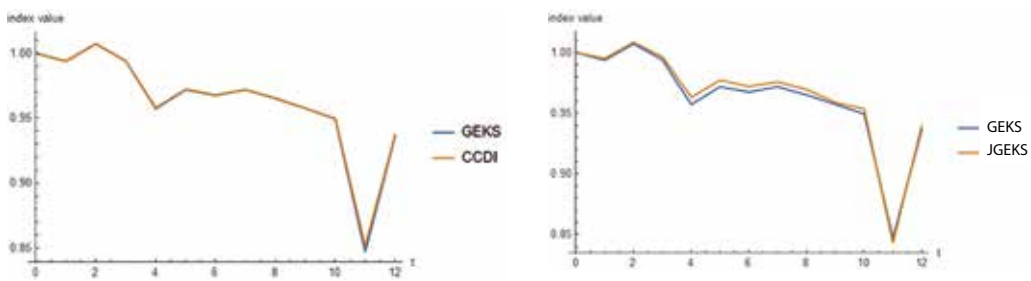
Source: Own calculations based on scanner data from one retailer chain

Figure 8 Comparison of the GEKS index with the CCDI and JGEKS indices calculated over the whole period of 13 months for plain flour, milk and rice

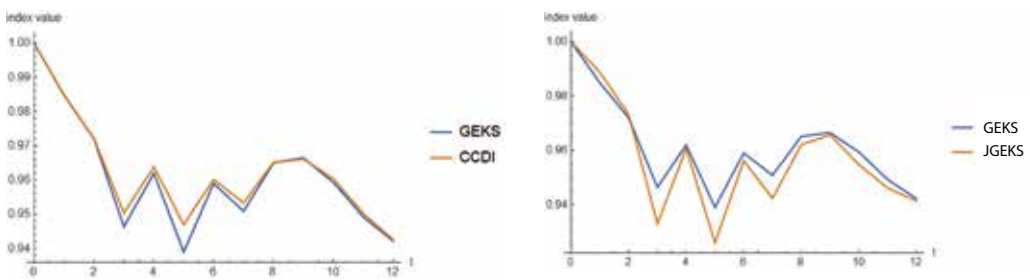
a) plain flour



b) milk



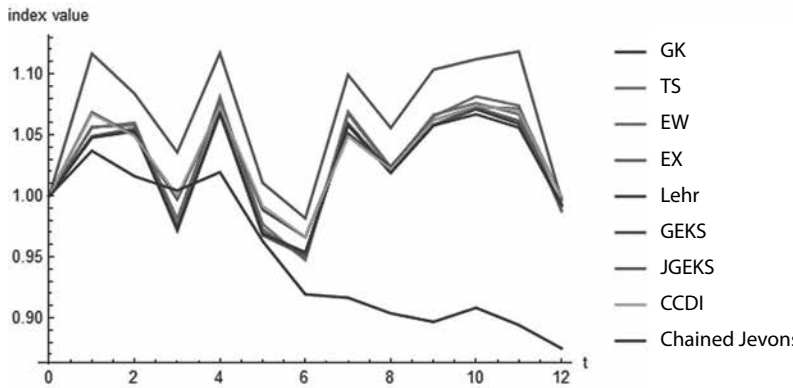
c) rice



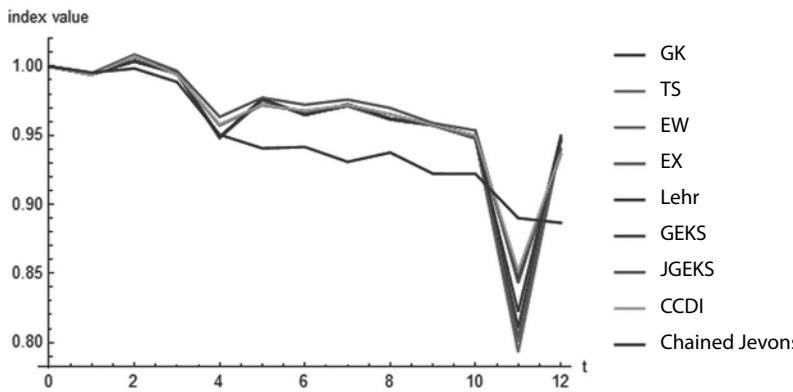
Source: Own calculations based on scanner data from one retailer chain

Figure 9 All considered multilateral indices together with the chained Jevons index calculated over the whole period of 13 months for plain flour, milk and rice

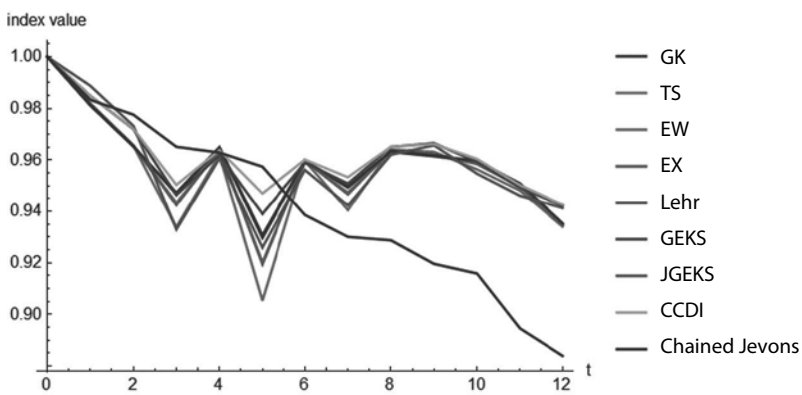
a) plain flour



b) milk



c) rice



Note: For coloured figure see the online version of Statistika journal No. 1/2020.
Source: Own calculations based on scanner data from one retailer chain

CONCLUSIONS

Our empirical study provides the following conclusions: (a) when we have no historical data from supermarkets and we start using scanner data sets, then the application of multilateral indices for the "currently" available time window (from the beginning of cooperation with supermarkets till the current month) is justified since differences between selected indices (CCDI, GK) for the fully and "currently" available time window are not too big, i.e. these differences are decreasing functions of time and, as a rule, after 6–8 months they are negligible (see Figure 7); (b) In practice, there are no substantial differences between the GEKS and CCDI indices and it is not surprising since superlative indices (Fisher, Törnqvist) approximate each other (Diewert, 1976). Nevertheless, the differences between the GEKS and JGEKS indices are crucial and, in our opinion, it confirms that the movements of quantities may not be (rationally) correlated with price movements (see Figure 8); (c) Differences between multilateral indices and the chained Jevons index may be very big (see Figure 9 for plain flour or rice), and as a rule they are. Thus, switching the chained Jevons index to one of multilateral indices does matter in the CPI measurement based on scanner data sets; (d) The choice of the weighting schemes in the QU method does matter – differences in results may be crucial (in our study time moments for which the differences between the TS, EW and EX indices exceeded 3 percentage points were observed).

ACKNOWLEDGEMENTS

This publication is financed by the National Science Centre in Poland (Grant No. 2017/25/B/HS4/00387).

References

ABS. *Making Greater Use of Transactions Data to Compile the Consumer Price Index*. Information Paper 6401.0.60.003, Australian Bureau of Statistics, Canberra, 29 November 2016.

BALK, M. B. Price Indexes for Elementary Aggregates: The Sampling Approach. *Journal of Official Statistics*, 1995, 21(4), pp. 675–699.

CARLI, G. Del valore e della proporzione de'metalli monetati. *Scrittori Classici Italiani di Economia Politica*, 1804, 13, pp. 297–336.

CAVES, D. W., CHRISTENSEN, L. R., DIEWERT, W. E. Multilateral comparisons of output, input, and productivity using superlative index numbers. *Economic Journal*, 1982, 92, pp. 73–86.

CHESSA, A. G. *Towards a generic price index method for scanner data in the Dutch CPI*. Room document for Ottawa Group Meeting, Urayasu City, Japan, 20–22 May 2015.

CHESSA, A. G. A New Methodology for Processing Scanner Data in the Dutch CPI. *Eurona*, 2016, 1, pp. 49–69.

CHESSA, A. G. AND GRIFFIOEN, R. *Comparing Scanner Data and Web Scraped Data for Consumer Price Indices*. Report, Statistics Netherlands, 2016.

CHESSA, A. G. *Comparisons of QU-GK Indices for Different Lengths of the Time Window and Updating Methods*. Paper prepared for the second meeting on multilateral methods organised by Eurostat, Luxembourg, Statistics Netherlands, 14–15 March 2017.

CHESSA, A. G., VERBURG, J., WILLENBORG, L. *A comparison of price index methods for scanner data*. Paper presented at the 15th Meeting of the Ottawa Group on Price Indices, Eltville am Rhein, Germany, 10–12 May 2017.

Consumer Price Index Manual. Theory and practice. IMF, OECD, UNECE, Eurostat, The World Bank, International Labour Office (ILO), Geneva, 2004.

DIEWERT, W. E. Exact and superlative index numbers. *Journal of Econometrics*, 1976, 4, pp. 114–145.

DIEWERT, W. E. *Consumer price statistics in the UK*. Office for National Statistics, Newport, 2012.

DIEWERT, W. E. AND FOX, K. J. *Substitution Bias in Multilateral Methods for CPI Construction using Scanner Data*. Discussion paper 17–02, Vancouver, Canada: Vancouver School of Economics, The University of British Columbia, 2017.

DE HAAN, J. The re-design of the Dutch CPI. *Statistical Journal of the United Nations Economic Commission for Europe*, 2006, 23, pp. 101–118.

DE HAAN, J. AND VAN DER GRIENT, H. A. Eliminating chain drift in price indexes based on scanner data. *Journal of Econometrics*, 2011, 161, pp. 36–46.

DE HAAN, J. AND KRSINICH, F. *Time Dummy Hedonic and Quality-Adjusted Unit Value Indices: Do They Really Differ?* Paper presented at the Society for Economic Measurement Conference, Chicago, U.S, 18–20 August 2014.

DE HAAN J. *A Framework for Large Scale Use of Scanner Data in the Dutch CPI*. Paper presented at the 14th Ottawa Group meeting, Tokyo, Japan, 2015.
DE HAAN, J., WILLENBORG, L., CHESSA, A. G. *An Overview of Price Index Methods for Scanner Data*. Paper presented at the Meeting of the Group of Experts on Consumer Price Indices, Geneva, Switzerland, 2–4 May 2016.
EICHHORN W. AND VOELLER J. *Theory of the Price Index. Fisher's Test Approach and Generalizations.* Berlin, Heidelberg, New York: Springer-Verlag, 1976.
ELTETÖ, Ö. AND KÖVES, P. On a Problem of Index Number Computation Relating to International Comparisons (in Hungarian). *Statisztikai Szemle*, 1964, 42, pp. 507–518.
EUROSTAT. Practical Guide for Processing Supermarket Scanner Data [online]. In: *Harmonised Index of Consumer Prices*, 2017. <https://ec.europa.eu/eurostat/web/hicp/overview>.
FEENSTRA, R. C., MA, H., PRASADA RAO, D. S. Consistent comparisons of real incomes across time and space. *Macroeconomic Dynamics*, 2009, 13(S2), pp.169–193.
FISHER, I. *The Making of Index Numbers.* Boston: Houghton Mifflin, 1922.
GEARY, R. G. A Note on Comparisons of Exchange Rates and Purchasing Power between Countries. *Journal of the Royal Statistical Society*, Series A, 1958, 121, pp. 97–99.
GINI, C. *On the Circular Test of Index Numbers.* Metron 9:9, 1931, pp. 3–24.
GRIFFIOEN, A. R. AND TEN BOSCH, O. *On the Use of Internet Data for the Dutch CPI*. Paper presented at the UNECE-ILO Meeting of the Group of Experts on Consumer Price Indices, Geneva, Switzerland, 2–4 May 2016.
HILL, R. J. Measuring substitution bias in international comparisons based on additive purchasing power parity methods. *European Economic Review*, 2000, 44, pp. 145–162.
INKLAAR, R. AND DIEWERT, W. E. Measuring Industry Productivity and Cross-Country Convergence. *Journal of Econometrics*, 2016, 191, pp. 426–433.
IVANCIC, L., DIEWERT, W. E., FOX, K. J. Scanner Data, Time Aggregation and the Construction of Price Indices. *Journal of Econometrics*, 2011, 161(1), pp. 24–35.
JEVONS, W. S. The variation of prices and the value of the currency since 1782. *J. Statist. Soc. Lond.*, 1865, 28, pp. 294–320.
KHAMIS, S. H. *A* New System of Index Numbers for National and International Purposes. *Journal of the Royal Statistical Society*, Series A, 1972, 135, pp. 96–121.
KRSINICH, F. *The FEWS Index: Fixed Effects with a Window Splice – Non-Revisable Quality-Adjusted Price Indices with No Characteristic Information*. Paper presented at the UNECE-ILO Meeting of the Group of Experts on Consumer Price Indices, Geneva, Switzerland, 2–4 May 2016.
LAMBORAY, C. *The Geary Khamis index and the Lehr index: how much do they differ?* Paper presented at the 15th Ottawa Group meeting, Eltville am Rhein, Germany, 10–12 May 2017.
LASPEYRES, E. Die Berechnung einer mittleren Waarenpreissteigerung. *Jahrbücher für Nationalökonomie und Statistik*, 1871, 16, pp. 296–314.
LEVELL, P. Is the Carli index flawed? Assessing the case for new retail price index RPIJ. *J. R. Statist. Soc*., 2015, A, 178(2), pp. 303–336.
LOON, K. V. AND ROELS, D. *Integrating big data in the Belgian CPI.* Paper presented at the meeting of the group of experts on consumer price indices, Geneva, Switzerland, 8–9 May 2018.
MADDISON, A. AND RAO, D. S. P. *A Generalized Approach to International Comparison of Agricultural Output and Productivity*. Research memorandum GD-27, Groningen Growth and Development Centre, Groningen, The Netherlands, 1996.
MARTINI, M. A General Function of Axiomatic Index Numbers. *Journal of the Italian Statistics Society*, 1992, 1(3), pp. 359–376.
PAASCHE, H. Über die Preisentwicklung der letzten Jahre nach den Hamburger Borsennotirungen. *Jahrbücher für Nationalökonomie und Statistik*, 1874, 12, pp. 168–178.
SUMMERS, R. International Price Comparisons Based Upon Incomplete Data. *Review of Income and Wealth*, 1973, 19, pp. 1–16.
SZULC, B. Indices for Multiregional Comparisons (in Polish). *Przeglad Statystyczny*, 1964, 3, pp. 239–254.
SZULC, B. Linking Price Index Numbers. In: DIEWERT, W. E. AND MONTMARQUETTE, C. eds. *Price Level Measurement*, 1983, pp. 537–566.
TÖRNQVIST, L. The Bank of Finland's Consumption Price Index. *Bank of Finland Monthly Bulletin*, 1936, 10, pp. 1–8.
VAN DER GRIENT, H. A. AND DE HAAN, J. *The use of supermarket scanner data in the Dutch CPI*. Paper presented at the Joint ECE/ILO Workshop on Scanner Data, Geneva, 10 May 2010.
VON AUER, L. Processing scanner data by an augmented GUV index. *Eurostat Review of National Accounts and Macroeconomic Indicators*, 2017, 1, pp. 73–91.
VON AUER, L. *The Nature of Chain Drift*. Paper presented at the 17th Meeting of the Ottawa Group on Price Indices, Rio de Janeiro, Brasil, 8–10 May 2019.
WALSH, C. M. *The Measurement of General Exchange Value.* New York: The MacMillan Company, 1901.
WILLENBORG, L. *Chain Indexes and Path Independence*. Report, Statistics Netherlands, 2010.
WILLENBORG, L. AND VAN DER LOO, M. *Transitivizing Price Index Numbers Using the Cycle Method: Some Empirical Results*. Report, Statistics Netherlands, 2016.
WILLENBORG, L. *Transitivizing Elementary Price Indexes for Internet Data using the Cycle Method*. Discussion Paper, Statistics Netherlands, 2017.