Impact of the partial time coverage of retail chain data on the accuracy of the price index calculation

Petra Mazureková[1], Helena Glaser-Opitzová[1]

[1]Statistical Office of the Slovak Republic

Abstract

In the framework of modernization and improvement of the quality of price statistics by using new data sources, the Statistical Office of the Slovak Republic (SOSR) has commenced using transaction data from retail chains, also called scanner data. SOSR collects data for the food and non-alcoholic beverages sector from the five largest retail chains that contain sales and quantities sold that are aggregated on a weekly level for each individual item. The implementation of the new data source in the production environment required a significant methodological change. The calculation of price indices no longer involves the price determined at a specific time (as in the traditional collection), but the average price per unit of goods for the observed period. This type of price more accurately reflects the prices that consumers pay throughout the entire observed period, taking into account discounts and the impact of these discounts on the quantity of goods sold. Consequently, weekly data sets are aggregated on a monthly level due to the frequency of the compilation of price indices. These monthly files contain aggregated values of sales and quantities sold for individual product items for selected weeks of the month. The accuracy of the average price is affected by the length of the time period considered within the month, i.e. the number of weeks used in the monthly aggregation. Theoretically, the best practice would be to use a complete month for the compilation of the Harmonized Index of Consumer Prices (HICP). However, in practice, due to the HICP publication timeliness, the time span of price data only covers two weeks for the reference month. The aim of this paper is to assess the impact of insufficient time coverage, i.e. to examine the impact of using price data with different lengths of time span on the values of average prices and price indices. We use data from the Slovak market to compare different types of indices, namely, Jevons, Törnqvist, and GEKS-Törnqvist.

Keywords: price indices, transaction data, partial time coverage

1. Introduction

One of the innovative data sources in price statistics is electronically recorded data on retail transactions, called scanner data. They are a source of detailed information on sales of consumer goods at the EAN/GTIN level because they represent complete information on all transactions made by a particular retailer over a given time period. The use of this information makes it possible to calculate unit prices for each product item, which take into account the different quantities that are sold at different prices during the reference period. It also takes into account the real impact of discounts.

The data used in the calculation of the HICP/CPI should cover the whole reference period for which the index is calculated. However, in practice, statistical offices normally use a sub-

period of the reference period for reasons of trade-off between timeliness and accuracy in the production of price indices. In view of the deadline for the delivery of transaction data by retail chains and subsequent publication practices, SOSR uses the first two full weeks of the reference month for the determination of the unit prices.

The aim of this paper is to analyze the impact of the partial time coverage. The calculation of price indices was carried out on a file with a partial time coverage of 2 weeks (2W), 3 weeks (3W), and a full time coverage of a whole month (4W). This means that the average unit price of goods was calculated from data for 2, 3 weeks, and the full month. The indices calculated for the full time coverage served as a benchmark for comparison. In practice, performing the calculation using the entire reference period is unrealistic due to publication deadlines.

A dynamic approach was taken to select individual products for the calculation of price indices. The dynamic method automatically selects a representative sample of item codes for each consecutive set of two months (t and t+1, t+1 and t+2, t+2, and t+3, and so on) by selecting all matched item codes that have a turnover above a certain threshold and will include new and sufficiently important items whilst dropping less important items (EUROPEAN COMMISSION, EUROSTAT, 2017). Items with extreme price changes and dumping products are also excluded from the price index calculation.

2. Theoretical framework

From the bilateral indices we select the unweighted Jevons index and the weighted Törnqvist symmetric superlative index, and their chained version.

The Jevons index is defined as (EUROPEAN COMMISSION, EUROSTAT, 2017):

$$I_{Jevons}^{m-1,m} = \prod_{i \in N} \left(\frac{p_i^m}{p_i^{m-1}} \right)^{\frac{1}{N}} \quad (1)$$

where *m* is the current period, *m-1* is the previous period and p_i is the price of the i-th product item.

The Törnqvist index is defined as (ILO/IMF/OECD/UNECE/Eurostat/The World Bank, 2020):

$$I_{Tornqvist}^{m-1,m} = \prod_{i \in N} \left(\frac{p_i^m}{p_i^{m-1}} \right)^{\frac{s_i^{m-1} + s_i^m}{2}} \quad (2)$$

where $s_i^{m-1} = p_i^{m-1} q_i^{m-1} / \sum_{i \in N} p_i^{m-1} q_i^{m-1}$ a $s_i^m = p_i^m q_i^m / \sum_{i \in N} p_i^m q_i^m$ are the shares of expenditures in periods *m-1* and *m,* respectively, and q_i are the quantities sold in a given month.

Both of these indices compare the price change between two consecutive periods. For the publication practice, the index must be referenced to a base period 0, i.e. to December of the previous year. It is therefore necessary to chain the month-on-month index as:

$$I^{0,m} = I^{0,1} . I^{1,2} \dots I^{m-1,m} \tag{3}$$

A drawback of chained indices is that they generate a chain drift, which occurs when an index does not return to 1 when prices and quantities in the current period return to their levels in the base period. To eliminate this negative effect, multilateral methods are recommended. In multilateral methods, the aggregate price change between two compared periods is obtained from prices and quantities observed in multiple periods. One of the representatives of these types of indices is the GEKS-Törnqvist index. The GEKS method uses bilateral Törnqvist indices for its compilation and it is defined as (EUROSTAT, 2022):

$$I^{0,m}_{GEKS-Törnqvist} = \prod_{l \in T} \left(I^{0,l}_{\text{Törnqvis}} \times I^{l,m}_{\text{Törnqvist}} \right)^{\frac{1}{T}} \tag{4}$$

where T is the window length, in our case T=13 months and the length of the interval for calculation is 25 months.

A disadvantage of multilateral methods is that they suffer from revision, i.e. each time a new index is calculated for period T+1 and so on, all previous indices are recalculated and changed for the defined time window. This is an undesirable phenomenon for statistical offices. Therefore, splicing techniques must be used that link the latest multilateral index to previous results in order to avoid revisions of already published results. Technically, the splicing of two series operates via a link month. There are different possibilities for splicing when compiling the results for T+1 and so on. We use the HASP (half splice - link to published) splicing method, which is defined as (EUROSTAT, 2022):

$$I^{0,t}_{pub} = I^{0,t-1}_{pub} \times \prod_{k=t-T+1}^{t-1} \left(I^{t-1,k}_{pub} \times I^{k,t}_{[t-T+1,t]} \right)^{\frac{1}{T-1}} = \prod_{k=t-T+1}^{t-1} \left(I^{0,k}_{pub} \times I^{k,t}_{[t-T+1,t]} \right)^{\frac{1}{T-1}} \tag{5}$$

Note that the above indices are applied at the elementary level, which, in our case, means the ECOICOP6[1]. A Laspeyres-type index is used to calculate price indices at a higher level of aggregation of the ECOICOP classification, which is defined as (EUROSTAT, 2018):

$$P_A^{y,m/y-1} = \frac{\sum_{a\in A} w_a^{y-1} I_a^{y,m/y-1}}{\sum_{a\in A} w_a^{y-1}} \quad (6)$$

where w_a^{y-1} are weights based on annual expenditure for all items belonging to the elementary aggregate of ECOICOP6.

We quantify the error that would be caused by the partial time coverage through the mean absolute percentage error - MAPE over an interval length of 12 months, which is defined as:

$$MAPE = \frac{1}{12}\sum_{i=1}^{12} \frac{\left|I_i^{4w} - I_i^{*w}\right|}{I_i^{4w}} * 100 \quad (7)$$

where I_i^{4W} is the index calculated from full month data in the i-th period and I_i^{*W} is the index calculated from data for either 2 or 3 weeks of the i-th period.

3. Result

We analyze the evolution of the time series of price indices at the 6-digit national level (ECOICOP6) and at the 3-digit ECOICOP level, for the product groups of food and non-alcoholic beverages.

We first examine the actual unit prices of the selected goods. The following table shows how differently the prices of specific goods can change depending on the length of time span.

Table 1: Average unit prices of goods for different time coverage in selected periods (green colour marked price decrease and yellow colour marked price increase for a given product)

GTIN/EAN	Item description	Period	Average price in €		
			2W	3W	4W
4008671013004	Crystal sugar 1kg	04	0,9395	0,9147	0,9099
4008671013004	Crystal sugar 1kg	05	0,9150	0,9129	0,9137
4008671013004	Crystal sugar 1kg	06	0,8386	0,8902	0,9179
4008671013004	Crystal sugar 1kg	07	1,0497	1,0357	0,9961
00401301	RAJO UHT semi-skimmed milk 1,5 % 1 L	01	0,9446	0,9374	0,9390
00401301	RAJO UHT semi-skimmed milk 1,5 % 1 L	02	0,9476	0,9359	0,9387

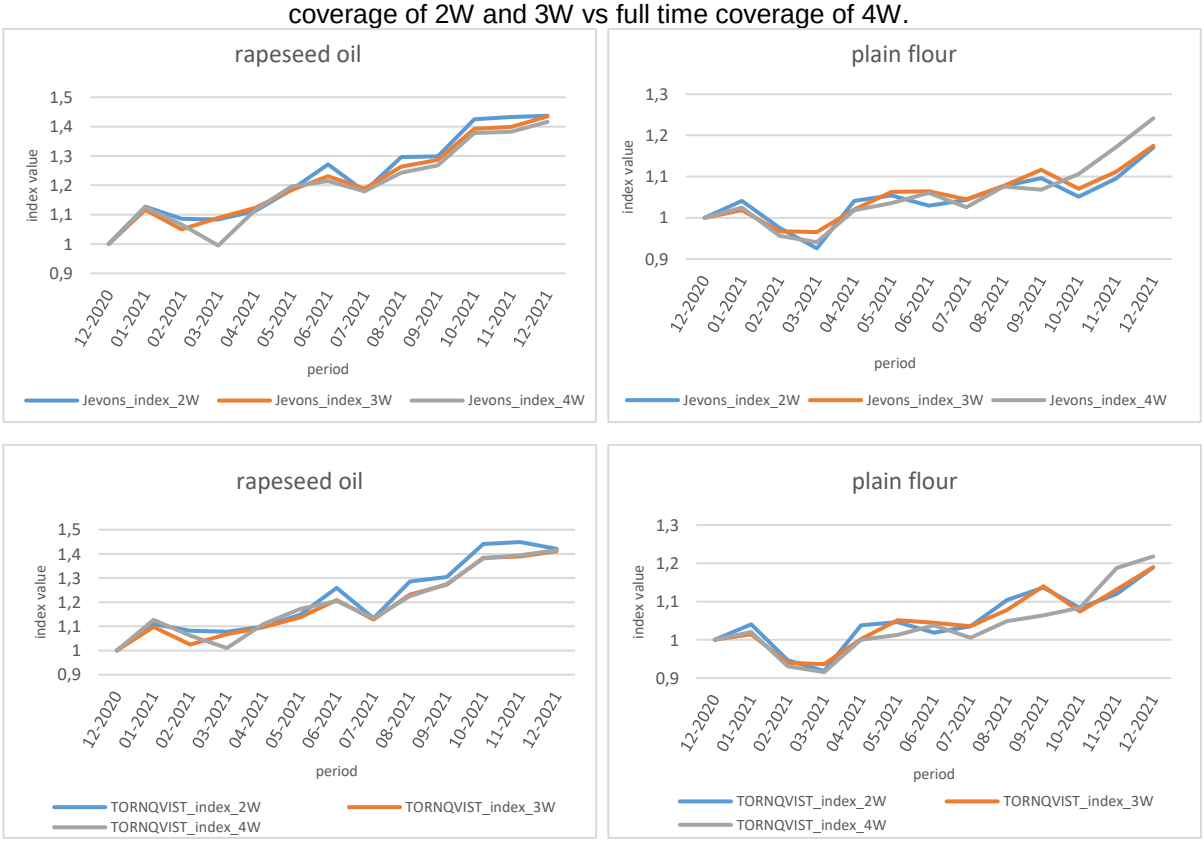
[1] ECOICOP6 - the national lower level of the international ECOICOP classification. It consists of homogeneous product groups, the same for all retail chains.

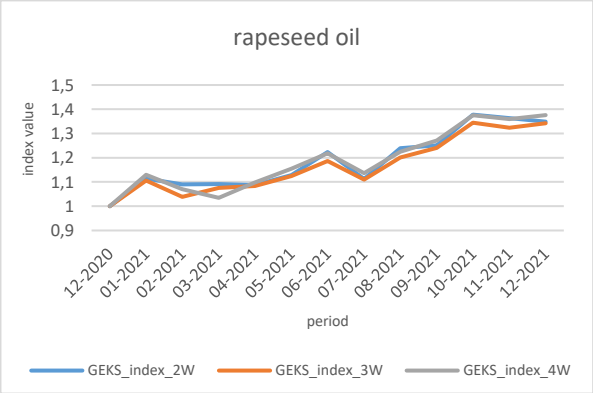
00401301	RAJO UHT semi-skimmed milk 1,5 % 1 L	03	0,9236	0,9415	0,9305
00401301	RAJO UHT semi-skimmed milk 1,5 % 1 L	04	0,9868	1,0061	1,0169
00401301	RAJO UHT semi-skimmed milk 1,5 % 1 L	10	1,4422	0,9475	0,9553
00401301	RAJO UHT semi-skimmed milk 1,5 % 1 L	11	1,4762	1,4980	1,4832
00401301	RAJO UHT semi-skimmed milk 1,5 % 1 L	12	1,5491	1,2081	1,2057
2002006552802	Rapeseed oil RACIOL 2l	08	6,8671	6,7935	6,7628
2002006552802	Rapeseed oil RACIOL 2l	09	6,7874	5,8253	5,7718
2002006552802	Rapeseed oil RACIOL 2l	10	6,2477	6,3722	6,5170
2002006552802	Rapeseed oil RACIOL 2l	11	7,2538	7,2438	7,2702

Table 1 shows that the average price of crystal sugar calculated from 2 or 3 weeks decreases but the average price of crystal sugar calculated from 4 weeks increases between the fifth and sixth month. Another example is that the average price of RAJO semi-skimmed milk calculated from 2 weeks increases but its average price calculated from 3 or 4 weeks decreases between the eleventh and the twelfth month.

The impact of partial time coverage on the values and trend of the price indices calculated for selected groups - rapeseed oil and plain flour is illustrated in Figures 1-6 and at the 3-digit ECOICOP level in Figures 7-12.

Figure 1-6: Comparison of the trend of monthly price indices on homogeneous product groups, partial coverage of 2W and 3W vs full time coverage of 4W.





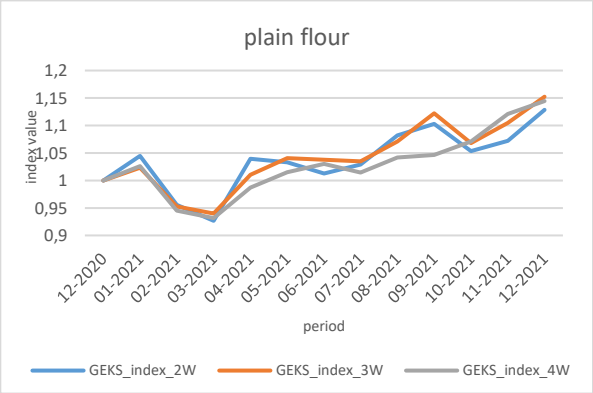
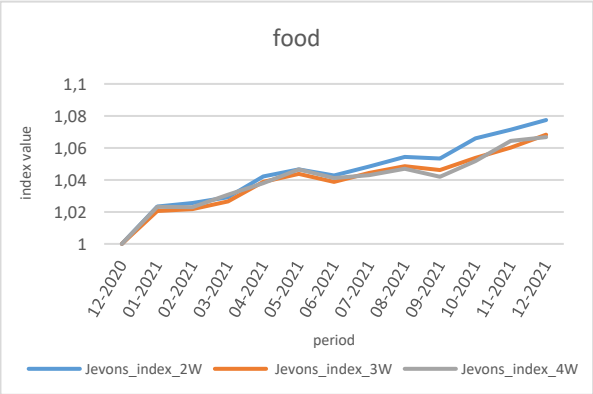
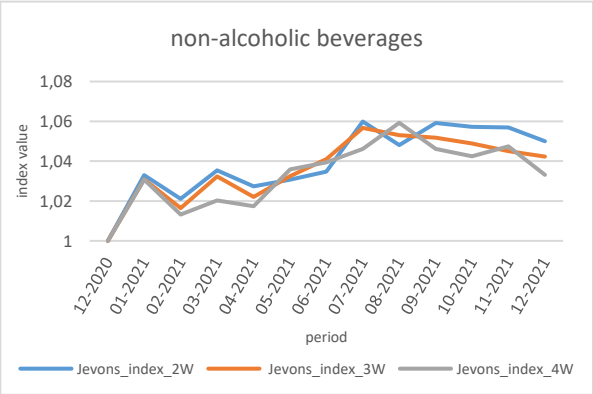
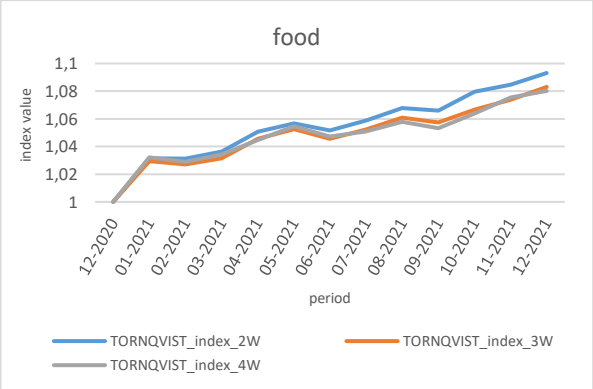
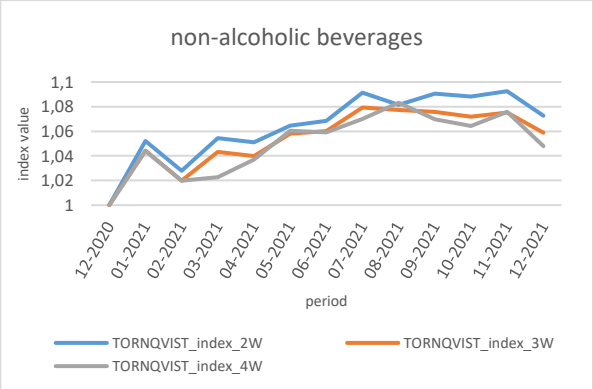


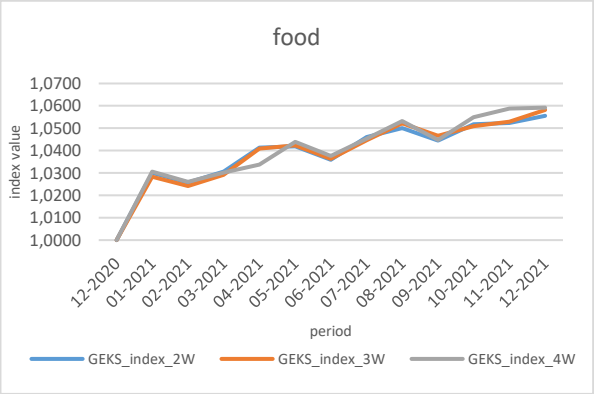
Figure 7-12: Comparison of the trend of monthly price indices on the product group - 01.1 Food and 01.2 Non-alcoholic beverages, partial 2W and 3W vs. full time coverage 4W

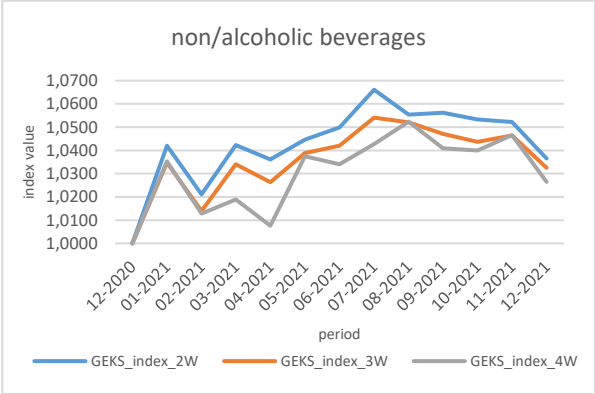












From the figures, we see that the time series of indices, which are calculated from partial time coverage of 2 weeks, are more or less overestimated at the ECOICOP6 level. The overestimation is also visible at the level of food and non-alcoholic beverages, too.

Table 2: Mean absolute percentage error caused by partial coverage

	Mean Absolute Percentage Error (MAPE)					
	Jevons 2W/4W	**Törnqvist 2W/4W**	**GEKS 2W/4W**	**Jevons 3W/4W**	**Törnqvist 3W/4W**	**GEKS 3W/4W**
01.1 Food	0,527%	0,672%	0,238%	0,230%	0,219%	0,236%
01.2 Non-alcoholic beverages	0,998%	1,462%	1,304%	0,525%	0,534%	0,590%

The results in Table 2 show that by using partial time coverage of 2 weeks, we accept an error of less than 0,7% for food and less than 1,5% for non-alcoholic beverages, with partial time coverage of 3 weeks the error is further reduced.

4. Conclusions

The aim of the paper was to find out the impact of the partial time coverage of the observation period (in our case 4 weeks vs. 2 weeks or 3 weeks). The analysis was performed for the bilateral Jevons and Törnqvist index and for the multilateral GEKS-Törnqvist index. The results of the analysis show that the different partial time coverage does not have a significant impact on the values of the price indices. Additionally, it was found that it is better and more accurate to use partial coverage for 3 weeks of the reference period to calculate price indices. This fact creates a premise for us to try to make our production processes more effective in the practice of price statistics so that we are able to implement the coverage of the reference period for 3 weeks.

References

de Haan, J.(2015). A framework for large scale use of scanner data in the Dutch CPI. *In 14th Meeting of the International Working Group on Price Indices: Tokyo, Japan, 20 – 22 May 2015 Tokyo: Ottawa Group,* http://www.ottawagroup.org/Ottawa/ottawagroup.nsf/4a256353001af3ed4b2562bb00121564/d012f001b8a1cf6cca257eed008074c9/$FILE/Jan%20de%20Haan%20(Statistics%20Netherlands)%20A%20Framework%20for%20Large%20Scale%20Use%20of%20Scanner%20Data%20in%20the%20Dutch%20CPI.pdf

Diewert, W.E. (1995). Axiomatic and Economic Approaches to Elementary Price Indexes. *Discussion Paper No. 95-01, Department of Economics, University of British Columbia,Vancouver, Canada* https://www.nber.org/system/files/working_papers/w5104/w5104.pdf

Diewert W.E. (2021). Scanner Data, Elementary Price Indexes and the Chain Drift Problem. *Discussion Paper 20-07,Vancouver School of Economics, The University of British Columbia, Vancouver, Canada V6T 1L4.* https://www.imf.org/-/media/Files/Data/CPI/companion-publication/chapter-6-chain-drift-problem-and-multilateral-indices.ashx

EUROPEAN COMMISSION, EUROSTAT (2017): HICP-Practical Guide for Processing Supermarket Scanner . https://circabc.europa.eu/ui/group/7b031f10-ac19-4da3-a36f-58708a70133d/library/8e1333df-ca16-40fc-bc6a-1ce1be37247c/details

EUROSTAT (2018). Harmonised index of consumer prices (HICP): methodological manual. *Luxembourg: Publications Office of the European Union*, https://data.europa.eu/doi/10.2785/68673

EUROSTAT (2022). Guide on Multilateral Methods in the Harmonised Index of Consumer Prices, *Luxembourg: Publication Office of the European Union*. ISBN 978-92-76-44354-4. https://ec.europa.eu/eurostat/documents/3859598/14503841/KS-GQ-21-020-EN-N.pdf/243796c9-f5ad-2155-e546-c94e17d9a7eb?t=1649074284236

Glaser-Opitzová H., Mazureková P. (2023). Vplyv parciálneho časového pokrytia údajov zo skenerov na presnosť cenových indexov. *Slovenská štatistika a demografia, vol.33 (3/2023), 39-54.* https://ssad.statistics.sk/SSaD/index.php/slovenska-statistika-a-demografia-3-2023/

Chessa, A.(2016). A new methodology for processing scanner data in the Dutch CPI. *In Eurostat Review on National Accounts and Macroeconomic Indicators 2016 [online]. Luxembourg: Publications Office of the European Union, ISSN: 1977-978X*. Https://ec.europa.eu/eurostat/documents/3217494/7556543/KS-GP-16-001-EN-N.pdf/70e246de-734c-42ba-bee2-bc0b3dd97faa

ILO/IMF/OECD/UNECE/Eurostat/The World Bank (2020). Consumer Price Index Manual: Concepts and Methodts*. IMF Publications, Washington, DC. ISBN 978-1-51354-298-0 (PDF).* https://unece.org/sites/default/files/2020-12/cpi-manual-concepts-and-methods.pdf