Using scanner data for sports equipment*

Paper written for the joint UNECE/ILOs Meeting of the Group of Experts on Consumer Price Indices, 7-9 May 2018, Geneva, Switzerland

Kjersti Nyborg Hov
Advisor, Division for Price Statistics
Kjersti.Nyborg.Hov@ssb.no

Randi Johannessen
Head of division, Division for Price Statistics
Randi.Johannessen@ssb.no

Statistics Norway

Using scanner data for sports equipment

In recent years a modernization of how to produce the consumer price index (CPI) has been in focus. Along with technical improvements and digitalization, new opportunities have emerged. One of these opportunities being access to alternative data sources. Technical improvements make it possible to collect actual transaction prices using electronic price data files from the retailers, the so-called scanner data. Using electronic price files such as scanner data makes price collectors or questionnaires redundant. On the other hand, new challenges emerge; how to process the data to properly produce the consumer price index. In this paper we discuss how scanner data for sports equipment has been implemented in the Norwegian CPI.

Introduction

According to Eurostat scanner data can be described as: "*transaction data obtained from retail chains containing data on turnover, quantities per item code based on transactions for a given period and from which unit value prices can be derived at item code level*"[1]. Statistics Norway has utilized scanner data since the late 1990s in the production of the Norwegian CPI. At first, scanner data from supermarkets were used to imitate the traditional price collection method for food and non-alcoholic beverages. The first approach has later become known as the *static approach*, namely using a fixed sample of items identified by item codes (Eurostat, 2017). Later, the use of scanner data has been through methodological developments, and the utilization has expanded in the Norwegian CPI. As of today, the static approach has for the most part been replaced by a more *dynamic approach*, using a larger share of the price data. In addition to the indices of food and non-alcoholic beverages, also the

* The authors would like to thank Ragnhild Nygaard and Ingvild Johansen, Statistics Norway, for conducting the analytical research prior to, and the actual implementation of scanner data for sports equipment in the Norwegian CPI. Their work form the basis for this paper.

[1] Eurostat (2017), "Practical guide for processing supermarket scanner data", URL: https://circabc.europa.eu/sd/a/8e1333df-ca16-40fc-bc6a-1ce1be37247c/Practical-Guide-Supermarket-Scanner-Data-September-2017.pdf

indices of alcoholic beverages, tobacco, home and personal care products, gasoline and diesel, and pharmaceutical products are calculated using scanner data. Although, Statistics Norway has long experience in how to utilize scanner data in the CPI, implementing scanner data for other consumer goods is not necessarily straight forward.

The consumer goods currently covered by scanner data are characterized by certain stability in the range of goods, identified by GTIN or other item codes[2]. Consumer electronics, clothing and footwear and sports equipment, on the other hand are subject to frequent changes in the range of goods., This challenging the current calculation methods, particularly the issue of not performing (automated) actual replacements of items leaving the market. Currently there is a strong focus on how to best utilize scanner data, especially the multilateral calculation methods, that is meant to better exploit the weight information on a more detailed level. At present no national or international recommendations have been agreed upon, nonetheless the need for scanner data calculation methods that are also able to handle the difficulties of goods being short lived are still present.

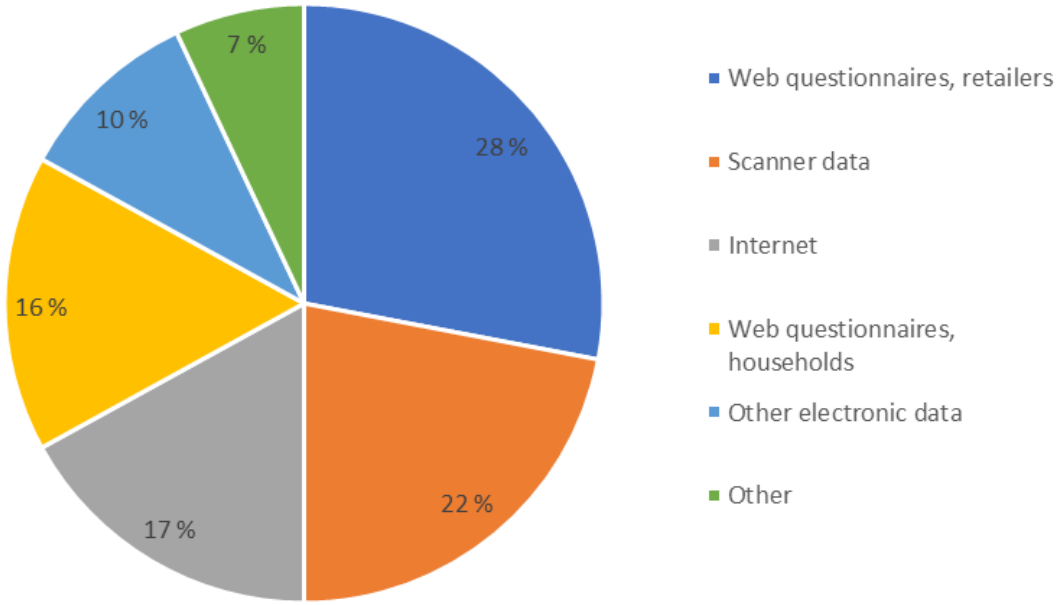
As of January 2018, a cautious approach to using scanner data for a sports equipment chain in the Norwegian CPI was implemented. This paper gives an overview of the experiences of the implementation of scanner data for sports equipment in the Norwegian CPI. First, an introduction and the utilization of scanner data in the Norwegian CPI before an overview of the Norwegian sports equipment market. Further a thorough examination of the sports equipment scanner data, the challenges it possesses and the solutions on how to best utilize the data. In the end a short description of the implemented calculation method before some concluding remarks.

Scanner data in general

In Statistics Norway, as well as in other national statics offices, the desire for scanner data is ever emerging. The technical improvements over the past decades have made it possible to achieve greater amount of data and greater information at a lower cost than the original price collection methods. The traditional price collection method for the Norwegian CPI has been (web) questionnaires sent to a sample of retailers. Covering close to 30 per cent of the CPI measured by the CPI weights, it is still a substantial part of the price collection methods for the Norwegian CPI, however the share is decreasing.

[2] An item code can be presented as the Global Trade Item Number (GTIN), a Price Look Up code (PLU) or other chain specific item codes. The term “Item code” will be used as the unique identifier code throughout the paper.

Figure 1: Price collection methods represented as share of the CPI weight



Source: Statistics Norway

Reasons for using alternative data sources such as scanner data is among others, to achieve greater amount of data at a lower cost, both for the statistics office and the data providers. The Norwegian Statistics Act states that Statistics Norway is in position to claim price data in any delivery form available, and at no cost[3]. For a national statistics office the legislation is a great privilege in the request for data. Even though there is no cost for the statistics office to claim the data, there is a noteworthy cost for the data providers, in this case the retailers that have to fill out the questionnaires[4]. The data collection method using questionnaires requires the data providers to manually fill out price data each month, i.e. occupying working hours for the data providers. For some, the number of items requested is low and the time spent filling out the questionnaires are negligible, for others however it requires considerable time spent. This is of course a cost that should not be ignored. In addition, using alternative data sources such as scanner data we receive substantially more price data, including additional metadata, which in turn should be able to give us better quality indices. Thus, in the Norwegian case, it is both in the statistical office and the data providers' interest to use alternative data sources such as scanner data. However, using scanner data as an alternative data source is not always straight forward. Excess data and metadata is only more worth if we are able to utilize the data adequately.

Up until January 2018 the consumer goods covered by scanner data in the Norwegian CPI were food and non-alcoholic beverages, alcoholic beverages, tobacco and other everyday commodities from supermarkets and gasoline stations, gasoline and diesel, and pharmaceutical products both from the

[3] The Norwegian Statistics Act, URL: https://lovdata.no/dokument/NL/lov/1989-06-16-54

[4] Unlike most countries, data collectors are not being used for the data collection in the Norwegian CPI. Instead a web questionnaire is sent to the retailers and the retailers themselves fill out the questionnaire.

supermarkets and pharmacies. The first indices fully covered by scanner data in the Norwegian CPI were the indices of food and non-alcoholic beverages. As of today, the indices of food and non-alcoholic beverages are calculated using a monthly chained un-weighted Jevons formula, in line with Eurostat recommendations on how to process supermarket scanner data (2017)[5]. To avoid having price data from marginal items influencing the indices there is a cut-off threshold based on turnover included in the calculation. Based on the cut-off threshold covering 80 per cent of the total turnover in period *t* and *t+1*, approximately 60 per cent of the items are excluded. For consumer goods such as food and non-alcoholic beverages the matched model approach at item code level has proved well, however, there is no replacement of items that are either temporarily or permanently missing. In general, the prices of missing items will be imputed based on the aggregated level. Permanently missing items can be subject to sales activity and/or drop in sales volume in the time period before exiting the market, thus causing a downward drift in the indices if included and not being replaced. A so-called "dumping filter" is therefore being used to counterbalance the difficulties of not being able to replace outgoing items. The dumping filter excludes items that show a sharp decline in both sales price and sales volume given certain thresholds (Eurostat, 2017).

Even though the cut-off threshold and dumping filter has worked well for the calculation of the food and non-alcoholic beverages indices, yet this method might not be as sufficient for other consumer goods such as clothing and footwear, consumer electronics and sports equipment, where the product life cycle and the price development proves different.

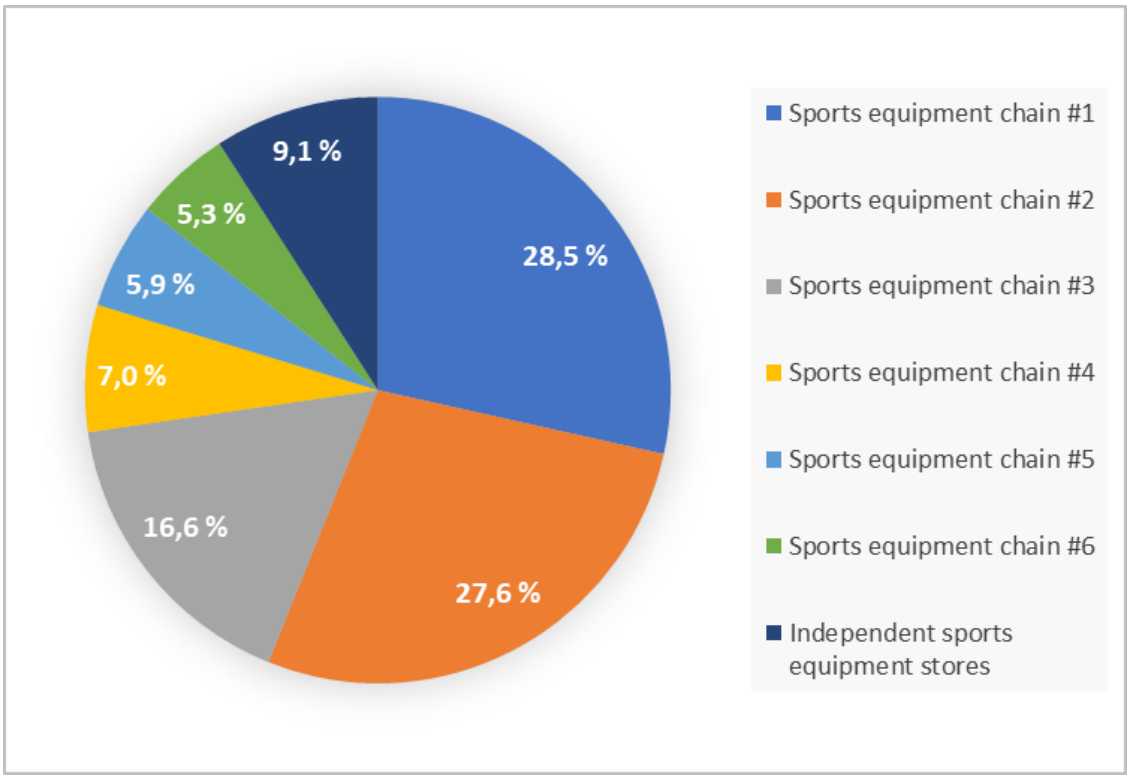
Scanner data for sports equipment

During the second half of 2016, a leading sports chain began providing scanner data test files including back-data, alongside the traditional web questionnaires. Thus, we were able to start testing the data with the aim of implementing scanner data also for sports equipment. These scanner data covers several consumer goods; clothing, footwear, sports equipment and items for recreational use, thus ideal for testing calculation methods. Further analysis of the data, may provide great insight on how to treat scanner data for e.g. clothing and footwear in general.

The Norwegian sports equipment market is characterized by chain concentration. The market is dominated by three leading chains, still also some smaller chains are well represented, see figure 2. The two largest chains, measured by market share, cover more than half the market.

[5] Superlative indices such as Fisher and Törnqvist are not recommended as these tend to cause drifting in the indices as the sales volumes after a sales period does not match the sales volume during the sales period.

Figure 2: The Norwegian sports equipment market, market shares



Source: Norsk Sportsbransjeforening[6]

The data variables included in the scanner data for the sports equipment chain is somewhat more detailed than the scanner data from supermarkets, gasoline stations and the pharmaceutical stores, meaning it provides more variables than usual. Not necessarily surprising as different consumer goods contains various characteristics. The level of detail in the scanner data from the sports equipment stores has been essential in the data analysis, and the end result. The data variables received on a weekly basis are shown in table 1.

Table 1: Scanner data sports equipment, variables

Variables	**E.g.**
Store concept identifier	"Main", "Warehouse", "Local"
Item code	99999999
Item text	Nike Air Max Thea Sneaker shoe, women
Chain specific product group code	9999
Chain specific product group text	Sneaker shoes, woman
Time periode	yyyy.mm.ww, e.g. 20180419
Transaction price in a week calculated as total turnover pr. item code pr. store divided by quantity	NOK 1199
Quantity sold in a week per item code and store concept	5
The color of the item	110-white/white, 033-desert sand
The size of the item	Small, medium, onesize, 37-39
The brand of the item	Nike, Adidas, Burton

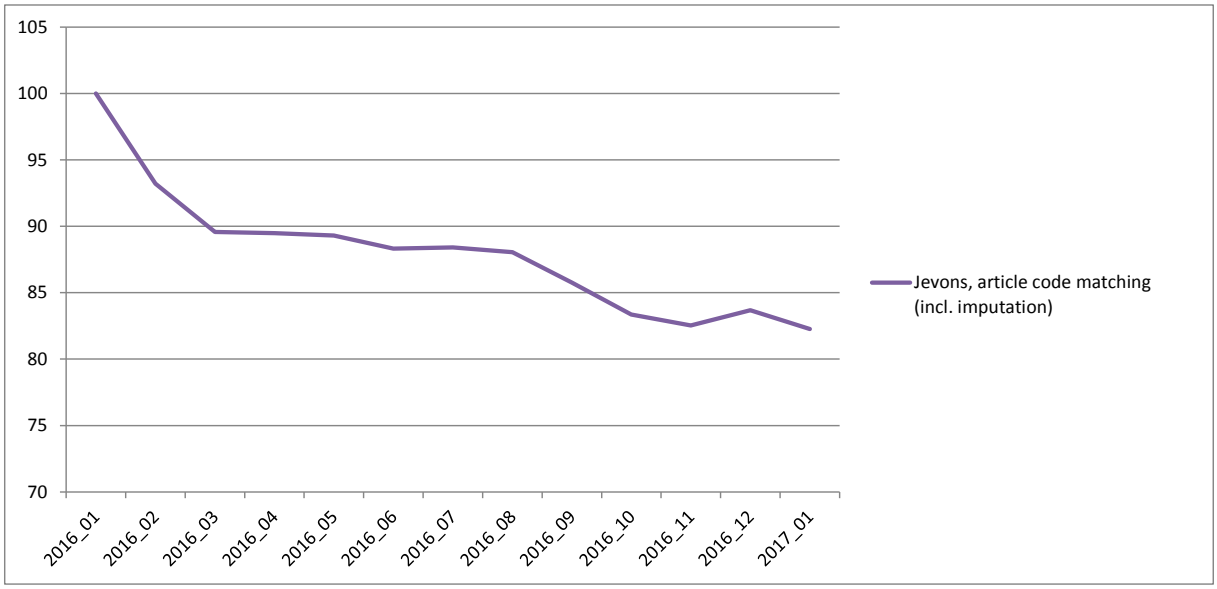
[6] Norsk Sportsbransjeforening (2016), URL: http://www.sportsbransjen.no/uploads/magazines/78/pdf/Sportbransjen%202016%20nedkortet%20versjon_d503067b33844bf7805e78bc3ff132cc.pdf

The received scanner data contains information for all stores within the sports equipment chain, separated by best described as *store concepts*. The specific sports equipment chain is separated into three different store concepts with its own concept/brand name; (1) the main concept covering a large share of the total turnover, (2) a larger warehouse concept characterized by larger stores and bulk items, and (3) a more local oriented concept where items are adjusted to local consumption and price levels tend to be somewhat higher than the former two.

Product life cycle

In the assessment of utilizing scanner data for sports equipment the product life cycle is of importance, especially for the choice of calculation method. While food and non-alcoholic beverages are staple household items, thus less subject to replacement, certain sports equipment items are shorter lived as trends, fashion and style are changing. This is not to say that food and non-alcoholic items not are subject to replacements, but to a lesser degree than for example clothing, footwear and other sports equipment. To verify these assumptions, the time period an item was being purchased was examined, using the item code as a unique identifier. Test calculations using a matched model approach at item code level showed the expected results; the indices proved to have a downward trend throughout the year. Items that are leaving the market are often facing a sales period prior to the exit, meaning that if exiting items are included in the calculation, and not being replaced by the new item, there will be a systematic downward drift in the indices. Hence, using a matched model approach on item code level without replacement as the calculation method will not be suitable.

Figure 3: Jacket's men. Matched model, item codes. Index. January 2016=100.



Traditional procedures for treating permanently missing items is to replace items no longer representative and/or leaving the market by a similar item that meets the same utility from a consumer perspective. Using scanner data however complicates the replacement strategy. First of all, the large quantity of data makes replacement strategies troublesome; an automated solution

ought to be in place for this to be possible to maintain. The matched model approach, that means items at item code level, creating parameters to detect an identical item is crucial, however not necessarily applicable. Using the traditional price collection method, namely using web questionnaires for the retailers to fill out, the retail clerk will identify and provide a replacement item that is assumed having the same utility from a consumer perspective. If the new and old item are not of same quality, the retailer will remark this. In scanner data however, the task of replacement will belong to the statistician, with more limited knowledge of the items besides the item description. Imagine a plain grey, cotton t-shirt whit a rounded neck is going out of stock. A different color, but otherwise identical shirt, will contain a different item code. We would then have to rely on the item description variables to find a match, however these are often not detailed enough. Thus, it could be difficult to identify the actual replacement item(s), at least in an automated way. Also, the point in time of a replacement is an issue, namely when to replace an outgoing item. Often an overlap occurs; the relaunch/replacement item is introduced in the market before the outgoing item has exited. On the other hand, due to for example seasonality, a replacement item might not be available on the market before months after the exiting item has gone out of stock. The difficulty is thus to both identify a replacement item, also finding the optimal point in time for when to perform the actual replacement, again in an automated way.

An alternative to using matched model at item code level, targeting the challenges explained above, could be to create a group of homogenous items. This procedure would allow for more than one replacement item and a gradual entering/exiting of items, hence not being forced to decide when the actual replacement should be carried out. This strategy requires creating well defined homogenous item groups, so called item clusters, where all the items within the cluster may be viewed as identical in both quality and utility. The unit value (all) the items in the item cluster will then be used as a price observation in the same manner as other traditional price observations.

It should be noted that hedonic calculation methods could be an alternative to homogenous item groups, however there are difficulties of not always having enough information (i.e. explanatory variables) to properly perform hedonic calculations. As discussed later, regression analysis of the price determining variables showed significant results. However the explanatory power was rather small, and given that the variables received are somewhat limited with respect to quality characteristic needed for hedonic regressions, we found it less suitable to use hedonic calculation methods.

Creating well defined item clusters are subject to several considerations. Too narrowly defined cluster will be no different than using the item code as the unique identifier. Due to the limited life span in combination with the price development of some specific sports equipment items, using item codes as the unique identifier would result in a downward drift in the indices if exiting items are not replaced. This is illustrated in figure 3. On the other hand, a widely defined cluster would allow for replacement items, however too widely defined the cluster would be subject to unit value index bias. E.g. that the items are not homogenous enough causing unwanted quality and price dispersion; items with higher (lower) price levels will overshadow (pull down) the prices and distort the price movements of items with lower (higher) price levels if they were in the same cluster.

Creating item clusters for sports equipment scanner data

The initial stages of creating homogenous item clusters for the sports equipment scanner data consisted of analyzing the price determining variables in the data. As mentioned above, the scanner data for sports equipment contains more explanatory variables than the scanner data received from the other scanner data providers, thus giving more insight on the price determining variables. Regression analysis of the price data show that the explanatory power of the variables is no higher than 30 to 40 per cent, however some conclusions can still be drawn.

Much as expected neither the color of an item nor the item size (e.g. clothing size) of the item seems to have a significant impact on the price. The significant price determining variables however proved to be *brand* and *store concept*. The brand of an item is assumed to represents both an actual and a perceived level of quality, the price following accordingly, thus the brand is not surprisingly a significant price determining variable. The variable store concept is a variable containing the three store concept names, hence showing at which store (concept) the item was being purchased. As explained earlier, the sports equipment chain is made up by three so-called store concepts, all having their own product and price profile. The locally oriented concept stores proved in general having higher price levels than the main concept stores and the warehouse concept stores, the latter having the lowest price levels. The difference in price levels is mainly due to the various store concept offering different assortments of brands, collections, and at various quality. The regression analysis proved that the price level differences across the store concepts were significant and should be taken into account when creating the item clusters. There are also online stores associated with the store concepts, however separated into only two online stores; one covering both the main and the warehouse concepts and one covering the locally anchored concept stores. The regression analysis showed that there was not a significant price level difference between the store concept and its corresponding online store. Since one of the online stores is being used for both the main and the warehouse store concept, the price data from the online store was combined with the price data of the main store concept due to it covering the largest overall turnover.

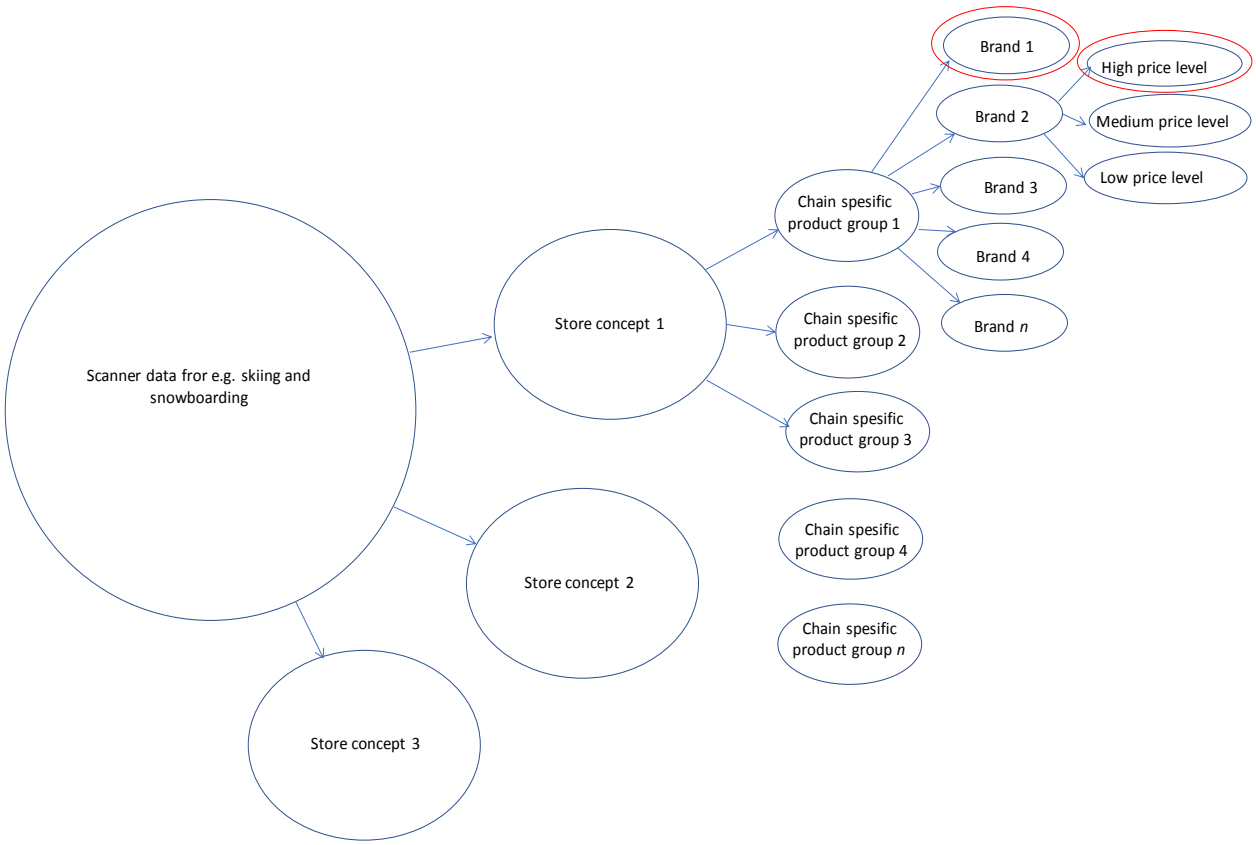
In the creation of clusters, we also made use of the chains *own classification system*. The sports equipment items are classified into rather specific product groups by the sports equipment chain itself, and by making use of this information together with brand and store concept we were able to create rather detailed item clusters. Nevertheless, the level of detail was not sufficient; the clusters were at times subject to unit value bias. Including *price* as a variable representing the quality, we were able to better create defined homogenous clusters. Using the price as a measure of quality is rather controversial as quality should be measured by the characteristics of an item. However, one reason for including price as a variable is because we saw that the various brands were frequently represented in both the higher and lower quality levels of product groups. Brands typically known for producing high quality items were also offering items of less complexity, i.e. poorer quality. Likewise, brands typically known for poorer quality were also seen offering items of higher quality. For this reason, we needed more information than just the brand and the store concept to create well defined homogenous item clusters. An example of sleeping bags illustrates the difficulties; while the higher quality brand offer items at higher price levels, they also offer some lower quality and/or less complex items. If we were to equate the simpler sleeping bag meant for backyard sleeping in the summer with a sleeping bag meant for a winter expedition as a result of both being sold in the same store concept, both are of the same brand and both belong in the “sleeping bag category”, it would

certainly cause unit value bias. Thus, as an attempt at separating the higher and lower quality items we found it useful to also include price, as an explanatory variable in the creation of homogenous item clusters.

The price variable is only used in the case of large price dispersion within the formerly defined clusters, which was generated by combining product group, brand and store concept. I.e. if the price variance within the initially created cluster is negligible, only one cluster will be created. To allow for as much variation as possible without distorting the calculations we decided on using three price levels within the clusters; low price, average price and higher priced items. It should be noted that in the actual calculations only a minor part of the data is subject to the additional price variable when creating the item clusters; most item clusters are not based on the additional price variable classification.

The creation, of clusters are performed in the following order; first, the clusters are created based on the combinations of the variables *the store concept*, *the brand*, and the chains own classification of *product group*. Further, if the price variation within the cluster is large, the item clusters is then separated into the three different price categories; low, medium and high.

Figure 4: Illustration of homogenous item cluster creation



Treatment of new items

One of the ideas behind using item clusters as opposed to item codes is to be able to capture new items that are entering the market, hence being able to replace items no longer on the market. New items entering the market will be included in the calculation the first month they enter, as long as they meet the criteria explained above. New items will be classified into an existing cluster according to the combination of the store concept, brand, item group and then the possible price cluster. Given the initial clustering based on store concept, brand and product group, and there being more than one price cluster, the new item(s) will enter the price cluster which shows the least deviation between the initial price of the item and the average price of the items within the price cluster. New items should serve the same utility in a consumer perspective and be of the same quality as the items within the clusters, and in effect replace items that are going out of stock. If a new item has a different utility, i.e. be of a different quality or serve a different purpose, the item will not be included in any of the cluster but should instead be treated as a possible new cluster. As of today, the clusters are fixed for one year, thus no new clusters are being generated throughout the year. If a new item does not meet the criteria it will be omitted from the calculation.

It should be noted that compared to the item code level, the item(s) represented as clusters are rather stable throughout the year, meaning that there is less desertion of item clusters throughout the year than what would have been the case if we were to use item codes. This supports the use of a fixed basket of clusters throughout the year.

Combining data sources

Since the scanner data is from one single sports equipment chain only, the data needs to be combined with other data sources for the indices in question. In this case it needs to be combined with data from (web) questionnaires provided to a sample of retailers[7]. Given the amount of data, scanner data can easily overshadow the data from other retailers if not adequately treated. In addition, the scanner data received is aggregated to store concept, while the web questionnaire data is at store level. Several options were discussed for combining the price observations from the two data sources together. The solution in place is to treat data from the questionnaires, using the traditional representative items[8], and the scanner data, represented by item clusters, separately, before combining them at a more aggregated level, i.e. COICOP 6-level.

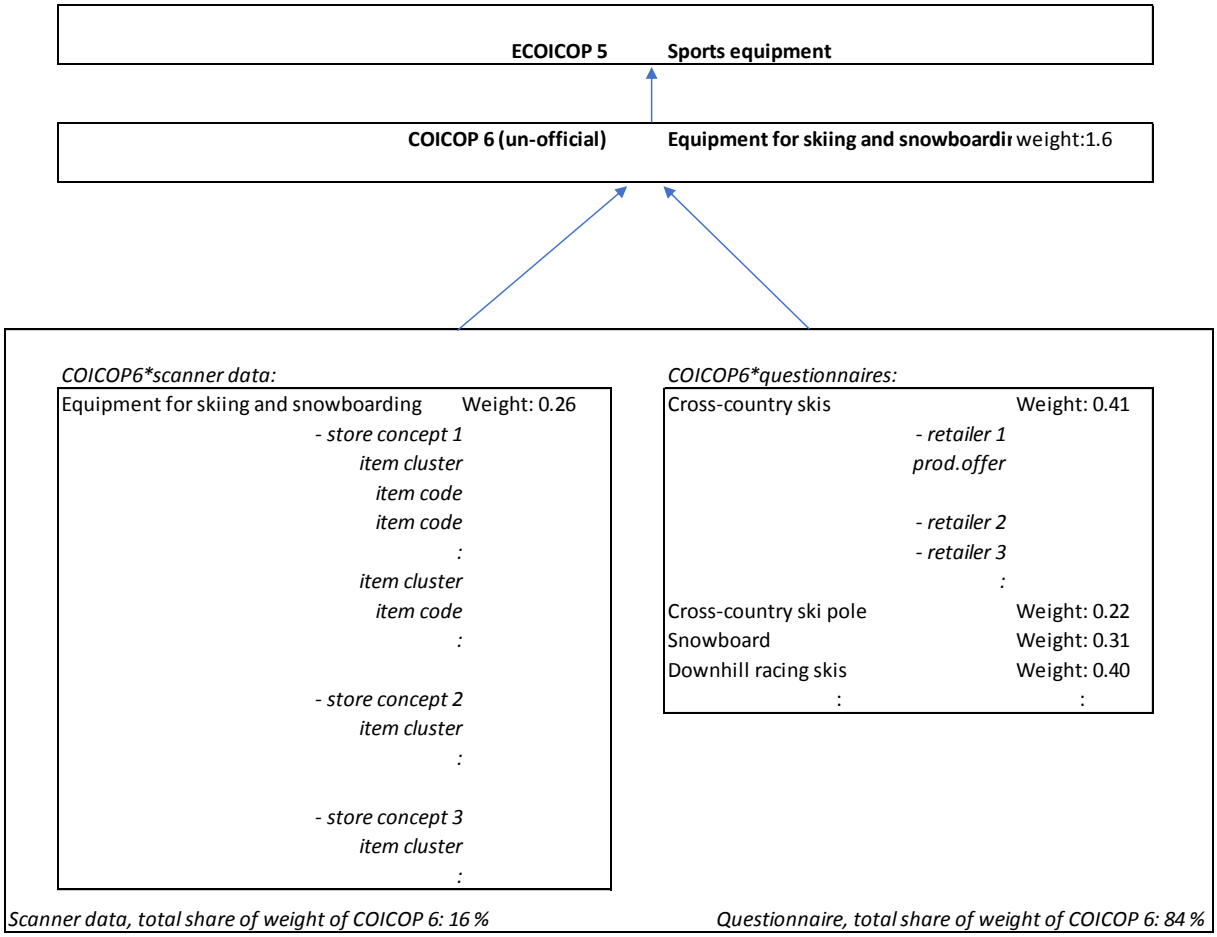
Data from the questionnaires are organized into several representative items, each with their corresponding representative item weight. Examples illustrated below; cross-country skis, ski pole, snowboard, down hill racing skis and the like. Scanner data is organized into a more generally defined group of items, containing the corresponding items (clusters). This gives COICOP 6 groups that consist of indices based on either scanner data or web questionnaires and is illustrated in figure 5 as COICOP 6*data source, e.g. COICOP 6* scanner data, COICOP 6* questionnaires.

[7] Note that this is the solution to combining scanner data from one singular sports equipment chain together with (web) questionnaire data and should not necessarily be treated as a general solution to combining data sources. We recognize that if more sports equipment chains, or clothing or footwear chains, began delivering data we would have to revise the method and create a more permanent solution.

[8] A representative item in this context is centrally specified items that make up the basket of consumer goods and services in the Norwegian CPI.

The weight of the COICOP 6 * scanner data is created based on information from both the Wholesale and retail trade sales statistics and the Wholesale and retail trade, breakdown of turnover by product statistics, both from Statistics Norway. The weight of the COICOP 6*questionnaires is simply the sum of the weights of each representative item within the specific COICOP 6 group.

Figure 5: Illustration of combining data sources



Calculation method

The calculation formula chosen for the sports equipment scanner data is the un-weighted Jevons formula at the lowest level, however using item clusters as the unique identifier instead of the item code as traditionally used in scanner data calculations. Since using item clusters can be seen as a mimic of using item codes, however including the benefit of being able to handle replacements, we found that the Jevons formula was a valid first step of implementing scanner data for sports equipment. Note that the Jevons formula is used for the item clusters within the store concept; the store concept item clusters are then weighted together with yearly fixed store concept weights. This can be viewed as using geographical weights for the treatment of representative items from the questionnaires. The Laspeyres formula is used for further aggregation.

Using equal weighting at cluster level within each store concept, marginal items will be in position to cause great impact on the indices. The reason being that scanner data contains data for all items

being sold, no matter the quantities. To offset this issue using cut-off thresholds are usually the solution to exclude marginal items from the calculations. As shown below, looking at one specific month, we find that for most product groups, about 20-30 per cent of the item clusters cover about 80 per cent of the total turnover. This means that omitting about 70 per cent of the clusters in the same month would give us price information of 90 per cent of the total turnover. As shown in the table below, the amount of price clusters needed to cover 90 per cent of the turnover varies across product groups (represented as COICOP6 * scanner data), thus the cut-off threshold will result in various amounts of clusters kept/excluded.

Table 2: selection of COICOP6*data source groups and the corresponding amount of item clusters needed to cover 90 per cent of total turnover

COICOP 6 * data source	Amount of item clusters per COICOP6 * data source	Share of item clusters needed to cover 90% of total turnover
Socks, men	486	21,4 %
Trousers, men	826	29,7 %
Pullover, men	540	24,3 %
Jackets, men	580	29,8 %
Undergarments, men	426	23,2 %
Trouser, women	650	26,0 %
Pullover, women	589	24,4 %
Jackets, women	614	28,8 %
Undergraments etc., women	591	17,6 %
Trousers, kids	542	31,0 %
Jackets, kids	436	36,0 %
Hatts, mittons and the like	752	22,5 %
Boots, men	161	34,8 %
Sneaker shoes, men	321	24,6 %
Bicycles	157	55,4 %
Backpacks and bags	567	36,7 %

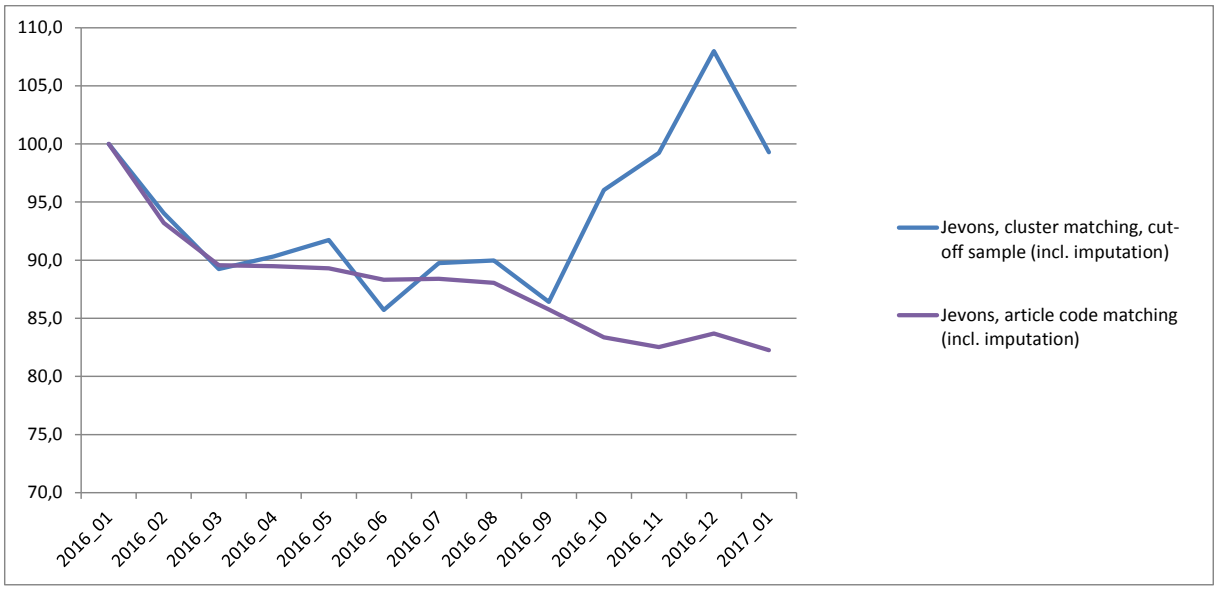
For practical purposes the basket of item clusters is fixed for one year. Test calculations show stability of clusters throughout the year, supporting the decision to using a fixed basket of item clusters. Since new item clusters are not generated throughout the year, we risk not capturing for example a new brand, or a new type of product, until the year after. This however is no different than the using yearly fixed weights in the general CPI. Still, given that we are using a fixed basket of item clusters throughout the year, and because this is a newly developed method, there is created a control in the production system that monitors the expenditure shares of the item clusters within product groups. If the expenditure shares were to drop drastically we would need to make alterations to keep the sample of item clusters representative.

Using homogenous item clusters that contain one or more unique items, the clusters are less vulnerable for temporarily missing items, thus less subject to imputation. However, there is still need

for price imputation if, and when, a situation of missing item (clusters) occurs. Since we are using clusters as the unique identifier, the price observations within the cluster will not be imputed, however the unit price of the cluster itself will be imputed if there are no price observations. A missing unit value price at cluster level will be imputed using the price movement of the other clusters within the same store concept within the item group. If neither of the clusters in the item group contains price observations, the cluster will then be imputed using an aggregation of COICOP*store concept. This is similar to how missing items generally are being treated in the CPI, the main difference being that the item cluster itself and the corresponding unit value is the unique identifier as opposed to one single item. However, calculations have shown an imputation ratio of about 5-6 per cent of the price data, well below the average imputation ratio seen for, for example the indices of food and non-alcoholic beverages.

Test calculations of using item clusters show that this method manage to capture the replacement items. The matched model index on item code level on the other hand, shows the expected downward trend.

Figure 6: Jacket's men. Calculations methods. Indices. January 2016=100.



Summary

Technical improvements and digitalization has created new opportunities for price statistic. New data sources such as scanner data contains significantly larger amounts of data and metadata than the more traditional data collection method of using questionnaires. It is in both Statistics Norway as well as the data providers' interest to replace the questionnaire-based data collection by scanner data. However, both the product life cycles and the price movements over the life cycles vary across consumer goods, thus challenging the traditional methods of using scanner data in the Norwegian CPI. A weakness of the current method is that there are no direct replacement of items leaving the market, instead various filters has been used to counterbalance the lack of a replacement strategy. It is evident that the nature of the sports equipment scanner data made it troublesome to apply the same methodology as previously used, much as a result of the product life cycle generally proving

short. Alternative methods that would allow for (automated) replacements were therefore sought. The quantity of data requires that a replacement strategy must be automated for it to be able to maintain. The solution to the replacement difficulties was to use homogenous item clusters, and the corresponding unit value of the clusters. The homogenous item clusters allow for items to both exit and enter the clusters throughout the time period, making the calculations less vulnerable for missing items and indirectly allowing for replacement. The implemented method is thought of as a cautious approach to using scanner data for goods identified by short product life cycles and rapid replacements. New calculation methods, such as for instance multilateral methods, might undertake scanner data calculation methods in the future. In the meantime, by using homogenous item clusters we were able to implement scanner data for the sports equipment chain as of January 2018.