Statistics Denmark,
Prices and Consumption

Jonas Mikkelsen
JOM@DST.dk

Extending the Danish CPI with scanner data - A stepwise analysis

Introduction

In 2011 Statistics Denmark (DST) got access to scanner data from the largest Danish supermarket chains. The main focus has been on understanding how scanner data can be incorporated into the current CPI production systems and processes. Scanner data is different from the information receive in the regular price collection. This paper focuses on these differences presenting a stepwise analytical approach moving from traditional price collection to a scanner data based CPI.

The first part of this paper focuses on the change of price concept that is inevitable when using scanner data. In order to analyse the isolated effect from this change the current product basket is identified in the scanner data. In this way the collected shelf prices have been exchanged with prices from scanner data defined by weekly unit prices.

The second part of this paper is concerning the definition of the item basket. The goal is to generate a representative and manageable basket that can be used in the current CPI production system. Using the turnover data makes it possible to obtain a set of representative products for each COICOP group. But high attrition rates of items require efforts when dealing with missing prices when items leave the scanner data on discount.

Creating a manageable basket is a challenge both regarding the large amount of prices and regarding the attrition of items. In this paper it is analysed whether a longer data collection period can limit the problems with missing values. Secondly, limiting the amount of prices in the basket is attempted by aggregating store specific prices to chain specific prices. Thirdly, the effect of reducing the number of products in the basket is investigated. This is carried out by limiting the basket only to include the top 3 bestselling product for each COICOP 8-digit category.

The data

Since January 2011 Statistics Denmark (DST) has received scanner data from the largest supermarket chains on a weekly basis accounting for approximately 60% of the Danish sales of food and beverages.

The product ID

In scanner data each product is identified by a product code either called European Article Numbering (EAN) or Product Lookup Code (PLU). The EAN number is defined by the producer and normally has 13 digits while the PLU number is often defined by the supermarket chain and often has less than 13 digits. When working with scanner data the EAN/PLU[1] number is used as the product identification which is used to secure matched prices.

Example of scanner data from the supermarkets

Date	Store	EAN number	Turn-over	Vol-ume	Unit	Quantity per unit	Product number	Product description
1104	7894	2920080800007	3402,70	211	Gram	300	910076003	Sliced bacon 2x150 G.
1104	7895	2920080800007	2119,65	163	Gram	300	910076003	Sliced bacon 2x150 G.
1104	7896	2920080800007	1516,05	108	Gram	300	910076003	Sliced bacon 2x150 G.
1104	7897	2920080800007	1478,13	105	Gram	300	910076003	Sliced bacon 2x150 G.
1104	7214	2921056000005	302,50	14	Gram	200	911056001	Chicken Fillet
1104	7215	2921056000005	102,50	5	Gram	200	911056001	Chicken Fillet

[1] Hereafter EAN/ PLUs are referred to as EAN.

The *date* field consists of a 2 digit year number and a 2 digit week number. The *store number* is unique for the specific supermarket store. From this number it is also possible to identify the supermarket chain it belongs to. The *prices* are derived from dividing the weekly turnover with the weekly volume for each EAN number for each store. The *quantity per unit* is often defining the size of the product in grams or kilograms. Unfortunately, it sometimes holds the information e.g. “10 apples” which is not a precise measure of the actual size of the product. The *product number* is very important as it links to the product hierarchy of the supermarket chain. This product hierarchy is indispensable when linking the EAN number to the COICOP. For each EAN there is a *product description* created by the supermarket chain.

The price concept

One important difference between traditional price collection and scanner data is the price concept. In traditional price collection you are dealing with shelf prices. In scanner data the price is a unit price calculated from turnover and sales volumes. Because of the huge amount of data that is generated in a supermarket each day we receive scanner data aggregated on a weekly basis. Scanner data consists of weekly aggregated turnover and volume for each specific EAN number for each store in the chain. One of the shortcomings with the weekly aggregation is that it is not possible to identify volume discounts e.g. 3 packages for 2’s price in the data. The price will be a mix of both single packages and volume discounts. The derived challenge with volume discount prices is that the matched item criteria may not be entirely met. A single package is not strictly the same product as a volume discount item of 3 packages.

Scanner data volatility

Another challenge of using scanner data for CPI purposes is the ability to continuously monitoring products over time. Every week there is a number of products entering or leaving the supermarket store. Many products have a short life cycle but also changes in the product packaging result in new EAN numbers. Consequently, there are many missing prices through the year. These missing prices can create a bias in the indices if they are not dealt with properly.

Possible bias deriving from the scanner data volatility

It is possible that the scanner data based CPI has a downwards or an upwards bias when either new products enter the item basket or when products are leaving the item basket.

We have observed biases in two ways:

- A product enters the basket on discount the first month and is then set to the normal price in the months after. This leads to an artificial increase in the index that will not get levelled out.
- A product leaves the basket on discount. This leads to a persistent decrease in the index.

When a large proportion of the above incidents happen the bias becomes problematic in the indices over time. The upwards or downwards biases often happen when the products are temporarily out of the scanner data. The many missing prices in data series are impossible to foresee and therefore must be dealt with ex post. If a product is temporarily out of the scanner data or persistently out of the scanner data the decisions on imputation then becomes important. The problem with imputing is that it can lead to a bias towards index 100. Historically, Statistics Denmark has treated missing prices manually securing the imputation bias is minimized. If a price is missing in month t0 (p(t0)) the previous month’s p(t-1) is evaluated with the p(t-2). If p(t-1) is different from the p(t-2) the missing price p(t0) is set to p(t-2). If the product price is also missing in the future months it is replaced by a new representative item. In the implementation of scanner data in the CPI it is attempted to continue processing the missing prices as described above.

The analysis

In the stepwise analysis it is emphasised to change as little as possible in each step in order to isolate the effect of each change. This paper does not deal with analysing effects from imputations or effects from filters.

The current production flow

In appendix A the current monthly production flow is shown together with the new scanner data processes. The current production process allows use of 2 weeks scanner data per month. Scanner data introduces new processes such as maintenance of the key between the COICOP and the EANs and maintenance of the item basket securing a suitable coverage of the total turnover. When introducing the scanner data it is important that these new processes can fit into the time frame of the production window of less than 2 weeks.

The on-going product update

The current production includes an on-going product update procedure. When a product is not sold by the supermarket chain it is changed to a new similar product either by DST or the supermarket. Traditionally this has been emphasised in order to allow for an item basket that is aligned with the actual sales in supermarkets and to minimize the use of imputations.

The incremental steps of the analysis

The basis for all analyses in this paper is the currently used index formula on elementary aggregate level – a weighted Jevons index. We have separated the analysis in two main steps going from the current data collection based CPI to the scanner data based CPI. Additional three steps are carried out attempting to limit issues with missing prices and to optimise the manageability of the scanner data in the production processes. The analytical steps are incremental which means that the parameters that seems to work are carried over to the next analytical step:

1. The change of price concept – CPI where products from the current publish CPI index are found in the scanner data and replaced in our current CPI production system.
2. The change of product selection.
3. The change of data collection period.
4. Chain aggregation.
5. Limiting the size of the item basket.

The products analysed in the paper

Five different COICOP groups each with different characteristics have been chosen for the analysis. In the current CPI production the chosen product prices are collected in different ways. These differences are important when comparing the current CPI with the scanner based indices. The current collection methods are described below:

1. **Rice** – presumably a generic product and price stable category.
   - *A shelf price list is provided by supermarkets.*
   - *DST chooses the rice products from the list.*
   - *The chosen rice types are based on the Household Budget Survey (HBS).*

2. **Red wine** – is typically bought on discount.
   - *A shelf price list is provided by supermarkets.*
   - *DST chooses the red wine products from the list and also collects some product prices from large wine dealers on the internet (all non-bag-in-box).*
   - *The red wines chosen from the supermarket are mainly from France, Italy and Spain whereas the products from the internet are collected without country specifications.*

3. **Coffee** – is based on 1 kg product.
   - *Products are chosen by price collectors*
   - *The coffee product prices are based on 1 kg.*
   - *The coffee products are collected by price collectors in the store. The price collectors choose products having the most shelf space.*

4. **Minced beef** – a product often bought on volume discount.
   - *The minced beef product prices are based on 1 kg.*
   - *The minced beef products are collected by price collectors in the store.*
   - *The price collectors choose the products based on the content of fat. Products with less than 15% fat are included.*

5. **Apples** – a seasonal good.
   - *The apple product prices are based on 1 kg. There is a mix of prices from one 1 kg bags and single apples that are weighted and recalculated to represent 1 kg prices.*
   - *The apple products are collected by price collectors in the store.*

In the current price collection new similar products are included in the basket when earlier products disappear. In all cases except coffee the price collection includes non-volume discounts. The differences in the data collection methods are essential when comparing the current published indices to the scanner data based indices.

In order to limit the size of this paper only the main trends from the analyses will be discussed in the chapters. The rest of the indices can be found in appendix C.

The current published CPI

All analyses are based on our current CPI index formula on elementary aggregate level. This is at the COICOP 8-digit level (hereafter referred to as c8 level). The index is calculated as described below:

$$(1)\qquad I_{0:t}^{Jv} = \prod \left(\frac{p_t^i}{p_0^i} \right)^{w^i} = \frac{\prod \left(p_t^i \right)^{w^i}}{\prod \left(p_0^i \right)^{w^i}}, \sum w^i = 1$$

The index formula is a weighted Jevons. The weighting is made on two levels; on the COICOP 8-digit level and at the store level. The c8 level weights are based on the household budget survey and the COICOP specific store weights are based on the supermarkets turnover. It has been chosen to use the same store weights on c8 level as in the current published CPI in order to make the stepwise analysis as "clear" as possible. The specific calculations for the store prices, product prices, and basis prices can be found in appendix B.

1. The change of price concept

In this analysis the question to be answered is:

- *"Are the current CPIs on elementary level identical with indices based on scanner data prices covering the exact same items from the exact same stores?"*

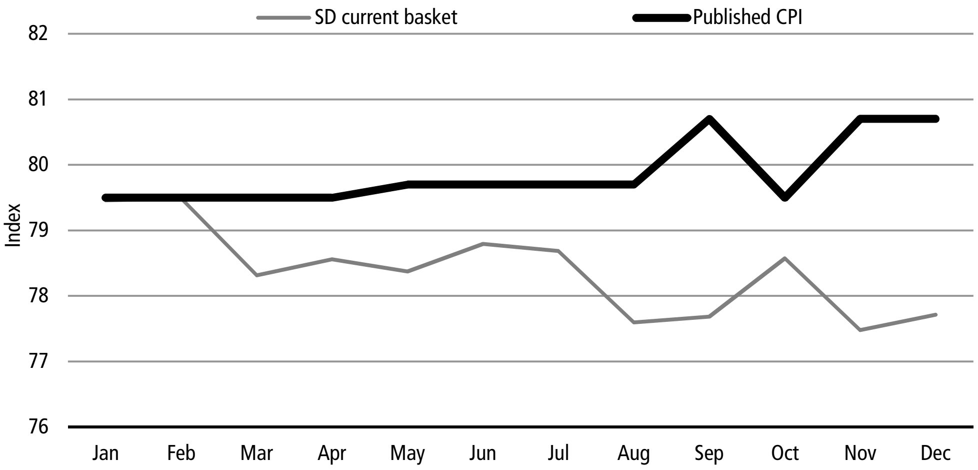
Matched item criteria

One obvious comment regarding the change in price concept is that a product from the scanner data will have a weekly average price which could level out the weekly price changes. This could be a problem since volume discounts would be mixed with non-discount within the same EAN code which implicate that the "matched item" criteria may not be entirely met.

Identifying the current item basket in the scanner data

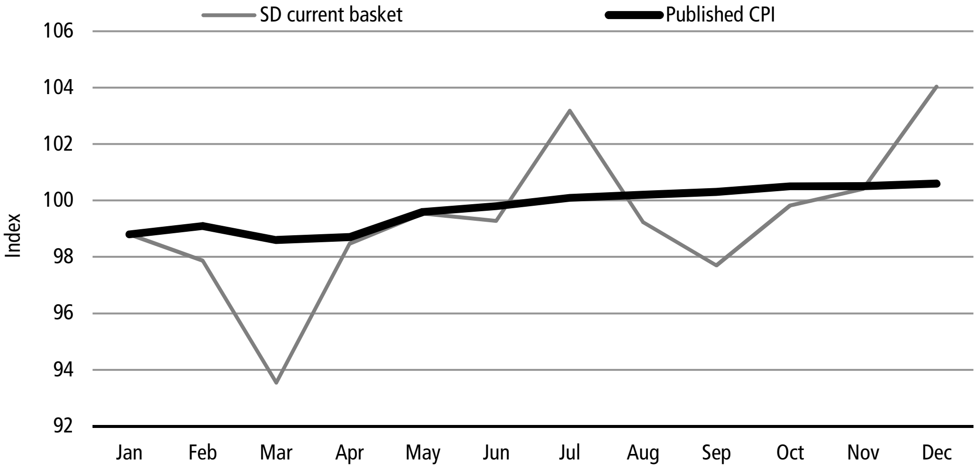
It was expected that rice products would be easy to identify when picking the current item baskets rice products in the scanner data. However, it was challenging to identify the exact same rice products as there is a huge variety of products. The actual rice market is differentiated with a very wide price span. Since the scanner data item descriptions are limited the method of just picking the current basket in the scanner data is time consuming. In general the item from the current basket has been picked in the exact same store in the exact same weeks from the scanner data. When there were many possible matching products we chose the ones that had the best price match. In this step of the analysis only rice, red wine and coffee has been analysed due to the time-consuming task of manually identifying the exact current basket items.

The Consumer Price Index on Rice from January to December 2011



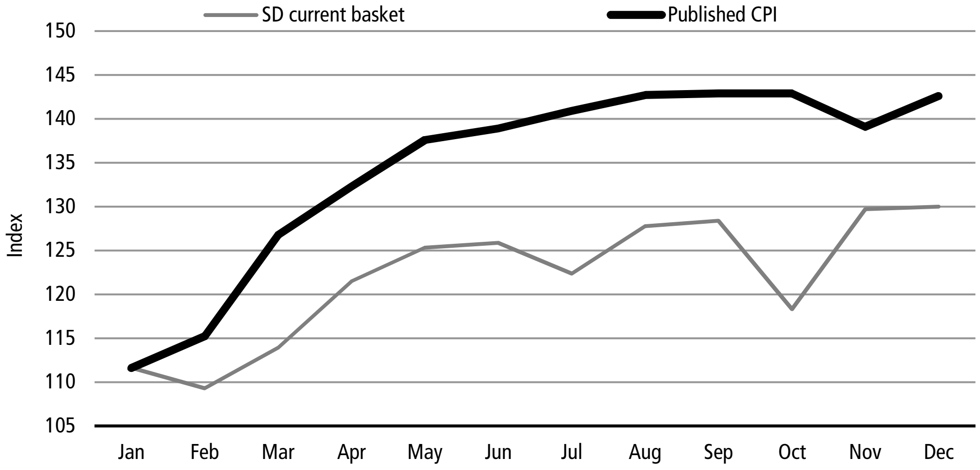
The published CPI on rice is relatively stable whereas the current basket picked in scanner data (SD current basket) on rice is more volatile. This is plausible since the current collected prices are shelf prices and the scanner data consists of actual sales prices that are per definition more volatile. The "SD current basket" seems to have a slightly negative trend whereas the published CPI is slightly increasing through 2011. We expect that the trends will show a pattern close to each other on a longer time frame.

The Consumer Price Index on Red Wine from January to December 2011



The published CPI on red wine has a stable slightly increasing trend through 2011 whereas the SD current basket on red wine is again more volatile. The "SD current basket" seems to be volatile around the "Published CPI" and therefore seems to have similar overall trends as expected. The main difference between the two sets of prices is that the scanner data include volume discounts whereas the "Published CPI" only includes prices for 1 bottle. It is currently not possible to separate non-volume discounts in the received scanner data.

The Consumer Price Index on Coffee from January to December 2011



Besides the increased volatility in the scanner data based coffee index the overall trend is diverging more extensively than the indices on rice and red wine. This divergence may be based on a change in consumer preferences when the general coffee prices went up in the first half of 2011. As the published CPI is based on prices on the shelf the price index has a clear increase whereas the weekly turnover based price index from the scanner data reflects a change on consumer preference towards buying a larger amount of coffee on discount.

Conclusion

In general the different price concept linked to scanner data introduces more volatility to the indices mainly due to discounts. In the short term the fluctuations can be large whereas the trend is expected to be similar on the longer term.

With product groups of rice or red wine the trend may be close to the published CPI by picking the same products in scanner data. However, large differences seem to appear when consumer product preferences change because of large price changes. Hence, a change of price concept introduces new index behaviours which are in reality closer to the actual market situation experienced by the consumer.

2. The change of product selection

Introducing scanner data in the CPI production eliminates the current feedback from the supermarket chains regarding the product selection. Scanner data presumably offers a superior way of selecting a proper item basket. The basket will reflect what is actually bought by the consumers when using the turnover from the scanner data. A consequence, however, is that the amount of selectable products is huge and the ways of defining the basket are many. The question to be answered in this step of the analysis is:

"What is the effect on the indices when defining a representative item basket by the scanner data turnover?"

All the presented analyses in this chapter will have a flexible item basket. It is attempted to benchmark how different selection parameters impact on the indices. The goal is to define a durable item basket that can be handled in the current production system. In the first part of the analysis only 1 week of data per month (1w/m) is used.

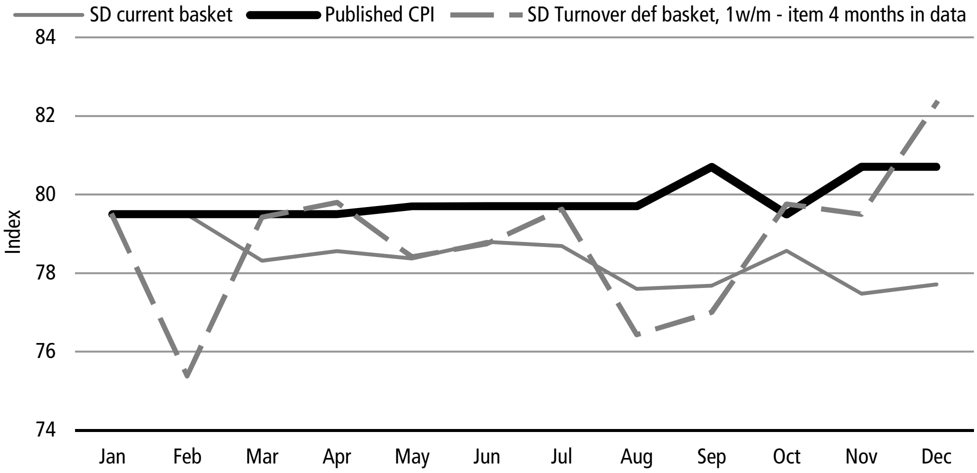
The basket is defined from month to month with the following criteria:

- The items must have been sold in the 4 previous months.
- The turnover regarded for each specific item in each store consists of 4 months aggregated turnover.

- The items selected for the item basket add up to 50% of the COICOP 8-digit turnover for each supermarket chain[2].

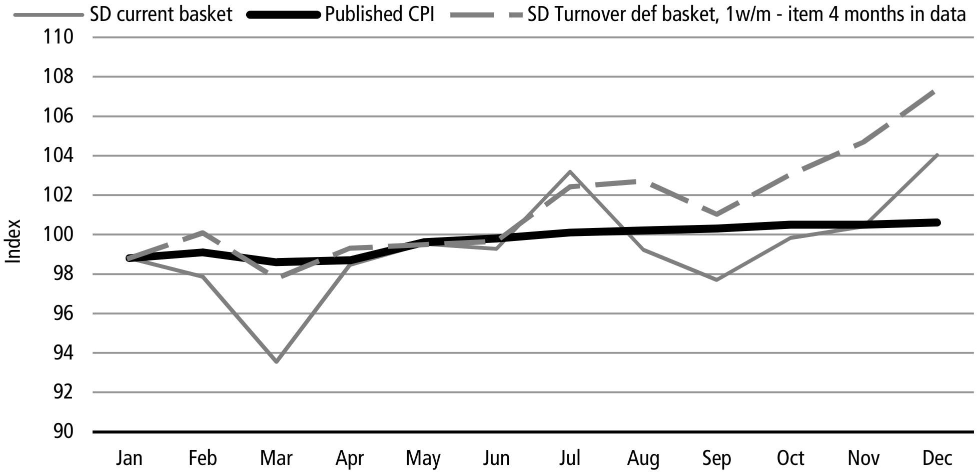
In the index calculation each supermarket chain on the COICOP 8-digit level is weighted securing that a supermarket chains with lower turnover have less impact on the index than stores with higher turnover[3].

The Consumer Price Index on Rice from January to December 2011



With the turnover defined item basket of rice (SD turnover def basket) the volatility is large especially in February and August. These extreme decreases are due to discounts. The longer term trend seems to go back around the published CPI whereas the current basket from scanner data (SD current basket) shows a downwards trend. The index with scanner data defined basket is in general more volatile because of discounts that are not captured in the shelf prices.

The Consumer Price Index on Red Wine from January to December 2011



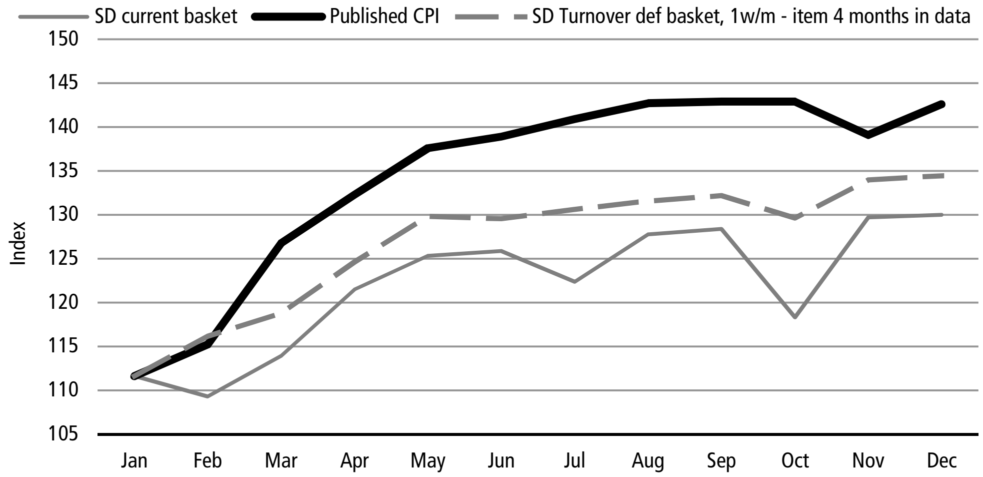
The turnover based basket on red wine is volatile as the current basket in scanner data. However, the last 4 month of 2011 it has a larger increase. This increase stems

[2] When defining the item basket the first goal was to be able to select the most important product within the COICOP group. This has been done with a selection identifying the product with the highest turnover of all the supermarket chains within the COICOP group on the 6-digit level. A derived challenge from this method is that some supermarket chains are excluded from the data series. In some cases a supermarket chain had discount on a product one month resulting in very large sales excluding the items in other supermarket chains. The next month another supermarket chain would have a discount resulting in an item basket with a very volatile selection of supermarket chains. The solution was to select the items with top 50% turnover within the COICOP 6-digit level within each supermarket chain.

[3] See appendix B for further information

from wines that are sold on a lower introductory price that in a later period is sold on a normal price before it again leaves the data (is out of stock). Looking into the micro data there are a lot of data series only having 3-5 months of sales in each store. The item exits combined with price increases in the month before results in an upwards bias.

The Consumer Price Index on Coffee from January to December 2011



The turnover based coffee index is closer to the published CPI than the “SD current basket”. The difference between the turnover based index and the “SD current basket” is the product selection. The coffee products with high sales turnover are different from the selected coffee products in the current basket from the published index. One of the main reasons is that the current collection of coffee is only based on single packages and not volume or bundled coffee packages.

The published coffee index is based on 1815 prices in 2011 whereas the turnover based item basket is based on 98,988 prices from all the different stores in the supermarket chains. The number of price observation for each analysis can be found in appendix D.

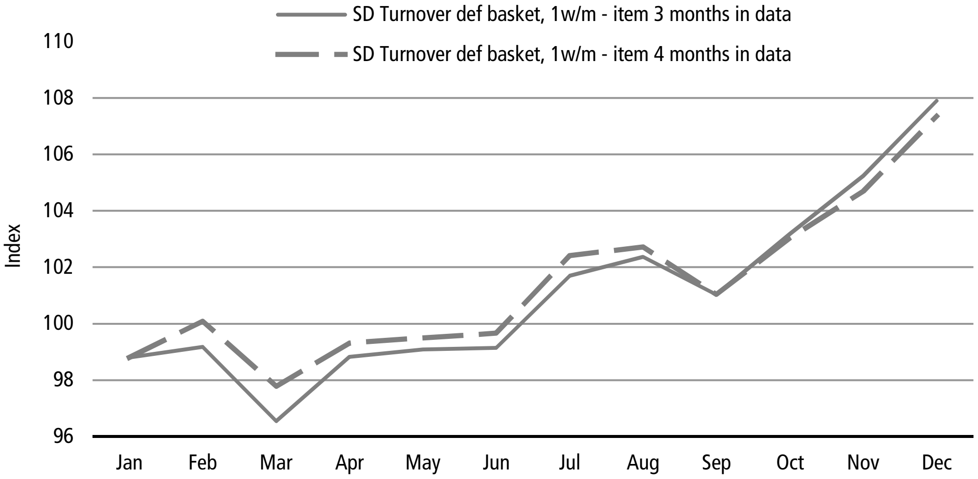
Scanner data volatility and the item basket

In order to secure that the scanner data item basket is more robust to the products entering and leaving the scanner data[4] we have added the criteria that an item in the basket must have been in the scanner data for at least 4 months. In this way products that are sold for a very short time are excluded limiting the problems with bias from ingoing and outgoing products on discount.

Even though the criterion of 4 months in data seems to be adequate it is analysed below what difference a 3 months in data criteria makes. This shows how sensitive the indices are to this type of criteria.

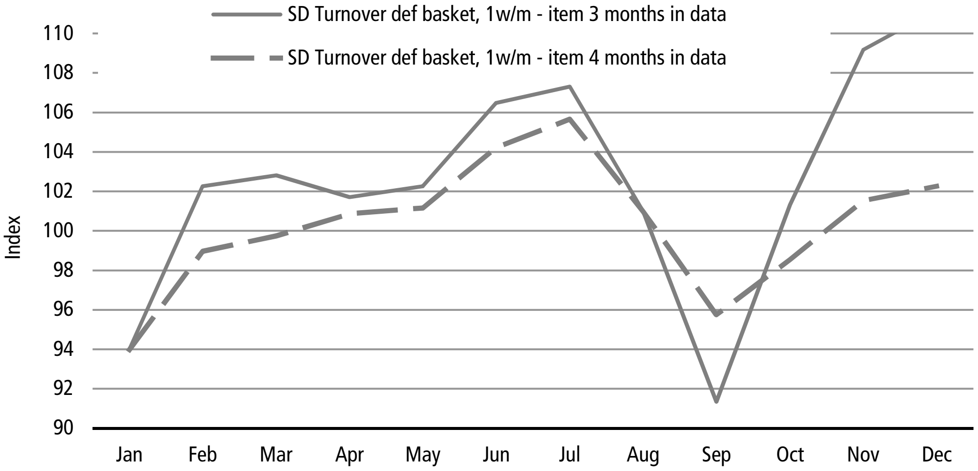
[4] Only 74.3% of the EAN numbers represented in January 2011 are still represented in the scanner data of December 2011.

The Consumer Price Index on Red Wine from January to December 2011



The red wine indices are quite similar. From January to September the index with the 4 months criteria is slightly higher than the index with 3 months criteria. The reason could be that the 4 month criteria lead to fewer products leaving the data on discounts.

The Consumer Price Index on Apples from January to December 2011



In the analysis of apples the 4 month criteria seems to result in less volatility in the index which is mainly because of the apples' seasonality. The most seasonal apples are excluded by the 4 months criterion which is not desirable. When dealing with seasonal items it seems to be best to have a less restrictive criterion than 3 or 4 months in the basket.

In most of the analysed product groups there is little impact on the index by going from the 3 to the 4 months criteria. The indices are slightly less volatile when requiring that the item basket candidates must have been traded at least four months instead of three months. The remaining product groups analysed regarding the 3 or 4 months criteria can be found in appendix C, page 21.

Flexible versus fixed basked item basked

There are advantages and disadvantages with a flexible basket. An advantage is that the monthly basket sample will include many new products entering scanner data every week. The supermarkets introduced 8,783 new EAN numbers during 2011 which is an increase of 32.9%. A disadvantage with the flexible basket is that many new products have a short lifetime which could lead to data series set with many missing prices.

In future works we will attempt to create a system that can support a semi-fixed basket with the opportunity to manually monitor and include new important upcoming products in the item basket.

Conclusion

It is challenging to make an item basket based on scanner data even though the turnover information is available. To answer the question regarding the impact of using turnover information as a selection parameter it generally leads to a different basket of items than the item baskets of the currently published indices.

In order to limit the number of missing prices in the data series it is important to pick products that are sold in a longer time period. The analyses show that the 4 months criteria is better than 3 months regarding missing prices. With non-seasonal goods the index differences are quite small. Seasonal goods, however, should be treated with less restrictive criteria.

The goal is to use a semi-fixed basket so that new important products are included in the CPI sample. Therefore it becomes essential to build a system that keeps track of the current basket and add new important products to the basket.

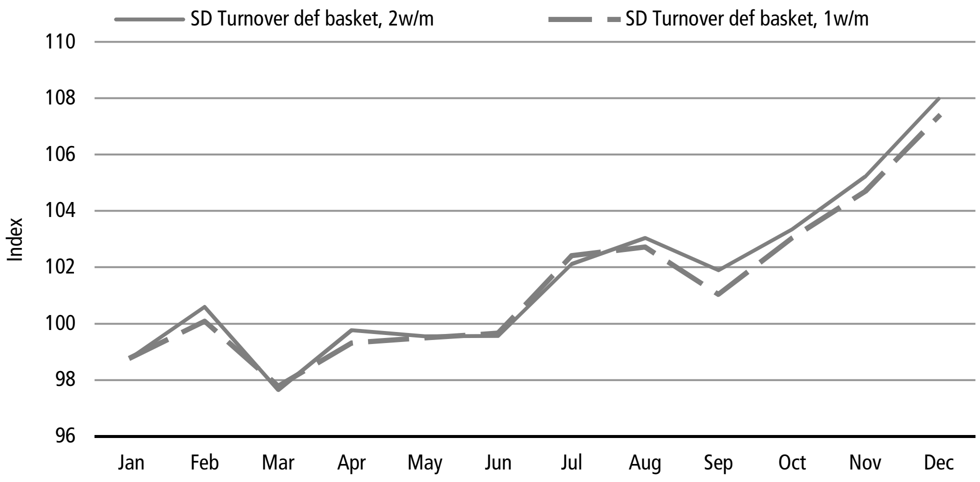
3. The change of data collection period

Using scanner data in the current CPI production flow has restrictions. First of all there is limited time for processing the data. Secondly, the weekly aggregated data are often split between months in the first and the last week of a month often have data that belong to both the previous and the following month. These restrictions allow using 2 weeks of data per month. In this step of the analysis the question is:

"How will change from 1 week of data per month to 2 weeks of data per month impact the indices?"

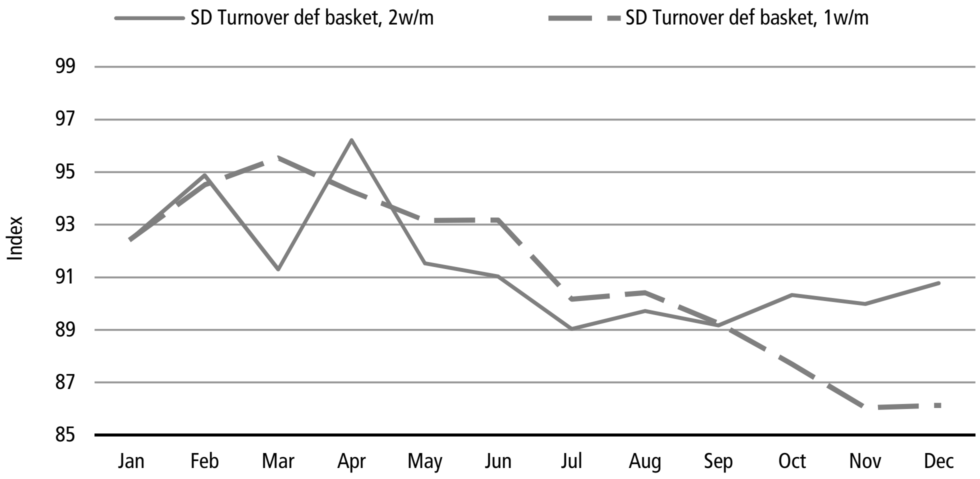
An obvious comment is that 3 weeks per month should be tested as well. Because the current production window is limited it is at this moment not realistic to incorporate 3 weeks. To limit this paper the 3 weeks analysis has been left out.

The Consumer Price Index on Red Wine from January to December 2011



Regarding red wine there is a limited impact on the index by changing from 1 week to 2 weeks of data per month. The reason is that the wines selected for the item basket in the particular month are often sold in the 2 weeks at similar prices. Using 2 weeks of data per month leads to fewer missing prices. One week of data per month includes 353,927 prices of red wine whereas two weeks of data per month include 410,795 prices.

The Consumer Price Index on Minced Beef from January to December 2011



There is a downwards trend in the index of minced beef when only including 1 week of data per month. The main explanation is that there are a lot of missing prices in the data series. When a product temporarily leaves the data on a discount in one week the index will not have the corresponding upwards impact when the product returns to the normal price. Often the item will be sold again in later months at the normal price and then again exit at volume discount.

Changing from one week to two weeks of data per month has a large effect on the minced beef index. Especially in the 4th quarter of 2011 the bias seems to be removed when using 2 weeks of data per month. The reason for this is that two weeks of scanner data per month leads to less missing prices. If the product is out of stock in one of the weeks and it is sold in the other week the product will be represented in the data series. The same trends are seen in the rice and apple indices[5].

Conclusion

In general two weeks of data limit the level of bias stemming from products that leaves the item basket on discount. The effect of increasing the data period is larger with rice, minced beef than with red wine and coffee. To limit the temporarily out of stock problem further and limit the missing prices it would be desirable to use 3 weeks per month which will hopefully be possible in the future.

4. Supermarket chain aggregation

In this step of the analysis the focus is to deal with the missing prices from e.g. temporarily stock outs in the specific store but also to attempt to limit the amount of prices to process in the production flow. Aggregating the supermarket chains for each EAN number could be a solution. The question is therefore:

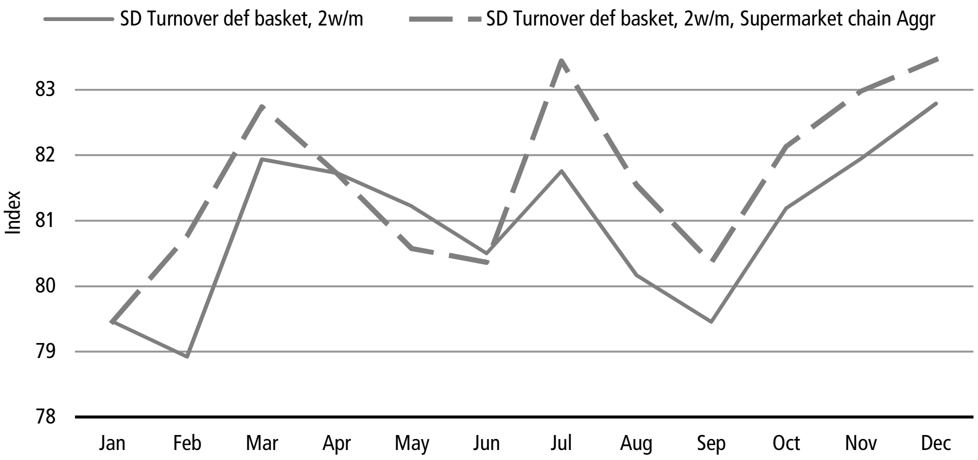
"What happens to the indices when aggregating product prices on the supermarket chain level? Can this aggregation limit the amount of missing prices?"

When aggregating the specific products prices per supermarket chain the number of prices in 2011 is reduced from 83,477 to 536 in the analysis with rice.

The aggregation is carried out by simply summing the turnovers and volumes of all the chains' stores for each EAN.

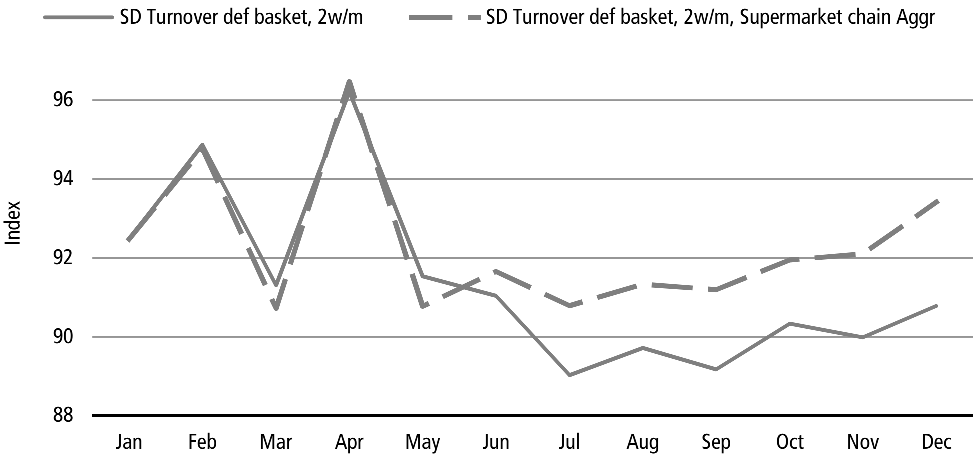
[5] The other indices can be found in appendix C.

The Consumer Price Index on Rice from January to December 2011



The two indices of rice have similar trends, however, the index levels are generally higher when comparing the store aggregated CPI's with the non-store aggregated. This is mainly because the supermarket chain aggregation eliminates many of the missing prices.

The Consumer Price Index on Minced Beef from January to December 2011



The supermarket chain aggregated minced beef index have similar trend as the non-aggregated index. However, the level of the chain aggregated index is higher than the non-store aggregated index. The difference between the levels starting in June is mainly due to difference in missing prices. When the June prices on discount are missing in July a bias is introduced. By aggregating prices on supermarket chain level the bias is limited.

The number of missing prices is especially decreasing in the analyses of minced beef and to a lesser extent with rice, red wine, and coffee.

Conclusion

In general it seems to be a good idea to aggregate each EAN on chain level as it limits the level of missing prices the potential derived bias. Moreover the amount of data to handle is limited which speeds up the performance of the IT systems and the handling of missing prices.

5. Limiting the size of item basket

Using 2 weeks per month of data instead of 1 week most likely leads to better indices and store aggregation secures that missing prices are more seldom. With these parameters it is interesting to observe if a smaller number of items in the basket would result in the same overall index trend. An incentive to limit the size of the item basket even further is the increase efficiency in the production processes.

"What is the impact on the indices when limiting the number of items per COICOP 8-digit group to three?"

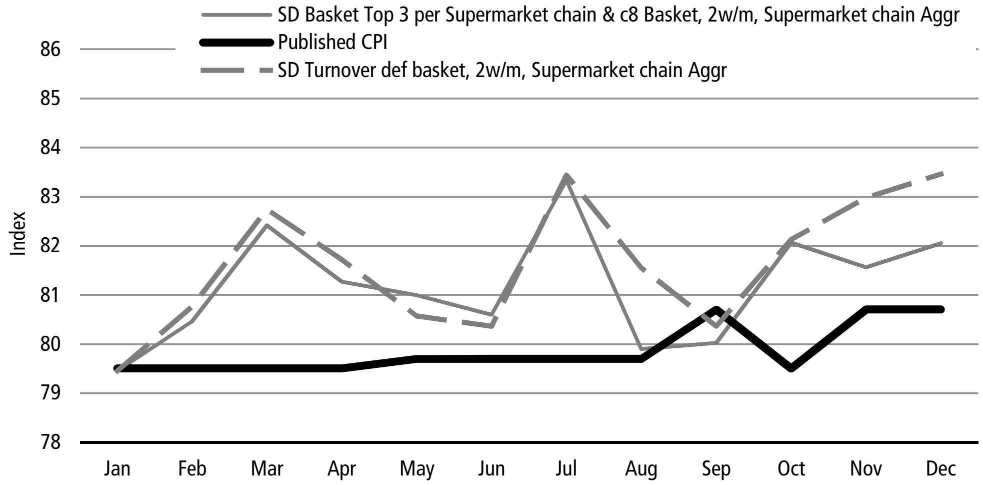
In this analysis the top 3 item indices have the following criteria:

- The data consist of the 2 middle weeks of scanner data per month.
- The items have to be present in the data in previous 4 months to candidate for the item basket.
- The 3 best-selling products in each supermarket chain per COICOP 8-digit level are included in the basket.

The indices named "SD Turnover def basket…" from the earlier analyses have the following criteria:

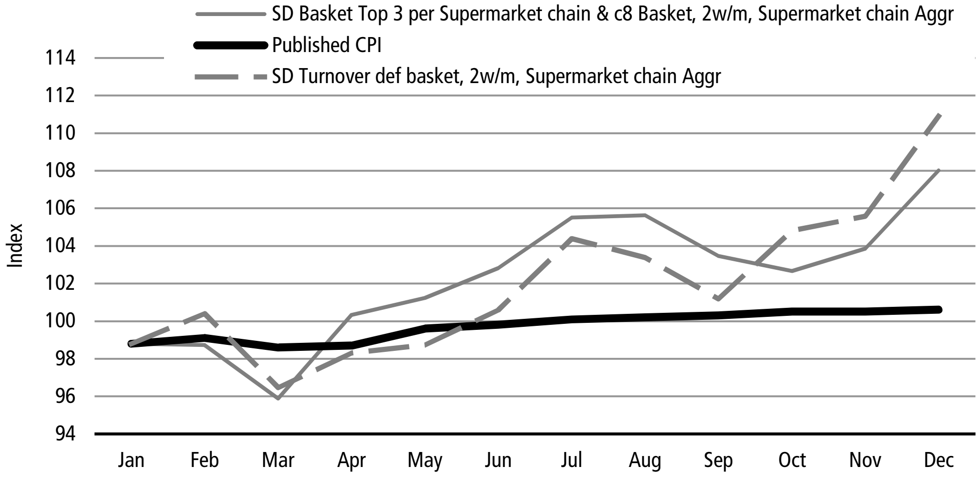
- The data consist of the 2 middle weeks of scanner data per month.
- The items must have been sold in the 4 previous months.
- The items selected for the item basket add up to 50% of the COICOP 8-digit turnover for each supermarket chain.

The Consumer Price Index on Rice from January to December 2011



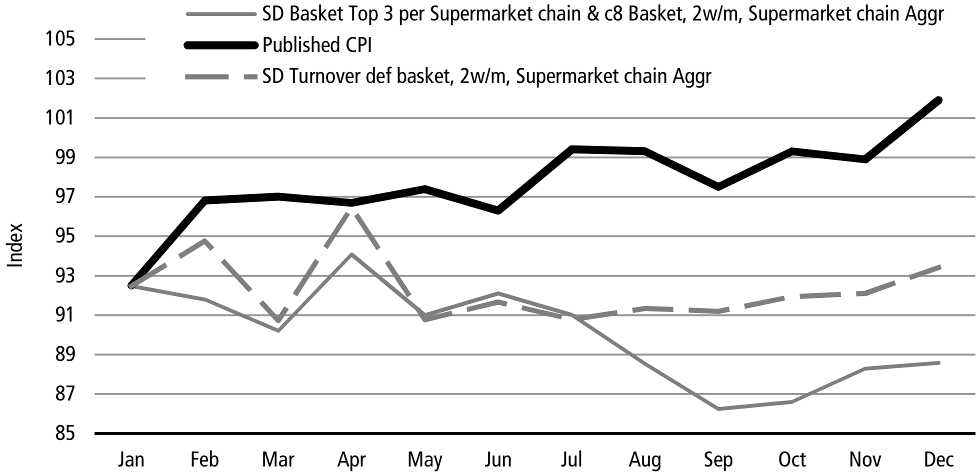
Instead of using 536 prices in 2011 with the 50% turnover criteria the use of top 3 criteria includes 324 prices in 2011. The reduction in the number of observations does not impact the index significantly before August. The differences in August, November and December mainly stems from items on discount. As the published CPI rice index consists of only single packages the price volatility is more limited. The scanner data prices include volume discounts which are the main explanation for the volatility in the indices.

The Consumer Price Index on Red Wine from January to December 2011



Using the top 50% criteria leads to 2,758 prices in 2011 whereas the top 3 criteria include 326 prices in 2011. The scanner data based indices on red wine have similar trends, but a level difference of 1-2 pct. The difference is caused by changes in the product selection and the embedded differences in price developments. The difference between the published CPI on red wine and the scanner data indices is that scanner data contains prices on “bottle in bags” as they have a large sales turnover. The published CPI only contains prices of single bottles of red wine. Hence, the main explanation for the difference is the product selection.

The Consumer Price Index on Minced Beef from January to December 2011



Using the top 3 criteria leads to a lower level minced beef index than the 50% turnover and 4 months in basket criteria. In general the minced beef and apple indices have significantly fewer prices when using the 50% criteria than when using the top 3 selling products. Typically there are 2 products per supermarket chain dominating the sales[6]. Therefore the analyses on these product groups do not fit the originally purpose of limiting the items in the basket. This opens an important discussion whether to set a fixed target on the number of products included in the basket for all COICOP groups. The different COICOP groups have different market dynamics. The amount of products available and their volatility are different from

[6] The minced beef in scanner data consists of 150 products with a total turnover of 388,770,820 DKK in December 2011. By using the 4 month in basket criterion only 74 items are included add up to 192,043,344 DKK of the turnover. When applying the top 3 bestselling products per supermarket chain only 24 items are included adding up to a turnover of 116,447,638 DKK. When applying the 50% turnover per supermarket chain only 13 items are included which account for 86,256,898 DKK.

product group to product group. Ideally the number of products selected for each COICOP group should be decided in relation to the COICOP characteristics. The further work will continue in the direction of defining appropriate selection criteria for the COICOP group.

Conclusion

Rice, Coffee and Red wine have similar trends when either using the top 3 selection or the top 50% turnover. An interesting issue appeared when analysing the minced beef and apple indices because of large changes EANs included in the basket over time. These very volatile data series requires specific inclusion criteria. In general it depends on the COICOP group whether the top 3 products by turnover or 50% turnover criterion are suitable to describe the COICOP group.

Conclusion

Scanner data is indeed different from the information receive in the regular price collection. This paper shows that large differences appear when only changing the price concept. This is partly due to volume discounts inevitably included in the scanner data whereas the current data collection excludes the volume discounts. Another change introduced with scanner data is the index volatility which is a natural element when going from the more stable shelf prices to the actual sales prices. The volatility of scanner data leading to somewhat divergent trends is plausible on the short and medium term whereas the long-term effect of changing price concept should be small. It will be interesting to analyse the long-term trends when two years of scanner data becomes available.

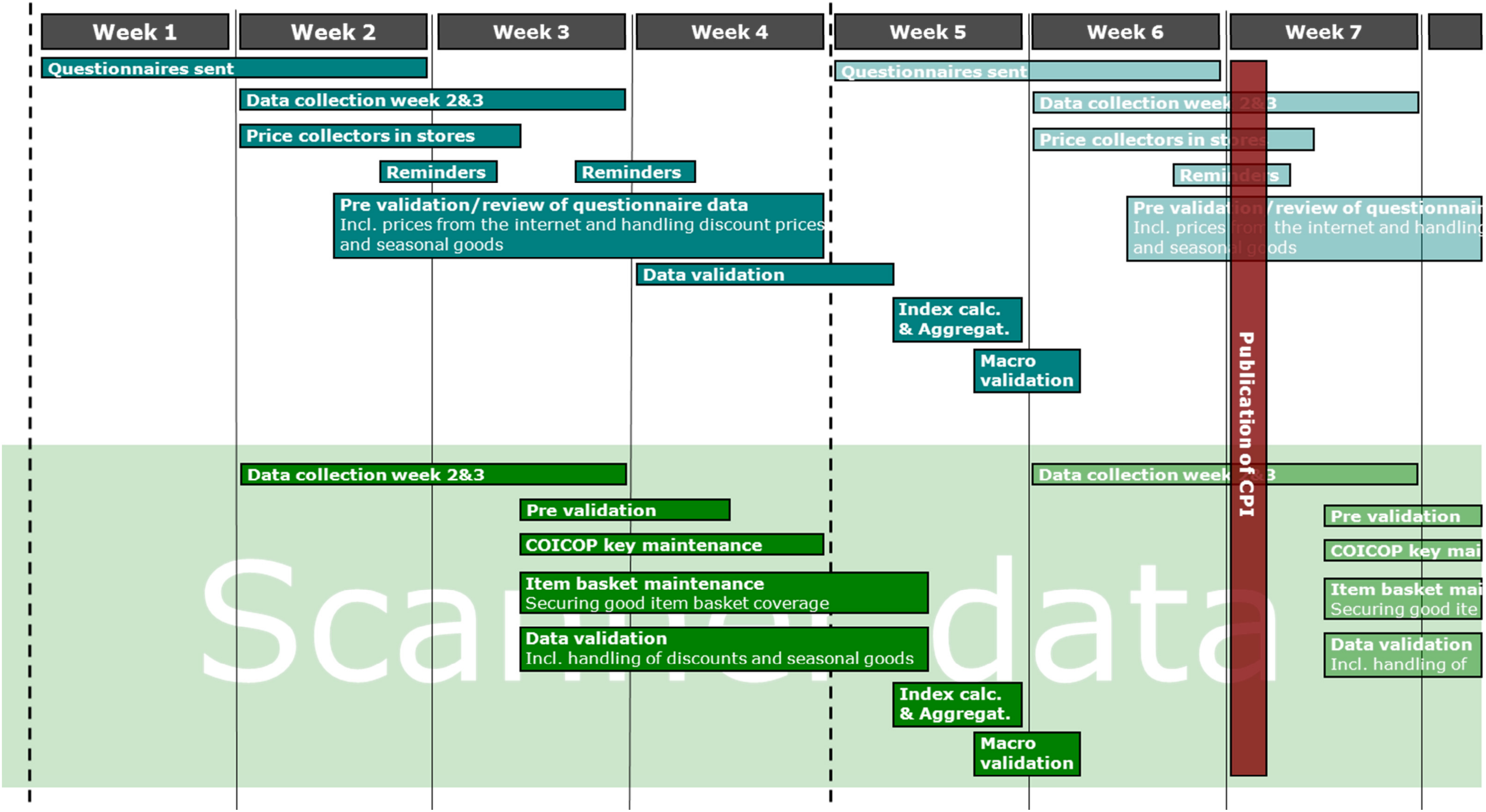
Defining a representative item basket based on turnover information from scanner data is not trivial. It became clear that a fixed basket accompanied with a flexible attributes during the year is desirable. In this paper only the flexible basket methodology is used in order to cope with the high attrition rates. In general this delivers a quite volatile but representative item basket. When introducing scanner data in the Danish CPI production especially two issues calls for solutions which are the missing prices and the number of items to process in the production systems.

It became evident that bias from missing prices could partly be dealt with by using a 2 weeks of data per month instead of 1 week per months. Also the supermarket chain aggregation removed some of the bias.

Regarding the specific parameters defining the item basket it became clear that the different COICOP groups have diverse market dynamics. Especially seasonal goods should ideally not have a restriction requiring the item to exist in scanner data the previously 3 or 4 month. Also products with a lot of discounts with temporarily stock-outs should be handled without the 3 or 4 month criterion. As this may not be sufficient to prevent thin series the turnover selection should include more items than a 50% turnover rule or the top 3 best-selling items would include. Further analyses of the COICOP group characteristics will be carried out in future studies which are expected to result in representative item basket suited for the current CPI production.

Appendices

Appendix A - The Danish monthly CPI Workflow



Appendix B – Index formulas

Store prices

(1) $$f_t = \prod_{i=1}^{r} (\rho_t^i)^{1/r} = (\rho_t^1)^{1/r} \cdot (\rho_t^2)^{1/r} \cdot .... \cdot (\rho_t^r)^{1/r}$$

f_t	: store price in period *t*
ρ_t^i	: item prices in period *t from store* i=1,…,r

Product prices

(2)

$$e_t = \prod_{g=1}^{m} (f_t^g)^{s^g} = (f_t^1)^{s^1} \cdot (f_t^2)^{s^2} \cdot .... \cdot (f_t^m)^{s^m}, \sum s^g = 1$$

e_t	: product price in period *t*
f_t^g	: store prices in period *t*
s^g	: weight for store group g = 1,…,m

Basis prices

(3)

$$p_t = \prod_{k=1}^{n} (e_t^k)^{v^k} = (e_t^1)^{v^1} \cdot (e_t^2)^{v^2} \cdot .... \cdot (e_t^n)^{v^n}, \sum v^k = 1$$

p_t	: price on elementary level in period *t*
e_t^j	: product prices in period *t*
v^k	: weight for product group k = 1,…,n

Monthly index on elementary level

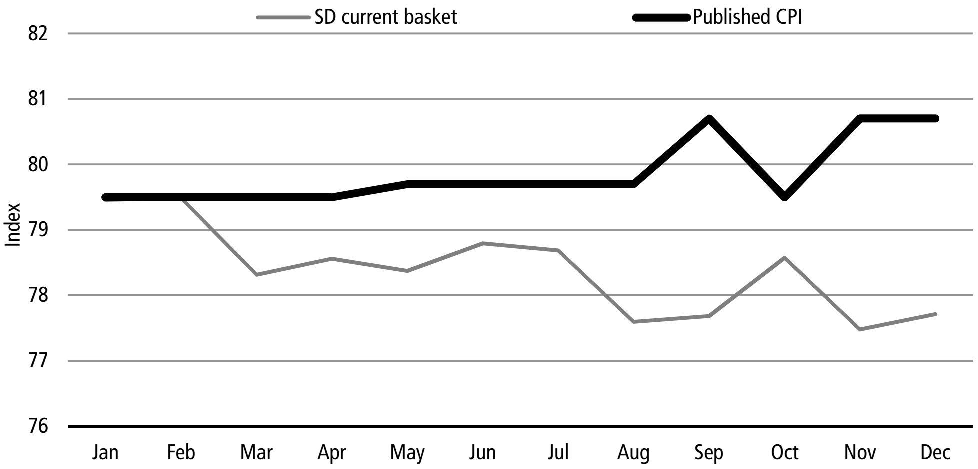
(4) $$I_{t-1:t} = \frac{p_t}{p_{t-1}}$$

$I_{t-1:t}$	: Index on elementary level from *t-1* to *t*
p_t	: Index on elementary level in period *t*
p_{t-1}	: Index on elementary level in period *t-1*

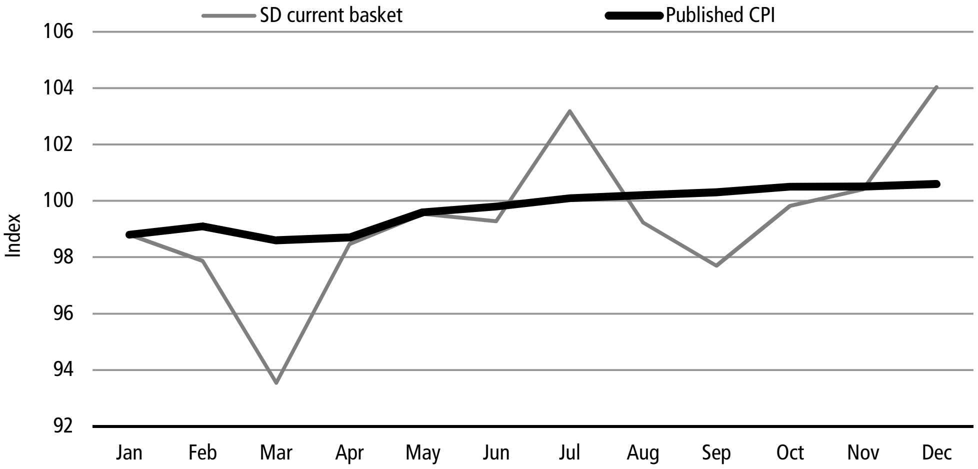
Appendix C – Results of the analyses

1. Change of Price Concept

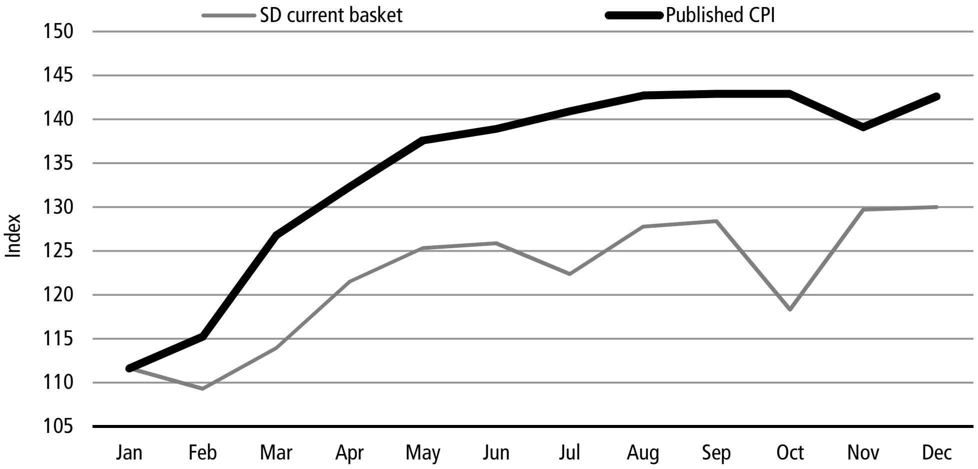
The Consumer Price Index on Rice from January to December 2011



The Consumer Price Index on Red Wine from January to December 2011

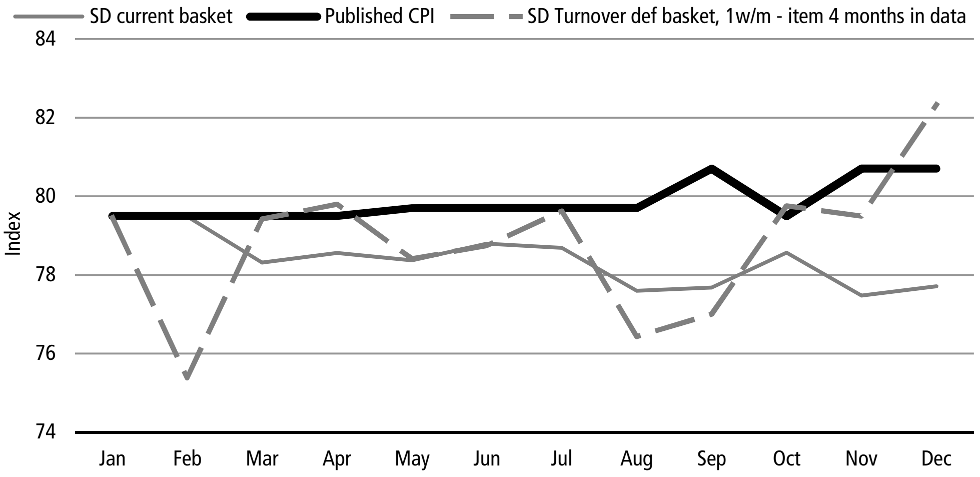


The Consumer Price Index on Coffee from January to December 2011

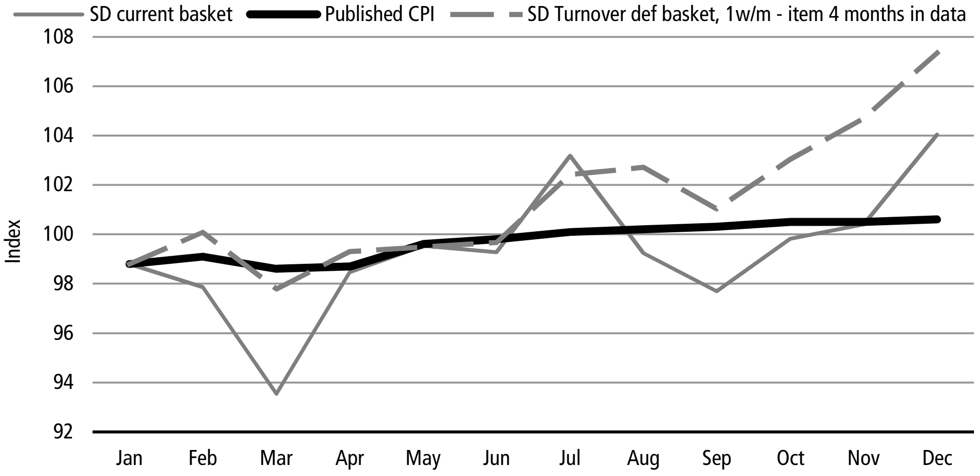


2a. Item basket definition, 1 week data per month

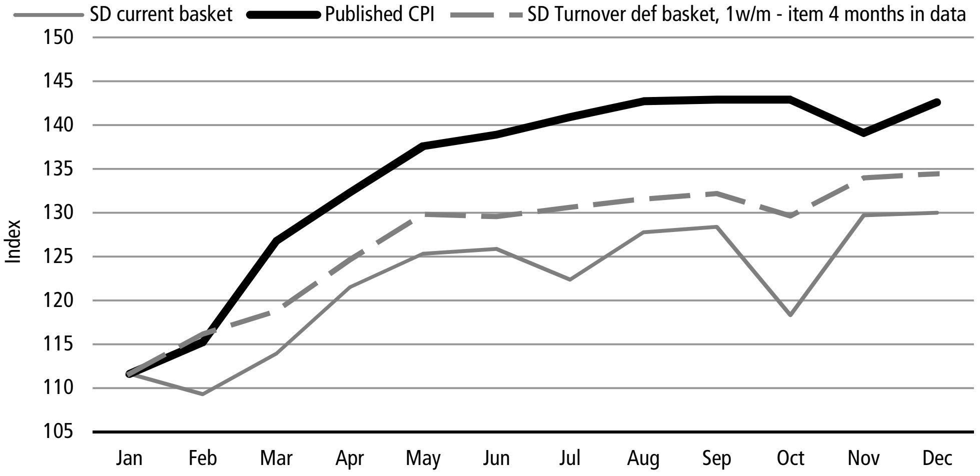
The Consumer Price Index on Rice from January to December 2011



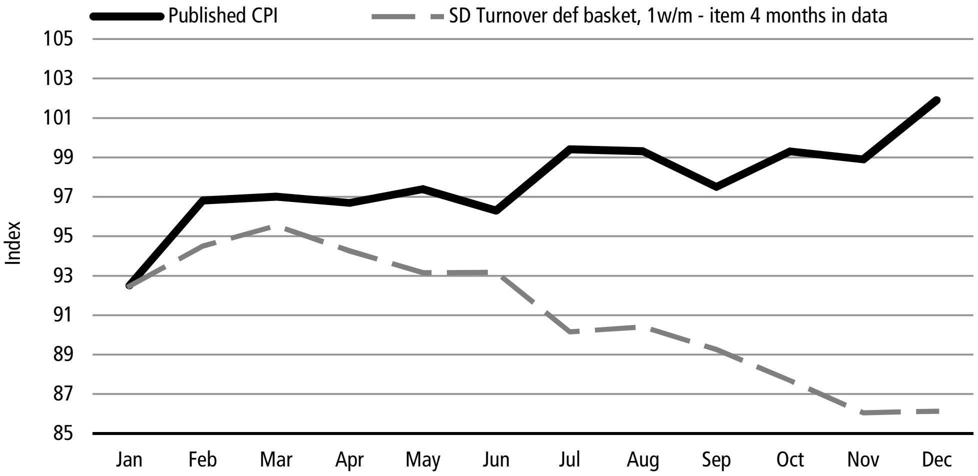
The Consumer Price Index on Red Wine from January to December 2011



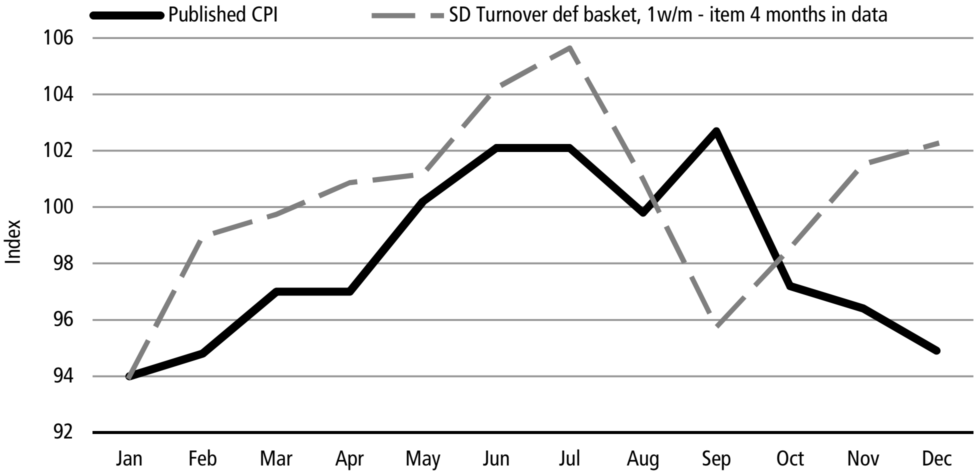
The Consumer Price Index on Coffee from January to December 2011



The Consumer Price Index on Minced Beef from January to December 2011

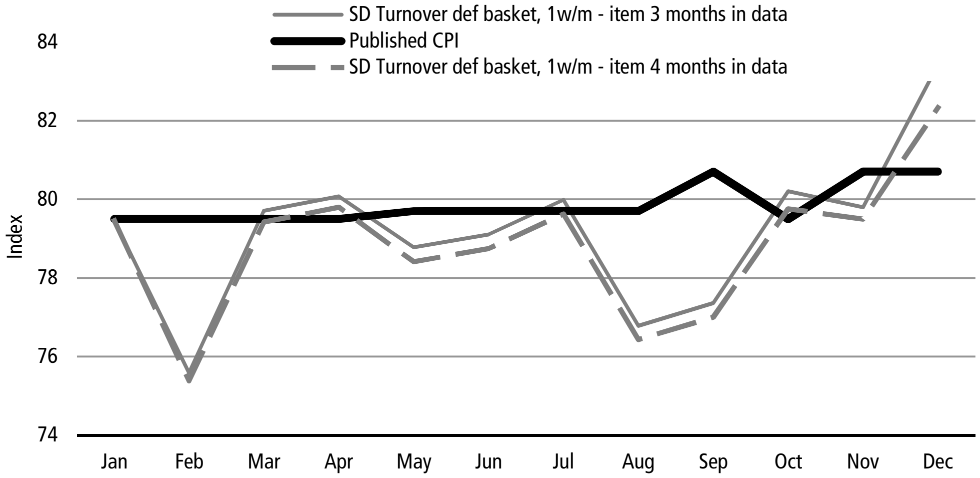


The Consumer Price Index on Apples from January to December 2011

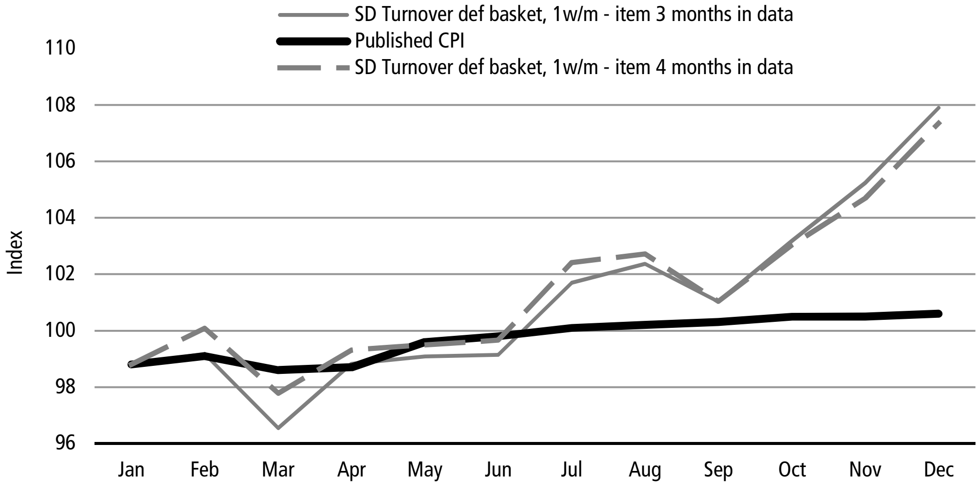


2b. Item in scanner data criteria; 3 months vs. 4 months

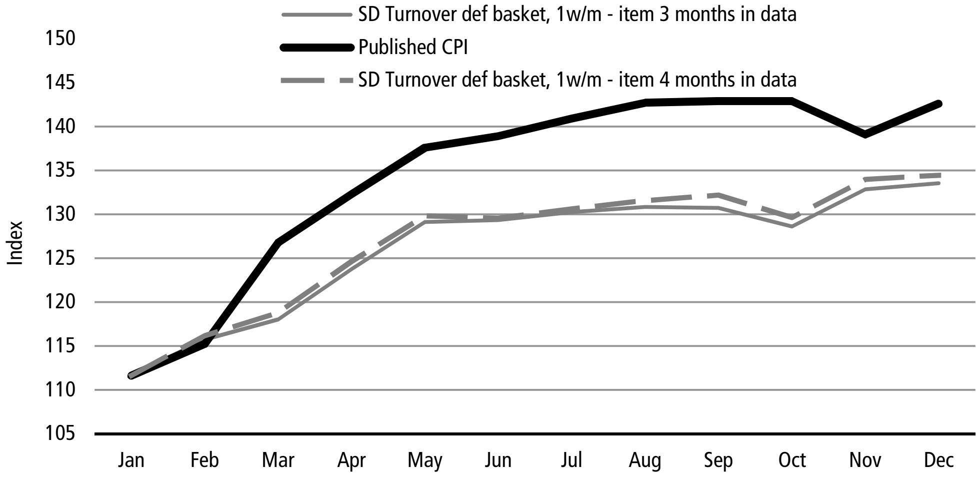
The Consumer Price Index on Rice from January to December 2011



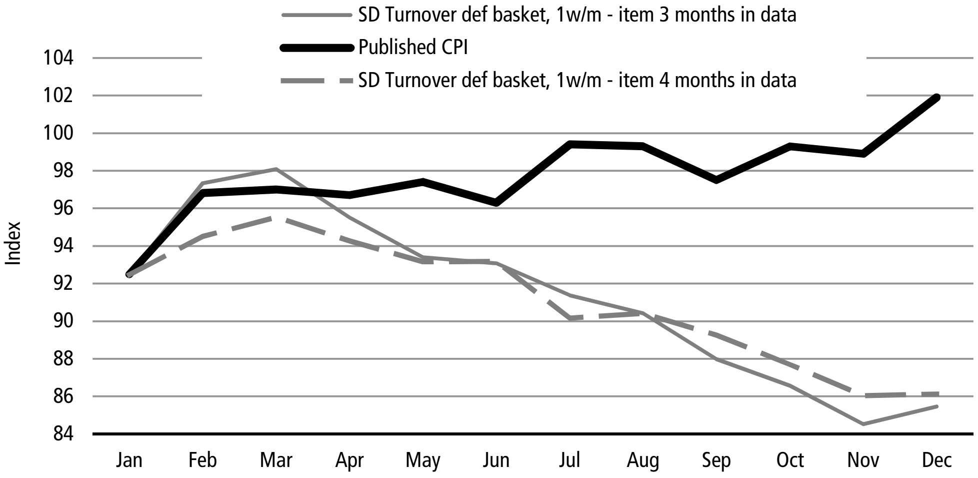
The Consumer Price Index on Red Wine from January to December 2011



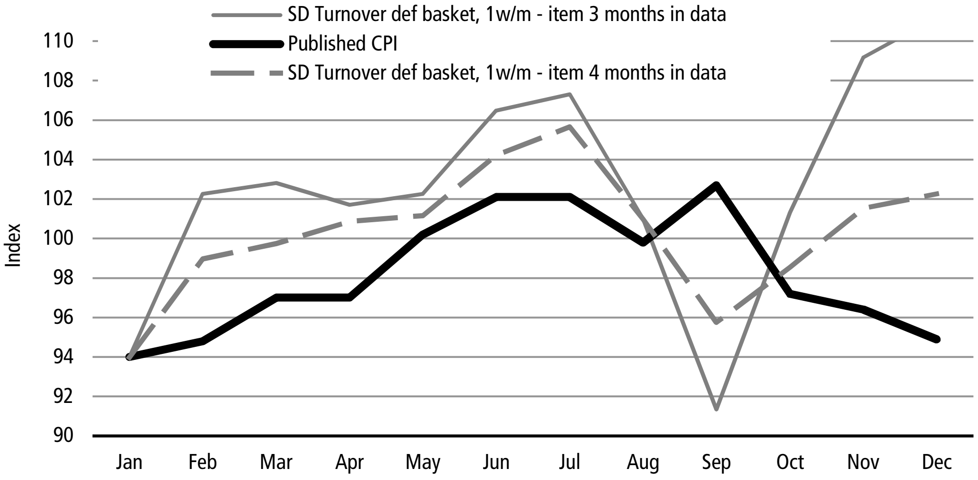
The Consumer Price Index on Coffee from January to December 2011



The Consumer Price Index on Minced Beef from January to December 2011

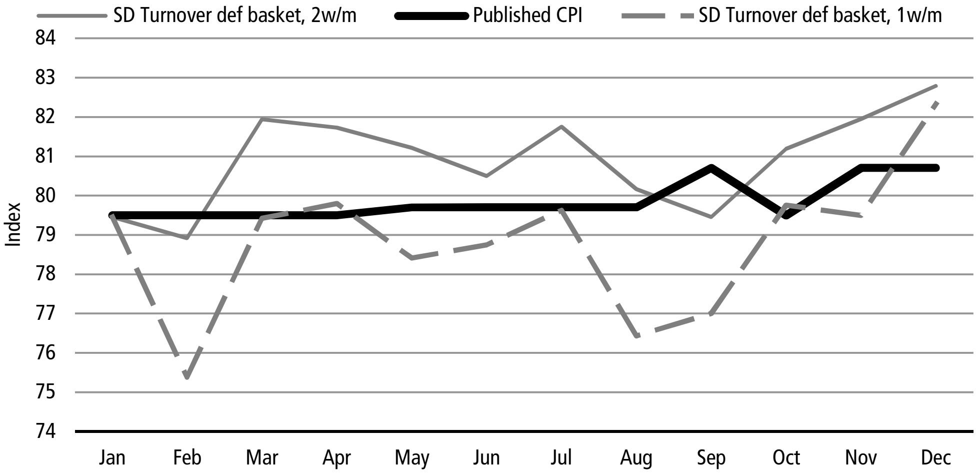


The Consumer Price Index on Apples from January to December 2011

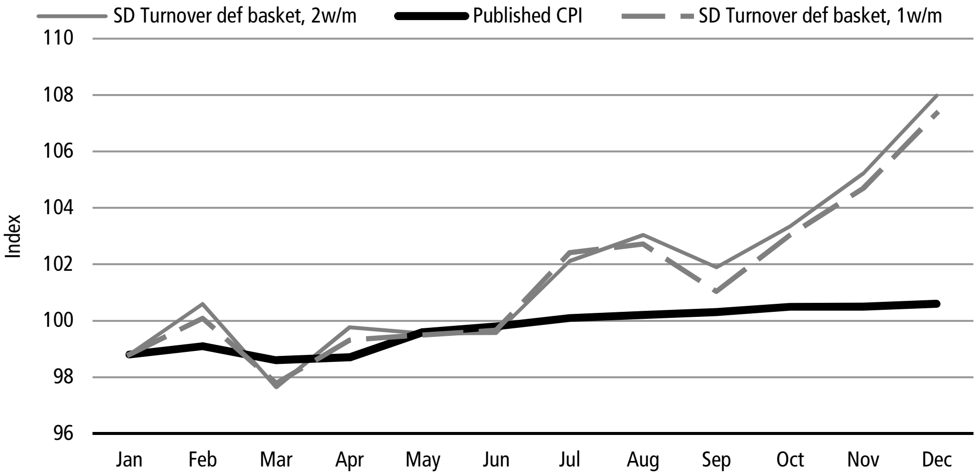


3. Data collection period; 1 week per month vs. 2 weeks per month

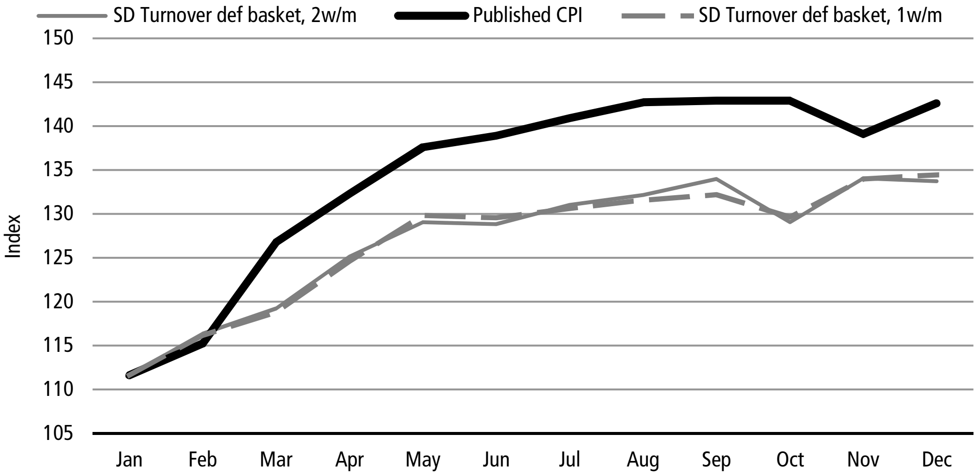
The Consumer Price Index on Rice from January to December 2011



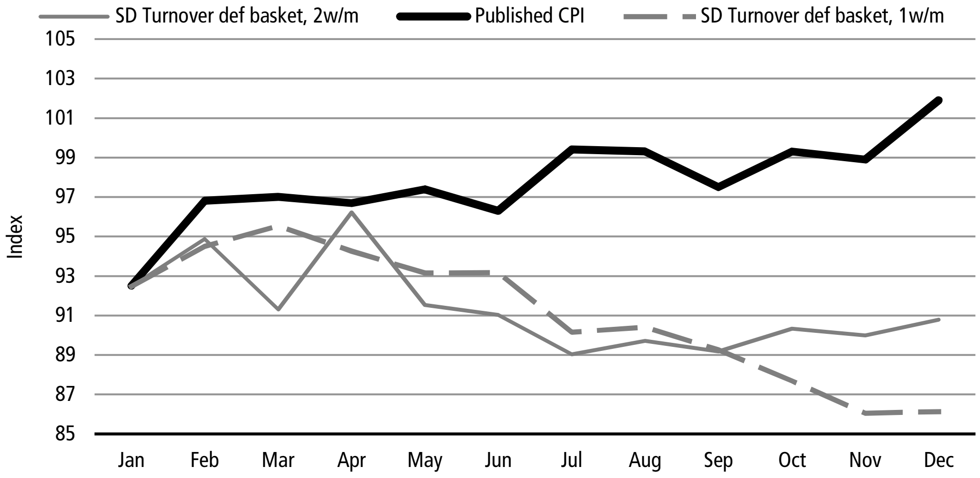
The Consumer Price Index on Red Wine from January to December 2011



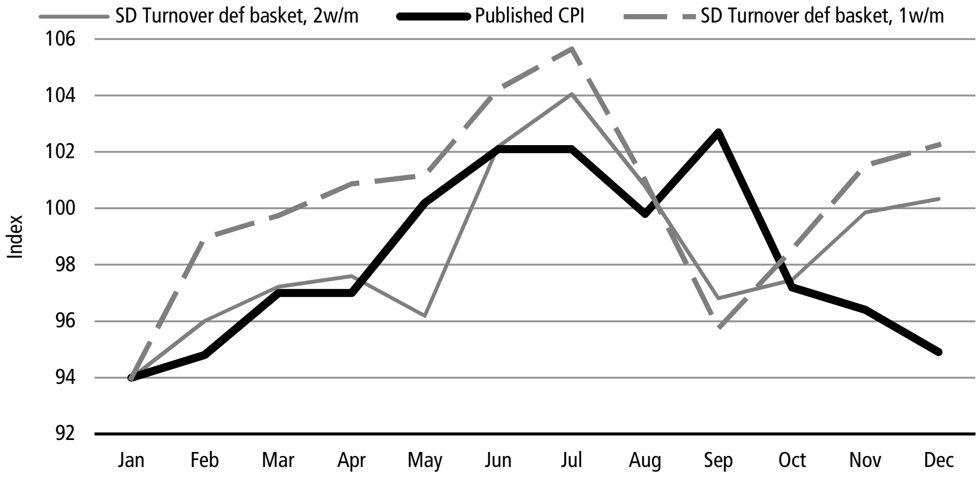
The Consumer Price Index on Coffee from January to December 2011



The Consumer Price Index on Minced Beef from January to December 2011

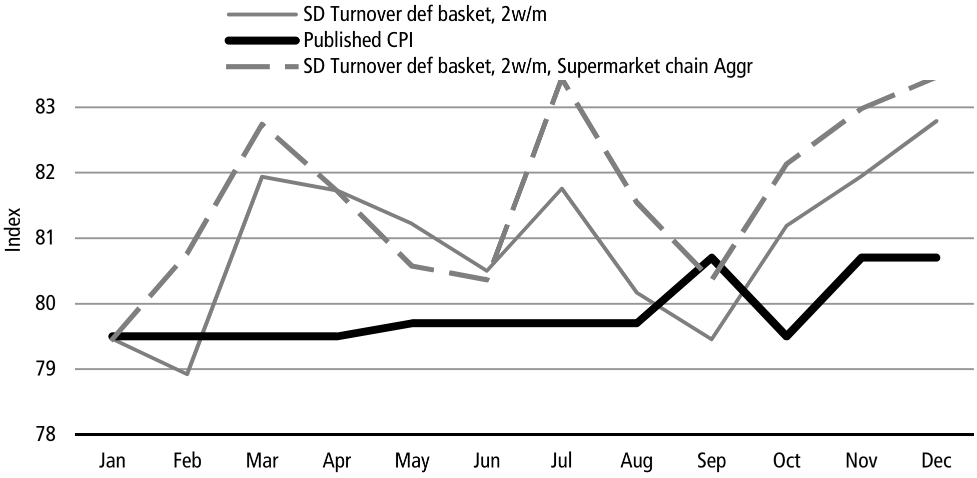


The Consumer Price Index on Apples from January to December 2011

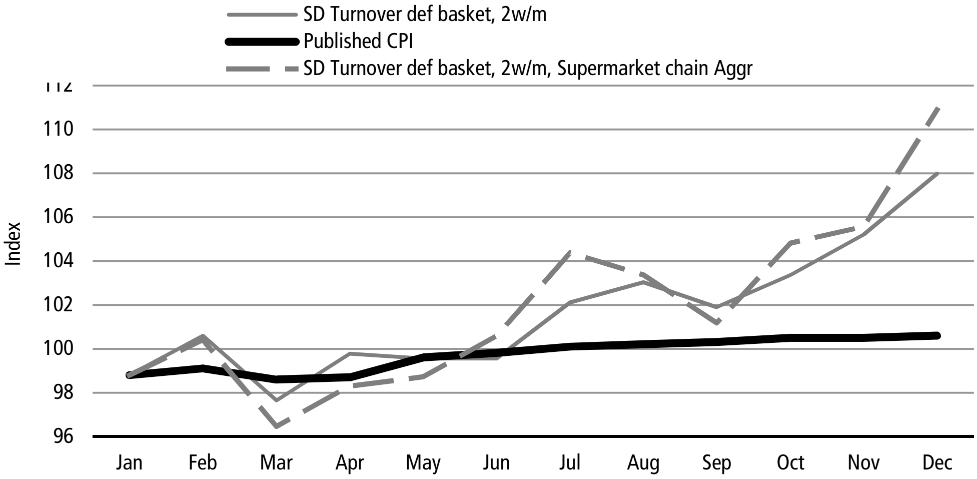


4. Supermarket chain aggregation

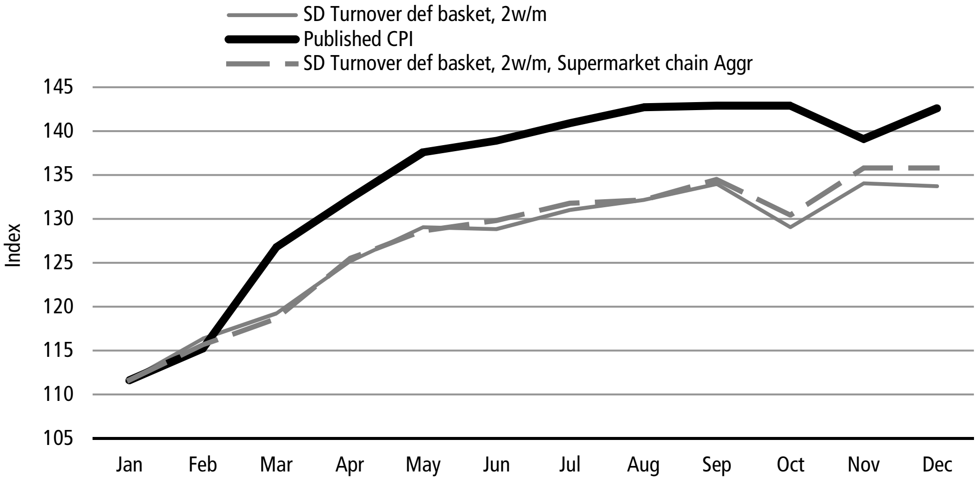
The Consumer Price Index on Rice from January to December 2011



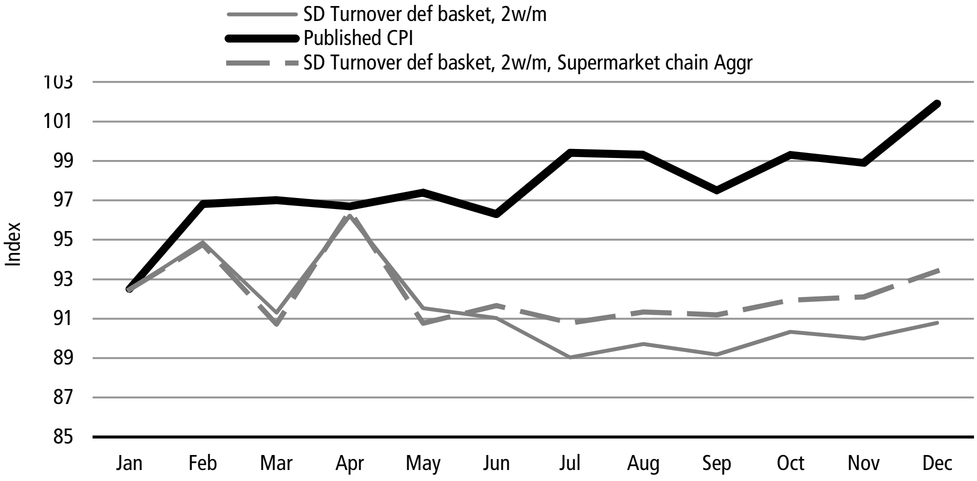
The Consumer Price Index on Red Wine from January to December 2011



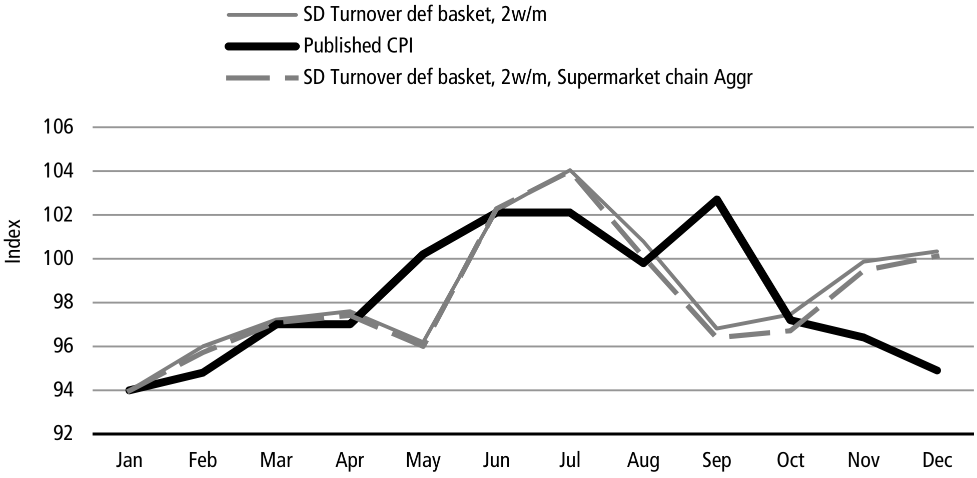
The Consumer Price Index on Coffee from January to December 2011



The Consumer Price Index on Minced Beef from January to December 2011

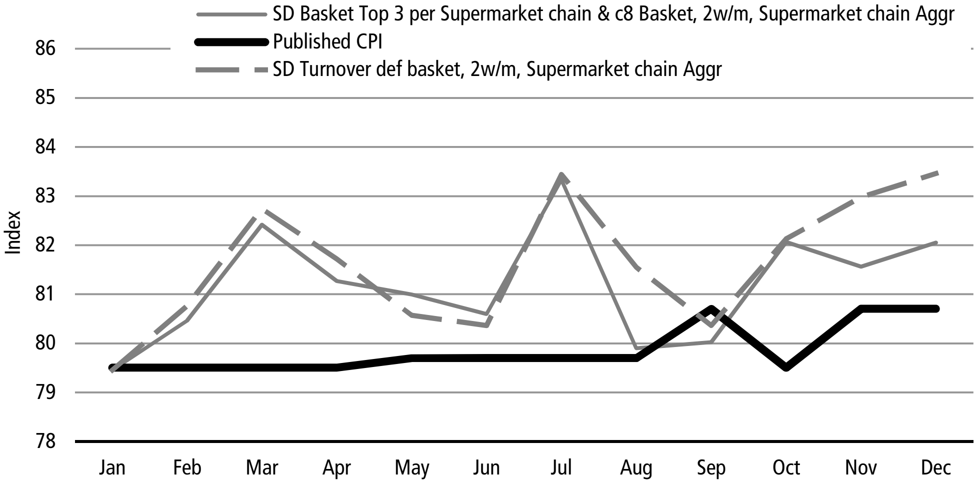


The Consumer Price Index on Apples from January to December 2011

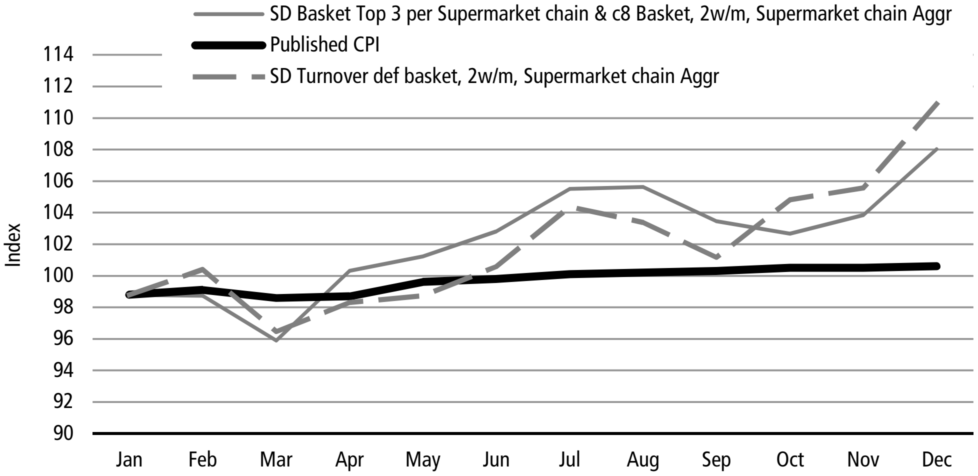


5. Top 3 products per supermarket chain in the COICOP group

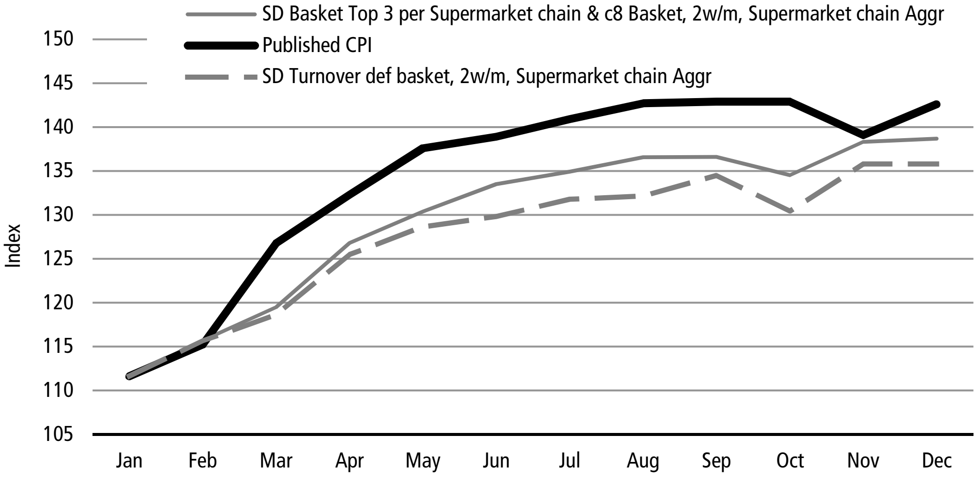
The Consumer Price Index on Rice from January to December 2011



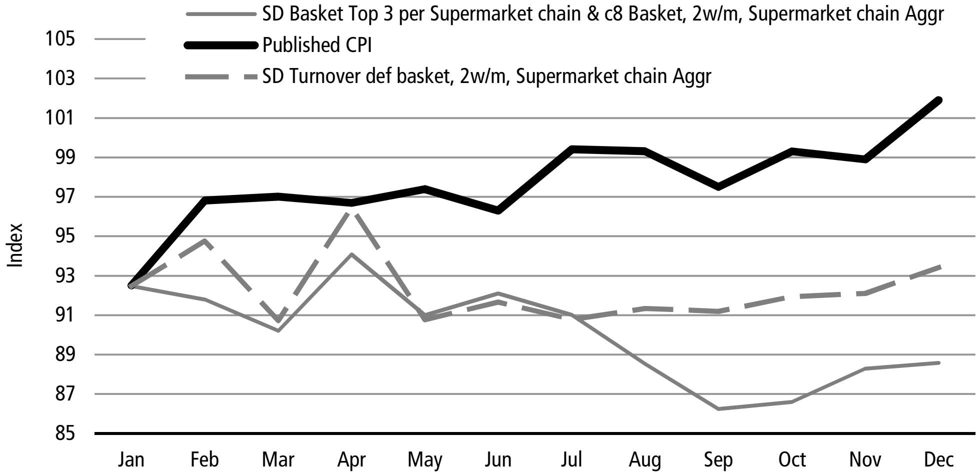
The Consumer Price Index on Red Wine from January to December 2011



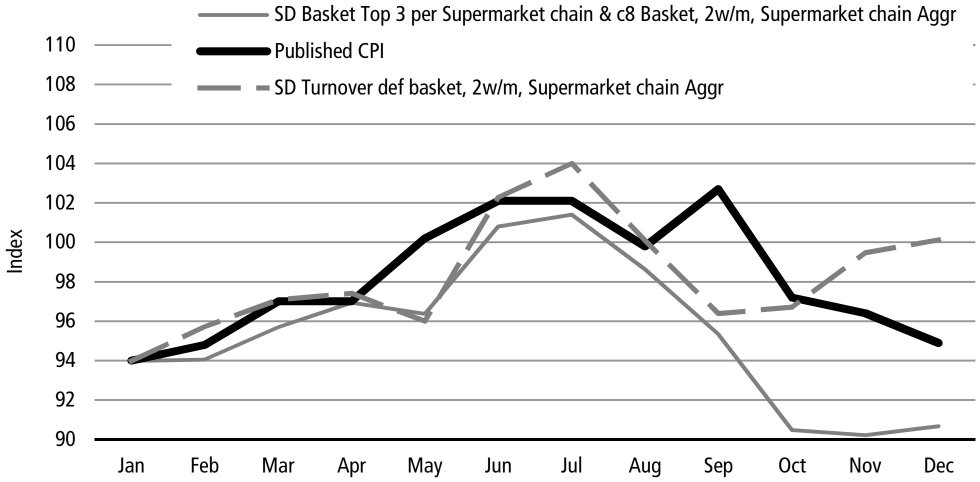
The Consumer Price Index on Coffee from January to December 2011



The Consumer Price Index on Minced Beef from January to December 2011



The Consumer Price Index on Apples from January to December 2011



Appendix D – Number of prices in the analyses

	Number of price observations	
Rice	**Min' - 'Max' obs per month**	**Total in 2011**
Publ CPI	12	144
SD current basket	1978-2726	26687
SD Turnover def basket, 1w/m - item 4 months in data	5884-7487	78846
SD Turnover def basket, 1w/m - item 3 months in data	5535-7810	82218
SD Turnover def basket, 2w/m - item 4 months in data	6359-7791	83477
SD Turnover def basket, 2w/m, Supermarket chain Aggr - item 4 months in data	40-50	536
SD Basket Top 3 per supermarket chain & c8 Basket, 2w/m, Supermarket chain Aggr	27	324

	Number of price observations	
Red wine	**Min' - 'Max' obs per month**	**Total in 2011**
Publ CPI	147-154	1815
SD current basket	14615-17760	194989
SD Turnover def basket, 1w/m - item 4 months in data	27755-30568	353927
SD Turnover def basket, 1w/m - item 3 months in data	28370-32084	361240
SD Turnover def basket, 2w/m - item 4 months in data	32666-35205	410795
SD Turnover def basket, 2w/m, Supermarket chain Aggr - item 4 months in data	220-239	2758
SD Basket Top 3 per supermarket chain & c8 Basket, 2w/m, Supermarket chain Aggr	27-28	326

	Number of price observations	
Coffee	**Min' - 'Max' obs per month**	**Total in 2011**
Publ CPI	110-115	1351
SD current basket	2823-3016	35412
SD Turnover def basket, 1w/m - item 4 months in data	7717-8433	98988
SD Turnover def basket, 1w/m - item 3 months in data	7732-9232	102092
SD Turnover def basket, 2w/m - item 4 months in data	8308-9103	105678
SD Turnover def basket, 2w/m, Supermarket chain Aggr - item 4 months in data	51-53	621
SD Basket Top 3 per supermarket chain & c8 Basket, 2w/m, Supermarket chain Aggr*	53-54	645

*3 normal coffee products and 3 instant coffee products

	Number of price observations	
Minced beef	**Min' - 'Max' obs per month**	**Total in 2011**
Publ CPI	115-121	1407
SD current basket		
SD Turnover def basket, 1w/m - item 4 months in data	832-2423	21676
SD Turnover def basket, 1w/m - item 3 months in data	828-2580	22121
SD Turnover def basket, 2w/m - item 4 months in data	1013-2561	23098
SD Turnover def basket, 2w/m, Supermarket chain Aggr - item 4 months in data	10-15	155
SD Basket Top 3 per supermarket chain & c8 Basket, 2w/m, Supermarket chain Aggr	23-26	291

	Number of price observations	
Apples	**Min' - 'Max' obs per month**	**Total in 2011**
Publ CPI	106-117	1358
SD current basket		
SD Turnover def basket, 1w/m - item 4 months in data	1827-2656	27066
SD Turnover def basket, 1w/m - item 3 months in data	1477-2710	29175
SD Turnover def basket, 2w/m - item 4 months in data	1857-2704	27806
SD Turnover def basket, 2w/m, Supermarket chain Aggr - item 4 months in data	10-13	145
SD Basket Top 3 per supermarket chain & c8 Basket, 2w/m, Supermarket chain Aggr	26-28	325