Statistics Denmark
Prices and Consumption

Nina Gustafsson*

Drawing a Sample from Scanner Data to use in the Danish CPI

Paper for poster session at the 13th Ottawa Group Meeting

Copenhagen, Denmark

May 2013

* The work on the system for receiving scanner data and linking EAN to COICOP, was done by Jonas Mikkelsen before he left the office for Prices and Consumption. I just describe it in this paper in order to create coherence with my work.

Introduction

In Statistics Denmark we are currently working on integrating scanner data into the Danish consumer price index (CPI). To begin with we will use the scanner data only for food, beverages and tobacco. In order to do so we have decided not to use all scanner data available but rather draw a sample from the scanner data to use in the CPI calculations. The focus of this paper is the issues we have to deal with in order to draw a representative yet stable sample to use in the Danish CPI from the massive amount of data scanner data supplies.

The first section of this paper presents the scanner data as well as the work we have put into the reception and initial treatment of the data. Some critical issues we need to deal with when introducing scanner data into the CPI are presented in the second section. The third section describes the system for drawing and maintaining a representative sample from the scanner data while the fourth section presents our initial experiences on drawing the sample. Finally, the fifth section of this paper sums up as well as briefly presents the future work we will do on this project.

The scanner data

The data

Since January 2011 Statistics Denmark has received scanner data from the largest supermarket chains in Denmark on a weekly basis. These supermarket chains account for approximately 60% of the Danish sales of food and beverages providing us with good coverage of data. Currently we are in the process of retrieving scanner data from one more supermarket chain making the total coverage of the scanner data for food and beverages 80%. The data contains the following variables for each sold item

- Date
- Store number
- EAN (or PLU) number
- Turnover
- Volume
- Unit
- Quantity per unit
- Product number
- Product description

The *date* is 4 digits and consists of a 2 digit year number and a 2 digit week number, e.g. week number 2 in 2012 would have the date 1202. The *store number* is a unique number for the specific supermarket store in which the item is sold; each supermarket chain covers many different stores. In scanner data each item has its own product code called *European Article Numbering (EAN)* or *Product Lookup Code (PLU)*. The EAN number is defined by the producer of the product whereas the PLU number often is defined by the supermarket chain. When working with scanner data, the EAN/PLU number is used as the product identification which enables us to secure matched prices. The price of the item is derived from dividing the weekly *turnover* with the weekly *volume* for each EAN number. The *product number* is a very important variable to us as it reflects 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.

Reception, storing and validation of the data

Statistics Denmark has been very fortunate that the supermarket chains are willing to share their scanner data, providing us with massive amounts of data. At Statistics Denmark we have now created a system able to receive and store the data as well as a system to make various tests on the received data in order to make sure there are no obvious errors in the data file, e.g. File name, record lengths, numeric variables only containing numbers, no null values.

Linking the EAN numbers to COICOP

A great task in the scanner data project has been to link the EAN numbers to the COICOP classification thus making the data compatible with the CPI. This has resulted in a key between the EAN numbers and COICOP on a 6-digit level. The main part of the key was made by linking the supermarket chains' own product structures, based on information retrieved in relation to the product numbers mentioned above, with the COICOP. In cases where this was not possible, EAN numbers was linked to the COICOP by matching the product description with the COICOP description.

Maintenance of the key

A key linking the EAN numbers to the COICOP, however, is not static since many new EAN numbers occur in the scanner data on a regular basis. Therefore a pilot IT system supporting the maintenance of the key is currently being tested. The system is based on the linking between the supermarket chains' product structures and the COICOP where possible and a search word process on the product description for any other EAN numbers. The search word process is only applied to a group of EAN numbers covering a share of the weekly turnover in the corresponding 4-digit COICOP group of more than 5%. The search word process is only done once per EAN number since the result is stored and used in the key from then on.

Build in the IT system for maintenance of the key is also some quality assurance processes. For one thing tests for changes in the supermarket chains own classification is set up as well as a test for one EAN number having different product descriptions.

Critical issues when using scanner data in the CPI

Using scanner data in the CPI is not an uncomplicated task. Many issues have to be considered. The main issue is avoiding possible bias in the CPI.

Scanner data volatility

For the CPI we want to continuously monitor products over time. However, every week there are a number of products entering or leaving the supermarket store, i.e. all products are not available in the scanner data in every period. 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 not properly dealt with.

Possible bias deriving from the scanner data volatility

The scanner data based CPI may have an upward or downward bias when either new products enter the item basket or when products leave the item basket. We have observed biases in two ways:

1. A product enters the item basket on discount the first month. The next month the product has its normal price. This leads to an artificial increase in the index which will not be levelled out.
2. A product leaves the item basket on discount. This leads to a persistent decrease in the index.

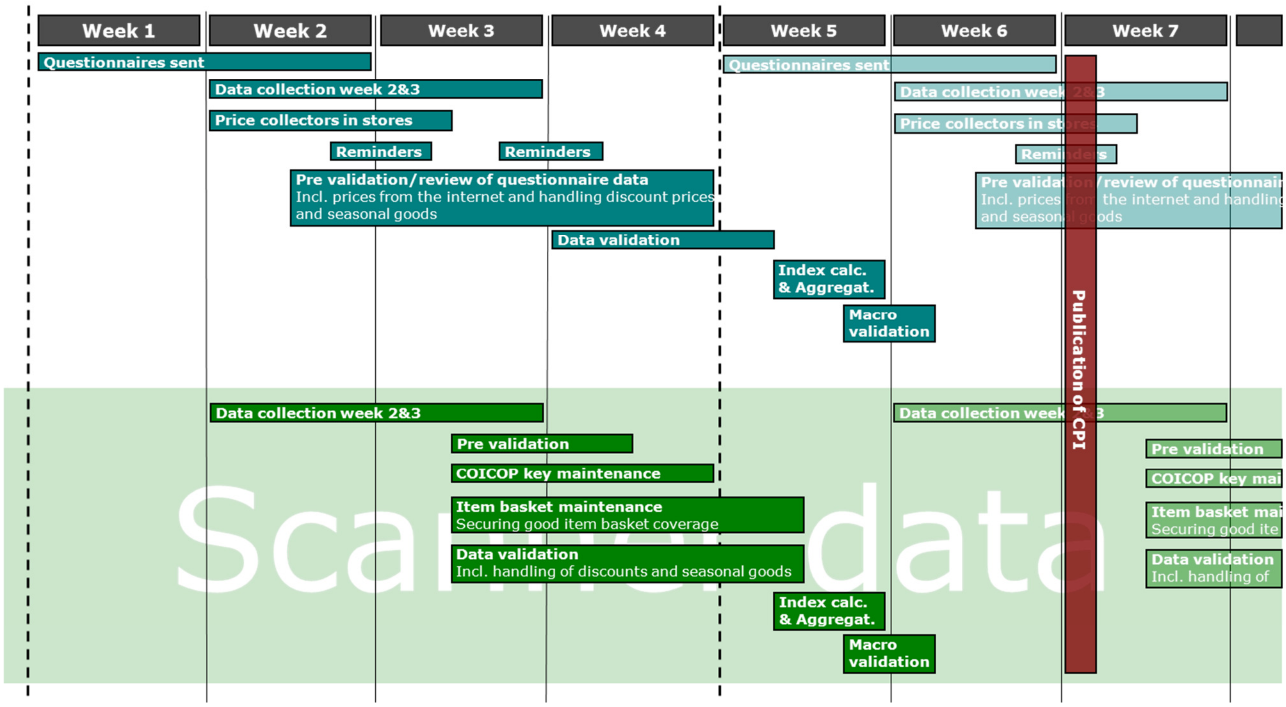
When the incidents described above happen to a larger proportion of the items, the bias becomes problematic in the indices over time. The bias is more likely to happen when the products are often temporarily out of the data. The many missing prices in the scanner data series are impossible to foresee, thus has to be handled ex post. The decision on how to treat missing prices is important and may change depending on the product being temporarily or permanently out of the scanner data.
Historically Statistics Denmark has treated missing prices manually. This ensures that any imputation bias towards index 100 is minimized. In the implementation of scanner data in the CPI it is attempted to continue processing the missing prices with the current method, which therefore will include some manual handling.

Limiting missing prices by aggregating on chain level

A way to minimize the number of missing prices when dealing with scanner data, is to aggregate each item volume and turnover on chain level. This way temporary stock outs in a specific store is no longer a problem as long as the item is sold in any other store in the supermarket chain. Moreover, the amount of data to handle is limited when aggregating on chain level which speeds up the performance of the IT systems as well as allow more manual monitoring. Aggregation on chain level limits the weekly data from 5.1 million observations to 41,000 observations on average. Therefore, we have decided to aggregate each item turnover and volume on chain level when dealing with scanner data.

Data processing within the current production flow

Another critical issue for us when using scanner data in the CPI is allowing the data processing to be within the time span of our monthly production flow. This means there is limited time for processing the scanner data. Furthermore, the weekly aggregated data can be split between months, that is, data for the first (last) week of a month often includes data belonging to both the previous (following) month and the month in question. Below the current monthly production flow is shown as well as the projected scanner data processes.



The current production process and the restriction regarding weekly data containing observations for two different months allow use of 2 weeks of scanner data per month. Scanner data introduces new processes such as maintenance of the key between the COICOP and EAN numbers and maintenance of the item basket securing a suitable coverage of the total turnover. These processes are in addition to traditional data validation where we deal with missing prices, check extreme price developments and handle seasonal goods which will also be necessary with the use of scanner data.

The system for drawing and maintaining a representative sample from scanner data

Scanner data can basically be used in the compilation of price indices in two different ways. Either you can use the scanner data sets to draw a sample of price observations and then calculate the price index using this sample in a traditional way i.e. a representative basket methodology. Otherwise you can use more or less the full population in the calculations including quantities at the micro level. The full population method is, however, prone to drift and bias problems due to difficulties in taking proper account of seasonal goods and goods on discount leaving the sample. We have therefore decided as our starting point to turn to the representative basket methodology using a sample size that we are able to properly monitor and control for seasonal goods and goods leaving the sample.

Pilot IT system

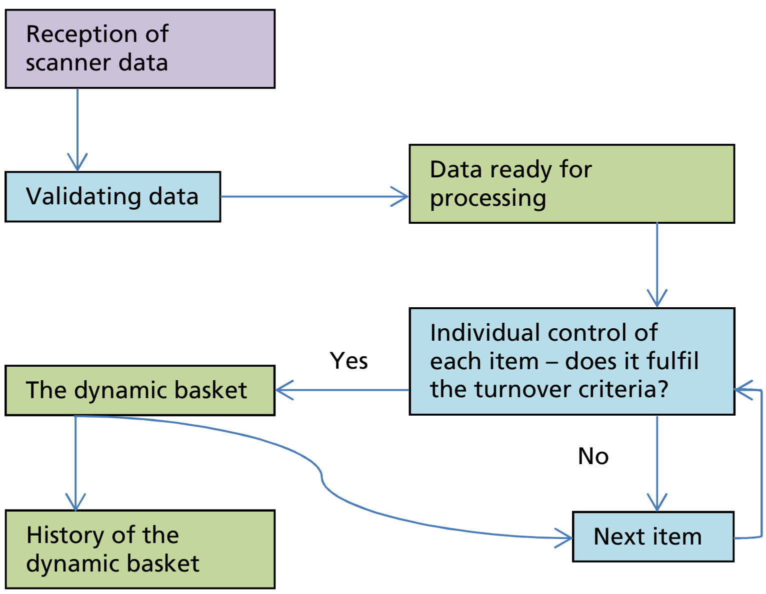
To utilize this method in our production of the CPI it will therefore be necessary to build a pilot IT system that can draw and maintain a representative basket using scanner data. The basket should be updated every month and the precise rules for updating the basket will have to be determined (e.g. how many price observations for each product should be included in the basket, for how long should it have been available on the market etc.).

The dynamic basket

The system will be based on drawing two types of item baskets. The first is the dynamic basket. This basket of items is drawn from the total scanner data by filtering out items that may be candidates for the representative basket. The filter for the dynamic basket will be based on turnover, i.e. items with the largest turnover in each 6-digit COICOP group are selected. The practical way we do this is to take the items in the monthly data and determine their aggregated turnover of the latest four months relative to the 6-digit COICOP group aggregated turnover of four months. Only the items with the highest relative turnover of four months are selected to enter the dynamic basket. The criteria for the dynamic basket will be individualized for certain COICOP groups depending on e.g. how prone the group is to seasonal goods.

History of the dynamic basket

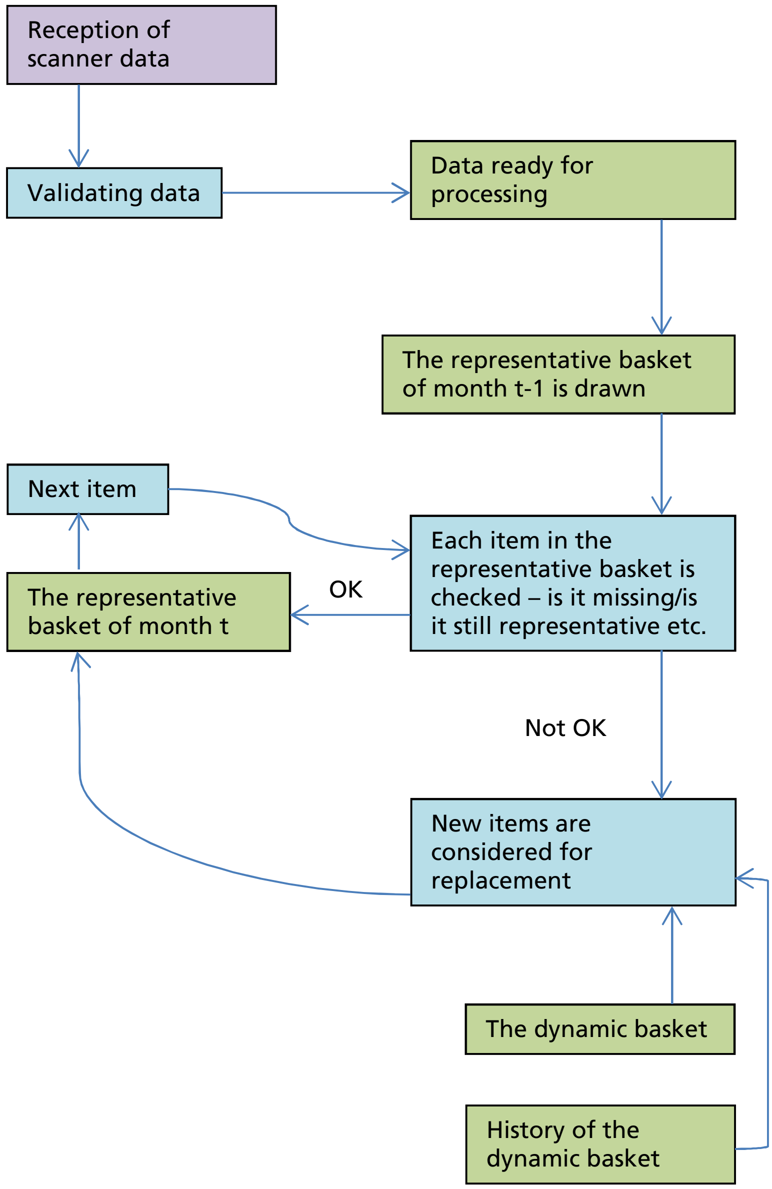
A system will also monitor the history of the dynamic basket. This means that information on how long the item has fulfilled the turnover criteria, i.e. been in the dynamic basket, will be available. This gives us an idea of the stability of each item, enabling us to choose a representative basket with more stability. The system for drawing the dynamic basket is illustrated below.



The representative basket

The second basket is the actual representative basket. This basket includes the items that will enter the index calculations for the CPI. The purpose of this basket is to collect a sample that is stable and representative. The initial sample or basket of items is drawn from the scanner data based on a number of criteria. This basket is of course not static since items leave the data and items may become unrepresentative over time. In such cases the representative basket is therefore updated with items from the dynamic basket.
For the representative basket a system will make sure that, items with large price changes, items that have become too unrepresentative or items that are no longer

available are made available for manual handling. When removing an item from the representative basket a new item will be chosen from the dynamic basket. Most of the changes in items will happen automatically, however, it will be possible to monitor the changes as well as do them manually. For example, items temporarily not available could be kept in the basket in spite of the systems suggestions of changing the item. The system for the monthly handling of the representative basket is illustrated below.



Experiences on drawing the sample

In this section experiences on drawing the sample from scanner is presented. Recall that we are only dealing with COICOP groups 1 and 2. Drawing the sample is still a work in progress making the results shown here preliminary. But first our current sample is described briefly.

The current sample

The current sample for food, beverages and tobacco covers about 8.200 price observations in total on 153 sub-groups. Each price observation covers one unique item in one unique store. Therefore, there will be multiple prices for the same item gathered from different stores. The prices are collected manually in the stores or reported to Statistics Denmark by the stores.

Drawing the initial sample

To begin with we have to draw an initial representative basket. For this we use scanner data for 2011 where monthly datasets have been generated using 2 weeks of data per month. Furthermore, all items (EANs) have been aggregated on chain level, as mentioned above, limiting the amount of data considerably.

Two selection criteria: available in 12 months and 50% of turnover

When selecting items for the representative basket we realise that no single selection criteria will fit all 153 COICOP sub-groups we have on a 6-digit level. However, as a starting point we look at items that are present in all twelve months of 2011 and that constitute the highest share of turnover within the COICOP sub-group. More precisely, we look at items that within their COICOP sub-group constitute the top 50% of the yearly turnover for each supermarket chain. Due to major differences in the sizes of the chains (Dansk Supermarked and COOP are much larger in terms of turnover than Rema1000) looking at items constituting top 50% of turnover within their sub-group overall, i.e. without the chain level, would not ensure representation of all chains in the sample. By looking at the top 50% within each supermarket chain we make sure that all three chains are represented in the sample.
These two criteria – available in data in all 12 months of 2011 and within top 50% of turnover for each chain – form the sample shown in table 1 below. The table also shows the number of prices in the current sample, the number of prices the CPI weights dictates as well as the average monthly observations in scanner data without the two selection criteria. Note that the number of observations in the sample drawn from scanner data in reality covers several prices, because each observation is an aggregate over multiple price observations.

From table 1 we see how the amount of observations in scanner data is massively reduced from 44.381 a month on average to 3.545 when using the two selection criteria. Moreover, we see that the availability and turnover criteria combined produce a sample of observations of less than half the current sample. This, however, does not necessarily mean that the sample covers fewer prices since, as mentioned above, each observation in the scanner data sample is an aggregate of price observations from multiple stores within the supermarket chain.

Table 1. Number of observations in sample drawn from scanner data with items constituting top 50% of COICOP sub-group turnover by chain and available in all 12 months of 2011

COICOP	Description	No. of observations in sample from SD	No. of observations in current sample	No. of observations dictated by CPI weight	Avr. no. of observations in monthly SD without selection criteria
	Total sample	**3.545**	**8.217**	**3.514**	**44.381**
11110	Rice	22	12	12	188
11121	Flour	41	20	15	535
11122	Oats	25	12	10	301
11131	Rye bread	43	103*	51	409
11133	Whole grain bread	16	104*	19	131
11134	White bread	40	98*	19	505
11135	Rolls	38	103*	53	383
11136	Flute, sausage bread and pita bread	34	49	15	281
11137	Crisp bread	36	12	14	347
11141	Pastry	32	46*	22	348
11142	Cream cakes	98	62	15	1.217
11143	Cakes and pies	16	12	21	184
11144	Cookies	21	12	14	168
11145	Biscuits	41	12	16	384
11150	Spaghetti, pasta and noodles	49	12	21	465
11160	Cereal	26	49	27	195
11170	Pizza etc. frozen	19	31	10	180
11211	Ground beef	6	54*	52	76
11212	Shoulder of beef	2	75*	12	10
11213	Beef minced	13	92*	21	277
11214	Beef tenderloin	42	195*	42	907
11215	Ground beef, organic	3	69*	2	15
11221	Veal minced	4	57*	7	68
11222	Veal topside	11	66*	16	152
11231	Pork minced	7	95*	3	105
11233	Pork loin without bacon	16	61*	31	297
11234	Pork tenderloin	5	104*	14	112
11236	Pork loin with bacon	25	97*	22	632
11237	Ground pork	10	111*	20	124
11238	Ground pork, organic	2	32	0	6
11240	Lamb and game meat	22	44	15	393
11251	Chicken, fresh and frozen	6	20	18	66
11252	Ducks	18	15	16	194
11253	Turkey breast	4	77*	8	20
11254	Chicken fillet	25	86	40	318
11255	Meat salads	20	11	6	96
11261	Meat offal	7	46*	6	126
11271	Cold cuts, ham	7	12	19	49
11272	Cold cuts, salted meat	4	12	7	22
11273	Cold cuts, pork fillet	3	12	9	19
11274	Prepared dishes	48	13	21	419
11275	Prepared dishes, cans	28	22	3	313
11276	Ham	4	114	12	25
11278	Cold cuts, mortadella	72	24	28	859
11280	Liver pate	21	115	36	371
11282	Cooked meats	6	110	13	159
11283	Sausages and bacon	44	106	47	689
11284	Cold cuts, salami	14	12	25	143
11311	Cod	9	22*	4	97

11312	Flounder	4	90*	7	14
11313	Herring fillet	3	6*	5	20
11314	Salmon fillet	15	86*	9	155
11321	Cod fillet, frozen	2	11	5	11
11322	Flounder fillet, frozen	18	13	3	154
11331	Smoked mackerel	3	0*	2	32
11332	Smoked salmon, seafood and caviar	18	90	16	154
11333	Mackerel, cans	15	23	13	180
11334	Cod roe, cans	10	12	6	93
11335	Marinated herring	30	11	10	215
11336	Fish salads	13	12	7	86
11338	Stuffed fish fillets	15	67*	12	131
11339	Shrimps	21	12	8	175
11411	Whole milk and infant formula	5	51	14	148
11412	Semi-skimmed milk	5	52	20	20
11413	Skimmed milk	5	50	14	34
11414	Buttermilk	6	47	10	59
11415	Low fat milk	5	53	18	27
11416	Whole milk, organic	4	44	3	28
11417	Semi-skimmed milk, organic	3	47	5	28
11418	Skimmed milk, organic	3	51	8	11
11419	Low fat milk, organic	5	50	11	39
11431	Whipping cream	9	51	20	104
11432	Sour cream	10	46	9	66
11433	Yogurt	52	146	52	482
11434	Chocolate milk	8	12	11	129
11441	Cream cheese	15	12	29	182
11442	Brie cheese	22	50	22	297
11443	Cheese	93	92	90	1.274
11450	Eggs	8	50	25	140
11452	Eggs, organic	3	50	8	16
11511	Butter	6	53	24	89
11512	Butter mixture	7	53	21	48
11521	Margarine	12	13	8	75
11522	Vegetable margarine	4	13	9	6
11531	Food oils	12	12	16	206
11611	Apples and pears	12	112*	30	183
11612	Citrus fruits	6	107*	24	172
11613	Soft fruits	15	63*	45	299
11614	Bananas	4	92*	20	35
11615	Grapes and melons	12	112*	25	236
11621	Dried fruit	46	24	17	623
11622	Nuts, almonds	29	45	26	497
11630	Canned fruit, frozen berries	21	34	8	134
11711	Carrots	7	106*	11	84
11712	Root vegetables	8	98*	12	71
11713	Tomatoes	7	108*	25	132
11714	Cucumber, eggplant, zucchini	4	94*	17	104
11715	Onions	9	111*	12	164
11716	Mushrooms	6	110*	7	72
11717	Lettuce	12	91*	19	216
11718	Peppers	10	109*	13	153
11719	Cabbage	3	83*	4	11
11721	Potatoes	8	112*	33	151

11724	Potatoes, organic	3	91*	5	21
11731	Frozen vegetables	31	71	14	275
11732	Potato chips	15	12	8	99
11751	Canned vegetables	75	94	25	817
11752	Crisps	56	12	11	550
11753	Roasted onions	3	11	1	12
11754	Curry salad, Italian salad	16	24	9	96
11791	Cauliflower	17	98*	21	221
11794	Carrots, organic	0	101*	2	
11810	Sugar	10	12	14	185
11821	Jam	69	24	24	570
11822	Honey	10	12	6	159
11831	Chocolate	162	125	95	1.541
11832	Candy	239	99	121	2.790
11840	Ice cream	75	49	61	747
11911	Salt	7	12	4	64
11912	Spices	22	24	13	202
11913	Vanilla sugar	3	12	3	19
11914	Herbs	33	10	10	871
11921	Vinegar	14	11	3	102
11922	Mustard	20	12	7	139
11923	Tomato ketchup	14	12	11	112
11924	Ready sauce	77	12	23	868
11925	Salad dressing	30	12	9	266
11926	Baking ingredients	41	23	11	433
11931	Mayonnaise	8	12	6	36
11932	Remoulade	10	12	6	50
11945	Soups, baby food	47	20	13	542
12110	Coffee	33	118	87	725
12120	Tea	60	12	17	507
12130	Cocoa	10	13	4	37
12210	Mineral water	11	11	17	184
12221	Soft drinks	38	291	109	1.029
12222	Lemonade	26	12	11	422
12231	Orange juice	23	50	52	272
12232	Apple juice	22	11	19	268
21101	Snaps	16	38	16	214
21103	Gin, vodka, rum	16	79	13	259
21104	Whiskey, cognac	15	29	14	199
21211	Red wine	158	155	158	2.178
21212	White wine	59	74	38	778
21221	Vermouth, champagne	32	43	17	357
21222	Port wine, sherry	7	27	3	163
21301	Beer	32	141	85	1.458
21302	Strong beer, cider	25	150	28	348
21303	Light beer	2	19	10	9
22010	Cigarettes	5	23	400	699
22021	Cigars, cigarillos	20	10	5	151
22022	Pipe tobacco	27	8	46	680
22023	Cigarette paper	4	12	3	17

* In addition to this number of observations there are a number of prices collected from specialized stores

Individualized selection criteria

Even though the two selection criteria reduces the scanner data to a sample easier to handle, it does not provide the most desirable amount of observations for each sub-groups nor does it take into account that some sub-groups need individualized selection criteria. Therefore we look at the sub-groups individually. The aim is of course to end up with a sample of a size we can handle and that is representative. This means that at this point anyway we do not want a sample bigger than our current sample.

When looking at the 153 sub-groups in table 1 we see that some of them actually have good representation in the sample with the two selection criteria – 12 months in scanner data and within top 50% of turnover – compared to the current sample and the number of observations dictated by CPI weights. For example, the three sub-groups of cheese, 11441, 11442 and 11443, end up with 15, 22 and 93 observations with the two selection criteria compared to 12, 50 and 90 observations respectively in the current sample. Groups like these where the two criteria provide a number of observations close to the current or close to the number of prices dictated by CPI weights (in total 54 sub-groups) will therefore not be treated further in this next part.

Leaving out the 54 sub-groups where the two selection criteria are sufficient, we are left with 99 sub-groups in need of further treatment. Based on the sample drawn from scanner data using the two selection criteria, these groups can roughly be divided into two categories:

1) Sub-groups with too many observations compared to the current sample and CPI weight.
2) Sub-groups with too few observations compared to the current sample and CPI weight.

For all the 99 sub-groups, the items discarded because they were not available 12 months in the data are assessed in order to determine whether it is reasonable to discard them on the basis of their share of turnover (this will also be done for the 54 sub-groups left out at another time). Luckily, nearly all discarded items have a low share of turnover making it reasonable to discard them in the sample.

Sub-groups with too many observations

The first of the two categories holding 29 sub-groups is the easiest to deal with. The sub-groups' number of observations are chosen from the two-criteria-sample based on highest turnover. This means that each supermarket chain's share of the COICOP sub-group turnover is multiplied with the total number of observations desired for the sub-group, determining the number of observations per chain. Then the observations are chosen based on highest turnover.

Sub-groups with too few observations

Dealing with the second of the two categories holding 70 sub-groups is more complex. The two selection criteria discard too many observations in these sub-groups which means that we have to look at the full scanner data and make individualized selection criteria for these groups.

What we do is that, for each sub-group all of 2011's scanner data is collected and each item's share of the sub-groups total turnover is calculated. Then we examine which criterion we can set for the sub-group for how many months the item is available in data, i.e. the stability of the item. The stability criterion is set so that generally the best-selling items become part of the sample.
For 16 sub-groups a stability criterion – for most of the groups 12 months in data, for other groups less – is enough to get a desired number of observations which in turn are stable and representative. This means that no turnover criterion is applied to these sub-groups. For the rest of the sub-groups the individual stability criterion provides too many observations. Therefore we examine which turnover criterion allows the desired amount of observations. For 16 sub-groups a turnover criterion of 75% - meaning that the observations are within the top 75% of the sub-group total turnover – suffices. For another 16 sub-groups a 90% turnover criterion is required. And yet for two sub-groups a criterion of 60% and 50% respectively is sufficient to get the desired amount of observations. Please note, however, that this does not

necessarily mean that the sub-group sample covers that turnover share since the stability criterion most likely has discarded some observations beforehand.

A few sub-groups have even more specialized selection criteria. For two sub-groups, 11412 Semi-skimmed milk and 11415 Low fat milk, simply the top 3 best-selling items are chosen for each supermarket chain covering almost all sales of the two sub-groups. For the sub-groups, 11215 Ground beef, organic and 11133 Whole grain bread, where the 12 months in data criterion is otherwise applied there are no items from the supermarket chain Rema1000 in data for 12 months. However, two items available in the latest three months of 2011 is available constituting 24% and 100% respectively of the chain's sub-group turnover. These two items are then selected for the sample as well.

Table 2 shows the sample as it is after individualizing the selection criteria for certain COICOP sub-groups as described above.

Seasonal items

The sub-groups marked in grey in table 2 cover items with extensive seasonal patterns. We have not yet determined how to treat such items in the sample. Most likely we will choose items for these sub-groups based on monthly turnover. Furthermore, when the items are no longer available in the data we wish to treat them as we do now, that is we impute the price with the price change of comparable seasonal goods.

Table 2. Number of observations in sample drawn from scanner data of 2011 with individualized selection criteria

COICOP	Description	No. of observations in sample from SD with individualized selection criteria	No. of observations in current sample	No. of observations dictated by CPI weight	Selection criteria
	Total sample	**3.817****	**8.463**	**3.514**	
11110	Rice	12	12	12	12 months in data and 50% turnover produces too many observations. The top-selling items are selected to reach the desired number of observations
11121	Flour	20	20	15	12 months in data and 50% turnover produces too many observations. The top-selling items are selected to reach the desired number of observations
11122	Oats	12	12	10	12 months in data and 50% turnover produces too many observations. The top-selling items are selected to reach the desired number of observations
11131	Rye bread	91	103*	51	Top 3 from each supermarket chain
11133	Whole grain bread	56	104*	19	Top 3 from each supermarket chain
11134	White bread	40	98*	19	12 months in data and 50% turnover
11135	Rolls	83	103*	53	12 months in data and 75% turnover
11136	Flute, sausage bread and pita bread	34	49	15	12 months in data and 50% turnover
11137	Crisp bread	14	12	14	12 months in data and 50% turnover produces too many observations. The top-selling items are selected to reach the desired number of observations
11141	Pastry	32	46*	22	12 months in data and 50% turnover
11142	Cream cakes	62	62	15	12 months in data and 50% turnover produces too many observations. The top-selling items are selected to reach the desired number of observations
11143	Cakes and pies	16	12	21	12 months in data and 50% turnover
11144	Cookies	14	12	14	12 months in data and 50% turnover produces too many observations. The top-selling items are selected to reach the desired number of observations
11145	Biscuits	16	12	16	12 months in data and 50% turnover produces too many observations. The top-selling items are selected to reach the desired number of observations
11150	Spaghetti, pasta and noodles	21	12	21	12 months in data and 50% turnover produces too many observations. The top-selling items are selected to reach the desired number of observations
11160	Cereal	26	49	27	12 months in data and 50% turnover
11170	Pizza etc. frozen	19	31	10	12 months in data and 50% turnover
11211	Ground beef	45	54*	52	12 months in data and 90% turnover. For one chain items are only available in sept.-dec. 2011, these are also selected
11212	Shoulder of beef	6	75*	12	12 months in data and 75% turnover
11213	Beef minced	62	92*	21	12 months in data
11214	Beef tenderloin	96	195*	42	12 months in data and 90% turnover
11215	Ground beef, organic	12	69*	2	12 months in data and 75% turnover
11221	Veal minced	28	57*	7	12 months in data for two chains. Only items available in nov-dec. 2011 for third chain, these are selected also
11222	Veal topside	39	66*	16	12 months in data and 90% turnover
11231	Pork minced	23	95*	3	12 months in data and 90% turnover
11233	Pork loin without bacon	56	61*	31	12 months in data
11234	Pork tenderloin	56	104*	14	12 months in data and 90% turnover
11236	Pork loin with bacon	25	97*	22	12 months in data and 50% turnover
11237	Ground pork	66	111*	20	12 months in data
11238	Ground pork, organic	3	32	0	12 months in data and 90% turnover
11240	Lamb and game meat	22	44	15	12 months in data and 50% turnover

11251	Chicken, fresh and frozen	17	20	18	12 months in data and 75% turnover
11252	Ducks	18	15	16	12 months in data and 50% turnover
11253	Turkey breast	11	77*	8	12 months in data and 90% turnover
11254	Chicken fillet	60	86	40	12 months in data and 75% turnover
11255	Meat salads	11	11	6	12 months in data and 50% turnover produces too many observations. The top-selling items are selected to reach the desired number of observations
11261	Meat offal	24	46*	6	12 months in data and 90% turnover
11271	Cold cuts, ham	16	12	19	12 months in data
11272	Cold cuts, salted meat	7	12	7	12 months in data
11273	Cold cuts, pork fillet	9	12	9	12 months in data and 75% turnover
11274	Prepared dishes	21	13	21	12 months in data and 50% turnover produces too many observations. The top-selling items are selected to reach the desired number of observations
11275	Prepared dishes, cans	22	22	3	12 months in data and 50% turnover produces too many observations. The top-selling items are selected to reach the desired number of observations
11276	Ham	18	114	12	12 months in data and 75% turnover
11278	Cold cuts, mortadella	28	24	28	12 months in data and 50% turnover produces too many observations. The top-selling items are selected to reach the desired number of observations
11280	Liver pate	48	115	36	12 months in data and 90% turnover
11282	Cooked meats	40	110	13	12 months in data and 75% turnover
11283	Sausages and bacon	44	106	47	12 months in data and 50% turnover
11284	Cold cuts, salami	14	12	25	12 months in data and 50% turnover
11311	Cod	9	22*	4	12 months in data and 50% turnover
11312	Flounder	9	90*	7	12 months in data and 90% turnover
11313	Herring fillet	12	6*	5	12 months in data and 90% turnover
11314	Salmon fillet	32	86*	9	12 months in data
11321	Cod fillet, frozen	10	11	5	12 months in data
11322	Flounder fillet, frozen	18	13	3	12 months in data and 50% turnover
11331	Smoked mackerel	3	0*	2	12 months in data and 50% turnover
11332	Smoked salmon, seafood and caviar	18	90	16	12 months in data and 50% turnover
11333	Mackerel, cans	15	23	13	12 months in data and 50% turnover
11334	Cod roe, cans	10	12	6	12 months in data and 50% turnover
11335	Marinated herring	10	11	10	12 months in data and 50% turnover produces too many observations. The top-selling items are selected to reach the desired number of observations
11336	Fish salads	13	12	7	12 months in data and 50% turnover
11338	Stuffed fish fillets	15	67*	12	12 months in data and 50% turnover
11339	Shrimps	12	12	8	12 months in data and 50% turnover produces too many observations. The top-selling items are selected to reach the desired number of observations
11411	Whole milk and infant formula	20	51	14	12 months in data and 90% turnover
11412	Semi-skimmed milk	9	52	20	12 months in data and 60% turnover
11413	Skimmed milk	15	50	14	12 months in data and 75% turnover
11414	Buttermilk	11	47	10	12 months in data
11415	Low fat milk	9	53	18	12 months in data and 75% turnover
11416	Whole milk, organic	7	44	3	12 months in data and 75% turnover
11417	Semi-skimmed milk, organic	8	47	5	12 months in data
11418	Skimmed milk, organic	10	51	8	12 months in data and 75% turnover
11419	Low fat milk, organic	29	50	11	12 months in data and 75% turnover
11431	Whipping cream	38	51	20	12 months in data and 75% turnover
11432	Sour cream	10	46	9	12 months in data and 50% turnover
11433	Yogurt	52	146	52	12 months in data and 50% turnover
11434	Chocolate milk	12	12	11	12 months in data and 75% turnover

11441	Cream cheese	15	12	29	12 months in data and 50% turnover
11442	Brie cheese	22	50	22	12 months in data and 50% turnover
11443	Cheese	93	92	90	12 months in data and 50% turnover
11450	Eggs	15	50	25	12 months in data and 90% turnover
11452	Eggs, organic	10	50	8	12 months in data
11511	Butter	11	53	24	11 months in data
11512	Butter mixture	15	53	21	10 months in data and 90% turnover
11521	Margarine	12	13	8	12 months in data and 50% turnover
11522	Vegetable margarine	5	13	9	10 months in data and 90% turnover
11531	Food oils	12	12	16	12 months in data and 50% turnover
11611	Apples and pears		112*	30	
11612	Citrus fruits		107*	24	
11613	Soft fruits		63*	45	
11614	Bananas		92*	20	
11615	Grapes and melons		112*	25	
11621	Dried fruit	24	24	17	12 months in data and 50% turnover produces too many observations. The top-selling items are selected to reach the desired number of observations
11622	Nuts, almonds	29	45	26	12 months in data and 50% turnover
11630	Canned fruit, frozen berries	21	34	8	12 months in data and 50% turnover
11711	Carrots		106*	11	
11712	Root vegetables		98*	12	
11713	Tomatoes		108*	25	
11714	Cucumber, eggplant, zucchini		94*	17	
11715	Onions		111*	12	
11716	Mushrooms		110*	7	
11717	Lettuce		91*	19	
11718	Peppers		109*	13	
11719	Cabbage		83*	4	
11721	Potatoes		112*	33	
11724	Potatoes, organic		91*	5	
11731	Frozen vegetables	31	71	14	12 months in data and 50% turnover
11732	Potato chips	15	12	8	12 months in data and 50% turnover
11751	Canned vegetables	75	94	25	12 months in data and 50% turnover
11752	Crisps	11	12	11	12 months in data and 50% turnover produces too many observations. The top-selling items are selected to reach the desired number of observations
11753	Roasted onions	3	11	1	12 months in data and 50% turnover
11754	Curry salad, Italian salad	16	24	9	12 months in data and 50% turnover
11791	Cauliflower		98*	21	
11794	Carrots, organic		101*	2	
11810	Sugar	12	12	14	12 months in data and 50% turnover
11821	Jam	24	24	24	12 months in data and 50% turnover produces too many observations. The top-selling items are selected to reach the desired number of observations
11822	Honey	10	12	6	12 months in data and 50% turnover
11831	Chocolate	125	125	95	12 months in data and 50% turnover produces too many observations. The top-selling items are selected to reach the desired number of observations
11832	Candy	121	99	121	12 months in data and 50% turnover produces too many observations. The top-selling items are selected to reach the desired number of observations
11840	Ice cream	61	49	61	12 months in data and 50% turnover produces too many observations. The top-selling items are selected to reach the desired number of observations
11911	Salt	7	12	4	12 months in data and 50% turnover

11912	Spices	22	24	13	12 months in data and 50% turnover
11913	Vanilla sugar	3	12	3	12 months in data and 50% turnover
11914	Herbs	10	10	10	12 months in data and 50% turnover produces too many observations. The top-selling items are selected to reach the desired number of observations
11921	Vinegar	14	11	3	12 months in data and 50% turnover
11922	Mustard	12	12	7	12 months in data and 50% turnover produces too many observations. The top-selling items are selected to reach the desired number of observations
11923	Tomato ketchup	14	12	11	12 months in data and 50% turnover
11924	Ready sauce	23	12	23	12 months in data and 50% turnover produces too many observations. The top-selling items are selected to reach the desired number of observations
11925	Salad dressing	9	12	9	12 months in data and 50% turnover produces too many observations. The top-selling items are selected to reach the desired number of observations
11926	Baking ingredients	23	23	11	12 months in data and 50% turnover produces too many observations. The top-selling items are selected to reach the desired number of observations
11931	Mayonnaise	8	12	6	12 months in data and 50% turnover
11932	Remoulade	10	12	6	12 months in data and 50% turnover
11945	Soups, baby food	20	20	13	12 months in data and 50% turnover produces too many observations. The top-selling items are selected to reach the desired number of observations
12110	Coffee	97	118	87	9 months in data
12120	Tea	17	12	17	12 months in data and 50% turnover produces too many observations. The top-selling items are selected to reach the desired number of observations
12130	Cocoa	10	13	4	12 months in data and 50% turnover
12210	Mineral water	11	11	17	12 months in data and 50% turnover
12221	Soft drinks	109	291	109	9 months in data and 75% turnover
12222	Lemonade	12	12	11	12 months in data and 50% turnover produces too many observations. The top-selling items are selected to reach the desired number of observations
12231	Orange juice	56	50	52	9 months in data and 90% turnover
12232	Apple juice	22	11	19	12 months in data and 50% turnover
21101	Snaps	16	38	16	12 months in data and 50% turnover
21103	Gin, vodka, rum	16	79	13	12 months in data and 50% turnover
21104	Whiskey, cognac	15	29	14	12 months in data and 50% turnover
21211	Red wine	158	155	158	12 months in data and 50% turnover
21212	White wine	59	74	38	12 months in data and 50% turnover
21221	Vermouth, champagne	32	43	17	12 months in data and 50% turnover
21222	Port wine, sherry	7	27	3	12 months in data and 50% turnover
21301	Beer	103	141	85	9 months in data and 50% turnover
21302	Strong beer, cider	102	150	28	8 months in data
21303	Light beer	6	19	10	6 months in data
22010	Cigarettes	46	23	400	6 months in data
22021	Cigars, cigarillos	10	10	5	12 months in data and 50% turnover produces too many observations. The top-selling items are selected to reach the desired number of observations
22022	Pipe tobacco	27	8	46	12 months in data and 50% turnover
22023	Cigarette paper	4	12	3	12 months in data and 50% turnover

* In addition to this number of observations there are a number of prices collected from specialized stores

** The total sample is not including observations for fruit and vegetables, groups marked in grey, as they have not been treated yet.

Drawing the dynamic basket

As mentioned above our sampling system will consist of two baskets – the dynamic basket filtering out items from the total scanner data that may be candidates for the representative basket, and the representative basket which is the sample used for index calculations. When items fall out of the representative basket, either because of missing prices or non-representativeness, new items are to be selected from the dynamic basket. The initial sample described above is aimed to constitute the representative basket. However, some of the selection criteria for the initial sample will be used as filters for the dynamic basket.

Turnover selection criteria for the dynamic basket

The filter for the dynamic basket will be based on turnover. This means that only items with the largest turnover are selected for the dynamic basket and thus are candidates for the representative basket. The turnover selection criteria we are working with at this point are the same as the individualized turnover criteria used for drawing the initial sample presented in table 2. However, the criteria are based on turnover aggregated for the latest four months. This way, the dynamic basket will always contain the most representative items, hence representativeness is ensured in the representative basket. An overview of how many observations are drawn on a monthly basis from scanner data to the dynamic basket is presented in appendix 1.

Conclusion and future work

Even though we have come a long way in the process of integrating scanner data into the CPI, we still have a long way to go.
We have established a system for receiving and storing the scanner data and have a good corporation with the supermarket chains delivering the data. We have established a system for linking the EAN numbers to the COICOP classification making it possible to use the data for CPI calculations. Furthermore, we have decided to use the representative basket methodology when integrating scanner data into the CPI in order to avoid possible bias as well as to have a better opportunity of manual control. We have developed a model for drawing and maintaining the sample and we have made our first experiences on using the model.

The next steps in this process is to find a way to treat seasonal goods in the sample, finish developing the IT-system for maintaining the sample, testing the system, make further examinations of the sub-groups within the CPI and of course make test calculations of the indices.
For seasonal goods we intend to look at turnover on a month-to-month basis in order to ensure representativeness in the sub-groups. With regard to the IT-system especially the way each item in the representative basket is checked for missing prices, representativeness etc. needs to be developed so that it becomes user friendly. Even though we have developed individualized selection criteria as described in this paper, we still have to examine some sub-groups more thoroughly to ensure that the items in the sample cover the variety of products we want. And last but not least on our list to do, we have to do test calculations of the indices based on the sample. Hopefully we will have some good results and better indices, but that is to be presented in another paper.

Appendix 1

Table A1. Number of observations drawn from scanner data from 2011-2012 to the dynamic basket

COICOP	Description	Minimum no. of observations	Maximum no. of observations	Average no. of observations
	Total sample	**4.250**	**6.423**	**5.015**
11110	Rice	18	22	20
11121	Flour	36	49	42
11122	Oats	24	30	26
11131	Rye bread	87	113	92
11133	Whole grain bread	52	66	57
11134	White bread	38	45	40
11135	Rolls	72	107	80
11136	Flute, sausage bread and pita bread	31	41	34
11137	Crisp bread	25	55	36
11141	Pastry	24	32	28
11142	Cream cakes	75	112	89
11143	Cakes and pies	12	26	20
11144	Cookies	18	27	21
11145	Biscuits	35	45	40
11150	Spaghetti, pasta and noodles	46	61	51
11160	Cereal	25	34	27
11170	Pizza etc. frozen	19	28	23
11211	Ground beef	64	92	81
11212	Shoulder of beef	7	14	10
11213	Beef minced	60	106	76
11214	Beef tenderloin	93	122	103
11215	Ground beef, organic	10	22	16
11221	Veal minced	26	43	29
11222	Veal topside	35	62	45
11231	Pork minced	22	41	27
11233	Pork loin without bacon	68	90	82
11234	Pork tenderloin	73	164	120
11236	Pork loin with bacon	24	38	31
11237	Ground pork	113	156	130
11238	Ground pork, organic	3	7	6
11240	Lamb and game meat	19	34	25
11251	Chicken, fresh and frozen	17	30	21
11252	Ducks	7	39	18
11253	Turkey breast	11	33	20
11254	Chicken fillet	65	83	76
11255	Meat salads	16	25	19
11261	Meat offal	22	42	26
11271	Cold cuts, ham	12	19	15
11272	Cold cuts, salted meat	9	12	11
11273	Cold cuts, pork fillet	9	13	11
11274	Prepared dishes	39	56	49
11275	Prepared dishes, cans	20	41	28
11276	Ham	22	41	26
11278	Cold cuts, mortadella	69	87	73
11280	Liver pate	45	65	47
11282	Cooked meats	36	77	50
11283	Sausages and bacon	39	62	50
11284	Cold cuts, salami	13	19	15

11311	Cod	9	14	11
11312	Flounder	11	20	14
11313	Herring fillet	16	29	20
11314	Salmon fillet	29	48	35
11321	Cod fillet, frozen	9	16	11
11322	Flounder fillet, frozen	18	24	21
11331	Smoked mackerel	3	6	4
11332	Smoked salmon, seafood and caviar	15	24	18
11333	Mackerel, cans	11	19	14
11334	Cod roe, cans	8	14	10
11335	Marinated herring	26	36	29
11336	Fish salads	12	17	13
11338	Stuffed fish fillets	15	21	16
11339	Shrimps	20	25	22
11411	Whole milk and infant formula	17	27	23
11412	Semi-skimmed milk	9	12	9
11413	Skimmed milk	12	15	14
11414	Buttermilk	7	14	12
11415	Low fat milk	9	12	9
11416	Whole milk, organic	7	10	8
11417	Semi-skimmed milk, organic	9	11	10
11418	Skimmed milk, organic	9	15	11
11419	Low fat milk, organic	35	55	40
11431	Whipping cream	37	45	38
11432	Sour cream	9	13	10
11433	Yogurt	50	68	53
11434	Chocolate milk	10	17	14
11441	Cream cheese	13	27	18
11442	Brie cheese	21	29	22
11443	Cheese	90	126	96
11450	Eggs	13	23	16
11452	Eggs, organic	13	19	16
11511	Butter	11	18	13
11512	Butter mixture	11	19	14
11521	Margarine	11	17	12
11522	Vegetable margarine	5	7	6
11531	Food oils	11	18	13
11611	Apples and pears			
11612	Citrus fruits			
11613	Soft fruits			
11614	Bananas			
11615	Grapes and melons			
11621	Dried fruit	42	63	46
11622	Nuts, almonds	26	41	29
11630	Canned fruit, frozen berries	19	26	22
11711	Carrots			
11712	Root vegetables			
11713	Tomatoes			
11714	Cucumber, eggplant, zucchini			
11715	Onions			
11716	Mushrooms			
11717	Lettuce			
11718	Peppers			
11719	Cabbage			

11721	Potatoes			
11724	Potatoes, organic			
11731	Frozen vegetables	28	39	32
11732	Potato chips	14	21	16
11751	Canned vegetables	60	79	72
11752	Crisps	53	69	58
11753	Roasted onions	3	4	3
11754	Curry salad, Italian salad	14	18	15
11791	Cauliflower			
11794	Carrots, organic			
11810	Sugar	9	13	10
11821	Jam	56	71	66
11822	Honey	9	13	11
11831	Chocolate	137	205	151
11832	Candy	205	272	232
11840	Ice cream	59	90	80
11911	Salt	7	11	8
11912	Spices	17	31	20
11913	Vanilla sugar	3	4	3
11914	Herbs	32	69	40
11921	Vinegar	11	19	13
11922	Mustard	17	24	19
11923	Tomato ketchup	14	18	15
11924	Ready sauce	69	92	75
11925	Salad dressing	24	36	30
11926	Baking ingredients	33	52	38
11931	Mayonnaise	7	9	8
11932	Remoulade	8	12	9
11945	Soups, baby food	42	61	50
12110	Coffee	93	119	104
12120	Tea	54	75	63
12130	Cocoa	9	13	9
12210	Mineral water	11	17	12
12221	Soft drinks	105	145	127
12222	Lemonade	23	33	26
12231	Orange juice	53	78	57
12232	Apple juice	17	31	23
21101	Snaps	15	20	16
21103	Gin, vodka, rum	14	23	16
21104	Whiskey, cognac	13	22	15
21211	Red wine	137	173	148
21212	White wine	51	70	60
21221	Vermouth, champagne	26	42	30
21222	Port wine, sherry	5	15	7
21301	Beer	95	193	126
21302	Strong beer, cider	87	155	104
21303	Light beer	6	14	9
22010	Cigarettes	29	49	38
22021	Cigars, cigarillos	16	24	18
22022	Pipe tobacco	43	86	57
22023	Cigarette paper	4	6	4

* Selection criteria for fruit and vegetables, groups marked in grey, have not been established yet.