Implementing scanner data in the Danish CPI - Paper to be presented at UN- Group of experts on the consumer price indices meeting

Geneve May 2014

Workshop 4- Scanner Data

Table of contents

Introduction

This paper tries to outline Statistic Denmark's system for the reception, processing and index calculation of COICOP groups 1 and 2 based on scanner data. The plan is for Statistics Denmark to use scanner data in production from January 2015 onwards.

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, which is to be used for the calculation of the CPI.

The supermarket chains in question are:

Supermarket	**Weekly reception of scanner data since**	**Approx. sales percentage of the total Danish Food and beverage market**	**Supermarkets product structures and store info**
Chain 1	January 2011	25%	Both
Chain 2	January 2011	25%	Both
Chain 3	January 2011	10%	Both

The received scanner 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 following gives an example of the structure of the scanner data:

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

The *date* is 4 digits and consists of a 2 digit year number and a 2 digit week number, the *store number* is a unique number for the specific supermarket store in which the item is sold. The price of the item is derived from dividing the weekly *turnover* with the weekly *volume* for each EAN number. Finally the *product number* can be used to reflect 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.

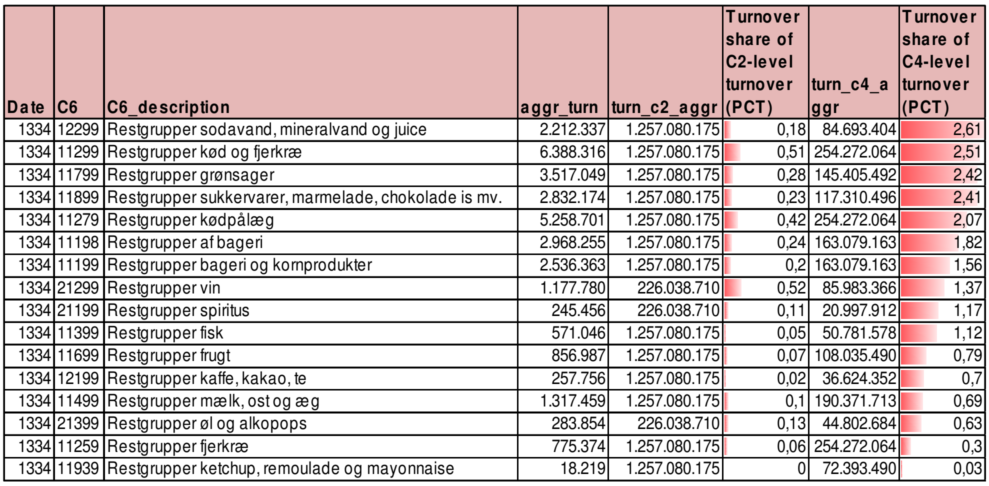
IT-system for maintenance of the link between EAN/PLU and COICOP

We have created an IT-system that supports the weekly maintenance of the link between the codes in the scanner data and COICOP div. 1 and 2. The system is made so that the supermarket chains weekly delivered scanner data are processed on a week to week basis.

The system is based on 3 levels of interference on the scanner data:

1. The use of the supermarket chains own classifications manually assigned to COICOP. This manual assignment is only done once and only changed if the supermarket chain decides to change their classification.
   Where the supermarket classification is not precise enough (e.g. when covering several 6-digit COICOP categories) the classification is assigned to a special residual 4-digit COICOP group. These residual groups are monitored and handled by a search-word process.

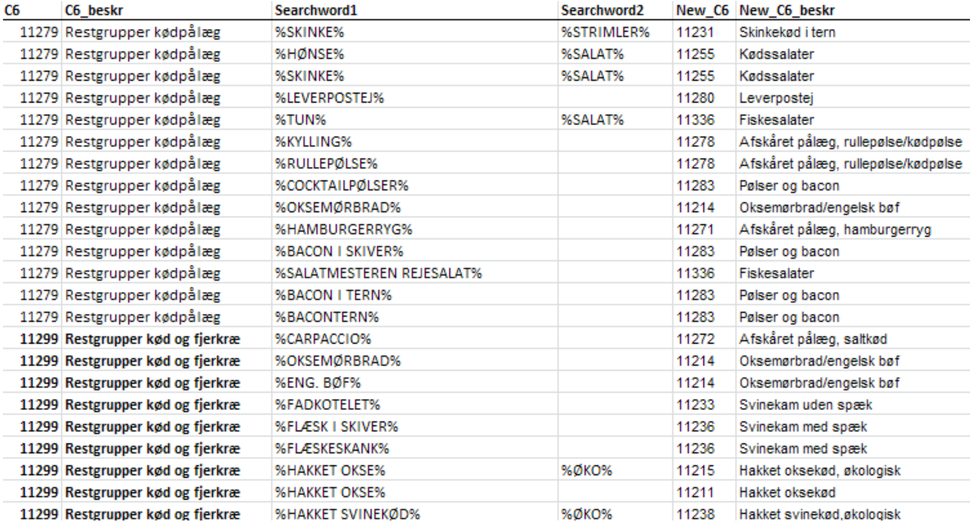
2. The search-word process starts by monitoring the turnover from sold products in the residual groups and their ratio to the turnover of the already assigned COICOP 6-digit groups aggregated to the 4-digit level. In that way the residual groups are evaluated and prioritized for the production team (see illustration below):



Date	C6	C6_description	aggr_turn	turn_c2_aggr	Turnover share of C2-level turnover (PCT)	turn_c4_aggr	Turnover share of C4-level turnover (PCT)
1334	12299	Restgrupper sodavand, mineralvand og juice	2.212.337	1.257.080.175	0,18	84.693.404	2,61
1334	11299	Restgrupper kød og fjerkræ	6.388.316	1.257.080.175	0,51	254.272.064	2,51
1334	11799	Restgrupper grønsager	3.517.049	1.257.080.175	0,28	145.405.492	2,42
1334	11899	Restgrupper sukkervarer, marmelade, chokolade is mv.	2.832.174	1.257.080.175	0,23	117.310.496	2,41
1334	11279	Restgrupper kødpålæg	5.258.701	1.257.080.175	0,42	254.272.064	2,07
1334	11198	Restgrupper af bageri	2.968.255	1.257.080.175	0,24	163.079.163	1,82
1334	11199	Restgrupper bageri og kornprodukter	2.536.363	1.257.080.175	0,2	163.079.163	1,56
1334	21299	Restgrupper vin	1.177.780	226.038.710	0,52	85.983.366	1,37
1334	21199	Restgrupper spiritus	245.456	226.038.710	0,11	20.997.912	1,17
1334	11399	Restgrupper fisk	571.046	1.257.080.175	0,05	50.781.578	1,12
1334	11699	Restgrupper frugt	856.987	1.257.080.175	0,07	108.035.490	0,79
1334	12199	Restgrupper kaffe, kakao, te	257.756	1.257.080.175	0,02	36.624.352	0,7
1334	11499	Restgrupper mælk, ost og æg	1.317.459	1.257.080.175	0,1	190.371.713	0,69
1334	21399	Restgrupper øl og alkopops	283.854	226.038.710	0,13	44.802.684	0,63
1334	11259	Restgrupper fjerkræ	775.374	1.257.080.175	0,06	254.272.064	0,3
1334	11939	Restgrupper ketchup, remoulade og mayonnaise	18.219	1.257.080.175	0	72.393.490	0,03

If a residual group (restgrupper) has more than 5 % of the weekly turnover compared to the corresponding COICOP 4-digit group it has been decided to apply the search-word process.

The search-word assigning process is initiated by a list of the actual EANs in the residual group which are sorted by importance (turnover). From this list the product descriptions must be "translated" into suitable search-words that can identify the products on the 6-digit COICOP level. This is only done once. Future EANs will be processed by the actual search-word list and will therefore not show up in the residual groups anymore. Hence, the production team's work is stored properly and automatically applied in future COICOP assignments (see the search word list illustration below):



C6	C6_beskr	Searchword1	Searchword2	New_C6	New_C6_beskr
11279	Restgrupper kødpålæg	%SKINKE%	%STRIMLER%	11231	Skinkekød i tern
11279	Restgrupper kødpålæg	%HØNSE%	%SALAT%	11255	Kødssalater
11279	Restgrupper kødpålæg	%SKINKE%	%SALAT%	11255	Kødssalater
11279	Restgrupper kødpålæg	%LEVERPOSTEJ%		11280	Leverpostej
11279	Restgrupper kødpålæg	%TUN%	%SALAT%	11336	Fiskesalater
11279	Restgrupper kødpålæg	%KYLLING%		11278	Afskåret pålæg, rullepølse/kødpølse
11279	Restgrupper kødpålæg	%RULLEPØLSE%		11278	Afskåret pålæg, rullepølse/kødpølse
11279	Restgrupper kødpålæg	%COCKTAILPØLSER%		11283	Pølser og bacon
11279	Restgrupper kødpålæg	%OKSEMØRBRAD%		11214	Oksemørbrad/engelsk bøf
11279	Restgrupper kødpålæg	%HAMBURGERRYG%		11271	Afskåret pålæg, hamburgerryg
11279	Restgrupper kødpålæg	%BACON I SKIVER%		11283	Pølser og bacon
11279	Restgrupper kødpålæg	%SALATMESTEREN REJESALAT%		11336	Fiskesalater
11279	Restgrupper kødpålæg	%BACON I TERN%		11283	Pølser og bacon
11279	Restgrupper kødpålæg	%BACONTERN%		11283	Pølser og bacon
11299	**Restgrupper kød og fjerkræ**	%CARPACCIO%		11272	Afskåret pålæg, saltkød
11299	**Restgrupper kød og fjerkræ**	%OKSEMØRBRAD%		11214	Oksemørbrad/engelsk bøf
11299	**Restgrupper kød og fjerkræ**	%ENG. BØF%		11214	Oksemørbrad/engelsk bøf
11299	**Restgrupper kød og fjerkræ**	%FADKOTELET%		11233	Svinekam uden spæk
11299	**Restgrupper kød og fjerkræ**	%FLÆSK I SKIVER%		11236	Svinekam med spæk
11299	**Restgrupper kød og fjerkræ**	%FLÆSKESKANK%		11236	Svinekam med spæk
11299	**Restgrupper kød og fjerkræ**	%HAKKET OKSE%	%ØKO%	11215	Hakket oksekød, økologisk
11299	**Restgrupper kød og fjerkræ**	%HAKKET OKSE%		11211	Hakket oksekød
11299	**Restgrupper kød og fjerkræ**	%HAKKET SVINEKØD%	%ØKO%	11238	Hakket svinekød,økologisk

3. The third step deals with the COICOP-key quality. The different EAN-COICOP connections between the different supermarket chains should be the same and is checked for consistency. Secondly in particular the smallest supermarket chain has a very broad classification uses classifications from the other supermarket chains to the extent that the supermarkets have similar EANs.

This process produces an EAN-COICOP key for each week of scanner data.

It is furthermore noted that a manual check system for the rightful distribution of the EANs to the COICOP has been put in place. Thereby it is possible to manually alter the COICOP-group assigned to an EAN. This system is based upon Excel and SAS.

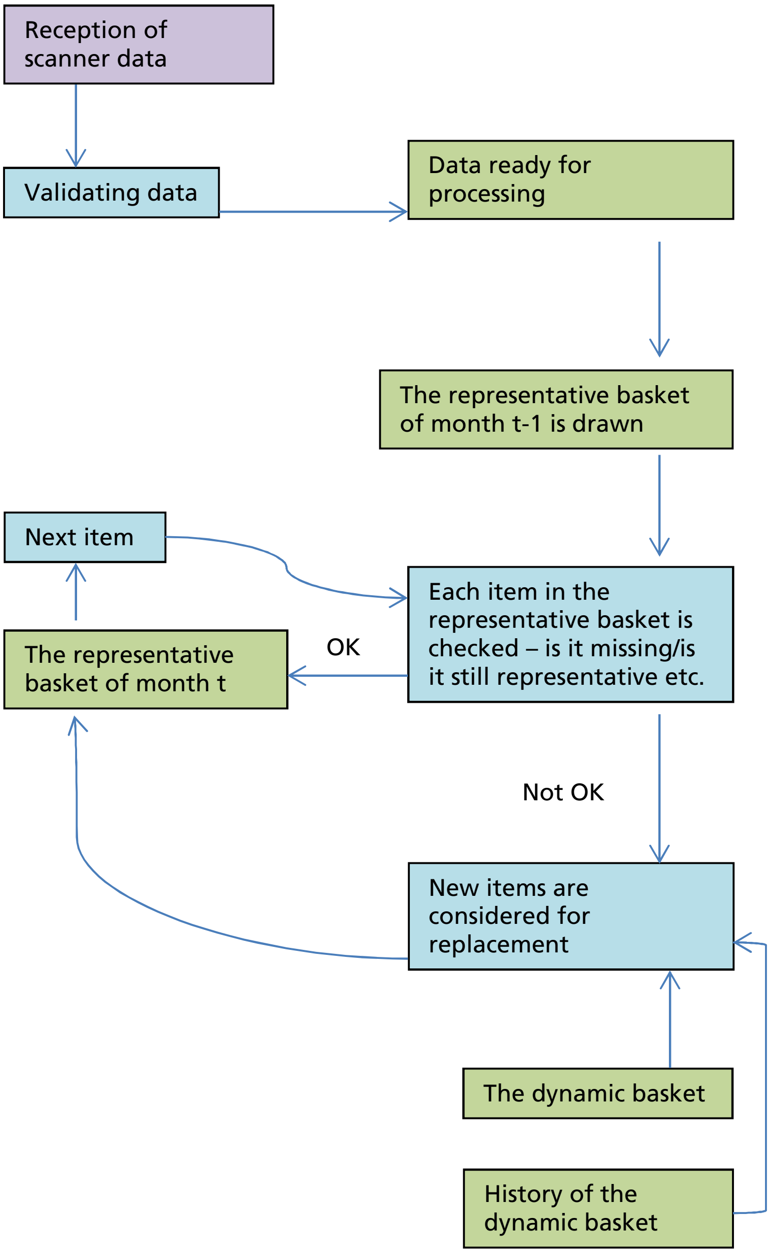
Finally it is noted that PLU-codes in general are treated like EAN-codes. However for some Coicopgroups (especially meat), one PLU-code may cover several amounts of one good. Therefore it is decided to convert the turnover into the turnover for 1 kilo, litre or piece of a good for these goods. In doing this, unrealistic disaggregated unit prices are furthermore removed.

When the weekly key has been applied to the weekly turnover data the scanner data is ready for data processing, data selection and index calculations.

IT system for drawing and maintaining a representative basket using scanner data

Statistics Denmark has decided to use a representative basket methodology for the calculation of the CPI/HICP. This is in contrast with a full population method, which is prone to drift and bias problems due to difficulties in taking proper account of seasonal goods and goods on discount leaving the sample.

The production system is outlined in the following:


Reception of scanner data
Validating data
Data ready for processing
The representative basket of month t-1 is drawn
Next item
Each item in the representative basket is checked – is it missing/is it still representative etc.
OK
The representative basket of month t
Not OK
New items are considered for replacement
The dynamic basket
History of the dynamic basket

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 and store level limiting the amount of data considerably.

When selecting items for the representative basket we realise that no single selection criteria will fit all 154 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 and their stores (the major types of stores) respectively. Due to major differences in the sizes of the chains (two of the chains are much larger than the third in terms of turnover) looking at items constituting top 50% of turnover within their sub-group overall, i.e. without the chain level and store level, would not ensure representation of all chains in the sample. By looking at the top 50% within each supermarket chain and store we make sure that all three chains are represented in the sample.

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.

This means that we are left with two categories in need of further treatment:

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.

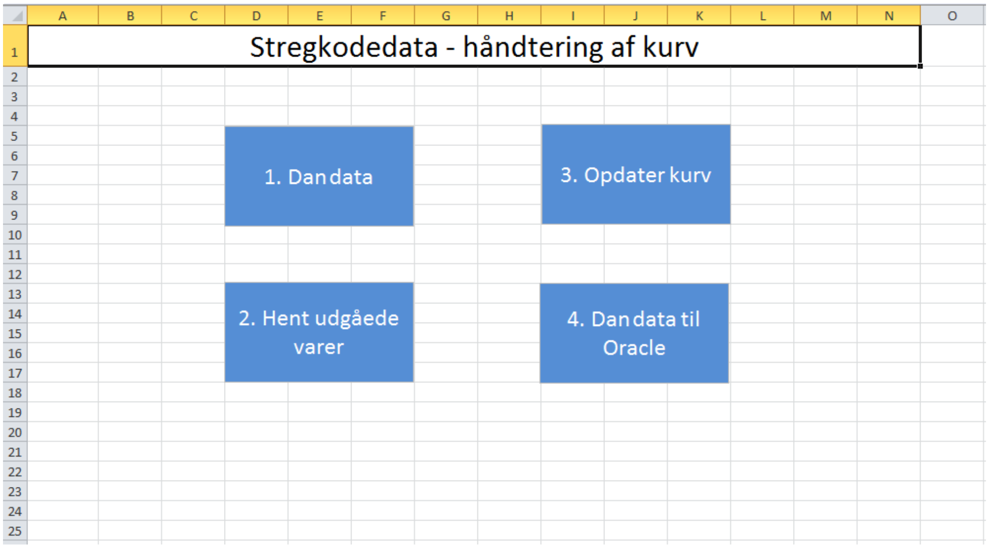
The first of the two categories is the easiest to deal with. The sub-groups' number of observations are chosen from the two-criteria-sample based on highest turnover. This means that each supermarket chain's (store level) share of the COICOP sub-group turnover is multiplied with the total number of observations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

Dealing with the second of the two categories 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.

Maintaining the sample

As for the system for maintaining a representative basket, It-staff have set up an Excel-SAS interface system which is to be monthly maintained by the HICP-section. For the different COICOP subgroups the section is presented with goods that have entered the scanner data, and that are to replace EANs, that are no longer present (generally the EANs that are new candidates for the representative basket cover the last 4 months and they are sorted by shops and highest turnover).



Stregkodedata - håndtering af kurv

The system is based on four pillars:

1. The section sets the desired period for which the replacement of scanner data should be carried out
2. Goods that are no longer present are then fetched and replaced
3. The basket is then updated
4. Finally the data that are to be implemented in the Oracle-production-system are the formed.

It should be noted that step number 3 is only done after step 2 has been thoroughly done.

This second step is done by taking every EAN, that is a possible candidate for the representative basket, and judging whether it is suited, based on (relative) turnover, product description etc.

These two systems, the representative basket (starting in December 2011) and the EXCEL-SAS interface system to maintain the representative basket, have then been implemented in our ORACLE-system for the production of the CPIs/HICPS.

It is noted that the ORACLE-system has been implemented with semi-automatic check-systems. It is thereby possible for the HICP-section to check whether scanner data have the right amounts, taxes for the HICP-CT etc.

The system is currently in a beta-version. This means that remaining possible errors are still to be removed.

IT-system capable of producing pilot HICPs with the developed representative basket methodology using scanner data

The plan is, when the errors from the ORACLE-production-system have been eliminated, for Statistics Denmark to be able to, on a test basis, calculate the CPI with the use of scanner code data in parallel with the ongoing monthly production of the existing CPI, from mid-2014 and the rest of the year. This calculation will also include calculation of historical CPIs with all the data going back to 2011.

Work plan and challenges

The following table summarizes the work plan for the implementation of the scanner data:

Work process	*Expected deadline*
Finalization of IT-system	1. July 2014
Investigate when at fourth chain will be able to deliver data on a weekly basis	1. July 2014
Test calculations of CPI with the use of scanner data	1. July 2014-31. December 2014
Binding agreements with Supermarkets on delivering scanner data	1. August 2014
Calculation of historical CPI´s back to 2011 using the scanner data	1. September 2014
Decision on suitability of CPI/HICP calculated with the use of scanner data	1. September 2014
Final decision on whether scanner data will be implemented in the CPI from January 2015	10. September 2014
Fully implemented system	1. Jan 2015-

It should be noted, that the CPI-section has been in contact with a fourth chain with regard to the possible delivery of scanner data. Since the chains' shops currently accounts for ca. 20 % of the turnover of food and drinks, it will improve the CPI/HICP, if they are to be included. Though they have generally been positive towards the possibility of delivering data, and have supplied a couple of weeks of test-data, no weekly delivering is in place. In 2014 we will try to get a final agreement with the fourth chain on when to start delivering weekly scanner data to us. In any case it is has been decided that we will not be able to implement the data from the fourth chain in the CPI from January 2015, but January 2016 could be a target.

The weekly delivery of scanner data from the 3 current chains is based on informal agreements. Since the calculation of the CPI is to be based on scanner data, it is necessary to establish contact and agree on written contracts, so as to ensure that Statistics Denmark will continue to receive the scanner data on a weekly basis. The works of developing these contracts have been set in motion.

Finally it is noted that according to the current contract for the collection of prices with Wilke A/S, they have to be notified 3 months in advance before the reduction of their price collection due to substitution of currently collected data with scanner data. It is therefore, by the 10 of September 2014, to be agreed whether the CPI can, on a reasonable basis, be calculated with the use of scanner data from January 2015 onwards.

Closing remarks

We recommend implementing scanner data in the CPI from January 2015 onwards if the IT-systems are finalized and running without problems. As described in the paper we have decided to only use a sample drawn from the scanner data. This method closely resembles the methods used today. For instance all product replacements in the sample will be watched and decided on manually. If the IT-systems are ready, then the change to scanner data should be smooth and the implementation should not be at stake.