Statistiska centralbyrån
Statistics Sweden

PCA/MFOS
Anders Norberg

Sampling of Scanner Data Product Offers in the Swedish CPI

1 Introduction / summary

For about 30 years, Statistics Sweden, SCB, has used probability samples of specific products for daily necessities except for fresh food such as vegetables, fruits, fish and meat. The advantages of probability sampling in general are well known, as the method has a strong scientific basis. A problematic point however, is one that is also present for scanner data: the risk that price changes are hidden in the index if sampled products cannot be replaced in price collection due to strict sampling methodology aspects. Consequently, a method is used to replace products that no longer are available on the market.

The Swedish CPI uses sales data to construct sampling frames of products. CPI is provided with aggregated retail sales from all outlets of the three major retail groups annually, based on retail scanner data. Such data are estimated to be 80% of all goods sold in supermarkets, not including fresh foodstuffs like vegetables and meat. Since 2001, SCB produces the annual statistics "Food Sales in the trade" based on scanner data and a questionnaire survey of fresh food sales, which complies with National Accounts and KPI's needs. These statistics are a high quality complement to household budget surveys, which suffer from severe bias due to missing data.

Sampling frames based on scanner data have first undergone a comprehensive coding of more than 200 000 different products into CPI product groups. The first step is a mechanical coding, based on nomenclatures of each outlet chain. Thereafter, the products are reviewed and coded manually, using Pareto principles for prioritizing the work.

Three samples are drawn by sequential Pareto πps selection within strata. Negative coordination of samples between the three groups of outlet chains is accomplished through the permanent random numbers (PRN) that are used. Positive coordination is applied over years.

2 Sampling of outlets

2.1 Allocation

Allocation of the outlet sample is done for each industry stratum. A measure of intermediate outlet variance is estimated. The sizes of the product samples are hence assumed to be fixed. This means that the variation that exists between products manifests itself in variation between outlets. Since the design does not currently allow a "total allocation", where sizes of both outlet sample and product sample are determined simultaneously, this assumption is reasonable. Arvidsson (2004) shows that if it wasn't practically unsuitable, the sample sizes for products would generally be larger and the sample sizes for outlets smaller in CPI.

The estimation of price changes involves stepwise aggregation, both of outlets and of products via product sub-groups to product group and then to CPI in total. Weights are present in all the aggregations. A non-parametric method has been used to estimate the variance (Jack-Knife) There are about 800 outlets in the sample. The same number of estimates of the price trend from December the previous year to the reference month as there are outlets in the sample is calculated. In each such estimate, one outlet has been removed from the data material.

Let y_{hi} be the estimate of the total price change when the outlet / associated outlet's industry stratum *h* is not included in the calculation. Let n_h be the number of outlets in stratum *h*. Let *y* be the corresponding estimate with all the outlets included in the calculations.

Establish a measure of within stratum variation: $d_{hi} = (y_{hi} - y) \cdot n_h / k$

where *k* is the sum of the product group weights for the interviewer systems, about 50%. The division means that the effect on total CPI of removing one outlet will be calculated.

The variation measure is $a_h^2 = \frac{1}{n_h - 1} \sum_{i \in h} (d_{hi} - \bar{d}_h)^2$

Sample size is determined by

$$n_h \propto a_h / \sqrt{c_h}$$

where c_h is the calculated cost per outlet on average, give the product sample that is to be searched in the outlets in the industry stratum. The sample sizes are determined so that the total cost for the interviewer systems will be the same as the year before or adapted to the budget.

2.2 Frame population

The sample is drawn within the framework of the economic statistics sampling system, SAMU, see Statistics Sweden (2008). The SAMU method is based on the sample frame being given a variable for permanent, uniformly distributed random numbers in the interval (0.1), known as permanent random numbers, PRN. New units, births, are given new PRNs, randomly, uniformly distributed and independent of numbers that already exist. Discontinued units disappear from the frame completely.

In SAMU, about 20 percent of the sample is rotated using a method called RRG, Random Rotation Group. Every unit in the sample frame is not just allocated a PRN but also randomly allocated one of five RRG numbers, 1-5. In Year 1, the PRN number for units in RRG 1 is decreased by 0.1, while the PRN numbers that then become negative are simultaneously increased by 1.0 so that they are again in the interval (0.1). In Year 2, the PRN number for units in RRG 2 is decreased by 0.1 and the PRN numbers than then become negative are simultaneously increased by 1.0. After five years, all PRN number have been decreased by 0.1 or increased by 0.9. The small units that have an inclusion probability of less than 10 percent will in all likelihood be in the samples for a

maximum of five years, while larger enterprises may be in them several years running.

The PRN numbers for outlets that are to be measured using cash register data are not to be rotated. There is no reason to successively replace the outlets. The reason why samples in Statistics Sweden business statistics are rotated is to relieve the burden on respondents. That reason does not exist when using cash register data. Rotation is detrimental to the accuracy of the change estimates.

2.3 Sampling method

The outlet sampling method is what is known as rotated, stratified, sequential Poisson sampling with inclusion probabilities proportional to the size of each respective outlet, according to Ohlsson (1990).

For every unit in the sample frame, the quotient between the size measurement and the allocated PRN is calculated. The sample frame is sorted by stratum and within the stratum according to these quotients in descending order. The desired gross sample is made up of the first four units in each stratum respectively in the number requested.

The gross sample is to be large enough so that after purging for overcoverage, we have at least the desired net sample size for most of the strata. Gross samples that are too large should not be ordered since the inclusion probabilities for the net sample will then not be correctly proportional to the size. In particular, we will obtain too many units selected with certainty. The size of the gross samples should be determined following analysis of gross samples used in previous years to achieve the desired net sample.

The result of the net sample is 47 industries containing a total of 866 establishments. Some industries are supplemented with postal order/online sales that are not drawn from the outlet sample and the occasional industry in the outlet sample is examined centrally instead of locally in the outlet.

For the strata hypermarkets and supermarkets there is a need to select two samples each, one according to regular SAMU methods for manual price collection in outlets and one without rotation for scanner data. SCB still need to collect prices for fresh food by manual collection at visits.

Table - Overview of outlet sample where daily necessities are priced

Stratum description	Industry (NACE)	Sample size Net	Collection method
Hypermarkets, broad assortment	47111	7	Scanner data
Hypermarkets, broad assortment	47111	9	Visit
Supermarkets with broad assortment	47112	52	Scanner data
Supermarkets with broad assortment	47112	32	Visit
Tobacconists	47260	5	Telephone
Health food shops	47291	5	Telephone
Pharmacies	47730	9	Visit
Pet shops	47762	3	Telephone

3 Sampling of scanner data products

3.1 Target population

"Daily necessities" mean food, beverages, tobacco, household maintenance products, personal hygiene articles, etc. Since 2012, outlet scanner data has been used for many "daily necessities" instead of data collected on outlet visits. For perishable fruit, vegetables, fish and meat, prices are still collected on visits to hypermarkets and supermarkets. Tobacco is also measured in tobacconists, health food in health food shops, hygiene articles in pharmacies and pet food in pet shops. All these products are daily necessities. It is therefore relevant to define the volume of products for which the prices are measured using scanner data to some degree at least, even though they are also measured on personal visits to some outlets.

Scanner data are mainly used in COICOP groups 01 (Food except for perishable fruits, vegetables, fish and meat) and 02.2 (Beer and tobacco) and some coverage in COICOP 02.1.3, 05.5, 05.6, 06.1, 09.3 and 12.1. The population is defined by the product groups given in Appendix 1.

3.2 Food Sales in the trade

Swedish Board of Agriculture (SJV) ceased its calculations of food consumption from accounting year 2000. SJV has made calculations of food consumption since the 1940s. Since 2001, SCB produces Food Sales in the trade, based on scanner data and a questionnaire survey of fresh food sales, which complies with National Accounts and KPI's needs. This statistics replace the Household Budget statistics for food and beverages.
http://www.scb.se/en_/Finding-statistics/Statistics-by-subject-area/Trade-in-goods-and-services/Domestic-trade/Food-sales/

3.3 Frame populations

Every year Statistics Sweden receives annual statistics from twelve hyper/super/ minimarket chains. The statistics refer to sales per EAN products during the latest calendar year. The purpose is to help Statistics Sweden in producing both the abovementioned Food Sales and the CPI. The twelve registers are grouped into three groups depending on main wholesaler. The three thus collapsed registers contain around 100 000 – 200 000 records each. Many of the records concern other products than those belonging to the target population for scanner data. Nevertheless, this must be established. After many years of coding product group code to products, there is a key available from last year. Moreover there is a procedure, described in detail below, to set product group codes to new products, based on the nomenclatures of each retail chains own products groups. After having extracted products belonging to the target population the three files now contain approx. 50 000 – 100 000 records each.

For the CPI year *y* the sales statistics concern the year *y-2*. There is both over- and under-coverage in the files. By comparing the registers with scanner data for individual outlets at the end of year *y-1*, we get an estimated overcoverage of approximately 25 percent in value. These coverage problems are due to vanishing products from year *y-2* to November year *y-1*, when the sampling frame is being prepared. The under-coverage can be a bit lower as some of the overcoverage consists of products leaving the market early in year *y*, i.e. almost 24 months before the date of sampling frame preparation. However, under-

coverage is caused by new products appearing in the market during the 12 months year *y-1*. Unweighted, the vanishing and new products constitute approximately 70 percent.

Note: Statistics Sweden should consider adding new products that have entered the market by November year y-1. This requires an additional process of product group coding which is time consuming. To save time addition of product could be restricted to big seller.

The scanner data for daily necessities cover 94 product groups of the Swedish CPI. Four of these are fresh vegetables and are more or less an experiment.

3.4 Stratification and product group coding

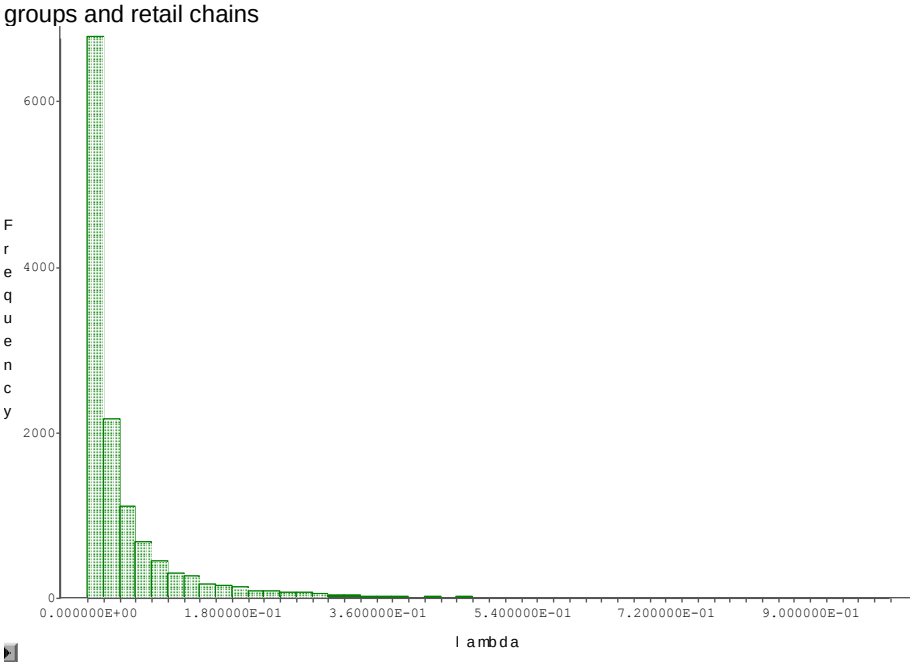
Statistics Sweden and each outlet chain classify the product varieties in different ways[1]. The registers from the trade are large and it is not possible to classify by machine or by hand all the product varieties in the registers in an entirely reliable way. Since statistics are to be created that do not necessitate this process to be perfect but to maintain coding error on such a level that is acceptable in relation to other sources of error and the total quality of the statistics.

We want to set COICOP code but also the more detailed Swedish CPI product group code.

- First automatic coding is made independently for all retail chains, using the product classification nomenclatures of each retail chain. For some products this is made successfully, for others with poor result.
- Secondly we have a SAS-script that searches for parts of product names to improve the automatic coding.
- Thirdly the files with preliminary codes for the chains are joined by EAN, making comparisons of product group code possible. The Pareto principle is used for products with ambiguous coding in the chains, thus sending big sellers to manual expertise coding. The most frequent preliminary code is applied for the products of less value.

[1] The companies that currently manage and analyse EAN codes have also defined product groups. We have yet to discuss with them whether they have used EAN codes to categorise the product varieties in a way that might be of interest to Statistics Sweden.

Figure 1 Distribution of turnover / inclusion probabilities for products in one of the three groups and retail chains



3.5 Sampling method

The sampling is positively coordinated over years. Thus the samples are as current as possible (considering coverage problems) and at the same time changed as little as possible from last year. Three stratified samples of product variations are drawn with inclusion probabilities proportional to turnover, a sample for each block in the supermarket trade. The samples for the three blocks are negatively coordinated.

Since it becomes a question of a mixture of "statistical methods" and "practical procedures" methods are still described from A to H. The main method is by sequential Pareto πps selection within strata. Co-ordinations of samples between the three groups of outlet chains and between years are accomplished by permanent random numbers (PRN).

Procedure

A. Match the year's sample frame with last year's register, by EAN, and apply the permanent random number (PRN), which was stored for each item in the register at that time.

B. A uniformly distributed random number (PRN) between zero (0) and one (1) is created for new items. It is easiest to allocate independent random numbers to the products that are to have one. Statistics Sweden has chosen the following method to create negative coordination between similar products:

B.1 Count the number of new products per stratum *h* that have the same brand *b*, and call it n_{hb}.

B.2 Order the new products with stratum *h* and brand *b* randomly and allocate them a serial number.

B.3.. Assign a new product a serial number *r, r=1* – n_{hb} a PRN in the interval $(r-1)/n_{hb} - n_{hb} - r/n_{hb}$

C. Assign the PRN to group number one, regardless if the product is sold by that group of outlets. Assign another permanent random number to group two as

PRN+1/3 or PRN-2/3, whichever is in the interval 0 – 1. Assign a third random number to group three as PRN+2/3 or PRN-1/3, whichever is in 0 – 1. This will make the three samples negatively coordinated, i.e. as different from each other, as possible.

From here on, the selection procedure is performed per group of outlet chains.

D. Compute target inclusion probabilities $\lambda_{hi} = n_h \cdot s_{hi} / \sum_{j=1}^{N_h} s_{hj}$

where n_h is the desired inclusion probability in product group (stratum) h and

s_{hi} is the size (turnover during year *y-2*) for product *hi*, *i*=1, 2, ..., N_h in stratum *h*=1, 2, ..., L. If $\lambda_{hi} > 1$ then let. $\lambda_{hi} = 1$

E. A purposive cut-off is implemented by deletion of products with $\lambda_{hi} < 0.01$, thus deleting a mass of more than half of all products, but only five percent of the total turnover. The distribution is very skewed as shown for one group of chain:

F. Calculate the value of the ranking variable $Q_{hi} = \frac{U_{hi} \cdot (1-\lambda_{hi})}{\lambda_{hi} \cdot (1-U_{hi})}$, i=1, 2, ..., N_h and stratum h=1, 2, ..., L, where U_{hi} is the PRN

G. Sort the register by stratum (h) and ranking variable Q_{hi}.

H. The n_h first items relating to CPI-measurable product variations are selected for each stratum.

I. The chosen items are scrutinised at a few visits to well-stocked outlets. If the net sample is insufficient, further products can be chosen, in the order they are in the sorted sample frame.

<u>Note:</u> Statistics Sweden finds great interest in the production of the annual statistics "Food Sales in the trade", based on scanner data. There is a great synergy effect to reuse the sales files, coded with COICOP and CPI product groups. If this had not been the case, we would consider to select big samples of products without stratification, say 5 000 - 10 000 products with sequential Pareto πps sampling. These samples are then coded with COICOP and CPI product group codes, a much smaller workload than the coding of all records. Now the sample of desired sizes can be extracted.

3.6 Weights

Since the products chosen are to be weighed proportional against the turnover when aggregating indices for product groups, we have chosen those with inclusion probabilities proportional to the turnover, the samples will be self-weighted. The exception is products that have been chosen with certainty. These are given weights

$$n_h \cdot \ s_{hi} / \sum_{j=1}^{N_h} s_{hj}$$

Varying weights are obtained even for a small number of strata that have been divided into sub-strata.

4. Short on use of scanner data information for daily necessities

4.1. Comments on two-dimensional sample design

The sample of products is relatively large: 800 products. The design of common product samples for all outlets in the outlet sample per block is called two-dimensional sampling. This design gives greater variance in the price trend estimation than with corresponding sample sizes and two-step samples, i.e. with independent product samples per outlet. The two-dimensional design works least well when it is not the retailers (the outlets) who determine the consumer prices without the manufacturers/suppliers and when these see to it that prices are changed in a uniform manner at the same time over the entire country and for all types of outlets.

4.2. Disappearing products in the market (Attrition)

We do a "Market analysis" when a product show a sharp reduction in sales from week to week in total over the sample of outlets. Approximately or less than 30 products in the πps sample of two thousand products leave the sample each month. Approximately one third of these can be replaced by products of the same brand and nearby size etc. Quality adjustment is made whenever quantity is changed. Two thirds of the products are not replaced, but deleted from the annual sample. We consider the gains in spending more resources on trying to replace products.

4.3. Variance contribution

Now that a lot of scanner data are processed in CPI, the contribution to sampling variance from food, beverages and the other stuff is estimated to be less than 35 per cent of total CPI variance, when the share of weights is 17 percent.

4.4. Weekly data

SCB has established a secured data transmission channel with retail chains through an FTP account. Input files are encrypted before transmission to ensure security. The routines give SCB enough time to reconnect in cases of a failed data transmission. Our scanner data system has six main stages of production:

1. Initiate a production month
2. Checking of the scanner data set
3. Selecting the data for the product-offer sample and reviewing it
4. Aggregating weekly average prices for three weeks to monthly average, per product and outlet
5. Sending data to the CPI production system
6. Analysing product life and replacing as many vanishing products in the sample as possible before next month

The technology involves applications based on the SAS System, a dot.Net solution interface and a robot-based file delivery system.

4.4. Elementary aggregate

Elementary aggregates constitute of product groups. Industry or region is not considered explicitly in the Swedish CPI, beside that sampling is supposed to deliver representative samples. Geometric means of price ratios, weighted by outlet weights and product weights are computed. For outlets the sample is self-weighted. Several products have individual weight due to the skewed distribution of turnover, see chapter 3.6.

5 References

Arvidson, J. (2004) Designutredning för KPI: Effektiv allokering av urvalet för prismätningarna i butiker och tjänsteställen, Background facts, SCB

Dalén, J. (2001) The Swedish Consumer Price Index, A handbook of methods, http://www.scb.se/statistik/PR/PR0101/handbok.pdf

Ohlsson (1990) "Sequential Poisson Sampling from a Business Register and its Application to the Swedish Consumer Price Index", R&D Report 1990:6.

Sammar, M., Norberg, A. and Tongur, C. (2012) Discussion on the Treatment of Discounts in the CPI and the Swedish experience on the use of Scanner Data. Paper presented at the Workshop on Scanner Data, Stockholm, June 7-8 2012 http://www.scb.se/Statistik/PR/PR0101/_dokument/Discussion-on-the-Treatment-of-Discounts-in-the-CPI-and-the-Swedish-Experience-on-the-use-of-Scanner-Data.pdf

SCB (2008) Urval – från teori till praktik (Sampling – from Theory to Practice), Handbok / Handbook 2008:1

APPENDIX 1 Product groups for which scanner data are collected in the Swedish CPI

COICOP	Product group	Weight (‰)	NamnSV	Sample size
01.1.1	1 113	1.08	WHEAT FLOUR	4
01.1.1	1 114	0.14	DANISH PASTRY	1
01.1.1	1 116	0.47	BISCUITS/COOKIES	2
01.1.1	1 122	1.05	CRISPBREAD	5
01.1.1	1 125	0.63	RICE	3
01.1.1	1 127	1.71	PASTA	11
01.1.1	1 130	1.66	CEREAL-BASED DISHES	8
01.1.1	1 131	1.47	CRACKERS AND RUSKS	5
01.1.1	1 132	0.11	GRUEL POWDER AND MIXER	2
01.1.1	1 133	2.28	BREAKFAST CEREALS AND SNACKS	10
01.1.1	1 136	0.33	FLOUR	2
01.1.1	1137	0.36	GRAIN	2
01.1.1	1 141	3.27	COARSE BREAD	20
01.1.1	1 142	4.17	FRENCH BREAD	24
01.1.1	1 143	1.81	PASTRIES	11
01.1.1	1 144	0.05	POTATO FLOUR	1
01.1.2	1 232	5.12	UNMIXED CURED MEATS	31
01.1.2	1233	5.38	MIXED CURED MEATS	31
01.1.2	1 235	1.31	TINNED MEATS	9
01.1.2	1 237	0.1	VEAL	1
01.1.2	1 238	0.92	OTHER MEAT	6
01.1.2	1 240	2.77	POULTRY	15
01.1.2	1245	1.81	FROZEN PROCESSED MEAT	6
01.1.3	1 315	1.41	FROZEN FISH	9
01.1.3	1 316	0.05	FROZEN SHELLFISH	1
01.1.3	1 318	0.66	TINNED HERRING	4
01.1.3	1 319	0.63	CAVIAR	3
01.1.3	1 323	0.77	MARINATED/SMOKED SALMON	3
01.1.3	1 324	2.8	FISH AND SHELLFISH PROD	18
01.1.4	1 410	1.9	EGGS	7
01.1.4	1 412	4.88	MILK	39
01.1.4	1 413	4.19	FERMENTED MILK AND YOGHURT	33
01.1.4	1 414	1.52	CREAM	15
01.1.4	1 417	1.02	SOUR CREAM; CRÈME FRAICHE	9
01.1.4	1 418	5.25	HARD CHEESES	28
01.1.4	1 419	3.32	SOFT CHEESES, ETC	20
01.1.5	1 504	0.97	BREGOTT	3
01.1.5	1 506	0.64	BUTTER	2
01.1.5	1 508	0.64	FOOD; FRYING AND BBQ OILS	4
01.1.5	1 509	0.62	COOKING MAGARINE	4
01.1.5	1 510	0.55	LOW-FAT MAGARINE	2
01.1.7	*1 611*	*0.2*	*POTATOES, PACKAGED*	*3*
01.1.7	*1 615*	*1,7*	*TOMATOES*	*19*
01.1.7	*1 671*	*0.2*	*FRESH HERBS AND SPICES*	*3*
01.1.7	*1 612*	*0.3*	*CARROTS, WASHED*	*4*
01.1.7	1 623	2.48	JUICES AND CONCENTRATES	14
01.1.7	1 625	1.68	TINNED VEGETABLES	9

01.1.7	1 626	2.21	POTATO PRODUCTS	11
01.1.7	1 635	0.89	FROZEN BERRIES AND FRUIT	7
01.1.7	1 636	1.44	FROZEN VEGETABLES	9
01.1.7	1 641	1.57	NUTS AND DRIED FRUIT	10
01.1.7	1 642	0.23	FRUIT AND BERRY PRODUCTS	2
01.1.7	1 643	0.16	PEAS AND BEANS	1
01.1.8	1 646	0.81	JAMS AND MARMELADES	7
01.1.8	1 650	0.92	FRUIT DRINKS, ETC	4
01.2.1	1 706	3.13	COFFEE	17
01.2.1	1 708	0.5	TEA	2
01.2.1	1 710	0.24	COCOA AND CHOCOLATE DRINKS	2
01.2.9	1 802	0.14	SALT	1
01.1.8	1 815	0.52	SUGAR	2
01.1.8	1 819	2.62	ICE-CREAM	11
01.1.8	1 823	3.88	CHOCOLATE	21
01.1.8	1 824	6.72	CONFECTIONERÝ	27
01.2.9	1 827	3.97	SPICES AND SAUCES	17
01.1.8	1 838	0.2	HONEY	2
01.1.8	1 839	0.05	SYRAP	1
01.2.9	1 841	2.33	YEAST, SOUPS, BABY FOOD, ETC.	11
02.1.3	1 905	0.12	LOW-ALCOHOL BEER	2
01.2.2	1 906	1.18	MINERAL WATER	5
01.2.2	1 907	5.25	SOFT DRINKS	20
01.2.2	1 908	0.34	CIDER	2
02.1.3	2 104	1.93	BEER, CLASS 2	6
02.2	2 310	9.46	CIGARETTES	20
02.2	2 311	5.81	OTHER TOBACCO PRODUCTS	18
05.5	5 414	1.6	LIGHT BULBS, BATTERIES, FUSES	7
05.6	5 513	2.47	KITCHEN PAPER	13
05.6	5 524	1.15	WASHING POWDER	8
05.6	5 525	0.61	WASHING-UP LIQUID	4
05.6	5 526	0.64	DETERGENTS, ETC	3
05.6	5 527	0.6	CLEANING EQUIPMENT	4
05.6	5 528	1.14	OTHER CONSUMABLES	7
09.3	7 711	4.54	PET FOOD	20
09.3	7 715	0.46	PLANTING SOIL	2
09.3	7 716	0.27	PLANT NUTRIENT	2
06.1	9 103	7.45	MEDICAL PRODUCTS	14
06.1	9 111	0.98	HEALTH CARE PRODUCTS	5
06.1	9 117	3.03	NATUROPATHIC MEDICINES AND VITAMINS	12
12.1	9 205	0.4	SOAP	2
12.1	9 208	2.51	SHAMPOO	12
12.1	9 219	0.62	DIAPERS	3
12.1	9 220	0.86	DURAB LE TOILETRIES	4
12.1	9 222	2.03	DEODORANT; SKIN CREAM, ETC	9
12.1	9 223	0.77	OTHER NON-DURABLE TOILETRIES	3
12.1	9 224	1.28	TOILET PAPER, ETC	7
Sum	Sum	173.74		891