Sampling Schemes for Price Index Construction:
A performance comparison across the COICOP Food Groups

by

Saeed Heravi and Peter Morgan, Cardiff Business School, Cardiff, Wales, UK

Abstract
Five sampling schemes for price index construction - one cut-off sampling technique and four probability proportional to size (*pps*) methods - are evaluated by comparing their performance on a homescan market research data set across 21 months for each of the 13 COICOP food groups. Classifications are derived for each of the food groups and the population index value is used as a reference to derive performance error measures such as RMSE, bias and standard deviation for each food type. Repeated samples are taken for each of the *pps* schemes and the resulting performance error measures analyzed using regression of three of the pps schemes to assess the overall effect of sampling scheme and COICOP group whilst controlling for sample size, month and population index value. Cut-off sampling appears to perform less well than *pps* methods and multistage *pps* seems to have no advantage over its single stage counterpart. The jackknife resampling technique is also explored as a means of estimating the standard error of the index and compared with the actual results from repeated sampling.

Acknowledgements
This paper arises out of work with the U.K. Office for National Statistics (ONS) and we are most grateful for their permission to publish this work and to thank Jo Woods, Mat Berger, Lewis Conn, Kat Pegler, Matthew Powell and Richard Campbell from ONS and John Doyle from Cardiff Business School for many helpful discussions. The opinions expressed here are ours and, of course, not necessarily those of the ONS. We also thank Mick Silver for his most helpful advice. Data supplied by TNS UK Limited. The use of TNS UK Ltd data in this work does not imply the endorsement of TNS UK Ltd. in relation to the interpretation or analysis of the data. All errors and omissions remain the responsibility of the authors.

1. INTRODUCTION

1.1 Sampling and the Consumer Prices Index

The Consumer Prices Index (CPI) is a macroeconomic indicator which attempts to summarize the change in price of a 'typical' basket of goods. It is widely used for formulating economic policy and indexing pensions and welfare benefits. Hence, its accurate measurement is critical and it is clearly of interest to know how various sampling schemes perform in the context of such a price index construction.

Sampling of representative items for a CPI is often judgmental (the US is an exception) with an implicit cut-off design (collecting prices from typically purchased items). There is general support in the international standards (ILO, 2004) and the literature (See, for example, De Haan *et al*. (1999) , Dorfman *et al*. (2006)) for cut off sampling. The accuracy of a CPI therefore depends on the selection of representative items. What is generally not recognised is that the price quotes which form the building blocks of the CPI cover a relatively small proportion of possible representative products in terms of expenditure (for example minced meat may cover all of beef). The traditional CPI methodology draws up a list of product types with product specifications (the classification scheme). These specifications, may be tight or loose. Tight specifications may adversely affect representativity since no products falling outside the specifications will enter the index. On the other hand, loose specifications give price collectors the freedom to adjust the sample which may or may not lead to greater representativity. However, combining this with the "most sold" criterion systematically under-represents the smaller brands and products that may be bought by important minorities (ILO, 2004). In the UK, the criterion used by the Office for National Statistics (ONS) is to choose a representative sample, or basket, of items that give a reliable measure of price movements for a wide range of goods. The sample chosen by the ONS (currently numbering over 680 items) is judgemental, stratified by region and shop for the CPI and RPI (Retail Prices Index) and stays in place for a whole year.

Since prices vary widely between and within different types of good, for practical sampling we need to adopt stratification and hence it is necessary to adopt some product classification as a framework within which to choose samples. The classification used in U.K. Consumer Price Index methodology (ONS, 2010) is the COICOP system standing for **C**lassification **O**f **I**ndividual **CO**nsumption by **P**urpose.[1] This system is one of a number which were approved by the United Nations at the 30th session of its Statistical Commission in 1999.

The analysis of sampling error and bias is problematic in practice since CPIs, by their nature, only have data from a single sample and not the population. The jackknife method (Efron and Tibshirani, 1994) for estimating the variance of a population from a representative sample is therefore a natural technique to examine in the context of estimating the dispersion of a (price) index. Another route to investigating how well a sampling scheme performs is to simulate index construction using repeated sampling from a model population. Hence, in this work we have carried out repeated sampling from a homescan data set as a model for the actual population of products and also examined the behaviour of the jackknife method.

Sampling schemes for representative items are multi-stage by nature and it is a central tenet of multi-stage sampling that more efficient designs sample higher number of units at the first stage. This presupposes that there is higher variation between these first level groups than

[1] The mandatory use of COICOP/HICP (Harmonized Index of Consumer Products) was established via an EU Commission regulation in 1999. (A. Zoppe, 2007).

within, Cochran (1977). This paper considers cut-off versus probabilistic sampling, and, given the somewhat forced nature of the multi-stage sampling within COICOP, evaluates multi-stage sampling strategies according to their accuracy and precision. Indices produced by a purposive (cut-off) scheme and three probabilistic schemes are compared with each other and the population index (which is the true value) from the whole data set. For other studies, comparing price indices constructed from scanner data with those based on official data, see, for example Fenwick et al. (2006).

The bases of the comparison are the month-by-month root-mean-square errors, biases and standard deviations for a 21 month series from January 2004 to September 2005[2] of indices derived from randomly drawn baskets matched across the months from January 2005 as the base month. The measures of performance used were the Root Mean Squared Error (RMSE), Bias and Standard Deviation. The three probabilistic schemes employed probability-proportional-to-size (*pps*) sampling, and the purposive (selective) scheme used cut-off sampling where only larger expenditure items were considered. In the former case, repeated sampling was used to estimate the variances of the sampling schemes and in, the latter case, only the bias could be used as a performance measure since, as will be seen below, it was not possible to carry out repeated sampling for the cut-off scheme.

Since the index has to reflect a vast range of different products within any COICOP group, a hierarchical classification is inherent to the sampling schemes chosen here. The remainder of the paper will cover the data used, the means of constructing such classification for dairy products, the sampling schemes used and regression analyses for the results obtained by repeated sampling together with discussion of the performance of the sampling schemes used as well as a comparison of the jackknife variance estimates with the actual variances from repeated sampling.

The main contributions of this paper are, thus, a) to provide a systematic comparison of the performance of different sampling schemes across a range of product categories backed up by a regression analysis whilst controlling for other variables such as sample size and index value and b) to evaluate the delete-d jackknife method for estimating the variance of an index across the same range of goods Though only one classification is shown, 13 classifications were devised for this paper – one for each of the COICOP food groups.

1.2 The data

Homescan data has become a valuable source of data for economic research (see Leicester and Oldfield, 2009, for a critical evaluation). Here we are using a homescan data set as a population from which to sample in the knowledge that the sampled indices so formed can be compared with a population value using all the products in the data set. Thus, to a limited extent, we are simulating the price collection process. The data arose from a market research data set supplied by Taylor Nelson Sofres (hereafter described as the TNS data set and now part of the Kantar World Panel).

The data set originally used here results from a survey of about 35,000 households between January 2004 and December 2005 and consists of barcode scanned records of all their food purchases.

2 The whole two years of data was originally used, but subsequently it was found that the last three months' data was not complete and so it was omitted from the analysis..

The base month for the indices was taken to be January 2005 so that the resulting price index went back 12 months to January 2004 and forward to December 2005. However, the last three months were subsequently found to be unreliable and were dropped from the study and the final results spanned the 21 months from January 2004 to September 2005 inclusive.

The annual expenditures for each COICOP group are presented in Table 1 and further divided by COICOP+ group which is the next level of classification by subtype of food.

Table 1 Expenditures in £millions for 2004 by COICOP and COICOP+ Group

BEER		**BREAD (& Cereals)**		**DAIRY**		**FISH**		**FRUIT**	
Stout	90	Flour	70	Eggs	559	Processed fish	530	Nuts	175
Cider	292	Rice	184	Milk products	1394	Fresh/Frozen fish	1574	Dried	297
Ale	370	Pasta	378	Cheese	2004			Fresh fruit	2560
Lager	1507	Chocolate biscuits	496	Milk	2471				
		Cereals	2607						
		Bread	4509						
TOTAL	2259	TOTAL	8244	TOTAL	6428	TOTAL	2104	TOTAL	3032
MEAT		**OIL (& Fat)**		**SOFT drinks**		**SPIRITS**		**SUGAR**	
Lamb	547	Butter	328	Fruit Juice	1447	Fabs	188	Sugar	538
Pork	679	Oils	678	Other soft drink	1538	Vodka	413	Ice cream	839
Beef	1470					Whisky	827	Confectionary	2130
Chicken	3461					Other drinks	941		
Other meat	5392								
TOTAL	11549	TOTAL	1006	TOTAL	2985	TOTAL	2369	TOTAL	3507
TEA		**VEGETABLES**		**WINE**					
Other hot drink	37	Processed Potato	472	Other	693				
Tea	453	Crisps	565	White	1287				
Coffee	658	Potato	843	Red	1469				
		Vegetable	4886						
TOTAL	1148	TOTAL	6766	TOTAL	3449				

Some interesting features emerge. For example, not only does Meat have the highest overall expenditure but it is notable that the miscellaneous Other Meat COICOP+ category has the highest expenditure within that group. This has importance for this study in that any sampling scheme which does not sample within a ‘dump’ category clearly ought not to perform well.

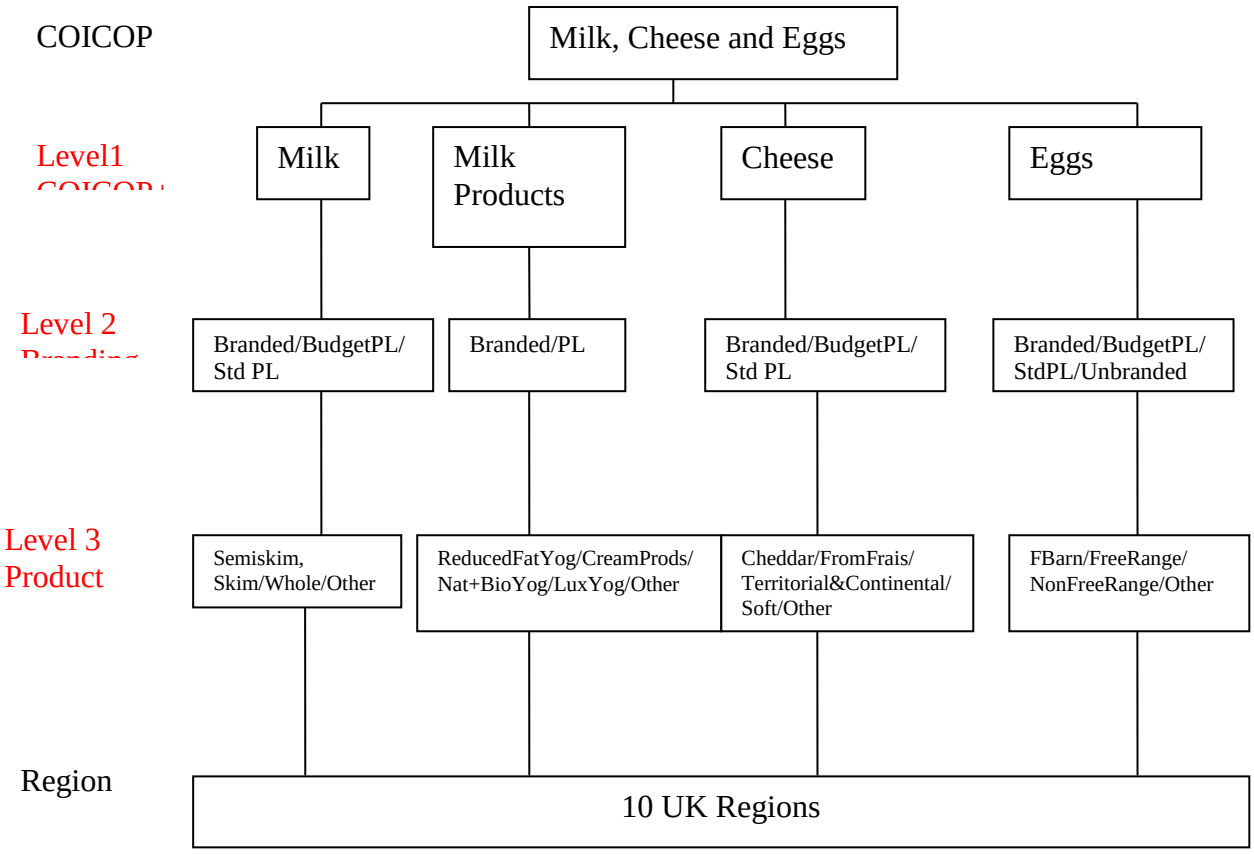
2 INDEX CONSTRUCTION

2.1 Classification

The TNS data set includes biographical information on the participating families (which was not used in this study) and product attribute information in addition to price and quantity. Product attributes, such as type of branding (Branded, Private Label, etc.), method of preservation (Fresh, Frozen, Chilled, etc.) etc., were used to construct classifications for each of the food groups. COICOP+ was used as the next classification level within each food group and one or two subsequent levels were typically constructed from these other attributes such as branding. At the bottom level we had the UK Government Office Regions.

Classification was fairly straightforward in most cases but, due to a few attributes having ambiguous or over-numerous categories or a substantial proportion of missing values, some recourse had occasionally to be made to cross-tabulation, aggregation and disaggregation to establish a set of unambiguous classifications with close to complete coverage of the data and a reasonable number of categories at each level. A typical classification is shown in Figure 1 which shows the products in the COICOP group containing Milk, Milk Products, Cheese and Eggs. In it we can see three levels before the regional level.

Figure 1 A classification for Dairy Products



Generally, most of the COICOP groups have a COICOP+ level which represents the subtypes of the particular food group. Throughout this paper, we define this COICOP+ level as being Level 1 with following levels being Level 2, Level 3, etc. For example in Figure 1, Branding is Level 2, Product Subtype is Level 3. In the Dairy example above we see four such items. In the case of wine, however, we have no such established COICOP+ level so subtype categories were made into pseudo-COICOP+ items (i.e. White Wine, Red Wine and Other (including Rose)).

2.2 Sampling Schemes

Table 2 A summary of the four sampling schemes

Sampling Scheme (SS)	Type	Description
CUT	**cut-**off	The two Level 1 items with the highest expenditure are selected. The two highest expenditure items at Level 2 within each of the selected Level 1 items are then selected, and so on with the two highest expenditure products within each region eventually being selected.
L1i	*pps*	Select all **L**evel **1** items **i**ncluding "All other" followed by a *pps* sample within each Level 1 item over all regions.
L1x	*pps*	Select all **L**evel **1** items e**x**cluding "All other" followed by a *pps* sample within each Level 1 item over all regions.
LLL	*pps*	Select all **L**evel 1 items, including "All other", then select all **L**evel 2 items, all **L**evel 3 items, and so on with *pps* sampling at the product level.

Table 2 gives a descriptive summary of the four sampling schemes used in this study. There were originally six schemes. A multistage sampling scheme was also deployed but, since the results were (very surprisingly) virtually identical to those of SS-L1i, results for this are not separately shown. This comprised selection of all Level 1 items including "all other" followed by *pps* selection of all Level 2 items and *pps* sampling of products within the lowest level items.) Another more complicated cut-off scheme based on an expenditure threshold was also tried but its performance was no better. For the sake of clarity and on the suggestion of an anonymous reviewer, the two schemes were excluded from the regression analyses and figures.

SS-L1i, -L1x, -L1x were used to draw repeated samples from the TNS population whereas SS-CUT is not susceptible to this being constrained by the requirement to choose the largest expenditure items. (There were insufficient quotes at the lowest level to carry out repeated sampling for this cut-off scheme.) The number of repeat samples is 500 in every case. Each sample of products in the base month is matched in other months and the Laspeyres expenditure weighted price index calculated across 21 months.

2.3 Index calculation

Price and quantity information was used to construct Laspeyres price indices - the Laspeyres monthly price index being defined as

$$Index_{Laspeyres} = \frac{\sum_{j=1}^{j=N} Q_{Base,j} P_{Current,j}}{\sum_{j=1}^{j=N} Q_{Base,j} P_{Base,j}} = \frac{\sum_{j=1}^{j=N} E_{Base,j} \left(\frac{P_{Current,j}}{P_{Base,j}} \right)}{\sum_{j=1}^{j=N} E_{Base,j}} = \sum_{j=1}^{j=N} w_j \left(\frac{P_{Current,j}}{P_{Base,j}} \right)$$

where w_j are the expenditure weights for the product basket

P, Q and E denote Price,Quantity and Expenditure

The base for the index was taken as January 2005 – over mid-way through the time period. The population index is over the whole data set matched from the products surveyed at January 2005 – the base month.

The data manipulation and calculation for the index construction were carried out using the SAS system (SAS, 2003).

3. EXPLORATORY DATA ANALYSIS: RMSE, Standard Deviation and Bias

3.1 Overall Error Measures

Table 3 gives the RMSE, Bias and Standard Deviation averaged over the 21 months (20 months in practice since the index base was set at 100 for January 2005). Only the Bias is appropriate for SS-CUT since repeated sampling was not possible in for cut-off sampling schemes. The Bias was calculated as an average bias from the population index over 500 repeated sample indices for the *pps* schemes. The average bias was also calculated across the months (excluding the base month) for the purposes of Table 3. The standard deviation (SD) and root mean squared error (RMSE) were likewise calculated from the deviations from the means of the 500 samples and the population index respectively. The overall measures of RMSE and SD in Table 3 were calculated as the square root of sum of squared RMSE or SD summed over the months.

$$Average\,RMSE = \sqrt{\frac{1}{m}\sum_{month}\left(\frac{\sum_{i=1}^{i=n}\left(I_i^{month} - I_{pop}^{month}\right)^2}{n}\right)^2} \quad Average\,SD = \sqrt{\frac{1}{m}\sum_{month}\left(\frac{\sum_{i=1}^{i=n}\left(I_i^{month} - \bar{I}^{month}\right)^2}{n}\right)^2} \quad Average\,Bias = \sqrt{\frac{1}{m}\sum_{month}\left(\frac{\sum_{i=1}^{i=n}\left(I_i^{month} - I_{pop}^{month}\right)}{n}\right)^2}$$

where m=number of months, n= number repeat samples, I_i^{month} = Index for any repeat sample in any month, I_{pop}^{month} = Population Index for any month

Table 3 Errors Averaged across the 20 Months

COICOP	Average RMSE			Average SD			Bias (SS-CUT) and Average Bias (SS-L1i, -L1x, -LLL)			
	SS-L1i	SS-L1x	SS-LLL	SS-L1i	SS-L1x	SS-LLL	SS-CUT	SS-L1i	SS-L1x	SS-LLL
beer	3.58	3.98	3.62	1.18	1.06	3.05	5.09	3.10	3.50	1.52
bread	2.17	2.12	2.46	1.99	1.97	2.24	0.94	-0.02	-0.13	0.40
dairy	1.10	1.29	1.25	0.71	0.64	0.72	0.39	-0.04	0.11	-0.53
fish	3.88	4.36	4.54	2.52	2.23	3.21	1.86	2.50	2.93	2.84
fruit	4.78	3.96	3.94	3.95	3.19	3.27	-5.25	-2.13	-1.79	-1.32
meat	2.98	3.01	4.42	2.66	2.58	4.17	-3.27	0.60	0.78	0.07
oil	0.59	2.28	0.70	0.48	0.48	0.64	-0.41	-0.17	-1.74	-0.12
soft	5.98	8.08	5.08	1.86	1.19	2.41	3.39	5.48	7.67	4.23
spirits	4.89	5.04	5.62	0.84	0.76	2.48	5.00	4.51	4.60	4.68
sugar	4.75	6.58	4.06	3.13	3.38	3.15	3.43	3.33	4.53	2.30
tea	2.27	2.05	2.76	1.54	1.15	2.52	2.2	1.21	1.25	0.65
veg	8.95	9.06	9.48	3.16	3.44	4.30	8.17	7.33	7.07	7.03
wine	6.20	5.18	5.76	2.36	2.39	3.47	5.13	5.38	4.28	4.36

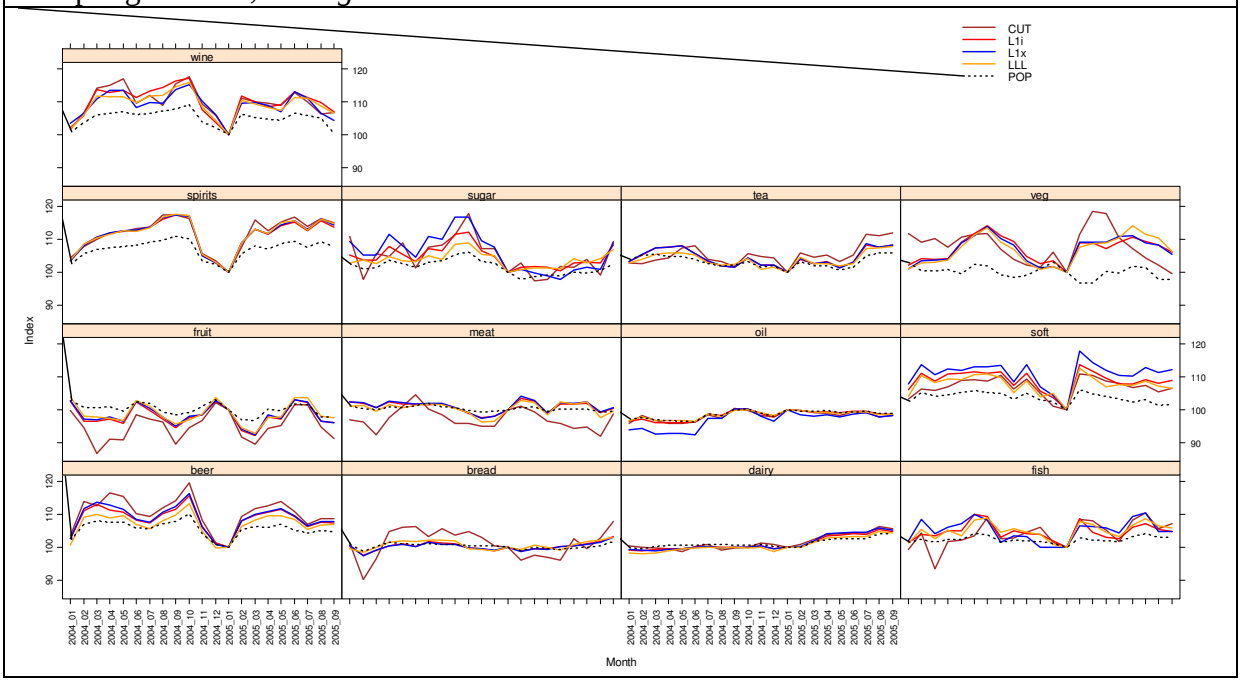
In terms of RMSE, the staple groups Bread, Dairy, Oil and Tea have relatively low errors whereas luxury items like Spirits and Soft drinks tend to be much higher. The exception here is the Vegetables group which has the poorest performance but does include 'luxury' items such as Potato Crisps. For standard deviation Oil and Dairy stand out as having extremely low error compared with the rest. In bias terms, Fruit has overall the most negative bias and Vegetables the most positive with Oil, Dairy and Bread having a low overall absolute bias.

3.2 Time series of indices

Figure 2 shows the time series (index base = January 2005) of indices together with the population index and SS-CUT (the cut-off scheme) is represented by single index series and SS-L1i, -L1x, -L1x (the *pps* schemes) are illustrated by the average of 500 indices.

It is clear from this plot that the cut-off scheme tends to be more volatile and far out of line with the other schemes for Vegetables, Fruit, Meat, Bread and Fish. (Another cut-off scheme based on an expenditure threshold also suffered from this problem.) It is also very noticeable that the sampled indices are far more volatile than the population index for a number of the staple goods (such as Tea, Meat, Bread and Dairy).

Figure 2 Conditioning plot of Index value vs. Month by COICOP group and grouped by Sampling Scheme, SS *Figure re-done*



3.3 Parallel Coordinates Plots

Altogether, there are 20 months × 3 *pps* schemes × 13 COICOP groups resulting in 780 error values for the RMSE, SD and Bias results. There is some sample size, index, percentage matching of the base sample (basket of products) and population index variation across the COICOP groups and months so to gain some initial insights into the performance of these schemes, parallel co-ordinates plots were made and the poorly performing (high RMSE or very positively or negatively biased) instances were highlighted. Parallel coordinates plots are a useful way of visualizing multivariate data (Inselberg and Dimsdale, 1990) and Figures 3 and 4 show two of these based on the RMSE and Bias results. In these plots, the data for each variable has first been scaled between zero and one. A number of equidistant vertical axes (not shown here for reasons of clarity), one for each variable, are erected side by side.

For each case (data table row), the scaled values for each variable are plotted and joined together to form a line and the diagram is a superposition of these lines so that clusters of cases, outliers, etc. become easily recognized. It is customary to join the points for each case by a polygonal line but in this case we have used spline interpolation to aid the viewer in following cases and groups of cases through the plot (Graham and Kennedy, 2003). The plots were produced in R and have been brushed in different colours to show the more extreme values with high RMSE, etc. A small amount of transparency was also introduced to alleviate visibility problems arising from overplotting.

It can easily be seen that the poorly performing instances are associated with a few COICOP groups and that they tend to be associated with lower values of the actual population value. The reason for this latter observation becomes quite clear when we look at the average bias of the sample indices across the months. The biases are overwhelmingly positive. Conversely, the poor RMSE instances are also associated with high values of the observed index. All three *pps* schemes were involved in these poorly performing instances and hence a regression analysis was used to quantify some of these effects and to examine the impact of using different sampling schemes whilst controlling for other factors such as sample size, month, population index, etc.

Figure 3 Parallel Coordinates Plots for RMSE Results (High RMSE in Red)

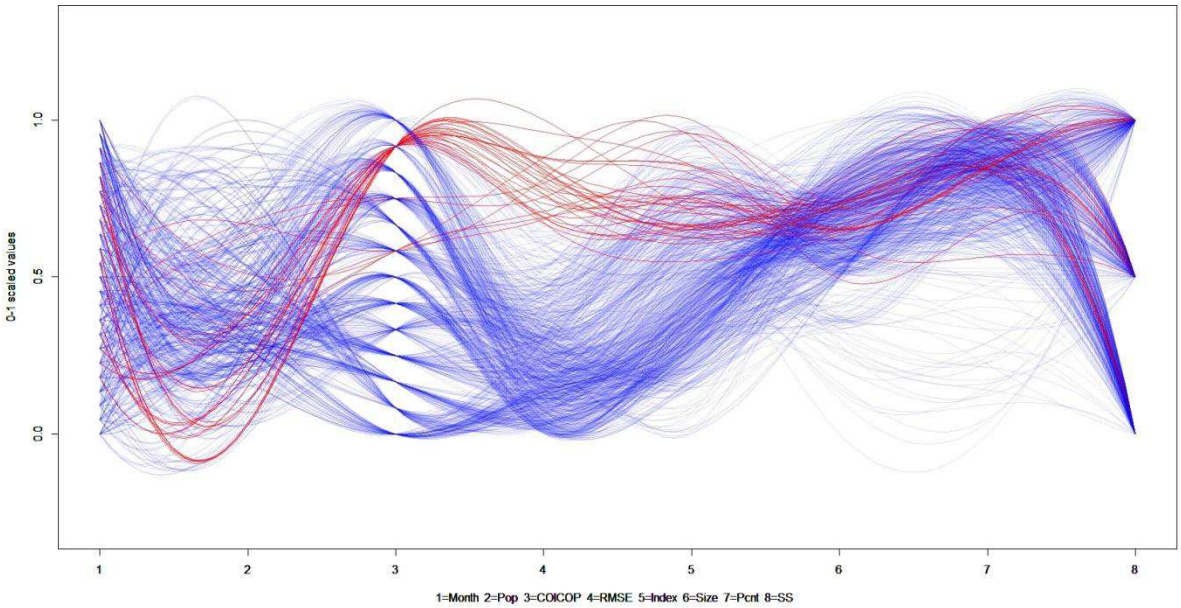
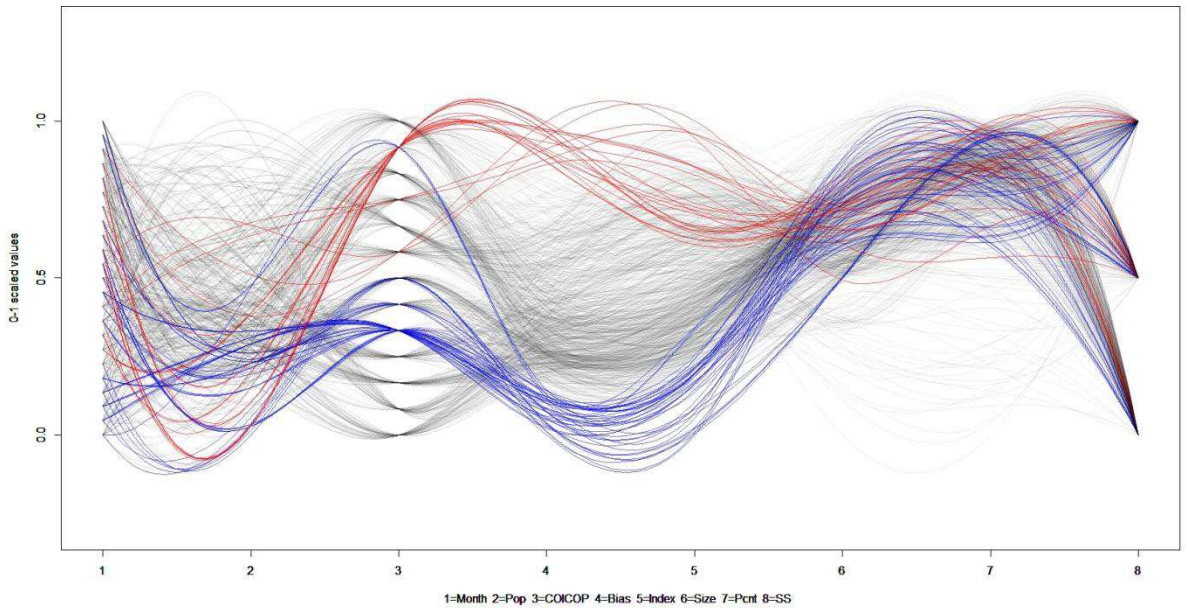


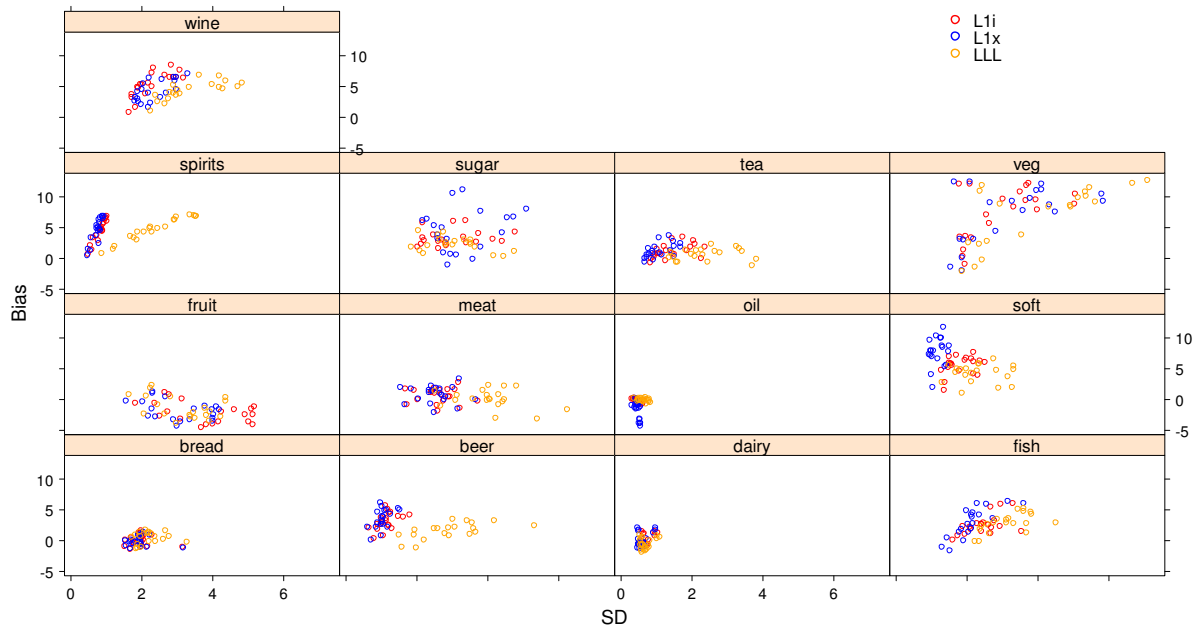
Figure 4 Parallel Coordinates Plots for Bias
Most Positive Bias in Red vs. Most Negative Bias in Blue



3.4 Plots of Bias vs. SD

Since the RMSE is decomposable into a bias and standard deviation component, it is clearly of interest to examine these aspects of the error of performance separately. The coplot in Figure 5 summarizes the behaviour of Bias and SD by *pps* sampling scheme and by COICOP food group.

Figure 5 Plots of Bias vs. SD by Sampling Scheme (colour-coded) for different COICOP groups
SS-L1i= Red , SS-L1x= Blue, SS-LLL= Orange

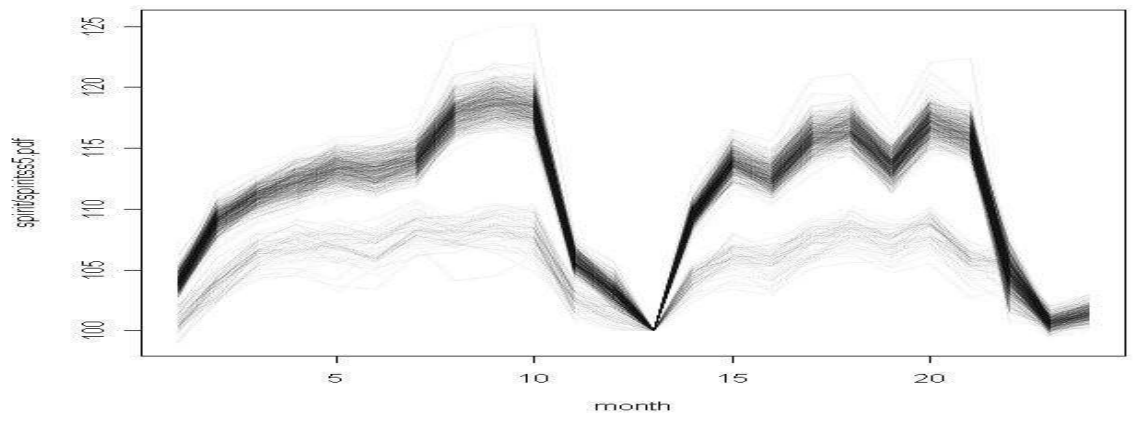


This conditioning plot has two notable features.

1. Method SS-LLL often stands out from all the rest and is generally less prone to bias than the others.
2. Some groups – notably spirits but also Wine, Fish and Beer show a degree of correlation in the relationship between Bias and SD.

Examination of the Spirits group results shows why this second feature comes about.

Figure 6 Superposition of 500 plots of the index for the Spirits group using SS-LLL



The superposed indices from the repeated sampling of the Spirits data using SS-LLL display a distinct cluster of indices with a much lower average value across the whole range of months. What is more, the baskets sampled for this lower-priced cluster have a similar, but proportionally lower, profile of price changes across the time period. Hence, the dispersion for the indices as a whole increases proportionally as the index value (and hence the bias) increases.

4. REGRESSION ANALYSES ON THE PERFORMANCE MEASURES

4.1 Variables and Transformations

The reference month was taken as January 2004 and the reference categories for COICOP group and Sampling Scheme were Bread and Sampling Scheme 3 (SS-L1i) respectively. Bread was chosen as the reference category since it is a staple food group with relatively little absolute bias, dispersion and volatility. All regressions have been carried out using mean centred variables and dummies. OLS and Robust Regressions were implemented in 'R' using the lm( ) and rlm( ) commands from the base and MASS packages respectively (http://cran.r-project.org/). The p-values for the robust regressions were obtained using built-in 'R' functions.

Initial regressions and the fact that it is lower bounded by zero suggested that SD as a dependent variable needs logarithmic transformation. Use of the bias component of the error as a dependent variable is somewhat less straightforward. Attempts to use (signed) Bias, absolute bias (|Bias|) or log|Bias| produced very non-normal residuals. However, referring to Figure 7 we can see that, since the standard deviation and the bias are plotted horizontally and vertically respectively and the MSE is equal to the sum of the variance and the squared bias, the RMSE is represented as the radial distance from the origin for each plotted point. We can thus perform a polar decomposition of the error such that the angle, arctan(Bias/SD), between the standard deviation axis and the radius tells us about how the error is distributed between the dispersion and the bias. As the angle approaches $\pm\pi/2$ the error becomes purely

due to positive or negative bias respectively and at zero angle we have purely dispersion error. Furthermore, this angle is independent of any transformations of RMSE since positive stretching of the radius will have no effect on the angle it makes with the standard deviation axis. Accordingly, we regressed this angle on the same variables to investigate the relative importance of the bias error. Since, the error in the angle for a given error at a point in the Bias-SD plane is proportional to 1/radius, we use a regression weighted by the RMSE to place less importance on the angle computed from small values of Bias and SD. Since we are also interested in the amount of bias irrespective of sign, we also use arctan(|Bias|/SD) as a variable which corresponds to reflecting the negatively biased data in the SD axis. These angles of bias and absolute bias, when used as dependent variables produced Q-Q plots which showed much greater normality than the attempts to use Bias, Absolute Bias or log(Absolute Bias). However there was still not complete normality of residuals so robust regression was used for these regressions and the coefficients and their significances were very similar if not identical to those from OLS. As a measure of the usefulness of this transformation we looked at the Median Absolute Percentage Difference (MAPD) between the OLS and Robust Regression coefficients and t-values for the angular vs. the raw bias regressions as shown in Table 4. It is clear that the log(SD) and absolute bias angle performed clearly better than the raw SD or Bias in terms of the regression since the use of robust regression made less difference than for their raw SD and Bias counterparts. Bias angle performed much better as a dependent variable than the raw bias but was still not terribly satisfactory. However, this signed measure is only being used to indicate the direction of the bias to compare with its unsigned counterpart so this was used as such.

Figure 7 Relationship between Bias Angle (θ), RMSE, SD and Bias in the SD-Bias Plane

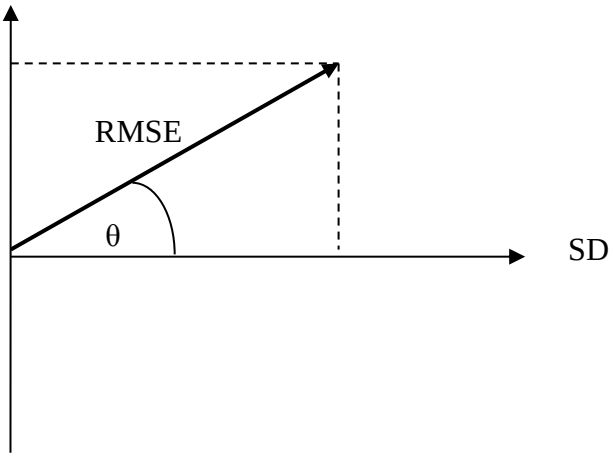


Table 4 Median Absolute Percentage Differences between the OLS and Robust Regressions

Dependent Variable	SD	Log(SD)	Absolute Bias Angle	Bias Angle	Absolute Bias	Bias
MAPD for coefficient estimates	8.2	8.6	4.8	20.3	28.2	48.3
MAPD for t-values	12.3	9.3	7.7	23.0	40.7	40.5

Figure 8 shows the results obtained for an OLS regressions of log(SD),

|θ| =arctan(abs(Bias)/SD) and θ = arctan(Bias/SD) on the variables shown - Population Index (PopC), Sample Size (SizeC), Sampling Scheme (SS), COICOP group (COICOP) and Month together with the interaction between SS and COICOP (and shaded by significance as shown). Including this interaction over the main variables improves the fit as measured by adjusted R squared as shown in Table 5.

Table 5 Variation of adjusted R squared for regression with and without an interaction term

Regression Equations with and without SS & COICOP interaction	OLS Adjusted R-sq Main variables only	OLS Adjusted R-sq and SS × COICOP interaction
log(SD) = PopC + SizeC + SS + COICOP + Month (+ SS×COICOP)	0.825	0.886
|Θ| = PopC + SizeC + SS + COICOP + Month (+ SS×COICOP)	0.682	0.749
Θ = PopC + SizeC + SS + COICOP + Month (+ SS×COICOP)	0.788	0.822

4.2 Standard Deviation as log(SD)

Cumulative normal probability plots show normality and near normality for log(SD) regressions with and without the SS×COICOP interaction. The values of the main variable coefficients vary little between the two regressions. SD tells us about the of error component due to the variability of the sampled indices around their mean irrespective of the bias from the population index value. The regression coefficients for SS-L1x and SS-LLL show how much worse or better these schemes are relative to SS-L1i controlling for all the other variables. We can see that SS-LLL is notably worse than the others and SS-L1x somewhat better. However the sampling scheme has an effect which is marginal by comparison with the COICOP group suggesting that issues related to COICOP group (such as classification) are of greater importance in determining this component of the error. The standardized coefficients for the COICOP groups range from 1 to 10 times larger in magnitude over those for the sampling schemes, SS, and are of the same order as those for the Month dummies. Sample size increase has a small but significantly beneficial effect on the error as expected and there is a significant but small increase in SD with the value of the population index, PopC.

When we look at the interaction coefficients, we see that SS-LLL worsens the SD in the majority of cases. Overall the interactions worsen the SD with the exception of Soft drinks and to a lesser extent for Fruit and Tea.

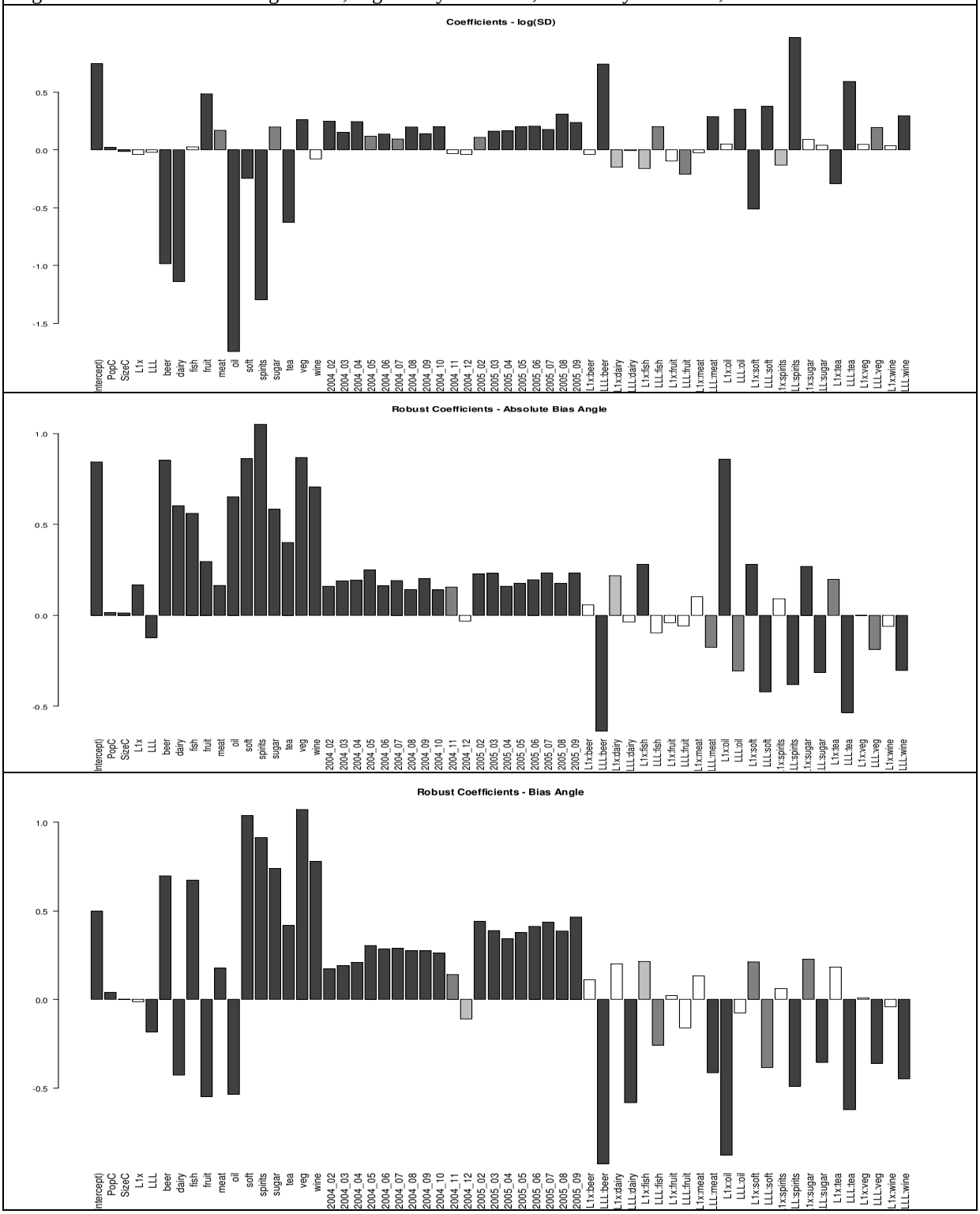
To check for autocorrelation, the autocorrelation function for the residuals from the regression of log(SD) were calculated for each combination of COICOP and Sampling Scheme (SS). In only two COICOP groups – Spirits and Vegetables - was an autocorrelation found to be significant at the 5% level.

4.3 Bias Angle (Θ) and Absolute Bias Angle (|Θ|)

The absolute bias angle tells us about the proportion of the error which is due to the bias as opposed to the standard deviation irrespective of whether that bias is positive or negative. Once we have established whether a worsening has taken place by finding a positive coefficient, we can use the corresponding bias angle coefficient to show whether this worsening has been brought about by making the bias more positive or negative. (Making the bias more positive results in a worsening if the average bias is positive or in a betterment if the overall bias is negative.) Hence we can compare the coefficients for the |Θ| and Θ

regressions to see the impact of individual variables whilst controlling for the others. Some points are immediately apparent.

Figure 8 Regression Coefficients for log(SD), Absolute Bias Angle and Bias Angle
Significances – White =Not significant, Light Grey=5% level, Dark Grey=1% level, Black=0.1% level



1. The impact of Month worsens the bias across the months except for December 2004 which makes it marginally better. Looking at the bias angle tells us that, save for December 2004, it does this by making the bias more positive.
2. Method SS-LLL improves the bias relative to SS-L1i by making it more negative. SS-L1x worsens the bias by making it slightly more negative.
3. As for the SD, the overall effect of the sampling scheme is insignificant compared with the impact of individual COICOP food groups.
4. The pattern of interaction coefficients for the signed and unsigned bias regressions are very similar except for the Oil and Dairy groups showing that, in the main, the interactions between SS and COICOP group either serve to increase the bias error by making it more positive or to decrease it by making the bias more negative which accords with the biases being predominately positive.

   The proportion of variation accounted for by the explanatory variables are shown in Table 6. We can see that the biggest impact comes from the **COICOP** variable whereas the sampling scheme very much takes second place – the sample size being really important only in the case of the dispersion error as would be expected.

Table 6 Analyses of Variance for log(SD) and Absolute Bias Angle - OLS Regressions

Analysis of Variance Table for log(SD)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
PopC	1	0.3	0.3	6.1	**0.01382**	*
SizeC	1	15.4	15.4	296.8	**<2e-16**	***
SS	2	30.7	15.4	295.2	**<2e-16**	***
COICOP	12	241.6	20.1	387.0	**<2e-16**	***
Month	19	7.5	0.4	7.5	**<2e-16**	***
SS:COICOP	24	21.8	0.9	17.5	**<2e-16**	***
Residuals	720	37.5	0.1			

Signif.codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1**

Analysis of Variance Table for Absolute Bias Angle

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
PopC	1	58.0	58.0	390.5	**<2.2e-16**	***
SizeC	1	4.7	4.7	31.4	**2.99e-08**	***
SS	2	42.6	21.3	143.4	**<2.2e-16**	***
COICOP	12	201.4	16.8	113.1	**<2.2e-16**	***
Month	19	13.6	0.7	4.8	**9.21e-11**	***
SS:COICOP	24	32.8	1.4	9.2	**<2.2e-16**	***
Residuals	720	106.9	0.1			

5. JACKKNIFE ESTIMATES OF STANDARD ERROR OF INDEX

Since, in practice, we have only one sample basket from which to construct a price index, one technique which has been suggested as a means of estimating the standard error of an index is to resample from that basket to reconstruct an estimate of the index sampling distribution. We have carried out a jackknife procedure on a series of sample baskets for all 13 COICOP Groups to assess its reliability.

We have used SS-L1i as the scheme to test the jackknife. Since all Level 1 items are chosen (samples of equal size taken from each to a total of as close to 100 items for most groups). The original leave-one-out jackknife method is adequate for linear statistics (Efron and

Tibshirani, 1994) but we have chosen to adopt the delete-d method where more than one datum is deleted from the sample – effectively sampling (*n-d*) data without replacement (where *n* is the whole sample size). To respect the sampling scheme, we delete the same number from each Level 1subsample. For the Meat group for example, there are 5 items at Level 1 and hence we take 20 product subsamples in each of those to construct the main sample and delete two from each to give the delete-10 jackknife resamples.

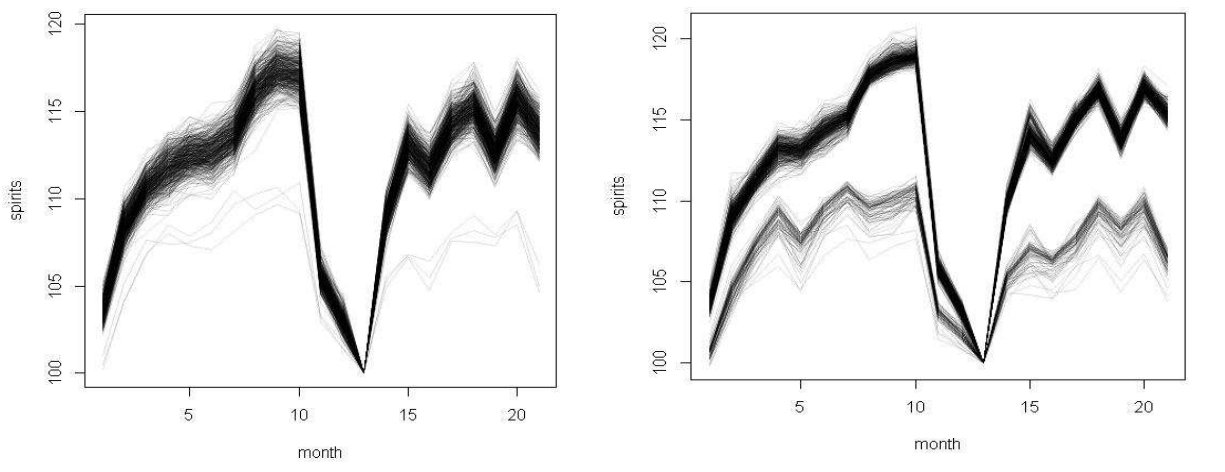
Following Wu (1990) and Efron and Tibshirani (1994) we resample a large number (500) of times (the number of possible data combinations being too great for complete enumeration unlike the leave-one-out simple jackknife method) and use the formula

$$I_{pseudovalue} = I_{full} + \sqrt{\frac{n-d}{d}}\left(I_{full} - I_{subset}\right)$$

I_{subset} and I_{full} are Index values calculated from the full and subsetted (partially deleted) sample and a large sample of $I_{pseudovalue}$ form an estimate for the empirical distribution function of *I*. The expression $\sqrt{\frac{n-d}{d}}$ can be thought of as an 'inflation factor' to account for the subsamples being very alike the parent sample. Then standard deviation of this estimated sampling distribution is thus an estimate for the standard error of the index.

A preparatory trial with a single sample from each COICOP Group showed that the jackknife is a conservative estimate for the standard error (since we know the actual standard error fairly closely from the repeated sampling experiments. In the main, the jackknife estimate was found to be between 1 and 2 times the actual value. The one exception being for the Spirits group where the jackknife result was 6 times its counterpart from actual repeated sampling. This was found to be due to the presence of very strong outlier within the Other Spirits Level 1 subgroup. Figure 9 shows the impact of outliers on the resampled indices. Here we can see a few outliers below the main cluster of index series. The 500 jackknifed index series all result from resamples of a single one of the 500 repeated samples. When this outlier was removed, the jackknife results fell into line. This posts a warning over the use of the jackknife method in this case since it appears to be very sensitive to the presence of outliers.

Figure 9 Comparison of 500 repeated sample indices (left) and 500 Jackknife resampled indices (right) showing the increase in outliers for the Spirits Group and SS-L1i

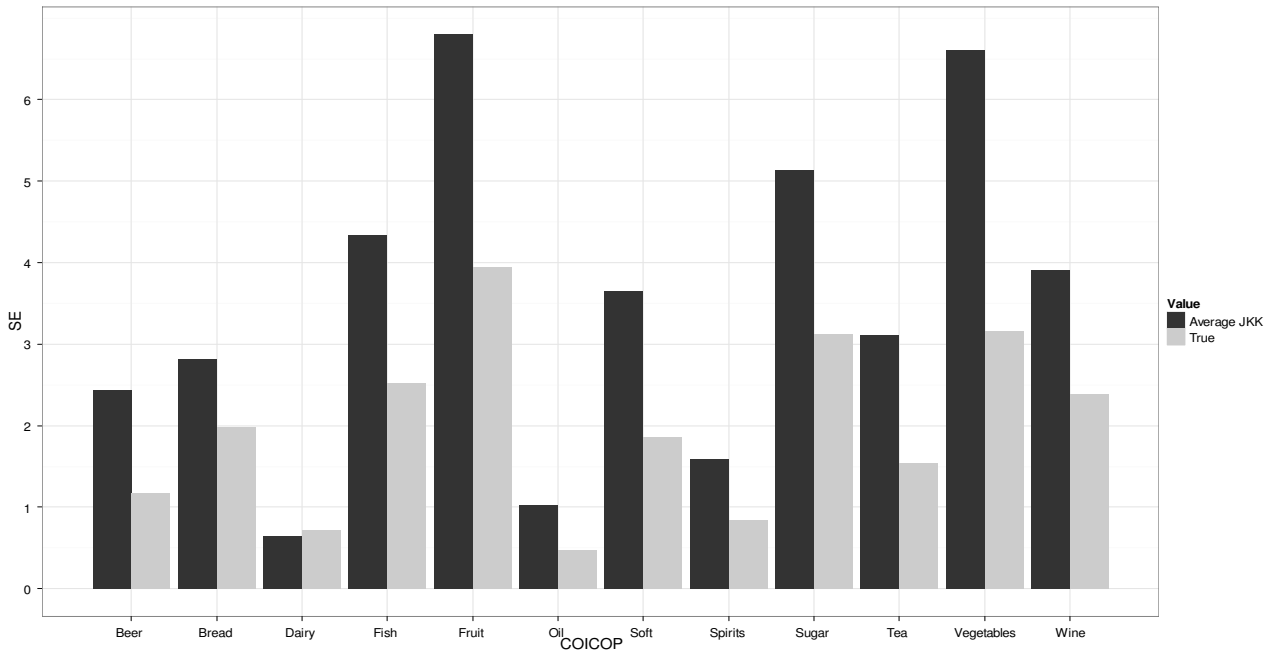


A series of 10 samples was taken from each of the 13 COICOP groups using SSL1i. The jackknife estimation was carried out on each of the sample sets and the results are shown in Table 7. The standard errors are taken over the 21 months of the data and the Jackknife results are averaged over 10 jackknife estimations each from a different sample. The final column presents the standard deviation of the 10 jackknife standard errors. These results are also depicted graphically in Figure 10.

Table 7 Comparison of the Actual and Jackknife Standard Errors for All COICOP Groups

COICOP Group	Actual Standard Error from repeated samples	Average Jackknife Standard Error of Index over 10 sets	Ratio of Average Jackknife SE to Actual SE	SD for 10 Jackknife estimates
Beer	1.18	2.44	2.07	0.11
Bread	1.99	2.82	1.42	0.84
Dairy	0.71	0.64	0.90	0.24
Fish	2.52	4.33	1.72	0.33
Fruit	3.95	6.80	1.72	0.77
Oil	0.48	1.03	2.15	0.04
Soft	1.86	3.66	1.96	0.50
Sugar	3.13	5.13	1.64	0.59
Spirits (without outlier)	0.84	1.59	1.88	0.07
Spirits (with outlier)	0.84	5.55	6.57	0.55
Tea	1.54	3.11	2.02	0.40
Vegetables	3.16	6.61	2.09	1.81
Wine	2.39	3.91	1.65	0.51

Figure 10 Actual and Jackknife Standard Errors for All COICOP Groups – averaged over 20 months and for 10 samples (Spirits group without outlier) *Plot re-done*



Notwithstanding the issue of outliers, we can see that the jackknife overestimates (with the one exception of Dairy) the actual standard error.

5. DISCUSSION

The main points that emerge from this study are as follows.

The choice of COICOP group has a great impact on the performance of the four schemes. As the ANOVA results for the *pps* schemes in Table 6 shows, there is variability of the schemes within COICOP groups as well as between them. With the notable exception of Fruit, the overall bias is positive and in many cases it is positive across all months and all schemes with respect to January 2004 and SS-LLL respectively.

The cut-off scheme is well out of line with the *pps* schemes except for Soft drinks, Spirits and Dairy. Generally, the cut-off scheme shows more volatility than the *pps* ones. Also, the population index is much less volatile than the others except for some staple items as mentioned above. These observations could be explained by a) the fact that the population index derives from all items rather than just the high expenditure lines and b) that the cut-off scheme targets the goods with highest expenditure and these are likely to be the ones with the shortest shelf life and hence more subject to frequent price changes. An alternative explanation is that the effect on the index of a subset of goods with high expenditures and very different prices is more evident in the smaller sample than in the population index. Eichenbaum et al. (2013) have used scanner data to examine price changes and their impact on the CPI.

The Jackknife technique shows some promise for estimating the standard error of the index, though it is clear that it is largely a conservative estimate and outliers can produce some undesirable effects. Hence, more work is needed to investigate this issue. The bootstrap might be an alternative means of assessing variability (see, for example, Patak and Beaumont,2009) and hence this will be considered in further work. There is, of course, a philosophical point to be raised in the context of bootstrapping for sampling price indices in that none of the bootstrap resamples could ever be a genuine basket of products since one would not select the same product more than once. However, from a purely pragmatic point of view, it will be important to compare the performance of the two methods of variance estimation.

For the *pps* schemes, SS-LLL performs significantly worse in terms of standard deviation and significantly better in terms of bias than SS-L1i. However, this effect is swamped by the effect of COICOP group. SS-L1x performs less well overall in terms of bias than the other schemes but slightly better in terms of standard deviation. SS-LLL forces selection of all Level 2 items as well as those in Level 1 and hence is the antithesis of a cut-off scheme at the upper levels. The interaction terms for the absolute bias angle show that in most cases using SS-LLL tends to reduce the bias. It therefore covers a much wider variety of products with the additional variability that implies. It may be this feature which makes this a bias reducing scheme. SS-L1x avoids selection of the 'All Other' categories and this may explain its better performance in terms of standard deviation and its worse performance in terms of bias.

One of the key features of the results is that schemes SS-L1i and the multistage scheme mentioned in Section 2.2 above, are virtually indistinguishable in terms of performance and it is very surprising that this multistage sampling seems to offer no advantage over its simpler

counterpart. However, it is more structured than SS-L1i and hence could offer some logistic advantages in terms of sampling cost.

The influence of classification hierarchy on the performance of the sampling schemes is not discussed here and is an issue we wish to explore in further work. The ordering of the levels in our classification trees was very much influenced by what might be the priority of customers' criteria for product choice and it is therefore an important question to ask what the effect might be of swapping classification levels, e.g. referring to Figure 1, if branding were less important to a customer than whether the eggs were free range or not, what impact would this have on the variability and level of the index for the different sampling schemes?

6. REFERENCES

1. Zoppe, Alice (2007): 'Use of COICOP in the European Union', Meeting of the Expert Group on International Economic and Social Classifications, New York.

2. Cochran, W. (1977), ' Sampling Techniques', Wiley

3. De Haan, J., Opperdoes, E. Schut, C.M. (1999) 'Item selection in the Consumer Price Index: Cut-off versus probability sampling' Survey Methodology, June, vol.25 no. 1, Product Classification

4. Dorfman, A.H., Lent, J., Leaver, S.G. and Wegman, E. (2006) 'On sample survey designs for Consumer Price Indexes', Survey Methodology, vol 32, no. 2 pp 197-216

5. Efron, B., and R.J. Tibshirani (1994) 'An Introduction to the Bootstrap' Chapman & Hall/CRC

6. Eichenbaum, M.S., Jaimovich, N., Rebelo, S. And Smith, J. (2013) , "How frequent are small price changes?", NBER Working Paper Series #17956, National Bureau of Economic Research, Cambridge MA

7. Fenwick, D. Melser, D. and Moran, P. (2006) 'Consumer Price Indices: real world quality measures', *9th Meeting International Working Group On Price Indices ,* The Ottawa Group

8. Graham, M. and Jessie Kennedy, J. (2003) 'Using Curves to Enhance Parallel Coordinate Visualisations' Seventh International Conference on Information Visualization, IV 2003, 16-18 July 2003, London, UK. IEEE Computer Society, ISBN 0-7695-1988-1

9. ILO(2004) 'Consumer Price Index Manual :Theory and Practice'

10. IMF(2004) 'Sampling Issues in Price Collection', *Producer Price Index Manual: Theory and Index* Chapter 5

11. Inselberg, A. and B. Dimsdale (1990) 'Parallel coordinates: A tool for visualizing multidimensional geometry.' In *Proc. Of IEEE Conference on vis'90*, pages 361–378.

12. Leicester, A. and Oldfield, Z. (2009) "An analysis of consumer panel data." (IFS Working Papers W09/09). Institute for Fiscal Studies: London, UK.

13. Office for National Statistics (2010) 'Consumer Price Indices Technical Manual, 2010 Edition', London: Office for National Statistics, London May 2010

14. Patak, Z., and Beaumont, J.-F. (2009), "Generalized Bootstrap for Prices Surveys," in Proceedings of the 57th Session of the International Statistical Institute , Durban, South Africa.

15. R Development Core Team (2010) 'R: A Language and Environment for Statistical Computing', R Foundation for Statistical Computing, Vienna, Austria

16. SAS (2003) 'Statistical Analysis System', SAS Institute Inc., Cary, North Carolina, USA

17. Wu, C. F. J. (1990) ‘On the Asymptotic Properties of the Jackknife Histogram’ Annals of Statistics, Vol. 18, No. 3, pp. 1438-1452