A Comparison of Six Sampling Schemes for Price Index Construction in a COICOP Food Group

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

ABSTRACT

This paper compares the behaviour of sampling techniques for price indices by using a scanner data set as a model population. Indices produced by two purposive deterministic cut-off designs and four probabilistic sampling schemes are compared with each other and with the 'true' population index from the whole data set. We found that the two deterministic cut-off sampling schemes show much different behaviour from the probabilistic sampling schemes. This is not unexpected as the former schemes have a very restricted focus with respect to the variety of products. We also found that the probabilistic schemes are generally closer to each other and the 'true' value than the deterministic cut-off designs. 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.

Key Words: price indices, product classification, sampling, jackknife

1. Introduction

Sampling of representative items for a CPI is often judgmental (the US is an exception) with an implicit cut-off design (typically purchased). The accuracy of a CPI 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 and with tight specifications, representativity will suffer 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 - leading 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 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/RPI and stays in place for a whole year.

The analysis of sampling error and bias is problematic since CPIs, by their nature, only have data from a single sample and not the population. However, scanner data sets offer a model data set of prices which can be repeatedly sampled to produce empirical distributions of price indices for different sample schemes enabling these schemes to be compared. There is general support in the international standards (ILO Manual) and the literature (See, for example, De Haan et al. (1999), Dorfman et

al. (2006)) for cut off sampling. Scanner data has been shown to be a very useful ancillary source of data for the measurement of inflation (see, for example, Fenwick et al., 2003; Silver & Heravi, 2001). This paper uses an extensive homescan data set to explore the behaviour and suitability of alternative sampling schemes.

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 (see 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.

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 within, Cochran (1977). In this work we are fortunate to be able to calculate a 'gold standard' index by considering every product in the data set. Hence we can evaluate the performance of any probabilistic sampling scheme by using repeated samples and comparing them with this true population index to measure bias and standard deviation (SD) as indicators of accuracy and precision together with root mean squared error (RMSE) as a gross measure of error.
This in line with other approaches in the literatures to test if commonly used empirical methods give results similar to the data generating process, Schaefer *et al.(2008).*
This paper considers two cut-off and four probabilistic sampling schemes. In the case of the cut-off schemes, the only error measure we can employ is the bias since, up to this last level, only the largest two expenditure items are considered and the schemes are essentially deterministic. For other studies comparing price indices constructed from scanner data with those based on official data see, for example, Fenwick et al. (2006).

Traditionally, lack of bias has been emphasised in the performance of sampling schemes for consumer prices indices. At the same time, there has been an ongoing debate about the relative merits of judgemental and probabilistic sampling. Indeed there is considerable variation in sampling techniques worldwide. More recently, there has been more interest in the variance of indices as well as their bias. The construction of prices indices involves choosing a (necessarily) limited but representative basket of goods and hence the choice of sampling scheme is critical. Cost also precludes repeated sampling from the population of good in practice in order to assess variability of indices, but the jackknife method of variance estimation could provide a way to get index variability information by resampling from a single sample. Hence, the key motivation for this study is to learn more about the performance of sampling schemes, both judgemental and probabilistic, across a variety of goods with a view to seeing which give good performance from both a bias and variance point of view and also to experiment with the jackknife method as above. To this end, we use a large homescan data set as a testbed population from which to sample.

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

It is hoped that such an approach can enable an understanding of which sampling schemes perform best and hence provide some indications for improving the performance of index construction. Though this paper is restricted to one staple COICOP food group (the Meat group) this is being done in the context of all 13 COICOP food groups; comparisons across these groups will form the basis of future publications.

We calculate the error measures month-by-month 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 four probabilistic schemes employed probability-proportional-to-size sampling (*pps*) and the two purposive (selective) schemes used cut-off sampling where only larger expenditure items were considered.

In practice, it is not possible to do repeated sampling so it pays therefore to make the best use of the information obtained through sampling. One set of tools which has been achieving more prominence in recent years uses resampling (see, for example, Efron and Tibshirani, 1994) in the form of the bootstrap and jackknife procedures. In the case of the bootstrap, we attempt to examine the sampling distribution for a statistic (such as an index value) by resampling with replacement from a single whole sample taken from our population (of goods and services in the case of the CPI) to create a host of resamples each leading to a value of the statistic under examination. The jackknife method takes subsamples of the whole sample without replacement to calculate alternative values (known as pseudovalues) of the statistic being studied. The original jackknife procedure was also known as the 'leave one out' method as it used the n subsamples arising from leaving each datum out in turn to create the necessary subsamples and hence the values of the statistic. This turns out to have disadvantages with non-linear statistics and those (such as the median) with an inherent discontinuity. More recently (see, for example,Wu, 1990), the delete-d jackknife has overcome these problems by taking subsets of size (n-d) rather than (n-1) of the original whole sample. In this paper, therefore, we also examine the viability of the jackknife technique for estimating the dispersion of the index estimate produced by one of the sampling methods by comparing it with the true estimate of the standard error from the repeated sampling results.

Since the index has to reflect a vast range of different products within any COICOP group, a hierarchical classification is inherent in the sampling schemes chosen here. The paper will cover the data used, the means of constructing such classification for meat, the sampling schemes used and regression analysis for the results obtained by repeated sampling together with discussion of the performance of the sampling schemes used.

2. Data

The data set employed was supplied by Taylor Nelson Softres (TNS) and contains sixty million transactions from a sample panel of 35,000 households for about 400,000 products. The households are chosen that they cover all ages, gender and

[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..

social class and in every region of the UK, thus thoroughly representing the whole population. Households are required to scan their shopping purchases within their own home. The main data set contains the details of the transactions and includes the bar codes, household number, product codes, shop code, product descriptions, market categories, year/month/week/day of transaction and the price of the goods and the number of packs bought. The second file contains the product attributes. For example for the meat data we have around eighteen product attributes.

The first step in this work is to merge the two files, containing the price and product attributes respectively, by the product code. This is a challenge in itself due to the massive size of the data bases and the number of variables concerned. However, the attribute variables have considerable redundancy and we have chosen only few important attributes to include in the analysis and sampling procedure.

The coverage of the data is impressive, in terms of number of quotes and details of transactions and attributes. Table 1 shows the number of quotes and expenditure for each of the COICOP+ level for meat for the year 2004. It shows that Lamb has the lowest number of quotes and expenditure followed by pork, beef, chicken and other meat. 'Other Meat' consists of a huge variety of products ranging from sausages to liver pâté and thus has the highest expenditure.

Table 1: Expenditure for 2004 for COICOP+

COICOP+	Quotes	Expenditure (£ Millions)
Lamb	49108	547
Pork	96715	679
Beef	194905	1470
Chicken	547684	3460
Other meat	1370256	5392

3. Approach to classification

As mentioned above, the basis for the classification used here is COICOP which is a demand based and functional classification of consumption.

3.1 Selecting Attributes

Since, the TNS data set contains a large number of product attributes – the first step was to select a useful subset of these as potential classifier variables. The COICOP system has 13 food and drink groups each of which is subdivided at a further level (known colloquially as the COICOP+ level). In the cases of the Meat group this comprises Beef, Lamb, Pork, Poultry and Other Meat.

The attributes for Meat contained a considerable number of anomalous categories which had presumably become 'fossilized' into the attribute list by accretion of custom and practice. For example, there were a number of portmanteau categories such "prepackaged". Also, some inconsistency persisted where, for example, the COICOP+ Lamb category contained an attribute referring to Chicken.

Our main criteria for including an attribute variable were to

1. Provide a clean (mutually exclusive) separation of the products across a natural dimension of the product description, e.g. Type of Preservation - "cooked", "frozen", "fresh", etc.
2. Have a relatively small proportion of missing values
3. Have, or be capable of having, relatively few values, i.e. does not split the data up into so many categories that their individual frequencies would be too low or a large 'dump' or Other category would need to be created

The strategy in constructing the classification tree for Meat was to look for common attributes across and below the COICOP+ level where possible, but to tailor the classification to reflect the expected impact of attributes on price. The number of levels needs to be workable in that too many will result in an unacceptably low number of quotes for categories at the lowest level and too few will not adequately reflect the structure of the market for that product group.

Where an attribute has categories which were ambiguous or seemingly out of line with the other categories, use was made of crosstabulation with other attributes in an attempt to better identify these 'rogue' categories.

Some attributes contained information on more than one dimension of the product. For example, there might be values such as 'frozen chicken', 'frozen turkey', 'fresh duck', etc. so that information on both the type of meat and the method of storage/preservation would be carried by the same attribute variable. It was occasionally possible to split such variables into two. Likewise, some attributes with excessive numbers of categories could be indentified and coalesced using keywords to provide a more workable set. For example, an attribute containing a number of chicken, turkey and duck products can be transformed into a variable identifying the type of poultry concerned.

3.2 Subclassification below COICOP+ level

In all COICOP+ categories for Meat, it was possible to use an attribute to identify 'brandedness' (i.e. 'No Brand Name', 'Private Label' and 'Unbranded' - the last category including loose, prepacked, etc. goods) on the basis that a brand name ought to imply a premium price for a product as opposed to an 'own label' or an unbranded product. However, over the COICOP+ level the usefulness of these categories varied greatly. For example, more than 99% of the beef products are either 'No Brand Name' or 'Private Label'. Thus, in the final classification, the use of Brandedness as a classification variable was restricted to the Other Meat COICOP+ group.

The classification by region provided the last level above the quote level for each of the COICOP+ groups.

Beef, Pork and Lamb
Apart from Brandedness, three attributes could be developed – Prepreparation/Preservation method ('fresh', 'frozen', 'cooked', etc.), Cut ('stewing steak', 'mince', etc.) and Origin (British vs. Non-British).

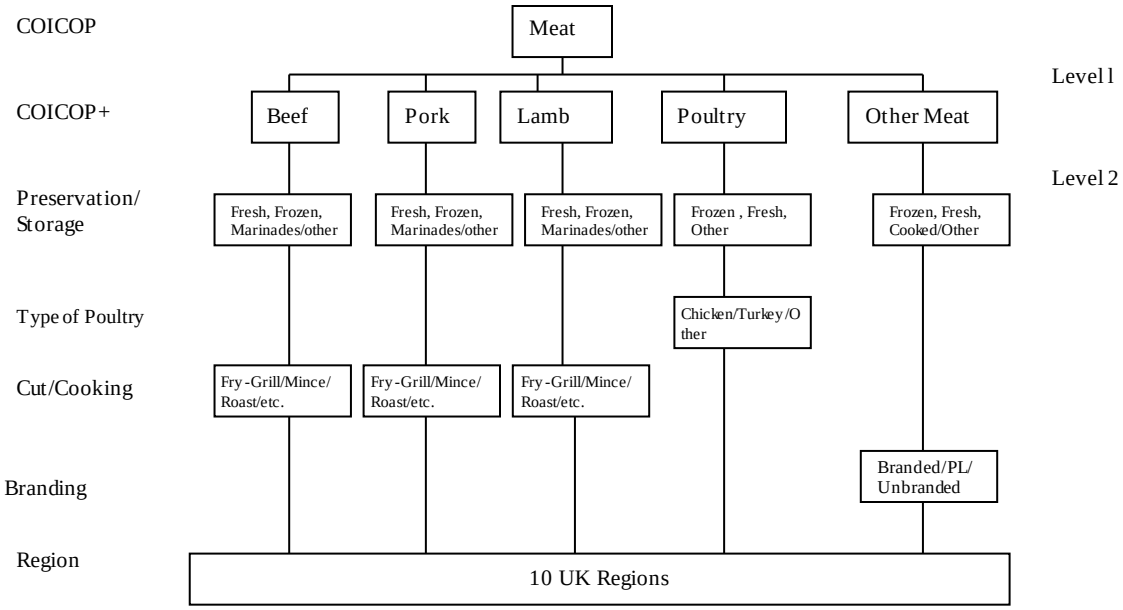
Poultry
Again, aside from Brandedness and Region, it was only possible to develop Pre-preparation, Type ('chicken', 'turkey', etc.). Origin was not identifiable for a large proportion of the products even if we had wanted to use it. Ambiguity in Pre-preparation was resolved, for the major part, by cross-tabulation with other attributes.

Other Meat
Apart from Brandedness, the only usable attribute for Other Meat was Pre-preparation. It is very noteworthy that Other Meat carries, by far, the largest expenditure.

Classification for Meat
The use of the Origin variable resulted in an underpopulation of the lowest levels and so was omitted from final classification. The Brandedness variable was only thought to be a useful differentiator on price for the Other Meat group which contains such items as sausages. The final classification for meat is shown in Figure 1.

Figure 1 Classification Hierarchy for Meat

`



4. Sampling Schemes and the Sampling Process

To reflect the practical variation in current sampling techniques we include both cut-off and probabilistic methods. Given the plethora of available goods, sampling schemes need to operate on multi-level classifications such as described in the previous section. Hence, six basic sampling schemes were carried out on the Meat data. These were

SS1 – cut-off sampling using two largest items at all levels
Select the two Level 1 (COICOP+ level) items with the highest expenditure. Select two highest expenditure items at the next level (Level 2) within each of the selected Level 1 items, etc. down to selection of the products with the highest expenditure within each region.[3]

SS2 – cut-off sampling using only the first level
Select all COICOP+ items with an estimated annual National Expenditure of more than £400 million and the largest item for any COICOP groups left without such an item. The most popular product within each region is then chosen.

SS1 and SS2 are examples of 'cut-off sampling' as defined by Dorfman et al. (2006) whereby we sample the larger expenditure items and ignore the remainder. By way of illustrating the way that the sampling scheme interacts with the classification, Figure 2 shows the path taken through the classification tree by Sampling Scheme 1 (SS1).

SS3 – equal sampling at the first level and *pps* at product level
Select all COICOP+ (Level 1) items including "all other" followed by a *pps* sample within each category over all regions.

SS4 – equal sampling at first level excluding 'Other' category and pps below
As SS3 but excluding the "all other" category at COICOP+ level.

This exclusion of the miscellaneous category was done to examine the effect of omitting products which could be considered to be 'outliers' in the sense that they fail to fit into one of the main categories. In the case of Meat, the expenditure on products in this category is the largest of all the COICOP+ items and this shows, perhaps, that the specification of items at this level has not kept up with consumer choice.

SS5 – equal sampling at all levels with *pps* at product level
Select all COICOP+ (Level 1) items including "all other" followed by selection of all level 2 items and so on with *pps* sampling of products down to the lowest level items.

SS6 – equal sampling at first level with *pps* at next level and at product level
Select 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.

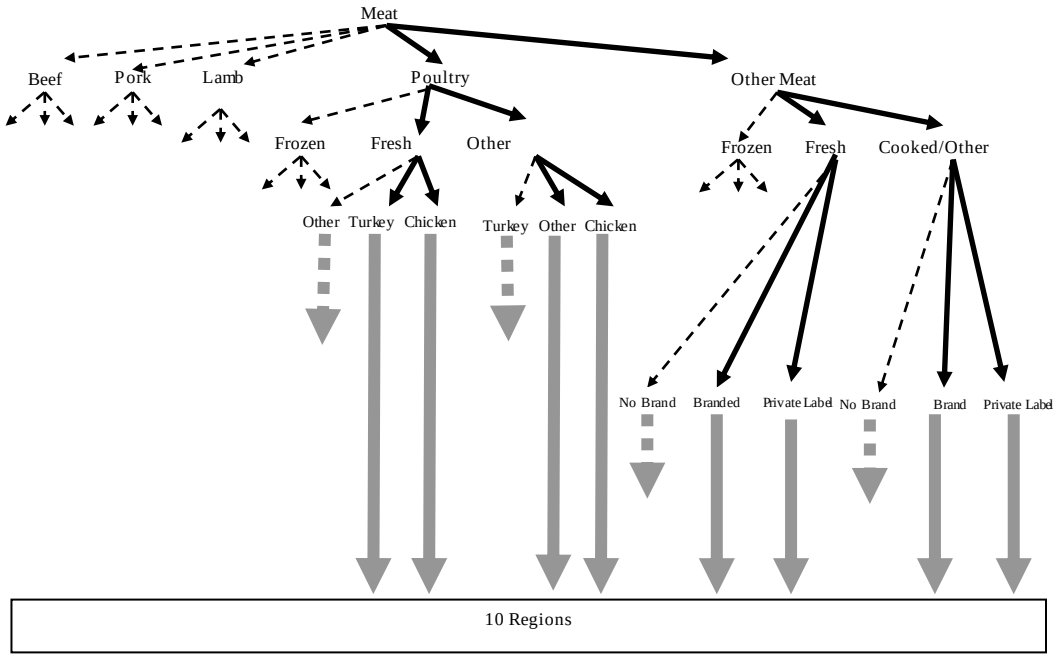
SS1 and SS5 force us to choose products at levels below Level 1 (the COICOP+ level) while SS2, SS3 and SS4 forces choice at Level 1 (SS2 is an expenditure-based

[3] In the general case of course we may wish to choose more than 2 items at any level. Techniques exist to combine cut-off sampling with *pps* whereby the number, k, of items to be sampled is pre-determined. The mean cumulative expenditure per sample item is calculated as E_{total}/k and the j items with expenditure greater than this are chosen with certainty. After removal of these chosen items, the remaining mean cumulative expenditure is recalculated as $E_{total\ remain}/(k-j)$ and the process repeated. If it cannot be repeated and we have not yet chosen k items, the remaining items are sampled by *pps*, IMF (2004)

cutoff sampling schemes which, in the case of meat, effectively forces choice at Level 1 because all COICOP+ groups have national expenditure in excess of £400M). SS6 is a multistage sampling scheme which also forces choice at Level 1. It is likely that the more restrictions are applied in a sampling scheme by way of forced choice ensure that a greater weight is given to outlier products.

SS3 – SS6 were used to draw repeated samples from the TNS population whereas SS1 and SS2 are not susceptible to this since they are constrained by the requirement to choose the largest items. (There were insufficient quotes at the lowest level to carry out repeated sampling for these first two schemes.) 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.

Figure 2 Path through the classification tree taken by SS1



The Laspeyres monthly price index is 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 denotes price and Q denotes quantity

This was calculated for each of these sampling schemes using the SAS package (SAS, 2003) through a SAS program linking the classification scheme above to the TNS database. Originally, the index was calculated from January 2004 to December 2005 using January 2005 as the base month. However, the data from October to December 2005 was found to be incomplete and so these last three months' results were excluded from the final index series which were then input to the 'R' Package (R

Development Core Team (2010)) to derive summary statistics and carry out regression analyses.

In addition to sampling from the TNS population, a population index was constructed from every item in the database which served as a 'true' value from which to calculate the Root Mean Squared Error (RMSE) of each index series.

For SS3-6, 500 samples of product baskets were drawn in the base month using the classification above. The index series was calculated through matching from the database back to January 2004 and forwards to September 2005. No imputation was carried out and the % matching is displayed for each sampling scheme in Figure 3. RMSE, Standard Deviation and Bias values were then calculated as follows.

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

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

For SS1 & SS2, only the Bias could be calculated as there was no repeat sampling in these cases.

5. Results

5.1 Summary measures and plots

Table 2 Summary measures across the price index time series

Measure	Sampling Scheme					
	SS1	SS2	SS3	SS4	SS5	SS6
RMSE			2.98	3.01	4.42	3.03
Bias	-3.27	-0.66	0.6	0.78	0.06	0.65
SD			2.66	2.58	4.16	2.7
RMS Bias	4.19	2.78	1.35	1.54	1.49	1.38
Mean Sample Size	61	50	80	78	66.6	80.1

Mean Sample Size includes the base month whereas error measures exclude it. RMS Bias is the Root Mean Square of the biases across the 20 independent months and is introduced to provide a comparison across both the repeatedly sampled and 'deterministic' schemes.

Figure 3 shows the results of overlaying repeated time series based on January 2005 which is Month 13 (where the time series plots all pass through the same point at Index=100. The overlay incorporated transparency in order to reduce overplotting problems and give a better idea of the density of observations at any one time and index value. The other plots for SS4-6 show substantially the same behaviour.

Several features are shown by these plots.

1. Clear bimodality - in Months 6, 11 and 20 for example and confirmed in density plots for each month.
2. Coexisting inflation and deflation for subclusters of index series from Month 6 to 7 where two subclusters of plots undergo a 'see-saw' behaviour.
3. A great deal of correlation between successive months is evident between March, April and May 2005 (Months 15 to 17).
4. Outliers are prominent at various times.

Below this we can see the value of the average index by month on the same horizontal scale. SS1 and SS2 are essentially deterministic with insufficient quotes at the lowest level to allow repeated samples. These are thus included together with SS3-6 and with the population value but are absent in the uppermost plot. Key features of these indices are as follows.

1. The two 'deterministic sampling schemes SS1 & 2 show much different behaviour from the repeatedly sampled schemes SS3-6. This is not unexpected as the former schemes have a very restricted focus with respect to the variety of products. If we compare the % of products matched across time for each of the sampled indices we see that SS1 and SS2 behave very differently as well.
2. The volatility in the population index is far smaller than that in any of the sampled indices.
3. The repeatedly sampled indices are very close to each other – especially SS3, SS4 and SS6

In the bottom plot we see the % of products matched across from month to month which are fairly high. These values are even higher for the percentage of the expenditure covered by the matched samples since we select the high expenditure items.

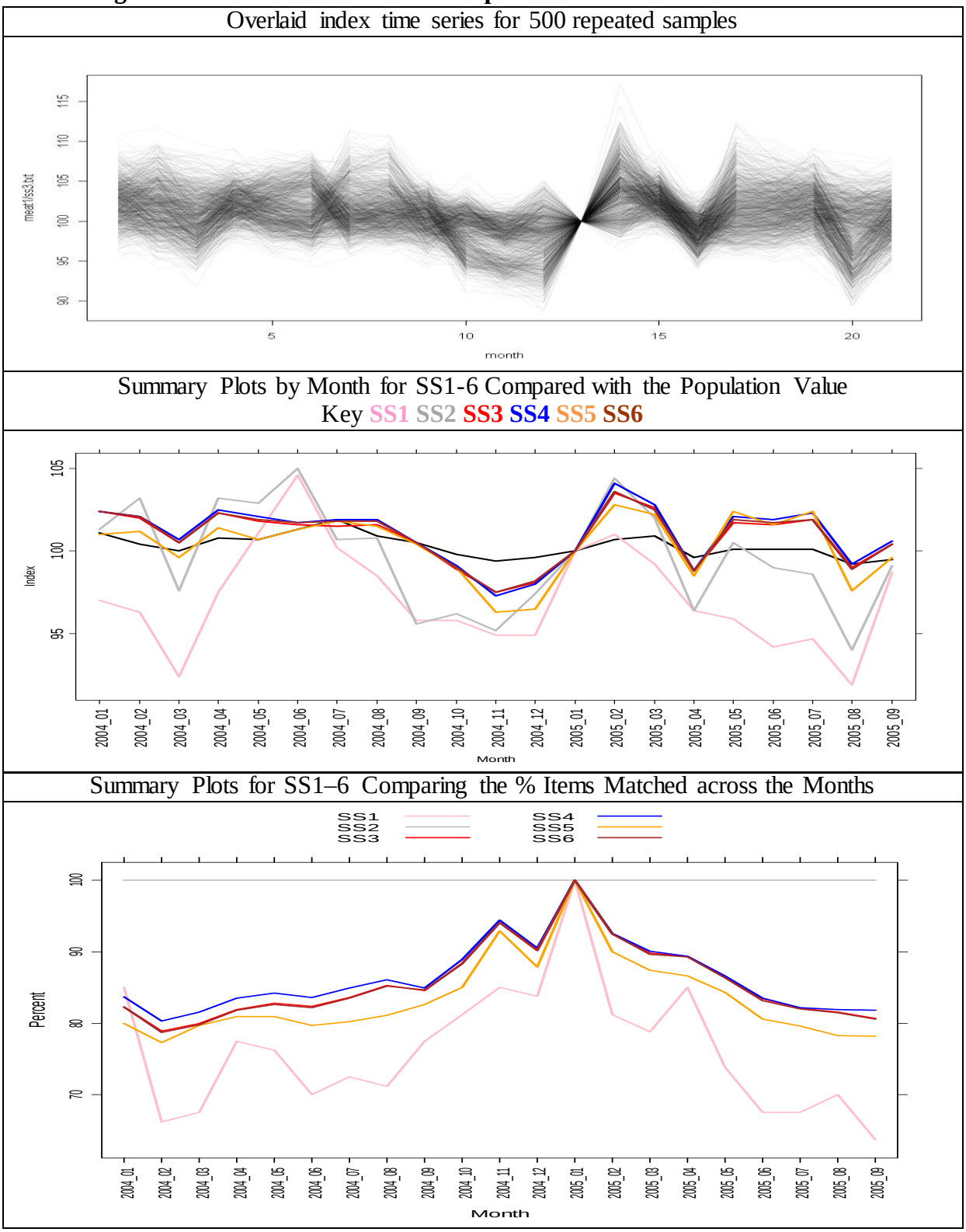
In Table 3 we see indices for SS1- 2 (as values from a single purposive sample) and SS3-6 (as averages for 500 repeat samples)

Table 3 Values of the Sampled and Population Price Indices for Meat

Month	SS1	SS2	SS3	SS4	SS5	SS6	Population
2004/01	97.0	101.3	102.4	102.4	101.0	102.4	101.1
2004/02	96.3	103.2	102.0	102.1	101.2	102.1	100.4
2004/03	92.4	97.6	100.5	100.7	99.6	100.5	100.0
2004/04	97.5	103.2	102.3	102.5	101.4	102.3	100.8
2004/05	101.1	102.9	101.8	102.1	100.7	101.9	100.7
2004/06	104.6	105.0	101.6	101.7	101.3	101.7	101.3
2004/07	100.2	100.7	101.5	101.9	101.8	101.8	101.9
2004/08	98.5	100.8	101.6	101.9	101.5	101.8	100.9
2004/09	95.8	95.6	100.5	100.5	100.4	100.5	100.5
2004/10	95.8	96.2	98.9	99.1	99.0	99.0	99.8
2004/11	94.9	95.2	97.5	97.3	96.3	97.5	99.4
2004/12	94.9	97.4	98.1	98.0	96.5	98.2	99.6
2005/01	100.0	100.0	100.0	100.0	100.0	100.0	100.0
2005/02	101.0	104.4	103.5	104.1	102.8	103.6	100.7
2005/03	99.2	102.0	102.6	102.8	102.2	102.5	100.9
2005/04	96.4	96.4	98.8	98.8	98.5	98.8	99.6
2005/05	95.9	100.5	101.7	102.1	102.4	101.9	100.1
2005/06	94.2	99.0	101.6	101.9	101.6	101.7	100.1

2005/07	94.7	98.6	101.9	102.3	102.4	101.9	100.1
2005/08	91.9	94.0	99.0	99.2	97.6	98.9	99.2
2005/09	98.7	99.1	100.4	100.6	99.6	100.4	99.5

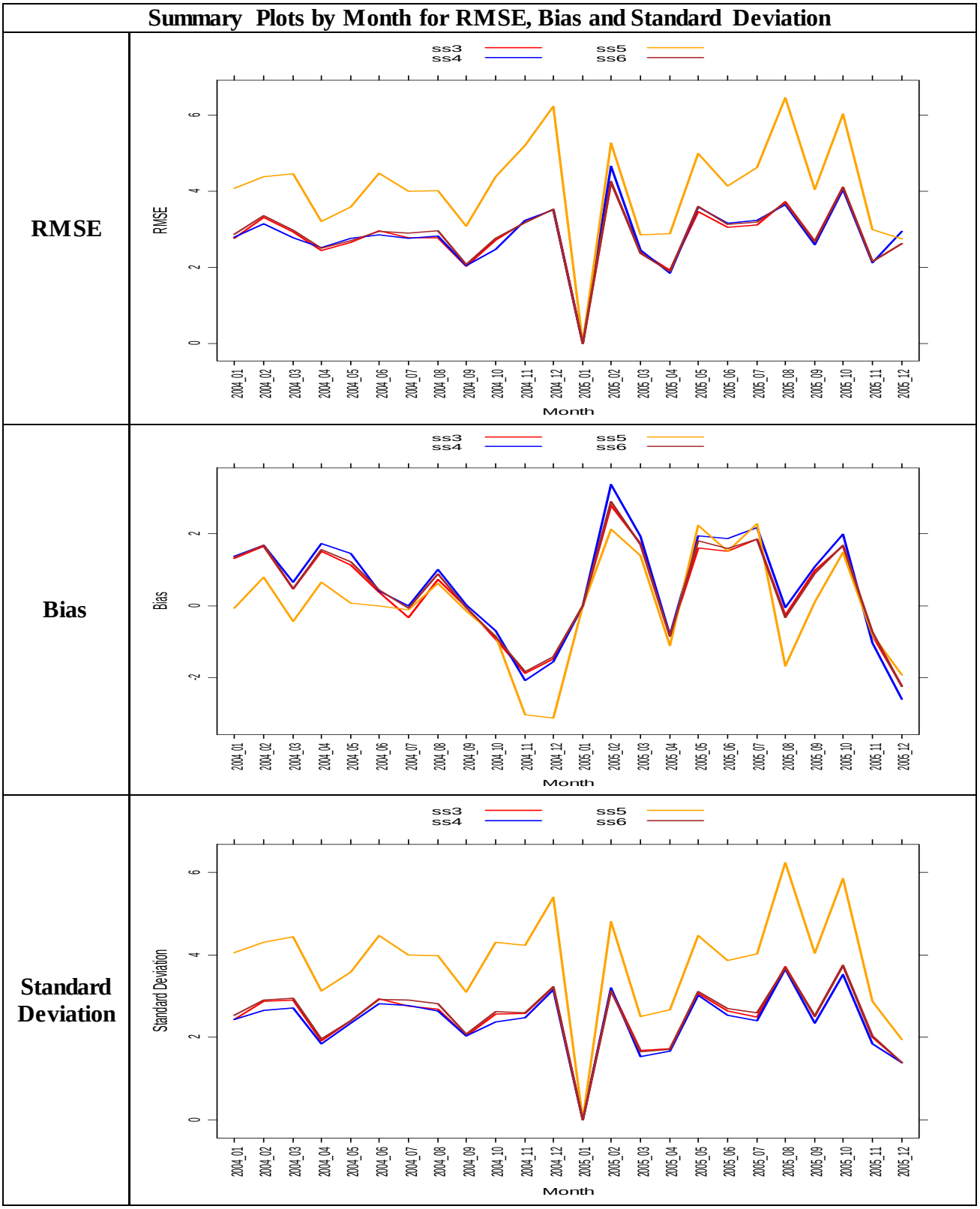
Figure 3 Overlaid price indices from repeated samples using SS3 together with the average values for SS1 – 6 and their sample sizes



In Figure 4, we quantify the variability we see in the repeatedly sampled indices (SS3-6) and further break this down into its bias and standard deviation (SD) components. The bias is calculable for all the six schemes whereas the RMSE and SD are only obtainable for the four repeatedly sample ones. Concentrating on the latter, we see

that RMSE and SD plots are very similar whereas the pattern of variation in the bias looks very different – though SS3 and SS6 are almost indistinguishable at times.

Figure 4 RMSE, Bias and Standard Deviation for repeatedly sampled indices SS3-6



5.2 Regression Analyses

To examine the dependence of the RMSE (as a measure of gross error) on sample size (rather than % coverage), month and sampling scheme, log(RMSE), log(Absolute Bias) and log(SD) were regressed on Month, Sampling Scheme and Sample Size with results as follows in Table 4. Centred variables were used and hence the main variable coefficients for SS3–6 (SS3 as reference category) have effectively been corrected for Sample Size (actual as opposed to the % matching shown in Figure 3) and Month effects.

This table shows the OLS results for the logarithm of the Root Mean Square Error, Bias and Standard Deviation in the meat index for the four repeatedly sampled schemes. Adjusted R-squared values for the three regressions were 0.94, 0.62 and 0.99 for log(RMSE), log|Bias| and log(SD) respectively.[4] Month was used as an independent variable to control for temporal variation but, for clarity, the coefficients have not been reported in Tables 4, 5 and 6.

Table 4 Regression of log(RMSE), log|Bias| and log(SD) on Month, Sample Size and Sampling Scheme

Dependent	log(RMSE)			log|Bias|		log(SD)		
Coefficient	Estimate	Standard Error		Estimate	Standard Error	Estimate	Standard Error	
Const	1.155	0.014	***	0.007	0.184	1.055	0.008	***
SS4	-0.030	0.038		-0.087	0.458	-0.029	0.020	
SS5	0.172	0.260		0.545	3.136	0.527	0.139	***
SS6	0.022	0.026		-0.027	0.309	0.019	0.014	
SizeC	-0.015	0.018		0.021	0.221	0.005	0.010	
SS4:SizeC	0.002	0.006		0.025	0.066	0.002	0.003	
SS5:SizeC	-0.001	0.006		0.107	0.071	0.003	0.003	
SS6:SizeC	-0.001	0.005		0.002	0.057	-0.002	0.003	
Adjusted R-squared	0.945			0.621		0.987		

For the Bias results, two zero values were excluded prior to the logarithmic transformation.

The pattern of coefficients for the Sampling Scheme show a consistency across the three measures of error. (We also have to remember that RMSE is an aggregate measure of SD and Bias taken together). There is very little to choose between SS3, 4 and 6 but scheme SS5 is always worse and very significantly so in terms of SD. The mean log(SD) is 1.05 and going from the reference scheme SS3 to SS5 results in an increase of 0.53 in log(SD). In unlogged terms this means a change from to 2.86 to 4.85 or an increase in SD of 69%.

[4] All models are significant with very low p-values for the F Test. The histogram of residuals shows symmetry and approximate normality. Since some outliers we evident in the Q-Q plot, robust regression was also carried out (see main text). First order serial correlation was insignificant at the 5% level for the residuals for each of the four sampling schemes.

There are a number of outlying values in these results and so robust regression was carried out using the R language 'rlm' (Robust Linear Model) command (R Development Core Team (2010)).

Table 5 Robust Regression of log(RMSE), log|Bias| and log(SD) on Month, Sample Size and Sampling Scheme

Dependent	log(RMSE)		log|Bias|		log(SD)	
Coefficient	Estimate	Standard Error	Estimate	Standard Error	Estimate	Standard Error
Const	1.181 ***	0.009	0.052	0.048	1.056 ***	0.006
SS4	- 0.016	0.023	0.167	0.120	- 0.026	0.018
SS5	0.380 *	0.160	1.104	0.822	0.562 ***	0.128
SS6	0.021	0.016	0.051	0.081	0.020	0.013
SizeC	-0.004	0.011	0.050	0.058	0.008	0.009
SS4:SizeC	0.002	0.003	0.000	0.017	0.002	0.002
SS5:SizeC	0.009	0.003	0.084 ***	0.018	0.004	0.003
SS6:SizeC	-0.001	0.002	-0.004	0.014	-0.002	0.002

Examination of these robust regression coefficients shows that the absolute values of the coefficients differ from the OLS model but the overall pattern is the same. These two tables taken together carry the relatively simple message which is that SS5 is by far the worst method overall of SS3-6 for the meat group and that the other three are relatively similar. This backs up the qualitative observations drawn from the time series plots already discussed.

The signed bias regressions shows very similar behaviour to the other error measures as Table 6 shows. The adjusted R-squared measure is 0.945 for this regression and we can see that the bias is positive on average which again accords with the appearance of Figure 3 above. Again, scheme SS3, 4 and 6 differ little from each other but method SS5 makes the bias more negative on average and compensates somewhat for this tendency to overestimate the index. Increasing sample size tends to make the bias more negative in contrast with the unsigned error measures where it has no effect.

Table 6 Regression of Signed Bias on Month, Sample Size and Sampling Scheme

Coefficient	OLS Coefficient		Robust Coefficient	
	Estimate	Standard Error	Estimate	Standard Error
Const	0.451 ***	0.073	0.430 ***	0.038
SS4	- 0.096	0.196	0.195	0.102
SS5	-3.098 *	1.355	-3.917 ***	0.709
SS6	0.058	0.140	0.058	0.073
SizeC	-0.181	0.096	-0.241 ***	0.050
SS4:SizeC	-0.025	0.029	-0.041 *	0.015
SS5:SizeC	-0.033	0.031	-0.044 *	0.016
SS6:SizeC	0.004	0.025	0.006	0.013

6. Jackknife Resampling

The jackknife technique is a powerful general method for bias correction and variance estimation (Tibshirani and Efron, 1994, Wu, 1994, Leaver and Larson, 2001). Here we restrict our investigations to the latter and explore its use with sampling scheme SS3.

Following, Wu (1994), given a statistic, θ, (in our case a price index), we can define pseudovalues

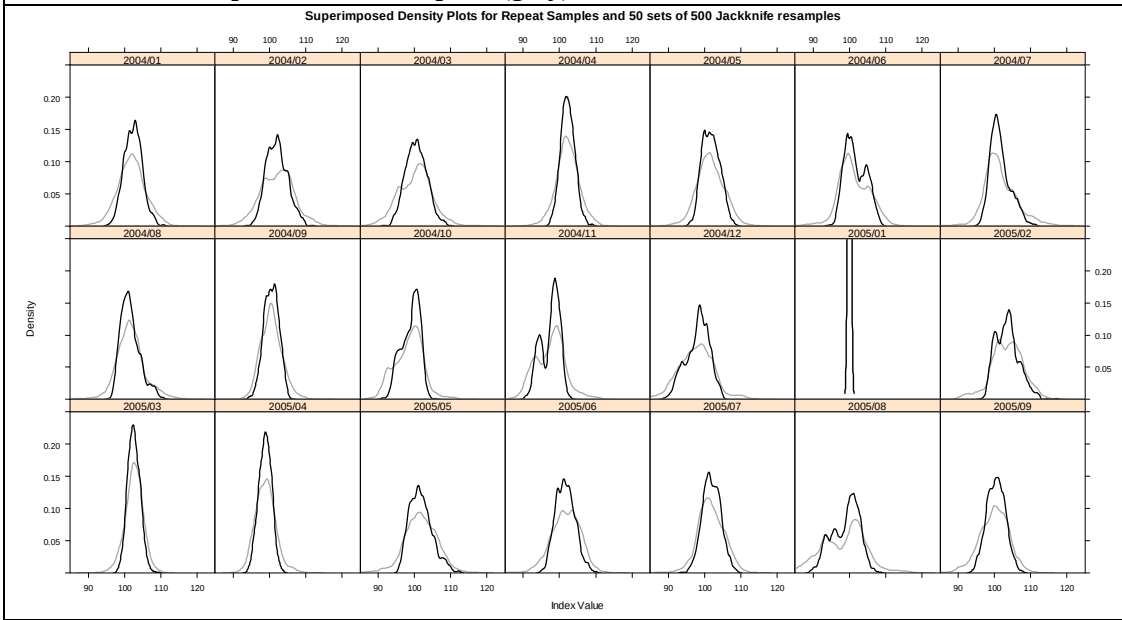
$\tilde{\theta}_{sub} = \hat{\theta}_{whole} + \sqrt{\frac{n-d}{d}}\left(\hat{\theta}_{sub} - \hat{\theta}_{whole}\right)$ which are weighted sums of the estimates of θ from the whole sample and subsamples. The assumption is that, for a representative whole sample, the distribution of these values will mirror the actual sampling distribution of the statistic, θ. The weighting factor $\sqrt{\frac{n-d}{d}}$ is needed to reflect the fact that, unlike bootstrap resamples and for small values of d, the jackknife resamples are very like the whole sample and, without it, the histogram of resamples would be artificially narrower than that of the actual sampling distribution.

In sampling scheme SS3, all the COICOP+ levels are selected and the *pps* sampling takes over below that with 20 products being selected from each of the 5 COICOP+ categories (Beef, Lamb, Pork, Poultry and Other Meat). We thus take a sample of 5 × 20 products (the whole sample) and calculate the index across the months using again January 2005 as the base month. Wolter (2003) suggests applying the jackknife technique to stratified samples by simultaneous deletion of one item from each stratum. To compute the jackknifed resamples, we need to honour the structure of the sampling scheme and, extending Wolter's scheme, two of the products are deleted at random from the whole sample in each COICOP+ category and the index again computed across the months. In this work we produce 500 such resamples. In a small minority of cases we may fail to match the product from the base month and, in the absence of imputation, a reduced whole sample is used but this effect is relatively minor.

Given that we have access to the actual (empirical) distribution of index values from the repeated sampling experiments described above, we are in the fortunate position of being able to use the standard errors of these empirical index sampling distributions as a check on the validity of the jackknifing procedure. Also, we can use each or any of the repeat samples as a whole sample upon which to base the jackknifing procedure which enables us to look at the standard error of the jackknife standard errors and, in the results that follow, we repeat the jackknife procedure for 50 such repeat whole samples.

Figure 5 shows the repeat sampling distribution for 500 repeat samples using SS3 compared with the jackknife resamples pooled from 50 different original samples. Though resamples from a single whole sample may reflect the idiosyncrasies of that sample, this confirms that indeed the jackknife resampling procedure, is indeed reproducing the empirical sampling distribution for SS3 in a general sense.

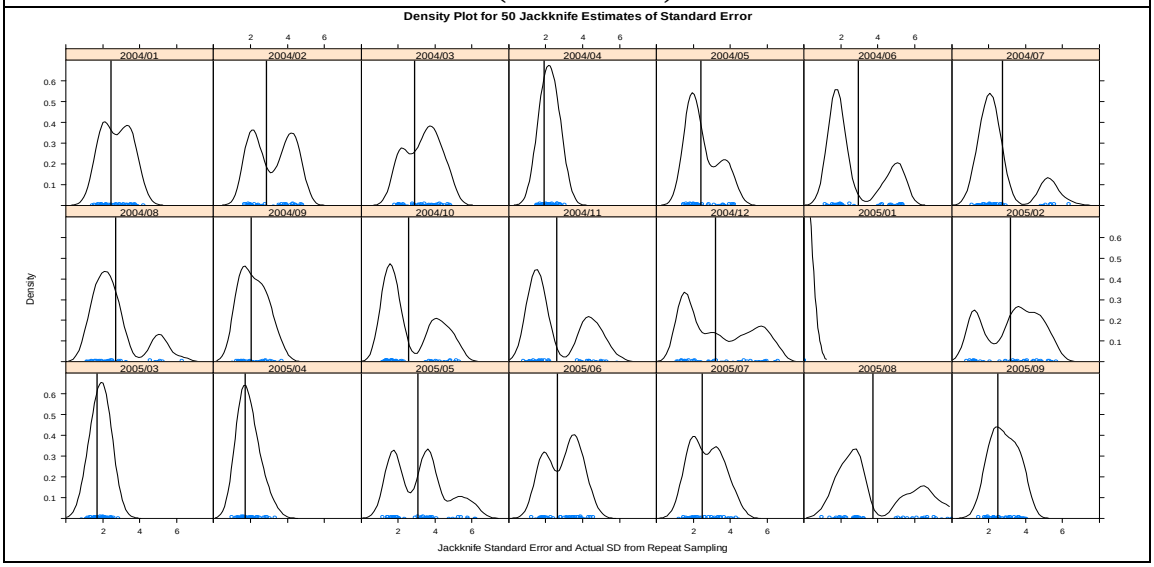
Figure 5 Comparison of the Density Plots for Repeat Samples (black) and Pooled Jackknife Resamples for 50 Samples (grey) – 2005/01 is Base Month



What remains to be seen is how well the jackknife standard error estimates the true standard error as estimated from the repeat sampling. In other words, how much does the jackknife standard error depend on the choice of sample?

Figure 6 shows the density plot of the jackknife standard errors for the 50 different samples compared with the actual values from repeated sampling. What is quite clear from this diagram is that the mode for unimodal (or nearly unimodal) densities is very close to the real value. The question then arises as to the cause of any additional modes in these density profiles.

Figure 6 Density of Standard Errors from Jackknife Estimation compared with Actual Standard Deviation of Index for SS3 (vertical bars)



Comparison of Figures 3, 5 and 6 shows that, where the distribution of index values in any month is noticeably bimodal, this is reflected in the bimodality of both the

jackknife resample index value density and the density of standard error values from jackknife standard error estimates on multiple samples. We see, especially in the overlaid index time series plot of Figure 3, that there are often two or even three clusters of index values - depending perhaps on which route through the classification tree was taken by the sampling process. It is natural, therefore, that these clusters might have different dispersions and, if the jackknife process happens to resample from a sample in one of these clusters, the dispersion of the jackknife histogram will reflect the characteristics of that cluster rather than another. In 2005/4 for example, there is only one such cluster and the jackknife estimate of standard error is closely matched to actual value from repeated sampling.

Table 7 shows the ratio of jackknifed to repeat sampled interquartile range (IQR) and the ratio of jackknifed to repeat sampled standard deviation (SD) for these empirical distributions of index number value.

Month	Ratio of Jackknifed to Repeat Sampled IQR	Ratio of Jackknifed to Repeat Sampled SD
2004/01	1.45	1.60
2004/02	1.54	1.59
2004/03	1.56	1.59
2004/04	1.49	1.65
2004/05	1.39	1.56
2004/06	1.23	1.45
2004/07	1.62	1.64
2004/08	1.31	1.52
2004/09	1.34	1.48
2004/10	1.39	1.57
2004/11	1.38	1.63
2004/12	1.43	1.53
2005/01	Base Month	NA
2005/02	1.32	1.42
2005/03	1.32	1.50
2005/04	1.48	1.62
2005/05	1.40	1.46
2005/06	1.46	1.51
2005/07	1.38	1.45
2005/08	1.54	1.63
2005/09	1.48	1.59

Table 7 shows that the jackknife technique is a conservative estimate of the true variability as might be expected (Efron and Tibshirani, 1994). This finding is echoed in Figure 5 where we can see that the jackknife and repeated sampling distributions are broadly the same shape but that the former are somewhat wider. However, it is clear that the level of overestimation of the actual variability of the index by the jackknife method is quite similar across the months for this single Meat group index.

7. Discussion and Conclusions

As mentioned above, the purposive (deterministic) schemes SS1 and SS2 are examples of 'cut-off sampling'. The work of de Haan et al. (1999) (discussed in

Dorfman et al. (2006)) concluded that such sampling methods could outperform simple random and probability proportional to size sampling in terms of mean square error.

1. In terms of bias, as Figures 3 and 4 and Table 2 show, the cut-off schemes are worse than the *pps* ones.

   This could be naturally explained by the fact that the *pps* schemes are more flexible and take better account of the price variations within the COICOP+ groups. It should be stressed that the cut-off schemes may well outperform probabilistic methods for other populations.

2. The *pps* schemes are very alike in terms of average index level but SS5 is worse than the others in terms of RMSE, and breaking down this error into bias and standard deviation shows that dispersion is the dominant error for this scheme.

   SS4 does not sample the 'Other Meat' category which is the largest expenditure item at Level 1, and hence it is quite surprising that SS4 has very similar performance to the other *pps* schemes. Reference to Figure 3 shows that the size of the sample can be seen to be very similar across the 21 months. This would be expected if the baskets are roughly similar and, though the sample size was increased to take account of this, we would still expect this to have some effect on the index yet it seems not to. On the other hand, around 80% of the basket of meat products will be the same for SS3 and SS4 and so this may account for this similarity between these two *pps*-based methods in this case.

   The difference between SS5 and the other *pps* schemes could be explained by the fact that the sampled indices – *pps* and cut-off alike - target high expenditure items. Method SS5 compels sampling across items at all levels of the classification. Therefore, it covers a wider range of products (relative to the whole population which contains both high expenditure popular lines and low expenditure miscellaneous products with individually low expenditure). This may explain why there is greater dispersion in this index and concomitantly less bias.

3. Surprisingly, there is very little difference between SS3 and SS6 so the additional stage of sampling in the latter offers no advantage. As yet, there seems to be no obvious explanation for this result.

4. The jackknife method overestimates the standard error by 50% on average and repeating the jackknife for different starting samples shows that these estimates can vary. However, when the distribution (empirical) of jackknife estimates is strictly unimodal so is the distribution of index values estimates and the actual standard error is very close to the mode of the former.

.

   The jackknife results show that this technique can give good estimates of the standard error of the index provided that the sample is a representative one. For those months where the repeated samples are relatively homogeneous, we

see a very good correspondence between the actual standard error of the index from repeat sampling the mode of the repeated jackknife estimates.

In future work, will extend these results to include other COICOP groups and also to explore the comparison between the jackknife and bootstrap techniques for the standard error estimation. Additionally, we intend to explore the relationship between the ordering of the classification levels and the sampling performance.

Acknowledgments:

This study is part of a wider project funded by the U.K. Office for National Statistics (ONS). The authors are most grateful to the ONS for permission to reproduce some of this work in the form of this paper. The views expressed in the paper are those of the authors and not the ONS. Helpful advice and stimulating discussions during the writing of this paper was received from Matt Berger, Richard Campbell, Lewis Conn, Kat Pegler, Matthew Powell and Jo Woods. We are most grateful to Mick Silver, John Doyle and participants at the 2012 ICES Conference in Montreal for their valuable comments. 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.

8. References

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

2. 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

3. 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

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

5. Fenwick, D. , Ball, A., Morgan, P. and Silver, M. (2003) 'Price Collection and Quality Assurance of Item Sampling in the Retail Prices index: How Can Scanner Data Help?', *'Scanner Data and Price Indexes' Studies in Income and Wealth, ed, Feenstra, R.C. and Shapiro, M.D.,* vol 64, pp 67-87, NBER University of Chicago Press

6. 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

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

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

9. Leaver, S.G. and Larson, W.E. (2001) 'Estimating Variances for a Scanner-Based Consumer Price Index' *Proceedings of the American Statistical Association, Government Statistics Section*.

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

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

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

13. Schaefer, K.C., Anderson, M. A. and Ferrantino, M.J. (2008) 'Monte Carlo Appraisals of Gravity Model Specifications', *Global Economy Journal*, vol 8, No.1, pp. 1-26

14. Silver, M. and Heravi, S. (2001) 'Scanner Data and the Measurement of Inflation', *Economic Journal*, vol 111, pp 384-405

15. Wolter, K. M. (2003) 'Introduction to Variance Estimation', Springer Series in Statistics, 2nd Edition

16. Wu, C. F. J. (1990) 'On the Asymptotic Properties of the Jackknife Histogram' *Annals of Statistics*, vol. 18, No. 3, pp. 1438-1452

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

Appendix 1

Classification procedure for Meat attributes

The COICOP Meat group has the highest national expenditure of all the 13 COICOP Food Groups across the time period of the study. There are 5 'COICOP+' subgroups within Meat - Beef, Pork, Lamb, Poultry and Other Meat. This last category is unusual in that it has the highest expenditure of these 5 subgroups despite being the miscellaneous or 'dump' category. We discuss our classification for Meat under each of these headings as follows. The data presents us with anything up to 22 different attributes many of them with a substantial proportion of missing values and/or ambiguity. To resolve ambiguities we resorted to cross-tabulation between attributes. Some attributes had compound values such as 'frozen chicken', 'fresh duck', etc. where we could use keywords to split the attribute into two simpler ones. Generally, we aimed to avoid constructing or adopting attributes with too many levels and/or too many missing values.

Beef

One of the attributes marks the **Preservation/Storage** of a product as fresh, frozen, cooked, etc. There is also an attribute which provides a useful classification of products by **Cut/Cooking** method, e.g. 'stewing steak', mince, etc. A further attribute segments the product by country of origin – British vs. Non-British in this case. However, the first attempt at a classification produced too many levels and an underpopulation of quotes at the lowest level and hence this 'country of origin' variable was left out of the final classification.

Pork

The **Preservation/Storage** attribute is again useful here and another attribute again gave the same information on cut of meat/ intended cooking method as **Cut/Cooking** gave for Beef. 'Country of origin' was again not used in the final classification for the same reason.

Lamb

This was treated in the same way as for the two previous subgroups – the 'country of origin' being again withheld.

Poultry

There is an attribute relating to branding which was not seen to be useful. However, there is again another attribute related to **Preservation/Storage**. This attribute contained some ambiguities which were resolved by cross-tabulation with other attributes. It turned out, for example, that 'fresh' referred to 'pies', 'pasties',etc. and 'frozen' was equivalent to 'sausage meat'. In similar fashion, we were able to construct an attribute identifying the **Type of Poultry** with the values 'chicken', 'turkey' and 'other'. 'Country of origin' was not identifiable for a large proportion of the products even if we had wanted to use it.

Other Meat

The only usable attribute for **Other Meat** apart from **Branding** (whether 'Branded', 'Private Label' (PL) or 'Unbranded') was again a **Preservation/Storage** attribute. Ambiguity of category was again resolved using cross-tabulation as above.

Appendix 2:

Table 8: Summary measures for cut-off sampling with different classification levels across the price index time series

Measure	Sampling Scheme			
	SS1 Uses Levels 1, 2 and 3	SS1 modified to use Levels 1 and 3 only	SS1 modified to use Levels 1 and 2 only	SS2 Uses only Level 1
Bias	-3.27	-0.38	-3.10	-0.66
RMS Bias	4.19	3.12	4.84	2.78

From this table we can see that dropping the Level 2 (Storage/Preservation) reduces the bias greatly (though the bias is still negative) and the RMS Bias is much the same.

Figure 7: Summary Plots by Month for cut-off sampling with variation of classification levels compared with the Population Value