UNITED NATIONS

E

Economic and Social Council

Distr.
GENERAL

ECE/CES/GE.22/2008/5
26 February 2008

Original: ENGLISH

ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Joint UNECE/ILO Meeting on Consumer Price Indices

Ninth meeting
Geneva, 8-9 May 2008
Item 3 of the provisional agenda

COLLECTION AND PROCESSING OF PRICE DATA

THE ROLE OF THE TUKEY ALGORITHM IN VALIDATION PROCEDURES FOR PRICES DATA IN A CONSUMER PRICES INDEX: THE UNITED KINGDOM EXPERIENCE ON THIS AND MORE GENERAL ASPECTS OF DATA EDITING

Note by the Office for National Statistics, United Kingdom

Summary

The editing procedures that are used to detect and correct errors in price quotes recorded in shops by price collectors are a non-trivial issue as they can have a systematic numerical impact on measured inflation which can lead to bias. The adopted procedures also have operational consequences. If applied correctly, editing procedures can not only improve the quality of the price index but also result in operational efficiencies in the compilation of the Consumer Prices Index and the Retail Prices Index. This paper reports on the results of some work undertaken by the United Kingdom Office for National Statistics in relation to the application of the Tukey algorithm and considers the issues which arise for compilers of the Consumer Price Indices. It also reports on a more detailed study undertaken into the impact of editing on the price index for clothing. Both pieces of work were undertaken a few years ago and resulted in the introduction

of revised and improved editing procedures. The work was originally undertaken in the context of the Retail Price Index but applies equally to the Consumer Price Indices [1].

Keywords: CPI, Tukey, data validation processes, application, efficiency, accuracy, clothing index.

I. INTRODUCTION

1. For the United Kingdom (UK) Consumer Prices Index (CPI) and Retail Prices Index (RPI), two distinct computer algorithms operate at headquarters to identify outliers amongst the prices collected locally from shops, that is, extreme prices which could have a relatively large effect on their respective item indices. The prices are weighted together to produce published sub-indices of the All Items Index. At the time of the study, the presumption held that an outlier was incorrect, and therefore declared to be invalid, unless subsequently validated by editing, that is, it would be accepted as a legitimate price quote and determined to be correct only after verification from an examination of all the meta-data sent in by the price collector and by appropriate follow-up action. The latter included asking the price collector to recollect the price quote and telephoning the shop to confirm that a price was correct but could also involve no more than headquarters staff making a judgment based on the metadata submitted by the price collectors.

2. Because prices in shops can change and the CPI/RPI production timetable is very tight, editing needs to be done in a short period of time. In ideal circumstances, auditing and editing should be close to "real time". Prolonged and indiscriminate examination of prices is extremely time consuming and not an operational option. "Real time" auditing and editing in the field has an important role to play. Price collection in the UK benefits from the use of handheld computers, which allows interactive editing of prices at the time of price collection[2] and before the data arrives at headquarters. Editing at headquarters follows editing in the field.

3. The first algorithm used to detect outliers at headquarters comprises of two tests which are applied to each "quote". These are the minimum-maximum test and the percentage change test which identify respectively prices or price changes which lie outside a pre-determined range The latter is determined implicitly, according to previous months' average price for the item being examined, and explicitly by the setting of the corresponding parameters. This editing focuses on the price level and change, for a particular item being priced in a particular shop against the average price level and change for the "same/similar" item in all shops. It differs from interactive editing in the field which is done by reference to the price history of that

[1] The CPI is the main measure of domestic inflation for macroeconomic purposes and in an international context is referred to as the Harmonised Index of Consumer Prices. It was launched in 1997 and became the UK inflation target in 2003. The RPI has a much longer history. It began life as a compensation index in the First World War and only much later came to be used as the main domestic measure of inflation for macroeconomic purposes. More details can be found in "The new inflation target: the statistical perspective", Roe and Fenwick, Economic Trends, January 2004.

[2] For example, using outer-bounds of the price change compared with the previous month.

particular item in that particular shop and by following up any apparent discrepancies by asking shop staff to confirm the correct price. This operational procedure at headquarters- the first algorithm- is referred to as "scrutiny".

4. The second algorithm which is used, the Tukey algorithm, is applied to those price quotes which are not identified as outliers by the process of "scrutiny". The Tukey algorithm is, in essence, a more sophisticated version of "scrutiny". More details are given in the next section.

5. Price quotes which are identified as outliers by either algorithm are subject to further checking and editing. At the time of the study an average of around 2000 price quotes, from a total of over 100,000, were being identified as outliers by "scrutiny"- due overwhelmingly to the percentage change test- and around 4000 price quotes were identified as outliers by the Tukey algorithm. All "scrutiny" outliers and half of the Tukey outliers were subject to further checking, for example, by re-pricing in the shop or seeking verification from the shop keeper over the telephone or by looking at the metadata and making a judgement.

6. Editing of "scrutiny" outliers usually persists after the running of Tukey, so the two processes often take place in parallel. In the event most outliers (from either algorithm), which were subsequently checked, were manually accepted, that is the CPI/RPI compiler looked at the available evidence and decided the price quote was correct and should not be revised. They then manually over-rode the automatic editing. Thus, having gone through these rigorous procedures very few outliers were rejected. The number of quotes that were explicitly accepted at this point in the editing process was around 100 times the number explicitly rejected.

7. The advantage of using such algorithms as filtering mechanisms is that they avoid the price analyst from having to examine an extremely large number of price observations. But the experience in the UK supports indicatively the view taken by the ILO Manual on Consumer Price Indices that the use of automated deletion systems without the benefit of manual checking is to be avoided because price changes can vary significantly between months, for example due to sales, and between different varieties of a product within a month, because of seasonality for instance. This applies as much to the statistical checking of input data where, for a particular time period, each price change is compared with the change in prices in the complete sample for the particular product under examination as to less sophisticated systems, based say on the use of predetermined limits which are not automatically reviewed during the editing process. It is because of these concerns and the results of investigative work which was undertaken that the UK abandoned the presumption that an outlier was incorrect.

II. THE TUKEY ALGORITHM

8. The Tukey algorithm is an example of a filtering system based on the statistical checking of the input data.

9. To apply the Tukey method, the price quotes are ordered by the corresponding price ratios and the highest and lowest 5 per cent are flagged for further investigation and excluded. Price ratios equal to one (i.e. where there has been no price change) are also excluded. The

arithmetic mean of the remaining price ratios is calculated – equivalent to a type of trimmed mean- and this mean is used to divide the remaining price ratios into two groups and their respective means (referred to as the upper and lower "mid-means"- AM_U and AM_L) are calculated. The upper and lower Tukey limits used to flag those price observations which warrant attention are then calculated as follows:

$T_U = AM + 2.5 (AM_U – AM)$
$T_L = AM – 2.5 (AM – AM_L)$

where AM_L is the lower trimmed mean and AM_U is the upper trimmed mean.

10. The particular attraction of the Tukey algorithm is that, meta data to one side, it maximises the use of the immediate price history and provides more practical and realistic parameters since it excludes cases where there has been no price change. It does, of course, rely on a sufficiently large number of price observations. Tukey can be applied to any time period and therefore may be used to examine both the monthly and annual change. The upper and lower limits are determined by the data and can be regularly re-calculated from the current prices data set. This and the use of the immediate price history can be a particular advantage when prices and the inflation situation are changing quickly.

III. THE ISSUES

11. Two main issues arise:

(a) The efficiency of the editing procedures;
(b) The impact on the accuracy of the CPI/RPI.

The latter includes the potential for bias. These two issues of efficiency and impact are, of course, inter-related. Each is considered in turn.

Efficiency of the editing procedures

12. The efficiency of the editing procedures is a function of:

(a) The overall system for data handling, including the full processes in place for data validation and editing;
(b) The role played by the Tukey algorithm in the overall system;
(c) The precise application of the algorithm.

13. Looking at the data validation process as a whole, it is instructive to note that there is a significant overlap between the three main editing processes: interactive editing in the field; "Scrutiny"; and Tukey.

(a) The interactive editing in the field comes from a number of real time data checks built in to the handheld computers. After the price collector has inputted the price, together with useful meta data (including whether the price is a special "sale" price and notification of a

replacement item where the previous one is no longer stocked), the collection programme checks the input by means of a series of in-built rules. These include minimum and maximum price ranges, based on the price inputted in to the handheld computer the previous month for exactly the same item in the same shop, and a series of logistical checks, for instance that a "recovery from sale" price only follows a price inputted the previous month as a "sale" price. No amount of editing at headquarters can replace interactive editing in the field- it is the most effective means for ensuring the accuracy of the prices recorded.

(b) "Scrutiny", unlike "interactive data editing", is not done in "real time" and is a less sophisticated post-collection editing process than Tukey. The relative merits of "scrutiny" can depend, in part, on its interaction with the other two, most particularly with Tukey. It is instructive to note that "scrutiny" was in place before the adoption either of handheld computers for price collection or more sophisticated editing processes at headquarters. The latter two innovations were introduced together as part of a series of measures to improve the accuracy and reliability of the CPI/RPI and to tighten up the quality management of the compilation processes to reduce the risk of errors. Of the three editing processes, "Scrutiny" makes the least "intelligent" use of the data. "Scrutiny" in large part survived for historical reasons- prices used to be collected on paper and were received in "batches" so an outlier detection programme was needed that, unlike Tukey, could be run without reference to the main body of prices data. Data is still, to a large extent, received in batches, before the initial "prices" data set is loaded on to the computer but the benefits of "Scrutiny", in terms of timeliness, were more marginal, so the question at the time the study was undertaken was whether "Scrutiny" sufficiently added to the quality of the final prices data. Its main advantage was seen to be its ability to identify extreme outliers and big differences in quality when replacements are chosen for items which have disappeared. It can also quickly deal with prices queried by the collectors themselves.

(c) The Tukey algorithm is more sophisticated than either the interactive editing of prices in the field via handheld computers or the process of "scrutiny" at headquarters but does not benefit from the inherent advantages of interactive editing in the field. There is a prima facia case that editing at headquarters should be concentrated on those outliers defined by Tukey. This is particularly so as a significant number of quotes defined by "Scrutiny" to be outliers and subsequently confirmed as correct would not be classified as outliers by the Tukey algorithm if they had been subjected to it. Some editing activity is, therefore, misdirected. This "inefficiency" can be accentuated by the disproportionate number of outliers identified by "Scrutiny" which are subsequently found to be correct. On the other hand, scrutiny can edit out extreme price quotes which could "skew" the operation of Tukey.

14. It is clear that the role played by the Tukey algorithm can be undermined and its performance can be impaired by "Scrutiny". Most particularly, the prices which are initially defined as outliers by "Scrutiny", yet which are validated in the intervening period, are excluded from the data set subjected to Tukey, even though their inclusion would better inform the overall measure of price movements (or prices) for each item, from which are determined the acceptance parameters for Tukey - and, hence, the Tukey outliers. Moreover, "Scrutiny" outliers, which are confirmed as being invalid, might reasonably be included in the data set used for Tukey, since Tukey, as applied in the editing of the CPI/RPI, trims ten per cent of the data in any case. Thus, it could be argued that little would be lost operationally in dropping "Scrutiny" although it would

mean that data checking and editing would need to be deferred a few days until enough price quotes are received to calculate the Tukey parameters. This was considered but rejected. The view was taken that despite it reducing the efficiency of the Tukey algorithm there was a net benefit from "Scrutiny" insofar as it can very quick and simply identify extreme outliers. But the view was also taken that the efficiency of "Scrutiny" could be improved by displaying the quote "expenditure" weight alongside other useful information to guide the scrutineer in their decision making.

15. On the issue of application, the presumption that an outlier is "wrong until proven right" may be challenged and was challenged as being erroneous. The fact that quotes defined as outliers by either algorithm were excluded from the index unless subsequently validated by editing, is legitimate only to the extent that the set of outliers overlaps the set of incorrect prices. For "Scrutiny" outliers this does not matter, as they are all checked. In the case of a Tukey outlier, the implicit referencing of its price movement (or price) to a measure of overall price change (or price) for the given item, by which the quote has been defined to be an outlier, could constitute sufficient additional information, further to that available to the collector, to meaningfully invalidate the price. Yet, an examination of the data showed that the proportion of Tukey (and "Scrutiny") outliers which were explicitly rejected by re-checking prices or by using informed judgement was very small. Assuming the editing decisions to be correct, the presumption to exclude an outlier would seem to be inappropriate. In view of this, the set of outliers may not adequately overlap the set of incorrect prices[3], in which case attention might reasonably focus more on price quotes for particular items in particular outlets that have remained unchanged for an unusually long period and on the more extreme outliers.

Impact on the accuracy of the CPI/RPI

16. Research by ONS has indicated that a theoretical effect on the CPI/RPI of the Tukey algorithm, as previously applied at the time the study was undertaken, could be to depress the index by rejecting more upward than downward price changes and which, unless confronted in the editing process, can lead to bias in the published index. This is the consequence of two things:

(a) The skewed distribution of price changes;
(b) The fact that no algorithmic use is made of metadata relating to whether levels of price changes are unusual, for example because of sales and special offers. This information was stored on the prices data base using indicator codes, which were easily accessible.

17. The potential bias can be exaggerated by failing to make use of the metadata submitted with the prices. This is particularly so where the item being priced is associated with particularly large price fluctuations due, for instance, to sales or seasonality.

[3] This might be further explored by matching the set of outliers with a database of locally collected quotes which have failed the back-check accuracy test.

18. The presumption held that an outlier was incorrect, and therefore declared to be invalid, unless subsequently validated by editing was not only unjustified and had the capacity to introduce errors in prices and bias as mentioned above, but it was also in breach of the EU Regulations for the Harmonised Index of Consumer Prices (HICP). Staff at head office were re-trained to always revalidate a price unless there was firm evidence of the need not to.

19. Additionally, during the course of the study investigations broadened and attention focused on:

(a) An automated computer procedure, within the CPI/RPI prices processing programme, that reset all quotes tagged non-comparable to comparable if the price change from the item being replaced was considered sufficiently small.

(b) Giving price checkers at headquarters the authority to over-ride decisions made by price collectors in the field on such issues as whether a replacement item is comparable or not. The judgement by price checkers is based on the metadata provided by the price collector and, where there are concerns over whether a replacement is comparable, also by looking at comparative price and using market knowledge.

These two latter issues were the focus of a second investigation.

20. To summarise thus far, like many data handling processes used by national statistics institutes, the CPI/RPI system is a product of history which has evolved over the years. It is argued that advances in technology - most particularly the use of handheld computers for price collection and the electronic transfer to headquarters of files containing prices - have meant that the same data checks are incorporated into different stages of editing leading to repetition. This lack of coherence can lead to operational inefficiencies. But of greater concern is the potential impact on the measurement of inflation.

21. The section that follows reports on the second investigation as mentioned above. This relates to a more detailed study undertaken into the potential impact of editing on the price index for clothing where sales and seasonality can lead to large short-term fluctuations in prices. It provides a more detailed analysis and a useful insight into some of the issues that can arise.

IV. CLOTHING

22. The study which ONS undertook into the clothing sub-index of the All Items Index was originally spawned by the interest generated in the index from the apparent downward trend in prices, which was mirrored in some but not all other countries in the European Union.

The underlying analysis

23. In July 2002, the Clothing and Footwear index fell below 100 for the first time since the summer sales of 1987. Chart 1 below shows the index level since January 1987, with the moving average plot accentuating the steep decline since 1998, which followed seven years of

steady prices. Between January 1998 and January 2002, the moving average[4] index plot in Chart 1 had decreased by around 14%. The question was raised of whether the prices for clothing really were below the level of 15 years ago for similar items.

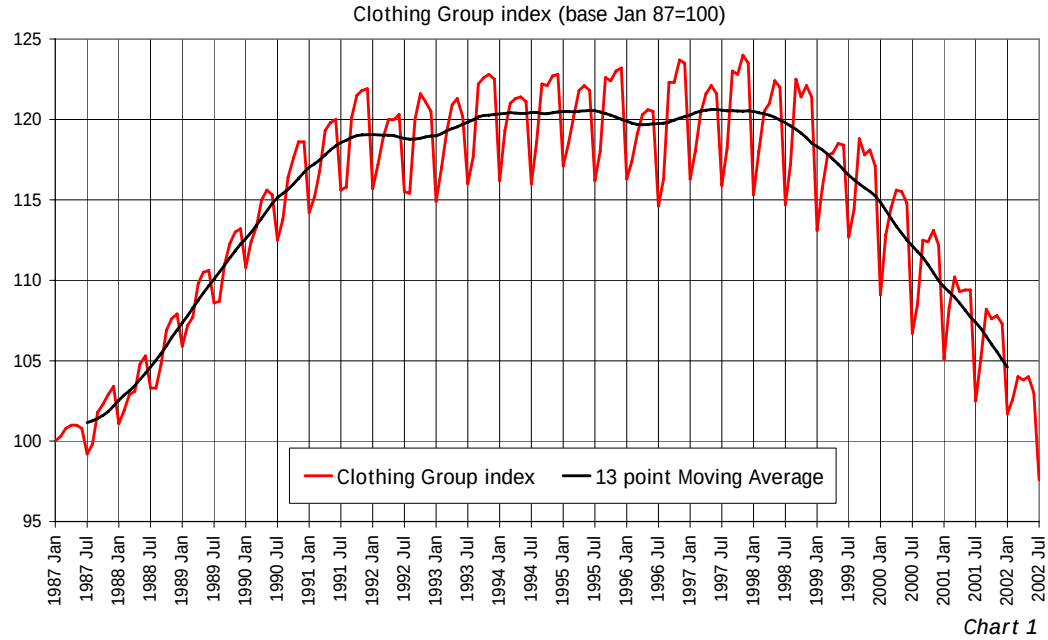


Chart 1

24. Chart 2, below, shows the Apparel Index for the USA. This is the most comparable series to the UK Clothing and Footwear Index. It covers the whole of the USA, is the "non-seasonally adjusted" version and includes both men's and women's clothes and footwear. The series has been re-based to January 1987 so that it can be directly compared to the UK data. The chart shows a similar pattern to the UK index, with seasonal sales and recoveries. The increase in prices decelerates in the early 90's and towards the end of the decade prices begin to fall. But the latter decrease is much less marked than in the UK index, the January 1998 to January 2002 moving average decrease being around 6%, compared with the 13% for the UK series. This difference in the rate of fall in prices remains even after economic factors, such as the significant depreciation of the £ sterling towards the end of the period is taken into account. The issue was unresolved from analysis, even more so, given the fact that some other countries, such as Sweden, did not experience any lowering of prices.

[4] The moving average is a 13 point *centred* version with both end points contributing half the amount of the weight of other points.

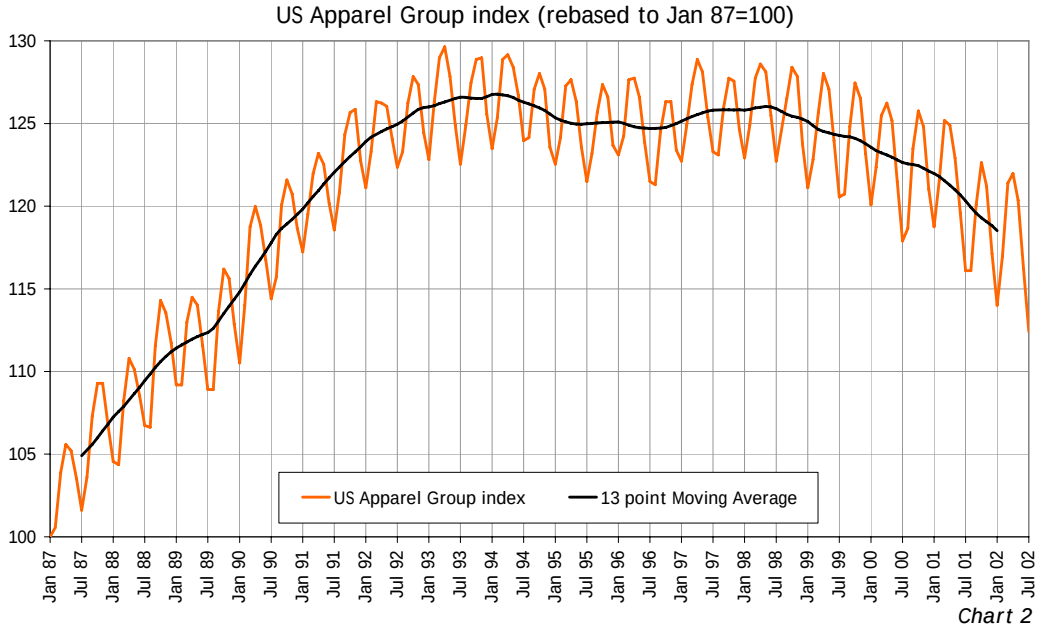


Chart 2

25. Chart 3 illustrates the time series for all sections within the UK clothing and footwear group. Four of them have a common trend pattern – initial increase in prices followed by a period of price stability in the mid 1990's before prices fell steadily during the latter part of the decade. The exception is "other clothing"- a very heterogeneous grouping- where the pattern of increase continued into the late 1990's and only in the last few years experienced a downturn.

26. Particularly worth noting are the sales periods, traditionally around January and July in the UK, and the recoveries that follow these. Chart 3 suggests that in more recent years there have been deeper seasonal sales and shallower recoveries for most of the sections. For the women's outerwear section especially, even between sale/recovery cycles, prices had fallen steadily over the last few years of the period being studied, which suggests that the decline of the index is not just due to seasonality and "fashion".

27. Looking at individual items within men's and women's outerwear, it becomes apparent that there are certain items that exhibited greater downward price movements than did others. Preliminary investigations for women's outerwear, for example, showed that the CPI/RPI indices for items including jacket, blouse and formal dress had all decreased by around 40% between 1995 and 2001/2. On the other-hand, for certain items prices appeared to be stable, or even exhibited small increases.

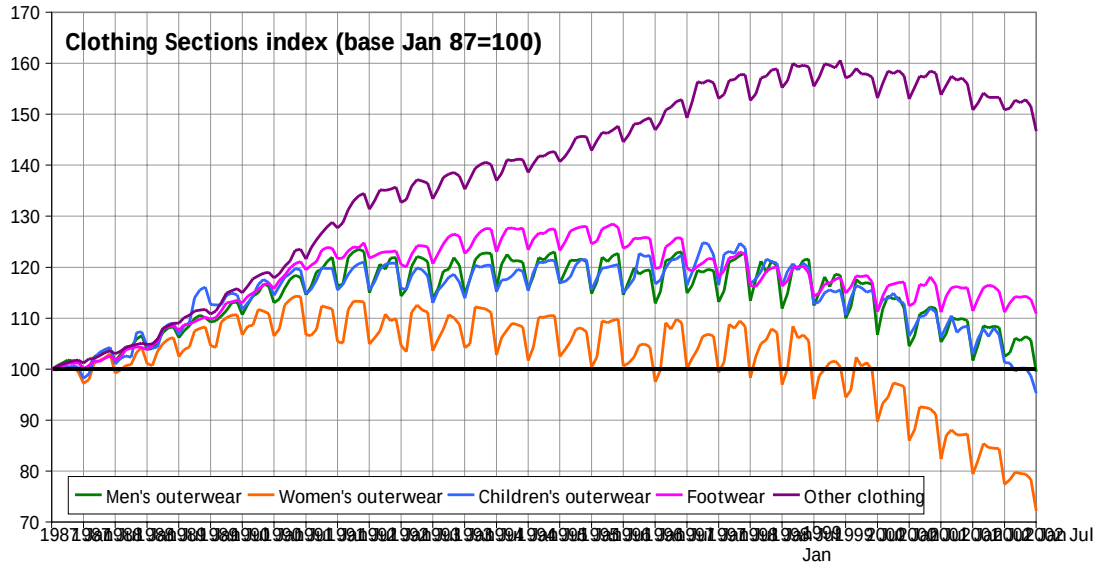


Chart 3

The research

28. A number of more in-depth investigations were undertaken to try and understand what was going on and whether editing and its potential to introduce bias was an issue. The research attempted to address a number of specific questions.

29. Is the judgement of the data editor at headquarters more reliable than that of the price collector?

30. Clearly this is a legitimate question to ask whatever algorithm is used to identify outliers but it is perhaps most relevant for the process of automatically over-riding price quotes for replacement items from “non-comparable” to “comparable” where the price difference is small and to “Scrutiny” which makes less use of the knowledge and general pointers that can be gained from the data to assist in setting the co-ordinates for the algorithms. Perhaps “Scrutiny” warranted the greatest concern in the sense that a prices analyst at headquarters could overwrite in a second, a decision that the collector may have spent several minutes on, with price change being heavily influential on the decision. The question is also relevant in the context of automated data editing facilities in the field.

31. An initial analysis indicated the extent of the potential problem. For example, for women’s outerwear alone, during 2002, on average there were only 10% as many recoveries as sales. Between February and December on average there were 812 prices per month tagged as sales, compared with only 81 recoveries from sales. The concern arises where a “non-comparable” or new item replaces an item in a sale. In these cases a base price must be imputed, so the price rise does not actually contribute to the index, but with the result that there

is no “recovery” price to match the sale price. Thus, it can be strongly argued that the correct identification of whether a replacement of an item at the end of a sale is comparable or non-comparable is just as an important as the correct recording of the price. Automated data editing algorithms, such as Tukey, will not test for this unless part of a broader package of editing procedures.

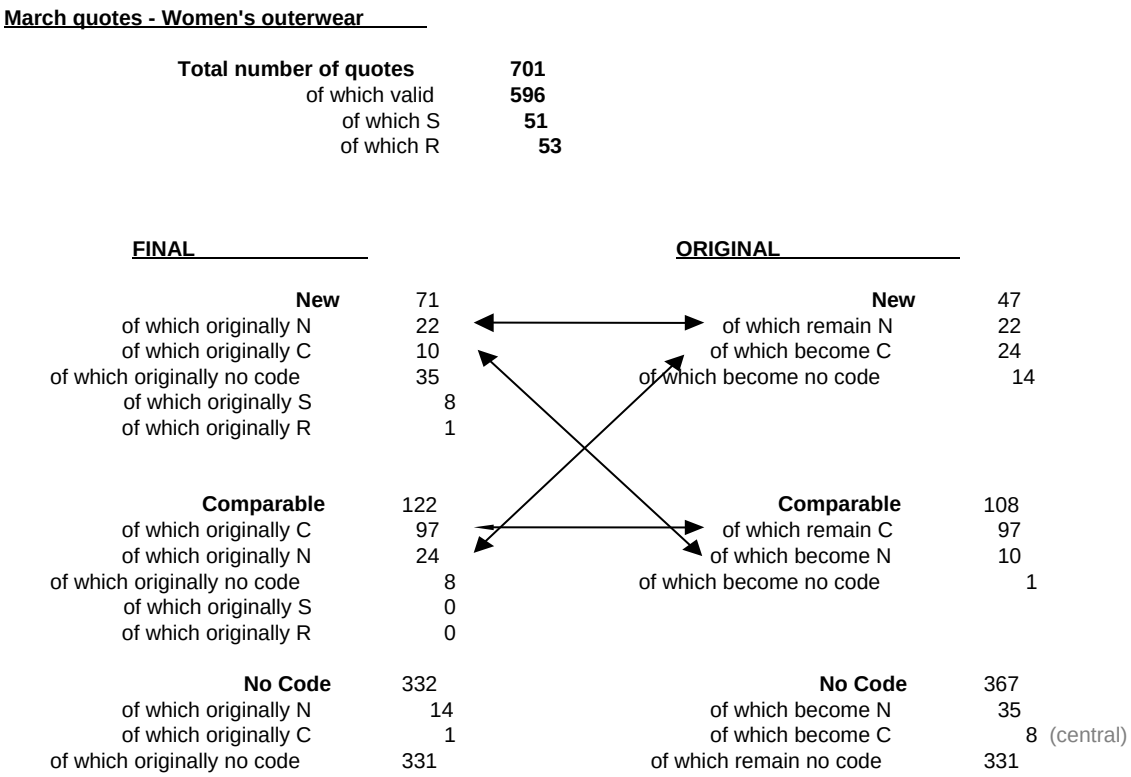
32. The diagram below illustrates the “Scrutiny” process described above for women’s outerwear. The indicator “S” stands for a sale price and the indicator code “R” represents a price which has reverted to its normal price, i.e. it has recovered[5].

33. The top section shows simply the number of quotes, those valid and the number tagged ‘sale’ or ‘recovery’. Below that, on the left, there is a breakdown of final indicator codes, that is after scrutiny and after the prices analysts have made their decisions. On the right shows the analysis of original indicator codes assigned by the collectors and the ones assigned by headquarters. Arrowed lines merely represent similar measures – eg. The number of final N codes that were originally C must equal the number of originally C codes that became N. Note in the table, ‘New’ means a replacement item which is ‘Non-comparable’.

34. It is interesting to note that at the time of the study over half of the quotes that came in tagged N were changed at headquarters, with half of them being re-classified as C. Around 10% of those quotes that arrived tagged C were later changed to N meaning that a new base price would be imputed for them.

[5] An indicator code is assigned to price quotes to aid the validation procedures and provide additional information on the quote by giving it a “status” (such as comparable, non-comparable, sale, recovery etc.). They are initially assigned in the field, though automated computer procedures and office staff can change them when studying the prices. An automated computer procedure resets all quotes tagged non-comparable to comparable if the price change from the item being replaced is “sufficiently small”. Initial indicator codes assigned in the field are recorded so it is possible to see where these have been changed. Comparable (C) is a replacement item chosen for an item that has disappeared from stock. It should be of the same quality as the previous item. Non-comparable (N) is a replacement item that is not the same in quality as the item it replaces. These should only be chosen if a comparable is not available. A new base price will be “imputed” for non-comparables.

UDiagram 1



35. The table below takes the analysis to another level. It looks at the price change from the previous month (or two months ago where there is no price last month) for each of the quotes tagged N (non-comparable) or C (comparable). For example, the top line considers the price changes since last month for all women's outerwear quotes that the collector tagged N and headquarters' staff did not change. It is interesting to note that the average price increase since the previous month was £18.30, with 150 prices increasing and 60 decreasing. The major point of interest from the table is to compare the two rows that have original indicator codes of **C**. It can be observed that of those that:

(a) <u>were changed to **N**</u>
   (i) The average price relative was 75% compared to the previous month's price;
   (ii) The average increase was around £10;
   (iii) Around 80% of those matched to previous prices showed an increase.

(b) <u>remained C</u>
   (i) The average price relative was only 12% compared to the previous month's price;
   (ii) The average increase was around £2.40;
   (iii) Around 70% of those matched to previous prices showed an increase.

36. Contrasting the two bullet points above tends to suggest that within both categories, most of the quotes show a price increase - indicating, based on the earlier discussion, that they might be "comparable items on recovery from sale". However, those that were changed to **N** show a greater increase in price in general. Although it is possible that this price rise is due to an increase in quality and the item being fundamentally different, it is quite possible that in at least some of the cases the price change has influenced scrutiny and a comparable item has actually been excluded from the index and the price rise has been missed. If this phenomena consistently takes place throughout the impact on the index will be cumulative.

Table 1. Price changes by indicator code

Original	Final	Quotes	Matched in last 2 months	Ave price increase	Min price increase	Max price increase	No. of price inc's	No. of price dec's	Ave price rel (%)
New	New	224	214	18.30	-60.00	417.00	150	60	73%
Comp	New	109	102	**9.99**	-119.01	92.95	67	17	75%
No code	New	354	353	-0.33	-30.00	0.00	0	7	-1%
Comp	Comp	978	892	**2.43**	-49.00	145.50	288	120	12%
New	Comp	240	161	1.46	-81.00	45.00	74	43	10%
No code	Comp	8	8	-1.25	-10.00	0.00	0	1	-3%

37. Putting to one side whether they are correct or not, it is possible to check on the impact of these decisions, as we record the indicator code recorded by the collector and the final indicator code assigned to the item as it enters the index. The ONS did so for the clothing group. As noted earlier, it is apparent that when the price quotes enter the published index (final indicator code), there will be fewer non-comparables in the dataset compared with immediately after the prices are collected (original indicator code), due to the automated processes then in place. As many higher priced comparable quotes were being re-coded as non-comparable the original indicator codes might yield a higher index.

38. For comparison purposes, two separate clothing indices, for a four year period were compiled:

(a) The first using the indicator codes as assigned in the field;
(b) The second using the final indicator code after computer validation, scrutiny etc.

39. The results are given in Charts 3-5 below, for the years 2000, 2001 and 2002 respectively. In line with expectations, all three charts depicted a higher index if we were to use the original indicator code, instead of the code assigned after scrutiny. In general the difference over a year between the indices was between three and four index points, although it was smaller in 2000.

Chart 3

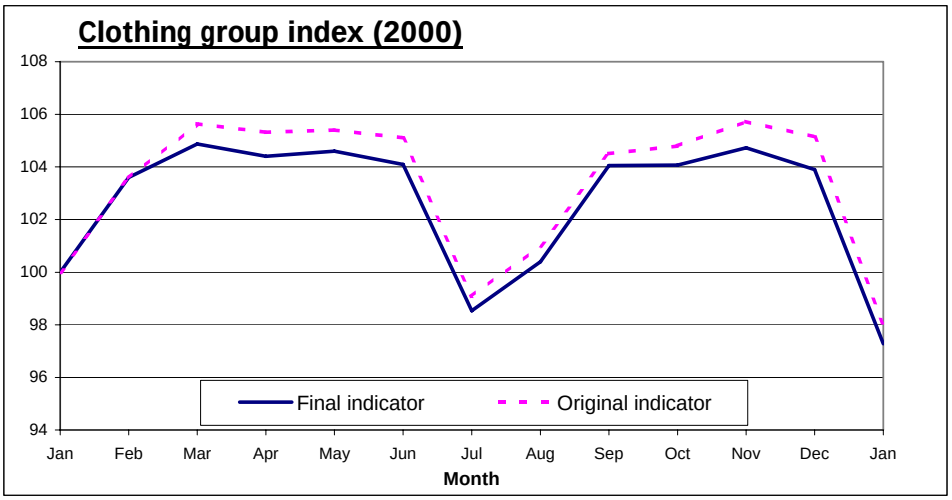


Chart 4

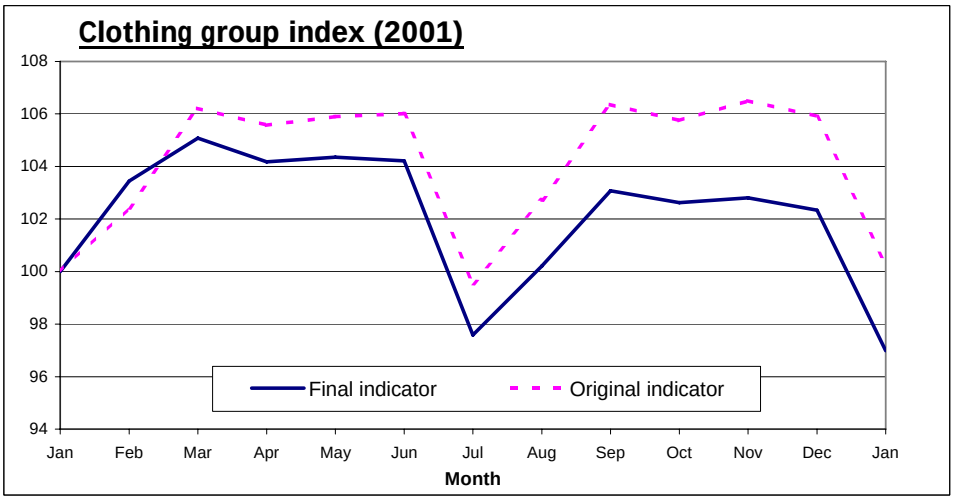
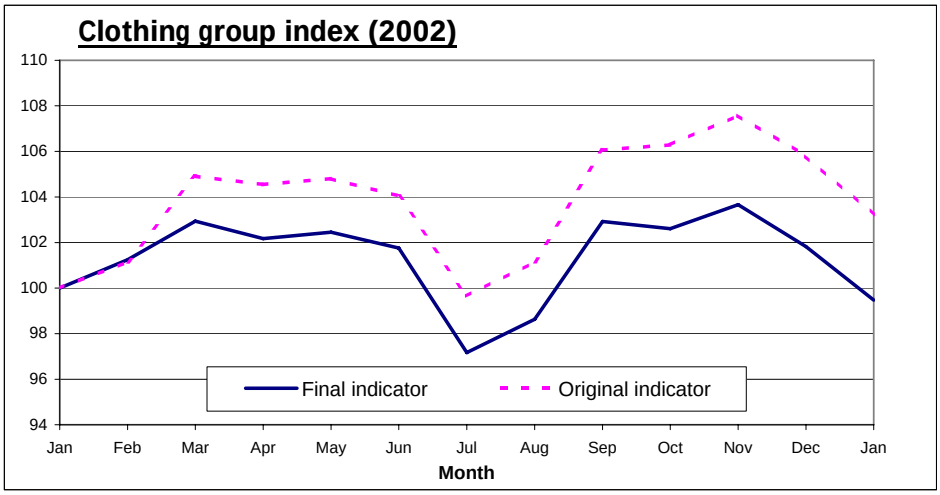
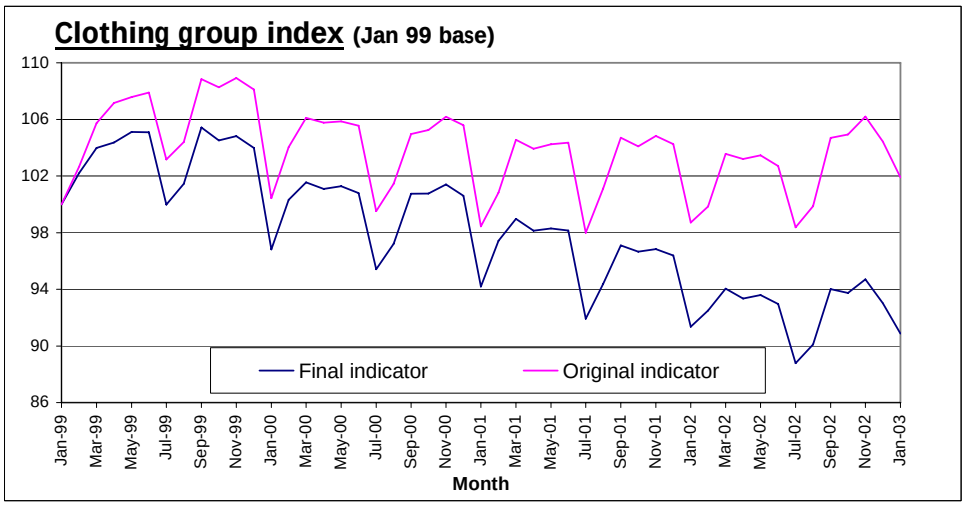


Chart 5



40. The aggregate effect of all the charts can be seen in Chart 6 below, with the chained indices from January 1999. It shows the difference between the two indices over the four-year period was 11 index points - 101.9 instead of 90.9.

Chart 6



41. Although the evidence of the simulations for four years indicates that re-classifying indicator codes can cause a downward pressure on the index for clothing this is not an issue if the reclassifications are correct, i.e. that the judgement at headquarters is better than the judgement of price collectors in the field. A small scale exercise was undertaken to test whether this was the case.

The performance of price collectors: auditor back-check of indicator codes

42. ONS auditors were instructed to back-check the accuracy of the N (non-comparable) and C (comparable) codes at a sample of purposively selected locations. Although the exercise covered three months only, due to the costs involved, it does provide a useful indication of the reliability of the price collectors' judgements. The back-checks were conducted within a few

days of the prices being first collected. The auditors were each asked to judge from both the item descriptions they had and from viewing the item, whether the item was comparable or non-comparable and to provide information which informed their judgement. An excerpt of one of the back-check forms can be seen in Annex.

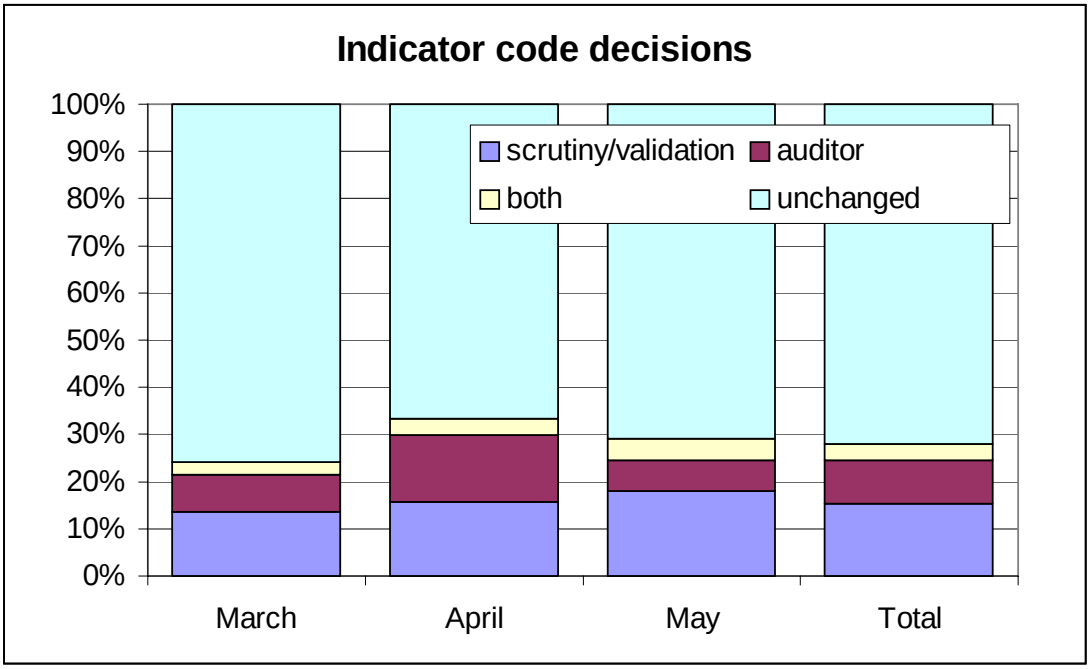
43. The results suggested that, in general, collector decisions on indicator codes were accurate. While the auditors changed some indicator codes, this could be argued to be more a question of the scope for differences in judgement at the margins from an unavoidable element of subjectivity, though a small proportion of these will inevitably be down to wrong decisions by the collector and for this reason there has to be, and in the case of the UK is, a constant collector training program in operation.

44. Table 2 below shows the overall results. These are also represented graphically in Chart 7. Although the majority of price quotes examined by the auditors were left unchanged, it was of concern that where auditors did indicate that the price collector's coding was not valid the change in coding that they recommended did not generally coincide with the editing changes made by the validation procedures at headquarters.

Table 2

	March	April	May	**TOTAL**
Total quotes	256	147	155	558
Total changed by scrutiny	35	23	28	86
Total changed by auditor	20	21	10	51
Total changed by both *(not incl above)*	7	5	7	19

Chart 7



45. The two tables below break the results down by original indicator code (the indicator assigned by collector), focussing on the changes from comparable to non-comparable or vice versa.

46. The first of these tables, table 3, shows that of 425 clothing quotes that the collector tagged "comparable"; only 4% were changed during scrutiny/validation, compared with 14% by the auditor. Most interestingly, of the 4% changed in the office, the average price change for the particular item from last month was an increase of £17.86. This compared with an average price rise of only £2.69 for those changed by the auditor.

Table 3

		(month on month) *Ave price change*
ORIGINAL C	425	
of which scrutiny changes to N	4%	*+ £17.86*
of which auditor changes to N	14%	*+ £2.69*

47. The table below shows that of 134 quotes that were tagged "non-comparable" in the field, 63% were changed during scrutiny to comparable. The average price rise associated with these was just £1.28 (which suggests that much of this 63% will have come from automated

procedures – mentioned earlier- that amend "non-comparable" prices with only a small price change). Conversely the auditor changed only 9% of these 134.

Table 4

ORIGINAL N	134	
of which scrutiny changes to C	63%	*+ £1.28*
of which auditor changes to C	9%	*+ £2.13*

48. The although the extent to which generalisations can be made will depend on local circumstances, including the quality of training provided to the price collectors, the overall conclusion to be drawn is that in normal circumstances the judgements made by price collectors and auditors in the field are better informed and preferable to those made at headquarters.

V. FURTHER THOUGHTS ON TUKEY

49. So far, the paper argues for greater emphasis to be given to greater coherence and integration in outlier detection and editing procedures and for greater use to be made of metadata combined with greater confidence being given to the judgements made by price collectors. The paper also argues that Tukey is a preferred algorithm as long as it is used efficiently and intelligently.

50. Ideally, the compiler of the CPI would receive all data in a single, timely delivery, and perform the Tukey algorithm on all of the data, otherwise its performance is impaired by the omission of (valid) data. But for most index compilers this is not the position. The question then arises as to whether a surrogate could be realised to identify from the first data set those quotes which would be most likely to be defined as outliers if the Tukey algorithm were to be applied to all of the data. As we have seen in the case of the CPI/RPI this isn't achieved by the process of "Scrutiny", although the latter does have operational advantages.

51. The Tukey algorithm is robust in the sense that the implicit thresholds which define outliers do not change much with the addition of a relatively small dataset. Thus, most quotes provisionally defined as outliers by the first operation of Tukey will be likely to be confirmed as outliers after the second operation of Tukey. Moreover, the number of provisional outliers may be limited, and attention focused automatically on the more extreme price movements in the first data set, by adjusting the explicit parameters used in the initial operation of Tukey. Indeed, different parameters might be used for different items, to reflect the respective proportions of quotes which are collected centrally (and therefore included in the second data set).

52. The Tukey algorithm performed twice - the first operation would apply to the initial, main set of locally collected quotes; the second operation would be on all quotes, both locally and centrally collected – has some attractions. Any validation decisions that are made after the

first Tukey operation could be suspended, to be enacted only if the quote is defined to be an outlier after the second operation of Tukey. The option would remain to preserve or ignore decisions on any quotes that are not defined as outliers after the second operation.

VI. THE ISSUE OF CENTRAL PRICE COLLECTION AND CENTRALLY CALCULATED INDICES

53. The observations made so far in this paper relate to local price collection, is used for most items. Prices are obtained from outlets in about 150 locations around the country. Some 110,000 quotations are obtained by this method. Normally, collectors must visit the outlet, but prices for some items may be collected by telephone.

54. A discussion about editing procedures is incomplete without a reference to central collection. Central collection is used for items where all the prices can be collected centrally by the ONS with no field work. These prices can be further sub-divided into two categories, depending on their subsequent use:

(a) Central shops, where the prices are combined with prices obtained locally; and
(b) Central items, where the prices are used on their own to construct centrally calculated indices.

55. Central shop prices are obtained from major chains of shops with national pricing policies. Branches of these chains are excluded from the local collection. Some chains fill in paper forms; others enter price data on spreadsheets via emails, or the data is obtained from the company's internet website. Mail order catalogues are also treated as central shops: prices are recorded as and when the catalogues are issued (generally twice a year). These prices are combined with those for the same items from the local collection.

56. In most cases, the retailers choose the products within the item specifications for which they send prices, but in some cases they send a complete price list, from which ONS staff choose the product to price. The choice is based on experience of what makes a good indicator.

57. There are about 130 items for which the prices are collected centrally and the index calculation is carried out separately from the main method of index production. Around 110 of these are used in the RPI, the remainder are for use in the CPI (e.g. unit trust charges and in the CPI new car index). Selecting this type of collection and calculation is usually dependent on one or more of the following considerations: sources of data; data presentation; frequency of price changes; national pricing policies and the possibility of future fundamental changes to pricing methods.

58. For most of these items, the method of collection and calculation is based on a generic model. Indices are aggregated from the lowest level up, with weights often available at the level of individual price quotes. The weights data used in the centrally calculated indices come from a variety of sources, which are usually specific to a particular index. Where feasible, price data are collected over the internet. Otherwise, price data are collected from one central source (trade associations, Government departments etc) whenever possible although more often than not the

collection of prices requires contact with regional or a number of competing companies. Prices may be requested in writing, by telephone or by e-mail, or may come automatically because the ONS is on a provider's mailing list. Providers may send either a full price list or tariff sheet from which the relevant prices will be extracted. Some travel fares data are provided in the form of price indices. All price quotations must be confirmed by some form of written documentation. Frequency of enquiry varies across the range of items and depends on when prices are known or expected to change. The most common frequencies are monthly or quarterly but thrice (e.g. some travel fares), twice (e.g. local authority rents) and once a year (e.g. football admissions) as well as 'when necessary' (e.g. when changes to national rail fares are announced) are also included in the timetable.

59. The importance of these centrally collected or centrally calculated indices in connection with a discussion on editing is threefold:

(a) Although relatively small in number they represent about 40 to 50 per cent of the CPI/RPI in terms of household expenditure;
(b) Because of the relatively high weight given to any individual price quote in centrally collected or centrally calculated indices there is more of an inherent risk that a pricing error could lead to an error in the published index;
(c) Inevitably more time spent editing locally collected prices will mean less time spent on editing and checking centrally collected prices and centrally calculated indices.

60. The Tukey algorithm cannot be applied to centrally collected prices and centrally calculated indices in isolation given the limited number of price quotes. Reliance is made on scrutiny procedures at headquarters, based a similar process to those adopted in the field for local price collection, i.e. the minimum and maximum price ranges based on the price collected in the previous period for exactly the same item and a series of logistical checks, for instance that a "sale recovery" price only follows a price inputted the previous month as a "sale" price. In addition, the prices obtained are checked against expectations- for instance, checking against when an annual price review can be expected and prior reports on planned price increases. Centrally collected prices are subjected to the Tukey algorithm after the prices are combined with those for the same items from the local collection.

61. From this it can be seen that centrally calculated indices represent the biggest inherent risk. This is particularly so given the fact that many such indices relate to services where there is the added challenge of dealing with changes in complex tariff structures. For instance, utility prices, fares and mobile telephones. To check that the indices have been calculated properly requires:

(a) Collection of the previous prices and a note of the reasons for the price change;
(b) A check of the pricing information received against pre-announced price changes;
(c) An ongoing review of methodology, most particularly to ensure that the average price change generated by a new tariff together with an allowance for changes in service provision are properly reflected in the calculation;
(d) An independent check by another index compiler.

62. Implementation of an optimally efficient procedure means spending more time per quote on centrally collected prices and centrally calculated indices than on locally collected prices.

63. The ONS introduced a three-tier validation process for centrally collected and calculated price indices:

(a) The compiler at headquarters who collects the prices checks for completeness and accuracy, for example by comparing with previous prices and pre-published price increases;
(b) Another index compiler will check to see that the index looks sensible and that there is a rational interpretation. At a minimum, if the index doesn't "look sensible" all prices will be re-checked, although their goal will be to check all prices;
(c) The Price Statistician will check all associated spreadsheets for the logic behind the calculation and for calculation errors and odd-looking prices. This is the final check before publication.

VII. SUMMARY AND CONCLUSIONS

64. Earlier on in this paper it was stated that the advantage in using algorithms, such as Tukey, as data editing "filtering" mechanisms is that they avoid the price analyst at headquarters from having to examine an extremely large number of price observations over a very short period of time and making snap judgements. But in practice their performance in doing so and in providing an effective process for identifying and correcting true errors in the price quotes which enter the CPI, is not guaranteed and depends on a number of factors relating to the specific way in which they are applied.

65. The United Kingdom studies, undertaken a few years ago, indicated that;

(a) Although Tukey may be the preferred outlier detection algorithm, its performance can be impaired by the omission of (valid) data as a result of earlier editing (the "scrutiny" process).

    (i) Tukey is further undermined if there is a presumption that an outlier is incorrect (i.e. invalid) unless subsequently validated by editing.

        a. The legitimacy of the presumption to exclude an outlier is challenged by the very small number of outliers whose status as such is explicitly confirmed during active editing. The presumption could, in fact, introduce bias.

    (ii) Tukey can also be undermined by the automatic re-coding of non-comparable "replacements" as comparable "replacements" where the price difference is very small and by the lack of reference to the corresponding "indicator codes" (C for comparable, N for non-comparable, S for sale price etc.).

(b) The judgement of the price collector is generally more reliable than that of staff at headquarters (although there are exceptions).

(c) Centrally calculated indices are relatively complex and account for about 40 per cent of the CPI/RPI "basket" by expenditure weight. They represent a greater inherent risk of a pricing error leading to an error in the published index.

66. A number of actions were taken as a result of the studies described in the paper:

(a) "Scrutiny" was retained due to its ability to identify at an early stage extreme outliers and big differences in quality from pricing replacement goods, but was improved by displaying quote "weights" alongside other useful information.
(b) The presumption that the outlier is wrong until proven correct was modified-scrutineers now adopt the approach of revalidating a price quote initially identified as an outlier unless evidence already exists to the contrary.
(c) The collector judgement is only over-ruled where evidence dictates. Part of this revised process is taking more notice of "indicator" codes, for example whether a price represents a recovery from a sale.
(d) There is no automatic re-coding of "non-comparable" replacements to "comparable" where the price differential is small.
(e) A greater focus was placed on quality assuring and editing price quotes relating to centrally collected prices and centrally calculated indices where there is a greater inherent risk of errors impacting on the published indices.

67. The study also led to the following generalised conclusions:

(a) Editing is important because it can affect the measured inflation rate and is a non-trivial issue which needs to be carefully managed. Editing processes should be regularly reviewed, most particularly, to reflect the evolution in price collection methods and in the compilation process and the structure of the index.
(b) Editing efficiency, *defined as the potential impact on the accuracy of the CPI/RPI for a given volume of editing*, will depend on the use of optimally efficient procedures:

(i) There is no substitute for "real-time" editing of prices in the field.
(ii) The price collector is normally based placed to make informed judgements.
(iii) There should be a relatively greater focus on the quality assurance and editing of prices which exert the most influence on the published index, either because of the expenditure weight or the number of price quotes. Account should also be taken of the complexity of the calculation and the corresponding inherent risk of making an error.

68. The research undertaken in this paper was carried out by the CPI research team at ONS. This particular project was lead by Damon Wingfield, to whom I am grateful for the stimulating discussions we had at the time and for his comments on an earlier draft of this paper. The views expressed in this paper are those of the author.

ANNEX

[ENGLISH ONLY]

Location 1552 - Bluewater **CONFIDENTIAL** **Clothing Indicator Backcheck**

Shop	February Last month		March This month
HOBBS LTD, L101 & L102 GUILDHALL (LOWER)	Women's skirt: casual HOBBS/NCO BLACK/PLAIN"18.5 SAIGON SK STRETCH CORD/2 FRONT/2 SLIT BACK POCKS/BELT LOOPS 97 COTTON 3 LYCRA £49.00 Notes: _____	Item 1 2 3 4 5 Price	Women's skirt - casual HOBBS/ITALY BLUE DISTRESS DENIM/"19" CATH" PANELLED A LINE WITH SEWN DOWN FRONT FLAP/NO WAISTBAND 98.5 COTTON 1.5 ELASTANE £59.00 **Comparable** ☐
CECIL GEE U137 GUILD HALL	MEN'S CASUAL S/SLEEVE SHIRT CECIL GEE/NCO LT BLUE/PLAIN KNITTED JERSEY COLLAR/6 BUTT FRONT/TURN UP ON SLEEVE 50 COTTON 50 POLY £0.00 out of stock Notes: _____	Item 1 2 3 4 5 Price	Men's casual s/sleeve shirt with collar ARMANI/ITALY NAVY/POLO SHIRT COLLAR/5 BUTTONS/CHEST POCKET WITH LOGO/WHITE STRIPE ON COLL 100 COTTON £69.00 **New** ☐
CECIL GEE	MEN'S CASUAL SHIRT-LONG SLEEVE ARMANI JEANS/ITALY LIGHT DENIM BLUE/PLAIN COLLAR,ONE CHEST POCKET WITH LOGO/FINE BLUE STITCHING 80 COTTON 20 LINEN £79.00 Notes: _____	Item 1 2 3 4 5 Price	Men's Casual Shirt, long sleeved ARMANI JEANS/ROMANIA WHITE/LIGHT BLUE 1 CM CHECK BUTTON COLLAR/CHEST POCKET WITH LOGO 100 COTTON £85.00 **New** ☐
RIVER ISLAND U32 ROSE GALLERY	MENS SUIT-READY MADE -1 RIVER ISLAND/ROMANIA/6898 MID GREY/LINEY RIBBED FEEL S/B 4 BUTTONS/3 FRONT POCKETS 4 BUTT CUFF/FLAT FRNT BOOT TRS 67 POLY/33 VISCOSE £99.00 Notes: _____	Item 1 2 3 4 5 Price	Mens suit, ready made RIVER ISLAND/ROMANIA/6900 MID GREY/LINEY RIBBED FEEL S/B 4 BUTTONS/3 FRONT POCKETS 4 BUTT CUFF/FLAT FRNT BOOT TRS 67 POLY/33 VISCOSE £99.00 **Comparable** ☐
RIVER ISLAND	MEN'S FORMAL SHIRT-LONG SLEEVE RIVER ISLAND/NCO/E 7107 WHITE WITH BLUE/BLACK LINE CHK CUFF/1BREAST POCKET/WHITE BUTT IN CUBBY HOLE ON WALL 100 COTTON £0.00 out of stock Notes: _____	Item 1 2 3 4 5 Price	Men's Formal Shirt, long sleeved RIVER ISLAND/HONG KONG/E 0017 WHITE WITH THIN GREY 2 CM STRI COLLAR/SINGLE CUFF/BUTTON FRONT/ON HANGER BY SUITS 100 COTTON £24.99 **Comparable** ☐