WP. 5
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ottawa, Canada, 16-18 May 2005)

Topic (i): Editing administrative data and combined data sources

DETECTION OF OUTLIERS IN THE CANADIAN CONSUMER PRICE INDEX

Supporting Paper

Submitted by Statistics Canada[1]

I. INTRODUCTION

1. The Canadian Consumer Price Index (CPI) is a measure of changes in retail prices paid by consumers for goods and services in a predetermined and fixed basket of commodities. For each item in the basket, we calculate the 'price relative', which is the ratio of the price of an item in the current period to the price of the same item in the previous period. The estimated index is a weighted sum of means of price relatives. Central activities in data editing for the CPI are the detection of outliers and the identification of influential observations. In this study, we concentrate on determining appropriate outlier detection methods and measures of influence for the CPI.

2. Outlier detection is performed on the price relatives. In choosing suitable methods, we take into account, among other practical, theoretical and empirical considerations, the skewness of the distribution of the price relatives, the small size of the CPI sample and the existence of 'specials' in many components of the CPI. Indeed, a price for an item in an outlet can be its usual or 'regular' price, or it can be a temporarily reduced or 'special' price. On any month, a price can be returning from 'special' or going into special causing a steep decline or raise in the corresponding price relative. Even though in general these are correctly recorded price relatives, they are the first outlying observations to be detected by the usual automatic editing methods. Thus, detection methods with higher breakdown points are favoured and when possible, price relatives from products going into 'special' or returning from 'special' are grouped into separate editing classes.

3. A comparison of five different methods was conducted on many periods in 2003 and 2004, and several commodities. These were chosen to provide a wide variety of price relative behaviours across commodities and periods, including commodities that were susceptible to change and periods with some type of economic upheaval. For example, the data for meat cuts during the mad cow scare in May-June 2003 were particularly examined. Gas, bread, sugar, men's clothing, and other commodities were also studied. The quartile method with asymmetric fences applied to the log of the price relatives was recommended. Implementation is soon to start in a parallel run with the old detection method (a combination of manual detection and fixed stationary bounds). Measures of influence on the index were also studied to further reduce the number of items sent for verification or imputation.

[1] Prepared by Abdelnasser Saïdi abdelnasser.saidi@statcan.ca and Susana Rubin Bleuer Susana.rubin@statcan.ca .

II. BRIEF DESCRIPTION OF THE APPLICATION

4. The CPI measures price changes by comparing through time, the cost of a predetermined basket of commodities. The basket is assumed to contain commodities of unchanging or equivalent quality and quantity, and thus the index reflects "pure" price movement only (see Statistics Canada catalogue no. 62-557-XPB). There are eight major components in the Canadian CPI, including "Food", "Shelter", "Household operations and furnishings", "Clothing and footwear", "Transportation", "Health and Personal Care", "Recreation, Reading and Education" and "Alcoholic beverages and tobacco products". Within each major component, there are elementary commodity groups, or basic classes. The lowest level of aggregation is the class. There are 169 basic classes.

5. Information on the spending habits of Canadian households is obtained periodically from Survey of Household Spending, and expenditure weights are obtained. Through the use of these expenditure weights, the CPI reflects the spending behaviour of average Canadians living in urban and rural private households. The CPI is a weighted arithmetic average of price indices over all basic classes. The price index at the basic class level I for a certain geographic area is, in turn, a weighted sum of the micro indices I_k over all items k in the class, $k=1,\ldots,K$:

$$I = \sum_{k=1}^{K} I_k \times w_k, \ \sum_k w_k = 1,$$

where w_k are the expenditure weights mentioned above, representing the relative importance of item k in the class. For most classes, the micro indices I_k are geometric means of price changes (also called price relatives):

$$I_k = \prod_{i=1}^{n} (p_{i,current} / p_{i,previous})^{1/n}$$

where $p_{i,current}$ is the price of a specific product of item k in outlet i in the current period. For example, if the product refers to white bread of certain characteristics and current period refers to the month of May, then $p_{i,current}$ is the price in May of a loaf of white bread appropriately standardized, in outlet i. The price $p_{i,previous}$ is that of the same product of item k in outlet i in the previous period, e.g., the price of the same type of bread in outlet i in April. The number n corresponds to the number of products priced, which in our case coincides with the number of outlets surveyed for this micro-index.

6. The current evaluation procedures regarding prices and related data can be considerably improved. A large amount of the resources are used to correct a small number of errors. Actually, verifications and corrections are made manually. This does not improve accuracy: verification mistakes can occur due to the heavy workload that evaluators have. The goal of this project is to determine outlier detection methods for the Canadian CPI that are efficient in terms of resources and accuracy. We first describe briefly some of the current evaluation procedures in the production of the CPI: specifically, those procedures that have relevance in the detection of outliers. For other editing procedures having to do with missing values, missing outlets and quality changes, see Allard Saulnier and Beyrouti. (2003).

7. For a great majority of the products, data are collected at selected outlets by interviewers, by means of a portable computer. The data are transmitted to headquarters at Statistics Canada and prices are automatically standardized (to the standard unit of measurement) and taxes deducted (if they were previously included). The system selects all records that have a price with status code = 'special', and all records with price relatives (i.e., $p_{i,current} / p_{i,previous}$) that have decreased or increased by 15%. These records are then evaluated manually. We first note that many records with status code = 'special' had their previous price also coded as 'special'. They usually carry the same price in both the previous and current period, yielding a price relative equal to 1. Thus, many records that are not outlying observations are manually evaluated with this system. On the other hand, those observations with status code =

'regular', whose previous price had status code = 'special' are not selected by the current rules to be checked by the manual evaluators if their price relatives have not increased by 15%, so many possible outlying observations are not inspected.

8. The data available to us for the study are:

- the record and outlet identification;
- the current month;
- $p_{i,\,previous}$;
- $p_{i,current}$ after appropriate standardization;
- the status code, taking values "regular", "special" or "verified". The value "verified" corresponds to a price that was suspect at one time before the creation of this file and that was either verified to be correct or changed. If the price was a special to start with, this characteristic cannot be traced. Thus, this variable is not reliable in the determination of the type of price. Hence, we use this variable to define the editing groups but we do not try to edit it. The observations with current status code = previous status code = 'regular' or current status code = previous status code = 'special' will be called RP's. The observations with current status code = 'regular' and previous status code = 'special' or current status code = 'special' and previous status code = 'regular' will be called SP's.

9. The previous prices are supposed to have been already verified and used for the calculation of the previous month index, and hence our aim is to detect outliers on the current prices. Since we have only one quantity of interest that we can verify, the detection method is univariate. We choose to do the outlier detection on the price relatives, since the parameter of interest is an average (whether geometric or arithmetic) of price relatives. Thus, we have built-in an automatic 'historical edit'.

10. As we said in the introduction, we must take into account the existence of specials in many components of the CPI and their effect on the index. The price relatives corresponding to SP observations are respectively higher or lower than for the rest of the items and they are very different from the underlying price change distribution. This means that SP's are, in general, outliers. Since there is no reliable code to separate the SP's outright, detection methods with higher breakdown points are favoured (e.g. Median Absolute Deviation (MAD) and quartile method variants as opposed to Tukey's methods). However, price relatives from SP's are grouped into separate editing classes when it is possible to identify them and when there are enough of them to edit them in a separate group.

11. The distribution of the price relatives is very skewed for many items in the CPI basket. In that sense, methods that take into account the skewness of the data, whether by means of a transformation that would make it more symmetric (like the log transformation or the Hidiroglou-Berthelot transformation (see Hidiroglou and Berthelot 1986)) or by the use of asymmetric bounds (asymmetric fences), are more efficient than those appropriate for Gaussian and symmetric distributions.

12. The editing group is the smallest set of observations where the outlier detection is performed. In the CPI, it consists in all the observations pertaining to a particular item (or a set of items which define a basic class) in a particular geographical area. There are four possible levels of geographical areas: strata within provinces (17 categories), provinces and territories (13 categories), regions (5 categories) and Canada (1 category). Depending on the item, we define the geographical level for the editing group as the smallest level that contains at least 15 regular observations, though this parameter can be changed by the user. The choice between a particular item and a collection of items within a class to define the editing group is not easy. We compared editing groups consisting in a number of items together versus separate items for the same geographic level on bread, men's shirts, gas and beef, and results were inconclusive as to which group the split was better.

III. DESCRIPTION OF THE OUTLIER PROBLEM

13. Outliers in the CPI arise from:

- Procedural errors, data entry errors, or mistakes in coding.
- Correctly recorded observations that are SP's or a result of unusual economic events:

i) SP's are outliers that are mostly correct price relatives, but they are the first outlying observations to be detected by usual editing methods. Under the current system, we cannot obtain the exact proportion of SP observations in the sample, but using the "status code", we could calculate a probable overestimate. But we can confidently say that for many items and periods of time, the total count of SP's is at least 15% of the total number of observations. Even if the CPI sample is selected in a representative way, the proportion of SP in the sample will not be the same as that in the population. This happens because the sample size is very small and extreme samples do occur frequently.

(ii) Extraordinary observations are due to once-in-a-time phenomena. For example, after the first mad cow scare, in May 2003, beef prices were down at a time that they are usually high. These observations could be either representative or not of other units in the population. Also, they could be influential or not.

IV. TECHNIQUES STUDIED FOR DETECTING OUTLIERS

Tukey's algorithm and it's variants

14. We included this method because it is one of the methods recommended in the International Labour Office (ILO) manual, though the breakdown point is very low and it is more appropriate for symmetric distributions.

The tolerance interval is given by $[mid_m - 2.5 * \Delta\ell_m \ ; \ mid_m + 2.5 * \Delta u_m]$.

OUT_TUK is a variant where the mid-mean mid_m is the 10% trimmed mean of the price relatives, the lower mid-mean ℓ_m is the mid-mean of the price relatives smaller than the mid-mean (that is, ℓ_m is the 10% trimmed mean of the price relatives smaller than the mid-mean mid_m), then $\Delta\ell_m = mid_m - \ell_m$; the upper mid-mean u_m is the mid-mean of the price relatives larger than mid_m, (that is, u_m is the 10% trimmed mean of the price relatives larger than the mid-mean mid_m) and $\Delta u_m = u_m - mid_m$. The breakdown point is lower than that of the mid-mean, which is 5%.

OUT_TUK_CPI is the same algorithm as OUT_TUK, except that it is applied to the distribution of the price relatives after the lowest 5%, the highest 5% price relatives are discarded and all price relatives equal to 1. This is the variant described in the ILO manual. Even though this method has a breakdown point under 9.5%, it is a bit extreme, since it will automatically reject 10% of the observations even if there are no outliers in the group.

OUT_TUK_1 is the same algorithm as OUT_TUK, but replacing the lower and upper mid-means (ℓ_m and u_m) respectively by the mean of the price relatives smaller than the mid-mean mid_m and the mean of the price relatives larger than mid_m.

Quartile method or Asymmetric Resistant Fence method (QM)

15. The price relatives are first transformed so that their distribution is more symmetric. We usually use the log transformation. The tolerance interval is given by $[q_{0.50} - c * d_L \ ; \ q_{0.50} + c * d_U]$ where $q_{0.50}$ is the median of the transformed price relatives; $q_{0.25}$ and $q_{0.75}$ are respectively the first and third quartiles; and the lower (d_L) and the upper (d_U) quartiles ranges are defined as $d_L = q_{0.50} - q_{0.25}$ and

$d_U = q_{0.75} - q_{0.50}$. Here, we used c = 4, but this constant may have to be adjusted for some classes of items not yet studied. In general d_L is different from d_U, which reflects some left over skewness in the data, after the transformation. Often, it happens that either d_L or d_U is too small. In that situation, to avoid that points with very small deviations from the median be detected as outliers, d_L and d_U are modified by $d_L = \max\left(q_{0.50} - q_{0.25}, |A|\right)$ and $d_U = \max\left(q_{0.75} - q_{0.50}, |A|\right)$ where A = 0.03 seems adequate in most items studied.
This method has a breakdown point of 25%. Using simulated data modeled on collected economic data, Thompson (1998) recommends the use of the quartile method for small samples (<50), without transforming the data to account for the skewness. When the sample size is reasonable (probably greater than 50), she recommends to examine the effect of the potential symmetrizing power transformations on the data (using the log or the square root functions).

Resistant Fences Method (RFM)

16. The tolerance interval is given by $\left[q_{0.25} - k*(q_{0.75} - q_{0.25})\ ;\ q_{0.75} + k*(q_{0.75} - q_{0.25})\right]$. Thompson (1998) notices that this variant of the quartile method gives good results when the distribution is symmetric. The value used for this study is k = 4.

MAD

17. Once the same transformation as for the quartile method is applied to the prices relatives, the tolerance interval is given by $\left[q_{0.50} - c*MAD\ ;\ q_{0.50} + c*MAD\right]$ where $q_{0.50}$ is the median of the transformed price relatives y_i, $MAD = median_i\{|\ y_i - q_{0.50}\ |\}$ and $c = 2.575$ is an appropriate constant. MAD has the highest breakdown point of the methods described here, and it is often too sensitive, rejecting many more observations we can afford to verify.

Hidiroglou-Berthelot variant (HB)

18. The Hidiroglou-Berthelot (HB) method is basically the same as the quartile method (with the same breakdown point), but applied to a different transformation of the data. The method transforms the price relatives R_{it} to reduce skewness in their distribution as follows:

$$S_{it} = \begin{cases} 1 - \dfrac{q_{0.50}}{R_{it}} & \text{if} \quad 0 < R_{it} < q_{0.50} \\ \dfrac{R_{it}}{q_{0.50}} - 1 & \text{if} \quad R_{it} \geq q_{0.50} \end{cases}$$

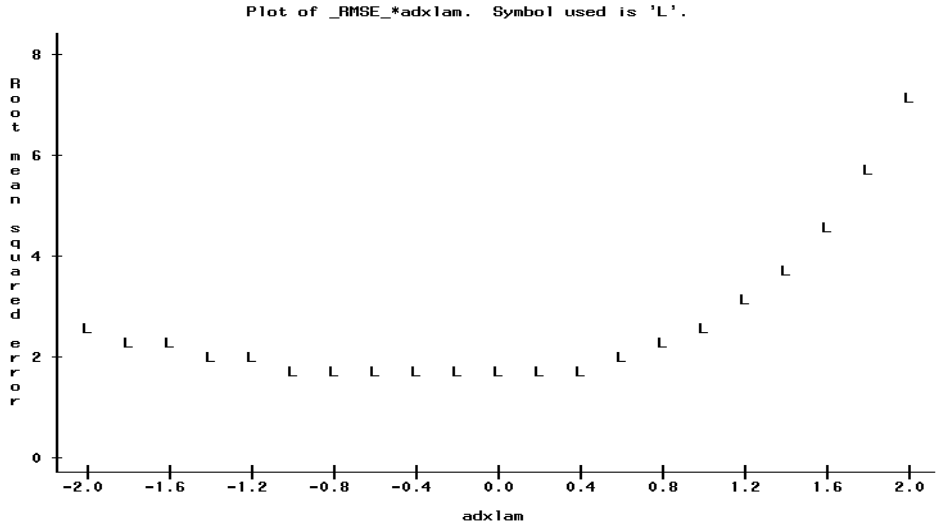
The quartile method is applied to S_{it} to detect outliers. In the classical application of the HB method, there is a second transformation applied to the data. Its aim is to give higher influence to the ratios belonging to businesses with higher revenues. This does not apply to price relatives in the calculation of micro-indices. The expenditure weights take care of the economic importance of an item in later aggregations of the index.

V. AN APPLICATION TO THE SURVEY OF CONSUMER PRICE INDEX

19. The goal is to compare the results of the different methods, and to define the edit groups and to adjust the different constants involved in the implementation of the methods. Finally, we compare our results with the corrections made under the current evaluation system. We cannot compare the outliers detected under our system with those detected under the current system, but we can verify if the values corrected under the current system were at least detected by the automatic procedure recommended.

20. We compared five different methods (QM, RFM, MAD, HB and the Tukey's variants) in many periods of 2003 and 2004, and on several commodities. We looked at gas, bread, sugar, men clothing, etc. We present here, as an example, the analysis done on the item hand towels. The log transformation was preferred to the square root function using a test based on the Cramer von Mises statistic. We confirm this result with the Box Cox method using the macro ADXTRANS of SAS®. The following plot shows that the root mean square error is minimum for the value $l = 0$; this justifies using the log transformation. We further ensured that the skewness coefficient of the transformed price relatives is lower than the one of original price relatives.

Plot of RMSE against l (adxlam)



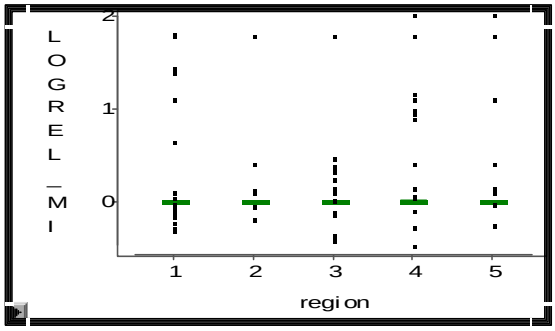
21. When defining an editing group, the geographic level is determined as the lowest level that includes at least 15 RP observations. For the item 'hand towels', the geographic level is the region. The variable GOODSTAT is an indicator of whether the observation is an SP or an RP. We also form two different editing groups for RP's and SP's if the proportion of SP's is greater than 15% of the observations. This criteria is based on the assumption that we will not have more than 10% of outlying observations that are not SP's. The criteria can be changed by the user. In the 'hand towels' example, each region carries a proportion of SP's less than 15%, hence we do not separate them.

Table 1 - *Frequencies by geographic zone and status*

GOODSTAT	region	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	**1**	93	20.09	93	20.09
0	**2**	63	13.61	156	33.69
0	**3**	93	20.09	249	53.78
0	**4**	99	21.38	348	75.16
0	**5**	65	14.04	413	89.20
1	**1**	13	2.81	426	92.01
1	**2**	7	1.51	433	93.52
1	**3**	13	2.81	446	96.33
1	**4**	9	1.94	455	98.27
1	**5**	8	1.73	463	100.00

22. The box plot of prices shows the log price relatives distribution by region. In addition, there is a lot of points detected by the RFM[2] method (which is the classical Tukey's method described in SAS®) where the median, first and third quartiles are identical.

Table 2 – *Box plot of the log of price relatives (LOGREL_MI) by region*



23. The results obtained with the QM, MAD, RFM and HB are very similar. The majority of outliers detected by these four methods are the same. The proportion of outliers (14%) is bigger than the breakdown point of the Tukey's variants.

Table 3 - *Frequencies of outliers for QM, MAD, RFM, and HB*

outlier_QM	outlier _MAD	outlier _RFM	outlier _HB	Frequency
0	**0**	**0**	**0**	354
0	**0**	**1**	**0**	2
0	**1**	**1**	**0**	1
0	**1**	**1**	**0**	9
0	**1**	**1**	**1**	5
1	**1**	**1**	**1**	52

Table 4 - *Frequencies of outliers for the Tukey's variants*

Method	OUT_TUK	OUT_TUK1	OUT_TUK_CPI	OUT_TUK_W
Frequency	**7**	**0**	**40+11**	**0**

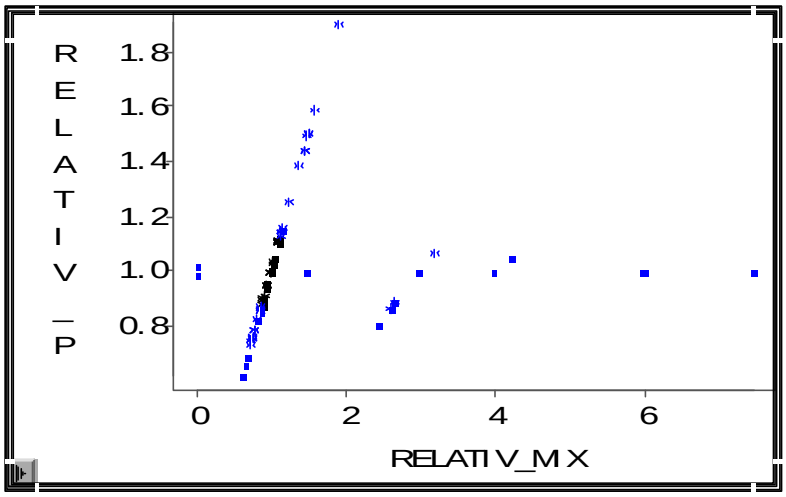
Table 4 shows that Tukey's variants are not appropriate here. The OUT_TUK_CPI yielded as outliers the 40 (10% of total observations) automatically selected whether their values are extreme or not, plus 11 outliers more. Even in the case of gas and other items where there are few or no specials, OUT_TUK_CPI didn't yield to a reasonable set of outliers. The other two Tukey's variants have detected few outliers compared with the QM, RFM, HB and MAD methods.

24. Based on the results obtained for several items and periods, we decided to recommend the QM with the logarithm transformation, since it is the simplest of the methods that yield similar good results.

25. The following plot shows that all the points not on the first bisector (or on the same line) were corrected. The blue points are the outliers detected by the QM, the stars are the SP's, and the squares are the price relatives that are not SP. We notice that all the corrected points had been detected as outliers by the recommended method.

[2] In our RFM application we added the protection of A=0.03 for the case where median and quartile are too close.

Table 5 – *Price relatives after (RELATIV_P) and before correction (RELATIV_MIX) by the current procedure*



VI. INFLUENCE OF OUTLIERS DETECTED BY QUARTILE METHOD

26. The detection of outliers on the items of the basic class « gas » yielded 14 outliers among 1662 observations. The influence analysis consists in evaluating the contribution of each outlier to the basic class index. The following table shows the percentage of contribution. The question is how to define a cut-off to decide if an outlier is influent or not.

Table 6 – *Percentage contribution of the outlier on the basic class index*

ID outlier	CPI in the basic class*	CPI without the outlier**	difference in %	Direction
1	0.974712	0.975571	0.088192	-
2	0.97931	0.979023	0.029261	+
3	0.980203	0.980995	0.080797	-
4	0.980203	0.981065	0.087940	-
5	0.99359	0.993202	0.039037	+
6	0.980203	0.981277	0.109513	-
7	0.980203	0.980476	0.027850	-
8	0.980203	0.981107	0.092157	-
9	0.99359	0.993194	0.039784	+
10	0.985128	0.984380	0.075923	+
11	0.985128	0.984874	0.025738	+
12	0.967881	0.967853	0.002836	+
13	0.985128	0.984853	0.027903	+
14	0.985128	0.985018	0.011162	+

27. The largest deviation in Table 6 is for the sixth outlier. The basic index changes from 0.981 to 0.980 yields a small difference of 0.109%. It is interesting to note that no point was corrected by the evaluators.

VII. CONCLUSION

28. When comparing the methods QM, MAD, RFM and HB, based on the results obtained for several items and periods, we found that both the quartile and the Hidiroglou-Berthelot methods yield

similar results, and meet our expectations in the sense that they detect a reasonable number of outliers that we can verify: MAD is too sensitive and even after the application of the transformation logarithm, the data remains somewhat asymmetric and quartile method performed better than the RFM. Since the quartile method with the logarithm transformation is the simplest to apply and to explain visually, this is the method we recommend to be used with the Canadian CPI.

References

Cotton, C.M .(1982): "Outlier detection in the revised ISPI", *Statistics Canada*.

Berthelot J.M. (1983): "Wholesale-Retail Redesign, statistical edit proposal", *Statistics Canada.*

Hidiroglou,M.A. and Berthelot, J.M. (1986) "Statistical Editing and Imputation for periodic business surveys", *Survey Methodology, June 1986, Vol 12, N1, pp73-83.*

Lee H., P.D Ghangurde, L. Mach and W. Yung (1992): "Outliers in sample surveys", *Statistics Canada.*

Kovar J.G., MacMillan J.H. and P.Whitridge (1991): "Overview and strategy for the generalized edit and imputation system", *Statistics Canada.*

Rubin-Bleuer S. (2001): "Consumer Price Index traveller accommodation component", *internal document BSMD Statistics Canada.*

Thompson .K.J. (1998): "Ratio edit tolerance development using variations of exploratory data analysis (EDA) resistant fences methods", *Economic Statistical Methods and Programming Division, United States Bureau of the Census.*

Allard Saulnier M., and Beyrouti M. (2003) " Revue des procédures d'évaluation", *internal document. Prices Division, Statistics Canada.*

SAS (1999) " Procedures Guide version 8".
