

OUTLIER DETECTION FOR THE CONSUMER PRICE INDEX

Saad Rais¹

ABSTRACT

The processing system for the Consumer Price Index (CPI) is undergoing a redesign. Therefore, the current methods used by the system for methodological tasks such as outlier detection are being reviewed to determine if changes can be made to enhance the efficiency of the system and the quality of the estimates. In this paper, we study different non-parametric outlier detection methods that could be implemented in the redesigned CPI system.

KEY WORDS: Outlier Detection, CPI, Quartile Method, Hidiriglou-Berthelot Method, Resistant Fences Method, Tukey Algorithm

RÉSUMÉ

Le système de traitement de l'Indice des prix à la consommation (IPC) de Statistique Canada fait à l'heure actuelle l'objet d'un remaniement. Les méthodes utilisées par le système actuel telles que pour la détection des valeurs aberrantes sont évaluées pour déterminer si on peut les changer pour améliorer l'efficacité ou la qualité des estimations. Dans cet article, on compare quelques méthodes non paramétriques pour la détection des valeurs aberrantes qui pourraient être utilisées dans le système remanié.

MOTS CLÉS : Algorithm de Tukey; l'Indice des prix à la consommation (IPC); détection des valeurs abérrantes; méthode des clôtures résistantes; mthode des quartiles; méthode Hidiroglou-Berthelot.

1. INTRODUCTION

The processing system for the Consumer Price Index (CPI) is undergoing a redesign. Therefore, the current methods used by the system for methodological tasks such as outlier detection are being reviewed to determine if changes can be made to enhance the efficiency of the system and the quality of the estimates. In this paper, we study different non-parametric outlier detection methods that could be implemented in the redesigned CPI system.

Outliers are observations that appear to be inconsistent with the rest of the data set. They may appear in the data either due to some error in collection or reporting, due to inherent variability in the data resulting from forces such as seasonality or item substitution, or due to some observations following a distribution different from the majority of observations. The degree of distortion in an estimate due to the presence of outliers will depend upon the level of inconsistency and weight of the outlying units. One procedure to address the presence of outliers is to first employ some method to detect the outliers, and then to perform some outlier treatment to rectify the impact of the detected outliers upon the estimates.

A good outlier detection method will identify all the 'true' outliers in the data, and avoid the occurrence of false positives – falsely declaring an observation to be an outlier. Too many false positives may result in resources being put into manually verifying flagged outliers, while missing the true outliers may yield an unreliable index. In the context of monthly and quarterly surveys, such as a CPI survey, where there are operational constraints due to short production time, an ideal outlier detection method will also be automated, fast, versatile, simple to execute and understand, yet robust and based on accepted principles.

Outlier detection in the context of the CPI is an editing tool for identifying errors in price relative data so that they can be corrected by reviewing officers. Currently, the outlier detection method employed by CPI processing system involves the use of tolerance intervals whoses upper and lower bounds are defined by subject matter experts using their experience and

¹ Saad Rais, Industrial Organization, Finance, and Prices Section, BSMD

knowledge of the industry. Units with values falling in the interval are accepted as legitimate values, whereas the remaining units are identified as potential outliers that undergo review. The bounds for these intervals are specific to each item or a set of related items, and typically remain static over long periods of time.

Methods that are frequently seen in the literature and used by different statistical agencies for price indices include the *Quartile Method (QM)*, the *Hidiroglou-Berthelot (HB) method*, the *Resistant Fences (RF) method*, and the *Tukey algorithm (TA)*. In this paper, we review and study these four methods to determine the most appropriate method to implement into the redesigned CPI system. In section 1, the four methods are introduced, and their properties are compared. Section 2 describes and presents the results of an empirical study comparing the four methods. Section 3 concludes the paper with a summary, a note on some of the study's limitations, and a recommendation.

1. The Four Outlier Detection Methods

There appears to be no consensus among various national statistical agencies on a preferred outlier detection method. The United Kingdom's Office of National Statistics (ONS) has been using the TA to detect outliers in their CPI since 1987. Thompson et al. (1999) from United States Bureau of Bureau of the Census (USBC) favors the RF method for ratio data. At Statistics Canada (StatCan), Saïdi et al. (2005) shows that the QM performs at least as well as the other methods. Rais (2007) also prefers the QM, but also found the HB method to be acceptable. The IMF (ILO) Consumer Price Index Manual (2004) mentions the QM, the TA, and the HB method but does not favor one method over another.

These varied opinions reflect the reality that non-parametric outlier detection methodology requires an element of subjectivity and therefore is not an exact science. Additionally, the methods are very similar in some respects and therefore the difference in performance may be negligible under certain conditions. They all involve the construction of a tolerance interval that defines the range of acceptable observation values. These tolerance intervals are based on formulas that are a function of the data that they will be applied to, as well a function of user-defined parameters. Except for the TA, all of the methods require that the data be transformed. The transformation ensures that the transformed data follow an approximately symmetric distribution.

Below, we describe each of the methods and compare their properties when applied to price relative data. A price relative is a ratio of prices for item k comparing the current month t with the previous month $t-1$, expressed as:

$$P_k^{(t/t-1)} = P_k^t / P_k^{t-1}.$$

1.1 The Quartile Method (QM) and the Hidiroglou-Berthelot (HB) Method

The procedure for constructing the bounds for the QM tolerance interval is the following: Let $q_{0.25}$, $q_{0.5}$, and $q_{0.75}$ represent the first quartile, median, and third quartile respectively, for the set of observations s . Also, let l_s and u_s represent the lower and upper quartile ranges respectively, where $l_s = q_{0.5} - q_{0.25}$, and $u_s = q_{0.75} - q_{0.5}$. Then, the tolerance interval is defined by the following expression:

$$(q_{0.5} - c_l l_s, q_{0.5} + c_u u_s)$$

where c_l and c_u are some predetermined constants that are often held equal i.e., $c_l = c_u = c$.

In some instances, the data may be clustered closely together, resulting in a tight tolerance interval that will flag observations that are slightly deviating from the median as outliers. The following modification of the quartile ranges helps account for this possible occurrence by imposing a minimum tolerance interval spread:

$$l_s = \max(q_{0.5} - q_{0.25}, |aq_{0.5}|)$$

$$u_s = \max(q_{0.75} - q_{0.5}, |aq_{0.5}|)$$

where $0 \leq a \leq 1$. Lee et al. (1992) remarks that a reasonable value for a is 0.05, for most applications.

When $c_l = c_u = c$, the QM assumes that the data follow a symmetric distribution. With right-skewed data such as price relative data, the QM may create a *masking effect* where the tolerance interval is more sensitive to the right tail of the distribution and less sensitive to the left tail. To address this occurrence, the data can undergo some symmetrizing transformation. Saïdi et al. (2005) and Thompson et al. (1999) both prefer the natural logarithmic transformation to symmetrize right-skewed data.

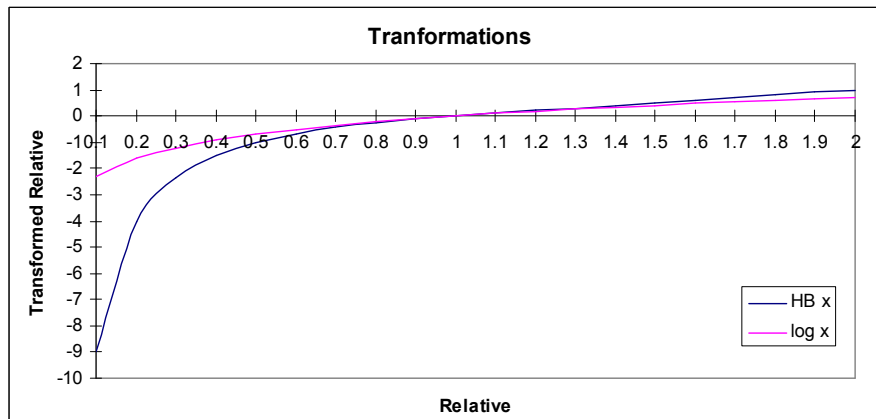
While the log function may be applied to both whole numbers and ratios, Hidiroglou et al. (1986) proposed an alternative transformation function that is specific to observations of ratios. The function is expressed as the following:

$$s_k = \begin{cases} 1 - \frac{q_{0.5}}{r_k} & \text{if } 0 < r_k < q_{0.5} \\ \frac{r_k}{q_{0.5}} - 1 & \text{if } r_k \geq q_{0.5} \end{cases}$$

where $r_k = y_k^t / y_k^{t-1}$ is a ratio of y -values from two consecutive time periods t and $t - 1$ for the k^{th} unit. For the purposes of our discussion, we refer to the HB method as the QM applied to this transformation².

If we assume that $q_{0.5} = 1$ (which is almost always the case with price relative data), and $r_k \approx 1$ for all k , it can be shown that the HB transform is equivalent to the first order Taylor expansion of the log function, i.e., $s_k \approx \log(r_k)$. Therefore, with stable price relative data centered around 1, the two transformations should yield very similar results. However, if the data set contains extreme values, the symmetrizing power of the HB transform will be less effective and the masking effect will prevail. The log transform, on the other hand, performs well even with acutely right-skewed distributions.

The graph below illustrates the behavior of the two transformations at different values of relatives:



The graph confirms that the two transformations lead to almost identical values when the true value is centered around 1 (between 0.7 and 1.4), and for more extreme relatives, the HB transformation leads to larger transformed values.

1.2 The Resistant Fences (RF) Method

This RF method is fairly new to price index data, and therefore is not mentioned in the IMF CPI manual (2004). It is not known if the RF method is used by any statistical agency for any price index survey, though it has been recommended by Thompson et al. (1999) from the USBC. The method resembles the QM but uses the interquartile range in the tolerance interval formula. Let $\Delta q_s = q_{0.75} - q_{0.25}$ represent the interquartile range, where $q_{0.25}$ and $q_{0.75}$ represent the first quartile and third quartile respectively, for the set of observations s . Then, the RF tolerance interval is defined by the following expression:

² The HB method also involves a second ‘weighting’ transformation, causing the sensitivity of the method to be proportional to the size of the unit. However, this transformation is irrelevant for the CPI since the index is a function of unweighted relatives.

$$(q_{0.25} - c_l \Delta q_s, q_{0.75} + c_u \Delta q_s)$$

where c_l and c_u are some predetermined constants that are often made equal i.e., $c_l = c_u = c$.

As with the QM, the RF method requires that the distribution of the data be symmetric when $c_l = c_u = c$. Thompson et al. (1999) recommends using the log transform with the RF method.

It can be shown that the RF method is essentially an extension of the QM method. If we let QM_l and QM_u represent the QM lower and upper bounds respectively, the RF tolerance interval formula can be rewritten as:

$$(QM_l - H_l, QM_u + H_u)$$

Where $H_l = c(q_{0.75} - q_{0.5}) + (q_{0.5} - q_{0.25})$ and $H_u = c(q_{0.5} - q_{0.25}) + (q_{0.75} - q_{0.5})$. Since $H_l \geq 0$ and $H_u \geq 0$ (when $c \geq 0$), the RF tolerance interval will always be at least as wide or wider than the QM tolerance interval for the same c-value. Hence, at the same c-value, the RF method will be less sensitive than the QM and will therefore flag fewer units as outliers.

1.3 The Tukey Algorithm

There are a few variants of the Tukey Algorithm that are mentioned in the literature, but we only present the version described in the ONS CPI Technical Manual (2006):

1. Sort the price relatives in ascending order.
2. Remove price relatives of 1.
3. Remove the top and bottom 5% of price relatives. Let d represent the set of remaining units.
4. Let \bar{x}_d represent the arithmetic mean of the set d , and let \bar{x}_l and \bar{x}_u respectively represent the arithmetic means of observations above and below the median for the set d . Then, let $\Delta\bar{x}_l = \bar{x}_d - \bar{x}_l$ and $\Delta\bar{x}_u = \bar{x}_u - \bar{x}_d$.

The tolerance interval according to this variant of the Tukey Algorithm is:

$$(\bar{x}_d - c_l \Delta\bar{x}_l, \bar{x}_d + c_u \Delta\bar{x}_u)$$

where c_l and c_u are some predetermined constants that are usually made to be equal.

The TA is a unique method in that it removes the price relatives of 1 during the calculation of the tolerance interval, which effectively removes the stability in the data so that the tolerance interval is not tightly bounded to the mean. Therefore, it is not required to establish a minimum tolerance interval spread, as in the case with the QM. However, such an interval may not be representative of the entire data set, since it is calculated using a subset of the data i.e. all non-one relatives. Thus, when the interval is applied to the complete data set, the outlier detection may not provide accurate results, especially when the sample size is small. Since some of the stratum sample sizes in CPI data *are* small, the TA may not be appropriate to employ with CPI data.

2. THE 'HIT-RATE' STUDY

An ideal study for comparing outlier detection methods would require a data set in which all the true outliers are known in advance. The different methods would be applied to this data set to determine which method best identifies the true outliers, i.e., achieves the most 'hits', and minimizes the flagging of legitimate values as outliers (i.e., the false positives). We undertook such a study, assuming that the outliers identified by the Prices CPI system were the true outliers.

Under each method, for each item, tolerance intervals were computed by province at each time period. Using these tolerance intervals, observations were either flagged as outliers or identified as legitimate values. The counts of outliers were then used to compute two measures: the hit-rate - the proportion of true outliers detected by a given method, and the

ratio of outliers detected over the true outliers - a measure of the number of false positives produced by the outlier detection method.

Each method requires the user to specify a value for the parameter c that is present in the tolerance formula. The c -value used will affect the performance of each method; large c -values will result in wide tolerance intervals, which could reduce the hit-rate but also reduce the number of false positives, while small c -values will increase the sensitivity of the method, thereby potentially increasing the hit-rate but also increasing the number of false positives. Hence, we tested each method under different c -values. In this manner, the study not only served to help determine the method that is the most appropriate for our CPI data, but it also helped identify the c -value that allowed each method to perform well.

2.1 Data Used and Set-up

Our data set consisted of monthly CPI micro-data for 1023 items spanning the months July 2007 to January 2008, in the 13 provinces and territories³. The data file contained prices for each product both before and after outlier detection and treatment by the CPI system, which enabled us to determine the observations that the system flagged as outliers. In order to ensure comparability of prices in our study, we limited our population of prices to regular prices only, therefore excluding specials. The prices used were pre-adjusted for quality and quantity changes. These prices were then used to calculate price relatives for each item. Outlier detection was then performed on these price relatives.

For the purposes of a preliminary investigation, our study was limited to data for eight specific items⁴ that were representative in terms of the CPI basket share, and in terms of varying degrees of variability (high/medium/low). To classify items into these variability levels, the items were arranged in descending order in terms of their variance⁵ (at the national level, across time) and then split into three equally sized groups representing the high/medium/low categories.

The eight items are the following (by variability group):

Variability	Items
Low	Carton Cigarettes, Dinner Meal
Medium	Women's Briefs, Paint, Men's Dress Socks
High	Regular Unleaded Gasoline With Service, Deodorants or Antiperspirants, Beef Rib Roast

2.2. Results

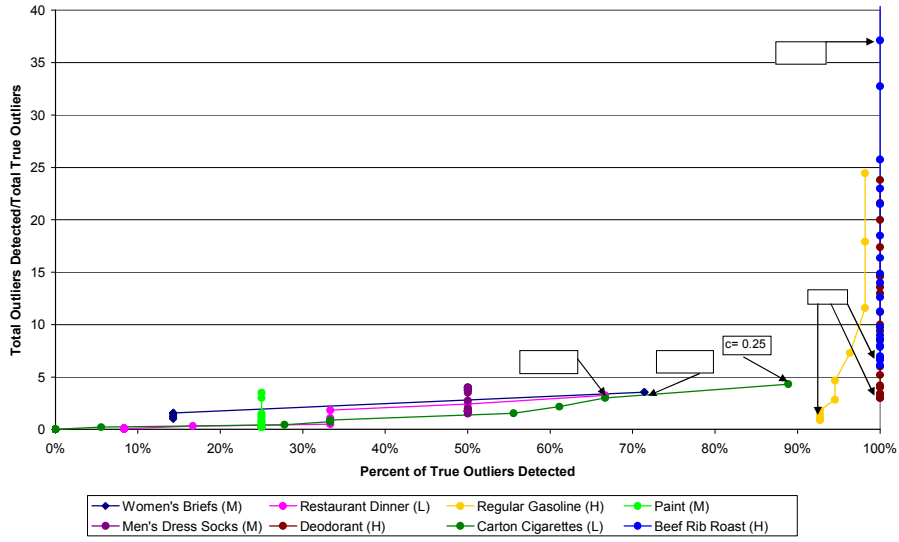
A graph for each outlier detection method is presented below. Each line within a graph represents an item. Within each graph, for each item, the hit-rate (x-axis) is plotted against the ratio of detected outliers over true outliers (y-axis). The c -values for some significant points have been highlighted. The purpose of these graphs is not only to compare the performance of the different methods, but also to determine the number of extra false outliers that would require manual verification to achieve a certain hit rate. The graphs also serve to highlight the differences in results between the three variability groups. For the sake of presentability, the results presented are an aggregation of the results from all the time periods at the national level.

³ A description of the fields for the micro-data can be found in the [IDIS dictionary](#).

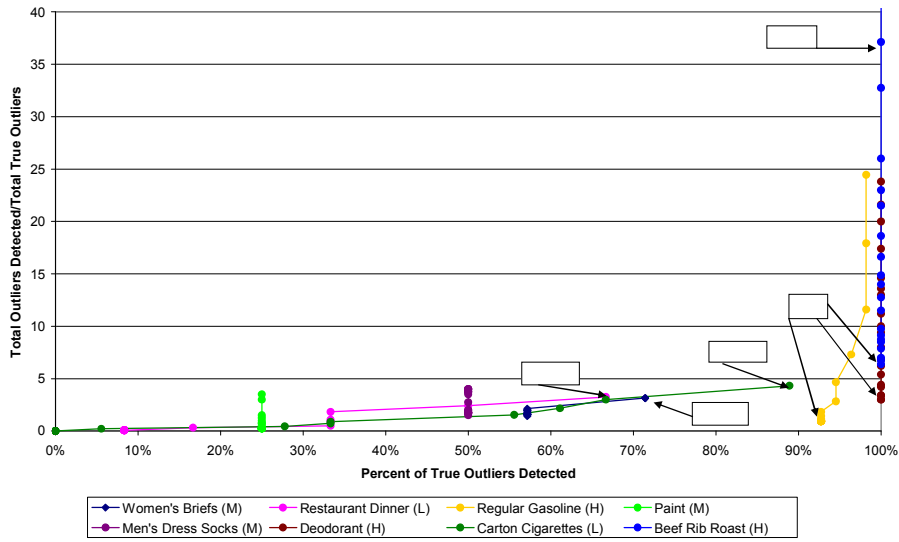
⁴ These items were reviewed by the subject matter team in Prices Division, Statistics Canada.

⁵ In the future, we plan on ranking items in terms of variability using the interquartile range, which would be a more robust measure than the variance.

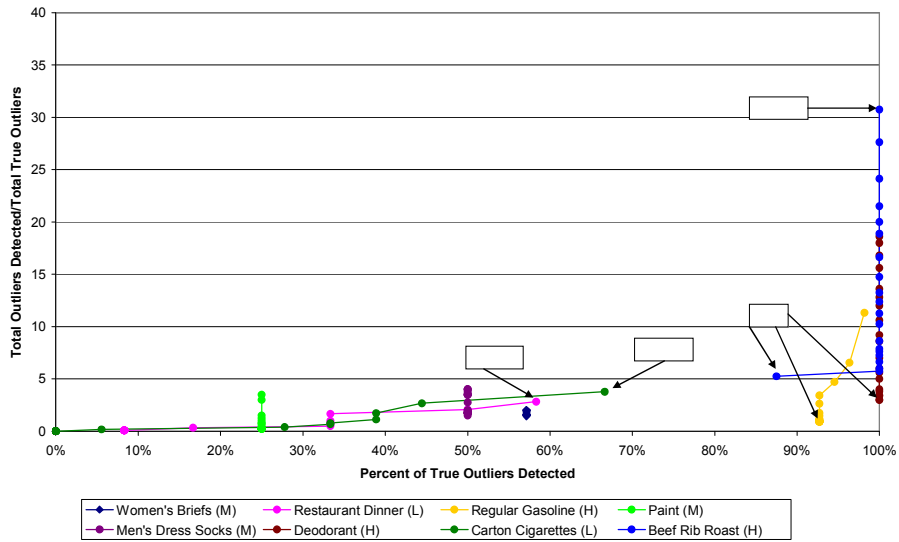
Quartile Method with Log Transformation



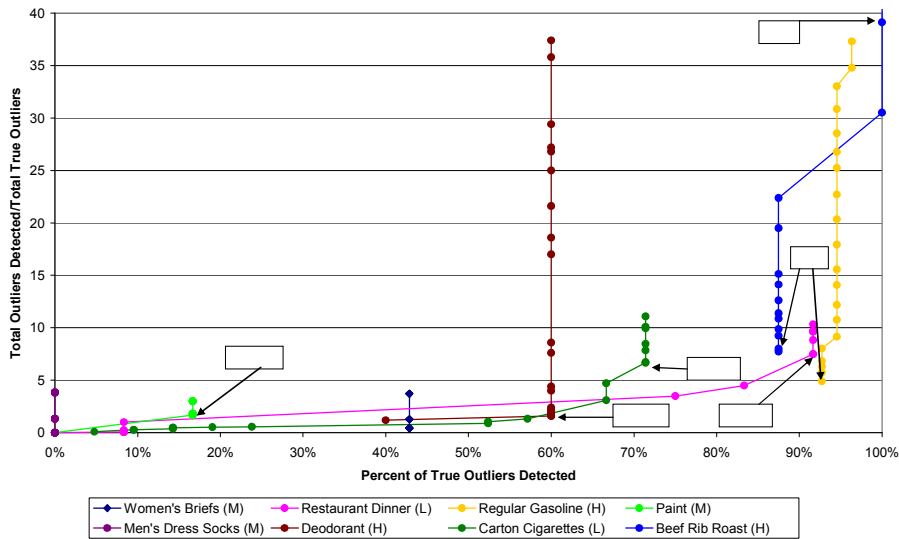
Quartile Method with HB Transformation



Resistant Fences with Log Transformation



Tukey Algorithm



A comparison of the graphs shows that the QM applied to both the log transform and the HB transform produce almost identical results, which confirms our knowledge of the similarity between the two transformations when the range of relative values is not too wide and is centered around 1. The RF method displays a graph similar to the QM and HB graphs, but the RF hit-rates for some of the items (across all variability groups) are not as strong as the hit-rates resulting from the QM and HB methods, at the same c-value. This observation is consistent with the fact that the RF tolerance interval is wider than the QM counterpart, and therefore is less sensitive to the presence of outliers. However, it is possible that the RF method could achieve the same results as the QM with a smaller c-value.

The TA graph shows results that are inconsistent with the results of the other graphs, which is indicative of the dissimilarity of the TA with the other methods, specifically in the manner that the TA removes relatives of 1 from the tolerance interval computation. Some examples illustrate the reduced performance of the TA compared to the other methods. With the 'deodorant' item, the TA produces a 60% hit-rate, while the other methods all achieve a 100% hit-rate. Also, with the 'socks' item, the TA scores a 0% hit-rate, while the other methods all achieve a 50% hit-rate. In summary, the empirical results from our study, coupled with the potentially unfavorable properties of the TA, suggest that the TA should not be recommended for use with our CPI data, though this method could perform reasonably in other surveys with large sample sizes.

There are some trends that are common to all of the graphs. The graphs for all of the methods illustrate that good hit-rates often require a significant increase in the number units flagged for manual outlier detection. A number of reasons were found to explain this occurrence, which are highlighted below:

- The assumption that the outliers detected by the CPI system are the ‘true’ outliers and the only true outliers in the data did not always hold. We found that some stable relatives (~ 1) were flagged as outliers by the CPI system but not by the automated methods, and many observations with extreme relatives were flagged by the automated methods but not by the CPI system. These cases would have to be discussed with the CPI analysts.
- The sample size for some of the cells (item by province by month) were extremely small – less than five units⁶. Thus, the tolerance intervals for such cells were very wide, and therefore insensitive to extreme relatives. This problem can either be resolved through collapsing cells by region or time to get a sufficient sample size, or through establishing a maximum spread for a tolerance interval.
- Some of the outliers that the QM detected appeared to be outliers in terms of a relative, but in absolute dollars, the change in price was negligible. For example, if $p_k^t = 2.20$ and $p_k^{t-1} = 2.40$, the relative is $p_k^t / p_k^{t-1} = 2.20 / 2.40 = 0.92$, and the absolute difference is $|p_k^t - p_k^{t-1}| = |2.20 - 2.40| = 0.20$. The relative may be flagged as an outlier, but it is the granularity of the currency denominations that created the apparent outlier. To reduce the flagging of such observations, a minimum constraint on the absolute price change could be established. Alternatively, the minimum constraint could be increased by increasing the a parameter in the tolerance interval formulas when possible.

Other measures can be taken to reduce the number of outliers to manually verify. Outliers may be ranked in terms of the degree of consistency, so that the reviewer can prioritize the flagged outliers that he/she would like to verify. Items can also be ranked in terms of basket share, so that only flagged outliers belonging to items with large shares are verified.

Another trend visible in all of the graphs is that the lines for some of the items are vertical. These vertical lines indicate that for these items, the hit-rate is not improved by increasing the number of outliers detected i.e., by decreasing the c -value. Thus, with such items, a relatively large c -value reduces the number of false outliers without sacrificing the hit-rate, and therefore will be more resource-efficient.

The graphs also show that the items belonging to the high variability group - gasoline, beef, and deodorant – all seem to exhibit a similar pattern in that hit-rates of exactly or close to 100% are achieved, even with large c -values. The medium and low variability items exhibit a hit rate that increases significantly with little change in the increase of the number of detected outliers. This observation suggests that a large decrease in the c -value significantly enhances the hit-rate without significantly increasing the number of detected outliers that would have to be verified. In summary, these graphs provide strong evidence that the parameter values (c -values) should vary depending on the variability of the item. Further study would be required in determining a sufficient number of variability groups, and the parameter value appropriate for each group.

3. SUMMARY

The purpose of our study was to determine the outlier detection method that is most appropriate for CPI data. The performance of each method was measured by observing how well each method captured the true outliers via the ‘hit-rate’, and by noting how many false outliers each method identified. In general, it was found that the automated methods detected significantly more outliers than the true outliers. The plausible causes for such a discrepancy were highlighted, along with proposals to address these causes.

The empirical results confirmed some known properties of each of the methods. Under the QM, the log and HB transformations produced almost identical results. The RF method also produced similar results, but the hit-rates were not as strong, which is reflective of the wider intervals that the RF method produces in comparison to the QM tolerance intervals. The results for TA significantly differed from the other methods, and often produced lower hit-rates. This result was not surprising, given that the theory behind the TA significantly differs from the theory of other methods.

⁶ When quartiles are being used, there should be a minimum of 15-20 observations.

Secondary questions relating to the primary purpose were also studied. One such question was to determine the optimal symmetrizing transformation to use with the QM (and the RF method). We showed mathematically and empirically that the HB transform is an approximation of the log transform under certain conditions. However, it can also be shown that with strongly skewed distributions, the log transform will outperform the HB transform. Hence, in the context of price relative data, there does not appear to be any advantage in preferring the HB transform over the log transform.

Another secondary question was to help determine the ideal c parameter values to input into each method. The art of determining the c -value involved finding a balance between maximizing the hit-rate and minimizing the resources (involved in manually verifying false outliers) required to achieve that hit-rate. In relation to this question, we observed that there was a relationship between the c parameter and the variability of each item. This relationship may be used to determine c -values for groups of items exhibiting the same variability, rather than establishing c -values for each item, which would be resource-intensive, difficult to maintain, and unreliable in the presence of small sample sizes. This issue requires further study using simulations in a controlled environment.

3.1. Limitations of Our Study

The study assumes that the method for outlier detection programmed into the current system correctly identifies outliers. We found that this assumption may not always be valid. The CPI system did not always flag “obvious” extreme observations, and sometimes it flagged observations that were very stable. This occurrence may explain the large discrepancy in the outliers detected between the automated methods and current system method. This issue is under investigation.

Due to the high volume of items in the CPI, it was impossible to present detailed results for the entire data set. Instead, we narrowed down our analysis to eight items. The expectation was that the results obtained from these items would be indicative of the results for other items belonging to the same item class and variability. However, for more representative results, further analysis incorporating more items could be performed. We are currently performing a volume test on all of the items in the CPI basket to obtain an overall assessment on the hit-rate and number of false outliers detected.

This study can be expanded on several fronts in order to obtain more comprehensive results. The range of c -values studied could be expanded to better observe the hit-rate and the occurrence of false outliers. Breakdowns of results by province and period could be studied to explore the possibility of establishing province-specific or period-specific parameter values. While non-parametric methods are simple and fast, more sophisticated parametric outlier detection methods could also be researched and may prove to be more accurate for identifying aberrant units if an appropriate model can be found. Access to at least 24 month’s worth of data would have been more appropriate to deduce a trend or draw some strong conclusions, especially for seasonal items and specials.

3.2. Conclusion

Our preliminary study showed that the QM applied to transformed data produced the best hit-rate in general. We recommend using the log transform over the HB transform to symmetrize the data, since the log transform is more robust in the presence of strongly skewed distributions, although in our study both transforms performed approximately the same. To use the QM to its full capability, appropriate c -values must also be determined. Our study showed that the c -value should be proportional to the variability of the data, so that highly variable data should make use of large c -values, while low c -values should be used with low-variability data. These c -values could be updated every one or two years using recent data.

3.3. Discussion Questions

1. Are there any other considerations in our study that we should take into account?
2. Are there other tests or approaches that we may take to compare the different methods?
3. How could we determine c -values and create groups of items with similar variability?

REFERENCES

- Saïdi, A., and Rubin-Bleuer, S., “*Detection of Outliers in the Canadian Consumer Price Index*”, Business Survey Methods Division, Statistics Canada, May 2005.
- Fuller, W. A., “*Simple Estimators for the Mean of Skewed Populations*”, *Statistica Sinica* 1(1991), 137-158.
- International Labor Organization, International Monetary Fund, Organization for Economic Co-operation and Development, United Nations Economic Commission for Europe, The World Bank, “*Consumer Price Index Manual Theory and Practice*”, 2004.
- Lee, H., Ghangurde, P.D., Mach, L., and Yung, W., “*Outliers in Sample Surveys*”, Statistics Canada, August 1992.
- United Kingdom’s Office of National Statistics, “*Consumer Price Indices Technical Manual*”, 2006.
- Rais, S., “*Outlier Detection for Statistics Canada’s Consumer Price Index*”, Business Survey Methods Division, Statistics Canada, 2007.
- Thompson, K., and Sigman, S., “*Statistical Methods for Developing Ratio Edit Tolerances for Economic Data*”, *Journal of Official Statistics*, Vol. 15, No 4, 1999.