

Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

A cautionary note on Clark Winsorization

by Mary H. Mulry, Broderick E. Oliver, Stephen J. Kaputa
and Katherine J. Thompson

Release date: December 20, 2016



Statistics
Canada Statistique
Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at STATCAN.infostats-infostats.STATCAN@canada.ca

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-877-287-4369

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under “Contact us” > “Standards of service to the public.”

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard table symbols

The following symbols are used in Statistics Canada publications:

- . not available for any reference period
- .. not available for a specific reference period
- ... not applicable
- 0 true zero or a value rounded to zero
- 0^s value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
- ^P preliminary
- ^r revised
- X suppressed to meet the confidentiality requirements of the *Statistics Act*
- ^E use with caution
- F too unreliable to be published
- * significantly different from reference category ($p < 0.05$)

Published by authority of the Minister responsible for Statistics Canada

© Minister of Industry, 2016

All rights reserved. Use of this publication is governed by the Statistics Canada [Open Licence Agreement](#).

An HTML version is also available.

Cette publication est aussi disponible en français.

A cautionary note on Clark Winsorization

Mary H. Mulry, Broderick E. Oliver, Stephen J. Kaputa and Katherine J. Thompson¹

Abstract

Winsorization procedures replace extreme values with less extreme values, effectively moving the original extreme values toward the center of the distribution. Winsorization therefore both detects and treats influential values. Mulry, Oliver and Kaputa (2014) compare the performance of the one-sided Winsorization method developed by Clark (1995) and described by Chambers, Kokic, Smith and Cruddas (2000) to the performance of *M*-estimation (Beaumont and Alavi 2004) in highly skewed business population data. One aspect of particular interest for methods that detect and treat influential values is the range of values designated as influential, called the detection region. The Clark Winsorization algorithm is easy to implement and can be extremely effective. However, the resultant detection region is highly dependent on the number of influential values in the sample, especially when the survey totals are expected to vary greatly by collection period. In this note, we examine the effect of the number and magnitude of influential values on the detection regions from Clark Winsorization using data simulated to realistically reflect the properties of the population for the Monthly Retail Trade Survey (MRTS) conducted by the U.S. Census Bureau. Estimates from the MRTS and other economic surveys are used in economic indicators, such as the Gross Domestic Product (GDP).

Key Words: Outlier; Masking; Monthly retail trade survey.

1 Introduction

Recently we studied methods of detecting and treating verified influential values with the goal of finding an objective method for identification and treatment of influential values in a highly skewed business population (Mulry et al. 2014). An observation is considered influential if its value is correct but its weighted contribution has an excessive effect on the estimated total or period-to-period change. Although influential values occur infrequently in economic surveys, if one appears and is not “treated,” it may introduce substantial over- or under-estimation of survey totals or period-to-period change. In turn, this can impact other measures of the economy. For example, monthly estimates of sales and inventories from the U.S. Census Bureau’s Monthly Retail Trade Survey (MRTS) are inputs to the Gross Domestic Product (GDP). With any outlier detection and treatment method, one aspect of particular interest is the range of values that methods designate as influential, called the detection region. The size of the region and its boundary directly impact the number of identified values and the minimum amount by which the value(s) will be adjusted. Consequently, it is important to understand how to “manipulate” the method used, to ensure that (1) true influential values are always identified and receive the minimum treatment needed to ameliorate their impact on totals without overly perturbing the sample’s distribution and (2) values that are not influential are rarely identified and are consistently associated with trivial adjustments.

One approach for detecting and treating influential values is called Winsorization. These procedures replace extreme values with other, less extreme values, effectively moving the original extreme values toward the center of the distribution. Winsorization procedures may be one-sided by treating only extreme values that are too high, or they may be two-sided by simultaneously treating high and low values. Values

1. Mary H. Mulry, Broderick E. Oliver, Stephen J. Kaputa, and Katherine J. Thompson, U.S. Census Bureau, Washington, DC 20233, U.S.A.
E-mail: mary.h.mulry@census.gov.

designated as influential are modified (“treated”) by replacing them with values chosen to minimize the mean squared error (MSE) of the estimate of the total. For further discussion, see Chambers (1986), Chambers et al. (2000), and Martinoz, Haziza and Beaumont (2015).

In this note, we focus on the Clark Winsorization, a one-sided method developed by Clark (1995) and described by Chambers et al. (2000). The Clark Winsorization method assumes a data model and then uses an algorithm to detect and treat influential values. The detected and treated values form the detection region. Our studies found the Clark Winsorization algorithm can be effective, but the resultant detection region is highly dependent on the number of influential values in the sample. If the sample contains no influential values, the procedure is anti-conservative, meaning it makes very small changes to several values not considered influential thus reducing the variance and mean square error but essentially leaving the estimated total unchanged (trimming). On the other hand, the procedure can become very conservative if the sample contains a single influential value, depending on the distance of the value from the remainder of the distribution. When the sample contains two or more influential values, Clark Winsorization detects and adjusts only the influential values and does not trim any values that are not influential. However, the procedure can be prone to masking (Barnett and Lewis 1994). Trimming observations when no influential value is present does not appeal to subject matter analysts in a production setting where time is limited. The cost of examining a “false positive” can be prohibitive and treated values might be categorized as imputed in response rate computations. However, the algorithm has the advantage of being straightforward to implement and not requiring prior knowledge of the population. Certainly there are situations where these advantages of Clark Winsorization may outweigh the disadvantages.

We examine the influential value detection regions from Clark Winsorization using a simulated dataset that realistically reflects the population of the MRTS and was first used in (Mulry et al. 2014). We illustrate how the presence of one versus two high influential values can affect the detection region under several scenarios. Our objective is *not* to advocate for or against this method; the purpose of this note is to make potential users aware of aspects of this procedure that can affect its outcome.

Section 2 contains background on monthly business surveys including an overview of the sample design and weighting. A description of the Clark Winsorization methodology and its implementation using MRTS data appears in Section 3. The discussion in Section 4 concentrates on the detection region for influential values with Section 4.1 addressing the scenario of one influential value in a sample and Section 4.2 focusing on the scenario when two influential values are present. Section 5 contains a summary.

2 Business survey setting

As is typical of many business surveys, the MRTS is sampled from a highly skewed population of companies. The MRTS selects a sample every five years using a stratified simple-random sample design. Primary strata are determined by major industry as reported by the company, whose units are further sub-stratified by estimated annual sales (U.S. Census Bureau 2014). When the sample is introduced, small businesses are generally sampled at a low rate and have large sampling weights, whereas the larger businesses are sampled at a higher rate and have small sampling weights, again typical of many business

survey designs (Smith 2013). Originally, all businesses in the same sampling strata have the same sampling weight. However, weights for individual businesses may be adjusted as the sample matures due to persistent increases in sales for some businesses and decreases for others. For this reason, simulating a realistic weighting structure for a matured sample is challenging. When influential values do appear, it is the combination of the weight and the reported sales that produces the unusually large weighted value.

Sampling weights for small units can be very large, so examining the unweighted values to identify influential values would be quite misleading. We illustrate the combined effect of weight and sales with a single sample (replicate) throughout this note. Figure 2.1 presents plots of sampling weights against unweighted and weighted values of sales, respectively from this sample. Certainty businesses – those sampled with probability equal to 1 – are marked by hollow circles. The graph on the left shows that the units that have the smallest values of sales tend to have the highest sampling weights. By design, as the observed (unweighted) value of sales increases, the sampling weight likewise decreases. However, as shown in the graph on the right, the weighted value from both the small and large businesses contribute similarly to the estimated total. Indeed, a small value of sales multiplied by a large sampling weight can easily affect the estimated total.

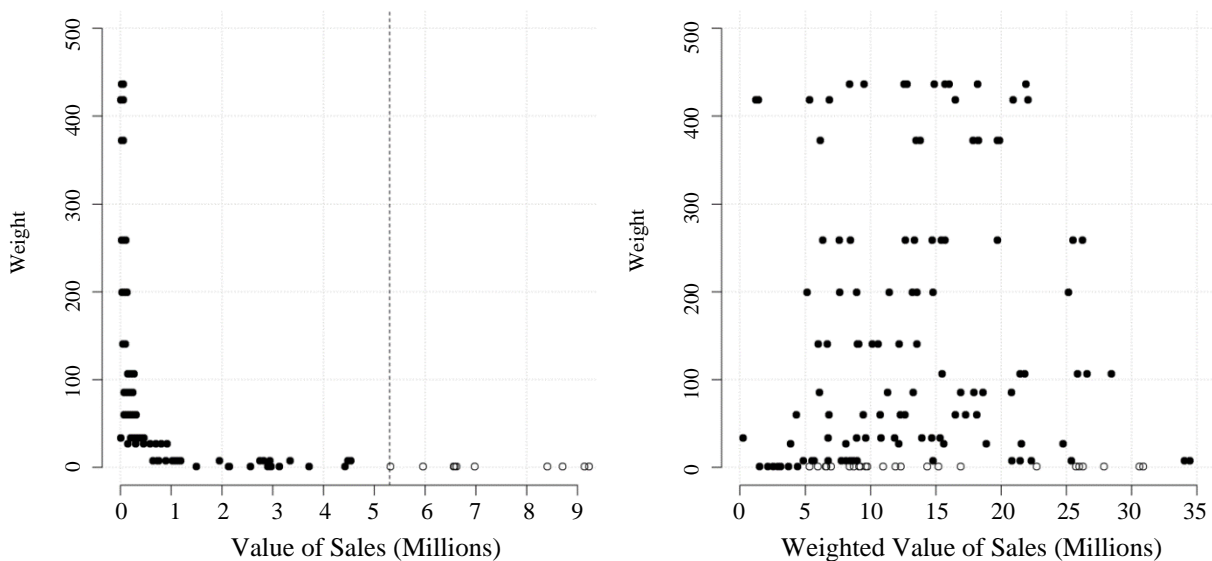


Figure 2.1 On the left, sampling weight versus *unweighted* value of sales. On the right, sampling weight versus *weighted* value of sales for unit. Units selected with certainty are shown as hollow circles.

Economic surveys publish totals and period-to-period change estimates. Influential values are examined with respect to their weighted impact on the total. If the estimates of total sales vary greatly by period, the change estimates are affected accordingly. Currently, when an influential value is detected, the mitigation strategy depends on whether the subject matter experts believe the observation is a one-time phenomenon or a persistent shift. If the influential value appears to be an atypical occurrence for the business, then the

influential observation is replaced with an “imputed” value that is more consistent with the remainder of the distribution whether or not it fails an edit [Note: if the replacement value were obtained via Clark-Winsorization, then it would technically be an “adjusted” value]. If the influential value persists, indicating a permanent change, then its sampling weight is adjusted.

3 Method

We first introduce notation needed to describe the Clark Winsorization, which follows Mulry et al. (2014). For the i^{th} business in a survey sample of size n for the month of observation t , Y_{it} is the collected characteristic (e.g., sales), w_{it} is its survey weight (which may or may not be equivalent to the inverse probability of selection), and X_{it} is a variable highly correlated with Y_{it} , such as previous month’s revenue. The monthly total Y_t is estimated by \hat{Y}_t defined by $\hat{Y}_t = \sum_{i=1}^n w_{it} Y_{it}$.

For ease of notation, we suppress the index for the month of observation t in the remainder of this section. In MRTS, the survey weight w_i is the (possibly modified) sampling weight since the missing data treatment is imputation.

The general form of the one-sided Winsorized estimator of the total is designed for large values and is written as $\hat{Y}^* = \sum_{i=1}^n w_i Z_i$ where $Z_i = \min\{Y_i, K_i + (Y_i - K_i)/w_i\}$.

Detection of observation i as an influential value by Clark Winsorization occurs when $Z_i \neq Y_i$. More than one observation may be identified. Note that using $Z_i = \min\{Y_i, K_i\}$ would ensure bounded influence and a robust estimator. However, this may lead to a large bias in \hat{Y}^* .

To implement the method, Clark assumes a general model where the Y_i are characterized as independent realizations of random variables with $E(Y_i) = \mu_i$ and $\text{var}(Y_i) = \sigma_i^2$. Then the approach approximates the K_i that minimizes the MSE under the model by setting $K_i = \mu_i + L(w_i - 1)^{-1}$, which requires estimating μ_i and L . Clark’s approach builds on a method developed by Kocic and Bell (1994) that derived a K for each stratum rather than for each individual unit.

For an estimate of μ_i , Chambers et al. (2000) suggest using the results of a robust regression. In our application, we used the SAS Procedure ROBUSTREG (SAS 2014) to implement the weighted least median of squares (LMS) robust regression method. The LMS robust regression uses weights to compensate for the heteroscedasticity visible in Figure 2.1. Other considered methods appeared too sensitive with our data, designating some observations as influential when they were not large enough to have an excessive effect on the estimated total in our empirical data sets. In different applications, different robust regression methods could exhibit superior performance and should be considered. Our prediction model estimates μ_i with bX_i where b is the regression coefficient and X_i is the previous month’s observation, chosen because X_i and Y_i tend to be highly correlated and no administrative data are available on a monthly basis. To estimate L , the Clark Winsorization procedure uses the estimate of μ_i to estimate weighted residuals

$$D_i = (Y_i - \mu_i)(w_i - 1) \text{ by } \hat{D}_i = (Y_i - bX_i)(w_i - 1).$$

Certainty units have weighted residual values of zero, assuming that no other weight adjustments are performed (e.g., for unit nonresponse, for post-stratification). Next, the method sorts the estimates of the residuals in *decreasing* order $\hat{D}_{(1)}, \hat{D}_{(2)}, \dots, \hat{D}_{(n)}$. Then the Clark method finds the largest value of k , called k^* , such that $(k+1)\hat{D}_{(k)} - \sum_{j=1}^k \hat{D}_{(j)}$ is positive, then estimates L by $\hat{L} = (k^* + 1)^{-1} \sum_{j=1}^{k^*} \hat{D}_{(j)}$. Finally, the estimate of K_i is formed by $\hat{K}_i = bX_i + \hat{L}(w_i - 1)^{-1}$, which is used to determine the values of Z_i for the estimate of the total \hat{Y}^* . Chambers et al. (2000) recommend forming the estimate of L for the procedure by using an average of estimates of L from several previous months of data. However, our examples in Section 4 use only the previous month because we use data from a simulated *stationary series* constructed to reflect the different means and variances in the sampling strata for an industry in the MRTS. The stationary series was created by constructing a simulated population from MRTS data and applying an ARMA model to generate the time series. Thus, additions and deletions to the MRTS sample over time (i.e., births and deaths) are not incorporated in the simulation design. Consequently, averaging over several previous months offers no advantage over the point estimate from the previous month. In addition, we used the Winsorized values as auxiliary values (X_i) in the application of the procedure to the subsequent month in order to study the propagation of the effects of the adjustment in the production setting. Although influential values were induced by adding a large amount to an observation selected at random from a stratum with one of the largest weights, the calculation of the value of L used all the sample observations with weights greater than one. More details on the construction of the series may be found in Mulry et al. (2014). We have not explored using an average of estimates of L from several previous months with simulated MRTS data that incorporated seasonality, volatility, and changes in economic conditions or with empirical MRTS data. Such an average of estimates of L may be useful in other designs and surveys that exhibit more stable behavior, such as annual rather than monthly implementations.

4 Detection regions

We examine the range of influential values that Clark Winsorization designates as influential, called the detection region, under three scenarios. One scenario has a single high influential value present in the sample. In the other two scenarios, the sample contains two high influential values.

Figures 4.1 and 4.2 use grids of *unweighted* data to illustrate the detection regions for the application of the Clark Winsorization algorithm on a single sample from a simulated MRTS industry with low volatility, a monthly revenue of \$2.5 billion and a sample size of 147. In these figures, each (x, y) point on the grid represents a possible influential value where x represents the unweighted value for the previous month and y represents the current month's unweighted value. Since the weights for the same business rarely change from month-to-month, the scatterplots of weighted values are similar and therefore are not shown. We use sampling weights for the points on the grid and do not modify the weights in our simulation. All the points on a *vertical* line have the same weight with the sample weights lower for units that have higher values of sales. The detection regions are constructed by inserting each pair of (x, y) coordinates from the grid into the sample and then running the Clark Winsorization algorithm with the parameter settings described in Section 3 to see if the weighted y value in the inserted pair is designated as influential.

4.1 Results for one influential value

In this section, we illustrate the effect on the detection region of a sample containing a single influential value, hereafter referred to as *Scenario 1*. In Figure 4.1, the unweighted sample observations used to form the detection regions are shown in black with the x -axis representing the previous month's value and the y -axis representing the current month. The robust least median of squares regression line used in the prediction model has been included for reference. For the given sample, a single observation that falls in the light gray hashed region (detection region) is flagged as influential and adjusted by the Clark Winsorization method. The broken vertical line marks the largest sampled observation with a weight greater than one; that is, all observations to the right of this asymptote are guaranteed to have a weight of one.

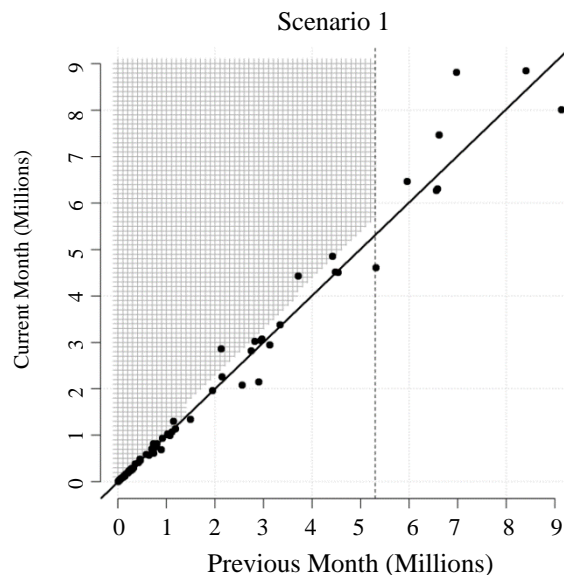


Figure 4.1 Detection region for Clark Winsorization for a single influential value. All sample points are in black.

The close proximity of the lower boundary line of the detection region to the regression line reflects the trimming done by the method to minimize the MSE by lowering the variance at the cost of introducing a small bias. Therefore, several non-influential cases in this detection region will be trimmed slightly nonetheless. We observed this phenomenon repeatedly in several other (different) empirical data sets.

4.2 Results for two influential values

Now we turn to investigating the detection region when the sample contains two induced high influential values. Our approach holds one induced observation at a fixed value and weight in the sample and allows the second induced observation to vary in value with its corresponding weight, permitting the identification of the detection region for the second observation, conditional on the first. This approach allows us to assess whether the procedure is subject to *masking* which occurs when a large value prevents the identification of other extreme values. We consider two scenarios for the fixed value. In *Scenario 2*, the contribution of the

fixed influential value to the estimate of total sales is 667 million higher than the previous month. In *Scenario 3*, the fixed influential value is less severe since its contribution is 334 million higher, half of the increase in *Scenario 2*.

The graph on the left in Figure 4.2 presents the detection region (light gray area) under *Scenario 2*. Here, the fixed (unweighted) value is 350,000 in the previous month and 8.2 million in the current month with a weight of 85. Regardless of where the second observation was placed throughout the graph, the fixed observation was always designated as influential. Notice that the observations that would have been *falsely* designated as influential and slightly trimmed in *Scenario 1* (see Figure 4.1) would *not* have been changed in this scenario. Here, the detection region is restricted to identifying only *similar* severe observations, which are supposed to be atypical.

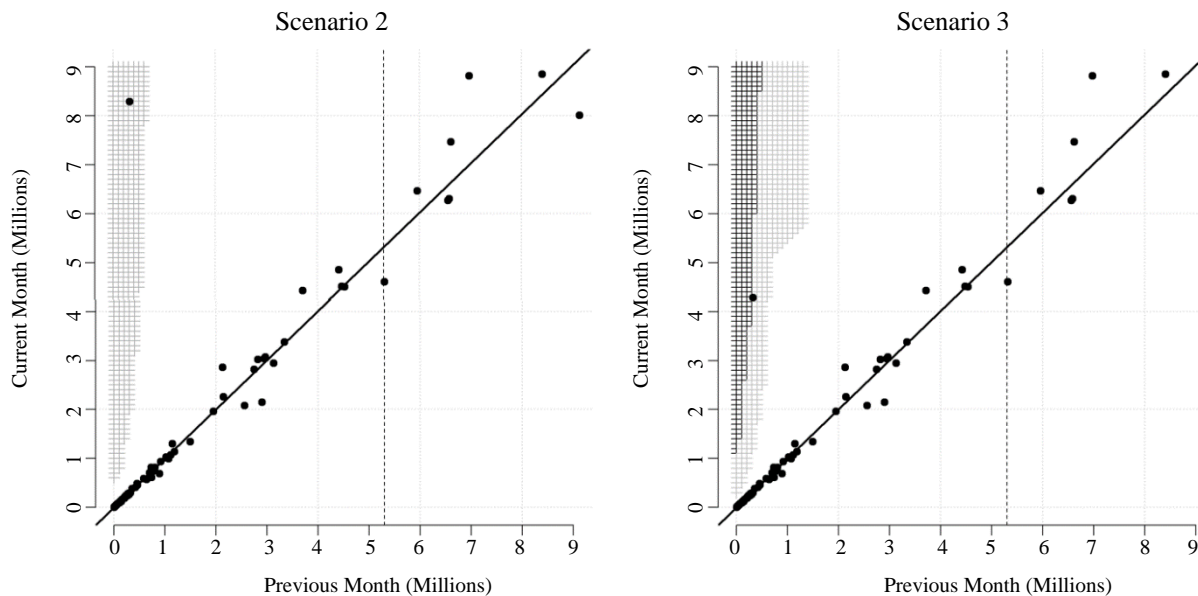


Figure 4.2 Detection regions for Clark Winsorization for the second of two influential values. On the left, the first one is held fixed and is extreme. On the right, the first one is held fixed but is less extreme. All sample points are in black.

The dramatic difference in the relative sizes of the detection regions between *Scenario 1* and *Scenario 2* could indicate that this procedure – as applied – is vulnerable to masking. Masking occurs when one influential value causes a failure to identify the presence of another (Barnett and Lewis 1994). We explore this possibility in *Scenario 3*, halving the unweighted value of the fixed influential value in the current month (now 4.1 million instead of 8.2 million) while allowing the weight to retain the same value of 85. The graph on the right in Figure 4.2 shows two different shaded regions: a light grey area where *both* the fixed influential value and the second (variable) value can be detected, and a dark grey region to the left of the light grey region where the algorithm detects the *variable value* as influential but misses the fixed value.

Adjustments in the light gray area reduce both the bias and the MSE. In the dark gray area, the adjustment reduces the MSE, but may not reduce the bias substantially. The white area to the right of the light gray region shows where only the fixed influential value is identified. However, the white area contains large observations with small weights so these observations are not representing much more than themselves, and consequently adjustments in this range have small impact on the bias.

This preliminary exploration validates our concern about the potential for masking. One approach that may alleviate masking when a stationary series has a high level of noise is to average L over several previous months as suggested in Chambers et al. (2000). The sampling design may be a factor. The graph on the left in Figure 2.1 shows that the weights decline rapidly as the unweighted observations increase for observations between 0 and 1 million. In this range, the weight of the unit has more impact than its observed value on the size of its weighted residual used in calculating the k^* . A relatively small change in the variable value may trigger a much larger change in its weighted residual and cause the k^* to change, which affects the number of influential values detected. The weights used in this example reflect the weights used in the MRTS for the industry and were not constructed artificially to create an illustration for the Clark Winsorization methodology.

5 Summary

The usage of Clark Winsorization is very appealing for the simplicity of its implementation and lack of parameters as long as one can build a viable robust regression model. However, as with many outlier detection procedures, the method has certain vulnerabilities that are not always obvious. This note demonstrates how the procedure can be effective at identifying and treating influential values, but is also highly sensitive to the number of influential values in the sample and their magnitude with respect to the regression line used to determine the detection region bounds. The properties of the detection region vary by whether an influential value is present and by the number and severity when one or more appear. If the sample contains no influential values, the procedure is anti-conservative in that it trims values not considered influential to minimize the MSE (by reducing the variance). In contrast, the procedure can become very conservative depending on the degree of difference of the weighted influential value from the others in the sample. When the sample contains two or more influential values, Clark Winsorization detects and adjusts only the influential values and does not trim any values that are not influential. However, our results demonstrate a potential for masking which should be considered when implementing the procedure.

If the occurrence of an influential value is truly a rare event and large influential values are of interest, then the small trimming of a handful of values that are not influential is a disadvantage. However, in applications where influential values are common or where historic data are not available for modeling, implementing Clark Winsorization definitely requires an assessment of the amount of trimming to determine if the aggregated small changes greatly affect the estimated total. If not, then this is an appealing approach. If yes, then other methods such as M -estimation – which give more control over the detection region – may be advantageous.

Acknowledgements

This report is released to inform interested parties and encourage discussion of work in progress. The views expressed on statistical, methodological, and operational issues are those of the authors and not necessarily those of the U.S. Census Bureau. The authors thank Lynn Weidman, Eric Slud, Scott Scheleur, William C. Davie Jr. and Carma Hogue for their helpful reviews of previous versions of the manuscript. The authors also thank Ray Chambers for his comments during presentations of our work in progress. The authors appreciate the comments from the Associate Editor and the anonymous Referees.

References

- Barnett, V., and Lewis, T. (1994). *Outliers in Statistical Data*. 3rd Edition. New York: John Wiley & Sons, Inc.
- Beaumont, J.-F., and Alavi, A. (2004). Robust generalized regression estimation. *Survey Methodology*, 30, 2, 195-208. Paper available at <http://www.statcan.gc.ca/pub/12-001-x/2004002/article/7752-eng.pdf>.
- Chambers, R.L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81, 1063-1069.
- Chambers, R., Kokic, P., Smith, P. and Cruddas, M. (2000). Winsorization for identifying and treating outliers in economic surveys. *ICES II, The Second International Conference on Establishment Surveys, Survey Methods for Businesses, Farms, and Institutions*, American Statistical Association, 717-726.
- Clark, R. (1995). *Winsorization Methods in Sample Surveys*. Masters Thesis. Department of Statistics. Australia National University. <http://hdl.handle.net/10440/1031> (accessed September 29, 2016.).
- Kokic, P.N., and Bell, P.A. (1994). Optimal winsorising cut-offs for a stratified finite population estimator. *Journal of Official Statistics*, Stockholm, Sweden, 10, 419-435.
- Martinoz, C.F., Haziza, D. and Beaumont, J.-F. (2015). A method of determining the winsorization threshold, with an application to domain estimation. *Survey Methodology*, 41, 1, 57-77. Paper available at <http://www.statcan.gc.ca/pub/12-001-x/2015001/article/14199-eng.pdf>.
- Mulry, M.H., Oliver, B.E. and Kaputa, S.J. (2014). Detecting and treating verified influential values in a monthly retail trade survey. *Journal of Official Statistics*, 30(4), 1-28.
- SAS (2014). *Help and Documentation*. SAS Institute, Inc. Cary, NC.
- Smith, P. (2013). Sampling and estimation for business surveys. In *Designing and Conducting Business Surveys*, (Eds., G. Snijkers, G. Haraldsen, J. Jones and D. Willimack), Hoboken, NJ: John Wiley & Sons, Inc., 219-52.
- U.S. Census Bureau (2014). *Monthly Retail Trade Survey Methodology*. U.S. Census Bureau, Washington, DC. http://www.census.gov/retail/mrts/how_surveys_are_collected.html (accessed September 29, 2016).