DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 30, No. 4, 2014, pp. 721–747, http://dx.doi.org/10.2478/JOS-2014-0045

Detecting and Treating Verified Influential Values in a Monthly Retail Trade Survey

Mary H. Mulry[1], *Broderick E. Oliver*[2], *and Stephen J. Kaputa*[3]

In survey data, an observation is considered influential if it is reported correctly and its weighted contribution has an excessive effect on a key estimate, such as an estimate of total or change. In previous research with data from the U.S. Monthly Retail Trade Survey (MRTS), two methods, Clark Winsorization and weighted M-estimation, have shown potential to detect and adjust influential observations. This article discusses results of the application of a simulation methodology that generates realistic population time-series data. The new strategy enables evaluating Clark Winsorization and weighted M-estimation over repeated samples and producing conditional and unconditional performance measures. The analyses consider several scenarios for the occurrence of influential observations in the MRTS and assess the performance of the two methods for estimates of total retail sales and month-to-month change.

Key words: Outlier; Winsorization; M-estimation.

1. Introduction

In survey data, an observation is considered influential if its value is correct but its weighted contribution has an excessive effect on an estimated total or period-to-period change. To be clear, our focus is on influential values that remain after all the data have been verified or corrected, so these unusual values are true and not the result of reporting or recording errors. Failure to "treat" such influential observations may lead to substantial over- or under-estimation of survey totals, which in turn may lead to overly large increases or decreases in estimates of change.

The presented research was motivated by a request from the methodologists and subject matter experts who supervise the U.S. Census Bureau's Monthly Retail Trade Survey (MRTS) to find a method that improves or replaces current methodology for identifying and treating influential values. New methodology would need to use the influential observations, but in a manner that assures their contribution does not have an excessive effect on the monthly totals or an adverse effect on the estimates of month-to-month

[1] U.S. Census Bureau, Washington, DC 20233, U.S.A. Email: mary.h.mulry@census.gov
[2] U.S. Census Bureau, Washington, DC 20233, U.S.A. Email: broderick.e.oliver@census.gov
[3] U.S. Census Bureau, Washington, DC 20233, U.S.A. Email: stephen.kaputa@census.gov

Acknowledgments: The authors thank Katherine Jenny Thompson for her useful suggestions. The authors are very grateful to Jean-Francois Beaumont for providing the SAS code for the M-estimation algorithm and for all his advice and consultations. The authors thank Lynn Weidman, Eric Slud, Scott Scheleur, William Davie, Jr., Paul Smith, two referees, and the associate editor for their helpful comments on earlier versions of this manuscript. This article 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.

change. The tight time schedule for producing MRTS estimates monthly means that the preference is for a new methodology for detecting and treating influential values that is automated, but is implemented in a manner that allows for a final (manual) review. Therefore, the objective of this research is to find an automated statistical procedure to replace the current subjective procedure performed by analysts.

Each month, the MRTS surveys a sample of about 12,000 retail businesses with paid employees to collect data on sales and inventories. The MRTS is an economic indicator survey whose monthly estimates are inputs to the Gross Domestic Product estimates. Moreover, significant changes in levels are important to monetary and budgetary decision makers, economists, business analysts, and economic researchers in assessing the health of the economy, and in making corporate investment decisions. The MRTS sample design is typical of business surveys, employing a one-stage stratified sample with stratification based on major industry, further substratified by the estimated annual sales. The sample design requires the sampling rates to be higher in the strata with the larger units than in the strata with the smaller units and companies that have been determined to comprise a large portion of the total are included with certainty. The sample is selected every five years after the Economic Census and then updated as needed with a quarterly sample of births (new businesses) and removal of deaths (businesses no longer in operation). MRTS publishes Horvitz-Thompson estimates of totals, as well as month-to-month change. Because of its typical sample design and characteristic data, the results that we obtain by studying the program in detail should be applicable to other similar programs.

In the MRTS, when an influential observation appears in a month's data, the current corrective procedures depend on whether the subject-matter experts believe the observation is a one-time phenomenon or a permanent shift. If the influential value appears to be an atypical occurrence for the business, then the influential observation is replaced with an imputed value. If the influential value persists for a few months and appears to represent a permanent change, then methodologists adjust its sampling weight using principles of representativeness or move the unit to a different industry when the nature of the business appears to have changed (Black 2001). Prior to influential value detection, the MRTS processing already includes running the algorithm by Hidiroglou and Berthelot (1986), often called the HB edit, to identify (and – on occasion – treat) within-imputation-cell outliers and create the imputation base (Hunt et al. 1999). Treatment of influential values is the final step of the estimate review process. Hence, the methods described here are developed to complement, not replace, the HB edit.

The research reported in this article builds on several previous studies on methods of addressing influential values in the MRTS. Initial work (Mulry and Feldpausch 2007a) examined a variety of outlier detection and treatment methods from the literature on empirical data from one month of a volatile MRTS industry with an obvious influential value. Of the considered methods, Clark Winsorization (Clark 1995) and M-estimation (Beaumont and Alavi 2004; Beaumont 2004) emerged as the most promising. This study examined several methods, including a second type of Winsorization that developed the cut-off value for the observations by stratum (Kokic and Bell 1994) (instead of specifying an individual cut-off value for each observation as in Clark Winsorization) and a combination of robust estimation and reverse calibration to address influential values

(Ren and Chambers 2003; Chambers and Ren 2004). Mulry and Feldpausch (2007a) concluded that the MRTS data was too volatile for the other methods, which may perform very well in other situations. One might also consider the robust estimators studied by Hulliger (1995) or Farrell and Salibian-Barrer (2006) for other applications.

Subsequent work (Mulry and Feldpausch 2007b) with 38 months of empirical MRTS data from the same industry confirmed the potential for both methods (Clark Winsorization and weighted M-estimation) to address influential values in MRTS data. The infrequent appearance of influential values in empirical data made it difficult to evaluate the performance of the considered methods with respect to relative magnitude of identified influential observation(s) or to examine the statistical properties of the considered methods over repeated samples. Consequently, Mulry and Oliver (2009) conducted a simulation study and presented some preliminary but inconclusive results.

The focus of this article is the use of simulation methodology to investigate these two robust statistical methods for identifying and treating influential observations: Clark Winsorization (Clark 1995) and M-estimation (Beaumont and Alavi 2004; Beaumont 2004). In a sample survey setting, robust methods are especially appealing since they are valid for a variety of probability distributions and therefore are less sensitive to model misspecifications. This is especially important for economic data that generally have skewed populations where the assumption of a normal distribution, or even symmetry, is unlikely to hold.

Building on past research, we developed simulation methodology to obtain decisive results about the statistical properties of Clark Winsorization and weighted M-estimation when applied to data like that collected for industries in the MRTS. The methodology includes simulation of a stationary time series for the population data and the development of performance measures. This simulation examines the effectiveness of the methodologies when seasonal effects are *not* present to illuminate the properties of the methods.

This article describes the simulation methodology and includes performance results for Clark Winsorization and M-estimation in several scenarios for influential values. Both methods were designed for totals estimates, but the most important measure for MRTS is month-to-month change. Therefore, our analysis emphasizes the simulation's estimates of relative bias for estimates of total sales and month-to-month change, both when an influential value is present and when it is not. Additional evaluation criteria include the number of true and false detections.

2. Detection and Treatment Methods

In this section, we present the studied methods. Subsection 2.1 describes the Clark Winsorization methodology for modifying an influential value, and Subsection 2.2 discusses the M-estimation methodology that provides the choice of adjusting the influential value or its weight. Figure 1 illustrates how Clark Winsorization and M-estimation adjust an influential observation.

Before describing the methods, we first introduce the notation. For the i^{th} business in a survey sample of size n for the month of observation t, Y_{ti} is the characteristic of interest (revenue in our application), w_{ti} is its survey weight (which may be equivalent to the

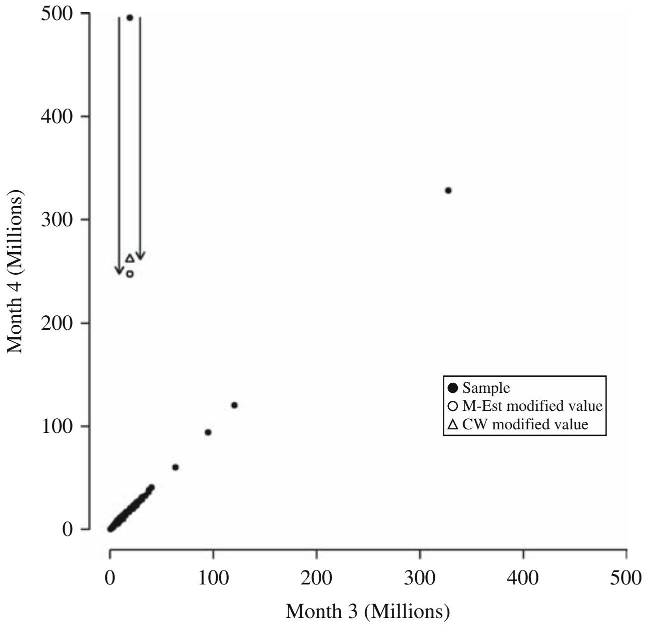

Fig. 1. Illustration of an influential value and its adjustments from Clark Winsorization and weighted M-estimation

inverse probability of selection but can include poststratification, generalized regression, or calibration adjustments), and X_{ti} is a variable highly correlated with Y_{ti}, such as the previous month's collected revenue or the frame revenue value. Note that the more general formulations allow X to be a vector of auxiliary variables. We restrict our analysis to a single covariate and set X_{ti} equal to the unit's previous month's revenue, paralleling the MRTS ratio imputation and outlier-detection (HB edit) procedures. The total monthly revenue Y_t is estimated by

$$\hat{Y}_t = \sum_{i=1}^{n} w_{ti} Y_{ti}.$$

In MRTS, the missing data treatment is imputation (Thompson and Washington 2013), and consequently, the survey weight w_{ti} is the design weight. For ease of notation, hereafter we suppress the t index. Both Clark Winsorization and weighted M-estimation methodologies use a comparison to the prior month's value to detect observations with influential values in the current month.

2.1. Clark Winsorization

Winsorization procedures replace extreme values with less extreme values, effectively moving the original extreme values toward the center of the distribution. Winsorization methods offer adjustments for the observed influential value but could be used to derive an adjustment for the survey weight if that is needed instead. Winsorization procedures may be one-sided or two-sided, but the method developed by Clark (1995) and described by Chambers et al. (2000) is one-sided.

The general form of the one-sided Winsorized estimator of the total is designed for large values and is written as

$$\hat{Y}^{*} = \sum_{i=1}^{n} w_i Z_i \quad \text{where } Z_i = \min\{Y_i, K_i + (Y_i - K_i)/w_i\}. \tag{1}$$

Detection of observation i as an influential value by Clark Winsorization occurs when $Z_i \neq Y_i$. To implement the method, Clark suggests approximating the K_i that minimizes the mean squared error under the general model by $K_i = \mu_i + L(w_i - 1)^{-1}$, using a general model where the Y_i are characterized as independent realizations of random variables with $\mathrm{E}(Y_i) = \mu_i$ and $\mathrm{var}(Y_i) = \sigma_i^2$. To estimate μ_i and L, Clark's approach builds on a method developed by Kokic and Bell (1994) that derived a K for each stratum rather than for each individual unit.

Chambers et al. (2000) suggest using the results of a robust regression to obtain the estimate of μ_i as bX_i where b is the regression coefficient and X_i is the auxiliary variable (the previous month's observed revenue in our application). We used the least median of squares (LMS) robust regression method (Rousseeuw 1984; Rousseeuw and Leroy 1987) because other robust regression methods that we considered, including the least median trimmed (LMT), appeared too sensitive in that they flagged many non-influential values (Mulry and Feldpausch 2007a). To estimate L, the Clark Winsorization first uses the estimate of μ_i to estimate weighted residuals

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

which are sorted in decreasing order $\hat{D}_{(1)}, \hat{D}_{(2)}, \ldots \hat{D}_{(n)}$. The Clark method finds the last value of k, called k*, such that $(k+1)\hat{D}_{(k)} - \sum_{j=1}^{k} \hat{D}_{(j)}$ is positive, and then estimates L by $\hat{L} = (k^{*}+1)^{-1} \sum_{j=1}^{k^{*}} \hat{D}_{(j)}$. Last, 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 $\hat{Y}^{*}$.

2.2. *Weighted M-Estimation*

M-estimators (Huber 1964) are robust estimators that come from a generalization of maximum likelihood estimation. The application of M-estimation examined in this investigation is regression estimation. The weighted M-estimation technique proposed by Beaumont and Alavi (2004) uses the Schweppe version of the weighted generalized technique (Hampel et al. 1986, 315–316). The estimator of the total using this approach is consistent for a finite population since it equals the finite population total when a census is conducted (Särndal et al. 1992, 168).

A key assumption of the M-estimation approach is that y_i given x_i is distributed under the prediction model m with

$$E_m[y_i | x_i] = x_i'\beta \text{ and } V_m[y_i | x_i] = v_i \sigma^2. \tag{1.1}$$

In our application, y_i is the current month's value; x_i is the previous month's value, and the regression model does not include an intercept. With retail trade, the regression of current month's sales on the previous month's sales tends to go through the origin (Huang 1984). We use the unbiased sampling weights w_i to maintain parallel estimation with the MRTS.

Briefly, the method estimates $\hat{B}^M$, which is implicitly defined by

$$\sum_{i \in S} w_i^*(\hat{B}^M)(y_i - x_i\hat{B}^M)\frac{x_i}{v_i} = 0 \tag{2}$$

where

$$\begin{gathered} v_i = \lambda x_i \\ w_i^* = w_i \psi\{r_i(\hat{B}^M)\}/r_i(\hat{B}^M) \\ r_i(\hat{B}^M) = h_i e_i(\hat{B}^M)/Q\sqrt{v_i} \\ e_i(\hat{B}^M) = y_i - x_i\hat{B}^M \end{gathered}$$

and Q is a constant that is specified. The variable h_i is a weight that may or may not be a function of x_i. The variable λ, possibly a constant, is chosen to ensure the correct specification of the form of the variance in the underlying prediction model.

Section 4 contains a discussion of the investigation to determine the settings for these parameters.

The role of the function ψ is to reduce the influence of units with a large weighted residual $r_i(\hat{B}^M)$. We focus on two choices for the function ψ, Type I and Type II Huber functions, and describe their one- and two-sided-forms. The one-sided Type I Huber function is

$$\psi\{r_i(\hat{B}^M)\} = \left\{ \begin{array}{l} r_i(\hat{B}^M), r_i(\hat{B}^M) \leq \varphi \\ \varphi, \text{otherwise} \end{array} \right\} \tag{4}$$

where φ is a positive tuning constant. This form is equivalent to a Winsorization of $r_i(\hat{B}^M)$. Detection of observation i as an influential value by M-estimation with the Huber I function occurs when $r_i(\hat{B}^M) > \varphi$. In the two-sided Huber I function $r_i(\hat{B}^M)$ is replaced by its absolute value $\left|r_i(\hat{B}^M)\right|$.

The weight adjustment corresponding to the Type I Huber function ψ above is

$$w_i^*(\hat{B}^M) = \left\{ \begin{array}{l} w_i, r_i(\hat{B}^M) \leq \varphi \\ \dfrac{\varphi}{r_i(\hat{B}^M)}, \text{otherwise} \end{array} \right\} \tag{5}$$

an undesirable feature of using the Type I Huber function is that the unit's adjusted weight may be less than one if the influential value is very extreme, thereby not allowing the influential value to represent itself in the estimation. The Type II Huber function ψ ensures that all adjusted units are at least fully represented in the estimate. The one-sided Type II Huber function is

$$\psi\{r_i(\hat{B}^M)\} = \left\{ \begin{array}{l} r_i(\hat{B}^M), r_i(\hat{B}^M) \leq \varphi \\ \dfrac{1}{w_i} r_i(\hat{B}^M) + \dfrac{(w_i - 1)}{w_i}\varphi, \text{otherwise} \end{array} \right\} \tag{6}$$

where φ is a positive tuning constant. Detection of observation i as an influential value by M-estimation with the Huber II function occurs when $r_i(\hat{B}^M) > \varphi$. In the two-sided Type II Huber function $r_i(\hat{B}^M)$ is replaced by its absolute value $\left|r_i(\hat{B}^M)\right|$. This form is equivalent to a Winsorization of $r_i(\hat{B}^M)$, cf. the Type I Huber function.

An interesting feature of using the one-sided Type II Huber function in the M-estimation method is that the parameters can be set to mimic the assumptions of the Clark Winsorization outlined in Subsection 2.1 (Beaumont 2004). However, the results will not be identical because the method used to estimate $\hat{B}^M$ is different.

Solving for $\hat{B}^M$ requires the Iteratively Reweighted Least-Squares algorithm in many circumstances, although for certain choices of the weights and variables, the solution is the standard least-squares regression estimator.

The weight adjustment for the Type II Huber function above is

$$w_i^*(\hat{B}^M) = \left\{ \begin{array}{l} w_i, r_i(\hat{B}^M) \leq \varphi \\ 1 + (w_i - 1)\dfrac{\varphi}{r_i(\hat{B}^M)}, \text{otherwise} \end{array} \right\}. \tag{7}$$

The adjusted value corresponding to the Type II Huber function is

$$y_i^* = \frac{1}{w_i} y_i + \frac{(w_i - 1)}{w_i} \left\{ x_i \hat{B}^M + \frac{\sqrt{v_i}}{h_i} Q\varphi \right\}. \tag{8}$$

We use an adjusted value Beaumont and Alavi (2004) derived by using a weighted average of the robust prediction $x_i\hat{B}^M$ and the observed value y_i of the form

$$y_i^* = a_i y_i + (1 - a_i) x_i \hat{B}^M \quad \text{where} \quad a_i = \frac{w_i^*(\hat{B}^M)}{w_i}. \tag{9}$$

Beaumont (2004) finds an optimal value of the tuning constant φ by deriving and then minimizing a design-based estimator of the mean-square error via numerical analysis. Unlike the Clark Winsorization algorithm, the Beaumont version of M-estimation does not require a model to hold for all the data, or for the influential value, in particular, and therefore relies on less stringent assumptions.

Since the algorithm is an iterative procedure, convergence is not guaranteed. Failure of convergence appears to be more problematic with the use of two-sided Huber functions than with one-sided Huber functions. Section 4 contains more discussion of the possible consequences when convergence is not achieved.

3. Methodology

3.1. Research Approach

To assess how well M-estimation and Clark Winsorization identify and treat influential values in MRTS data, we conduct a simulation study using different – but realistic – influential value scenarios, considering detection and treatment effects on estimates of totals and of current-to-prior period change.

To do this, we generated two separate time-series populations of monthly sales data, modeled from two MRTS industries with different natures. We generate a stationary time series for each industry to avoid potential confounding of the influential value detection methods and other patterns such as trends or seasonality. Industry 1 has monthly sales of approximately 46.1 billion and one of the most volatile industries. Industry 2 has a more stable pattern and has monthly sales of approximately 2.5 billion. The sample sizes in our simulations are 1,161 for Industry 1 and 147 for Industry 2. Subsection 3.2 describes the procedure used to generate these simulated populations.

Our simulation evaluation approach is two-fold: an *unconditional analysis* where a small subset of the samples (replicates) contain an induced influential value and the majority do not; and a *conditional analysis* that employs only the subset of samples that contain the induced influential value. The objective of the unconditional analysis is to evaluate the performance of Clark Winsorization and M-estimation over a realistic survey setting, where it is not expected that each sample will include an influential value. The objective of the conditional analysis is to evaluate the respective performance of each approach when the sample does contain an influential value.

In practice, the most common scenario pertaining to influential values is an observation whose measurement is much higher than previous measurements and whose high weight greatly amplifies its impact on the estimates. Failure to address this scenario properly can have far-reaching consequences in interpreting the state of the economy, so we focus on this scenario.

3.2. Simulation Methodology

Recall that the MRTS is a stratified sample, with strata defined by unit size within industry where the measure of size is sales. An exploratory empirical analysis of the simulated data for both studied industries confirmed that the stratum-level means differ by within-industry-strata as shown in the examples in Figure 2, and that a realistic within-stratum prediction model is given by the stationary series.

$$\hat{y}_{hi,t} = \beta_h \hat{y}_{hi,t-1} + \varepsilon_{hi}, \varepsilon_{hi} \sim (0, \sigma^2_{hi}), t > 1$$

where h indexes the strata as illustrated in the examples in Figure 3.

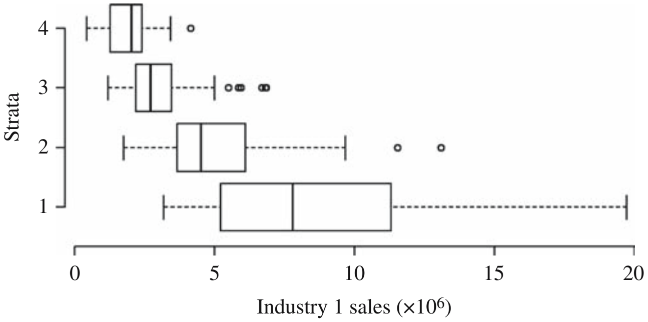

Fig. 2. Stratum-level Box-plots for simulated retail trade Industry 1

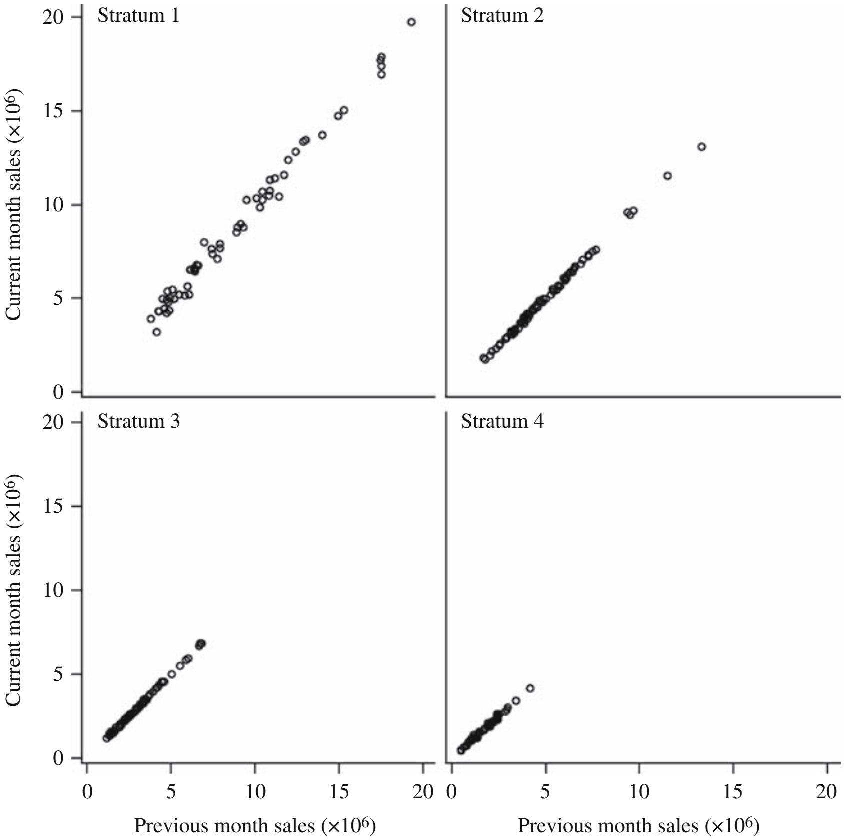

Fig. 3. Scatter plots of current month to previous month sales at the stratum level for simulated retail trade Industry 1

In the notation provided in Subsection 2.2, the “true” prediction model for the simulated data is $E_m\left[y_{hi,t}|y_{hi,t-1}\right] = y'_{hi,t-1}\beta_{h,t-1}$ and $V_m\left[y_{hi,t}|y_{hi,t-1}\right] = \sigma_h^2,$ so that $v_h \equiv 1$ within stratum.

To obtain a series 20 months in length, we generated the population for the first month and then generated the next 19 months as a stationary time series essentially as a forecast going forward from Month 1. The population data for the first month were generated using the SIMDAT algorithm (Thompson 2000) with modeling cells equal to sampling strata and population size equal to the original frame size in each cell. The stationary time series was generated using historical standard errors and autocorrelations to develop the AR(1) model within stratum for Months 2 to 20 given by

$$y_t - m = \Phi^*(y_{t-1} - m) + a_t, \quad \text{for } t = 2, . . ., 20 \tag{10}$$

where

$y_1 - m = 0$ and m is the series mean,

$a_t \sim N(0, \sigma^2)$ (white noise process where s is estimated empirically by the observation for the unit in the first month times the median of percent difference between observations in the historical first and second months),

Φ = the sample-based estimate of lag one autocorrelation for the selected industry.

The time series algorithm written in SAS creates an AR(1) series so that each new observation is set equal to Φ times the previous value $+$ a_t, where a_t is generated from the $N(0, \sigma^2)$ distribution. The initial value of the series is set to zero so that each subsequent point has an expected value of zero – which is necessary for series to be stationary. After all 20 observations for a unit have been created, the initial value (first month value) is added to them so that this number is actually the mean over the time series (in short, it shifts the mean from zero to the first month value).

Generating the series in this manner assures that each of the two populations (one for each industry) is a stationary series within strata, but not at the industry level. Our simulated population data follow directly from the stratification model and mimic the conditions under which the influential observation procedures would be implemented (i.e., after micro-data automatic editing/imputation and HB outlier detection). However, the stratification model diverges greatly from the prediction models assumed by Clark Winsorization (industry-level models, with one population model describing the industry data) and by M-estimation (also, industry-level, with the underlying weighted regression model using the v_i term to account for expected increasing variability with unit size). The funnel shape of the plot in Figure 4 illustrates how the variance of the observations of the retail trade industry data increases as the values of the observations increase. However, Figure 5 illustrates that neither the assumption $v_i = 1$ nor the assumption $v_i = x_i$ for the v_i in the prediction model in Equation (1.1) fits the data well at industry-level, but at the same time, both assumptions appear to have comparable weaknesses. Therefore, we defer the choice of the setting for v_i until we view the detection error rates as defined later in this section and discussed further in Subsection 4.1.

To assess the statistical properties of each influential value treatment method (M-estimation and Clark Winsorization), we induce an influential value into the

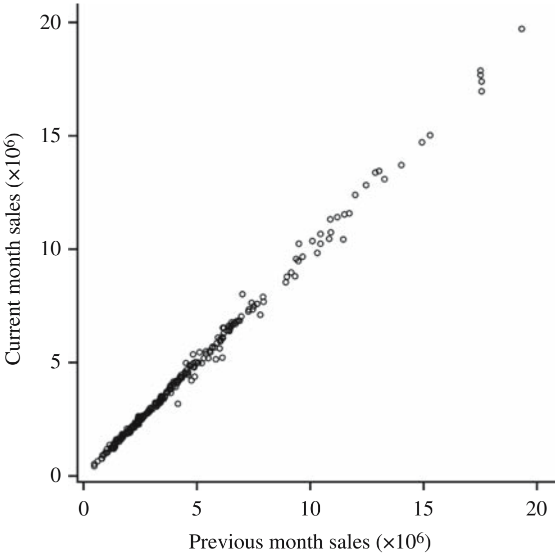

Fig. 4. Industry-level scatter plot of current month to previous month sales for simulated retail trade Industry 1

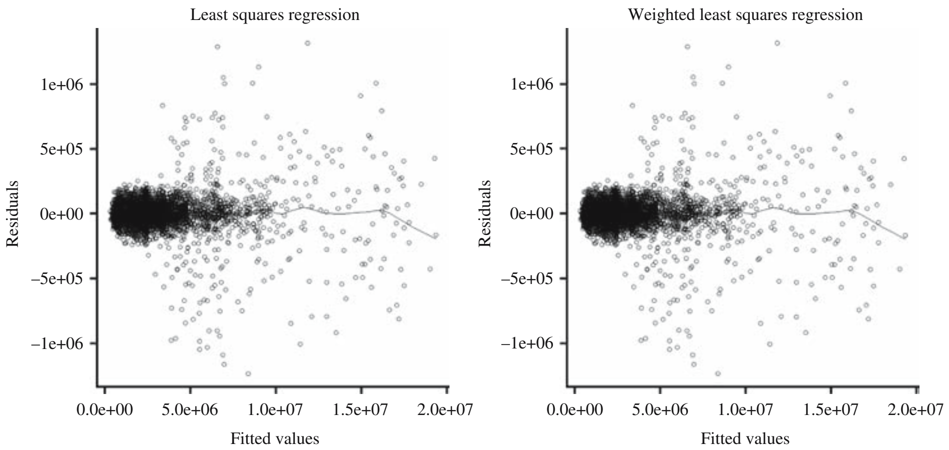

Fig. 5. Residual versus Fitted Values with LOESS curve from models for predicting Industry current month sales using previous month sales with Least Square Regression corresponding to $v_i = 1$ (left) and Weighted Least Square Regression where the weight $= 1/x_i$ corresponding to $v_i = x_i$ (right)

population in Month 4. The choice of Month 4 allows gauging the performance in the months before as well as after the influential value appears which is particularly important for estimates of month-to-month change. The induced influential value does not have an undue effect on the population total, but does have undue influence on the estimated population total if selected in sample. The details of constructing the time series for the population follow using Industry 1 for illustration; the same procedure generated the Industry 2 population.

First, we generate a time series for the Industry 1 population of length 20 months using the methodology described in the first paragraph of this section. We let $Y_1, Y_2, \ldots, Y_{20}$ represent the population totals for this stationary series.

Next, we create one influential unit in the population in Month 4 in a stratum with a sampling rate of approximately 1/50 by adding eight million to the unweighted value of a randomly selected unit in this stratum. Hence, the population total for Month 4 is now eight million larger than its initial value. This influential value *does not* have an undue effect on the population at approximately 46.1 billion in Month 4, but it can have an undue influence on the estimated population total if selected in sample since its weighted value is 400 million larger than its initial weighted value. With this design, we can expect the unit to be selected for one of every 50 samples and when selected, increase the estimated total by about one percent. The induced influential value in the simulation is based on influential values that occurred during the 38 months of the MRTS examined in Mulry and Feldpausch (2007b).

After creating the population time series, we select stratified simple random without replacement (SRS-WOR) samples of size comparable to the MRTS sample from Month 1 until 200 of these samples contain the unit that has the induced influential value in Month 4. The choice of 200 samples was a function of the processing requirements for M-estimation because the required number of samples to achieve 200 with the influential value was quite large and the algorithm had to be run on the total number of samples in the unconditional analysis. For Industry 1, the necessary number of samples is 10,742, and for Industry 2, the necessary number of samples is 11,931. By requiring the same unit to be included in all samples in the conditional analysis, we effectively reduce the size of the probability sample by one, but continue to give the influential value its stratum weight. This results in a small bias in the months without the induced influential value, and the magnitude of the bias is a function of how close the unadjusted unit's value is to the stratum mean in these months.

In each independent sample, we apply the M-estimation and Clark Winsorization algorithms to Month 2 using Month 1 as the auxiliary data and then continue to apply both methods to each month through Month 20 using the previous month as the auxiliary data. Modified values in a given month are used as auxiliary data in the next month. After repeating these procedures on each independent sample, we conduct the two analyses mentioned in Subsection 3.1, a conditional analysis that uses only the 200 samples with the influential value and an unconditional analysis using all the samples.

3.3. Estimators and Evaluation Criteria

To define the estimators, we first need some notation:

δ = u for the unconditional analysis,
c for the conditional analysis.
$S(\delta)$ = the total number of samples selected for analysis δ
$S(u)$ = 10,742 for the unconditional analysis in Industry 1
11,931 for the unconditional analysis in Industry 2
$S(c)$ = 200 for the conditional analysis in Industry 1 and Industry 2
ε = the outlier detection method
m = M-estimation
w = Clark Winsorization, none for the untreated estimate
Y_t = the true population total of the simulated data for month t
$\hat{Y}_{t,i}$ = the untreated estimate of Y_t for month t in sample i
$\hat{Y}^{\varepsilon}_{t,i}$ = the treated estimate of Y_t for month t in sample i with ε = M-estimation or Clark Winsorization.

The mean of the simulated values for month t, analysis δ, method ε is an estimate of Y_t

$$\hat{Y}^{\varepsilon}_t(\delta) = \frac{\sum_{i=1}^{S(\delta)} \hat{Y}^{\varepsilon}_{t,i}}{S(\delta)}.$$

The population values of the change are:
$\frac{Y_t}{Y_{t-1}}$ = true month-to-month change for the simulated data in month t, $t = 2$ to 20.
The estimates of this change are:
$\frac{\hat{Y}^{\varepsilon}_t(\delta)}{\hat{Y}^{\varepsilon}_{t-1}(\delta)}$ = estimate of month-to-month change for month t, analysis δ, method ε.

Now, let E^{ε}_t be a month t true population value, namely Y_t (total sales) or $\frac{Y_t}{Y_{t-1}}$ (month-to-month change). Also, let $\widehat{E}^{\varepsilon}_{ti}(\delta)$ be the estimate of total sales or month-to-month change for month t, analysis δ, method ε from replicate i. Then the relative bias (RB) of $\widehat{E}_t^{\varepsilon}(\delta)$ is

$$\mathrm{RB} = \frac{\sum_{i=1}^{S(\delta)} \left[\frac{100(\widehat{E}_{ti}^{\varepsilon}(\delta) - E_t^{\varepsilon})}{E_t^{\varepsilon}}\right]}{S(\delta)}. \tag{11}$$

We expect that the RB of the treated estimate is less than or equal to the RB of the untreated estimate in most circumstances.

The relative root mean square error (RRMSE) of $\widehat{E}_t^{\varepsilon}(\delta)$ is

$$\mathrm{RRMSE} = \sqrt{\frac{\sum_{i=1}^{S(\delta)} \left[\frac{100(\widehat{E}_{ti}^{\varepsilon}(\delta) - E_t^{\varepsilon})}{E_t^{\varepsilon}}\right]^2}{S(\delta)}}. \tag{12}$$

We expect that the RRMSE of the treated estimate is less than or equal to the RRMSE of the untreated estimate since the methods minimize MSE.

Mirroring Thompson and Sigman (1999), to evaluate the outlier detection performance of each method, we view each application as a hypothesis test, in which the null hypothesis is "the data item's value is *not* an influential value". One rejects the null hypothesis when the item's value is flagged as influential. Under this framework, two types of errors can occur:

- **Type I error rate** equals the percentage of observations that were *not induced* influential values that were designated as influential (false positive). If a method adjusts values that are not induced influential values, then the Type I error rate will be positive.
- **Type II error rate** equals the percentage of *induced* influential values that were not detected (false negative). The Type II error rate applies only to samples containing the induced influential value. So, the Type II error rate is equal to 0 in Months 1–3 and 5–20 since no influential values were induced in these months.

4. Results

In this section, we examine the simulation results regarding the performance of the two treatments and the quality of the estimates they produce. The Clark Winsorization algorithm does not require parameter settings, but the M-estimation algorithm does. First, we investigate the settings of the parameters for the M-estimation algorithm to determine which options produce the best estimates. Then we use those settings for M-estimation in the comparison with Clark Winsorization. As we will see in the simulation results, the choices of the M-estimation parameter settings affect whether the algorithm converges in some situations and therefore are important. For the Winsorization, we developed the software in SAS. For the M-estimation, we used SAS software developed by Jean-Francois Beaumont (personal communication), with minor modifications.

4.1. M-estimation Algorithm Settings

The M-estimation algorithm discussed in Subsection 2.2 requires settings for Q, h_{i}, v_i, the function ψ, and an initial value of the tuning constant φ. We use the default settings of

$Q = 1$ and $h_i = (w_i - 1)\sqrt{x_i}$, but explore different settings for the other parameters, as summarized in Table 1. We also consider whether to include the observations selected with certainty in fitting the regression line.

Our investigation considers two values of the weighting parameter for the residuals $v_i = \lambda x_i$ namely $v_i = x_i$ and $v_i = 1$. The choice $v_i = 1$ corresponds to $\lambda = 1/x_i$ so that $V_m\left[y_i | x_i\right] = \sigma^2$ (equal variances) and the choice $v_i = x_i$ corresponds to $\lambda = 1$ so that $V_m\left[y_i | x_i\right] = x_i \sigma^2$. Ideally, the choice of the setting for v_i should be a data-driven decision because v_i essentially specifies the variance of the model errors underlying the regression estimator for M-estimation. In our (realistic) setting, neither $v_i = x_i$ nor $v_i = 1$ provide a good model for the studied *industry* level estimates from the MRTS data. Indeed, this model misspecification is an inherent challenge with economic data.

Notice that when we used the default settings for Q and h_i along with setting $v_i = x_i$ for all units in sample, $r_i = (w_i - 1)(y_i - x_i \hat{B}^M)$ has the same form as $\hat{D}_i$ in the Clark Winsorization. However, recall that the b in the Winsorization estimation method and the $\hat{B}^M$ in the M-estimation method are not usually going to be equal because they use different estimation methods. With $Q = 1$ and $h = (w_i - 1)\sqrt{x_i}$ (the default settings), setting $v_i = 1$ tends to give the residuals for large weighted values of x_i more influence in fitting the M-estimation regression line than when $v_i = x_i$.

The M-estimation algorithm detects and adjusts influential values through finding an optimal value of the tuning constant φ, which is the cut-off value for the weighted regression residuals. The user sets an initial value for the tuning constant φ, and the algorithm finds the value of φ that minimizes the mean squared error (MSE). Setting the algorithm parameters in a manner appropriate for the MRTS data requires considerable investigation. We consider two options for the function ψ, the one-sided Huber I and II functions described in Subsection 2.2 and two options for the initial value of φ, one high and the other low. After exploring the application of M-estimation to samples that included and excluded the units selected with certainty, we found little difference and included the certainty units in our simulation. The units selected with certainty contribute to fitting the regression line but cannot be designated as influential because $r_i(\hat{B}^M)$ equals zero for a certainty unit with the default setting $h_i = (w_i - 1)\sqrt{x_i}$.

Selecting the high and low initial values of φ for the simulation depends on the data for the industry. If there are no weighted residuals larger than the initial value of φ, the M-estimation algorithm runs for only one iteration and does not offer any adjustments.

Table 1. M-estimation algorithm parameters

Parameter	Parameter function	Values
Q	Constant	1 (default)
h_i	Unit weight	$(w_i - 1)\sqrt{x_i}$ (default)
v_i	Model error underlying regression estimator	1 or x_i
ψ	ψ function	Huber I or Huber II
φ	Tuning constant (determines starting point for critical region)	User provides initial value and program calculates optimal value

Therefore, for low initial φ we choose a value that tended to be lower than the highest weighted residual in a sample since we wanted the algorithm always to run in the simulation. For the high initial φ, we want only to assure that the algorithm detects the induced influential value when it appears in Month 4. Consequently, we choose a value that is lower than the weighted residual for the induced influential value but higher than the weighted residuals for the other values. For Industry 1, the low initial φ is 4.8 million and the high value initial value is 150 million. The low and high initial values of φ for Industry 2 are 1.5 million and 150 million, respectively.

Table 2 summarizes the results for Type I and Type II errors for the parameter settings for Industry 1 and Industry 2 and offers results for the different parameter settings and functions using Type I and Type II errors as the evaluation criteria. A Type I error (false positive) may occur in all the months in all the samples, but a Type II error (false negative) may occur only in Month 4 of the 200 samples with the induced influential value in Month 4.

Both settings for the parameter v_i display some Type I errors when the initial setting of φ is the low value of 4.8 million while there are no Type I errors when the initial φ is the large value of 150 million. The Type I errors occur because the algorithms for Clark Winsorization and M-estimation when the initial φ is low (4.8 million) make small adjustments to several observations to achieve the minimum MSE although the reduction in MSE is small.

Remember that neither $v_i = 1$ nor $v_i = x_i$ is an appropriate error model for the simulated data for either of the two industries. The Type I and Type II errors are very similar for the two choices of the function ψ, Huber I and Huber II, when the same high or low initial φ is used in the unconditional analysis. The Type II error rate for $v_i = 1$ is zero for both options for the initial φ in Month 4 for Industry 1 for both Huber I and Huber II. However, when $v_i = 1$ for Industry 2, the Type II error rate is 0.0065 for the high initial φ, and 0.04 for Huber I and 0.05 for Huber II for the low initial φ. The Type II error rate when $v_i = x_i$ is always zero for all combinations of the options.

Table 2. Summary of M-estimation results for the unconditional analysis with Industry 1 and Industry 2 data in the scenario of one high influential value for two settings of the parameters v_i, two settings of the initial φ, and two options for the function ψ

	ψ function			
v_i	Huber I	Huber II	Type I error	Type II error
x_i	Option 1	Option 2	• Small Type I error rate when initial φ small at 4.8 million • No Type I errors when initial φ large at 150 million	Industry 1 rate: zero Industry 2 rate: zero
1	Option 3	Option 4	• Very small Type I error rate when initial φ small at 1.5 million • No Type I errors when initial φ large at 150 million	Industry 1 rate: zero Industry 2 rates: • when initial φ small, 0.04 for Huber I, 0.05 for Huber II • when initial φ large, 0.0065 for Huber I & II

Since there is some Type II error when $v_i = 1$ and none when $v_i = x_i$, and the two settings produce about the same results regarding Type I error, we decided to pursue only $v_i = x_i$.

4.2. *One High Influential Value*

4.2.1. Industry 1 Estimates and Quality

First, we focus on the simulation results for Industry 1, the more volatile of the two simulated industries and the larger of the two (in terms of sample size and total sales). We show results for only the Huber II function ψ because results for Huber I and Huber II functions are approximately equal. Since the M-estimation algorithm is an iterative procedure, convergence is not guaranteed. We used the default convergence criterion of a difference of 0.001 between the current and previous iterations and did not explore other options. In this simulation, the algorithm did not converge for about two percent of the samples in the unconditional analysis. Usually a researcher puts a limit on the number of iterations that the algorithm may run. We chose a limit of five iterations. When the limit is reached, the program choses the larger of the last two values of φ. The results for the performance measures include the consequences of this choice. In the conditional analysis, the algorithm converged for Month 4 in all 200 samples, and the convergence properties in other months were similar to those in the same months in the rest of samples in the unconditional analysis.

The relative bias estimates of total sales in Months 2 to 7 in the unconditional and conditional analyses are shown in Table 3 while Table 4 shows the RRMSE estimates for the same months. The population value of total sales in these months varies slightly around \$46.1 billion. Tables 3 and 4 only show the results involving Months 2 through 7 because the results for the rest of the 20 months parallel those involving Month 7. This is to be expected since the series is stationary and only Month 4 has an induced influential value.

In the unconditional analysis, the untreated estimate of the total for Month 4 has a relative bias of 0.012 percent, corresponding to approximately \$4.6 million, and an even smaller relative bias in the other months, corresponding to $-$\$1.7 million to $-$\$3.6 million. Since the reported estimates of total sales are in millions, this level of bias does appear in the reported estimates and is within the survey sampling error where the coefficient of variation is approximately two percent. In Month 4, the treated estimates do reduce the bias even further, with M-estimation with a high initial φ having the lowest absolute relative bias. In the other months, estimates of total from M-estimation with a high initial φ have a relative bias equal to that of the untreated because no observations are adjusted in those months. However, in months other than Month 4, Clark Winsorization and M-estimation with the low initial φ tend to introduce additional negative relative bias, about -0.01 percent, because they tend to trim about 0.5 percent of the observations to achieve a minimum MSE. Interestingly, Table 4 shows that the three methods produce estimates of total sales for Month 4 with approximately the same RRMSE of 1.261 in the unconditional analysis. Since Table 3 shows that Clark Winsorization and M-estimation with a low initial φ have more relative bias than M-estimation with a high initial φ, we

Table 3. Relative bias (percent) for one high influential value scenario (Industry 1, 1-Sided Detection)

	Unconditional				Conditional			
		M-est. Huber II				M-est. Huber II		
	Untreated	High φ	Low φ	Clark Winsorization	Untreated	High φ	Low φ	Clark Winsorization
Total								
2	−0.005	−0.005	−0.015	−0.015	0.246	0.246	0.236	0.236
3	−0.004	−0.004	−0.016	−0.015	0.256	0.256	0.244	0.244
4	0.012	0.003	−0.008	−0.008	1.166	0.716	0.716	0.721
5	−0.006	−0.006	−0.018	−0.018	0.237	0.237	0.222	0.225
6	−0.005	−0.005	−0.016	−0.016	0.233	0.233	0.222	0.222
7	−0.007	−0.007	−0.018	−0.018	0.238	0.238	0.226	0.226
Month-to-Month change								
2 to 3	0.001	0.001	−0.001	−0.0001	0.010	0.010	0.008	0.008
3 to 4	0.016	0.007	0.008	0.008	0.908	0.460	0.471	0.475
4 to 5	−0.017	−0.009	−0.010	−0.010	−0.918	−0.476	−0.491	−0.493
5 to 6	0.001	0.001	0.002	0.002	−0.004	−0.004	<0.001	−0.003
6 to 7	−0.002	−0.002	−0.002	−0.002	0.004	0.004	0.004	0.005

Table 4. RRMSE (percent) for one high influential value scenario (Industry 1, 1-Sided Detection)

	Unconditional				Conditional			
		M-est. Huber II				M-est. Huber II		
	Untreated	High φ	Low φ	Clark Winsorization	Untreated	High φ	Low φ	Clark Winsorization
Total								
2	1.255	1.255	1.255	1.255	1.229	1.229	1.227	1.227
3	1.257	1.257	1.257	1.257	1.223	1.223	1.220	1.220
4	1.267	1.261	1.261	1.261	1.675	1.400	1.400	1.403
5	1.257	1.257	1.257	1.257	1.221	1.221	1.218	1.218
6	1.255	1.255	1.256	1.256	1.233	1.233	1.231	1.231
7	1.256	1.256	1.256	1.256	1.222	1.222	1.219	1.219
Month-to-Month change								
2 to 3	0.106	0.106	0.106	0.105	0.097	0.097	0.097	0.097
3 to 4	0.165	0.125	0.126	0.126	0.914	0.471	0.482	0.486
4 to 5	0.167	0.128	0.129	0.128	0.925	0.489	0.503	0.505
5 to 6	0.108	0.108	0.108	0.107	0.109	0.109	0.109	0.109
6 to 7	0.108	0.108	0.108	0.108	0.108	0.108	0.107	0.107

conclude that these estimates achieve a comparable RRMSE by reducing the variance through trimming several observations. We are observing a classic bias versus variance trade-off and since the bias is a small component of the RRMSE, changes to the variance have a larger impact.

When we turn to month-to-month change in the unconditional analysis, the induced influential value in Month 4 causes a positive bias in the untreated estimate of change from Months 3 to 4 and a negative bias of comparable size in the untreated estimate of change from Months 4 to 5. All the treated estimates reduce the relative bias by about half in the change from Months 3 to 4 and from Months 4 to 5. The treatments reduce the RRMSE in the untreated estimate by about 24 percent. As with the estimates of total, the relative bias and RRMSE for the untreated and treated estimates of change are comparable in the months not involving Month 4.

For the conditional analysis, Table 3 shows that the relative bias is approximately equal for all the estimates of total sales in Months 2, 3, and 5 to 7. In Month 4, the relative bias in both versions of M-estimation and Clark Winsorization is approximately 60 percent of the relative bias in the untreated estimate. Recall that the simulation design introduces a small amount of bias in the conditional analysis. Table 4 shows that Clark Winsorization and both versions of M-estimation produce estimates with approximately 84 percent of RRMSE for the untreated estimate in Month 4, but the RRMSEs are comparable in the other months.

In the conditional analysis in Table 3, we see that untreated and treated estimates of change from Months 3 to 4 have a positive relative bias and an approximately offsetting negative relative bias for the change from Months 4 to 5. The relative bias for the estimates of change that do not involve Month 4 is very small and does not appear in estimates of change which are reported in tenths of percent. When Month 4 is involved, the untreated estimates of change would be apparent in the reported estimates. All treatments reduce the relative bias by approximately one-half with M-estimation with a high initial φ having slightly less relative bias than Clark Winsorization and M-estimation with a low initial φ. The treatments also reduce RRMSE in the untreated estimates of change by about one-half with M-estimation with a high initial φ having the lowest as shown in Table 4. Apparently, the trimming by the latter two methods to reduce the variance in the estimates of total sales creates additional bias in the estimates of change when Month 4 is involved. Clark Winsorization and M-estimation with a low initial φ appear to have some residual effect in the estimate of change from Months 5 to 6 since each has a lower relative bias than the untreated estimate and M-estimation with a high initial φ. However, the RRMSEs of all four estimates of change are approximately equal.

4.2.2. Industry 2 Estimates and Quality

Now we turn our attention to the simulation results for Industry 2, which has a less volatile pattern of change and a smaller sample size than Industry 1. The population value of total sales in these months is about \$2.5 billion and the sample size is 147.

The patterns in the performance measures for the unconditional analysis for Industry 2 shown in Tables 5 and 6 are very similar to the results for Industry 1. The effect of the induced influential value in Month 4 is larger because its size relative to the population total is larger as is the effect of adjusting it. The M-estimation algorithm converged for all

Table 5. Relative bias (percent) for one high influential value scenario (Industry 2, 1-Sided Detection)

	Unconditional				Conditional			
		M-est. Huber II				M-est. Huber II		
Month	Untreated	High φ	Low φ	Clark Winsorization	Untreated	High φ	Low φ	Clark Winsorization
Total								
2	0.007	0.007	−0.028	−0.044	0.103	0.103	0.068	0.053
3	0.010	0.010	−0.029	−0.075	0.173	0.173	0.120	0.089
4	0.330	0.172	0.132	0.122	19.021	9.607	9.600	9.737
5	0.006	0.006	−0.060	−0.053	0.079	0.079	−1.307	0.024
6	0.011	0.011	−0.034	−0.050	0.191	0.191	0.133	0.127
7	0.011	0.011	−0.030	−0.047	0.134	0.134	0.094	0.079
Month-to-Month change								
2 to 3	0.004	0.004	<0.001	−0.030	0.071	0.071	0.053	0.037
3 to 4	0.319	0.162	0.161	0.197	18.829	9.424	9.475	9.647
4 to 5	−0.272	−0.151	−0.175	−0.160	−15.922	−8.697	−9.957	−8.856
5 to 6	0.005	0.005	0.026	0.003	0.112	0.112	1.461	0.104
6 to 7	<0.001	<0.001	0.004	0.003	−0.056	−0.056	−0.038	−0.048

Table 6. RRMSE (percent) for one high influential value scenario (Industry 2, 1-Sided Detection)

	Unconditional				Conditional			
		M-est. Huber II				M-est. Huber II		
Month	Untreated	High φ	Low φ	Clark Winsorization	Untreated	High φ	Low φ	Clark Winsorization
Total								
2	2.783	2.783	2.783	2.784	2.620	2.620	2.616	2.618
3	2.773	2.773	2.773	2.772	2.593	2.593	2.588	2.587
4	3.712	3.043	3.043	3.051	19.196	9.948	9.942	10.075
5	2.770	2.770	2.775	2.772	2.609	2.609	2.912	2.608
6	2.770	2.770	2.771	2.771	2.615	2.615	2.612	2.611
7	2.782	2.782	2.782	2.783	2.607	2.607	2.604	2.603
Month-to-Month change								
2 to 3	0.277	0.277	0.276	0.293	0.287	0.287	0.283	0.289
3 to 4	2.454	1.251	1.257	1.281	18.838	9.432	9.482	9.654
4 to 5	2.080	1.160	1.320	1.181	15.928	8.704	9.970	8.863
5 to 6	0.276	0.276	0.339	0.283	0.287	0.287	1.555	0.294
6 to 7	0.271	0.271	0.269	0.277	0.281	0.281	0.273	0.283

samples with the high initial φ, but with the low initial φ, each month experienced a failure to converge in approximately ten percent of the samples. However, this does not appear to change the pattern observed in the unconditional analysis for Industry 1. Yet, M-estimation with the low initial φ experienced convergence problems in Month 5 in 106 of the 200 samples in the conditional analysis, which is the focus of this section. The reason for the failure to converge is a combined effect of a low influential value in Month 5 that is the consequence of an induced very high influential value in Month 4 and the small sample size in Industry 2.

In Month 4 in the conditional analysis, M-estimation with a high initial φ reduces the relative bias in the untreated estimate of total sales by 49 percent. The reduction in relative bias using M-estimation with a low initial φ is 50 percent while for Clark Winsorization the reduction is 49 percent. Viewing the results for the estimates of total sales for the other months, the relative bias and RRMSE from M-estimation with a high initial φ equal those for the untreated. In months other than Month 5, M-estimation with a low initial φ reduces the relative bias in the untreated estimate by 30 to 34 percent while Clark Winsorization achieves reductions ranging from 35 to 49 percent. Both methods appear to be trimming as in their application to Industry 1 although the percentage reductions are greater than seen for Industry 1. However, Month 5 is different – the relative bias for M-estimation with a low initial φ has a much bigger absolute value than the untreated and is negative which makes the *RRMSE* twelve percent higher than the untreated.

When we turn to month-to-month change, we see more anomalies when Month 5 is involved. First, for estimates involving neither Month 4 nor Month 5, the relative bias for M-estimation with a high initial φ equals the relative bias for the untreated while the trimming by M-estimation with a low initial φ and Clark Winsorization achieves reductions of 17 to 48 percent, but the RRMSEs for all four estimates are comparable. The relative bias in the untreated estimate of change for Months 3 to 4 continues to offset the relative bias for Months 4 to 5. All three treatments achieve a reduction of approximately 50 percent in RRMSE of the untreated estimate of change from Months 3 to 4. For the change from Months 4 to 5, both M-estimation with a high initial φ and Clark Winsorization reduce the relative bias by about 45 percent while M-estimation with a low initial φ produces a 30 percent reduction. The reductions in the RRMSE for the untreated estimate are comparable to the percentage reductions in the relative bias for the three treatments. For the change from Months 5 to 6, the relative bias in the untreated and M-estimation with a high initial φ are equal but slightly larger than Clark Winsorization. However, the relative bias for M-estimation with a low initial φ is 1.461 percent, an order of magnitude higher than for the other three estimates.

An examination of the data provides insight about what happens with the M-estimation algorithm when using the low initial φ in Month 5 for some samples with the induced influential value in Month 4. The algorithm identifies and treats the influential value in Month 4. However, in Month 5 the sample unit returns to a range closer to its value in Month 3. In some samples, but not all, the Month 5 value is small enough to create an unusually large negative weighted regression residual as illustrated in Figure 6.

Because the version of the M-estimation algorithm used in the simulations uses a one-sided Huber II function ψ, it does not treat unusually low values, and therefore, the MSE

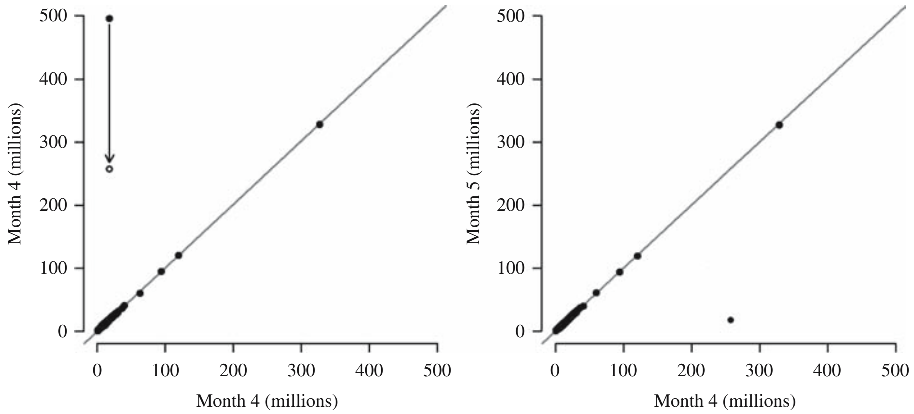

Fig. 6. Scatterplots of Month 4 versus Month 3 (left) and Month 5 versus Month 4 (right) with robust regression line when applying M-estimation with a low initial φ in a sample from Industry 2. The unusually high influential value in Month 4 was adjusted but not enough to avoid producing an unusually low influential value in Month 5 when the unit returned to its routine range

can be a strictly decreasing function of φ, which causes the algorithm not to converge. In the case of a strictly decreasing MSE, the algorithm does not converge by the limit on the number of iterations (five in our study) and instead selects the larger φ of the last two iterations, which is usually very small. This small φ causes the program to flag many observations in the sample as influential and to adjust them in over 50 percent of the samples. When using the one-sided version of the M-estimation algorithm, the adjustments reduce only observations larger than their previous month's values and thereby introduce a negative bias in the estimates of total sales. If the limit for the number of iterations increases beyond five, in some applications the algorithm converges to a local minimum that is usually very small. Therefore, increasing the number of iterations does not solve the problem.

To gauge whether a two-sided function ψ would perform better than a one-sided function ψ with a low initial value of φ, we applied the M-estimation algorithm to Months 4 and 5 to the 200 replicates that contained the influential value, but also found convergence problems. In Month 4, the algorithm failed to converge for eleven samples, but 96 of the 189 that achieved convergence produced a final value of φ that was very small and therefore, not helpful because it designated a large number of observations as influential. Results in Month 5 also were problematic since the algorithm did not converge for 39 samples and of the 161 achieved that convergence, 21 converged to nearly zero. In one other sample where the algorithm converged, it flagged more than ten percent of the observations as influential, which we consider to be many.

The samples with convergence problems caused by the induced high influential value returning to its routine range and producing a particularly low residual (Figure 6) illustrate the situation where the most desirable option probably is no adjustment. With the high initial φ setting, no residual is larger than the initial φ so the M-estimation algorithm does not run for any of the samples, and therefore, it produces no adjustment, and achieves the desirable option. This highlights the importance of choosing the initial φ to be a value low enough that an observation with a larger weighted residual requires an adjustment, but high enough for the algorithm not to run when no adjustment is needed.

5. Summary

Our investigation finds both weighted M-estimation and Clark Winsorization to be effective in identifying and treating influential values; however, each method has advantages and disadvantages that may affect a decision about which to employ. Although the simulation procedure was designed to produce data similar to the Census Bureau's MRTS, the studied problem and context are broadly applicable to other programs.

A big advantage of Clark Winsorization is the ease of implementation of its straightforward formulas. By design, the method identifies and treats only influential values that are unusually high so it does not identify or treat values that are influential because they are unusually low. However, the major concern in economic surveys regarding influential values usually is the occurrence of high ones. When an influential value is present, Clark Winsorization always identifies it and offers an adjustment.

On the other hand, the Clark Winsorization trims about 0.5 percent of the observations when no influential value is present in the sample, introducing adjustments that achieve a very small reduction in MSE for estimated totals and month-to-month change. The trimming increases the bias of the Winsorized estimate over that obtained with M-estimation with a high initial φ. Since the Clark Winsorization trimming reduces the variance in the treated estimates, the RRMSEs of the two studied methods are comparable. The trimming is also disadvantageous because the staff usually researches whether observations flagged as influential are accurate. The tight time schedule for production of monthly estimates requires avoiding unnecessary investigations. However, in some situations, the ease of implementation of Clark Winsorization and the protection that it offers against unusual influential values could outweigh the small amount of bias introduced by trimming a few falsely identified observations by a small amount. These would be situations where knowledge of the population is limited and/or where verification of values designated as influential could be restricted to focus only on those with treated values exhibiting large changes relative to the remainder of the units.

The weighted M-estimation methodology identifies and treats both high and low influential values. Our investigation focused on high influential values because they usually are the major concern in the studied programs although low influential values do occur and can introduce bias. The M-estimation algorithm has flexibility in setting parameters to make assumptions appropriate for the underlying data. In addition, weighted M-estimation with a high value of the initial tuning constant φ performed the best overall of the three options considered.

An attractive feature of M-estimation is that the algorithm allows an analyst to set the value of the initial tuning constant φ and thereby determine the minimum size of the weighted regression residuals that will be considered as potential influential values. This facilitates the efficient use of staff time in examining proposed adjustments. However, setting the initial φ is important to the effectiveness of the algorithm and needs to be a data-driven decision based on exploratory analysis. Some further refining may occur as the procedure is used in practice. In addition, there is a need to have a back-up strategy for situations when the algorithm does not converge and for situations when the algorithm converges but does not provide helpful results. In the latter cases, an influential value is present, but the MSE is either a strictly decreasing or a strictly increasing function of the tuning constant φ resulting in adjustments for almost all or none of the observations. If the MSE does have a global minimum but the algorithm does not converge, then changing the initial φ to be close to the value of φ corresponding to the minimum MSE usually results in the algorithm converging.

Research is currently underway on how to set the initial φ in an ongoing monthly survey that may or may not be subject to seasonal effects, but the approaches under study require at least minimal prior knowledge of the population. If one has no prior knowledge of the population, one could take the approach of applying Clark Winsorization. If Clark Winsorization produces no adjustment or merely trimming, then no adjustment is an acceptable choice.

Other research on M-estimation and Winsorization methods have either supported or not contradicted our findings. In a recent study with the U. S. Census Bureau's Annual Survey of Public Employment and Payroll, M-estimation also performed better than Clark Winsorization (Barth et al. 2012). In another study, Lewis (2007) attempted to formulate methodology for Winsorization of estimates of change, but did not find a satisfactory method in spite of making more restrictive assumptions than presented here.

Ultimately, we believe that the trimming of some observations by Clark Winsorization that introduces some bias for a small reduction in MSE is a less than desirable feature and instead choose to focus on M-estimation applications, with the full endorsement of the MRTS program managers. Implementing M-estimation in MRTS requires investigating the remaining issues, such as seasonality, data-driven methods of optimizing the selection of the initial tuning constant φ, and – most important – a changing economy. The flexibility of M-estimation makes the approach particularly appealing given these challenges.

6. References

Barth, J., J. Tillinghast, and M.H. Mulry. 2012. "Treatment of Influential Values in the Annual Survey of Public Employment and Payroll." In Proceedings of the 2012 Research Conference of the Federal Committee on Statistical Methods. Office of Management and Budget. Washington, DC. Available at: https://fcsm.sites.usa.gov/files/2014/05/Barth_2012FCSM_III-D.pdf (accessed October 10, 2014)

Beaumont, J.-F. 2004. *Robust Estimation of a Finite Population Total in the Presence of Influential Units*. Report for the Office for National Statistics, dated July 23, 2004. Office for National Statistics, Newport, U.K.

Beaumont, J.-F., and A. Alavi. 2004. "Robust Generalized Regression Estimation." *Survey Methodology* 30: 195–208.

Black, J. 2001. "Changes in Sampling Units in Surveys of Businesses." In Proceedings of the Federal Committee on Statistical Methods Research Conference. Office of Management and Budget. Washington, DC. Available at: http://www.fcsm.gov/files/2014/05/2001FCSM_Black.pdf (accessed October 20, 2014)

Chambers, R.L. and R. Ren. 2004. “Outlier Robust Imputation of Survey Data.” In Proceedings of the American Statistical Association, Section on Survey Research Methods [CD-ROM]. American Statistical Association. Alexandria, VA. 3336–3344. Available at: http://www.amstat.org/sections/SRMS/Proceedings/y2004/files/Jsm2004-000559.pdf (accessed October 20, 2014)

Chambers, R.L., P. Kokic, P. Smith, and M. Cruddas. 2000. “Winsorization for Identifying and Treating Outliers in Business Surveys.” In Proceedings of the Second International Conference on Establishment Surveys. Statistics Canada. Ottawa, Canada. 717–726.

Clark, R. 1995. “Winsorization Methods in Sample Surveys.” Masters Thesis. Department of Statistics. Australia National University. Available at: http://hdl.handle.net/10440/1031 (accessed October 21, 2014)

Farrell, P.J. and M. Salibian-Barrera. 2006. “A Comparison of Several Robust Estimators for a Finite Population Mean.” *Journal of Statistical Studies* 26: 29–43.

Hampel, F.R., E.M. Ronchetti, P.J. Rousseeuw, and S.A. Werner. 1986. *Robust Statistics. An Approach Based on Influence Functions*. New York: John Wiley & Sons.

Huang, E. 1984. “An Imputation Study for the Monthly Retail Trade Survey.” In Proceedings Joint Statistical Meeting, Survey Research Methods Section, American Statistical Association. Alexandria, VA. 610–615.

Huber, P.J. 1964. “Robust Estimation of a location parameter.” *Annals of Mathematical Statistics. Institute of Mathematical Statistics* 35: 73–101.

Hidiroglou, M.A. and J.-M. Berthelot. 1986. “Statistical Editing and Imputation for Periodic Business Surveys.” *Survey Methodology* 12: 73–83.

Hulliger, B. 1995. “Outlier Robust Horvitz-Thompson Estimators.” *Survey Methodology* 21: 79–81.

Hunt, J.W., J.S. Johnson, and C.S. King. 1999. “Detecting Outliers in the Monthly Retail Trade Survey Using the Hidiroglou-Berthelot Method.” In Proceedings of the Section on Survey Research Methods. American Statistical Association. Alexandria, VA. 539–543. Available at: http://www.amstat.org/sections/SRMS/Proceedings/papers/1999_093.pdf (accessed October 20, 2014)

Kokic, P.N. and P.A. Bell. 1994. “Optimal Winsorising Cut-Offs for a Stratified Finite Population Estimator.” *Journal of Official Statistics* 10: 419–435.

Lewis, D. 2007. “Winsorisation for estimates of change.” *Proceedings of the Third International Conference on Establishment Surveys*. American Statistical Association. Alexandria, VA. 1165–1172.

Mulry, M.H. and B. Oliver. 2009. “A Simulation Study of Treatments of Influential Values in the Monthly Retail Trade Survey.” *JSM Proceedings*, Survey Research Methods Section. American Statistical Association. Alexandria, VA. 2979–2993. Available at: http://www.amstat.org/sections/SRMS/Proceedings/y2009/Files/304284.pdf (accessed October 20, 2014)

Mulry, M.H. and R. Feldpausch. 2007a. “Investigation of Treatment of Influential Values.” *Proceedings of the Third International Conference on Establishment Surveys*. American Statistical Association. Alexandria, VA. 1173–1179.

Mulry, M.H. and R. Feldpausch. 2007b. “Treating Influential Values in a Monthly Retail Trade Survey.” *Proceedings of the Survey Methods Section, SSC Annual Meeting*.

Statistical Society of Canada. Ottawa, Ontario, Canada. Available at: http://www.ssc.ca/survey/documents/SSC2007_M_Mulry.pdf (accessed October 20, 2014)

Ren, R. and R.L. Chambers. 2003. "Outlier Robust Imputation of Survey Data via Reverse Calibration." S3RI Methodology Working Paper M03/19. Southampton Statistical Sciences Research Institute, University of Southampton, U.K. Available at: http://www.eprints.soton.ac.uk/8169/1/8169-01.pdf (accessed October 20, 2014)

Rousseeuw, P.J. 1984. "Least Median of Squares Regression." *Journal of the American Statistical Association* 79: 871–880.

Rousseeuw, P.J. and A.M. Leroy. 1987. *Robust Regression and Outlier Detection*. New York: John Wiley & Sons.

Särndal, C.-E., B. Swensson, and J. Wretman. 1992. *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Thompson, J.R. 2000. *Simulation: A Modeler's Approach*. New York: John Wiley and Sons. 87–110.

Thompson, K.J. and K.T. Washington. 2013. "Challenges in the Treatment of Unit Nonresponse for Selected Business Surveys: A Case Study." *Survey Methods: Insights from the Field*. Available at: http://surveyinsights.org/?p=2991 (accessed October 20, 2014)

Thompson, K.J. and R.S. Sigman. 1999. "Statistical Methods for Developing Ratio Edit Tolerances for Economic Data." *Journal Official Statistics* 15: 517–535.

Received November 2012
Revised October 2014
Accepted October 2014