DETECTING OUTLIERS IN THE MONTHLY RETAIL TRADE SURVEY USING THE HIDIROGLOU-BERTHELOT METHOD

James W. Hunt, Jennifer S. Johnson, and Carol S. King, Bureau of the Census
James W. Hunt, Bureau of the Census, SSSD, Washington, DC 20233

Key Words: Editing, Outliers

This paper describes the development of the Hidiroglou-Berthelot (H-B) edit as an alternative to the share of the market and ratio edits traditionally used to identify possible outliers for the Monthly Retail Trade Survey (MRTS). Using traditional methods, a reporting unit is an outlier if its share of the market statistic or ratio of current to prior month data exceeds a preset limit. These edits were devised at a time when the Census Bureau collected data for the MRTS on an establishment basis. Data is now collected on a company or sub-company basis for the MRTS. Changing the unit for collection exposed some previously undetected problems with these methods. The share of the market edit identifies industry leaders as outliers due to large dollar increases from month to month even though the trends are reasonable. Additionally, the ratio edit does not discriminate between companies with vast differences in dollar volume. As a result, reporting units with small dollar volume and high month to month trend will be identified as outliers even though they have little effect on a particular industry. As an alternative, the H-B edit performs a transformation of the month to month ratio that makes size a consideration. This method detects cases found by the traditional methods that appear to be genuine outliers. In addition, it detects some cases that the traditional methods fail to identify that also appear to be outliers.

1. INTRODUCTION

The U.S. Census Bureau conducts the Monthly Retail Trade Survey (MRTS) to estimate monthly sales for the retail industries in the United States.

Prior to the April 1999 data, the share of the market and ratio edits were used to identify possible outliers in the MRTS. Using these methods, a reporting unit is an outlier if its share of the market statistic or ratio of current month to prior month sales exceeds a preset limit. When devising these methods, the Census Bureau collected data for the MRTS on an establishment basis. Data is now collected on a company or sub-company basis. This change in data collection exposed some previously undetected problems. The share of the market statistic tends to identify industry leaders as outliers due to large dollar increases even with reasonable month to month trends. The ratio edit fails to discriminate between companies with vast differences in dollar volume.

After describing the traditional edits, this paper will describe the development of the Hidiroglou-Berthelot edit as a successful alternative to them.

2. SHARE OF THE MARKET EDIT

The share of the market statistic, z_i, is defined as

$$z_i = \left| \frac{w_i x_i}{x_e} - \frac{w_i y_i}{y_e} \right|$$

where

- w_i = weight of the i^{th} reporting unit,
- x_i = dollar value of the i^{th} unit for the current month,
- x_e = expected dollar volume for the current month,
- y_i = dollar value of the i^{th} unit for the previous month, and
- y_e = expected dollar volume for the previous month.

The expected dollar volume, y_e, is the estimate of total sales published for the previous month. The expected dollar volume for the current month, x_e, is computed by mulitplying y_e by the seasonal factor for the current month.

At tabulation time, we compute z_i for each reporting unit, and compare it to a predetermined parameter corresponding to its industrial classification code (IC). This predetermined parameter, KKP, is defined as

$$KKP = 2.5 * \sqrt{\frac{\sum_{i=1}^{n} z_i^2}{n}}$$

where n is the number of reporting units for each IC.

Within each IC, cases are identified for review when KKP $< z_i < 3.2*KKP$, and data is suppressed and imputed for a case when $z_i \geq 3.2*KKP$. This is basically the Chebychev inequality edit with mean of the z_i's equal to zero.

By design, the share of the market edit identifies reporting units showing the most significant change from one month to the next. It was designed for when data was collected and reviewed at the establishment level.

Currently, each unit represents a particular company or sub-company that may include several different establishments. As a result, companies with large dollar volumes and significant market shares but reasonable month-to-month trends may fail this edit test. Industry leaders repeatedly fail the edits despite accurate reporting. As a result, these cases are excluded from the imputation base, and flagged for imputation. Upon review, the analysts determine the data to be valid, and restore the reported data. However, they remain out of the imputation base. Graphical analysis further reinforces the idea that these cases are not true outliers.

Another drawback is that the expected dollar volumes used to compute the z_i are estimated using historical data based on the latest sample which may have been drawn as long as five years ago. King (1997) provides more detail on the share of the market edit.

3. RATIO EDIT

In addition to the share of the market edit, prior to April 1999, the MRTS also used the ratio edit. The ratio edit compares the ratio of current month to prior month data for the reporting unit, R_i, against the ratio of the expected dollar volumes for the current and prior months, R_e, having the same IC. The ratios are defined as

$$R_i = \frac{x_i}{y_i}$$

and

$$R_e = \frac{x_e}{y_e}$$

where x_i, y_i, x_e, and y_e are all defined as for the share of the market. Cases are suppressed and imputed when $R_i > 5*R_e$ or $R_i < 0.2*R_e$.

This test does not distinguish between companies with vast differences in dollar volume and will identify cases with small dollar volumes and high month-to-month ratios, but these cases may have little effect on an IC. Internal documentation contains more details regarding the ratio edit.

4. THE CHALLENGE

The challenge, then, is to identify cases with unusual month-to-month trends that have significant market share for a particular industry and, thus, are outliers.

As a first attempt, we examined the share-of-the-market statistic using the quartile edit. This edit identifies those cases whose share of the market statistics fall outside the range (P25 - a*IQR, P75 + a*IQR). P25 is the 25th percentile or first quartile, P75 is the 75th percentile or third quartile, IQR is the interquartile range (P75 - P25), and a is a constant set at 2 and 3 in our tests. This method identified the same cases as the share of the market edit as well as a number of other cases that did not appear to be data problems.

We then considered the Hidiroglou-Berthelot (H-B) edit.

5. HIDIROGLOU-BERTHELOT EDIT

The Hidiroglou-Berthelot (H-B) edit uses a variation of the quartile test. Once again, edits are done within the IC. The H-B edit begins with the month-to-month ratio, previously defined as R_i, and transforms it twice. The following material comes from the papers by Grandquist, by Hidiroglou and Berthelot, and by Hoglund. The first transformation is the following:

$$S_i = \begin{cases} 1 - \dfrac{R_{med}}{R_i} & 0 < R_i < R_{med} \\ \dfrac{R_i}{R_{med}} - 1 & R_i \geq R_{med} \end{cases}$$

where

S_i = the transformed ratio, and
R_{med} = the median of the R_i.

Half the S_i's are less than zero and half are greater than zero. Hoglund points out that while both tails of this transformation provide equally good detection of outliers, the transformation does not provide a symmetric distribution of the observations. Rewriting S_i as

$$S_i = \begin{cases} \dfrac{R_i - R_{med}}{R_i} & 0 < R_i < R_{med} \\ \dfrac{R_i - R_{med}}{R_{med}} & R_i \geq R_{med} \end{cases}$$

allows better observation of this lack of symmetry. Perhaps more importantly, S_i is undefined when R_i equals zero. This occurs when the current month data equals zero, and these cases are automatically flagged for review as this may indicate a reporting unit has gone out of business.

S_i still disguises the size of the reporting unit just as R_i did. Therefore, calculate the following:

$$E_i = S_i * \{\max(w_i x_i, w_i y_i)\}^u$$

where

E_i = the H-B statistic and
u = the size parameter.

The size parameter, u, may assume any value in the range, $0 \leq u \leq 1$. If this parameter is set to 0, the transformation reverts to the original S_i which, as previously discussed, masks the size of the company. On the other hand, if u is set to 1, the weighted sales provide a larger influence on the determination of the outliers. Iterative exploration led to the selection of u = .5 as the size parameter best suited for MRTS.

To use the H-B statistic to detect outliers, calculate the following:

$$D_{Q1} = \max\{E_{med} - E_{Q1}, |A * E_{med}|\}$$

and

$$D_{Q3} = \max\{E_{Q3} - E_{med}, |A * E_{med}|\}$$

where

E_{med} = the median value of the H-B statistic for a particular kind of business,
E_{Q1} = the first quartile,
E_{Q3} = the third quartile, and
A = .05.

The second term in the maximization function guards against the possibility of observations clustered tightly about the median with few variations. The constant of .05 is selected to keep this second term smaller than the interquartile distance in almost all cases (particularly for the MRTS).

The outliers then fall outside this range:

$$\{E_{med} - c * D_{Q1}, E_{med} + c * D_{Q3}\}$$

where c is the constant that determines the width of the acceptance interval. We selected three different values of c. The smallest value (20) determines which cases require analyst review.

The middle value (40) determines which cases are suppressed from the imputation base in addition to requiring analyst review. Suppression from the imputation base indicates that these cases will not be included in calculations of industry averages when imputing other data points. These cases will not, however, be imputed, and they also will not influence cases that require imputation.

Finally, the largest value of c (50) determines the cases considered for imputation as well as suppression from the imputation base and analyst review. That is, in addition to being excluded from the imputation base, the data reported for these cases will be ignored in favor of imputed data. An analyst may bypass this and restore the reported data if it is, in fact, correctly reported. At all levels of edit failure, the subject matter analyst will review the case to determine the accuracy of the reported data.

Under the traditional edits, we relied upon a two-tiered review system. The traditional edits would identify cases for review only and cases to suppress and impute.

We determined these values of c through iterative exploration and discussion with the subject matter analysts. The iterative exploration used to determine the u and c parameters consisted of examining values of u between 0 and 1 in increments of 0.1 in conjunction with values of c ranging from 5 to 75 in increments of 5.

Initially, we compared the H-B edit to the current edits for April, May, June, November, and December 1997 and January 1998 data. We found that the H-B method identified the cases found by the share of the market that we considered true outliers. Additionally, the H-B method did not identify as many cases that appeared to have good month-to-month trends, and also identified cases not found by the share-of-the-market or ratio edits that appeared to be outliers. Many of the cases detected using this method have a large effect on their particular

industry. The H-B method tended to identify more cases for review but fewer cases requiring imputation which, if implemented, would result in fewer cases needing restoration by the analysts.

Table 1 compares the share of the market and ratio edits to the H-B edit. The table displays the number of cases identified for review and for suppression and imputation for each method as well as the count of cases missed by each method. The counts come from the March 1998 tabulations. These counts exclude cases with current month or prior month data equal to zero. Counts are across all industries. See Hunt (1998), Johnson (1998), and King (1998) regarding the March 1998 data. MRTS is discussed by King (1998).

Figures 1 and 2 show the retail sales data from March 1998 for Automotive Dealers. The graphs plot the fourth root of the weighted current month data against the fourth root of the weighted prior month data. The graphs indicate a regression line and 95% confidence intervals. As in Table 1, these graphs exclude cases with current month or prior month data equal to zero. These graphs illustrate that cases detected by H-B for review and imputation, but missed by the current edits, are genuine outliers.

Figure 1 displays the outliers detected by the H-B method. Cases selected for review are indicated by diamonds. Cases marked for imputation are indicated by asterisks.

Figure 2 displays the same graph as Figure 1, but now the share-of-the-market and ratio edit outliers are highlighted. Cases selected for imputation are indicated by x's. There were no cases marked for review only by the traditional methods.

6. WHERE WE ARE

Beginning with the January 1999 data, we ran the H-B edits in production alongside the share of the market and ratio edits to ensure correct implementation. Beginning with the April 1999 data, we replaced the traditional edits with the H-B edits.

We have done additional research on the distance measurement algorithm for the selection of outliers (D-MASO), and may add this as an additional analytical tool for MRTS later in 1999. Refer to Hunt (1999) for more information regarding D-MASO.

In the future, we would also like to include exploratory data analysis (EDA) techniques in reviewing MRTS. The EDA techniques require extensive training and will most likely not be added to the monthly review during the 1999 calendar year.

DISCLAIMER

This paper reports the general results of research undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau Publications. This report is released to inform interested parties of research and to encourage discussion.

REFERENCES

Grandquist, L. (1990), "A Review of Some Macro-Editing Methods for Rationalizing the Editing Process," *Proceedings of Statistics Canada Symposium 90.*

Hidiroglou, M. A. & Berthelot, J. M. (1986), "Statistical Editing and Imputation for Periodical Business Surveys," *Survey Methodology*, **12**, 1, 73-83.

Hoglund, D. E. (1988), *A Method for Editing Periodical Data.* Master of Science thesis, University of Stockholm.

Hunt, J. W. (1998), "Monthly Wholesale Trade Survey Edits," Memorandum for the Record May 20, 1998.

Hunt, J. W. (1999), "D-MASO," Memorandum for the Record February 24, 1999.

Johnson, J. S. (1998), "H-B Edit for March RIS Cycles," Memorandum for the Record May 18, 1998.

King, C. S. (1997), "Consistency Edits of Current BSR-97 Data with Historical Levels and Trends," BSR-97 Action Memo 2K3 June 2, 1997. Updated by J. S. Johnson November 16, 1998.

King, C. S. (1998), "Use of the Hidiroglou-Berthelot (H-B) Edit for March 1998 Data Month - Retail Sales," Memorandum for the Record May 15, 1998.

Table 1. MRTS data across all industries.				
Traditional Edits / H-B Edits	Review	Suppress and Impute	Not Identified as Failures	Total
Review (c=20)	3	11	32	46
Suppress (c=40)	0	4	2	6
Impute (c=50)	0	30	1	31
Not Identified as Failures	0	2	N/A	2
Total	3	47	35	85

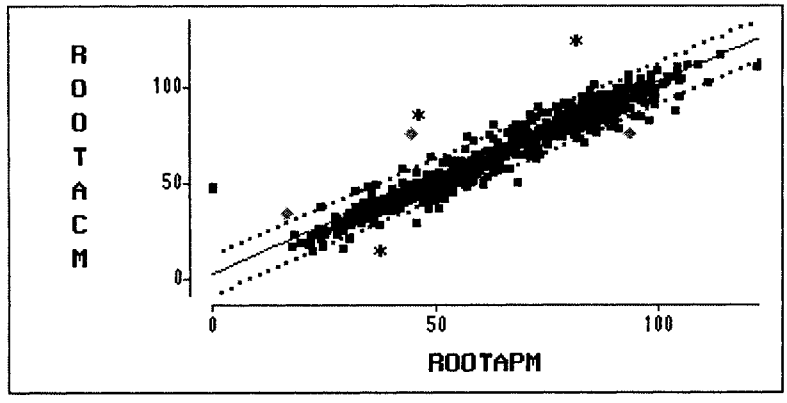

Figure 1. Current Month v Prior Month Data for Automotive Dealers. H-B outliers indicated. * represents suppress and impute. ♦ represents review.

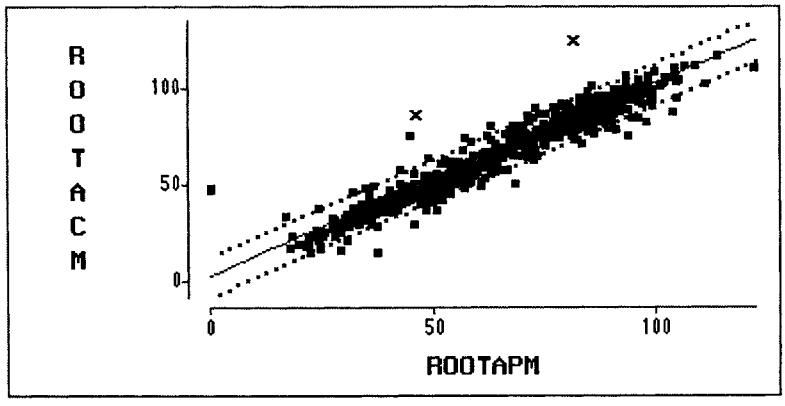

Figure 2. Current Month v Prior Month Data for Automotive Dealers. Share-of-the-market and ratio edit outliers indicated. X represents suppress and impute.