

# APPLICATION OF THE HIDIROGLOU-BERTHELOT METHOD OF OUTLIER DETECTION FOR PERIODIC BUSINESS SURVEYS

Richard Belcher <sup>1</sup>

## ABSTRACT

Detecting outliers in business surveys can be particularly difficult due to the extreme variation in the size of respondents. A well-known method for detecting outliers in periodic business surveys was created by Hidiroglou and Berthelot (1986). The strength of this method is that it allows us to include the size of the unit as being an important factor in declaring outliers. This is done via the inclusion of parameters that allow for manipulation of the size and shape of the acceptance region. Selection of appropriate values for these parameters, however, is not straightforward. This article presents a tool that can be useful in the specification of parameters for this method. We will also show an application of the method to the General Index of Financial Information, a census of Canadian corporate financial information.

KEY WORDS: Outlier detection; Periodic surveys.

## RÉSUMÉ

La détection des valeurs aberrantes dans les enquêtes auprès des entreprises peut être particulièrement difficile à cause de la très grande variation dans la taille des répondants. Une méthode très connue pour identifier les valeurs aberrantes dans les enquêtes-entreprise périodiques a été créée par Hidiroglou et Berthelot (1986). La force de cette méthode est qu'elle nous permet d'inclure la taille de l'unité comme un facteur important dans l'identification des valeurs aberrantes. Ceci est fait par l'inclusion de paramètres qui nous permettent d'ajuster la dimension et la forme de la région d'acceptation. Cependant, la sélection des valeurs appropriées pour ces paramètres n'est pas évidente. Le présent document propose un outil qui peut être utile dans la spécification des paramètres pour cette méthode. Nous montrons aussi une application de la méthode à l'Index général des renseignements financiers, un recensement de renseignements financiers d'entreprises canadiennes.

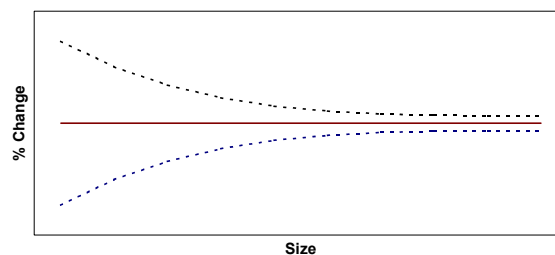
MOTS CLÉS : Détection des valeurs aberrantes; enquêtes périodiques.

## 1. INTRODUCTION

### 1.1 Background and Objective

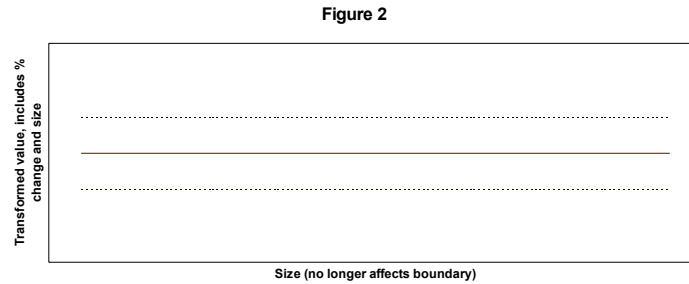
The Hidiroglou-Berthelot method of outlier detection is commonly used for periodic business surveys at Statistics Canada. The main idea behind this method is to have an acceptance boundary that varies according to the size of a unit. The larger the size of the unit, the smaller the percent change we allow from one period to the next (see Fig.1).

Figure 1



<sup>1</sup> Richard Belcher (Richard.Belcher@statcan.ca), Business Survey Methods Division, 11<sup>th</sup> Floor, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario K1A 0T6,

In practice, however, we create transformed values for each record that include a factor for percent change, and a factor for size that is adjusted by our parameters. We then set a fixed boundary that is exactly the same for all records, regardless of size. So, it is in the transformed data points themselves that we factor in the importance of the size of a record, not in the boundary (see Fig. 2).



Because the outlier detection is done using these transformed values, the boundary is neither expressed as an allowed percentage of change, nor in dollars, but as a combination of the two. It can therefore be difficult to assess whether or not the calculated boundaries are appropriate.

We would like to find a way to express these outlier boundaries in the units of the variable of interest. This would greatly simplify the analysis of our results, and would aid in the specification and adjustment of values for the parameters that control the method.

## 2. REVIEW OF THE HIDIROGLOU-BERTHELOT METHOD

### 2.1 Advantages

There are a number of reasons to favour this method for periodic business surveys. First, the size of the unit is an important factor in declaring outliers. Even better, we can control the level of importance assigned to size, with a parameter ( $U$ ) that we will examine later. Additionally, the method automatically adjusts in consideration of the distribution of the ratios, it finds both large and small outliers (in terms of our variable of interest), and it finds outliers on both tails of the distribution (positive and negative changes).

### 2.2 Calculating Effects

As set out in Hidiroglou and Berthelot (1986), outliers will be those observations whose ratio (or trend) from the previous period to the current one differs significantly from the corresponding overall trend of other observation belonging to the same subset of the population.

For each observation, we calculate a test statistic, which we call an effect, denoted  $E_i$ . We start by defining  $r_i$  as the ratio for observation  $i$  of variable  $x$  at time  $t$  to variable  $x$  at time  $t-1$ .

$$r_i = x_i(t) / x_i(t - 1)$$

That is, for a given unit, the ratio of its current period value to its previous period value. Note that the distribution of  $r_i$  values is non-symmetric, and centred at the median ratio,  $r_M$ , the value of which is very often close to 1.

Now, we transform this ratio to get a value that is easier to work with. We call the transformed value  $s_i$ , and define it as:

$$s_i = \begin{cases} 1 - r_M / r_i, & \text{if } 0 < r_i < r_M \\ r_i / r_M - 1, & \text{if } r_i \geq r_M \end{cases}$$

The important thing to note here is that the distribution of the  $s_i$  values is centred at 0, and is symmetric.

One final transformation is required to get our desired  $E_i$  value. We add a size term, which for any observation is the maximum of the observed values from time  $t$  and  $t-1$ , to the power of  $U$ . We define  $E_i$ , then, as:

$$E_i = s_i * \{ \text{Max} (x_i(t), x_i(t-1)) \}^U$$

The value of the  $U$  parameter controls the shape of our acceptance region by emphasizing or de-emphasizing the importance of size in our  $E_i$  values. We can acceptably set the value of  $U$  between 0 and 1. If we set it to 0, the size term goes to 1, and the boundary is uniform regardless of size. Conversely, if we set it to 1, the size term will overpower the  $s_i$  term, and we will find that our outlier population will simply consist of all our largest units. In practice, then, we will generally use values in the range of 0.3 to 0.5.

### 2.3 Defining a Boundary

With an effect calculated for each observation, the next step is to define the boundaries to which we will compare them.

First, we define  $d_{Q1}$  as essentially the lower half of the inter-quartile distance of the effects. Technically, it is the maximum of that value and a second term, but that is only to protect us from strange distributions. The actual formula is  $d_{Q1} = \text{Max} (E_M - E_{Q1}, |AE_M|)$ . If we think about this as basically a portion of the inter-quartile distance, then  $d_{Q3}$  represents the other portion and is defined as:  $d_{Q3} = \text{Max} (E_{Q3} - E_M, |AE_M|)$ . Note the new parameter  $A$ , which is generally set to 0.05 as this has worked well in practice. Our acceptance interval will now be defined by:

$$(E_M - Cd_{Q1}, E_M + Cd_{Q3}).$$

Note how the third and final parameter used in the method,  $C$ , will control the overall width of our acceptance region.

### 2.4 Challenge

When using this method, how can we best determine appropriate values for parameters  $U$  and  $C$ ? How can we decide whether the calculated outlier boundaries are reasonable? In practice, the acceptance intervals are fixed, and do not vary with the size of the unit. Rather, the size term is embedded into our  $E_i$  value for each observation. As a result, we get a fixed acceptance interval with no units, an example of which might be (-1600, 2000). Clearly, it is not straightforward to analyse and interpret an interval of this type.

## 3. INTERPRETING THE BOUNDARIES

### 3.1 Concept

Our goal is to find a way to express the acceptance interval in the appropriate units, for example, in dollars. In so doing, we would also like to clearly show how the intervals will vary with size. Boundaries of this form will be straightforward to interpret. This will be helpful if we wish to present the proposed boundaries to non-statisticians, such as subject matter experts, who will then be able to provide the most useful feedback on their suitability.

By definition, an observation lying directly on the lower outlier boundary would have  $E_i = E_M - Cd_{Q1}$ . Similarly, an observation on the upper outlier boundary would have:  $E_i = E_M + Cd_{Q3}$ . Using these facts, and the formula for  $E_i$ , we can find formulas that express  $x_i(t)$  as a function of  $x_i(t-1)$  for points on the boundaries. This will allow us to choose a desired value for  $x$  at time  $t-1$ , and to calculate the exact acceptance interval for the unit at time  $t$  for a given set of parameters.

### 3.2 Lower Boundary

To create our general formula for the lower boundary, we start by looking at the formula for the effects  $E_i$ . Since we are limiting ourselves to looking at the lower boundary, we know that the unit is decreasing in value from time  $t-1$  to time  $t$ . In other words, we know that  $x_i(t)$  is less than  $x_i(t-1)$ .

The expression for the effects is  $E_i = s_i \{ \text{Max} (x_i(t), x_i(t-1)) \}^U$ . Since we know that  $x_i(t)$  is less than  $x_i(t-1)$  for points on the lower boundary, we have  $\text{Max} (x_i(t), x_i(t-1)) = x_i(t-1)$ . This simplifies the effect equation to:

$$E_i = s_i (x_i(t-1))^U.$$

We know also that  $s_i = 1 - r_M / r_i$  when  $0 < r_i < r_M$ . Again, because we are looking for a lower bound, we know that this condition is true, and we can substitute for  $s_i$  in the effect equation, giving us

$$E_i = (1 - r_M / r_i)(x_i(t-1))^U.$$

Finally, we know that  $r_i = x_i(t) / x_i(t-1)$ , and we can again substitute into our equation to arrive at the following, which expresses effects only in terms of our observed values, plus constants:

$$E_i = (1 - r_M / (x_i(t) / x_i(t-1)))(x_i(t-1))^U.$$

Now, the lower bound on the effects is given by  $E_M - Cd_{Q1}$ , where  $d_{Q1} = \text{Max}(E_M - E_{Q1}, |AE_M|)$ .

This first value is a constant for a given data set, and our chosen parameters. It does not depend in any way on the theoretical points on the boundary that we are trying to locate. Since it has no relationship to the  $x_i(t)$  or  $x_i(t-1)$  values of these theoretical points, we can simply replace the expression by a single term, called  $LB (=E_M - Cd_{Q1})$ .

Equating our new expression for the effects to the expression for the lower boundary gives us the following relationship, which is held by any point lying on the lower boundary:

$$(1 - r_M / (x_i(t) / x_i(t-1)))(x_i(t-1))^U = LB.$$

Solving this equation for  $x_i(t)$  will give us an expression for the lower boundary (in the units of interest) for an observation at time  $t$ , given the value for the same observation at time  $t-1$ . The solved equation is the following:

$$x_i(t) = (r_M)(x_i(t-1)) / (1 - (LB / (x_i(t-1))^U))$$

Using this equation, it is possible to choose a value for  $x_i(t-1)$ , and then calculate the lowest  $x_i(t)$  that would be allowed without flagging the record as an outlier. Some examples will be seen in the table in section 4.1.

### 3.3 Upper Boundary

The upper boundary proves slightly more difficult to find. In searching for the upper boundary, we let  $x_i(t-1)$  be our chosen value, and  $x_i(t)$  be the value that would put the record exactly on the upper outlier boundary. As before, we have  $E_i = s_i \{ \text{Max}(x_i(t), x_i(t-1)) \}^U$ , but this time, we know that  $\text{Max}(x_i(t), x_i(t-1)) = x_i(t)$ . So, we can say that

$$E_i = s_i(x_i(t))^U.$$

We also know that  $s_i = r_i / r_M - 1$ , if  $r_i > r_M$ . Since we are looking for the upper bound, we know that this condition is true, and we substitute for  $s_i$ , giving us

$$E_i = (r_i / r_M - 1)(x_i(t))^U.$$

Finally, we know that  $r_i = x_i(t) / x_i(t-1)$ , and we can substitute into our equation to arrive at the following, which expresses effects only in terms of our observed values, and constants:

$$E_i = (((x_i(t) / x_i(t-1)) / r_M) - 1)(x_i(t))^U$$

The upper bound on the effects is given by  $E_M + Cd_{Q3}$ , where  $d_{Q3} = \text{Max}(E_{Q3} - E_M, |AE_M|)$

Like the lower bound,  $LB$ , this value is a constant, and does not depend on the  $x_i$  values of our theoretical points on the boundary. We replace the expression by a single term,  $UB (=E_M + Cd_{Q3})$ .

Equating these two expressions gives us

$$(((x_i(t) / x_i(t - 1)) / r_M) - 1)(x_i(t))^U = UB.$$

Unfortunately, this expression proves impossible to solve explicitly for  $x_i(t)$ . However, we can use Newton's method to find an approximate solution. This is an iterative method that allows us to approximate the zeroes of a differentiable function, and is described in Adams (1986). Using this method allows us to avoid solving the upper boundary equation directly. Instead, we simply find an approximate solution to the following equivalent expression

$$(((x_i(t) / x_i(t - 1)) / r_M) - 1)(x_i(t))^U - UB = 0.$$

Since we will need to differentiate the function to use Newton's method, we perform some algebra to arrive at

$$((((x_i(t))^{U+1} / x_i(t - 1)) / r_M) - (x_i(t))^U) - UB = 0.$$

And then, multiplying through by  $r_M$ , and then by  $x_i(t - 1)$ , we get the following easily differentiable function, which we will now call  $f(x_i(t))$ :

$$f(x_i(t)) = (x_i(t))^{U+1} - (x_i(t - 1))(r_M)(x_i(t))^U - (UB)(x_i(t - 1))(r_M) = 0.$$

This function, differentiated, is

$$f'(x_i(t)) = (U + 1)(x_i(t))^U - (U)(x_i(t - 1))(r_M)(x_i(t))^{U-1}.$$

The iteration procedure is  $x_{n+1} = x_n - (f(x_n) / f'(x_n))$ . Our iteration procedure then, becomes

$$x_{n+1} = x_n - ((x_n)^{U+1} - (x_i(t - 1))(r_M)(x_n)^U - (UB)(x_i(t - 1))(r_M) / (U + 1)(x_n)^U - (U)(x_i(t - 1))(r_M)(x_n)^{U-1}),$$

where  $x_n$  represents the approximate solution for  $x_i(t)$ , at the  $n^{\text{th}}$  iteration. To start the process, we set  $x_0 = x_i(t - 1)$ , and then iterate until we achieve the desired level of precision. This method is extremely efficient. In tests with the GIFI data seen in the next section, we can approximate the boundary to within \$1 (by processing until we have  $x_{n+1} - x_n \leq 1$ ) in approximately 5 or 6 iterations.

## 4. APPLICATION OF THE METHOD

### 4.1 The General Index of Financial Information (GIFI)

We will now look at an example of an application of the method to the General Index of Financial Information, a census of Canadian corporate tax information. The following table, created using the formulas developed in section 3, shows the acceptance intervals for 1999 assets for various chosen levels of 1998 assets. For example, a unit with \$1 million in assets in 1998 would be allowed to report assets between \$153,391 and \$5,407,050 in 1999 without being flagged as an outlier.

GIFI outlier detection parameter check				
Bounds on 1999 assets given 1998 assets (approximate)				
	Lower Bound		Upper Bound	
% change	Assets 1999	Assets 1998	Assets 1999	% change
	$x_i(t)$	$x_i(t - 1)$	$x_i(t)$	
-100%	0	100	6,558	6458%
-99%	11	1,000	34,313	3331%
-97%	279	10,000	181,229	1712%
-93%	6,723	100,000	974,128	874%
-85%	153,391	1,000,000	5,407,050	441%
-69%	3,131,385	10,000,000	31,742,850	217%
-47%	53,491,866	100,000,000	203,699,200	104%
-26%	744,969,606	1,000,000,000	1,473,408,800	47%
-12%	8,830,086,166	10,000,000,000	12,076,824,000	21%
-5%	95,333,334,556	100,000,000,000	108,985,858,000	9%

These intervals depend on the distribution of our data, but also on our chosen parameters,  $U$  and  $C$ . The  $C$  parameter allows us to narrow or widen the intervals across all size levels. Meanwhile, the  $U$  parameter allows us to change the shape of our acceptance region by narrowing the boundary for large units while tightening it for small ones, or vice versa.

Tables of this type proved very useful in the process of editing GIFI data in 2000. A number of tables were produced showing the effects of different proposed sets of parameters, and subject matter experts were asked to choose the set of parameters that seemed the most appropriate based on the resulting acceptance intervals.

Another table that can be helpful in the assessment of a set of parameters for this method follows:

Assets 1998	low outliers	high outliers	outlier %
0-100	0	1554	38.05
101-1,000	27	763	12.20
1,000-10,000	227	651	2.94
10,000-100,000	1818	629	1.29
100,000-1,000,000	2357	533	0.92
1,000,000-10,000,000	1254	301	1.77
10,000,000-100,000,000	738	155	10.83
100,000,000-1,000,000,000	198	66	25.71

This table takes the outliers identified with a particular set of parameters and groups them by size of assets in 1998. That is, by size of  $x_i(t - I)$ . In this example, we can see that the percentage of outliers being found among the very smallest and very largest units is quite high. For example, we flagged over 38% of the units with assets between \$0 and \$100 in 1998 as outliers in 1999. However, through the bulk of our distribution, for units with assets between \$1,000 and \$10,000,000, we are flagging 1% to 3% of the units as outliers, which seems reasonable. If we found that this was an unacceptably high or low percentage, we could simply make an adjustment to our  $C$  parameter to increase or decrease the number of outliers across all size categories.

In this case, note that the pattern of the outlier percentages is somewhat symmetric with respect to size. If, for example, we saw instead that our outlier percentages were generally much higher for the large units than for the small units, this could be seen as an indication that an adjustment to our  $U$  parameter would be in order.

## 5. CONCLUSION

### 5.1 Conclusion

For periodic business surveys, the Hidiriglou-Berthelot method is a very desirable method for detecting outliers. The main challenge in applying this method comes in the selection of appropriate values for the parameters, due to the inherent difficulties in analyzing the resulting outlier boundaries. In this paper, we have shown a way to convert the boundaries into a format that is directly understandable. It is hoped that users of the method will find that the use of analytical tables such as those presented in section 4, made possible by the formulas presented in section 3, can be extremely helpful in simplifying this process.

### 5.2 Acknowledgments

The author would like to thank Chantal Grondin and Nathalie Hamel for their help in reviewing the manuscript, and Mike Hidiriglou for his support and assistance, particularly in the selection of the iterative method and the development of the formulas in section 3.4.

## REFERENCES

Adams, Robert A. (1986). *Single-Variable Calculus* (2nd. ed.). Addison-Wesley.

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