Evaluating the Hidiroglou-Berthelot Method for Survey Data Collected by the US EIA

Michael J. Winkler[1], Preston A. McDowney[1], Makayla S. Cowles[1], Orhan Yildiz[1], Caitlin I. Steiner[2] and Pushpal K. Mukhopadhyay[1]

[1]Office of Statistical Methods and Research, Energy Information Administration, Washington, DC 20585

[2] U. S. Department of State, Washington, DC 20585[2]

Abstract

The U.S. Energy Information Administration (EIA) conducts weekly, monthly, quarterly, and annual establishment surveys to promote the understanding of energy and its interaction with the economy and the environment. These establishment surveys typically measure characteristics of skewed business populations based on volumetric totals. EIA currently relies on traditional editing methods, such as range and ratio edits, based on fixed edit. For selected surveys conducted by EIA, we evaluate potential improvements to our editing methods by comparing the results from our current edits to those using the statistical ratio edit proposed by M.A. Hidiroglou and J.-M. Berthelot. The Hidiroglou-Berthelot edit uses a data-driven approach to determine edit parameters and is based on a transformed edit statistic that detects outliers at both tails of its distribution and takes a unit's size into account.

Key Words: statistical edit, establishment survey, skewed distribution, energy statistics

1. Background

The Energy Information Administration (EIA) is a principal federal statistical agency that conducts numerous establishment surveys for its energy supply and demand data collection programs. To ensure the quality of surveys, EIA uses several statistical editing techniques for outlier identification. This paper will evaluate the Hidiroglou-Berthelot method (HB method) on selected surveys conducted by EIA. The HB method is a flexible ratio method for outlier identification, not currently used by EIA.

The HB method was first proposed by Hidiroglou and Berthelot (1986) to improve on some of the common issues in editing data from establishment surveys conducted by Statistics Canada. Information collected in establishment surveys such as revenue, sales, or production data are typically dominated by a few large establishments. Data editing methods that do not account for the size of the establishment commonly flag a large number of smaller establishments as potential outliers while missing some of the larger establishments. The HB method modifies the traditional edit methods by using the size of the establishment.

[1] The analysis and conclusions expressed in this paper are those of the authors and do not represent the official position of the U.S. Energy Information Administration (EIA) or the U. S. Department of Energy (DOE).

[2] Dr. Steiner conducted this research when she was a Mathematical Statistician at the U.S. Energy Information Administration, though she currently is a Data Scientist at the U.S. Department of State.

The two most commonly used edit methods at EIA are the range and ratio edits. The range edit method flags all observations that fall outside a predetermined range. Ratio edit methodologies take the ratio of two correlated variables and flags all observations in which the ratio falls outside a range that is based on the distribution of the ratio. See De Waal (2011) for commonly used statistical data editing methods. Compared to range edits and other ratio edits, the HB method is data-driven and uses the distribution of the transformed ratios when determining a boundary. Ratio methods can improve on the range edits since they rely on the distribution of the data; however, there are drawbacks for certain ratio methods as well. First, some ratio methods will identify outliers better in one tail or the other but not both. Other ratio methods can have a masking effect, meaning the method's large outliers can hide smaller outliers. Alternatively, a final phenomenon is the size masking effect, where too many small outliers can be identified. Due to the skewed nature of establishment surveys, all of these can present issues. To correct these issues, the HB method begins by taking the ratio of two correlated survey items. Next, we apply two transformations. The first transformation makes the ratios symmetric and the second transformation accounts for the size of the possible outlier. Finally, using the transformed ratios, an analyst can create upper and lower boundaries to an acceptance region.

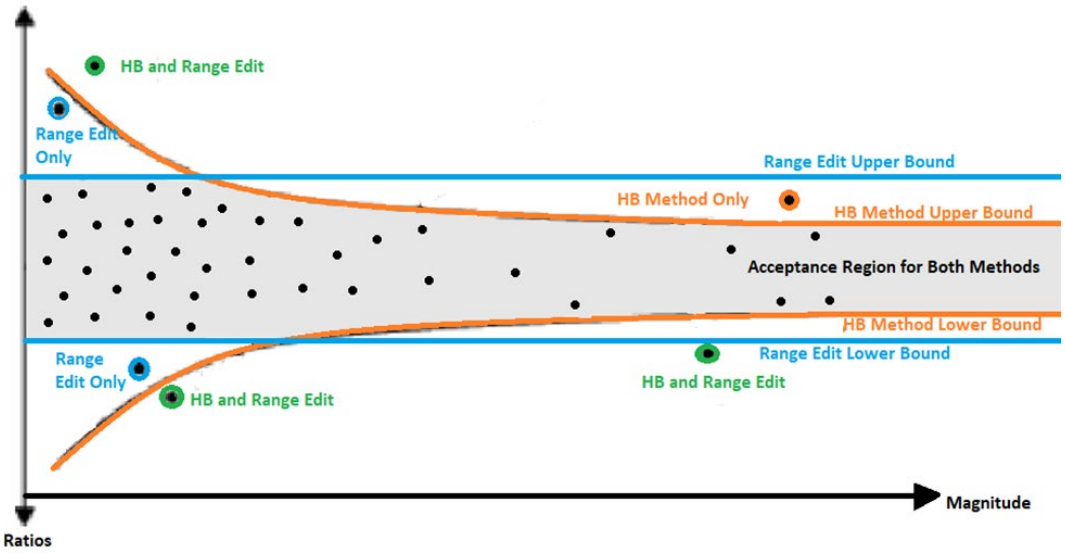

Figure 1. Upper and lower bounds from the HB and range edit. The HB bounds account for the magnitude of the observations.

Figure 1 is a modified version of Figure 4 found in Hidiroglou and Lavallée (2009), showing how the HB method compares to range edits. Figure 1 shows that the HB method tightens the acceptance region as the magnitude (maximum of the numerator and denominator or the ratio) increases. This means the magnitude of a potential outlier is emphasized when determining if it is an outlier. Additionally, because the boundaries are symmetric, the outliers are detected equally well in both tails. The curvature also eliminates both the masking and size masking effects, making the method ideal for skewed establishment survey data.

1.1 Applications from Other Countries and Agencies

The HB method is widely used by U.S. and international statistical agencies for editing data collected from establishment surveys. For example, the Annual Capital Expenditures Survey: Actual Preliminary Estimate and Intentions (CAPEX) survey conducted by Statistics Canada for collecting capital and repair expenditures data uses the HB method for outlier detection (https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=2803).

The U.S. Census Bureau's Monthly Retail Trade Survey (MRTS) and Annual Survey of Government Finances (ASGF) also use the HB method. The MRTS collects information on monthly

retail trade sales and inventories. The MRTS uses the HB method for identifying units with unusual month-to-month trends with significant market share for a particular industry. The HB method performs well compared to the ratio and month-to-month trend edits by identifying a similar number of mid-level outliers but a fewer number of high-level outliers (Hunt et. al, 2002). The ASGF collect annual data on revenue, expenditures, debt, and assets of state and local government. The survey used ratio and historical edit methods in the past, but these methods were not possible to implement because of a questionnaire redesign in 2005 requiring some data items to be consolidated or split. Cornett et al. (2006) successfully applied the HB method but noted that the HB method could perform better if it is applied at the state level instead of the national level. At the national level the bounds created by the HB method were largely influenced by data from a few large states such as New York.

1.2 The HB Method

Like other ratio edit methods, the HB method creates a vector of ratios, r_k, by dividing each of the current period's responses, y_k, by either a final version of the prior period's responses or the current value of a correlated variable, x_k, as follows:

$$r_k = \frac{y_k}{x_k}, k = 1,2, \dots, n$$

where n is the number of observation units.
To detect outliers in both tails, the HB method applies a centering transformation producing non-negative and symmetric transformed ratios, s_k , defined by

$$s_k = \begin{cases} 1 - \frac{r_{Q2}}{r_k}, if\ 0 < r_k < r_{Q2} \\ \frac{r_k}{r_{Q2}} - 1, \quad otherwise \end{cases}$$

where r_{Q2} is the median of $\{r_k\}$. Then, to account for the size of the observation, the HB method creates an effects vector, e_k, by scaling the symmetric ratios as follows:

$$e_k = s_k \max(x_k, y_k)^u, \text{ where } 0 < u < 1$$

The parameter u controls for the importance associated with the magnitude of the data by scaling the maximum value. More information on parameters and calibrating the HB method are discussed in Belcher (2003). To help with the intuition, a u close to 0 will approximate range edits, while u close to 1 will create a more curved boundary which will tighten the boundary for high magnitude points while loosening the boundary around lower magnitude observations. Multiple sources (Hidirgolou 1986 and Hunt 1999) in the literature indicate that 0.5 is the most common and often provides a reasonable edit boundary.

Finally, the upper and lower HB edit bounds are given by:

$$\text{Lower Bound: } e_{Q2} - c \max\{e_{Q2} - e_{Q1}, |ae_{Q2}|\}$$
$$\text{Upper Bound: } e_{Q2} + c \max\{e_{Q3} - e_{Q2}, |ae_{Q2}|\}$$

Where e_{Q1}, e_{Q2}, and e_{Q3} are the first, second, and third quartiles of $\{e_k\}$, respectively. The parameter a, usually set to 0.05, ensures that the boundaries are not arbitrarily small. This is a problem that arises when the effects $\{e_k\}$ are clustered around the median with modest deviations.
The value of c is set via iterative exploration and discussion with subject matter analysts.

Although the first and third quartiles of are used in constructing the HB bounds, they do not need to be. If more than ¼ of the ratios are the same or there are too many false outliers, this number can be changed. Hidirgolou and Emond (2018) recommend using the 10th and 90th percentiles in such cases.

The HB outlier-detection region resembles a confidence interval that is centered around the median, e_{Q2}, and has a width created using the distance of the first or third quartile to the median. Below is a reference table of the parameters, what the parameter controls, and a typical value of the parameter.

Table1: HB Parameters

Parameter Name	What the Parameter Controls	Common Value(s)
u	Controls the curve of the final boundaries	0.5
Quantiles	Allows boundaries to be calculated based on quantiles from $\{e_k\}$	0.25 and 0.75
a	Ensures the upper and lower bounds are not arbitrarily close to the median (e_{Q2})	0.05
c	Controls the width of the acceptance region	4

2. Applications to EIA Surveys

We applied the HB edit method in two establishment surveys conducted by EIA: The Monthly Biofuels, Fuel Oxygenates, Isooctane, and Isooctene Report (Section 2.1) and The Monthly Electric Power Industry Report (Section 2.2).

2.1 Monthly Gross Production of Denatured Fuel Ethanol

The Monthly Biofuels, Fuel Oxygenates, Isooctane, and Isooctene Report (EIA-819) is a survey that collects information on the production capacity of fuel alcohol, biodiesel, renewable diesel fuel, heating oil, jet fuel naphtha, gasoline and renewable fuels, isooctane, isooctene, and fuel oxygenates. It collects information about the beginning stocks, receipts, production, inputs, shipments, plant use and losses, ending stocks of biofuels and oxygenates at a plant level. Additionally, consumption of fuels and feedstocks are collected.

For this research, we focused on the responses from December 2019 of Gross Production for Denatured Fuel Ethanol (GPDFE). Figure 2 displays the density plot and Table 2 shows some summary statistics for GPDFE. Like data from most establishment surveys, the estimated density for GPDFE is skewed.

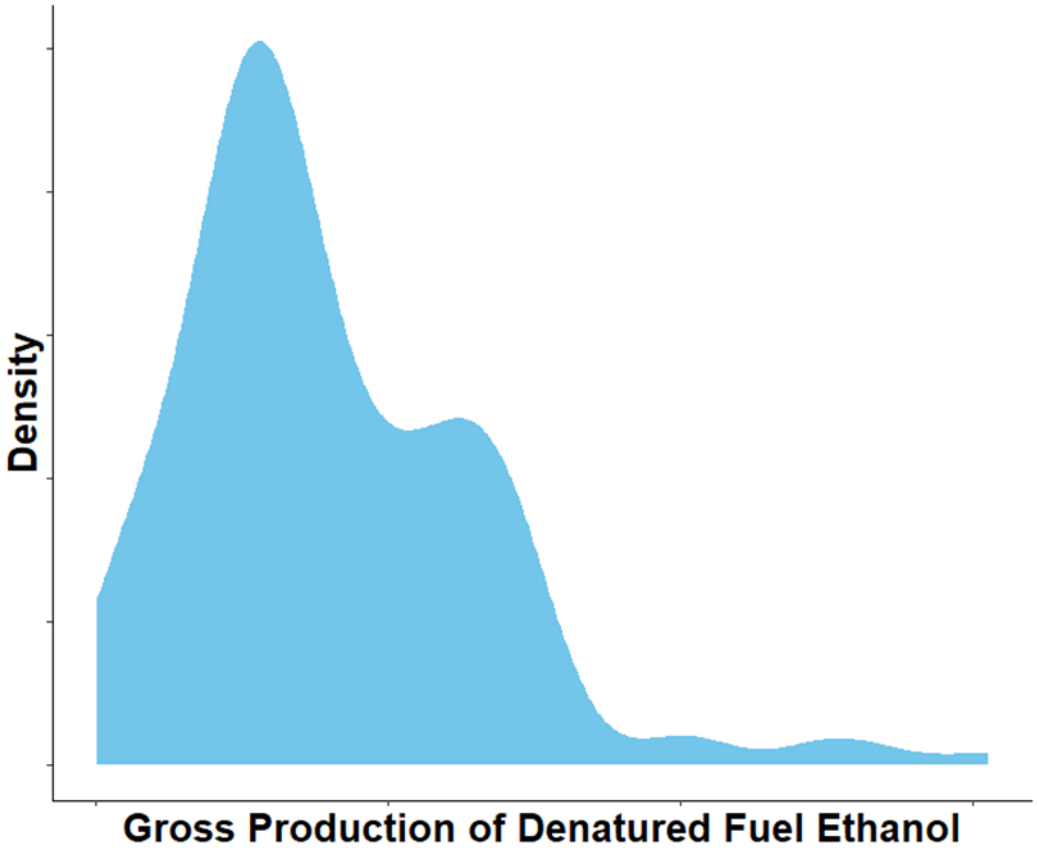

Figure: 2: Density for gross production for denatured fuel ethanol for December 2019.

Table: 2 Gross Production for Denatured Fuel Ethanol Summary Statistics

Observations	Mean	Median	Skewness
187	158	131	1.27

We applied the HB method for editing the reported December 2019 GPDFE using the previous period's finalized GPDFE (November 2019). These two variables are highly correlated with a Pearson correlation coefficient of 0.95. Using the initial values for the tuning parameters (Table 1), the HB method flagged approximately 12% of observations for editing/review. Because of the high number of flagged observations and exploring the boundary plots, we decided to tune the method. Increasing the *c* parameter (width) to 10 and the *u* parameter (curvature) to 0.75, approximately flagged 5% of the observations. Looking at Figure 3 and the other diagnostics, these parameters appear to be much more reasonable.

Figure 3 shows a scatter plot for GPDFE 2019 November and GPDFE 2019 December. Eleven observations are flagged as outliers out of a total of 175 observations. As expected, the HB method flagged observations away from the 45-degree line as outliers, it detects outliers on both tails, and accounts for the size of the observations.

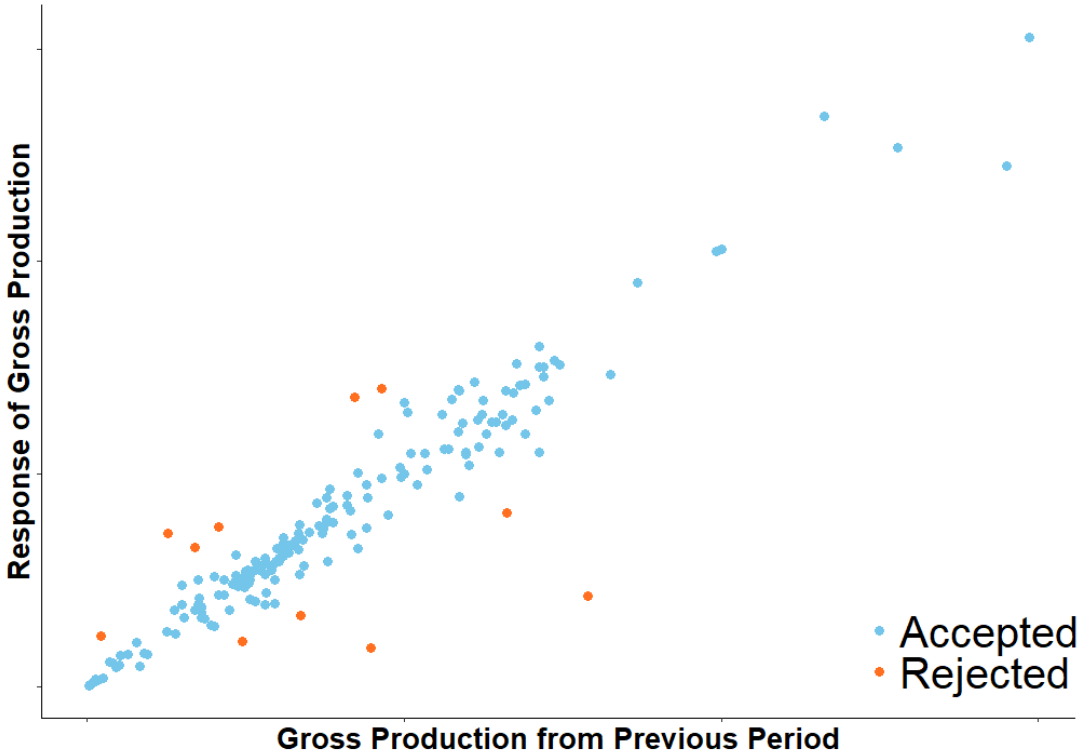

Figure 3: Gross production for Denatured Fuel Ethanol for December 2019 vs November 2019. The HB method identifies the orange observations as potential outliers.

2.2 Residential Electricity Revenue

We applied the HB method to a key survey item from the Monthly Electric Power Industry Report (EIA-861M) as well. The EIA 861M is a survey that collects information from utilities and nonutility companies that sell or deliver electric power to end users, including electric utilities, energy service providers, and distribution companies. Data collected include retail sales and revenue for all end-use sectors (residential, commercial, industrial and transportation). For this research, we focused on responses from October 2021 residential revenue. Figure 4 shows the density plot and Table 3 shows some summary statistics for October 2021 residential revenue. Both demonstrate the skewness of the residential revenue.

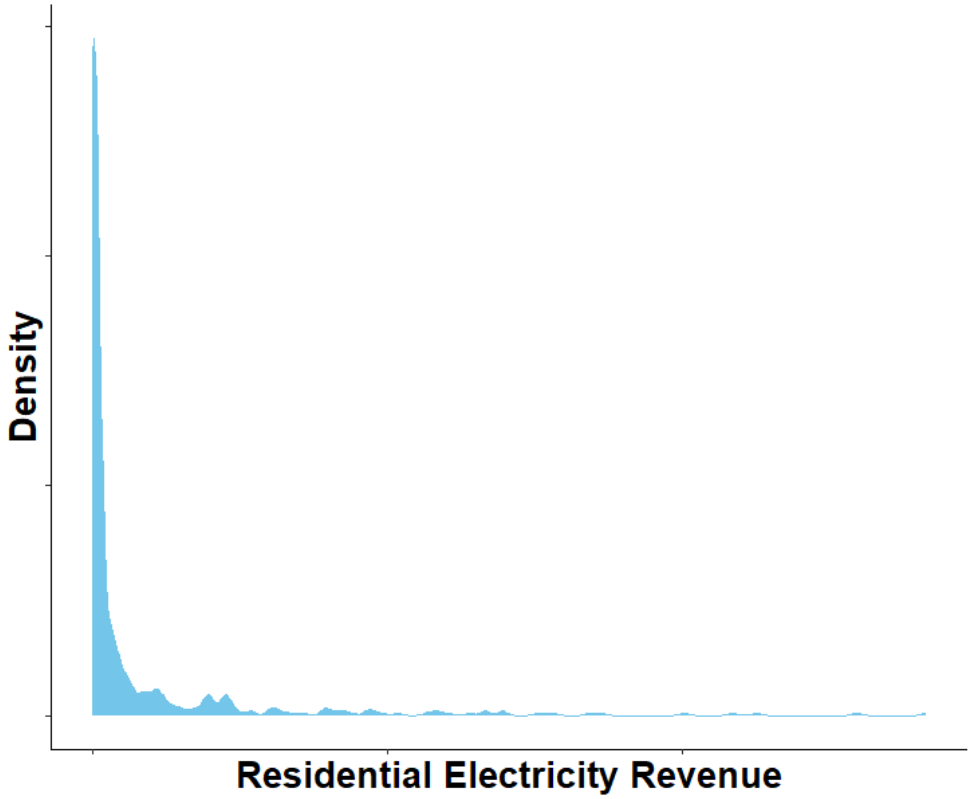

Figure: 4 Density for residential electricity revenue for October 2021.

Table 3: Residential Revenue Summary Statistics

Observations	Mean	Median	Skewness
497	14301.54	789.45	4.02

The current range edit has an acceptance region of ±35% of the rolling average maximum and minimum response. We applied the HB method using two different correlated variables to compare to the range edits. The first variable correlated with October 2021 residential revenue that we researched was October 2020 residential revenue. Because electricity revenue has a high degree of seasonality, we used finalized responses from October 2020 instead of those from September 2021.The revenue from October 2021 and October 2020 are highly correlated with a Pearson correlation coefficient of 0.99.

For this data, $c=40$ and $u=0.5$ provide reasonable bounds for the HB method. Table 4 displays the number of observations accepted and rejected by the HB edit and range edit methods. The HB method flagged considerably fewer observations (18) compared to the range edit (38). There are 13 observations that are flagged by both the HB method and range edit.

Table 4: The number of observations accepted and rejected by the current range edit and HB edit.

	HB Method	
Range +/- 35%	Accepted	Rejected
Accepted	454	5
Rejected	25	13

Figure 5 displays a scatter plot for the log of 2021 revenue and the log of 2020 revenue. The gray rectangles are accepted by both edit methods. All observations that are flagged by the HB method are concentrated at the top of the distribution, while the observations that are flagged by the range edit are concentrated at the bottom of the distribution. This is expected from the HB method, because this method uses the size of the observation to determine larger and more influential outliers. This allows analysts to prioritize their time and for respondents to have less burden overall.

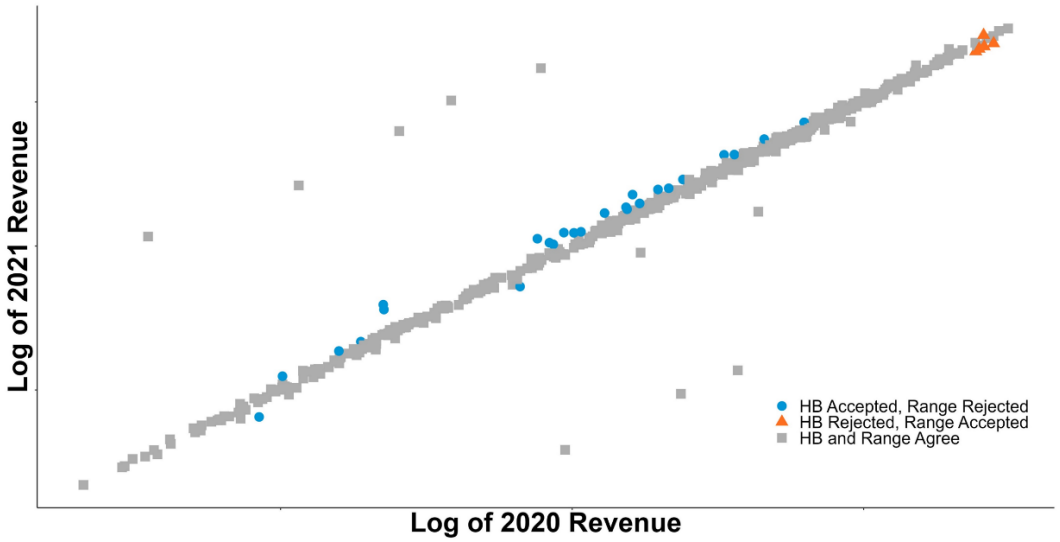

Figure 5: Logarithm of 2020 and 2021 Electricity Revenue. Observations identified by orange triangles are rejected by the HB edit but accepted by the range edit. Observation identified by blue circles are accepted by the HB edit but rejected by the range edit.

For October 2021 revenue, the first correlated variable was October 2020 residential revenue. Because electricity revenue has a high degree of seasonality, we used finalized responses from October 2020. The revenue from October 2021 and October 2020 are highly correlated with a Pearson correlation coefficient of 0.99.

The second variable correlated with October 2021 residential revenue that we researched was October 2021 sales, since sales are also expected to be highly interrelated with revenue. October 2021 revenues and sales has a Pearson correlation coefficient of 0.86.

We used c=40 and u=0.5 as before. Figure6 displays a scatter plot for the log of 2021 sales and the log of 2021 revenue.

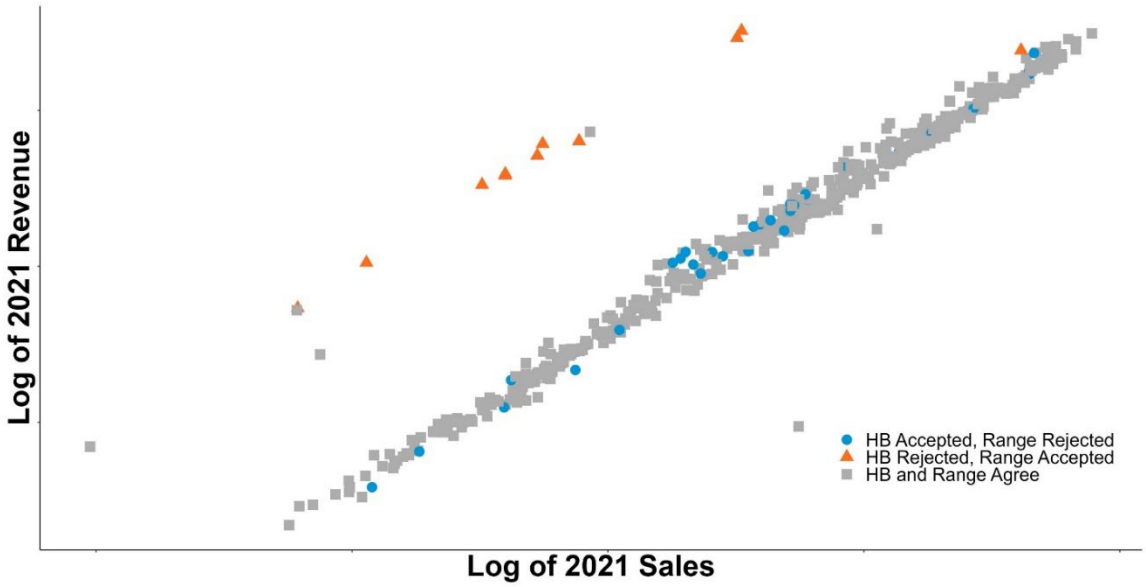

Figure 6: Logarithm of 2021 Electricity Sales and 2021 Electricity Revenue. Observations identified by orange triangles are rejected by the HB method but accepted by the range edit. Observation identified by blue circles are accepted by the HB method but rejected by the range edit.

Using sales instead of the previous period's revenue identifies additional potential outliers. Most of these additional flagged observations seem likely to be outliers and should be checked by subject matter experts. The HB method can identify additional outliers by using different variables that are more or less correlated with the analysis variable.

3. R Shiny App

3.1 Development

We designed the R Shiny app to allow analysts to explore the data and the HB method by accessing relevant statistics and data visualizations easily, changing variables and parameters quickly, and showing all relevant diagnostics. We developed the app for EIA energy analysts to use and compare the HB edit method graphically with other range and ratio edits. We developed a graphical user interface (GUI) that the analysts can use to fine tune the parameters for the HB method. No programming or statistical knowledge is required to use the application.

First, an analyst loads a CSV or Excel file and specifies the survey variable to edit, and a correlated variable being used to edit. The analyst can then change the default parameter values. The application will display the results in five pages; Input Data and Output Data (separated into Scatter Plot, Boundary Graph, Statistics, and Method); as described below.

3.2 Input

The data tab has two options. First is "Full Data" which is a fully sortable table with all the data the analyst provided, and an HB Flag indicator and the following diagnostic values defined in Section

1.2: *r* values, magnitude, *s* values, and *e* values. The second option is "Flagged Observations", which shows only the flagged observations with the flags and diagnostics.

3.3 Output

The output data is broken into three different sections: scatter plots, boundary graphs, and summary statistics.

The Scatter Plot page displays three outputs: a density plot of the original data being edited, a scatter plot in the original data scale, and a scatter plot in the log-transformed scale. The two scatter plots highlight the flagged observations. These plots can be screenshotted and zoomed, and values can be hidden or hovered over if an analyst would like to see something closer.

The Boundary Graph page displays a scatter plot showing the upper and lower HB boundaries in the symmetric ratio scale and a scatter plot showing the upper and lower HB boundaries in the ratio scale. This is similar to the plot displayed in Figure 1. The ratios (or symmetric ratios) are displayed on the y-axis and the magnitude of the ratios are displayed on the x-axis with the upper and lower HB boundaries from the parameters chosen. The ratio scale is easier to interpret but the symmetric scale makes it easier to visualize the HB boundaries for different parameters values.

The Statistics page displays two tables: a "Summary of Flags" and a "Summary of Statistics." The Summary of Flags shows the number of accepted observations, number of observations in below the lower HB bound, number of observations above the upper HB bound, and the number of excluded observations (due to the denominator being 0 or either value being missing). The Summary of Statistics provides detailed diagnostics such as the percentage of observations flagged, skewness of the data, correlation between the two variables, and a summary for the *e* values.

4. Summary

Data collected by EIA's establishment surveys are often skewed. The HB edit method works well for skewed data by identifying observations with high magnitude more often than observations with low magnitude. Reviewing observations with higher magnitude will increase data collection efficiency by prioritizing callbacks, reduce respondent burden, and improve overall data quality.
We applied the HB method in one natural gas and one electricity survey data conducted by EIA and found that the HB method works well for both surveys.
Finally, we developed an R Shiny application that analysts can use without programming knowledge and can fine tune the HB parameters by visualizing the flagged observations.

For future research, we would like to apply the HB method for editing micro data when the parameter of interest is a weighted mean and compare the HB method with regression-based techniques.

Acknowledgements

The authors gratefully acknowledge the contributions of EIA colleagues Samson Adeshiyan, Kenneth Pick, David Kinyon, Janice Lent, and Aidan Whitis to this research.

References

Belcher, R. (2003). Application of the Hidiroglou-Berthelot Method of Outlier Detection for Periodic Business Surveys. In *Proceedings of the Statistical Society of Canada Annual Meeting, June 2003*. Retrieved from https://ssc.ca/sites/default/files/survey/documents/SSC2003_R_Belcher.pdf

Czaplicki, N.M., Thompson, K.J. (2013) Outlier Detection for the Manufacturing, Mining, and Construction Sectors in the 2012 Economic Census. In *JSM* Proceedings, Survey Research Methods Section. Alexandria, VA: American Statistical Association. Retrieved from http://www.asasrms.org/Proceedings/y2013/files/309457_82715.pdf

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

Hunt, J.W., Johnson. J.S., King C.S. (1999). Detecting Outliers in the Monthly Retail Trade Survey using the Hidiroglou-Berthelot Method. In JSM Proceedings, Survey Research Methods Section. Alexandria, VA: American Statistical Association. Retrieved from https://api.semanticscholar.org/CorpusID:195851415

McKenzie, L.A.,Craig, T.L, Whitaker, J.N. (2009). Developing Macro Edits for the Census and Annual Survey of Governments Finance. In JSM Proceedings, Survey Research Methods Section. Alexandria, VA: American Statistical Association. Retrieved from http://www.asasrms.org/Proceedings/y2009/Files/303722.pdf

Washington, K., Burton J., Detlefsen R. (2010). "Frame Construction and Sample Maintenance for Current Economic Surveys." In JSM Proceedings, Survey Research Methods Section. Alexandria, VA: American Statistical Association. Retrieved from http://www.asasrms.org/Proceedings/y2010/Files/307773_59307.pdf

US Energy Information Administration. "Individual Survey Form Descriptions." Accessed September 1, 2023. https://www.eia.gov/survey/

Hidirogloy, M. A. (2018). Divisional Research - Business Survey Methods Division (BSMD). In Statistical Methodology Research and Development Program Annual Report 2017/2018, edited by Fortier, S., 18 – 20. Statistics Canda. Retrieved from https://www150.statcan.gc.ca/n1/en/pub/12-206-x/12-206-x2018001-eng.pdf?st=Ligudb2m.

Hidiroglou, M.A. Emond, N. (2018). "Modifying the Hidiroglou-Berthelot (HB) method". Unpublished note, Business Survey Methods Division, Statistics Canada, May 18, 2018.

Hidiroglou, M.A., Lavallée P. (2009). Sampling and Estimation in Business Surveys. In Sample Surveys: Design, Methods, and Applications, edited by D. Pfeffermann and C. R. Rao, 441–70. Vol. 29A of Handbook of Statistics. Amsterdam: North-Holland.

Wall, T. D. (2009). Statistical Data Editing. In Sample Surveys: Design, Methods, and Applications, edited by D. Pfeffermann and C. R. Rao, 187–213. Vol. 29A of Handbook of Statistics. Amsterdam: North-Holland.

Cornett, E., McLaughlin, J.F., Hogue, C.R. (2006). A Comparison of Two Ratio Edit Methods for the Annual Survey of Government Finances. In JSM Proceedings, Survey Research Methods Section. Alexandria, VA: American Statistical Association. Retrieved from http://www.asasrms.org/Proceedings/y2006/Files/JSM2006-000199.pdf