Application of Efficient Sampling with Prediction for Skewed Data

James R. Knaub, Jr.[1]
1890 Winterport Cluster, Reston, VA 20191-3624, USA

Abstract:
Official Statistics from establishment surveys are not only the basis for routine monitoring of markets and perhaps systems in general, they are essential for discovering problems for which innovative approaches may be needed. Statistical agencies remain the workhorses for regularly providing this information. For establishment surveys one needs to collect data efficiently, making an effort to reduce burden on small establishments, and reduce costs to the Government, while promoting accuracy of results in terms of total survey error. For over three decades, these demanding standards have been met using quasi-cutoff sampling and prediction, applied extensively to some of the repeated official US energy establishment surveys. This success may be duplicated for other applications where sample surveys occur periodically, and there is an occasional census produced for the same data items. Sometimes stratification is needed, but sometimes the borrowing of strength, as in small area estimation/prediction, may be used. References will be given to help avoid pitfalls. The idea is to encourage expanding this elegant approach to other applications. The material here is an expanded version of a poster for the 2022 Joint Statistical Meetings. **This is a tutorial/guide.** Appendices are written in stand-alone form.

Keywords:
Applications, Cost/Resources, Establishment surveys, Model-based approach, Official Statistics, Prediction Approach, Quasi-cutoff sampling, Skewed data, Total survey error

1.0 Introduction

The purpose of this paper is to demonstrate that simple modeling, used with quasi-cutoff sampling for multiple attribute establishment surveys, with generally highly skewed data, can perform very well for repeated surveys when the goal is Official Statistics for markets such as for energy statistics. There this has been very helpful for over 30 years, for tens of thousands or more aggregations, and should be helpful in other such cases of Official Statistics on production. The idea is to take advantage of the fact that the same data item in a previous census is generally an excellent size measure in a current sample. (See Cochran(1953), pages 205-206.) Multiple regression may sometimes be helpful. The US Energy Information Administration (EIA) has had various applications, and papers have been written on development there. An invited presentation in September 2017, Knaub(2017b), at the EIA, describes and references a good deal of what occurred there. Further, Knaub(2016), and Knaub(2022), respectively, provide information (1) on how to perform cutoff or quasi-cutoff sampling in general, and (2) on a comparison of variance estimators with regard to the prediction errors associated with the resulting predicted totals. However, to demonstrate that this also applies to other applications where there are repeated surveys, where the author has not had access, there was a need to find a source which is not confidential, and is available. In searching for such data, Toxic Release Inventory (TRI) data from the US Environmental Protection Agency (EPA) was brought to my attention, as noted later, below. These are annual data only, but for demonstration purposes, data for the year 2014 are used here as the predictor data for lead released from

[1] https://www.researchgate.net/profile/James-Knaub

paper manufacturing, and samples are drawn from year 2015. (See the "Other Resources" section.) This is just cutoff sampling with one attribute (data item) but serves to show that this works, and not just for energy data. (See Appendix 1.) A second example was performed using populations of cities data from two sets of data, ten years apart, originally – it appears – found in Cochran(1953), page 70. In Cochran(1977), on page 93, as described below, William Cochran notes that these data behave similarly to business populations, and thus they are used here for another example.

Even though point estimates may be close, Lohr(2010), page 148, notes that for the probability-of-selection-based approach, it is the design of the sample which governs how we estimate variance, but in the prediction approach, it is the model which provides the method of estimating variance, and the model needs to be appropriate [for the population or segment of the population to which we wish to apply that model]. To her comments we will add that for the probability approach one would like to have a sample size large enough and selected in such a manner that there is a reasonable chance that the mix in the sample "represents" the population well, and that weights[2] reflect this reality. Perhaps a stratified random sample would be desired to achieve this. This is why it is important to know something about the population, as discussed later. One can use a model-assisted approach, as in Särndal, Swensson, and Wretman(1992), or work to combine probability and model approaches, as in Brewer(2002), to overcome this difficulty. However, in the case of highly skewed establishment surveys, especially those which are repeated and predictor data are provided from an occasional census to use with the frequent sample surveys, the accuracy/efficiency and resource effectiveness of a quasi-cutoff sample can be very impressive. For such production-type Official Statistics there may be numerous attributes and small populations which may mean a great many aggregate values are required throughout the year, at short intervals. Brewer(2013) noted that a purely model-based approach may be helpful for small samples. Ken Brewer had a broader interest, but did tell this author he recalled a cutoff sample on tailor shops. In an earlier work, Brewer(1963), as in Royall(1970), noted cutoff sampling with ratio modeling. Both were discussed in Cochran(1977), pages 158 and 160. Both Brewer and Royall subsequently looked into broader approaches. Valliant, Dorfman, and Royall(2000), Chambers and Clark(2012), and others have done other work with the model-based approach. (See Royall(1992) for the gist of the model-based approach.) However, for the current topic, thirty years of practice with applications in the tens of thousands or more at the US Energy Information Administration, and Knaub(2017b), indicate that for this kind of Official Statistics, under these circumstances, quasi-cutoff sampling with prediction is likely by far the best method, both in terms of accuracy/total survey error, and resources.

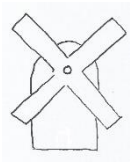
Earlier work on modeling at the EIA can be seen in Ahmed and Kirkendall(1981). Previously it occurred to this author that the choice of an unweighted regression in that paper might have come from scatterplots which presumably had data quality issues for the smallest respondents. On page 675 they say that "... one would expect the spread of the residuals ... to be independent of company size." However, Brewer(2002), mid-page 111 explains why this is not true. This is discussed in Knaub(2017c). Knaub(2021) shows that for more complex models this may sometimes appear to be true, though that is problematic, but heteroscedasticity should be found for such a simple

[2] Survey weights are meant here. With auxiliary data we consider calibration weights, and for the model-based approach we have regression weights which are also part of calibration weights. See Knaub(2012).

model. For weighted least squares (WLS) regression, using regression weights, this is modeled without transformations.

Ahmed and Kirkendall(1981) referenced Royall(1970). That article covered homoscedasticity and two levels of heteroscedasticity, respectively equivalent to considering coefficient of heteroscedasticity values of 0, 0.5, and 1.0, when 0.5 is reflected in the classical ratio estimator. Ken Brewer covered these three levels of heteroscedasticity earlier in his article, Brewer(1963). In Brewer(2002), mid-page 111 shows that the value should be between 0.5 and 1. This is consistent with observed empirical values noted in Cochran(1953), at the top of page 212, as was once pointed out to this author by another.

2.0 Monthly and Annual Miller Surveys in Canada



These surveys are collected by Statistics Canada. In both cases they say that the purpose of these surveys is to collect "…information on grains milled, production of flour and offal as well as stocks." Larger mills are on the monthly survey, and smaller ones are only collected annually, as has been done for electric generation and fuels data at the US Energy Information Administration (EIA) for electric power plants. So with the existing data collections in Canada, quasi-cutoff samples could be employed each month using the current Monthly Miller Survey, and predictor data could be taken from the previous Annual Miller Survey combined with the total of twelve months of the previous year's Monthly Miller Surveys to make a complete annual census of the mill data as predictor data. The current samples would likely be quasi-cutoff samples for multiple attributes as some mills will have larger amounts of different types of grain, flour, offal, and stocks than other mills, and smaller amounts of other such products and stocks, which generally leads to tradeoffs in size by item. One would construct the previous census, and use the current monthly samples, and apply model-based ratio predictions, just as has been done for an electric data survey of many parts, noted just above, at the EIA, for many years. Multiple regression may be useful when relative volumes of products may change, as in the case of "fuel switching" for electric generation. See Knaub(2017b), slides 39 and 40. Although the current make-up of the Monthly Miller Survey may already be best, improvements might be found.

3.0 Weighted Least Squares Regression

Much experience has shown the author that for frequently repeated establishment sample surveys (with an occasional census for predictor data – see Cochran(1953), pages 205-206 on size measures), a default coefficient of heteroscedasticity consistent with the classical ratio estimator is often useful for accommodating lower data quality for responses from the smaller ("Ma & Pa") sample members, who lack resources and expertise to provide high quality data on a frequent basis. In general, these applications have higher levels of heteroscedasticity, but data quality issues in cases corresponding to smaller predicted-y_i will artificially increase variance there, which lowers the effective value of the coefficient of heteroscedasticity, γ, shown below. (See Knaub(2017b), slides 13-17. Also, note that γ is approximate there, as e is an approximation for ϵ. See Knaub(2021).)

For $Y_i = \beta x_i + \epsilon_{0_i} x_i^{\gamma}$ we predict a total, and estimate the variance of the prediction error associated with that total, as follows:

$\boldsymbol{T^* = \sum_{i=1}^{n} y_i + \sum_{i=n+1}^{N} b x_i}$ where $b = \sum_{i=1}^{n} w_i\, x_i y_i / \sum_{i=1}^{n} w_i x_i^2$, and $w_i = x_i^{-2\gamma}$;
$V_L^*(T^* - T) = V^*(b)\left(\sum_{i=n+1}^{N} x_i\right)^2 + \sigma_{e_0}^{*2} \sum_{i=n+1}^{N} x_i^{2\gamma}$, where
$V^*(b) = \sigma_{e_0}^{*2} / \sum_{i=1}^{n} x_i^{2-2\gamma}$, $\sigma_{e_0}^{*2} = \sigma_i^{*2} w_i = \sum_{i=1}^{n} e_{0_i}^2 / (n-1)$, $e_{0_i}^2 = w_i e_i^2$;
and here $\gamma = 0.5$, which gives us the model-based version of the classical ratio estimator.

Note that for a given application, because b is a function of γ, T^* will be changed by a change in γ, although the impact of γ, the coefficient of heteroscedasticity, on $V_L^*(T^* - T)$ is likely much more pronounced.

See Knaub(2022) on variance, Knaub(2016) on the overall methodology, Knaub(2021) on heteroscedasticity in regression, and for more on v_D see pages 130 – 134 in Valliant, Dorfman, and Royall(2000), and see the part of the bibliography for that book regarding Royall with coauthors.

One place the preferred variance estimator here, $V_L^*(T^* - T)$, may be found is in Thompson(2012), in Section 7.6, "Models in Ratio Estimation," pages 105 - 109. On page 107, he calls this variance "The mean square prediction error ...." The format is a little different, but it is equivalent to what we have here. He also considers the situation where γ = 0.5, so that, as Thompson notes on page 107, "...the variance of Y_i is proportional to x_i ...." (See Cochran(1977), pages 158-160.) Notice that the predicted total at the bottom of page 106 in Thompson(2012) is still the sum of the collected Y_i, plus the sum of the predicted-y_i cases. Note we are not restricted to $\gamma = 0.5$. On page 109, Thompson shows three alternatives which correspond to the same three in Brewer(1963) and Royall(1970), such that $\gamma = 0, 0.5, and\ 1$.

On pages 107-108, Thompson(2012) notes that Royall and others have found this variance estimator to be in need of help when the model departs too much from the standard with $\gamma = 0.5$, but Knaub(2022) indicates that this has not actually been a problem in practice here. In the Toxic Release Inventory example below, very like the Official Energy Statistics of Knaub(2017b) and probably the Canadian Miller's Surveys, results for a "robust" estimator, v_D, barely differ from the basic variance estimator.[3] Perhaps this is an indication that $\gamma = 0.5$ works well enough here, though scatterplots in Section 5.0 indicate γ may be a little larger in this example. $V_L^*(T^* - T)$ for $\gamma = 0.5$ and v_D perform similarly for the second example here also, but in that example, for populations of cities, performance is not as good for either variance estimator used. This is true even though in the case of N=63, γ is estimated to be just below 0.5, but larger than 0.75 when all data are considered. However, the data set size is not very large. (Brewer(2002), page 87, stated that a large amount of data is needed to estimate γ well.) This second example appears to be a more sensitive situation, though that can be expected on occasion.

Note, in general, that if you do not have to worry about surprise data quality issues which lower the effective value of γ, results might be found to be better using a larger value for γ, say 0.8, or some other value found by experiment to work best, or using some other default value of γ when the sample size is not large. See Brewer(2002), mid-page 111, and pages 87, 137, 142, and 203. On page 87 he suggests using $\gamma = 0.75$ for "business populations."

[3] v_D in Valliant, Dorfman, and Royall(2000), pages 130-134, appears to be derived from a design-based format.

4.0 Toxics Release Inventory (TRI) Data from the US Environmental Protection Agency (EPA)

4.1 Toxics Release Inventory used as an example to demonstrate methodology

Data showing the volume of toxic substances released from industrial processes is made available by the US EPA in their Toxic Release Inventory as noted under "Other Resources" below the reference and bibliography section, along with acknowledgments for help obtaining the data. Here, data for the years 2014 and 2015 for total releases of lead by paper manufacturers were used to see how well 2014 data could predict for 2015 data. In this set of skewed data, the largest five cases in 2014 corresponded to uncharacteristically small 2015 data. They could be treated separately. There could be a reason such as a special incentive for those plants to reduce lead releases in 2015, but whatever the reason, they are considered special cases for test purposes here, and would be collected and added on as a separate censused stratum. (Units, such as **pounds** of lead, are unimportant here, as long as they are consistent.)

There were also three cases where the 2014 number was zero, and the corresponding 2015 values were positive numbers, and though two added very little in total (8.1 and 175.9), one was of substantial size (3245.6). These were considered to be "births." New 2015 data such as these would also be collected. I refer to them as "add-ons." Here they were removed from the test data set, but would be collected in addition to the sample used for prediction. (The smallest one might be ignored, but the largest one should definitely be added to the predicted total.) There could be other "births" in the sense that a paper manufacturing unit may only be operating and/or releasing for a small part of 2014, and though there lead releases could be atypically small, they might not be zero for 2014. Not knowing this information, and not accounting for the corresponding 2015 volume of lead released as an "add-on," such a data point would have a large residual, perhaps artificially so. The data point noted in the next paragraph could be such a point, but not having expertise with these data, the objective here was just to obtain a reasonable test set to demonstrate that the methods used at the US Energy Information Administration could be useful in similar settings. Large residuals in the other direction would occur if a paper manufacturing unit were to operate normally in 2014, but cease operations for part of 2015. There were five cases which had positive 2014 lead release values and zero values in 2015, and such "deaths" were left in the test data. (They totaled 3079 in 2014, with one a little over half of that total.) The idea is to find a regression which will show how the 2014 respondents morphed into the corresponding 2015 respondents (noting what b makes $y^* = bx$ generally close to Y), which would still need to include/account for cases going to zero volume. Any births, however, would be add-ons to the new population. Shifts in products or volume categories, such as electric power fuel switching or changes in types of flour milled, may be handled with multiple regression. See Knaub(2017b), slides 39 and 40. Still we want y^* generally close to Y. The more you know about your data populations, the better you can handle these issues.

One other data point was considered extreme enough that it was highly suspicious. Its y-value (for year 2015) divided by its substantial x-value (for year 2014) was over 8.5, when this is generally not far from 1 in this particular example. Because the x-value was substantial, and y and x are generally comparable in these data, this does not seem reasonable. In real situations, this would need to be investigated. For a similar situation, see slide number 55 in Knaub and Douglas(2010) for a scatterplot (provided by then JPSM

Intern Lisa Guo) which showed the kind of data investigated at the US Energy Information Administration, which almost invariably was found to have been submitted in error. In practice data should not be removed without investigating and determining if an error has actually been made. In addition, the respondent here could be an establishment which did not release for some part of 2014, using an environmental-friendly option, as also noted in Section 13.0, "Closing." Less extreme versions of this may be a routine part of the process one is modeling, but for unusual cases it would be best if the reporting process were to identify them. For purposes of demonstration here, this point was removed from this test data set for reasonable testing purposes.

Therefore, for the test data population used, after the few "births", which would be "add-ons," and the five large cases were removed as well as the one other arguable data point, the population size for the final **test data** set became N = 152. Changes erroneously made or not made for lack of the author's knowledge of the data would be part of the nonsampling error. When comparing the predicted total, T^*, to a known total T in a test data set/population, the relative error $(T^* - T)/T$ may be made somewhat larger than variance alone would explain, except that the variance estimates themselves could be inflated by the nonsampling error. In any case, below $(T^* - T)/T$ is compared to relative standard errors in test cases. (Besides the one questionable case described above, note that when the three "birth" add-ons, and the five 'large' cases are added, the sample size goes up by 9, but the values of T^*, and T each become larger by over 100,000, and therefore the relative error, $(T^* - T)/T$, and the relative standard error estimates, shown below, become smaller. That is, accuracy of predicting the total becomes greater because of separately collected data points, as long as nonsampling error for these added values is not problematic. These added values are typically just add-ons for births, but a censused stratum which might not be used in the regression might be a possibility.[4]

The variance estimator is compared to a "robust" variance estimator. See Knaub(2022).

In multiple attribute surveys, red and blue lines, such as those below, would likely not strictly hold for each item. A quasi-cutoff sample results from compromises as to which respondents to include when some are 'larger' for some items/attributes than others.

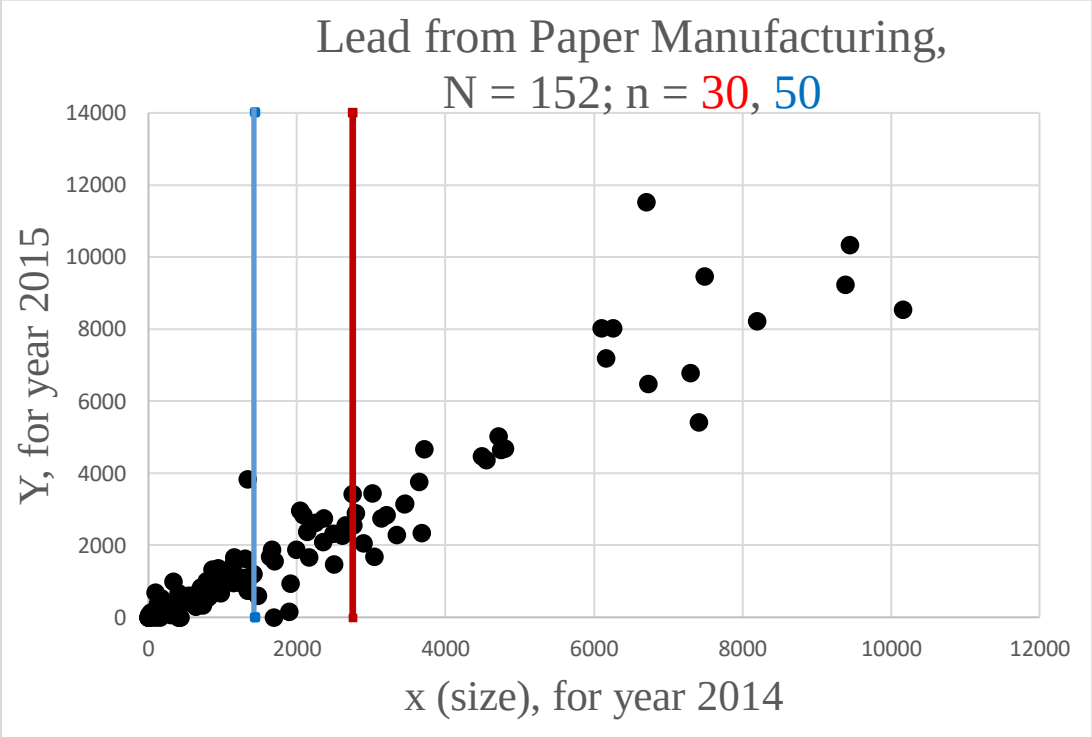
For various regions for an attribute/item, "borrowing strength" might occur.

4.2 How this relates to the Miller Surveys:

Monthly Miller Survey data would provide the multiple-attribute, quasi-cutoff samples. Though there would not be a specific cutoff in a graph such as the one below for a given item, most of the large cases (judged by the previous census) would be included, and perhaps some of the smaller ones. This worked very well for energy data. See Knaub(2017b).

[4] Contrary to this, the first case where the author used quasi-cutoff sampling and prediction at the US Energy Information Administration, *circa* 1990, was to replace a stratified random sample of electric sales and revenue data by end-use sector, which had a censused stratum for the largest entities, by **using that censused stratum as the sample,** and dropping the smaller strata. Results compared favorably to the previous design-based method.

US EPA Toxics Release Inventory (TRI)
Lead Released by Paper Manufacturers



Test Results:

N	n	RSE*	RSE* from vD	T*	T	(T*-T)/T	coverage by sample
152	(top) 30	2.5%	2.7%	247,682	243,536	1.7%	64.7%
152	(top) 50	1.6%	1.7%	241,461	243,536	-0.9%	82.1%

("Coverage by sample" is the percent of T^* actually collected in the sample.)

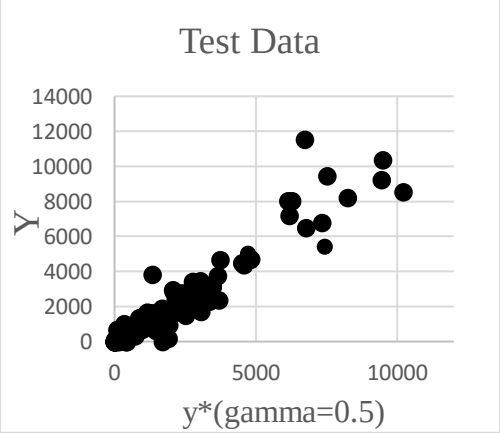
5.0 Heteroscedasticity in Regression

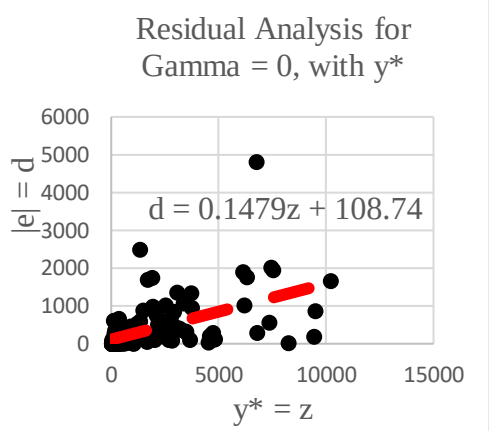
Example Check on Coefficient of Heteroscedasticity, Gamma -
Lead Released from Paper Manufacturing, Test Data, N = 152

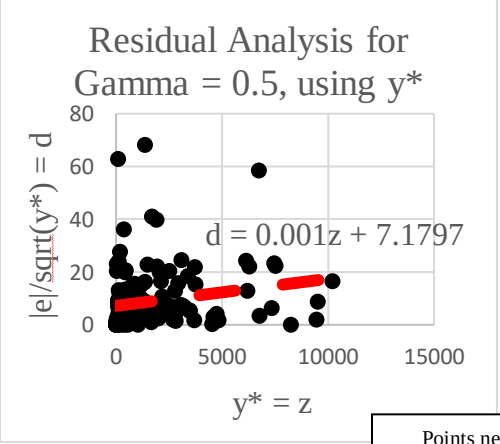
Gamma estimated for e, approximate for true gamma for ϵ:

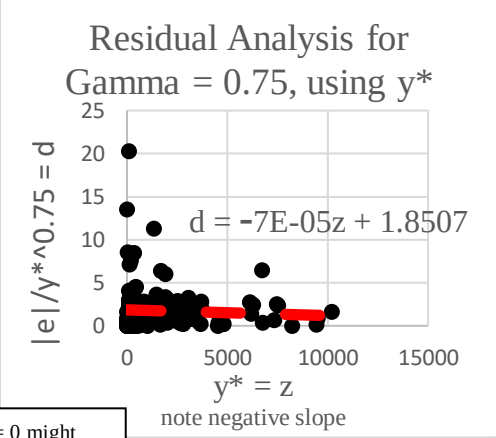
$Y_i = bx_i + e_{0_i} y_i^{*\,\sim\gamma} = y_i^* + e_{0_i} y_i^{*\,\sim\gamma}$

$|e_i| = |e_{0_i}|\, y_i^{*\,\sim\gamma}$ … See Knaub(2019) and Knaub(1993). In both, γ is actually for e, so this is an approximation.







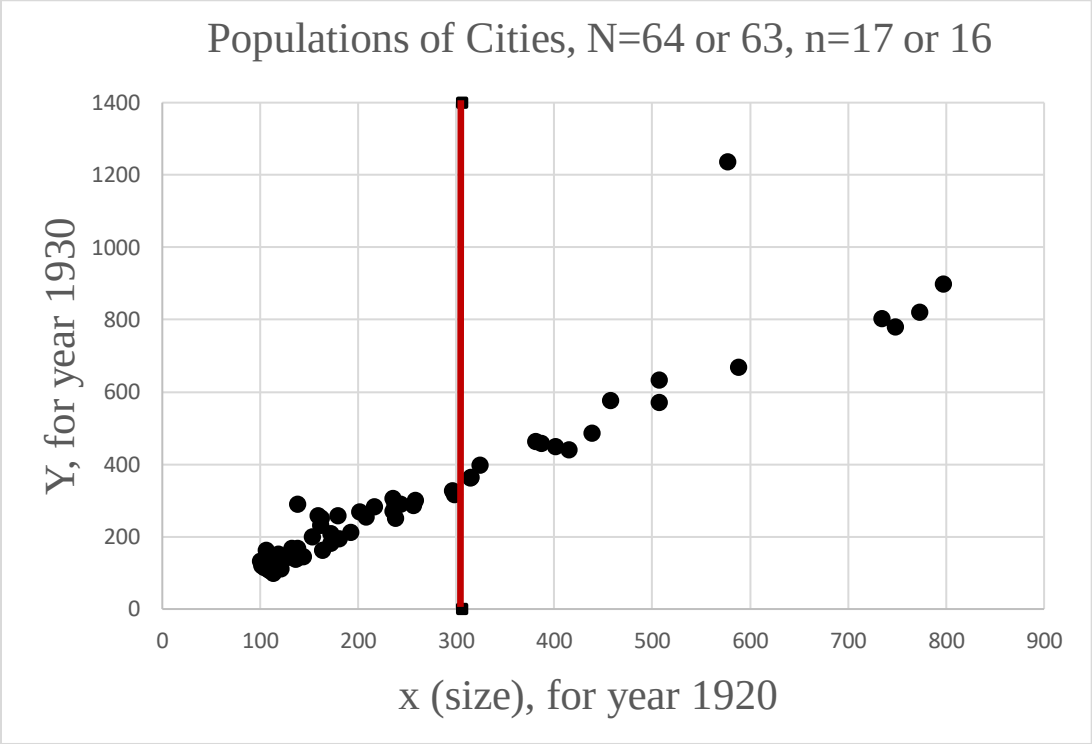


Points near y* = 0 might have disproportionately large measurement error in Y.

We can factor $|e_i|$ into $|e_{0_i}|$ and $y_i^{*\,\sim\gamma}$, where y_i^* is predicted-y_i, using the asterisk as did G.S. Maddala to designate weighted least squares, and y_i^* is the size measure. The coefficient of heteroscedasticity is only approximate here, as we are not considering h_{ii} from the hat matrix, as found, for example, in Weisberg(2014), pages 204-208. Also see Knaub(2021), and for the graphs above, see Knaub(2019). Note that when a simple linear

regression line is drawn by Excel in a broken red line as in each of the residual analysis graphs above, that if the slope is zero, the estimated γ value is found. Here it is indicated to be between 0.5 and 0.75. See Knaub(2019).

6.0 Populations of Cities
As found in Cochran(1977), page 94, and in Chambers and Clark(2012):



Test Results:

N	n	RSE*	RSE* from vD	T*	T	(T*-T)/T	coverage by sample
64	(top) 17	3.7%	4.0%	19,620	19,568	0.3%	53.2%
63	(top) 16	1.0%	1.1%	17,870	18,330	-2.5%	51.5%

- We look at these data with and without the point which has the largest y-value, which might be an outlier. Here again, we would like to know if there is a reason for this point.
- Results for these data appear rather sensitive to this change. Coverage (by the sum of the collected y-values, as a percent of T^*) is low in either case.
- (The point at approximately (140, 300) may also have been important to have authenticated.)

6.1 Populations of Cities used as an example

Data are found in Cochran(1977), page 94, and in Chambers and Clark(2012) see the table on page 32, and a graph on page 53. The case of N=64 keeps the obvious potential "outlier." Thus perhaps accuracy is greater than indicated by the RSE estimates because they assume the suspicious point is typical (*i.e.*, there are others like it, not in the sample).

N=63 is for the case that we know the reason for the odd data point and treat it separately. It would be an additional point we would need to observe, an "add-on" which would be added to the total, and thus it would reduce the RSE estimates and $(T^* - T)/T$. However, here we calculate without it, as an exercise, using the remaining population as a test population.

6.2 How this relates to the Miller Surveys:

In Cochran(1977), on page 93, he says that this population is like many business entity populations in that there are [a few] large units which are a sizable part of the total, and more variable than the smaller ones. This variability would, in general, apply to the estimated residuals as well. The big difference in small and large units means heteroscedasticity is usually more obvious. See Brewer(2002), mid-page 111, and Knaub(2017c, 2021). These data are skewed, as in establishment surveys, which may be highly skewed. Note that to show heteroscedasticity on a graphical residual analysis, estimated residuals do not need greater y-axis range for larger predicted-y on the x-axis, *i.e.*, a 'fan shape,' but only lower density is needed. (An example of this is seen in Knaub(2019).) Here again, the model-based classical ratio estimator is used for predictions.

7.0 Bias versus Variance for cutoff sampling with prediction

When careful not to model populations or strata under one model which do not belong together (see the California hydroelectric example, Appendix 2 here, and Knaub(1999)), little bias will result from cutoff or quasi-cutoff sampling under a ratio model, compared to the vast reduction in variance. This is aided by the fact that when a ratio model is appropriate, one expects that there is a zero intercept with no interpretation for a non-zero intercept, given the subject matter. Points near the origin should not stray very far. When previous census survey predictor data are already available, it is very efficient to use it.

In a comment to the poster for this paper on ResearchGate, I noted the following:

> Apparently many people may think that the tradeoff between the prediction-based approach and the probability-of-selection-based approach is only that models are more efficient but that they are also not completely accurate. But one should also consider that without the auxiliary/predictor data and a model, one may be completely in the dark about the population, making a probability design possibly full of unknown risks that are falsely assumed not to exist at all. Ken Brewer liked to combine the two approaches, see Brewer, K.R.W.(2002), Combined Survey Sampling Inference: Weighing Basu's Elephants, Arnold: London and Oxford University Press, and another good book along this line is

Särndal, C.-E., Swensson, B., and Wretman, J.(1992), Model Assisted Survey Sampling, Springer-Verlang. In the case of model-assisted design-based sampling, the emphasis may be on probability sampling where it seems they address the flaw, of a lack of knowledge of the population, by the use of models. (Ken Brewer once told me that his mentor, Ken Foreman, gave him advice relevant here: You need to know your data/population. He said "There is no substitute," or similar wording.)

See Brewer(2013). … [See balanced versus cutoff sampling in Knaub(2013).]

For many repeated small sample surveys on small populations, a simple model with a quasi-cutoff sample, for many years, has been shown at the US Energy Information Administration (EIA) to be extremely effective - accurate and feasible. That is what this poster and the upcoming paper suggests be used for other Official Statistics from repeated establishment sample surveys with an occasional census.

For other lessons learned see the following ResearchGate project: https://www.researchgate.net/project/Cutoff-and-quasi-cutoff-sampling-with-prediction-for-Official-Statistics, with updates and references.

(Relatedly, heteroscedasticity in regression has been given much consideration here, see https://www.researchgate.net/project/OLS-Regression-Should-Not-Be-a-Default-for-WLS-Regression, and the various updates and references there, but a default coefficient of heteroscedasticity of 0.5, as described there, which is smaller than would often be measured, generally seems useful when smaller respondents are prone to lower data quality. That tends to artificially increase residual sigma associated with smaller predictions, offsetting some of the natural, "essential," heteroscedasticity.)

8.0 Model-based/Prediction approach versus Probability-of-selection-based approach

In the model-based/prediction approach we find a relationship between predictor data and response data, and when simple models are followed well, and we have predictor data for the universe, and results can be tested with repeated surveys, we may be fairly confident of our results. In a probability-of-selection-based approach, we hope that the randomly selected sample members are representative of ones not selected, but we may not know much about the population. How many random selections are needed under a given design to be reasonably certain that the sample resembles or represents (through weighting) the population distribution? That is apart from sample size requirements 'calculated' the usual way by assuming there is no kind of bias problem, and using a variance formula. But that calculation assumes you have not left out any extreme part of the population. If surveys are repeated, we may have a chance to learn more about the population, *especially* if there is an occasional census, but we could still draw 'unfortunate' sample selections (say, mostly small or mostly large members). Ken Brewer advocated a combination of these approaches, Brewer(2002, 2013), and Särndal, Swensson, and Wretman(1992) advocates a probability-of-selection-based approach which is model-assisted, to take advantage of various strengths. However, when you have a relatively small, highly skewed establishment population, it is advantageous to use a quasi-cutoff sample where the largest

members of the population for each of the attributes/data items collected are included. Also, in Brewer(2013), he notes that models may be helpful for small populations, and establishment populations can be small for each of many data items. It has been suggested that the Government do more multivariate regression, *i.e.*, look at the same predictor or predictors for multiple response variables, but in the Official Statistics covered here, we are not so much looking for such relationships as trying to provide baseline information on production or sales for a given market. Data are collected quickly and may be noisy. We need simple relationships which reliably work. Ratio models where the predictor is the same data item from a previous census work well. In the case of fuel switching for energy data (Knaub(2017b), slides 39-40), or for the Miller's Surveys if a mill switches to other grains from the predictor data year, one may need multiple regression, but not multivariate regression. The multiple predictors would be of changing relevance, and may not satisfy the usual increase in variance for an additional (unneeded) predictor noted in Brewer(2002), pages 109-110. If there has been no change in prevalence of fuels used or grains milled, etc., then a variance increase may occur, but when there is 'switching,' accuracy improves when you have included a more relevant predictor. This is a very important point. If there is a chance that, say, the type of wheat milled will change substantially from what was milled in the previous year, then a second regressor/predictor which may perhaps be a sum for all other milled grains, may make a substantial improvement for some months, and this may make a slight increase in variance for other months, where it is not needed, worthwhile. The idea is that during the monthly production of Official Statistics, one will not have the opportunity to tweak modeling throughout the information production system, so deciding on the predictor or predictors and coefficient of heteroscedasticity, whether or not stratification should be done, or conversely, borrowing of strength may be advisable, has to be done for all aggregations being addressed throughout the system, in advance. Luckily for the Monthly and Annual Miller's Surveys, there is a great deal of past data that have been collected and if available would be very useful in deciding what to do. Of course, any subject matter changes would need to be taken into account as they occur, but the massive bulk of modeling decisions can be tested prior to the first data publications from such a system. This may be far superior to most other instances where you have inference from a sample survey, probabilistic or otherwise.

9.0 Quasi-Cutoff Sampling Compared to Unequal Probability Sampling for Multiple Attributes

The vertical lines found on the graphs for the environmental and populations of cities examples are representative of cutoff values on the x-axis for single attribute (one question) surveys. But how many of those do you see? Certainly the Miller Surveys are typical for their many attributes (variables/questions). One mill may process large volumes of oats, but little rye, and another may process large volumes of rye, but fewer oats, as a simple example. If a mill is in the Monthly Miller's Survey because it processes a great deal of oats, they likely report whatever rye they process as well.

Consider now a scatterplot where the x-axis is for the previous year (made up of the Annual Miller's Survey and the sum of the 12 corresponding Monthly Miller's Surveys). So if we look at the rye reported on the current Monthly Miller's Survey, and form a scatterplot as has been done here, there will not be an exact vertical line separating the previous year responses into those with a corresponding Monthly Miller's Survey response to the right of the line, and cases not sampled to the left. There will be some 'smaller cases' in the

sample, and perhaps a marginally large case which may not be in the sample. This is a quasi-cutoff sample.

An advantage to quasi-cutoff sampling as opposed to unequal probability sampling is that each data item has its own size measure. The size measure is **the annual value as described, which corresponds to a current value**, either in the Monthly Miller's Survey or for cases not in the sample which need to be 'predicted' (where prediction means estimation for a random variable). In Cochran(1953), on pages 205-206, Section 9.9, "Measures of the size of a unit," the same data item in a recently performed census is often noted to be "... the best auxiliary ..." variable that one may find, and this choice of size measure has been used extensively for three decades for energy sales and production data at the US Energy Information Administration (EIA).

For an unequal probability sample, which also might be used when handling highly skewed establishment survey data, perhaps in relatively small populations as for the ones which we consider here, the problem is that one size does not fit all. You have to use one size measure, and randomly draw units with probability weighted according to that one size measure. This size measure will then be used to impact estimation of totals for each attribute, though it may not be very appropriate for many of those attributes/data items. For a quasi-cutoff sample, we use a compromise among best samples for each attribute, and a customized size measure for prediction for each one of them.

Many argue that a weakness of such modeling is that the smallest cases are not represented, however with repeated surveys, one has a chance to know the population better. We want to know when one model may cover a population or subpopulation or should more than one model be used by stratum/group, or can there be borrowing of strength. That is, you want to know your data well enough to know what to put under a given model. (As previously noted, Ken Brewer once told this author that his mentor, Ken Foreman, said it is very important to be knowledgeable about 'your' data.) Further, the smallest cases here, as you approach the origin, are not likely to stray very much. If you use unequal probability sampling, few small cases will be included anyway, and we may not know the population so well, which could be problematic as demonstrated by Basu (Basu(1971)) as referenced in Brewer(2002) and shown in Appendix A there ("Basu's Original Elephant Fable"). **Thompson(2012) notes at the bottom of page 103 that the point of that 'elephant story' is that the Horvitz-Thompson estimator will not do well when the inclusion probabilities and the Y_i are not sufficiently well related.** When there are various data items/attributes, what works well for one of them may not work well at all for another.

10.0 Testing and related: Individuals, Totals, Variance, Sample Size

In the case of the Canadian Monthly and Annual Miller's Surveys there should be a large cache of previously collected data available which can be used to test this methodology: determine where it might be advisable to stratify or set subpopulations, and where one might borrow strength, and determine what value or values of γ to use, when 'predicting' for monthly totals and estimating relative standard errors.

10.1 Individual Predictions

To test this methodology, one might first examine some individual cases, as if one were imputing. You can remove a value that was collected, and see how closely you come when 'predicting' it. You can remove several cases randomly, or several small ones only, when using one annual census as if you drew a sample from it, and a previous census for predictor/regressor data. See Knaub(1999) where that was done, though it was the total impact which was considered. (Note that one might think of this as imputation with variance estimation,)

10.2 Predicted Totals

To test prediction of totals, one can take 12 months of predicted totals, by item, for a given year, add results to have predicted totals for the entire year, and compare those to the annual census which later occurs. This was done by Brett Foster and Lisa Guo, JPSM (Joint Program in Survey Methodology) summer interns, as noted in Knaub and Douglas(2010), slide 39.

10.3 Variance Estimation

It is also possible to test variance estimation. See Knaub(2001) and Knaub(2022). One can see the distribution of z-values when z is the predicted total minus the known total, all divided by the estimated variance. Please note that Knaub(2022) demonstrates why the traditional variance was used, as opposed to the "robust" one considered there, and that Knaub(2022) complements Knaub(2016) in showing "When and How to Use Cutoff Sampling with Prediction." Reiterating, strictly speaking, cutoff sampling is for one attribute, and compromises in the sample selection to obtain near cutoff samples for each attribute/data item is something Joel Douglas and I called "quasi-cutoff" sampling at the US Energy Information Administration (EIA). Having some smaller cases for a given data item because they were reported by a respondent which is larger for another data item, however, can also help verify that the model is still good in the lower range. Note that in Knaub(1999, 2001), a further approximation was done for variance estimation with the idea of making it more reproducible, and more easily possible to rearrange the aggregations, and 'borrow strength,' all given the software limitations and problems with data storage and revisions which had occurred. Knaub(2014b) shows simpler examples of borrowing of strength, where this part of the variance estimate was not used.

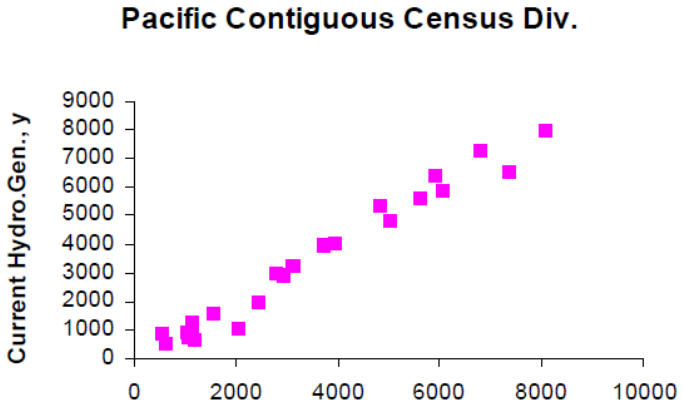
10.4 Additional: Borrowing strength, Heteroscedasticity, Seasonality

10.4.1 Borrowing Strength: Note that in Knaub(1999), it was shown that borrowing strength from incompatible geographic regions will, as one would expect, degrade results. There, for hydroelectric generation in California, it was shown that for purposes of 'borrowing strength,' it was better to model data in the same regional groups of States designated by the National Climatic Data Center, than those States designated by the Bureau of the Census. For purposes of publishing aggregate data, Census Divisions may sometimes be used, but for purposes of modeling, my experiment indicated that the NOAA/NCDC divisions were preferable, as expected, and they were then used in the

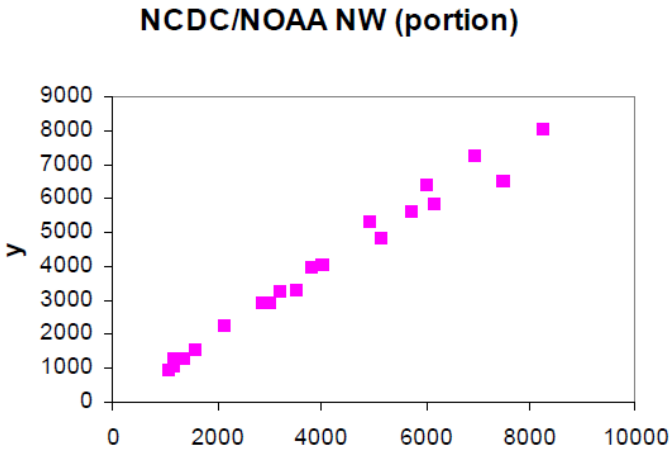
programming for generating State level estimated (technically "predicted") totals, and estimated relative standard errors.

In Knaub(1999), one can see graphically how grouping data matters. Also, please see Knaub(2011b, 2014b). Consider the following graphs from Knaub(1999).

From page 17, the following is hydroelectric generation taken from two annual sets of data prior to 1999. The predictor data on the x-axis is from a multiple regression using previous hydroelectric generation, and nameplate capacity for the reporting establishment generator. Here a cutoff sample for the Pacific Contiguous Census Division is reported, which means that the States involved were California, Oregon, and Washington. (This was experimentation which showed that this was not advisable.)



Next, from page 16 in Knaub(1999), we have a similar cutoff sample, except that the largest three responses are not shown so that the range is comparable to the graph above. For this graph we have the NOAA/NCDC North West Region, which means the States reporting were Washington, Oregon, and Idaho.

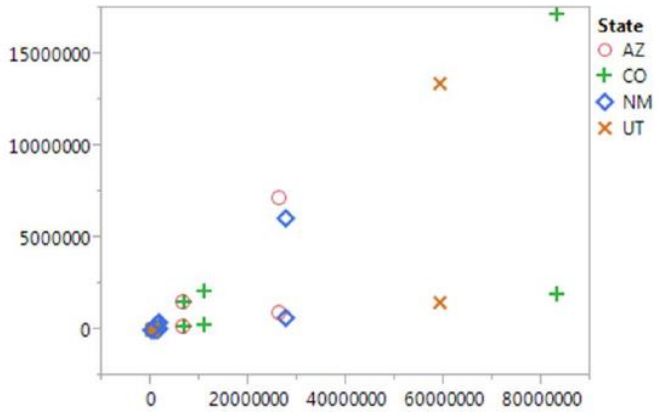


The problem apparently was that California belonged more with Nevada, and that Oregon, Washington, and Idaho belonged together, just as the NCDC/NOAA regions suggested, as opposed to Census Divisions.

Note that what is large for one State may not be large for another State, just as what is a large volume for one item may not be for another. Sample selection must include the very largest respondents for every item and region which will be published. Accuracy for published aggregates at various levels is important.

10.4.2 Heteroscedasticity: Individual monthly data may be tested, but in most cases we use annual census data from which to draw a test sample, with another, previous annual census as predictor data. That isn't ideal, as one wishes to know data quality for the smallest cases. It is suspected that smaller cases in a frequently drawn sample may have data quality issues, as smaller companies may have inadequate resources for providing data frequently. (This is a good reason for asking the smallest establishments to only provide data annually.) This may be a good reason for lowering the estimated coefficient of heteroscedasticity, gamma. In Brewer(2002), mid-page 111, commented upon in Knaub(2017c) and Knaub(2021), we see that we should have $0.5 < \gamma < 1.0$, and often tests for these establishment surveys may show $0.80 < \gamma < 0.95$. However, it may be best to use $\gamma = 0.5$ in frequent, say monthly, data collections, because data quality issues can artificially increase variance of estimated residuals near the origin where variance should be smallest under these models. (See the scatterplots in Section 5.0, and the note in the box there.)

10.4.3 Seasonality: Further, though validation for individual predictions is possible for values we know, and we can compare the sum of 12 monthly predicted totals to corresponding, later obtained annual census totals, we do not have a direct check on individual monthly predicted values. The 12-month totals can be close to the later obtained annual totals, but is the seasonality distribution adequate? An indication of this can be obtained as found in Knaub(2014b), where the same annual data are used as predictor data, with two monthly samples of energy consumption data on the same plot, one for a winter month and one for a summer month, just done as two separate regressions. Following is part of Figure 2 on page 5 of Knaub(2014b):



The impact of seasonality is easily seen in the scatterplot immediately above. For each application one may decide if small missing cases are likely to make much difference.

10.5 Sample Size

With regard to sample size, the current sample size for each data item/attribute in the Miller's Surveys would be whatever is found in the Monthly Miller's Survey. However, if experimenting with old data shows that the accuracy would be too low for a given data item, then it might be advisable to move one or more mills from the annual survey to the monthly survey. One could judge this accuracy both by the variance estimated, and the estimation of the accuracy of the variance estimator, considering its z-value performance, as described in Knaub(2022). That performance may be degraded not only by inaccuracy of the variance estimator, but also by bias due to model-failure. To consider how much larger the sample size should be, consider Knaub(2013), which puts sample size requirement estimation for the model-based version of the classical ratio estimator (i.e, where $\gamma = 0.5$) in the same format as that of estimating sample size requirements under simple random sampling shown in Cochran(1977), at the bottom of page 77. This is just for a population, or one subpopulation, or one stratum, say, one data group being modeled, but it may be helpful. So when we can show that model-failure is not a concern, because we are not mixing data groups which should be modeled separately, then quasi-cutoff sampling as already collected in the Monthly Miller's Surveys may be very efficient, using regression to predict totals. Note in Knaub(1999), "EGs" are "estimation groups" (technically, prediction groups) because they are considered strata or subpopulations which are covered by one model. "PGs" in that paper are "publication groups" which may be made up of EGs, or part of an EG, or there may be overlaps, as illustrated in Knaub(1999). Sample size determinations may therefore be complex, but with all of the past Monthly and Annual Miller's Survey data, one may experiment and see if any adjustments are needed. If the Monthly Miller's Survey is more than adequate as a quasi-cutoff sample here, then it might even be possible to move some mills from the monthly to the annual survey.

Knaub(2013) also shows the extreme advantage cutoff or quasi-cutoff sampling has over balanced sampling, as explained there, with regard to variance estimation. Balanced sampling almost guarantees as 'representative' a sample as one may be hoping for with random sampling. With good auxiliary data one may be more certain of this. However, the cost in loss of efficiency (i.e., higher variance) is astronomical. And with repeated surveys, where testing may be done as described earlier, and the industry may be under real time scrutiny (see Appendix 2), there is good reason to expect you will know if there is substantial model-failure. In addition, with quasi-cutoff sampling, the smaller responses from mills with other data items more important to them, may provide yet another check on model performance as noted earlier.

11.0 Editing using scatterplots

A bonus for using the model-based approach for generation of Official Statistics from repeated establishment surveys is that scatterplots may easily be generated for purposes of data editing. For example, see Knaub and Douglas(2010), slides 54 and 55. Often, for energy data, it has been found that someone reported a number in the wrong units, or from the wrong file, and it was instantly noticeable when scatterplots of preliminary data were generated, as illustrated in those slides.

12.0 Considering a more complex model used in another application

In Guadarrama, Molina, and Tillé(2020) we see a similar situation, except for the absence of a long history of repeated surveys for testing purposes. The example which is noted there was in Section 9, "Estimation of total sales in Spanish provinces." One method used was a model to be described. There the predictor is three months of revenue for a single product for a population consisting of most relevant shops in Spain, across 48 provinces. The Y_i data are collected for a succeeding/current month, from a cutoff sample of large shops, by use of a special device issued to those large shops only. So, we have the desirable case of a data item in a sample survey whose size measure is the same data item in a previous census. Perhaps an annual census instead of 90 days would have been better, but the sample month immediately followed the census data period, so this should be a very good place to use a ratio model, perhaps without a lot of "births" to consider. The model actually used also had an intercept term which varied by province, which is discussed below.

The size measure, z_{ij}, is revenue for a single product for a three month period for the entire population of interest, with "i" being one of the 48 geographic regions noted. Instead of Y_{ij}, they use v_{ij} for the revenue for a sample of shops selling this product in the current month, with the apparent goal of having an early prediction of total revenue for that new month. Because one of the methods they used for estimation or prediction was a model-based approach, we say we "predict" totals, as that is what is of interest here. That model is *basically* a ratio model, where, again as Cochran(1953), pages 205-206 notes, the same data element in a previous census is a very nice size measure for that data element in a current sample. However, they address small area predictions by using the same slope across all 48 geographic regions, but adding a different intercept term for each region. It seems that in the case of revenue for this application, as well as applications at the US EIA in Knaub(2017b), and for the Canadian Monthly Miller's Survey, the ratio of current to past data for a given data item/attribute is important, but an intercept appears to have no real meaning. The difference between the Spanish provinces could be different ratios, even if often only slightly different ratios/slopes by region/province, but differentiating them by various nonzero intercepts would not seem to be very helpful. Apparently this did perform well in this application, but it may be preferable to better consider the meanings of the slopes and intercepts. If instead, as in Knaub(1999), we grouped similar, perhaps adjacent, sets of provinces, and modeled each group of provinces separately (a separate estimated ratio for each group) we might do better than a separate intercept for every single province with a single slope for all provinces, or trying to have a different ratio for each province. At any rate, the borrowing of strength by groups of small areas is what I suggest for the Miller's Surveys, as I did at the US EIA. - As usual, sample sizes for areas, and differences between areas, are considerations which may require compromise.

Even if one were to need to put all areas under one ratio to have a sufficient sample size, it may still be better not to distinguish them by a random intercept, as that could just be noise, as noted in Brewer(2002), pages 109-110. Please see Knaub(2011b) for suggestions on this for some weekly petroleum surveys, which might since have been implemented to some degree.

Note that Guadarrama, Molina, and Tillé(2020) used a transformation to handle heteroscedasticity. This may damage interpretation and is not recommended.

13.0 Review and Closing

It was stressed that Cochran(1953), pages 205-206, notes that a given data item/attribute in a previous census can be a very desirable measure of size for that same data item in a current sample survey. This has performed well for energy data with prediction (versus randomization, and tested using test data from censuses, and by adding monthly results to compare to a later census), and should work well for establishment surveys in general.

For a mid-size to smaller survey on natural gas and oil wells, a colleague, Joel Douglas, was asked to evaluate a problematic methodology being used, and apparently one issue he found troubling was that the sample was being revised on a frequent basis, based on recent sampling. However, sampling must be based on x (*i.e.*, bx or predicted-y, say, y^*), not Y. To drop a member of the sample because it falls under a 'threshold' will tend to artificially/incorrectly increase the predicted total. (If this is just consistent with a lower b, then this action will cause no change in general, except for a larger variance.) Selection should only depend on the predictor data, not obtained sample values. (Note that balanced sampling is also based on predictor data, as would be any model-based method.)

Multiple regression may be used when fuels used or grains milled or whatever volume item is of interest changes, but will actually degrade accuracy a small amount when not needed. However, when such 'switching' does occur, multiple regression may help a great deal. For production of Official Statistics from frequent establishment survey sampling, especially with the huge number of aggregate results required at the US Energy Information Administration, it is not practical to determine the best models post hoc. If one has determined through experimentation that there may be a need for multiple regression because of the possibility of a changing predominance of data items such as fuels used or types of grain milled, then one might just need to put that in place, and reevaluate as the opportunity permits.

At the EIA, the coefficient of heteroscedasticity, γ, was pre-established as well. In most cases for these establishment surveys, experimentation showed that for $Y_i = \beta x_i + \epsilon_{0_i} x_i^{\gamma}$, one might find that $0.7 < \gamma < 0.95$. However, to guard against data quality issues for smaller respondents which could artificially reduce γ due to measurement error induced larger than expected variance near the origin, $\gamma = 0.5$ was often found useful. A step function for γ was even considered in one or more especially problematic cases.

Borrowing strength works only when the groups combined are modeled well with the same model, which may be investigated graphically. See, for example, Knaub(2012), Figure 3, page 13.

Note that EPA TRI data have two kinds of releases (off-site and on-site) which might be considered in the same manner as fuels in fuel switching for electric generation, or different grains in milling, so multiple regression might possibly be helpful when trying to publish each type of release separately. Tim Antisdel (EPA) has studied the releases, the ratio of one type to another and how they change, and he also noted that some entities may reduce releases by some other method. (Perhaps we could call those 'green' methods.) Thus it appears that multiple regression might be a possibility which could be explored in the context there. Also, for the TRI data used in this paper, all geographic regions were collapsed, though the EPA did provide geographic information. If collapsing regions here

was not a good idea, artificially increased heteroscedasticity would be expected. However, the levels shown seem reasonable for essential heteroscedasticity alone.

In Blair and Blair(2015), on pages 169 and 170, cutoff sampling is discussed. On page 170 it is noted that there may be more nonresponse among the smallest members of the population anyway. It has already been noted here that data quality issues may impact the smaller respondents. If one can obtain good annual census data from the smaller and smallest members of the population, then perhaps better predictions may be made for the smallest members than could have been collected in the frequent samples anyway.

Some Background Notes:

J. Knaub, J. Douglas, and others put quasi-cutoff sampling and prediction to work for the US Energy Information Administration (EIA):

1) "Cutoff Sampling and Estimation [Prediction] for Establishment Surveys," June 2010, presentation to EIA found on ResearchGate, *i.e.*, Knaub and Douglas(2010).

2) “Quasi-Cutoff Sampling and the Classical Ratio Estimator - Application to Establishment Surveys for Official Statistics at the US Energy Information Administration - Historical Development,” September 2017, presentation to EIA Math/Stats Lunch found on ResearchGate, i.e., Knaub(2017b).

Dedicated to the memory of Ken Brewer, mentor and friend.

Ken’s interests were more diverse, as shown in his book, *Combined Survey Sampling Inference*, 2002, Arnold. However, he was very encouraging and helpful to me, and to many others.

Thank you Ken.

Ken Brewer
August 2002
Arlington, Virginia, USA

References and Bibliography

Ahmed, Y.Z. and Kirkendall, N.J. (1981). Results of model-based approach to sampling. *JSM Proceedings*, Survey Research Methods Section, American Statistical Association, 674-679.

Basu, D.(1971), "An essay on the logical foundations of survey sampling, part one," (and discussion) in , V.P. Godambe and D.A. Sprott (eds.), *Foundations of Statistical Inference*, Toronto, Ontario, Canada: Holt, Rinehart, and Winston, 203-242.

Blair, E., and Blair, J.(2015), *Applied Survey Sampling*, Sage Publications.

Brewer, K.R.W.(1963), "Ratio estimation in finite populations: some results deducible from the assumption of an underlying stochastic process," *Australian Journal of Statistics*, 5, 93-105.

Brewer, K.R.W.(2002), *Combined Survey Sampling Inference: Weighing Basu's Elephants*, Arnold: London and Oxford University Press.

Brewer, K.R.W.(2013), "Three controversies in the history of survey sampling," *Survey Methodology*, Dec 2013 - Ken Brewer wrote this article as requested when he received the Waksberg Award:
https://www150.statcan.gc.ca/n1/pub/12-001-x/2013002/article/11883-eng.htm

Carroll,R.J. and Ruppert, D.(1988), *Transformation and Weighting in Regression*, Chapman & Hall, Ltd. London, UK.

Chambers, R, and Clark, R(2012), *An Introduction to Model-Based Survey Sampling with Applications*, Oxford Statistical Science Series.

Chaudhuri, A., Stenger, H.(1992), *Survey Sampling: Theory and Methods*, 1st ed, Marcel Dekker, lnc., New York, Basel, Hong Kong. (2nd ed, 2005, CRC Press.)

Cochran, W.G.(1953), *Sampling Techniques*, 1st ed, John Wiley & Sons

Cochran, W.G.(1977), *Sampling Techniques*, 3rd ed, John Wiley & Sons

Guadarrama, M., Molina, I., and Tillé, Y.(2020), "Small area estimation methods under cut-off sampling," *Survey Methodology*, Catalogue no. 12-001-X ISSN 1492-0921, Release date: June 30, 2020,
https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X202000100004, and also see
https://www.researchgate.net/publication/342657185_Small_area_estimation_methods_under_cut-off_sampling

Karmel, T.S., and Jain, M. (1987), "Comparison of Purposive and Random Sampling Schemes for Estimating Capital Expenditure," *Journal of the American Statistical Association*, Vol.82, pages 52-57.

Knaub, J.R., Jr. (1993), "Alternative to the Iterated Reweighted Least Squares Method: Apparent Heteroscedasticity and Linear Regression Model Sampling," *Proceedings of the International Conference on Establishment Surveys*, American Statistical Association, pp. 520-525.
https://www.researchgate.net/publication/263809034_Alternative_to_the_Iterated_Reweighted_Least_Squares_Method_-_Apparent_Heteroscedasticity_and_Linear_Regression_Model_Sampling

Knaub, J.R., Jr.(1999), "Using Prediction-Oriented Software for Survey Estimation," *InterStat*, August 1999, http://interstat.statjournals.net/, partially covered in "Using Prediction-Oriented Software for Model-Based and Small Area Estimation," in *JSM Proceedings*, Survey Research Methods Section, American Statistical Association, 1999, and partially covered in "Using Prediction-Oriented Software for Estimation in the Presence of Nonresponse," presented at the *International Conference on Survey Nonresponse*,1999.
https://www.researchgate.net/publication/261586154_Using_PredictionOriented_Software_for_Survey_Estimation

Knaub, J.R., Jr. (2001), "Using Prediction-Oriented Software for Survey Estimation - Part III: Full-Scale Study of Variance and Bias," *InterStat*, June 2001, http://interstat.statjournals.net/. (Note another version in *JSM Proceedings*, Survey Research Methods Section, American Statistical Association, 2001.)
https://www.researchgate.net/publication/261588075_Using_PredictionOriented_Software_for_Survey_Estimation_-_Part_III_FullScale_Study_of_Variance_and_Bias

Knaub, J.R., Jr. (2010), "On Model-Failure When Estimating from Cutoff Samples," *InterStat*, June 2010,
https://www.researchgate.net/publication/261474154_On_ModelFailure_When_Estimating_from_Cutoff_Samples

Knaub, J.R., Jr.(2011a), "Cutoff Sampling and Total Survey Error," Letter to the Editor, March 2011, *Journal of Official Statistics*, 27(1):135–138,
https://www.researchgate.net/publication/261757962_JOS_Letter_-_Cutoff_Sampling_and_Total_Survey_Error

Knaub, J.R., Jr.(2011b), "Some Proposed Optional Estimators [Predictors] for Totals and their Relative Standard Errors for a set of Weekly [Quasi-]Cutoff Sample Establishment Surveys," *InterStat*, July 2011.
https://www.researchgate.net/publication/261474159_Some_Proposed_Optional_Estimators_for_Totals_and_their_Relative_Standard_Errors_for_a_set_of_Weekly_Quasi-Cutoff_Sample_Establishment_Surveys

Knaub, J.R., Jr. (2012), "Use of Ratios for Estimation of Official Statistics at a Statistical Agency," *InterStat*, May 2012,
https://www.researchgate.net/publication/261508465_Use_of_Ratios_for_Estimation_of_Official_Statistics_at_a_Statistical_Agency

Knaub, J.R., Jr. (2013), "Projected Variance for the Model-Based Classical Ratio Estimator: Estimating Sample Size Requirements," *JSM Proceedings*, Survey Research Methods Section, American Statistical Association, pp. 2885-2889, https://www.researchgate.net/publication/261947825_Projected_Variance_for_the_Model-based_Classical_Ratio_Estimator_Estimating_Sample_Size_Requirements

Knaub, J.R., Jr.(2014a), "Efficacy of Quasi-Cutoff Sampling and Model-Based Estimation [Prediction] for Establishment Surveys and Related Considerations," *InterStat*, January 2014, Revised April 2014, https://www.researchgate.net/publication/261472614_Efficacy_of_QuasiCutoff_Sampling_and_ModelBased_Estimation_For_Establishment_Surveys_and_Related_Considerations

Knaub, J.R., Jr.(2014b). "Quasi-Cutoff Sampling and Simple Small Area Estimation [Prediction] with Nonresponse," *InterStat*, May 2014, https://www.researchgate.net/publication/262066356_Quasi-Cutoff_Sampling_and_Simple_Small_Area_Estimation_with_Nonresponse

Knaub, J.R., Jr. (2014c), "A Note on Regression Through the Origin: What to do with missing or zero for x or y," ResearchGate, https://www.researchgate.net/publication/262336211_A_Note_on_Regression_Through_the_Origin_What_to_do_with_missing_or_zero_for_x_or_y

Knaub, J.R., Jr.(2015), "When Prediction is Not Time Series Forecasting," Research found on ResearchGate, https://www.researchgate.net/publication/275365705_When_Prediction_is_Not_Time_Series_Forecasting

Knaub, J.R., Jr.(2016), ""When and How to Use Cutoff Sampling with Prediction," Method found on ResearchGate, https://www.researchgate.net/publication/359204410_Variance_of_the_Prediction_Error_for_Totals_Under_Cutoff_or_Quasi-Cutoff_Sampling

Knaub, J.R., Jr.(2017a), "Comparison of Model-Based to Design Based Ratio Estimators," Prepared for *JSM Proceedings*, Survey Research Methods Section, American Statistical Association, https://www.researchgate.net/publication/319434665_Comparison_of_Model-Based_to_DesignBased_Ratio_Estimators.

Knaub, J.R., Jr.(2017b), "Quasi-Cutoff Sampling and the Classical Ratio Estimator - Application to Establishment Surveys for Official Statistics at the US Energy Information Administration - Historical Development," Invited Presentation, Presentation: EIA Math/Stats Lunch, September 2017, DOI: 10.13140/RG.2.2.33300.60803/1, Presentation found on ResearchGate, https://www.researchgate.net/publication/319914742_Quasi-Cutoff_Sampling_and_the_Classical_Ratio_Estimator_-_Application_to_Establishment_Surveys_for_Official_Statistics_at_the_US_Energy_Information_Administration_-_Historical_Development

Knaub, J.R., Jr. (2017c), "Essential Heteroscedasticity," November 2017, Research on ResearchGate,
DOI: 10.13140/RG.2.2.20928.64005, ResearchGate,
https://www.researchgate.net/publication/320853387_Essential_Heteroscedasticity

Knaub, J.R., Jr.(2018), "Nonessential Heteroscedasticity," April 2018, Research found on ResearchGate,
https://www.researchgate.net/publication/324706010_Nonessential_Heteroscedasticity

Knaub, J.R., Jr.(2019), "Estimating the Coefficient of Heteroscedasticity," June 2019, Method found on ResearchGate,
https://www.researchgate.net/publication/333642828_Estimating_the_Coefficient_of_Heteroscedasticity,
with a tool for implementing this found at
https://www.researchgate.net/publication/333659087_Tool_for_estimating_coefficient_of_heteroscedasticityxlsx

Knaub, J.R., Jr.(2021), "When Would Heteroscedasticity in Regression Occur?"
Pak. J. Statist., Vol. 37(4), 315-367, https://www.pakjs.com/,
http://www.pakjs.com/wp-content/uploads/2021/07/37401-1.pdf,
https://www.researchgate.net/publication/354854317_WHEN_WOULD_HETEROSCEDASTICITY_IN_REGRESSION_OCCUR

Knaub, J.R., Jr.(2022), "Variance of the Prediction Error for Totals Under Cutoff or Quasi-Cutoff Sampling," Research found on ResearchGate,
https://www.researchgate.net/publication/359204410_Variance_of_the_Prediction_Error_for_Totals_Under_Cutoff_or_Quasi-Cutoff_Sampling

Knaub, J.R., Jr., and Douglas, J.R.(2010), "Cutoff Sampling and Estimation for Establishment Surveys," EIA Seminar, DOI:10.13140/RG.2.1.1001.3282, Presentation found on ResearchGate,
https://www.researchgate.net/publication/263927238_Cutoff_Sampling_and_Estimation_for_Establishment_Surveys

Lohr, S.L(2010), *Sampling: Design and Analysis*, 2nd ed., Brooks/Cole.

Maddala, G.S.(1977), *Econometrics*, McGraw-Hill.

Maddala, G.S.(2001), *Introduction to Econometrics*, 3rd ed., Wiley. (There is a fourth edition with Lahiri.)

Royall, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.

Royall, R.M.(1992), "The model based (prediction) approach to finite population sampling theory," *Institute of Mathematical Statistics Lecture Notes* - Monograph Series, Volume 17, pp. 225-240. – Note that information on this is found at https://www.researchgate.net/publication/254206607_The_model_based_prediction_approach_to_finite_population_sampling_theory. -
The paper is available under Project Euclid, open access:
https://www.doi.org/10.1214/lnms/1215458849

Särndal, C.-E., Swensson, B., and Wretman, J.(1992), *Model Assisted Survey Sampling*, Springer-Verlang

Sweet, E.M., and Sigman, R.S. (1995). "Evaluation of Model-Assisted Procedures for Stratifying Skewed Populations Using Auxiliary Data," *JSM Proceedings*, Survey Research Methods Section, American Statistical Association, pages 491-496.

Thompson, S.K.(2012), *Sampling*, 3rd ed, John Wiley & Sons.

Valliant, R, Dorfman, A.H., and Royall, R.M.(2000), *Finite Population Sampling and Inference: A Prediction Approach*, Wiley Series in Probability and Statistics.

Weisberg, S.(1980), *Applied Linear Regression*, 1st ed., John Wiley & Sons.

Weisberg, S.(2014), *Applied Linear Regression*, 4th ed., John Wiley & Sons.

Other Resources

U.S. Energy Information Administration Resources:
Main page: https://www.eia.gov/
List of Surveys: https://www.eia.gov/survey/
Electric Power Monthly: https://www.eia.gov/electricity/monthly/
Electric Power Annual: https://www.eia.gov/electricity/annual/
Electricity: https://www.eia.gov/electricity/data.php
Natural Gas Monthly: https://www.eia.gov/naturalgas/monthly/
Natural Gas Annual: https://www.eia.gov/naturalgas/annual/
Natural Gas: https://www.eia.gov/naturalgas/data.php
Weekly Petroleum Status Report: https://www.eia.gov/petroleum/supply/weekly/
Petroleum Supply Annual: https://www.eia.gov/petroleum/supply/annual/volume1/
Total Energy Monthly: https://www.eia.gov/totalenergy/data/monthly/
EIA-914 Monthly Crude Oil and Lease Condensate, and Natural Gas Production Report Methodology,

https://www.eia.gov/petroleum/production/pdf/eia914methodology.pdf -
Problems in here which colleagues and/or this author have found: (1) Sample should be based only on x (bx) or predicted-y, not Y. (2) Might not be able to mix different types of wells, which may be separable by depth.

Source for US Environmental Protection Agency (EPA) Toxic Release Inventory (TRI) data:
https://www.epa.gov/toxics-release-inventory-tri-program/tri-data-and-tools

This data source was suggested to me by Prof. Wayne B. Gray, Executive Director, Boston Research Data Center, NBER, and Professor of Economics, Clark University. The US EPA was helpful when I obtained data for the example application used here. In particular, Timothy Antisdel, EPA, was very helpful in answering my questions and sharing some of his knowledge of these data. Any misunderstandings would be my own. Good test data were obtained.

Canadian Mill surveys:
Of course all Canadian/agriculture contacts were gracious and helpful. Thank you.
Monthly Miller's Survey:
https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3403
https://www.statcan.gc.ca/en/survey/agriculture/3403
Annual Miller's Survey:
https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3443
https://www.statcan.gc.ca/en/survey/agriculture/3443

Research Projects on ResearchGate:

1. CUTOFF AND QUASI-CUTOFF SAMPLING WITH PREDICTION FOR OFFICIAL STATISTICS, https://www.researchgate.net/project/Cutoff-and-quasi-cutoff-sampling-with-prediction-for-Official-Statistics
2. OLS REGRESSION SHOULD NOT BE A DEFAULT FOR WLS REGRESSION, https://www.researchgate.net/project/OLS-Regression-Should-Not-Be-a-Default-for-WLS-Regression - Note the "Update" there arguing against transformations.

Appendix 1
Anecdote: On Quasi-Cutoff Sampling and Prediction – *Circa* 2011

At a presentation over ten years ago, an internationally known statistician was visiting the United States in the Washington DC area, and addressed local statisticians. On this particular occasion, I commented on the usefulness of (quasi-)cutoff sampling with prediction, but the speaker did not agree. Another member of the audience, who, as in the case of the speaker, was also a well-respected and well-known statistician and university professor, noted that where I worked, the US Energy Information Administration (EIA), there appeared to be an exception to the speaker's objection. This statistician had been a member of the American Statistical Association's Committee on Energy Statistics, and had worked one summer at the EIA. I led major development of this approach, with my first application to a survey beginning around 1990, and application of "borrowing of strength" for some small area predictions beginning in 1999. Perhaps tens of thousands of results over more than 30 years now have proven the continuing utility of this approach. Many users of different parts of the EIA data from EIA surveys have apparently scrutinized results thoroughly and continuously, over these years.

At the beginning, the first results were obtained by using a certainty stratum from a stratified random design as a quasi-cutoff sample, dropping the non-certainty strata, and similar results were obtained by predicting for all data not collected.

Quasi-cutoff sampling means that some smaller cases for some items/attributes will be included in the sample, and some larger ones will not. For each item, the size measure will be the value for the same item in the previous census, a size measure highly endorsed in Cochran(1953), pages 205-206. At the prediction stage, each regression weight will have its own size measure, unlike the case of unequal probability sampling where one size measure is used for every item. One could start with a quasi-cutoff sample based on a cutoff for some total volume, but may need to adjust the sample to accommodate the various categories/items collected. What is a large volume for some items would be small for others, and various establishments may have greatly differing interests in different items. Thus to approach a sample of establishments with an emphasis on collecting most of the largest values for each item, some adjusting may be done, and the real goal is to obtain reasonable relative standard error estimates for the prediction errors associated with the predicted totals, at least for the most important items.

The distinguished member of the audience that day, over ten years ago, as I recall, noted that the EIA represented a special case. But what is special about these data? Well, they are highly skewed establishment survey data, with frequent sample surveys (say monthly and even weekly), with an occasional census (say annually or perhaps less often) collecting data on the same items. The population basically remains the same, but "births" are considered new to the population, which I called "add-ons," and "deaths" would be kept within the predictions as that shows what happened to the original population, morphing into the new one. Thus x-values are always positive (or, in multiple regression, at least one predictor will be positive), but Y-values can be zero or particularly small when a business shuts down for any reason, temporary or not.

One thing special about this situation is that the predictor data are reliably well correlated with the variable of interest in each case. This can change, for example, if fuel sources are

switched. Then I found that using multiple regression to cover the possibility of "fuel switching" could make a great improvement when there actually was fuel switching, though it did not do quite as well when there was no fuel switching. See Knaub(2017b).

Further, one cannot mix data which should not be modeled under the same model, and the repeated surveys at the EIA provide data for testing to avoid this problem when borrowing strength. This is the subject of Appendix 2. One also needs to consider subject matter news of changes to the populations.

Test data have been used to prove the results useful. Further, taking 12 monthly sample predictions for totals and adding them has matched well with later collected annual census data. Scatterplots based on modeling have repeatedly indicated data subsequently found to have been collected in error, such as from the wrong file, or collected in the wrong units.

So are these conditions unique to the US EIA? Are there any other statistical agencies which collect such repeated establishment surveys? I would think so. And perhaps the successes at the EIA do not require all of these conditions, where quasi-cutoff sampling with prediction just have not been tried. Highly skewed populations are compatible with this methodology, and a long line of repeated sample surveys with occasional census surveys provides good test data, but when there are too much data being collected too frequently, data quality is a concern, and one cannot make changes such as considering multiple regression if it is not already included. (Please keep in mind, however, that the sample is based on x (*i.e.*, bx or y^*) , not Y.) Perhaps a less hectic environment might produce even better results.

Looking for other such surveys, from other sources, to show that this methodology has applications beyond energy data meant finding data that were not confidential, and this proved difficult. In Canada they have a Monthly Miller Survey, which is basically a cutoff or likely a quasi-cutoff sample, and then they collect an Annual Miller Survey on the small mills, both for various products and stocks. An annual census may be formed by the aggregation of these, as is the case for electric generation, and consumption and stocks of fuels in the survey discussed in Appendix 2. (In the US, for electric sales by economic end-use sector surveys, at least in the past, the annual survey covered the entire population so that nonsampling error was able to be examined for the monthly versus annual survey collections of numerous large establishments in the monthly survey. However, for electric generation, fuel consumption and stocks, as in the case of the Canadian Miller Surveys, the large entities surveyed monthly are not surveyed again annually.)

However, though this should work well for the mill surveys, using data already collected, those data were not available to me to examine. It is my hope that this methodology will be used with those surveys to provide monthly predicted totals, using data already being collected, where older data may be used for experimental/testing purposes.

The US Environmental Protection Agency (EPA) data used in this paper were employed for demonstration purposes. However, if the EPA wanted to predict, using cutoff or quasi-cutoff sampling for every other year for similar data, this should be possible. With, say, perhaps even years collecting census surveys, and odd years each being a sample, that may free resources for other purposes, though the EPA may not be able to wait two years for some census data. If some monthly samples were desirable, that might also be an option.

The populations of cities data were also just used for demonstration purposes, as

Cochran(1977), page 93, said that these data resembled business survey data, and it was difficult to find other available data. These city data are also used in Chambers and Clark(2012), where I saw them presented in a scatterplot graph.

It seems obvious that this has been a neglected area of survey statistics. Many statisticians have been led to believe that cutoff sampling is fatally flawed with bias, such that it is only a quick, cheap alternative when the smallest establishments are very little and do not add too much to totals. However, decades of research and vast practical application has shown that this assessment need not be so. For thousands of applications of energy data at the US EIA, over three decades, very little bias is added in exchange for extreme reductions in variance,[5] so that cheaper can also be more accurate, and perhaps even make some survey results possible which otherwise would not have been possible.

An advantage with prediction here is that you can make full use of census data which are already available. With probability-of-selection-based (design-based) approaches in general applications, one may not be as familiar with a population, and not know that there are different parts of the population, such that either a missed or an included special case may greatly impact inference when you assume the data collected are like the data not collected. Even to properly stratify means knowing something about the population. Ken Brewer did tell me that his mentor, Ken Foreman, noted There is no substitute for knowing your data, or words to that effect. (Dr. Brewer worked in other areas and combined probability-of-selection-based and model-based approaches, but he once told me that he had been involved with a survey of tailor shops for which he thought the type of sampling and prediction used here was satisfactory.) Model-assisted design-based sampling and inference could make use of auxiliary data, which are the same as the predictor data here, but perhaps not as efficiently for these highly skewed, numerous, small populations. (Note comments on small populations near the end of Brewer(2013).)

Quasi-cutoff sampling with prediction, I have found, can be the most accurate alternative, and the most cost effective one, as well as sometimes the only viable course of action. (Sometimes collecting data from very small establishments on a frequent basis is extremely problematic.) That it can sometimes be most accurate is often not realized, and this may be far more frequent an occurrence than generally imagined. It may not be so much that the US EIA is such a special case, as that few statisticians have worked very much with quasi-cutoff sampling and prediction, and perhaps stayed with unequal probability sampling in similar circumstances, or stratified random sampling, with strata by size. There are serious concerns with all methods. One should consider all aspects of total survey error for various applications.

[5] See the difference between balanced and cutoff sampling shown in Knaub(2013).

Appendix 2
Anecdote: Augmentation of Survey of Electric Generation, Fuels and Stocks – *Circa* 1999 – On Scrutiny of Published Data

In 1999 the US Energy Information Administration (EIA) promised to provide finer levels of details for aggregate data published monthly for electric generation, fuels and stocks by geographic location, to its eager data user community. State level was the finest geographic level, though one past type of more aggregate region could actually cut through a State's boundaries, and past data had been difficult to regenerate because of numerous micro-level data revisions. When I became aware that this promise had been made, and for the near term without considering the need to increase the overall sample size, I had to very quickly consider how to meet the new demands with the limits of the available software. The number of new (sub)aggregate categories was to be, as I recall, several times as many as in the past. I managed to convince management to slightly increase the sample size, and I experimented with borrowing strength for an elementary small area prediction approach. To handle future data revisions and possible documentation issues, software limitations, and possible changes in groupings for aggregation, I developed an estimate for variance which was quite satisfactory, though accuracy of the variance estimate was somewhat reduced from the usual method. See Knaub(1999). A member of the American Statistical Association's Committee on Energy Statistics did not like this, but I am not certain that he knew all of the restrictions under which I was laboring. Later, under simpler conditions for monthly natural gas publications, I did not use that estimation in the estimated variance of the prediction errors associated with the small area predicted totals. See Knaub(2014b).

At any rate, we prepared for the 1999 monthly surveys of electric generation, fuels and stocks at tremendous speed. The sample selections were made, and predictions were programmed quickly. That is where this anecdote will show the scrutiny under which the published results are generally placed. I had decided, after testing for the West Coast, to model hydroelectric generation, which needed to be published by State, by borrowing strength, not across California, Oregon, and Washington, even though they were considered an aggregate group at a higher level for publishing, but to model California with Nevada in one group, and Oregon, and Washington, with Idaho in another. This is because I found that the US National Climatic Data Center (NCDC) for the National Oceanic and Atmospheric Administration (NOAA), has regions with similar weather patterns, and I used those instead of the publication regions derived from the Bureau of the Census, which EIA called Census regions. I experimented with the data, as just noted, and decided that borrowing strength within NOAA/NCDC regions was better than borrowing strength within Census regions, even though nothing was published at the NOAA/NCDC region level.

However, the first month that this methodology was in place, a data user complained that the hydroelectric generation ('predicted') (sub)total which we published for California did not seem correct. We went back and found that the programming, which was exceedingly hurried to make the deadline, had mistakenly been written to combine California, Oregon, and Washington data for modeling purposes, contrary to the experimental results I had found (Knaub(1999)). When that was corrected, there were no more issues. This illustrates, however, the scrutiny under which EIA data were subjected.

This highlights an issue when using this approach: A model must only be applied to data which should be modeled together. When repeated surveys are used for Official Statistics, we have a good chance of finding what goes together by (1) experimenting with the past data, and (2) being alert for new changes to the subject matter (such as news that certain entities are doing business differently across geographic regions than they had previously). When borrowing strength, as above, we need to study where that is and is not appropriate.