An Iterative Cutoff Sampling Method Applied to EIA's Annual Survey of Domestic Oil and Gas Reserves

Jason Worrall, Samson Adeshiyan
U.S. Energy Information Administration, Forrestal Bldg., Washington DC 20585

1. Introduction

This paper presents a procedure for selecting a cutoff sample of multiple survey data items using non-enveloped estimation and publication groups (i.e., estimation groups are not contained within publication groups, nor are publication groups contained within estimation groups). The procedure is intended to select the minimal sample size needed to obtain given target Relative Standard Errors (RSEs) for each publication item. It builds on work by Jim Knaub at the Energy Information Administration in model based sampling to present a generalizable framework for cutoff sample selection.

Cut-off samples using a model-based estimation approach were introduced to EIA surveys in the 1980s. References [1], [4] and [7] noted that this sampling approach is useful for highly skewed establishment surveys with periodic samples where the target population is surveyed periodically by a census survey, or reliable third-party data are available. Since the respondent data from the sample and census surveys that collect the same data items from the same companies are expected to have high positive correlation, a model-based estimation approach that capitalizes on this relationship typically yields small and efficient samples. This saves cost and reduces respondent burden.

The theoretical underpinnings of this methodology were presented in seminal papers by references [3] and [10]. Essentially, the samples are comprised of relatively large establishments. Because smaller establishments have, in the past, been responsible for a greater number of reporting errors and nonresponse, cut-off sampling may also reduce the levels of non-sampling error affecting the published estimates (see references [5], [6]) for more examples of how cut-off sampling has been applied to several EIA surveys).

References [10], [12] and [9] describe how the model-based estimation approach is derived from the superpopulation model, which is assumed inherent in a given finite population. Consequently the values of the survey variable, denoted by $Y_1, Y_2, ., Y_N$, are assumed to be random variables having a joint probability distribution. In any given survey period the actual quantities $y_1, y_2, ., y_N$ form a realization of the random variables $Y_1, Y_2, ., Y_N$, respectively. The joint distribution is the superpopulation model, and it describes the randomness of the Y-values. In our application, the superpopulation model is specified by a regression model, for which we possess information on a positive-valued auxiliary variable X whose values $x_1, x_2, ., x_N$ on the population units are fixed and known in advance. In our application, these X values are from a third party data source. The regression model M that yields the Classical Ratio Estimator (CRE) is of the form:

$$E_M(Y_i|X_i = x_i) = \beta x_i, \quad V_M(Y_i|X_i = x_i) = \sigma^2 x_i$$

$$Cov_M(Y_i, Y_j|X_i = x_i, X_j = x_j) = 0, \quad \forall i \neq j$$

Suppose we want to estimate $T = \sum_i^N Y_i$. If we denote the set of sample units by s, then $T = \sum_s Y_i + \sum_{s^c} Y_i$. So, T can be estimated as $\hat{T} = \sum_s y_i + \sum_{s^c} \hat{Y}_i$, where $\hat{Y}_i$ is the best linear

unbiased predictor. We can use the asymptotic normality of the standardized error, $\frac{\hat{T}-T}{V_M(\hat{T}-T)}$ under the regression model M to set approximate confidence intervals for T. This provides the basis for 90 percent confidence intervals of the form $\hat{T} \pm 1.645\sqrt{V_M(\hat{T}-T)}$.

Knaub 2013 ([8]) presents a methodology in the case of the Classical Ratio Estimator (CRE) for estimating the approximate survey sample coverage needed to achieve a target RSE, but no general procedure has been described for optimizing sample size across multiple data items or estimation and publication groups, nor for models aside from the CRE. This paper attempts to do that.

The case study for this methodology is EIA's Form EIA-23, "Annual Survey of Domestic Oil and Gas Reserves". This survey collects annual oil and gas production and year ending oil and gas reserves from a sample of oil and gas well operators. Estimates of total reserves are then published in EIA's "Annual Domestic Oil and Gas Reserves Report" by state, subdivision, and reservoir type. The sampling frame is provided by "DrillingInfo" (DI), a subscription database of monthly oil and gas production by well, which has information on well characteristics and current operator identification.

2. The Problem

Respondents to Form EIA-23 are oil and gas well operators. They may operate many different wells in many different regions all across the country. They report annual production of oil and gas by field, and year-end reserves of oil and gas by field on the EIA-23. Reserves are defined as "Proved reserves of oil and gas ... are the estimated quantities of oil and/or gas, which geological and engineering data demonstrate with reasonable certainty to be recoverable in future years from known reservoirs under existing economic and operating conditions." Additional details are available in the Form EIA-23 instructions. EIA estimates reserves for non-sampled operators using the gamma super population model (Equation 1):

$$y_i = \beta x_i + x_i^{\gamma}\epsilon_i \tag{1}$$

Where y_i represents an operator's estimated year-end reserves, x_i annual production, and ϵ_i is a random disturbance with mean zero and variance σ^2. β is a model parameter to be estimated. This model is estimated by weighted least squares with weights equal to $x_i^{-2\gamma}$. The reason for expecting an approximately linear relationship between reserves and production is mostly a empirical matter, but can be very loosely tied to decline curve theory from petroleum engineering. Reference [2] was a landmark paper in petroleum engineering that used fluid dynamics to establish a theoretical relationship between production and reserves for an individual well, given in equation 2.

$$q_t = \frac{q_i}{(1+At)^{\frac{1}{B}}} \tag{2}$$

A, B, q_i are all well-specific constants that result from petroleum engineering principles, q_t is the production rate at time t. Define reserves at time T as the total cumulative production of a well over its lifetime, minus the cumulative production at time T, i.e. $R_T = \sum_{t=0}^{\infty} q_t - \sum_{t=0}^{T} q_t$. Plotting reserves versus production yields Figure 1 ($B = 1$, $q_i = 100$).

Figure 1 shows a curvilinear relationship between reserves and production from a single well. In reality this relationship is distorted for many reasons. Respondents to the Form EIA-23 report total reserves which they produce from, and in practice accessing all of these reserves may require drilling multiple wells. The technical skill of the operator in question (which may be correlated with total operator production) can affect recoverable reserves. An operator that is not producing from all their

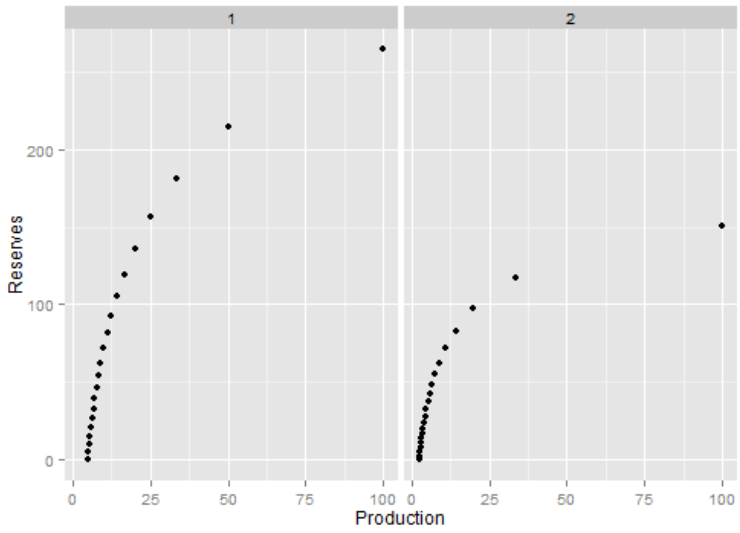


Figure 1: Reserves Vs Production Predictions Using Decline Curve Theory For Different Values of 'A'

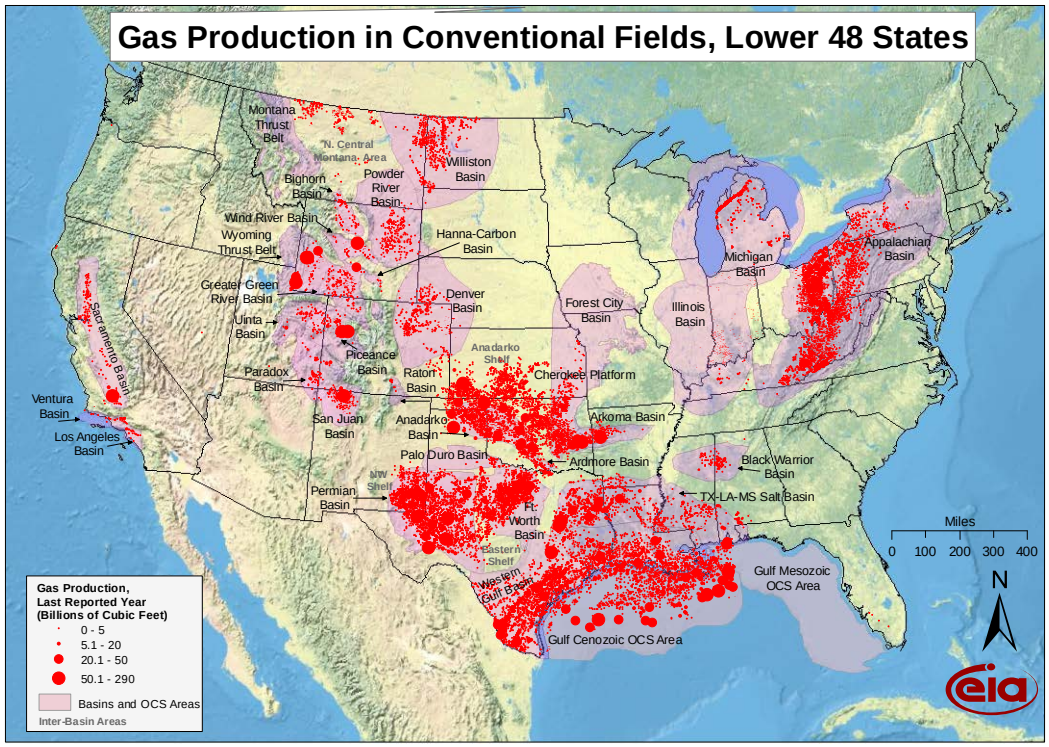
wells for the entire report period will have their production artificially deflated. The result of all this is that higher production operators often report substantially higher reserves than predicted by figure 1, making a linear model more appropriate.

Oil and gas wells are not homogenous, i.e. different regions and reservoir types have different geologic characteristics that affect the relationship between reported production and reserves. So Equation 1 is estimated separately for each region and reservoir type (distinct estimation groups). In addition, the Annual Reserves Report publishes reserves by state, subdivision, and reservoir type (distinct publication groups). These groups are not enveloped, meaning estimation groups are not necessarily contained within a single publication group, nor are publication groups contained within a single estimation group. The situation is illustrated in figure 2, where the estimation group regions are the different basins.

Suppose $y_{i,e,p}$ represents the reserves in estimation group e and publication group p from operator i, with operators $i \in [1, n]$ sampled and operators $i \in [n+1, N]$ unsampled. The total reserves to be estimated for a given publication group p made up of estimation groups $e \in (1, E)$ is given by equation 3.

$$\hat{T_p} = \sum_{e=e_1}^{e_E} \left(\sum_{i=1}^{n} y_{i,e,p} + \sum_{i=n+1}^{N} \hat{\beta}_e x_{i,e,p} \right) \tag{3}$$

If there were only a single estimation group and publication group, least squares theory allows us to



Source: Energy Information Administration based on data from HPDI, IN Geological Survey, USGS
Updated: April 8, 2009

Figure 2: Conventional Gas Fields Map

estimate the variance of the error on the total as:

$$V(T-\hat{T}) = \hat{\sigma}^2 \left(\sum_{i=n+1}^{N} x_i^{2\gamma} + \left(\sum_{i=n+1}^{N} x_i \right)^2 \left(\sum_{i=1}^{n} x_i^{2-2\gamma} \right)^{-1} \right) \tag{4}$$

with

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{x_i^{2\gamma}(n-1)}$$

And the Relative Standard Error (RSE) is given by

$$RSE = \frac{\sqrt{V(T-\hat{T})}}{\hat{T}}$$

With non-enveloped estimation groups, the variance of a particular publication group will be given by the sum of the variance of each piece, as we are assuming independence in ϵ_i across observations, and so independence of $\hat{\beta}$ across different estimation groups.

$$V(T_p - \hat{T}_p) = \sum_{e=e_1}^{e_E} \left(\hat{\sigma_e}^2 \left(\sum_{i=n+1}^{N} x_{i,e,p}^{2\gamma} + \left(\sum_{i=n+1}^{N} x_{i,e,p} \right)^2 \left(\sum_{i=1}^{n} x_{i,e}^{2-2\gamma} \right)^{-1} \right) \right) \tag{5}$$

With this formula in hand, the question becomes: what is the smallest sample that will achieve RSE targets? Given a target RSE R, reference [8] shows in the case of $\gamma = 0.5$ and a single estimation group that the needed sample coverage to achieve a target RSE is a function of only three quantities: the anticipated value of of σ^2, the anticipated total y_i across all population units (T_y), and the total x_i across all population units (T_x):

$$coverage = \left(1 + \frac{1}{T_x}\left(\frac{RT_y}{\sigma}\right)^2\right)^{-1} \tag{6}$$

One option is to combine estimation groups. But the reason for different estimation groups is that there are important differences between them that the researcher wishes to account for. One could simply apply equation 6 to each estimation/publication (ep) group combination individually, using the RSE target for the publication group for each ep group, and then combine the samples from each ep group to get the sample for the publication group. This will yield suboptimal results; a particular estimation group may make up a relatively small portion of the publication group, and so sampling to achieve a low RSE in such a small area is not required. Also, if one wishes to apply a more complicated model or variance formula (for instance, reference [11] found that equation 4 is unstable and they propose several alternatives), the method proposed in reference [8] is not applicable. In the next section we present a procedure for selecting a cutoff sample in this general case.

3. Proposed Procedure

Examining equation 5, we may follow in the footsteps of reference [8] and consider the optimal sample S to be a function of the target RSE, R, the expected total dependent variable for the publication groups T_{yp}, the anticipated σ_e^2, and the full regressor data, $\mathbf{X}$. In other words there exists a function, $f : (R, T_{yp}, \sigma_e^2, \mathbf{X}) \rightarrow S$. However, this function will be completely intractable in most cases. So, we propose a methodology that iteratively adds one establishment at a time to the sample for a given publication group, always adding the operator that will reduce the anticipated RSE by the most. Explicitly:

1. Obtain anticipated σ_e^2 and β_e from Equation 1 for each estimation group.
2. Add to the sample the largest d units from each estimation group $e_1, ...e_E$, where d is the number of degrees of freedom in the estimation model. (In the present context, d=2.)
3. Consider a particular publication group p.
4. Calculate what the anticipated RSE (using Equation 5 or whatever variance formula is appropriate to the model) on T_{yp} would be if the largest unsampled unit from estimation group e_1, publication group p, (e_1, p) were added to the sample. Repeat for $(e_2, p)...(e_E, p)$.
5. Add to the sample the unit that lowered the RSE by the most.
6. Repeat 3-5 until the RSE has been brought below the cutoff.
7. Repeat 2-6 for every publication group.

This procedure attempts to minimize inefficiencies by accounting for all estimation groups within a publication group while building the sample, but does not make considerations across publication groups.

Step (1) is necessary in order to make any prospective statements about necessary sample sizes. Typically this information will come from a previous survey. Step (2) is necessary to ensure that there

are enough units sampled to estimate the model parameters. Steps (2) and (4) rely on some measure of size. In a single regressor case as in the present context, the measure of size will be x, and x is the variable that is used to identify the cutoff sample. In a multiple regression setting, one may use whatever measure of size is convenient for step (2), and then consider the largest units according to each regressor step (4). The smallest size unit sampled for a given region defines the cutoff for that region. In the case of multiple regression, the smallest size unit sampled for a particular regressor in a particular region defines the cutoff for that regressor and region, and a cutoff is defined for each regressor in each region. "Sampled for a particular regressor in a particular region" means that at some iteration of step (4) the unit was considered for sample selection because it had the largest regressor value in a particular region, and was subsequently added to the sample in step (5).

4. Results

First consider a test case of simulated data for demonstration. We simulate 100 units, with each unit having a value for each of two publication groups and two estimation groups. Data is simulated as equation 1 with $\gamma = 0.5$. β is 1 for both estimation groups, and $\sigma^2 = 1$ for $e = e_1$ and $\sigma^2 = 3$ for $e = e_2$, and ϵ_i is normally distributed. Simulated data for a single unit is shown in table 1.

ID	Estimation Group	Publication Group	X	Weight	Y
1	1	1	32.06	0.03	2.40
1	2	1	38.51	0.02	70.61
1	1	2	39.33	0.03	2.9
1	2	2	99.80	0.01	182.99

Table 1: Example Simulated Data

We use a relative standard error target of 50% for the first publication group, and 5% for the second publication group. The sample selected using the algorithm described in the previous section is of size 8. The selections are:

1. We begin with an empty sample.
2. From step (2) above, we add to the sample the top two operators from each estimation group, which adds 4 units to the sample.
3. From steps (3)-(7), no new units are required to lower RSEs to targeted levels for regions e_1, p_1, e_2, p_1, or e_1, p_2. Four units are added from group $e = 2, p = 2$ due to the combination of the low RSE target for group $p = 2$ and the high σ for group $e = 2$.

If each publication group is targeted individually, then the sample size is 22.

Next consider the case of sampling Colorado gas reserves. The EIA-23 divides Colorado into eight geological provinces/estimation groups, labeled as $A-H$ here. Estimates of β and σ^2 based on historical data are presented in Table 2, where T_{COx} is the total Colorado x for the estimation group.

Looking at the numbers in Table 2, groups B and H make up the majority of the production. It stands to reason that more respondents from those groups will be necessary. In addition, group H has both higher σ^2 and β estimates, which will increase the needed sample size. Applying the procedure

Estimation Group	T_{COx}	$\hat{\sigma}^2$	$\hat{\beta}$
A	1,225	68.2	11.08
B	832,684	55.8	9.0
C	365,067	115.6	13.0
D	6,129	25.3	7.3
E	27,906	74.9	13.3
F	14,688	99.0	10.2
G	9,347	100.0	13.1
H	764,569	90.5	13.2

Table 2: Estimation Groups

outlined in the last section, a sample of 14 respondents is necessary to achieve an expected RSE of 5% in Colorado. Group B has 5 respondents sampled, group C has 2, and group H has 4. All the remaining groups have only a single operator sampled (remember that the largest operator from each group is always sampled). Two operators sampled had production in two regions.

If each ep group is targeted at the 5% RSE level separately, adding respondents until the RSE for each falls below 5%, the sample size calculated is 51. This is a dramatic difference, understood as a result of the five smallest estimation groups making up only 3% of the Colorado total. If the analysis is restricted to the two largest estimation groups (B, H), separately sampling from the two groups yields a sample size of 12 while the proposed procedure yields a sample size of only 9.

5. Discussion

In practice, statisticians selecting cutoff samples will often use a cutoff that has been shown to give satisfactory RSEs, but the cutoff is often *not determined by the RSE*. It is specified exogenously or determined based on some other criteria. The proposed approach allows statisticians to ground cutoff sample selection explicitly on the expected variance of the final estimate for arbitrary estimation/publication groupings. Furthermore, the procedure described does not rely on a particular structure of the model or variance estimates. The critical component is that reliable estimates of model parameters are available, whether from a previous survey or otherwise.

Bibliography

[1] Y.Z. Ahmed and N.J. Kirkendall. Results of model-based approach to sampling. In *Proceedings of the Survey Research Methods Section*. American Statistical Association, 1981.

[2] J. J. Arps. Analysis of decline curves. *Trans. AIME*, 1945.

[3] K.R.W. Brewer. Ratio estimation in finite populations: Some results deducible from the assumption of an underlying stochastic process. *Australian Journal of Statistics*, 1963.

[4] N.J. Kirkendall. When is model-based sampling appropriate for eia surveys. In *Section on Survey Research Methods*. American Statistical Association, 1992.

[5] James R. Knaub. Cutoff sampling and inference. 2007. `http://interstat.statjournals.net/YEAR/2007/abstracts/0704006.php?Name=704006`.

[6] James R. Knaub. Cutoff sampling. 2008.

[7] James R. Knaub. On model failure when estimating from cutoff samples. 2010.

[8] James R. Knaub. Projected variance for the model-based classical ratio estimator: Estimating sample size requirements. In *Joint Statistical Meeting 2013*. American Statistical Association, 2013.

[9] S.L. Lohr. *Sampling, Design and Analysis*. Brooks/Cole, 2010.

[10] R.M. Royall. On finite population sampling theory under certain linear regression models. *Biometrika*, 1970.

[11] R.M. Royall and W.G. Cumberland. Variance estimation in finite population sampling. *Journal of the American Statistical Association*, 1978.

[12] Dorfman A.H. Valliant, R. and R.M. Royall. *Finite Population Sampling and Inference, A Predictive Approach*. John Wiley and Sons, 2000.