The Application of Dual System Principles to Estimate the Number of Missing Frame Elements.

Howard Bradsher-Fredrick

Energy Information Administration
1000 Independence Ave., S.W. Washington, D.C. 20585
Howard.Bradsher-Fredrick@eia.doe.gov

Introduction

The Energy Information Administration (EIA) spends significant resources evaluating and maintaining the survey frames it uses for conducting its large number of establishment surveys. Evaluating survey frames can involve many factors, including conducting element by element comparisons to existing survey frames used by other Federal agencies and organizations.

One such approach is to directly compare the frame presently employed by EIA with a frame attempting to contain some or all of the same frame units. Such a comparison has several benefits:

1. Effectively updating EIA frames based upon the information available from the alternate frame.
2. Effectively updating the frame of the cooperating agency based upon EIA's frame information.
3. Providing sufficient information so that estimates can be made regarding the number of units missing from both frames.

A significant literature has developed regarding benefit #3 and the evolution of the methodology for estimating the number of elements missing from both frames. This methodology was originally known by various names including, "Capture-Recapture Methodology." It originated in wildlife biology (Seber, 1982) and demography (El-Khorzaty et al., 1977). It has been adapted in the field of epidemiology to provide estimates based upon two incomplete sources and to refine incidence estimates. It has also been used more broadly to estimate the completeness of apparently exhaustive survey frames.

This paper will present the methodology, required assumptions, computations and results of the application of the dual system principles to the NREL/EIA frame comparison.

Methodology

In its efforts to effectively evaluate and maintain its own survey frames, EIA has also employed dual system estimation techniques. One such recent effort was to compare a listing of renewable electric plants compiled by the National Renewable Energy Laboratory (NREL) with a similar frame employed by EIA. Based upon the number of units common to both frames, or in the EIA frame but not the NREL frame, or in the NREL frame but not the EIA frame, an estimate of the number of units in neither frame could be estimated. Moreover, the total frames could be stratified by fuel type and similar analyses could be conducted to estimate the number of units missing from both frames for each individual fuel type.

Assumptions

It should be remembered that in order to properly apply this methodology, there are two necessary assumptions that probably do not apply to either of the examples:

- The frames need to be independently assembled. Since NREL largely uses secondary sources for their frame and their data, they probably use EIA for some of their frame information. Due to frequent staff turnover at NREL, they were unable to estimate their reliance on EIA data and frame information at the time of this comparison study.
- The missing frame information should be random rather than systematic. This appears not to be the case. For example, 140 out of 214 of the missing elements (65%) in the EIA frame are "Timber Residues."

Example 1

However, as a first exercise the principles were employed to all of the aggregated frame elements as follows.[1]

NREL Frame # of Elements:		5,146
EIA Frame # of Elements:	4,960	
Elements Common to Both:		4,932
Elements in EIA frame not in NREL:	28	
Elements in NREL not in EIA frame:		214
Total number of Elements:	X	

The total number of known elements can then be computed as follows:

Total Count of # of Elements: 4,932 + 214 + 28 = 5,174

In tabular form, the data would appear as follows:

Frame Elements	**In NREL**	**Not in NREL**	**Totals**
In EIA Frame	4,932	28	4,960
Not in EIA Frame	214	?	?
Totals	5,146	?	X

Using the error rates in the assembly of both frames, we can solve for X[2],

X = Known Total in NREL * Known Total in EIA / Known Elements in Common
X = 5146 * 4960 / 4932 = 5,175.21
Rounding X (discrete elements) = 5,175

Thus, we estimate that one facility was missed in both frames (a total of 5,175 elements versus 5,174).

Example 2

We can apply the principle to all of the renewable fuels individually. The only fuel where we obtain a value other than the total known count is with agricultural residues. Thus, we can apply the dual system principles to a second example. The known data are as follows:

NREL Frame # of Elements:	33
EIA Frame # of Elements:	27
Elements Common to Both:	19
Elements in EIA frame not in NREL:	14
Elements in NREL not in EIA frame:	8
Total number of Elements:	X

The total number of known elements can then be computed as follows:

Total Count of # of Elements: 19 + 14 + 8 = 41

In tabular form, the data would appear as follows:

[1] Only renewable facilities (including hydro) above 1 MW nameplate capacity in both cases are being analyzed, although NREL collects information on all facilities without the 1 MW cutoff.

[2] For a derivation of the formula, see Report No. 22 (2005) of the Statistics Commission of the U.K., Appendix C: Derivation of the Dual System (DSE) Estimator.

Frame Elements	In NREL	Not in NREL	Totals
In EIA Frame	19	14	33
Not in EIA Frame	8	?	?
Totals	27	?	X

X = Known Total in NREL * Known Total in EIA / Known Elements in Common
X = 27 * 33 / 19 = 46.89
X = Rounding X (discrete elements) = 47

This estimate of 47 facilities in the frame is to be contrasted with the known total of 41 (19 + 14 + 8). Thus, an estimated 6 facilities are missing from both frames.

Summary and Conclusions

The assumptions implicit in employing the dual system estimation technique are sufficiently strict to severely limit its appropriate usage. However, the real-world examples shown provide a useful illustration as to how the technique might be employed if the stringent assumptions were in fact met. It appears as though reasonable results could be obtained through the use of the methodology.

Future research on the subject should involve the use of the technique when the assumptions appear to hold. This would provide more substantial support for the utility of the method in the frame size estimation process. Additionally, basic research should involve developing modified techniques to enable the analyst to employ the technique under less stringent conditions. For example, if the assumption on independence of the two frame element collection efforts could be relaxed, the technique could find more widespread appropriate usage. Moreover, if the assumption on randomness regarding the missed elements in the frame could be relaxed, this would also enable the analyst the opportunity for more frequent application.

References

Berger, J.O., 1985. Statistical Decision Theory and Bayesian Analysis, 2nd edition, Springer-Verlag, New York, N.Y.

El-Khorzaty M., Imreg P.B., Koch G.G., Wells, H.B. 1977. "Estimating the Total Number of Events with Data from Multiple-Record Systems: A Review of Methodological Strategies." International Statistical Review 45: 129 – 157.

Seber, G.A., 1982. The Estimation of Animal Abundance and Related Parameters, 2nd edition. London, Charles Griffin.