EIA 914 Survey – Investigation of Alternate Methods and Data Sources

Prepared for:

Meeting of the American Statistical Association (ASA)
Committee on Energy Statistics with the
Energy Information Administration (EIA)
October 1 and 2, 2009
Washington, D.C.

Prepared by:

ICF International
Bill Pepper, Asa Janney, and Harry Vidas

This is a working document prepared for the Energy Information Administration (EIA) in order to solicit advice and comment on statistical matters from the American Statistical Association Committee on Energy Statistics. This topic will be discussed at EIA's fall 2009 meeting with the Committee to be held on October 1, 2009.

This report summarizes our progress to date on improving the sampling and estimation procedures of the EIA 914 survey and provides a list of next steps. The memorandum is divided into four major sections: Terminology and Data Sources, Sampling, Estimation, and Conclusions and Next Steps. The memorandum presents alternative approaches in the Sampling and Estimation sections.

Terminology and Data Sources

The primary source of data used in this analysis is the HPDI production data summaries provided by EIA. EIA provided two sets of data, one of which includes HPDI's most recent reports of production from 1997 to 2008 for six states and an aggregate number for the rest of the lower 48 state. The other data set includes reports received from October 2007 through 2009 of six states' individual production during every month from January 2000 to July 2009. Each month-by-state production report changes as the operators complete their filings (which could take up to two years) and as the ownership of difference production properties change over time.

Change of ownership creates problems that require special attention, and the problems encountered in our analysis here (see below) may be addressed through alternative and perhaps more precise handling of the ownership relationships in the more detailed HPDI data available to EIA.

For this analysis we use the follow terminology to describe the processing and some of the issues associated with the HPDI data:

- Sampling month – the month used to choose the sample for the estimation month
- Production month – this is the month when the production occurred.
- Reporting month – this is the month in that the HPDI reports production is made. In our data sets, all historic production from any given property (e.g., a well or lease) is listed under the current operator in a specific reporting month even if the operator did not own the property in the production month.
- Estimation month – this is the month for which EIA is estimating production.
- Reported production – the quantity of production (million cf/day) in the production month that has been reported to the State or MMS and is contained in the HPDI data
- Sampling Age – the time difference (in months) from the sampling month to the estimation month
- Reporting Age - the months of delay in the operators reporting production to the state/MMS plus the delay in HPDI getting those data into their database

Sampling

There are a number of issues associated with the selection of the sample. Currently, EIA annually uses the HPDI data from 2 years before to select the operators to sample for the coming calendar year for the existing cutoff survey. Approximately 25 operators (about 10 percent of the total sample) are replaced in the sample annually and the share of production covered by the sampled operators has often shown a significant decline over those two years (that is, "drift"). Furthermore, the large change in the operators sampled in January can create discontinuities in the production estimates of adjacent months (see estimation below).

The selection of the operators can be improved in a number of areas.

Sampling Option 1: Use More Recent Data for Selecting the Cutoff Sample

Our evaluation of the HPDI data provided by EIA suggests that more recent data, or lower sampling age, can be used to select operators in the cutoff sample, thus reducing drift in the estimates. In 94% of the cases, an operator selected to be in the cutoff sample using reported production for 6 months ago would also be in the cutoff sample based on data for 24 months ago. Also, 99% of the time if it was not in the cutoff sample using a 6-month report age, it would not be in the cutoff sample using a 24-month reporting age.

The volatility of the set of operators in the cutoff sample also does not increase significantly if a shorter sampling age is used. For example, the number of operators that are in the cutoff sample and stay in the cutoff sample increaes by only 0.8% if the sampling age increases from 6 months to 18 months.

In addition, 90% of the most recently reported production for the six states for which we have data is captured by the sample using a 6-month sampling age. In fact, increasing the sampling age does not significantly improve the coverage of the sample but it does improve the completeness and consistency with final reported values by about four to five per cent. In the month where the cutoff sample is selected. However, the reported production for production months less than 6 months from the reporting month are significantly incomplete and not suitable for use in selecting the cutoff sample.

Sampling Option 2: More Frequent Sample Selection for the Cutoff Sample

One option to improve the coverage of the cutoff sample is to increase the frequency of sample updating. We did tests with the HPDI data where, on average, 200 operators are included in the cutoff sample. If we use a sampling age for sample selection from 14 to 23 months we achieve a minimal change month to month in the sample with around two operators leaving and two operators entering the sample on a monthly basis. Using a lower sampling age for sample selection increases the number of operators entering and leaving the sample each month to an average of over 4. This increase is due primarily to the inaccuracies of the reported production with the lower sampling age. It is likely that this number could be significantly reduced if an operator were not allowed to enter or leave the sample until the pattern in production persisted for a set number of months (perhaps three or four), but we did not test this.

Sampling Option 3: Probability Proportional to Size Sampling

This option is to be addressed in the Estimation Section of the Memorandum.

Sampling Option 4: Stratified Sampling

Another option considered is a stratified sample with 3 strata, the first consisting of a cutoff sample similar to the existing procedure, the second stratum including samples down to 3 million cubic feet per day, and the third stratum including all other operators. For the second stratum, the sampling can be random or proportional to size.

The purpose of the stratified sampling is to provide an alternative to the scaling of the reported production for the sampled operators or forecasting of the production for the some of the nonsampled operators and to reduce the drift in the final reported production estimates. The results of this approach are provided in the Estimation section of the Memorandum.

Sampling Option 5: Modeling to Estimate the Final Reported Production from the HPDI Production Data

Another option we considered is to use modeling to estimate the final reported production numbers for HPDI production data for the sample months when the sampling age is low and the HPDI reported numbers are still incomplete. This may provide a better basis for estimating production in the estimation month and reduce the scaling that is required to estimate production for nonsampled operators in the estimation month and should help improve the selection of the operators to be in the sample each month..

This modeling is possible because the HPDI reported production for selected production months follows a predictable pattern by state as a function of the reporting age. The reported production data are revised a number of months which varies by operator and state until the final reported production stabilizes for a selected production month. The pattern in terms of the reporting lag and how quickly the final production values are achieved varies by operator but can be modeled by operator using the following equation which reflects the asymptotic behavior of the pattern:

- $Dly_d = 0$ if $d<MD$
- $Dly_d = (1-Dly_{d-1})*Shp + Dly_{d-1}$ if $d \geq MD$

Where

- Dly_d is reported production as a fraction of final reported value (fraction), d is the reporting lag (months);
- MD is the estimated typical reporting lag until any production is reported (months); and
- Shp is a shape variable between 0 and 1 which characterizes how quickly the estimates converge to the final value.

Figures 1 and 2 show the coverage by state of the HPDI estimates for 4 select months, for the six states where data where provided, and for a number reporting lags. As seen in Figure 2, the estimates for the Federal Gulf (FG) region for operators not in the cutoff sample are highly variable, which is likely due to changes in ownership.

Figure 1: HPDI Reported Production as a Percentage of the Final Reported Production and the Reporting Lag for the Cutoff Sample

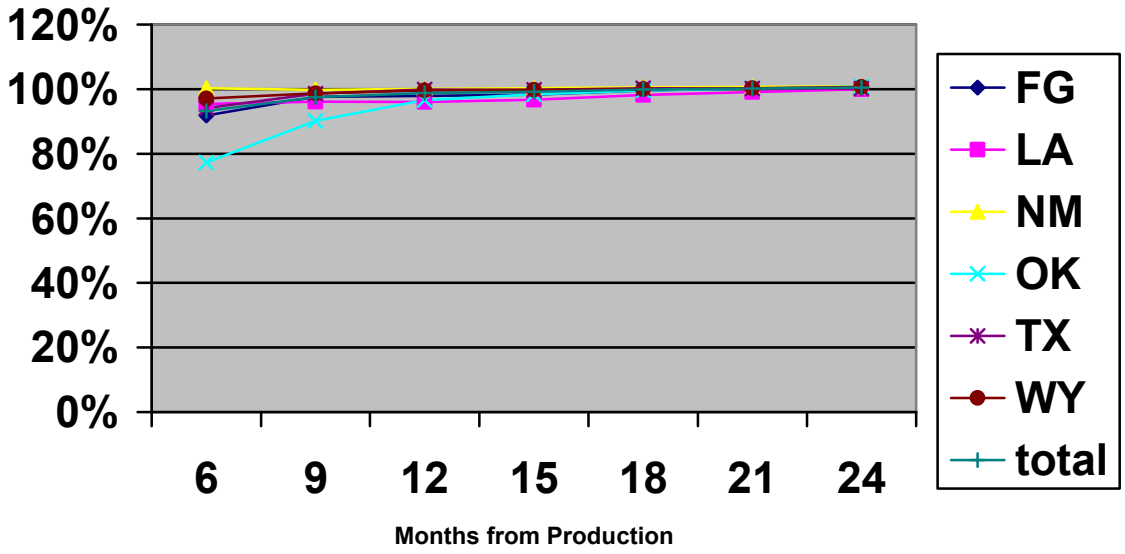
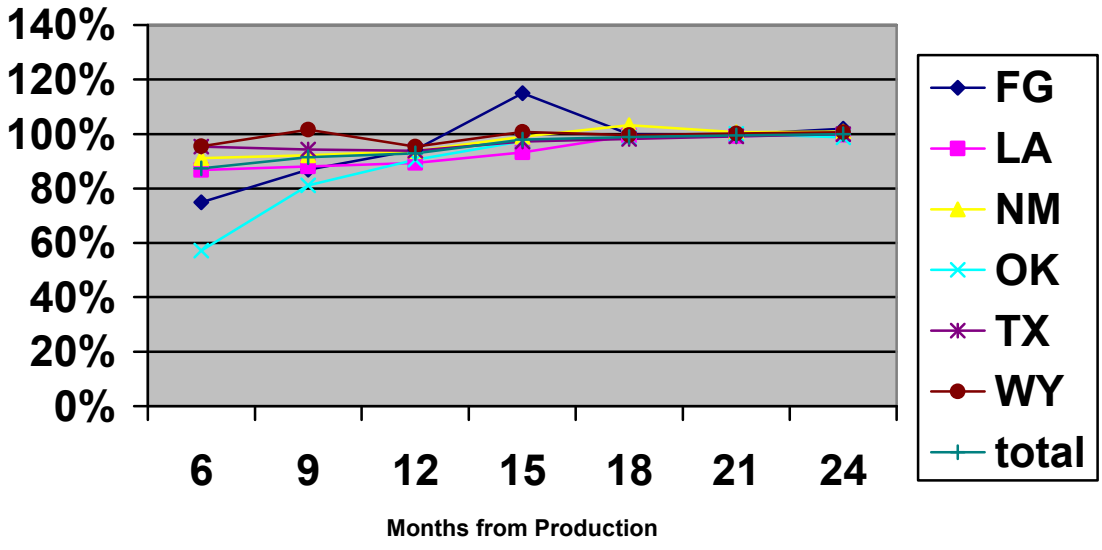


Figure 2: HPDI Reported Production as a Percentage of the Final Reported Production and the Reporting Lag for Operators not in the Cutoff Sample



For testing, we modified the equation to address ownership changes in the HPDI data which we feel can be improved with more detailed modeling or data adjustments using the underlying and more detailed HPDI

data. We calculate the values of the MD, $Scale_f$ and Shp variables for each operator for each state that minimizes the following:

- $\sum_{t,t'}(RProd_{pm,rm} - Size_{pm} * Dly_d * Scale_{t'})^2$

where:

- pm is the production month (number of months from December 1993).
- rm is the reporing month.
- d is the number of months from the production month to the reporting month and is equal to is set to rm – pm.
- $RProd_{pm,rm}$ is ithe HPDI reported production in production month pm for reporting month rm.
- $Size_t$ is the maximum production for the operator in production month t for any reporting month regardless of ownership.
- Dly_d, MD, and Shp are defined earlier.
- $Scale_{rm}$ is a scaling factor used to account for changes in ownership and changes in ownership that occur from one reporting month to the next.

Table 1 shows the number of operators that solves the equation with different values of MD by state. For example, in Texas 1,744 operators would be expected to report production 2 months after the production takes place. As the table shows, the reporting of data starts within 4 months from over 96% of the operators for the six states. The reporting lags are the highest in Oklahoma.

Table 1: Number of Operators by State with Various Months of Delay Before Reports of Production

Delay, MD	FG	LA	NM	OK	TX	WY
2	129	499	214	520	1744	118
4	29	39	48	287	48	78
6		1		33	3	1
8	1			10	3	
10				5	2	1
12			1	4	3	
14	1	1		6	3	1
16		2		10	2	
18	1			24	17	

Table 2 shows the number of operators that have different values of the SHP variable by state. A SHP value of 1.0 indicates that once production is reported, the reported value is complete while a SHP value of 0.125 indicates that the reported production is only 12.5% of what eventually will be reported. Again, the pattern suggests the largest delays in full reporting are in Oklahoma.

Table 2: Number of Operators with Solved Values of SHP by State

Shape	FG	LA	NM	OK	TX	WY
1	21	114	46	17	582	40
0.875	29	57	87	6	206	36
0.75	20	58	30	77	195	31
0.625	14	64	16	32	105	22
0.5	14	50	12	75	94	13
0.375	13	34	15	129	78	6
0.25	14	27	8	220	103	10
0.125	36	138	49	343	462	41

The production volume for operators that solves with different SHP values is provided in Table 3. Reporting in New Mexico converges the quickest once reporting starts, and reporting in Oklahoma is the slowest.

Table 3: Share of Production Volume for Operators with Selected SHP Variables

Shp	FG	LA	NM	OK	TX	WY
1.000	7%	35%	7%	0%	13%	14%
0.875	12%	18%	81%	0%	40%	21%
0.750	25%	9%	6%	3%	16%	29%
0.625	10%	11%	1%	6%	6%	21%
0.500	22%	11%	4%	6%	7%	6%
0.375	8%	11%	1%	39%	17%	0%
0.250	5%	4%	0%	35%	1%	8%
0.125	11%	1%	0%	10%	1%	0%

The accuracy of the models is fairly good with the standard deviation of the error between the HPDI reported production and the estimated production from the adjusted model for reporting ages of at least six months ranging from 0.5% for New Mexico to 5.5% for Oklahoma with a six state average of 2.2%[1].

Estimation

We looked at a number of different approaches to estimating total production including variations on the current approach which involve scaling of the cutoff sample, PPS sampling, stratified sampling, and use of the cutoff sample with estimation of the residual production.

Finally, we found a solution to the discontinuity (or "lurch") problem of bridging production estimates from December to January across a resample with the current estimation process used by EIA.

In all of the simulations of the estimation approaches, we used the adjusted model from sampling option 5 above to randomly expand the historic coverage of our reported production numbers to months before October 2007.

Estimation Option 1: Scaling the Cutoff Sample

This option is the most similar to the current process, but it allows for monthly resampling and uses a historic series to estimate the scaling factor to apply to the sampled production values. In this case, we pick

[1] All averages are volume weighted.

a sampling age from six to twenty four months, revise the operator sample, and revise the factor used to scale the sampled production to the final production for each state. The scaling uses a historic series of scale factors to scale the sampled production to get total production but accounts for the share of production reported in the sample month from the operators in the sample.

The general scaling factors and production are given as follows:

- $SR_{m,d} = (RProd_{m-d,m,S} + RProd_{m-d,m,NS})/ RProd_{m-d,m,S}$
- $FR_{m,em} = (RProd_{m,em,S} + RProd_{m,em,NS})/RProd_{m,em,S}$
- $AF_{em,d} = \sum_{em-24 \leq m < em-12}\{FR_{m,em}\}/\sum_{em-24 \leq m < em-12}\{SR_{m,d}\}$
- $EProd_{em,NS} = SProd_{em,S}*(AF_{em,d}*SR_{em,d} -1)$

Where

- em is the estimation month.
- m is a month prior to the estimation month.
- d is the sampling age, and m-d would be the sampling month for the month *m.*
- S indicates the reported production or estimate production is for the group of operators in the sample.
- NS indicates that the reported production or estimated production is for operators not in the sample.
- $RProd_{m-d,m,S}$ is the reported production for the sampling month *m-d* made in the reporting month *m* for the sampled operators.
- $RProd_{m-d,m,NS}$ is the reported production for the sampling month *m-d* made in the reporting month *m* for the nonsampled operators.
- $SR_{m,d}$ – is the ratio of the reported production for all operators to the reported production for the sample operators with production in month m and with a sampling age of d, and is used to adjust the scale factor from month to month.
- $FR_{m,em}$ – is the ratio of the reported production for all operators to the reported production for the sample operators for production in the month *m* with the reporting month equal to the estimation month *em.*
- $AF_{em,d}$ is an average scalar for 24 to 12 months prior to the estimation month of the FR scalar divided by the SR scalar.
- $SProd_{em,S}$ is the reported production from the operators in the sample for the estimation month.
- $EProd_{em,S}$is the estimated production for all operators for the estimation month.

This option tends to be of the most accurate overall, although Estimation Option 2, below, is better for some states such as New Mexico (NM) and Wyoming (WY) and shows improvement over the current process except for Oklahoma. Table 4 shows the simulated standard error of the estimates as a percentage of the actual volumes. Oklahoma would benefit by using a sampling age of more than 6 months, such as 9 months, to select the sampled operators and determine the scaling factor. Also, the standard error of the forecast for Oklahoma may be overstated by the overall error in estimating the actual production for the operators in the state which uses the adjusted model from sampling option 2.

Table 4: Estimation Option 1 Standard Error of Forecast

Delay	FG	LA	NM	OK	TX	WY	Starting Month the Simulation which last 20 Months[2]
6	1.9%	1.7%	0.5%	2.9%	1.8%	1.4%	Apr-07
7	1.7%	1.7%	0.5%	2.6%	1.7%	1.4%	Mar-07
8	1.8%	1.4%	0.4%	2.6%	1.8%	1.3%	Feb-07
9	2.2%	1.4%	0.5%	2.2%	1.9%	1.4%	Jan-07
10	1.6%	1.4%	0.5%	2.2%	2.0%	1.5%	Dec-06
11	1.6%	1.6%	0.5%	1.8%	1.9%	1.5%	Nov-06
12	2.0%	1.6%	0.6%	2.2%	2.0%	1.7%	Oct-06
13	2.1%	1.8%	0.7%	1.9%	1.9%	1.4%	Sep-06
14	1.9%	2.0%	0.7%	1.6%	1.9%	1.5%	Aug-06
15	2.1%	2.2%	0.6%	1.6%	1.7%	1.5%	Jul-06
16	2.1%	1.9%	0.6%	1.6%	1.7%	1.4%	Jun-06
17	2.0%	1.9%	0.7%	1.5%	1.7%	1.5%	May-06
18	1.9%	2.2%	0.7%	1.7%	1.7%	1.8%	Apr-06
19	1.8%	2.1%	0.8%	1.6%	1.5%	1.7%	Mar-06
20	2.2%	1.9%	0.8%	1.5%	1.5%	1.9%	Feb-06
21	2.2%	2.1%	0.8%	1.4%	1.4%	2.0%	Jan-06
22	1.7%	1.9%	0.8%	1.8%	1.3%	2.0%	Dec-05
23	2.0%	1.5%	1.0%	1.7%	1.2%	2.1%	Nov-05
24	1.9%	1.9%	1.0%	1.7%	1.3%	2.2%	Oct-05
Current Cutoff Sample	2.2%	3.9%	2.1%	1.5%	2.0%	1.9%	

Estimation Option 2: Estimating Production for Nonsampled Operators Using HPDI Reported Production for these Operators

In estimation option 2, we use the HPDI reported production for nonsampled operators for the sampling month to estimate production for nonsampled operators in the estimation month. This approach uses a simple scaling of the production from nonsampled operators based on historic estimates of the ratio of the production from nonsampled operators in the estimation month to the nonsampled operators in the sampling month.

This approach uses the following equations:

- $ANF_{em,d} = \sum_{em-24 \leq m < em-12}\{ RProd_{m,em,NS} \} / \sum_{em-24 \leq m < em-12}\{RProd_{m-d,m,N}\}$
- $EProd_{em,NS} = RProd_{em-d,em,NS} * ANF_{em}$

Where

- em is the estimation month.
- d is the sampling age .
- $RProd_{m-d,m,NS}$ is the reported production for the sample month *m-d* made in the reporting month *m*for the nonsampled operators.
- $ANF_{em,d}$ is a (ratio of reported production) of nonsampled operators for the estimation month divided by the reported production of nonsampled operators for the sampling month which is the *em-d*.

[2] In these tests, the sampled operators are updated each month

- $EProd_{em,NS}$ is the estimated production for the nonsampled operators for the estimation month.

Table 5, below shows the forecast error for this approach. In general, the approach does slightly worse than option 1 except for states such as New Mexico and Wyoming but shows improvement over the current process. Option 1 does better for the Federal Gulf, Louisiana, and Oklahoma primarily because of the increased error in the HPDI reported production for the nonsampled operators for shorter lags from the sampling month to the reporting month.

Table 5: Estimation Option 2 Forecast Error

Delay	FG	LA	NM	OK	TX	WY	Starting Month[3]
6	2.1%	2.0%	0.5%	3.3%	1.9%	1.3%	7-Apr
7	1.8%	2.0%	0.5%	2.9%	1.8%	1.3%	7-Mar
8	1.9%	1.9%	0.5%	3.2%	1.9%	1.2%	7-Feb
9	2.4%	2.0%	0.5%	3.0%	2.0%	1.3%	7-Jan
10	1.8%	1.8%	0.5%	2.7%	2.0%	1.3%	6-Dec
11	2.0%	1.9%	0.5%	2.1%	2.0%	1.3%	6-Nov
12	2.4%	2.1%	0.5%	2.7%	2.1%	1.5%	6-Oct
13	2.6%	2.3%	0.5%	2.5%	2.0%	1.3%	6-Sep
14	2.3%	2.4%	0.5%	2.2%	2.0%	1.4%	6-Aug
15	2.6%	2.7%	0.5%	2.2%	1.9%	1.4%	6-Jul
16	2.7%	2.3%	0.4%	2.2%	1.9%	1.3%	6-Jun
17	2.4%	2.4%	0.5%	2.1%	1.9%	1.4%	6-May
18	2.6%	2.6%	0.5%	2.1%	2.0%	1.7%	6-Apr
19	2.2%	2.6%	0.5%	2.2%	1.9%	1.5%	6-Mar
20	2.7%	2.6%	0.5%	2.0%	1.9%	1.7%	6-Feb
21	2.8%	2.6%	0.5%	2.1%	1.9%	1.8%	6-Jan
22	2.3%	2.6%	0.5%	2.1%	1.8%	1.9%	5-Dec
23	2.6%	2.3%	0.6%	2.1%	1.8%	1.9%	5-Nov
24	2.3%	2.5%	0.6%	2.2%	1.9%	2.0%	5-Oct
Current Cutoff Sample	2.2%	3.9%	2.1%	1.5%	2.0%	1.9%	

Estimation Option 3: PPS Sampling

Sampling with probability proportional to size (PPS) is good alternative sample design to investigate whenever the population is skewed, as the distribution of gas production certainly is. A PPS sample design was considered when the 914 survey was set up, but it was not accepted because of the lack of an complete, accurate sampling frame. At that time the only data available were from the EIA-23 annual survey on gas reserves, which lacked recency and data on production. EIA agreed that the availability of the HDPI data make it worthwhile to reexamine PPS sampling.

[3] In these tests, the sampled operators are updated each month for 20 months, and this date represents the starting month of the simulation

We ran two simulations of PPS sampling using the HPDI data with lags of 24 and six months between the sampling month and the estimation month. Twenty four months was chosen because that is the gap currently used by the DFO for the first January after a resample. Six months was chosen because that is the minium lag that one could use and expect to have HPDI data that were complete and stable enough to be used as a sampling frame. At six months the HPDI data are clearly not mature, but we used the mature HDPI data in the simulation to get a lower bound for the error rate with that lag time. If the error rate with the mature data is higher than that of alternative designs, there is no point in adjusting the error further upward to account for the incompleteness or inaccuracy of the sampling frame. If the error rate comes in below that of some other methods, the vintage HDPI data can be used in the simulation (with additional work) to adjust it upward to the level that would obtain in practice.

Cutoff sampling would have an advantage when sampling from the immature six-month-old HPDI data because it needs mature data only for the largest producers, which seems to be available after six months. In some applications PPS has the advantage of not needing an inflation factor, but the simulation results give us a way to see whether this seems to be one of those cases.

The simulation results for the 24-month case are rather startling at first look. For the Lower-48 forecast (the only one we examined) the coefficient of variation (ratio of standard error to actual) in the error is 8.92 percent, which compares poorly to the 0.73 percent of the current design. So PPS looks like a loser just on this comparison.

It is interesting to investigate how the PPS error rate could be so large. It seems that the migration of some small producers to large ones over the 24 months from the sample data to data collection causes most of the problem. Occasionally, just often enough to really affect the results, a really small producer becomes a really big producer in that interim. There are cases of production going up almost 1000 times. For example,that could happen when production goes from 0.01 mmcf/day to 10 mmcf/day, which would not be atypical for a new producer who was just starting to produce in the sampling month.

One such producer can easily make the L48 estimate have over 100-percent error, even after its effect is averaged in with 224 other contributors. The problem is that he is not typical. Each such small producer in the sample represents lots of similar producers, which was correct in the sampling month; the population is skewed and has lots of small producers. But a producer with a sharply rising production may be the only one of the producers it represents who ran its production up to 1000 times as much over the last two years. The problem is the lack of recency in the data used to compute the weights. PPS samples in real life are usually based on an available contemporaneous variable that is highly correlated with the variable of interest. For example, to estimate the total income of a community you could use assessed value of houses as a size variable for your PPS sample. That value is relatively stable and current. We know of no such counterpart that is available for gas producers who are not currently in the 914 sample.

That result should kill the notion of using PPS with a 24-month lag, but we document another drawback for completeness. The simulation also shows that this PPS design also suffers from drift, just like the cutoff sample. Again, the culprit is the lag time, but in this instance it is the drift of the big producers that causes underestimation. At 24 months, the estimated L48 total averages 0.60 percent low. If the sample is used for a year, at 35 months, this slips further to 0.97 percent low. This is caused by the decline in some big producers between the sampling month and the estimation month. PPS favors the large producers and picks lots of them. As we know from our other research, the largest group tends to lose members to rising stars. So some of the big producers have weights at sample time that are too large for what they are producing by estimation time. These big weights get applied to reduced production amounts, and the projection tends to come in low, and gets lower over time. So, PPS with a long lag does not enjoy not needing an inflation factor.

With a six-month lag, PPS produces a lower-bound estimate of the standard error of 1.41 percent for the L48 with very little bias. Despite the large reduction in the error rate, it is still much bigger than the 0.73 percent of the current method using a 24-month lag. Of course, using the vinataged HPDI data at six months would show an even greater disadvantage.

In conclusion, We think a PPS design is still inferior to a cutoff sample, especially now that we have discovered that the cutoff sample can be based on more recent, immature HPDI data. The PPS sample is at a disadvantage using the immature HPDI data because it needs a complete and accurate list of the population to obtain a low error rate.

Estimation Option 4: Stratified Sampling

We also tested a stratified sampling approach in a way that is comparable to the other tests above. In this case, we used the 12 years of HPDI data provided by EIA with a delay-from-sampling age of up to 24 months. Our approach used three strata. The first stratum is a cutoff sample with the cutoff being 30 million cf/day for total production and 15 million cf/day for Oklahoma. These higher limits were chosen to reduce the number of samples in the first stratum (cutoff sample). The second stratum used a PPS approach on operators with volumes over 3 million cubic feet per day (not in the first stratum) for the state (and 1.5 million cubic feet per day for Oklahoma). The third stratum includes all the other operators and is not sampled. The estimation for the third stratum is a simple scaling of the reported production from the sample month to get the production for the estimation month.

The average number of operators sampled equaled 221. The overall results of the estimation are provided in table 7 and are compared to the average error from the current cutoff sampling process that uses a similar number of operators in the sample:

Table 7: Stratified Sample Forecast Error

	Average Error Over 10 years	Standard Deviation	Current Cutoff Sample Process
TX	-0.35%	2.85%	2.01%
OK	-0.91%	3.20%	1.52%
NM	-0.52%	2.11%	2.08%
LA	-5.20%	3.51%	3.86%
WY	-0.38%	2.26%	1.89%
FG	-0.57%	1.49%	2.23%
OT	-0.94%	1.73%	2.25%
Total	-0.90%	1.06%	0.73%

Estimation Option 5: Solution to the Discontinuity Problem for the Current Method

The discontinuity problem happens when a new annual sample is drawn, and the published estimate of total production changes in a direction or magnitude that is inconsistent with a reconstruction of what a constant sample would have done. After the fact, one can obtain HPDI data on the historical production of producers who entered the sample and get continuing reports on those who left the sample, and combine them to get the total production of those who stayed, entered, and left – a constant sample. The inflation factor that one

would apply to such a sample in January would be little changed from December, so the change in the production of a constant sample should be closely parallel total production of the population. Thus, the two being inconsistent is suspicious.

In a memo entitled "EIA-914 Sample and Model Issues", Gary Long described the discontinuity between the December 2008 and the January 2009 Texas estimates. Total reported production rose by 400 MMcf/day, but estimated total production, based on the inflation factors, fell by 360 MMcf/day. Oddly, if reported production had been based on a constant set of producers (constructed in retrospect), that would have yielded a rise of only 60 MMcf/day in reported production, which indicates that the true rise in total production was neither a sharp rise or sharp fall. Such inconsistencies happen because of errors in one or both of the inflation factors – the old-sample factor used in December and the new-sample factor in January. Most of the error is probably in the old factor as it is based on sample data with a greater lag.

The old inflation factor can have errors because, for example, the rapidly rising production of fewer producers than expected could have put them above the cutoff while fewer than expected producers in the sample could have drifted below the cutoff. In such a case the inflation factor would be higher than needed to compensate for these drifts. When the new sample swapped the new big producers for the fading ones, the large reduction in the inflation factor offset the rise in reported production too much, and the estimate of total production fell instead of rising slightly.

In our memo entitled "Derivation of the Ratio of Inflation Factors" (30 July 2009), we derive a formula that relates what the ratio of the old inflation factor, IF_8, to the new inflation factor, IF_9, should be:

- $IF_9 = IF_8 (SGS + SGL) / (SGS + NSL)$

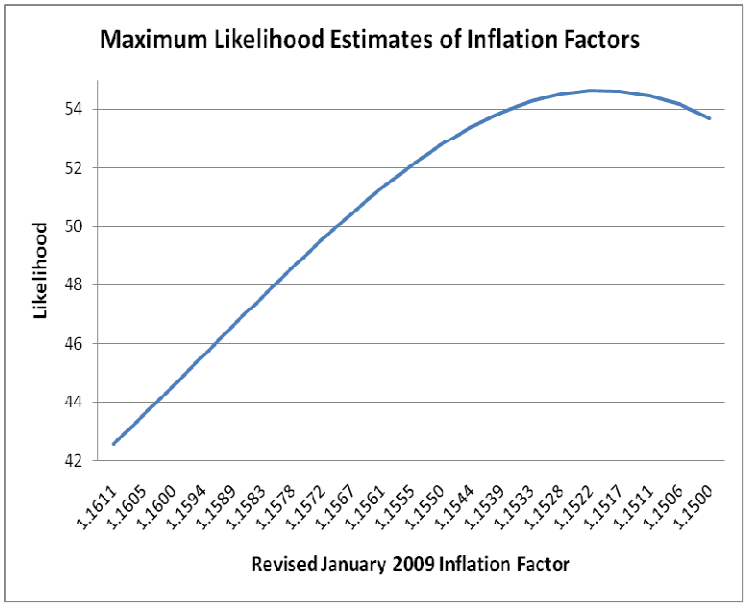
where

- SGS = Sample Group that Stays
- SGL = Sample Group that Leaves
- NSL = Nonsample that Leaves

Notice that nonsampled producers that remain unsampled, whose production is unknown in the short run, are not in the equation. Thus. every variable in the equation is observed.

This equation represents the extra information we can bring to bear on the problem. It shows how we expect the estimates to behave across the change of sample. We now have a problem that has three equations in two unknowns (the true values of the old and new inflation factors). The problem is to combine those three equations into one set of two estimated inflation factors. We did that using a maximum-likelihood technique. This can be applied because the two IFs are estimated with regressions, so they come with standard deviations. For the Texas example above, they rise at the rate of 0.000485 MMcf/day/month. The old IF is 29.5 months from the center of the data it is based on, so its standard deviation is 0.0143 MMcf/day, and the January IF is 18.5 months old, and has a standard deviation of only 0.0090 MMcf/day. We can reduce the value of the old IF and increase the new one and find a combination that satisfies the equation above and maximizes the likelihood based on the estimated standard errors of the IF's. The figure below illustrates the solution.

Figure 3



The solution is that the old IF should fall from 1.19 to 1.171 while the new one should rise from 1.15 to 1.1522. The application of these factors revised the estimated totals to 22.07 MMcf/day in December and 22.18 MMcf/day (a slight rise) in January. So, the solution is consistent with the implication of a constant sample and seems satisfactory. It shoud be possible to automate this procedure if it is adopted for routine use.

Conclusions and Next Steps

As shown in Table 8 below. our interim conclusions from this analysis is that some combination of estimation options 1 and 2 would be best. The selection of the sample should be more frequent than current . The sampling age should be less than 24 months and as low as 6 months but should vary by state. The selection of the sample should allow for an operator to be above the cutoff production level for a set number of months before it enters the sample and below the cutoff production level for a set number of months before it is removed from the sample to reduce the chance that operators will “bounce” into and out of the sample.

The modeling of the HPDI reported data can be improved and help with the sample selection but the evaluation of this possibility requires more analysis and simulation using the more detailed HPDI data. ICF is also reviewing the simulations to see where they can be improved.

Table 8
Form 914 Sampling and Estimation: Current Process and Alternatives

Estimation Option	Sample Type	Resampling Frequency	Age of Data Used for Sample	Production as Reported or Estimated	Population Total Projection Method	Frequency	Standard Error of L48 Estimate
Current Process	Cutoff	Annual	24 to 36 months prior	Reported	Inflation factor based on trend regression	Inflation factor estimated once a year and includes a coefficient to account for changes over the year	0.73 Percent for the lower 48 forecast with State errors ranging from 1.5 Percent to 3.9Percent
Option 1 - Scaling of the Cutoff Sample	Cutoff	Monthly	Anywhere from six months to twenty four months and varying by state	Reported	Scale 914 sampled production by scalar estimated using historic HPDI or other data where appropriate	Scaling factor is recalculated each month	Anywhere from 0.3 Percent to 2.0 Percent improvement over current process except for Oklahoma where it is 0.5 to 1.0 Percent worse

Option 2 - Estimating Production for Nonsampled Operators Using HPDI Reported Production for these Operators	Cutoff	Monthly	Anywhere from six months to twenty four months and varying by state	Reported	Scale the reporting production for the nonsampled operators based on a scaling factor calculated using historic HPDI or other data	Scaling factor is recalculated each month	Anywhere from 0.1 Percent to 2.0 Percent improvement over current process except for Oklahoma where it is 0.5 to 1.0 Percent worse
Option 3: PPS Sampling	PPS	Not part of analysis	24 & 6 months prior	Reported	PPS Weights	Monthly	1.41 Percent for the National Forecast
Option 4: Add extra stratum	Cutoff for stratum 1; PPS for stratum 2; Scaling for stratum 3	Not part of analysis	Up to 24 months prior	Reported	Scale factor & PPS Weights	Scaling factor for stratum 3 is recalculated each month	1.06 Percent for the National Forecast
Recommended Process	Cutoff	Monthly with constraints on the number of months an operator either exceeds or is below the production target before they are added or removed respectively from the sample	As low as 6 months prior	Reported	Mix between Option 1 and Option 2 based on the suitability by state	Scaling factor is recalculated each month	N/A