

Background on Form EIA-914: The Monthly Natural Gas Production Survey

(Draft Posted: Thursday, April 21, 2005)

NOTE: Revisions *may* be made available at a later date. Please contact Kara Norman (kara.norman@eia.doe.gov) if you have any questions.

Abstract

At our meeting last fall, members of the American Statistical Association (ASA) Committee on Energy Statistics may recall that the Energy Information Administration (EIA) introduced the new and *approved* Natural Gas Production Survey, Form EIA-914. This spring meeting will provide us the opportunity to discuss EIA's preliminary impressions and the chance to preview the initial results from the first and second reporting months of this new survey. EIA will discuss the challenges we faced during the early months of operation, as well as our plans for future analysis to compare the results of this survey with our previous methods of estimating natural gas production. This session will be presented by John Wood of the Reserves and Production Division (RPD) of EIA.

Summary

Through the end of 2004, natural gas production estimates were based on data EIA collected from gas producing states and data collected by the U. S. Minerals Management Service (MMS) in the Department of Interior. The states and MMS collect this information from producers of natural gas for various reasons, most often for revenue purposes. However, reporting was neither complete nor timely for all states.

Thus is born Form EIA-914, a monthly survey that has been designed to fill this data need. Survey respondents are a cut-off sample of well operators who report on Form EIA-23, "Annual Survey of Domestic Oil and Gas Reserves". The respondents to the EIA-23 are selected annually from the survey frame for the EIA-23, and the sample of the EIA-914, also selected annually, is a subset of the sample for the EIA-23.

The survey began collecting production in the January 2005 reporting month. Data is collected by and reported for the six largest gas-producing areas (Texas, the Federal Gulf of Mexico, New Mexico, Louisiana, Oklahoma, and Wyoming) and the U.S. Total. Respondents are currently permitted 60 days to report, but that will be reduced to 40 days in the near future. The accuracy target for the survey is 1% at the U.S. level and 2.5% by area. By the end of 2005, data are expected to be available for publication 60 days after the close of the reference month, but currently no data has been published.

EIA's Response to Committee Suggestions from Fall 2004

In the Fall 2004 ASA Energy Committee Session, EIA sought feedback on our interpretation and analysis used to select our estimation procedure. We also requested

recommendations for handling non-response or apparent errors in reporting for the large operators. Lastly, we asked if the committee could offer any suggestions for detecting or handling outliers and overly influential operators and how we should deal with these when making estimates.

The main recommendation by the committee was that EIA use the Hajek estimator as an unbiased alternative to the variations of the Horowitz-Thompson estimator that EIA was utilizing, which were biased. The committee also suggested that EIA eliminate using any estimators that used both certainty and non-certainty operators in the estimation procedure, which we did. EIA responded to the former committee suggestion by incorporating the Hajek estimator into the simulation program, but ultimately we opted for a hybrid heuristic approach for estimating natural gas production due to the dynamic nature of the data.

The committee also had concerns with the 90% cut-off sample. There was discussion about how EIA was going to estimate variance. One proposed method was to look at a window of time and analyze how the estimation actually performed when the data actually came in. The committee was also apprehensive about the random perturbation of the 90% cut-off sample. EIA responded to the committee suggestions by recognizing that EIA's estimation procedure attempts to calibrate for the random perturbation. EIA also closely monitors the issue and calibrates the procedure as needed.

Recent History of Form EIA-914

October 2004: OMB approved Form EIA-914.

November 2004: A cut-off sample of the top 286 operators based on the operator's natural gas production in the lower 48 states is selected.

December 2004: EIA makes first contact with the respondents via letters requesting respondent information.

January 2005: A trained staff of six begins to call operators who did not respond to the contact information request.

February 2005: Form EIA-914 is sent to respondents by email and mail. Of the nine respondents who submitted their January data in February, only one respondent got their submission correct and that was because they only needed to report zeros!

March 2005: 240 more of the expected total of 286 operators filed this month. The majority of them had errors of one kind or another (usually due to units). A reminder is sent to those who have not filed their January submissions.

April 2005: 20 more respondents finally filed for January. 82 respondents have already filed for February. Of the remaining respondents left to file for January, one says we will get [his submission] when we get it. Another says to send the marshals otherwise he is

not filing. Many will not answer the phone anymore. We continue to try to get the data. They are mostly small companies with the exception of two respondents.

May 2005: EIA may release January and February estimates at the end of May, but we hope to first convince ourselves of the reasonability of these new estimates.

Methodology (Review of the Current Estimation Procedure)

This model uses a sample of approximately 300 operators, less any non-respondent operators and operators with extraordinary changes discovered by pre modeling/estimating edits, to model and estimate the non-sampled group of operators. (The non-respondent operators and operators with large changes could be included after any imputations and individual estimates.)

The basic and relatively general relationship is this.

$$T_i = \sum_{j=1}^{N_i} y_{i,j} = \sum_{j=1}^{m_i} y_{i,j} + \sum_{j=m+1}^{n_i} y_{i,j} + \sum_{j=n+1}^{N_i} y_{i,j}$$

T_i is the total production for month i (the sum of the production of all N_i operators). This summation can be broken into three summations: the first summation is the group of operators in the sample used to estimate the non-sampled operators ($j=1$ to m) in month i ; the second summation is the group of operators in the sample that are not used to estimate the non-sampled operators ($j=m+1$ to n) in month i ; the third summation is the group of operators that are not sampled ($j=n+1$ to N_i) in month i .

Together, the first two summations are the sampled group of operators. The N_i is not constant from month to month. If all sampled operators are used in the first summation, then the second summation is zero. A different group of the sampled operators can be used for each month i . Companies can enter and leave the data set as they start or stop producing.

The third summation, the production of the non-sampled operators, needs to be estimated. The following model assumes this monthly production estimate for non-sampled operators depends on the monthly production from the sampled operators.

$$\sum_{j=n+1}^{N_i} y_{i,j} = R_{i,j} * \sum_{j=1}^{m_i} y_{i,j}$$

$R_{i,j}$ can be a simple ratio or a more complex function which considers the changing production percentages of both the sampled and non-sampled operator groups over time. In the calibration year we could simply divide the sample by 0.9 and get the total. However, two years later the production percentage of the sample group has declined while the non-sampled group has increased.

The function used for $R_{i,j}$ in this model is the following.

$$R_{i,j} = R_m * (1 + A_m * t)$$

If the calibration year is 2000 and the survey is the 12 months of 2002 (a two year lag), the parameter t (time) ranges from 13 to 24 for the 12 months of the survey year. The parameters R_m and A_m are in the range of 0.12 and 0.01 respectively.

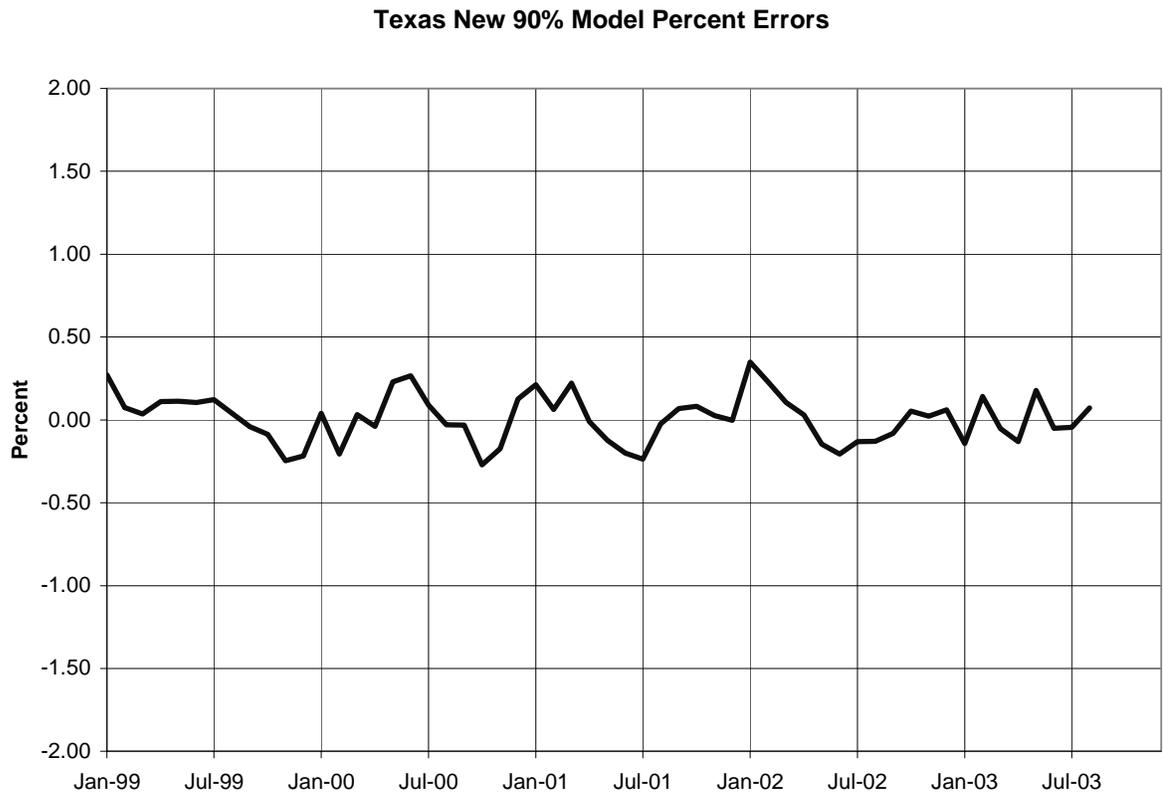
The model can be written as follows.

$$\hat{T}_i = \left(\sum_{j=1}^{m_i} y_{i,j} \right) (1 + R_{i,m}) + \sum_{j=m+1}^{n_i} y_{i,j}$$

For a 90 percent sample in Texas for each of the calibration years 1997 through 2001 this model was fit for the survey years 1999 through August of 2003. The fit yielded values for the parameters R_m and A_m for the survey years as follows.

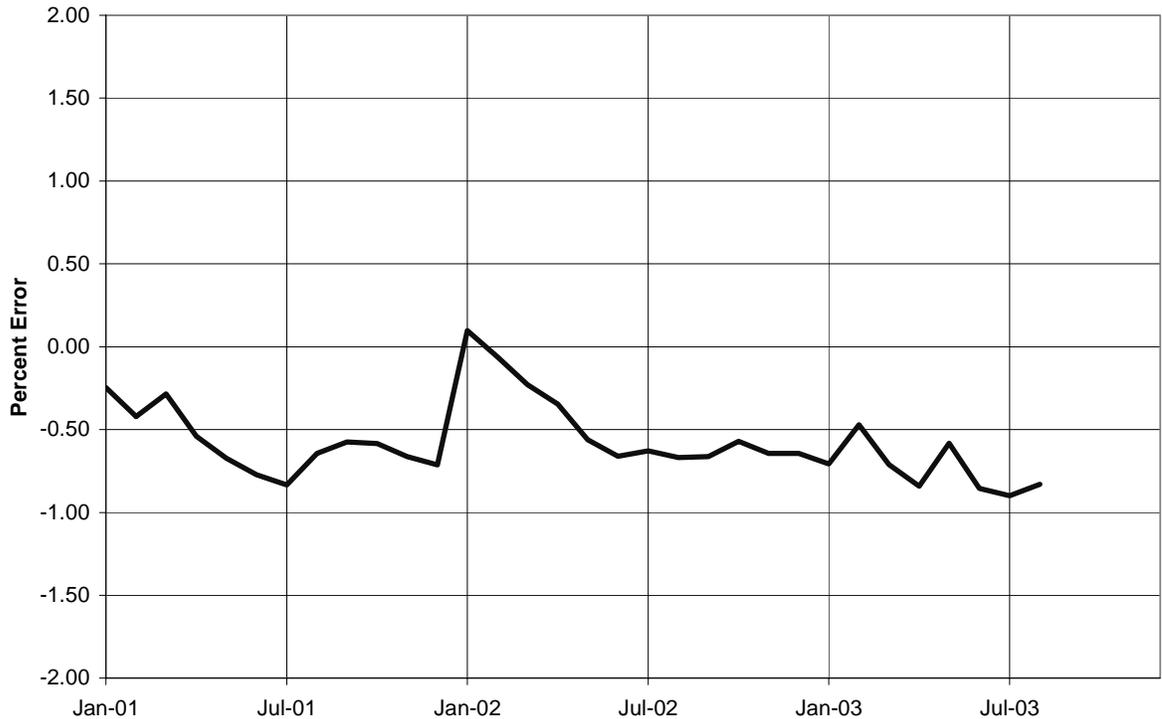
	R_m	A_m
1999	0.1185	0.0033
2000	0.1219	0.0035
2001	0.1180	0.0066
2002	0.1141	0.0099
2003	0.1191	0.0102 (partial year)

The percent errors for this fit (calibration) to the actual data are shown in the following graph.



In order to test the model we used the parameters determined two years prior to estimate current production. In Texas we may only need a one year lag but a two year lag is a good test. The resulting production estimates are slightly low but within 1 percent. The average error is -0.6 percent. We can change the apparent systematic bias which results from the changing A_m 's over time by shortening the lag or using a more complex $R_{i,j}$ function.

Texas New 90% Model Test



Implementation of Form EIA-914

As you have read above, this is a brand new survey only recently implemented. Because of this, we wish to save the latest and great data for our presentation. To give you a preview of what we plan to discuss, though, we will outline some of those topics now.

There were interesting challenges, inherent to the early phase of any survey, to overcome. These issues include response rates, submission errors, and other survey startup matters. We are also currently examining and preparing to analyze the newest 'unofficial' estimates provided by Form 914. To discuss these estimates, our presentation will make use of historical data and absolute error ranges. We will present the current suggestions for methods we plan to employ to determine the reasonability of this new data, emphasizing limitations of comparable data availability, specifically the problem posed in Texas, a state which has now implemented a new method of data collection via internet, which renders the previous multinomial estimation methodology – 'Cristal's Model' – inadequate, at least for the present time.