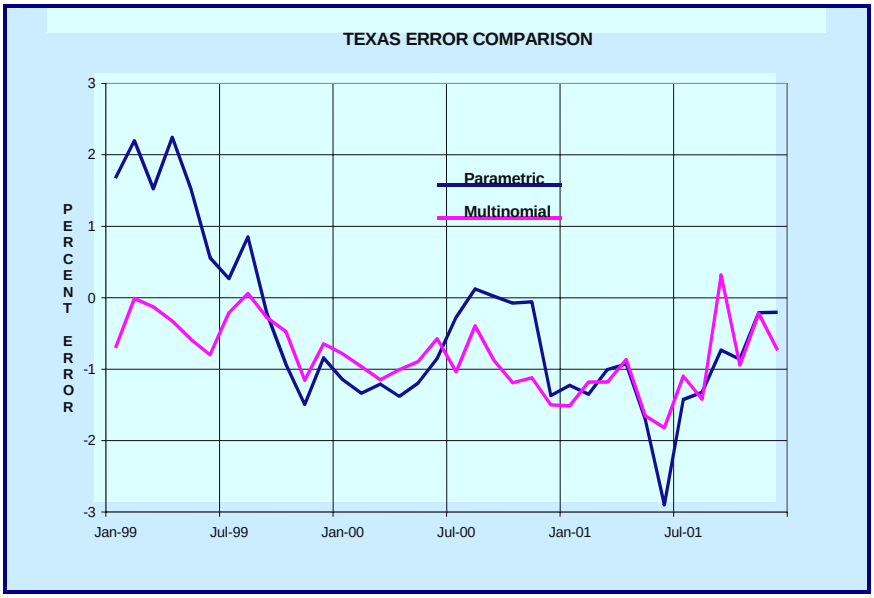
Comparative Evaluation of Two Methods to Estimate Natural Gas Production in Texas

December 2003



Office of Oil and Gas
Energy Information Administration

Table of Contents

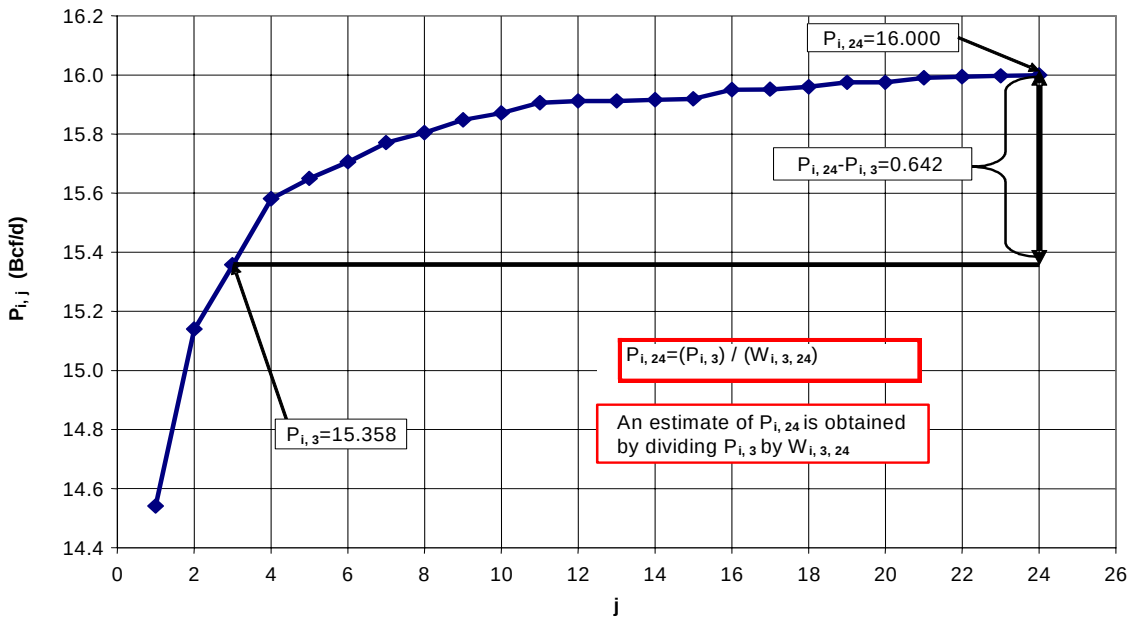
Summary

This report describes an evaluation conducted by the Energy Information Administration (EIA) in August 2003 of two methods that estimate natural gas production in Texas. The first method (parametric method) was used by EIA from February through August 2003 and the second method (multinomial method) replaced it starting in September 2003, based on the results of this evaluation.

EIA publishes State-level natural gas production estimates monthly and annually. Texas is the largest producing state, (27% of US production in 2001) and timely Texas production information is very important to EIA and its customers. The principal source of Texas natural gas production data (measured as gross withdrawals) is the Texas Railroad Commission (TRC). Natural gas production values (collected and processed by the TRC) are posted on the TRC website between 45 and 60 days after the close of a report month. The initial values are then regularly revised monthly for about 24 months, and sporadically thereafter.

Posted production values P_j (for a given report month j) typically start out low and approach their "final" values after many months. Figure 1 illustrates this reporting

Figure 1. Texas Gross Natural Gas Production in May 2001 (As Reported Over 24 Months)



pattern for May 2001. The diamonds indicate the sequence of reported values from the initial low level (for j=1) to the almost final value two years later (j=24). EIA's goal has been to obtain estimates of Texas production for publication in the *Natural Gas Monthly* (*NGM*) 120 days after

the close of the reference month. For this purpose, the value after three months of reports (j=3), $P_{i,3}$, is used as the basis for estimating the "final" value $P_{i,24}$ (j=24). Both methods that were evaluated estimate the weight $W_{3,24}$ (defined in Figure 1) in making the estimate.

The previous estimation technique uses a parametric model which produces estimates that usually have been within one or two percent of the "final" production value. A second method was proposed (multinomial model) that is statistically rigorous and showed promise of improving estimation performance. Both methodologies use the same data and require the same data preparation. The two methods are described detail in Appendices 2 and 3, respectively.

The evaluation approach is contained in Appendix 1 and the detailed evaluation results are contained in Appendix 4. The evaluation results were the basis for the recommendations at the end of this summary, which were accepted. The American Statistical Association Advisory Committee on Energy Statistics provided recommendations, which are contained in Appendix 5.

Previous Method (Parametric Model)

This estimation model uses the 24-month historical data revision pattern as a template to estimate final production values from preliminary data. Reported volume data approach their final reported values according to a relatively stable pattern (exemplified in Figure 1). This revision pattern is determined from history, for which 24 months of revisions are available for a given production month. The modeled historical pattern is then applied to recent production months, for which reported production may have been revised up to 23 times (i.e., every subsequent month for two years). The model also attempts to account for changes in the relationships between the preliminary data and final data over time. The estimates provided by the parametric model (120 days following the production month) have usually been within one or two percent of the "final" production value (obtained after 24 months of data have been received). A detailed discussion of the parametric model is in Appendix 2.

Alternative Method (Multinomial Model)

An alternative method was proposed to estimate Texas natural gas production that is based on a multinomial distribution model of the reporting patterns observed in the data. The model assumes that all the gas produced in a month will be reported in one of the subsequent 24 reporting months. This multinomial distribution provides a rigorous basis for computation of maximum likelihood estimates for the final production in a month given preliminary data, for calculation of prediction intervals, and for model improvement should reporting patterns change. The model assumes that the reporting patterns remain constant over the most recent m months. The model was run with m = 6 and 9, with more accurate results occurring with m = 6. A detailed discussion of the multinomial model is in Appendix 3.

Evaluation Results

The multinomial model was compared to the parametric method to assess improvements in accuracy and predictability. Both methods were run for three years' worth of data – calendar years 1999, 2000, and 2001. The key results of the comparison are summarized in Table S-1 and shown graphically in Figure S-1; details are provided in Appendix 4. Expert review comments on this work were provided by the American Statistical Association Advisory Committee on

Energy Statistics (Appendix 5) and by Dr. Fred Joutz, professor of statistics at the George Washington University (Appendix 6).

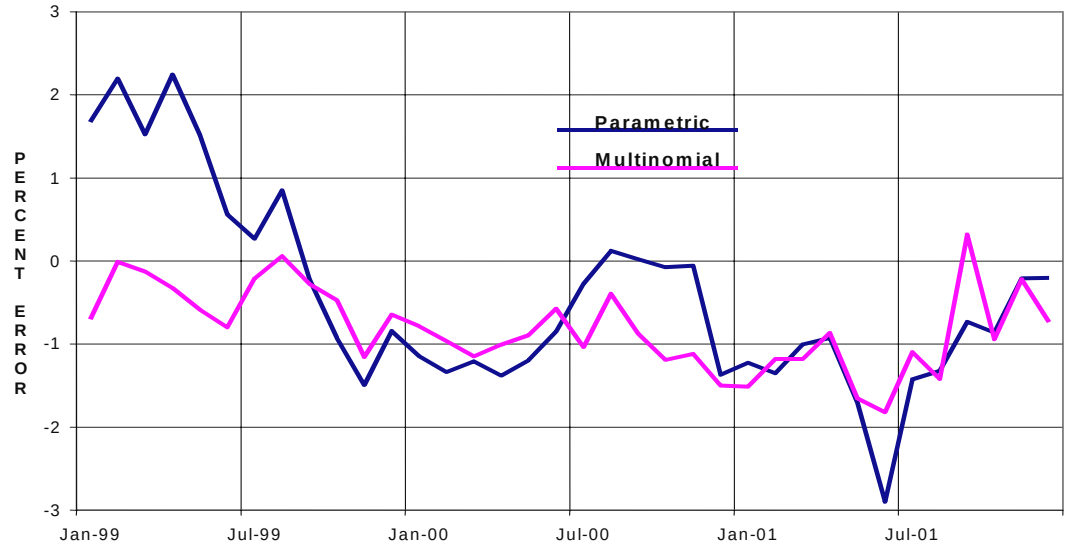
According to EIA's Information Quality Guidelines[1] all estimation methods are to be transparent and reproducible, and provide high quality estimates in a timely manner. The summary statistics from the comparison address the quality of estimates produced by the two methods. Transparency can be assessed by reviewing the descriptions of the two methods (Appendix 2 and Appendix 3) for clarity and understandability. Reproducibility is achieved by maintaining archived versions of the exact code and input data used to produce estimates. Using the same procedure regularly without manual intervention enhances reproducibility.

Table S-1. Summary of Evaluation Results

Time Period	Statistical Measures of Error	Multinomial Model	Parametric Model
1999	Average Error (%)	-0.46%	0.42%
	Mean Absolute Deviation (%)	0.50%	1.01%
	Root Mean Squared Error (%)	0.61%	1.12%
	Max Error	0.16%	1.64%
	Min Error	-1.15%	-1.49%
2000	Average Error (%)	-0.89%	-0.76%
	Mean Absolute Deviation (%)	0.89%	0.77%
	Root Mean Squared Error (%)	0.99%	0.96%
	Max Error	-0.23%	0.10%
	Min Error	-1.77%	-1.45%
2001	Average Error (%)	-0.81%	-1.17%
	Mean Absolute Deviation (%)	0.83%	1.17%
	Root Mean Squared Error (%)	0.92%	1.35%
	Max Error	0.09%	-0.20%
	Min Error	-1.53	-2.90%
1999-2001	Average Error (%)	-0.72%	-0.50%
	Mean Absolute Deviation (%)	0.74%	0.98%
	Root Mean Squared Error (%)	0.86%	1.15%
	Max Error	0.16%	1.64%
	Min Error	-1.77	-2.90%

[1] http://www.eia.doe.gov/smg/EIA-IQ-Guidelines.html

Figure S-1. Percent Error of Parametric and Multinomial Methods



Both models provide natural gas production estimates for gross withdrawals that are usually accurate to within one percent. For the 36 months from 1999 to 2001, the multinomial model with m=6 had 28 estimates with errors less than 1%, 34 with errors less than 1.5% and 36 months with errors less than 2%. In contrast, the parametric model had 19 estimates with errors less than 1%, 31 with errors less than 1.5%, and 35 with errors less than 2%. (Data are shown in Appendix 4.) The summary results show that the multinomial model has a lower mean squared error and a lower mean absolute deviation, and the magnitude of the largest error is smaller. However, the multinomial model appears to have a slight negative average error of about -0.72%. That is, the multinomial model tends to underestimate final production by a small amount.

The smaller estimates of variation (mean squared error and mean absolute deviation) indicate that the multinomial model is more accurate. In addition, the multinomial model provides mathematical theory for the reporting pattern that allows for the estimation of prediction intervals. The assessment showed that all 36 estimates for the multinomial model were within the 90% prediction intervals.

It is suspected that the small bias associated with the multinomial model is due to the assumption that reporting probabilities stay constant over six months. The data clearly show that there are increasing delays in the reporting of production in the State of Texas. In the future, EIA plans to investigate alternative methods to reduce the bias. EIA hopes that a relatively simple enhancement to the multinomial method can be developed to remove the bias from the estimates.

The multinomial model can be executed in about five minutes and is not expected to require application of expert judgment. Prior to August 2003, use of the parametric model required the setting of model parameters and about an hour for execution. Setting these model parameters

required expert judgment, and the model produces different results depending on the parameters chosen.

Recommendations

The following recommendations were made at the conclusion of the evaluation and accepted by EIA.

- Implement the multinomial model with m=6 to estimate Texas monthly natural gas production, starting with *Natural Gas Monthly* published in September 2003.
- Evaluate methods to minimize the slight negative bias of the multinomial model and determined a recommended methodology by January 30, 2004 (or sooner if the bias becomes statistically significant.)
- Evaluate model performance annually in conjunction with the preparation of the *Natural Gas Annual.* Any resulting model changes would be approved according to the Evaluation Guidelines described in Appendix 1.
- If model results for a particular month appear adversely affected by unexpected events (e.g. missing data or revisions, changes in Texas reporting procedures), EIA will modify the estimates only with the approval of a review team consisting OOG's Reserves & Production and Natural Gas Divisions, Statistics and Methods Group, Energy Markets and End Use, and Integrated Analysis and Forecasting.

Appendix 1. Evaluation Guidelines

EIA's Information Quality Guidelines (whttp://www.eia.doe.gov/smg/EIA-IQ-Guidelines.html) state that all estimation methods are to be *transparent*, *reproducible*, and provide *timely* and *accurate* results. The goal of the evaluation was to determine which of competing methods more closely meets these requirements. If in the future new or improved methods are proposed, they will be evaluated in the same manner.

The process of an evaluation had several steps. First, the Office of Oil and Gas would prepare the documentation described below, and compute the summary statistics described below. The information would be assembled into an evaluation report. Second, the report would go through EIA Category I clearance to obtain peer review within EIA and to gain concurrence on the preferred method. The results and review findings would be provided to the Administrator for final approval.

The requirements specified in the Information Quality Guidelines would be assessed as follows:

- **Transparent:** As part of the Category I review of the model, reviewers would be asked to review model documentation to provide an assessment of the transparency and reproducibility of the two methods.

- **Reproducible**: The chosen methodology would be used to produce estimates of natural gas production for publication in the *Natural Gas Monthly* until such time as a new methodology is adopted via the procedure outlined in this section. Methodology would be documented for each estimate and the code and data used to generate each monthly estimate would be archived.

- **Timely:** Estimates for State level data should be completed within an agreed-upon number of days after the close of the reference month to allow timely publication in the *Natural Gas Monthly*. Documentation should demonstrate that this requirement is satisfied and may be improved upon in the future. The current timeliness goal is publication in the *Natural Gas Monthly* 120 days after the close of the reference month.

- **Accurate**: Accuracy would be assessed by comparing volumes "estimated" for a month to the best final monthly data for the three most recent years for which reasonably final data are available. In addition, the estimates published in the *Natural Gas Monthly* and the *Natural Gas Annual* would be compared to the best final data available for the same three-year period. As of July 2003 the final monthly data for January 1999 through December 2001 would provide the basis for comparison. The following specific guidelines would be followed in assessing accuracy:

    1. To the extent possible, volumes "estimated" for a month would be computed using only the data that would have been available at the time the estimate would have been prepared for use in the *Natural Gas Monthly*. For example, for the state of Texas, the "estimated" data for the month of "January 1999" could include only the "P_1, P_2, and P_3"

data from the State of Texas for January 1999, the "P_1, P_2, P_3 and P_4" data for December 1998, etc.

2. If it were impossible to use only the data available at the time the estimate would have been prepared, all parties would agree to alternative data sources in advance.

3. Alternative estimation methods would be run using exactly the same data sets (a separate data set is needed for each month from January 1999 through December 2001), and predicted or estimated values of natural gas production will be computed for each month.

4. For each alternative estimation method, the estimated monthly values from step 3 would be compared to the final monthly values from Step 1. The "error" would be computed as the final value minus the estimate. The percent error is the error, multiplied by 100 and divided by the final value.

5. The following summary statistics would be computed:

Error By Year:	Average, Mean Absolute Deviation, Mean Squared Error, Max, Min
Percent Error By Year:	Average, Mean Absolute Deviation, Mean Squared Error, Max, Min
Error for 3 years:	Average, Mean Absolute Deviation, Mean Squared Error, Max, Min
Percent Error for 3 years:	Average, Mean Absolute Deviation, Mean Squared Error, Max, Min

6. Time series plots comparing the final data with estimates prepared using alternative methods would also be prepared. A comparison of these statistics by the Category I reviewers would allow reviewers to assess which method produces more accurate results.

Appendix 2. Parametric Model

Background

The Texas Railroad Commission posts gross natural gas production data on its website (www.rrc.state.tx.us/divisions/og/information-data/stats/ogismcon.html) and revises the aggregated data regularly over 24 months, with small revisions occurring sporadically for years. The parametric estimation model is a spreadsheet model that uses this 24-month historical data revision pattern as a template to estimate final production values from preliminary data. Reported volume data generally start out low and approach their final reported values according to a relatively stable pattern (curve). The revision pattern is based on 24 months of revision history for a given production month. The historical pattern is then applied to recent production months to estimate potential total production volumes. The model also attempts to account for changes in the relationships between the preliminary data and final data over time.[2]

For production months with at least seven pieces of information (six revisions), the model works very well. For production months with fewer than seven pieces of information (the six most recent months with 0 - 5 revisions) some additional controls or parameters are used.

Data Preparation

The data are entered into a sheet in columns of monthly vintages. Each month a new column is entered with the first report for the current month and revised reports for all previous months. The data are then organized into columns of first reported data, second reported, third reported . . . (i.e., first preliminary, second preliminary, third preliminary . . .columns of data) referred to as $P_{i,1}$, $P_{i,2}$, $P_{i,3}$, etc. These data, organized by "P's," are the fundamental input for the model.

Basic Model

The fundamental model equation is below.

$$BF_{i,j} = \frac{P_{i,j}}{1 + MC_{i,j} + E_{i,j}}$$

where $MC_{i,j} = -\left(1 - \frac{P_{i,j}}{BF_{i,j}}\right)$

$BF_{i,j}$ = Best Final estimated production
$P_{i,j}$ = Preliminary reported production data
$MC_{i,j}$ = From the smoothed lagged 6 month median model of $MC_{i,j}$
$E_{i,j}$ = Error, amount not accounted for by $MC_{i,j}$

[2] Prior to the development of the model described in this section (in February 2003), an average historical month-to-month change was used to estimate Texas monthly gas production. The average was taken from 2 - 8 years of historical month-to-month changes for the particular month for which the estimate was being determined. The analyst selected the specific years used to calculate the average month-to-month change. The average was applied to last month's estimate or successive averages were applied beginning with the latest close-to-final reported production data.

i = Production month
j = Number of the preliminary estimate for production month i

$MC_{i,j}$ Model

The $MC_{i,j}$ model is fit first. Since this term is based on a smoothed six month lagged, six month median, calculated value of $MC_{i,1}$, the $BF_{i,j}$ fit parameter can be used to calculate the $MC_{i,j}$ term which is then used to determine a later $BF_{i,j}$ parameter and so on. This "cascading" through the historical data carries the revision pattern forward through the current month's estimate, i.e., $BF_{i,1}$ from $P_{i,1}$.

The $MC_{i,j}$ model is based on the smoothed six month lagged, six month median, calculated value of $MC_{i,1}$ term as a starting point for the revision pattern. The Z term allows the revision pattern to change over time as the relationships between the preliminary values and the final values change.

$$MC_{i,j} = \frac{MC_{i,1}}{\left(1+\left(j-1\right)*Z\right)^{C}} \qquad \text{For j = 2 to 24} \qquad \text{Where } Z_i = A*\left(1+B*MC_{i,1}\right)$$

A = 0.678
B = 8.367
C = 3

The $MC_{i,j}$ model is fit over about six years of historical data where the $P_{i,24}$ values are available. For this historical period $BF_{i,j}$ is equal to $P_{i,24}$. The fit parameters A, B, and C are determined by a least squares fit. The $MC_{i,j}$ model with its determined fit parameters A, B, and C is cascaded from the historical data fitting period down through the most recent months of reported data.

$BF_{i,j}$ Model

The $BF_{i,j}$ model determines the Best Final production value for up to 24 simultaneous equations for each production month. The Basic Model equation is rearranged as follows and a least squares fit is used to minimize the difference between modeled Ps and actual Ps. The $BF_{i,24}$ becomes a fit parameter. The additional control mentioned above for the first six months appears here as the error term $E_{i,j}$ (see 2ndSheet tab in TexasModel29.xls workbook).

$$P_{i,j} = BF_{i,24}*\left(1+MC_{i,j}+\sum_{j}^{24} dE_{i,j}+E_{i,j=1\text{to }6}\right)$$

$$dE_{i,j,j+1} = FC_{i,j} - FC_{i,j+1} + MC_{i,j+1} - MC_{i,j}$$

$$FC_{i,j} = \frac{P_{i,j}}{BF_{i-1,24}} - 1$$

This model is used without the error term from $P_{i,7}$ through $P_{i,24}$ and with the error term from $P_{i,1}$ through $P_{i,6}$. Where $P_{i,24}$ is available, $BF_{i,24}$ is set equal to $P_{i,24}$. Otherwise, $BF_{i,24}$ is a fit parameter.

$P_{i,1}$ through $P_{i,6}$ Models

For $P_{i,1}$ through $P_{i,6}$, a separate model is used for each $P_{i,j}$ to estimate the error term. For each estimate the current $P_{i,j}$ and all previous $P_{i,j}$'s and previous $BF_{i,j}$ estimates are used. Most of the fit parameters are in the error term (described below). $BF_{i,j}$ is a fit parameter and minimized in a least squares fit everywhere a $P_{i,24}$ is available in each $P_{i,j}$ model and all modeled $P_{i,j}$'s are minimized against actual $P_{i,j}$'s in the same least squares fit (see 3rdSheet tab in TexasModel29.xls workbook).

For example, the $BF_{i,3}$ estimate for the $P_{i,3}$ reported data uses $P_{i,1}$, $P_{i,2}$, and $P_{i,3}$ with an error term model to determine the $E_{i,3}$ term. The $P_{i,1}$, $P_{i,2}$, and $P_{i,3}$ model equations, listed below, and the error functions are simultaneously fit.

$$P_{i,1} = BF_{i,3} * \left(1 + MC_{i,1} + dE_{i,1,2} + dE_{i,2,3} + E_{i,3}\right)$$

$$P_{i,2} = BF_{i,3} * \left(1 + MC_{i,2} + dE_{i,2,3} + E_{i,3}\right)$$

$$P_{i,3} = BF_{i,3} * \left(1 + MC_{i,3} + E_{i,3}\right)$$

The error term is defined as follows:

$$E_{i,3} = FE_{i,3} * \left(FC_{i,3} - MC_{i,3} - PC_{i,3}\right)$$

Where

$$PC_{i,3} = \frac{BF_{i,2} - BF_{i-1,24}}{BF_{i-1,24}}$$

If

$$W_{i,3} = \left(FC_{i,3} - MC_{i,3} - PC_{i,3}\right)$$

For $W_{i,3} < 0$

$$FE_{i,3} = 1 - \frac{1 - A_{i,3}}{1 + \left(BN * |W|\right)^2}$$

For $W_{i,3} >= 0$

$$FE_{i,3} = \frac{A_{i,3}}{1 + \left(BP * |W|\right)^2}$$

Where

$$A_{i,3} = A * e^{\left[D^* \left(1 - \frac{0.005 + |dE_{i,1,3}| + |W|}{0.005 + |dE_{i,1,3}|} \right) \right]}$$

A = 1.244
BN = 73.36
BP = 156.4
D = 0.395

The $P_{i,1}$ through $P_{i,6}$ Models are fit or minimized in sequence as part of an iterative process with the $BF_{i,j}$ Model. The $P_{i,1}$ through $P_{i,6}$ models are fit sequentially because each one depends on the $BF_{i,j}$ from the previous one. The last error term from each $P_{i,1}$ through $P_{i,6}$ model is used in the last six terms or months of the $BF_{i,j}$ model described above. The $BF_{i,j}$ Model is then fit using the supplied error terms ($E_{i,1}$ through $E_{i,6}$). Because the $P_{i,1}$ through $P_{i,6}$ Models are also dependent on the results of the $BF_{i,j}$ Model the $P_{i,1}$ through $P_{i,6}$ Models are fit again sequentially. The last error term from each $P_{i,1}$ through $P_{i,6}$ model is used again to revise the last six terms or months of the $BF_{i,j}$ model. Approximately five iterations of the $P_{i,1}$ through $P_{i,6}$ Models and the $BF_{i,j}$ Model are necessary to optimize the resulting monthly production rate estimates.

Appendix 3. Multinomial Model

Background

The initial formulation of this methodology was presented in the master's thesis of Crystal Linkletter,[3] whose goal was to prepare timely estimates of natural gas production given the available data structure. The work was conducted under a research fellowship jointly sponsored by the American Statistical Association and the Energy Information Administration. The methodology had been used for product warranty estimation and in AIDS research.[4, 5] The model theory is based upon determining maximum likelihood estimates for the parameters of a multinomial distribution.

Data Preparation

The data are prepared in the same way as is done for the parametric methodology. Data from the Texas Railroad Commission website are monthly updates of aggregate gross natural gas withdrawals for the most recent and all previous months. Data for the most recent month, denoted month t, are first available between 45 and 60 days after the close of the reference month.

These data are extracted and added to the historical data in a spreadsheet. The data are entered into a sheet in columns of monthly vintages. Each month a new column is entered with the first report for month t, denoted $P_{t,1}$, and revised reports for all previous months, denoted $P_{t-k,k+1}$ for k=1, … 96 (or the number of months from the first value included in the spreadsheet.) The data are then arranged into columns, one for each value of k, from k=1, …, 24. These data are the fundamental input for the model.

Basic Model

The theory for the multinomial model is based upon maximum likelihood estimates for certain parameters of a multinomial distribution. Gas that is produced in month t will be included in either $P_{t,1}$ (the first report from the state of Texas), or $P_{t,2}$ (the second report from the state of Texas), or … $P_{t,24}$ (the 24th report from the State of Texas).[6] The partitioning of the gas produced into one of 24 reporting months can be viewed as defining a multinomial distribution with 24 possible report months for each tcf of gas produced. The basic probabilities in the multinomial distribution are the probabilities that a given tcf of gas will be reported in a given month, k.

---

[3] Crystal Linkletter, "Predicting Natural Gas Production in the Presence of reporting Delays", Simon Fraser University, MSc Project, 2002. Abstract available at http://www.stat.sfu.ca/alumni/Theses/Linkletter.abs.shtml

[4] Brookmeyer, R. and Liao, J. (1990). "The Analysis of Delays in Disease Reporting: Methods and results for the Acquired Immunodeficiency Syndrome." *American Journal of Epidemiology*, **132**, 355-365.

[5] Kalbfleisch, J.D., Lawless, J.F. and Robinson, J.A. (1991). "Methods for the Analysis and Prediction of Warranty Claims." *Technometrics*, **33**, 273-285.

[6] The number of months defining the multinomial distribution is a parameter of the model. Currently the value 24 is being used. In earlier years, 12 months might have been sufficient. However, delays in company level reporting to the State of Texas seem to be increasing.

Based on the assumption that the multinomial distribution holds, a likelihood function can be written. At this step, the model is quite general, and the basic probabilities may change over time. However, to make it possible to compute a maximum likelihood estimate, the assumption is made that the probabilities remain constant (stationary) over the recent past (m reporting periods). With this assumption, maximizing the likelihood function with respect to the specific parameters needed to estimate the total production in a month at any point in the reporting process yields the expressions below. In particular, the model estimates $g_{t,k,}$ the conditional probability that gas produced in month t is reported in the kith report from the state of Texas given that it was reported on or before the k^{th} report for k=1, …, 24

The stationarity assumption is that the reporting patterns have remained stable over the most recent m months, where m is a chosen time period (which can be specified parametrically). The model has been run with m=6 and with m=9. Larger values of m are preferred if the stationarity assumption holds because averaging more values results in a smaller variance. Smaller values of m are better if the assumption of stationarity does not hold. For the data currently available, results for m=6 appear to be somewhat better than for m=9 because there are increasing delays in company level reporting to the state.

The stationarity assumption is that $g_{t,k} = g_k$ over the most recently available m time periods. Under this assumption, maximum likelihood estimates for the conditional probabilities, $g_{t,k}$, are given by $g_{t,1} = 1$ and

$$g_{t,k} = \sum_{j=t+1-m-k}^{t+1-k} (P_{j,k} - P_{j,k-1}) \Big/ \sum_{j=t+1-m-k}^{t+1-k} P_{jk} \quad \text{for k>1.}$$

The $g_{t,k}$ are used to provide an estimate of the factor used to "weight up" a current report from the State of Texas, $P_{t,k}$ to prepare an estimate for the final reported production volume in month t.

The weight, which is used to adjust the k^{th} estimate from the State of Texas for production at time t is the product of the conditional probabilities a unit of natural gas **not** being reported by time t+k

$$\hat{W}_{t,k} = \prod_{i=k+1}^{24} \left[1 - g_{t,i}\right]$$

Hence, the estimate for the final value of production for month t based on knowing the k^{th} preliminary value is obtained by dividing $P_{t,k}$ by $\hat{W}_{t,k}$, or

$$\tau_{t,k} = P_{t,k} / \hat{W}_{t,k}$$

For publication in the *Natural Gas Monthly* in its current production cycle, the third estimate for production in month t is used as the basis for estimation. Hence $\tau_{t,3}$ provides the estimate for

publication. As the Natural Gas Monthly moves its production cycle forward $\tau_{t,2}$ or $\tau_{t,1}$ may be used to provide more timely estimates.

Prediction intervals

The variance of $\hat{W}_{t,k}$ is given by

$$\hat{V}(\hat{W}_{t,k}) = \hat{W}_{t,k}^{\,2} \sum_{s=k+1}^{24} \frac{(\hat{g}_{t,s})}{(1-\hat{g}_{t,s}) \sum_{j=t+1-m-k}^{t+1-k} P_{jk}} .$$

The approximate variance for the prediction interval $Y_{t,24} - \tau_{t,k}$ is given by

$$V(Y_{t,24} - \tau_{t,k}) = \frac{P_{t,k}^{\,2}}{\hat{W}_{t,k}^{\,4}} V(\hat{W}_{t,k}) .$$

The reader is referred to the references for more detailed information about the methodology and the derivation of the estimates and variances.

Appendix 4. Comparative Evaluation Results

In August 2003, the multinomial model was compared to the then-current parametric method to assess improvements in accuracy and predictability. Both methods were run for three years of monthly data – calendar years 1999, 2000, and 2001. To make the evaluation valid, only data that were available at the time the estimates would have been produced were used. For example, in predicting the final natural gas production value for January 1999, only data available from the first three reports of January production from the state of Texas and earlier data were used. [7]

The statistics described in Appendix 1 are shown in Figure A4-1 and in Table A4-1 to provide a comparison of the parametric model (Appendix 2) with the proposed multinomial model (Appendix 3). Both models provide production estimates that are usually accurate to within one percent. For the 36 months from 1999 to 2001, the multinomial model with m=6 had 28 estimates with errors less than 1%, 34 with errors less than 1.5% and 36 months with errors less than 2%. The parametric model had 19 estimates with errors less than 1%, 31 with errors less than 1.5%, and 35 with errors less than 2%. The summary results show that the multinomial model has a lower mean squared error, a lower mean absolute deviation, and the magnitude of the largest errors is smaller. However, the multinomial model appears to have a slight negative average error of about -.72%. That is, the multinomial model tends to underestimate final production by a small amount.

The smaller estimates of variation (mean squared error and mean absolute deviation) mean that the multinomial model is more accurate. In addition, the multinomial model provides mathematical theory for the reporting pattern, that allows for the estimation of prediction intervals. The assessment showed that all thirty-six estimates for the multinomial model were within the 90% prediction intervals.

It is suspected that the small bias associated with the multinomial model is due to the assumption that reporting probabilities stay constant over six to nine months. The data clearly show that there are increasing delays in the reporting of production from the State of Texas. It is hoped that a relatively simple enhancement to the multinomial method can be developed to remove the bias from the estimates.

[7] Estimates described here are based on the data available within the current publication schedule, namely using the P_3 values. More timely estimates can an be obtained using only data available at P_1, or P_2 however, estimates will not be as accurate as those described in this report, based upon the P_3 values.

Figure A4-1. Percent Error of Parametric and Multinomial Methods

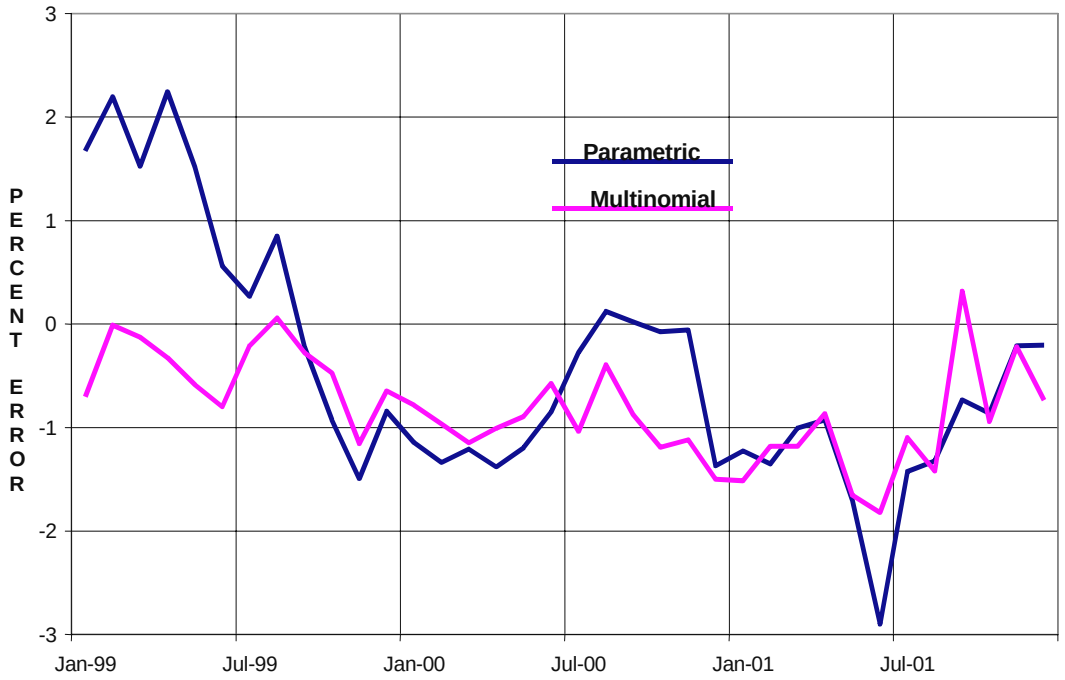


Table A4-1. Monthly Texas Estimation Results for Multinomial and Parametric Models (All volumes measured in billion cubic feet/day)

Production Month	P1	P3	Multinomial Model Estimate, m=9	% Diff from Final Value	Lower PI	Upper PI	Multinomial Model Estimate, m=6	% Diff from Final Value	Lower PI	Upper PI	Parametric Model	% Diff from Final Value	Final Value as of 7/30/03
Jan-99	14.771	15.261	15.408	-0.70%	15.200	15.615	15.359	-1.02%	15.200	15.615	15.760	1.57%	15.517
Feb-99	14.780	15.408	15.556	-0.01%	15.346	15.765	15.507	-0.32%	15.346	15.765	15.812	1.64%	15.557
Mar-99	14.522	15.175	15.310	-0.13%	15.111	15.51	15.261	-0.45%	15.111	15.510	15.482	0.99%	15.329
Apr-99	14.519	15.135	15.255	-0.33%	15.067	15.444	15.227	-0.51%	15.067	15.444	15.563	1.69%	15.305
May-99	14.347	15.060	15.170	-0.58%	14.990	15.35	15.203	-0.37%	14.990	15.350	15.413	1.01%	15.259
Jun-99	14.296	15.002	15.133	-0.80%	14.936	15.33	15.158	-0.63%	14.936	15.330	15.340	0.56%	15.255
Jul-99	14.386	15.103	15.228	-0.21%	15.034	15.422	15.284	0.16%	14.997	15.570	15.299	0.25%	15.260
Aug-99	14.351	14.982	15.145	0.06%	14.924	15.365	15.151	0.11%	14.875	15.428	15.264	0.85%	15.135
Sep-99	14.437	15.106	15.261	-0.28%	15.045	15.478	15.295	-0.05%	15.001	15.590	15.269	-0.23%	15.303
Oct-99	14.191	15.153	15.338	-0.47%	15.100	15.575	15.346	-0.42%	15.048	15.645	15.268	-0.93%	15.411
Nov-99	14.418	15.244	15.384	-1.16%	15.176	15.591	15.384	-1.15%	15.129	15.640	15.332	-1.49%	15.564
Dec-99	14.346	15.173	15.335	-0.65%	15.111	15.558	15.300	-0.87%	15.057	15.543	15.299	-0.88%	15.434
1999 Average				**-0.44%**				**-0.46%**				**0.42%**	
Jan-00	14.380	15.186	15.351	-0.78%	15.125	15.577	15.335	-0.89%	15.072	15.598	15.291	-1.17%	15.472
Feb-00	14.330	15.185	15.348	-0.96%	15.124	15.571	15.325	-1.11%	15.071	15.578	15.291	-1.33%	15.497
Mar-00	14.459	15.361	15.498	-1.15%	15.292	15.705	15.491	-1.20%	15.246	15.736	15.487	-1.22%	15.678
Apr-00	14.161	15.388	15.569	-1.01%	15.332	15.807	15.563	-1.05%	15.278	15.848	15.507	-1.40%	15.728
May-00	14.715	15.428	15.611	-0.89%	15.372	15.85	15.667	-0.54%	15.333	16.001	15.562	-1.20%	15.752
Jun-00	14.735	15.599	15.802	-0.57%	15.550	16.055	15.850	-0.27%	15.506	16.194	15.755	-0.87%	15.893
Jul-00	14.829	15.452	15.651	-1.04%	15.403	15.899	15.695	-0.76%	15.359	16.032	15.771	-0.28%	15.815
Aug-00	14.828	15.464	15.727	-0.39%	15.441	16.013	15.752	-0.23%	15.385	16.119	15.805	0.10%	15.789
Sep-00	14.633	15.386	15.636	-0.88%	15.359	15.914	15.683	-0.58%	15.313	16.054	15.774	0.00%	15.775
Oct-00	14.380	15.327	15.585	-1.19%	15.303	15.866	15.594	-1.13%	15.243	15.945	15.749	-0.15%	15.772
Nov-00	14.596	15.305	15.545	-1.12%	15.274	15.817	15.534	-1.19%	15.202	15.857	15.703	-0.12%	15.721
Dec-00	14.151	15.366	15.619	-1.50%	15.341	15.898	15.577	-1.77%	15.267	15.887	15.628	-1.45%	15.857
2000 Average				**-0.96%**				**-0.89%**				**-0.76%**	
Jan-01	14.542	15.393	15.632	-1.52%	15.361	15.902	15.629	-1.53%	15.361	15.902	15.671	-1.27%	15.872
Feb-01	14.656	15.459	15.726	-1.18%	15.439	16.014	15.706	-1.31%	15.439	16.014	15.691	-1.40%	15.914
Mar-01	14.523	15.385	15.729	-1.18%	15.403	16.055	15.773	-0.91%	15.403	16.055	15.754	-1.02%	15.917
Apr-01	14.677	15.433	15.801	-0.86%	15.462	16.140	15.838	-0.63%	15.462	16.140	15.790	-0.93%	15.939
May-01	14.541	15.358	15.735	-1.65%	15.393	16.078	15.848	-0.95%	15.393	16.078	15.728	-1.70%	16.000
Jun-01	14.075	15.339	15.781	-1.82%	15.410	16.152	15.920	-0.96%	15.410	16.152	15.608	-2.90%	16.074
Jul-01	13.949	15.350	15.795	-1.10%	15.423	16.167	15.929	-0.26%	15.423	16.167	15.743	-1.42%	15.970
Aug-01	14.595	15.254	15.739	-1.42%	15.351	16.127	15.814	-0.95%	15.351	16.127	15.755	-1.32%	15.966
Sep-01	14.323	15.549	16.092	0.32%	15.676	16.507	16.056	0.09%	15.676	16.507	15.923	-0.73%	16.041
Oct-01	14.328	15.345	15.916	-0.94%	15.493	16.340	15.887	-1.13%	15.493	16.340	15.929	-0.86%	16.068
Nov-01	14.528	15.243	15.831	-0.22%	15.402	16.260	15.807	-0.37%	15.402	16.260	15.833	-0.21%	15.866
Dec-01	14.331	15.234	15.746	-0.74%	15.347	16.144	15.727	-0.86%	15.347	16.144	15.830	-0.20%	15.862
2001 Average				**-1.03%**				**-0.81%**				**-1.17%**	

Table A4-2. Summary Comparison of Model Results

Time Period	Statistical Measures of Error	Multinomial Model m=9	Multinomial Model m=6	Parametric Model
1999	Average Error (%)	-0.44%	-0.46%	0.42%
	Mean Absolute Deviation (%)	0.45%	0.50%	1.01%
	Root Mean Squared Error (%)	0.56%	0.61%	1.12%
	Max Error	-0.01%	0.16%	1.64%
	Min Error	-1.16%	-1.15%	-1.49%
2000	Average Error (%)	-0.96%	-0.89%	-0.76%
	Mean Absolute Deviation (%)	0.96%	0.89%	0.77%
	Root Mean Squared Error (%)	1%	0.99%	0.96%
	Max Error	-0.39%	-0.23%	0.10%
	Min Error	-1.50%	-1.77%	-1.45%
2001	Average Error (%)	-1.03%	-0.81%	-1.17%
	Mean Absolute Deviation (%)	1.08%	0.83%	1.17%
	Root Mean Squared Error (%)	1.18%	0.92%	1.35%
	Max Error	0.32%	0.09%	-0.20%
	Min Error	-1.82%	-1.53	-2.90%
1999-2001	Average Error (%)	-0.81%	-0.72%	-0.50%
	Mean Absolute Deviation (%)	0.83%	0.74%	0.98%
	Root Mean Squared Error (%)	0.95%	0.86%	1.15%
	Max Error	0.32%	0.16%	1.64%
	Min Error	-1.82%	-1.77	-2.90%

Tables 1 and 2 contain the results from weak efficiency tests of the "forecast" estimates using the two models. The data value we are considering almost final, P_{24}, is regressed on a constant and the "Forecast" of the Final Estimate using the two methodologies. For both the parametric and multinomial model one cannot reject the null hypothesis that the constant is zero (indicating that the bias is not statistically significant). For both the parametric and multinomial model one cannot reject the null hypothesis that the coefficient is one. Hence both models perform reasonably well. However, there is significant autocorrelation at lag one in both regressions, suggesting that it should be possible to improve both methodologies.

Table A4-3. Weak Efficiency Tests for Parametric Model

Dependent Variable: FINAL				
Method: Least Squares				
Date: 09/14/03 Time: 12:04				
Sample: 1999:01 2001:12				
Included observations: 36				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.230745	2.069189	-0.111515	0.9119
MOD1	1.019931	0.132610	7.691234	0.0000
R-squared	0.635017	Mean dependent var		15.68242
Adjusted R-squared	0.624282	S.D. dependent var		0.273719
S.E. of regression	0.167778	Akaike info criterion		-0.678395
Sum squared resid	0.957084	Schwarz criterion		-0.590421
Log likelihood	14.21110	F-statistic		59.15509
Durbin-Watson stat	0.324434	Prob(F-statistic)		0.000000
Wald Test:				
Equation: WE_MOD1				
Test Statistic	Value	Df		Probability
F-statistic	4.126491	(2, 34)		0.0249
Chi-square	8.252983	2		0.0161
Null Hypothesis Summary:				
Normalized Restriction (= 0)		Value		Std. Err.
C(1)		-0.230745		2.069189
-1 + C(2)		0.019931		0.132610
Restrictions are linear in coefficients.				

Table A4-4. Weak Efficiency Tests for Multinomial Model, m=6

Dependent Variable: FINAL				
Method: Least Squares				
Date: 09/14/03 Time: 12:04				
Sample: 1999:01 2001:12				
Included observations: 36				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.443877	0.765621	-0.579760	0.5659
M6	1.035812	0.049170	21.06577	0.0000
R-squared	0.928836	Mean dependent var		15.68242
Adjusted R-squared	0.926742	S.D. dependent var		0.273719
S.E. of regression	0.074085	Akaike info criterion		-2.313251
Sum squared resid	0.186613	Schwarz criterion		-2.225278
Log likelihood	43.63852	F-statistic		443.7667
Durbin-Watson stat	1.207189	Prob(F-statistic)		0.000000
Wald Test:				
Equation: WE_M6				
Test Statistic	Value	Df		Probability
F-statistic	42.63701	(2, 34)		0.0000
Chi-square	85.27402	2		0.0000
Null Hypothesis Summary:				
Normalized Restriction (= 0)		Value		Std. Err.
C(1)		-0.443877		0.765621
-1 + C(2)		0.035812		0.049170
Restrictions are linear in coefficients.				

Appendix 5. Recommendations of the American Statistical Association Advisory Committee on Energy Statistics

During the month of October 2003, the *Comparative Evaluation of Two Methods to Estimate Natural Gas Production in Texas* was presented to a panel of experts at a meeting of the American Statistical Association Advisory Committee on Energy Statistics. Panel members were briefed on the evaluation and asked to offer their comments. Below are their suggestions.

- The panel suggested running the model with different values for the assumed period of stationarity (m), as well as running preliminary statistics to compare the 2002 data with the most recent information.
- The panel would like to see experiments with estimates predicted when only P_1 or P_2 are available. This could improve timeliness, but may detract from the overall accuracy; documenting those comparisons could be beneficial.
- The panel suggested that Crystal Linkletter publish her master's thesis on which the Multinomial Model is based. The panel members felt that this will provide EIA with additional independent expert reviews, thus making this method even more credible.
- Finally, the panel recommended that in future comparisons of estimation methods, EIA should compare and contrast those models over a longer period of time, utilizing as much data as possible.