

## EIA-914 Gas Production Survey And Gas Production Estimates

The 914 survey will collect monthly gas production by State from the 250 to 350 top producing companies. Monthly volumes for the Federal Gulf of Mexico, Louisiana, New Mexico, Oklahoma, Texas, Wyoming, and other States excluding Alaska will be collected on this survey. The goal is to be able to estimate the Total US monthly gas production within 1 percent and within 1 to 5 percent (preferably 1 percent) for the six areas listed above (excluding other States). The following table shows how many operators have to be sampled to have 90 percent of the production in each area included.

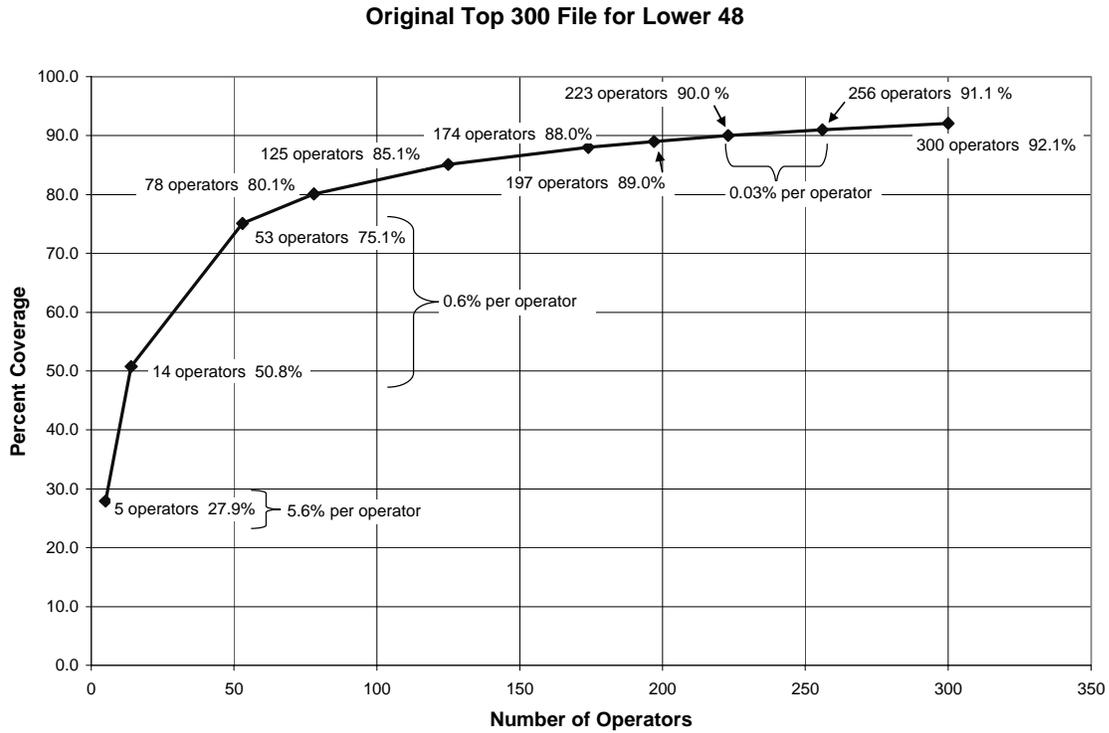
# of Operators	L-48	GOM	LA	NM	OK	TX	WY	OT
222	90.04	98.26	87.93	92.65	83.17	85.97	96.69	83.50
254	91.21	98.26	88.00	92.70	83.25	90.07	96.88	83.51
280	91.87	98.26	88.03	92.71	90.10	90.24	96.94	83.69
284	92.03	98.26	90.33	92.71	90.10	90.26	96.94	83.69
337	93.43	98.27	90.54	92.77	90.42	90.36	97.46	90.04

### Questions for the Committee

1. We would like your comments on our interpretation and analysis of the data and our approach to modeling or estimating the non-sampled portion.
2. Do you have any recommendations on handling non-response or apparent errors in reporting for the large operators? (Top 5 operators in Texas have roughly 25 percent of production.)
3. Any suggestions on detecting or handling outliers and overly influential operators and how we should deal with these when making estimates?

## Top 300 Operators ~ Lower-48 only

Using the original top 300 operator file, the following graph has been created to show how many operators are needed to obtain a certain percentage production for the Lower-48. Note that 14 operators achieve 50.8% coverage while 223 achieve 90% coverage. As the number of operators increases, production added per operator decreases.

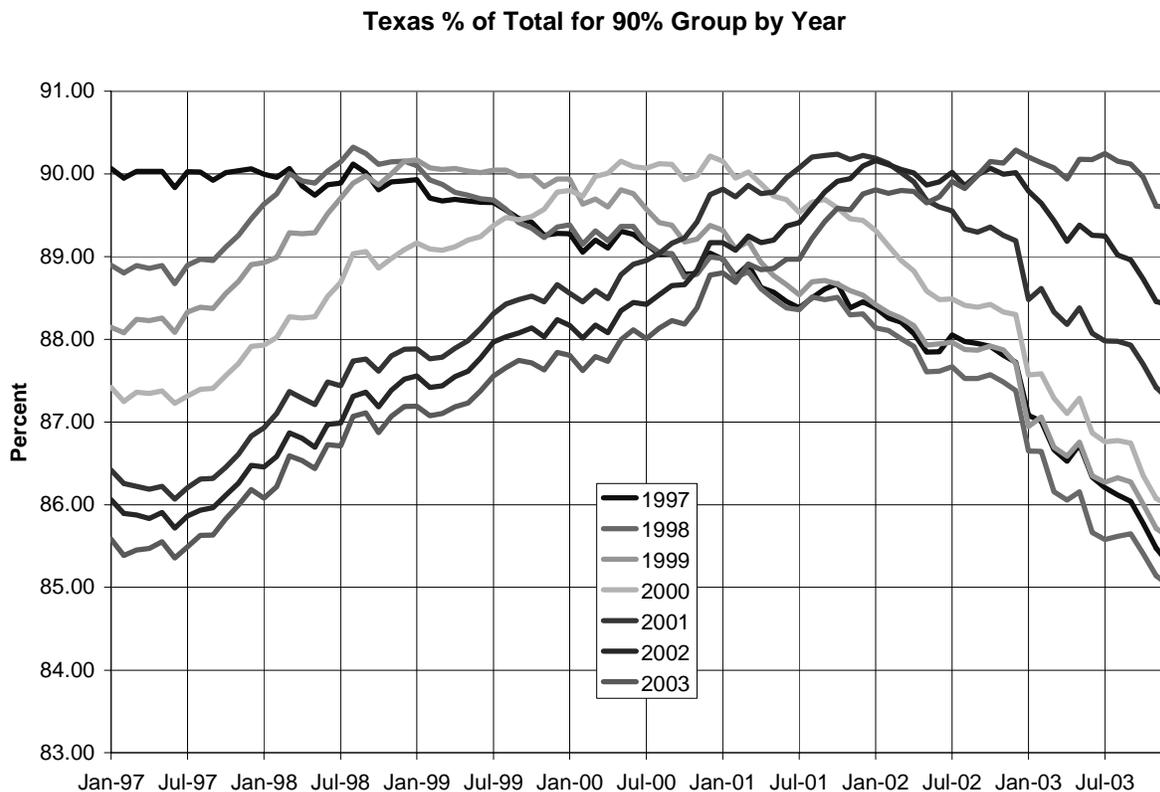


## Estimating Gas Production from a Cut Off Sample Survey

### Test Files

Test files have been generated for several areas for the initial model development and testing. Test files for Texas, Oklahoma, GOM, New Mexico, Louisiana, and Wyoming were generated in order to test several models. The test files were created by merging IHS Energy operator codes with EIA operator codes to yield a file of monthly total IHS gas production by EIA operator for 1997 – 2003. (IHS Energy is a company that collects well level data from State agencies and sells this data to the oil and gas industry and others.)

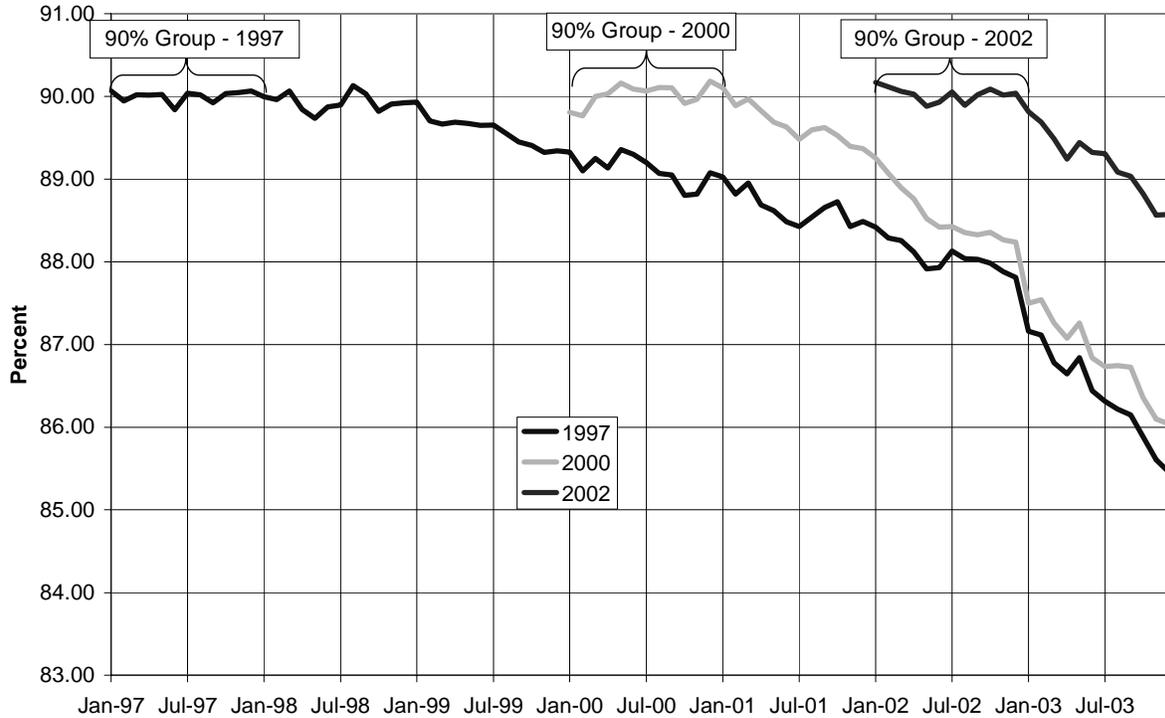
The annual daily average production for each operator in each calendar year (1997 – 2003) was calculated and used to select the cut off sample of operators for each year. Cut off samples were determined for 85 percent and 90 percent or 90 percent and 95 percent of the total production depending on the area and the number of operators. The monthly production data by operator was sorted by the annual daily average production 7 times, once for each year. Monthly production for each cut off sample (7, one for each calendar year) was plotted. The plot below shows the production for each 90 percent cut off sample for Texas.



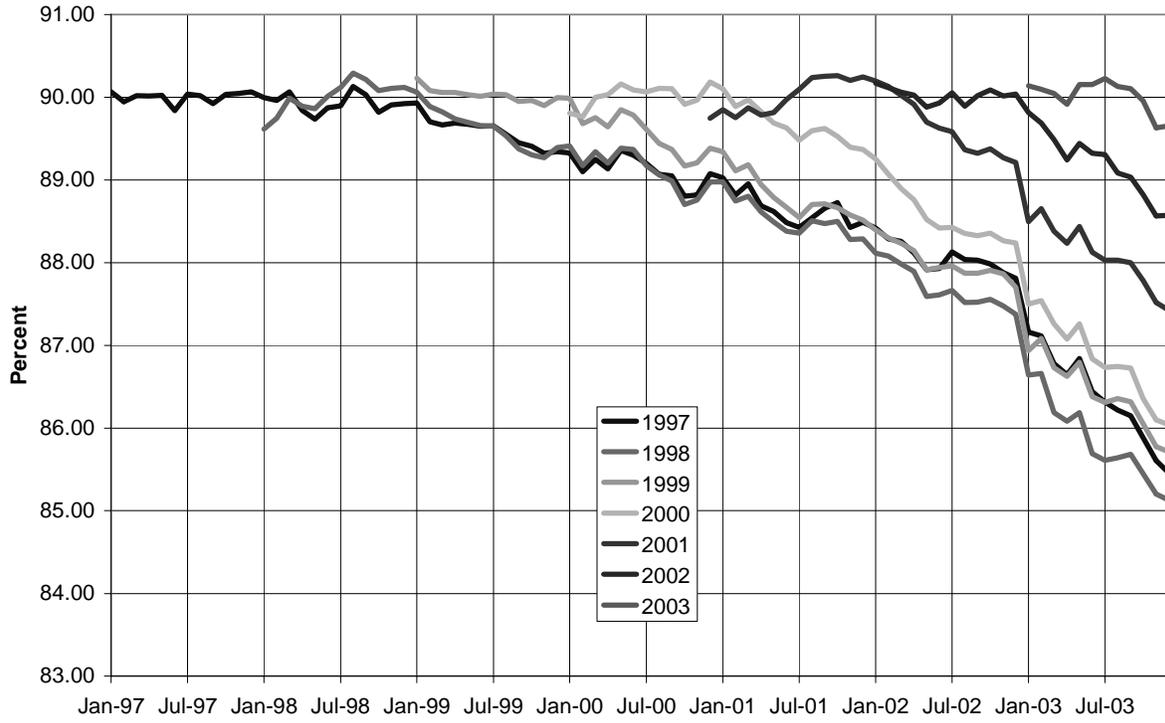
Notice that before and after the year of the cut off sample the percentage of total production for that cut off sample is lower. There are few instances in other areas where this is not true. Production for the cut off sample of operators can be higher in other years, but the percentage of

the total for the group is usually lower. The cut off sample group of operators in a given year is not the same as the prior or following years. The top companies appear to be continuously changing. The following plots are the same as the one above except that the monthly production prior to the cut off sample year is not shown. The first graph shows only 3 years 1997, 2000, and 2002 for the 90 percent group. The next 2 graphs show all 7 years for the 90 and 85 percent cut off operators.

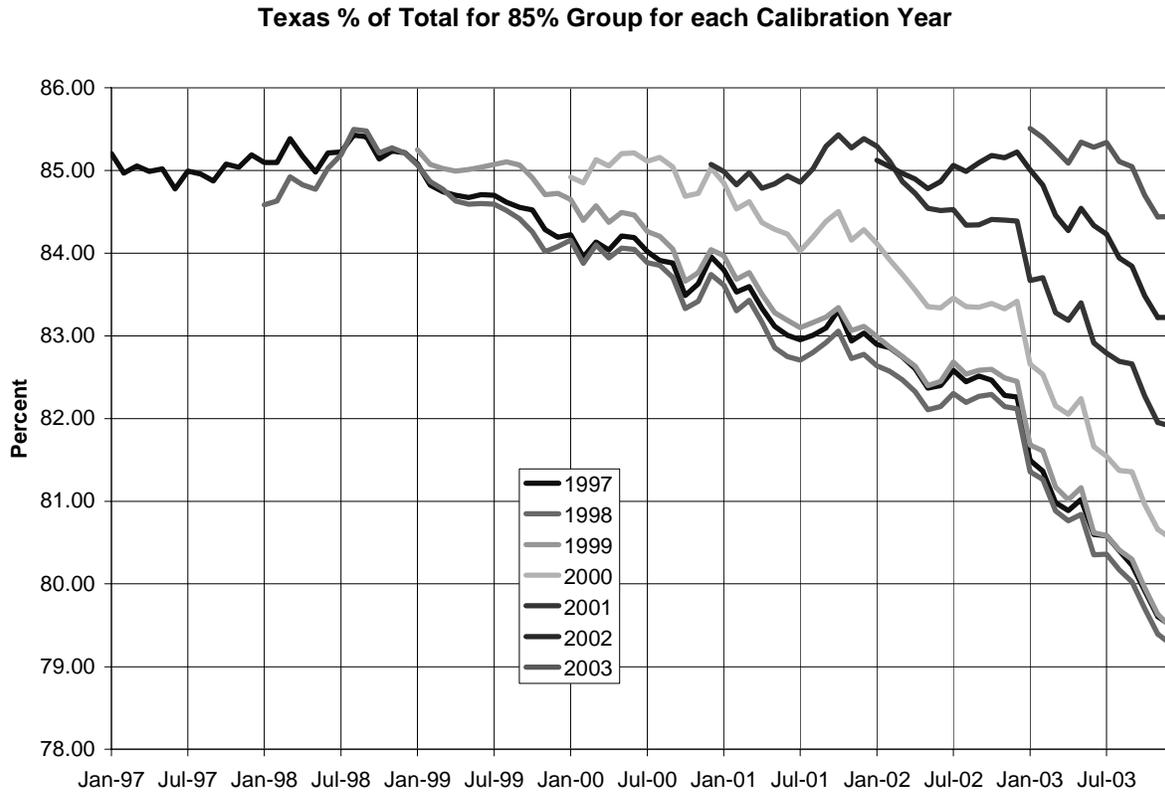
**Texas Monthly % of Total for 90% Group for Selected Calibration Years**



Texas % of Total for 90% Group for each Calibration Year

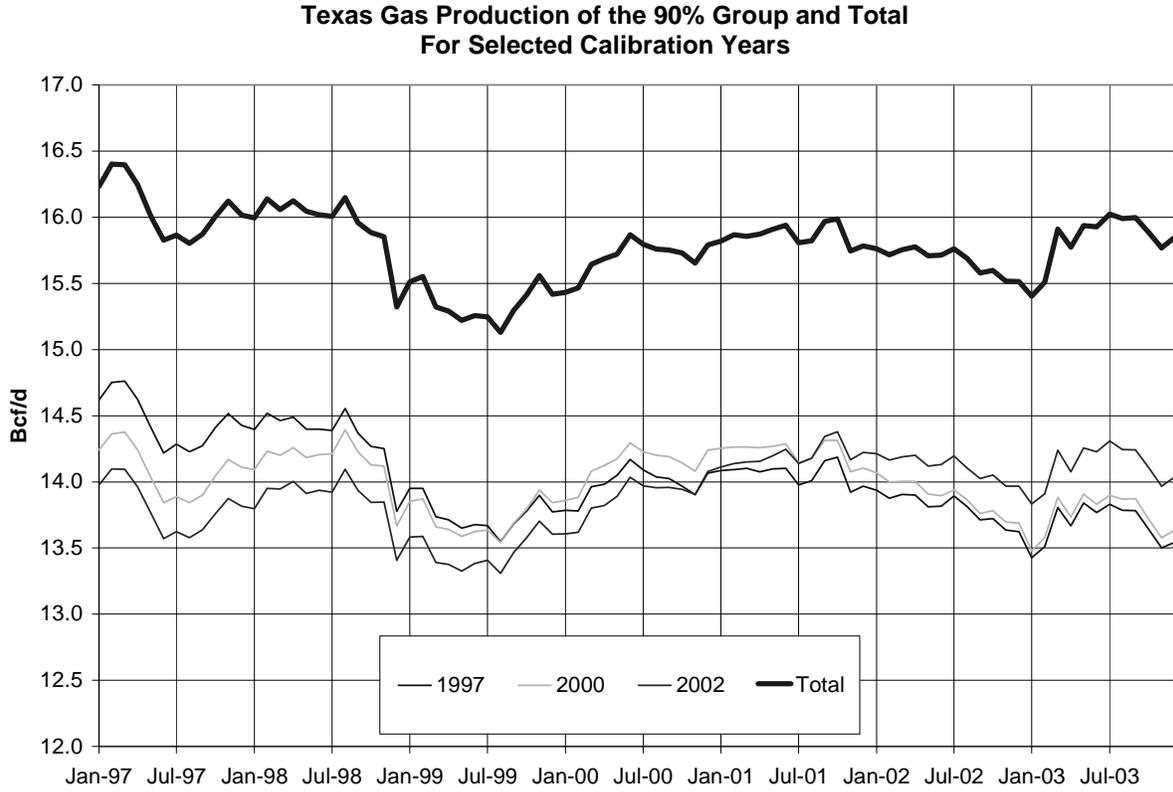


A similar plot for the 85 percent cut off sample of operators is shown below.

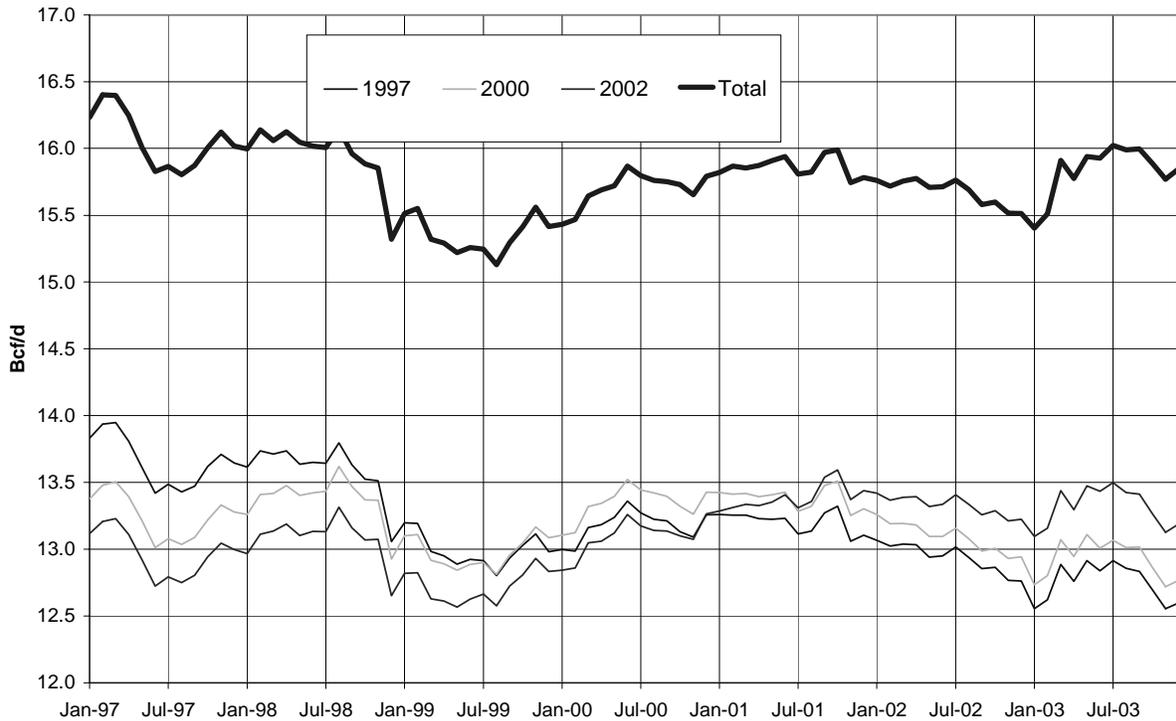


Notice that the 90 percent cut off sample operators are generally smoother than the 85 percent group in Texas. In Texas, the number of operators in the 85 percent cut off sample varies from 120 to 128 with a cut off rate that varies from 12.116 to 14.497 MMcf/d. For the 90 percent cut off sample, the number of operators varies from 195 to 223 with a cut off rate that varies from 5.840 to 7.391 MMcf/d.

The following graphs show monthly production for the 90 and 85 percent cut off operators and the total production. Notice that the character of the cut off sample companies is similar to the total.



**Texas Gas Production of the 85% Group and Total  
For Selected Calibration Years**

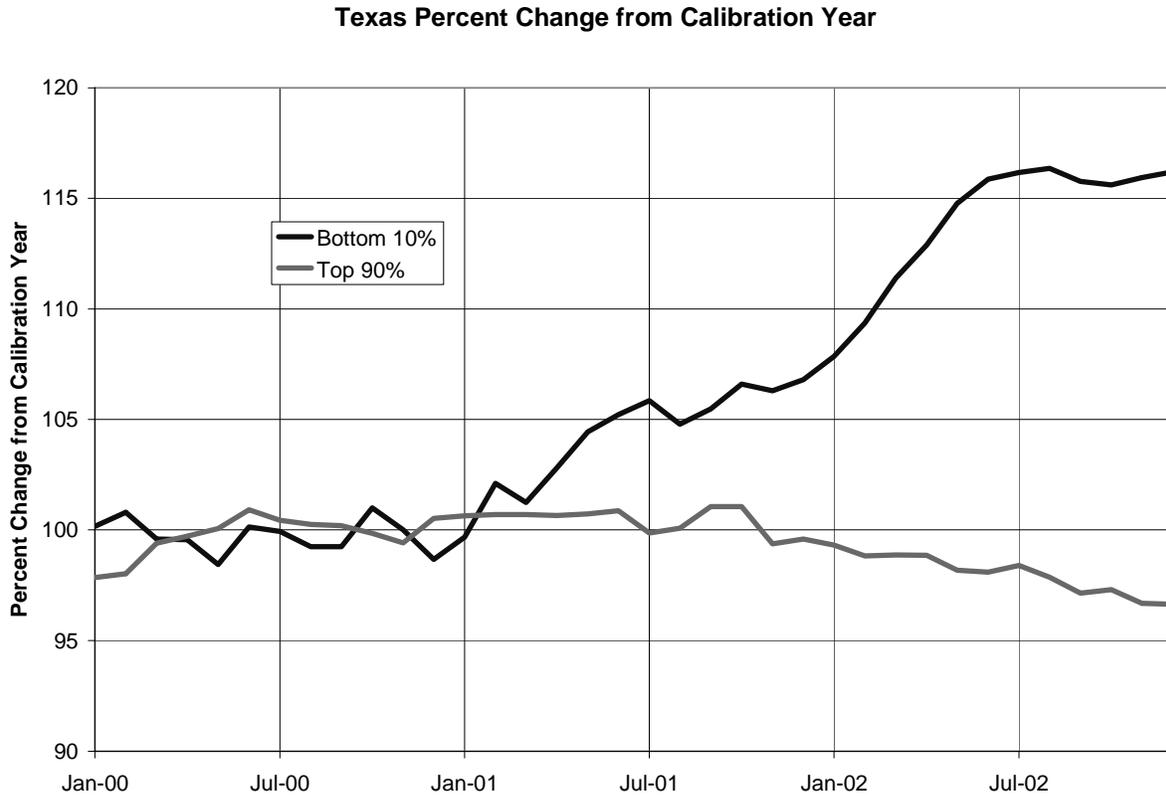


Some preliminary findings for Oklahoma and the GOM are listed here for comparison. In Oklahoma, the number of operators in the 85 percent cut off sample varies from 101 to 115 with a cut off rate that varies from 3.885 to 4.806 MMcf/d. For the 90 percent cut off sample, the number of operators varies from 171 to 199 with a cut off rate that varies from 2.001 to 2.316 MMcf/d.

For the GOM, the number of operators in the 90 percent cut off sample varies from 29 to 33 with a cut off rate that varies from 72.789 to 97.483 MMcf/d. For the 95 percent cut off sample, the number of operators varies from 41 to 44 with a cut off rate that varies from 38.460 to 45.134 MMcf/d.

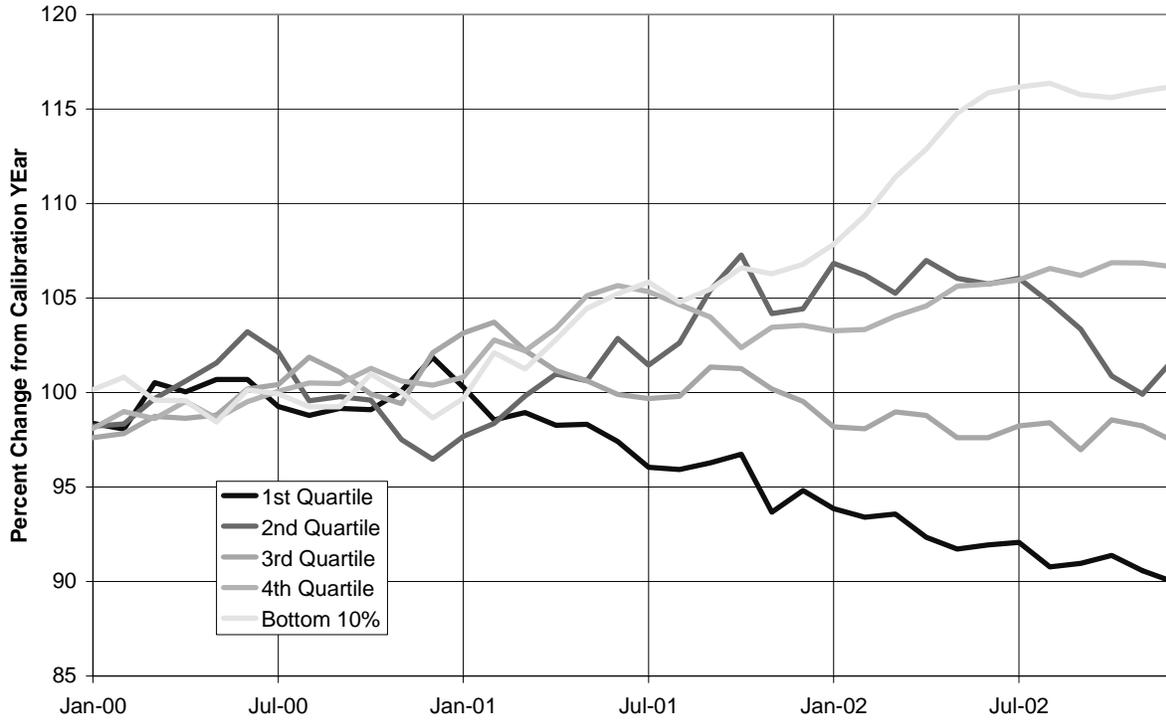
## Model Development

A simple model would use the sampled data to estimate the non-sampled data. We looked at the Texas data for 2000 through 2002 selecting a 90 percent sample in 2000 and then following these operators through the next 2 years. The following graph highlights the different behavior of the bottom 10 percent non-sampled group compared to the 90 percent sampled group. The two curves are normalized to the calibration year (2000).

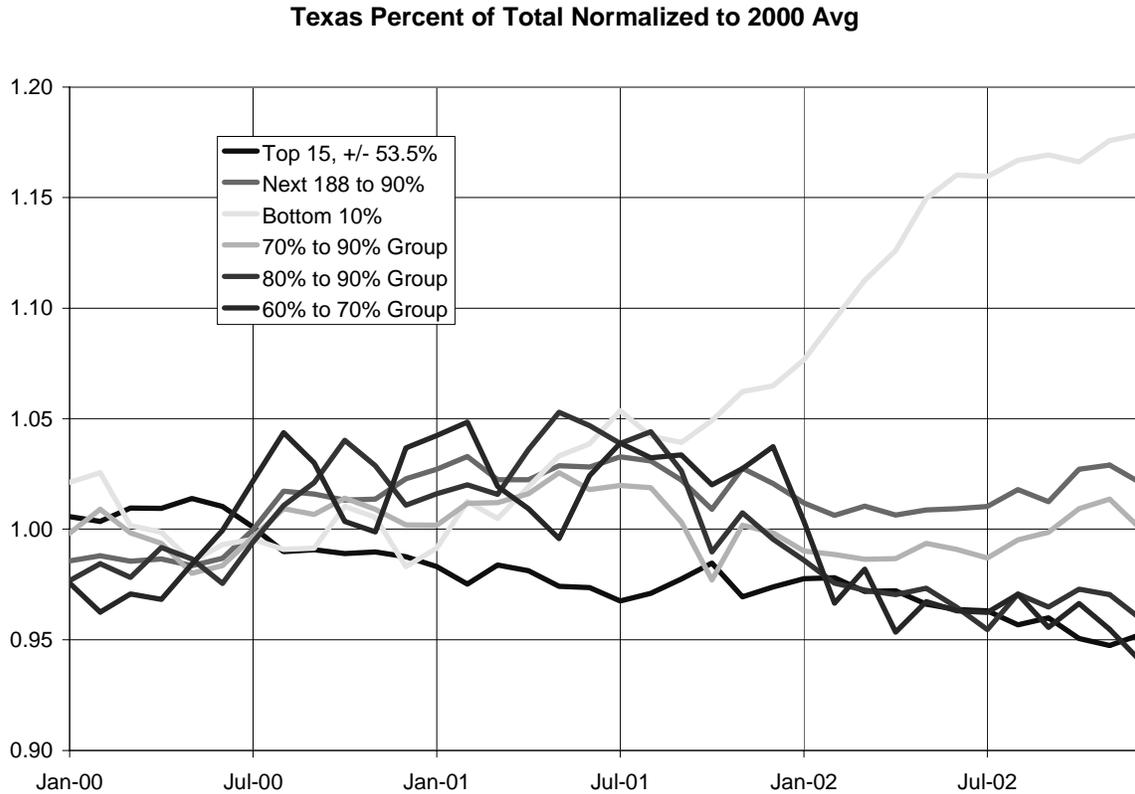


Thinking that there might be a sub-group within the sample group that will have a similar behavior to the non-sampled bottom 10 percent we constructed the following graph of quartiles. The fourth quartile has the most similarity with the bottom 10 percent, but it includes the bottom 10 percent.

Texas Percent Change from Calibration Year



Assuming that there might be some part of the sampled 90 percent group that behaves similarly to the bottom 10 percent (perhaps somewhere close to the bottom of the 90 percent group) we constructed the following graph with several sub groups. The bottom 10 percent non-sampled group seems to be unique.



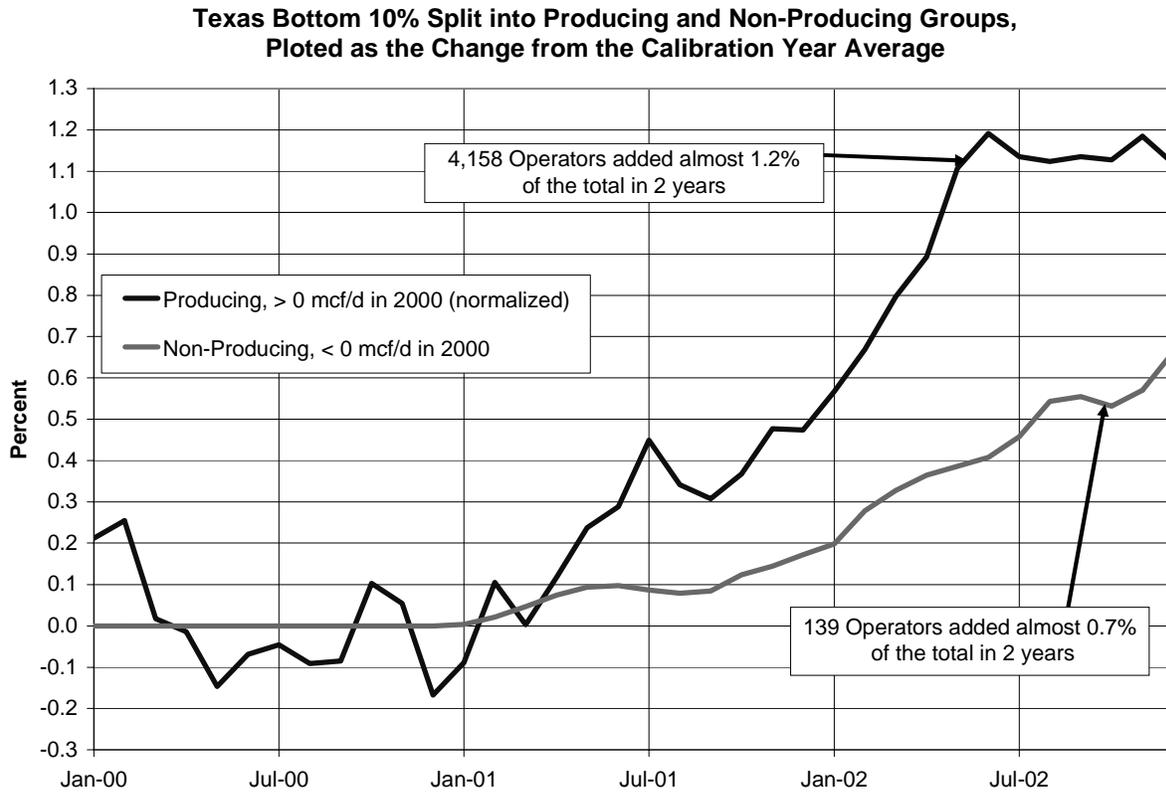
The operator data set consisted of 5,676 operators that produced natural gas in one or more months from January 1997 through March 2004. Of these, in 2000, the top 203 operators accounted for 90 percent of Texas production. The next 4,158 operators produced 10 percent and the remaining 1,315 operators had no production in 2000.

Of the 1,315 operators that had no production in 2000, 139 started producing and added nearly 0.7 percent of the total by the end of 2002. The 4,158 producing operators in 2000 added almost 1.2 percent of the total by 2002. These two groups combined added nearly 2 percent which is about what the top 90 percent declined.

The bottom 10 percent group in 2000 had 4,158 producing operators that collectively produced 1,571 mmcf/d in 2000. By the end of 2002, 488 ceased production. However in 2002, the remaining 3,670 operators collectively produced 1,722 mmcf/d.

Of the 1,315 operators producing zero in 2000, 139 operators collectively produced 69.1 mmcf/d in 2002. Taken with the 488 operators lost from the producing group, the bottom 10 percent group had a net loss of 349 operators from 2000 through 2002.

The graph below shows how these two groups of the bottom 10 percent change over the next two years. Each group was normalized by subtracting the percentage of the same group in 2000 each month.



### Outliers and Overly Influential Operators

A statistical determination of operators that are outliers or overly influential indicated that many of the top operators in Texas may not be used to estimate the non-sampled operator's production in a given month. The top 5 operators account for roughly 25 percent of production and were excluded in a test described later to see what effect their exclusion may have on production estimates.

## New and Improved Model

As a result of the efforts described above a model is proposed. This model uses the sample, 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: 1) the first summation is the group of operators in the sample used to estimate the non-sampled operators ( $j=1$  to  $m$ ) in month  $i$ ; 2) 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  $i$ ; 3) 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. (The second summation in the equation above is zero.)

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

$R_{i,j}$  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 a total if all sampled respondents are used in the first summation (and the second summation is zero). However, two years later the production percentage of the sample group has declined while the non-sampled group has increased indicating the need for more than a simple ratio.

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 0.11 and 0.01 respectively assuming a 90 percent sample is used to estimate the non-sampled portion. The  $R_m$  parameter is calculated by the following equation.

$$R_m = \frac{\sum_{i=1}^{12} \sum_{j=1}^{N_i} y_{i,j}}{\sum_{i=1}^{12} \sum_{j=1}^{m_i} y_{i,j}} - 1$$

Since the  $R_m$  parameter is calculated, there is only one fit parameter, the  $A_m$ . It is determined using a least squares procedure comparing the estimated total monthly production to the actual monthly production. This fit is done for the calendar year (12 monthly values) two years after the calibration year (2 year lag, i.e.  $t = 13$  through 24).

The model can be written as follows.

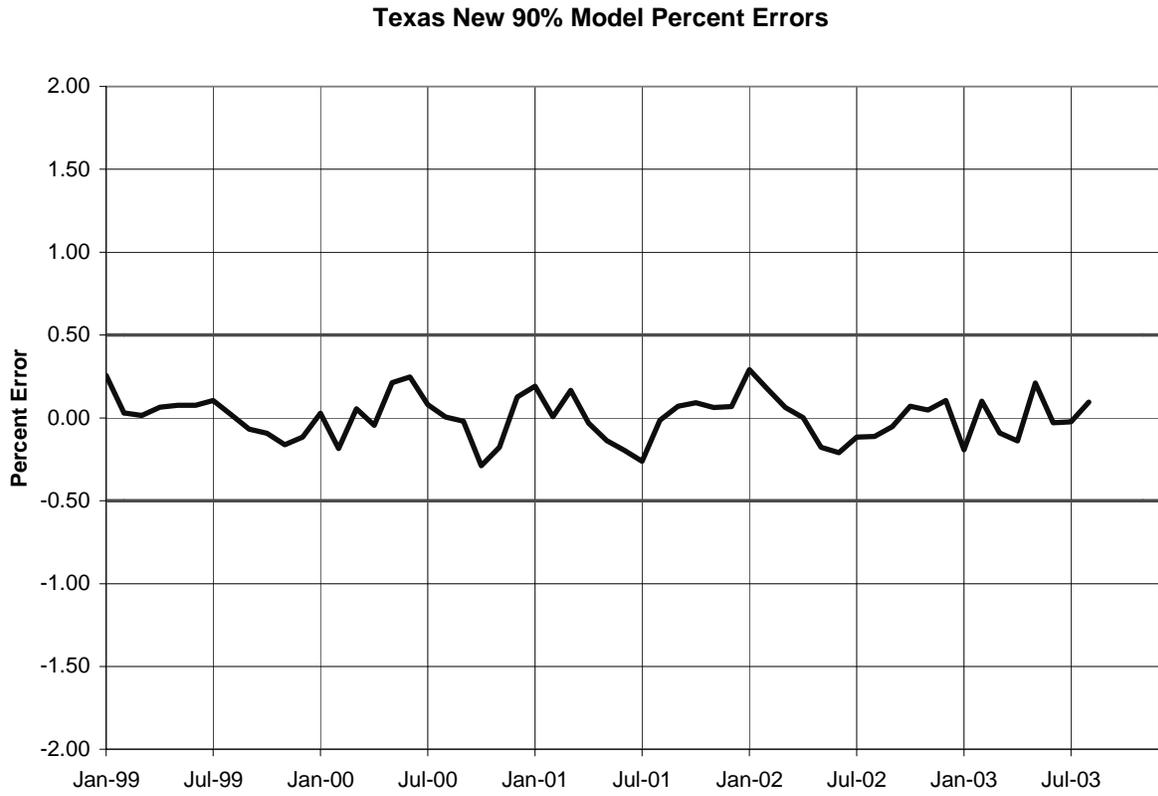
$$\hat{T}_i = \left( \sum_{j=1}^{m_i} y_{i,j} \right) (1 + R_{i,j}) + \sum_{j=m+1}^{n_i} y_{i,j}$$

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. A value for  $R_m$  was calculated for each year and a value for  $A_m$  was determined by least squares fit for the survey years as follows.

#### Calibration

Year	$R_m$	$A_m$
1997	0.1111	0.0026
1998	0.1110	0.0053
1999	0.1107	0.0074
2000	0.1109	0.0089
2001	0.1107	0.0119
2002	0.1108	
2003	0.1109 (partial year)	

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



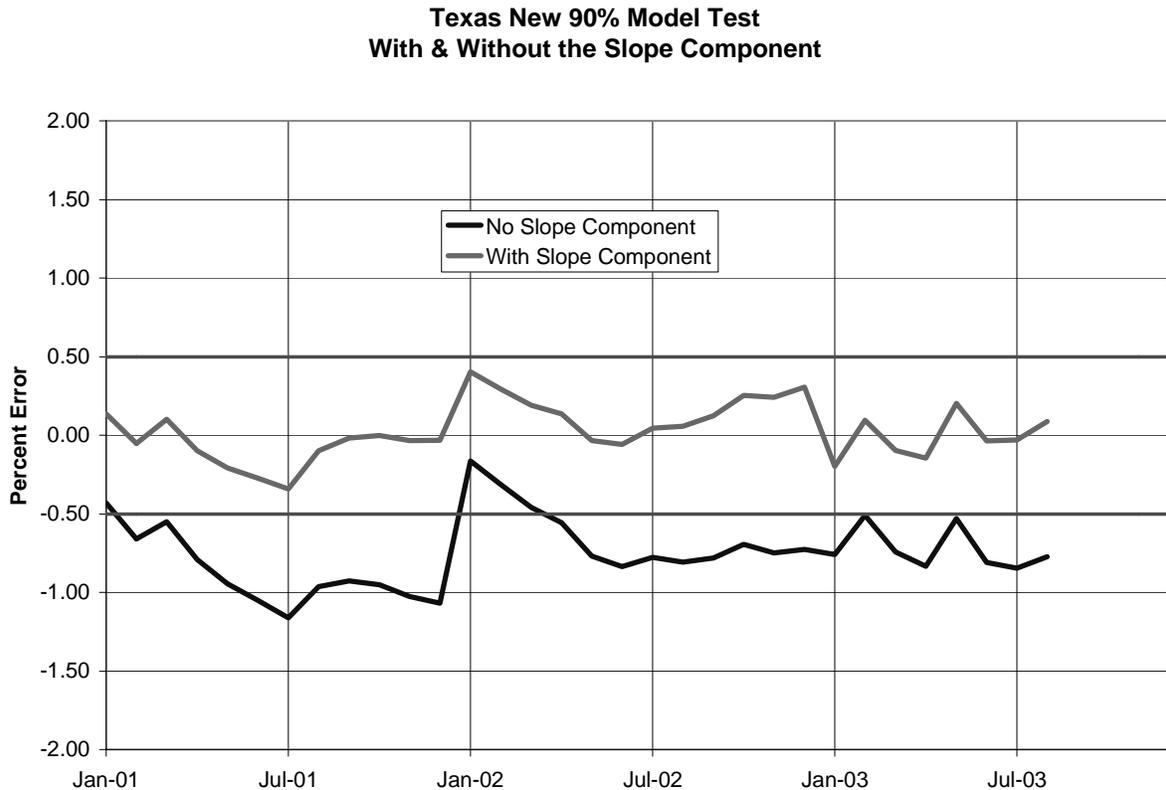
## Model Testing

In order to test the model we used the  $A_m$  parameter determined two years prior to the estimated (forecast) current production. In Texas we may only need a one year lag but a two year lag is a more stringent test. We can change the apparent systematic bias due to the systematically changing  $A_m$ 's over time by shortening the lag or by adding a slope component to the  $A_m$ 's to adjust the forecast  $A_m$  rather than carry forward an  $A_m$  from two years prior. The graph below shows both cases; i.e, a case carrying forward for two years an  $A_m$  without a slope adjusting component and a case including a slope adjusting component.

Adding the slope component changes the  $R_{i,j}$  equation to the following.

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

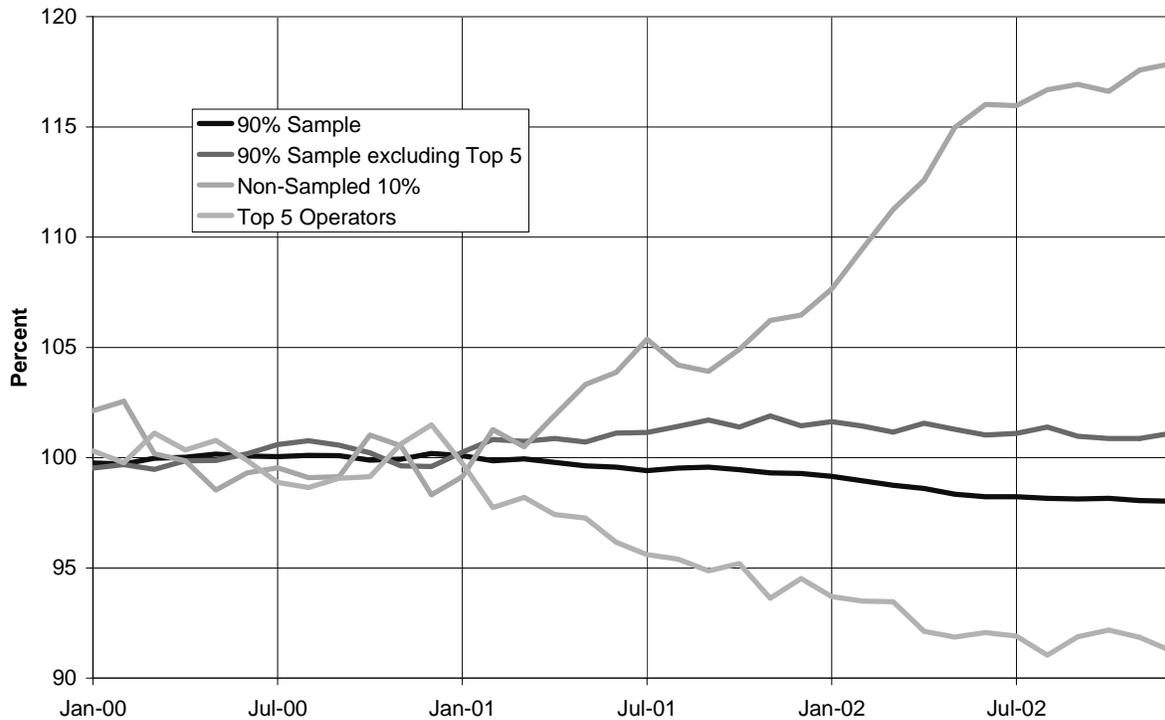
For the case with no slope component, the resulting production estimates are slightly low but mostly within 1 percent. The average error is -0.75 percent and the average absolute error is 0.75 percent. The addition of the slope component (0.0022 per year) eliminates the low bias, lowers the average error to 0.03 percent, and lowers the average absolute error to 0.14 percent. Percent errors are well within plus or minus 0.5 percent.



In the event that outliers and overly influential operators are identified and subsequently not used to estimate the non-sampled group, a test scenario was run to determine the effect of excluding these operators on the model results. The top five operators, which account for roughly 25 percent of the total production, were excluded from the 90 percent sample used in the model test above. For the purpose of this test the top five operators are assumed to be outliers and/or overly influential. The following test then uses roughly a 65 percent sample to estimate the non-sampled group.

The graph below shows how several groups of operators change over time as a percentage of total production. Three years are shown, 2000 through 2002, with 2000 being the calibration or sample year. The 90 percent sample in 2000 loses about 2 percent by the end of 2002 while the 10 percent non-sample group gains about 2 percent. The sample used for this outlier and influence test (90 percent minus the top five companies) is about 62 percent of the total production in 2000. The excluded top five companies are roughly 28 percent of the total in 2000. The top five companies drop roughly 3 percent while the group excluding the top five companies (outlier and influence sample group) gain about 1 percent by the end of 2002.

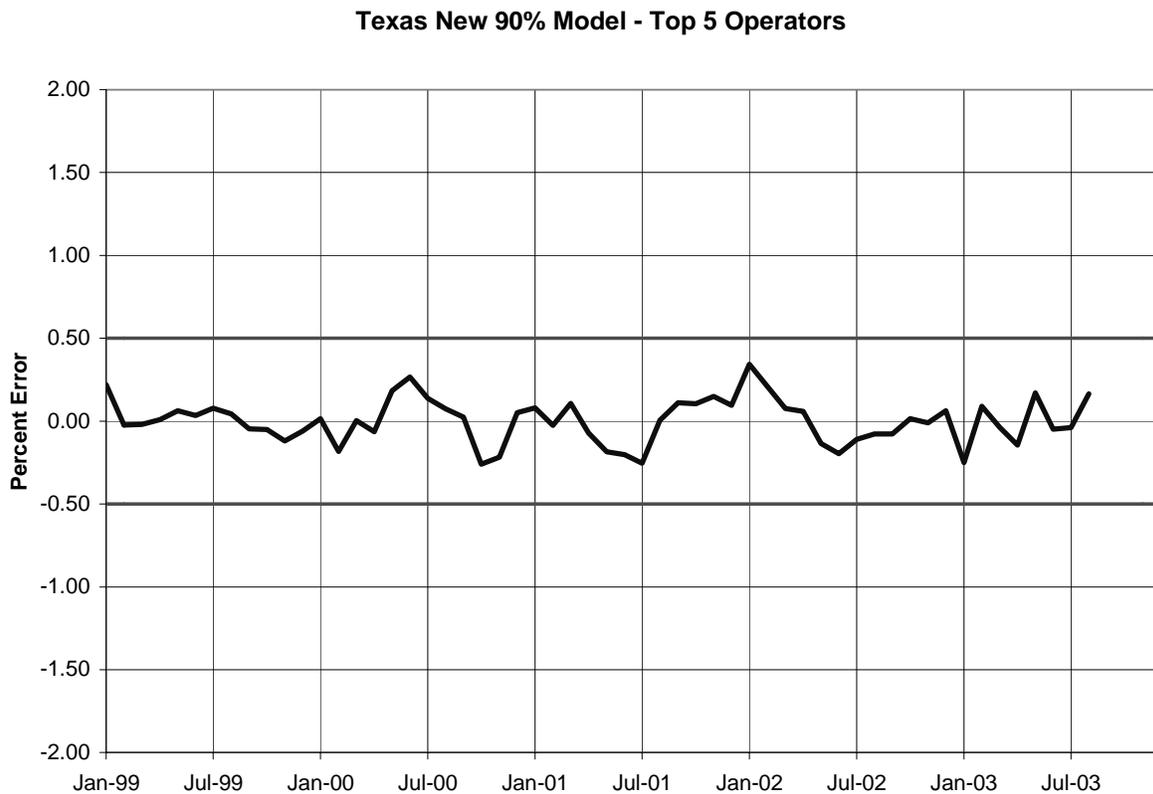
**Percent of Total Production Normalized to the Calibration Year**



$R_m$  is now calculated by the following equation which excludes the top 5 companies.

$$R_m = \frac{\sum_{i=1}^{12} \sum_{j=1}^{N_i} y_{i,j} - \sum_{i=1}^{12} \sum_{j=1}^5 y_{i,j}}{\sum_{i=1}^{12} \sum_{j=5+1}^{m_i} y_{i,j}} - 1$$

The outlier and influence test follows the same procedure as in the test above, i.e., we ran two cases; one without a slope component and one case with a slope component on the  $A_m$  parameter. As before, we first performed the calibration run with all data known to determine the annual  $A_m$  parameters and their slope. In this case the  $R_m$  parameter is in the range of 0.16 while the  $A_m$  parameter is in roughly the same range mentioned above, 0.01. The graph below shows this result of the calibration run which assumes all data are known.



The next graph shows both test cases; i.e, a case carrying forward for two years an  $A_m$  without a slope adjusting component and a case including a slope adjusting component.

The resulting production estimates are slightly low but mostly within 1 percent for the case with no slope component. The average error is -0.47 percent and the average absolute error is 0.48 percent. The addition of the slope component (0.0015 per year) eliminates the low bias, lowers the average error to 0.07 percent, and lowers the average absolute error to 0.18 percent. Percent errors are within plus or minus 0.5 percent.

**Texas New 90% Model - Top 5 Operators Test  
With & Without the Slope Component**

