Journal of Official Statistics, Vol. 23, No. 2, 2007, pp. 239–251

Predicting Natural Gas Production in Texas: An Application of Nonparametric Reporting Lag Distribution Estimation

Crystal D. Linkletter[1] *and Randy R. Sitter*[2]

There exists a delay between the production of natural gas by states in the U.S. and the reporting of that production. As a result, the true volume of natural gas produced in a given month is not known until a year or more after production. It is important to have earlier and more accurate production estimates. In this article we apply nonparametric reporting-lag-distribution estimation originally developed to analyze epidemiological, insurance, and product warranty data, and demonstrate its simplicity and effectiveness in producing better early estimates of natural gas production in Texas.

Key words: AIDS; prediction intervals; right-truncated data; survey sampling.

1. Introduction

Maintaining current records of energy resources is vital for balancing supply, demand, and cost. The United States has a particular interest in tracking natural gas production as the U.S. and worldwide consumption of natural gas is expected to grow more rapidly than that of any other fuel source over the next 20 years (Report #:DOE/EIA-0484, 2005). With this growth in demand will come added pressure on natural gas supplies, increasing the importance of maintaining accurate estimates of production.

The Energy Information Administration (EIA) is the agency of the U.S. Department of Energy responsible for the collection, analysis, and reporting of energy-related data. Traditionally, the EIA collects data on natural gas production from each state through a survey completed by state agencies. The rates reported shortly after production are known to underestimate the true monthly production rates. The EIA's problem of interest is finding a way to get early, yet more accurate, estimates of production rates.

The agency which reports Texas natural gas production, the Railroad Commission of Texas (RCT), maintains particularly good records. Within one month of each production month, the RCT reports an initial production rate for that month. This rate, however, is subsequently revised monthly until there are no further changes reported

[1] Simon Fraser University, Department of Statistics and Actuarial Science, 8888 University Drive, Burnaby, BC V5A 1S6, Canada. Email: cdlinkle@stat.sfu.ca
[2] Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC V5A 1S6, Canada. Email: sitter@stat.sfu.ca
Acknowledgments: This research was funded by a grant from the Natural Science and Engineering Research Council of Canada. The authors thank the members of the ASA Advisory Committee to the Energy Information Administration for comments, in particular Jay Briedt, Nick Hentgartner, and Moshe Feder. The authors also thank John Wood for his critique. Special thanks to John Vetter of EIA for his comments and helpful suggestions.

and it is assumed that the true production rate for the month is known. Generally a month's true production rate is not known until approximately a year after production. Initial attempts by the EIA to deterministically calibrate early reported rates to the final reported rates showed some promise, but were complex and ad hoc (see Linkletter 2002). Straightforward parametric modelling with time-variant parameters was also tried, but lacked motivation, involved overly complicated updating and did not perform as well (see Linkletter 2002). Through further exploration of the monthly data and subsequent discussions with the EIA, we consider adjustments to initial values as reflecting a distribution of plausible delays in reporting to the RCT. This would rationalize the observed pattern of updated production totals. As a result, the purpose of this article is to present natural gas production prediction as an important new application of reporting-delay-adjustment methodologies.

The notion of adjusting for reporting delays to more accurately monitor an event has already been studied in detail for epidemiological, insurance, and product warranty applications. Examples include delays in general practitioners' reporting AIDS cases to central collecting agencies (e.g., Harris 1990; Zeger, See, and Diggle 1989), delays between infection with the AIDS virus and the onset of clinical AIDS (Lagakos, Barraj, and De Gruttola 1988), incubation times between blood transfusions and AIDS infection (Kalbfleish and Lawless 1989), lag times between repair claims and the entry of claims into a product warranty database (Kalbfleish, Lawless, and Robinson 1991), and the delays in reporting of liability claims in casualty insurance (Kaminsky 1987).

In this study, we apply versions of these ideas to the Texas natural gas production data. Unlike what is the case in many related applications, the individual observed responses are unavailable; only the cumulative reported total over time is available. But the performance of the methods appears to be satisfactory for many purposes. The value of this work is three-fold: first, the method has been implemented at EIA for the past year with good performance; second, the implication of reporting delays in Texas combined with the even poorer data from other producing states has prompted the EIA to work toward a more direct monthly natural gas survey of producers instead of the current procedure of using data reported by the state, which should eventually yield richer data with individual producer-level reporting times from which to adjust for lags, as well as more direct control over the frame, selection procedures and estimation; and third, it is quite possible that other government agencies which rely on reporting procedures for collecting information may also benefit from the methodology.

An outline of this article is as follows. In Section 2, we present the Texas natural gas production data in more detail, highlighting trends observed in reporting. To better predict natural gas production that has occurred but has not yet been reported, we focus on using a nonparametric estimate of the reporting lag distribution (e.g., Lagakos, Barraj, and De Gruttola 1988; Brookmeyer and Liao 1990; Lawless 1994). This methodology is reviewed in Section 3 within our context of predicting Texas natural gas production. Specifically, we show that it yields reasonably accurate estimates of production rates within three months of production and is simple to implement. In addition, a method for constructing pointwise prediction intervals is presented. Finally, we conclude in Section 4 by discussing some of the interesting features of the Texas data that do not perfectly fit into the nonparametric reporting lag estimation framework and that suggest room for future research.

2. Texas Natural Gas Production

For this study, monthly production data collected by the RCT is available for the production months January 1994 ($t = 0$) to October 2001 ($t = 93$). For each production month $t = 0, 1, \ldots$, a report of the month t production rate is given at lags (in months) $x = 1, 2, \ldots, T - t$, where T is the month in which the study ends. Here, $T = 94$, i.e., one initial reported rate is available for October 2001, while 94 reports are available for January 1994. Let Y_{tx} denote the cumulative rate of month t production reported by month $t + x$, where by "month $t + x$" we mean x months after t. Production rates are reported in units of billion cubic feet per day (Bcf/d). Owing to the nature of the data collection, the rate Y_{tx} that is reported may in fact be an estimated amount. We will return to this point later. In this notation, the last reported cumulative production rate for month t is $Y_{t(94-t)}$ for this data set, or $Y_{t(T-t)}$ in general. It is immediately obvious that the data is right-truncated, since Y_{tx} is only observed for $x \leq T - t$, i.e., the available data, $\{Y_{tx}\}$, can be arrayed in a triangular pattern where $1 \leq t + x \leq T$. An excerpt from the data is given in Table 1.

As previously mentioned, relatively few revisions occur beyond a year after production (see Figure 1). However, not knowing production rates until a year after production runs counter to the EIA mandate of maintaining current records. It is desired to have instead a good estimate of production rates by three months after production at the latest. Figure 1 shows the rates reported 1, 2, 3, and 12 months after productions (Y_{t1}, Y_{t2}, Y_{t3}, Y_{t12}) relative to $Y_{t(94-t)}$ over the study period. We note here that in the following we deal with missing reports by imputing an average. This is meant to be a quick solution that could be

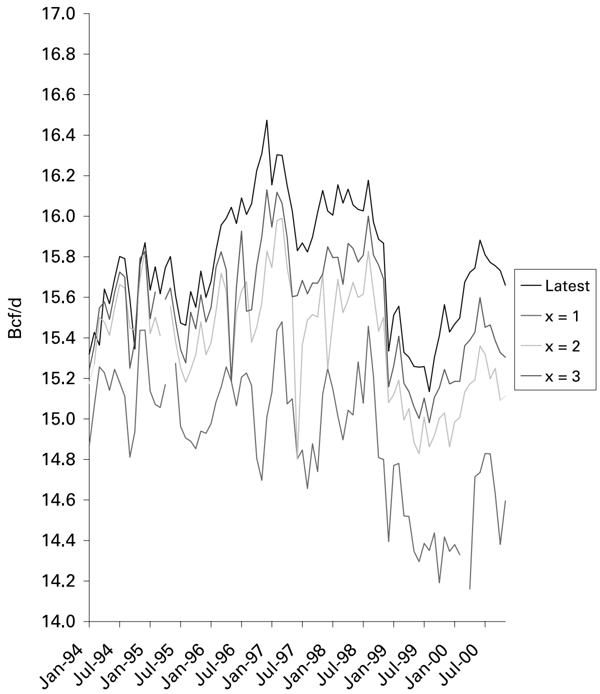


Fig. 1. Y_{tx} for $x = 1, 2$, and 3 with $Y_{t(94-t)}$ for production months January 1994 to November 2000

Table 1. Cumulative rates, Y_{tx}, reported in the last year of the study

	$x = 1$	2	3	4	5	6	7	8	9	10	11	12
Nov-00	14.60	15.11	15.31	15.40	15.46	15.51	15.59	15.62	15.63	15.65	15.66	15.66
Dec-00	14.15	15.11	15.37	15.51	15.58	15.68	15.74	15.75	15.78	15.78	15.78	
Jan-01	14.54	15.21	15.39	15.58	15.64	15.69	15.73	15.78	15.79	15.79		
Feb-01	14.66	15.17	15.46	15.53	15.58	15.70	15.77	15.78	15.81			
Mar-01	14.52	15.15	15.39	15.50	15.62	15.73	15.75	15.77				
Apr-01	14.68	15.23	15.43	15.56	15.69	15.74	15.79					
May-01	14.54	15.14	15.36	15.58	15.65	15.71						
Jun-01	14.08	14.62	15.34	15.46	15.56							
Jul-01	13.95	15.17	15.35	15.44								
Aug-01	14.60	14.97	15.25									
Sep-01	14.32	14.95										
Oct-01	14.33											

refined. It is clear that early reports underestimate true production rates, although the general production trends over time are captured quite well within three months of production. Table 2 gives the pointwise average percent relative bias and MSE between Y_{tx} and $Y_{t(94-t)}$ for $x = 1, 2, 3$, and 12. The challenge is to find a reasonable way to shift or adjust the early reported rates to obtain predictions that are still early, but more accurate.

Evidence of a reporting delay distribution can be seen in Figure 2, which shows two years of revisions for production months in the calendar year 1999. The horizontal axis represents the lag at which each revised production rate was given. It can be seen here that the revision pattern for each month is very similar to that of any other month (and although not displayed here, this pattern is seen for all calendar years). Knowing that the state estimated totals for each month are derived from lower-level data suggests interpreting this pattern as resulting from reporting lags of individual producers. Estimating this reporting lag distribution and using it to adjust early reports might correct the reporting underestimation problem. This can be done without individual values. Figure 2 also shows that the cumulative reported rates Y_{tx} are generally increasing with the reporting lag x, since most new reports are of more production. The fact that the reported rates are not strictly increasing over time is likely due to the fact that reported rates are actually estimates, and subject to error, though the method of estimation is hidden from the EIA. In the following, we do not make special considerations for this feature of the data. We will consider some of the implications of this in the final discussion.

3. Adjusting for the Reporting Lag Distribution

Although there are several unique approaches to adjusting for reporting delays in the literature (Lagakos et al. 1988; Zeger et al. 1989; Harris 1990), we focus on the nonparametric reporting lag distribution estimation as outlined in Lawless (1994). Advantages of this approach include that it can be implemented very easily and it allows for the calculation of pointwise prediction intervals. A subtlety in adapting the method to the present context is that the approach was originally developed to predict counts of events (e.g., AIDS cases or warranty claims) that have occurred but have not yet been reported. For simplicity, however, we view the current problem of predicting natural gas production as discrete by calling an event the production of one cubic foot of gas per day, for example.

3.1. Nonparametric Estimation of the Reporting Delay Distribution

In this section, we briefly describe a general method for estimating the reporting lag distribution and using it to make predictions. Given that reporting ends in month T,

Table 2. Summary of bias and error between Y_{tx} and $Y_{t(94-t)}$ for $x = 1, 2, 3$ and 12 using reported rates made for production months January 1994 to November 2000

	$x = 1$	$x = 2$	$x = 3$	$x = 12$
Bias (%)	−5.5	−2.4	−1.4	−0.2
MSE (Bcf/d)^2	0.828	0.181	0.062	0.004

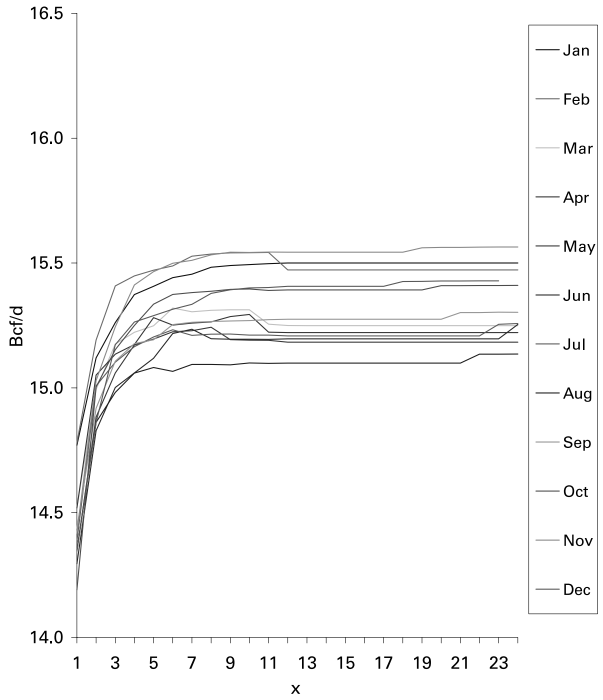


Fig. 2. Y_{tx} over two years of lag, $x = 1, \ldots, 24$ for production months in 1999

the overall goal is to predict the production rate for a month t that will be reported by some month $T^* > T$. For illustration, suppose the latest available revised rate for month t is Y_{t3}, i.e., the study ended at $T = t + 3$. Then, one might wish to predict what rate would be reported by time $T^* = t + 12$. Following Lawless (1994), let $F_t(x)$ be the cumulative probability that natural gas produced in month t is reported by month $t + x$. To predict $Y_{t(T*-t)}$ from $Y_{t(T-t)}$, Brookmeyer and Liao (1990) and Lawless (1994) suggest using

$$\hat{Y}_t(T^* - t) = Y_{t(T-t)} \frac{\hat{F}_t(T^* - t)}{\hat{F}_t(T - t)} \tag{1}$$

provided the necessary estimates $\hat{F}_t(x)$ are available. The nonparametric estimation of $F_t(x)$ used by Lagakos et al. (1988) and Lawless (1994) proceeds as follows.

Let y_{tx} be the month t rate of natural gas production reported in month $t + x$, i.e., reported in the interval $(t + x - 1, t + x)$. The amount y_{tx} can be easily found for the Texas production data by letting $y_{t1} = Y_{t1}$ and $y_{tx} = Y_{tx} - Y_{t(x-1)}$ for $x > 1$. Let $f_t(x)$ be the probability that production occurring in month t is reported in month $t + x$. Note that $F_t(x) = \sum_{r=0}^{x} f_t(r)$.

Neither $f_t(x)$ nor $F_t(x)$ are observable from the data, so consider the conditional probability

$$g_{tx} = f_t(x)/F_t(x)$$

Lawless (1994) notes the following properties of g_{tx}. First, $g_{t1} = 1$ for all t. Also, for any δ such that $0 < x < \delta$

$$\frac{F_t(x)}{F_t(\delta)} = \prod_{r=x}^{\delta} \{1 - g_{tr}\} \tag{2}$$

This can be seen by noting that $\{1 - g_{tr}\} = F_t(r-1)/F_t(r)$. The probabilities g_{tx} are observable for $x \le T - t$, but interest lies in estimating

$$\frac{\hat{F}_t(T^* - t)}{\hat{F}_t(T - t)} = \prod_{x=T-t+1}^{T^*-t} \{1 - \hat{g}_{tx}\}$$

which requires estimates of g_{tx} for $x > T - t$. Brookmeyer and Liao (1990) note that this is similar to the Kaplan-Meier estimator used in standard survival analysis, which turns out to be a useful point for variance estimation.

To obtain the necessary estimates, a relationship among the g_{tx} over time must be assumed so that information from past production months can be used. As does Lawless (1994), assume the g_{tx} are stationary over the last m time periods and into the future, i.e.,

$$g_{tx} = g_x \quad \text{for } t + x \ge T - m + 1 \tag{3}$$

This assumption allows the g_{tx} to vary over time, but through appropriate choice of m minimizes the dependence on production months far in the past. To obtain maximum likelihood estimates of g_{tx} for $x > T - t$, use only data $\{y_{tx}\}$ collected in the last m months along with the final cumulative reports $Y_{t(T-t)}$. Assume that, given $Y_{t(T-t)}$, the amounts y_{tx} for a particular month t, with x ranging from $\max\{T - m + 1 - t, 1\}$ to $T - t$, follow a multinomial distribution with probabilities $f_t(x)/F_t(T - t)$. Then, the appropriate conditional likelihood for month t is

$$L_t \propto \prod_{x=\max\{T-m+1-t,1\}}^{T-t} \left\{\frac{f_t(x)}{F_t(T-t)}\right\}^{y_{tx}} = \prod_{x=max\{T-m+1-t,1\}}^{T-t} g_{tx}^{y_{tx}} \{1 - g_{tx}\}^{Y_{tx}-y_{tx}}$$

The final step above can be seen by noting that

$$\frac{f_t(x)}{F_t(T-t)} = \frac{f_t(x)}{F_t(x)} \frac{F_t(x)}{F_t(T-t)} = g_{tx} \prod_{r=x+1}^{T-t} \{1 - g_{tr}\}$$

using the property of g_{tx} given in (2). Incorporating two additional assumptions yields the required estimates of g_{tx} for $x > T - t$. First, include the stationarity assumption given in (3). Second, assume that the reporting delays for different production months are independent. Thus, the likelihood function over all production months is

$$L = \prod_{t=0}^{T} L_t \propto \prod_{x=1}^{T} g_x^{y_{.x}} \{1 - g_x\}^{Y_{.x} - y_{.x}}$$

where

$$y_{.x} = \sum_{t=max\{T-m+1-x,0\}}^{T-x} y_{tx} \text{ and } Y_{.x} = \sum_{t=max\{T-m+1-x,0\}}^{T-x} Y_{tx}$$

Using this likelihood (which is a product of binomials), the maximum likelihood estimates for g_x are simply

$$\hat{g}_x = y_{.x}/Y_{.x}, \quad x = 1, \ldots, T$$

Recalling (1), predictions for $T^* \leq T + t$ can be made for a production month t using

$$\hat{Y}_{t(T^*-t)} = \frac{Y_{t(T-t)}}{\prod_{x=T-t+1}^{T^*-t} \{1 - \hat{g}_x\}} \tag{4}$$

We emphasize that predictions can only be made for $T^* \leq T + t$. This is sensible for the following reasons. First, it is risky to assume stationarity into the future, but the assumption is milder if predictions are not made too far into the future, i.e., if T^* is not much larger than T. Second, it implies that in order to make a prediction for lag $x = 12$, for example, one must go back in time until at least one previous month has had production reported at $x = 12$.

In the next section we show how this simple reporting-delay adjusted prediction methodology can be successfully tailored to the natural gas problem.

3.2. *Predicting Natural Gas Production*

In the natural gas problem, suppose that for a production month t reported rates are given until $T = t + i$, for $i = 1, 2$, or 3. The aim is to predict the cumulative rate that will be reported by $T^* = t + 12$. First, given the data $\{Y_{tx}\}$ for $1 \leq t + x \leq 94$, obtain the incremental reports y_{tx}. Then, under the stationarity assumption given in (3), the conditional probabilities g_x, $x = i + 1, \ldots, 12$, can be estimated by

$$\hat{g}_x = \frac{\sum_{j=\max\{t+i-m+1-x,0\}}^{t+i-x} y_{jx}}{\sum_{j=\max\{t+i-m+1-x,0\}}^{t+i-x} Y_{jx}} = \frac{\sum_{j=\max\{t+i-m+1-x,0\}}^{t+i-x} y_{jx}}{\sum_{j=\max\{t+i-m+1-x,0\}}^{t+i-x} \sum_{r=1}^{x} y_{jr}} \tag{5}$$

and the desired predictions using

$$\hat{Y}_{t12} = \frac{Y_{ti}}{\prod_{x=i+1}^{12} \{1 - \hat{g}_x\}} \tag{6}$$

Various values of m can be tried to see which is the most reasonable stationarity assumption for the data and yields the best predictions. We found $m = 6$ to $m = 9$ performed best.

Although the expression given in (5) appears complicated, the predictions can be calculated easily. Figure 3 illustrates the data, $\{y_{tx}\}$, required to obtain the estimate $\hat{g}_4$ for November 2000 when $m = 9$; $\hat{g}_4$ is merely the ratio of the sum of cells in the more darkly shaded box over the sum of all shaded cells.

The performance of the method when applied to the natural gas production data is demonstrated in Figure 4. In Figure 4(a), the initial reported rate, Y_{t1}, is shown along with the predictions $\hat{Y}_{t12}$ found by adjusting Y_{t1} for the reporting delay using (6) with $i = 1$, for several production months. For comparison, $Y_{t(94-t)}$ is also shown. Figure 4(b) shows the same using the rates reported three months after production. Clearly, the method reduces the underestimation problem. Figure 4 also reveals, however, that the poorest predictions are made in late summer to early fall in 1996. This is a byproduct of the nonparametric estimation procedure and the size of the assumed window of stationarity. The point predictions during this time are influenced by the lower production rates in the preceding 6–9 month window. This demonstrates the need to be aware when choosing a window size. The 6–9 month window we chose performed best overall, though it is not ideal at every prediction point.

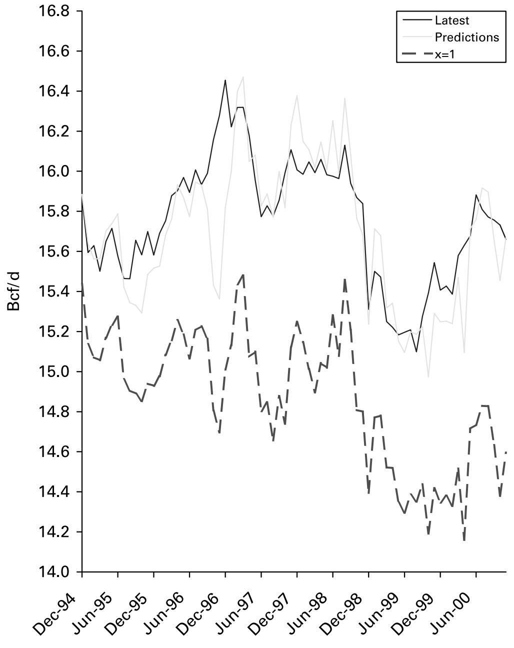
In Table 3 the pointwise average percent relative bias and MSE are given between $\hat{Y}_{t12}$ and $Y_{t(94-t)}$ when predictions are made at 1, 2, and 3 months after production (using $m = 6$ and $m = 9$). The values with no adjustment (UN in Table 3) are repeated from Table 2 to aid comparison. These results are averaged over the number of production months for which predictions were made. The number of months varied slightly depending on the amount of previous data required for each calculation. Clearly, choosing how many months after production to wait before making predictions depends on the preferred trade-off between speed and accuracy. In practice it would be ideal to continuously update predictions as more reports become available. This can be done easily, and has been implemented at EIA.

3.3. Variance Estimation and Prediction Intervals

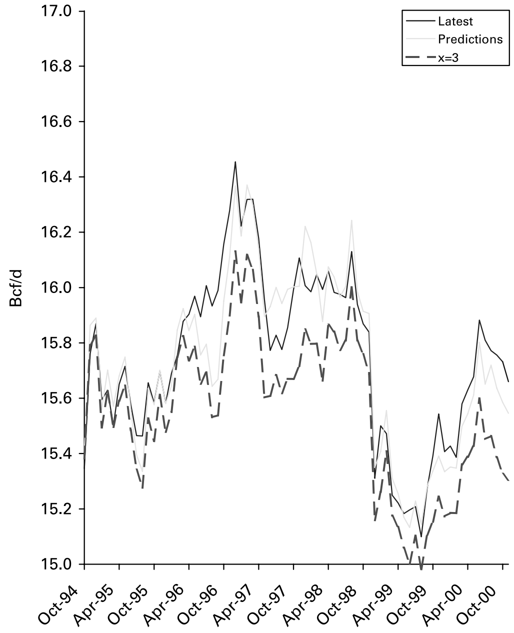
An advantage to making predictions by assuming a probability distribution for the reporting lag as described in the previous sections is that it yields a simple way to obtain pointwise prediction intervals for $Y_{t(T^*-t)}$. From Lawless (1994), an estimate of the

	x=1	2	3	4	5	6
Jan-00	14.380	0.606	0.200	0.046	0.038	0.046
Feb-00	14.330	0.677	0.178	0.073	0.041	0.077
Mar-00	14.516	0.617	0.228	0.082	0.089	0.068
Apr-00	14.161	1.012	0.216	0.134	0.087	0.019
May-00	14.715	0.477	0.236	0.145	0.033	0.061
Jun-00	14.735	0.624	0.240	0.059	0.072	0.047
Jul-00	14.829	0.491	0.132	0.180	0.054	0.024
Aug-00	14.828	0.370	0.266	0.093	0.064	0.015
Sep-00	14.633	0.616	0.137	0.122	0.043	
Oct-00	14.380	0.710	0.236	0.141		
Nov-00	14.596	0.518	0.191			
Dec-00	14.151	0.962				
Jan-01	14.542					

Fig. 3. The data required to calculate $\hat{g}_4$ for November 2000 when $m = 9$. Summing over the whole shaded rectangle gives $Y_{\cdot 4}$ and summing the boxed column gives $y_{\cdot 4}$. The value $\hat{g}_4$ is $y_{\cdot 4}/Y_{\cdot 4}$. The height of the rectangle is $m = 9$



(a) Predictions made at time t + 1 along wih Y_{t1}



(b) Predictions made at time t+3 along wih Y_{t3}

Fig. 4. Predictions $\hat{Y}_{t12}$ using $m = 9$ along with $Y_{t(94-t)}$ for a range of production months

Table 3. Summary of bias and error for predictions found at t + i, and unadjusted (UN)

	$i = 1$			$i = 2$			$i = 3$		
	$m = 6$	$m = 9$	UN	$m = 6$	$m = 9$	UN	$m = 6$	$m = 9$	UN
Bias (%)	−.27	−.35	−5.5	0.026	0.025	0.18	0.012	0.011	0.06

asymptotic covariance matrix for $\hat{\boldsymbol{g}} = (\hat{g}_1, \ldots, \hat{g}_T)$ using multinomial large sample theory is the $T \times T$ matrix

$$\hat{V}_{\hat{g}} = diag(\hat{g}_x\{1 - \hat{g}_x\}/Y_{.x}) \tag{7}$$

For ease of notation, let

$$\hat{W}_t = \prod_{x=T-t+1}^{T^*-t} \{1 - \hat{g}_x\}$$

Then, using (7), the variance of $\hat{W}_t$ can be approximated by

$$\hat{V}_{Wt} = \hat{W}_t^2 \sum_{x=T-t+1}^{T^*-t} \left(\frac{\hat{g}_x}{Y_{\cdot x}\{1 - \hat{g}_x\}} \right)$$

which is essentially equivalent to Greenwood's formula for the variance of the product limit estimator used in survival analysis (Kaplan and Meier 1958).

The prediction interval calculation given in Lawless (1994) proceeds as follows. Consider

$$Z_t = Y_{t(T^*-t)} - \hat{Y}_{t(T^*-t)}$$

Using the predictor $\hat{Y}_{t(T^*-t)}$ given by (4), the variance of Z_t can be approximated by

$$\hat{V}_{Z_t} = \frac{Y^2_{t(T-t)}}{\hat{W}_t^4} \hat{V}_{W_t}$$

assuming that the totals $Y_{t(T^*-t)}$ and $Y_{t(T-t)}$ are fixed (not random) amounts. An approximate $1 - 2\alpha$ prediction interval for $Y_{t(T^*-t)}$ is then given by

$$\hat{Y}_{t(T^*-t)} \pm z_\alpha \hat{V}_{Z_t}^{1/2} \tag{9}$$

where z_α is the standard normal $1 - \alpha$ quantile.

To illustrate with the Texas production data, if a prediction for month t production, $\hat{Y}_{t12}$, was made at $t + 3$, the interval of interest would be $\hat{Y}_{t12} \pm z_\alpha \hat{V}_{Z_t}^{1/2}$, with $\hat{V}_{Z_t} = \{Y_{t3}^2/\hat{W}_t^4\}\hat{V}_{Wt}$. Figure 5 shows the predictions $\hat{Y}_{t12}$ made at $t + 3$ (using $m = 9$) for several production months, along with the latest reported totals $Y_{t(94-t)}$ and the corresponding pointwise 90% prediction intervals. Only the final production rates for July, August, and September 1996 fall outside their prediction interval (recall this was the period with poor predictions).

We caution that the variance estimation and prediction interval calculation described here is only appropriate for one production month at a time. See Lawless (1994) for a discussion on constructing simultaneous prediction intervals for a block of predictions.

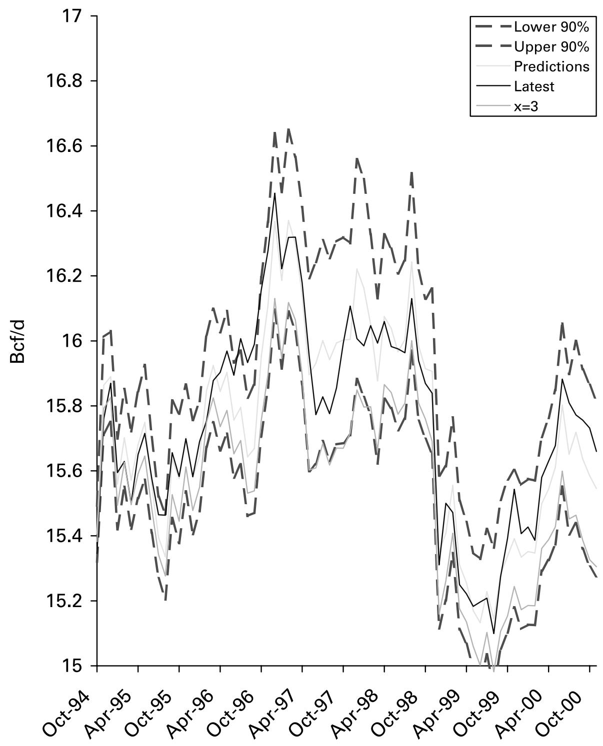


Fig. 5. 90% prediction intervals for $\hat{Y}_{t12}$ (found using $m = 9$) along with Y_{t3} and $Y_{t(94-t)}$

4. Discussion

The goal is to find a method for making early yet accurate predictions of natural gas production. By hypothesizing the existence of individual reporting lags, we are able to do so using a straightforward reporting-delay-adjustment methodology. In return, we were able to present an important new application for the methodology. It is hoped that other government agencies which encounter similar problems may also benefit.

Although the nonparametric delay-distribution estimation procedure is easy to use and quite powerful, it is important to point out existing weaknesses and potential areas for future investigation. As previously mentioned, reported rates are estimated, and as a result sequential revisions are not strictly increasing as would be expected. Consequently incremental reports y_{tx} can be negative, and therefore estimates of the conditional probabilities g_x can be slightly less than zero. Allowing the negative estimates or truncating them to zero are simple solutions, but slightly less than desirable. An alternative approach may be to assume a parametric form for the delay distribution. As noted by Kaminsky (1987), one reasonable assumption might be that the conditional reporting lag distribution is exponential, since reporting is highest in the first few reporting periods and then gradually tapers off. This would also eliminate the need to discretize the problem, and may make it easier to incorporate possible serial correlation over time so as to provide a mechanism for creating tighter prediction intervals.

We also note that the nonparametric distribution estimation does depend on having monthly revisions. Most states do not currently maintain that level of detailed records, although the results in this article suggest there are many benefits to doing so. In the end, our main motivation for using the nonparametric approach despite the concerns that may exist is the fact that it does work quite well and, most importantly, is extremely easy to implement and has been implemented by the EIA for the past two years.

Finally, in part, the process and results of this study motivated the EIA to implement a direct monthly survey of producers. This survey returned its first preliminary results recently and verified the hypothesized individual reporting lags. As this richer data set develops over time, more sophisticated methods of correcting for potential lagged reporting may be possible.

5. References

Brookmeyer, R. and Liao, J. (1990). The Analysis of Delays in Disease Reporting: Methods and Results for the Acquired Immuno Deficiency Syndrome. American Journal of Epidemiology, 132, 355–365.

Harris J.E. (1990). Reporting Delays and the Incidence of AIDS. Journal of the American Statistical Association, 85, 915–924.

Kalbfleisch, J.D. and Lawless J.F. (1989). Inference Based on Retrospective Ascertainment: An Analysis of the Data on Transfusion-related AIDS. Journal of the American Statistical Association, 84, 360–372.

Kalbfleisch, J.D., Lawless, J.F., and Robinson, J.A. (1991). Methods for the Analysis and Prediction of Warranty Claims. Technometrics, 33, 273–285.

Kaminsky, K.S. (1987). Prediction of IBNR Claim Counts by Modelling the Distribution of Report Lags. Insurance: Mathematics and Economics, 6, 151–159.

Kaplan, E.L. and Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association, 53, 457–481.

Lagakos, S.W., Barraj, L.M., and De Gruttola, V. (1988). Nonparametric Analysis of Truncated Survival Data, with Application to AIDS. Biometrika, 75, 515–523.

Lawless, J.F. (1994). Adjustments for Reporting Delays and the Prediction of Occurred But Not Reported Events. Canadian Journal of Statistics, 22, 15–31.

Linkletter, C. (2002). Predicting Natural Gas Production in the Presence of Reporting Delays. M.Sc. Thesis, Simon Fraser University, Burnaby, BC, Canada.

Report #:DOE/EIA-0484 (2005). International Energy Outlook 2005. DOE/EIA Report. http://www.eia.doe.gov/oiaf/ieo/nat__gas.html

Zeger, S.L., See, L.-C., and Diggle, P.J. (1989). Statistical Methods for Monitoring the AIDS Epidemic. Statistics in Medicine, 8, 3–21.

Received July 2005
Revised October 2006