ACCOUNTING FOR BUSINESS BIRTHS AND DEATHS IN CES: BIAS VS. NET BIRTH/DEATH MODELING

Jurgen Kropf, Sharon Strifas, Monica Traetow, U.S. Bureau of Labor Statistics
Monica Traetow, 2 Massachusetts Avenue N.E., Room 4860, Washington, DC 20212

Key Words: bias adjustment, net birth/death model

1. Introduction

The Bureau of Labor Statistics (BLS) cooperates with State employment security agencies in the Current Employment Statistics (CES) survey to collect data each month on employment, hours, and earnings from a sample of nonfarm establishments (including government). This sample includes approximately 350,000 reporting units. From this data, a large number of employment, hours, and earnings series in considerable industry and geographic detail are prepared and published each month[1]. The employment data includes series on all employees, women workers, and production (or nonsupervisory) workers. Hours and earnings data include average weekly hours, average weekly overtime hours, and average hourly and weekly earnings. For many series, seasonally adjusted data are also published.

In June of 1995, BLS announced plans for a comprehensive sample redesign of its monthly payroll survey, from a quota-based sample to a probability-based sample. The initial research phase for the CES sample redesign was completed in 1997, and BLS launched a production test of the new sample design at that time. In June 2000 the first estimates from the redesign were published for the major industry division of wholesale trade. Next, in June 2001, the mining, construction, and manufacturing divisions were published under the redesign estimates. Then, in June 2002, the estimates for the TPU, retail trade, finance, insurance, and real estate divisions were published under the redesign. The remaining industry, services, will be phased in with the 2003 June benchmark release[2].

Since business births and business deaths cannot be captured in a timely manner in the CES survey, a model is needed to account for these in the CES estimates. The probability sample uses imputation, as well as the net birth/death model, while the quota sample uses a bias adjustment factor model.

1.1 Universe

The Longitudinal Data Base (LDB) is the universe from which BLS draws the CES sample and is derived from state unemployment insurance (UI) records submitted by employers who are subject to state unemployment insurance laws. The LDB contains data on approximately 7.5 million records for the year 2001, representing nearly all nonfarm business establishments of the U.S. economy. These reports provide information on the number of people employed and the wages paid to employees each quarter[3]. This database is also the main source for the annual CES benchmarks.

1.2 Non-sampling error in the CES survey

As in all surveys, non-sampling error exists in the CES survey. The inability to capture business births on a timely basis is believed to be the largest component of this non-sampling error. This is due to the fact that a lag exists between an establishment opening for business and its appearance on the universe frame, where it would be available for sampling. Another significant component of non-sampling error is the inability to capture business deaths in a timely manner. If a sampled firm goes out of business, it typically does not report for that month rather than report zero employment. Follow-up calls may prove that the firm closed down, but this information is received too late to be incorporated into the monthly employment estimates. Due to the size of the CES sample, the immediate investigation and determination of the status of a non-responding reporter is impossible[4]. In the event that zero employment is reported by a sample respondent, it is used in the estimation process.

Other sources of non-sampling error in the CES survey are non-response, response, and processing errors[5]. Non-response problems include failure to locate establishments in the sample, a respondent who is unable to provide the information the interviewer asks for, refusal by the establishment to participate in the survey, or a reporter who reports late. Furthermore, response errors can occur when a respondent does not give accurate information. Finally, processing errors are human errors that may occur when editing, coding, and tabulating the results of the survey.

[It is also important to note that the universe from which the CES sample is drawn contains some non-response and processing errors as well, even though it serves as the major benchmark source for the CES survey.]

2. Bias Adjustment Factors for the Quota Sample

2.1 Reasons for the bias adjustment factor model

A bias adjustment is needed for the current CES quota estimates primarily to account for the employment growth of new firms and to reduce other components of non-sampling error in the survey. If this adjustment was not made, substantial underestimation of employment levels would occur and the "bias" would be for the estimates to vary in one direction systematically and persistently.

2.2 Methodology of bias adjustment factor model

The model used to compute the bias adjustment factors is a "regression-adjusted mean-error model," shown in Table 1. The primary input to the model is the *mean-error* (ME3). With this component, other elements of non-sampling error in the survey are adjusted for as well.

The bias adjustment factor model has been in effect since the early 1960s in the CES survey, in conjunction with the quota sample. Prior to 1983, the model was solely a mean-error model, determined by the variable ME3. During those years, economic cycle changes were not incorporated into the bias adjustment factors. If an economic downturn happened, for example, the number of business births was slightly overstated. After 1983, the model was expanded to include recent economic swings as shown by the other remaining variables in Table 1.

The bias adjustment factors are computed one quarter at a time. The employment added from the use of bias adjustment factors may differ slightly from the time the factors are initially computed to the time of the monthly press release. This is due to the multiplicative nature of the bias factors, and which closing estimates are used in the model and when. Often the difference is referred to as the "target" bias versus the "actual" bias.

Table 1. Regression-Adjusted Mean-Error Model

$$BF_c = b(ECH_{c-1} - AECH3) + ME3 \qquad (1.)$$

BF_c	The current quarter bias factor
ME3	The average difference calculated as a ratio between the sample-based estimates (without bias adjustments) and the benchmark levels over the past three years.
AECH3	The average benchmark seasonally adjusted employment over-the-year change over the past three benchmark years.
ECH_{c-1}	The seasonally adjusted over-the-quarter change in employment from the previous two quarters. This number is annualized, or multiplied by 4, to be comparable to AECH3 units.
b	A coefficient derived from a regression model: $ME3=b_1i(AECH3_1i) + b_2i(PCTCOV_2i) + e$ Inputs to the model are ME3, $AECH3_1i$, and $PCTCOV_2i$. 'b_1' is calculated by major industry level and 'i' is the regression applied to each division. 'b_1' is updated annually for each industry division. The error term is 'e'.

3. Net Birth/Death Model for the Probability Sample

3.1 Reasons for the net birth/death model

The redesign methodology accounts for business births in the estimates through imputation of those reporters that appear to have zero employment. For such reporters, the corresponding cell-specific link (see formula 4) is applied to the previous month's employment to obtain the current month's employment count. The links are calculated using the actual reported data from all other reporters in the estimating cell. A CES estimating cell is defined as a group of reporters with similar industry and regional characteristics. This procedure serves two purposes, (1) to account for business births through the imputation of out-of-business reporters and (2) to eliminate the need to determine the real status of a reporter in the current month. Therefore, the imputed employment from the out-of-business reporters is used to account for a portion of the business births.

However, this method can only *approximate* the actual employment generated by business births; recall that there is a lag between the time that a business opens and its availability for sampling (when it appears on the universe). It is important to note that the growth of new firms is often significantly different than the growth of established firms in the sample. To minimize the imputation error generated by this methodology, a net birth/death model has been developed which calculates the effect of the imputation, measures the imputation error, and generates a forecast of this error to adjust the current estimate. The following section discusses the structure and components of the net birth/death model.

3.2 Methodology of the net birth/death model

This methodology assumes that the probability sample accurately represents the universe (the LDB). By using the universe data as input in the net birth/death model and applying the same methodology to the data as is applied to the sample when accounting for business births, the historical imputation error can be measured with certainty.

The input data from the LDB reflects the different levels of information regarding business births available to the sample over time. The sample is selected from this database in the first quarter of every year with a lag of 12 months because UI information is not available for more current months. At the time of sample selection all business births, up to the selection quarter, are known and these units are eligible to be selected for the sample. At this time (March of every year) the monthly CES estimates are benchmarked as well and under the redesign methodology a new estimation cycle begins for the post benchmark months. The post benchmark period is defined as the April following the benchmark through the following March. In order for the LDB to reflect these information levels, 25-month frames are created beginning in March of every year; for example, the frame for the benchmark year 2001 starts in March of 1999 and ends in March 2001.

These frames are used as input for the net birth/death model which iterates through the following steps to calculate the imputation error for each cell:

1. The model creates a subset from the universe of continuous establishments (CE), defined as establishments with employment (EMP_{BMK}) greater than zero for the benchmark month. These establishments are eligible for selection in the sample.

$$CE \text{ where } EMP_{BMK} > 0 \qquad (2.)$$

2. Each record in this subset is scanned from the first to the last month of the frame for non-zero employment. When a record displays zero employment (EMP_t), an out-of-business firm is detected and the employment for this month is imputed ($EMP_{t,imp}$). The imputed value is calculated by applying a link (LK_t) to the previous month's employment and replacing the zero value with the calculated value for this month. The link is derived from the sum of the over-the-month change of every record in the cell (C) which does not have zero employment ($\Sigma EMP_{t,EC}$).

$$EMP_{t,imp} = EMP_{t-1} * LK_t \qquad (3.)$$

where t is the month with $EMP_t = 0$ and

$$LK_t = \Sigma EMP_{t,EC} / \Sigma EMP_{t-1,EC} \qquad (4.)$$

where EC are all other establishments in the cell with $EMP_t > 0$.

This process is continued for every month when an establishment displays zero employment.

3. The records with continuous employment ($EMP_{t,EC}$) for each month of the frame and the records for which employment is imputed ($EMP_{t,imp,C}$) are summed for each cell. This series represents the estimate ($EST_{t,c}$)if the sample were the universe.

$$EST_{t,C} = \Sigma\ EMP_{t,EC} + \Sigma\ EMP_{t,imp,C} \quad (5.)$$

4. This step calculates the imputation error ($ERR_{imp,C}$) by subtracting the universe-based employment estimate from the employment of the total population ($EMP_{pop,C}$) in the LDB for each cell and month. The total population contains the actual business births and deaths for the time period each frame covers.

$$ERR_{t,imp,C} = EMP_{t,pop,C} - EST_{t,C} \quad (6.)$$

5. The resulting imputation error series ($ERR_{imp,C}$) contains 25 months of data for each production cycle from the benchmark month (when the error is zero because the business births are completely captured), to the last month of the cycle, which is March of the second year after the benchmark year. The first year of the cycle represents the post-benchmark year and the second year is the current production year. The cumulative imputation error for the production cycle is generally larger than the cumulative imputation error from the post-benchmark cycle because of the increased time difference from the benchmark month. Therefore, the forecast from the net birth/death model has to be based on two different time spans, the post-benchmark months ($ERR_{PB,C,BMK}$) and the production months ($ERR_{PR,C,BMK}$). The frames of 25 months are split into 12 post-benchmark months and 13 production months.

$$ERR_{imp,C,BMK} = \{\ ERR_{BMK,t,C}\ ,\ ERR_{BMK,t+1,C}\ ,\ \ldots,\ ERR_{BMK,t+12,C}\ ,\ \ldots,\ ERR_{BMK,t+24,C}\ \} \quad (7.)$$

For the post-benchmark forecast:

$$ERR_{PB,C,BMK} = \{ERR_{BMK,t+1,C},\ldots,ERR_{BMK,t+12,C}\} \quad (8.)$$

For the production forecast:

$$ERR_{PR,C,BMK} = \{\ ERR_{BMK,t+12,C}\ ,\ldots,\ ERR_{BMK,t+24,C}\ \} \quad (9.)$$

This procedure is repeated for each benchmark year and the corresponding sections from each frame are joined to form a time series, which consists of over-the-month changes of the errors, as input for the net birth/death forecast ($NETBD_{PB/R,C}$). The input series consists of 5 benchmark years. This example shows the inputs for the benchmark-2001 cycle.

For post-benchmark forecast:

$$NETBD_{PB01,C} = ERR_{PB,C,BMK97} \parallel ERR_{PB,C,BMK98} \parallel ERR_{PB,C,BMK99} \parallel ERR_{PB,C,BMK00} \parallel ERR_{PB,C,BMK01} \quad (10.)$$

For production forecast:

$$NETBD_{PR02,C} = ERR_{PR,C,BMK97} \parallel ERR_{PR,C,BMK98} \parallel ERR_{PR,C,BMK99} \parallel ERR_{PR,C,BMK00} \parallel ERR_{PR,C,BMK01} \quad (11.)$$

6. During this step two series are forecasted from the input series, $NETBD_{PB01,C}$ (for the post-benchmark period) and $NETBD_{PR02,C}$ (for the production period.) The forecasts are generated by the standard statistical software program X12-ARIMA[6] which also tests the series for significant outliers. The forecasts are applied to the cell estimates for each month. In addition the forecasts for the production period are updated quarterly as more recent information from the UI program becomes available. As shown in Table 2, the forecast differs for the post-benchmark and production period because of the differences in elapsed time from the benchmark month.

Table 2. Bias Adjustment Effects Versus Net Birth/Death Model Effects for the Mining, Construction, and Manufacturing Industries (in thousands)[7]

Year and Month	Mining, Construction, and Manufacturing Industries	
	Bias adjustment for published series	Net birth/death adjustment for the post-benchmark period
	Monthly amount	
2000:		
April	20	45
May	20	64
June	20	43
July	19	15
August	19	30
September	19	20
October	23	6
November	23	-10
December	23	-13
2001:		
January	20	-114
February	20	23
March	20	45
Cumulative Total (4/2000 to 3/2001)	246	154
2001:		
April		44
May		55
June		41
July	Bias not calculated because of the transition to the net birth/death model.	8
August		28
September		16
October		10
November		-4
December		-6
2002:		
January		-99
February		20
March		43
Cumulative Total (4/2001 to 3/2002)		156

3.3 Interpretation of the net birth/death model

The forecasted monthly amounts from the net birth/death model should not be interpreted as an independent time series or even be characterized as displaying a distinct seasonal behavior. Tests for seasonality using the X12-ARIMA software did not support the presence of consistent seasonal behavior in the series, especially when taking into consideration that the post-benchmark and production forecasts are based on different input series, representing different information levels. However, the forecasts from the net birth/death model should be interpreted as an error adjustment to the imputed employment for business births included in the initial estimates. The evaluation of the net birth/death forecasts should only be conducted in conjunction with the final estimates, as the model values are complimentary to the sample information. However, the net birth/death forecast does have an influence on the seasonal behavior of the final series that can result in increases or decreases in seasonal fluctuations, depending on the month and the industry. BLS publishes the net birth/death adjustment monthly at an aggregate level to inform the user about the effect the model has on the final estimates.

4. Final Remarks

The bias adjustment is a model-based adjustment designed to account for business birth absence, as well as other errors in the quota-based sample estimates. The new probability sample accounts for part of business births through the imputation of business deaths. The remaining portion that must be modeled is a small component. The net birth/death model addresses *only* this component of the non-sampling error, and not the other components that may exist in a survey. Because of its improved design, the probability sample methodology does lessen the risk of bias found in the quota sample. The redesign adjustment from the net birth/death model under the probability sample is expected to be smaller overall than the bias adjustment factors under the quota sample (see Table 2).

As a final note, the most significant potential drawback to any model-based approach is that time series modeling assumes a predictable continuation of historical patterns and relationships. Therefore, a model-based approach is likely to have some difficulty producing reliable estimates at economic turning points or during periods in which there are sudden changes in trend. With the net birth/death model, this difficulty is significantly reduced, as imputation allows for such trends or turning points to be partially captured in the estimates. These imputed amounts are based on the active, reporting establishments which reflect changes in the economy. Therefore, contributions from the imputation for business births will also vary during economic changes. Furthermore, the imputation error from the net birth/death model is forecasted from the errors of the last five years. The error amounts from the most recent year are weighted more heavily, thus allowing recent changes to be more effectively captured. Even so, accurate estimation of the business birth component of total nonfarm employment will continue to be the most difficult issue in CES employment estimation.

Notes

[1]John F. Stinson, Jr. (eds.), "Establishment data," *Employment and Earnings*, July 2001, pp.181.

[2]John F. Stinson, Jr. (eds.), "Establishment data," *Employment and Earnings*, July 2002, pp.184.

[3]John F. Stinson, Jr. (eds.), "Establishment data," *Employment and Earnings*, July 2001, pp.193.

[4]John F. Stinson, Jr. (eds.), "Establishment data," *Employment and Earnings*, June 2000, pp. 187.

[5]Patricia M. Getz and Theodore C. Logothetti, "BLS establishment estimates revised to incorporate March 1999 Benchmarks," *Employment and Earnings*, June 2000, pp. 4.

[6]*X-12-ARIMA Reference Manual, Beta Version 1.0* (1996) Bureau of the Census.

[7]*CES Bias Adjustment and Net Birth/Death Model*. Retrieved May 13, 2002, from http://stats.bls.gov/web/cesbd.htm/.