Numerical Comparison of Small Domain Estimators Computed from Current Employment Statistics Data October 2002

R.M. Harter, J. Gershunskaya, J.L. Eltinge, L. Huff and S. Lehan
R.M. Harter, National Opinion Research Center, 55 East Monroe, Suite 4800, Chicago, IL 60603
harter@norcmail.uchicago.edu

Key Words: **Covered Employment and Wages Program (ES-202); Empirical evaluation; Mismatched units; Outlier detection; Small area estimation; Synthetic estimator.**

1. Introduction: Small Domain Estimation in the Current Employment Statistics Program

The Current Employment Statistics Program (CES) uses a large-scale sample survey to produce monthly estimates of total employment, earnings and hours worked in the U.S. labor force. For some general background on this program, see Bureau of Labor Statistics (1997), Werking (1997), Butani et al. (1997), West et al. (1997) and references cited therein. Historically, methodological work for the CES has focused on production of national- or state-level estimates within specified industries. However, many stakeholders recently have expressed strong interest in the development of methodologically sound estimates of total employment for relatively small domains. For the current study, these domains are defined by the intersection of major industrial divisions (MIDs, as defined under the Standard Industrial Classification system) with Metropolitan Statistical Areas (MSAs).

For some general background on small domain estimation for the CES, see Harter et al. (1999), Eltinge et al. (2001) and references cited therein. For the current discussion, three points are of principal interest. First, the CES is closely related to the Covered Employment and Wages (ES-202) program of the BLS, which is a nominal census of relevant employment, and is based on administrative records in the unemployment insurance (UI) system. For the current discussion, the UI data for a given establishment j and month t, denoted $x_{(j)t}$, will be treated as true (i.e., error-free) values. A partial limitation of the ES-202 data is that they become available four months or longer after the end of the applicable quarter.

Second, researchers have considered several possible CES-based estimators of employment for small domains, including the following.

a. A direct estimator based primarily on CES data from the domain of interest, denoted $\hat{Y}_1$.
b. A predictor based on a time series model for the underlying ES-202 data, denoted $\hat{Y}_2$.
c. A synthetic estimator based on a simple model for state x MID -level growth, denoted $\hat{Y}_4$
d. A weighted least squares estimator, denoted $\tilde{x}_{wls}$, that is a weighted sum of $\hat{Y}_1$, $\hat{Y}_2$ and $\hat{Y}_4$.
e. A somewhat more complex estimator known as the NORC/Illinois small domain engine, denoted here as $\hat{Y}_5$. For a given domain (iat), this estimator is computed as a sum,

$$\hat{Y}_{5iat} = \sum_{j \in C_{iat}} y_{(j)t} + \sum_{j \in R_{iat}} y_{(j)t} + \sum_{j \in N_{iat}} \hat{y}_{(j)t}$$

where C_{iat} is the set of certainty units in MID i and MSA a, which responded in month t; R_{iat} is the set of sampled noncertainty units in MID i and MSA a, which responded in month t; N_{iat} is the set of all units that are contained in the ES-202 files for MID i and MSA a, and are not contained in C_{iat} or R_{iat}; and $\hat{y}_{jt}$ is a synthetic estimator based on estimated state-level rates of change generally computed at the two-digit SIC level for noncertainty units.

Third, under different sets of specified conditions and optimality criteria, any one of these estimators may be approximately optimal. One very important feature of small domain estimation work for the CES program is that the availability of the ES-202 true values $x_{(j)t}$ on a lagged basis allows one to carry out a direct empirical evaluation of the performance of these estimators for a given set of domains and months. Consequently, the CES program has the potential to provide some insights into the empirical properties of small domain estimators in a real-data setting, and thus to provide a very valuable complement to traditional theoretical and simulation-based approaches to evaluation of small domain estimator properties. In work not reported here, the authors have explored a large number of properties of the abovementioned small domain estimators. In the interest of space, the current discussion will focus on one specific topic: the effects of editing CES data for possibly outlying or mismatched units. Section 2 discusses some methods for identification of outlying or mismatched units. Section 3 presents some empirical results.

2. Methods to Identify Possibly Outlying or Mismatched CES Observations

2.1 Initial screening, editing, review and adjustment methods

The microdata files used for this study passed through several screening, editing, and possibly review & adjustment steps. When the interviewers initiate data collection for a given sample unit j, the BLS compares initial reported CES employment level ($y_{(j)1}$, say) with the employment level ($x_{(j)B}$, say) provided by the sample frame for unit j in a baseline month. That edit step requires the data collector to clarify and code observed differences between $y_{(j)1}$ and $x_{(j)B}$. The possible clarifications offered by a data collector include, but are not limited to, the following categories: employment growth or decline; addition of an expansion worksite; or merger or acquisition. This initial screening test is designed to determine whether the unit collected is the same as the unit sampled. There are several tests that are used and the unit only has to pass one of these tests to pass the edit. The parameters of the edit are designed to allow a maximum number of units pass this test and fail only the most egregious cases. The reasoning behind this rule is that the analyst is the best judge as to whether observed differences are attributable to true changes in employment level, or to problems with reporting error, mismatching of units or other nonsampling errors.

After the initiation of the responding units there is an edit step that attempts to identify whether the monthly reported AE values are in line with data reported for earlier months. As in the edit described above, this test is designed to allow a maximum number of reports pass this test. Only the most egregious cases are failed. Again, there are actually several tests and the AE value has to pass only one of the tests to pass the edit.

The third cause for a reported value $y_{(j)t}$ to be flagged for review and possible correction is a comment code of 90 or 91. These comment codes indicate that the unit has changed its basis for reporting, i.e., it is not reporting for the same unit that it had previously been reporting. Probably the most common event to cause a unit to be coded in this manner is the merger of a company causing the report to include the merged company report rather than the previously reported portion of the merged company.

When a unit fails based on the first two edits described above they are supposed to be sent to statistical methods division for review and disposition. If, after the review by the statistical methods division, it is clear what the unit is reporting and the reporting meets acceptable CES reporting options, then corrections are made in the files to reflect exactly what the unit is reporting with respect to what was sampled. These corrections adjust the files to indicate a respondent's current method of reporting – such as a mult-RUN UI respondent giving all data one report when it initially reported data with several reports. If needed, the unit may also be reweighted to better reflect the probability of selection of the reporting unit. If it cannot be determined what the reporting unit is including in its report with respect to the sample frame, then the unit may be returned to the data collection center for re-contact and clarification. The unit may again be returned to the statistical methods division for review and correction. If, in the final analysis, it cannot be determined what the unit is reporting, then the unit may be coded unusable. If the unit is reporting in a manner inconsistent with acceptable CES reporting options, then it may also be coded unusable.

Units that are coded with a comment code of 90 or 91 are reviewed by the statistical methods division in a manner similar to that described in the above paragraph. A complicating factor in the initial data collection is that the edits were not run properly in all cases and in some cases the units were used in estimation and sent to states even though they failed edits. The statistical methods division did not receive the failed cases on a flow basis and approximately 6,000 cases were sent to the statistical methods division for review with little time for the review. Many of the cases that were passed and not reviewed are almost certainly problem cases that will need review at some time in the future.

2.2 Additional related methods

The NORC estimator $\hat{Y}_5$ required matching of CES and ES202 employment at the micro level, generally at the establishment level. The smaller the domain of interest, the more noticeable the effect of differences in CES and ES202 employment. Some major differences are real business changes over time, either changing fortunes or seasonal patterns. Some of these changes are reflections of economic trends in the industry as a whole or in the broader geographical area. Other changes are restricted to a single establishment, and it would be misleading to extrapolate this pattern to nonsample units. NORC and the Illinois Department of Employment Security developed a set of pre-processing rules that operate on comment codes to classify changes as generalizable or nongeneralizable. A few examples of such rules are shown in Table 1.

Not all differences between the CES employment report $y_{(j)t}$ and an ES202 report $x_{(j)0}$ from a previous "benchmark" period 0 reflect true changes in employment levels within establishment j. Many of these differences are caused by mismatches between the CES and ES202 files. For example, an ES202 reporter may be reporting only production workers, causing the ES202 employment figure to be substantially lower than the CES figure. Alternatively, a CES sampling unit thought to include all of the firm's establishments in the state may, in fact, be including some unknown subset of establishments. Typically the mismatches cause the most egregious differences in employment between CES and ES202. To keep these phenomena from contaminating the models, it is important that as many of these mismatches as possible be analyzed and either corrected or removed from the sample.

There are many ways that data differences can be flagged for review. One quick and easy check is to rank the cases by a function of the CES vs. ES202 employment differences. In construction of such a function, note that very large establishments will dominate functions of absolute differences, while very small establishments may dominate functions of relative differences. For the current evaluation, we used a diagnostic statistic,

$$T_{jt} = (y_{(j)t} - x_{(j)0})^2 / \{(y_{(j)t} + x_{(j)0})/2\}$$

which can be thought of as the product of the absolute difference and the relative difference between $y_{(j)t}$ and $x_{(j)0}$, and relative large values of T_{jt} the statistic indicate something out of the ordinary that should be investigated. Routine Taylor expansion arguments show that this quantity is a first-order approximation to a standard measure of entropy. This diagnostic, and related quantities, have been used in previous work with data review by the A.C. Nielsen company; for some general background, see, e.g., Strobel (1982), Harter (1987) and references cited therein. Once the CES/ES202 discrepancies are sorted according to T_{jt}, the largest values can be investigated for data problems. In general, we recommend a careful evaluation incorporating local knowledge of the establishment involved. For the current evaluation, however, without benefit of any local knowledge, we arbitrarily treated any observations with a T_{jt} value of 50 or greater as an erroneous match. That is, we assumed that the CES data could not be clearly matched to ES202 data and ignored the sample CES values. In effect, we treated these cases as nonsample units.

3. Empirical Results

Table 2 reports empirical results for estimated levels (employment totals within a specified domain) in $I = 5$ industries (wholesale, durables, nondurables, construction and mining) and $A = 5$ relevant MSAs in Oregon covering the $T = 11$ months from January, 2000 through November, 2000.

Specifically, for a given estimator $\hat{Y}_{iat}$ and the corresponding ES-202 true value x_{iat}, Table 2 reports the empirical seventy-fifth percentile of the absolute relative error $\left|\hat{Y}_{iat} / x_{iat} - 1\right|$, denoted $Q_{.75}$. Thus, within the specified group of domains, one-quarter of the domains have an absolute relative error that is larger than $Q_{.75}$. Table 2 reports results for the aggregate of all MIDs, MSAs and months using the expression given above, and separately for individual months, MIDs and MSAs using the corresponding simplified expressions. For a given estimator, the column labeled "before" refers to an estimator based on data subject only to the steps outlined in Section 2.1; and the column labeled "after" refers to an estimator based on data subject to the steps outlined in both Sections 2.1 and 2.2. Thus, comparison of the "before" and "after" columns indicates the incremental empirical effect of the additional steps in Section 2.2. For $\hat{Y}_5$, the estimator evaluated in the the "after" column also incorporates some additional editing steps based on comment codes recorded for specific responding units.

To study the effect of the deletion step for some specific cases, Figures 1 through 3 display results for, respectively, Wholesale Trade in Portland, Oregon; Nondurables Manufacturing in Eugene-Springfield, Oregon; and Nondurables Manufacturing in Salem, Oregon. Each figure presents point estimates $\hat{Y}_5$ and nominal 95% confidence bounds $\hat{Y}_5 \pm t_{6,0.975} se(\hat{Y}_5)$ based on the CES data before the deletion step; and point estimates Y_5^* and nominal 95% confidence bounds $Y_5^* \pm t_{6,0.975} se(Y_5^*)$ based on the CES data after the deletion step, where $t_{6,0.975}$ is the 0.975 quantile of a t distribution on six degrees of freedom. The standard errors $se(\hat{Y}_5)$ and $se(Y_5^*)$ were computed through balanced repeated replication with Fay factors equal to 1.5 and 0.5. In addition, each figure displays the ES-202 nominal true employment levels.

In Figure 1, the effect of the deletion step is fairly mild, leading to small changes in the point estimates and relatively small reductions in the associated standard error estimates. Figure 2 displays two potential benefits of the deletion step. First, the original estimates $\hat{Y}_5$ decrease markedly between February and March, 2000, while the alternative estimates Y_5^* and the ES-202 values remain approximately constant. Second, for November, 2000, the standard error of $\hat{Y}_5$ is large relative to the standard error of Y_5^*, and relative to the standard error of $\hat{Y}_5$ in other months. This phenomenon was observed for $\hat{Y}_5$ in Nondurables Manufacturing for several other MSAs in Oregon, and is attributable to the effects of the outlying or mismatched units on the synthetic estimation component of $\hat{Y}_5$ in this month.

Figure 3 displays a potentially problematic effect of the automated deletion step. Both the original estimator $\hat{Y}_5$ and the ES-202 values display a strong pattern of seasonal growth and decline in May through November, 2000. However, the alternative estimator Y_5^* is relatively constant during this period. Examination of the microdata identified several establishments that displayed a parallel pattern of rapid employment growth and decline during this period, and which also had large values of the test statistic T_{jt} in these months. Consequently, the automated deletion rule employed here excluded these units from the calculation of the alternative estimator Y_5^*. Thus, Figure 3 illustrates the fact that a purely automated implementation of a simple deletion rule can lead to substantial biases in the resulting point estimators. In keeping with generally accepted survey practice, test statistics like T_{jt} are primarily of value in initial screening of data to identify potentially problematic observations that warrant further analyst review.

Table 1: Comment Codes Used by NORC and Illinois

		Action	
Code	Comment	Code Applied to ES202	Code Applied to CES
01-04, 06	General business conditions	Ignore code, keep data.	Ignore code, keep data.
05, 07-11	Labor disputes, remodeling, special projects	Do not use data for estimating models. Use to predict, but report codes.	Do not use data for estimating models. Treat as certainty unit.
12-19	Employment shifts	Ignore code, keep data.	Ignore code, keep data.
20-36	Pay shifts	Ignore code, keep data.	Ignore code, keep data.
40-49	Time and vacation (includes shift of part-time workers)	Ignore code, keep data.	Ignore code, keep data.
50	Adverse weather	Assuming the weather affects large portions of the state, ignore the code and keep the data.	Assuming the weather affects large portions of the state, ignore the code and keep the data.
51	Fire disruption	Do not use data for estimating models. Use to predict, but report codes.	Do not use data for estimating models. Treat as certainty unit.
52-53	Disaster disruption	Assuming the disaster affects large portions of the state, ignore the code and keep the data.	Assuming the disaster affects large portions of the state, ignore the code and keep the data.
54-58	Other external factors, including secondary effects.	Ignore code, keep data.	Ignore code, keep data.
59-60	Defense related codes	Ignore code, keep data.	Ignore code, keep data.
61-64	Temporary codes	Insufficient information. Ignore code, keep data.	Insufficient information. Ignore code, keep data.
65-74	State specific codes	Insufficient information. Ignore code, keep data.	Insufficient information. Ignore code, keep data.
75-79	Tax or coverage changes	Ignore code, keep data.	Ignore code, keep data.
80-82	Coding and classification changes	Ignore code, keep data.	Ignore code, keep data. Use CES codes and classifications.
83-84	Different person reporting or adjusted for summer	Ignore code, keep data.	Ignore code, keep data.
85	New establishment or work-site	Ignore code, keep data.	Drop the record from the matched files; do not use. Treat as birth.
86	Permanently out of business	Drop the record.	Drop the record from the matched files. Treat as death.
87	Reactivated account (ES202); first month reporting (CES)	Ignore code, keep data.	Ignore code, keep data.
88	Establishment dissolution	Drop the record.	Drop the record from the matched files but not the large file. Treat as death.
89, 93	Merger, predecessor/successor	Ignore code, keep data.	Drop the record from the matched files but not the large file. Treat as death and birth.
90	Reporter changed basis of reporting employment	Ignore code, keep data.	Drop the record from the matched files but not the large file.
91	Other reporting issues (related to hours and wages)	Ignore code, keep data.	Ignore code, keep data.
92	CES cancellation	Not applicable	Drop the record from the matched files but not the large file. Treat as non-sample unit.
94-99	Verification	Ignore code, keep data.	Ignore code, keep data.

Acknowledegements

The authors thank Pat Getz, Richard Clayton, Shail Butani, Kenneth Shipp, Kirk Mueller, Gregory Erkens, Kirk Wolter and Michael Macaluso for many helpful comments on the CES and ES-202 programs. The views expressed in this paper are those of the authors and do not necessarily represent the policies of the U.S. Bureau of Labor Statistics or the National Opinion Research Center.

Table 2: Empirical Seventy-Fifth Quantiles of the Absolute Relative Errors of Point Estimators Y1, Y4, Y5 and Xwls Before and After Deletion of Identified Outlying or Mismatched Units: Overall, by Month, by Major Industrial Division (MID) and by Metropolitan Statistical Area (MSA)

	Y1 before	Y1 after	Y4 before	Y4 after	Y5 before	Y5 after	X-WLS before	X-WLS after
Total	0.1539	0.1436	0.0873	0.0844	0.0914	0.0906	0.0741	0.0728

Month	Y1 before	Y1 after	Y4 before	Y4 after	Y5 before	Y5 after	X-WLS before	X-WLS after
JAN2000	0.1046	0.1013	0.0696	0.0609	0.0706	0.0697	0.0617	0.0590
FEB2000	0.1170	0.1205	0.0769	0.0746	0.0810	0.0869	0.0718	0.0652
MAR2000	0.1410	0.1324	0.0824	0.0767	0.0731	0.0745	0.0802	0.0721
APR2000	0.1599	0.1454	0.0839	0.0742	0.0811	0.0777	0.0749	0.0697
MAY2000	0.1376	0.1385	0.0827	0.0844	0.1080	0.0821	0.0664	0.0673
JUN2000	0.1463	0.1343	0.0841	0.0867	0.1046	0.0987	0.0752	0.0752
JUL2000	0.1608	0.1346	0.1122	0.0989	0.0992	0.1098	0.0753	0.0743
AUG2000	0.1662	0.1613	0.1343	0.0966	0.0969	0.1165	0.0786	0.0786
SEP2000	0.1480	0.1453	0.1225	0.0995	0.0958	0.1175	0.0833	0.0854
OCT2000	0.1699	0.1699	0.1311	0.0883	0.0707	0.0997	0.0695	0.0819
NOV2000	0.1688	0.1627	0.0831	0.0837	0.1075	0.0938	0.0618	0.0698

MID	Y1 before	Y1 after	Y4 before	Y4 after	Y5 before	Y5 after	X-WLS before	X-WLS After
10	0.1931	0.1931	0.1622	0.1622	0.2121	0.2121	0.1783	0.1728
15	0.1546	0.1336	0.0693	0.0640	0.0586	0.0771	0.0719	0.0692
20	0.2372	0.1818	0.1292	0.0744	0.0950	0.0590	0.0697	0.0684
24	0.1037	0.1381	0.0720	0.0763	0.0706	0.0712	0.0550	0.0597
50	0.1259	0.1167	0.0835	0.0927	0.0769	0.0993	0.0847	0.0819

MSA	Y1 before	Y1 after	Y4 before	Y4 after	Y5 before	Y5 after	X-WLS before	X-WLS after
1890	0.2050	0.2050	0.0814	0.0749	0.1122	0.0982	0.0772	0.0687
2400	0.1441	0.1423	0.0907	0.0794	0.1034	0.0712	0.0760	0.0755
4890	0.1417	0.1468	0.1590	0.1595	0.1060	0.1311	0.1378	0.1543
6440	0.0457	0.0558	0.0306	0.0328	0.0539	0.0553	0.0243	0.0244
7080	0.1478	0.1176	0.0800	0.0841	0.0811	0.1001	0.0675	0.0633

References

Bureau of Labor Statistics (1997). *BLS Handbook of Methods*. U.S. Department of Labor, Bureau of Labor Statistics Bulletin 2490, April, 1997. Washington, DC: U.S. Government Printing Office.

Butani, S., Harter, R. and Wolter, K. (1997). Estimation Procedures for the Bureau of Labor Statistics Current Employment Statistics Program. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 523-528.

Butani, S., Stamas, G. and Brick, M. (1997). Sample Redesign for the Current Employment Statistics Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 517-522.

Eltinge, J.L., J. Fields, R.C. Fisher, J. Gershunskaya, P. Getz, L. Huff, R. Tiller and D. Waddington (2001). Small Domain Estimation in the Current Employment Statistics Program. Paper presented to the Federal Economic Statistics Advisory Committee, June 8, 2001.

Harter, R.M. (1987). Exception Reporting: Judging What is Significant. *Nielsen Marketing Trends,* January, 1987.

Harter, R.M., K.M. Wolter and M. Macaluso (1999). Small Domain Estimation of Employment Using CES and ES202 Data. *Federal Committee on Statistical Methodology Research Conference: Complete Proceedings, Part 1 of 2*, 195-204.

Strobel, D. (1982). Determining Outliers in Multivariate Surveys by Decomposition of a Measure of Information. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 31-35.

Werking, G. (1997). Overview of the CES Redesign. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 512-516.

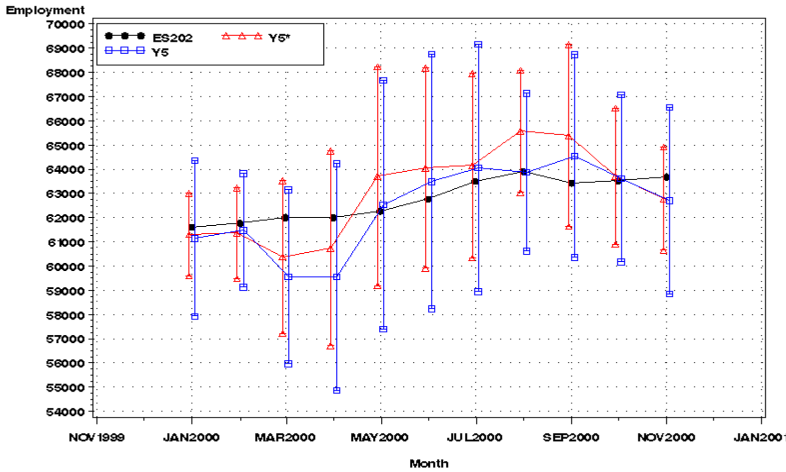

Figure 1: Point Estimates Y5 and Y5* Confidence Bounds and ES—202 True Values for Total Employment in Wholesale Trade for Portland, Oregon (Between 74 and 101 Responding UI Accounts Per Month.)

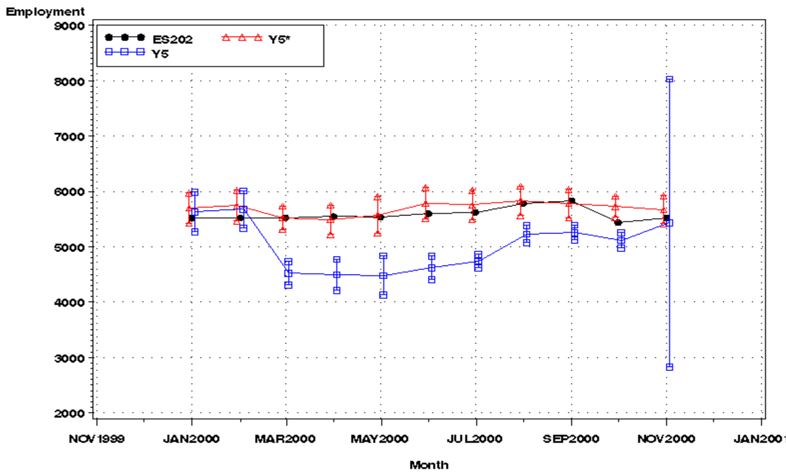

Figure 2: Point Estimates Y5 and Y5* Confidence Bounds and ES—202 True Values for Total Employment in Nondurables Manufacturing for Eugene—Springfield, Oregon (Between 11 and 18 Responding UI Accounts Per Month.)

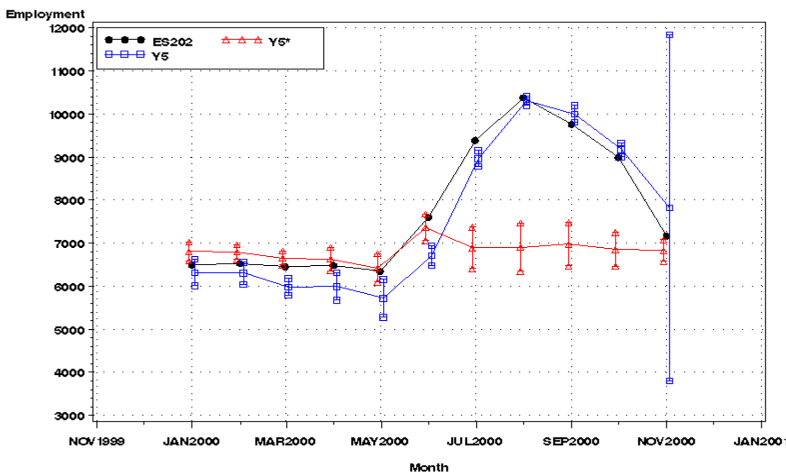

Figure 3: Point Estimates Y5 and Y5* Confidence Bounds and ES—202 True Values for Total Employment in Nondurables Manufacturing for Salem, Oregon (Between 13 and 21 Responding UI Accounts Per Month.)